NOVA
IMS
Information
Management
School

MDSAA

Master’s degree Program in
Data Science and Advanced Analytics

DEMOGRAPHICS IMPUTATION IN MARKETING SECTOR BY MEANS OF MACHINE LEARNING

Margarita Venediktova

Internship Report

presented as partial requirement for obtaining the Master Degree Program in Data Science and Advanced Analytics

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

DEMOGRAPHICS IMPUTATION IN MARKETING SECTOR BY MEANS OF MACHINE LEARNING

by

Margarita Venediktova

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics, with a Specialization in Data Science

Supervisor: Prof. Doutor Flávio Luis Portas Pinheiro

November, 2022

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, 25th November 2022

ABSTRACT

The goal of this project is to develop a predictive model in order to impute missing values in data collected through surveys (demographics data) and evaluate its performance. Currently there are two existing issues: demographics data for each user is either incomplete or missing entirely. Current POC is an attempt to exploit the capabilities of machine learning in order to impute missing demographics data.

Data cleaning, normalization, feature selection was performed prior to applying sampling techniques and training several machine learning models. The following machine learning models were trained and tested: Random Forest and Gradient Boosting. After, the metrics appropriate for the current business purposes were selected and models' performance was evaluated.

The results for the targets 'Ethnicity', 'Hispanic' and 'Household income' are not within the acceptable range and therefore could not be used in production at the moment. The metrics obtained with the default hyperparameters indicate that both models demonstrate similar results for 'Hispanic' and 'Ethnicity' response variables. 'Household income' variable seems to have the poorest results, not allowing to predict the variable with adequate accuracy. Current POC suggests that the accurate prediction of demographic variable is complex task and is accompanied by certain challenges: weak relationship between demographic variables and purchase behavior, purchase location and neighborhood and its demographic characteristics, unreliable data, sparse feature set. Further investigations on feature selection and incorporation of other data sources for the training data should be considered.

KEYWORDS

Machine Learning; Missing Values Imputation

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS AND ACRONYMS

POC	Proof of concept
EM	Expectation-Maximization
KNN	K- nearest neighbors
MAR	Missing at random
MCAR	Missing completely at random
MNAR	Missing not at random
RF	Random Forest
ML	Machine Learning
MI	Mutual Information

1. INTRODUCTION

1.1. Problem Statement

Gaps in the data affect businesses' insights, leading companies to lose clients and revenues. According to Dun&Bradstreet's report *(Dun&Bradstreet, 2019)* 1 out of 5 companies have lost a customer due to incomplete or inaccurate data. Needless to say, that it is crucial for the companies to build a production pipeline which will adjust for the missing data.

The project is carried out at a marketing research company, at the department of product innovation and development and focuses on the demographics' data imputation. The basis of the product is data (purchase receipts) which is collected through mobile application by the recruited users (panelists) who voluntarily submit their paper or digital receipts to the platform. The collected data is then aggregated and projected to the population to produce meaningful insights into consumer behavior within certain timeframes. The complete demographics data is vital for existence of the final product. However, there are certain issues with obtaining the full set of demographic variables for each user.

In the marketing research, traditionally, data is collected through surveys offered to users (panelists). In our particular case, panelists who voluntarily participate in receipts' submission via mobile application are required to provide some basic demographics information about the household (such as ***age***, ***gender,*** and ***zip code***). They are only encouraged to fill in additional demographics data (such as household ***income, presence and age of children, race, hispanic identity, and education***) at the next step of a process. This second round of data collection suffers from non-responses or incomplete responses when the user either does not know how to answer the question or refuses to do so. This results in some demographics data being missed and does not allow the company to use the available information since it only can be used in conjunction: even though the participants qualify to be eligible for use of the data they provide (they purchase receipts) based on other criteria such as frequency of the submission, and the amount spent per period, they are disqualified from the panel unless the provide full demographics data. Only the presence of all variables makes the data received from panelists usable. Therefore, the missing fields, incomplete is a major concern and requires to be addressed appropriately.

1.2. Objective of the present project

Given the problem stated above the primary objective of the present work is to investigate the possibility of imputing multivariate missing demographics data using machine learning approach and data available. Many studies have addressed the issue of imputing missing values and have come up with successful approaches (*Wang et al., 2022; Sterne et al., 2009, Emmanuel et al., 2021. etc.*)

In order to proceed with the project, first of all, the possibility of using the input data for production purposes is evaluated and then the methodology overview is done, followed by training and testing several machine learning models. The results, conclusions and business decisions are based on several criteria, such as models' performance and production feasibility.

1.3. STRUCTURE OF THE PRESENT PROJECT

Initially, the literature review was conducted (Chapter 2), in order to understand relevant approaches used in imputation of the missing data. Chapter 3 gives a theoretical overview of methodologies used in the present project: possible data sources and appropriate variables to build a model; data pre-processing activities, normalization (when required), as well as balancing the dataset (if required); evaluation criteria is defined based on business needs. Next step involves building a model and evaluation of its performance. Results and discussions are presented in Chapter 4. After all, the final conclusion about usage of machine learning imputation is made based on models' performance, relevant business requirements and limitations can be found in Chapter 5.

1.4. CONTRIBUTION TO THE COMPANY

Data collection, as well as the data quality are one of the most cumbersome parts of the marketing research business. Both of these processes would greatly benefit from the ability to omit some data collection parts and having an ability to accurately generate missing data based on available resources. The possibility of restoring missing data by the means of accurate imputation has a valuable contribution to the quality of a company's products overall.

Therefore, the current project aims to benefit the company by providing a method to generate required data in-house or impute the missing values by the means of machine learning. The current project has a POC status, and the possibility of production deployment will be discussed in the final chapters.

2. LITERATURE REVIEW

2.1. LITERATURE REVIEW OF DEMOGRAPHICS IMPUTATION TECHNIQUES

2.1.1. Missing data, missing data types and missing patterns.

Missing data is a common issue for many datasets, especially those where participation of a human is an integral part of data collection (e.g. surveys) as a result of non-response. Ideally, design of a study should be done in such a way that minimizes in data collection. However, it is a difficult task to anticipate all possible scenarios that would lead to errors and, as such, it is common that datasets include errors of different types and different sources. Missing data very often leads to limitations in the possible applications in analytical projects *(Sterne et al., 2009*). For example, issues related to missing data arise when developing a machine learning model or application: Clustering of data will be corrupted since missing values are treated as equal therefore providing incorrect information about actual distances between vectors *(Oba et al., 2003).* Oba et al. (2003) also named support vector machines (SVM) classifiers, PCA (principal component analysis) and singular value decomposition (SVD) as models and methods that could potentially suffer from missing values. Similarly, artificial neural networks (ANN) or logistic regression cannot handle missing values.

There are several types of missing data each characterized by the relationship between measured variables and the probability of missing data *(Baraldi and Enders, 2010)*, namely: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) *(Chinomona and Mwambi, 2015).* Data is considered to be missing completely at random (MCAR) if the probability of missingness is the same across all items or, in other words, does not depend on the observed value or the missing ones *(Gelman et al, 2021; Rubin, 1976).* In this scenario it is probable that missing values cannot be predicted from any other value in the dataset.

However, most of the time, is not completely at random, therefore a more general assumption would be that missing data depends only on available information (other measured or observed variables) but does not relate to the underlying observed values of the incomplete variable (Gelman et al, 2021) Finally, the missing data that is not MCAR or at least not MAR is considered to be not missing at random (NMAR), but on the other hand relate to unobserved variables or observed variables, e.g. on the variable itself. In the latter case, the process of missing data can be a case of self-censoring (e.g. sensitive question in a survey). In case of data being missed-not-at-random imputation may result in biased data *(Sterne et al., 2009)*. However, it is impossible to distinguish between MAR and MNAR based only on the observed data, a field expert should use his knowledge of the study and subject matter to decide whether MAR is plausible *(Hughes et al., 2019).*

According to Hughes et al., 2019 multiple imputation (MI) gives unbiased results for the MCAR or MAR data and is biased when the data is generated under MNAR. There are different types of missing patterns which can occur in the data. Missing data pattern is univariate when there is only one variable containing missing data *(Demirtas, 2018).* If variable J is missing then Variable K is also missing for all K>J and is referred to as monotone pattern and tends to occur in longitudinal studies *(Chen, 2022).* In a non-monotonic pattern, the unavailability of one variable does not affect another variable *(Chen, 2022).*

Restoring data when complete non-response occurs vs item non-response has different approaches. In case of complete non-response deletion, weighting adjustments and imputation methods are used in order to deal with missing data. For item nonresponse imputation is the most common way to approach it (*Chinomona and Mwambi, 2015).* However, since most statistical analysis requires a complete-case approach and it is not advisable to eliminate records from the data, it is preferable that item non-responses should be restored using the imputation techniques.

2.1.2. Main methods of accounting for missing values.

As described in the previous section the most common methods of handling missing value in the data are deleting cases with missing values (*complete-case analysis*), weighting adjustment methods and missing values imputation *(Wang et al., 2022; Sterne et al., 2009).* The latter is believed to be the most advanced method widely used in various studies (*Sun and Kardia, 2008).*

Also, traditional techniques of handling missing data include statistical methods to processing missing data, e.g., imputation with calculated statistics value drawn from the dataset (*mean substitution* method), hot-deck imputation, k-nearest neighbors (KNN), expectation maximization and multiple imputation *(Sterne et al., 2009).*

Mean substitution is the simplest approach used to account for missing values, used for numerical values, while mode substitution is a method to impute missing data, used for non-numerical variables.

Hot-deck imputation matches key variables of records with missing data with complete records and afterwards the substitution of missing values on random basis using complete data *(Emmanuel et al., 2021)*

KNN algorithm is another common method used for imputation of missing values based on the mining the similarity between samples using not missing data by distance measurement and then estimate missing values using complete data of neighboring data points (median, mean, or other calculated statistics) (*Wang et al., 2022*). KNN imputation is possible for discrete and continuous data and can also be implemented for multiple missing variables *(Emmanuel et al., 2021)*. KNN can be regarded as an extension to a hot deck approach. Hot deck imputation techniques similarly use observed values from “the donor” close to the record with incomplete data (*Fouad et al., 2021).*

Multiple imputation relies on observed data in order to approximate values of missing records and the calculation is performed fixed number of times and then the results from all rounds are combined for the final imputation result *(Emmanuel et al., 2021)*

2.2. MACHINE LEARNING FOR MISSING DATA IMPUTATION.

The development of machine learning methods had a significant effect on the feasibility of imputing missing data as the machine learning models can restore the true distribution of data from missing data sets more accurately than the traditional missing data processing models (*Wang et al., 2022*).

Mostly, researchers tend to use supervised machine learning models to predict missing values, however, there are studies where researchers used unsupervised learning, namely *clustering techniques* such as hierarchical clustering and k-means clustering. However, clustering methods are reported to not be robust enough to handle the missing data problem *(Emmanuel et al., 2021).* Gajawada and Toshniwal (2012) in their study proposed a missing value imputation method based on K-means and KNN where they used imputed values to iteratively impute other values. The result of their study demonstrates that whilst this method has its potential, must be carefully implemented as errors in the earlier rounds of the imputation process might propagate further. Nevertheless, their method performed better than the method that didn’t not use previously imputed records.

Logistic regression was used in imputing missing values by Wang (2022) and demonstrates competitiveness with other models such as a*rtificial neural network* or *ensemble models*.

Decision tree algorithm and ensemble version of decision tree algorithm *Random forest* appear to be appealing approaches for imputing missing data as they are able to handle multiple data types, handle complex interactions and nonlinearity *(Tang & Ishwaran, 2017).* Tang and Ishwaran (2017), in their extensive research, using a diverse collection of datasets they evaluated various random forest based missing values imputation techniques and given different missing value types. They concluded that performance improved with increase of the correlation of the attributes, specifically the missForest algorithm performed better; in addition, they discovered that traditional KNN was outperformed by RF based methods. Previously, Shah et al. (2014) also suggested that multivariate imputation by chained equations (MICE) performed better when the predictor was random forest. Hong & Lynn, (2020) after conducting their research on RF based methods, insist RF based algorithms do not always perform best for missing data imputation and careful analysis of the underlying mechanism of missing data and the relationship between variables in the dataset should be conducted. Nevertheless, they confirmed that RF based imputation methods can have good prediction accuracy, but when the imputed variables are used in subsequent imputation it might lead to biased results. Wang (2022) among other techniques adopted *RF (Random Forest) and SVM,* which also showed good results. All the models were successfully reaching >80% level of sensitivity and AUC. Overall, study by Wang (2022) confirmed the initial assumption that machine learning methods would have better imputation performance compared to traditional methods.

Imputation methods based on RF have become popular due to their abilities to handle data without the need to specify the distributions of the variables like most standard methods

Other researchers experimented with ensemble models*:* Zhu et al. (2021) in their research proposed the construction of ensemble classifier based on one of the most commonly used imputation methods (KNN, LLS (local least squares), ILLS (iterated local least squares), SVD (singular value decomposition)) using bootstrapping sampling and then weighting was applied to get the final result. Experimental results confirmed the advantage of the proposed method over other tested methods.

Khan et al. (2019) investigated ensemble of mean imputation (MI), Gaussian Random Imputation (GRI) and expectation-maximization (EM) imputation combined in different ways. According to Khan et al., 2019 ensemble-based imputations perform better than their single imputation counterparts for missingness ratio of 10% or more.

Another group of techniques uses *artificial neural networks* classifiers for imputation of missing values. Cheng et al. (2020) experimented with MLP (multilayer perceptron) in order to impute missing data in their Attention-Deficit Hyperactivity Disorder study, with Rectified Linear Unit (ReLU) as hidden-layers activator and softmax function for the output layer and SVM was used to evaluate imputation quality. Their findings provide evidence that neural networks can help impute missing values with high accuracy. Mishra et al. (2018) proposed MLP approach in conjunction with genetic algorithms for optimization. In another study by Sun and Kardia (2008) they investigated *feed-forward neural network* performance of imputing missing data in a genotype dataset where they are able to achieve >86% of accuracy imputing missing values.

Recently, some fusion techniques were proposed in order to improve the performance of imputation of missing values. Among them research made by Nikfalazar et al. (2019) where they propose a method which combines the benefits of decision trees and fuzzy clustering. The experimental results of the proposed approach outperform the other five effective imputation methods. Moreover, the proposed method demonstrates robustness when experimenting with various types of missing data. Rahman and Islam (2013) introduced a model that uses Decision tree and an EM imputation technique as they are convinced that EM technique would perform better on partitions of the data (identified using decision trees) with higher correlation among variables. Afterwards they applied an EM algorithm on various segments of data where correlation of attributes was higher. The performance of their newly proposed algorithms was higher than standard EM based algorithms.

It is argued that, while traditional evaluation criteria are mainly based on the difference between the actual values and the imputation values, this approach might be limiting the feasibility of application of imputation models since it is rather complicated to make the imputed data distribution completely consistent with the underlying true distribution. Even though there are differences between the two distributions (imputed vs underlying) the imputation models can be beneficial unless they undermine the accuracy of decision-making *(Wang et al., 2022).*

2.3. Features pre-processing and selection

The primary objective of feature selection is to select relevant variables from a dataset which are believed to efficiently describe the target variable, avoid curse of dimensionality, ease the interpretation of the results and even reduce computational effort *(Brank et al., 2011, Chandrashekar and Sahin, 2014)*. There are several known groups of approaches to feature selection: filter methods, wrapper methods, embedded methods, ensemble methods and other.

Filter methods are ranking techniques where an appropriate ranking criterion is used to rank variables and based on the predefined threshold to remove less relevant attributes. The feature is considered relevant if the feature is not independent of the target variable, otherwise, it can be disregarded. The feature can be independent of the other variables, however, still can be affecting the target label (*Chandrashekar and Sahin, 2014).* Some of the most common ranking methods are Pearson correlation and Mutual Information. Pearson correlation, while being the most simple and straightforward, can only reveal linear relationships. Mutual information is a measure based on Shannon's entropy and if a value of the measure is equal to zero, then the variables are independent, otherwise the variables are dependent.

The major issue of filter methods is that features that are not considered informative on their own can be informative in conjunction with other features, thus removing those variables can result in the situation when finding a suitable learning algorithm can become difficult since the underlying learning algorithm is ignored (*Chandrashekar & Sahin, 2014, John et al., 1994).*

Wrapper methods concept implies sampling variables and using the learning algorithm as the objective function to evaluate the subset of variables. Given that with a large number of initial attributes the problem tends to be computationally difficult to do the exhaustive search and therefore heuristic methods are used in conjunction. For example, Genetic Algorithm and Particle Swarm Optimization can yield local optimum with reasonable computational effort and produce good results (*Chandrashekar & Sahin, 2014, Xuan et al., 2011*)

Embedded methods group of methods can be split into two: regularization and algorithm-based approach. Algorithm based approach’s main idea is to incorporate features selection process in the training of the model, e.g., the classification algorithm has its own feature selections in them. The most known algorithm-based approach would be Random Forest classifier (as well Decision tree, XGboost, etc.)

2.4. Performance evaluation metrics

M & M.N (2015) in their review of the evaluation metric they emphasize that in order for the metrics to be reliable the following aspects of the metrics to be considered: metrics should accommodate for multiple classification problem, computational cost, metrics must be able to produce distinctive and discriminable value in order to able to search for optimal solution, and finally, informativeness, specifically in the case of massively imbalanced data.

Nevertheless, the most used metrics are presented here, and their advantages and disadvantages are discussed.

Accuracy metrics measures the ratio of correct predictions over all the predictions and therefore is highly susceptible to imbalance in the data. In spite of this, accuracy is still the most widely used metrics *(M & M.N, 2015),* probably due the simplicity in interpretability. *Sensitivity* and *specificity*, in their turn, are meant to provide for that weakness. Sensitivity is a ratio of positive predictions that are correctly classified, similarly specificity is a ratio of negative predictions that are correctly classified. Additionally, *precision metrics* shows the fraction of positive predictions which are correctly predicted from the pull of all positive predictions. *Recall* metrics is used to measure the ratio of correct predictions of positive class in the pool of positive class (correctly or incorrectly classified). Recall equals sensitivity. For multi-class problems the above-described metrics could be used in a weighted manner.

Other measures of performance of machine learning models include F1 score, and area under the ROC Curve (AUC ROC).

AUC is a universal metric that can be used to compare several learning algorithms. The AUC was proven to be better than the accuracy metric for both evaluating the classifier performance and discriminating an optimal solution during the classification training *(Jin Huang & Ling, 2005).*

3. DATA AND METHODS

In this section the utilized methods for data pre-processing, data normalization, feature selection are explained and also more theoretical background is given regarding each of the machine learning methods used in this project.

3.1. DATA

Data consists of census data, demographics of users of an application and paper and electronic purchasing receipts collected within a time window of one month (September 2021). The possibility of using more than one period of data and therefore more data concerning purchasing behavior was considered but declined due to the business requirement to enable the use of the data within a shorter period of time. Paper and electronic purchasing receipts in the present research were collected by a third company through mobile application which processes the images of the receipts submitted by the users.

Census data used in this project was retrieved from United States Census Bureau. Demographics data (age, gender, zip-code, ethnicity (race), education, Hispanic identity, age and presence of children, household income) are collected through mobile application. Paper and electronic purchasing receipts give us information about purchasing behavior of the respondent. Firstly, this information describes the share of volume of each product category in the basket and share of monetary value spent in certain product categories, and secondly represents the share of volume of each channel and share of monetary value spent per market channel. Category is the second hierarchy level of products which represents a group of products with similar characteristics (i.e., chocolate milk, white milk, and skim milk are all the products found in the milk category). Channel is the general classification of shops, e.g., Grocery, Pharmacy etc. Shops are entities linked to the channel, and each shop can belong to only one channel. For the training and testing purposes only a complete set of data is used.

3.2. DATA PRE-PROCESSING AND NORMALIZATION

For the purpose of this project the dataset is divided into training, validation, and test sets. Training set comprises roughly 70% of the dataset, while the test set represents 30%; in its turn, the validation set roughly accounts for 30% of the training data. Training and validation sets are used in the training and benchmarking process while obtaining the best model by training and testing different models as well as their hyperparameters. The test set is only used for final evaluation of optimal hyperparameters of the model and is completely independent of the training and validation set.

There are 86.000 rows in the data set, 483 of them contain missing values which corresponds to roughly 0.5% of the dataset: no imputation strategy is applied due to the low number of such cases - these entries are removed.

The scale and distribution of the data can be different for each variable, which can lead to difficulties in modeling due to skewness when, for example, dealing with models based on distance measuring (k-nearest neighbor) or artificial neural networks. Therefore, it is necessary to apply scaling transformation on our dataset. Even though two approaches use decision tree algorithms as a base model and are not affected by the scale of the variables, the normalization can be applied since it does not affect the results. Most of the variables in the dataset concerning purchasing behavior of each user, such as quantity bought, total spend or frequency per each channel and super category, are represented as a proportion, therefore, there is no need to transform these variables. Nevertheless, for other variables, such as total spend, median income families, median income non-families, etc., max-min scaler is used to normalize variables with absolute value range exceeding [0,1] interval by using the following formula:

$$y = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Equation 1 - Min-Max Scaler

where x_{max} and x_{max} values are drawn from the observable variables. Standardization (subtracting the mean from each value and dividing the result by the standard deviation) is not used in this project since it assumes normal distribution of the data. Min-max scaling fit is performed only on training set, in order to avoid information leakage to the validation and test sets. Transformation is applied to all the sets of the data.

3.3. FEATURE SELECTION

Information redundancy, as well as feature relevance, are significant factors in data modeling. It is desirable to reduce the number of input variables both due to the computational costs as well as improving the performance of the model.

It is assumed in the beginning of the training process that the results for decision-based algorithms will not be affected by the features in the dataset since they have feature selection as an embedded process, while for some other models (artificial neural networks) it is expected that the results might be worse. Also, ´no feature selection´ approach serves as a baseline and will demonstrate if some feature selections have weakened or, on the other hand, have strengthened the algorithm performance.

Another approach would be to apply correlation-based feature selection such as Pearson's correlation in order to filter out highly correlated and therefore redundant features. In the current project the threshold of 0.85 was used. This step is applied regardless of any other combination of other techniques and parameters. Pearson's correlation is calculated the following way:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}},$$

Equation 2 - Pearson's correlation

Where r -correlation coefficient, x_i- values of the x-variable in a sample, $\bar{x}$ - mean of the values of the x-variable,y_i - values of the y-variable in a sample, $\bar{y}$- mean of the values of the y-variable.

Mutual information feature selection method selects the top specified % of features with the highest score to be used in the training process, the score is determined by the calculation of mutual information between label/feature.

Mutual information is calculated the following way:

$$I(X;Y) = H(X) - H(X|Y)$$

Equation 3 - Mutual information

where $H(X)$ is the entropy for X and $H(X|Y)$ is the conditional entropy for X given Y.

3.4. Data sampling techniques for imbalanced data

Given that the dataset is poorly balanced for all the three target features in consideration (‘HHIncome’, ‘Hispanic’, ’Ethnicity’), it is vital to address this issue by applying some of the data balancing techniques. Imbalanced data often means that there are few examples of a certain class or classes in order for the model to effectively learn the decision boundary. In our dataset target ‘HHIncome’ has the following distribution: 27%, 22%, 21%,16%, 13% shares of the following levels of household income respectively <$25,000, $25,000-$49,999, $50,000-$74,999, $75,000-$99,999 and $100,000+. Target ‘Hispanic’ is binary with the following distribution of response variable: 86% - non Hispanic origin and 14% - Hispanic origin. And finally, ‘Ethnicity’ target is split the following way: 80%, 9%, 9%, 1.98% and 0.02% share of ‘white’, ‘black’, ’asian/pacific islander’,’native american’ and ‘other’ respectively.

Some of the possible solutions to deal with this problem could be to perform oversampling by duplicating the existing entries or under sampling randomly the existing entries. Both of these approaches are not advisable since the first does not add any new information to the algorithm to learn from while the other is reducing the amount of information contained in the data already. However, data augmentation is one of the solutions and was used in the present project.

Synthetic minority over-sampling technique (SMOTE) is one of the most widely used for balancing the datasets. SMOTE selects examples that are close in the feature space using k-nearest neighbor algorithm. Then a randomly selected neighbor is chosen, and a synthetic example is created at a randomly selected point between the two examples in feature space. Also, ‘no sampling’ technique is applied in order to provide the baseline results for the models in consideration.

3.5. Machine learning methods

3.5.1. Random forest

Random forest algorithm is a stack of decision trees as base predictors combined using bootstrap aggregation and makes predictions based on all the votes of all the base predictors. Decision tree, even though easy to use and interpret if-then algorithm, tends to overfit the training data and its performance is usually inferior then its ensemble variation - random forest.

Random forest is built in the following way: for each tree in the forest the observations are sampled from all the available observations with replacement and the sample size is equal to the main sample size. Further, for each sample the classification or regression decision trees are built and later the prediction for previously unseen observation is made by averaging the result of the trees in the forest or by the majority vote in case of classification decision trees.

The summary of steps in Random forest is the following:

1. Select a random sample of observations with replacement;
2. A set of variables is selected randomly;
3. A variable providing with the best split is selected based on Gini impurity index or Information gain;
4. The tree is continued to be split by choosing the next best variable until the maximum depth is reached;
5. The steps above are repeated until the certain number of trees is built;
6. The prediction is done using the majority vote;

In our project for the best possible split of nodes in decision trees Gini impurity index is used.

The computation of the Gini impurity index for a set of objects with J classes is:

$$Gini = 1 - \sum_{i=1}^{J} {p_i}^2$$

Equation 4 - Gini impurity index

where p_i is the denotes the probability of an element being classified as a distinct class i.
For the calculation RandomForestClassifier from sklearn ensemble library is used with the following default parameters: number of trees in the forest - 100, split is based on Gini impurity index, and the maximum depth is not defined (the nodes are expanded until all leaves are pure or until all the leaves contain less than 2 samples).

3.5.2. Gradient Boosting

Gradient Boosting is an ensemble technique from a family of boosting methods, which unlike other ensemble techniques, like, for example, widely used random forest which averages the results of many classifiers, instead, relies on a different strategy of ensemble formation. The main principle of boosting ensembles is iterative addition of new models to the ensemble. A new weak base-learner is sequentially added and trained with respect to the error of the whole ensemble learnt so far *(Natekin & Knoll, 2013).* In gradient boosting algorithm learning procedure consecutively fits new models to provide a more accurate estimate of the response variable.
The principal idea behind this algorithm is to construct the new base-learners to be maximally correlated with the negative gradient of the loss function, associated with the whole ensemble. At first the size of the ensemble, loss function and base learner are determined. At the next step of the process the estimated value is initialized and later the residuals based on the initial estimated value are calculated and the base learner is built in order to predict those residuals. Later, it is needed find the step value that would minimize the loss function and update the function estimate *(Chen & Guestrin, 2016).*
There are several techniques which could be used to improve the performance of gradient boosting:

1. Tree constraints: it is important to keep the base learners weak. It can be achieved by imposing some constraints such as tree depth, number of nodes or number of leaves, etc.
2. Shrinkage: the learning rate can be slowed down by weighting the contribution of each base-learner.
3. Random sampling: each base learner is created from a random subsample of the data (subsample of variables or subsample of observations).

In this present project XGBoost library was used in order to train and test the Gradient boosting model. In this project the default parameters are the following: the base learner used is a decision tree, the loss function is a SoftMax function and learning rate equal to 0.1.

3.6. Models' performance evaluation criteria

The evaluation of the models was done using the validation set (30 % of the initial training set) and a test set (30 % of the initial dataset). The metrics considered for models' assessment were: Accuracy, F1 score, Precision, Recall, Sensitivity, Specificity and AUC ROC. The metrics are calculated based on confusion matrix which represents a contingency table in which each row represents the instances of the actual class while the columns represent the instances of predicted class. Any binary confusion matrix would have the following outcomes: TP (True Positives) Number of elements belonging to Class1 and that are classified as Class1; TN (True Negatives) Number of elements that don't belong to Class1 and that are not classified as Class1; FP (False Positives) Number of elements that don't belong to Class1 but are classified as Class1; FN (False Negatives) Number of elements belonging to Class1 and that are not classified as Class1; Positives = TP+FN, Negatives = FP+TN and Total Population = P+N.

Table 1 - Binary classification contingency table

		observed	
		1	0
predicted	1	TP	FP
	0	FN	TN

Table 2 - Performance metrics

Metrics	**Formula**
Accuracy	$\frac{TP + TN}{P + N}$
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$
Sensitivity	$\frac{TP}{TP + FN}$
Specificity	$\frac{TP}{TN + FP}$
F1 score	$\frac{2 * Precision * Sensitivity}{Precision + Sensitivity}$
True positive rate	$\frac{TP}{P}$
False positive rate	$\frac{FP}{N}$
AUC ROC	$\frac{1 + TPR + FPR}{2}$

For each response variable different thresholds of the predicted probabilities were used. Initially, the default thresholds were considered and were the following: 0.5 for binary response variable and 0.2 for multilabel variables. However, since we are interested in high probability predictions, higher probabilities were also tested in order to understand the share of high probability predictions, their accuracy and therefore the possibility to use them in case of the poor results within default probability predictions. The threshold for high probability was chosen to be 0.9 for binary variables and 0.4 for multiple classes.

4. RESULTS AND DISCUSSION

4.1. PRESENTATION OF THE RESULTS

The figures below show the results for each target for each model in consideration together with feature selection and data augmentation methods. All the scenarios include data scaling using min-max scaler technique. All the machine learning models presented below have default hyperparameters.

In the table below (Tab. 3) the results for target 'Hispanic' are given for high probability thresholds. For both thresholds the 'XGBoost' model seems to be performing slightly better based on Accuracy, Recall and AUC, however, precision metrics produced with Random Forest are outperforming the XGBoost's precision. Models trained with unbalanced data (14 % minority class, 86% majority class) slightly outperforms the model which was trained with data balanced using SMOTE technique, which does not align with our expectations: the accuracy and AUC ROC is higher.

The application of the feature selection mechanism is not adding any advantages when applied to data used for XGboost: the difference is maximum 1 percentage point across different metrics. However, the positive effect of the feature selection is observed when used together with Random forest classifier (F1 score is 57% compared to 54%). Please see all the figures related to the 'Hispanic' variable in appendix.

Table 3 - 'Hispanic' target high probability results

'Hispanic'; High probability threshold results							
Oversampling	Model	Accuracy	F1 score	Precision	Recall	AUC	Share of high probability predictions; threshold 0.9
unbalanced data	RF	96%	49%	48%	50%	64%	64%
unbalanced data	XGBoost	95%	62%	90%	58%	73%	75%
SMOTE	RF	98%	61%	81%	55%	60%	62%
SMOTE	XGBoost	95%	67%	84%	62%	74%	65%

In the figure below (Fig.1) the most important features for the model used to predict Hispanic identity (Random Forest) are presented. The top features seem adequate and align with our understanding of demographics within the Hispanic group.

The share of high probability predictions (0.9 probability) is quite significant for the 'Hispanic' variable and is in the range of 63-75% of the initial data which is sufficiently high. However, the results for high probability predictions only outperform in terms of precision and accuracy, but fall behind in F1 score, recall and AUC metrics.

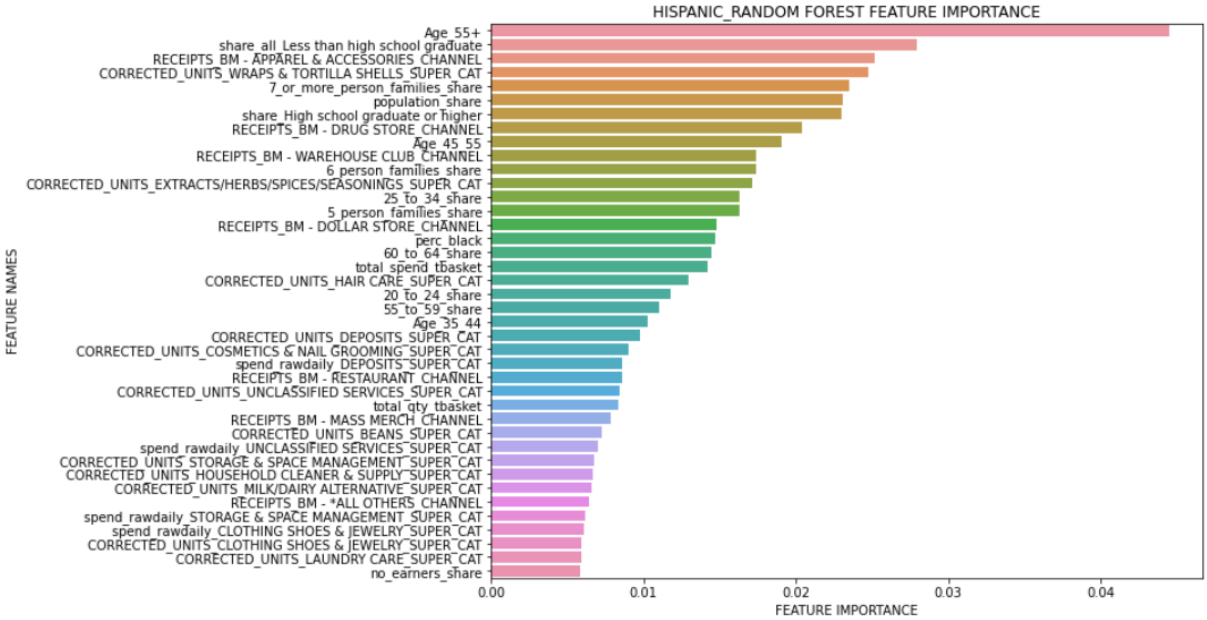

Figure 1 - Feature importance for ‘Hispanic’ variable, Random Forest

Table 3 - ‘Hispanic’ Contingency table, Gradient Boosting

Hispanic						
	predicted					
		non hispanic	hispanic	recall	Total	Predicted
actual	non hispanic	16,084	666	96%	16,750	17,542
	hispanic	1,458	728	33%	2,186	1,394
	precision	92%	52%			

Contingency table (Tab. 4) for the response variable ‘Hispanic’ demonstrates the low ability of a model to make correct predictions for the minority class even for the balanced training dataset.

In table 5, the results for another important variable ‘Household Income’ are presented. Overall, the results for this variable are poor, with recall reaching maximum 30% (XGboost, SMOTE and Mutual Information feature selection) and AUC ROC reaching 59% for variable ‘$75,000-$99,999’ for RF and XGBoost, which is still a very low value and close to random selection. The share of high probability predictions (0.9 prob) is quite low for the ‘Household Income’ variable and is in the range of 7-37% of the initial data and the results are not close to the sufficient values: accuracy barely reaches 53%, and F1 score being 26% (please see the appendix for the results of high probability threshold results).

Table 4 - ‘Household Income’ target default probability results

‘Household Income’, Default probability threshold										
Oversampling	Model	Accuracy	F1 score	Precision	Recall	AUC <$25,000	AUC $25,000-$49,999	AUC $50,000-$74,999	AUC $75,000 -$99,999	AUC $100,000+
none	RF	34%	21%	36%	28%	30%	41%	52%	30%	50%
none	XGBoost	33%	28%	30%	30%	31%	43%	52%	34%	50%
SMOTE	RF	32%	29%	29%	31%	29%	50%	51%	59%	31%
SMOTE	XGBoost	32%	30%	30%	30%	30%	51%	51%	59%	30%

Table 5 - ‘Ethnicity’ target high probability results

‘Ethnicity’, High probability threshold results												
Feature Selection	Oversampling	Model	Accuracy	F1 score	Precision	Recall	AUC White	AUC Black	AUC Asian/Pacific Islander	AUC Native American/Aleut Eskimo	AUC Other	share of high probability predictions; threshold 0.4
MI	none	RF	83%	31%	48%	28%	83%	85%	87%	66%	45%	100%
MI	none	XGBoost	85%	38%	50%	35%	85%	88%	89%	67%	60%	100%
MI	SMOTE	RF	81%	39%	39%	42%	83%	84%	89%	67%	74%	81%
MI	SMOTE	XGBoost	83%	39%	42%	38%	83%	86%	88%	65%	73%	99%

Contingency table (Tab. 7) for ‘Household income’ confirms the figures above and suggests the inability of a model accurately perform multiclass classification.

Table 6 - ‘Household Income’ Contingency table, Gradient Boosting

HHIncome									
actual		predicted							
		<$25,000	$25,000-$49,999	$50,000-$74,999	$75,000-$99,999	$100,000+	recall	Total	Predicted
	<$25,000	657	1,102	304	136	229	27%	2,428	2,125
	$25,000-$49,999	806	2,261	914	486	710	44%	5,177	6,560
	$50,000-$74,999	439	1,484	824	509	938	20%	4,194	3,231
	$75,000-$99,999	223	918	584	414	971	13%	3,110	2,090
	$100,000+	179	795	605	545	1,903	47%	4,027	4,751
	precision	29%	34%	26%	20%	40%			

In Table 6, the results for another important variable ‘Ethnicity’ are presented. Overall, the results for this target variable are adequate, with F1 score reaching 39% (XGboost, SMOTE regardless of feature selection) and AUC reaching 86%, 88%, 89%, 71% and 73% for variables ‘White´, ‘Black’, ‘Asian/Pacific Islander’, ’Native American/Aleut Eskimo’ and ‘Other’ respectively. Most of the predictions for this target have a high probability (the share is in the range of 78-100% of the initial data) and therefore the values of the metrics do not differ much from default probability predictions. XGBoost is performing better than RF, oversampling increases recall and decreases precision; feature selection is having a positive effect on the performance of XGBoost only in terms of AUC metrics: when MI feature selection is not used the results for variables ‘Asian/Pacific Islander’ and ‘Native American/Aleut Eskimo’ drop from 89%, 67% to 61%, 62% respectively.
Similarly to ‘Hispanic’ response variable, ‘Ethnicity’ is largely affected by the majority class and defaults to ‘White’ (Tab. 8)

Table 7 - ‘Ethnicity’, Contingency table, Gradient Boosting

Ethnicity							
actual		predicted					
		white	black	other	recall	Total	Predicted
	white	14,112	490	600	93%	15,202	15,990
	black	893	728	127	42%	1,748	1,329
	other	985	111	890	45%	1,986	1,617
	precision	88%	55%	55%			

In the figure below (Fig.2) the most important features for the model used to predict ‘Ethnicity’ (Random Forest) are presented. The top could theoretically be related to the response variable; however, we don’t see any particular features which could determine the ethnicity, which is aligned with our understanding that variables have a weak relationship with demographic features.

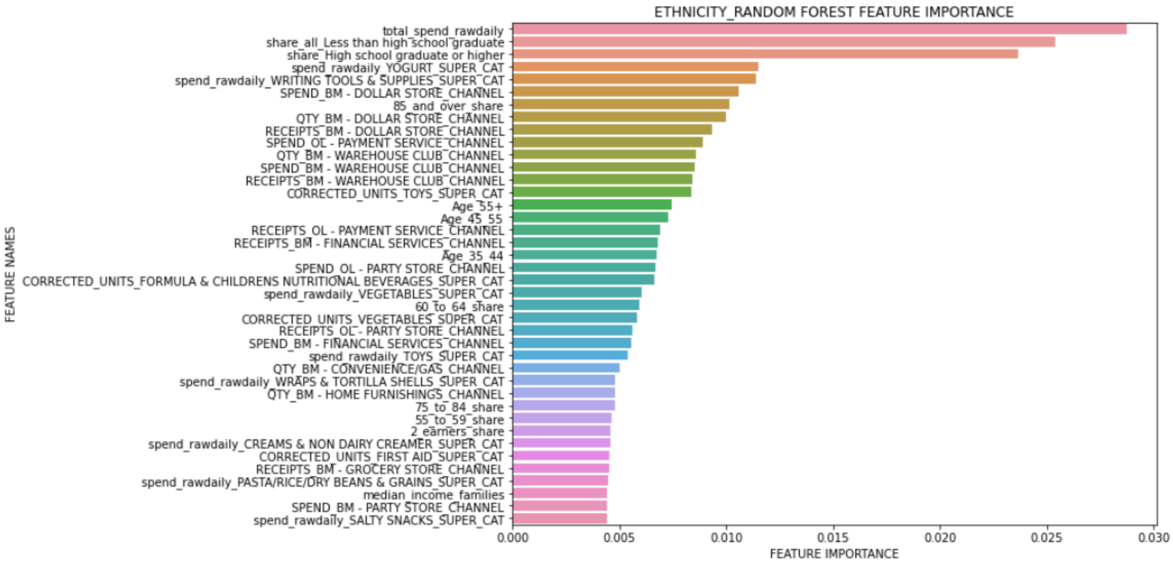

Figure 2 - Feature importance for ‘Ethnicity’ variable, Random Forest

Similarly, top important features are obtained when predicting ‘Household income’ (Fig.3). Despite the poor performance of the model, the top features seem to be logically very well related to the response variable: median income, ‘dollar store’ and ‘warehouse club’ channels purchasing behavior as well as ‘total spent per trip’ to the store.

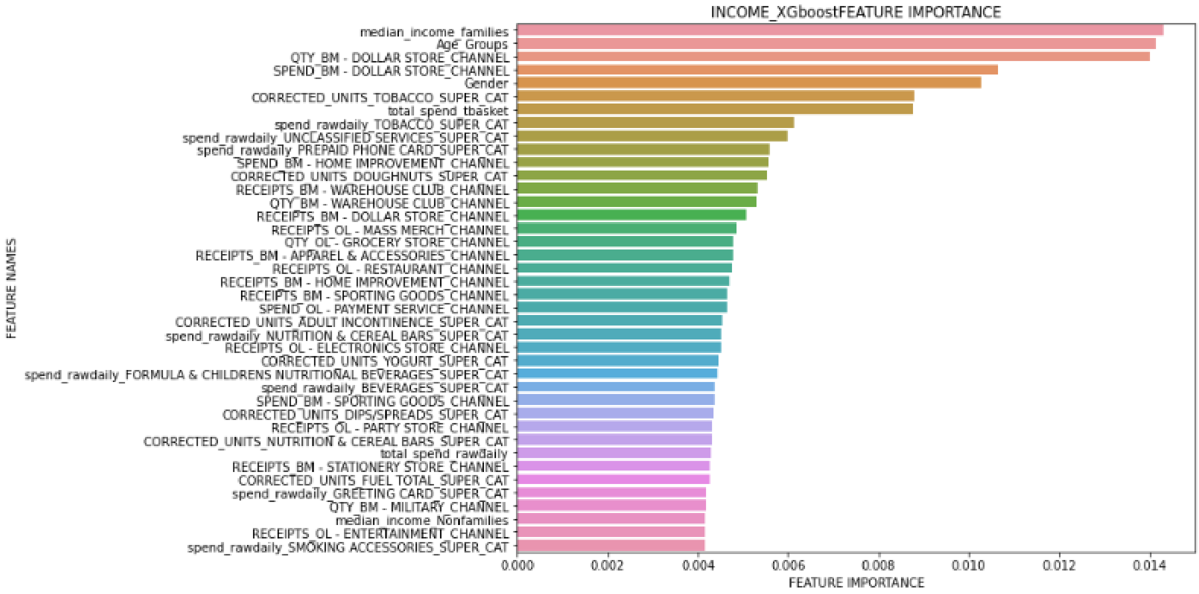

Figure 3 - Feature importance for ‘Household income’ variable, Gradient Boosting

4.2. Discussion of the results

‘Household income’ variable seems to be the most complex among all the targets to predict. This is, most probably, the result of the incorrect data provided by the respondent in the first place and is a common issue in data collection processes. The respondents tend to overstate or understate their income due to various reasons. Also, perhaps due to the arbitrary income breaks we are not able to label the data correctly and it might be the reason for the inaccuracy in predictions we observe.

The results for the targets ‘Ethnicity’ and ‘Hispanic’ are not within the acceptable range and therefore could not be used in production either. The metrics obtained with the default hyperparameters indicate that the best model to predict ‘Hispanic’ identity is XGboost, and Random forest ‘s performance being similarly acceptable and not significantly different. Similarly, there no significant difference between RF and XGBoost models for response variable ‘Ethnicity’. However, the overall performance is poor and even though the accuracy metrics are quite high (around 80-95%) for the ‘Hispanic ‘ and ‘Ethnicity’ response variables, other metrics such as precision, recall, F1 score and AUC ROC suggest that the models cannot be used if it is important to have accurate predictions for all available classes.

Overall, the predicted values obtained via the machine learning process are not reliable enough to be used for the creation of the product. Other possible ways of tackling this problem are discussed in the section of ‘Limitations and recommendations for future work’.

5. CONCLUSIONS

The missing data problem is an indispensable part of any real-world big data collection and therefore needs to be approached with the most advanced techniques of data imputation - machine learning. Even though the ability of ML to produce high quality predictions for the missing parts is non-negotiable, sometimes the problem in accuracy of the model arises due to the quality of available training data and cannot be mitigated without significant changes in the data collection processes or data preparation. The difficulties to predict 'Household income' with adequate accuracy suggest that the training data is not truthful. Another hypothesis is that the income breaks are too rigid in order to accommodate for the complexity of the dependencies. Also, the features which have been used to try to predict the income might not be informative enough for the model to learn. We have observed the inability of a model to predict correctly 'Hispanic' and 'Ethnicity' variables, even though their performance demonstrates better results, predictions still default to the majority class.

Since the primary goal of the present project was to investigate the possibility of prediction of the missing values using machine learning approach and highlight the challenges related to it, we can conclude that at this stage it is impossible to implement the current approach in production without tackling significant challenges and, undoubtedly, further work is needed to eliminate those. And, of course, there are some ways which could mitigate the risk of having missing values in the first place: one of them is to suggest that data collection strategies should be designed differently in order to obtain complete data in the first place or by re-contacting & incentivizing users to provide missing data.

The most significant limitation of the present research is the reliability of the self-reported data (demographic surveys) and lack of instruments to verify the accuracy of the data.

Nevertheless, future work around sources of data and feature selection is required. For example, it might be beneficial to use more granular data (actual products purchased by the user) for training the model. This would inevitably raise dimensionality issue, which would be necessary to address by applying more elaborate feature selection process. There are still other ways of improving the models, for example using more sophisticated feature selection methods and finding ways to clean the data to ensure its reliability during the training process. Potentially, some additional work that can be done on hyper- parameters tuning of the models as well as exploration of the possibility of using Artificial neural networks.

6. REFERENCES

Demirtas, H. (2018). Flexible Imputation of Missing Data. *Journal of Statistical Software*, 85(Book Review 4).

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal Of Big Data*, *8*(1).

Chen, Y.-C. (2022). Pattern graphs: A graphical approach to non monotone missing data. *The Annals of Statistics*, 50(1).

Chinomona, A. and Mwambi, H. (2015). Multiple imputation for non-response when estimating HIV prevalence using survey data. *BMC Public Health*, 15(1).

Baraldi, A.N. and Enders, C.K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1), pp.5–37.

Autor: Andrew Gelman, Hill, J. and Aki Vehtari (2021). Regression and other stories. Editorial: Cambridge Cambridge University Press.

Hughes, R.A., Heron, J., Sterne, J.A.C. and Tilling, K. (2019). Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *International Journal of Epidemiology*, 48(4), pp.1294–1304.

Sterne, J.A.C., White, I.R., Carlin, J.B., Spratt, M., Royston, P., Kenward, M.G., Wood, A.M. and Carpenter, J.R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, [online] 338(jun29 1), pp.b2393–b2393. Available at: https://www.bmj.com/content/338/bmj.b2393.

Wang, H., Tang, J., Wu, M., Wang, X. and Zhang, T. (2022). Application of machine learning missing data imputation techniques in clinical decision making: taking the discharge assessment of patients with spontaneous supratentorial intracerebral hemorrhage as an example. *BMC Medical Informatics and Decision Making*, 22(1).

Sun, Y.V. and Kardia, S.L.R. (2008). Imputing missing genotypic data of single-nucleotide polymorphisms using neural networks. European Journal of Human Genetics, 16(4), pp.487–495.

Lai, W.Y. (2019). A Study on Sequential K-Nearest Neighbor (SKNN) Imputation for Treating Missing Rainfall Data. *International Journal of Advanced Trends in Computer Science and Engineering*, 8(3), pp.363–368.

Gajawada, S. and Toshniwal, D. (2012). Missing Value Imputation Method Based on Clustering and Nearest Neighbours. *International Journal of Future Computer and Communication*, pp.206–208.

Fouad, K. M., Ismail, M. M., Azar, A. T., & Arafa, M. M. (2021). Advanced methods for missing values imputation based on similarity learning. *PeerJ Computer Science*, *7*, e619.

Lin, W. C., & Tsai, C. F. (2020). Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, *53*(2), 1487-1509.

Nikfalazar, S., Yeh, C.-H., Bedingfield, S., & Khorshidi, H. A. (2019). Missing data imputation using decision trees and fuzzy clustering with iterative learning. *Knowledge and Information Systems*, *62*(6), 2419–2437.

Rahman, Md. G., & Islam, M. Z. (2013). Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques. Knowledge-Based Systems, 53, 51–65.

Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *10*(6), 363–377.

Hong, S., & Lynn, H. S. (2020). Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. BMC Medical Research Methodology, 20(1).

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. American Journal of Epidemiology, 179(6), 764–774.

Rubin, D. B. (1976). Inference and missing data. Biometrika.

Oba, S., Sato, M. ., Takemasa, I., Monden, M., Matsubara, K. , & Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. Bioinformatics, 19(16), 2088–2096.

Zhu, X., Wang, J., Sun, B., Ren, C., Yang, T., & Ding, J. (2021). An efficient ensemble method for missing value imputation in microarray gene expression data. BMC Bioinformatics, 22(1).

Khan, S. S., Ahmad, A., & Mihailidis, A. (2019). Bootstrapping and multiple imputation ensemble approaches for classification problems. Journal of Intelligent & Fuzzy Systems, 37(6), 7769–7783.

Cheng, C.-Y., Tseng, W.-L., Chang, C.-F., Chang, C.-H., & Gau, S. S.-F. (2020). A Deep Learning Approach for Missing Data Imputation of Rating Scales Assessing Attention-Deficit Hyperactivity Disorder. Frontiers in Psychiatry, 11.

Mishra, A., Naik, B., & Srichandan, S. K. (2018). Missing Value Imputation Using ANN Optimized by Genetic Algorithm. International Journal of Applied Industrial Engineering, 5(2), 41–57.

M, H., & M.N, S. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process, 5(2), 01-11.

Jin Huang, & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on Knowledge and Data Engineering, 17(3), 299–310.

Brank, J., Mladenić, D., Grobelnik, M., Liu, H., Mladenić, D., Flach, P. A., Garriga, G. C., Toivonen, H., & Toivonen, H. (2011). Feature Selection. Encyclopedia of Machine Learning, 402–406.

Chandrashekar, G. and Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), pp.16–28.

John, G. H., Kovahi, R., & Pfleger, K. (1994). Relevant features and the subset selection problem [Review of Relevant features and the subset selection problem]. Machine Learning: Proceedings of the Seventh International Conference, 121–129.

García-Torres, M., Gómez-Vela, F., Melián-Batista, B., & Moreno-Vega, J. M. (2016). High-dimensional feature selection via feature grouping: A Variable Neighborhood Search approach. Information Sciences, 326, 102–118.

Xuan, P., Guo, M. Z., Wang, J., Wang, C. Y., Liu, X. Y., & Liu, Y. (2011). Genetic algorithm-based efficient feature selection for classification of pre-miRNAs. Genetics and Molecular Research, 10(2), 588–603.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in Neurorobotics, 7.

Chen, T., & Guestrin, C. (2016, June 10). XGBoost: A scalable tree boosting system. arXiv.org. Retrieved October 27, 2022, from https://arxiv.org/abs/1603.02754

APPENDIX

Table 1 The results of the target 'Hispanic'. Default and High probability threshold

Feature Selection	Oversampling	Default probability threshold						High probability threshold					
		Model	Accuracy	F1 score	Precision	Recall	AUC	Accuracy	F1 score	Precision	Recall	AUC	Share of high probability predictions; threshold 0.9
MI	unbalanced data	RF	89%	57%	83%	55%	80%	96%	49%	48%	50%	64%	64%
MI	unbalanced data	XGBoost	90%	65%	77%	62%	81%	95%	62%	90%	58%	73%	75%
MI	SMOTE	RF	87%	65%	63%	55%	80%	98%	61%	81%	55%	60%	62%
MI	SMOTE	XGBoost	88%	66%	68%	65%	78%	95%	67%	84%	62%	74%	65%
none	unbalanced data	RF	89%	54%	84%	54%	80%	96%	49%	48%	50%	64%	63%
none	unbalanced data	XGBoost	90%	64%	76%	61%	81%	95%	63%	91%	58%	73%	75%
none	SMOTE	RF	86%	65%	65%	64%	80%	98%	61%	84%	57%	63%	10%
none	SMOTE	XGBoost	89%	67%	71%	64%	81%	95%	66%	89%	61%	72%	67%

Table 2 The results of the target ‘Household income’. Default probability threshold

Feature Selection	Oversampling	Model	Accuracy	F1 score	Precision	Recall	AUC <$25,000	AUC $25,000-$49,999	AUC $50,000-$74,999	AUC $75,000-$99,999	AUC $100,000+
MI	none	RF	34%	21%	36%	28%	30%	41%	52%	30%	50%
MI	none	XGBoost	33%	28%	30%	30%	31%	43%	52%	34%	50%
MI	SMOTE	RF	32%	29%	29%	31%	29%	50%	51%	59%	31%
MI	SMOTE	XGBoost	32%	30%	30%	30%	30%	51%	51%	59%	30%
none	none	RF	34%	20%	28%	28%	30%	41%	52%	30%	50%
none	none	XGBoost	33%	29%	31%	30%	30%	42%	51%	34%	50%
none	SMOTE	RF	32%	28%	29%	31%	29%	42%	51%	29%	50%
none	SMOTE	XGBoost	33%	30%	30%	31%	30%	43%	50%	32%	50%

Table 3 The results of the target 'Household income'. High probability threshold

Feature Selection	Oversampling	Model	Accuracy	F1 score	Precision	Recall	AUC <$25,000	AUC $25,000-$49,999	AUC $50,000-$74,999	AUC $75,000-$99,999	AUC $100,000+	share of high probability predictions; threshold 0.4
MI	none	RF	51%	25%	30%	31%	71%	31%	58%	58%	47%	7%
MI	none	XGBoost	40%	30%	33%	34%	22%	66%	44%	53%	62%	37%
MI	SMOTE	RF	50%	26%	39%	38%	21%	72%	40%	66%	53%	6%
MI	SMOTE	XGBoost	38%	32%	33%	35%	22%	53%	52%	61%	25%	37%
none	none	RF	53%	23%	32%	28%	36%	83%	53%	38%	49%	5%
none	none	XGBoost	39%	30%	34%	34%	22%	66%	56%	29%	63%	37%
none	SMOTE	RF	52%	26%	20%	38%	24%	86%	57%	27%	49%	5%
none	SMOTE	XGBoost	39%	32%	34%	36%	66%	24%	42%	53%	62%	37%

Table 4 The results of the target ‘Ethnicity. Default probability threshold

Feature Selection	Oversampling	Model	Accuracy	F1 score	Precision	Recall	AUC White	AUC Black	AUC Asian/Pacific Islander	AUC Native American/Aleut Eskimo	AUC Other
MI	none	RF	83%	31%	48%	28%	83%	85%	87%	66%	45%
MI	none	XGBoost	85%	38%	48%	35%	85%	88%	89%	67%	62%
MI	SMOTE	RF	75%	36%	35%	40%	80%	82%	87%	69%	73%
MI	SMOTE	XGBoost	83%	39%	41%	38%	83%	86%	88%	66%	72%
none	none	RF	82%	28%	48%	26%	84%	85%	62%	71%	60%
none	none	XGBoost	85%	39%	52%	35%	86%	88%	61%	62%	66%
none	SMOTE	RF	74%	35%	34%	39%	79%	83%	86%	61%	58%
none	SMOTE	XGBoost	83%	39%	42%	38%	83%	87%	88%	65%	57%

Table 5 The results of the target 'Ethnicity. High probability threshold

Feature Selection	Oversampling	Model	Accuracy	F1 score	Precision	Recall	AUC White	AUC Black	AUC Asian/Pacific Islander	AUC Native American/Aleut Eskimo	AUC Other	share of high probability predictions; threshold 0.4
MI	none	RF	83%	31%	48%	28%	83%	85%	87%	66%	45%	100%
MI	none	XGBoost	85%	38%	50%	35%	85%	88%	89%	67%	60%	100%
MI	SMOTE	RF	81%	39%	39%	42%	83%	84%	89%	67%	74%	81%
MI	SMOTE	XGBoost	83%	39%	42%	38%	83%	86%	88%	65%	73%	99%
none	none	RF	82%	28%	48%	26%	84%	85%	62%	71%	60%	100%
none	none	XGBoost	85%	39%	52%	35%	86%	88%	61%	62%	66%	100%
none	SMOTE	RF	81%	39%	37%	42%	83%	86%	89%	59%	58%	78%
none	SMOTE	XGBoost	83%	40%	42%	38%	83%	87%	88%	64%	57%	99%