NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Predicting healthcare high-cost users using data mining methods

Bernardo Neves Pantaleão
M20190042

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management, with specialization in Knowledge Management and Business Intelligence

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

PREDICTING HEALTHCARE HIGH-COST USERS USING DATA MINING METHODS

by

Bernardo Neves Pantaleão

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management, with specialization in Knowledge Management and Business Intelligence

Advisor: Roberto Henriques

October 2021

ACKNOWLEDGMENT

I would like to thank my wife Fernanda, the person that encouraged me the most to accept the challenge of applying to a Master program and that came with me to Portugal during the two years of the course, making this journey delightful.

This study was developed under the sponsorship of the Central Bank of Brazil Post Graduation Program and I would like to thank this initiative's team, specially Claudinei José dos Santos, and all managers that supported my application.

ABSTRACT

The increase in healthcare costs is, perhaps, one of the most important issues that governments and organizations face nowadays. An ageing population and technological advancements are the key reasons for this phenomenon. In this scenario, proactive measures are very important. This work aimed to improve the effectiveness of the prevention by helping the identification of the most probable high-cost users of health services in future years. Data from 2015 to 2019 of approximately 30,000 Central Bank of Brazil's Health Program's enrollees were used to train, validate and test four types of models, considering the kind of high-cost users (simple or cost-bloomers, *i.e.*, non-high-cost in previous periods) and the time-span between predictors and the dependent variable (none or one year), an innovation suggested by other authors. Different percentual cut-off points to define high-cost were used, and up to 67% of high-risk users' expenses could be correctly captured. Results confirmed the importance of previous costs data for this kind of prediction and showed that cost-bloomers and one-year time-span approaches reach good performance, creating opportunities to improve users' health outcomes while contributing to the fiscal sustainability of private and public health systems.

Keywords

Healthcare Costs; Costs Prediction; Health Management; Predictive Methods; Data Mining

INDEX

LIST OF ABBREVIATIONS AND ACRONYMS

ANN	Artificial Neural Networks
ANOVA	Analysis of Variance
ANS	Agência Nacional de Saúde – National Health Agency
AUC	Area Under Curve
BCB	Banco Central do Brasil – Central Bank of Brazil
CART	Classification and Regression Tree
DCG	Diagnoses Costs Groups
DRG	Diagnoses Related Groups
IBGE	Instituto Brasileiro de Geografia e Estatística - Brazilian Institute of Geography and Statistics
GDP	Gross Domestic Product – the sum of all products and services made or rendered in a given year
MLP	Multi-layer Perceptron
PASBC	Programa de Assistência à Saúde dos Servidores do Banco Central do Brasil – Central Bank of Brazil employees Health Program
RFE	Recursive Feature Elimination
ROC	Receiver Operating Characteristic
VIF	Variance Inflation Factor

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

In 2017, health expenditures, including medical services and medicines provided by the government and private companies, accounted for 9.2% of the Brazilian GDP. This number seems impressive, but, in countries with an older population, it may reach much more: 11.3% in Germany and France, 12.4% in Swiss and 17.3% in the US (IBGE, 2019). In these countries, the numbers are even more impressive, as, despite their older population, their GDPs per capita are also much higher than in Brazil. This scenario shows the importance of this matter for governments and organizations all over the world.

Although these numbers are already very high, they are increasing at a fast pace. New technologies are the main reason for it: every day, new products and services are created in this sector that is so important, and that has a low elasticity (Smith & Freeland, 2009), as people will try to pay everything they can for the sake of their health and their beloved ones. Nonetheless, these advanced technologies have a price and, usually, it is not low. Besides its price, the new technologies have another effect: they impact the population ageing, as more modern they are, they tend to be more effective and to increase the life expectancy of an already ageing population, which is, by itself, another important reason for the growth of health costs. The older the people, the bigger the probability of demanding healthcare services and, with these new technologies, the age limits are being pushed further and further (Blumenthal et al., 2016; Caley & Sidhu, 2010; Peixoto et al., 2004).

Few people tend to be responsible for most of these costs (Cohen, 2001). Healthcare costs are very concentrated, not only among the elders but mainly in people with chronic health issues, like diabetes, cancer and diseases of the circulatory system (Kim & Park, 2019), all very correlated with the age of the health system users. This scenario demands proactive creative actions, which, nowadays, are basically being targeted at prevention and customized care management (Dove, Duncan & Robb, 2003). If we can identify some of these future high-cost users of healthcare services, preventive measures can be taken and customized care and follow up may be proposed to mitigate a very high growth of expenditures (Tamang et al., 2017; Blumenthal et al., 2016).

Many studies have already advanced on this task of future high-cost users' identification. They usually use a mix of cost and utilization variables along with check-ups, diagnostic and clinical data. Bertsimas et al. (2008) were one of the first to use more modern methods of predictive data mining, including a holdout strategy for time series, training the model with data from one year and testing with unseen data from another period. This study created a baseline model and, by the addition of more variables, tried to improve its accuracy on the classification of health services' users. Kim and Park (2019) used more recent data of the South Korean national health system, including clinical and diagnostic information. Tamang et al. (2017) also tried to predict high-cost users, using data from more than 1,5 million Danish from 2004 to 2011. All these studies have used different sources of variables. However, all of them concluded that the previous year's cost variables are the best features to classify future years' high-cost users of healthcare services correctly.

These studies used modern algorithms, like Random Forests, Decisions Trees and Artificial Neural Networks, besides more classical algorithms, like Logistic Regression. All of them took advantage of how medical claims are presented to governments and health insurers, with every item being charged separately in a classical payment model called fee-for-service. It is also important that these claims are presented electronically, so their processing and analysis are made much easier.

1.1. Study Objectives

This study aims to replicate some of the models presented so far using data from the more than 30 thousand insurees of PASBC (Programa de Assistência à Saúde dos Servidores do Banco Central do Brasil), the Central Bank of Brazil's employees' Health Program. This insurance provides healthcare for employees and close relatives and, for the same reasons presented in the beginning of this introduction, has been seeing increases on its expenditures. So, the idea of this dissertation is to answer the following research question:

How can previous utilization, costs and clinical data, besides demographics, be used to identify future high-cost users of healthcare services?

In the end, the main idea is to evaluate several models, using different data and algorithms, and identify the ones that make the best predictions, after reviewing the main literature produced about the related topics. In order to achieve this objective, some intermediary goals need to be attained:

1. A Review of the Literature regarding healthcare utilization prediction.
2. Discussing with PASBC employees and designing queries to extract the data needed for the analysis to be made.
3. Data analysis in order to better understand the problem and the population.
4. Pre-processing data and engineering new features.
5. Models' development and comparison through different metrics.
6. Identification of the best model regarding pre-selected performance metrics.

It's believed that, by attaining all these goals, the main objective of this project will be achieved and it will be possible to test and evaluate different models, identifying the best ones among them to classify and predict high users.

1.2. Study Relevance and Importance

It's been already showed the importance of studies related to healthcare costs and utilization. Numbers show that expenditures have been steadily increasing worldwide and the trend doesn't show to be changing (IBGE, 2019).

In this context of almost unstoppable growth in healthcare costs, if governments and organizations manage to correctly predict which ones are the users who will present biggest expenditures in the future, they can take preventive measures and customize care management. This way, not simply letting the user by himself in the health system, looking for different physicians and treatments by his own, may demonstrate to be an important action. Besides, by identifying possible future cases of chronic diseases, organizations may adopt preventive measures, much less expensive than the reactive ones.

This way, this study may have a practical and theoretical application, as the identification of high-users, so important for the ones involved in healthcare practice, will be discussed and implemented theoretically, through the use of different variables and algorithms.

According to an extensive research conducted, it will be just the second time that a study like this will be executed with data of Brazilian users of health services (Galdino, 2019), the first one with data from private insured users. As it has already been presented, healthcare accounted for almost 10% of the Brazilian economy in 2017, what represents a huge opportunity for academics, governments and health companies.

Nonetheless, the importance of this study should not be restricted to the cost reduction, although this fact by itself should be celebrated by the entire population, as, due to resources limitations, the more is saved, the more will be available for other users. An early identification of potential high-cost users of health services also means that these people may have better and customized care in early stages of their health issues. This way, this study doesn't only contribute for the cost management in healthcare, but may also improve the quality of life of users and the patients' outcomes, depending on the actions taken after their identification.

1.3. METHODOLOGY SUMMARY

This project will embrace many steps and different kinds of knowledge. First, a broad literature review will be conducted. Then, data will be obtained from corporate systems of the Central Bank of Brazil, specifically the Benner system, an Online Transactional Processing (OLTP) system used in the PASBC to manage members enrolment, procedures authorizations and payments processing, among other functions. Its databases are stored in SQL, what allows direct queries to get the necessary data.

Payments processing and PASBC's chronic patients' program management is done in this system, so all data regarding clinical information, services' utilization and costs will be obtained from its database. It's interesting to understand that health providers send XML files in formats specified by ANS, Brazilian National Health Agency, with its claims, including services provided, their costs, date and other characteristics of the service, like if it was surgical or clinical, for instance. This data is vital for this study and may be considered, according to previous studies presented in the introduction, the core features of this classification. The first step, then, will require SQL skills for query building and database understanding.

After obtaining the data, it will be pre-processed. This step usually requires considerable effort, encompassing data cleaning, outliers' and missing values treatment. At this moment, new features based on the ones directly obtained from the database will be created, like the sum of expenses in one, two or more years and the growth percentage in recent years, for example.

Data analysis to better understand the observations and its distributions will be necessary, including here the construction of charts, like boxplots and histograms for different variables.

Having finished the entire pre-processing, the models' building *per se* will be executed. Creating predictive models require splitting data according to best practices of the holdout method. It is necessary to randomly define which observations will be used to train the model and which ones will be used to validate and fine-tune it, with the rest, not used so far, being used to test the model. It is important to test the model in unseen data to ensure that no information from these observations affected the construction of the algorithm.

In the case of this project, an out-of-time sampling strategy will be used, training the model with data from year 0, year 1 being used to validate the model predictions and year 2 being good for testing.

During the validation, fine-tuning of hyperparameters will take place in order to improve the accuracy of each one of the models created.

Different algorithms will be used with different sets of features to create multiple comparable models using Python's Sklearn package (Pedregosa et al., 2011). Decision trees, logistic regressions, random forests and neural networks are some of the possibilities, besides ensemble methods that use different algorithms simultaneously, trying to increase the precision through the simultaneous use of multiple predictions.

The last step will be the evaluation of the results with the use of different metrics to assess the models' predictions quality in the test set (data from the available last years). A specific metric for this purpose was presented by Tamang et al. (2016). The cost capture tries to balance the evaluation by increasing the weight of correct positive predictions, once it is better to identify one high-cost user than many low-cost users correctly.

The last chapters will discuss the results and present this dissertation's conclusion, final considerations and suggestions for future studies.

2. LITERATURE REVIEW

2.1. COST PREDICTION IN HEALTHCARE

Many authors have studied and developed different methods to predict healthcare costs, which shows the importance of this subject. Despite not being a new field of study, it is consistently evolving with new predictors and more modern and powerful algorithms.

Morid et al. (2018), while systematically reviewing the literature about this topic, identified three types of cost prediction approaches in healthcare: rule-based methods, developed based on the knowledge of experts; multiple regression models, which they defined as statistical models; and supervised learning methods. All these methods are divided between actual costs' predictions and classes' predictions (high-cost users or users' multiple buckets) based mainly on medical or costs data, besides other features.

In sections 2.1.1 and 2.1.2, respectively, previous studies about actual costs estimations and high-cost users classification studies are presented. While reviewing them, it was possible to enhance Morid et al.'s systematization, listing other characteristics that allow us to differentiate them, as presented in Table 1.

2.1.1. Risk Adjustment Models: Estimating Actual Healthcare Costs

Many studies have approached the problem of identifying high-cost users in healthcare. However, many of them were not classification problems, but rather estimations of the actual costs in the following year, which can also be used to predict the costliest users by simply ranking the estimations, like done by Ash et al. (2001) and Meenan et al. (2003). These estimations are usually called Risk Adjustment Models and are used to predict actual costs related to each user in the following period (Yang et al., 2018). They have this name because they are a tool to establish the fixed value that will be paid by the government or a health plan to a hospital or a medical group in a capitation system (Ash et al., 2000), a payment model in which a provider is responsible for predefined medical services of a group of users during a determined period. In this scenario, the payment is fixed, adjusted according to the risk of the group of users the provider will be responsible for.

On their broad literature review about Risk Adjustment Models, Cucciare and O'Donahue (2006) state that the first of these tools relied solely on sociodemographic variables, like age and gender. With the introduction of more powerful predictors, like clinical and costs features, R^2 increased from less than 5% to more than 15%, tremendously incrementing the predictive power of these models.

Traditionally, these types of models used multiple linear regressions, as in Ash et al. (2000, 2001), Powers et al. (2005) and Meenan et al. (2006), and a huge set of variables, mainly sociodemographics, like age, gender, income group and location, and clinical data, extracted from the previous periods claims, although other kinds of variables could also be used, like pharmacy data (drugs' costs and disease groups) used by Powers et al. (2005).

Ash et al. (2000), for example, used 118 hierarchized health condition categories features, created by grouping diagnoses codes extracted from medical claims to predict costs in the following period, a Risk Adjustment Model called Diagnoses Cost Groups (DCG). They split same years (Y0 predictors/Y1

costs) data for model training and validation, testing the model in a separate sample. The best R^2 for the regression models developed was 21%, but it's important to note that they did not use cost-based predictors, nor made any transformation in the dependent variable. Although this result may not seem impressive, it showed a huge increase in predictive power when compared to a R^2 lower than 2% for a baseline model that used only demographic features.

2.1.1.1. Estimation-based Users' Classification

As stated earlier, these traditional regression estimators were also used to predict high-cost users by ranking the estimations. Meenan et al. (2003) trained five different Risk Adjustment Models using diagnoses and costs data to estimate expenses, considered the predicted top 0.5% and 1% values as high-cost users, tested them on 7% of the sample and compared their predictive power. Authors reached best AUC of 0.86 and 0.85 for top 0.5% and 1% high-cost users, while correctly capturing 24% and 26% of the costs for a sensitivity of 18% and 21%, respectively. According to them, traditional performance metrics for this kind of classification are not the best option, since they value equally every correct prediction, while cost capture weights every prediction by their actual monetary value, so, if a model accurately predicts the highest cost users among the high-cost users, it may have the same sensitivity than others, while capturing much more of the cost.

Ash et al. (2001) also made estimation-based classifications on their study comparing a DCG Risk Adjustment Model with a costs-based model to predict the top 0.5% high-cost user of the following year. They used multiple linear regression to estimate costs based on DCG predictors, selected the top 0.5% predictions and compared with the top 0.5% in year 0. They argue that, with the improvements in diagnoses-based models, predictions of these kinds are better or, at least, as good as cost-based predictions, because the DCG model could "capture" more of the actual top 0.5% total cost than the latter (although the comparison was made with the simplest possible cost-based model). According to their study, due to randomness, the persistence of high users is not so considerable. Nevertheless, both models were capable of capturing at least 7.5% of the total cost in year 1, despite selecting only 0.5% of the observations.

Table 1: Different Characteristics of Healthcare Costs Predictions Studies

Characteristic	**Options**
Prediction Type	Costs Estimations Estimation-based Classifications • High-cost users • Cost buckets Users Classification • High-cost Users • Cost Bloomers
Variables	Sociodemographic Clinical Data • Diagnoses and Diseases • Procedures • Number of Conditions Self-reported Condition Check-up Data Services Utilization Previous Costs
Algorithms	Multiple Linear Regression Logistic Regression

	Clustering Classifier Decision Trees Artificial Neural Networks Random Forests Support Vector Machines Gradient Boosting AdaBoost
Data Splitting	Non-split Train/Test same years Train/Test different years Train/Validate/Test different years
Performance Measures	R^2 Accuracy (hit ratio) Sensitivity (Recall) Specificity Precision (Positive Predictive Value) Area Under ROC Curve (AUROC) Penalty Errors (sum of weighted confusion matrix) Cost Capture (prediction costs/actual costs) Average Absolute Prediction Error
Methodological steps	Base Model New Models • Combination of Variables • Different Algorithms • Different Thresholds
Sampling	Imbalanced Sample Combination of Over and Under Sampling

More recently, modern estimation methods began to be used. Maybe the best example is Bertsimas et al. (2008), who conducted one of the broadest studies about healthcare costs prediction so far. Using costs, sociodemographic and medical data (like diagnoses, procedures and drug groups), they developed clustering classifier estimators and decision trees models with over 1,500 features for a time span of 3 years (2 for predictors and 1 for result). The clustering classifier was developed in two steps, first determining costs predictions for each cluster and then classifying each observation in one of the clusters. Interestingly, they used 22 costs variables, including "trend" (the slope in monthly expenditures) and "acute" (number of months above average), enhancing the predictive power of the costs' variables, which they found to be the best predictors.

Splitting data in training, validation and test sets (all for the same 3 years), they predicted the actual costs and classified observations among 5 possible cost buckets, comparing the models results with a baseline one (costs from the previous year). Results showed a good improvement in performance measures compared to the baseline model, with better results for the clustering classifier, with a R^2 of 0.18 and an accuracy of 42% for the top 0.5% high-cost users prediction. Complex medical data were useful to increase performance for the highest cost bucket slightly.

Like Bertsimas et al. (2008), Morid et al. (2018) also created five buckets of same total value and used multiple algorithms and many costs features to classify users in one of them based on actual expenditures estimation. They used data from the same years to train and test the models, splitting the sample 30/70 and using a 20-fold cross-validation to test the results. Gradient boosting showed the best overall performance, with a R^2 of 0.46 and a 92.9% sensitivity, although ANN had the best metrics for the highest cost bucket (around 2% of the test dataset), 0.45 and 49.6%, respectively.

Yang et al. (2018) also used machine learning methods to predict expenditures in the following period, creating a sort of Risk Adjustment Model. They worked with a created continuous dependent variable, which is the rank of the observation regarding costs in the following period divided by the population size (so, in a population of 100, the 14th highest cost user would have the value of 0.14). They used clinical data (diagnoses and procedures group codes), medication codes and demographic features from 2011 to 2014 of their Texas Medicaid sample. Using linear and regularized regressions, gradient boosting and recurrent neural networks, they achieved a R^2 greater than 55%.

2.1.2. High-Cost Users and Multiclass Classification

Nonetheless, these modern predictive data mining models were not used only for costs estimation. Lavange et al. (1986) seem to have been the first to develop models to classify high-cost users in healthcare using data mining methods. More interested in studying the logistic regression algorithm, they used health status data, like chronic conditions, to classify top high-cost users in the same year. No data splitting strategy nor performance metric were calculated, but reading this seminal study was very important to see for how long authors have been developing healthcare high-cost users classification models using data mining techniques.

Since that study, many authors have developed classification models to predict high-cost users in healthcare, using an extensive and diverse set of predictors. Better data mining practices, like data splitting, cross-validations and even under and oversampling techniques, began to be used, enhancing the models' predictive power.

Chechulin et al. (2014) ran a logistic regression with utilization and clinical (diagnoses, especially for chronic conditions) data from 2007 to 2009 (2006 to 2008 in validation dataset) to predict top 5% high-cost users in 2010 (2009 in validation dataset), reaching an AUC of 0.865 and a sensitivity of 42%. Past utilization features were among the strongest predictors. Interestingly, they did not use a threshold for classification, considering all the top 5% highest probabilities predicted high-cost users.

Other authors tried different features. Fleishman and Cohen (2010), for instance, used self-reported health condition data, besides diagnoses cost groups and number of chronic conditions based on a medical survey to predict top 10% high-cost users in the following period, testing different logistic regression models while increasing the number of variables and comparing the results with a baseline model only with sociodemographic features. Best models captured 63% of the cost with a sensitivity of 78% and a precision of 29%, approximately, and an AUROC around 0.86. Their results showed that, although not as good as DCG features, chronic conditions counts are also valuable predictors.

Kim and Park (2019) added check-up data (laboratory test, self-reported medical history and health behaviour) to the traditionally used medical (diagnosis groups), costs and utilization data to predict top 10% high-cost users. Training multiple models with three different algorithms (logistic regression, random forests and artificial neural networks), they captured up to 66% of the cost from the actual top 10% highest users in the test dataset, reaching an AUC of 0.843. Cost and utilization features were considered the best predictors, once again showing the importance of these variables for high-cost users classification.

It can be seen that authors used a great variety of algorithms, predictors, evaluators and time-series classification approaches. However, only Moturu et al. (2010) tackled an important aspect of these

classifications: unbalancing. They developed a mixed approach between clinical and utilization data to predict high-cost users (>$50.000 and >$25.000) in the following period. They rebalanced the sample using a combination of under and oversampling techniques since this kind of classification is a very imbalanced one. They used the number of inpatient, outpatient and emergency procedures and visits related to each of 20 disease groups, created from the thousands of possible codes of diseases presented in the providers' claims. Besides the total number, they also created dummy variables for each one of these 20 groups and 136 drug categories prescribed to the users.

Training multiple algorithms with data from 2002/2003, they evaluated the cost capture and other performance metrics in a 2003/2004 test set. Comparing their results with a base model that assumed the same high-cost users in both years, authors could improve F from 0.43 to 0.79 when rebalancing the training dataset with 60% of high-cost users, reaching the best AUC of 86%. This approach seems promising for increasing sensitivity, although precision is extremely penalized since we end with much more predicted high-cost users than expected, which is a problem for targeted health programs.

Another important aspect of this kind of prediction is the time span between dependent and independent variables. Meenan et al. (2003) and Dove et al. (2003) suggest new studies with wider time gaps between risk assessment and future expenses, which would be very important, so healthcare providers and payers could take more efficient preventive actions.

Despite the importance of widening the time gap, only one study using an approach close to the one suggested was found. Rosella et al. (2018) developed a logistic regression model based on self-reported health and risk behavioural data to predict users who were going to become top 5% high-cost users ("cost bloomers") in one of the following 5 years. In the validation cohort, using data for a five-year period different than the one used for training the model, they reached an AUC of 0.82 and a R^2 of 8%.

2.1.2.1. Cost Bloomers Identification

According to Dove et al. (2003), high-cost users prediction models usually overlook the regression to mean phenomenon, when high-cost users expenditures in year 0 tend to decrease in the following period. For that reason, they developed a regression model to predict the probability of low-cost users (costs < $2000 in 1998) become high-cost users in the following year.

Using a compound dependent variable based on actual costs in year 1 and their concentration, as, for the authors, the higher the concentration, the higher the risk, Dove et al (2003) developed a regression model with medical, behavioural and utilization data, like the number of visits and chronic conditions, the existence of diseases and users' compliance pattern. In the test dataset (1999/2000), they reached a precision of 39.7%.

More recently, Tamang et al. (2016) developed multiple models with up to 1,059 features (costs, utilization, diagnoses and procedures group codes) to predict top 10% high-cost users and "cost bloomers" (users that became top 10% high-cost users in the following period). Training logistic regressions in 2008/2009, validating them in 2009/2010 and testing in 2010/2011 data, best models captured 60% of costs from top 10% whole population and 49% from cost bloomers, while reaching AUC of 83.6% and 78.6%, respectively.

Something interesting regarding cost bloomers studies is that consistent high-cost users usually don't have manageable diseases, so organizations shouldn't expect preventive care programs to bring noticeable outcomes (Dove et al., 2003). For that reason, cost bloomers tend to be better candidates for case management programs, increasing the importance of these models.

2.1.3. Literature Summary

Table 2 shows how characteristics listed in Table 1 were used by each one of the studies analysed during this systematic literature review. Every study used sociodemographic variables, and only Moturu et al. (2010) developed a model with a combination of under and oversampling methods. Basically, all reviewed studies used traditional evaluation metrics, like precision, recall and the area under the ROC curve, although not always calling them by these names (recall, for instance, was usually called sensitivity).

Table 2: Literature Review Summary

Study	Prediction Type	Variables	Algorithms	Data Splitting	Performance Metrics
Ash et Al (2000)	Cost Estimation	Diagnoses Groups Conditions Groups	Linear Regression	Train years 0/1 Validation years 0/1 Test years 0/1	R^2 = 21.1% Average Prediction/Actual Costs for users with 1 or more chronic condition = 92%
Ash et Al (2001)	Estimation-based top 0.5% high-cost user classification	Diagnoses Groups	Linear Regression	Non-Split	High-cost user average cost/Population average cost = 16.5 Total cost captured = 7.8%
Bertsimas et al (2008)	Cost Estimation User's cost bucket (5) classification	Diagnoses Groups Procedures Groups Drugs Groups Utilization Costs	Clustering Classifier Decision Tree	Train years 0,1/2 Validation years 0,1/2 Test years 0,1/2	R^2 = 18% Overall Accuracy = 84% Top 0.5% Accuracy = 43% (80% and 19% on baseline model, respectively) Average Penalty Error for top 0.5% decreased 36% from baseline model Mean absolute prediction error decreased 58% from baseline model
Tamang et al (2016)	Top 10% High-cost users Classification Top 10% Cost Bloomers	Diagnoses Groups Utilization Costs	Logistic Regression	Train years 0/1 Validate years 1/2 Test years 2/3	Top 10%: • Cost capture = 60% • AUC = 0.84 • Precision = 33% Cost bloomers: • Cost capture = 49% • AUC = 0.79
Kim and Park (2019)	Top 10% High-cost users Classification	Diagnoses Groups Self-Reported Health Status Check-up Utilization Costs	Logistic Regression Artificial Neural Networks Random Forests	Train years 0/1 Test years 1/2	Cost Capture = 66% AUROC = 0.84
Meenan et al (2003)	Cost Estimation Estimation-based top 0.5% and 1% High-cost user Classification	Diagnoses Groups Chronic Conditions	Linear Regression	Train years 0/1 Test years 0/1	Top 0.5%: • AUROC = 0.86 • Sensitivity = 18% • Correctly captured cost = 24% Top 1%: • 0.85 • Sensitivity = 21% • Correctly captured cost = 26%
Powers et al (2005)	Cost Estimation	Drugs Groups	Linear Regression	Train years 0/1	R^2 = 11%

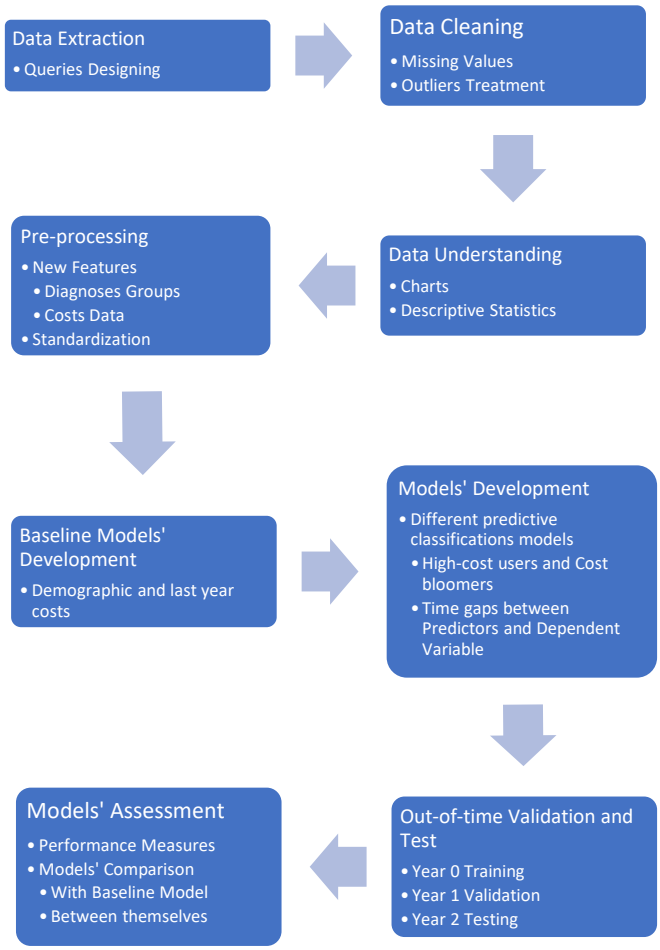
	Estimation-based top 1% High-cost user Classification	Drugs Costs	Log Transformed Linear Regression	Test years 0/1	Mean absolute prediction error decreased 13% from baseline model Precision = 14%
Rosella et al (2018)	Top 5% Cost Bloomers (in the following 5 years)	Self-Reported Health Status Self-Reported Health Behaviour	Logistic Regression	Train period 0/1-5 Test period 2/3-7	AUROC = 0.82 Pseudo R^2 = 8%
Chechulin et al (2014)	Top 5% High-cost users Classification	Diagnoses Groups Chronic Conditions Utilization	Logistic Regression	Train years 1/2 Test years 0/1	Sensitivity = 42% Accuracy = 94% Precision = 43% AUROC = 0.87
Morid et al, (2018)	Cost Estimation User's cost bucket (5) classification	Costs	Linear Regression Artificial Neural Networks Random Forests SVM Gradient Boosting Bagging Decision Trees	Train years 0/1 Test years 0/1 (20-fold cross validation on 70% of data)	R^2 = 46% Mean Absolute Percentage Error = 0.65 Overall Accuracy = 93% Top 2% Sensitivity = 50% Top 2% Average Penalty Error = 0.96
Dove et al (2003)	Probability Estimation-based Cost Bloomers	Diagnoses Groups Costs	Linear Regression	Train years 0/1 Test years 1/2	AUROC = 0.73 Precision = 39.7%
Moturu et al (2010)	Top 0.69% (>50k) High-cost user Classification	Disease-related utilization Drug groups utilization	Logistic Regression SVM AdaBoost	Train years 0/1 Test years 1/2	Balanced 10/90 Sample: • Correctly predicted cost = 30% • Sensitivity = 28% Balanced 60/40 Sample: 1 AUROC = 0.86
Lavange et al (1986)	Non prediction High-cost Classification	Self-Reported Costs No. of Conditions	Logistic Regression	Non-split	Pseudo R^2 = 0.25
Fleishman and Cohen (2010)	Top 10% High-cost users Classification	Self-Reported Health Status Diagnoses Groups Chronic Conditions	Logistic Regression	Train period 0/1 Test period 2/3	Precision = 29% Sensitivity = 78% AUROC = 0.86

3. METHODOLOGY

In this study, different healthcare high-cost users predictive classification models will be developed. The decision of working with predictive classification methods and not predictive estimation methods, i.e., to predict if an observation belongs to the class of top users and not the actual value of each observation, was taken because the primary intent of this study is to identify high-cost users for future programs of primary preventive care, as explained in the Introduction. For that reason, the prediction of actual expenditures would help only to rank observations and identify the top users, as some authors have done, which seems counterproductive in this project as classification algorithms allows this identification automatically.

Based on the literature reviewed in the previous section, it was decided that cost bloomers models would also be developed, and not only high-cost users classifications, which was the first idea for this thesis. The latter are more commonly found in the literature. However, the former seem more interesting for the purpose of this study, as identifying future high-cost users that do not belong to this group yet allows healthcare payers and providers to take preventive actions before critical events happen or health conditions become chronic.

Figure 1: Methodological Steps



In this section, the different models will be presented, considering the type of predictive classification (simple high-cost or cost bloomer), the time span between the dependent and independent variables

and the sample selection and balancing. Besides, the different algorithms and variables, including the created features, will be explained. Furthermore, the processes of data understanding, pre-processing and model selection and assessment will be presented.

3.1. Methodological Steps

1. In order to develop the models to be tested in this study, the first step is to understand and extract the data from the Benner OLTP system's database using SQL queries;

2. Next step will be looking for inaccurate and missing values. As null values mean that the user did not have any healthcare expenditure, ICD code in one of their claims or utilization data in the period, they can simply be replaced by 0, but all features with missing values will be analysed to check the best way to have them filled;

3. Treating outliers is the following step, although, as the objective of the models is to predict the highest-cost users, the outliers are exactly the instances that will be looked for, so it wouldn't make sense to have them removed;

4. Then it will be necessary to develop new features, transforming ICD codes in grouped Diagnoses Codes, as explained in 3.5.3, and summing claims' values to create 2 and 6-months and 1 and 2-years costs features;

5. Next, some data visualizations will be developed, analysing the relationship of different variables and better understanding the sample;

6. The following methodological step will be will be the creation of dummy variables and the standardization of some features;

7. Creating the different datasets for each one of the four types of models listed on section 3.5 will also be necessary;

8. Finally, multiple models will be run with different methods and classifiers using Python's Sklearn package and the results will be evaluated and compared between themselves and with the metrics of the baseline model.

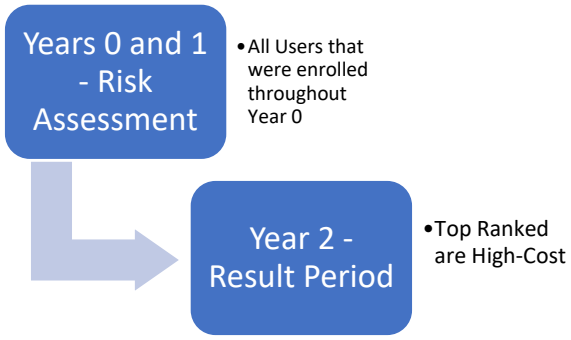
3.2. Different Models to be Developed

Four types of high-cost users classification models will be tested during this study regarding the classification purpose and the period between the risk assessment and the resulting period. Figures 2 to 5 organize the four possibilities of models according to these factors.

Two types of classification will be tested in multiple models. The simple top high-cost users classification aims to identify a pre-determined percentage of users according to their predicted probability of being a member of this class in the resulting period, independently if these users were already ranked inside this top percentile in the observation period. The second approach is the classification of cost bloomers, which also intends to identify users who will be among a top percentage in the future, except if they were already a high user in the second year of the observation period. So,

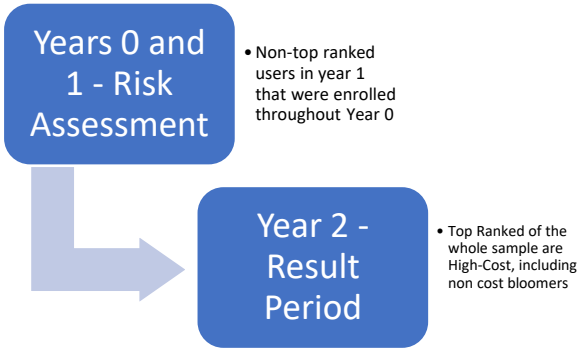
the cost bloomers approach, shown in Figure 6, drops these top ranked observations from the sample, but keeps considering them in the result period.

Figure 2: Years 0 and 1/Year 2 Simple High-Cost Classification



It is expected that the traditional approach will provide better performance measures than the cost bloomers, as observed in Tamang et al. (2016). This may happen because there is a considerable probability that some high-cost users will be the same in consecutive periods, so, by not being able to predict them (once they were dropped), the precision decreases. Nonetheless, the cost bloomers approach is interesting for healthcare users, payers and providers as it's a kind of early identification of high users, which could provide a good opportunity for primary preventive care.

Figure 3: Years 0 and 1/Year 2 Cost Bloomers Classification

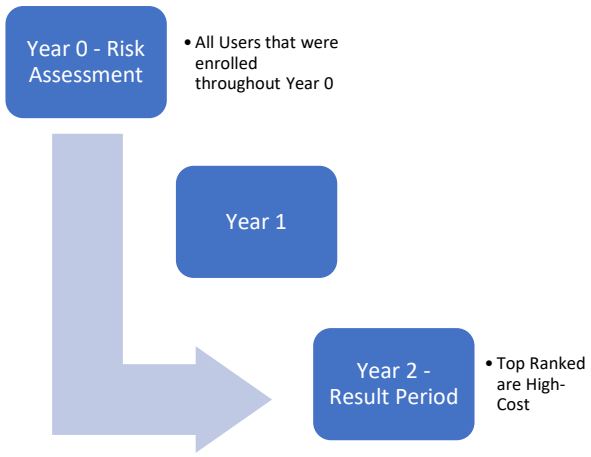


Besides the type of predictive classification, different time spans between the observation and result periods will also be considered when building the models. Basically, all literature reviewed considered consecutive years between the risk assessment predictors and the dependent variables, i.e., features in Year 1 predicted the probability of being a high-cost user in Year 2. Meenan et al. (2003) and Dove et al. (2003) suggested widening this time gap, but the only study analysed that tried a similar approach was Rosella et al. (2018). However, they did not try to predict top-ranked instances in a specific non-consecutive year but in anyone of the following five years, reducing the error probability.

In this study, two two-time gaps between risk assessment and result will be tested: the traditional consecutive periods (Y0 and Y1/Y2) and a wider one (Y0/Y2). It's expected that the performance metrics for the latter approach will be worse, as the uncertainty will increase with this extra "blind" period between the assessment and the result. Nonetheless, this approach is even more relevant than the cost bloomers one, as it is an actual early identification method, while the simple cost bloomers

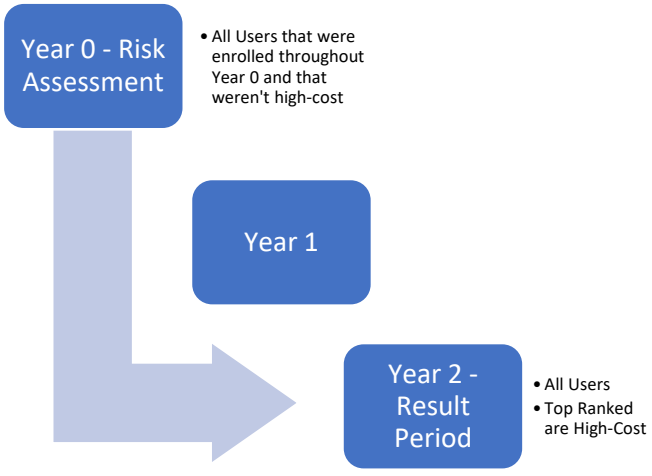
approach, despite classifying users that were not high-cost users, only identify them in the year they may become one, leaving little time for preventive primary care.

Figure 4: Year 0/Year 2 Simple High-cost Classification



The combination of an entire year interval between risk assessment and result with the cost bloomers approach will generate the most interesting models in this study. It will try to identify non-high-cost users in year 0 that will become top-ranked in year 2. Although it seems important for everyone involved in the healthcare sector, this type of model has two extra layers of uncertainty (the wider time-span and the exclusion of high-cost users from the risk assessment period) so, unfortunately, it might have the worst performance measures of all, although the approach of selecting only the historic percentage of cost bloomers, explained above and shown in Figure 6, may reduce the predictive error.

Figure 5: Year 0/Year 2 Cost Bloomers Classification



3.3. HIGH-COST USERS THRESHOLDS

In Predictive Data Mining, thresholds are usually the minimum calculated probabilities for considering an instance a member of a specific class. For instance, if an algorithm calculates a probability of 64%

for an observation belonging to a class and the defined threshold in this example is 50%, then this instance will be considered as a member of this class.

In this study, as the objective is identifying instances that will be part of a defined percentage of high-cost users in the following year, there will be two kinds of thresholds in the models: the traditional "probability thresholds" and the "classification thresholds", i.e., the percentage of users that will be considered members of the high-cost group. Different models will be developed, with classification thresholds varying from top 0.5% to top 10% highest cost users, as shown in Table 3. Meenan et al. (2003) and other authors used this approach. The objective is to evaluate the different performance metrics for these models with distinct top-ranked users' percentages, which could help the comparison between them for the selection of the best model.

3.3.1. The Probability Thresholds

Two probability thresholds' approaches will be used in the models to classify instances as high-cost users. The basic one will simply rank the algorithm calculated probabilities and consider the top x observations, being x the classification threshold explained before. This will be interesting because the objective of managers and health providers may be selecting a pre-determined number of users to participate in a primary care initiative, so it does not matter the probability but the number of users selected.

In cost bloomers' models, the number of instances that are high-cost in both periods will be excluded during the classification, that means, if 50% of top 5% high-cost users are cost bloomers, then the top 2.5% users with the highest probabilities, according to the cost bloomers models, will be considered high-cost users in the resulting year, as presented in Figure 6.

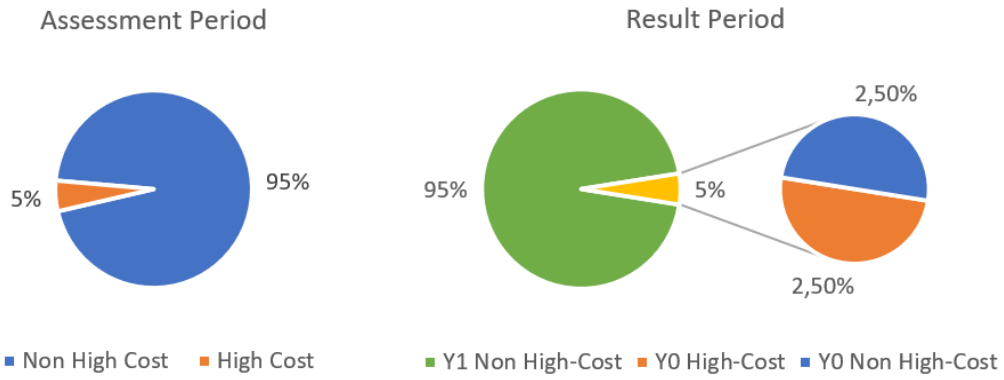
Table 3: Different Classification Thresholds

Different Classification Thresholds to be Tested
0.5%
1.0%
2.0%
5.0%
10.0%

Although it may appear that it would not make sense to consider the traditional probability thresholds in this study, as, if the models are classifying instances as members of a group with a predetermined size, selecting more or fewer users would be an *a priori* error, this is not true and depends on the model objectives. If the goal is to maximize the precision, for example, it makes sense to classify only part of the top x ranked as high-cost users, despite knowing beforehand that some high-cost users would be considered low-cost. On the other hand, if the goal is to maximize the model recall, it makes sense to select more than the top x ranked[1].

1 It's expected that these examples with performance measures will become more clear in section 3.9.

Figure 6: Cost Bloomers - Top 5% High-Cost Model Example[2]



For that reason, besides the simple top x% ranked instances approach, a probability threshold will also be used. A calculated theoretical threshold was developed, that is, considering hypothetical per capita costs matrix of a primary care initiative (Table 4), a calculated threshold to maximize the benefits of the classification model will be used when oversampling techniques are used, as explained in Sheng and Ling (2006) and Elkan (2001).

Table 4: Example of Primary Care Initiative Cost Matrix

Actual/Predicted	High-Cost	Low-Cost
High-Cost	0	42000
Low-Cost	7000	0

Considering this matrix and *Cost (h,l)* as the cost of predicting as high-cost someone that is actually low-cost, the expression for the calculated theoretical threshold would be the following:

$$\text{Threshold} = \frac{\text{Cost}\,(\text{h},\text{l}) - \text{Cost}\,(\text{l},\text{l})}{\text{Cost}\,(\text{h},\text{l}) - \text{Cost}\,(\text{l},\text{l}) + \text{Cost}\,(\text{l},\text{h}) - \text{Cost}\,(\text{h},\text{h})}$$

This would mean a threshold in the example of $\frac{7000}{49000} \sim 14\%$.

3.4. SAMPLE

This study will use PASBC insurees' data from 2015 to 2019. No actual identification of any kind will be made: only a random number automatically created by PASBC's OLTP during the enrolment of any insuree will be used to merge tables and extract grouped data. All data will be collected from the Benner OLTP system, used by PASBC in all its functions, from the enrolment of new insurees to the healthcare providers' payment processing. Different cohorts will be made up according to the model type, time gap and learning schema phase (training, validation and test).

All cohorts include only insurees with, at least, the last 12 consecutive months of the risk assessment period, that is, for the 1 Year Time Span Cost Bloomers Model Training Dataset, only users that were enrolled in PASBC during the whole 2016 year that were not high-cost in this period will be part of the

2 Cost Bloomers can't be among the high-cost users in the risk assessment period, but they will be only a proportion of high-cost users in the result period. In the example, classifying all top 5% most probable cost bloomers would increase the chance of error, as, historically, a percentage of high-cost users persist in consecutive periods. For that reason, instead of selecting top 5%, only top 2.5% ranked cost bloomers would be considered high-cost in an approach without pre-defined probability threshold.

sample. Tables 5 to 8 will provide a better understanding of the cohorts composition, and section 3.4 will fully present the sample's features.

Table 5: 1 Year Time Span Simple High-Cost Users Classification

	Training	Validation	Test
Number of Instances	30719	30641	30449
Risk Assessment Period	2015-2016	2016-2017	2017-2018
Result Period	2017	2018	2019
Rules	All Users enrolled during whole 2016	All Users enrolled during whole 2017	All Users enrolled during whole 2018

Table 6: 2 Years Time Span Simple High-Cost Users Classification

	Training	Validation	Test
Number of Instances	30235	30719	30641
Risk Assessment Period	2015	2016	2017
Result Period	2017	2018	2019
Rules	All Users enrolled during whole 2015	All Users enrolled during whole 2016	All Users enrolled during whole 2017

Table 7: 1 Year Time Span Cost Bloomers Classification

	Training	Validation	Test
Number of Instances[3]	27647 up to 30565	27577 up to 30488	27404 up to 30297
Risk Assessment Period	2015-2016	2016-2017	2017-2018
Result Period	2017	2018	2019
Rules	Users enrolled during whole 2016 and not high-cost in 2016	Users enrolled during whole 2017 and not high-cost in 2017	Users enrolled during whole 2018 and not high-cost in 2018

Table 8: 2 Years Time Span Cost Bloomers Classification

	Training	Validation	Test
Number of Instances[4]	27212 up to 30084	27647 up to 30565	27577 up to 30488
Risk Assessment Period	2015	2016	2017
Result Period	2017	2018	2019
Rules	All Users enrolled during whole 2015 and not high-cost in 2015	All Users enrolled during whole 2016 and not high-cost in 2016	All Users enrolled during whole 2017 and not high-cost in 2017

3.5. Predictors and Dependent Variable

This study will use four kinds of features (sociodemographic, clinical, cost and utilization data) to predict a binary dependent variable. Basically, all predictors have been used before in previous studies.

[3] Number of instances changes according to high-cost "classification" threshold explained in chapter 3.3, as previous high-cost users are excluded from the cohorts.

[4] Number of instances changes according to high-cost "classification" threshold explained in chapter 3.3

3.5.1. Dependent Variable

The predicted variable will be binary representing if the instance is a member of the top high-cost user class. Considering the different classification thresholds presented in 3.3, instances will be ranked by their costs in the resulting year and the top x%, being X the classification threshold) will get a value of 1 (high-cost user), against a value of 0 for the rest of the instances.

3.5.2. Sociodemographic

Gender and age at the end of the risk assessment period will be the two sociodemographic variables used in this study.

3.5.3. Clinical Data

Chronic ill PASBC insurees may enroll at their will in “Programa Vem Ser” (a pun with the verb “vencer”, “to win” in Portuguese, and the expression “vem ser”, that means “come be”), a program that monitors their treatment and quality of life. When they do so, their chronic condition(s) is(are) registered on the Benner OLTP system, used by PASBC. With this data, a dummy variable will be created for each one of the conditions. This approach to work with clinical data was used in many studies, like Bertsimas et al. (2008), Tamang et al. (2016) and Kim and Park (2019), among others listed in Table 2, although these authors grouped different diseases, usually extracted from healthcare providers claims, in Diagnoses Groups, while the chronic conditions in “Programa Vem Ser” are already grouped in 13 possible types listed in Table 9.

Another feature that will be used is the number of conditions in Vem Ser by the end of the risk assessment period. A predictor of this kind was already used by Fleishman and Cohen (2010). Besides these chronic conditions’ variables, the date of registration of each one of the conditions on Benner OLTP will also be used to create the feature “number of years with the condition X”.

Although the “Programa Vem Ser” data might already provide valuable predictor power to the models, this study will also use clinical data extracted from healthcare providers claims, the classical approach in high-cost users prediction studies. Claims are healthcare invoices sent by providers to the ones responsible for the payment and are one of the most important sources of data in healthcare, as they bring much information about the patient, the kind of care, procedures and treatments. Among this data, there is usually the International Classification of Diseases (ICD) code, which is the user's illness at the moment he looks for care.

There are more than 70.000 diagnosis codes in the current version of ICD (ICD-10), but in all analysed PASBC’s claims, there were only 4.780 different disease codes. This small number, when compared to the total available codes, may be due to several reasons: many claims do not bring an ICD code, providers may fill their claims in a simplified way, and there is a natural concentration in more common health problems (many specific conditions are rare and may never occurred for any of the PASBC’s enrollees). For that reason, it was even considered not to use this data, as the “Programa Vem Ser” could supply the clinical data. Nevertheless, it was preferred to work with these codes, as they could increase the model’s predictive power.

Table 9: Programa Vem Ser list of conditions

Systemic Arterial Hypertension (SAH)
Lipid Storage Diseases
Diabetes
Chronic Kidney Disease
Obesity
Cancer
Chronic Cardiopathies
Chronic Pneumopathies
Neurological Diseases
Post-stroke sequelae
Hepatitis C
HIV
Transplant Pacients

Even with such a small percentage of ICD-10 codes used, it would not be reasonable to use all these 4.780 codes as features. As explained in section 2.1, it is best practice to group these codes in Diagnoses Groups, as done by many authors listed in Table 2, like Chechulin et al. (2014), for instance. Usually, these groups are specified in healthcare claims or are defined according to tested DRG methodologies, like the Medicare Severity Diagnoses Related Groups (MS-DRG). However, using these techniques require specific software and more details regarding medical care, which were not available in the extracted data used in this study.

For that reason, it was necessary to group the used ICD codes in diagnoses groups manually. The 781 most relevant ICD Codes (regarding chronicity) were selected and grouped in 69 features, according to the methodology presented in the Appendix. Only major chronic conditions were selected to control the number of features and focus on the most important diagnoses. This way, it was possible to work with the clinical data available in the providers' claims.

In the same way that was done with the chronic conditions of the "Programa Vem Ser", a binary feature will be engineered for each one of these diagnoses groups, and a numeric variable "years since diagnosis" will be created for each one of the groups. Besides, the sum of groups will also be used as a predictor.

3.5.4. Costs Data

Other kinds of features that will also be used are cost-based predictors, i.e., variables based on users' previous periods' costs. This approach may seem "less scientific" from a clinical standpoint, as these features are based solely on previous costs without any kind of explanation (Ash et al., 2001). Nonetheless, costs predictors were found to be the most powerful ones by many authors (Morid et al., 2018). Besides, they are a pretty efficient surrogate for clinical information when not very dense and complex medical data is available (Bertsimas et al., 2008). Furthermore, they are much simpler to work with than diseases and diagnoses codes, that demand too much technical knowledge for data pre-processing.

Table 10 presents the main cost-based features that will be used. As the ICD codes explained in section 3.5.3, costs are also extracted from claims presented by healthcare providers to PASBC for payment. After being presented, they are audited and the amount understood as fair and correct, according to the user's condition, executed procedures, used medication and material and contracted values, is paid. It's this final amount that will be used to engineer new features.

Table 10: Cost-Based Features

Year 1 total Expenditure
Year 0 total Expenditure*
Last 2 months total Expenditure
Last 6 months total Expenditure
Last Month total Expenditure
12-Months Total Hospitals Expenditure (Inpatient and Outpatient)
12-Months Total Inpatient Expenditure (Hospitals and other providers)
Acuteness
Expenditures Trend

* 24-Months total Expenditure will be only used in 1 Year Time Span Models, when risk assessment period = 2 years

Two of these variables are not self-explanatory and need to be clearly explained.

Acuteness and Expenditures Tendency were developed by Bertsimas et al. (2008) and used by Morid et al. (2018). While the former tries to identify if the total annual expenditures were concentrated in one or more specific months, what could differ chronic patients from severe acute ones (that could have suffered an accident, for instance), the latter tries to measure if the user is on a trend of growing expenditures. Acuteness is measured by the number of months during the last year whose expenditures are greater than the monthly average, while the Expenditures Trend is the slope of the curve defined by linear regression of the monthly expenses by the number of the month (1 to 12) in the last year of the assessing period.

3.5.5. Utilization Data

Four[5] healthcare utilization variables will be used in this study:

Table 11: Healthcare Utilization Variables

Previous Years* Total Medical Visits
Previous Years* Total Emergency Rooms Visits
Previous years* number of Non-Intensive Care Unit inpatient days
Previous years* number of Intensive Care Unit inpatient days

* 1 Year Time Span Models will use data from last observation period's year only

Besides them, two other features will be engineered for the models with two-years assessing periods: the growth in the ICU and Non-ICU inpatient days from year 0 to 1 of the observation.

[5] In datasets with two-years assessing period, this number increases to 8.

3.6. Models' Learning Schema – Out-of-time Sampling

As presented in Tables 5 to 8, models will be trained, validated and tested on different two or three years datasets, respectively for 2 years and 1 year time spans. This approach was only used by Tamang et al. (2016) among all studies reviewed. As there's independence between data from different years and the whole sample has users' data from 2015 to 2019, it's possible to train the model with data from 2015/2016 (risk assessment/result), validate on data from 2016/2017 and test on data from 2017/2018, for example.

This approach is known as out-of-time sampling, as the natural factor that promotes the independence between the different datasets is time, here expressed in years. It is important to explain again that the categorical dependent variable is calculated from each sample's last year's costs, as this study tries to predict the high-cost users in a future year based on previous data, this way, each training, validation and test dataset will have predictors from one or a pair of years (risk assessment) and the predicted variable from a following period (result). The number of risk assessment years will depend on the model's time span, since the available data allow the split in three datasets for 2 years time-span models only using 1 year data for risk assessment.

3.7. Sample Rebalancing

Top-ranked predictive classifications like the ones proposed in this study are naturally imbalanced. If it is being predicted the top 0.5% high-cost users, the proportion of negative to positive instances is 199 to 1, which could make it very hard for models to identify positive observations, as, during training, they will learn that these instances are highly uncommon. This will not be a problem when using the first probability threshold strategy presented in section 3.3.1. After all, the model will be forced to select the desired top-ranked users' percentage, but may be a problem with the second approach, a pre-defined threshold, because less than the top percentage defined as high-cost users may be classified as so. It could increase the precision at the expense of decreasing the recall, as already stated in the same section.

An approach to solve this bias is rebalancing the sample, as done by Moturu et al. (2010), through over or undersampling. In this study, there will be developed some models with the Synthetic Minority Oversampling Technique-SMOTE, an approach that, based on the minority class instances, create similar pseudo-observations, artificially increasing their total number trying to increase the model's predictive power of this class. These model's results will be compared with others to check if this strategy can improve the metrics in healthcare high-cost users predictions.

3.8. Classifiers

Besides the different types of models, probability thresholds used and sample rebalancing approaches, diverse classifiers, listed on Table 12, will also be applied for comparison. Python's Sklearn package will be used during this study.

Table 12: Classifiers to be used

Logistic Regression
Decision Tree
Artificial Neural Networks (ANN)
Random Forest
Engineered Ensemble Method

The most basic one will be the Logistic Regression, that was first used in healthcare high-cost users classification in 1986 (Lavange et al., 1986). This algorithm calculates the probability of an instance belonging to a class by fitting a linear logistic model (*idem*) using the predictors and the categorical dependent variable. One of the interesting characteristics of the logistic regression is that this algorithm calculates coefficients for each one of the predictors, helping to understand the effect of each one of them in the classification process (although, as it's a logistic model, the interpretation is not that easy).

The second algorithm to be used is the decision tree, that, through the use of a selected algorithm (CART or C4.5), creates predictors-based rules to split instances (Breiman et al., 1984). The idea is basically the same of someone taking decisions on a road, for example, choosing to turn right or left, based on the features' values. According to the decisions made, the rules, the model classify the observation in one of the classes. The hierarchical representation of the rules resembles a tree, the reason for the name of the classifier, and the interpretation is easy (Hastie et al., 2009), depending on simple yes or no answers to the questions presented by each rule.

Artificial Neural Network algorithm will also be applied in this study using the Multilayer Perceptron classifier from Python's Sklearn package. This classifier tries to replicate our brain's neurons functioning, developing pseudo-synapses between the instances' data and artificial nodes, organized in a number of layers. Each one of these nodes receives these values multiplied by weights (whose values are defined during the training phase) and calculates, based on a specific function, new values (a kind of intermediary hidden features), that will be the input for the next layer of nodes, until the output layer is reached, with the model's calculated probability for each observation presented to the classifier (Larose, 2015).

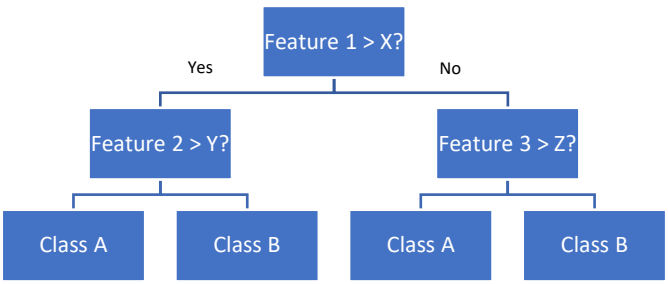


Figure 7: Representation of a Decision Tree

Another classifier to be used is the Random Forest, an ensemble of Decision Trees. This classifier randomly creates decision trees models with subsets of the samples and of the features (so the results of each one of them are different) and then use the most predicted class by the forest for each one of the instances, a kind of a voting system.

The last classifier to be used will be an engineered ensemble method similar to Stacking (which uses a classifier on top of predicted probabilities by other classifiers). The difference is that, on the engineered method, some first level models (classifiers that generate the probabilities that will be used by the top one) will use a SMOTE oversampled dataset, already explained in section 3.7.

3.9. Performance Measures

In order to evaluate and compare the different models, classical predictive data mining metrics will be calculated using Python's Sklearn package: precision, recall, area under the ROC curve and area under the Precision-Recall curve. Besides, the cost capture, a performance metric already used by other studies (see Table 2), will also be calculated.

These metrics will be compared not just between the different models but also with the performance measures of a baseline model created with only Age and last year total costs data. This is a best practice used by other authors trying to evaluate the increase in the predictive power of complex models against a pretty simple one.

Except for the last one, all performance metrics are measured on top of the numbers from the confusion matrix, which presents the count of true positives and negatives and false positives and negatives, according to the predicted and actual class of the observations. It's needless to say that this matrix depends directly on the adopted threshold explained in chapter 3.3.1.

Table 13: Confusion Matrix

Actual/Predicted	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

The precision tries to measure how good are the model's positive predictions, which means the ratio of true positives by all predicted positives. Two precisions will be calculated, a probability ranking-based, that will consider positive the top x% instances, according to the calculated probability of the model, and a regular precision, that will use the calculated theoretical threshold explained in 3.3.1.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

The recall or sensibility tries to measure the model's power to find all positives, the ratio of actual positives correctly predicted. Only one recall will be calculated, using the theoretical threshold of .14, because a ranking-based recall would have the same value of the ranking-based precision, as the total number of positive instances is predetermined by the classification threshold (the high-cost users cut-off point), so the number of False Positives would be the same of False Negatives, equalizing the equations.

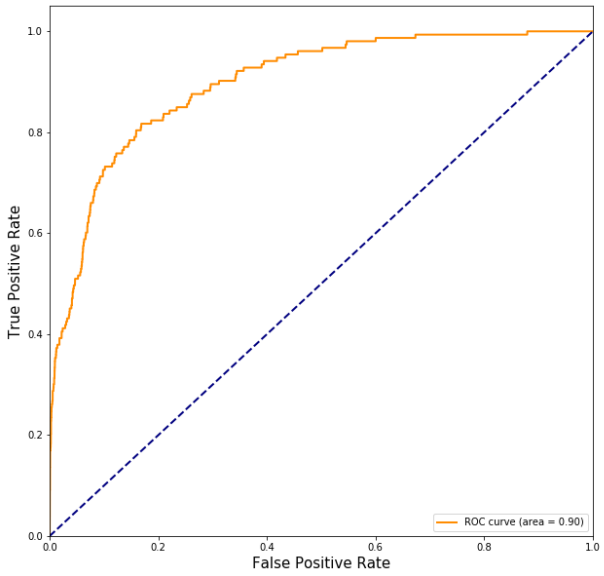
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

The precision and recall depend on the model's threshold. If we decide that any observation whose calculated probability is greater than 1% is "Positive", the model will probably have an excellent recall, because basically all actual positives will be correctly classified. On the other hand, it will not be very

precise because the number of false positives will probably be very large as well. The opposite is also true, and if a 99% threshold is chosen, the precision will probably be very high at the expense of the recall.

The ROC curve is another way to compare different models independently of the chosen probability threshold. It is the graphic expression of the relationship between the true positive and false positive rates, calculated by dividing true positives by the total number of actual positives and false positives by the actual number of negatives, respectively. This relationship is also a trade-off, so the ROC curve is plotted by changing the threshold values for the model, with each dot representing those rates for a different threshold, and then the area under this curve is calculated. Figure 8 shows an example of a ROC Curve. This metric was used in many healthcare high-cost users studies, which will allow an interesting evaluation of the models developed in this work, remembering that for the two simple high-cost users prediction models, the thresholds is not particularly important, as the ranked probabilities will define which instances will be classified as positive or negative.

Figure 8: ROC Curve Example



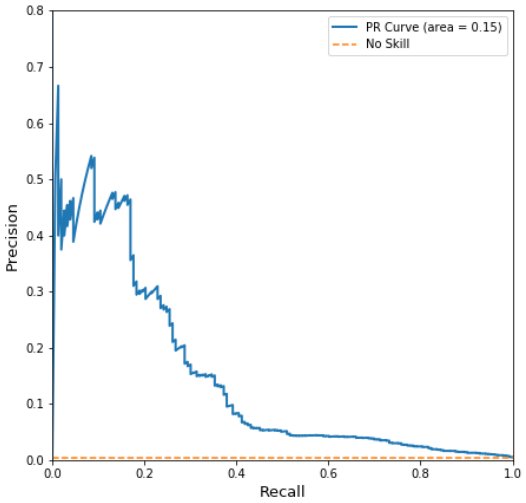
Despite having been heavily used in previous studies, the ROC Curve, according to some authors (He and Ma, 2013; Branco et al., 2015; Fernández et al., 2018), is not the best measure for imbalanced datasets as is the case when predicting top high-cost users in a sample. It occurs for two main reasons: the ROC Curve considers both the positive (true positive rate) and negative (false negative rate) classes while evaluating the model. Nonetheless, in an imbalanced sample, the positive class is usually the most important and much more difficult to predict. In this way, ROC Curves may be misleading, presenting excessively optimistic results thanks mainly to the very low false negative rate.

The second reason is that, exactly for its unbalancing and small percentage of positives, a reduced number of correct or wrong predictions can considerably change the shape and the area under the curve (Fernández et al., 2018), making it difficult to compare the models.

In order to solve this uncertainty related to the use of ROC Curves, He and Ma (2013) suggest the use of Precision-Recall Curves, a curve plotted with the values of precision and recall for each one of the possible thresholds, from 0 to 1. Like ROC Curves, the areas under Precision-Recall curves are also

calculated so the different models can be compared. As this curve is based solely on the predictions of the positive class, it is considered a better way to assess imbalanced classification models. For this reason, this metric will also be used in this study. Figure 9 is an example of a Precision-Recall Curve.

Figure 9: Precision Recall Curve Example



The last performance measure that will be used to evaluate and compare all models is the cost capture, which compares the cost of predicted positives during result period with the cost of actual positives. This metric is very interesting because it considers the real costs of users and not just the correct classification. According to Tamang et al (2016), in predictions like this, it may be better to correctly predict one extreme high-cost user than many not so critical high-cost users and that's why the cost capture is considered a so important metric in this study.

$$\text{Cost Capture} = \frac{\sum \text{Cost Predicted Positive}}{\sum \text{Cost Actual Positive}}$$

Differently from the area under the ROC curve, this metric will only be used to evaluate models that classified as positive the top x% ranked users, where "x%" is the high-cost user classification threshold, as it would only make sense to compare same number of instances' costs sum.

4. DATA UNDERSTANDING, EXTRACTION AND CLEANING

After the literature was reviewed and the methodology was designed, understanding PASBC's data and its structure was the next logical step in this study. PASBC uses an OLTP system called Benner, where all data is input and processed, from enrollees' identification and registration to providers' claims.

Claims are not the only but are the primary source of data in this project. Every healthcare service provided is charged in one or more claims presented by hospitals, labs and clinics to PASBC. In each of them, every material, medicine, exam and procedure are listed, with its quantity and price. Besides this data, claims may also bring the International Classification of Diseases-ICD code of the condition that caused the enrollee to seek healthcare.

Both costs and utilization data used in this study were extracted from these claims after being grouped. So, if someone sought healthcare 15 times during a period of time and PASBC received 22 claims related to the services provided, both utilization and costs data were aggrouped and summed to reach the total cost of the enrollee and the number of specific services, like emergency room visits and inpatient days.

The other type of data used was clinical data, and it was necessary to use two sources in the Benner system to get it. The first one was just explained, as some of the claims bring ICD codes (unfortunately, it is not mandatory to send the claims with this relevant information, which is not also carefully analyzed by PASBC or corrected by PASBC employees). To work this, all claims of each one of the enrollees were processed, and the ICD code and the service date were extracted. Codes were grouped according to the appendix 12.1, and the date was used to calculate the number of years since the enrollee's condition was first reported on a claim.

This way, if providers sent claims in 2014 and 2015 for an enrollee where the ICD code for Alzheimer's disease was reported, only the first date was considered to calculate for how long the person has this condition. The difference was calculated with the last year on the risk assessment period added of 1 (so when the ICD code is reported precisely during this last year, the feature did not have a value of 0). So, in this case, if data from 2015 and 2016 was being used to predict high-cost users in 2017, the Alzheimer's disease feature for this enrollee in this dataset would have a value of 3 (2016-2014+1).

The other source of clinical data was the program of the chronic disease (Vem Ser), which follows up the treatment of enrollees' chronic conditions, as explained in chapter 3.5.3. PASBC's insurees may decide if they want to participate in this program, so they have to enroll and specify which chronic conditions they want the program to help them manage and follow up. In the same way that was done with the ICD codes, the number of years since the enrollment was calculated for each one of the conditions of all the program's participants.

The only kind of missing values found in the datasets is the risk assessment period's first year's costs and utilization (for datasets with a two-years prediction period) for those instances who enrolled during this first year. As explained in section 3.4, only insurees who had been enrolled during the whole last year of the risk assessment period were selected in this study. Nonetheless, some models use data from two years to make predictions and cost and utilization data from the first year's months prior to enrollment is not known. In this study, whenever it happened, values equal to 0 were inputted, as if no healthcare was necessary, which is the most common situation.

Regarding inaccuracy, only some utilization features seemed to have inaccurate values (in a pretty small quantity). Inpatient days in non-intensive and intensive care unities could never be greater than 365 during a year, but this result was found for a few instances. The decision was to truncate these values in the upper limit of 365.

4.1. OUTLIERS

On the matter of outliers, the decision was also simple. As the objective of this thesis is to predict the highest-cost users, which will very probably be outliers themselves in the evaluation period (at least for the top .5% and top 1% high-cost classification), it's expected that outliers in the assessing period

will be of great importance. Many authors found big correlations between healthcare costs in different years (Morid et al., 2018; Cohen, 2001; Tamang et al., 2017), so removing cost outliers could mean removing some of the rare positive values of the dependent variable, the more important instances in this study, and could decrease the importance of these features, which are expected to be the most important ones.

An alternative used by other authors (Bertsimas et al., 2008) was to truncate the healthcare costs at some point. In this thesis, some datasets with clipped cost values will be created and tested, so it is possible to evaluate if this approach is capable of improving the results.

The decision on how to clip the cost values was taken after analyzing the outliers with three different methods: extreme percentiles, Tukey's fence (using the interquartile range) and 3 standard deviations. Table 14 shows these results for variable "Year 1 Total Costs" in the training dataset for 1 Year Time Span Simple High-Cost model.

It is possible to see that Tukey's fences method classifies too many instances as outliers, which shows that data is more distributed, so it wouldn't be reasonable to use this method. Both extreme percentiles and three standard deviations methods reach close values as upper limits, so it was decided to work with one of them. As standard deviations had a slightly higher limit, it was used to truncate cost features to create datasets with outliers' treatment.

Table 14: Outlier's detection methods

	Percentile Outliers	Tukey's Fence Outliers	Standard Deviation Outliers
Valid Interval (R$)	[0.0 – 125,792]	[-7,555 – 15,048]	[-119,310 – 137,934]
No. of Lower Outliers	0	0	0
No. Of Upper Outliers	308	3060	278
Total Outliers	308	3060	278

5. DATA EXPLORATION, VISUALIZATION AND SELECTION

5.1. DEMOGRAPHICS

Two demographic features were used: gender and age in the last year of the assessing period. Regarding gender, when exploring the training dataset for the 1-year span simple high-cost prediction (chapter 3.4), it was possible to see that most of the enrollees were female (Figure 10). This proportion slightly increased when considered just the top 10% high-cost users in the evaluation period of 2017 (Figure 11), what could characterize a small relationship between gender and high-cost risk (cost in the evaluation period). Nonetheless, for the top 0.5% (Figure 12), the distribution was exactly 50/50, with males slightly more represented than on the entire population.

When analyzing year-2 cost values (Table 15), mean and median for males and females were not very different, with male average value (R$ 11.700) a bit higher than the female one (R$ 11.367), and the opposite for the median value (R$ 2.533 vs R$ 3.728). This information could indicate that gender would not be a very relevant feature for the models that would be built.

Gender Distribution

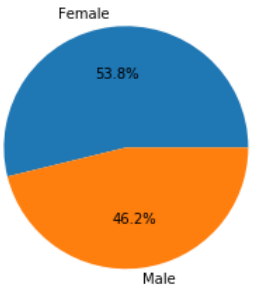


Figure 10: Training Dataset

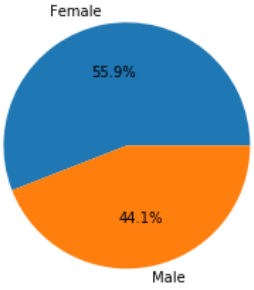


Figure 11: Top 10% High-Cost

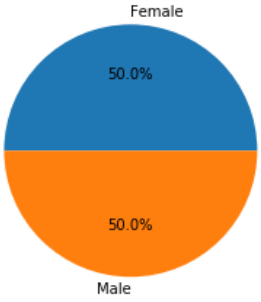


Figure 12: Top 0.5% High-Cost

On the other hand, age appeared to be an important feature. As shown in Figure 13, age distribution indicates clearly that older enrollees represent a much higher percentage of the top 0.5% high-cost users in year 2 (2017) when compared to the low-risk instances. The same happens with the top 5% high-cost users (Figure 14). Table 16: Age by Risk Group shows median and average ages according to the risk group classification of each one of the instances. Once again, it was possible to see that age is positively correlated to the costs in the evaluation period.

Table 15: Year 2 Mean and Median Costs by Gender

	Mean	Median
Female	R$ 11.367,02	R$ 3.728,75
Male	R$ 11.700,59	R$ 2.533,25
Total	R$ 11.520,98	R$ 3.162,09

To check the relationship between age and gender, in order to evaluate if the small correlation between gender and cost in the result year could be explained by the age distribution, a population pyramid (Figure 15) was built and did not present a relevant difference between the age structure by gender.

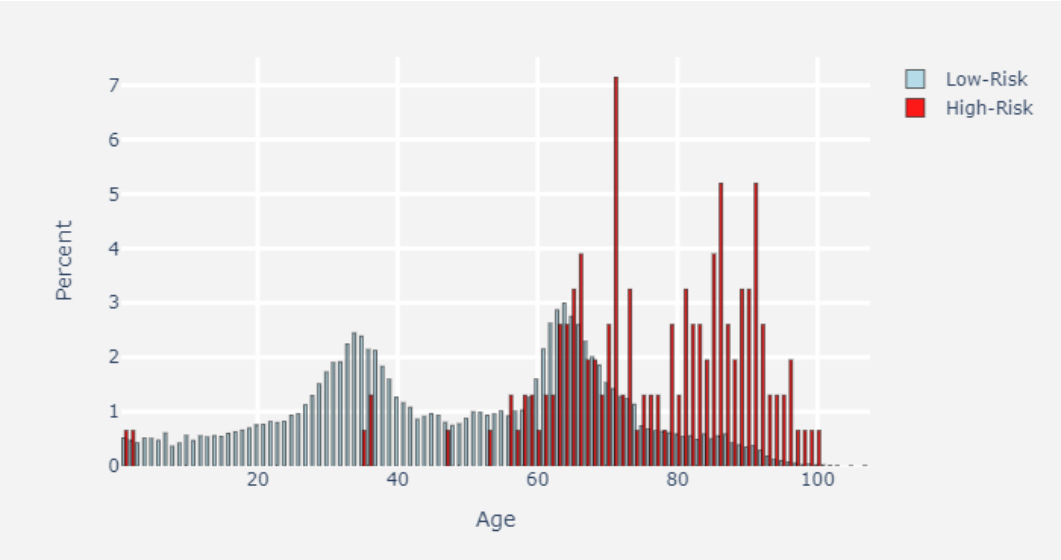


Figure 13: Age Histogram - Top 0.5% low and high risk

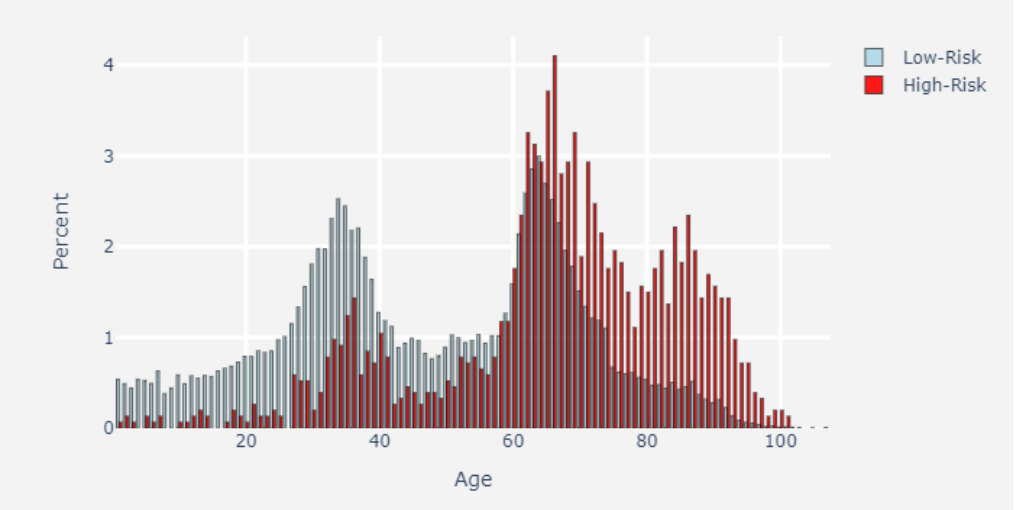


Figure 14: Age Histogram - Top 5% low and high risk

5.2. Cost Features

While exploring cost features, already explained in section 3.5.4, in the 1-year time span model training dataset, it was observed that these variables had a reasonable correlation with the evaluation period costs and, therefore, with the dependent variable derived from them. It was possible to observe on the heatmap that costs in Y0 and Y1 from the assessment period had a considerable Spearman correlation with costs in the evaluation period (.5 and .6, respectively). Nonetheless, a correlation between them was also medium-strong, suggesting possible collinearity between them. The cost trend (slope in monthly expenditures in year 1) and the acuteness, differently than expected from the mean and median high-cost classes analysis, presented low-medium and weak correlations with costs in year 2, respectively.

It's important to remember that many cost features are subsets of the two main cost variables (year 0 and year 1 costs) what explains the collinearity between some predictors.

Table 17 presents different averages and medians for six cost features by high-cost class and clearly shows that previous years' costs can help classify enrollees by its risk of being a high-cost user in the following year.

Table 16: Age by Risk Group

	Top 0.5%		*Top 1%*		*Top 2%*		*Top 5%*		*Top 10%*	
Risk	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Mean	48.46	76.03	48.3	73.4	48.14	70.81	47.66	66.48	49.92	63.66
Median	50	79	49	74	49	71	48	68	46	66

As an example, top 0.5% high-cost users in the evaluation period had, on average, healthcare expenses of R$ 118,500 and R$ 208,720 in years 0 and 1 of the assessment period, respectively, while low-cost enrollees' assessment years 0 and 1 costs were R$ 6,600 and R$ 8,310 on average. The same relationship was observed for all cost features, except for the acuteness, that which did not present a clear difference as the one observed for the other cost variables.

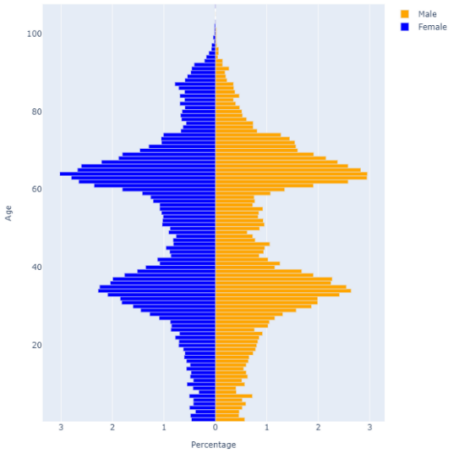


Figure 15: 2016 Population Pyramid

A heatmap (Figure 16) for all cost variables was also elaborated to analyse the Spearman correlation between themselves. The choice for this method was due to the nature of the problem, as the dependent variable is a class created from the result period costs and based on the rank of these values. This way, a method that makes use of the ranked values to calculate the correlation between variables, as Spearman's, seemed more interesting to analyse the relationship among the features and between each of the cost features and the year 2 expenses.

It was possible to observe on the heatmap that costs in Y0 and Y1 from the assessment period had a considerable Spearman correlation with costs in the evaluation period (.5 and .6, respectively). Nonetheless, a correlation between them was also medium-strong, suggesting possible collinearity between them. The cost trend (slope in monthly expenditures in year 1) and the acuteness, differently than expected from the mean and median high-cost classes analysis, presented low-medium and weak correlations with costs in year 2, respectively.

It's important to remember that many cost features are subsets of the two main cost variables (year 0 and year 1 costs) what explains the collinearity between some predictors.

Table 17: Mean and Median Cost Features by High-Cost class

(R$ 000)	*Threshold*	*Top 0.5%*		*Top 1%*		*Top 2%*		*Top 5%*		*Top 10%*	
Cost Feature	Metric	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Cost Y0	Mean	6.60	118.50	6.34	88.89	5.97	65.78	5.49	38.90	4.94	27.13
	Median	2.34	18.66	2.33	11.56	2.30	9.46	2.24	7.08	2.12	6.68
Cost Y1	Mean	8.31	208.72	7.76	162.85	7.14	115.89	6.43	64.07	5.85	40.44
	Median	2.75	77.73	2.73	37.19	2.71	25.81	2.61	13.07	2.45	9.97

Cost 6M	Mean	4.40	128.42	4.04	102.35	3.69	70.20	3.28	38.00	2.97	23.41
	Median	1.14	41.23	1.13	23.59	1.12	11.77	1.08	5.92	1.01	4.48
Trend	Mean	0.01	1.41	0.01	1.16	0.00	0.68	0.00	0.32	0.00	0.17
	Median	0.00	0.08	0.00	0.06	0.00	0.03	0.00	0.01	0.00	0.01
Cost Hospitals Y1	Mean	4.32	142.07	4.02	103.94	3.61	73.67	3.15	40.45	2.83	24.65
	Median	0.34	8.25	0.34	6.21	0.33	4.37	0.31	2.52	0.28	2.11
Cost Inpatient Y1	Mean	3.49	129.71	3.18	97.43	2.79	69.27	2.36	37.70	2.10	22.37
	Median	0.00	9.77	0.00	3.65	0.00	0.70	0.00	0.00	0.00	0.00

To complete the costs data exploration, two boxplots were generated, comparing the distributions of costs in both years of the assessment period by each one of the high-cost classes in the evaluation period. As the features were dispersed in a very big range, a logarithmic scale was used.

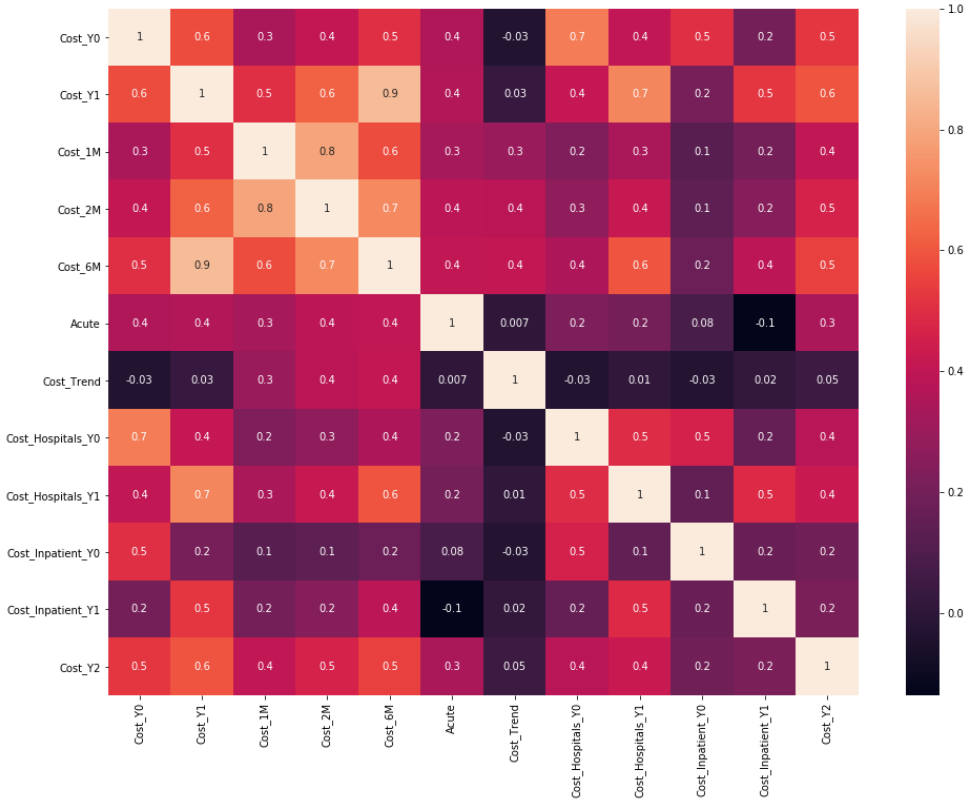


Figure 16: Costs Feature Heatmap

As one can see in Figure 17 and Figure 18, assessing period's years 0 and 1 costs have a positive relationship with the high-cost classification in the following year, the evaluation period. Nonetheless, the boxplots also show that many instances are considered outliers by Tukey's Fences method, showing that these variables alone are not capable of perfectly classifying the observations.

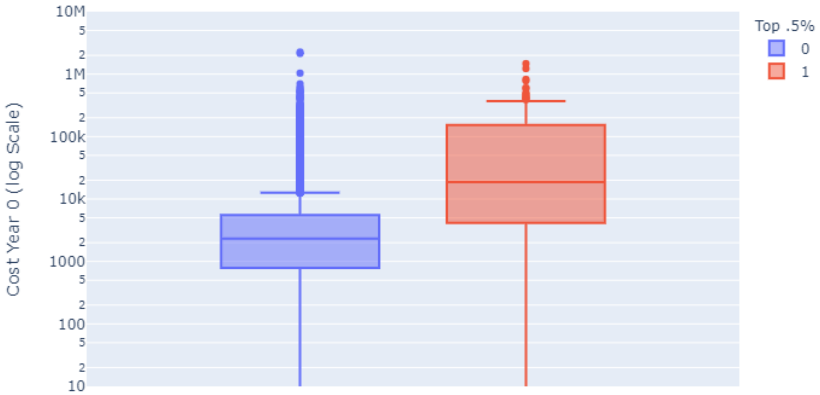


Figure 17: Year 0 Cost by high-cost class in evaluation period

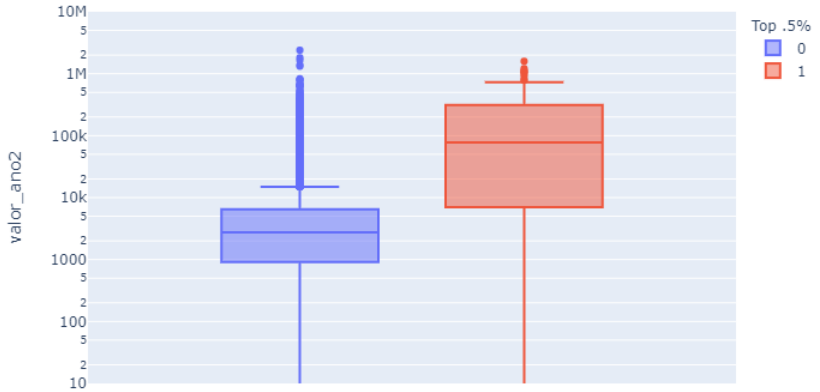


Figure 18: Year 1 Cost by high-cost class in evaluation period

Five engineered cost features were created and analysed. Cost Growth, which is the difference between assessing period's years 0 and 1 total costs and the percentage of hospital and inpatient costs to the total costs in the respective years (considering a 0 percentage if total costs were equal to 0). Figure 19 shows the Spearman's correlations heatmap for these variables and the total cost in the evaluation period.

5.3. Utilization Features

Utilization features' importance was checked through correlation analysis and descriptive statistics. As it was done with cost variables, a heatmap (Figure 20) was created to visually assess the correlation among them and between them and the cost in the evaluation period. Once again, Spearman's method was chosen, due to the way the dependent variable is created (ranking instances by their costs in the final year). No feature showed a strong correlation with costs in the evaluation period, but regular medical visits in years 0 and 1 presented the highest correlation values, although still low and high correlated between themselves.

Table 18 shows the average inpatient days in Non-Intensive Care Unit and in Intensive Care Unit during the assessing period. Although Spearman's correlation's heatmap doesn't suggest a strong relationship between these variables and costs in the evaluation period, this table presents a different scenario, with very high mean values for high-cost instances when compared to low-cost ones.

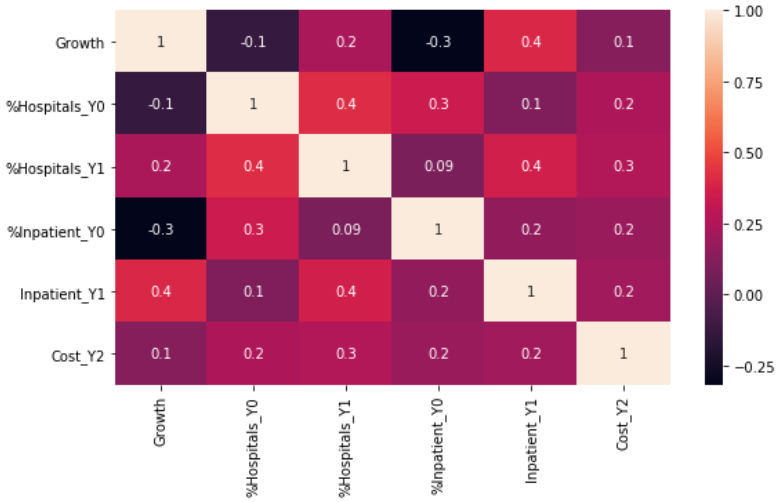


Figure 19: New Cost Features Heatmap

5.4. Clinical Data

There were high expectations for the two kinds of clinical data available: Chronic Diseases Program (VemSer) and the Diagnoses Related Groups developed according to the methodology in the Appendix. Nevertheless, correlations seemed low between them and the cost in the evaluation period. Once again, it's important to state that the original features were the number of years since the disease was first informed to PASBC (by the enrollment in the chronic disease program or in a provider's claim, for the DRGs). Table 19 shows Spearman's correlation for the Vem Ser features and Table 20, the top 14 for the DRGs.

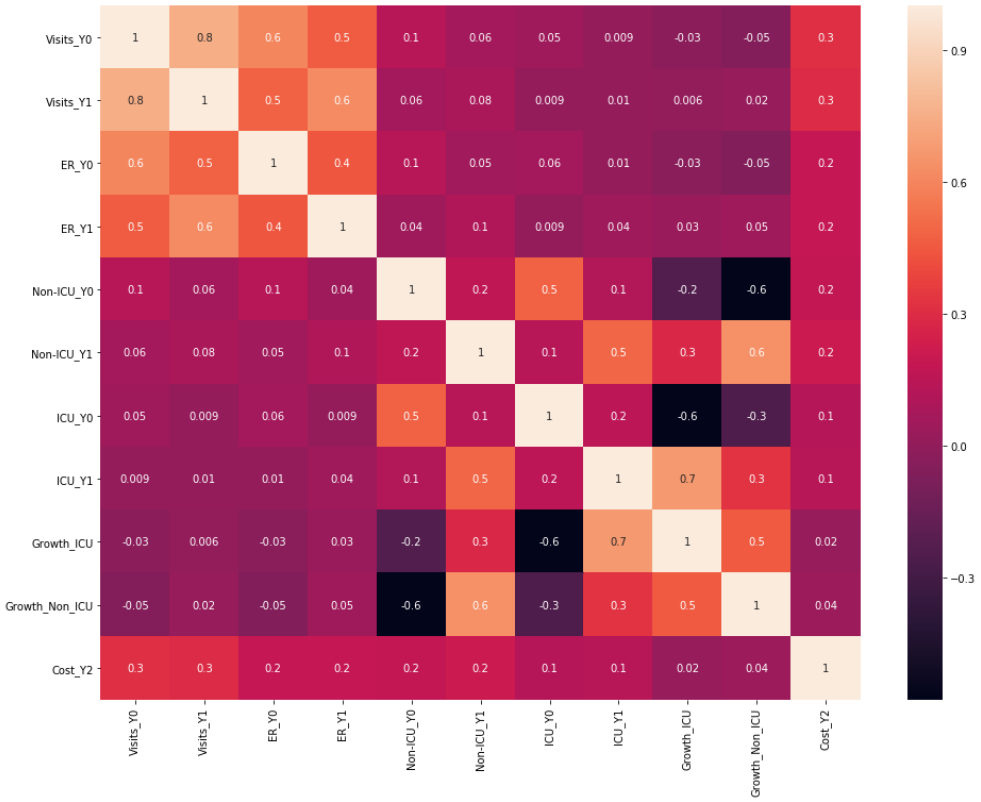


Figure 20: Utilization Features Heatmap

Table 18: Average Utilization Days

Threshold	*Top 0.5%*		*Top 1%*		*Top 2%*		*Top 5%*		*Top 10%*	
Cost Feature	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Non ICU Y0	0,8	24,8	0,8	18,6	0,7	16,1	0,5	9,5	0,4	5,7
Non ICU Y1	1,1	40,0	1,0	29,2	0,8	24,5	0,6	14,0	0,5	8,0
ICU Y0	0,1	6,3	0,1	4,9	0,1	3,1	0,1	1,6	0,1	0,9
ICU Y1	0,2	7,0	0,1	6,4	0,1	4,0	0,1	2,1	0,1	1,2

Another strategy was used, which was the creation of dummy variables for each one of the clinical features, so, instead of representing the number of years since the disease was first informed, they would just represent each one of the disease's existence or not. The sum of these dummy variables was also calculated to assess the importance of the number of conditions to predict high-cost users in the future. As shown on Table 21 and Figure 21, the number of conditions has a higher correlation than any dummy variable, suggesting that this may be a good predictor in the models.

Table 19: Chronic Diseases Program Features Correlation

	Cost Y2
Cost Y2	1.000000
Systemic Arterial Hypertension	0.189978
Dyslipidemia	0.166751
Diabetes mellitus	0.153214
Overweight / Obesity	0.125886
Malignant neoplasia	0.104452
Heart Diseases	0.100584
Alzheimer's disease	0.059742
Stroke	0.051147
Chronic Obstructive Pulmonary Disease	0.050995
Chronic Renal Failure	0.049737
Asthma	0.044784
Parkinson's Disease	0.041375
Transplantation	0.023210
Hepatitis C	0.018501
HIV/AIDS	0.017858

Table 20: DRGs Features Correlation

	Cost Y2
Cost Y2	1.000000
Primary hypertension	0.136963
Breast cancer	0.077765
Pneumonia	0.064296
Dyslipidemia.1	0.058622
Cardiopathy	0.058155
Diseases circulatory system	0.054805
Diabetes	0.054305
Osteoporosis	0.053922
Other neoplasms	0.053246
Prostate hyperplasia	0.052844
Anaemia	0.052621
Lymphoma	0.049890
Arrhythmia	0.049767
Prostate cancer	0.049253

5.5. Features Selection

After analyzing the relationship between all features and both the cost and the high-cost indicator in the evaluation period, it was time to define the datasets that would be used in the models to be tested. Besides the baseline model dataset (already explained in chapter 3.9), which only uses Age and observation period's last year total costs, three other datasets were created.

One of the datasets would use all features, including the engineered ones. Another one would use only cost features, the group of variables that showed the highest correlation with the costs in the evaluation period. Finally, a dataset was created with the best predictors identified during data exploration and features selection process. The last chapter had already shown that costs and utilization data were more correlated with the evaluation period's total costs (the variable that was ranked to create the high-cost user identifier). Nonetheless, more analytical methods were needed in order to identify the best features.

Table 21: Clinical Conditions Sum x Cost Y2 Correlations

	Cost Y2
Cost Y2	1.000000
No_Conditions_Total	0.158541
No_Conditions_DRG	0.145155
No_Conditions_VemSer	0.102585

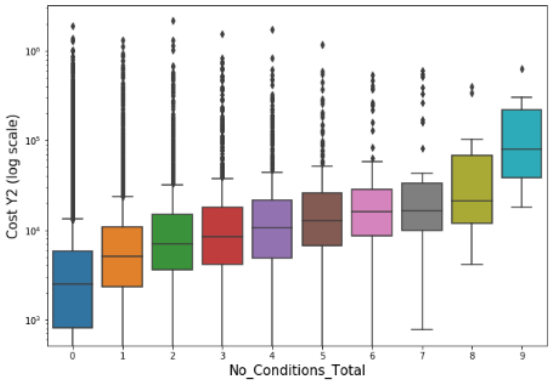


Figure 21: Total Conditions x Evaluation Period Total Costs

Initially, an Analysis of Variance (ANOVA) test was conducted for each one of the top x% high-cost indicator to split the sample. For the top .5%, 67 out of 115 features showed F-values lower than 2.5%, demonstrating that they have a statistically significant relationship with the dependent variable.

Table 22 presents the best 15 features by ANOVA F-value for this dependent variable in the training dataset. Nonetheless, ANOVA does not consider the hypotheses of multicollinearity, so, for that reason, the Variance Inflator Factor (VIF) was calculated for each one of the features. This statistic is the reciprocal of 1 - R^2 (using the R^2 of the feature regression by the other variables), so the higher the collinearity, the higher the R^2 and the VIF.

Table 23 shows that some variables are highly correlated. Cost growth was already expected to be, as it's just the difference between years 0 and 1 costs, but even after dropping this feature from the analysis, VIF kept high for other variables (Table 24).

Table 22: Best 15 ANOVA F-Values

	F_score	F_Value
Cost_1M	3561.59	0.0
Cost_2M	3111.07	0.0
Cost_6M	2658.92	0.0
Cost_Y1	2408.17	0.0
Cost_Hospitals_Y1	1631.08	0.0
Cost_Inpatient_Y1	1476.32	0.0
Non-ICU_Y1	1433.22	0.0
Cost_Y0	1369.80	0.0
Cost_Hospitals_Y0	1079.03	0.0
Cost_Inpatient_Y0	1056.24	0.0
ICU_Y1	761.79	0.0
Non-ICU_Y0	759.57	0.0
Cost_Trend	738.29	0.0
DRG - Renal failure	466.43	0.0
Growth	440.61	0.0

Table 23: Best ANOVA Features' VIFs

variables	VIF
Cost_1M	3.34
Cost_2M	4.61
Cost_6M	6.49
Cost_Y1	85.24
Cost_Hospitals_Y1	8.11
Cost_Inpatient_Y1	8.80
Non-ICU_Y1	3.13
Cost_Y0	37.25
Cost_Hospitals_Y0	7.98
Cost_Inpatient_Y0	8.48
ICU_Y1	1.34
Non-ICU_Y0	2.80
Cost_Trend	1.62
DRG - Renal failure	1.03
Growth	1625.12

Table 24: VIFs without Cost Growth

variables	VIF
Cost_1M	3.54
Cost_2M	4.94
Cost_6M	5.91
Cost_Y1	13.77
Cost_Hospitals_Y1	8.73
Cost_Inpatient_Y1	9.23
Non-ICU_Y1	3.15
Cost_Y0	8.93
Cost_Hospitals_Y0	8.61
Cost_Inpatient_Y0	8.87
ICU_Y1	1.30
Non-ICU_Y0	2.82
Cost_Trend	1.42
DRG - Renal failure	1.03

To select the best features, a method that considered multicollinearity needed to be used. Python's Sklearn library provides the Recursive Feature Elimination (RFE), which works as a backward selector, starting a model with all features and removing the least important ones until the desired number of features is reached. This method was used to find the 20 best features of four different algorithms: Logistic Regression, Decision Tree and Random Forests (without limit and with 4 levels of depth). Table 25 shows which features were selected for each one of them when this wrapped algorithm was used in the training dataset considering the top .5% high-cost indicator.

Table 25: 20 Best Features by Algorithm using Recursive Feature Elimination

Features/Algorithms	Logistic Regression	Decision Tree	Random Forest (complete)	Random Forest (4 levels of depth)
Age	X	X	X	X
Cost_Y0		X	X	X
Cost_Y1	X	X	X	X
Cost_1M	X	X	X	X
*Cost_2M**	X	X	X	X
Cost_6M		X	X	X
Acute	X	X	X	X
Cost_Trend		X	X	X
Cost_Hospitals_Y0		X	X	X
Cost_Hospitals_Y1			X	X
Cost_Inpatient_Y0	X	X	X	X
Cost_Inpatient_Y1				X
Visits_Y0		X	X	
Visits_Y1	X	X		
Non-ICU_Y0		X	X	X
Non-ICU_Y1		X	X	X
ICU_Y0				X
VS - HIV/AIDS	X			
*VS - Dyslipidemia**	X			
VS - Parkinson's Disease	X			
VS - Systemic Arterial Hypertension		X		
*DRG - Asthma**	X			
DRG - Epilepsy	X			
DRG - Renal failure	X	X		
DRG - Melanoma	X			
*DRG - Myeloma**	X			
DRG - Neopl malig of bronchi and lungs	X			
*DRG - Malignant neoplasm of the liver**	X			
Growth		X	X	X
%Hospitals_Y0		X	X	
%Hospitals_Y1			X	
*%Inpatient_Y0**	X	X	X	X
%Inpatient_Y1	X		X	
Growth_Non-ICU			X	X
Growth_ICU				X
No_VS_Conditions				X
No_DRG_Conditions	X			
No_Conditions_Total		X		

* Statistically non-significant in the Logistic Regression

Each model was analysed separately, with the 20 features selected with RFE for each one of them. The Logistic Regression, for instance, showed p-values greater than 0.05 for 6 of the 20 independent variables (Table 26). After removing these features, all remaining ones presented statistically significant p-values (Table 27).

Table 26: 20 Features Logistic Regression Stats Summary

	coef	std err	z	P>|z|	[0.025	0.975]
Age	-9.0904	0.278	-32.658	0.000	-9.636	-8.545
Cost_Y1	5.1599	0.472	10.931	0.000	4.235	6.085
Cost_1M	3.7893	0.718	5.275	0.000	2.381	5.197
Cost_2M	1.0866	0.812	1.338	0.181	-0.506	2.679
Acute	-10.9821	0.479	-22.910	0.000	-11.922	-10.043
Cost_Inpatient_Y0	4.4102	0.534	8.258	0.000	3.364	5.457
Visits_Y1	-7.5083	1.248	-6.016	0.000	-9.954	-5.062
VS - HIV/AIDS	9.2393	1.584	5.831	0.000	6.134	12.345
VS - Dyslipidemia	0.9044	1.797	0.503	0.615	-2.619	4.427
VS - Parkinson's Disease	3.6149	0.972	3.718	0.000	1.709	5.520
DRG - Asthma	0.5221	1.394	0.375	0.708	-2.210	3.254
DRG - Epilepsy	3.6344	1.445	2.515	0.012	0.802	6.467
DRG - Renal failure	3.3354	0.889	3.751	0.000	1.593	5.078
DRG - Melanoma	6.3246	1.760	3.594	0.000	2.876	9.774
DRG - Myeloma	1.9764	1.404	1.407	0.159	-0.776	4.729
DRG - Neopl malig of bronchi and lungs	4.0483	1.530	2.646	0.008	1.049	7.047
DRG - Malignant neoplasm of the liver	-108.4279	1.67e+08	-6.5e-07	1.000	-3.27e+08	3.27e+08
%Inpatient_Y0	-0.3592	0.447	-0.804	0.421	-1.235	0.516
%Inpatient_Y1	-4.0417	0.431	-9.383	0.000	-4.886	-3.197
No_DRG_Conditions	3.8827	0.892	4.355	0.000	2.135	5.630

Table 27: 14 Best Features Logistic Regression Stats Summary

	coef	std err	z	P>|z|	[0.025	0.975]
Age	-9.1174	0.275	-33.135	0.000	-9.657	-8.578
Cost_Y1	5.4234	0.422	12.843	0.000	4.596	6.251
Cost_1M	4.6269	0.374	12.360	0.000	3.893	5.361
Acute	-10.9688	0.476	-23.042	0.000	-11.902	-10.036
Cost_Inpatient_Y0	4.0937	0.328	12.493	0.000	3.451	4.736
Visits_Y1	-7.7345	1.245	-6.213	0.000	-10.174	-5.295
VS - HIV/AIDS	9.2679	1.557	5.954	0.000	6.217	12.319
VS - Parkinson's Disease	3.6342	0.954	3.808	0.000	1.764	5.505
DRG - Epilepsy	3.7649	1.415	2.661	0.008	0.992	6.538
DRG - Renal failure	3.2349	0.894	3.619	0.000	1.483	4.987
DRG - Melanoma	6.7744	1.634	4.145	0.000	3.571	9.978
DRG - Neopl malig of bronchi and lungs	4.1409	1.542	2.685	0.007	1.119	7.163
%Inpatient_Y1	-4.0815	0.423	-9.655	0.000	-4.910	-3.253
No_DRG_Conditions	4.0989	0.811	5.052	0.000	2.509	5.689

Feature importance was also calculated for both the Decision Tree and the Random Forests, based on the Gini criterion, that considers more important the least impure variable, that means, the feature that will create a splitting node that will help the classification process the most. Figures 22 to 24 present bar plots of these values.

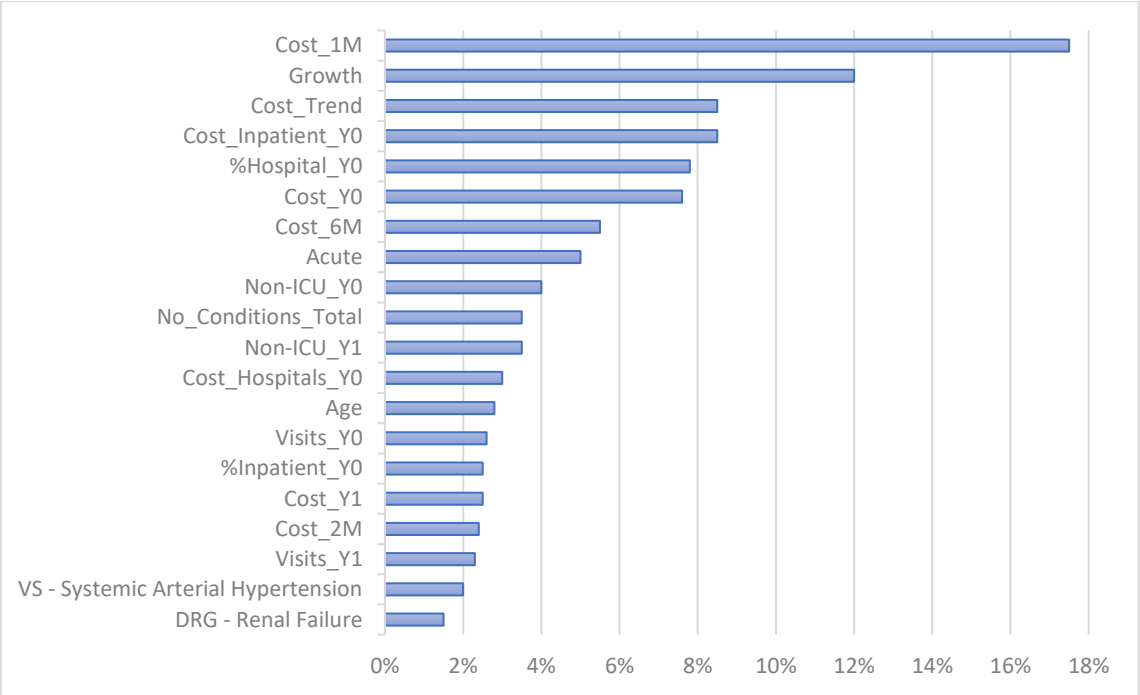


Figure 22: Decision Tree Feature Importance

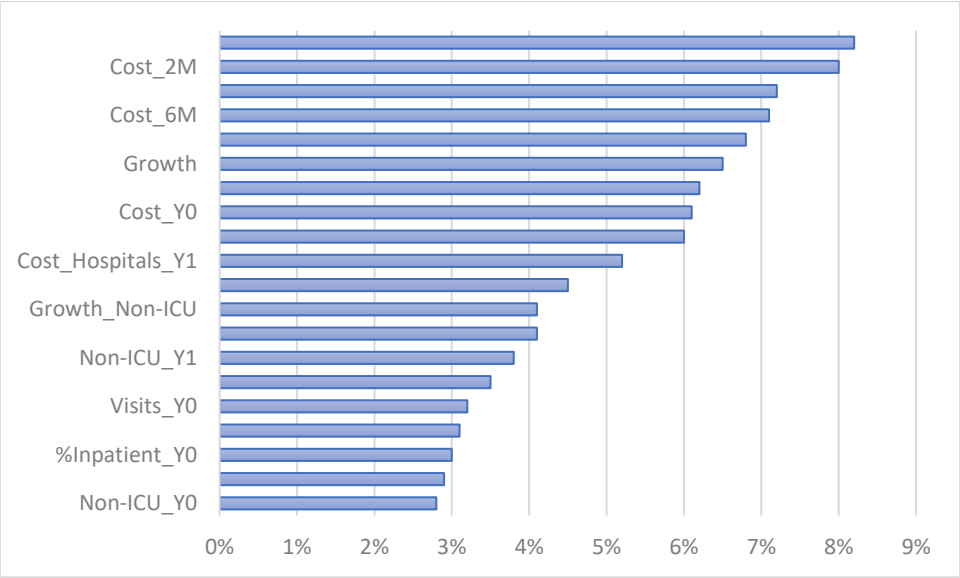


Figure 23: Complete Random Forest Feature Importance

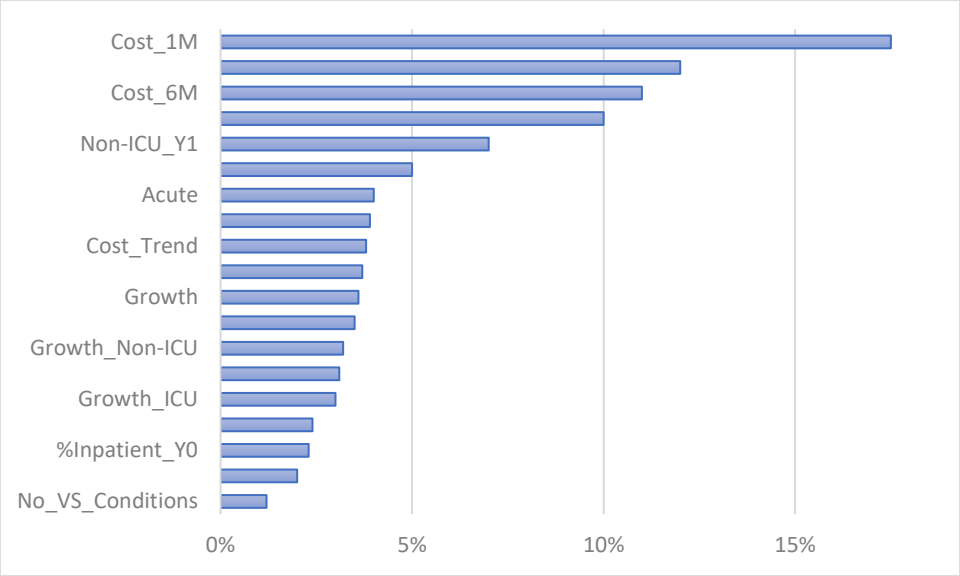


Figure 24: 4 Depth Levels Random Forest Feature Importance

According to the analysis of these metrics, considering the feature's importance, the ANOVA F-values, Logistic Regression's p-values and the consistency of selection by RFE for the different models, a 27 best features dataset was also created to be compared to the other models in the 1-year time span simple top .5% high-cost classification. Table 28 summarizes the 4 different datasets used for this problem.

Table 28: Datasets Composition

Dataset/Features	**Sociodemographics**	**Costs**	**Utilization**	**Clinical**
Baseline	Age	Last year total costs		
Complete	Age and Gender	11 Original* • Y0 and Y1 Costs • 1M, 2M and 3M Costs • Y0 and Y1 Hospitals and Inpatient Costs 5 Engineered* • Y1 Cost Growth	8 Original* • Y0 and Y1 ER and regular visits • Y0 and Y1 ICU and Non-ICU days 2 Engineered* • Y1 ICU and Non-ICU days Growth	15 chronic conditions in disease management program (years since enrolment) 69 DRG (years since first claim) 3 Engineered (sums of conditions)

		• Hospitals and Inpatient Cost %		
Costs		11 Original* 5 Engineered*		
Best	Age	10 Original* (all except Y1 Inpatient Cost) 3 Engineered • Cost Growth* • %Inpatient_Y0 • %Inpatient_Y1	4 Original* (Years 0 and 1 Physician visits and Non-ICU inpatient days) 1 Engineered* (Non-ICU inpatient days Growth)	7 Original (years since enrolment in Vem Ser or first claim) • VS - HIV/AIDS • VS - Parkinson's Disease • VS - Systemic Arterial Hypertension • DRG - Epilepsy • DRG - Renal failure • DRG - Melanoma • DRG - Neopl malig of bronchi and lungs 1 Engineered (Total number of conditions – chronic diseases program + DRG)

* One-year assessing period datasets didn't have Growth features nor Y1 features (just Y0)

The same best features selection strategy was used for all the others 19 classification problems. Independent variables used can be found in the Appendix 12.3.

5.6. Principal Components Analysis

Principal Components Analysis was the last method used to create datasets. This strategy creates uncorrelated pseudo-features composed by the original variables in a way that the new components explain most of the variance of all features in fewer dimensions, while reducing the multicollinearity (due to the uncorrelation). With 115 features, the use of PCA could be a good alternative to reduce the dimensionality of the problem while using information from all variables.

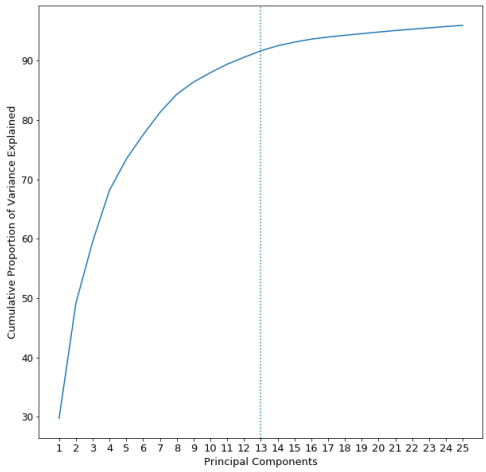


Figure 25: PCA Cumulative Variance Explained

Figure 25 shows that 12 components seemed a good number, as they could explain more than 90% of the variance of all 115 features and the next component's increment was lower than 1%. Table 29 presents the top 15 mean absolute coefficients' values for the 115 features in these 12 components. Cost variables, age and visits are the ones that explain most of the variance in the dataset according to this table. A dataset with these 12 principal components was also created to be evaluated in the next chapter. Coefficients for each feature in all 12 components can be found in the Appendix 12.2.

Table 29: Top 15 Mean Absolute 12 PCA Coefficients

	mean
%Inpatient_Y1	0.24
%Inpatient_Y0	0.23
%Hospitals_Y1	0.20
Acute	0.20
Age	0.18
%Hospitals_Y0	0.18
VS - Overweight / Obesity	0.16
Cost_Y0	0.15
Visits_Y1	0.13
No_Conditions_Total	0.13
Cost_2M	0.13
Cost_Hospitals_Y0	0.13
Cost_Inpatient_Y0	0.12
Visits_Y0	0.12
Cost_6M	0.12

6. MODELS' VALIDATION AND CHOICE

After the datasets were created, it was time to build and compare different models, using unseen data from the validation dataset to evaluate their metrics and choose the best ones to be used in the test dataset. The idea was to choose the best potential model for each one of the five datasets, so it was possible to compare their results in the test dataset.

6.1. GRID SEARCHES FOR MODEL TUNING

The first step was running grid searches to compare the four preselected algorithms (Logistic Regression, Decision Trees, Random Forests and Multilayer Perceptrons) with different parameters for each one of the five datasets. The tested parameters can be seen on Table 30.

Table 30: Grid Search Parameters

Classifier	Parameters
Logistic Regression	Single Model
Decision Tree	Maximum Depth per Tree*: • 3, 5 or 7 Maximum Nodes per Leaf*: • 3, 4 or 5 Tree's Split Criteria • Gini • Entropy
Random Forest	Number of Trees: 10, 50 or 100 Maximum Depth per Tree*: • 3, 5 or 7 Maximum Nodes per Leaf*: • 3, 4 or 5 Tree's Split Criteria • Gini • Entropy
Multilayer Perceptron	Hidden Layers • 5 – 5 – 5 • 10 – 10 • 15 – 15 – 15 Activation Function • Logistic • Relu • Tanh • Identity

* Max Depth and Max Leaf Nodes parameters weren't used simultaneously

After running the grid searches, the best parameters' combination for the two top classifiers were selected. This selection was made according to the precision by rank and the cost capture for the training and the validation datasets. The main metric used was the validation dataset cost capture. Then, the training dataset's measures were checked to avoid overfitting (if values were considerably greater than the validation set's ones, this option was discarded and the second better was evaluated).

As the parameters for an algorithm were defined, the second best algorithm would be found and the same process would be done for it, so the best two classifiers and its parameters would be chosen.

Table 31 is a slice of the 61 lines grid search's results table for the Top 10% Bloomers 1 Year Time Span classification model using baseline dataset. According to its results, a Decision Tree with maximum depth of 5 layers and Gini splitting criteria and a Multilayer Perceptron with the Tanh sigmoid activation function and three hidden layers of 5 nodes were chosen to be further evaluated for the baseline dataset. The same process was repeated for everyone of the five datasets for all the 20 classification models studied in this work.

Table 31: Top 10% Bloomers 1 Year Time Span Grid Search Results Slice

Model	Precision_Train	Precision_Test	%_Train	%_Test
DT_gini_depth_5	0.236	0.223	0.416426	0.384918
MLP_activation_tanh_layers_5_5_5	0.222	0.231	0.358939	0.379326
RF_10estimators_depth_3_entropy	0.226	0.227	0.388040	0.375910
MLP_activation_tanh_layers_10_10	0.228	0.226	0.356983	0.374784
MLP_activation_relu_layers_5_5_5	0.228	0.226	0.347354	0.374552
RF_100estimators_depth_3_entropy	0.230	0.226	0.315193	0.374419
MLP_activation_relu_layers_15_15_15	0.230	0.225	0.350657	0.374376
RF_100estimators_depth_5_gini	0.258	0.227	0.397876	0.374001

6.2. Customized Stacking Method

A customized method was also developed to be compared with the others. Using the best features dataset, a new one was created with the probalities calculated by the two best classifiers identified for the later, but not just that. In this dataset, the probabilities calculated by three SMOTE oversampled models were also added. After evaluating these SMOTE models (see Table 32), it was considered a good strategy to multiply by 2, 3 and 4 the positive class occurrence (for a top 0.5% high-cost users classification, for instance, models with 1%, 1.5% and 2% were developed) to create the three new probability-features. This 5 probability-features dataset was than evaluated by a grid search to tune the parameters for a Multilayer Perceptron as the final classifier. This algorithm and a logistic regression were then chosen for the final evaluation of this Engineered Dataset, as it was done for the other five datasets.

Table 32: Smote Models Evaluation Example

Model	Precision_Train	Precision_Test	%_Train	%_Test	Precision_Model	F1	ROC	PR_Curve
Smote_40.0	0.254	0.244	0.407470	0.407915	0.148	0.199	0.762	0.187920
Smote_30.0	0.254	0.239	0.406820	0.403299	0.182	0.195	0.761	0.188716
Smote_20.0	0.252	0.245	0.392593	0.402703	0.189	0.105	0.763	0.192932

6.3. Models' Final Validation

With the two best classifiers for each dataset already tuned, it was time for a final validation, using more metrics, to choose the best one to be tested with each dataset. Table 33 is an example of this final evaluation of the two best classifiers for each dataset. The columns for the PCA dataset bring

some metrics for a Random Forest and a Multilayer Perceptron and, according to them, the former was chosen as the best classifier for this dataset. As one can see, this choice is not always that easy (see the two first columns for the baseline dataset, for instance), but a single classifier has always been selected as the best one to be evaluated with each group of features in the test dataset.

Table 33: 1 year Time Span Top 5% Classification Models Final Evaluation

	BASELINE_MLP	BASELINE_RF	COST_MLP	COST_LOGIT	COMPLETE_RF	COMPLETE_MLP	BEST_MLP	BEST_RF	PCA_RF	PCA_MLP	STACK_SMOTE_MLP	STACK_SMOTE_LOGIT
PRECISION_RANK_TRAIN	37%	38%	38%	38%	41%	41%	39%	39%	34%	39%	41%	40%
PRECISION_TRAIN	29%	35%	37%	41%	42%	33%	32%	40%	43%	32%	35%	36%
AUC_TRAIN	84%	84%	80%	79%	84%	85%	85%	84%	69%	85%	85%	85%
RECALL_TRAIN	45%	39%	39%	36%	40%	48%	46%	39%	29%	45%	43%	43%
AUC_PR_TRAIN	32%	33%	39%	36%	44%	42%	39%	40%	34%	40%	41%	41%
%_TRAIN	57%	56%	58%	58%	60%	62%	61%	58%	55%	62%	60%	60%
PRECISION_RANK_TEST	38%	38%	40%	39%	40%	40%	40%	40%	32%	40%	41%	41%
PRECISION_TEST	29%	33%	35%	39%	35%	27%	30%	37%	22%	28%	33%	34%
AUC_TEST	85%	84%	81%	80%	84%	85%	85%	84%	70%	85%	86%	86%
RECALL_TEST	49%	42%	42%	38%	44%	52%	52%	42%	38%	53%	48%	47%
AUC_PR_TEST	34%	34%	37%	35%	39%	38%	38%	39%	28%	38%	38%	39%
%_TEST	57%	56%	57%	57%	57%	58%	60%	58%	51%	58%	59%	60%

7. MODELS' RESULTS

After having selected and tuned the best algorithm for each one of the 120 pairs of models and datasets (Baseline, Costs, Complete, Best Features, PCA and Probabilities for Stacking) to be tested, it was finally time to check the results in unseen data, using the test datasets that were not used so far. As seen on section 3.4, these sets bring the most recent data used on this study, classifying high-cost users by their expenses in 2019.

Models were not retrained using the validation dataset before being tested. It could have been done, but it wasn't understood as needed, as more recent data wouldn't necessarily make better predictions based on scaled features. For instance, the way the scaled inpatient costs from 2017 support top high-cost users classification in 2018 (validation set) would not necessarily promote better results in the 2019 classification (test set) than the way the 2016 costs support the 2017 classification (training set). Due to the different data, the patterns may change (and if the training data were much older than the validation set, these changes could be substantial and it would make sense to retrain the model with the validation data), but not in a way that necessarily would improve models' predictions.

A different approach would be merging both training and validation sets, but this approach wouldn't be correct as the dependent variable is derived from ranked costs and many of the features are also costs, so they suffer the influence of inflation over time. This way, all merged datasets top ranked users could be from the most recent period, despite the predictors' values, just because of the role of inflation over the resulting period costs.

Table 34 summarizes best metrics found for each one of the models tested and all results are detailed in Appendix 12.4.

The best of the six models developed using features from two previous years to predict the current year's top 0.5% high-cost users achieved a cost capture of 55% with an area under the ROC curve of .929, recalling 51% of the minority class' instances. These values were much better than the ones achieved by previous results found in the literature. Meenan et at. (2003) and Moturu et al. (2010) reached cost captures of 24% and 30% respectively (although the former used a monetary value threshold, focusing on top .69%, and not on top 0.5% as this study). The AUROC of .83 calculated by Moturu et al. (2010) was also considerably lower than the value reached in this study. Regarding the calculated probabilities ranking-based precision, the value of .408 reached was slightly lower than the .427 that Bertsimas et al (2008) achieved for their top bucket (0.5%) "hit ratio", what might be explained by more concentrated costs, as that study's top .5% high-cost users accounted for 27.9% of the total value, against 23.9% in this study, but also by the use of more features (1,523 vs 115) and, as it will be explained further in this section, by more precise clinical predictors.

Results for the top 10% high-cost users without one-year time span were algo interesting. The best model could capture 67% of the total costs of this group, reaching a probability ranking-based precision of .452 and an AUROC of .825. Tamang et al. (2016) and Kim and Park (2019) captured 60% and 66% of the top 10% costs, respectively, while both found an AUROC of .843. The cost captured in this study accounted for 46% of PASBC's 2019 total costs and the average cost for the top 10% predicted users was 7.6 times that of the bottom 90% ones.

Regarding the cost-bloomers models with no time-span between predictors and dependent variable, the top 0.5% high-cost bloomers best model could capture 31.5% of this group's 15% of 2019 total costs, classifying instances 8.5 times more expensive than the rest. It's important to state that cost-bloomers' models add one layer of complexity, as it drops the previous year's top high-cost instances from the sample, but still considers them while determining the top instances in the evaluating period. This approach is different than the one taken by Tamang et al (2016), who dropped last year's top 10% from the sample and defined the top 10% of the remaining instances as cost-bloomers, and by Dove et al (2003), who specified a monetary value as the border between low and high-cost, defining as bloomer the users who crossed this line from one year to another.

Table 34: Best Metrics by Model's types and Thresholds

Model/Metrics		Cost Capture	Ranking-Based Precision	AUROC	AU PR Curve
No time-span Simple High-Cost	.5%	.553	.408	.929	.292
	1%	.557	.418	.923	.364
	2%	.546	.432	.905	.412
	5%	.623	.420	.858	.432
	10%	.669	.452	.825	.471
No time-span cost-bloomers	.5%	.315	.156	.891	.061
	1%	.248	.137	.869	.068
	2%	.240	.131	.842	.071
	5%	.319	.206	.805	.138
	10%	.394	.251	.765	.197
One-year time-span Simple High-Cost	.5%	.332	.216	.890	.132
	1%	.394	.291	.874	.214
	2%	.436	.317	.860	.261
	5%	.511	.345	.827	.310
	10%	.576	.388	.794	.379
One-year time-span Cost-bloomers	.5%	.201	.094	.860	.036
	1%	.220	.126	.838	.064
	2%	.253	.150	.819	.087
	5%	.282	.195	.786	.139
	10%	.338	.230	.743	.188

The best top 10% high-cost bloomers model's results were also interesting, identifying instances 3.5 times more expensive than the ones in the bottom 90% group and capturing 39.4% of this group's cost, what accounts for 12% of PASBC's 2019 total costs. Although Tamang et al. (2016) reached a greater value (49% cost capture), the methodology difference cited before helps explaining it.

One-year time span top % high-cost users' best model reached a 51% cost capture, while top 1% and top 10% captured 39% and 58% respectively. It's important to remember that consecutive models (without time-spans) could capture 62%, 56% and 67% of the costs for the same thresholds, so, the higher the threshold, the lower the difference, what may be explained by the costs' concentration, what will be better explained ahead in the next chapter. Another relevant aspect of the findings is that the top 5% best model could capture 30% of PASBC's 2019 total costs, predicting instances 7.4 times riskier than the bottom 95% ones, despite a whole one-year time-span between predictions and actual expenses, giving great opportunities for preventive care.

Capturing, respectively, 28.2% and 33.8% of the target classes' total expenses, top 5% and top 10% best cost-bloomers models showed that, although hard, this kind of approach is not impossible, identifying instances 4 and 3 times more expensive than the classified as non-cost-bloomers (11 times for the top 0.5% model).

8. DISCUSSION

This study's findings support that the previous years' costs, utilization, and clinical data can predict with considerable precision healthcare high-cost users in future periods. Even when not correctly identifying the exact highest-cost instances, the created models can classify riskier instances that demand better healthcare and should be primary targets for preventive care programs. This would also contribute to improve user's quality of life and medical outcomes while helping to control the natural increase in costs due to population ageing and technical innovations.

According to Dove et al. (2003), consistent high-cost enrollees usually have chronic conditions that don't allow preventive care, the reason why cost-bloomers' identification is very important, as they can be the preferential target for primary care. For that reason, results achieved for this kind of models and presented in the last chapter are relevant and demonstrate that this kind of predictions can be an important tool for governments and healthcare insurers and providers.

One-year time-span models were the main novelties in this study. Meenan et al. (2003) and Dove et al. (2003) suggested this kind of approach, so healthcare providers and payers could take more efficient preventive actions. During the literature review, not a single model like this was found, with Rosella et al. (2018) being the only study that had an approach similar to this one while classifying instances as cost-bloomers in one of the five next years. Findings support this kind of model, as despite not reaching the same performance as the ones without time-span between the assessment and the evaluation year, metrics showed that they can be used to capture at least part of the high-risk instances' cost.

The last type of models added two layers of uncertainty: cost-bloomers classification with a time-span between predictors and the dependent variable. Despite knowing beforehand that this kind of model would present the worst performance, its results were also the most expected, as approaches like this weren't found before and because, according to previous authors (Meenan et al, 2003; Dove et al, 2003), it could promote the best results in healthcare, identifying low-cost users that will become high-cost ones a whole year before it happens, what creates an amazing opportunity for preventive care and improve of outcomes. Once again, results presented in last chapter show that, although not as precise as simpler models, this kind of predictions may be important to improve clinical outcomes while controlling healthcare costs' increase.

Another finding of this study concerns the predictive power of a baseline model. There isn't a rule to create a simple combination of features to make healthcare high-cost users predictions. In this work, age and last year' total costs were understood as easy to find variables that common sense would support as important predictors. It was supposed beforehand that they would reach a reasonable performance, but the results seemed even more impressive than expected: just once the baseline model captured less than 60% of the cost captured by the best model (one-year time span top 1% bloomers, what probably was due to the choice and tuning of the random forest algorithm used). It's true that for the lowest thresholds (top .5% and top 1%), baseline's performance was consistently worse than the best model's, what may be explained by the need of more features to correctly discriminate positive and negatives when the minority class is so small, nevertheless, the metrics achieved were still quite impressive on average.

This performance was very different than other studies' baseline models. Usually, the former reached very poor results, like in Ash et al. (2000), when the baseline regression achieved a R^2 lower than 2%, more than 90% less than the study's best model. The main reason for that difference lies in the variables used. Ash et al. (2000), Fleishman and Cohen (2010) and Powers et al. (2005) used only demographic features, like age and gender, while Kim and Park (2019) added to them disability codes and type of insurance, and Tamang et al (2016) worked also with some diseases' risk scores. From the reviewed studies, Yang et al. (2018) and Bertsimas et al. (2008) used last years' costs in the baseline model, although the former only to rank the instances, without pre-processing data or choosing and tuning a best algorithm, which may be the second reason for previous baseline model's poor performance when compared to the ones reached in many models of this work.

This study did not only choose two important features that common sense would incentive to use, but also completely pre-processed data, truncating values as a way to treat outliers and scaling, and validated the baseline's metrics the same way it was done for other datasets, choosing and tuning a best baseline algorithm. It was known beforehand that this approach would improve considerably the baseline model's results (lowering the comparative increase in performance between baseline and best models), but it was understood that this was the way to go, so the two features used could really be evaluated.

This solid approach supports another finding of this study: previous year's total costs (Year 0 or Year 1, depending on the time-span between independent and dependent variables) is a very strong predictor, able to create a simple model with feature "age" and reach a good performance if a solid algorithm choice, tuning and validation is executed.

This finding corroborates the one made by Kim and Park (2019) and Bertsimas et al. (2008), that point cost features as the most important for predicting healthcare high-cost users. Actually, as demonstrated in the Appendix 12.3, this study reached the same conclusion, as most of best variables for all 20 models are cost ones.

If by one hand cost features showed great importance, clinical data did not bring much more predictive power to the models, presenting low correlation with resulting period's costs and low feature importance for decision trees and random forests, except for the number of conditions, that regularly was selected as one of the best predictors to be included in the best features' models. Sadly, this may be explained by the clinical data available in PASBC. In section 3.5.3, medical features were presented and the problems in ICD codes were cited, as providers may send claims without codes or with a simple code to be used anytime. Besides, as the enrollment in "Vem Ser", PASBC's chronic conditions follow-up and case management program is optional, this data is also incomplete, so many instances with chronic conditions may bring negative values, despite having the disease or condition.

The surge in performance, mainly in cost capture and rank-based precision, as the high-cost classification threshold also increases was also something interesting found in this study. This can be explained by the costs' concentration and the minority class's size growth, respectively. Despite not correctly predicting an instance as top 0.5% high-cost, good algorithms will predict to it a higher probability. As the threshold increases and more expensive instances, that weren't considered top 0.5%, for example, are classified as top 5% (because the predicted probability wasn't high enough for the first case, but is for the second), a bigger share of the cost is captured, because few users represent

most of the total cost. So, if a top 1% model may identify only 50% of the minority class, the top 10% model may identify the same 50%, but it will correctly classify 90% of the top 1% minority class, for instance and, as costs are concentrated, these 90% may represent a disproportional share of the costs, increasing cost capture as the high-cost threshold grows.

9. CONCLUSIONS

In this study, it was examined the predictive power of demographic variables, besides previous utilization, costs and clinical data, to identify future healthcare services high-cost users. Data from more than 30,000 Central Bank of Brazil Health Program enrollees (including current and former employees and their relatives), from 2015 to 2019, was used to develop four types of models, crossing two dimensions: time-span and high-cost user's type. This way, models to predict top high-cost instances or cost-bloomers (non-previous high-cost enrollees that become one in the next time period) with or without one-year time-span between the assessing and evaluation years were created.

Five cut-off points to define high-cost were used (0.5%, 1%, 2%, 5% and 10%) and five different datasets were created combining the 115 predictors, four using distinct features' subsets and one using principal components method. Logistic Regressions, Decision Trees, Random Forests and Multi-layer Perceptron Neural Network were used, besides an engineered Stacking method using SMOTE. An out-of-time sample strategy was implemented, training, validating and testing the models with data from different years. A robust model choice and tuning was executed, grid searching multiple parameters and analyzing overfitting by the performance in the training data.

Up to 55% and 67% of the top 0.5% and 10% high-cost users could be correctly captured by the best simple models, while 31.5% of top 0.5% cost-bloomers costs could be correctly captured and the top 10% bloomers identified instances responsible for 12% of PASBC's total costs in 2019. One-year time-span models also reached interesting performances, with top 5% simple high-cost and cost-bloomers best models capturing 51% and 28% of these groups' costs respectively, while top 0.5% identified cost-bloomers had average expenses 11 times higher than the other instances.

Results showed the importance of previous years' cost data, predictors that were consistently considered the best ones for all models tested, and a two features baseline model (age and last year's total costs) presented good performance, mainly to classification thresholds greater than 2%. Unfortunately, clinical data used did not bring strong predictive power, what can be explained by factors presented in the last section. This doesn't mean that these features should not be used in future studies, that have to keep trying to use them, specially more qualitative, detailed and customized data, like check-up information and personal physicians' reports, remembering that privacy concerns must be a systematic priority.

Finally, following recommendations found in the literature but not followed yet by other authors according to the studies reviewed, one-year time-span models were developed and showed a performance not so lower than the ones without time span, specially for higher cut-off high-cost users points. These models' results and the ones achieved by cost-bloomers problems showed that these approaches, despite the layers of uncertainty added, can reach good metrics and be further explored by future research, mainly because one-year time lapses and the identification of users that weren't high risk ones in previous periods create opportunities for preventive and primary care that, if well taken, can consistently improve users' health outcomes and quality of life, while contributing to the fiscal sustainability of private and public health systems.

10. LIMITATIONS AND RECOMMENDATIONS

Although every training, validation and test datasets used in this study had more than 30,000 instances, basically all Central Bank of Brazil Health Program population, this number is not very impressive when compared to other studies, that used samples with up to 10 million users (Chechulin et al., 2014). It's true that, as almost the whole population of a health program was used, conclusions may be considered valid for it. Nevertheless, for future studies, it would be recommended trying to work with even larger datasets, using, if possible, data from big commercial health insurers in Brazil.

Another limitation of this study regards the clinical data used. As explained before, the inclusion of an ICD on claims sent to PASBC by healthcare providers is not mandatory and the data is not checked and validated before being read by Benner, the OLTP used by the health program, allowing general unspecific ICD codes to be sent, which don't provide any important clinical information to be used. Besides, the other clinical data used comes from PASBC's chronic conditions program, which brings interesting information but, as the enrolment is voluntary, many instances that have the listed conditions may have presented a value of 0 for these features in the dataset.

Trying self-reported clinical conditions and check-up data, like Kim and Park (2019), may be a good way to improve the performance of models, but other kinds of quali-quantitative data, more specific, may be used, like physicians reports and medical opinions. It's true that privacy is one of the major concerns and should be always considered carefully, making the most to work with anonymous information, but this kind of data may improve the predictive power considerably and is a possibility when talking about small targeted health plans, like PASBC.

Regarding cost-bloomers, the users that, as expressed by Dove et al. (2003), may be of high importance for not being high-cost ones previously, what brings good opportunities for preventive care, it's a recommendation trying to use this study's different approach. Although making it harder to make predictions, as considering instances that were removed from the sample part of the target class reduces the dataset's minority class, this approach may be more interesting than the one used by Tamang et al. (2016), that simply drop a percentage of top high-cost users and tried to classify the same percentage of instances, as if the ones dropped weren't part of the population. If the goal is to find users that will become part of a top percentage, it wouldn't make sense to artificially disregard a group of enrollees that may be part of it.

At last, the development of time-span models was also an innovation (considering the extensive literature research conducted earlier) that can bring new perspectives for healthcare high-cost users predictions and should be explored in future studies, as the performance achieved proved that it may be worth trying classifications like these to reach better practical preventive care results. Despite not being common in the literature, some of these model's results were very good when compared to models without a time interval (one-year time-span top 10% simple high-cost model, for instance, captured 58% of this group's cost, against 67% of the one without interval) and this interlude can be used to improve preventive care outcomes.

11. REFERENCES

Ash, A. S., Ellis, R. P., Pope, G. C., Ayanian, J. Z., Bates, D. W., Burstin, H., Iezzoni, L. I., MacKay, E., & Yu, W. (2000). Using diagnoses to describe populations and predict costs. Health care financing review, 21(3), 7–28.

Ash, A. S., Zhao, Y., Ellis, R. P., & Schlein Kramer, M. (2001). Finding future high-cost cases: comparing prior cost versus diagnosis-based methods. Health services research, 36(6 Pt 2), 194–206.

Bertsimas D, Bjarnadóttir MV, Kane MA, et al (2008). Algorithmic prediction of health-care costs. Operations Research 2008;56: 1382–92.

Blumenthal D., Anderson G., Burke S.P., Fulmer T., Jha A.K. & Long P. (2016). Tailoring Complex-Care Management, Coordination, and Integration for High-Need, High-Cost Patients. Vital Directions for Health and Health Care Series. Discussion Paper, National Academy of Medicine, Washington, DC.

Branco P., Torgo L., and Ribeiro R., "A survey of predictive modelling under imbalanced distributions," arXiv:1505.01658, 2015

Breiman, L., Friedman J., Olshen R., and Stone C., "Classification and Regression Trees", Wadsworth, Belmont, CA, 1984.

Breiman, L. and Cutler, A., "Random Forests", www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm accessed in December 2020.

Caley M. & Sidhu K. (2010). Estimating the future healthcare costs of an aging population in the UK: expansion of morbidity and the need for preventative care, Journal of Public Health, Volume 33, Issue 1, March 2011, Pages 117–122

Chechulin Y, Nazerian A, Rais S, Malikov K. Predicting patients with high risk of becoming high-cost healthcare users in Ontario (Canada). Healthc Policy. 2014 Feb;9(3):68-79.

Cohen SB (2001). The Concentration and Persistence in the Level of Health Expenditures over Time: Estimates for the U.S. Population, 2012-2013. In: Statistical Brief (Medical Expenditure Panel Survey (US)). Rockville (MD): Agency for Healthcare Research and Quality (US); 2001.

Cucciare MA, O'Donohue W (2006). Predicting future healthcare costs: how well does risk-adjustment work? J Health Organ Manag ;20(2-3):150-62

Dove HG, Duncan I & Robb A (2003). A prediction model for targeting low-cost, high-risk members of managed care organizations. American Journal of Managed Care. 2003;9(5):381-389.

Elkan, C (2001). The foundations of cost-sensitive learning. In Proceedings of the 17th international joint conference on Artificial intelligence - Volume 2 (IJCAI'01). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 973–978.

Fernández, A., García S., Galar M., Prati R. C., Krawczyk B., and Herrera F., Learning from Imbalanced Data Sets. Berlin, Germany: Springer, 2018.

IBGE - Instituto Brasileiro de Geografia e Estatística, Diretoria de Pesquisas, Coordenação de Contas Nacionais (2019). Conta-Satélite de Saúde: Brasil 2010-2017. Contas Nacionais n. 71 • ISSN 1415-9813.

Fleishman JA, Cohen JW (2010). Using information on clinical conditions to predict high-cost patients. Health Serv Res. Apr;45(2):532-52.

Galdino, R. (2019). Uso de Machine Learning para predição de pacientes de alto custo no Sistema Único de Saúde (SUS). Not published yet. https://medium.com/@renatagaldino/uso-de-machine-learning-para-predição-de-pacientes-de-alto-custo-no-sistema-único-de-saúde-sus-7bb1ffa046b0. Accessed in December 2020.

Hastie T., Tibshirani R. and Friedman J. "Elements of Statistical Learning", Springer, 2009.

He, H., Ma, Y.: Imbalanced Learning: Foundations, Algorithms, and Applications, 1st edn. Wiley-IEEE Press, New York (2013).

Kim Y. J. and Park H. (2019) Improving prediction of high-cost health care users with medical check-up data. Big Data 7:3, 163–175

LaVange, L. M., V. G. Iannacchione, S. A. Garfinkel. (1986). An application of logistic regression methods to survey data: Predicting high-cost users of medical care. Proc. Survey Research Methods Section, American Statistical Association.

Larose D. T., Larose, C. D. Data Mining and Predictive Analytics, Wiley, 2015

Meenan RT, Goodman MJ, Fishman PA, Hornbrook MC, O'Keeffe-Rosetti MC, Bachman DJ (2003). Using risk-adjustment models to identify high-cost risks. Med Care. 2003 Nov;41(11):1301-12.

Morid, M. A., Kawamoto, K., Ault, T., Dorius, J., & Abdelrahman, S. (2018). Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation. AMIA . Annual Symposium proceedings. AMIA Symposium, 2017, 1312–1321.

Moturu S.T., Johnson W.G., Liu H (2010). Predictive risk modelling for forecasting high-cost patients: A real-world application using Medicaid data. International Journal of Biomedical Engineering and Technology. -3 (1-2) , pp. 114-132.

Pedregosa et al (2011) Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830.

Peixoto, S.V., Giatti, L., Elmira Afradique, M., & Fernanda Lima-Costa, M. (2004). Custo das internações hospitalares entre idosos brasileiros no âmbito do Sistema Único de Saúde. Epidemiologia e Serviços de Saúde, 13(4), 239-246

Powers, C. A., Meyer, C. M., Roebuck, M. C., & Vaziri, B. (2005). Predictive modeling of total healthcare costs using pharmacy claims data: a comparison of alternative econometric cost modeling techniques. Medical care, 43(11), 1065–1072.

Rosella, L. C., Kornas, K., Yao, Z., Manuel, D. G., Bornbaum, C., Fransoo, R., & Stukel, T. (2018). Predicting High Health Care Resource Utilization in a Single-payer Public Health Care System: Development and Validation of the High Resource User Population Risk Tool. Medical care, 56(10), e61–e69.

Sheng, V. S. and Ling, C. S. (2006). Thresholding for making classifiers cost-sensitive. In Proceedings of the 21st national conference on Artificial intelligence - Volume 1 (AAAI'06). AAAI Press, 476–481.

Smith, S., Newhouse, J. & Freeland, M. (2009). Income, Insurance, And Technology: Why Does Health Spending Outpace Economic Growth? Health affairs (Project Hope). 28. 1276-84.

Tamang S., Milstein A, Sørensen H.T. et al (2017). Predicting patient 'cost blooms' in Denmark: a longitudinal population-based study. BMJ Open 2017;7

Yang, C., Delcher, C., Shenkman, E. et al (2018). Machine learning approaches for predicting high cost high need patient expenditures in health care. BioMed Eng OnLine 17, 131

12. APPENDIX

12.1. ICD CODES GROUPING TABLE

Diagnoses Group - Feature	CID Code	CID Description	Quantity
Primary Hypertension	I10	Essential (primary) hypertension	3486
Pneumonia	J12	Viral pneumonia, not elsewhere classified	10
	J12.2	Parainfluenza virus pneumonia	1
	J12.8	Other viral pneumonia	5
	J12.9	Viral pneumonia, unspecified	6
	J13	Pneumonia due to Streptococcus pneumoniae	16
	J15	Bacterial pneumonia, not elsewhere classified	356
	J15.0	Pneumonia due to Klebsiella pneumoniae	4
	J15.1	Pneumonia due to Pseudomonas	4
	J15.2	Pneumonia due to Staphylococcus	1
	J15.4	Pneumonia due to other streptococci	1
	J15.5	Pneumonia due to Escherichia coli	1
	J15.6	Pneumonia due to other aerobic Gram-negative bacteria	1
	J15.7	Pneumonia due to Mycoplasma pneumoniae	2
	J15.8	Other bacterial pneumonia	74
	J15.9	Bacterial pneumonia, unspecified	304
	J16	Pneumonia due to other infectious organisms, not elsewhere classified	4
	J16.8	Pneumonia due to other specified infectious organisms	3
	J17	Pneumonia in diseases classified elsewhere	1
	J17.0	Pneumonia in bacterial diseases classified elsewhere	1
	J17.8	Pneumonia in other diseases classified elsewhere	2
	J18	Pneumonia, organism unspecified	254
	J18.0	Bronchopneumonia, unspecified	355
	J18.8	Other pneumonia, organism unspecified	6
	J18.9	Pneumonia, unspecified	205
Calculus of kidney and ureter	N20	Calculus of kidney and ureter	367
	N20.0	Calculus of kidney	223
	N20.1	Calculus of ureter	188
	N20.2	Calculus of kidney with calculus of ureter	36
	N20.9	Urinary calculus, unspecified	38
Heart Diseases	I24.8	Other forms of acute ischaemic heart disease	28
	I24.9	Acute ischaemic heart disease, unspecified	9
	I25	Chronic ischaemic heart disease	229
	I25.0	Atherosclerotic cardiovascular disease, so described	34
	I25.1	Atherosclerotic heart disease	90
	I25.2	Old myocardial infarction	12
	I25.4	Coronary artery aneurysm	3
	I25.5	Ischaemic cardiomyopathy	59
	I25.6	Silent myocardial ischaemia	21
	I25.8	Other forms of chronic ischaemic heart disease	7
	I25.9	Chronic ischaemic heart disease, unspecified	19
	I30	Acute pericarditis	3
	I30.0	Acute nonspecific idiopathic pericarditis	3
	I31.2	Haemopericardium, not elsewhere classified	1

	I31.3	Pericardial effusion (noninflammatory)	3
	I31.9	Disease of pericardium, unspecified	1
	I32	Pericarditis in diseases classified elsewhere	1
	I33	Acute and subacute endocarditis	3
	I33.0	Acute and subacute infective endocarditis	4
	I33.9	Acute endocarditis, unspecified	1
	I34	Nonrheumatic mitral valve disorders	2
	I34.0	Mitral (valve) insufficiency	15
	I34.1	Mitral (valve) prolapse	13
	I34.2	Nonrheumatic mitral (valve) stenosis	2
	I34.9	Nonrheumatic mitral valve disorder, unspecified	1
	I35	Nonrheumatic aortic valve disorders	2
	I35.0	Aortic (valve) stenosis	24
	I35.1	Aortic (valve) insufficiency	2
	I35.2	Aortic (valve) stenosis with insufficiency	1
	I35.8	Other aortic valve disorders	3
	I35.9	Aortic valve disorder, unspecified	1
	I36	Nonrheumatic tricuspid valve disorders	1
	I37.0	Pulmonary valve stenosis	1
	I38	Endocarditis, valve unspecified	2
	I39.8	Endocarditis, valve unspecified, in diseases classified elsewhere	1
	I40	Acute myocarditis	3
	I40.9	Acute myocarditis, unspecified	2
	I41	Myocarditis in diseases classified elsewhere	1
	I41.1	Myocarditis in viral diseases classified elsewhere	1
	I42	Cardiomyopathy	2
	I42.0	Dilated cardiomyopathy	6
	I42.1	Obstructive hypertrophic cardiomyopathy	2
	I42.7	Cardiomyopathy due to drugs and other external agents	1
	I42.8	Other cardiomyopathies	4
	I42.9	Cardiomyopathy, unspecified	1
	I43.8	Cardiomyopathy in other diseases classified elsewhere	1
	I44	Atrioventricular and left bundle-branch block	2
	I44.0	Atrioventricular block, first degree	2
	I44.1	Atrioventricular block, second degree	15
	I44.2	Atrioventricular block, complete	28
	I44.3	Other and unspecified atrioventricular block	6
	I44.7	Left bundle-branch block, unspecified	5
	I45	Other conduction disorders	4
	I45.0	Right fascicular block	2
	I45.4	Nonspecific intraventricular block	1
	I45.6	Pre-excitation syndrome	1
	I46	Cardiac arrest	7
	I46.0	Cardiac arrest with successful resuscitation	3
	I46.1	Sudden cardiac death, so described	1
	I46.9	Cardiac arrest, unspecified	5
	I47	Paroxysmal tachycardia	27
	I47.0	Re-entry ventricular arrhythmia	4
	I47.1	Supraventricular tachycardia	34
	I47.2	Ventricular tachycardia	16

	I47.9	Paroxysmal tachycardia, unspecified	6
	I51	Complications and ill-defined descriptions of heart disease	1
	I51.1	Rupture of chordae tendineae, not elsewhere classified	1
	I51.3	Intracardiac thrombosis, not elsewhere classified	1
	I51.4	Myocarditis, unspecified	1
	I51.7	Cardiomegaly	1
	I52	Other heart disorders in diseases classified elsewhere	10
Malignant neoplasm of breast	C50	Malignant neoplasm of breast	273
	C50.0	Malignant neoplasm, nipple and areola	89
	C50.1	Malignant neoplasm, central portion of breast	4
	C50.2	Malignant neoplasm, upper-inner quadrant of breast	1
	C50.4	Malignant neoplasm, upper-outer quadrant of breast	1
	C50.5	Malignant neoplasm, lower-outer quadrant of breast	3
	C50.6	Malignant neoplasm, axillary tail of breast	239
	C50.8	Malignant neoplasm, overlapping lesion of breast	18
	C50.9	Malignant neoplasm, breast, unspecified	141
	D48.6	Neoplasm of uncertain or unknown behaviour, breast	36
Osteoporosis	M80	Osteoporosis with pathological fracture	73
	M80.0	Postmenopausal osteoporosis with pathological fracture	12
	M80.1	Postoophorectomy osteoporosis with pathological fracture	4
	M80.2	Osteoporosis of disuse with pathological fracture	3
	M80.3	Postsurgical malabsorption osteoporosis with pathological fracture	1
	M80.4	Drug-induced osteoporosis with pathological fracture	2
	M80.5	Idiopathic osteoporosis with pathological fracture	4
	M80.8	Other osteoporosis with pathological fracture	2
	M80.9	Unspecified osteoporosis with pathological fracture	5
	M81	Osteoporosis without pathological fracture	295
	M81.0	Postmenopausal osteoporosis	188
	M81.1	Postoophorectomy osteoporosis	4
	M81.2	Osteoporosis of disuse	4
	M81.3	Postsurgical malabsorption osteoporosis	2
	M81.4	Drug-induced osteoporosis	1
	M81.5	Idiopathic osteoporosis	32
	M81.6	Localized osteoporosis [Lequesne]	1
	M81.8	Other osteoporosis	18
	M81.9	Osteoporosis, unspecified	91
	M82	Osteoporosis in diseases classified elsewhere	7
	M82.0	Osteoporosis in multiple myelomatosis (C90.0+)	6
	M82.1	Osteoporosis in endocrine disorders (E00-E34+)	6
Lipidaemias	E78	Disorders of lipoprotein metabolism and other lipidaemias	341
	E78.0	Pure hypercholesterolaemia	184
	E78.1	Pure hyperglyceridaemia	13
	E78.2	Mixed hyperlipidaemia	93
	E78.3	Hyperchylomicronaemia	3
	E78.4	Other hyperlipidaemia	5
	E78.5	Hyperlipidaemia, unspecified	32
	E78.6	Lipoprotein deficiency	25
	E78.8	Other disorders of lipoprotein metabolism	16
	E78.9	Disorder of lipoprotein metabolism, unspecified	19
Hyperplasia of prostate	N40	Hyperplasia of prostate	659

	N41	Inflammatory diseases of prostate	19
	N41.0	Acute prostatitis	19
	N41.1	Chronic prostatitis	9
	N41.2	Abscess of prostate	2
	N41.3	Prostatocystitis	2
Angina	I20	Angina pectoris	337
	I20.0	Unstable angina	230
	I20.1	Angina pectoris with documented spasm	2
	I20.8	Other forms of angina pectoris	8
	I20.9	Angina pectoris, unspecified	63
Varicose	I83	Varicose veins of lower extremities	92
	I83.0	Varicose veins of lower extremities with ulcer	36
	I83.1	Varicose veins of lower extremities with inflammation	19
	I83.2	Varicose veins of lower extremities with both ulcer and inflammation	1
	I83.9	Varicose veins of lower extremities without ulcer or inflammation	470

Other Neoplasms	C00	Malignant neoplasm of lip	3
	C00.0	Malignant neoplasm, external upper lip	1
	C00.1	Malignant neoplasm, external lower lip	1
	C00.2	Malignant neoplasm, external lip, unspecified	1
	C00.8	Malignant neoplasm, overlapping lesion of lip	1
	C01	Malignant neoplasm of base of tongue	4
	C02	Malignant neoplasm of other and unspecified parts of tongue	1
	C02.0	Malignant neoplasm, dorsal surface of tongue	1
	C02.1	Malignant neoplasm, border of tongue	2
	C02.8	Malignant neoplasm, overlapping lesion of tongue	2
	C04	Malignant neoplasm of floor of mouth	1
	C04.9	Malignant neoplasm, floor of mouth, unspecified	1
	C05	Malignant neoplasm of palate	3
	C05.0	Malignant neoplasm, hard palate	1
	C05.8	Malignant neoplasm, overlapping lesion of palate	1
	C06	Malignant neoplasm of other and unspecified parts of mouth	1
	C06.9	Malignant neoplasm, mouth, unspecified	1
	C07	Malignant neoplasm of parotid gland	11
	C08	Malignant neoplasm of other and unspecified major salivary glands	1
	C08.0	Malignant neoplasm, submandibular gland	1
	C08.9	Malignant neoplasm, major salivary gland, unspecified	1
	C09.9	Malignant neoplasm, tonsil, unspecified	2
	C10	Malignant neoplasm of oropharynx	17
	C10.0	Malignant neoplasm, vallecula	1
	C10.9	Malignant neoplasm, oropharynx, unspecified	3
	C11.0	Malignant neoplasm, superior wall of nasopharynx	2
	C11.8	Malignant neoplasm, overlapping lesion of nasopharynx	1
	C11.9	Malignant neoplasm, nasopharynx, unspecified	1
	C14	Malignant neoplasm of other and ill-defined sites in the lip, oral cavity and pharynx	2
	C14.0	Malignant neoplasm, pharynx, unspecified	3
	C15	Malignant neoplasm of oesophagus	10
	C15.0	Malignant neoplasm, cervical part of oesophagus	2

	C15.5	Malignant neoplasm, lower third of oesophagus	1
	C15.8	Malignant neoplasm, overlapping lesion of oesophagus	1
	C15.9	Malignant neoplasm, oesophagus, unspecified	4
	C16	Malignant neoplasm of stomach	17
	C16.0	Malignant neoplasm, cardia	6
	C16.1	Malignant neoplasm, fundus of stomach	1
	C16.2	Malignant neoplasm, body of stomach	3
	C16.3	Malignant neoplasm, pyloric antrum	2
	C16.5	Malignant neoplasm, lesser curvature of stomach, unspecified	2
	C16.6	Malignant neoplasm, greater curvature of stomach, unspecified	1
	C16.8	Malignant neoplasm, overlapping lesion of stomach	3
	C16.9	Malignant neoplasm, stomach, unspecified	14
	C17	Malignant neoplasm of small intestine	8
	C17.0	Malignant neoplasm, duodenum	5
	C17.1	Malignant neoplasm, jejunum	1
	C17.2	Malignant neoplasm, ileum	3
	C17.3	Malignant neoplasm, Meckel's diverticulum	1
	C17.9	Malignant neoplasm, small intestine, unspecified	2
	C19	Malignant neoplasm of rectosigmoid junction	6
	C21	Malignant neoplasm of anus and anal canal	4
	C21.0	Malignant neoplasm, anus, unspecified	1
	C21.1	Malignant neoplasm, anal canal	2
	C21.8	Malignant neoplasm, overlapping lesion of rectum, anus and anal canal	5
	C22	Malignant neoplasm of liver and intrahepatic bile ducts	14
	C22.2	Malignant neoplasm, hepatoblastoma	4
	C23	Malignant neoplasm of gallbladder	5
	C24	Malignant neoplasm of other and unspecified parts of biliary tract	2
	C24.1	Malignant neoplasm, ampulla of Vater	1
	C24.9	Malignant neoplasm, biliary tract, unspecified	2
	C26	Malignant neoplasm of other and ill-defined digestive organs	2
	C30	Malignant neoplasm of nasal cavity and middle ear	2
	C30.0	Malignant neoplasm, nasal cavity	1
	C30.1	Malignant neoplasm, middle ear	1
	C31	Malignant neoplasm of accessory sinuses	3
	C31.0	Malignant neoplasm, maxillary sinus	1
	C31.9	Malignant neoplasm, accessory sinus, unspecified	1
	C37	Malignant neoplasm of thymus	2
	C38.0	Malignant neoplasm, heart	4
	C38.3	Malignant neoplasm, mediastinum, part unspecified	1
	C38.4	Malignant neoplasm, pleura	2
	C39.8	Malignant neoplasm, overlapping lesion of respiratory and intrathoracic organs	1
	C40	Malignant neoplasm of bone and articular cartilage of limbs	4
	C40.0	Malignant neoplasm, scapula and long bones of upper limb	1
	C40.1	Malignant neoplasm, short bones of upper limb	1
	C41	Malignant neoplasm of bone and articular cartilage of other and unspecified sites	4
	C41.0	Malignant neoplasm, bones of skull and face	3
	C41.2	Malignant neoplasm, vertebral column	3
	C41.9	Malignant neoplasm, bone and articular cartilage, unspecified	4
	C47	Malignant neoplasm of peripheral nerves and autonomic nervous system	1

	C47.9	Malignant neoplasm, peripheral nerves and autonomic nervous system, unspecified	3
	C48.0	Malignant neoplasm, retroperitoneum	1
	C48.2	Malignant neoplasm, peritoneum, unspecified	2
	C49	Malignant neoplasm of other connective and soft tissue	10
	C49.0	Malignant neoplasm, connective and soft tissue of head, face and neck	3
	C49.2	Malignant neoplasm, connective and soft tissue of lower limb, including hip	2
	C49.5	Malignant neoplasm, connective and soft tissue of pelvis	1
	C49.8	Malignant neoplasm, overlapping lesion of connective and soft tissue	2
	C49.9	Malignant neoplasm, connective and soft tissue, unspecified	8
	C51.8	Malignant neoplasm, overlapping lesion of vulva	2
	C52	Malignant neoplasm of vagina	1
	C55	Malignant neoplasm of uterus, part unspecified	4
	C60	Malignant neoplasm of penis	1
	C60.0	Malignant neoplasm, prepuce	1
	C60.2	Malignant neoplasm, body of penis	1
	C62	Malignant neoplasm of testis	2
	C62.0	Malignant neoplasm, undescended testis	1
	C62.9	Malignant neoplasm, testis, unspecified	3
	C63.0	Malignant neoplasm, epididymis	1
	C63.2	Malignant neoplasm, scrotum	1
	C63.9	Malignant neoplasm, male genital organ, unspecified	1
	C65	Malignant neoplasm of renal pelvis	2
	C68	Malignant neoplasm of other and unspecified urinary organs	1
	C68.9	Malignant neoplasm, urinary organ, unspecified	3
	C69	Malignant neoplasm of eye and adnexa	1
	C69.0	Malignant neoplasm, conjunctiva	6
	C69.1	Malignant neoplasm, cornea	23
	C69.2	Malignant neoplasm, retina	3
	C69.8	Malignant neoplasm, overlapping lesion of eye and adnexa	1
	C70	Malignant neoplasm of meninges	3
	C70.0	Malignant neoplasm, cerebral meninges	5
	C70.1	Malignant neoplasm, spinal meninges	1
	C70.9	Malignant neoplasm, meninges, unspecified	1
	C72	Malignant neoplasm of spinal cord, cranial nerves and other parts of central nervous system	2
	C72.0	Malignant neoplasm, spinal cord	2
	C72.9	Malignant neoplasm, central nervous system, unspecified	6
	C74	Malignant neoplasm of adrenal gland	3
	C74.9	Malignant neoplasm, adrenal gland, unspecified	1
	C75	Malignant neoplasm of other endocrine glands and related structures	1
	C75.0	Malignant neoplasm, parathyroid gland	1
	C75.1	Malignant neoplasm, pituitary gland	15
	C75.8	Malignant neoplasm, pluriglandular involvement, unspecified	1
	C76	Malignant neoplasm of other and ill-defined sites	2
	C76.0	Malignant neoplasm, head, face and neck	6
	C76.1	Malignant neoplasm, thorax	17
	C76.2	Malignant neoplasm, abdomen	7
	C76.3	Malignant neoplasm, pelvis	2
	C76.4	Malignant neoplasm, upper limb	3

	C76.5	Malignant neoplasm, lower limb	1
	C76.7	Malignant neoplasm, other ill-defined sites	1
	C77	Secondary and unspecified malignant neoplasm of lymph nodes	2
	C77.0	Secondary and unspecified malignant neoplasm, lymph nodes of head, face and neck	2
	C77.3	Secondary and unspecified malignant neoplasm, axillary and upper limb lymph nodes	3
	C77.9	Secondary and unspecified malignant neoplasm, lymph node, unspecified	2
	C78	Secondary malignant neoplasm of respiratory and digestive organs	2
	C78.0	Secondary malignant neoplasm of lung	2
	C78.4	Secondary malignant neoplasm of small intestine	1
	C78.5	Secondary malignant neoplasm of large intestine and rectum	3
	C78.6	Secondary malignant neoplasm of retroperitoneum and peritoneum	3
	C78.7	Secondary malignant neoplasm of liver and intrahepatic bile duct	3
	C79	Secondary malignant neoplasm of other and unspecified sites	4
	C79.0	Secondary malignant neoplasm of kidney and renal pelvis	1
	C79.1	Secondary malignant neoplasm of bladder and other and unspecified urinary organs	1
	C79.3	Secondary malignant neoplasm of brain and cerebral meninges	6
	C79.4	Secondary malignant neoplasm of other and unspecified parts of nervous system	1
	C79.5	Secondary malignant neoplasm of bone and bone marrow	8
	C79.8	Secondary malignant neoplasm of other specified sites	1
	C81	Hodgkin lymphoma	13
	C81.0	Nodular lymphocyte predominant Hodgkin lymphoma	2
	C81.1	Nodular sclerosis (classical) Hodgkin lymphoma	25
	C81.9	Hodgkin lymphoma, unspecified	4
	C88.0	Waldenstrom macroglobulinaemia	2
	C88.3	Immunoproliferative small intestinal disease	1
	C88.7	Other malignant immunoproliferative diseases	1
	C88.9	Malignant immunoproliferative disease, unspecified	4
	C96	Other and unspecified malignant neoplasms of lymphoid, haematopoietic and related tissue	1
	C96.1	Malignant histiocytosis	1
	C96.2	Malignant mast cell tumour	1

Diabetes	E10	Type 1 diabetes mellitus	48
	E10.0	Type 1 diabetes mellitus with coma	10
	E10.1	Type 1 diabetes mellitus with ketoacidosis	5
	E10.2	Type 1 diabetes mellitus with renal complications	1
	E10.3	Type 1 diabetes mellitus with ophthalmic complications	2
	E10.4	Type 1 diabetes mellitus with neurological complications	4
	E10.5	Type 1 diabetes mellitus with peripheral circulatory complications	2
	E10.6	Type 1 diabetes mellitus with other specified complications	2
	E10.7	Type 1 diabetes mellitus with multiple complications	7
	E10.8	Type 1 diabetes mellitus with unspecified complications	3
	E10.9	Type 1 diabetes mellitus without complications	14
	E11	Type 2 diabetes mellitus	100
	E11.0	Type 2 diabetes mellitus with coma	19
	E11.1	Type 2 diabetes mellitus with ketoacidosis	4
	E11.2	Type 2 diabetes mellitus with renal complications	9

	E11.3	Type 2 diabetes mellitus with ophthalmic complications	1
	E11.4	Type 2 diabetes mellitus with neurological complications	11
	E11.5	Type 2 diabetes mellitus with peripheral circulatory complications	1
	E11.6	Type 2 diabetes mellitus with other specified complications	1
	E11.7	Type 2 diabetes mellitus with multiple complications	10
	E11.8	Type 2 diabetes mellitus with unspecified complications	8
	E11.9	Type 2 diabetes mellitus without complications	78
	E13	Other specified diabetes mellitus	2
	E13.5	Other specified diabetes mellitus with peripheral circulatory complications	3
	E14	Unspecified diabetes mellitus	64
	E14.0	Unspecified diabetes mellitus with coma	16
	E14.1	Unspecified diabetes mellitus with ketoacidosis	8
	E14.2	Unspecified diabetes mellitus with renal complications	3
	E14.5	Unspecified diabetes mellitus with peripheral circulatory complications	6
	E14.6	Unspecified diabetes mellitus with other specified complications	1
	E14.7	Unspecified diabetes mellitus with multiple complications	2
	E14.8	Unspecified diabetes mellitus with unspecified complications	8
	E14.9	Unspecified diabetes mellitus without complications	48
Obesity	E66	Obesity	195
	E66.0	Obesity due to excess calories	133
	E66.2	Extreme obesity with alveolar hypoventilation	1
	E66.8	Other obesity	63
	E66.9	Obesity, unspecified	70
Vascular Diseases	I71	Aortic aneurysm and dissection	10
	I71.0	Dissection of aorta [any part]	3
	I71.1	Thoracic aortic aneurysm, ruptured	2
	I71.2	Thoracic aortic aneurysm, without mention of rupture	4
	I71.3	Abdominal aortic aneurysm, ruptured	3
	I71.4	Abdominal aortic aneurysm, without mention of rupture	20
	I71.6	Thoracoabdominal aortic aneurysm, without mention of rupture	4
	I71.9	Aortic aneurysm of unspecified site, without mention of rupture	8
	I72	Other aneurysm and dissection	4
	I72.0	Aneurysm and dissection of carotid artery	5
	I72.3	Aneurysm and dissection of Iliac artery	4
	I72.4	Aneurysm and dissection of artery of lower extremity	7
	I72.8	Aneurysm and dissection of other specified arteries	3
	I72.9	Aneurysm and dissection of unspecified site	4
	I73	Other peripheral vascular diseases	9
	I73.0	Raynaud's syndrome	2
	I73.8	Other specified peripheral vascular diseases	3
	I73.9	Peripheral vascular disease, unspecified	16
	I74	Arterial embolism and thrombosis	10
	I74.1	Embolism and thrombosis of other and unspecified parts of aorta	1
	I74.2	Embolism and thrombosis of arteries of upper extremities	2
	I74.3	Embolism and thrombosis of arteries of lower extremities	22
	I74.5	Embolism and thrombosis of Iliac artery	4
	I74.8	Embolism and thrombosis of other arteries	8
	I74.9	Embolism and thrombosis of unspecified artery	2
	I77	Other disorders of arteries and arterioles	4
	I77.0	Arteriovenous fistula, acquired	7

	I77.1	Stricture of artery	8
	I77.5	Necrosis of artery	1
	I77.6	Arteritis, unspecified	4
	I78.1	Naevus, non-neoplastic	4
	I79	Disorders of arteries, arterioles and capillaries in diseases classified elsewhere	1
	I79.2	Peripheral angiopathy in diseases classified elsewhere	1
	I80	Phlebitis and thrombophlebitis	25
	I80.0	Phlebitis and thrombophlebitis of superficial vessels of lower extremities	9
	I80.2	Phlebitis and thrombophlebitis of other deep vessels of lower extremities	24
	I80.3	Phlebitis and thrombophlebitis of lower extremities, unspecified	3
	I80.8	Phlebitis and thrombophlebitis of other sites	5
	I80.9	Phlebitis and thrombophlebitis of unspecified site	16
	I81	Portal vein thrombosis	8
	I82	Other venous embolism and thrombosis	70
	I82.0	Budd-Chiari syndrome	5
	I82.1	Thrombophlebitis migrans	3
	I82.2	Embolism and thrombosis of vena cava	4
	I82.8	Embolism and thrombosis of other specified veins	31
	I82.9	Embolism and thrombosis of unspecified vein	69
Goitre	E04	Other nontoxic goitre	210
	E04.0	Nontoxic diffuse goitre	59
	E04.1	Nontoxic single thyroid nodule	59
	E04.2	Nontoxic multinodular goitre	84
	E04.8	Other specified nontoxic goitre	5
	E04.9	Nontoxic goitre, unspecified	29
Anaemia	D46.0	Refractory anaemia without ring sideroblasts, so stated	3
	D46.3	Refractory anaemia with excess of blasts with transformation	1
	D46.4	Refractory anaemia, unspecified	11
	D50	Iron deficiency anaemia	159
	D50.0	Iron deficiency anaemia secondary to blood loss (chronic)	23
	D50.8	Other iron deficiency anaemias	9
	D50.9	Iron deficiency anaemia, unspecified	51
	D51	Vitamin B12 deficiency anaemia	5
	D51.0	Vitamin B12 deficiency anaemia due to intrinsic factor deficiency	1
	D51.8	Other vitamin B12 deficiency anaemias	2
	D51.9	Vitamin B12 deficiency anaemia, unspecified	1
	D53	Other nutritional anaemias	2
	D53.0	Protein deficiency anaemia	1
	D53.1	Other megaloblastic anaemias, not elsewhere classified	2
	D53.9	Nutritional anaemia, unspecified	3
	D57.0	Sickle-cell anaemia with crisis	1
	D57.1	Sickle-cell anaemia without crisis	1
	D59	Acquired haemolytic anaemia	1
	D59.0	Drug-induced autoimmune haemolytic anaemia	4
	D59.9	Acquired haemolytic anaemia, unspecified	2
	D61.0	Constitutional aplastic anaemia	1
	D61.1	Drug-induced aplastic anaemia	1
	D61.9	Aplastic anaemia, unspecified	5
	D62	Acute posthaemorrhagic anaemia	6

	D63	Anaemia in chronic diseases classified elsewhere	5
	D63.0	Anaemia in neoplastic disease (C00-D48+)	4
	D63.8	Anaemia in other chronic diseases classified elsewhere	7
	D64	Other anaemias	20
	D64.0	Hereditary sideroblastic anaemia	2
	D64.8	Other specified anaemias	3
	D64.9	Anaemia, unspecified	106

Asthma	J45	Asthma	233
	J45.0	Predominantly allergic asthma	46
	J45.1	Nonallergic asthma	8
	J45.8	Mixed asthma	18
	J45.9	Asthma, unspecified	120
Chronic sinusitis	J32	Chronic sinusitis	150
	J32.0	Chronic maxillary sinusitis	122
	J32.1	Chronic frontal sinusitis	4
	J32.2	Chronic ethmoidal sinusitis	8
	J32.3	Chronic sphenoidal sinusitis	4
	J32.4	Chronic pansinusitis	19
	J32.8	Other chronic sinusitis	16
	J32.9	Chronic sinusitis, unspecified	102
Acute bronchitis	J20	Acute bronchitis	80
	J20.0	Acute bronchitis due to Mycoplasma pneumoniae	1
	J20.1	Acute bronchitis due to Haemophilus influenzae	1
	J20.2	Acute bronchitis due to streptococcus	1
	J20.5	Acute bronchitis due to respiratory syncytial virus	2
	J20.6	Acute bronchitis due to rhinovirus	1
	J20.7	Acute bronchitis due to echovirus	1
	J20.8	Acute bronchitis due to other specified organisms	10
	J20.9	Acute bronchitis, unspecified	107
	J21	Acute bronchiolitis	28
	J21.0	Acute bronchiolitis due to respiratory syncytial virus	12
	J21.8	Acute bronchiolitis due to other specified organisms	12
	J40	Bronchitis, not specified as acute or chronic	53
	J41	Simple and mucopurulent chronic bronchitis	64
	J41.0	Simple chronic bronchitis	14
	J42	Unspecified chronic bronchitis	8
Arrhythmia	I48	Atrial fibrillation and flutter	194
	I49	Other cardiac arrhythmias	76
	I49.0	Ventricular fibrillation and flutter	9
	I49.2	Junctional premature depolarization	1
	I49.3	Ventricular premature depolarization	10
	I49.4	Other and unspecified premature depolarization	4
	I49.5	Sick sinus syndrome	19
	I49.8	Other specified cardiac arrhythmias	12
	I49.9	Cardiac arrhythmia, unspecified	52
Renal Failure	N17	Acute renal failure	31
	N17.0	Acute renal failure with tubular necrosis	5
	N17.8	Other acute renal failure	7

	N17.9	Acute renal failure, unspecified	32
	N18	Chronic kidney disease	79
	N18.0	End-stage renal disease	92
	N18.8	Other chronic renal failure	12
	N18.9	Chronic kidney disease, unspecified	66
	N19	Unspecified kidney failure	28
Heart failure	I50	Heart failure	116
	I50.0	Congestive heart failure	105
	I50.1	Left ventricular failure	5
	I50.9	Heart failure, unspecified	125
Hypothyroidism	E03	Other hypothyroidism	115
	E03.0	Congenital hypothyroidism with diffuse goitre	25
	E03.1	Congenital hypothyroidism without goitre	7
	E03.2	Hypothyroidism due to medicaments and other exogenous substances	1
	E03.4	Atrophy of thyroid (acquired)	3
	E03.8	Other specified hypothyroidism	18
	E03.9	Hypothyroidism, unspecified	152
Malignant neoplasm of prostate	C61	Malignant neoplasm of prostate	307
Malignant Neoplasms of Skin	C44	Other malignant neoplasms of skin	55
	C44.0	Malignant neoplasm, skin of lip	11
	C44.1	Malignant neoplasm, skin of eyelid, including canthus	12
	C44.2	Malignant neoplasm, skin of ear and external auricular canal	9
	C44.3	Malignant neoplasm, skin of other and unspecified parts of face	31
	C44.4	Malignant neoplasm, skin of scalp and neck	9
	C44.5	Malignant neoplasm, skin of trunk	13
	C44.6	Malignant neoplasm, skin of upper limb, including shoulder	5
	C44.7	Malignant neoplasm, skin of lower limb, including hip	6
	C44.8	Malignant neoplasm, overlapping lesion of skin	6
	C44.9	Malignant neoplasm of skin, unspecified	64
	D04.6	Carcinoma in situ, skin of upper limb, including shoulder	1
	D04.7	Carcinoma in situ, skin of lower limb, including hip	1
	D04.8	Carcinoma in situ, skin of other sites	6
	D04.9	Carcinoma in situ, skin, unspecified	32
	D48.5	Neoplasm of uncertain or unknown behaviour, skin	23
Disorders of Thyroid	E07	Other disorders of thyroid	188
	E07.0	Hypersecretion of calcitonin	21
	E07.8	Other specified disorders of thyroid	26
	E07.9	Disorder of thyroid, unspecified	35
Cerebrovascular Diseases	I60	Subarachnoid haemorrhage	18
	I60.0	Subarachnoid haemorrhage from carotid siphon and bifurcation	1
	I60.7	Subarachnoid haemorrhage from intracranial artery, unspecified	1
	I60.9	Subarachnoid haemorrhage, unspecified	7
	I61	Intracerebral haemorrhage	8
	I61.0	Intracerebral haemorrhage in hemisphere, subcortical	1
	I61.1	Intracerebral haemorrhage in hemisphere, cortical	2
	I61.6	Intracerebral haemorrhage, multiple localized	1
	I61.9	Intracerebral haemorrhage, unspecified	9
	I62	Other nontraumatic intracranial haemorrhage	6
	I62.0	Subdural haemorrhage (acute)(nontraumatic)	5
	I62.9	Intracranial haemorrhage (nontraumatic), unspecified	6

	I63	Cerebral infarction	32
	I63.0	Cerebral infarction due to thrombosis of precerebral arteries	5
	I63.2	Cerebral infarction due to unspecified occlusion or stenosis of precerebral arteries	1
	I63.3	Cerebral infarction due to thrombosis of cerebral arteries	3
	I63.4	Cerebral infarction due to embolism of cerebral arteries	4
	I63.5	Cerebral infarction due to unspecified occlusion or stenosis of cerebral arteries	1
	I63.6	Cerebral infarction due to cerebral venous thrombosis, nonpyogenic	1
	I63.8	Other cerebral infarction	2
	I63.9	Cerebral infarction, unspecified	21
	I65	Occlusion and stenosis of precerebral arteries, not resulting in cerebral infarction	11
	I65.0	Occlusion and stenosis of vertebral artery	6
	I65.2	Occlusion and stenosis of carotid artery	35
	I66.2	Occlusion and stenosis of posterior cerebral artery	1
	I67	Other cerebrovascular diseases	5
	I67.0	Dissection of cerebral arteries, nonruptured	2
	I67.1	Cerebral aneurysm, nonruptured	28
	I67.2	Cerebral atherosclerosis	3
	I67.5	Moyamoya disease	1
	I67.6	Nonpyogenic thrombosis of intracranial venous system	1
	I67.8	Other specified cerebrovascular diseases	7
	I67.9	Cerebrovascular disease, unspecified	4
	I68.0	Cerebral amyloid angiopathy (E85.-+)	1
	I68.8	Other cerebrovascular disorders in diseases classified elsewhere	1
	I69	Sequelae of cerebrovascular disease	8
	I69.0	Sequelae of subarachnoid haemorrhage	4
	I69.3	Sequelae of cerebral infarction	2
	I69.4	Sequelae of stroke, not specified as haemorrhage or infarction	15
Stroke	I64	Stroke, not specified as haemorrhage or infarction	240
Lymphoma	C82	Follicular lymphoma	28
	C82.0	Follicular lymphoma grade I	2
	C82.1	Follicular lymphoma grade II	1
	C82.7	Other types of follicular lymphoma	2
	C82.9	Follicular lymphoma, unspecified	5
	C83	Non-follicular lymphoma	37
	C83.0	Small cell B-cell lymphoma	15
	C83.1	Mantle cell lymphoma	1
	C83.2	Mixed small and large cell (diffuse)	1
	C83.3	Diffuse large B-cell lymphoma	17
	C83.5	Lymphoblastic (diffuse) lymphoma	1
	C83.6	Undifferentiated (diffuse)	23
	C83.8	Other non-follicular lymphoma	2
	C83.9	Non-follicular (diffuse) lymphoma, unspecified	14
	C84.0	Mycosis fungoides	1
	C84.4	Peripheral T-cell lymphoma, not elsewhere classified	3
	C84.5	Other mature T/NK-cell lymphomas	3
	C85	Other and unspecified types of non-Hodgkin lymphoma	8
	C85.0	Non-Hodgkin lymphoma, lymphosarcoma	3
	C85.1	B-cell lymphoma, unspecified	9

	C85.7	Other specified types of non-Hodgkin lymphoma	2
	C85.9	Non-Hodgkin lymphoma, unspecified	24

Malignant neoplasm of colon	C18	Malignant neoplasm of colon	77
	C18.0	Malignant neoplasm, caecum	26
	C18.1	Malignant neoplasm, appendix	2
	C18.2	Malignant neoplasm, ascending colon	16
	C18.3	Malignant neoplasm, hepatic flexure	1
	C18.4	Malignant neoplasm, transverse colon	5
	C18.5	Malignant neoplasm, splenic flexure	1
	C18.6	Malignant neoplasm, descending colon	6
	C18.7	Malignant neoplasm, sigmoid colon	11
	C18.9	Malignant neoplasm, colon, unspecified	37
Thyroiditis	E06	Thyroiditis	97
	E06.0	Acute thyroiditis	22
	E06.1	Subacute thyroiditis	2
	E06.2	Chronic thyroiditis with transient thyrotoxicosis	1
	E06.3	Autoimmune thyroiditis	35
	E06.4	Drug-induced thyroiditis	1
	E06.5	Other chronic thyroiditis	2
	E06.9	Thyroiditis, unspecified	8
Chronic Obstructive Pulmonary Disease	J44	Other chronic obstructive pulmonary disease	53
	J44.0	Chronic obstructive pulmonary disease with acute lower respiratory infection	16
	J44.1	Chronic obstructive pulmonary disease with acute exacerbation, unspecified	21
	J44.8	Other specified chronic obstructive pulmonary disease	7
	J44.9	Chronic obstructive pulmonary disease, unspecified	52
Atherosclerosis	I70	Atherosclerosis	55
	I70.0	Atherosclerosis of aorta	16
	I70.1	Atherosclerosis of renal artery	5
	I70.2	Atherosclerosis of arteries of extremities	41
	I70.8	Atherosclerosis of other arteries	17
	I70.9	Generalized and unspecified atherosclerosis	13
Melanocytic naevi	D22	Melanocytic naevi	18
	D22.0	Melanocytic naevi of lip	21
	D22.1	Melanocytic naevi of eyelid, including canthus	1
	D22.2	Melanocytic naevi of ear and external auricular canal	3
	D22.3	Melanocytic naevi of other and unspecified parts of face	2
	D22.5	Melanocytic naevi of trunk	5
	D22.6	Melanocytic naevi of upper limb, including shoulder	1
	D22.7	Melanocytic naevi of lower limb, including hip	1
	D22.9	Melanocytic naevi, unspecified	86
Malignant neoplasm of bronchus and lung	C34	Malignant neoplasm of bronchus and lung	60
	C34.0	Malignant neoplasm, main bronchus	17
	C34.1	Malignant neoplasm, upper lobe, bronchus or lung	8
	C34.2	Malignant neoplasm, middle lobe, bronchus or lung	4
	C34.3	Malignant neoplasm, lower lobe, bronchus or lung	3
	C34.8	Malignant neoplasm, overlapping lesion of bronchus and lung	4
	C34.9	Malignant neoplasm, bronchus or lung, unspecified	42

Acute myocardial infarction	I21	Acute myocardial infarction	54
	I21.0	Acute transmural myocardial infarction of anterior wall	13
	I21.1	Acute transmural myocardial infarction of inferior wall	10
	I21.4	Acute subendocardial myocardial infarction	11
	I21.9	Acute myocardial infarction, unspecified	46
Alzheimer	F00.0	Dementia in Alzheimer's disease with early onset (G30.0+)	12
	F00.1	Dementia in Alzheimer's disease with late onset (G30.1+)	3
	F00.2	Dementia in Alzheimer's disease, atypical or mixed type (G30.8+)	3
	F00.9	Dementia in Alzheimer's disease, unspecified (G30.9+)	5
	G30	Alzheimer's disease	49
	G30.0	Alzheimer's disease with early onset	15
	G30.1	Alzheimer's disease with late onset	9
	G30.8	Other Alzheimer's disease	5
	G30.9	Alzheimer's disease, unspecified	13
Malignant neoplasm without specification of site	C80	Malignant neoplasm without specification of site	113
Pulmonary embolism	I26	Pulmonary embolism	59
	I26.0	Pulmonary embolism with mention of acute cor pulmonale	13
	I26.9	Pulmonary embolism without mention of acute cor pulmonale	35
Malignant neoplasm of thyroid gland	C73	Malignant neoplasm of thyroid gland	106
Malignant neoplasm of bladder	C67	Malignant neoplasm of bladder	52
	C67.0	Malignant neoplasm, trigone of bladder	17
	C67.1	Malignant neoplasm, dome of bladder	3
	C67.2	Malignant neoplasm, lateral wall of bladder	7
	C67.4	Malignant neoplasm, posterior wall of bladder	1
	C67.6	Malignant neoplasm, ureteric orifice	2
	C67.8	Malignant neoplasm, overlapping lesion of bladder	1
	C67.9	Malignant neoplasm, bladder, unspecified	17
Leukaemia	C91	Lymphoid leukaemia	6
	C91.0	Acute lymphoblastic leukaemia [ALL]	4
	C91.1	Chronic lymphocytic leukaemia of B-cell type	19
	C91.4	Hairy-cell leukaemia	1
	C91.9	Lymphoid leukaemia, unspecified	1
	C92	Myeloid leukaemia	10
	C92.0	Acute myeloblastic leukaemia [AML]	10
	C92.1	Chronic myeloid leukaemia [CML], BCR/ABL-positive	33
	C92.4	Acute promyelocytic leukaemia [PML]	2
	C92.7	Other myeloid leukaemia	1
	C92.9	Myeloid leukaemia, unspecified	5
	C95	Leukaemia of unspecified cell type	1
	C95.0	Acute leukaemia of unspecified cell type	3
	C95.1	Chronic leukaemia of unspecified cell type	1
Malignant neoplasm of piriform sinus	C12	Malignant neoplasm of piriform sinus	96
Myeloma	C90	Multiple myeloma and malignant plasma cell neoplasms	32
	C90.0	Multiple myeloma	62
	C90.2	Extramedullary plasmacytoma	1
Transient cerebral ischaemic attacks	G45	Transient cerebral ischaemic attacks and related syndromes	31
	G45.8	Other transient cerebral ischaemic attacks and related syndromes	7
	G45.9	Transient cerebral ischaemic attack, unspecified	48
Schizophrenia	F20	Schizophrenia	35

	F20.0	Paranoid schizophrenia	23
	F20.1	Hebephrenic schizophrenia	1
	F20.3	Undifferentiated schizophrenia	1
	F20.4	Post-schizophrenic depression	1
	F20.5	Residual schizophrenia	2
	F20.6	Simple schizophrenia	2
	F20.9	Schizophrenia, unspecified	4
Malignant Melanoma	C43	Malignant melanoma of skin	27
	C43.0	Malignant melanoma of lip	5
	C43.2	Malignant melanoma of ear and external auricular canal	1
	C43.3	Malignant melanoma of other and unspecified parts of face	1
	C43.5	Malignant melanoma of trunk	4
	C43.6	Malignant melanoma of upper limb, including shoulder	2
	C43.7	Malignant melanoma of lower limb, including hip	3
	C43.8	Malignant melanoma, overlapping malignant melanoma of skin	5
	C43.9	Malignant melanoma of skin, unspecified	14
Hypertensive heart disease	I11	Hypertensive heart disease	19
	I11.0	Hypertensive heart disease with (congestive) heart failure	15
	I11.9	Hypertensive heart disease without (congestive) heart failure	4
	I12	Hypertensive renal disease	1
	I13.0	Hypertensive heart and renal disease with (congestive) heart failure	3
	I13.1	Hypertensive heart and renal disease with renal failure	5
	I13.9	Hypertensive heart and renal disease, unspecified	2
	I15	Secondary hypertension	5
	I15.0	Renovascular hypertension	1
	I15.2	Hypertension secondary to endocrine disorders	1
	I15.9	Secondary hypertension, unspecified	4
Parkinson's disease	F02.3	Dementia in Parkinson's disease (G20+)	3
	G20	Parkinson's disease	56
Dementia	F01.0	Vascular dementia of acute onset	3
	F01.3	Mixed cortical and subcortical vascular dementia	1
	F01.9	Vascular dementia, unspecified	1
	F02.0	Dementia in Pick's disease (G31.0+)	5
	F02.8	Dementia in other specified diseases classified elsewhere	5
	F03	Unspecified dementia	36
	F05.0	Delirium not superimposed on dementia, so described	1
	F05.1	Delirium superimposed on dementia	7
Malignant neoplasm of brain	C71	Malignant neoplasm of brain	18
	C71.0	Malignant neoplasm, cerebrum, except lobes and ventricles	14
	C71.1	Malignant neoplasm, frontal lobe	1
	C71.2	Malignant neoplasm, temporal lobe	1
	C71.3	Malignant neoplasm, parietal lobe	1
	C71.6	Malignant neoplasm, cerebellum	1
	C71.8	Malignant neoplasm, overlapping lesion of brain	2
	C71.9	Malignant neoplasm, brain, unspecified	20
Malignant neoplasm of pancreas	C25	Malignant neoplasm of pancreas	20
	C25.0	Malignant neoplasm, head of pancreas	13
	C25.1	Malignant neoplasm, body of pancreas	1
	C25.2	Malignant neoplasm, tail of pancreas	2
	C25.4	Malignant neoplasm, endocrine pancreas	1

	C25.7	Malignant neoplasm, other parts of pancreas	4
	C25.8	Malignant neoplasm, overlapping lesion of pancreas	2
	C25.9	Malignant neoplasm, pancreas, unspecified	15

Rheumatoid Arthritis	M06	Other rheumatoid arthritis	5
	M06.0	Seronegative rheumatoid arthritis	10
	M06.2	Rheumatoid bursitis	1
	M06.4	Inflammatory polyarthropathy	5
	M06.8	Other specified rheumatoid arthritis	1
	M06.9	Rheumatoid arthritis, unspecified	30
Malignant neoplasm of corpus uteri	C54	Malignant neoplasm of corpus uteri	14
	C54.0	Malignant neoplasm, isthmus uteri	2
	C54.1	Malignant neoplasm, endometrium	25
	C54.2	Malignant neoplasm, myometrium	1
	C54.8	Malignant neoplasm, overlapping lesion of corpus uteri	2
	C54.9	Malignant neoplasm, corpus uteri, unspecified	3
Malignant neoplasm of rectum	C20	Malignant neoplasm of rectum	46
Epilepsy	G40	Epilepsy	42
Carcinoma	C22.0	Malignant neoplasm, liver cell carcinoma	8
	D01	Carcinoma in situ of other and unspecified digestive organs	1
	D02	Carcinoma in situ of middle ear and respiratory system	1
	D04	Carcinoma in situ of skin	5
	D05	Carcinoma in situ of breast	4
	D05.1	Intraductal carcinoma in situ of breast	3
	D05.9	Carcinoma in situ of breast, unspecified	12
	D06	Carcinoma in situ of cervix uteri	2
	D09.7	Carcinoma in situ of other specified sites	1
	D09.9	Carcinoma in situ, unspecified	1
Malignant neoplasm of ovary	C56	Malignant neoplasm of ovary	38
Malignant neoplasm of kidney	C64	Malignant neoplasm of kidney, except renal pelvis	36
Alcoholic cirrhosis of liver	K70.3	Alcoholic cirrhosis of liver	7
	K74	Fibrosis and cirrhosis of liver	11
	K74.6	Other and unspecified cirrhosis of liver	14
Malignant neoplasm of cervix uteri	C53	Malignant neoplasm of cervix uteri	15
	C53.0	Malignant neoplasm, endocervix	5
	C53.8	Malignant neoplasm, overlapping lesion of cervix uteri	2
	C53.9	Malignant neoplasm, cervix uteri, unspecified	8
Emphysema	J43	Emphysema	16
	J43.2	Centrilobular emphysema	4
	J43.8	Other emphysema	2
	J43.9	Emphysema, unspecified	8
Ankylosing spondylitis	M45	Ankylosing spondylitis	25
Hydrocephalus	G91	Hydrocephalus	6
	G91.0	Communicating hydrocephalus	3
	G91.2	Normal-pressure hydrocephalus	7
	G91.8	Other hydrocephalus	1
	G91.9	Hydrocephalus, unspecified	6
Severe depressive episode	F32.2	Severe depressive episode without psychotic symptoms	7
	F32.3	Severe depressive episode with psychotic symptoms	3

	F33.2	Recurrent depressive disorder, current episode severe without psychotic symptoms	10
	F33.3	Recurrent depressive disorder, current episode severe with psychotic symptoms	2
Multiple sclerosis	G35	Multiple sclerosis	16
Malignant neoplasm of larynx	C32	Malignant neoplasm of larynx	7
	C32.0	Malignant neoplasm, glottis	1
	C32.8	Malignant neoplasm, overlapping lesion of larynx	2
	C32.9	Malignant neoplasm, larynx, unspecified	3
Hepatitis C	B18.2	Chronic viral hepatitis C	11
Malignant neoplasm of liver	C22.9	Malignant neoplasm, liver, unspecified	7

12.2. PCA COMPONENTS FEATURES COEFFICIENTS

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
%Inpatient_Y1	0.0182	0.3162	0.3919	0.1554	0.2688	0.0737	-0.23	0.4242	-0.5332	0.0768	0.2074	-0.1739
%Inpatient_Y0	0.0214	0.2449	-0.3512	0.3264	0.2235	-0.4657	-0.043	0.1331	-0.0735	-0.327	-0.5427	0.0222
%Hospitals_Y1	0.0321	0.5767	0.3068	-0.5734	-0.1974	-0.4077	0.0575	-0.0832	0.138	-0.0723	-0.013	-0.0004
Acute	0.0289	0.0113	-0.0786	0.039	-0.1531	-0.0378	0.378	-0.4752	-0.6417	-0.3619	0.21	0.0444
Age	0.0312	0.1149	0.041	0.3336	-0.6169	-0.0272	-0.6527	-0.2335	-0.0323	-0.0293	-0.0012	-0.0577
%Hospitals_Y0	0.0313	0.5042	-0.6189	-0.2292	0.0084	0.5353	-0.1088	0.0422	-0.033	-0.0146	0.064	-0.0152
VS - Overweight / Obesity	-0.0038	0.0319	0.0115	0.0955	-0.2819	0.0556	0.2803	0.3662	0.0629	-0.2079	-0.0055	-0.5149
Cost_Y0	0.0153	0.1568	-0.1271	0.2498	0.0493	-0.2219	0.117	-0.103	0.0988	0.2549	0.3533	-0.042
Visits_Y1	0.022	0.0574	0.0098	-0.0047	-0.0617	0.0064	0.1421	-0.1243	-0.2277	0.4624	-0.3333	-0.1555
No_Conditions_Total	0.001	0.0955	0.0183	0.1543	-0.3127	0.0572	0.257	0.1975	0.0087	0.163	-0.0832	0.27
Cost_2M	0.0076	0.112	0.1163	0.1579	0.0618	0.1416	0.1162	-0.2184	0.2168	-0.0618	-0.0572	-0.3453
Cost_Hospitals_Y0	0.0096	0.1511	-0.1531	0.1692	0.0661	-0.1819	0.0601	-0.0362	0.1293	0.211	0.3501	0.0243
Cost_Inpatient_Y0	0.0063	0.1307	-0.1407	0.2045	0.1095	-0.2475	0.0484	-0.013	0.0941	0.122	0.2788	-0.0062
Visits_Y0	0.0223	0.0521	-0.0317	0.0133	-0.0615	0.0005	0.1094	-0.0859	-0.183	0.4335	-0.267	-0.1419
Cost_6M	0.012	0.1553	0.1624	0.1945	0.0889	0.1711	0.1085	-0.2008	0.162	-0.0571	-0.0649	-0.0263
Cost_1M	0.0071	0.0919	0.0915	0.1355	0.0534	0.1174	0.1102	-0.207	0.2109	-0.0445	-0.0389	-0.3799
Cost_Y1	0.0153	0.1901	0.1861	0.2163	0.0991	0.1706	0.0967	-0.1536	0.0851	-0.0241	-0.055	0.1764
Cost_Hospitals_Y1	0.0099	0.1821	0.1779	0.1198	0.0952	0.1316	0.0372	-0.0737	0.1083	-0.0465	-0.0609	0.2752
No_VS_Conditions	-0.0052	0.0499	0.0094	0.1389	-0.3045	0.0497	0.2515	0.2792	0.0487	-0.0708	0.04	0.1063
Gender_F	0.9966	-0.0659	0.0061	-0.0112	0.0117	0.0047	0.0056	0.0323	0.0273	-0.0084	0.0093	0.0094
Cost_Inpatient_Y1	0.0063	0.1669	0.1881	0.1628	0.1503	0.1426	0.0038	-0.0149	0.0251	-0.083	-0.03	0.2247
No_DRG_Conditions	0.0056	0.0638	0.0124	0.052	-0.0854	0.0209	0.069	-0.0221	-0.0328	0.2453	-0.1286	0.2107
Growth	0.0021	0.0405	0.1606	0.0153	0.0351	0.1928	0.0045	-0.0429	0.0058	-0.1236	-0.1836	0.1184
VS - Systemic Arterial Hypertension	-0.0034	0.0255	0.0034	0.0722	-0.1656	0.0255	0.1235	0.1519	0.0254	-0.0482	0.0207	0.0355
VS - Heart Diseases	-0.0091	0.0226	-0.0015	0.057	-0.11	0.0082	0.0867	0.1149	0.0262	-0.0022	0.0405	0.2468
ER_Y1	0.0067	0.027	0.0064	-0.0182	-0.0058	-0.0031	0.0528	-0.0404	-0.0718	0.1499	-0.1245	-0.0637
VS - Dyslipidemia	-0.0015	0.0154	0.0012	0.0488	-0.1258	0.0168	0.0961	0.1159	0.0122	-0.0282	0.0166	0.056
VS - Malignant neoplasia	0.0015	0.015	0.0008	0.0368	-0.063	0.0117	0.0645	0.0486	0.0218	0.0042	0.0171	0.0651
DRG - Primary hypertension	0.0009	0.0136	-0.0004	0.0096	-0.0415	0.0041	0.0157	-0.0051	-0.0203	0.0874	-0.0525	0.0657
VS - Diabetes mellitus	-0.0019	0.0121	0.0043	0.0393	-0.0918	0.0167	0.0767	0.0846	0.0102	-0.0288	0.0116	0.0274
ER_Y0	0.0048	0.0192	-0.0152	-0.0066	-0.0048	0.0066	0.0296	-0.0188	-0.045	0.1204	-0.0745	-0.0446
Non-ICU_Y1	0.0012	0.0249	0.0191	0.0396	0.0352	0.0093	0.0141	-0.0234	0.0052	-0.0289	0.0231	-0.0073
Non-ICU_Y0	0.0006	0.0162	-0.0056	0.0304	0.0237	-0.0209	0.0113	-0.0117	0.0066	-0.0047	0.0349	-0.0156

VS - Parkinson's Disease	-0.0	0.0035	0.0012	0.0097	-0.0094	0.0026	0.0092	0.0028	0.0066	-0.0033	0.0091	0.0062
DRG - Asthma	0.0009	0.0038	0.0012	-0.0006	-0.004	0.0023	0.0096	0.0041	-0.007	0.0124	-0.0119	0.0278
VS - Alzheimer's disease	0.0011	0.0047	0.0017	0.0161	-0.0155	0.0038	0.0106	0.0036	0.0138	-0.0061	0.0103	0.0176
DRG - Other neoplasms	-0.0	0.0079	0.0002	0.0083	-0.0024	0.0016	0.0099	-0.0087	0.0052	0.0266	-0.0056	0.0184
DRG - Malignant neoplasm of the pyriform sinus	0.0007	0.003	-0.0012	0.003	-0.0048	0.0013	0.0057	0.0001	-0.0146	0.0371	-0.0292	0.0015
DRG - Acute bronchitis	-0.0002	0.0038	0.0007	0.002	-0.0038	0.0001	0.0088	0.0036	-0.0021	0.0142	-0.0075	0.0231
DRG - Renal failure	-0.0002	0.0064	0.0024	0.0101	-0.0052	0.0038	0.0118	-0.0068	0.0125	0.0106	0.0024	0.0115
ICU_Y0	0.0002	0.0052	-0.0012	0.011	0.0084	-0.0049	0.0054	-0.0064	0.0058	-0.0001	0.0205	-0.0087
VS - Asthma	0.0008	0.0025	-0.0019	0.009	-0.0243	0.0033	0.0245	0.028	0.0032	-0.0042	0.001	0.0052
VS - Stroke	0.0002	0.006	0.0018	0.0124	-0.0163	0.0034	0.0135	0.0147	0.0053	-0.0033	0.0108	0.0255
VS - Chronic Obstructive Pulmonary Disease	-0.0003	0.005	0.0038	0.0117	-0.0212	0.0052	0.0196	0.0174	0.0063	-0.0015	-0.0008	0.0234
Cost_Trend	0.0001	0.003	0.006	0.0055	0.003	0.0097	0.0049	-0.0147	0.0173	-0.0093	-0.0077	-0.0205
VS - Chronic Renal Failure	-0.0007	0.0035	0.0004	0.0073	-0.007	0.0034	0.0116	0.0007	0.0107	0.0018	0.0101	0.0084
Growth_Non-ICU	0.0004	0.005	0.0163	0.0046	0.0064	0.0205	0.0013	-0.0071	-0.0012	-0.0155	-0.0092	0.0061
DRG - Prostate cancer	-0.0011	0.0014	0.0	0.002	-0.0033	0.0002	0.0013	-0.0	0.0	0.0051	-0.0034	0.0045
DRG - Breast cancer	0.0018	0.0019	-0.0003	0.003	-0.005	0.0009	0.0063	-0.0008	0.0012	0.0136	-0.0013	0.005
DRG - Colon neoplasia	0.0001	0.0012	-0.0005	0.003	-0.0025	0.0003	0.0016	-0.0022	0.0007	0.0076	-0.0001	0.0113
DRG - Neopl malig of bronchi and lungs	0.0002	0.002	0.0012	0.0036	-0.0019	0.0028	0.0054	-0.0053	0.0047	0.0098	-0.0052	0.004
DRG - Neoplasia of the utero	0.0005	0.001	0.0002	0.0007	-0.0001	0.0003	0.0008	-0.0012	0.001	0.0049	-0.0011	0.0003
ICU_Y1	0.0003	0.0048	0.0046	0.0058	0.0044	0.0044	0.0018	-0.0053	0.0066	-0.0051	-0.0005	0.0063
DRG - Neopl malig of pancreas	0.0004	0.0014	-0.0011	0.002	-0.0002	-0.001	0.0024	-0.0031	0.0038	0.0069	0.0033	0.0015
DRG - Encephalo neopl malig	-0.0001	0.0011	0.0004	0.0016	0.0005	0.0008	0.0024	-0.0019	0.0024	0.0011	0.0006	0.0012
DRG - Neopl utero neck malig	0.0005	0.0006	-0.0002	0.0011	-0.0004	-0.0002	0.0017	-0.0013	0.0021	0.0034	-0.0005	0.0012
DRG - Neopl malig of the larynx	-0.0001	0.0005	0.0006	0.0004	-0.0003	0.0003	0.0003	-0.0	-0.0006	0.0019	-0.0009	0.001
DRG - Malignant neoplasm of the skin	-0.0003	0.0025	-0.0002	0.0027	-0.0059	0.0003	0.0012	0.0027	-0.0016	0.0006	-0.0019	0.0128
DRG - Neopl bladder malig	-0.0004	0.0012	0.0007	0.0011	-0.0023	0.0001	0.0008	0.0018	-0.0013	0.0034	-0.0035	0.0034
DRG - Malignant neoplasm of the thyroid gland	0.0006	0.0008	-0.0002	0.0011	-0.0014	-0.0005	0.0027	-0.0001	-0.0026	0.0066	-0.0044	0.0031
DRG - Pneumonia	0.0003	0.0054	0.0005	0.0039	-0.0024	-0.0001	0.0036	-0.002	0.0006	0.0108	-0.0063	0.0123
DRG - Malignant neoplasm of the liver	0.0	0.0005	0.0001	0.0005	-0.0006	0.0005	0.001	-0.0011	0.0012	0.0017	-0.0001	0.0012
DRG - Malignant neoplasm of the ovary	0.0005	0.0009	-0.0007	0.0016	-0.0005	-0.0004	0.002	-0.0021	0.0015	0.0067	-0.0004	0.0024
DRG - Malignant neoplasm of the rehest	0.0003	0.0013	-0.0	0.0017	-0.0007	0.0005	0.0021	-0.0015	0.0002	0.008	-0.0001	0.0058
DRG - Malignant neoplasm of the kidney	0.0001	0.0003	-0.0	0.0006	-0.0008	0.0001	0.0009	-0.0006	0.0007	0.0006	-0.0004	-0.0013
DRG – Malignancy, no location specification	-0.0	0.0017	0.0004	0.002	-0.0022	0.0005	0.0007	-0.0015	0.0012	0.0022	0.0	0.0052
DRG - Melanocytic nevus	0.0002	0.0008	0.0012	0.0012	-0.0014	0.0012	0.0018	-0.0014	-0.0023	0.0042	0.0012	0.0059
DRG - Obesity	0.0006	0.0004	0.0001	0.0004	-0.0013	0.0004	0.002	0.0005	-0.0013	0.0029	-0.0012	0.0059
DRG - Osteoporosis	0.0017	0.0011	-0.0003	0.0022	-0.0063	0.0012	0.002	-0.0017	-0.0017	0.013	-0.0069	0.0079
DRG - Melanoma	-0.0	0.0006	0.0	0.0003	-0.0006	-0.0003	-0.0002	-0.0006	-0.0	0.0027	-0.0009	0.0007
DRG - Chronic sinusitis	0.0001	0.001	0.0	0.0001	-0.001	0.0001	0.0016	-0.0	-0.0028	0.0061	-0.0044	0.0037
DRG - Thyroiditis	0.0003	0.0003	0.0001	0.0001	-0.0005	-0.0003	0.0003	-0.0001	-0.0007	0.0027	-0.001	0.0015
DRG - Thyroid disorders	0.0005	0.0001	-0.0001	0.0001	-0.0015	0.0002	0.0016	0.0005	-0.0002	0.0033	-0.0022	0.0026
DRG - Varicose vein	0.0012	0.0019	-0.0003	0.0009	-0.0019	0.0006	0.0004	0.0007	-0.003	0.005	-0.0061	0.0067

Growth_ICU	0.0001	-0.0003	0.0044	-0.0039	-0.0029	0.0069	-0.0027	0.0008	0.0006	-0.0037	-0.0156	0.0111
DRG - Myeloma	0.0007	0.002	0.0012	0.0032	-0.0019	0.0018	0.0048	-0.0023	0.0015	0.011	-0.0055	0.0085
DRG - Hypertension	-0.0001	0.0005	-0.0001	0.0003	-0.0013	0.0001	0.0003	0.0005	-0.0004	-0.0	-0.0004	0.0011
DRG - Lymphoma	0.0	0.0013	0.0005	0.0018	-0.0011	-0.0	0.0019	-0.0019	0.0005	0.0066	-0.0008	0.0018
DRG - Atherosclerosis	0.0001	0.0021	0.0002	0.001	-0.0056	0.0013	0.0025	-0.0004	-0.0013	0.0132	-0.0071	0.0092
DRG - Dementia	0.0004	0.0007	0.0003	0.0012	-0.0007	0.0006	-0.0005	-0.0008	-0.0	-0.0005	-0.0028	0.0022
DRG - Liver cirrhosis	-0.0001	0.0006	0.0001	0.0007	-0.0007	-0.0004	0.0005	0.0005	-0.0001	0.0031	-0.0009	0.0016
DRG - Cardiopathy	-0.0004	0.0041	-0.0	0.0043	-0.0039	-0.0008	0.0023	0.0009	-0.0002	0.0111	-0.0026	0.0172
DRG - Carcinoma	0.0002	0.0006	-0.0005	0.001	-0.0014	-0.0002	0.0016	-0.0009	0.0013	0.0047	0.0017	0.0017
DRG - Calculosis of kidney and/or ureter	-0.0006	0.0026	-0.001	0.0012	-0.0008	-0.0	0.0024	0.0014	-0.0036	0.0069	-0.004	0.0067
DRG - Goitre	0.0007	0.0011	-0.0005	0.0009	-0.0028	0.0001	0.0023	-0.001	-0.003	0.01	-0.006	0.0054
DRG - Stroke	0.0001	0.0013	0.0003	0.0019	-0.0009	0.0	0.0006	-0.0002	-0.0001	0.0017	-0.0003	0.003
DRG - Rheumatoid arthritis	0.0002	0.0007	0.0006	0.0005	-0.0	0.0007	0.0009	-0.0008	0.001	0.0005	0.0004	0.0026
DRG - Leukaemia	0.0001	0.0009	0.0001	0.0015	-0.0001	0.0001	0.0019	-0.0016	0.0012	0.0034	0.0007	0.0006
DRG - Arrhythmia	0.0	0.0017	0.0001	0.0017	-0.0017	0.0006	0.0006	-0.0006	0.0008	0.0049	-0.0014	0.004
DRG - Angina	-0.0003	0.0025	-0.0004	0.0033	-0.0046	-0.0017	0.0009	0.0014	-0.0012	0.011	-0.0034	0.012
DRG - Anaemia	0.0008	0.0015	-0.0	0.0021	-0.0008	-0.0002	0.0037	-0.0008	-0.001	0.0081	-0.0023	0.0071
DRG - Alzheimer 's	0.0001	0.0006	0.0001	0.0011	-0.0005	0.0003	0.0002	-0.0013	0.0003	-0.0002	0.0013	0.0013
VS - Transplantation	-0.0002	0.0005	0.0003	0.0013	-0.0014	0.001	0.0022	0.0006	0.0005	-0.0014	-0.0014	0.0013
VS - Hepatitis C	-0.0001	0.001	0.0005	0.0017	-0.0031	0.0013	0.0033	0.0023	0.0004	-0.0016	0.0001	0.0016
VS - HIV/AIDS	-0.0002	0.0005	-0.0004	0.0008	-0.0005	-0.0001	0.0008	0.0003	0.0006	-0.0006	0.0017	0.0006
DRG - Diabetes	0.0002	0.0021	-0.0006	0.0019	-0.0076	0.0007	0.003	0.0023	-0.0011	0.0056	-0.0	0.012
DRG - Dyslipidemia	0.0003	0.0022	-0.0018	0.0017	-0.0091	0.0011	0.0013	-0.0018	-0.0037	0.0127	-0.0005	0.0199
DRG - Parkinson's disease	-0.0001	0.0017	0.0004	0.0024	-0.0004	0.0009	0.0018	-0.0025	0.0024	0.0002	0.002	0.0004
DRG - Cerebrovascular diseases	-0.0	0.0016	-0.0002	0.002	-0.0012	0.0003	0.0009	-0.0001	-0.0007	0.0032	-0.0018	0.0053
DRG - Transient cerebral ischemia	0.0	0.0013	0.0006	0.0012	-0.0018	0.0007	0.0006	0.0021	-0.0011	0.0028	-0.0028	0.0058
DRG - Heart failure	-0.0001	0.0016	-0.0006	0.0021	-0.0016	-0.0004	0.0006	0.0009	0.001	0.0026	0.0006	0.0088
DRG - Infarction	-0.0003	0.0019	0.0019	0.002	-0.0005	0.0009	-0.0	0.0009	-0.001	0.0019	-0.0024	0.0086
DRG - Hypothyroidism	0.001	0.0007	0.0004	0.0009	-0.0036	0.0007	0.0021	0.0004	-0.0024	0.0038	-0.002	0.0076
DRG - Prostate hyperplasia	-0.0025	0.0021	0.0004	0.0021	-0.0051	0.001	0.0001	-0.0001	-0.0031	0.0057	-0.0051	0.0075
DRG - Hydrocephalus	0.0002	0.0009	0.0005	0.0013	0.0002	0.0006	0.0008	-0.0012	0.0019	0.0013	-0.0011	0.0016
DRG - Hepatitis c	-0.0003	0.0001	-0.0002	0.0002	-0.0003	-0.0003	0.0002	-0.0005	0.0002	0.0009	0.0008	0.0007
DRG - Schizophrenia	-0.0	0.0012	0.0002	0.0023	0.0024	-0.0012	0.0013	0.0	-0.0024	-0.0027	0.001	-0.0016
DRG - Hooking spondylitis	-0.0	0.0003	0.0002	0.0006	-0.0005	0.0	0.0016	-0.002	0.001	0.0022	0.0011	0.0012
DRG - Multiple sclerosis	0.0003	0.0003	-0.0001	-0.0001	0.0002	-0.0	0.0001	0.0002	-0.0008	0.0008	-0.0002	0.0004
DRG - Severe depressive episode	0.0001	0.0002	0.0001	-0.0002	0.0002	0.0005	0.0003	0.0001	-0.0007	0.0007	-0.0001	0.0002
DRG - Epilepsy	-0.0001	0.0009	-0.0006	0.0002	-0.0006	-0.0002	0.0012	-0.0012	0.0	0.0036	-0.002	0.0036
DRG - Emphysema	-0.0003	0.0005	0.0008	0.0007	-0.0007	0.0011	0.0004	-0.0	0.0001	0.0007	-0.0015	0.0038
DRG - Chronic obstructive pulmonary disease	0.0001	0.0021	0.0007	0.0025	-0.001	0.0012	0.0011	-0.0033	0.0014	0.0051	-0.0013	0.0047
DRG - Diseases circulatory system	0.0003	0.0033	-0.0	0.0031	-0.0037	0.001	0.0042	-0.0005	0.0011	0.0118	-0.0034	0.0079
DRG - Pulmonary embolism	0.0001	0.0006	-0.0002	0.0006	-0.0005	-0.0001	0.0007	-0.0004	-0.0006	0.003	-0.0015	0.0002

12.3. Models' Best Features

12.3.1. Years 0 and 1/Year 2 Top 1% Simple High-Cost

Sociodemographics	Costs	Utilization	Clinical
Age	Cost_Y0, Cost_Y1, Cost_1M, Cost_2M, Cost_6M, Acute, Cost_Trend, Cost_Hospitals_Y0, Cost_Hospitals_Y1, Cost_Inpatient_Y1, Cost Growth, %Hospitals_Y0, %Hospitals_Y1, %Inpatient_Y1	Visits_Y0, Non-ICU_Y1, Growth_Non-ICU, Growth_ICU	VS - Malignant neoplasia

12.3.2. Years 0 and 1/Year 2 Top 2% Simple High-Cost

Sociodemographics	Costs	Utilization	Clinical
Age	Cost_Y0, Cost_Y1, Cost_1M, Cost_2M, Cost_6M, Acute, Cost_Trend, Cost_Hospitals_Y0, Cost_Hospitals_Y1, Cost_Inpatient_Y0, Cost_Inpatient_Y1, Cost Growth, %Hospitals_Y0, %Hospitals_Y1, %Inpatient_Y1	Visits_Y0, Visits_Y1,Non-ICU_Y0,Non-ICU_Y1,Growth_Non-ICU,Growth_ICU	X

12.3.3. Years 0 and 1/Year 2 Top 5% Simple High-Cost

Sociodemographics	Costs	Utilization	Clinical
Age	Cost_Y0, Cost_Y1, Cost_1M, Cost_2M, Cost_6M, Acute, Cost_Trend, Cost_Hospitals_Y0, Cost_Hospitals_Y1, Cost_Inpatient_Y0, Cost Growth, %Hospitals_Y1, %Inpatient_Y0, %Inpatient_Y1	Visits_Y0, Visits_Y1, Growth_Non-ICU,	No_Conditions_Total

12.3.4. Years 0 and 1/Year 2 Top 10% Simple High-Cost

Sociodemographics	Costs	Utilization	Clinical
Age	Cost_Y0, Cost_Y1, Cost_1M, Cost_2M, Cost_6M, Acute, Cost_Trend, Cost_Hospitals_Y0, Cost_Hospitals_Y1, Cost_Inpatient_Y0, Cost_Inpatient_Y1, Cost Growth, %Hospitals_Y0, %Hospitals_Y1, %Inpatient_Y1	Visits_Y0, Visits_Y1, Non-ICU_Y0,Non-ICU_Y1,Growth_Non-ICU	No_Conditions_Total

12.3.5. Years 0 and 1/Year 2 Top 0.5% Bloomers

Sociodemographics	Costs	Utilization	Clinical
Age	Cost_Y0, Cost_Y1, Cost_1M, Cost_2M, Cost_6M, Acute, Cost_Trend, Cost_Hospitals_Y0, Cost_Hospitals_Y1, Cost Growth, %Hospitals_Y0, %Inpatient_Y0, %Inpatient_Y1	Visits_Y0, Visits_Y1, Growth_ICU, Growth_Non-ICU	VS - HIV/AIDS, DRG - Renal failure, DRG - Melanoma, DRG - Prostate hyperplasia

12.3.6. Years 0 and 1/Year 2 Top 1% Bloomers

Sociodemographics	Costs	Utilization	Clinical
Age	Cost_Y0, Cost_Y1, Cost_1M, Cost_2M, Cost_6M, Acute, Cost_Trend, Cost_Hospitals_Y0, Cost_Inpatient_Y1, Cost Growth, %Hospitals_Y0, %Hospitals_Y1, %Inpatient_Y0, %Inpatient_Y1	Visits_Y1, Growth_ICU, Growth_Non-ICU	VS - Malignant neoplasia, No_Conditions_Total

12.3.7. Years 0 and 1/Year 2 Top 2% Bloomers

Sociodemographics	Costs	Utilization	Clinical
Age	Cost_Y0, Cost_Y1, Cost_1M, Cost_2M, Cost_6M, Acute, Cost_Trend, Cost_Hospitals_Y0, Cost_Hospitals_Y1, Cost Growth, %Hospitals_Y0, %Hospitals_Y1, %Inpatient_Y0, %Inpatient_Y1	Visits_Y0, Visits_Y1, Growth_ICU, Growth_Non-ICU	VS – Chronic Renal failure, No_Conditions_Total

12.3.8. Years 0 and 1/Year 2 Top 5% Bloomers

Sociodemographics	Costs	Utilization	Clinical

Age	Cost_Y0, Cost_Y1, Cost_1M, Cost_2M, Cost_6M, Acute, Cost_Trend, Cost_Hospitals_Y0, Cost_Hospitals_Y1, Cost_Inpatient_Y0, Cost Growth, %Hospitals_Y0, %Hospitals_Y1, %Inpatient_Y0	Visits_Y0, Growth_ICU, Growth_Non-ICU	VS – Chronic Renal failure, No_Conditions_Total

12.3.9. Years 0 and 1/Year 2 Top 10% Bloomers

Sociodemographics	Costs	Utilization	Clinical
Age	Cost_Y0, Cost_Y1, Cost_1M, Cost_2M, Cost_6M, Acute, Cost_Trend, Cost_Hospitals_Y0, Cost_Hospitals_Y1, Cost Growth, %Hospitals_Y0, %Hospitals_Y1, %Inpatient_Y0	Visits_Y0, Visits_Y1, Non-ICU_Y0, Growth_Non-ICU	No_Conditions_Total

12.3.10. Year 0/Year 2 Top 0.5% Simple High-Cost

Sociodemographics	Costs	Utilization	Clinical
Age	Cost_Y0, Cost_Y1, Cost_1M, Cost_2M, Cost_6M, Acute, Cost_Trend, Cost_Hospitals_Y0, Cost_Inpatient_Y0, %Hospitals_Y0, %Hospitals_Y1, %Inpatient_Y0	Visits_Y0, Non-ICU_Y0, ICU_Y0	DRG - Renal failure, No_Conditions_Total

12.3.11. Year 0/Year 2 Top 1% Simple High-Cost

Sociodemographics	Costs	Utilization	Clinical
Age	Cost_Y0, Cost_Y1, Cost_1M, Cost_2M, Cost_6M, Acute, Cost_Trend, Cost_Hospitals_Y0, Cost_Inpatient_Y0, %Hospitals_Y0, %Inpatient_Y0	Visits_Y0, Non-ICU_Y0, ICU_Y0	DRG - Multiple sclerosis, No_VS_Conditions, No_DRG_Conditions

12.3.12. Year 0/Year 2 Top 2% Simple High-Cost

Sociodemographics	Costs	Utilization	Clinical
Age	Cost_Y0, Cost_Y1, Cost_1M, Cost_2M, Cost_6M, Acute, Cost_Trend, Cost_Hospitals_Y0, Cost_Inpatient_Y0, %Hospitals_Y0, %Inpatient_Y0	Visits_Y0, Non-ICU_Y0, ICU_Y0	No_Conditions_Total

12.3.13. Year 0/Year 2 Top 5% Simple High-Cost

Sociodemographics	Costs	Utilization	Clinical
Age	Cost_Y0, Cost_Y1, Cost_1M, Cost_2M, Cost_6M, Acute, Cost_Trend, Cost_Hospitals_Y0, Cost_Inpatient_Y0, %Hospitals_Y0, %Inpatient_Y0	Visits_Y0, Non-ICU_Y0, ER_Y0	No_Conditions_Total

12.3.14. Year 0/Year 2 Top 10% Simple High-Cost

Sociodemographics	Costs	Utilization	Clinical
Age	Cost_Y0, Cost_Y1, Cost_1M, Cost_2M, Cost_6M, Acute, Cost_Trend, Cost_Hospitals_Y0, Cost_Inpatient_Y0, %Hospitals_Y0	Visits_Y0, Non-ICU_Y0	No_Conditions_Total

12.3.15. Year 0/Year 2 Top 0.5% Bloomers

Sociodemographics	Costs	Utilization	Clinical
Age	Cost_Y0, Cost_Y1, Cost_1M, Cost_2M, Cost_6M, Acute, Cost_Trend, Cost_Hospitals_Y0, Cost_Inpatient_Y0, %Hospitals_Y0, %Inpatient_Y0	ICU_Y0, Non-ICU_Y0	VS - Alzheimers disease, DRG - Renal failure, No_Conditions_Total

12.3.16. Year 0/Year 2 Top 1% Bloomers

Sociodemographics	Costs	Utilization	Clinical
Age	Cost_Y0, Cost_Y1, Cost_1M, Cost_2M, Cost_6M, Acute, Cost_Trend, Cost_Hospitals_Y0, Cost_Inpatient_Y0, %Hospitals_Y0, %Inpatient_Y0	Visits_Y0, Non-ICU_Y0	VS - Alzheimers disease, DRG - Multiple sclerosis, No_VS_Conditions

12.3.17. Year 0/Year 2 Top 2% Bloomers

Sociodemographics	Costs	Utilization	Clinical

Age	Cost_Y0, Cost_Y1, Cost_1M, Cost_2M, Cost_6M, Acute, Cost_Trend, Cost_Hospitals_Y0, Cost_Inpatient_Y0, %Hospitals_Y0, %Inpatient_Y0	Visits_Y0, ICU_Y0, Non-ICU_Y0, ER_Y0	VS - Alzheimers disease, No_Conditions_Total

12.3.18. Year 0/Year 2 Top 5% Bloomers

Sociodemographics	Costs	Utilization	Clinical
Gender and Age	Cost_Y0, Cost_Y1, Cost_1M, Cost_2M, Cost_6M, Acute, Cost_Trend, Cost_Hospitals_Y0, Cost_Inpatient_Y0, %Hospitals_Y0, %Inpatient_Y0	Visits_Y0, Non-ICU_Y0	VS - Diabetes mellitus, No_Conditions_Total

12.3.19. Year 0/Year 2 Top 10% Bloomers

Sociodemographics	Costs	Utilization	Clinical
Age	Cost_Y0, Cost_Y1, Cost_1M, Cost_2M, Cost_6M, Acute, Cost_Trend, Cost_Hospitals_Y0, %Hospitals_Y0, %Inpatient_Y0	Visits_Y0	VS - Systemic Arterial Hypertension, No_Conditions_Total

12.4. Models' Detailed Results

12.4.1. Years 0 and 1/Year 2 Simple High-Cost Classification

12.4.1.1. Top 0.5% High-Cost Users

The engineered stacking model using a multi-layer perceptron as meta-model could capture more than 55% of the cost of the top 0.5% high-cost users in 2019 using predictors from 2017 and 2018, while reaching a ranking based precision of almost 41%. This model also presented the greater areas under ROC and Precision-Recall curves. A Random Forest with cost features and a MLP with the best features presented on Table 35 also showed good performances. Figures 28 and 29 show how the best model has a better Precision-Recall curve than the others, presenting better metrics for many possible thresholds.

These results were pretty impressive when compared to 24% and 30% of costs captured by Meenan et at (2003) and by Moturu et al (2010), respectively (although Moturu used a classification threshold of US$ 50k, approximately top .69% ranked). The best area under the ROC curve of .93 was also considerably higher than the .83 presented by the former.

The total costs of these top .5% ranked users in 2019 (152 enrollees) were around 24% of the total expenses, so capturing 55% of them would mean focusing on 152 users that had expenses of around 13% of the total costs of the Health Program in 2019. If preventive and primary care actions focused on these users were implemented, they could improve their medical outcomes and quality of life, while, if hypothetically decreasing this number by just 8%, it would already represent saving more than 1% of the total expenses in 2019.

As it was already explained in section 3.9, two precisions were calculated, a ranking-based one, considering the real number of top .5% users and a calculated one, using the calculated probability threshold of 14% (see section 3.3.1). The former was also used to calculate the recall and the engineered stacking method reached a .61 value for this metric, against .18 and .28 of the two studies cited before, a number that was also considered very good for the purpose of this study.

The baseline model could capture more than 46% of the total cost with an AUROC of .92 and a ranking-based precision of .35. These results were also better than the ones found in the literature, despite the

fact that this model used only two predictors. The algorithm selection and tuning methods during the validation phase may be responsible for part of this good performance, but most of it, undoubtfully, is due to the predictive power of the two features used, age and last year's total costs (mainly the former).

Table 35: Top 0.5% Simple High-Cost Test Metrics

	Precision_Rank	Precision	AUC_ROC	Recall	AUC_PR	Cost_Capture
Baseline_Logit	0.349	0.204	0.920	0.474	0.240	0.464
Cost_RF	0.382	0.325	0.866	0.454	0.262	0.535
Complete_RF	0.362	0.354	0.906	0.375	0.249	0.508
Best_MLP	0.395	0.187	0.920	0.533	0.250	0.515
PCA_MLP	0.375	0.151	0.916	0.559	0.272	0.476
Stack_Smote_MLP	0.408	0.151	0.929	0.612	0.292	0.553

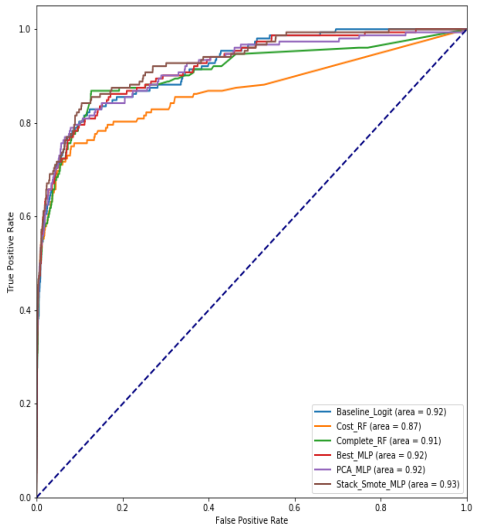


Figure 26: Top 0.5% Simple High-Cost ROC Curve

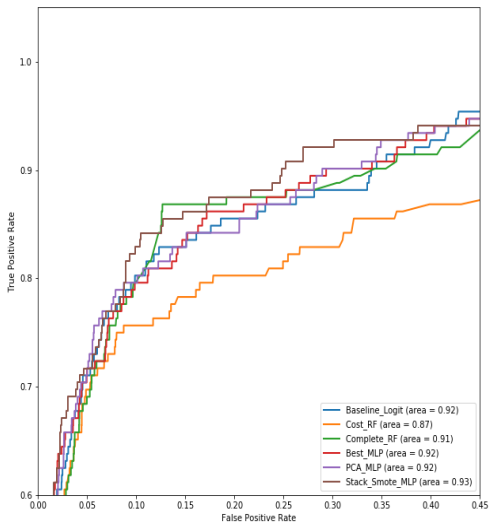


Figure 27: Top 0.5% Simple High-Cost Zoomed ROC Curve

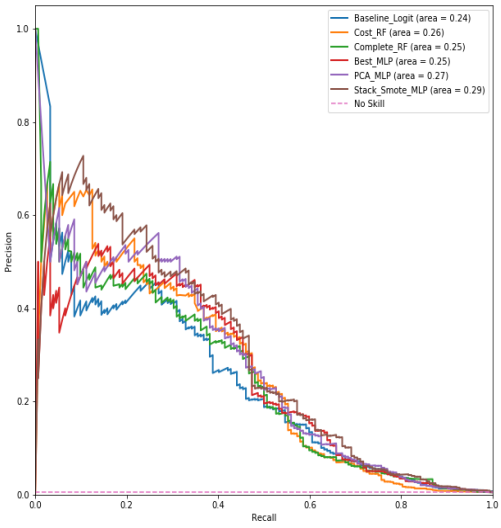


Figure 28: Top 0.5% Simple High-Cost PR Curve

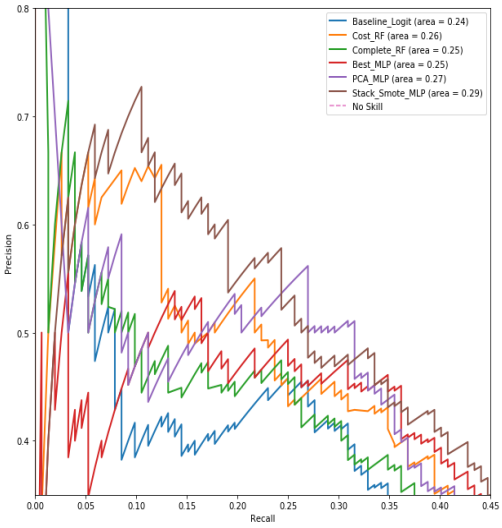


Figure 29: Top 0.5% Simple High-Cost Zoomed PR Curve

12.4.1.2. Top 1% High-Cost Users

Similar results were found for the top 1% high-cost users classification. Once again, the engineered stacking classifier presented the highest ranking-based precision and areas under the ROC and Precision-Recall curves, while an only cost features based Random Forest reached the best cost capture.1

The total expenses of the top 1% high-cost enrollees were 34% of the total, so the 304 instances classified as high-cost represented 18.8% of the total program's expenses. Again, doing a hypothetical exercise, if the primary care implemented achieved a decrease of 8% in total costs of this group, it would mean saving 1.5% of the resources spent, something extremely important in a moment of growing costs and fiscal constraints.

Although not capturing more than 5.5% of the costs and correctly classifying no more than 3% of instances than the baseline model, the best one achieved an area under the Precision Recall curve of .36, considerably greater than the baseline's .22, as shown in Figure 32.

Table 36: Top 1% Simple High-Cost Test Metrics

	Precision_Rank	Precision	AUC_ROC	Recall	AUC_PR	Cost_Capture
Baseline_RF	0.382	0.373	0.869	0.414	0.224	0.500
Cost_RF	0.411	0.399	0.861	0.418	0.351	0.557
Complete_MLP	0.388	0.247	0.891	0.477	0.296	0.512
Best_MLP	0.405	0.267	0.920	0.487	0.348	0.535
PCA_MLP	0.405	0.239	0.919	0.507	0.322	0.525
Stack_Smote_MLP	0.418	0.228	0.923	0.536	0.364	0.555

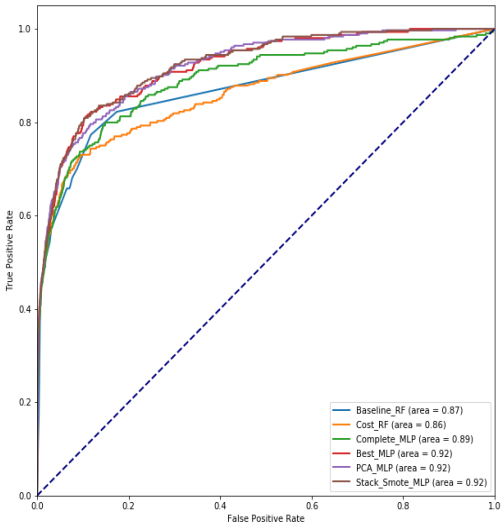


Figure 30: Top 1% Simple High-Cost ROC Curve

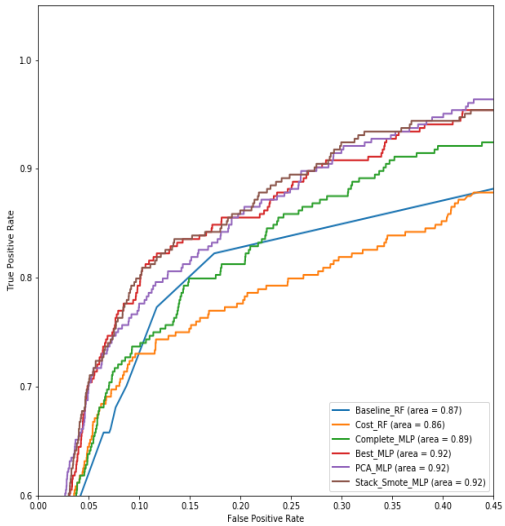


Figure 31: Top 1% Simple High-Cost Zoomed ROC Curve

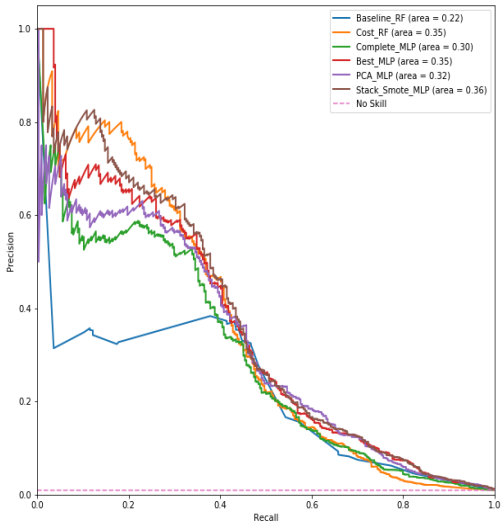


Figure 32: Top 1% Simple High-Cost PR Curve

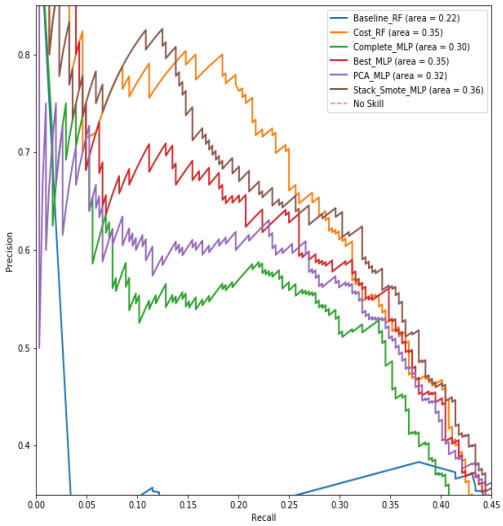


Figure 33: Top 1% Simple High-Cost Zoomed PR Curve

12.4.1.3. Top 2% High-Cost Users

Once again, results were similar to the prior ones, except for the baseline model's, which achieved the highest rank-based precision and the same cost capture as other models (being the area under the Precision-Recall curve the only performance metric which it still was considerably worse than the others).

The costs correctly captured accounted for 24.7% of the total PASBC's expenses. Despite consistently having a lower AUROC (Figure 34), the only costs-based model could classify a third of the 609 top 2% high-cost enrollees with a precision close to .9 (Figure 37), what could bring good results for a more focused preventive care program, targeted in fewer users.

Table 37: Top 2% Simple High-Cost Test Metrics

	Precision_Rank	Precision	AUC_ROC	Recall	AUC_PR	Cost_Capture
Baseline_RF	0.432	0.317	0.894	0.476	0.318	0.546
Cost_RF	0.424	0.438	0.821	0.414	0.412	0.542
Complete_MLP	0.430	0.370	0.903	0.475	0.401	0.551
Best_MLP	0.419	0.392	0.904	0.437	0.403	0.542
PCA_Logit	0.411	0.368	0.901	0.461	0.363	0.540
Stack_Smote_Logit	0.424	0.374	0.905	0.445	0.383	0.544

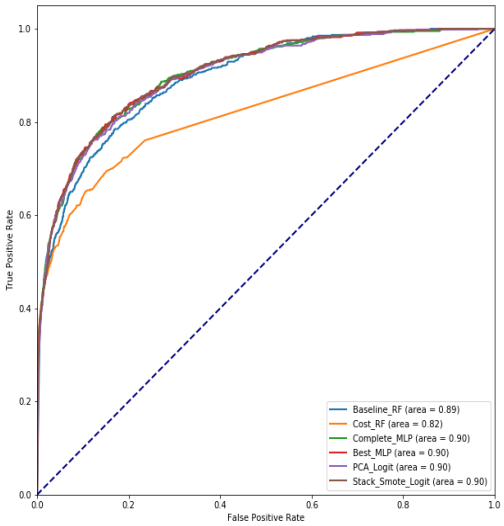


Figure 34: Top 2% Simple High-Cost ROC Curve

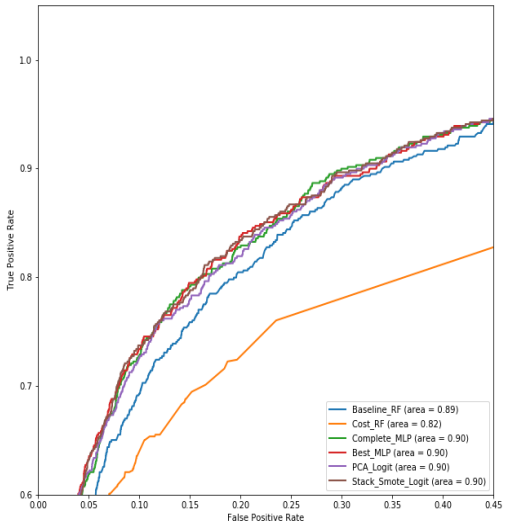


Figure 35: Top 2% Simple High-Cost Zoomed ROC Curve

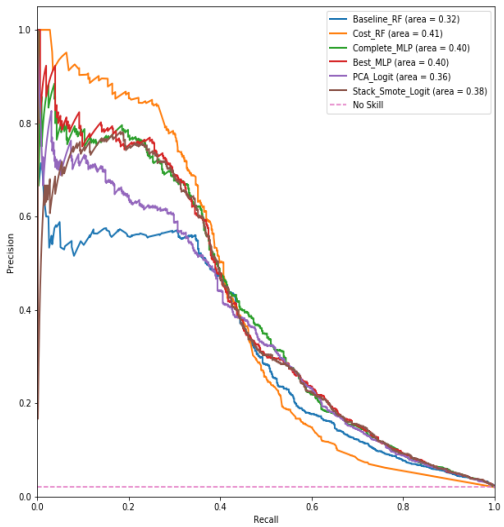


Figure 36: Top 2% Simple High-Cost PR Curve

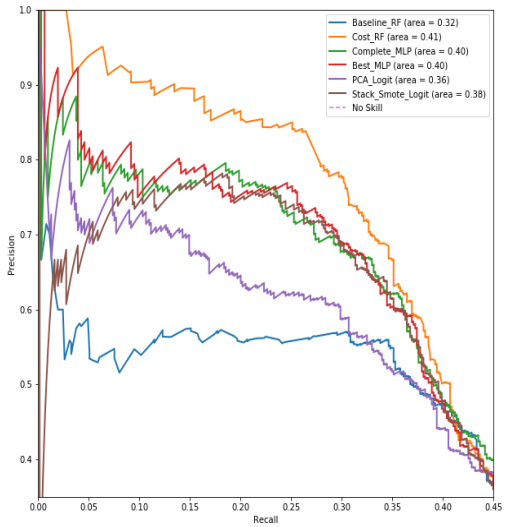


Figure 37: Top 2% Simple High-Cost Zoomed PR Curve

12.4.1.4. Top 5% High-Cost Users

A 12 Principal Components Logistic Regression was able to capture more than 62% of the top 5% high-cost users' total cost in 2019 using predictors from 2017 and 2018. As already said in section 12.4.1.2, baseline model reached very good results. It's true that data has been clipped, scaled and even for the baseline dataset the algorithm selection and tuning was extremely solid. Nonetheless, it was still interesting to see how this simple dataset could reach results almost as good as much more complex ones.

With the results for the top 5% high-cost users classification, a pattern started to become clear: the higher the classification threshold, the better the cost capture. This makes sense as most of costs are usually concentrated in few instances and, although the classifier may not correctly identify the top 0.5%, it probably will predict a higher probability for a lot of them than for regular instances. This way, despite not being classified as a top 0.5% high-cost user, it may be considered a top 5% and, this way, costlier instances will be correctly classified as the classification threshold increases, capturing a larger share of the top ranked high-cost users.

Top 1522 high-cost users in 2019 accounted for 58% of total costs, so capturing 62% of them means classifying as high-risk, instances that represented 36% of PASBC's total expenses. Highest AUROC of .858 was achieved for two models, slightly lower than the .865 area reached by Chechulin et al (2014).

Table 38: Top 5% Simple High-Cost Test Metrics

	Precision_Rank	Precision	AUC_ROC	Recall	AUC_PR	Cost_Capture
Baseline_Logit	0.386	0.391	0.841	0.377	0.350	0.593
Cost_RF	0.399	0.348	0.807	0.432	0.406	0.592
Complete_Logit	0.420	0.340	0.858	0.503	0.405	0.622
Best_RF	0.420	0.340	0.854	0.470	0.432	0.611
PCA_Logit	0.411	0.350	0.853	0.480	0.370	0.623
Stack_Smote_Logit	0.406	0.342	0.858	0.455	0.421	0.592

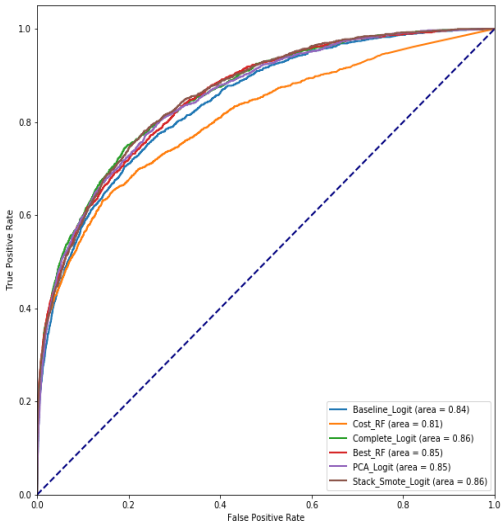


Figure 38: Top 5% Simple High-Cost ROC Curve

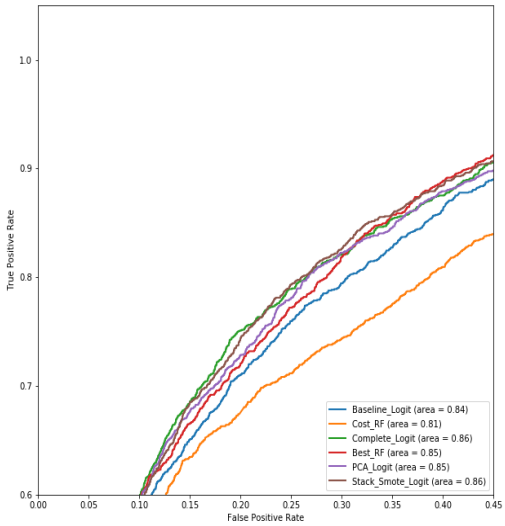


Figure 39: Top 5% Simple High-Cost Zoomed ROC Curve

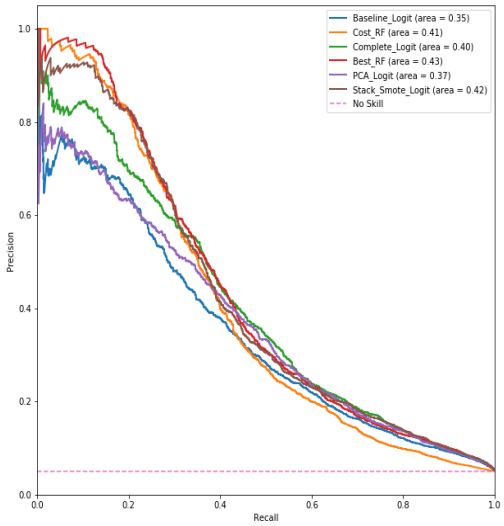


Figure 40: Top 5% Simple High-Cost PR Curve

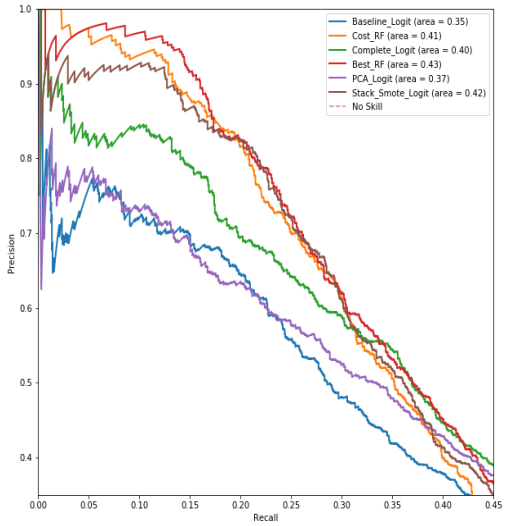


Figure 41: Top 5% Simple High-Cost Zoomed PR Curve

12.4.1.5. Top 10% High-Cost Users

Reaching a cost capture of 67% for the top 10% high-cost users model, the engineered stacking classifier was able to correctly identify 45% of the top 10% users in 2019, according to the ranking

based classification, and an reached an area under the Precision-Recall bigger of .47. This result is better than the 60% cost captured reached by Tamang et al (2016) and very close to the one reached by Kim and Park (2019), that could capture 66% of the top 10% high-cost users' total expenses. These authors reached an AUROC of .843, while the best models tested in this thesis achieved a performance of .825 for the top 10% simple high-cost classification.

When compared to the baseline model, the best one, developed with the engineered stacking method, could capture around 3% more of the top 10% total cost, reaching 46% of PASBC's total expenses.

According to Figure 42, the cost features based Random Forest presented a better area under the curve than for the previous classification thresholds (.5% to 5%), but still had the lowest of all models, while having the best area under the Precision-Recall up to a 25% recall, when it started dropping to have the worst precision for recalls closer to 1 (Figure 44).

Table 39: Top 10% Simple High-Cost Test Metrics

	Precision_Rank	Precision	AUC_ROC	Recall	AUC_PR	Cost_Capture
Baseline_RF	0.416	0.279	0.812	0.632	0.402	0.632
Cost_RF	0.433	0.277	0.797	0.623	0.462	0.647
Complete_RF	0.443	0.274	0.818	0.641	0.459	0.647
Best_Logit	0.437	0.285	0.821	0.650	0.449	0.662
PCA_MLP	0.439	0.270	0.825	0.692	0.438	0.664
Stack_Smote_MLP	0.452	0.308	0.825	0.622	0.471	0.669

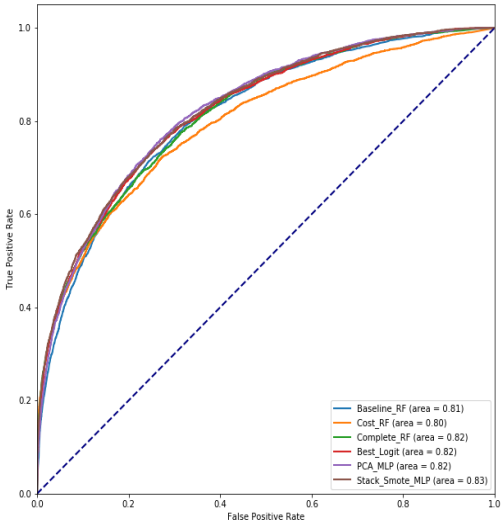


Figure 42: Top 10% Simple High-Cost ROC Curve

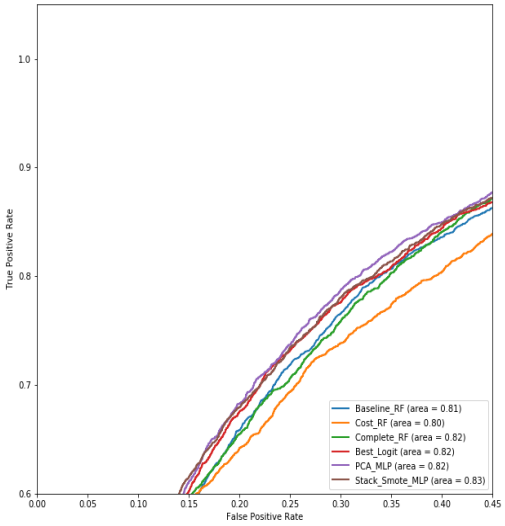


Figure 43: Top 10% Simple High-Cost Zoomed ROC Curv

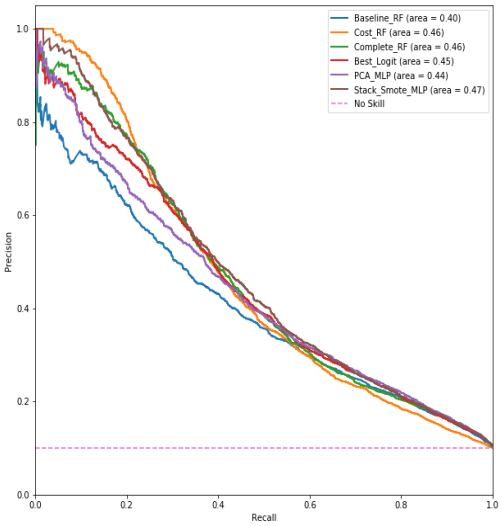


Figure 44: Top 10% Simple High-Cost PR Curve

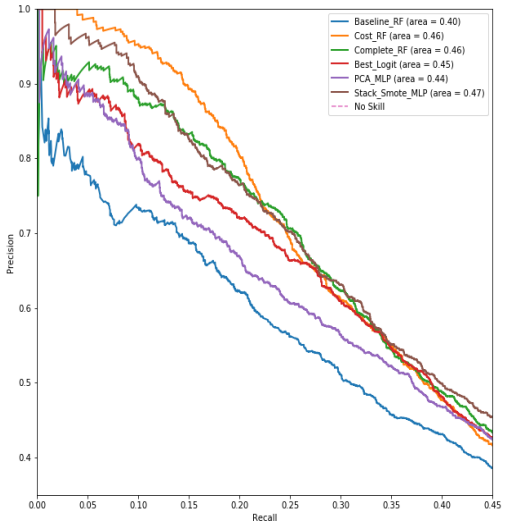


Figure 45: Top 10% Simple High-Cost Zoomed PR Curve

12.4.2. Years 0 and 1/Year 2 High-Cost Bloomers Classification

The second type of models tested was the classification of cost bloomers among the top ranked high-cost users in Year 2, using features from years 0 and 1. As it was already explained in section 3.2, this model tries to identify which instances that weren't high-cost ones in Y1 will become top users in Y2. This approach is relevant because bloomers might be usually neglected by primary care programs, as they are not current high-risk users, so their identification and management may bring good results.

12.4.2.1. Top 0.5% Bloomers

After dropping the 152 highest-cost users in 2018 from the sample, in order to develop the top 0.5% bloomers models, total cost in 2019 decreased 12%, what means that 2018 top .5% accounted for around this share of total expenses in 2019, with 56 of them being also high-cost users in 2019. The other 96 2019 top .5% were cost-bloomers and had total expenses of 15% of total.

A logistic regression using 12 principal components was able to correctly identify 15.6% of the cost bloomers among the top 0.5% high-cost users in 2019, capturing more than 31% of the total expenses of these 96 bloomers among the 152 top 0.5%. This performance is considerably better than the one achieved by the baseline model, that captured 25% of the top 0.5% high-cost bloomers' 15% of total expenses in 2019.

Table 40: Top 0.5% Bloomers Test Metrics

	Precision_Rank	Precision	AUC_ROC	Recall	AUC_PR	Cost_Capture
Baseline_Logit	0.094	0.073	0.874	0.281	0.044	0.252
Cost_Logit	0.135	0.073	0.787	0.354	0.052	0.313
Complete_Logit	0.135	0.077	0.888	0.333	0.059	0.291
Best_Logit	0.135	0.080	0.887	0.344	0.061	0.296
PCA_Logit	0.156	0.073	0.888	0.323	0.061	0.315
Stack_Smote_MLP	0.083	0.084	0.891	0.333	0.052	0.151

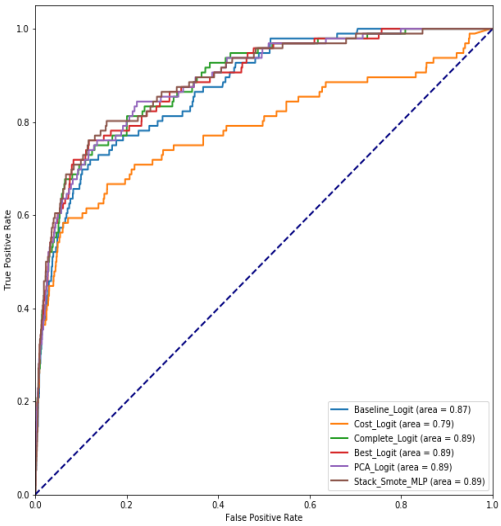


Figure 46: Top 0.5% Bloomer ROC Curve

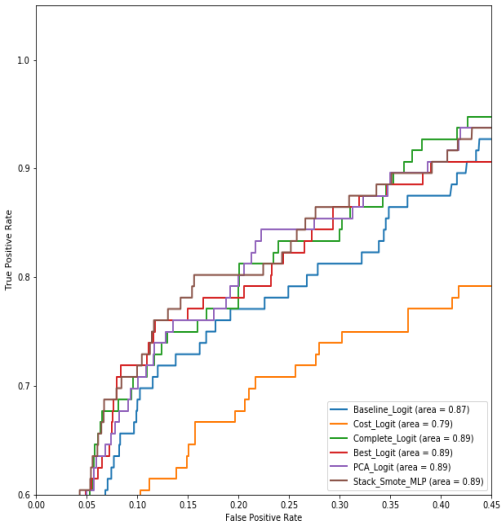


Figure 47: Top 0.5% Bloomer Zoomed ROC Curve

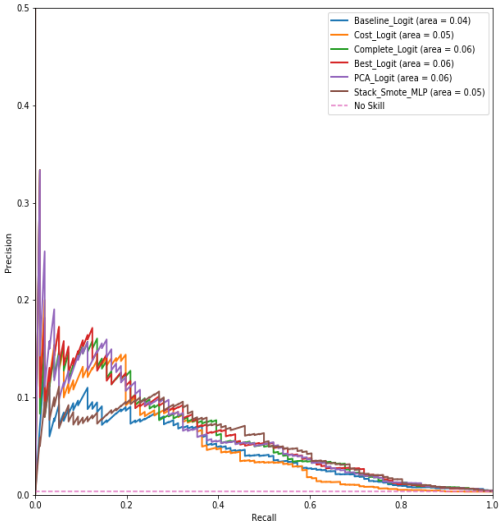


Figure 48: Top 0.5% Bloomer PR Curve

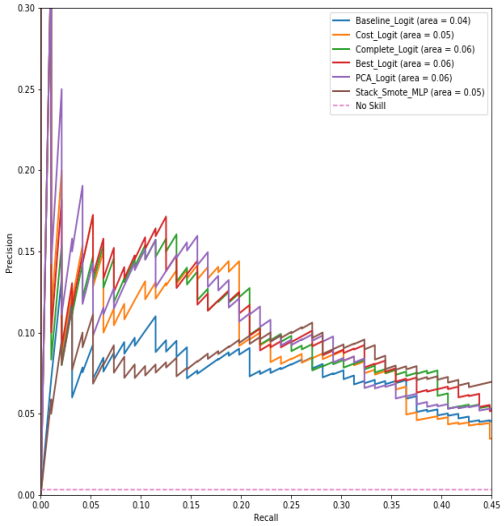


Figure 49: Top 0.5% Bloomer Zoomed PR Curve

12.4.2.2. Top 1% Bloomers

After dropping the 304 2018 top 1% high-cost users, total expenses in 2019 decreased 18%, with 121 of them being also a high-cost user in 2019. Although not presenting best results for the top 0.5% bloomers' classification, the baseline model, together with the best features one, reached the best performance classifying the 183 bloomers among the 304 top 1% high-cost users in 2019. A quarter of the total cost of this group was correctly captured, what may not seem much, but represented 4.8% of the total cost of PASBC in 2019, although these 183 enrollees were no more than 0.6% of the total number of users.

Table 41: Top 1% Bloomers Test Metrics

	Precision_Rank	Precision	AUC_ROC	Recall	AUC_PR	Cost_Capture
Baseline_MLP	0.137	0.080	0.861	0.251	0.060	0.248
Cost_RF	0.115	0.139	0.773	0.055	0.045	0.247
Complete_MLP	0.126	0.085	0.812	0.279	0.061	0.236
Best_RF	0.137	0.000	0.862	0.000	0.068	0.248
PCA_RF	0.109	0.082	0.709	0.180	0.052	0.226
Stack_Smote_Logit	0.109	0.091	0.869	0.251	0.052	0.219

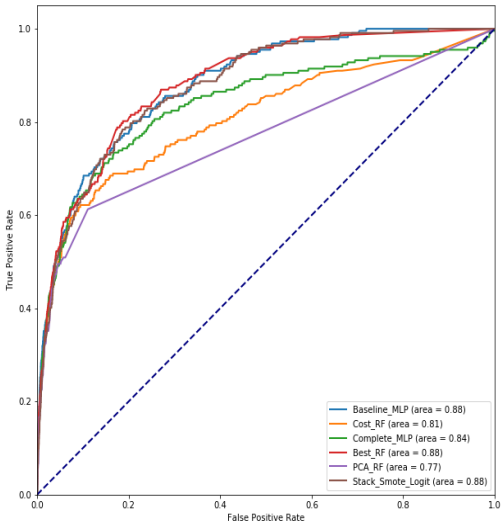


Figure 50: Top 1% Bloomer ROC Curve

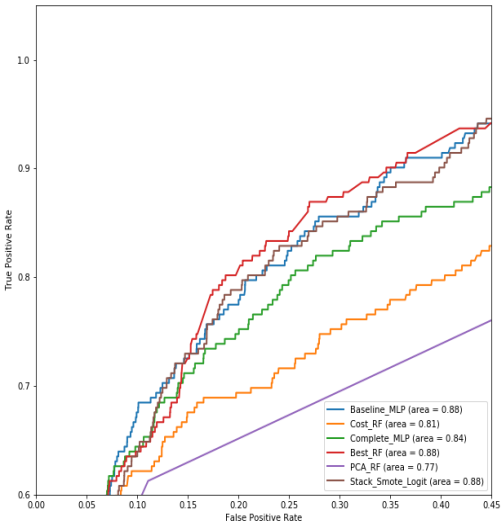


Figure 51: Top 1% Bloomer Zoomed ROC Curve

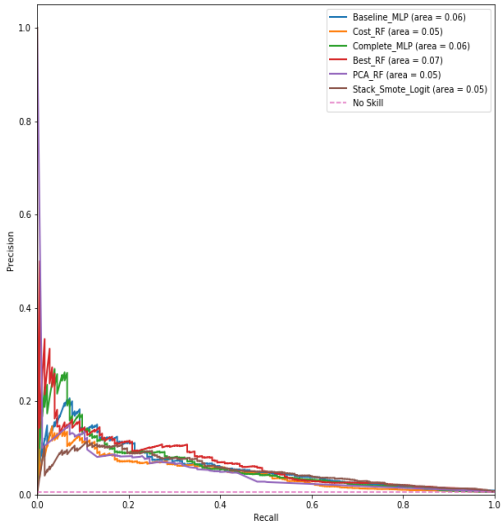


Figure 52: Top 1% Bloomer PR Curve

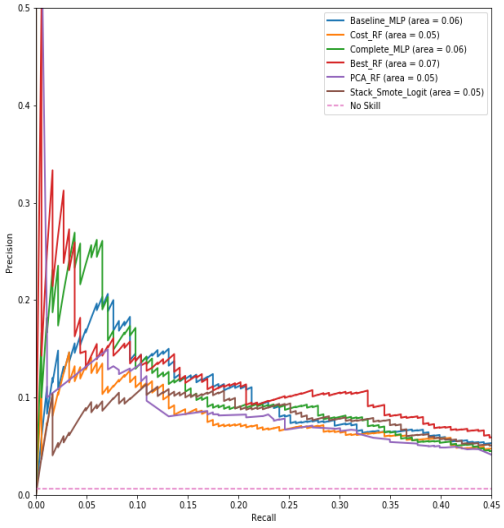


Figure 53: Top 1% Bloomer Zoomed PR Curve

12.4.2.3. Top 2% Bloomers

From the 609 top 2% high-cost users in 2019, 350 were cost bloomers and accounted for 31%, of the 2019 total cost (after dropping the top 2% from 2018) The engineered stacking method was the best classifier of bloomers among the top 2% high-cost users in 2019 using features from 2017 and 2018. As it happened with the top .5% problem, the performance achieved was significantly better than the

one reached by the baseline model and, once again, almost a quarter of the bloomers' total cost could be captured, representing around 6% of PASBC's 2019 total expenses.

Table 42: Top 2% Bloomers Test Metrics

	Precision_Rank	Precision	AUC_ROC	Recall	AUC_PR	Cost_Capture
Baseline_RF	0.106	0.087	0.825	0.023	0.056	0.195
Cost_MLP	0.094	0.074	0.735	0.220	0.048	0.199
Complete_MLP	0.120	0.101	0.811	0.203	0.062	0.202
Best_RF	0.131	0.149	0.832	0.109	0.068	0.224
PCA_Logit	0.117	0.098	0.843	0.206	0.068	0.214
Stack_Smote_MLP	0.131	0.000	0.842	0.000	0.071	0.240

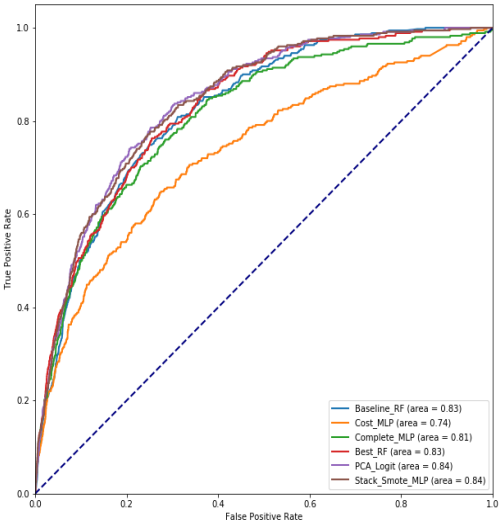


Figure 54: Top 2% Bloomer ROC Curve

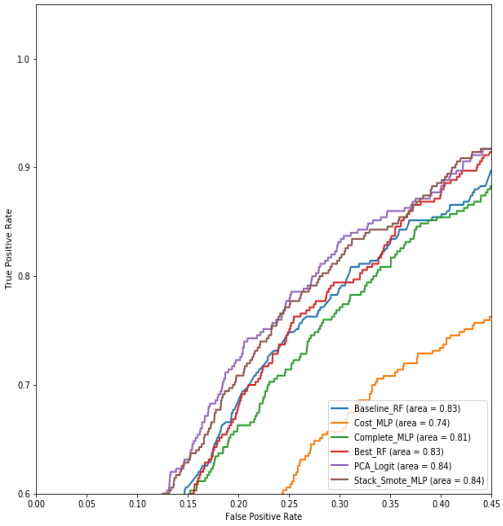


Figure 55: Top 2% Bloomer Zoomed ROC Curve

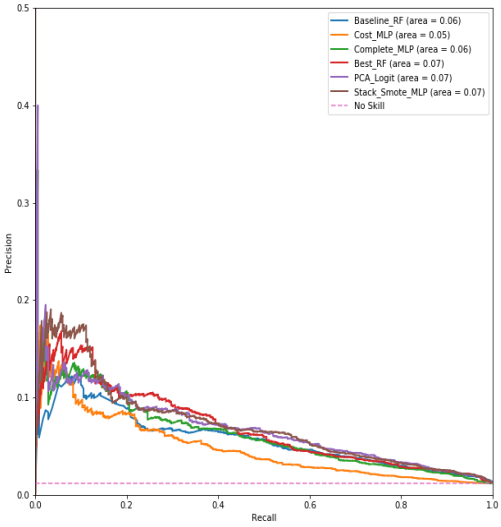


Figure 56: Top 2% Bloomer PR Curve

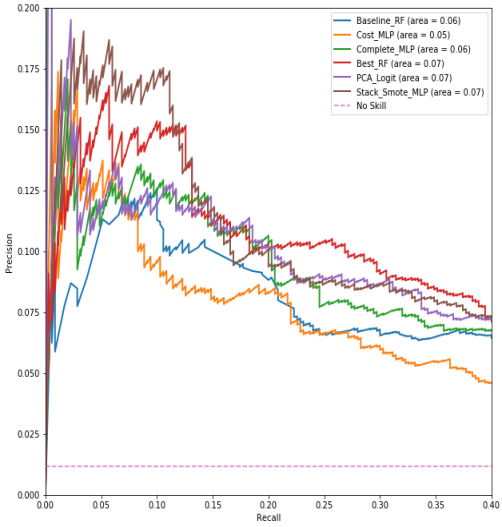


Figure 57: Top 2% Bloomer Zoomed PR Curve

12.4.2.4. Top 5% Bloomers

Bloomers represented 948 of the 1522 top 5% high-cost users in 2019, accounting for around 50% of this group's total costs. A multi-layer perceptron with the best features and the engineered stacking method (using a logistic regression as meta classifier) reached the best performances for the

identification of bloomers among the top 5% high-cost users in 2019, correctly capturing 32% of this group's expenses, around 9% of PASBC's 2019 total costs.

Table 43: Top 5% Bloomers Test Metrics

	Precision_Rank	Precision	AUC_ROC	Recall	AUC_PR	Cost_Capture
Baseline_MLP	0.176	0.164	0.795	0.213	0.123	0.277
Cost_Logit	0.158	0.152	0.715	0.214	0.099	0.279
Complete_MLP	0.165	0.145	0.786	0.333	0.120	0.266
Best_MLP	0.200	0.162	0.803	0.309	0.134	0.319
PCA_RF	0.148	0.054	0.739	0.003	0.094	0.223
Stack_Smote_Logit	0.206	0.163	0.805	0.303	0.138	0.305

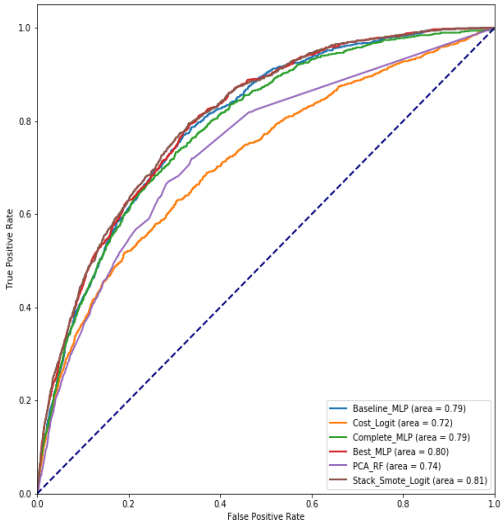


Figure 58: Top 5% Bloomer ROC Curve

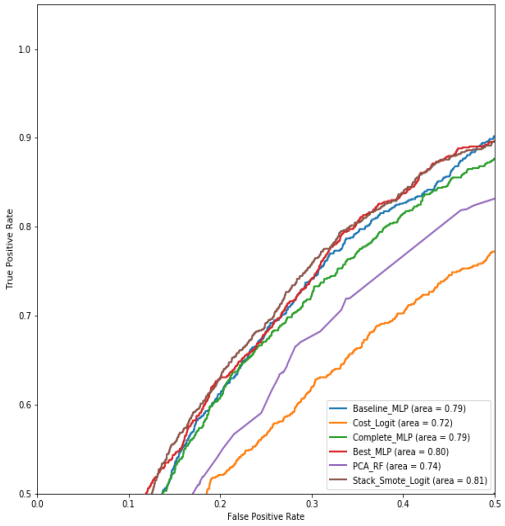


Figure 59: Top 5% Bloomer Zoomed ROC Curve

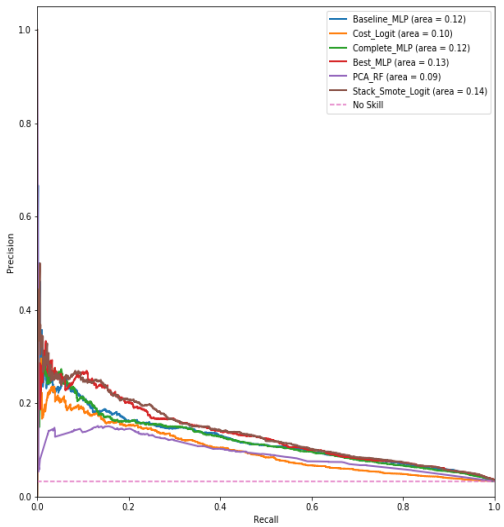


Figure 60: Top 5% Bloomer PR Curve

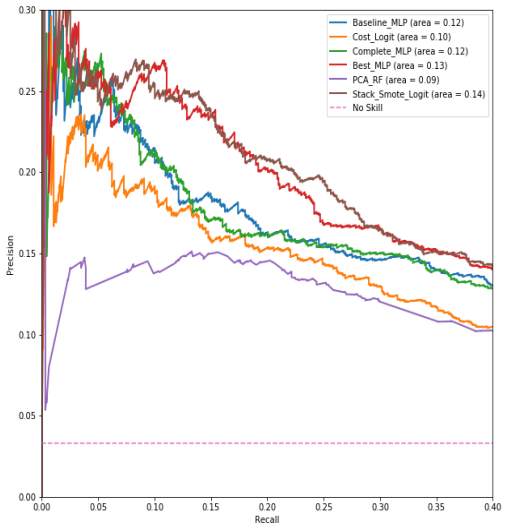


Figure 61: Top 5% Bloomer Zoomed PR Curve

12.4.2.5. Top 10% Bloomers

The multi-layer perceptron that was used with the PCA dataset captured more than 39% of the top 10% bloomers' costs, what accounted for 12% of the total expenses of PASBC in 2019, while this group was about 6% of the users. It is important to remember that, although harder to classify than simple

high-cost users, bloomers identification is important because, for not being high-risk users in the present, these enrollees might be neglected by primary care programs.

Table 44: Top 10% Bloomers Test Metrics

	Precision_Rank	Precision	AUC_ROC	Recall	AUC_PR	Cost_Capture
Baseline_MLP	0.225	0.176	0.747	0.444	0.177	0.338
Cost_MLP	0.225	0.188	0.716	0.356	0.169	0.349
Complete_RF	0.243	0.254	0.746	0.214	0.182	0.368
Best_MLP	0.247	0.180	0.765	0.488	0.196	0.387
PCA_MLP	0.251	0.168	0.763	0.560	0.194	0.394
Stack_Smote_MLP	0.246	0.199	0.764	0.431	0.197	0.377

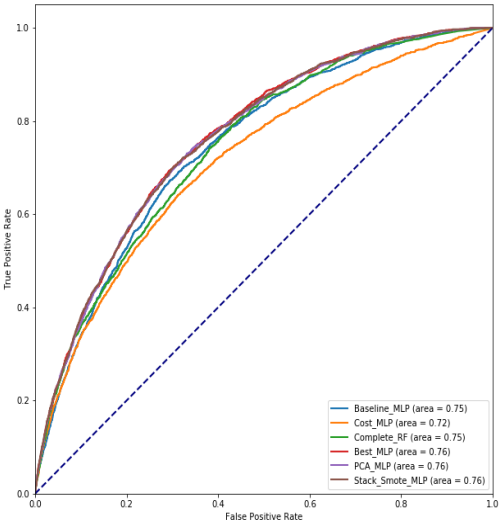


Figure 62: Top 10% Bloomer Zoomed ROC Curve

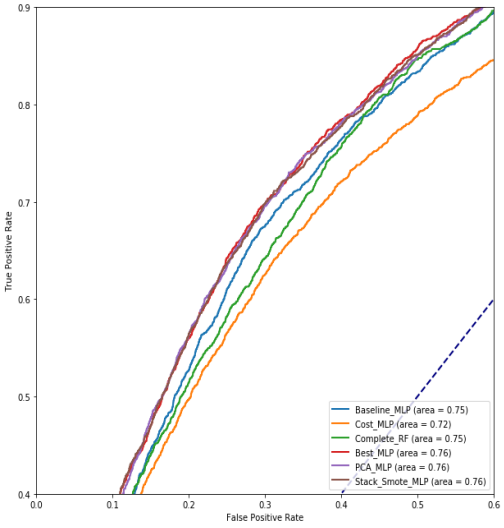


Figure 63: Top 10% Bloomer Zoomed ROC Curve

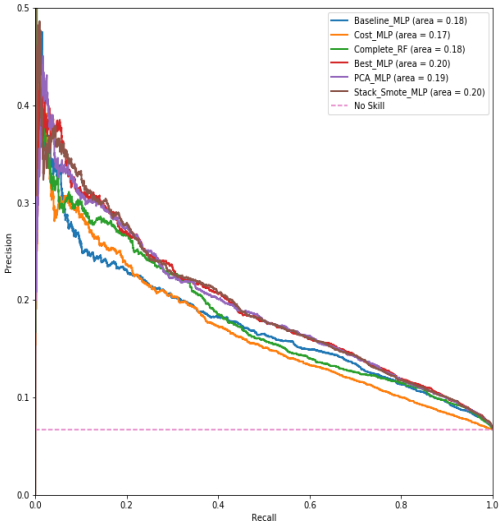


Figure 64: Top 10% Bloomer PR Curve

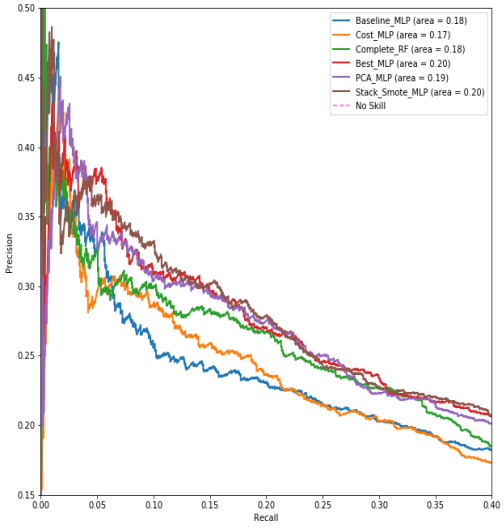


Figure 65: Top 10% Bloomer Zoomed PR Curve

12.4.3. Year 0/Year 2 Simple High-Cost Classification

This classification models use as predictors only data from Year 0, creating a 1-year time-span between predictions and evaluation. This approach is so relevant because, during this interval, intensive primary

care actions may be taken, reducing the probability of these users becoming high-cost in year 2 (or, at least, reducing their total expenses).

12.4.3.1. Top 0.5 High-Cost Users

Once again, the engineered stacking method was the best cost captor among the tested models, getting a third of the 153 top 0.5% users' total costs. It's interesting to observe that the other models, including the baseline one, had a good rank-based precision when compared to the engineered stacking. Nonetheless, the former seems to better classify more expensive instances, so capturing more of the total cost: around 8% of the 2019 total costs of this sample[6].

Table 45: 1Y Time-Span Top 0.5% Simple High-Cost Test Metrics

	Precision_Rank	Precision	AUC_ROC	Recall	AUC_PR	Cost_Capture
Baseline_Logit	0.216	0.158	0.890	0.307	0.119	0.297
Cost_MLP	0.196	0.176	0.780	0.288	0.078	0.301
Complete_DT	0.203	0.213	0.760	0.196	0.115	0.297
Best_DT	0.216	0.238	0.779	0.196	0.132	0.314
PCA_MLP	0.209	0.158	0.882	0.307	0.099	0.301
Stack_Smote_MLP	0.216	0.237	0.869	0.216	0.112	0.332

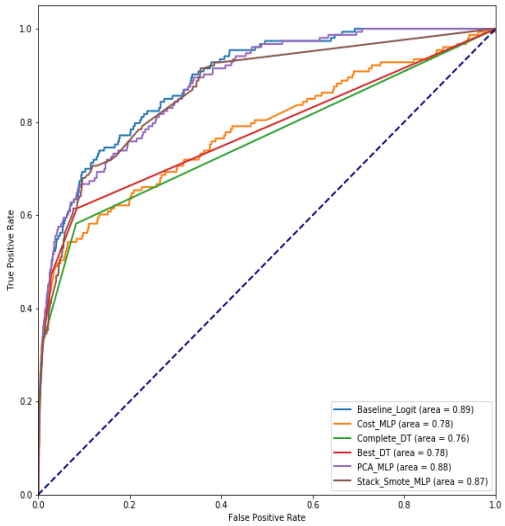


Figure 66: 1Y Time-span Top 0.5% ROC Curve

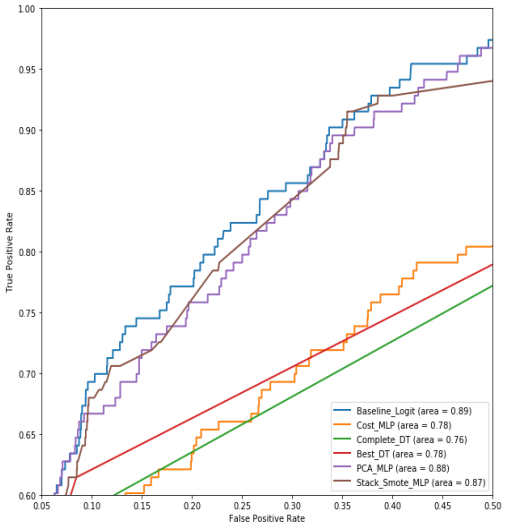


Figure 67: 1Y Time-span Top 0.5% Zoomed ROC Curve

[6] Sections 12.4.3 and 12.4.4 2019 total costs differ from samples in sections 12.4.1 and 12.4.2, because, while the former use all users that were enrolled during all year of 2018, the latter only use the ones that were enrolled during the complete year of 2017.

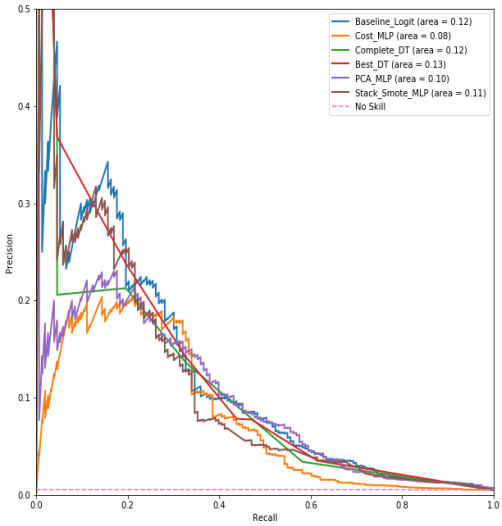


Figure 68: 1Y Time-span Top 0.5% Zoomed PR Curve

12.4.3.2. Top 1% High-Cost Users

A multi-layer perceptron trained with only costs features and as meta classifier in the engineered best features stacking method reached best results for the 1-year time-span top 1% classification problem, capturing almost 40% of this group total costs. Interestingly, as it has already happened before, the area under the ROC curve for the cost predictors' model was the smallest (Figure 69).

The 1% highest-cost users had expenses of 34% of the 2019 total costs for this sample. In a hypothetically exercise, if a one-year time-span preventive intensive care could manage to control expenses in a way that they were reduced by 20%, it would be possible to save up to 2.7% of PASBC's 2019 total costs and sensibly improve the outcomes for these 306 top 1% high-cost enrollees.

Table 46: 1Y Time-Span Top 1% Simple High-Cost Test Metrics

	Precision_Rank	Precision	AUC_ROC	Recall	AUC_PR	Cost_Capture
Baseline_RF	0.248	0.256	0.873	0.167	0.130	0.331
Cost_MLP	0.291	0.313	0.756	0.281	0.158	0.394
Complete_DT	0.242	0.394	0.855	0.199	0.214	0.306
Best_DT	0.255	0.394	0.855	0.199	0.214	0.346
PCA_MLP	0.284	0.238	0.868	0.327	0.159	0.373
Stack_Smote_MLP	0.288	0.174	0.874	0.343	0.161	0.389

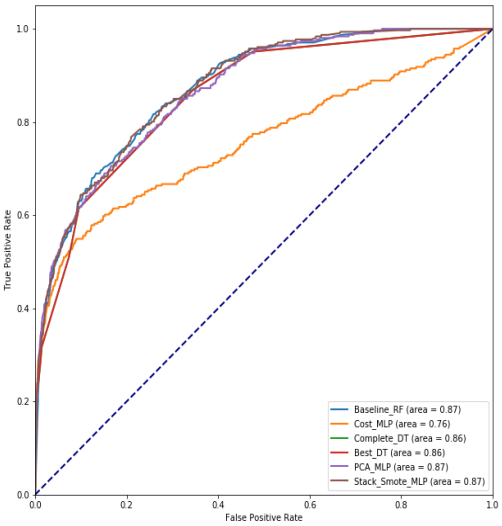


Figure 69: 1Y Time-span Top 1% ROC Curve

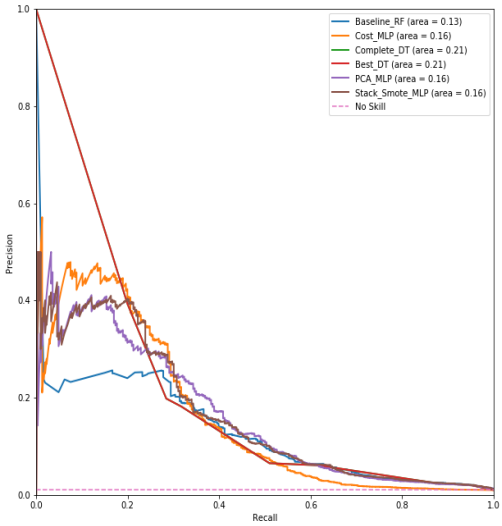


Figure 70: 1Y Time-span Top 1% PR Curve

12.4.3.3. Top 2% High-Cost Users

The engineered stacking method and the PCA model, both using multi-layer perceptrons, reached best performances for the 1-year time-span top 2% simple high-cost users, with an AUROC of .86 and a cost capture greater than 43%. These 613 instances accounted for 45% of PASBC's 2019 total costs, so the best model could correctly capture around 20% of them. Once again, doing a hypothetical exercise, a 20% reduction in these expenses thanks to preventive care could save up to 4% of the total budget of the health program, concomitantly improving outcomes focusing on enrollees that most need attention.

Table 47: 1Y Time-Span Top 2% Simple High-Cost Test Metrics

	Precision_Rank	Precision	AUC_ROC	Recall	AUC_PR	Cost_Capture
Baseline_DT	0.290	0.306	0.837	0.286	0.254	0.399
Cost_MLP	0.303	0.344	0.765	0.281	0.204	0.425
Complete_DT	0.278	0.321	0.837	0.271	0.255	0.386
Best_MLP	0.317	0.285	0.858	0.360	0.234	0.427
PCA_MLP	0.311	0.308	0.860	0.320	0.240	0.436
Stack_Smote_MLP	0.316	0.366	0.860	0.268	0.261	0.436

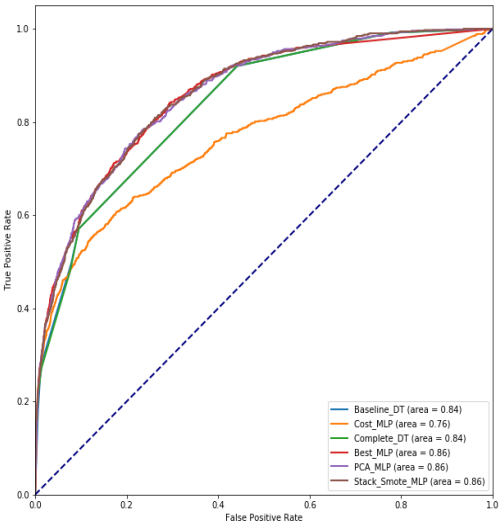


Figure 71: 1Y Time-span Top 2% ROC Curve

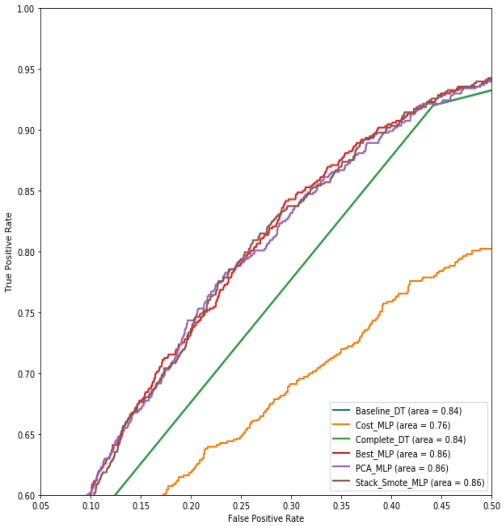


Figure 72: 1Y Time-span Top 2% Zoomed ROC Curve

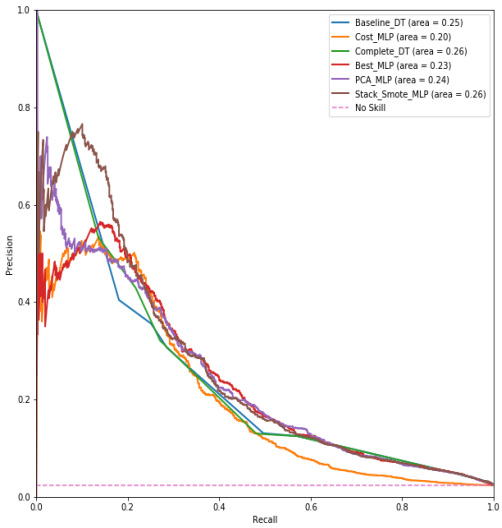


Figure 73: 1Y Time-span Top 2% PR Curve

12.4.3.4. Top 5% High-Cost Users

The best model for the top 5% high-cost classification used all features and a multi-layer perceptron, capturing more than 50% of the total cost. In this classification threshold, the baseline model approached the results of the others and the engineered stacking one, despite reaching the best area under the PR Curve, captured the lowest cost.

The cost of the 1532 enrollees classified as high-cost accounted for 30% of the health program's 2019 total costs, against 59% of the actual top 5%. Interestingly, the difference to the cost captured by the top 5% high-cost simple model (without the one-year time-span) presented in section 12.4.1.4 was just of 6%, as that model captured 36% of 2019 total expenses.

Table 48: 1Y Time-Span Top 5% Simple High-Cost Test Metrics

	Precision_Rank	Precision	AUC_ROC	Recall	AUC_PR	Cost_Capture
Baseline_RF	0.323	0.302	0.821	0.352	0.237	0.501
Cost_MLP	0.315	0.317	0.740	0.311	0.241	0.495
Complete_MLP	0.330	0.246	0.818	0.463	0.281	0.511
Best_MLP	0.345	0.252	0.827	0.473	0.288	0.508
PCA_MLP	0.337	0.253	0.826	0.459	0.279	0.503
Stack_Smote_MLP	0.336	0.242	0.825	0.482	0.310	0.479

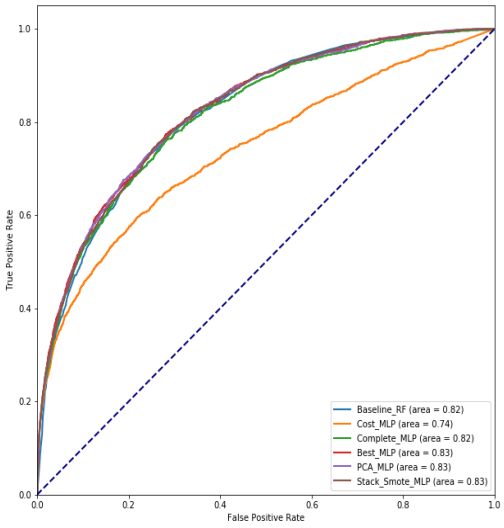


Figure 74: 1Y Time-span Top 5% ROC Curve

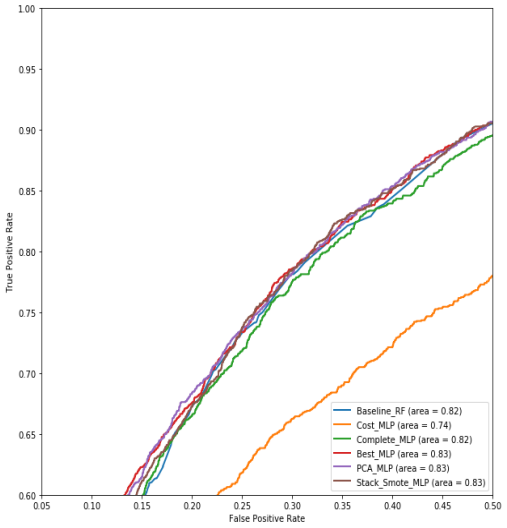


Figure 75: 1Y Time-span Top 5% Zoomed ROC Curve

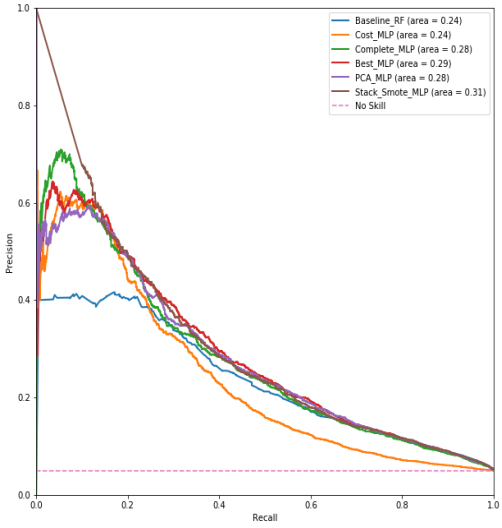


Figure 76: 1Y Time-span Top 5% PR Curve

12.4.3.5. Top 10% High-Cost Users

Models captured around 57% of the top 10% high-cost users' total expenses, 39% of PASBC's 2019 total costs. As seen on section 12.4.1.5, when using data from 2017 and 2018 to predict top 10% users in 2019, it was possible to capture 67% of those expenses (46% of PASBC's total). Of course, this difference is relevant, but it is important to consider the tradeoff between precision and having a time-

span to adopt intensive preventive care actions during this one-year interval, which can improve patients' outcomes and increase the costs' reduction.

The ranking-based precision reached around 39%, against 45% of the simple model that used two years' features as predictors. As usual, the engineered stacking method achieved the highest area under the Precision-Recall curve and could capture 4% more of the total costs than a simple MLP using the same features previously selected as best.

Table 49: 1Y Time-Span Top 10% Simple High-Cost Test Metrics

	Precision_Rank	Precision	AUC_ROC	Recall	AUC_PR	Cost_Capture
Baseline_MLP	0.366	0.237	0.786	0.677	0.315	0.567
Cost_RF	0.361	0.233	0.745	0.579	0.347	0.538
Complete_RF	0.375	0.248	0.784	0.619	0.360	0.559
Best_MLP	0.372	0.249	0.787	0.636	0.357	0.530
PCA_MLP	0.386	0.242	0.793	0.667	0.361	0.576
Stack_Smote_MLP	0.388	0.227	0.794	0.701	0.379	0.570

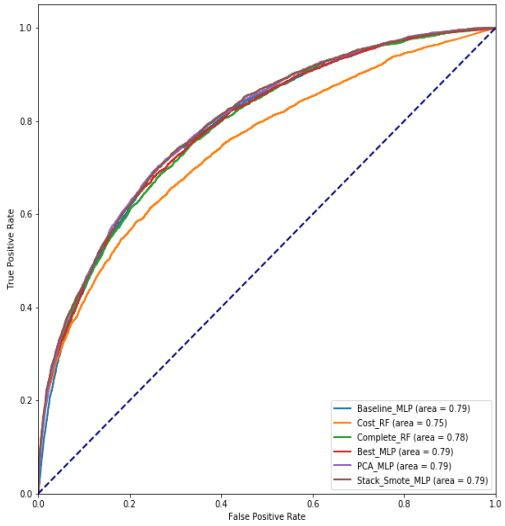


Figure 77: 1Y Time-span Top 10% ROC Curve

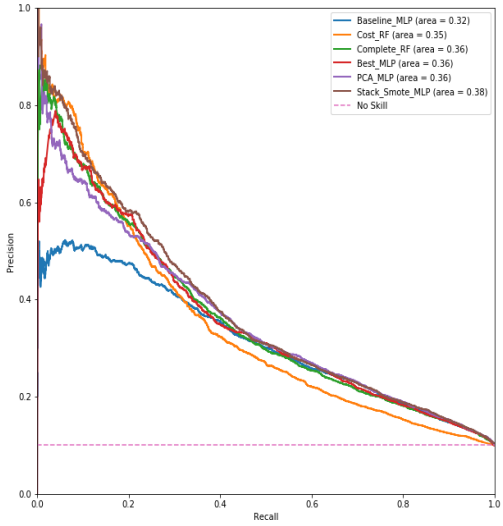


Figure 78: 1Y Time-span Top 10% PR Curve

12.4.4. Year 0/Year 2 High-Cost Bloomers Classification

It was known beforehand that the one-year time-span high-cost bloomers models would be the hardest ones, reaching the poorest performance metrics of all, because they work with two extra layers of uncertainty: one-year time span between predictors and the independent variable and the exclusion of top ranked instances in the year 0, as the goal is to predict bloomers, a harder task than predicting regular high-cost users. Nonetheless, exactly for this uncertainty, these are the models that may bring the best results for a health plan, as they try to identify bloomers, that might be neglected by primary care actions, a whole year before they become high-cost users, giving managers and physicians time to take good care of them, what can improve outcomes dramatically.

12.4.4.1. Top 0.5% Bloomers

As it was expected, performance for the one-year time-span top 0.5% bloomers model was poorer than the one reached for the top .5% bloomers with no time-span and for the top .5% high-cost users

with the one-year interval. Nonetheless, a multi-layer perceptron ran over the dataset with all features captured 20% of this group's costs, despite precisely classifying only 9.4% of the 117 bloomers among the 153 top 0.5% high-cost users.

36 users were among top .5% in 2017 and 2019 and accounted for 23% of this group's expenses in the former year, what means that bloomers represented 77% of this value. Considering the 2019 total costs, the new instances in top .5% accounted for 18.5%, so capturing 20% of this number means identifying users responsible for 3.7% of the program's whole budget.

As expected, it was lower than the 4.5% of 2019 total cost captured by the high-cost bloomers without time-span (what can also be explained by the higher number of bloomers: 117 against 96, as more instances are consecutively high-cost than one-year time-span high-cost) and also lower than the 8% captured by the simple model, that doesn't drop previous high-cost users. Nonetheless, this group is still responsible for a large share of total costs, considering that they are 117 enrollees of the 30641 PASBC's universe that were enrolled during the whole year of 2017 (including the 36 top .5% non-bloomers).

Table 50: 1Y Time-Span Top 0.5% Bloomers Test Metrics

	Precision_Rank	Precision	AUC_ROC	Recall	AUC_PR	Cost_Capture
Baseline_MLP	0.085	0.055	0.860	0.171	0.036	0.178
Cost_MLP	0.094	0.082	0.707	0.103	0.031	0.171
Complete_MLP	0.094	0.056	0.772	0.214	0.034	0.201
Best_MLP	0.060	0.073	0.857	0.154	0.033	0.158
PCA_MLP	0.034	0.000	0.835	0.000	0.031	0.155
Stack_Smote_Logit	0.043	0.042	0.831	0.026	0.024	0.118

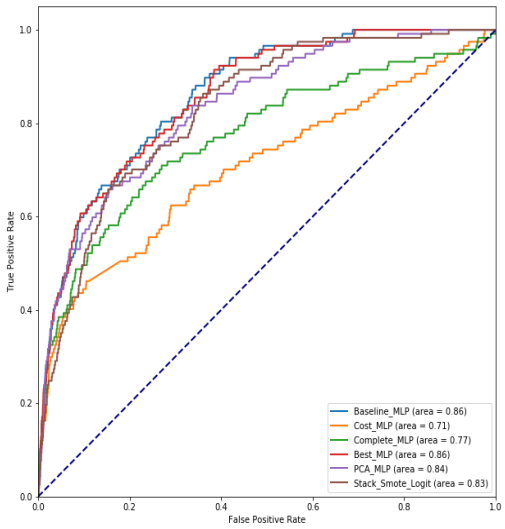


Figure 79: 1Y Time-span Top 0.5% Bloomer ROC Curve

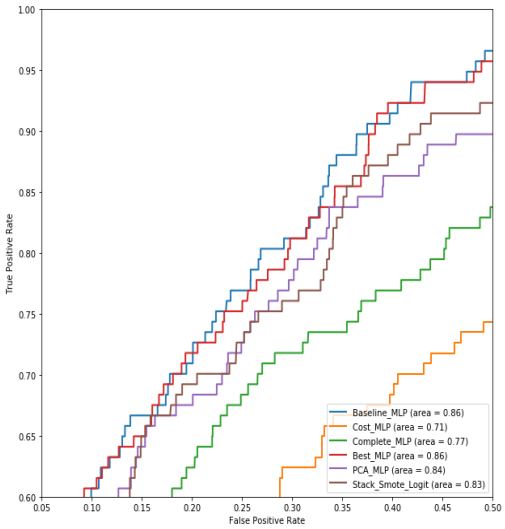


Figure 80: 1Y Time-span Top 0.5% Bloomer Zoomed ROC Curve

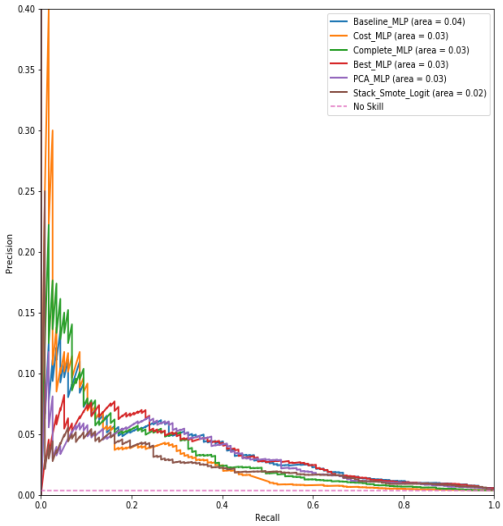


Figure 81: 1Y Time-span Top 0.5% Bloomer Zoomed PR Curve

12.4.4.2. Top 1% Bloomers

Out of the 306 2019 top 1% high-cost users, 223 were cost-bloomers, representing 25% of PASBC's 2019 total costs, with the 83 others accounting for 9% of them. For the top 1% model, a MLP ran over the 12 PCA dataset was the best model regarding cost capture, although the engineered stacking method precisely identified more bloomers, considering the ranking-based precision.

Capturing 5.5% of 2019 total expenses, the best model reached a greater share than the cost-bloomers model without a time-span, although, as previously stated, the former minority class was higher, as less instances were top 1% bloomers, due to a higher number of consecutive high-cost users when comparing 2018/2019 to 2017/2019.

Table 51: 1Y Time-Span Top 1% Bloomers Test Metrics

	Precision_Rank	Precision	AUC_ROC	Recall	AUC_PR	Cost_Capture
Baseline_RF	0.067	0.000	0.833	0.000	0.042	0.138
Cost_MLP	0.094	0.122	0.684	0.049	0.030	0.196
Complete_MLP	0.094	0.093	0.743	0.103	0.064	0.185
Best_RF	0.108	0.000	0.816	0.000	0.044	0.206
PCA_MLP	0.108	0.087	0.833	0.157	0.048	0.220
Stack_Smote_MLP	0.126	0.137	0.838	0.094	0.050	0.195

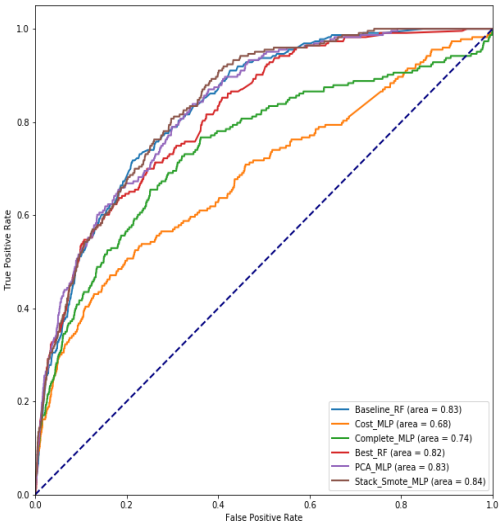


Figure 82: 1Y Time-span Top 1% Bloomer ROC Curve

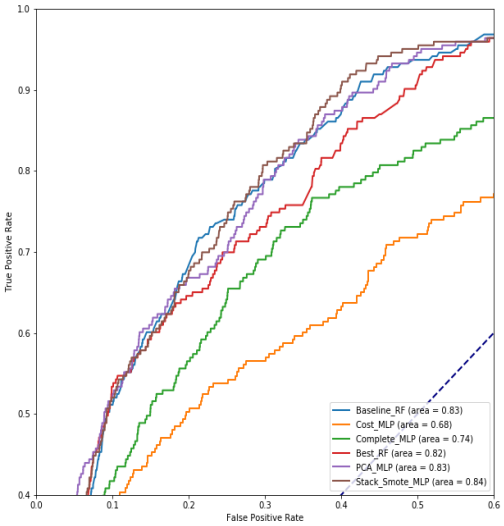


Figure 83: 1Y Time-span Top 1% Bloomer Zoomed ROC Curve

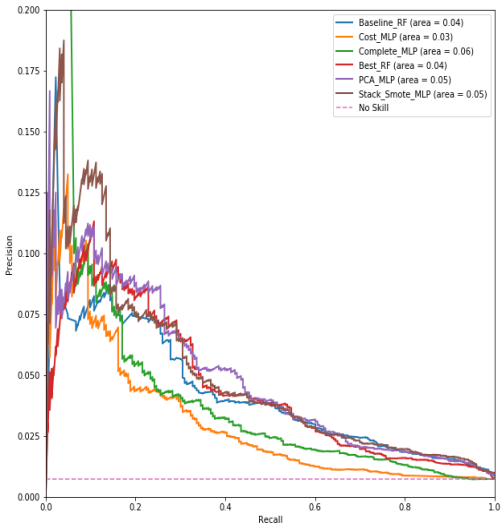


Figure 84: 1Y Time-span Top 1% Bloomer Zoomed PR Curve

12.4.4.3. Top 2% Bloomers

As expected, as the classification threshold increased, the performance metrics improved, as it has already happened before. The higher the minority class, the easier it became to detect these instances and, at the same time, costlier enrollees that weren't ranking-based classified as top 0.5%, but were close to, now will be considered positives and, as costs are highly concentrated, this will increase the cost capture metric. Best features' model was able to capture a quarter of top 2% bloomers' costs and, along with the baseline model, achieved a ranking-based precision of 15%.

Of the 613 users in top 1% slice of 2019, 425 were cost bloomer's (considering 2017) and they represented 29% of the program's total expenses (with non-bloomers accounting for 16% of the total), what means that more than 7% of the whole budget of PASBC was captured by the best model, providing a good opportunity to save the program's scarce resources.

Table 52: 1Y Time-Span Top 2% Bloomers Test Metrics

	Precision_Rank	Precision	AUC_ROC	Recall	AUC_PR	Cost_Capture
Baseline_MLP	0.150	0.160	0.819	0.120	0.083	0.233
Cost_Logit	0.104	0.115	0.685	0.088	0.050	0.192
Complete_Logit	0.142	0.153	0.814	0.142	0.087	0.235
Best_Logit	0.150	0.146	0.819	0.126	0.084	0.253
PCA_Logit	0.144	0.146	0.817	0.150	0.081	0.236
Stack_Smote_Logit	0.134	0.142	0.780	0.076	0.064	0.213

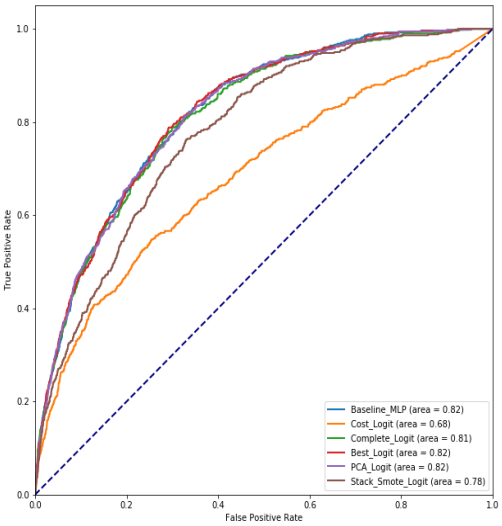


Figure 85: 1Y Time-span Top 2% Bloomer ROC Curve

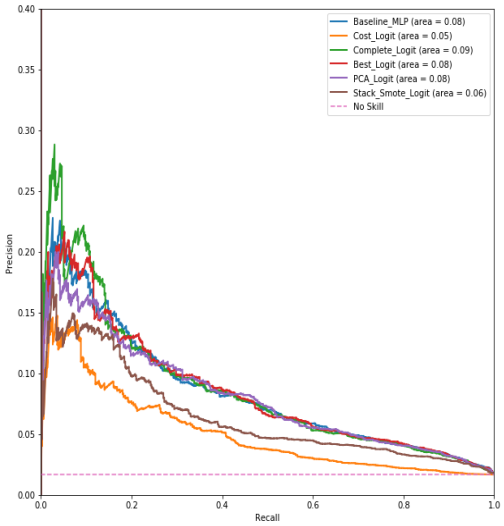


Figure 86: 1Y Time-span Top 2% Bloomer Zoomed PR Curve

12.4.4.4. Top 5% Bloomers

For the one-year time span top 5% bloomers model, the PCA and the baseline could capture more than 28% of this group's total costs, reaching an area under the ROC Curve greater than .78. Once again, although these values do not seem very solid, this problem is extremely complex and even an apparently poor performance may bring great outcomes if good data-driven decisions are taken.

Bloomers accounted for 1,081 of the 1,532 top 5% group in 2019 (considering users that were enrolled during all year of 2017), representing 34% of total costs, while the whole group accounted for 59%.

Table 53: 1Y Time-Span Top 5% Bloomers Test Metrics

	Precision_Rank	Precision	AUC_ROC	Recall	AUC_PR	Cost_Capture
Baseline_MLP	0.192	0.222	0.784	0.148	0.138	0.282
Cost_MLP	0.134	0.120	0.667	0.156	0.084	0.225
Complete_RF	0.186	0.000	0.771	0.000	0.128	0.257
Best_RF	0.179	0.339	0.779	0.019	0.130	0.256
PCA_Logit	0.195	0.172	0.786	0.291	0.139	0.282
Stack_Smote_MLP	0.128	0.123	0.745	0.137	0.098	0.194

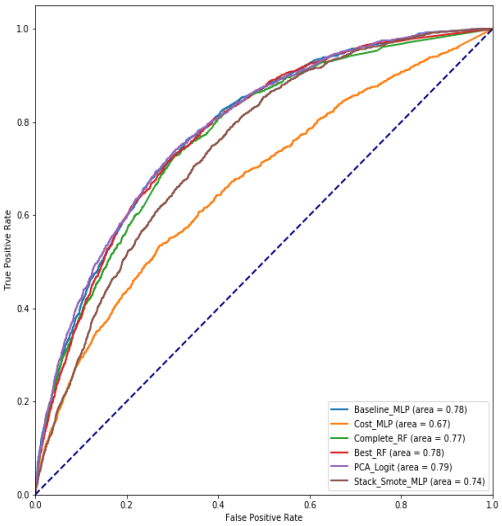


Figure 87: 1Y Time-span Top 5% Bloomer ROC Curve

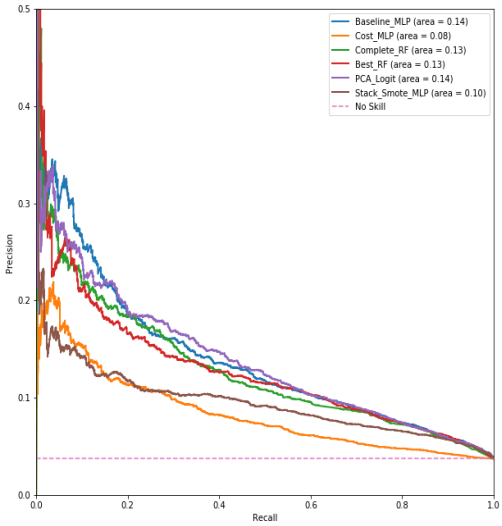


Figure 88: 1Y Time-span Top 5% Bloomer Zoomed PR Curve

12.4.4.5. Top 10% Bloomers

Of the 3,064 top 10% high-cost users in 2019, 1,994 were cost-bloomers and accounted for 36% of PASBC's total cost in that year, around half of that entire group's share. For their classification, except for the MLP used in the cost features dataset, all other models achieved similar performance, capturing around a third of top 10% bloomers' total costs and a ranking-based precision around 23%. As it was said before, although these numbers are not as good as desired, they represent 9% of the total costs of PASBC and, if the targeted preventive care actions could decrease expenses by at least 30% (something that may be feasible, considering a whole year interval to take measures), for example, it would mean saving almost 3% of the total costs of PASBC, while improving the enrollees' quality of life considerably.

Table 54: 1Y Time-Span Top 10% Bloomers Test Metrics

	Precision_Rank	Precision	AUC_ROC	Recall	AUC_PR	Cost_Capture
Baseline_MLP	0.217	0.160	0.736	0.577	0.174	0.324
Cost_MLP	0.179	0.155	0.669	0.343	0.139	0.253
Complete_Logit	0.230	0.178	0.742	0.482	0.186	0.338
Best_MLP	0.226	0.180	0.741	0.469	0.186	0.322
PCA_MLP	0.224	0.174	0.742	0.506	0.175	0.338
Stack_Smote_MLP	0.229	0.165	0.743	0.542	0.188	0.324

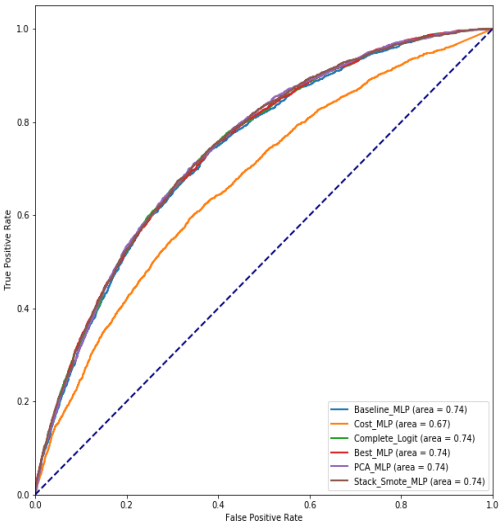


Figure 89: 1Y Time-span Top 10% Bloomer ROC Curve

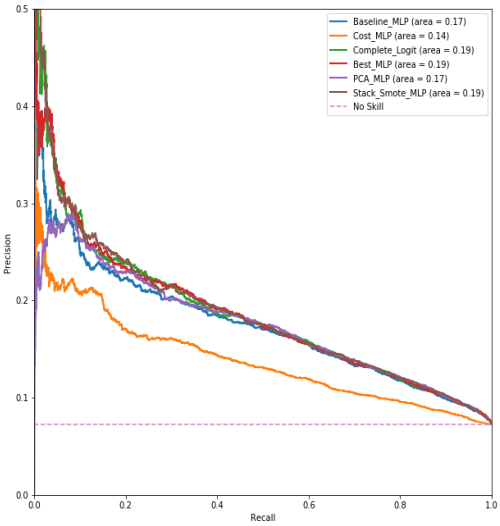


Figure 90: 1Y Time-span Top 10% Bloomer Zoomed PR Curv