International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue: International Conference on Advances in Engineering, Science and Technology – 2021
Organized by: Uttaranchal Institute of Technology, Uttaranchal University, Dehradun
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

A Review on Credit Card Default Modelling using Data Science

Harsh Nautiyal, Ayush Jyala, Dishank Bhandari

UIT, Uttaranchal University, Dehradun, Uttarakhand, India

How to cite this paper: Harsh Nautiyal | Ayush Jyala | Dishank Bhandari "A Review on Credit Card Default Modelling using Data Science" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Special Issue | International Conference on Advances in Engineering, Science and Technology – 2021, May 2021, pp.22-28, URL: www.ijtsrd.com/papers/ijtsrd42461.pdf

IJTSRD42461

1. INTRODUCTION

In the last few years, credit card issuers have become one of the major consumer lending products in the U.S. as well as several other developed nations of the world, representing roughly 30% of total consumer lending (USD 3.6 tn in 2016). Credit cards issued by banks hold the majority of the market share with approximately 70% of the total outstanding balance. Bank's credit card charge offs have stabilized after the financial crisis to around 3% of the outstanding total balance. However, there are still differences in the credit card charge off levels between different competitors.

Credit card is a flexible tool by which you can use bank's money for a short period of time. If you accept a credit card, you agree to pay your bills by the due date listed on your credit card statement. Otherwise, the credit card will be defaulted. When a customer is not able to pay back the loan by the due date and the bank is totally certain that they are not able to collect the payment, it will usually try to sell the loan. After that, if the bank recognizes that they are not able to sell it, they will write it off. This is called a charge-off. This results in significant financial losses to the bank on top of the damaged credit rating of the customer and thus it is an important problem to be tackled in todays world where financial risks are happening vigorously.

Predicting accurately which customers are most probable to default represents significant business opportunity and strategy for all banks. Bank cards are the most common credit card type in the U.S., which emphasizes the impact of risk prediction to both the consumers and banks. In a well-developed financial system, risk prediction is essential for predicting business performance or individual customers' credit risk and to reduce the damage and uncertainty.

Our client ITBCO Bank has approached us to help them to predict and prevent credit card defaulters to improve their bottom line. The client has a screening process, for instance, it has collected a rich data set of their customers, but they are unable to use it properly due to shortage of analytics capabilities.

The fundamental objective of the project is implementing a proactive default prevention guideline to help the bank identify and take action on customers with high probability of defaulting to improve their bottom line. The challenge is to help the bank to improve its credit card services for the mutual benefit of customers and the business itself. Creating a human-interpretable solution is emphasized in each stage of the project.

Even though plenty of solutions to the default prediction using the full data set have been previously done, but there lies a problem in the interpretability ,even in published papers, the scope of our project extends beyond that, as our ultimate goal is to provide an easy-to-interpret default mitigation program to the client bank.Which is done fairly easy by using gradient boosting LightGBM algorithm for prediction.

In addition to default prevention, the case study includes a set of learning goals. The team must understand key considerations in selecting analytics and machine learning methods and how these methodologies can be used efficiently to create direct business value. McKinsey also sets the objective of learning how to communicate complex topics to people with different backgrounds.

The project should include a recommended set of actions to mitigate the default and a clear explanation of the business implications. The interpretability and adaptability of our solution needs to be emphasized when constructing the solution. The bank needs a solution that can be understood and applied by people with varying expertise, so that no further outside consultation is required in understanding the business implications of the decisions.

2. RELATED WORK

There is much research on credit card lending, it is a widely researched subject. Many statistical methods have been applied to developing credit risk prediction, such as discriminant analysis, logistic regression, Knearest neighbor classifiers, and probabilistic classifiers such as Bayes classifiers. Advanced machine learning methods including decision trees and artificial neural networks have also been applied. A short introduction to these techniques is provided here.

K-nearest Neighbor Classifiers K-nearest neighbor (KNN) classifier is one of the simplest unsupervised learning algorithms which is based on learning by analogy. The main

idea is to define k centroids, one for each cluster. These centroids should be placed in appropriately because of different location causes different result. Therefore, the better choice is to place them as much as possible far away from each other. When given an unknown data, the KNN classifier searches the pattern space for the KNN which are the closest to this unknown data. This closeness is defined by distance. The unknown data sample is assigned to the most common class among its KNN.

Discriminant Analysis (DA) The objective of discriminant analysis is to maximize the distance between different groups and to minimize the distance within each group. DA assumes that, for each given class, the explanatory variables are distributed as a multivariate normal distribution with a common variance–covariance matrix.

Logistic Regression (LR) Logistic regression is often used in credit risk modeling and prediction in the finance and economics literature. Logistic regression analysis studies the association between a categorical dependent variable and a set of independent variables. A logistic regression model produces a probabilistic formula of classification. LR has problems to deal with non-linear effects of explanatory variables.

Classification Trees (CTs) The classification tree structure is composed of nodes and leafs. Each internal node defines a test on certain attribute whereas each branch represents an outcome of the test, and the leaf nodes represent classes. The root node is he top-most node in the tree. The segmentation process is generally carried out using only one explanatory variable at a time. Classification trees can result in simple classification rules and can also handle the nonlinear and interactive effects of explanatory variables. But they may depend on the observed data so a small change can affect the structure of the tree.

Artificial Neural Networks (ANNs) Artificial neural networks are used to develop relationships between the input and output variables through a learning process. This is done by formulating non-linear mathematical equations to describe these relationships. It can perform a number of classification tasks at once, although commonly each network performs only one. The best solution is usually to train separate networks for each output, then to combine them into an ensemble so that they can be run as a unit. Back propagation algorithm is the best known example of neural networks algorithm. This algorithm is applied to classify data. In back propagation neural network, the gradient vector of the error surface is computed. This vector points along the line of steepest descent from the current point, so we know that if we move along it a "short" distance, we will decrease the error. A sequence of such moves will eventually find a minimum of some sort. The difficult part is to decide how large the steps should be. Large steps may converge more quickly, but may also overstep the solution or go off in the wrong direction.

Naïve Bayesian classifier (NB) The Bayesian classifier is a probabilistic classifier based on Bayes theory. This classifier is based on the conditional independence which assumes that the effect of an attribute value on a given class is independent of the values of the other attributes. Computations are simplified by using this assumption. In practice, however, dependences can exist between variables. Comparing the results of the six data mining techniques, classification trees and K-nearest neighbor classifiers have the lowest error rate for the training set. However, for the validation data, artificial neural networks has the best performance with the highest area ratio and the relatively low error rate. As the validation data is the effective measurement of the classification accuracy of models, so, we can conclude that artificial neural networks is the best model among the six methods. However, the error rates are not the appropriate criteria for measuring the performance of the models. As, for example, the KNN classifier has the lowest error rate, while it does not perform better than artificial neural networks and classification trees based on the area ratio. While considering the area ratio in validation data, the results show that the performance of the six techniques is ranked as: artificial neural networks, classification trees, Naïve Bayesian classifier, kNN classifier, logistic regression, and Discriminant Analysis, respectively.

3. PROBLEM FORMULATION

With the growth of e-commerce websites, people and financial companies rely on online services to carry out their transactions that have led to an exponential and vigorous increase in the credit card frauds. Fraudulent credit card transactions lead to a loss of huge amount of money to banks as well as various other sectors.

The design of an effective fraud detection system is necessary in order to reduce the losses incurred by the customers and financial companies. Research has been done on many models and methods to prevent and detect credit card frauds. Some credit card fraud transaction datasets contain the problem of imbalance in datasets. A good fraud detection system should be able to identify the fraud transaction accurately and should make the detection possible in real-time transactions. Fraud detection can be divided into two groups: anomaly detection and misuse detection. Anomaly detection systems bring normal transaction to be trained and use techniques to determine novel frauds. Conversely, a misuse fraud detection system uses the labeled transaction as normal or fraud transaction to be trained in the database history. So, this misuse detection system entails a system of supervised learning and anomaly detection system a system of unsupervised learning. Fraudsters masquerade the normal behavior of customers and the fraud patterns are changing rapidly so the fraud detection system needs to constantly learn and update.

Background Timely information on fraudulent activities is strategic to the banking industry as banks have huge databases with variety. Valuable business information can be extracted from these data stores. Credit card frauds can be broadly classified into three categories, that is, traditional card related frauds (application, stolen, account takeover, fake and counterfeit), merchant related frauds (merchant collusion and triangulation) and Internet frauds (site cloning, credit card generators and false merchant sites)

Methodology Basically, there are five basic steps for the data mining process which defines the problem. 1) preparing data 2) exploring the data 3) development of the model 4) exploration and validation of the models 5) deployment and updation in the models. In this project, LightGBM is used as the data mining technique and it utilized above mentioned steps for accurate and reliable result. Moreover, Neural network was used as it has the capability of adaption and generalization. Moreover, python [3] is also a good option for

the experiment purpose. Jupyter is a notebook style open source interface for pyhon. It is an interactive web-based environment that allows persons to combine text, plot, mathematics, executable code in a single document.

4. OBJECTIVES

1. **Higher accuracy of fraud detection**. Compared to rule-based solutions, machine learning tools have higher precision and return more relevant results as they consider multiple additional factors. This is because ML technologies can consider many more data points, including the tiniest details of behavior patterns associated with a particular account.

2. **Less manual work needed for additional verification**. Enhanced accuracy leads reduces the burden on analysts. *"People are unable to check all transactions manually, even if we are talking about a small bank,"* Alexander Konduforov, *data* science competence leader at AltexSoft, explains. *"ML-driven systems filter out, roughly speaking, 99.9 percent of normal patterns leaving only 0.1 percent of events to be verified by experts."*

3. **Fewer false declines**. False declines or false positives happen when a system identifies a legitimate transaction as suspicious and wrongly cancels it.

4. **Ability to identify new patterns and adapt to changes**. Unlike rule-based systems, ML algorithms are aligned with a constantly changing environment and financial conditions. They enable analysts to identify new suspicious patterns

5. METHODOLOGY

DBSCAN

(Density Based Spatial Clustering of Applications with Noise) algorithm is a well-known data clustering algorithm, which is used for discovering clusters for a spatial data set. The algorithm requires the knowledge of two parameters. First parameter is eps which is defined as the minimum distance between two points. It simply means that if the distance between two points is smaller or equal to eps, these points are considered to be neighbors. The second is minPoints: the minimum number of points to form a dense region. For instance, if we define the minPoints parameter as 5, then at least 5 points are required to form a dense region. Based on the parameters Eps and MinPts of each cluster and at least one point from the respective cluster, the algorithm groups together the points that are close to each other [6]. **Gradient boosting** is a popular machine learning algorithm that combines multiple weak learners, like trees, into a one strong ensemble model. This is done by first fitting a model into the data. However, the first model is not likely to fit the model perfectly to the data points so we are left with residuals. We can then fit another tree to the residuals to minimize a loss function that can be the second norm but gradient boosting allows the use of any loss function. This can be iterated for multiple steps which leads to a stronger model and with proper regularization overfitting can be avoided [7]. The gradient boosting has many parameters that need to be optimized to find the best performing model for a certain problem. These parameters include both tree specific parameters like size limitations for leaf nodes as well as tree depth. There are also parameters considering the boosting itself, for example how many models are fitted in order to receive the final model and how much each individual tree impacts the end result. These parameters are usually optimized with a grid search that iterates through all the possible parameter combinations. This is usually computationally expensive since a large number of models have to fitted since the number of parameters needing to be tested increases rapidly as more parameters are introduced

Self-organizing map (SOM), also known as Kohonen network, is a type of artificial neural network that is used to produce low dimensional discretized mappings of an input space [9]. Self-organizing maps produce a grid that consists of nodes, which are arranged in a regular hexagonal or rectangular pattern. The training of a SOM works by assigning a model for each of the nodes in the output grid. The models are calculated by the SOM algorithm, and objects are mapped into the output nodes based on which node's model is most similar to the object, or in other words, which node has the smallest distance to the object on a chosen metric. For real-valued objects, the most commonly used distance metric is the euclidean distance,

although in this study, the sum of squares was used. For categorical variables, the distance metric used in this study is the Tanimoto distance.

The grid nodes' models are more similar to nearby nodes than those located farther away. Since it is the nodes that are being calculated to fit the data, the mapping aims to preserve the topology of the original space. The models are also known as codebook vectors, which is the term used in the R package 'kohonen' used to implement the algorithm [10]. Also, the Tanimoto distance metric is defined under the function supersom details in the package documentation.

In this project, multiple unsupervised self-organizing maps were trained using the demographic variables to produce a two-dimensional mapping serving as a customer segmentation. Different parameters and map sizes were tested to find the optimal mapping that would maximize quality of representation and distance to neighbouring clusters within the map. The maps were also compared on their ability to produce clusters with varying financial impact and default risk measured by the financial model and the default prediction algorithm. The two primary measures used to compare different mappings in this study was the quality (mean distance of objects from the center of node) and the U-matrix distances (mean distance of nodes to their neighbouring nodes). The name quality is used due to how it appears in the kohonen R package.

Preliminary data analysis Describing the data The data consists of 30,000 customers and 26 columns of variables. Each sample corresponds to a single customer. The columns consist of the following variables:

- Default (Yes or no) as a binary response variable
- Balance limit (Amount of credit in U.S. $)
- Sex (Male, Female)
- Education (Graduate school, University, High school, Others)
- Marital status (Married, Single, Others)
- Age (Years)
- Employer (Company name)
- Location (Latitude, Longitude)
- Payment status (last 6 months)
- Indicates payment delay in months or whether payment was made duly
- Bill amount (last 6 months)

- States amount of bill statement in U.S. $
- Payment amount (last 6 months)
- Amount paid by customer in U.S. $ 5

The variables Balance limit, Age, Sex, Education, Marital status, Employer, and Location are defined as demographic variables, since they describe a demography of customers and are available for new customers, unlike the historical payment data which is only available for existing customers.

The total proportion of defaults in the data is 22.12% which is 6,636 out of the total data set comprising of 30,000 samples. This could be due to a large bias and therefore not a realistic representation of the bank's customer base. However, the data was collected during a debt crisis which provides an argument for the assumption that the data represents a non-biased sample of the customer base. In any case, the high amount of defaults in should be taken into consideration when making generalizations about the results or methodology of this case study. The high number of defaults will especially have an effect on estimates of the bank's financials.

Default
This variable indicates whether or not the customer defaulted in their credit card debt payment. For the purpose of this project, predicting default is the main focus of the data analysis. A value of 1 indicates default, and a value of 0 indicates no default. It is unclear how long after the collection of the data this variable is measured. This means that default could have happened the following month or a longer time thereafter. Since this is unknown, no assumptions are based on the time of default. It is also not clear whether a value of 1 indicating default means the client missed only a single payment or multiple and whether or not the time of delay in payment was taken into account.

Balance limit
states the amount of given credit in US $. This is the maximum amount a customer can spend with their credit card in a single month. The amount of balance limit is dependent on the bank's own screening processes and other unknown factors.

Sex
This variable can obtain a value of 1 for male and 2 for female. In this study, sex and gender are used interchangeably to intend the same thing. It is unknown whether the difference between the two definitions were taken into account when the data was collected.

Education
The education level of a customer is represented as one of four values: 1 = Graduate school, 2 = University, 3 = High school, 4 = Other. For the purpose of analysing customer groups, this is assumed to indicate the highest level of education completed.

Marital status
Referred to as "married" in the analysis, this variable can obtain three values: 1 = Married, 2 = Single, 3 = Other such as divorced or widowed.

Age
of the customer is stated in years.

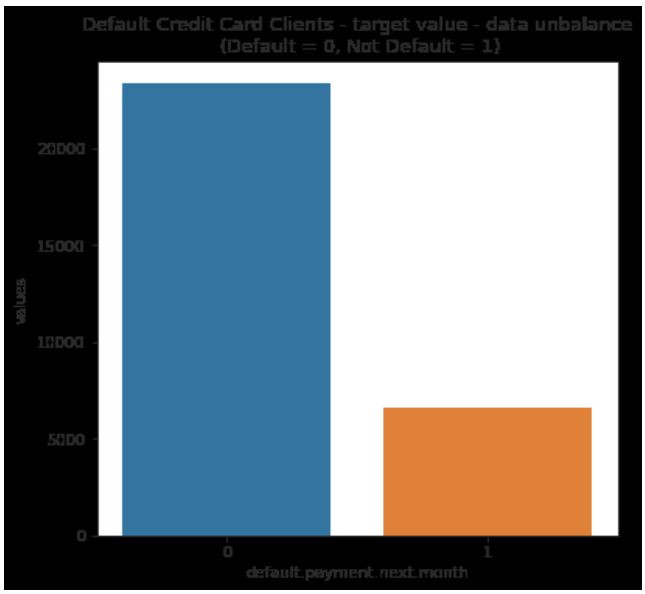
Location
This variable is composed of two different values for each customer. One is for the latitude, and the second one is for the longitude. In order to gain benefits from this data in predictions using only the demographic variables, we applied the DBSCAN algorithm.

Payment status
is represented as 6 different columns, one for each month. The value of payment status for a month indicates whether repayment of credit is was delayed or paid duly. A value of -1 indicates pay duly. 6 Values from 1 to 8 indicate payment delay in months, with a value of 9 defined as a delay of 9 months or more. Data collected from 6 months, April to September.

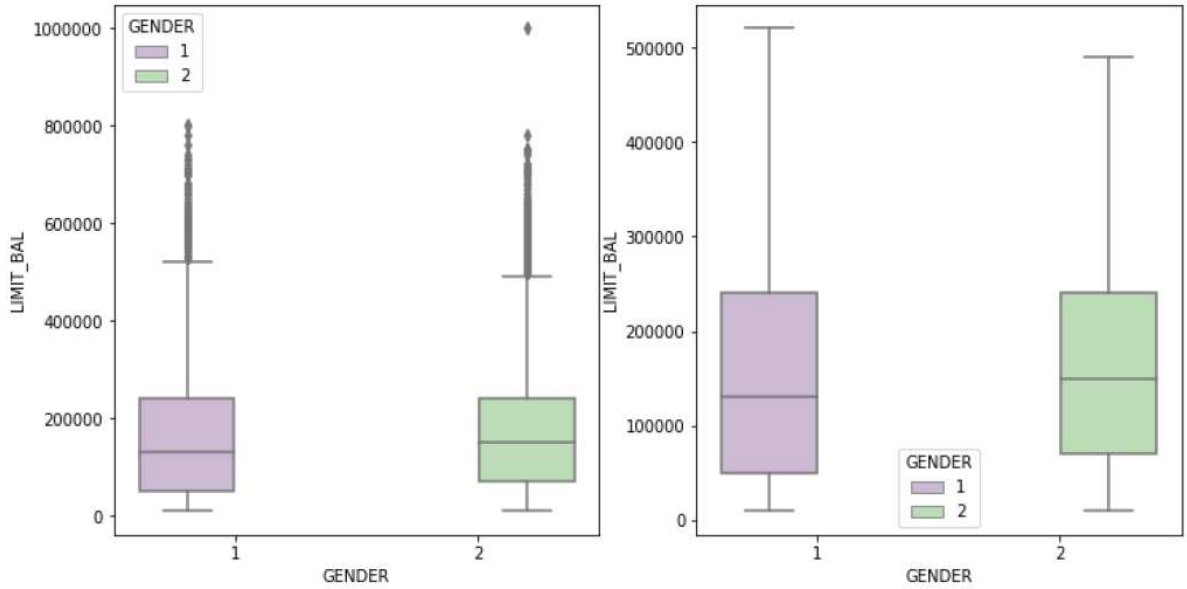
Bill amount
Amount of bill statement in U.S. $ is recorded in this variable. It is represented in the data as 6 columns, one for each month. Data collected from 6 months, April to September.

Payment amount
Amount of previous payment in U.S. $, stored in 6 different columns for each month, similarly to payment status and bill amount. The payment amounts correspond to the same months as payment status and bill amount. For example, the payment amount for April indicates amount paid in April.

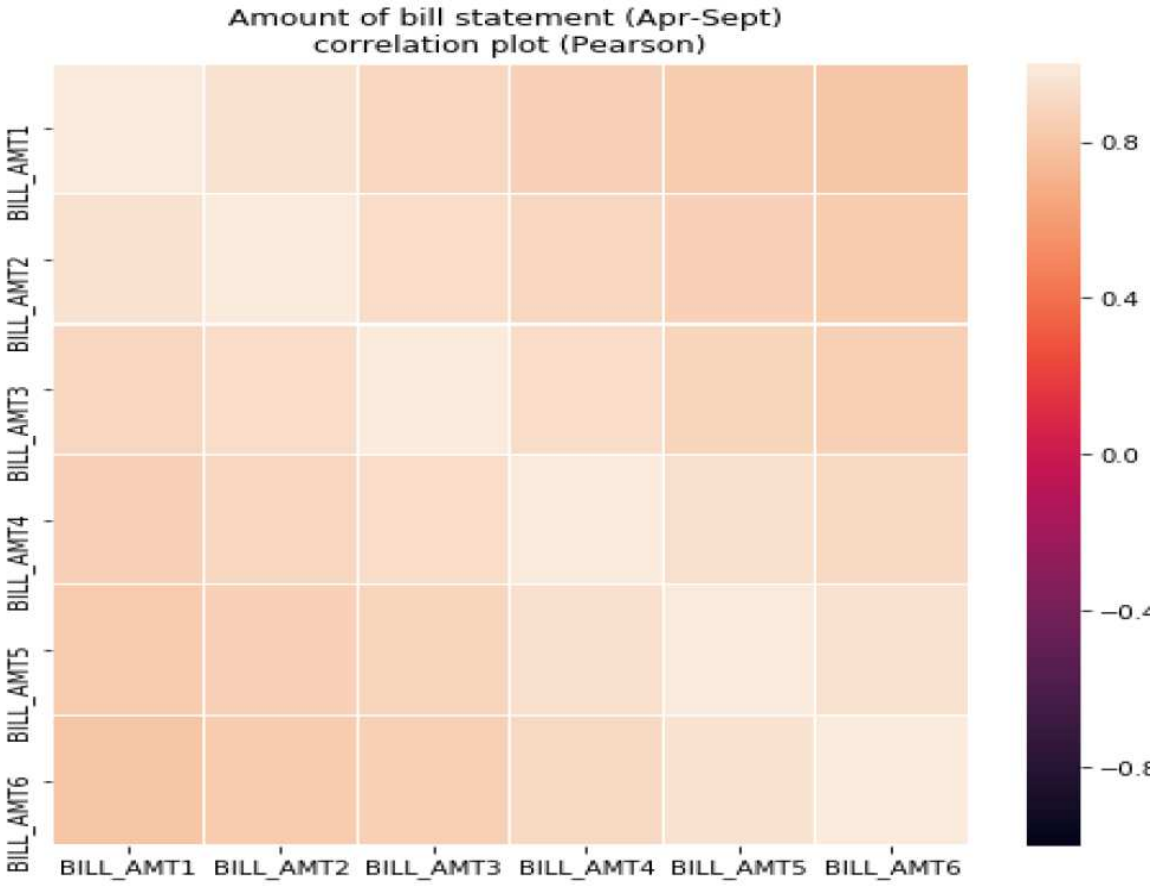
Checking data unbalance:



Preliminary data analysis



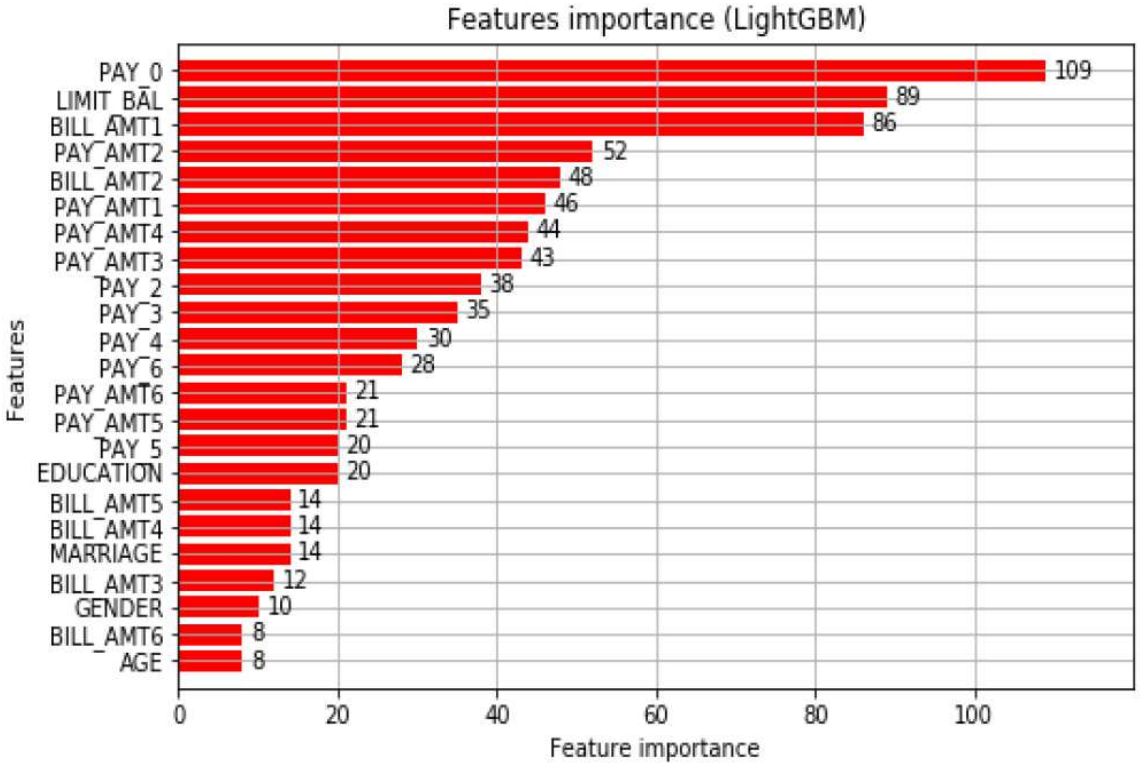
Features correlation



Using mainstream (LightGBM) algorithm:
Training the dataset:

[LightGBM] [Warning] Auto-choosing col-wise multi-threading, the overhead of testing was 0.006994 seconds.
You can set `force_col_wise=true` to remove the overhead. Training until validation scores don't improve for 50 rounds
23. train's auc: 0.778238 valid's auc: 0.771173
[100] train's auc: 0.789346 valid's auc: 0.782605
[150] train's auc: 0.794861 valid's auc: 0.784753 Early stopping, best iteration is:
[135] train's auc: 0.793452 valid's auc: 0.785154
Out[62]:
33
Best validation score was obtained for round 135, for which AUC ~= 0.78.

Plotting the variable importance



6. CONCLUSION

The results of analysis and predictive modelling show that neither directly measuring or using predicted proportion of defaults of a customer group to predict default is accurate. This is most likely due to multiple reasons. One of them being the limitations in accuracy of any machine learning algorithm caused by the small number of variables or due to missing values. Another reason is most likely the lack of specificity in customer segments, mixing up actual high risk customers with those of low risk. Comparing paying amounts, gender and maternal status in the training set and test set also showed large variation. This is most likely due to the high losses that a single customer can produce by defaulting with high amounts of debt. Much of the variation in the data could not be represented, since customer segmentation was only done using the demographic variables. Further analysis should be done in order to fully justify and support business decisions based on the customer segmentation in this study.

When it comes to default prediction, we have a model that is able to predict the defaults of customers with high enough certainty that the bank can utilize it in their functions. Assuming that the banks continues to receive customers that are represented in our dataset we could implement our model in the banks preliminary screening process and it would bring financial gain to the bank.

However, our solution is not viable to be used as a standalone system in its current form since it only considers part of the banks actions. Many factors that were not covered in this case study should be taken into consideration when taking any business action. For example young people could be preferable for the bank since they stay longer as a customer so it could be in banks interest to favor having them as a customer even if our model would suggest 26 otherwise.

Single customers should not be discriminated against especially based on the customer segmentation which relies on calculating averages over a group. A single customer defaulting with high debt can result in much higher losses than might be anticipated simply based on averages.

Similarly, the analysis does not go in-depth enough to justify assuming that the variables used in this study could explain or predict how reliable the customers are on the long run, especially considering that the data was collected during a debt crisis.

7. REFERENCES

[1] Wikipideahttps://www.8051projects.net/files/public/1259220442_20766_FT0_7380969-line-follower-using-at89c51.pdf

[2] Default Credit Card Clients Dataset, https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset/

[3] RandomForrestClassifier, http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[4] ROC-AUC characteristic, https://en.wikipedia.org/wiki/Receiver_operating_characteristic#Area_under_the_curve

[5] AdaBoostClassifier, http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html

[6] CatBoostClassifier, https://tech.yandex.com/catboost/doc/dg/concepts/python-reference_catboostclassifier-docpage/

[7] XGBoost PythonAPI Reference, http://xgboost.readthedocs.io/en/latest/python/python_api.html

[8] LightGBM Python implementation, https://github.com/Microsoft/LightGBM/tree/master/python-package

[9] LightGBM algorithm, https://www.microsoft.com/en-us/research/wp-content/uploads/2017/11/lightgbm.pdf

[10] Chauhan, N., Dhaundiyal, R., & Joshi, K. K-MEANS ON SEARCH ENGINE DATASET THROUGH WEKA. *International Journal of Research Fellow for Engineering (IJRFE)–Volume*, *4*.

[11] Joshi, K., Rawat, S., & Chaudhary, S. ANALYSIS OF DIFFERENT OPTICAL SWITCHING TECHNIQUES IN NOC ROUTER ARCHITECTURE. *International Journal of Research Fellow for Engineering (IJRFE)–Volume*, *4*.

[12] Joshi, K., Chaudhary, S., & Chauhan, N. HYBRID CLUSTERING ALGORITHM USING K-MEANS CLUSTERING ALGORITHM. *International Journal of Research Fellow for Engineering (IJRFE)–Volume*, *4*.

[13] Longkumer, M., Joshi, K. A Comprehensive Study on Recent Botnet. *International Journal of Science and Research (IJSR)- Volume*, 7.

[14] Joshi, K., Gupta, H., & Lamba, S. An Overview on Image Fusion Concept. Journal of Emerging Technologies and Innovative Research (JETIR)–*Volume*, *5, 873-879*.