MEF UNIVERSITY

RETENTION PERIOD PREDICTION FOR PENSION POLICIES

Capstone Project

Ömer Bayır

İSTANBUL, 2019

MEF UNIVERSITY

RETENTION PERIOD PREDICTION FOR PENSION POLICIES

Capstone Project

Ömer Bayır

Advisor: Asst. Prof. Evren Güney

İSTANBUL, 2019

MEF UNIVERSITY

Name of the project: Retention Period Prediction For Pension Policies
Name/Last Name of the Student: Ömer Bayır
Date of Thesis Defense: 09/09/2019

I hereby state that the graduation project prepared by Your Name (Title Format) has been completed under my supervision. I accept this work as a “Graduation Project”.

09/09/2019
Asst. Prof. Evren Güney

I hereby state that I have examined this graduation project by Your Name (Title Format) which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

09/09/2019
Prof. Dr. Özgür Özlük

Director
of
Big Data Analytics Program

We hereby state that we have held the graduation examination of __________ and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Asst. Prof Evren Güney	………………………..
2. Prof. Dr. Özgür Özlük	………………………..

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Ömer Bayır 09/09/2019 Signature

EXECUTIVE SUMMARY

RETENTION PERIOD PREDICTION FOR PENSION POLICIES

Ömer Bayır

Advisor: Asst. Prof. Evren Güney

September, 2019, Number pages (e.g. 45 pages)

Customer Retention in Pension market refers to the activities and actions companies and organizations take to reduce the number of customer defections. How long the customer will be with our company or will stay in the system is retention.

There are already workings in my company and other companies in the market about customer retention. Existing works generally contains how to measure customer retention and how to define distribution channels are successful in customer retention. Also existing predictive models are working on the feature set customer fund, total collection, un-paid premium frequency in general.

In pension market companies have small margin of profit from pension policies. To make a profit from pension policies the companies have to retain their customer for long years. It 's approximately nine year to make profit from a pension policy because of high sales costs. Therefore to gain a new customer is less profitable than retaining present customers in Pension Market.

In my project, I want to look retention in the pension application phase of customer. My main purpose is when the customer applied for pension product predict its retention period. If I produce an applicable model, It will be used in my company's sales channels.

Key Words: Pension, Retention, classification , multi classification

ÖZET

RETENTION PERIOD PREDICTION FOR PENSION
POLICIES

Ömer Bayır

Tez Danışmanı: Asst. Prof. Evren Güney

EYLÜL, 2019, sayfa sayısı (ör. 45 sayfa)

Bir Emeklilik Şirketi için yeni sözleşme/poliçe satışları ve çıkışlar şirket karlılığını etkilemektedir. Eğer müşteriler uzun süre şirkette kalırlar ve düzenli prim öderler ise şirketler emeklilik sisteminden bekledikleri faydayı elde edebiliyorlar.

Bu noktada daha satış aşamasında yapılan satışın kalitesini belirlemek önemli hale geliyor. Doğru müşteri ye doğru ürün satılması sözleşmenin emeklilik şirketindeki yaşam süresini etkiliyor. Satış ekiplerinin kazanç oranlarını artırmak üzere yaptıkları, sisteme ve şirket için fayda yaratmayan satışları belirlemek veya kazanç sisteminin bir parçası olarak satış aşamasında satışın kalitesini bir değişken olarak değerlendirmek emeklilik şirketlerinin karlılığını artıracak ve doğru müşteriye doğru ürün eşleşemesini destekleyerek müşteri memnuniyeti oluşturacatır. Bu amaçla sözleşmenin satış aşamasında retention periyodunu belirleyecek bir model oluşturmak amaç edinilmiştir.

Anahtar Kelimeler: Bireysel Emeklilik, Poliçe Yaşam Süresi, Çoklu Sınıflandırma

TABLE OF CONTENTS

.

1. INTRODUCTION

A Customer Retention program is in progress in one of the most famous Turkey's Pension Company. This program's scope contains every activities and actions to reduce the number of customer defections as every company does. Retention depicts how long the customer be with our company or stay in the system. Also, the activities to retain customers means retention.

To increase revenue and profitability every pension company needs to its customer stay with them and they want their customer to pay regularly. New customer accusation has high costs. Pension systems depends on large number law. Pension policies has low profit margin. For the companies in the market it is important to reduce operational cost because it is a scale economics market.

If the pension products penetrate enough and much more people enter to the system it will give the benefits as expected from it. Turkey pension system will penetrate more in the future because Turkey Government have been making new regulation for that aim.

Because of its low profitability, pension companies are trying to find new methods that how retain their customer much more with them. Campaigns to reduce churn and retention programs have been improving. Analytical models also have been running.

All models or programs are concentrating to predict churn for existing customers or tries to define customer retention period. These definitions contain KPI's to measure how salesperson or channels are successful to retain their customer. All the programs and studies generate valuable outcomes.

Besides the existing studies, there is an idea to define a pension policy's retention before the customer signs the proposal. The company needs to predict retention period when a customer applies to them or a salesperson find a customer to sell a pension policy. When it is successful company may have chance to deny non-profitable policy/contract or the policy's commission rates will be arranged depending on its profitability.

Predicting a policy how many years will stay in the system needs to know your customer and sales person much more. This project will use the existing structural data and will predict the retention period. According to program scope this approach will develop and at the next phases the models which will be developed in the capstone project will be use behavioral data also.

Predicting a pension contracts' proposal retention period has big challenges with existing conditions. Data quality and completeness of the data is low for proposal of pension contracts. Because of those challenges, the project divided into phases. Only the first phase of the company's project will be in the scope of this capstone project.

First phase handles sales made to an existing customer. If a customer has life insurance, pension policy or auto enrolment, it is an existing customer. If sales made to a customer and she is an existing customer the model will predict the new policy retention period.

This case covers much more cancelation or exit status when compared to other cases will be hold sooner.

The other challenge complexity of the insurer company's distribution channels. For the first phase, sales channels' complexity will not be will be in scope. Other channels will take into consideration next phases.

2. LITERATURE REVIEW

[1]Montserrat Guillén Estany Ana Maria Pérez-Marín Manuela Alcañiz(2018) summarizes why retention studies important for insurance companies as below :
Insurance companies in a more competitive environment than they in the past. Customers easily switch from one insurer to another or cancellations, lapses or exiting from the system have become one of the factors influencing the level of risk an insurance company and its position in the market. Now, the central problem for insurance companies is not only to create and launch new products for the market, but additionally to achieve commercial success by retaining customers. In the past the insurance business was only product-oriented, but now they must be also customer-oriented

[1]Montserrat Guillén Estany Ana Maria Pérez-Marín Manuela Alcañiz(2018) held profit loss in insurance market and they named the loss caused from unretained customers as business risk. They propose to divide every customer profit in three dimension which they are Historical profit, Prospective profit, Potential profit. Although their research was not about pension market their model built in the paper will give new ideas to me or my company. The model defines how to calculate profit loss due to the cancelation or lapses. The researcher built a logistic regression model to predict cancel or not.

[2]Leo Guelmana, Montserrat Guill´enb,∗, Ana M. P´erez-Mar´ınb(2015) tries to show that causal conditional inference trees and its natural extension to causal conditional inference forests can provide a solution to the purpose of selecting the best targets for cross-selling an insurance product. They also tells its usefulness in insurance pricing and retention.

[2]Leo Guelmana, Montserrat Guill´enb,∗, Ana M. P´erez-Mar´ınb(2015) they defined a model named "causal conditional inference trees". It predicts personel treatment as respond to retention or cross-sell activities. This tree based model prevents overfiting without pruning or cross-validation.

[3]Lawrence Ang and Francis Buttle Macquarie Graduate School of Management, Macquarie University, Sydney, Australia, Received (2005) paper tolds about customer retention below :

A top priority in any business is a constant need to increase revenue and profitability. One of the causes for a decrease in profits is when current customers stop transacting. When a customer leaves or churns from a business, the business loses the opportunity for potential sales or cross selling. When a customer leaves the business without any form of advice, the company may find it hard to respond and take corrective action. Ideally companies should be proactive and identify potential churners prior to them leaving. Customer retention has been noted to be less costly than attracting new customers. By analysing the data analytics, companies may analyse customer behavioural patters and gather insight on their customers. These insights will help to identify profitable customers and improvements in their business process thereby increasing customer retention.

[4]Enhancing Customer Retention Through Data Mining Techniques (2017) from the paper the researchers madea model to predict a supermarket customers churn. They uses Logistic Regression and Random Forest algorithms. They tell how they transformed and load the data. Their definitions about a customer life cycle is meaningful for my project because I want to build a model at the stage of Acquisiton but will define Retention stage.

There can be five stages of customer's life cycle in an organization:

Acquisition: Winning the prospect; making of a prospective customer.

Retention: Keeping the customer to obtain the economic benefit of a long term Relationship

Attrition: Breaking down of loyalty; unfulfilled demand; problems and complaints causing customer to reduce or terminate purchase

Defection Ending the relationship; the customer has gone to a competitor for product or services

Reacquisition Getting the customer back; new initiatives or problem correction resulting in the customer coming back to the company.

[5]Customer Retention Management In The Informatıon Era(2001) the paper underlines acquiring the right customers is highly important. My project's aim is also defining right customer. The documents important ideas are below as a summary:

Understanding customer retention is extremely important to the entire direct model of doing business with consumers. The secret to good customer retention is to acquire the right customers in the first place. So understanding customer retention is extremely important to the entire direct selling model of doing business with consumers, both for customer acquisition and retention. Good retention marketers have two objectives with any kind of customer retention marketing:

1. Hold on to the most valuable customers

2. Try to make fewer valuable customers more valuable

To retain and increase the value of customers, you have to create marketing promotions and execute them. To do this in the most efficient and effective way, you have to know the value of your customers and their likelihood to respond to a promotion, for these two reasons:

1. You don't want to waste money on promoting to low value customers because you can't make a profit.

2. You don't want to waste money promoting to customers who won't respond because this is just throwing money away.

At a time when the proactive companies are thinking to remain out of red and maintain a competitive edge it has to emphasize on Customer Retention Marketing. It's not a new paradigm; it's simply a better one. It's a path to choose with the added benefit that the map of the path is easy to follow.

[6] How to Project Customer Retention , Peter S. Fader Bruce G.S. Hardie1, May (2006) contains mathematicial models that defines distribution of customer lifetimes is that of the survivor function. It calculates the probability a customer has "survived" to time t. Retention rate, it's churn rate construct a model.

[7] Random Forests For Uplift Modeling: An Insurance Customer Retention Case(2011) they models of customer churn and predicts the probability of a customer switches to another company. They propose a new procedure that can be used to identify the target customers who are likely to respond positively to a retention activity.

They used random forests to anticipate the success of marketing actions aimed at reducing customer attrition.

3. INSURANCE TERMS

3.1. Pension Proposal Process

The process before contract is approved by the company and contract first contribution is payed is sales and proposal process in the pension insurance terminology. At that process sales persons completes all sales process and tasks and the customer signs proposal document and assures that he/she will pay the first contribution and contract will begins. Insurers’ responsibility begins at that point.

Customer has a right to cancel his contract in the first three months. This period is named as Cancelation Period and this type of contracts are cancelled contracts.

After cancelation period if customer wants to end his contract it is named as passive contract.

3.2. Distribution Channels and Sales Force

Insurers’ have different type customer type and channels. They have different type of process and products. Sales force in that channels meets customers with company’s products. Every channel has different type of benefit and commissioning system. Pension contract can be sold as Direct Sales, Bancassurance, Auto Enrolment, Group Contracting in Turkey’s Pension market. Due to the having different regulation or rules they can be different products. Auto Enrolment, Group Contracting don’t need a retention system depends individuals’ behaviors.

4. PROJECT DEFINITION AND TAGETS

The project is below to the company’s’ retention project and has four phases. First two phases contain existing customers and other two phases contain new customers. This capstone project covers the first two phases which related to existing customers. The first two phase will give intuition to the company about how to handle retention at contract proposal phase and will evolve from gatherings achieved in that project.

Therefore, analysis done, models built in the capstone covers existing customer which has a product in any branch (life, credit life, auto enrolment or second pension contract) in the company. The project covers also customer has a product in the company, but it is not active.

If a pension contract will cancel or not affects the retention analysis. First two phase contains below two period and scope

- Modeling the period begins with beginning of the contract and ends with end of cancelation period
- Modeling the period begins with end of cancelation period

In the project first two period will be hold with the assumptions below:

- Grouping contracts and auto enrolment contracts have not been included due to the need for different type of retention perspective
- Analysis and models made with a data set contains the last three years to reduce data size
- Due to the limited time new data sources or actions will to improve data quality so that models, built in the project, quality increases have not been included. They will be given as a recommendation list to the company.
- These models will be used as embedded analytic engine in the pension proposal process but now the needs are batch monitoring taking actions in a batch process. So that algorithms performance has not included as a target.
- Features about distribution and sales channels requires a different analytical modeling. To simplify project scope it was not included.

Having a two different type of process which first one is cancelation period and the second after cancelation period. Having two different type process two different model have been construct.

I. Target variable is “canceled” or “Not”. A binary classification model built

II. Target variables depicts which interval the contracts leave the company or exits. Required intervals 3-18 months, 19-30 months, 30+ months. Three classes are defined. A multi classification model has been built.

Model-I requires to predict if the contract will cancel or not. It is important not to say a contract will cancel and it will not cancel. In addition, saying it will not cancel and it cancel is important for business. F1 score is the target KPI for the first model. Model-II requires predicting true class. Accuracy is the target KPI for that model.

5. EXPLORATORY DATA ANALYSIS

5.1. Data

Data ingestion and wrangling was done with Oracle and SQL. Required data was in company's transaction system and data warehouse. All required data loaded to the data warehouse and data wrangling was done with Oracle PL/SQL to load a target table. Final feature columns set are below:

Feature	**Type**	**Definition**
RISKPUAN	float64	A questionnaire hold on in the proposal process to measure risk tendency of the customer and suggest fund type
YILLIK_KATKI_TUTAR	float64	The contribution will be paid for annually for the contract
SOZLESME_SURE	int64	Expected time for the retirement as year
AKTIF_BES_SAYISI	int64	Active Pension contract count
PASIF_BES_SAYISI	int64	Ended or passive pension contract count
AKTIF_HAYAT	int64	Active Life contract count
PASIF_HAYAT	int64	Ended or passive pension contract count
AKTIF_OK_SAYISI	int64	Active Auto Enrolment contract count
PASIF_OK_SAYISI	int64	Ended or passive Auto Enrolment contract count
NUMBER_OF_CHILDREN	int64	Children count
GELIR	float64	Yearly income fed on other analytic model

TARGET	Categorical	Target variable, defines cancelation or retention period
ODEME_DONEM	Categorical	Payment period; monthly, yearly..
ENDEKS_TERCIH	Categorical	Contribution increment applied every anniversary of contract
FON_DAGILIM_TIP	Categorical	Fund combination type chosen by customer
TERCIH_RISK_TIP	Categorical	After risk assessment the choice of customer
ONERILEN_RISK_TIP	Categorical	After risk assessment suggested risk type to customer
SIGORTA_ETTIREN_TIP	Categorical	Insured person type
LEHDAR_TIP	Categorical	Endorsee type
KATILIMCI_ODEYEN_ESITMI	Categorical	Insured person and contract owner are the same or not
BOLGE_KOD	Categorical	Sales region who sold the contract
ODEME_SEKLI	Categorical	Payment type; debit, credit card..
DUZENLI_ODEME_GUN	Categorical	Contribution payment day
DAGITIM_KANAL_KOD	Categorical	Distribution channel which Sales region belongs to
NATIONALITY_CODE	Categorical	Nationality Code of the insured customer
GENDER	Categorical	Gender of the insured customer
MARITAL_STATUS	Categorical	Marital Status of the insured customer

EDUCATION_LEVEL	Categorical	Education Level of the insured customer
SECTOR	Categorical	Sector of the insured customer
PROFESSION	Categorical	Profession of the insured customer
DGR	Categorical	Existing segment value of the insured customer which fed from another analytical model
POT_DGR	Categorical	Potential segment value of the insured customer which fed from another analytical model
YAS_KAT	Categorical	Age interval of the insured customer

Table 1: Features list to be used

FON_DAGILIM_TIP, ONERILEN_RISK_TIP, BOLGE_KOD, POT_DGR features are categorical variables and they have missing values. Their missing value imputation was done with SQL and they imputed as most frequent values in the same feature column.

5.2. Explanatory Data Analysis

Descriptive statistics gives the information, RISK_PUAN has missing values and skewness exists on most columns. A pair plot chart given below plotted with raw data to show an overview how the features related to each other with an eye look. From this plot, it seems that most of the features have skew data distribution and this requires making a transformation.

	RISK_PUAN	YILLIK_KATKI_TUTAR	SOZLESME_SURE	AKTIF_BES_SAYISI	PASIF_BES_SAYISI	AKTIF_HAYAT
count	345.668	346.677	346.677	346.677	346.677	346.677
mean	48	3.768	18	0,646628	0,476086	0,588562
std	10	4.648	8	0,996431	1	1
min	0	1.560	9	0	0	0
25%	50	1.800	10	0	0	0
50%	50	2.400	16	0	0	0
75%	50	3.000	24	1	0	1
max	96	600.000	38	13	23	20

Table 2: Descriptive statistics about numeric features

	PASIF_HAYAT	AKTIF_OK_SAYISI	PASIF_OK_SAYISI	NUMBER_OF_CHILDREN	GELIR
count	346.677	346.677	346.677	346.677	346.677
mean	2	0,042904	0,08906	0,027591	8.845
std	3	0,213362	0,321312	0,246911	8.694
min	0	0	0	0	3.375
25%	0	0	0	0	4.181
50%	0	0	0	0	5.221
75%	2	0	0	0	9.435
max	48	5	12	11	67.600

Table 2: Descriptive statistics about numeric features

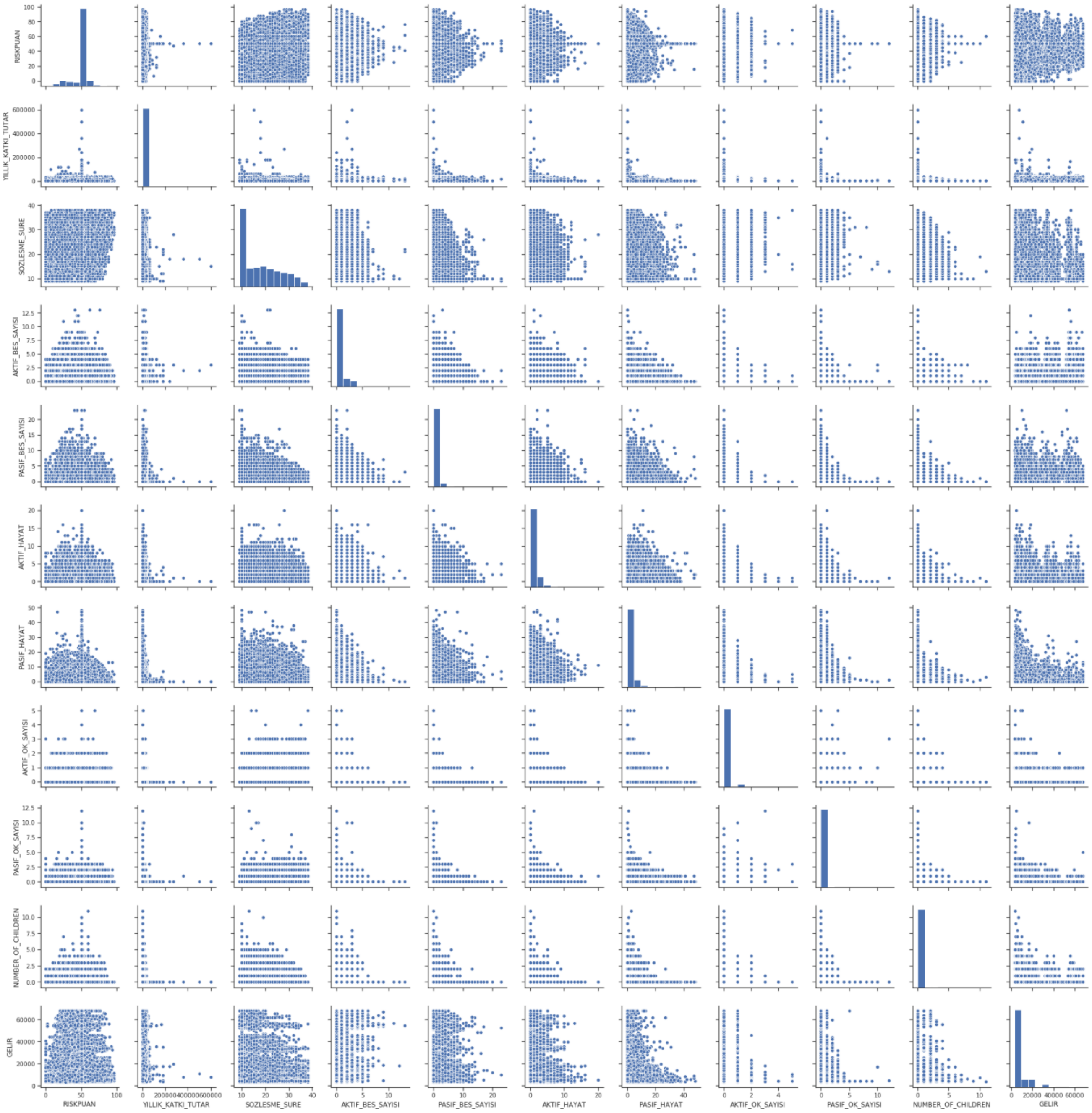


Figure 1: Pair plot for all numeric features

If we do not make a transformation for skewed data, algorithms results will not be better due to the difference from Gaussian Normal Distribution. Skewed data has side effect for machine learning algorithms because most of machine learning algorithm needs all variables have same variance and it is named as Homoscedasticity.

Box-cox transformation could not be used due to the do not having strictly positive values. Yeo-Johnson and log transformation has been tried. Before doing these transformations data distribution is checked with box-plot and violin plot it is shown in Figure 2.

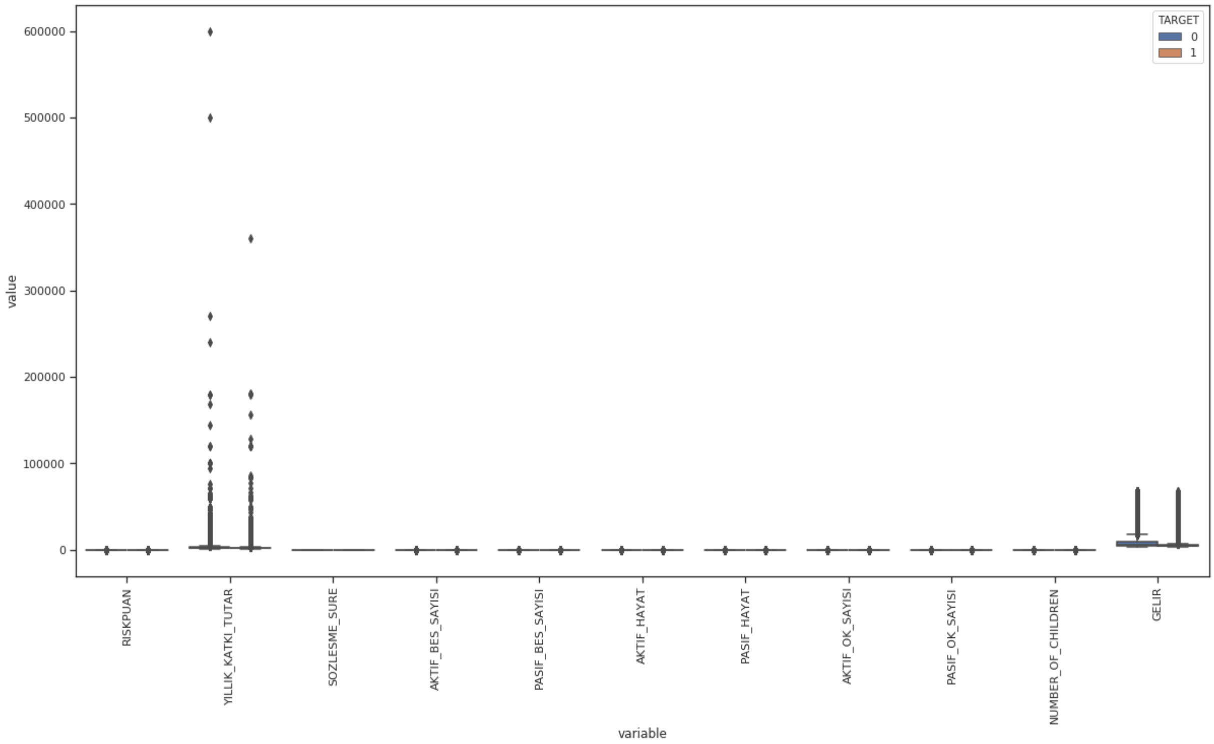


Figure 2: Box plot for numeric features without making any transformation

The steps have been repeated (make transformation and scaling, check with box plot and violin plot) to reach most similar normal distribution. Three methods have been applied (scaling, Log transformation and scaling, Yeo-Johnson and scaling) and the best result was Yeo-Johnson and scaling. Charts are in Figure 3-1 and Figure 3-2 below show the distributions.

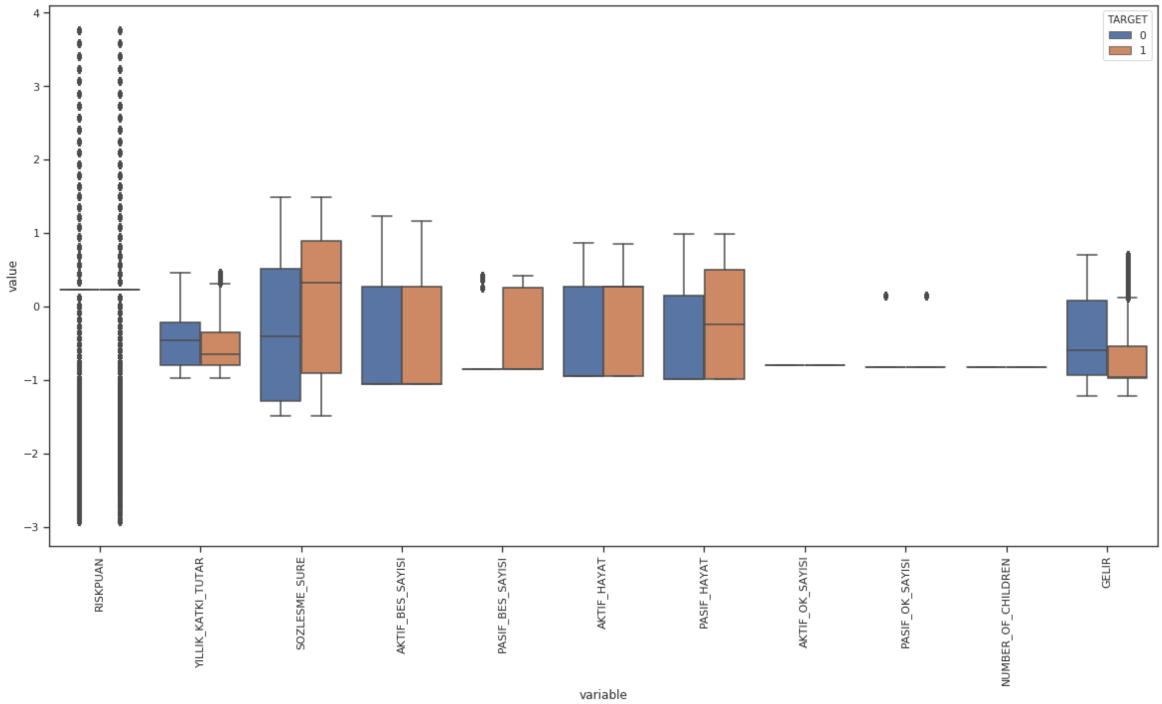


Figure 3-1: Box plot for numeric features after transformation and scaling

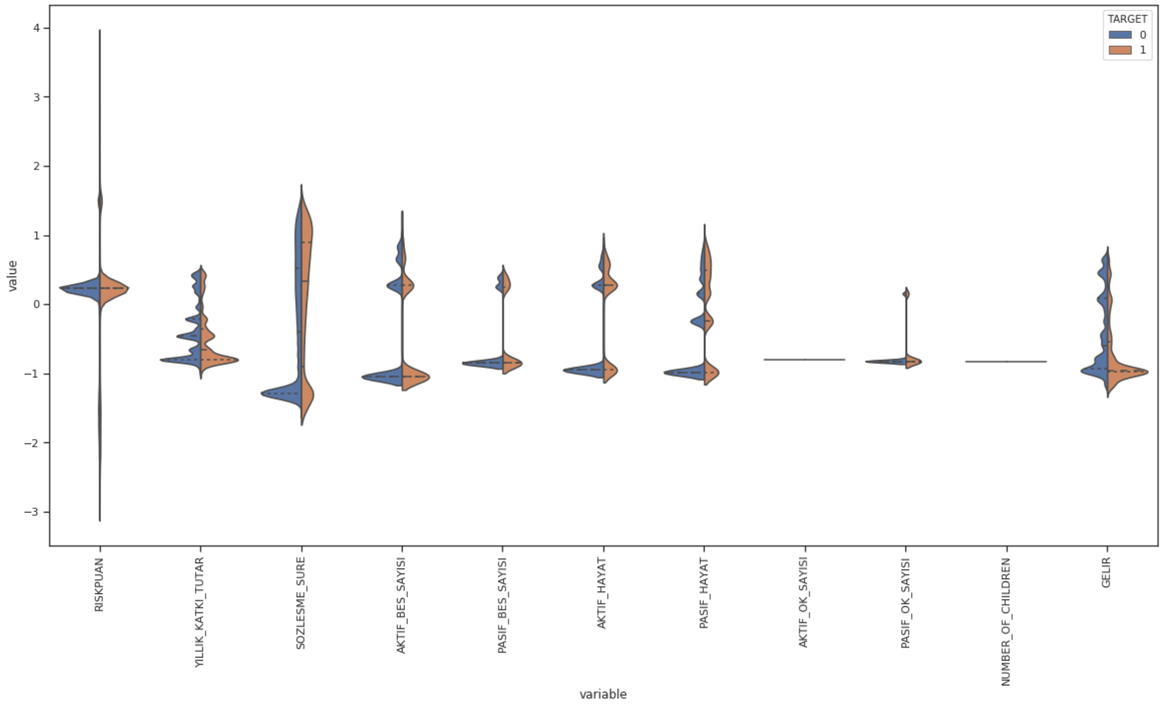


Figure 3-2: Violin plot for numeric features after transformation and scaling

In addition, that plots show RISK_PUAN, GELIR, PASIF_OK_SAYISI features have outliers.

For the Model-I above explanatory data analysis made to infer relations about target and features or between features.

- YILLIK_KATKI_TUTAR may have relation so that below plots checks if it exists. There is no significant effect GELIR and RISK_PUAN has on YILLIK_KATKI_TUTAR.

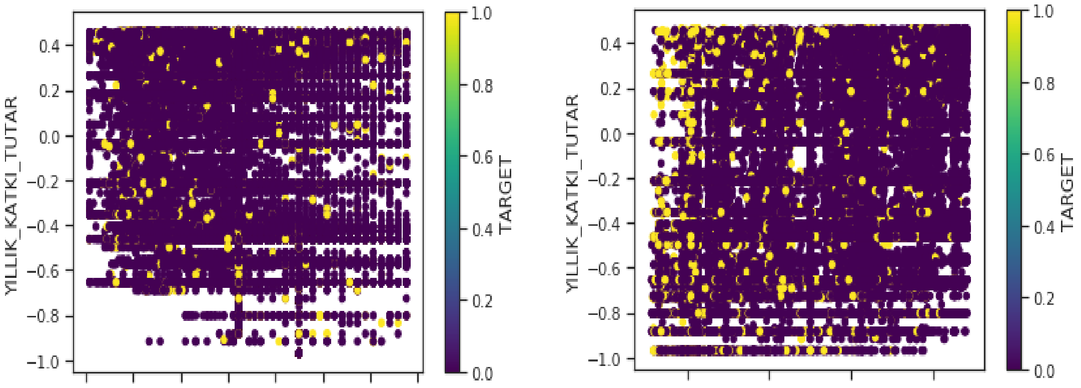


Figure 4: YILLIK_KATKI_TUTAR – RISKPUAN – GELIR and TARGET(Cancelation) Distribution

- AKTIF_BES_SAYISI and PASIF_BES_SAYISI; AKTIF_HAYAT and PASIF_HAYAT with TARGET(Cancelation) doesn’t have significant relation.

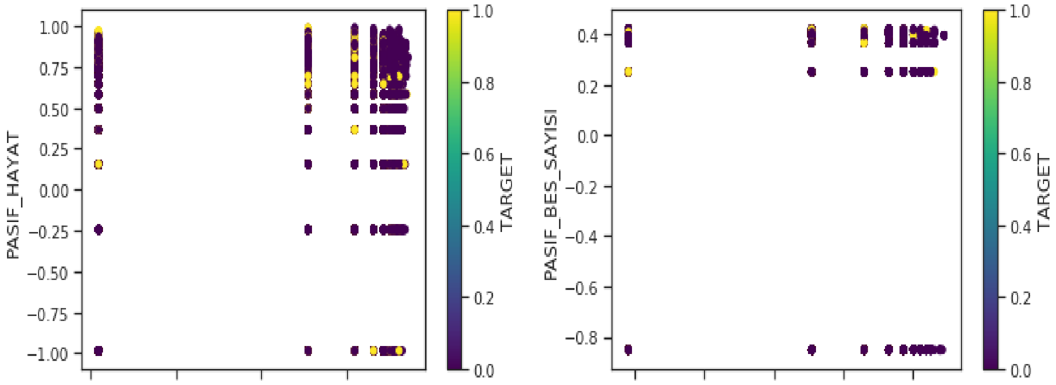


Figure 5: AKTIF_BES_SAYISI and PASIF_BES_SAYISI; AKTIF_HAYAT and PASIF_HAYAT with TARGET(Cancelation) Distribution

- Canceled contract and active contact counts distribution on BOLGE_KOD (Region) chart shows that there is not a region have the majority or canceled contracts.

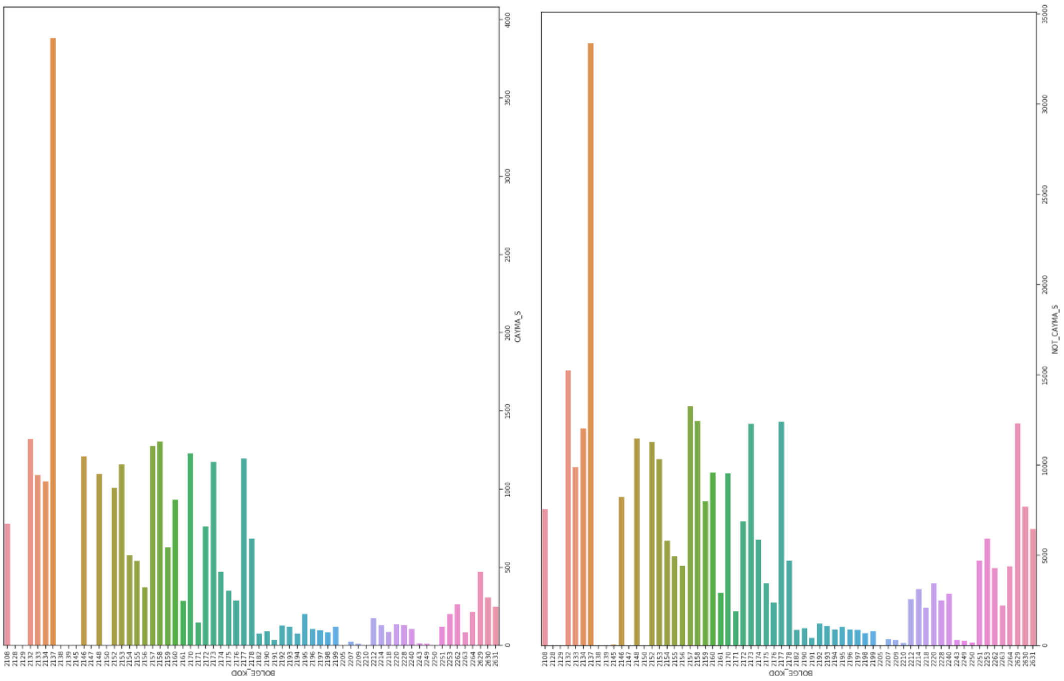


Figure 6: Canceled and active contacts (TARGET) over Regions (BOLGE_KOD) Distribution

- Regression plots for YILLIK_KATKI_TUTAR and GELIR; AKTIF_URUN_S(AKTIF_BES_SAYISI+ AKTIF_HAYAT+ AKTIF_OK_SAYISI) and YILLIK_KATKI_TUTAR;

PASIFF_URUN_S(.PASIF_BES_SAYISI + PASIF_HAYAT+ PASIF_OK_SAYISI) and YILLIK_KATKI_TUTAR shows again data does not contain a linear relation between these variables.

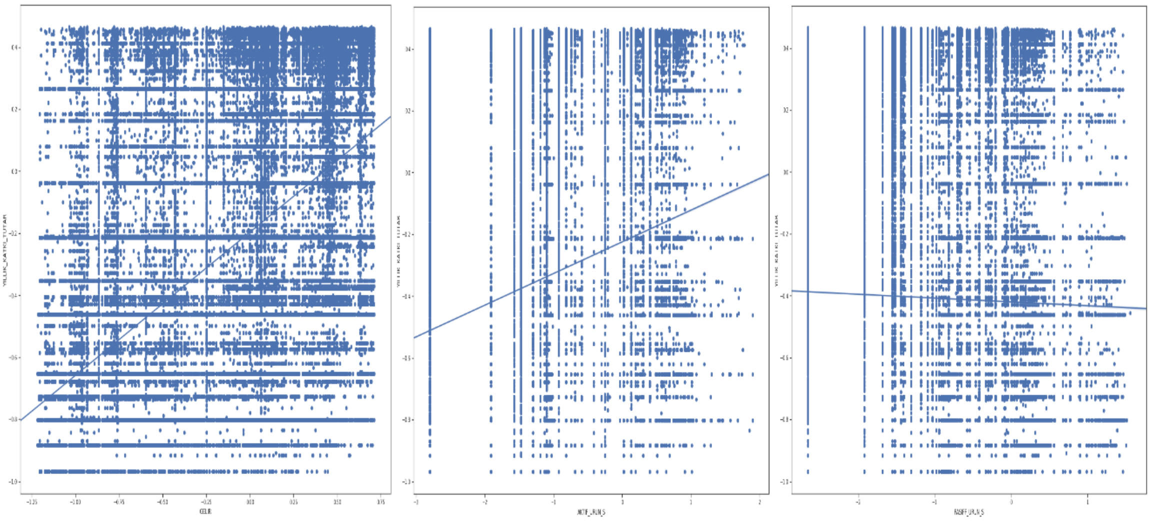
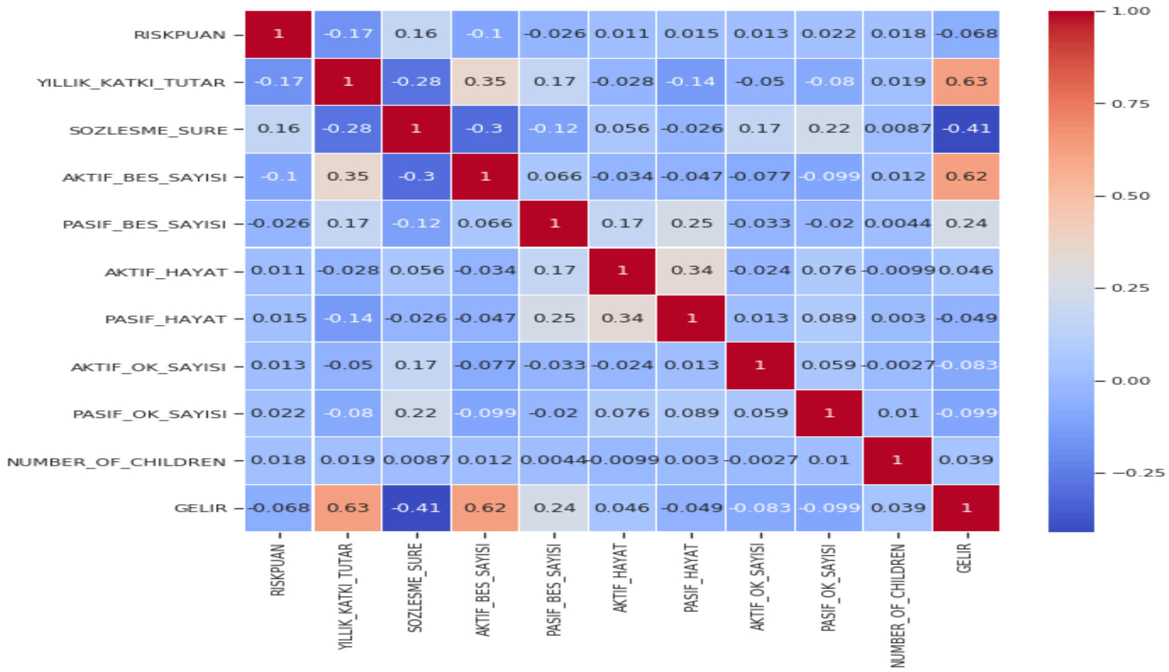


Figure 7: Regression plots for AKTIF_URUN_S, PASIF_URUN_S, YILLIK_KATKI_TUTAR and GELIR

- Heatmap graph for Correlation Matrix below show that YILLIK_KATKI_TUTAR and AKTIF_BES_SAYISI are correlated with a medium level.



	RISKPUAN	YILLIK_KATKI_TUTAR	SOZLESME_SURE	AKTIF_BES_SAYISI	PASIF_BES_SAYISI	AKTIF_HAYAT	PASIF_HAYAT	AKTIF_OK_SAYISI	PASIF_OK_SAYISI	NUMBER_OF_CHILDREN	GELIR
RISKPUAN	1	-0.17	0.16	-0.1	-0.026	0.011	0.015	0.013	0.022	0.018	-0.068
YILLIK_KATKI_TUTAR	-0.17	1	-0.28	0.35	0.17	-0.028	-0.14	-0.05	-0.08	0.019	0.63
SOZLESME_SURE	0.16	-0.28	1	-0.3	-0.12	0.056	-0.026	0.17	0.22	0.0087	-0.41
AKTIF_BES_SAYISI	-0.1	0.35	-0.3	1	0.066	-0.034	-0.047	-0.077	-0.099	0.012	0.62
PASIF_BES_SAYISI	-0.026	0.17	-0.12	0.066	1	0.17	0.25	-0.033	-0.02	0.0044	0.24
AKTIF_HAYAT	0.011	-0.028	0.056	-0.034	0.17	1	0.34	-0.024	0.076	-0.0099	0.046
PASIF_HAYAT	0.015	-0.14	-0.026	-0.047	0.25	0.34	1	0.013	0.089	0.003	-0.049
AKTIF_OK_SAYISI	0.013	-0.05	0.17	-0.077	-0.033	-0.024	0.013	1	0.059	-0.0027	-0.083
PASIF_OK_SAYISI	0.022	-0.08	0.22	-0.099	-0.02	0.076	0.089	0.059	1	0.01	-0.099
NUMBER_OF_CHILDREN	0.018	0.019	0.0087	0.012	0.0044	-0.0099	0.003	-0.0027	0.01	1	0.039
GELIR	-0.068	0.63	-0.41	0.62	0.24	0.046	-0.049	-0.083	-0.099	0.039	1

Figure 8: Correlation Heat-map chart

- Model-I Cancelation (TARGET) distribution graph shows that this is an unbalanced data set.

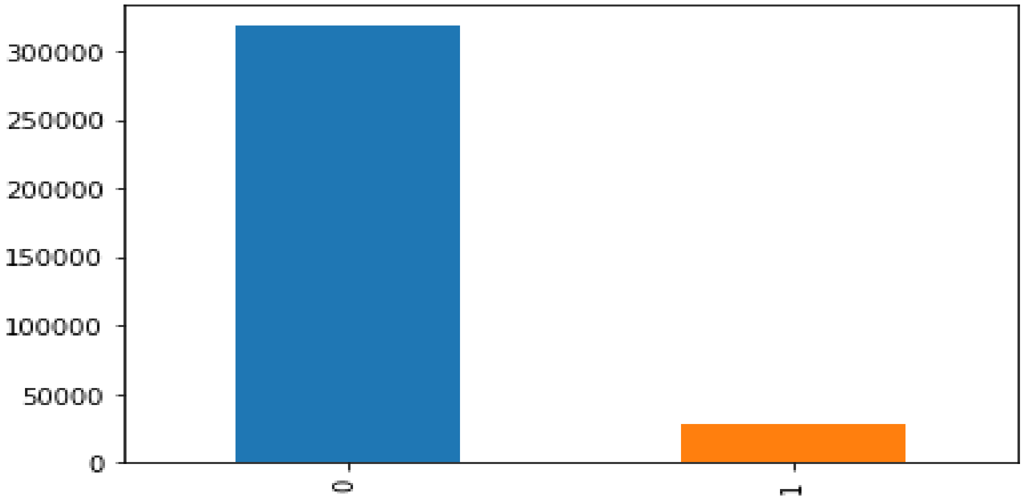


Figure 9: Target distribution for Model-I (Cancelation)

- Model-II (TARGET) distribution tried grouped for six class and their meaning is below

- Exiting between 3-12 months: 1
- Exiting between 13 and 18 months: 2
- Exiting between 18 and 24 months: 3
- Exiting between 25 and 30 months: 4
- Exiting between 31 and 36 months: 5
- 36+ months: 6

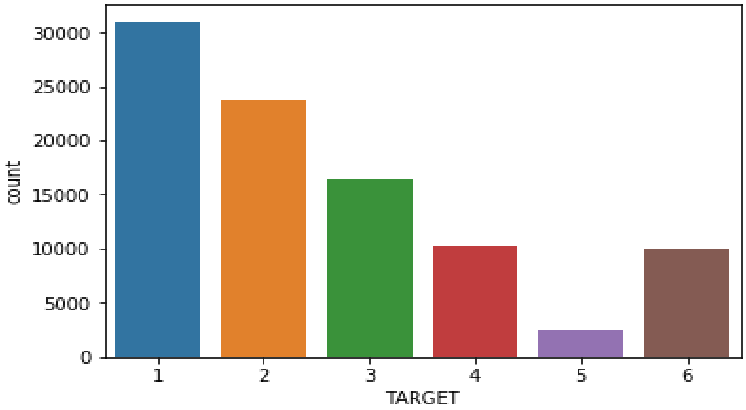


Figure 10 -1 : Target distribution for Model-II with six class

This graph showed that after 31 months of contracts retention period much more than other first months. Old contracts loyal than new contract. At that point, six class is not meaningful for this data. In addition, models built for six class was not successful enough. At that point, with domain experts consultancy class count decreased and retention interval changed. These classes are meaningful for business problem according to domain experts.

New classes are:

- Exiting between 3 and 18 months: 1
- Exiting between 19 and 30 months: 2
- 30+ : 3

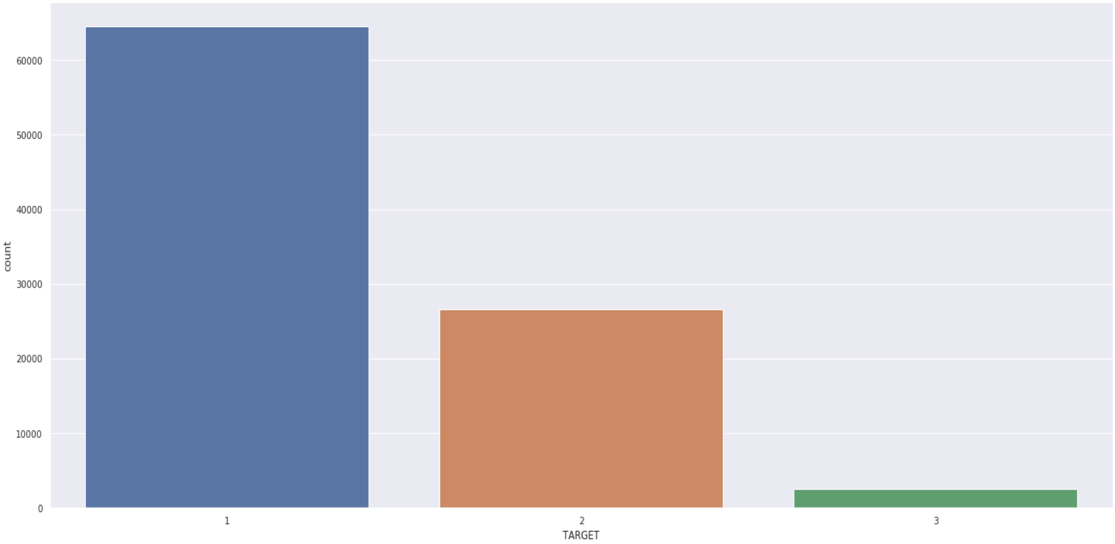


Figure 10-2: Target distribution for Model-II with three class

6. METHODOLOGY

6.1. Data Preprocessing

Descriptive statistics and explanatory analysis made earlier showed that data set has

- 4 unnecessary variables which have been dropped
- 11 numeric variable
- 21 categorical variable
- Target variable

For Model-I data set 346677 rows, for Model-II 93783 rows contains.

- Six categorical features have missing values, one numeric feature has missing value
- Variables have skewness need to transformation
- All variables are not same scale
- Outliers seen on the box plots need to be discard
- Categorical variables need to be encoding
- Model-I has unbalanced data needs to be sampled.

6.1.2. Missing Value Imputation

Imputations done with SQL for categorical variables with replacing most frequent values. Numeric values imputed with sklearn.impute packages' SimpleImputer function and variables imputed with their means.

6.1.3. Numerical Transformation for Skewness problem

Log transformation and Yeo-Jonson transformation results were compared and the Yeo-Jonson transformation has been chosen. It is done with sklearn.preprocessing packages' PowerTransformer function.

6.1.4. Scaling

Data has been scaled after the power transformation. sklearn.preprocessing packages' StandardScaler function has been used. Scaling strategy chosen by comparing a running a fast algorithm results. Mean and median results were not slightly different; then mean was chosen.

6.1.5. Discarding Outliers

Outliers have spoiling effect on machine learning algorithms so they must be handled. In the project, outliers have been discarded. Z score used to discard outliers. Z values calculated for every features and an interval chosen. Z<3 scores have been chosen comparing with other values and checking with box plot distributions.

6.1.6. Categorical Variables Encoding

Machine learning algorithms work with numbers so we need to encode categorical and ordinal variables. Pandas library get_dummies function and sklearn.preprocessing packages' LabelEncoder functions were used to encode variables.

6.1.7. Handling Unbalanced Data Set

Model-I has unbalanced data set. Major class has 93% proportion. Machine algorithms could not perform well do not generalize the data over unbalanced data. There are methods to handle it like over sampling, under sampling or doing both of them together. In the project over sampling made for minor class.

Sampling must be made over train data, due to this reason data split to two part as train and test. Train data's minor class is increased with oversampling. Oversampling made with imblearn library's SMOTE function. Sampling strategy chosen as 0.30 percentage. After doing over sampling base accuracy become 76%, decreased from 93%.

At the end of data preprocessing Imputed, Transformed/Scaled, encoded test and train data sets were ready. After encoding 340 features occurred in both test and train data set.

6.2. Feature Engineering and Machine Learning Models

6.2.1. Cross Validation, Grid Search, Pipeline Methods

Cross Validation is a method protects machine-learning models from over fitting by learning only from train data. Meanwhile it splits train data to folds and trains on a fold and test another one. It produces more reasonable train scores.

Grid Search is a process makes hyper-parameter tuning and finds optimum model and parameters for given machine-learning models.

Pipeline is process sequentially executes transforms and applies the output of the transforms to a final estimator algorithm.

This three method mostly used in the project to find best model and parameters with a similar simply flow below:

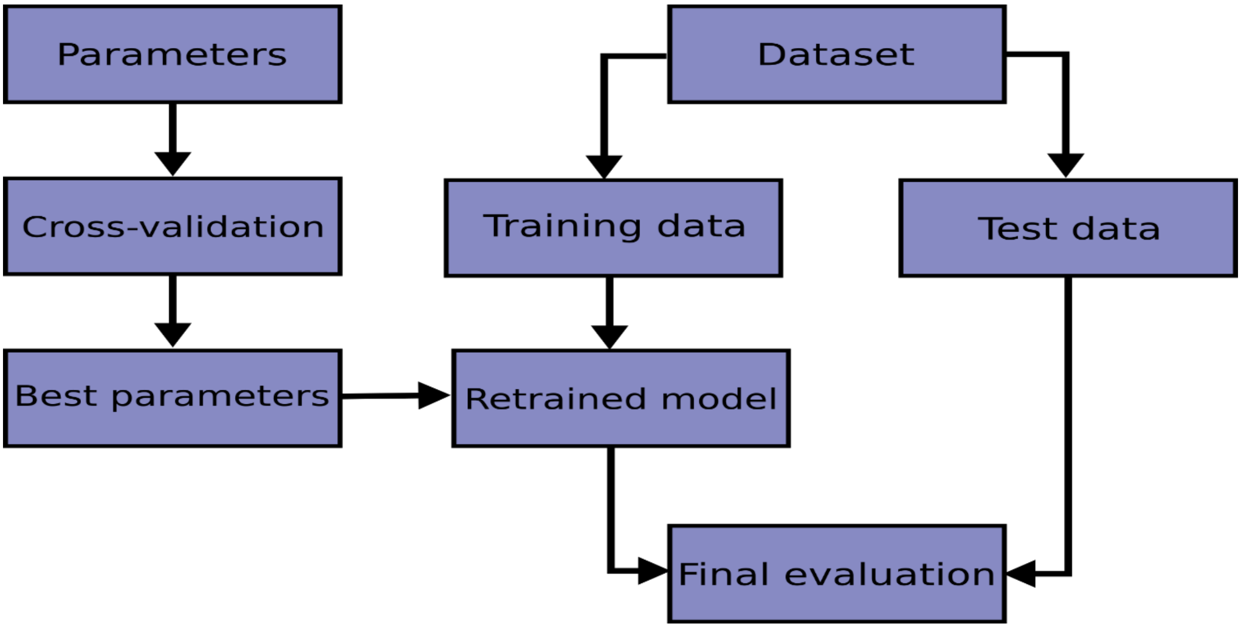


Figure 11: Cross validation/Grid search flowchart

6.2.2. Model Building

For Model-I (Cancelation Model) different algorithm have been tried with different hyper parameters. Decision Tree Classifier, SGD Classifier, Logistic Regression, VotingClassifier, Random Forest Classifier, ExtraTreesClassifier, Bagging Classifier, Gradient Boosting Classifier and XGB Classifier train F1 scores are shown below.

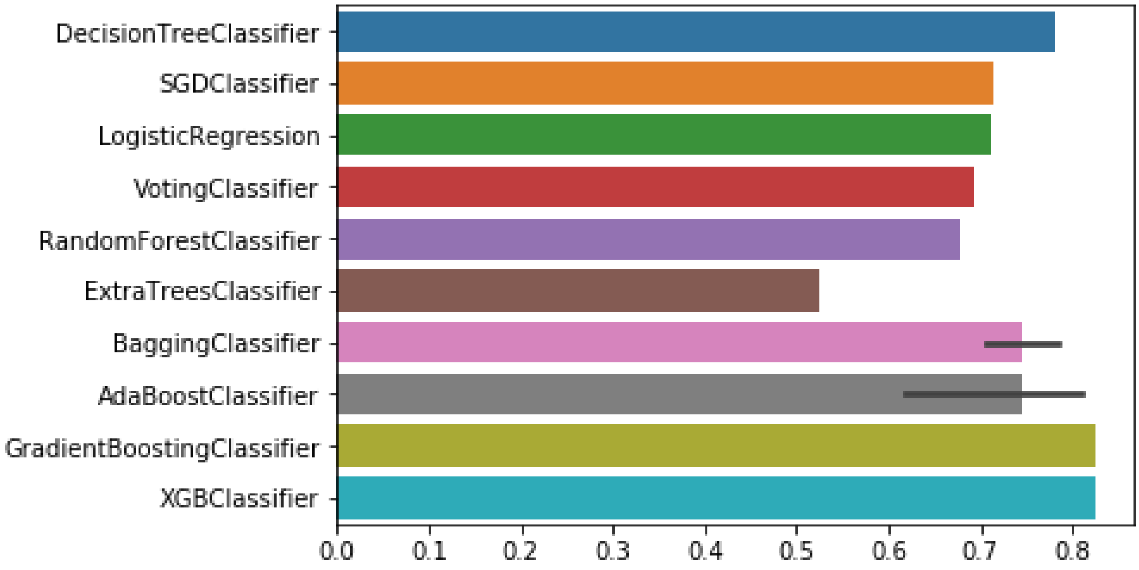


Figure 12: Model-I train F1 scores

Model-I (Cancelation Model) has been optimized for three machine-learning algorithms. Decision tree, XGB Classifier, multilayer perceptron (MLP) algorithm optimized to find the optimum prediction result. At the beginning of the project Decision Tree, Artificial Neural Network and Bagging/Boosting algorithms chosen to model the solution. ANN and Bagging/Boosting algorithms generally produce better results than other algorithms so that using them decided at the beginning. According to Figure-12 XGB Classifier has best outcome this because it has been optimized.

Model-II (Retention Period Model) was modelled with Random Forest was chosen to optimize and it is result will be presented.

In the first optimization iterations ANN and XGB over fitted. 340 features may have been result with over fit. To solve that problem two methods have been applied. PCA and new feature generation with clustering via K-Means algorithm.

6.2.3. PCA

PCA is a dimensionality reduction method and it optimizes variance and feature count. For the data set to see the cumulative variance while Principal Components increasing. This analysis made and it is output is below.

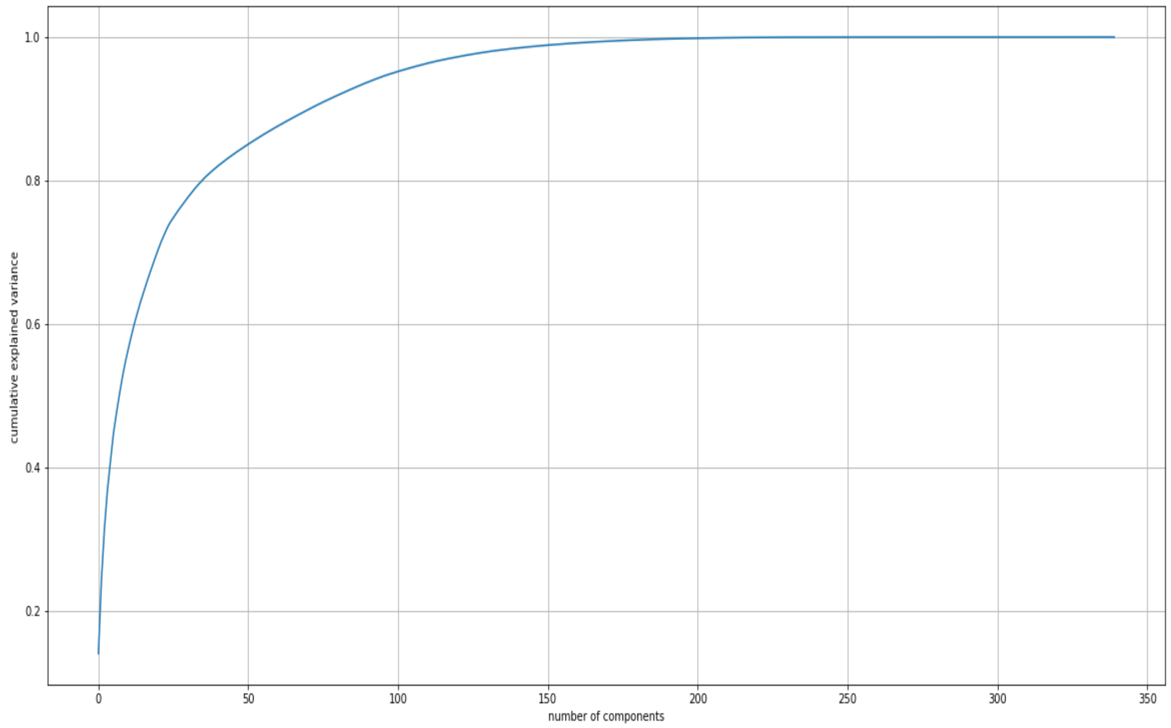


Figure 13: PCA analysis and cumulative variance plot

To variance maximum 180 principal components have been tried but the algorithms over fitted again. 50 principal component with a new feature explained below gave a better result and it is enough for the project scope.

6.2.4. Feature extraction with Clustering via K-Means

Machine learning algorithms give better result with better features which they have ability to clarify the target variable. Feature extraction done via generating new features. In the project to solve over-fit problem new feature generation has been tried.

K-Means is a unsupervised clustering algorithm which can be used generate new feature from existing data. K-Means generated classes were used as a new feature. To decide how many cluster to build an analysis made and it is output is below Figure 14. This technique compares cluster counts and inertia values. Inertia is a metric gives sum of squared distances of samples to their closest cluster center. It is needed to be as small as possible. The analysis below is an elbow method it is used to decide how many cluster is suitable for the data. From that elbow graph, cluster count is decided as 10 cluster. A K-means algorithm with ten cluster generated a new feature for the algorithms.

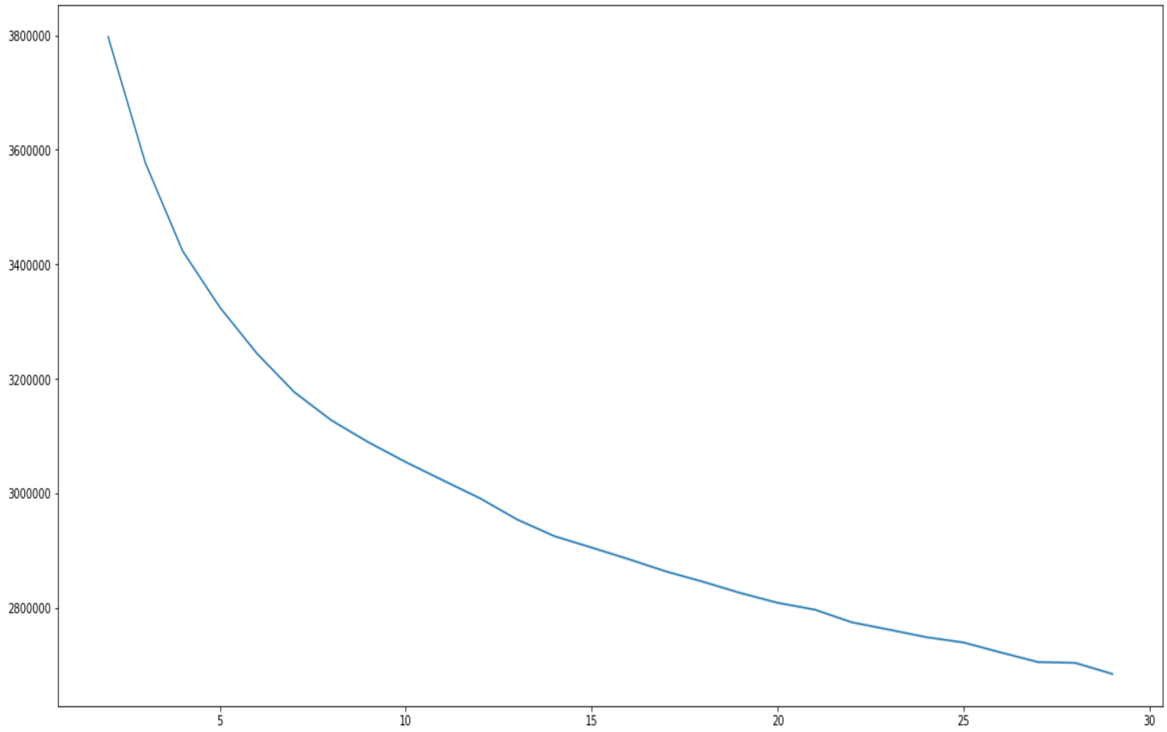


Figure 14: Elbow graph for K-means Clusters

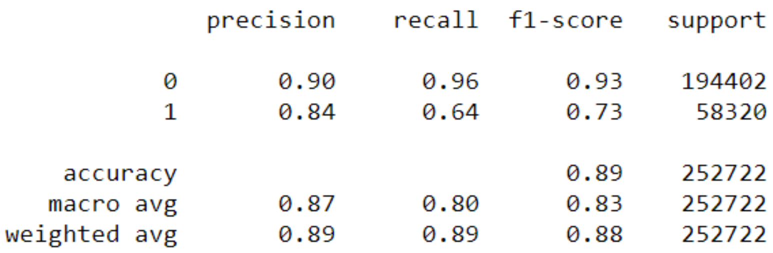
PCA with 50 component and a new feature extracted from K-means 51 feature was gathered. Algorithm we aimed to focus were optimized with that features and over fit problem was solved.

6.2.5. Model Results and Evaluation

Model-I (Cancelation Model)

ANN:

Train scores:

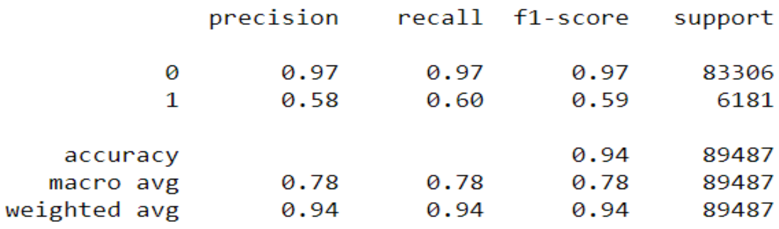


```
              precision    recall  f1-score   support

           0       0.90      0.96      0.93    194402
           1       0.84      0.64      0.73     58320

    accuracy                           0.89    252722
   macro avg       0.87      0.80      0.83    252722
weighted avg       0.89      0.89      0.88    252722
```

Test scores:



```
              precision    recall  f1-score   support

           0       0.97      0.97      0.97     83306
           1       0.58      0.60      0.59      6181

    accuracy                           0.94     89487
   macro avg       0.78      0.78      0.78     89487
weighted avg       0.94      0.94      0.94     89487
```

Learning Curve:

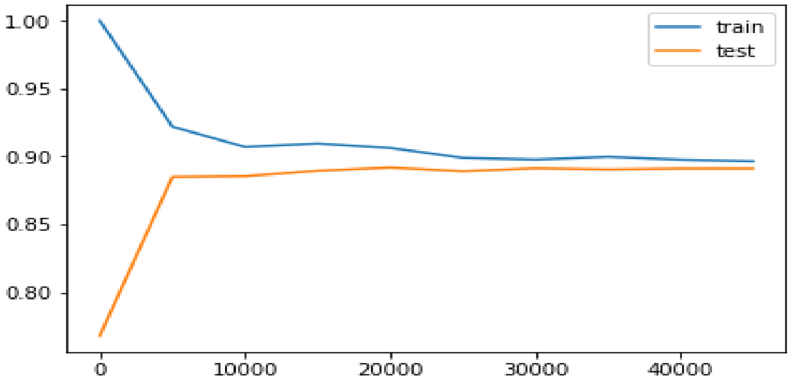
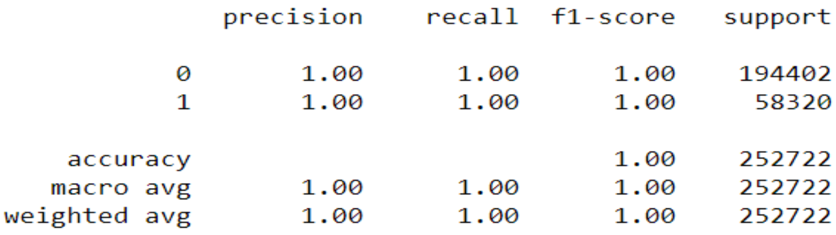


Figure 15: Learning Curve for ANN

XGB Classifier:

Train scores:

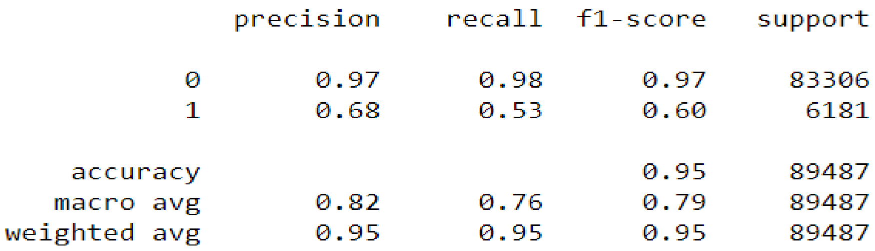


```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    194402
           1       1.00      1.00      1.00     58320

    accuracy                           1.00    252722
   macro avg       1.00      1.00      1.00    252722
weighted avg       1.00      1.00      1.00    252722
```

Test scores:



```
              precision    recall  f1-score   support

           0       0.97      0.98      0.97     83306
           1       0.68      0.53      0.60      6181

    accuracy                           0.95     89487
   macro avg       0.82      0.76      0.79     89487
weighted avg       0.95      0.95      0.95     89487
```

Learning Curve:

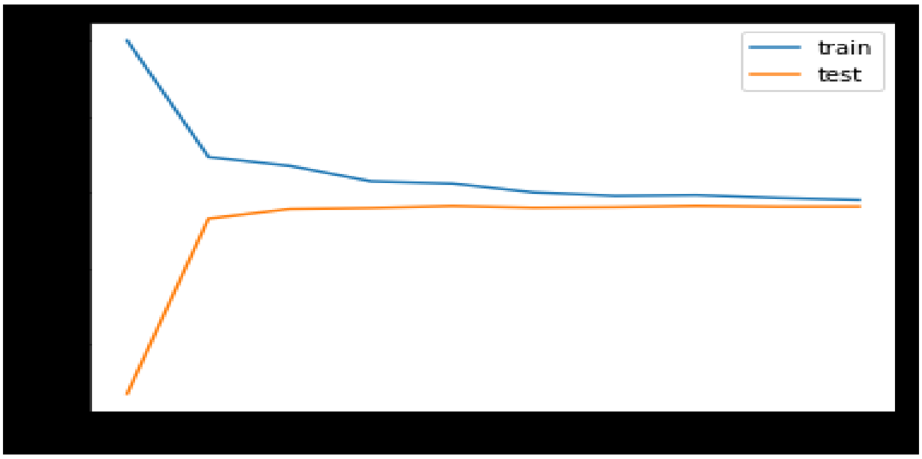
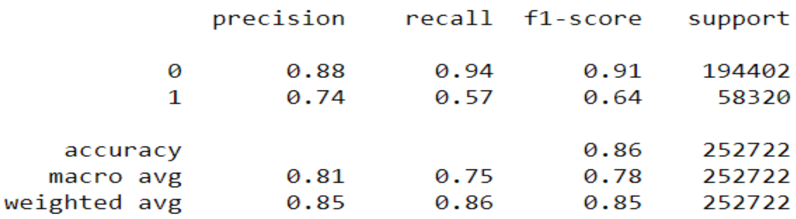


Figure 16: Learning Curve for XGB Classifier

Decision Tree:

Train scores:

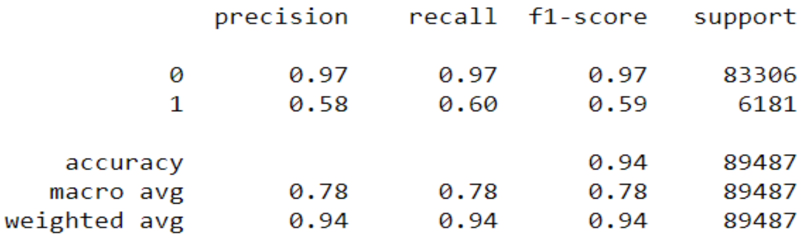


```
              precision    recall  f1-score   support

           0       0.88      0.94      0.91    194402
           1       0.74      0.57      0.64     58320

    accuracy                           0.86    252722
   macro avg       0.81      0.75      0.78    252722
weighted avg       0.85      0.86      0.85    252722
```

Test scores:



```
              precision    recall  f1-score   support

           0       0.97      0.97      0.97     83306
           1       0.58      0.60      0.59      6181

    accuracy                           0.94     89487
   macro avg       0.78      0.78      0.78     89487
weighted avg       0.94      0.94      0.94     89487
```

Learning Curve:

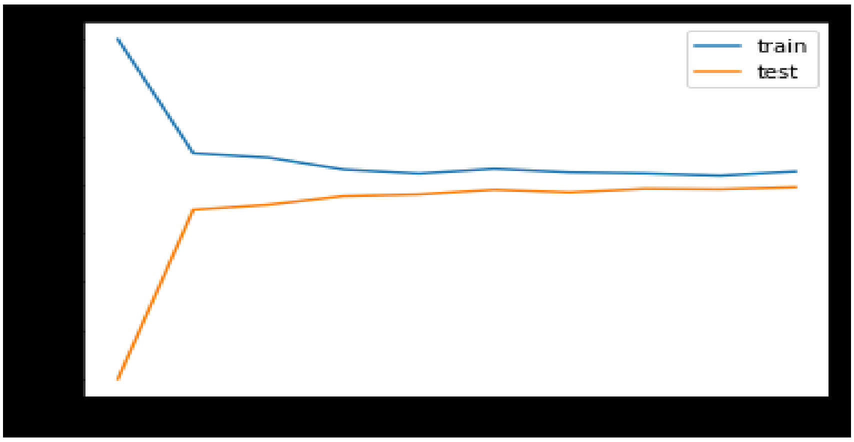


Figure 17: Learning Curve for D. tree

Precision Recall Curve:

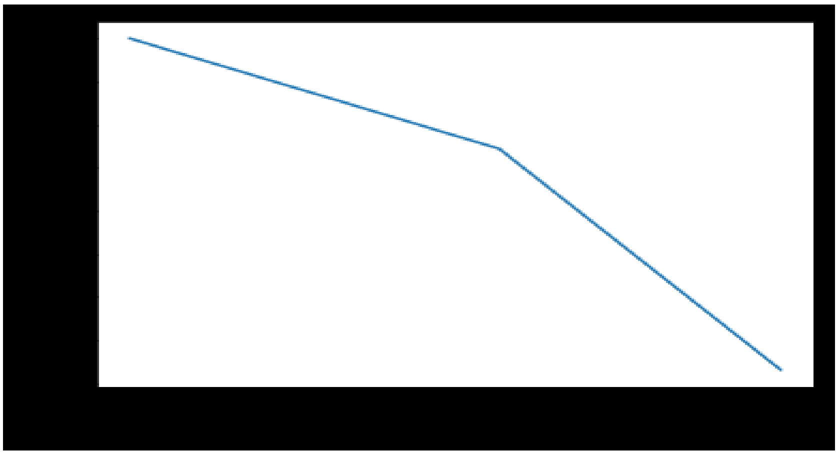
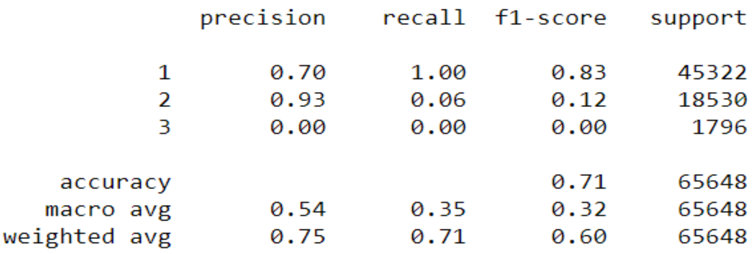


Figure 18: Precision Recall Curve for D. tree

<u>Model-II (Retention Periods Model)</u>

Random Forest

Train scores:

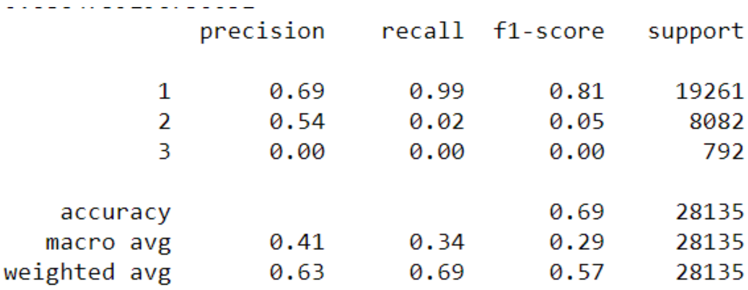


```
              precision    recall  f1-score   support

           1       0.70      1.00      0.83     45322
           2       0.93      0.06      0.12     18530
           3       0.00      0.00      0.00      1796

    accuracy                           0.71     65648
   macro avg       0.54      0.35      0.32     65648
weighted avg       0.75      0.71      0.60     65648
```

Test scores:



```
              precision    recall  f1-score   support

           1       0.69      0.99      0.81     19261
           2       0.54      0.02      0.05      8082
           3       0.00      0.00      0.00       792

    accuracy                           0.69     28135
   macro avg       0.41      0.34      0.29     28135
weighted avg       0.63      0.69      0.57     28135
```

Confusion Matrix:

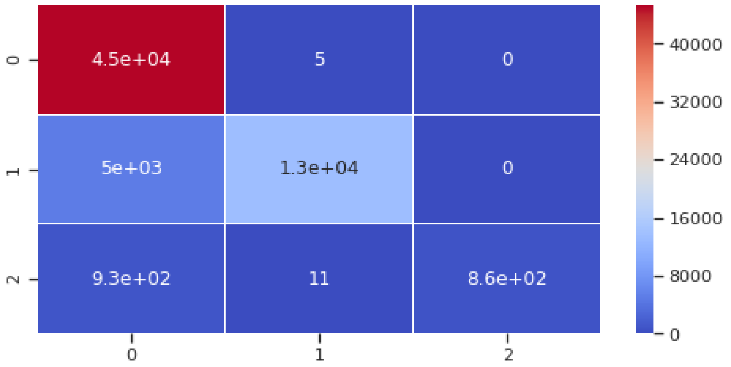


Figure 19: Confusion Matrix for Random Forest

```
array([[45279,    43,      0],
       [17327,  1203,      0],
       [ 1750,    46,      0]])
```

Table 3: Confusion Matrix

Learning Curve:

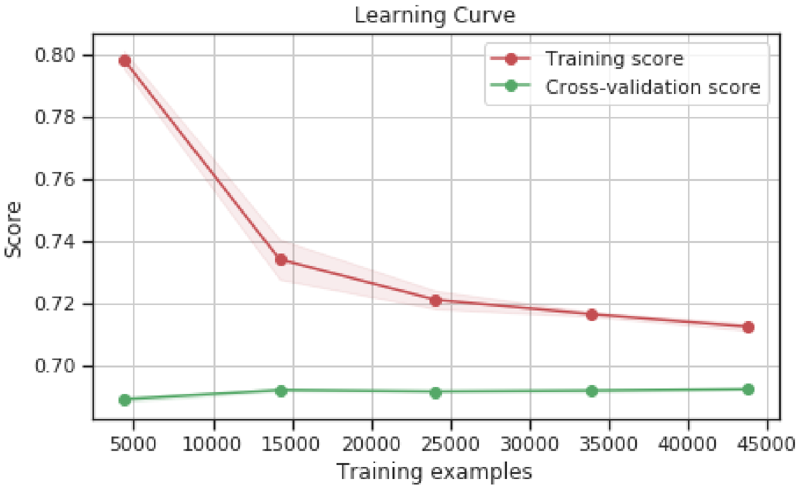


Figure 20: Learning Curve for Random Forest

Model – I results for ANN are enough and learning curve shows that over fit or under fit does not exists. XGB Classifier test scores and learning curve pattern is not bad but test and train scores has a difference more than a slight difference. It should be improved but it is enough for the projects scope. Decision tree also gave good results but precision/recall curve shows that there is opportunity to improve the algorithm.

Model-II worked two times for different class types. In six classes type over fit problem has not been solved. EDA also showed that for that analysis the data and features are not enough and there is a need to develop new data sources. According to domain expertise it has a better meaning with a tree class with data set used in that analysis. Therefore classes were changed and analysis done again. Over fit problem solved and with tree class over fit has not been observed. Learning curve and test train scores showed that it was handled. Although results are better accuracy score has improvement chance. These outcomes will be shared with the company that data quality and enrichment should be done to make stronger analysis.

6.3. Feature Importance

Feature importance depicts that which features have force on target more than others do. This analysis is meaningful with business domain explanation. The analysis made by Random Forest algorithm and its feature_importance_ attributes gives the important top n features. Top 50 feature plotted below

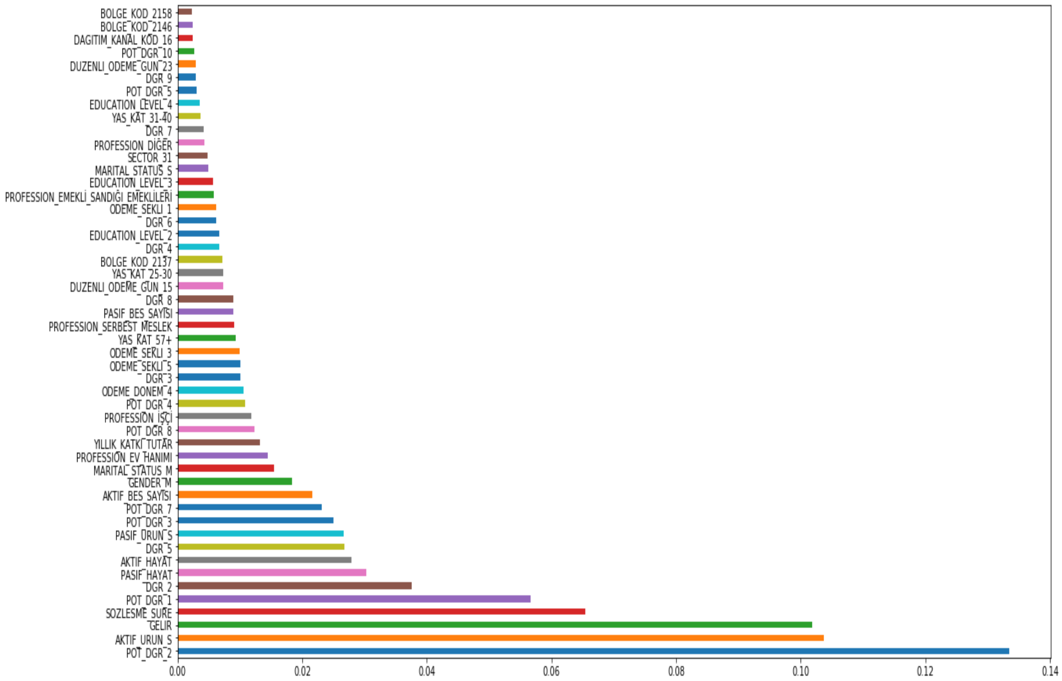


Figure 21: Feature Importance with Random Forest

Top five features are:

POT_DGR_2 , AKTIF_URUN_S, GELIR, SOZLESME_SURE, POT_DGR_1

7. CONCLUSION AND FURTHER STUDIES

The Project will made contribution to retention studies in pension market and the company. Pension proposal data sources searched to obtain features for models in a limited time in the project. Data collection, data preprocessing, explanatory data analysis made. Different type of machine learning algorithms tried and optimized with grid search technique or manually. ANN algorithm gave the best result. The most challenging part of the project was data quality. To enrich the data when exploration made in data sources there is no ready to use data exists. After the project, this area will be an improvement area. Insufficient quality or high cost of data wrangling many factors that impact cancelation or retention could not be added as features to the project.

Cancelation model performance is in acceptable limits and it can be a deployable model. Rather than using non-proactive reports, the cancelation model will generate new values in the company. It may reduce the cancelation rates, improves sales quality in the company.

The models and outputs in the project will generate a base and all these development areas will be in the scope of retention programs.

8. REFERENCES AND CITING

1: A logistic regression approach to estimating customer profit loss due to lapses in insurance (2018)

AUTHORS: Montserrat Guillén Estany Ana Maria Pérez-Marín Manuela Alcañiz

2: Decision support models for optimal personalized marketing interventions in insurance (2015)

AUTHORS: Leo Guelmana, Montserrat Guill´enb,∗, Ana M. P´erez-Mar´ınb

3: Customer retention Management processes A quantitative study (2005)

AUTHORS: Lawrence Ang and Francis Buttle Macquarie Graduate School of Management, Macquarie University, Sydney, Australia, Received July 2004 Revised January 2005

4: Enhancing Customer Retention Through Data Mining Techniques (2017)

AUTHORS: Alexiei Dingli1, Vincent Marmara and Nicole Sant Fournier

Department of Artificial Intelligence, University of Malta Faculty of Economics, Management and Accounting Department of Artificial Intelligence, University of Malta

5: Customer Retention Management In The Informatıon Era(2001)

AUTHORS: Devashish Das Gupta Snehashish Mukherjee

Delhi Business Review Vol. 2, No. 2, July - December 2001

6: How to Project Customer Retention

AUTHORS: Peter S. Fader Bruce G.S. Hardie1(May 2006)

7: Random Forests For Uplift Modeling: An Insurance Customer Retention Case(2011)

AUTHORS: Leo Guelman, Montserrat Guill´En And Ana M. P´Erez-Mar´In

8: Transforming Skewed Data:

https://towardsdatascience.com/transforming-skewed-data-73da4c2d0d16

9: An introduction to Grid Search
https://medium.com/datadriveninvestor/an-introduction-to-grid-search-ff57adcc0998

10: Cross-validation: evaluating estimator performance
https://scikit-learn.org/stable/modules/cross_validation.html

11: A Quick Guide to Feature Engineering
https://www.kdnuggets.com/2019/02/quick-guide-feature-engineering.html

9. TABLES AND GRAPHS