An Optimal Classifier discovery for diagnosing the account health in financial firms and a study of classifier performance on imbalanced data

R. Suguna
Vel Tech Rangarajan Dr Sagunthala R&D Institute of Science and Technology
R. Subhashini
Sathyambama Institute of Science and Technology: Sathyabama Institute of Science and Technology
Ramanathan Lakshmanan
Vellore Institute of Technology: VIT University
Ananda Kumar S (s.anandakumar@vit.ac.in)
VIT University https://orcid.org/0000-0002-9263-9623
Stavros N Shiaeles Stavros N Shiaeles S
Portsmouth Polytechnic: University of Portsmouth
R Sangeetha
Vellore Institute of Technology: VIT University

Research Article

Keywords: Customers, Linear, Non-linear, classifiers, Machine learning

Posted Date: July 11th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1242944/v1

Abstract

Customers are the backbone for any financial companies. The behaviour of customer changes over time and they disconnect when the services do not meet their expectations. Earning loyalty of the customer by providing remarkable services and adopting retention strategies are mandatory to run any user centric businesses. In view of the growth perception there is need for company to identify the churn and avoid them in time. The mechanism for churn prediction requires to explore the insight of data. Machine learning algorithms are capable of mining the patterns present in the data and able to discriminate between classes with statistical learning. A Standard bank dataset has been considered for the study and exploratory data analysis performed to understand the nature of the data. Suitable data pre-processing is done and training data split from dataset has been used to build classifier models. The dataset was found to be imbalanced and by adopting appropriate sampling the dataset was balanced. Linear, Non-linear and boosting classifiers were built and their performances on test data are summarized. A comparative study on the classifier performance for both imbalanced as well as balanced dataset was observed and an optimal classifier for diagnosing customer account health has been suggested.

1. Introduction

Financial sectors, particularly banking has undergone a steady transformation over the decades. New technology, change in demands and regulation have caused the bank to face challenge in securing dominant positions in society. Many banks today are with good customer size with a hope that they will not be tapped with their competitors. But existing intensified competition can pose threat in customer turnover. Customer turnover is also referred as customer churn which means that customer disconnects the association with company over a period of time. Retaining the existing customers and increasing their lifetime gains primary importance in banks. The cost of attracting new customers can be expensive than holding the existing customers. Long term customers provide higher profits and create new referrals. Availability of latest technology, customer friendly bank officials, location and services may be the reasons for customer switching. With the advent of internet banking, clear customer feedback is lacking and hence predicting customer churn at the right time becomes complex. Customer relationship can be endowed with early and accurate churn predictions. This makes the bank officials to be proactive in engaging the customers at right time and prevent the exit.

The practical issue is how to find signs for churn. Collecting feedback on customer experience may be time consuming task. Moreover surveys are infrequent and the response of customers to feedbacks cannot be guaranteed. Extracting early warnings from already existing data may provide a solution for this problem. Short prediction prospect in accuracy but it would be late if the customer has decided to quit. Generally churn can be defined at the product level like discontinuing particular product or at the relationship level like disengaging the bank itself. At relationship level customer's view can be understood clearly and engage them with complementary products that can strengthen the relationship. There is a growing demand in studying the set of characteristics to analyse and predict customer churn. Machine learning and data analysis algorithms have the capability to learn from past customer data and can

generate triggers on churn data. They try to capture all aspects of customer relationship with bank. Apart from churn prediction, they can help in recommendations of new or allied products and perform customer life time calculation. Prediction models built with machine learning techniques can identify customers who are likely to churn. The model analyses the historical data of past churners and finds similarity with existing customer data. If they match, current customers are labelled as potential churners.

Setting up a churn prediction model involves defining a churner, prediction attributes of the model and techniques used to build the model. Customer who closes the account or inactive for long time or decrease the number of transactions over a period of time are considered as churner. Normally churn is expressed as a degree of customer inactivity or disengagement, observed over a given time. This expresses within the account in various forms such as the frequency of account actions or change in the account balance. It is also useful to define churn based on the rate of decline of assets over a specified period. Four types of prediction attributes are commonly used in churn analysis: demographic, perception, behavioural and environment. Demographic variables include age, profession, gender, family and geographical data. Perception attributes reveal the customer appreciation to services. They include quality of service, locational convenience and pricing. Behavioural attributes exhibits how often services are used and which services are availed most. Environment variables refer to changes in world that could affect the customer.

Machine learning algorithms work on features and feature engineering is the process of creating features using domain knowledge of the data. Features play a pivotal role in capturing customer behaviour. Basic indicators like net balance outflow in the last few months, to more nuanced indicators like rate of change of average gap between bill payments made, can prove effective in providing early warning signals of impending churn.

Churn prediction belongs to classification problem category. Various techniques from simple logistic regression to complex tree-based techniques like XG Boost exist and it is important to identify the technique that provides the right balance of interpretability and performance. Complex algorithms such as random forest and XGBoost capture non-linear patterns in data and can handle null values comfortably. Logistic regression provides a more apparent and innate explanation of the impact of each variable on the predicted outcome.

Adoption of prediction models starts with demonstrating the model’s predictive power on the past data and running several simulations to measure the efficacy of the model and associated scheme. Testing involves selecting customers likely to churn and observe the model based predictions.

Banks acknowledge that customer churn is a critical problem, but there is no systematic and proactive methods to address it. Even after building robust churn prediction models, the challenge prevails in creating enablers in every phase. It involves setting up the churn likelihood scores periodically to offer right retention strategies.

2. Related Work

Banks use customer transaction data to predict the behaviour of customer who are likely to churn. Few works have concentrated on developing models using time series predictors and identification of rare events occurrence in transactions. One of the study used trend modelling to understand the customer change in behaviour over a period of time. They have used a three year old real time dataset of retail bank at Florida. Results show that the training data over a period of six months can predict the customer behaviour better than four months data[1]. Increase in competitions and business saturation force the companies to build strategies to retain the old customers than attracting new customers as it is an expensive effort. Customer retention focus on building efficient models to predict the behaviour pattern of customers. Various machine learning algorithms have been used for churn prediction. A hybrid method to predict churn was proposed and the work involves feature selection algorithms and effect on size of features on prediction were analysed. Two algorithms Linear Tree and decision tree algorithms were applied on the dataset and their results are combined using voting strategy. Results show that the proposed hybrid method performs effectively[2].

A new approach named data mining by evolutionary learning has been proposed to address the classification of churn problem. The process involves generation of possible rules using probabilistic induction technique and the likelihood of churn is estimated by representing the fitness value with the attributes of records. Experimental results prove that the proposed model performs well in churn prediction[3].To establish valued relationships with prevailing customers and to emphasize strong customer relationship management it is necessary to understand the customer expectations. Using customer samples of European financial company and application of random forest technique have shown better understanding of customer behaviour than linear regression techniques[4].

A mathematical model to analyse the key problem for churning has been proposed and analysed with different classifiers. Logistic regression and decision tree classifiers with different parameters were built to test the performance of churn prediction. Experimental results show that decision tree classifier with maximum depth of six perform well in churn classification[5].Recognising the value of existing customers all companies acknowledge the fact that minimizing churn rate will maximize their profit. Combination of SVM with Adaboost have been proposed to improve the accuracy of classification rules. The proposed churn model was able to identify the potential churners effectively[6]. An extreme learning machine has been proposed to forecast customer churning. The architecture is basically a feed forward neural network and follows random allocation of input weights and biases. They have used Portuguese bank dataset containing 20 attributes. They have experimented with different activation functions and estimated training and testing accuracies. They have proved that processing time for both training and testing was less for the proposed approach[13].

The rate of customer churn is increasing every year and at most care is required to retain the customers. Datamining provides classification techniques that can be applied for churn prediction problem. A study on applying five different classification techniques on a private Indonesia banking dataset was

performed and proved that SVM performs well in prediction of churners[7].Though commercial banks take measures to analyse the customer information in their records, it is found that prevention of churning is challenging. An improved Fuzzy C means clustering algorithm has been proposed to understand the customer behaviour which in turn helps in identification of churners[8].

Customers are valuable asset of any financial organization and customer retention is essential as well as basic requirement for the firms to run their business. Nowadays clients rely on electronic banking services and communications happen electronically which prevents to recognize the customer satisfaction. The characteristics of churners have been analysed using decision tree and feature selection algorithms were employed in the churn prediction model. Results show that backward elimination methods provide better accuracy[9].A broad review on identification of characteristics of customers who are likely to churn, finding the reason for churn, existing churn prediction models and how the efficacy of churn prediction model can be improved were discussed [10]. The task of determining the risk of customers leaving the bank has been analysed using neural networks. Neural networks are capable of handling large number of inputs and can provide approximation on inputs. Results have proved that clients who acquire more services are less likely to churn[11].

A comparative study on machine learning packages were conducted and results show that Spark ML package with its APIs perform better than MLLib packages. This comparative study done with customer churn prediction in banking as a cases study[12].Ensembling methods have been used in banking sector to improve the performance of classifiers. Bagging and boosting are most commonly used ensembling techniques. Random Forest, Extreme gradient boosting and Adaboost are some of the algorithms contributing in enhancement of classifier performances [14–15].A comparative study of machine learning algorithms on churn prediction has been evaluated in various research works [16-17]. Recorded customer complaints have been analysed used natural language processing and prediction models [18].

3. Machine Learning Classifiers

3.1. Linear Classifiers

Consider a binary classification problem where the output y=c when it belongs to a positive class and y=c when it belongs to a negative class. The representation of class is obviously two dimension and the classifier takes decision on class membership by comparing a linear combination of its input to a threshold. If the number of input features n is 2, the decision boundary is a line. When n > 2 the decision boundary is a hyperplane.

In its functional form this boundary can be represented as

$y(x)=w_1 x_1+ w_2 x_2+\cdots+w_nx_n+b$

The classification rule is to assign

y(x)=c if y(x)>b else y(x)=c̄

Generally hyperplane is defined as

$W^T X+b=0$

General Linear Classification Algorithm

```
LinearClassifier(w⃗,x⃗,b)
Compute s= ∑_(i=1)^n ⟦w_ix_i⟧
if s>b
    return 1
else
    return 0
```

The objective is to find the optimum value for the parameters $\vec{w}$ and b with the training examples. The efficiency of the linear classifier depends on evaluating the quality of the model with the new data.

For a probabilistic model, classification can be done with probability estimation. First it finds the probability distribution of the training data and helps to predict the new data.

Logistic Regression uses linear functions to build probabilities of observation belonging to the class.

3.2 Nonlinear Classifiers

Nonlinear functions are used in discriminating the classes. The general concept used finding distance/similarity between the training instances. K-nearest neighbors classifier uses distance metric in classification task.

NonlinearClassifier($\vec{x}$, d)

- Project input space in higher dimension (2d-1)
- Apply linear classification
- Some nonlinear classifiers such as Decision Tree make predictions based on features.

Some nonlinear classifiers such as Decision Tree make predictions based on features.

General idea:

1. Pick the feature that best distinguishes classes

- If you group the instances based on their value for that feature, some classes should become more likely

- The distribution of classes should have low entropy (high entropy means the classes are evenly distributed)

2. Recursively repeat for each group of instances

3. When all instances in a group have the same label, set that class as the final node.

- Multilayer Perceptron is basically a neural network algorithm mapping input to output by learning and adjusting the weights between the nodes.

3.3.Boosting Classifiers

Boosting algorithms work by training a sequence of weak models by compensating the weakness with the preceding models to improve the prediction rate.

```
Boosting Classifier(W, M)

    Initialize the weight W
    for each WeakClassifier[1..M]
    fit Classifier Ci with weight W
      Compute Error
      Find the coefficient of Ci
      Update the weights
      Normalize the weights
    Apply weighted voting on output of the model and find the
    classifier with the highest vote
```

4. Exploratory Data Analysis

The main objective of the work is to develop a model to help the bank officials in identifying the risk of the customers who are likely discontinue their services from banks. The dataset utilized for the study is obtained from kagglehttps://www.kaggle.com/shrutimechlearn/churn-modelling. The dataset contains 10000 customer details with no missing data. There are 14 attributes with 7 attributes revealing customer personal information, 4 attributes showing the status of the account and 2 attributes towards the details of products purchased. The target attribute holds a binary value signifying the exit status of the customer. Table 1 depicts details of the dataset.

Table 1
Customer Churn Data Attributes

Customer details	Attribute	Description
Personal Information	RowNumber	Row Numbers from 1 to 10000
	CustomerId	Unique Ids for bank customer identification
	Surname	Customer's name
	Geography	Customer’s nationality
	Gender	Male / Female
	Age	Customer age
	EstimatedSalary	Estimated salary of the customer in Dollars
Account information	CreditScore	Estimated credit score of the customer
	Tenure	Number of years customer associated with bank
	Balance	Current bank balance of the customer
	IsActiveMember	BInary value indicating the active utilization of bank services
Bank Products Utilization	NumOfProducts	Number of bank services utilized by the customer
	HasCrCard	Binary value indicating whether customer possess credit card
Target	Exited	0 – customer discontinued 1 – customer retained

Visualizing the distributions of variables

Categorical data are converted into numerical data. Figure 1 shows the distribution of categorical variables.

The distribution conveys certain information of variables

- Categorical variables such as Geography, IsMale (Gender), HasCrCard and IsActiveMember follow Bernoulli distribution.
- Tenure seems to be uniformly distributed and NumOfProducts declines for higher values
- Continuous variables such as CreditScore, Balance are normally distributed while EstimatedSalary is uniformly distributed

Nature of customers

To understand customer details the proportion of customers based on nationality and gender Figure 2 shows the proportion of customers based on nationality and gender as a whole as well as distribution of gender nationwide.

From the graph it is evident that 50% of the customers are French and the rest is equally shared between customers of Spain and Germany. Considering the gender, proportion men holding accounts are slightly higher than women, The same is reflected in nationwide observation.

Figure 3 shows the distribution of estimated salary with gender. There is no significant gap in salary and they have similar median.

There are no significant salary differences with respect to age. Figure 4 shows that few customers are senior citizen and there is slight difference in salary in their range.

Correlation between variables

Correlation matrix can be used identify the relationship between the variables. Figure 5 shows the correlation plot for the variables.

- There is high correlation between German and account balance
- Age and balance are correlated with exit variable
- Also the exit rate of Germans are higher

Sankey diagram is a visualization tool showing the flow from one variable to other. The width of the flow depicts the proportion of quantity flow between variables.

The flow diagram depicts that women are more disconnecting their services than men with bank. The churn rate of German customers are higher than other citizens.

The above exploratory analysis provides the insights on customers and suggests the direction to customer focus.

5. Methodology

The problem that is addressed belongs to binary classification category. First dataset is pre-processed by removing irrelevant attributes and converting all attributes to numerical data. Then data is split into training and test set. Various supervised classifiers are trained with training set and models are built. Classifier performance is assessed with the test data with suitable metrics. The dataset has 20% exited clients and 80% non-exited clients showing the imbalance in data distribution.

The proposed work has the following phases.

Phase-1: Analysing the performance of linear classifiers with the pre-processed data

Phase-2: Analysing the performance of nonlinear classifiers with the pre-processed data

Phase-3: Improving the classifier performance with boosting classifiers

Phase-4: Apply sampling to create balanced data

Phase-5: Classifier performance on balanced data set

Phase-6: Comparing the classifier performances and the one with highest accuracy is considered as an optimal classifier

5.1 Data Pre-processing

The steps involved in data cleaning is shown in Figure 7

- Attributes 'RowNumber', 'CustomerId', 'Surname' are removed as they do not relate with the target attribute.
- Categorical Attribute ‘Geography’, ‘Gender’ are converted to numerical data by applying one hot encoding.
- The values of attributes vary in magnitudes. The performance of machine learning algorithms will be better if data are normalized. Data are scaled using Standard Scalar normalization.
- Dataset is split into 75% training and 25% test set data.

5.2 Model Building

Figure 8 shows the process flow diagram of the proposed work.

5.3 Classifiers used for expeimentation

The classifiers used for model building are shown in Table 2

Table 2
List of Classifiers used for performance assessment

Linear Classifiers	**Logistic Regression**
	Naïve Bayes
	Linear SVM
Nonlinear Classifiers	Decision Trees,
	Random Forest
	K-Nearest Neighbour
	Multi-layer Perceptron
Boosting Classifiers	Gradient Boosting
	XGB Classifier
	LGBM Classifier

5.4 Performance Metrics

Confusion matrix is a mechanism to summarise the classification outputs. It clearly indicates how well the model behaves on positive and negative examples. Table 3 lists the metrics are used for assessing classifier performances.

Accuracy	**It is the ratio of correct predictions made to total predictions**
AUC	Area Under the Curve measures the ability of the classifier to distinguish between classes. Higher the AUC value, better is the classifier performance
Precision	It is the ratio of correctly predicted positive observations to the total predicted positive observations.
Recall	It is the ratio of correctly predicted positive observations to the all observations in actual class
F1-Score	It is the weighted average of Precision and Recall. This score takes both false positives and false negatives into account. F1 is more useful than accuracy, if the class distribution is uneven.

6. Results And Discussion

The performance of the classifiers are analysed with metrics such as accuracy, AUC, Precision, Recall and f-score.

6.1 Performance of linear classifiers with imbalanced churn dataset

The linear classifiers considered for the study are Logistic Regression, Naïve Bayes and Linear SVM.

Figure 9 Performance metric score of linear classifiers on imbalanced data

All three classifiers are providing same accuracy. The recall measure of SVM Classifier is better than other classifiers.

6.2 Performance of non-linear classifiers with imbalanced churn data

Non- linear Classifier performance metrics on the imbalanced churn data is shown in Figure 10. Observations reveal that neural network with multi-layer perceptron produces good accuracy compared to other classifiers.

A comparative study of best performing linear and non-linear classifier is shown in Figure 10

6.3 Performance of boosting classifiers with imbalanced churn data

Figure 11 shows that Gradient XGBoost classifier produces improved performance score compared to LightGBM and Gradient Boosting sklearn classifier.

A comparative study of MLP and XGBoost Classifier is shown in Figure 12. It is observed that both the classifiers yield same performance scores.

Performance of classifiers on balanced data set

It is observed that the churn dataset has 80% of instances belong to not exit category and 20% are in exit category. This influences bias in performance metric scores since classifiers cannot learn with minimal data in one category. This issue can be resolved by balancing the dataset. Here random sampling technique is applied to balance the data. More attention is paid towards data that is used to train the model.

6.4 Performance of linear classifiers with balanced churn data

Linear classifiers performance on balanced data is shown in Figure 12.

SVM Classifier is performing well on balanced and imbalanced data. The scores obtained on balanced and imbalanced data compared and shown in Figure 13.

It is evident that there is variation in performance scores of the same classifier with balanced and imbalanced data.

6.5 Performance of linear classifiers with balanced churn data

Non-linear classifiers performance scores on balanced data are shown in Figure 14.

The scores obtained by MLP Classifier on balanced and imbalanced data are compared and shown in Figure 15.

6.6 Performance of boosting classifiers with balanced churn data

Figure 16 shows the boosting classifiers performance scores on balanced data.Gradient Boosting (Sklearn) performs well on balanced data.

Figure 16 compares MLP performance with Gradient Boosting classifier on balanced data. Results show that MLP classifier outperforms on balanced data

7. Conclusion

Due to increase in private financial institutions the migration of customers from prevailing institutions has become a serious issue. Financial institutions are struggling to retain their customers. It has become mandatory to keep track of customer account health to take precautionary measures for retention. Machine learning algorithms can definitely help them to analyze the status of the customers and churn rate can be reduced. This paper has done elaborate exploratory analysis on churn data to visualize the correlations between the attributes in different dimensions. The performances of linear, non-linear and boosting classifiers are analyzed. The dataset considered for the study has imbalance in churn/not churn category. Hence random sampling is applied to make the dataset balanced. Then the various classifiers are applied on the balanced dataset to study the impact of imbalanced over balanced dataset. It is observed that imbalanced dataset has bias over the performance metrics. Results reveal that Multi layer Perceptron performs well on the churn data considered for the study.

Declaration

Conflict of Interest and Authorship Conformation Form

Please check the following as appropriate:

- All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.
- This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.
- The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript

*Funding –No Funds by any agency

*Conflicts of interest/Competing interests (include appropriate disclosures)

*Availability of data and material (data transparency)- no copyrighted data

References

1. H. C. Leung, “AIS Electronic Library (AISeL) Industry A Dynamic Classification Approach to Churn Prediction in Banking Industry,” *AMCIS 2020 Proc.*, vol. 28, pp. 0–5, 2020, [Online]. Available: https://aisel.aisnet.org/amcis2020/data_science_analytics_for_decision_support/28.
2. E. Jamalian and R. Foukerdi, “A hybrid data mining method for customer churn prediction,” *Technol. Appl. Sci. Res.*, vol. 8, no. 3, pp. 2991–2997, 2018, [Online]. Available: www.etasr.com.
3. W. H. Au, C. C. Chan, and X. Yao, “A novel evolutionary data mining algorithm with applications to churn prediction,” *IEEE Trans. Evol. Comput.*, vol. 7, no. 6, pp. 532–544, 2003, doi: 10.1109/TEVC.2003.819264.
4. B. Larivière and D. Van Den Poel, “Predicting customer retention and profitability by using random forests and regression forests techniques,” *Expert Syst. Appl.*, vol. 29, no. 2, pp. 472–484, 2005, doi: 10.1016/j.eswa.2005.04.043.
5. M. C. López-Díaz, M. López-Díaz, and S. Martínez-Fernández, “A stochastic comparison of customer classifiers with an application to customer attrition in commercial banking,” *Scand. Actuar. J.*, vol. 2017, no. 7, pp. 606–627, 2017, doi: 10.1080/03461238.2016.1209549.
6. A. S. Kumar and D. Chandrakala, “An Optimal Churn Prediction Model using Support Vector Machine with Adaboost,” *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 2, no. 1, pp. 225–230, 2017, [Online]. Available: www.ijsrcseit.com.
7. K. G. M. Karvana, S. Yazid, A. Syalim, and P. Mursanto, “Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry,” *2019 Int. Work. Big Data Inf. Secur. IWBIS 2019*, pp. 33–38, 2019, doi: 10.1109/IWBIS.2019.8935884.
8. S. Cui and N. Ding, “Customer churn prediction using improved FCM algorithm,” *2017 3rd Int. Conf. Inf. Manag. ICIM 2017*, pp. 112–117, 2017, doi: 10.1109/INFOMAN.2017.7950357.

9. A. Keramati, H. Ghaneei, and S. M. Mirmohammadi, “Developing a prediction model for customer churn from electronic banking services using data mining,” *Financ. Innov.*, vol. 2, no. 1, 2016, doi: 10.1186/s40854-016-0029-6.
10. V. Avon, “Machine learning techniques for customer churn prediction in banking environments,” 2016, [Online]. Available: http://tesi.cab.unipd.it/53212/.
11. A. Bilal Zoric, “Predicting Customer Churn in Banking Industry using Neural Networks,” *Interdiscip. Descr. Complex Syst.*, vol. 14, no. 2, pp. 116–124, 2016, doi: 10.7906/indecs.14.2.1.
12. H. Sayed, M. A. Abdel-Fattah, and S. Kholief, “Predicting potential banking customer churn using Apache Spark ML and MLlib packages: A comparative study,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 11, pp. 674–677, 2018, doi: 10.14569/ijacsa.2018.091196.
13. Mohanty R., Naga RatnaSree C. (2018) Churn and Non-churn of Customers in Banking Sector Using Extreme Learning Machine in Proceedings of the Second International Conference on Computational Intelligence and Informatics. Advances in Intelligent Systems and Computing, vol 712. Springer, Singapore. https://doi.org/10.1007/978-981-10-8228-3_6
14. Carmona P, Climent F, Momparler A. Predicting failure in the U.S. banking sector: An extreme gradient boosting approach. International Review of Economics and Finance. 2019;61: 304-323
15. Wan J, Yue Z-L, Yang D-H, Zhang Y, Jiao L, Zhi L, et al. Predicting non performing loan of business Bank with data mining techniques. International Journal of Database Theory and Application. 2016;9(12):23-34
16. Ilham A, Khikmah L, Indra A, Ulumuddin A, Iswara I. Long-term deposits prediction: A comparative framework of classification model for predict the success of bank telemarketing. Journal of Physics Conference Series. 2019; 1175(1):1-6
17. Manthoulis G, Doumpos M, Zopounidis C, Galariotis E. An ordinal classification framework for bank failure prediction: Methodology and empirical evidence for US banks. European Journal of Operational Research. 2020;282(2):786-801
18. Krishna GJ, Ravi V, Reddy BV, Zaheeruddin M, Jaiswal H, Sai Ravi Teja P, et al. Sentiment classification of Indian Banks’ Customer Complaints. In: Proceedings of IEEE Region 10 Annual International Conference. India; 17–20 October 2019. pp. 429-434

Figures

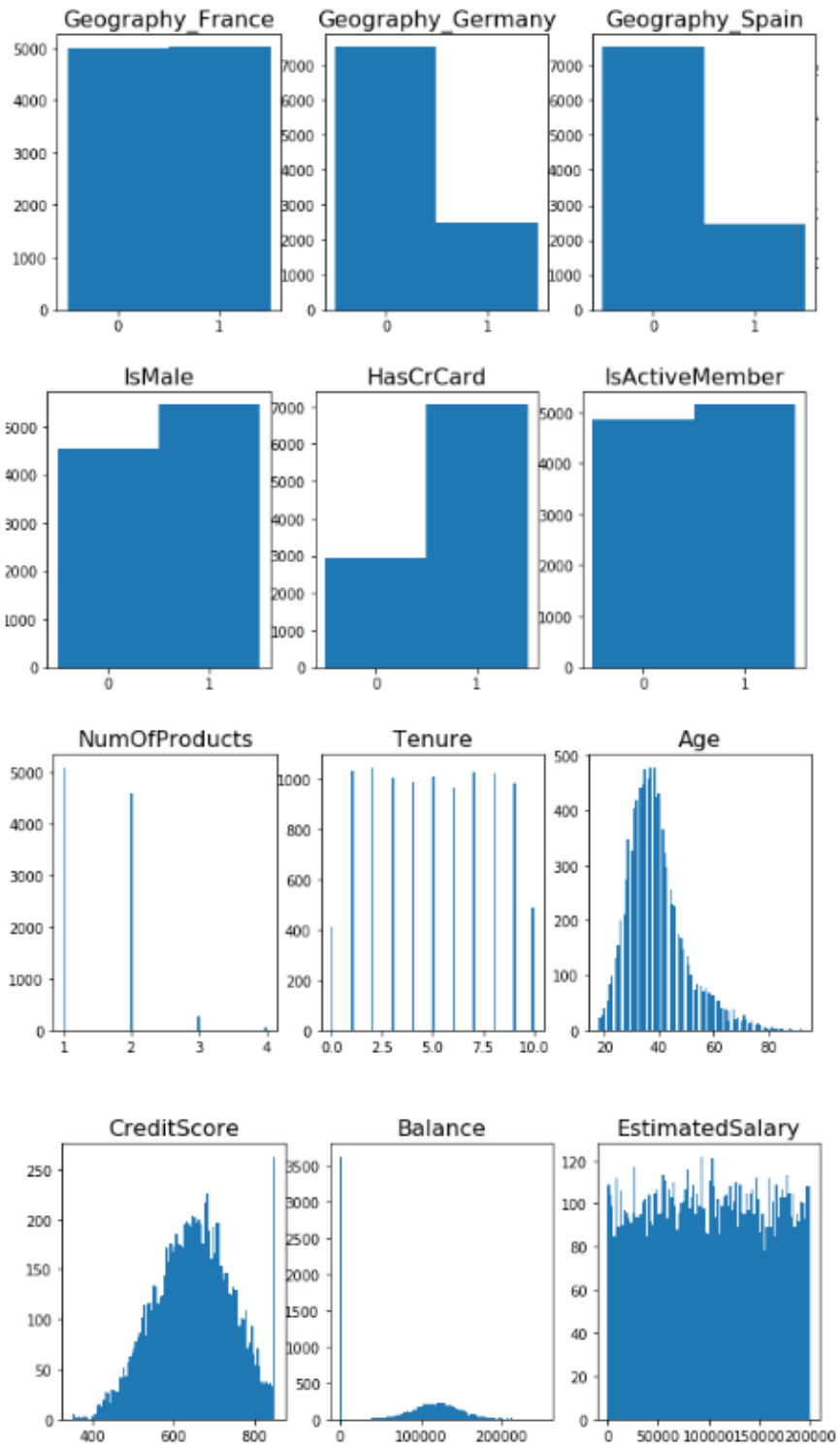

Figure 1

Distribution of categorical variables

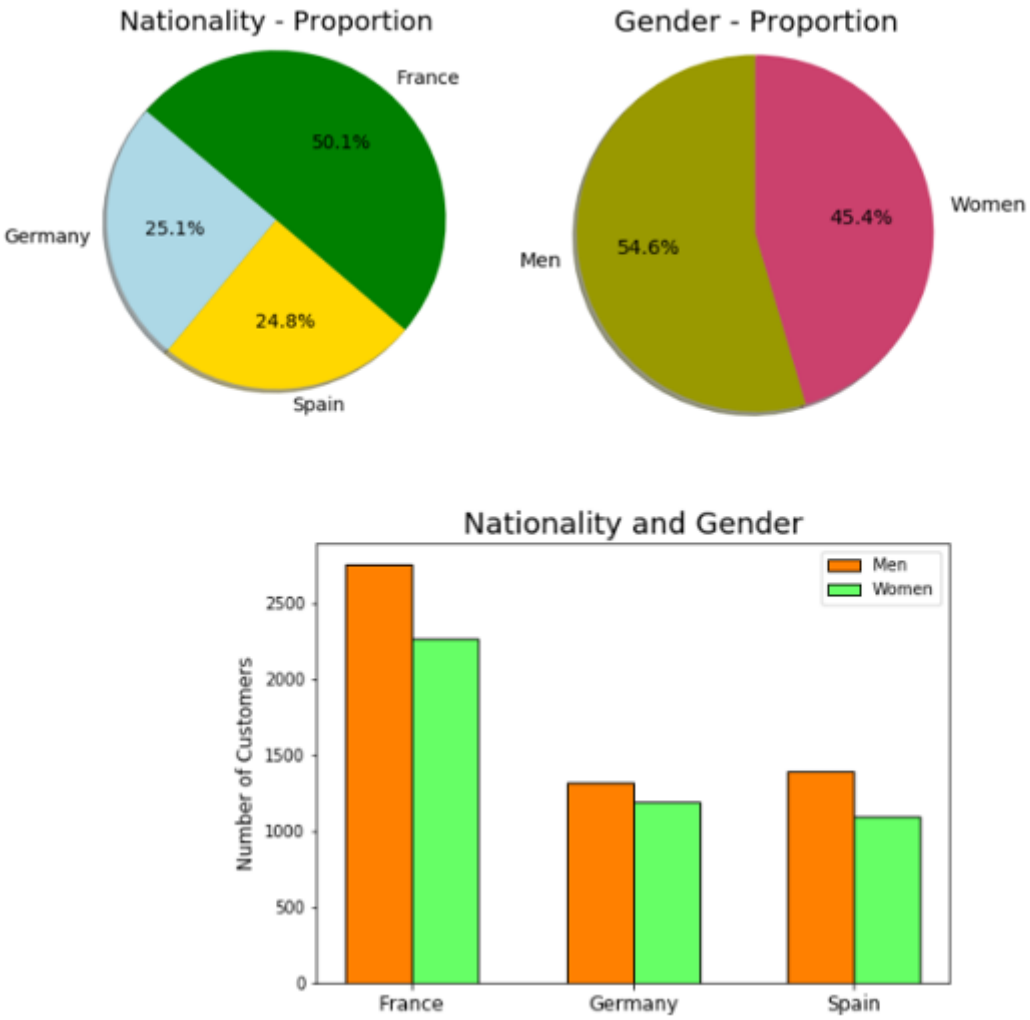

Figure 2

Nationality and Gender Distributions in the dataset

Figure 3

Distribution of salary with gender

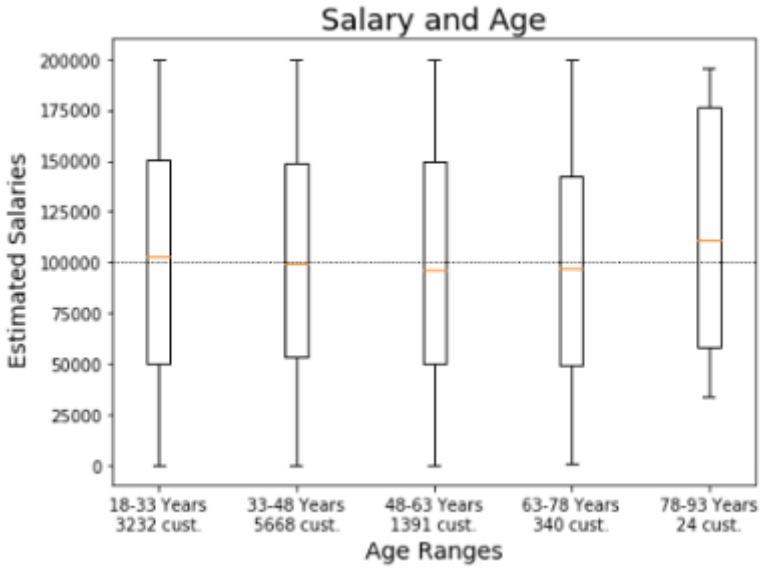

Figure 4

Distribution of salary with age

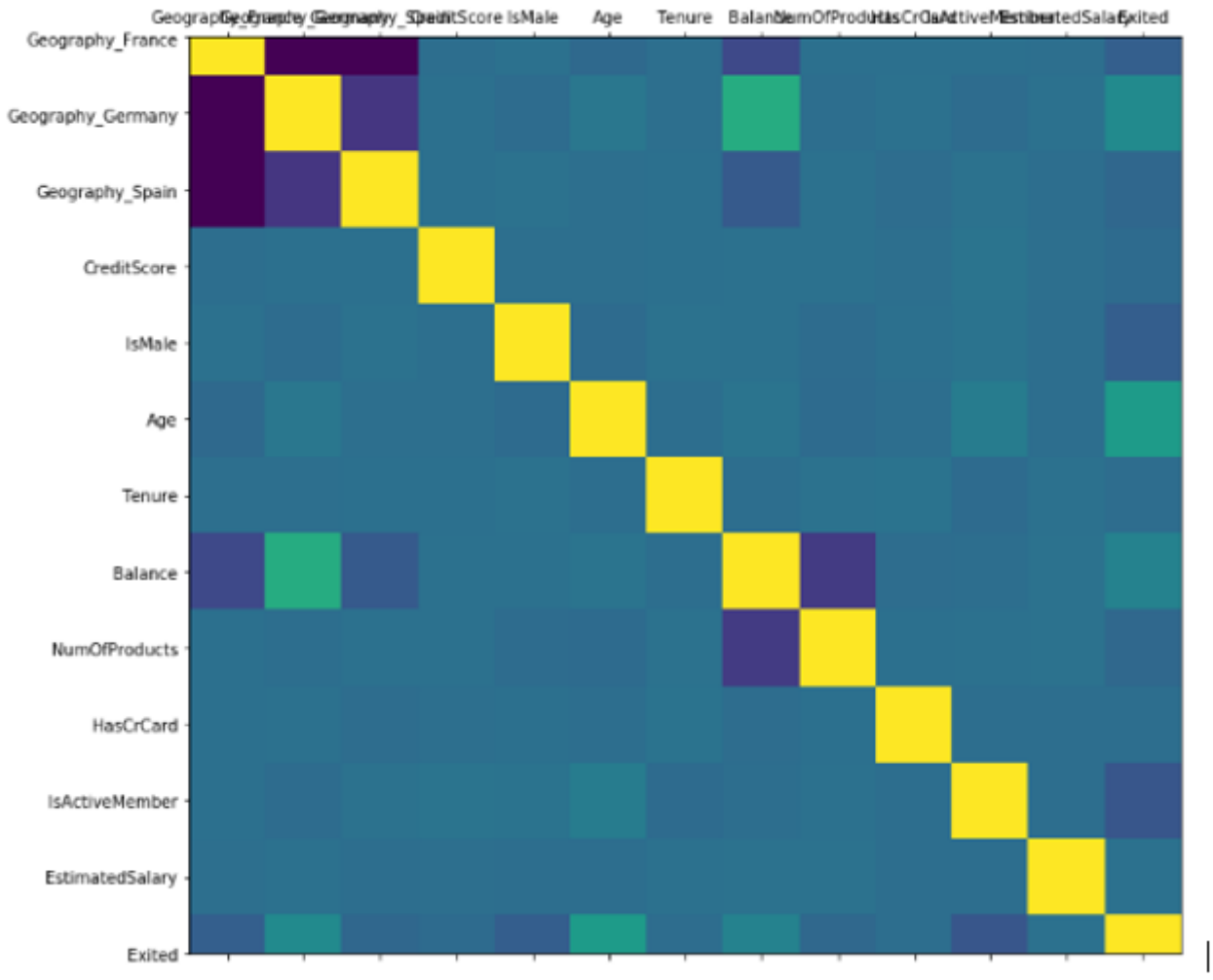

Figure 5

Correlation between variables

Figure 6

Flow Diagram of Gender and country with target value

Figure 7

Pre processing steps

Figure 8

Process Flow Diagram

Figure 9

Performance metric score of linear classifiers on imbalanced data

Figure 10

Figure 9 Performance metric score of non-linear classifiers on imbalanced data

Figure 11

Figure 10 Performance of linear and non-linear classifier on imbalanced data

Figure 12

Figure 11 Performance of boosting classifier on imbalanced data

Figure 13

Figure 11 Comparison o MLP with XGB classifier on imbalanced data

Figure 14

Figure 12 Performance of linear classifiers on balanced data

Figure 15

Figure 13 Performance of SVM on balanced and imbalanced data

Figure 16

Figure 14 Performance of Non-linear classifiers on balanced data

Figure 17

Figure 15 Performance of MLP on balanced and imbalanced data

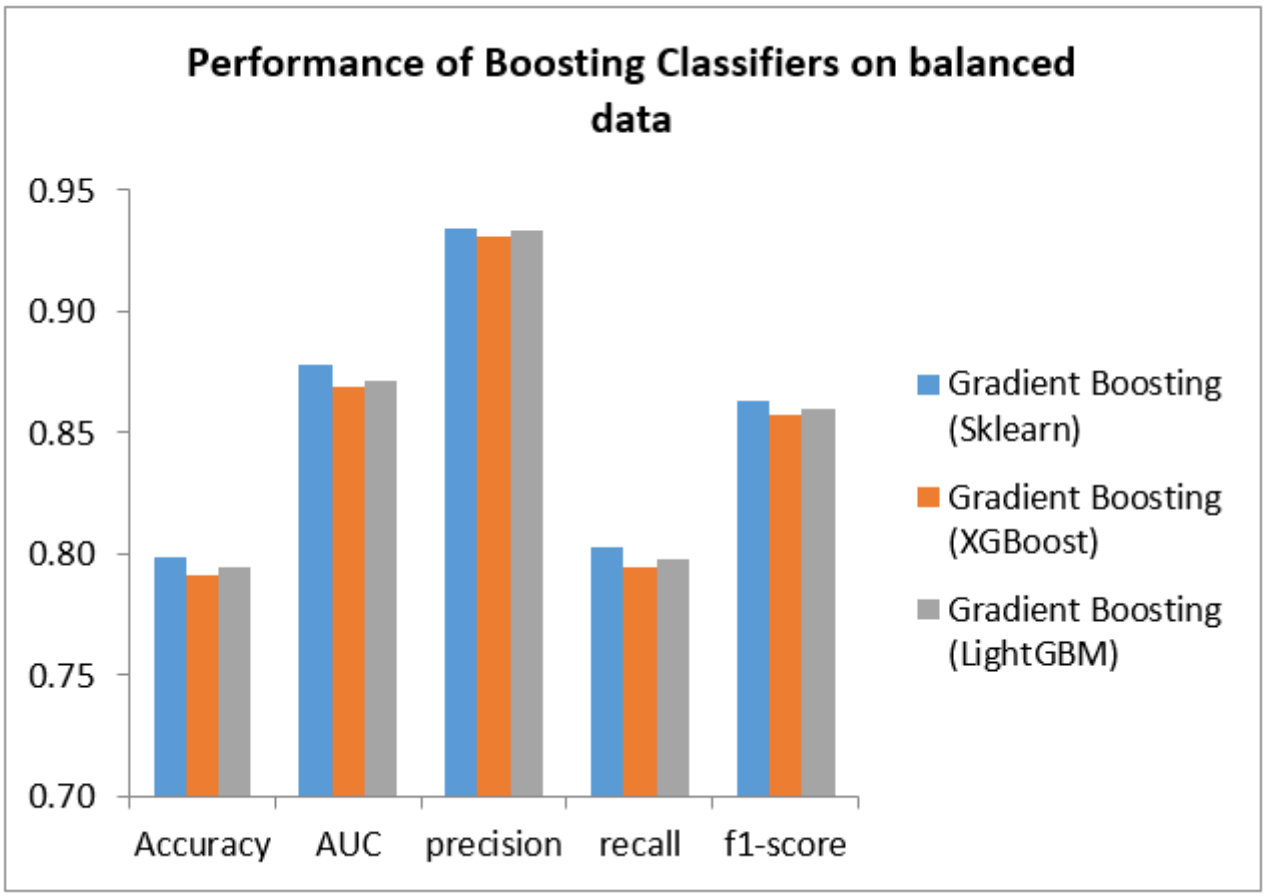

Figure 18

Figure 15 Performance of boosting classifiers on balanced data

Figure 19

Figure 16 Performance of MLP over Gradient boosting classifier on balanced data