Credit Card Lead Prediction using SMOTE, Random Forest, XGBoost, and LightGBM

*Abstract*— Every bank management contains the enormous amount of data of their customers. In this paper we are going to discuss about the problem of Happy Customer Bank which is a private bank and deals in all kinds of banking services such a saving accounts, current accounts, credit products and other offerings. Now the bank wants to segment those customers who are eligible and interested for taking credit cards as they wants to cross sell their credit cards to the existing customers. So, in this research we build the ensemble model to help the bank in identifying such customers based on the data available to us. The problem with the data is that it is a kind of imbalance data, so at first we just balance the dataset using the SMOTE function and after that applying bagging (Random Forest) and boosting (XGBoost & LightGBM) techniques to build the predictive model and at last use AUC score to evaluate the performance. The model outperformance with an AUC score of 90%, so this can be easily deployed in market to identify the interested customers in real time.

Keywords—Lead Prediction Model, Imbalance class, SMOTE, Random Forest, XGBoost, LightGBM, and AUC score.

# Introduction

To build any kind of Machine Learning model, we first need to identify its learning type. There are 3 kinds of Learning. 1. Supervised Learning, 2. Unsupervised Learning and 3. Reinforcement Learning. Now as we can see that the dataset contains target variable/labels so this falls under the criteria of supervised learning. With the supervised learning also we can solve 2 kind of problem. 1. Regression and 2. Classification. In this case we need to build the predictive model which can help us in classifying whether the customer is interested for credit card or not. In every machine learning pipeline we need to follow 4 steps. 1) Collect, read and describe the data. 2) Exploratory Data Analysis. 3) Data Pre-processing and Feature Engineering. 4) Model Building and Performance Evaluation. Two important concepts which is used in this research paper is SMOTE function which actually helps in balancing the imbalanced dataset and the second one is Ensemble Learning method which we use to build our predictive model. This model helps bank to segment the interested customer for selling their credit cards.

# the data set

The dataset is collected from the kaggle Machine Learning repository. The dataset contains 245725 examples in training set with 10 columns and 1 target variable and it contains 105312 examples in test set with 10 columns. The dataset is quite larger in size. Out of these 10 columns 1 column represent the ID of the customer which definitely put no impact on modelling. Then out remaining 9 features 6 are categorical and 3 features are numerical. The target variable “Is\_Lead” represents whether the customer is interested for credit card or not (0 or 1). 0 means customer is not interested and 1 means customer is interested. Now based on customer’s personal details (gender, age, region and occupation) and his/her relationship with the bank (Channel code, vintage, account balance, and activeness) we need to build the classification model. Now with the help of visualization graphs and statistical analysis we need to analyze the importance of each feature, and we also need to realize which feature contains an outliers or irrelevant which needs to be drop from the dataset in order to build the best model.

# Exploratory data analysis

In EDA, we analyze our dataset with the help of statistical graphs and data visualization methods. We can perform various Univariate and bivariate analysis to analyze our dataset. So, first we need to check the distribution of each class in the target variable.

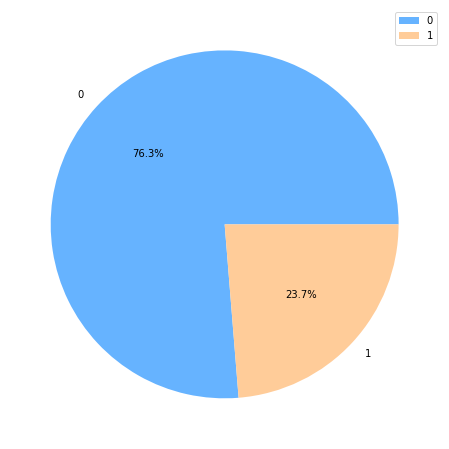


Fig 1. Class Imbalance

We can observe from the above pie chart that 76.3% people are not interested for credit card and 23.7% people are interested and thus the dataset is quite imbalanced. Now with bivariate analysis we can get some meaning full insights from data.

* The people who are not active and didn’t credit any product are less interested.
* Male customers are more interested for credit card in comparison with female customers, while percentage of non-interested customers are high in both the genders.

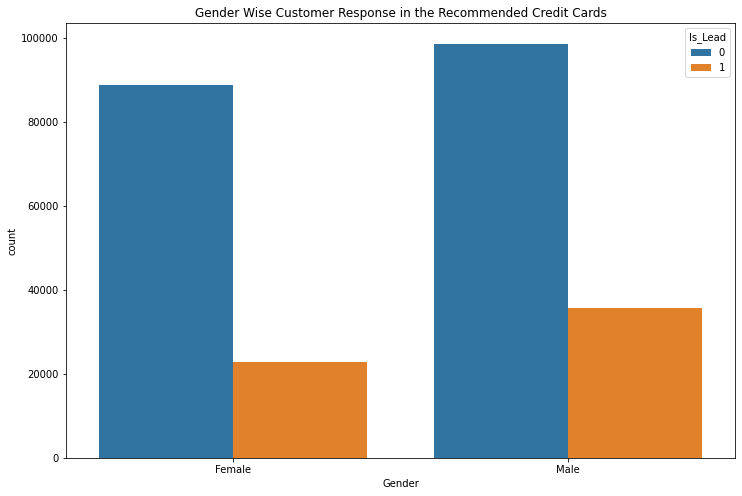


Fig 2. Gender vs. Lead

* Self-employed people are more interested for credit card.

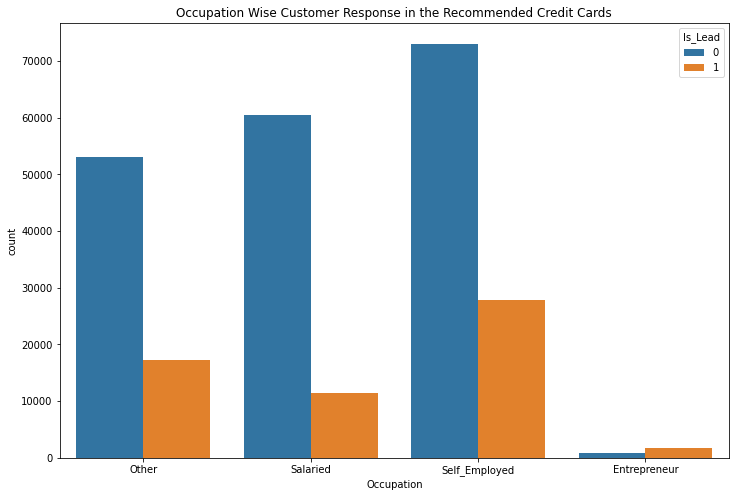


Fig 3. Occupation vs. Target

* We also observe that Avg.\_Account\_Balance column has very high outliers while Age column has no outlier.

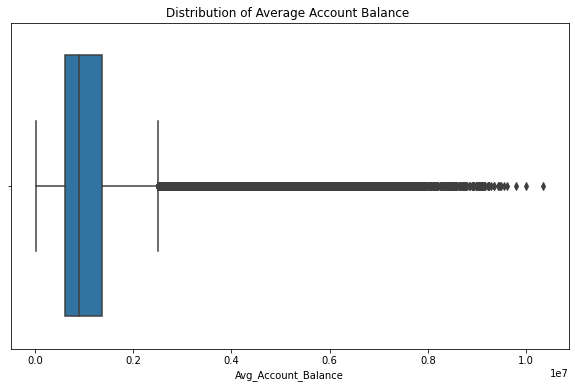


Fig 4. Boxplot of Account Balance

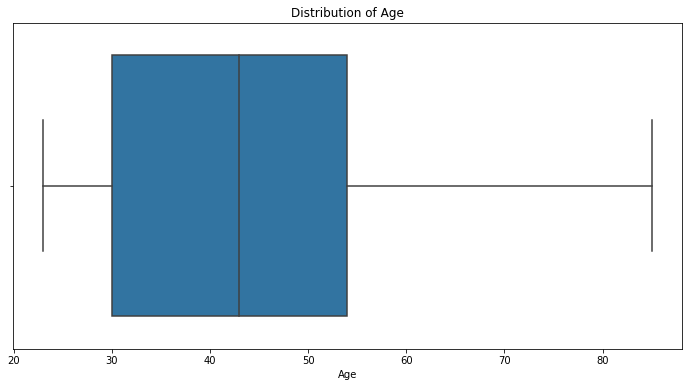


Fig 5. Boxplot of Age

* One strange observation is that in Credit\_Product column there are some missing values and that missing values have more leads.

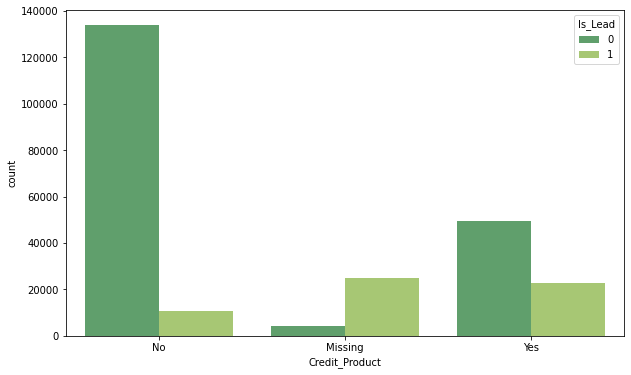


Fig 6. Credit Product vs. Target

Now our next step is Feature Engineering and Data Preparation for modelling.

# data prepataion

In this section we are going to discuss about the various feature engineering techniques such as filling the null values, categorical variable encoding, standardization, handling the imbalance data with SMOTE, and finally split the data into training and test set for modelling.

## Categorical Variable Encoding

The dataset contains 6 categorical variables which we need to handle. There are multiple ways to handle these categorical features but out of those methods 2 are more popular. First one is called one-hot encoding method and the second one is Label encoding. In one-hot encoding method we simply create variables and this is the effective way also but as we have 6 categorical columns and each column contain at least 2 class so it will increase the dimension of dataset and that’s a reason we use Label encoding method to handle these categorical columns.

## Data Standardization

As we can observe that the numerical columns are not normally distributed especially Avg\_Account\_Balnace column are highly skewed and it contains very large values also. So, in order minimize this skewness we first apply log on account balance column and then apply standardization on each numerical column. After standardization, the mean of the distribution becomes 0 and its standard deviation becomes 1.

## Handling Imbalnce Dataset

As we already observe that our dataset is imbalanced, so there are different techniques with which we can balance the dataset. They are 1) Random Under sampling, 2) Random Over sampling, and 3) SMOTE. Out of these three techniques SMOTE is considered to be the most effective one to solve this problem. So, let’s understand what is SMOTE and how it works.

SMOTE stands for synthetic minority over sampling technique. It creates new synthetic observations or new data points and the steps are:

* Plot every data point.
* Identify feature vectors along with its nearest neighbor and calculate the distance between these two.
* Then multiply these differences with random number between 0 and 1.
* Now identify the new point on the line segment by adding the random number to the feature vector.
* Repeat the process for identified feature vector.

Now the last step in data preparation is split the data into feature vector and targets and finally split the complete dataset into training and test set, where on 80% of the data, we train our algorithm to build the model and on remaining 20% of the data we evaluate our model’s performance.

# Model Building

## Random Forest

Random Forest is a kind of bagging technique which is built with the combination of several deep decision trees. Here deep decision trees means the decision trees whose depth is very large and definitely variance is very high. In this technique we combine several strong learners whose variance are very high and bias is low and then at a time of aggregation we minimize this variance and finally get the best model as per our need.

## Extreme Gradient Boosting

XGBoost is a decision tree based boosting technique. In several cases this algorithm outperforms even neural network. The algorithm is highly efficient, fast and portable. It applies the principle of boosting week learners and gradient descent algorithm. Generally, in boosting technique all the week learners are learn in a sequential manner, but XGBoost provide parallel tree boosting which helps in building the model in a more efficient, faster and accurate way.

## LightGBM

This is again a gradient boosting framework which works exceptionally very fast in comparison with any other tree based algorithm. Light GBM works exceptionally well when the dataset is very large as it takes less memory and gives faster result. This algorithm is widely used in the Hackathon and data science competitions where the datasets are very large and we want a fast and an accurate result. One very important thing about the LightGBM is that it is having more than 100 parameters.

# Result and discussion

In this section we are going to discuss the result of the model. Till now we build the model with 3 three different classification algorithms and now we need to evaluate the result the result with the help of different evaluation metrics. For this particular problem statement, since we have an imbalanced dataset we need to find the AUC score. AUC stands for Area Under the Curve and it measures the ability of classifier to distinguish between the 2 classes. It typically uses the summary of ROC curve which stands for Receiver Operating Curve in which on x-axis we have FPR (False Positive Rate) and on y-axis we have a TPR (True Positive Rate) and higher the AUC score means better the performance of the model and if the AUC score is 1, then it means the classifier have an ability to completely distinguish between 2 classes.

Now, when we make comparison between these models then out of these 3 models XGBoost perform excellent in comparison with Random Forest and LightGBM. XGBoost performs excellent as the AUC score of this model is 0.904, LightGBM attains an AUC score of 0.898 and Random Forest able to achieve the AUC score of 0.893.

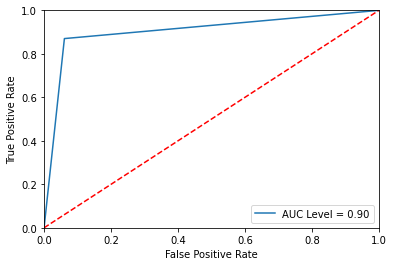


Fig 7. ROC Curve

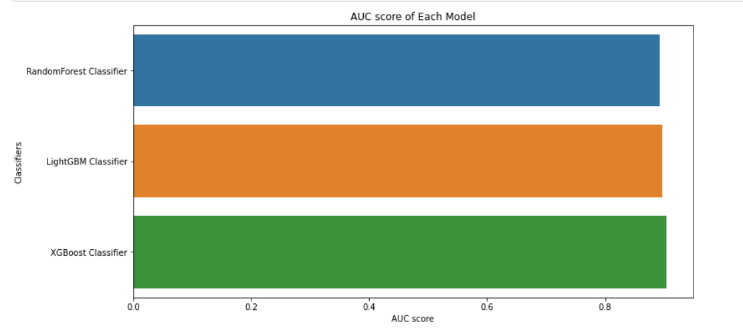


Fig 8. Performance Comparison

# Conclusion and Future research

SMOTE is use in this research paper to overcome the problem imbalance dataset. SMOTE helps in balance our dataset and there by helps in improve the result of the model. For this problem statement we didn’t use any single machine learning algorithm because of the reason that the dataset is quite large and none of the machine learning algorithm gives accurate and faster result with such large dataset. So, we directly go for the ensemble learning methods as they work faster and also helps in achieving an excellent result. We can observe that XGBoost model performs exceptionally well and give us the outstanding result.

In future we can also use Random Oversampling method to balance the dataset as in some cases it gives better result in comparison with SMOTE as well as we can also use Artificial Neural Network (ANN) model to improve the performance.

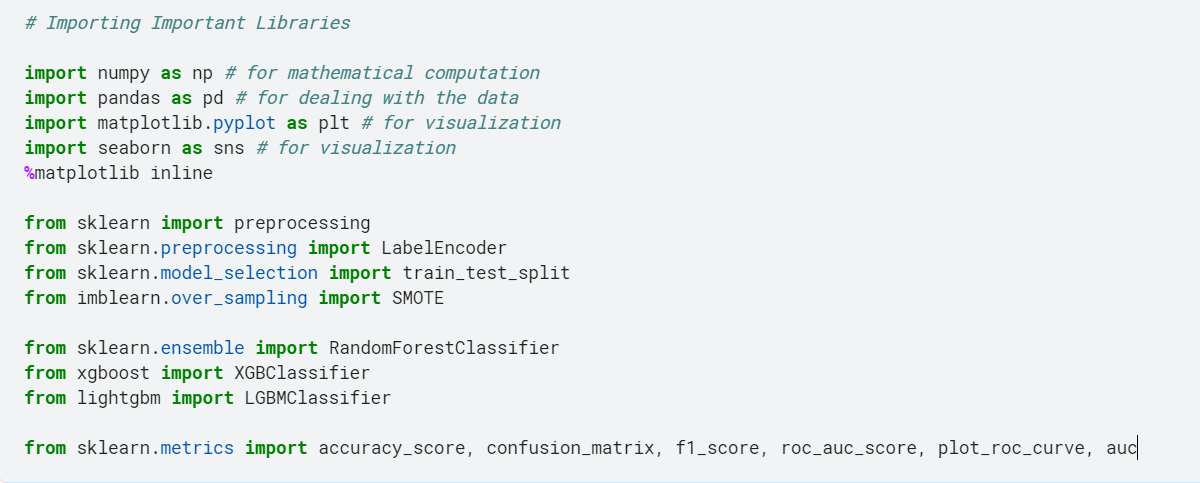
# Refereces

1. Machine Learning Applications in Complex Economics and Financial Networks. (2021). *Machine Learning Applications in Complex Economics and Financial Networks*, 1. https://www.hindawi.com/journals/complexity/2021/6618841/
2. Deep Learning Methods Applied to Complex Big Data Analysis. (2020). *Deep Learning Methods Applied to Complex Big Data Analysis*, 1. https://www.hindawi.com/journals/complexity/2020/8706285/
3. Predictive analysis of credit score for credit card defaulters. (2019). *Predictive Analysis of Credit Score for Credit Card Defaulters*, 1. https://www.researchgate.net/publication/332557433\_Predictive\_analysis\_of\_credit\_score\_for\_credit\_card\_defaulters
4. An Investigation of Credit Card Default Prediction in the Imbalanced Datasets. (2020). *An Investigation of Credit Card Default Prediction in the Imbalanced Datasets*, 1. https://www.researchgate.net/publication/344914401\_An\_Investigation\_of\_Credit\_Card\_Default\_Prediction\_in\_the\_Imbalanced\_Datasets
5. Predicting Credit Card Approval of Customers Through Customer Profiling using Machine Learning. (2020). *Predicting Credit Card Approval of Customers Through Customer Profiling Using Machine Learning*, 1. https://www.ijeat.org/wp-content/uploads/papers/v9i4/D7293049420.pdf
6. Ensemble Learning For Imbalanced Data Classification Problem. (2015). *Ensemble Learning For Imbalanced Data Classification Problem*, 1. https://www.researchgate.net/publication/300691822\_Ensemble\_Learning\_For\_Imbalanced\_Data\_Classification\_Problem

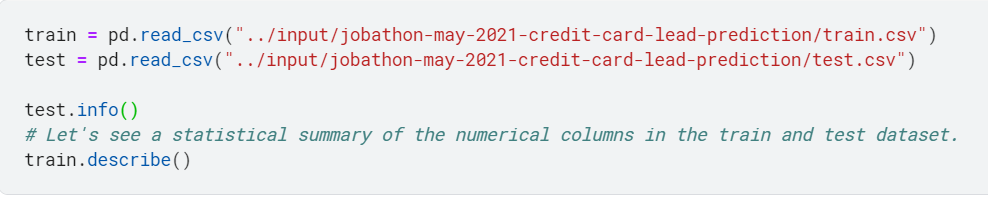
# Appendix

## Source Code

### Importing Important Libraries



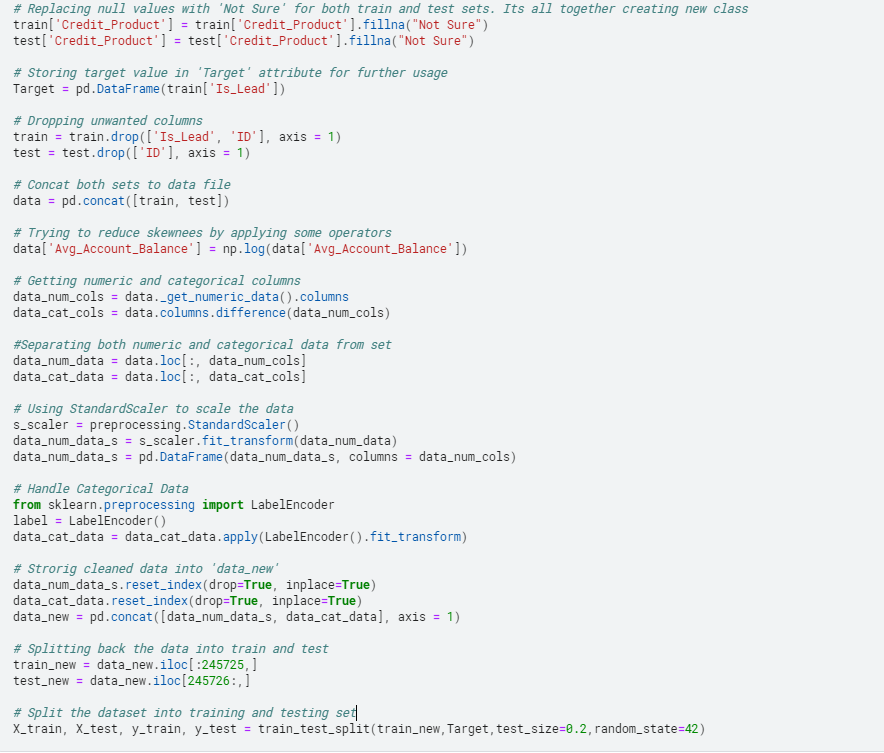
### Reading and Viewing the Dataset



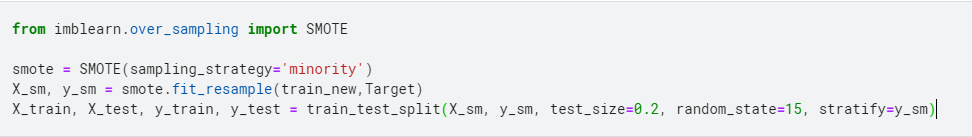
### Exploratory Data Analysis



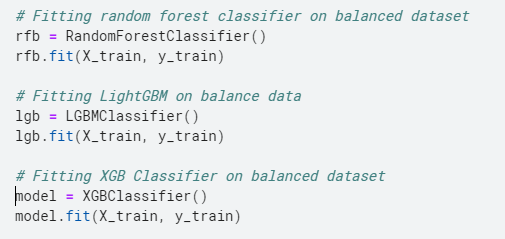
### Data Preprocessing and Preparation



### Balancing the Dataset



### Model Building (Random Forest, LightGBM, and XGBoost)



### Model Evaluation

