# **Credit card default Prediction using a classification model**

# **(Shubham Bhadouria)**

**Data science trainee,**

**AlmaBetter, Bangalore**

**Abstract:**

### Financial threats are displaying a trend about the credit risk of commercial banks as the incredible improvement in the financial industry has arisen. In this way, one of the biggest threats faces by commercial banks is the risk prediction of credit clients. Credit cards are now the most preferred way for customers to transact either offline or online. All our digital transactions through credit card statements are far easier and quicker as compared to cash transactions. One downside that has been witnessed over the past few years of this increasing digital phenomenon is the rise of fraud on credit cards. This project is aimed at predicting the case of customers default payments in Taiwan.

***Keywords: Exploratory Data Analysis, Train -Test split Classification, SMOTE Machine learning models (RF, KNN, SVM, XGB, LGBM), SHAP***

1. **Problem statement**

After understanding the criticality of the fraud situation worldwide, the question that comes to mind is how we prevent this fraud and the damage it causes to the overall economy and especially to the customer sentiments and trust in financial institutions. Some of the challenges that we face while dealing with credit card fraud dataset are:

**Data imbalance:** The fraud and non-fraud data are prone to skewness. Like in our dataset of Taiwan with here, we have **6636** frauds out of a total of 30000 transactions which is **22 %** of all the transactions. So, it is easy to achieve almost **80 %** accuracy with a naive model which just predicts all the transactions as non-fraud.

**Real-Time Detection:**Most of the fraud detection models in practice works under very strict timing conditions like transaction-level fraud detection model. Such models irritate the customer who is waiting to do the transaction, and if we process too fast, we may improve on customer experience, but it might lose out on accuracy.

**Customer friction:**The most likely outcome if a model predicts a current transaction as fraud is to decline the transaction outright to prevent any financial loss. It sometimes proves to be a bone of contention with genuine customers, who might get declined if the model has too many false positives or Type 1 errors.

**2. Introduction**

Fraud detection is a set of activities that are taken to prevent money or property from being obtained through false pretences. Fraud can be committed in different ways and in many industries. Most detection methods combine a variety of fraud detection datasets to form a connected overview of both valid and non-valid payment data to make a decision. **Credit Card Fraud Detection with Machine Learning** is a process of data investigation and the development of a model that will provide the best results in revealing and preventing fraudulent transactions.

## **3. Related Work**

In the prediction of “two classifications”, a few categories are called positive examples (default), and most categories are called counterexamples (no default). However, most of the credit card loan data are unbalanced. In this model SMOTE algorithm is used to balance the target feature classes and after that data has been feed to different ML supervised models. All subsections are structured as follows:

**3.1 Data Information and transformation**

We had to perform a few imputations and transformations on our dataset to create the desired visualizations. There were no null values and duplicate values in the dataset. We rename some columns and extract useful information from them.

The size of the dataset is 30,000 of which 6,636 are in the positive category (default) and 23,364 are in the negative category (no default). The dataset has a total of 25 features. In this project, considering that the variable ID has no relationship with the target variable, so we drop that column and finally we have 23 independent variables and 1 target variable.

# **3.2 Machine learning Models:**

**Credit Card Fraud Detection with Machine Learning** is a process of data investigation and the development of a model that will provide the best results in revealing and preventing fraudulent transactions. This is achieved through bringing together all meaningful features of card users. The information is then run through a subtly trained model that finds patterns and rules so that it can classify whether a transaction is fraudulent or is legitimate. In our project, we fitted our data to five ML classification models.

**4. Methodology:**

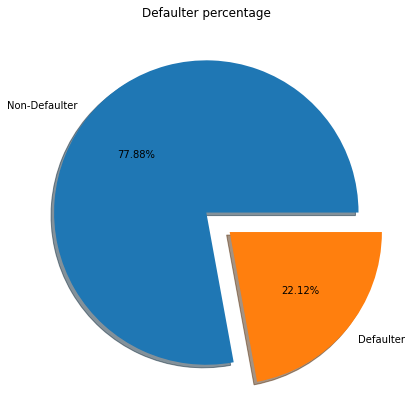
The existing methodologies for predictions are random forests, KNN, SVM, XGBoost, and LGBM. This research work allows having insight into the performance of various prediction algorithms and walks through the whole process of prediction.

* **Data pre-processing and transformation**
* **Developing and optimizing KNN model**
* **Developing and optimizing Random Forest model**
* **Developing and optimizing SVM (Support Vector Machine) model**
* **Developing and optimizing SVM (support vector machine)**
* **Developing and optimizing XGBoost (Xtream Gradient Boosting)**
* **Developing and optimizing LightGBM**

**5.1 Data Preprocessing and transformation**

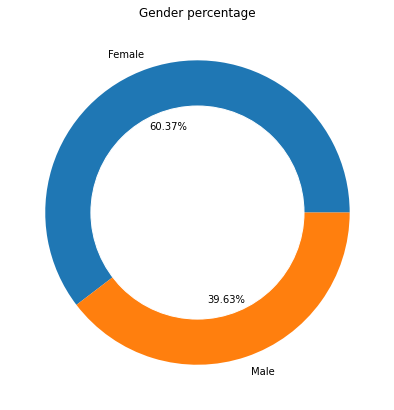
We can see that the problem of category imbalance is mainly solved from the following two perspectives: the first perspective is to balance the data by changing the number of samples. This method can also be divided into three aspects. First, oversampling method. Second, under-sampling to change the data distribution. Third, combination of oversampling and under sampling.

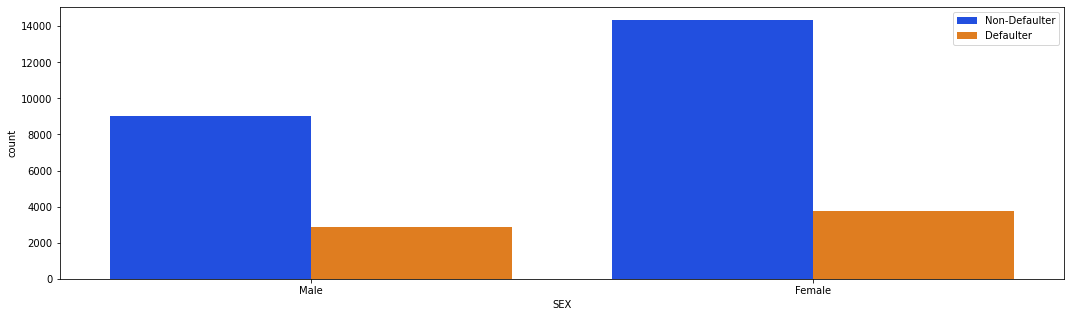
The second perspective is to improve the classifier algorithm to improve the prediction performance of the model and at the same time use relevant evaluation indicators to evaluate the prediction results.



The above pie chart shows that 77.86% are non-defaulters and 22.12% are defaulters. Thus, target variable is highly imbalanced.

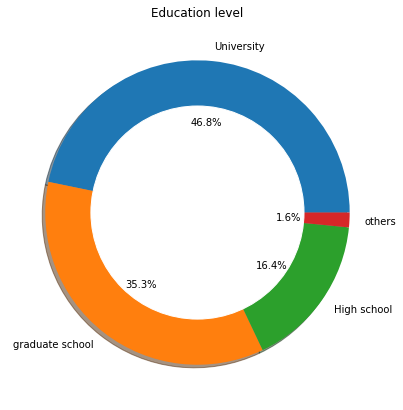
**Total number of defaulters based on gender**

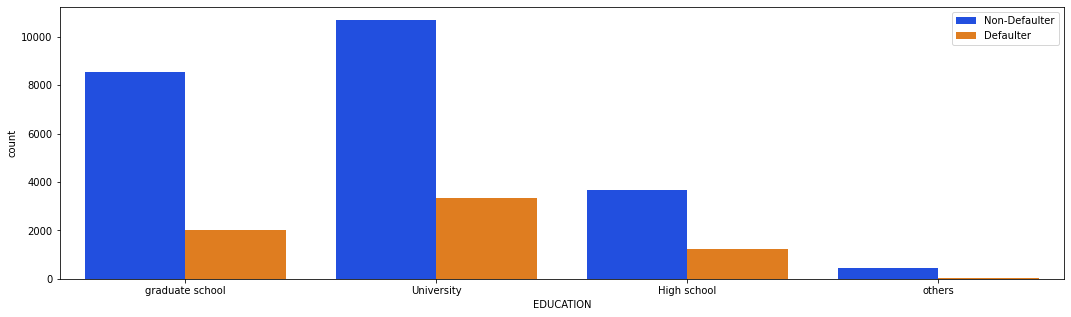
****

****

From the above plots, we can observe that the females are defaulters most of the time.

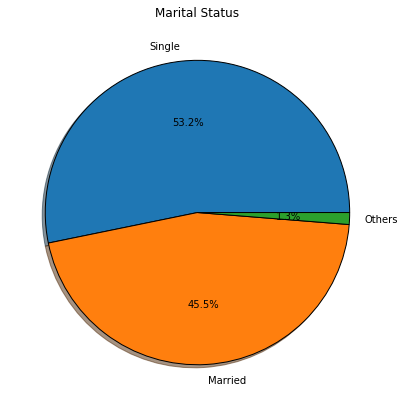
**Total number of defaulters based on educational qualification**

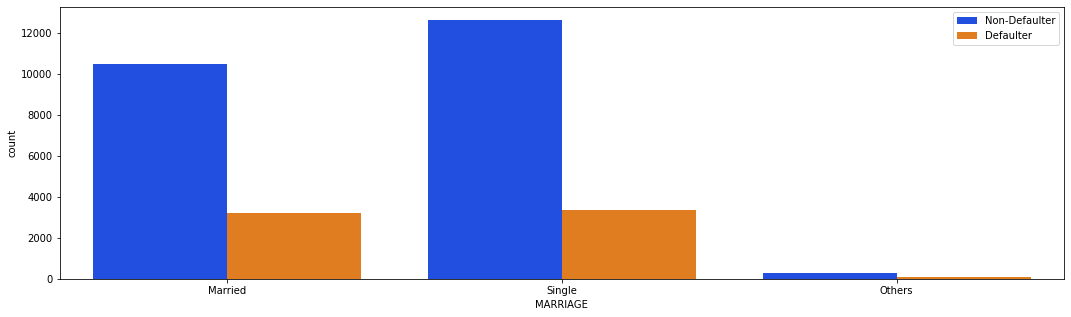




 Most of the defaulters are from university followed by graduate school, high school, and others. Variation of defaulters and non- defaulters are maximum in case of person has education level of university.

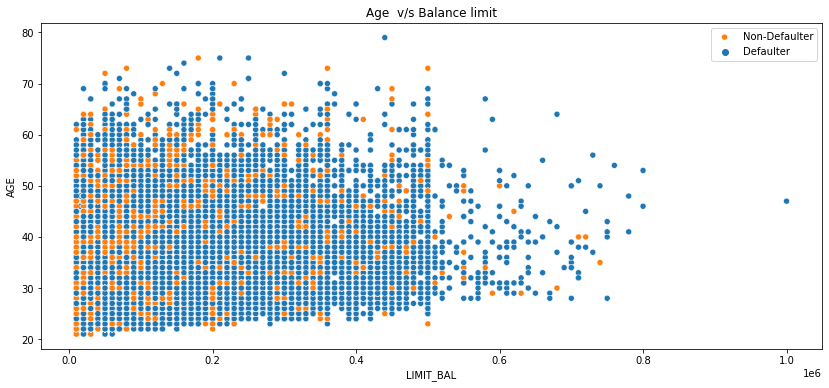
**Total number of defaulters based on Marital status**

****



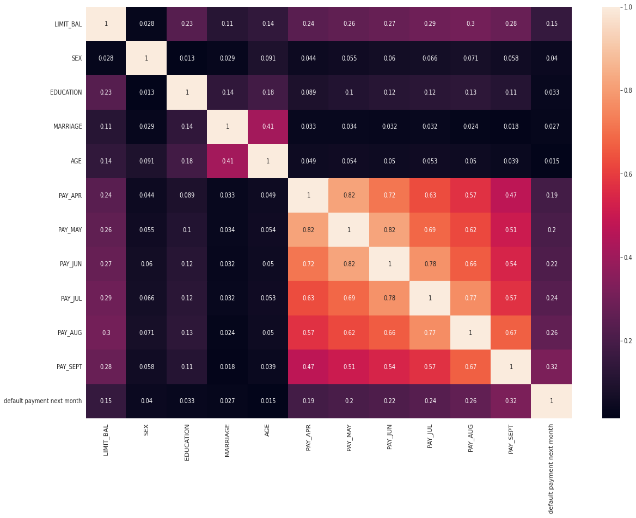
Most of the singles are non-defaulters followed by married person. total number of defaulters in case of married and single are almost same.

**Relation between age, balance limit and target variable**

****

This scatter plot is depicting that most of the credit card holder has balance limit less than 500000.

**Correlation analysis**

****

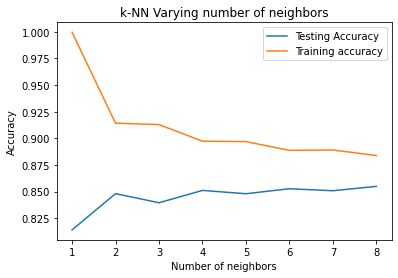
After plotting the correlation chart for all variables, we found that LIMIT\_BAL, SEX, EDUCATION, AGE and MARRIAGE are not highly correlated to each other. PAY\_APR','PAY\_MAY','PAY\_JUN','PAY\_JUL','PAY\_AUG','PAY\_SEPT are highly correlated with each other which means that people who can pay the bill on time will have high possibility to pay next bill on time. The probability of their late payment is mianly based on their previous behaviors instead of their characterics.

**Data set after the feature Engineering and dummy variable**

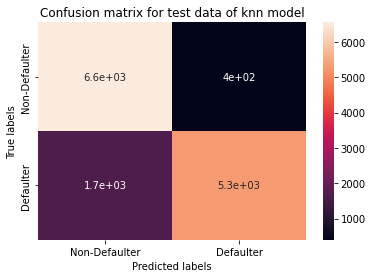
Feature engineering is the process that takes raw data and transforms it into features that can be used to create a predictive model. We applied Hot Encoding to produce binary integers of 0 and 1 to encode our categorical features.

**5.2 Developing and optimizing KNN model:**

K-NN algorithm stores all the available data and classifies a new data point based on the similarity which means that when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.



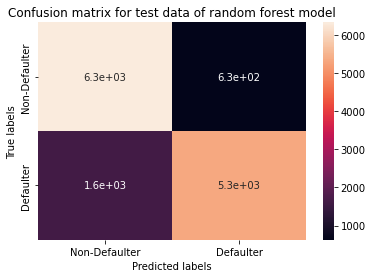
From the given plot we can take the best possible no of neighbors for our dataset. we can see here 8 is the best k value for further analysis.



From the above confusion matrix, the 1652 defaulters are predicted as non-defaulters and 398 non-defaulters are predicted as defaulters by the KNN model. The test accuracy is 0.85, F1 score is 0.83 and ROC score is 0.86.

**5.3 Developing and optimizing Random Forest model:**

Random Forest is a supervised learning algorithm, it creates a forest and makes it somehow random. The "forest” it builds, is an ensemble of Decision Trees.

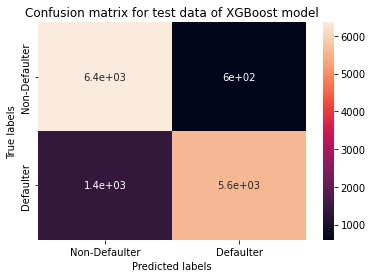
****

From the above confusion matrix, the 1614 defaulters are predicted as non-defaulters and 626 non-defaulters are predicted as defaulters by random forest model.

The test accuracy is 0.84, F1 score is 0.83 and ROC score is 0.84.

**5.4 Developing and optimizing XGBoost model:**

XGBoost classifier is a machine learning algorithm that is applied for structured and tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. It is an extreme gradient boost algorithm.

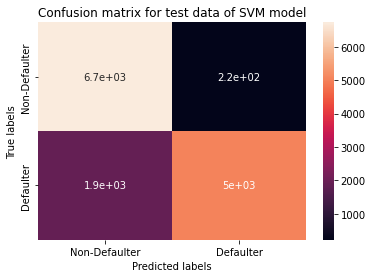


From the above confusion matrix, the 1380 defaulters are predicted as non-defaulters and 597 non-defaulters are predicted as defaulters by the xgboost model.

The test accuracy is 0.86, F1 score is 0.83 and ROC score is 0.87.

**5.5 Developing and optimizing SVM model:**

SVM algorithm is to create the best line or hyperplane that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.

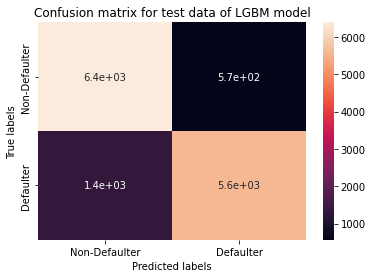
****

From the above confusion matrix, 1911 defaulters are predicted as non-defaulters and 224 non-defaulters are predicted as defaulters by the SVM model.

The test accuracy is 0.85, F1 score is 0.82 and ROC score is 0.87.

**5.6 Developing and optimizing LGBM model:**

**LightGBM** is a gradient boosting framework based on decision trees to increases the efficiency of the model and reduces memory usage.

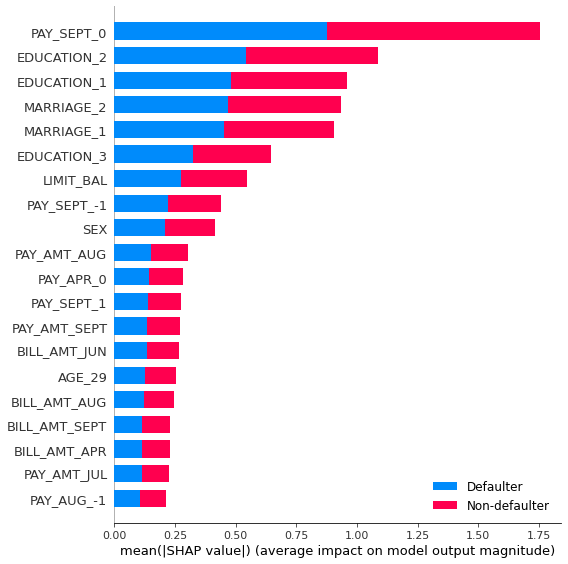
****

From the above confusion matrix, the 1356 defaulters are predicted as non-defaulters and 568 non-defaulters are predicted as defaulters by LGBM model.

The test accuracy is 0.86, F1 score is 0.85 and ROC score is 0.87.

**6. SHAP features**

The SHAP framework has proved to be an important advancement in the field of machine learning model interpretation. SHAP combines several existing methods to create an intuitive, theoretically sound approach to explain predictions for any model. SHAP values quantify the magnitude and direction (positive or negative) of a feature’s effect on a prediction.

****

In this plot, the impact of a feature on the target class is stacked to create the feature importance plot. From here we can see that PAY\_SEPT\_O has highest importance on the target 'Defaulter', while other features such as EDUCATION\_2, EDUCATION\_1 and subsequently all other features showed importance in decreasing order.

**7. Conclusion**

The most important takeaways from this project are:

* Most of the defaulters are university pass out.
* Most of the defaulter has the balance limit in the range 50000 to 200000.
* Credit cardholders who have no consumption or paid in full every month or delay 1 month, the number of no default is more than that of default. For those who use revolving credit, which means people who only pay the minimum amount every month, the non-default far exceeds the default. However, for those who delay the payment for more than one month, it turns out that the likelihood of default would then surpass the non-default, which also means the longer the payment delay, the higher risk for that person on default

We started with KNN Classifier and followed with Random Forest model, XGBoost model, SVM model and LGBM model, for which we obtained ROC score of 0.864, 0.845, 0.862, 0.868 and 0.866 respectively for the prediction of the test set target values. LGBM has highest test accuracy (0.861682), highest precision score (0.805032), highest F1 score (0.853379) and highest ROC score (0.866385) among all the models. So, overall LGBM is the winner among all the models.

**7. REFERENCES**

[1]L. Hu, Z. Peng, W. Xiang, and X. Rongze, “A new combination sampling method for imbalanced data,” in Proceedings of the 2013 Chinese Intelligent Automation Conference: Intelligent Information Processing, vol. 256, pp. 547–554, Yangzhou, China, 2013.View at: [Google Scholar](https://scholar.google.com/scholar_lookup?title=A%20new%20combination%20sampling%20method%20for%20imbalanced%20data&author=L.%20Hu&author=Z.%20Peng&author=W.%20Xiang&author=&author=X.%20Rongze)

[2]H. Han, W.-Y. Wang, and B.-H. Mao, “Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning,” Advances in Intelligent Computing, vol. 3644, no. 1, pp. 878–887, 2005.View at: [Google Scholar](https://scholar.google.com/scholar_lookup?title=Borderline-SMOTE:%20a%20new%20over-sampling%20method%20in%20imbalanced%20data%20sets%20learning&author=H.%20Han&author=W.-Y.%20Wang&author=&author=B.-H.%20Mao&publication_year=2005)

[3]S. Wang, W. Liu, and J. Wu, “Training deep neural networks on imbalanced data sets,” in Proceedings of the International Joint Conference on Neural Networks, pp. 4368–4374, IEEE, Vancouver, Canada, July 2016.View at: [Google Scholar](https://scholar.google.com/scholar_lookup?title=Training%20deep%20neural%20networks%20on%20imbalanced%20data%20sets&author=S.%20Wang&author=W.%20Liu&author=&author=J.%20Wu)

[4]J. Jiao, X. Zhang, F. Li, and Z. Niu, “Identically distributed multi-decision tree based on reinforcement learning and its application in imbalanced data sets,” Journal of Central South University (Science and Technology), vol. 50, no. 5, pp. 1112–1118, 2019.View at: [Google Scholar](https://scholar.google.com/scholar_lookup?title=Identically%20distributed%20multi-decision%20tree%20based%20on%20reinforcement%20learning%20and%20its%20application%20in%20imbalanced%20data%20sets&author=J.%20Jiao&author=X.%20Zhang&author=F.%20Li&author=&author=Z.%20Niu&publication_year=2019)

[5]D. Hong, L. Balzano, and J. A. Fessler, “Asymptotic performance of PCA for high-dimensional heteroscedastic data,” Journal of Multivariate Analysis, vol. 167, pp. 435–452, 2018.View at: [Google Scholar](https://scholar.google.com/scholar_lookup?title=Asymptotic%20performance%20of%20PCA%20for%20high-dimensional%20heteroscedastic%20data&author=D.%20Hong&author=L.%20Balzano&author=&author=J.%20A.%20Fessler&publication_year=2018)