CREDIT CARD DEFAULT PREDICTION

Akash Kagdelwar

Data science trainee,

AlmaBetter

Abstract:

Credit card default happens when you have become severely delinquent on your credit card payments. Defaults dose not happens in case of delay or missing few credit card bill payments. When anyone don't pay credit card bill(s) for an extended period of time their card may enter into default status. Default usually happens after six months in a row of not making at least the minimum payment due. Default is a serious credit card status that affects current credit standing and ability to get approved for other credit-based services.

From credit card lender perspective default in credit card payments lead huge losses. Bank and several financial institutions provide credit card to their customers. Several checks are there to issue a credit card to any customer such as credit score etc. So, to ensure profitability and reduce the risk of credit card default, prediction of credit card default plays very important role.

Thus, reducing risk by taking appropriate steps to counter defaults of credit card payments can be achieved by credit card default prediction.

Keywords: machine learning, credit card default, logistic regression, random forest, EDA.

1.Problem Statement

The main objective is aimed at predicting the case of customers default payments in Taiwan. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients.

2. Introduction

In order to get profitability and lower down the risk of credit card defaults in payments, credit card provider must do credit card default prediction. Credit card default depends on several variables. Understanding about all such variable and their relationship with credit card default is necessary.

This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature and used the following variables as explanatory variables:

- **ID:** Unique ID of each client
- **LIMIT_BAL:** Amount of the given credit (NT dollar).
  It includes both the individual consumer credit and his/her family (supplementary) credit.

- **Gender:** Gender of customer.
  (1 = male; 2 = female)
- **Education:** Education qualification of customers.
  (1 = graduate school; 2 = university; 3 = high school; 4 = others)
- **Marital Status:** Marital status of customer.
  (1 = married; 2 = single; 3 = others)
- **Age:** Age of customer in years.
- **History of Past Payment:**

We tracked the past monthly payment records from April to September, 2005.

PAY_1, PAY_2, PAY_3, PAY_4, PAY_5 and PAY_6 are repayment status in September, August, July, June, May and April 2005 respectively.

The measurement scale for the repayment status is:
-1 = pay duly;
1 = payment delay for one month;
2 = payment delay for two months; . . .;
8 = payment delay for eight months;
9 = payment delay for nine months and above.

- **Amount of Bill Statement**:
  (NT dollar)

BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5 and BILL_AMT6 are amount of bill statement in September, August, July, June, May and April 2005 respectively.

- **Amount of Previous Payment:**
  (NT dollar)

PAY_AMT1, PAY_AMT2, PAY_AMT2, PAY_AMT2, PAY_AMT2 and PAY_AMT2 are amount of previous payment in September, August, July, June, May and April 2005 respectively

- **Default Payment Next Month**: Default payment
  (1=yes, 0=no)

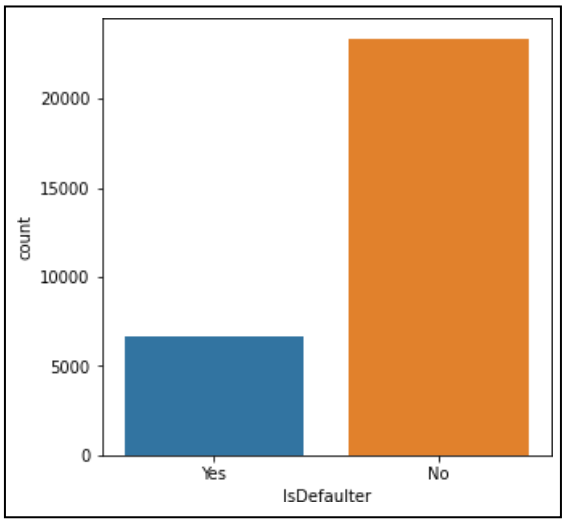
3. Steps involved:

- **<u>Null values Treatment</u>**
  Null values which might tend to disturb our accuracy hence we dropped them at the beginning of our project in order to get a better result. Our dataset dose not contains any null values.

- **<u>Duplicate Values</u>**
  Duplicate values dose not contribute anything to accuracy of results. Large duplicate may lead to slower computation and higher space requirement. Our dataset dose not contains any duplicate values.
- **<u>Data Preprocessing</u>**
  To be able to understand easily, name of some features replaced. Default payment next month to IsDefaulter. All features with PAY, BILL_AMT and PAY_AMT are replaced at ending with respective month name.

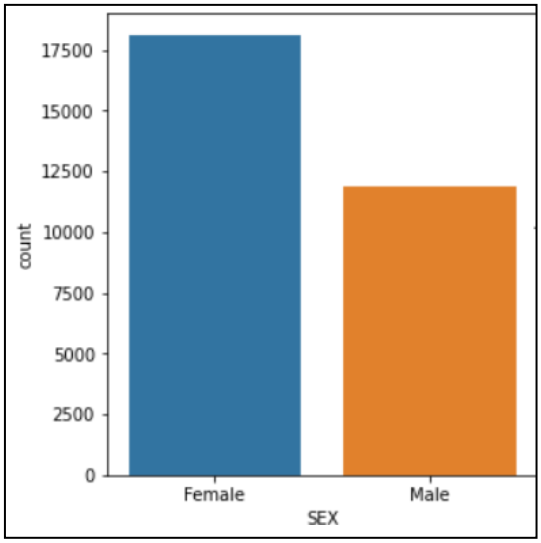
- **Exploratory Data Analysis**

IsDefaulter is our dependent variable. Weightage of each class in this feature can be seen from count plot.

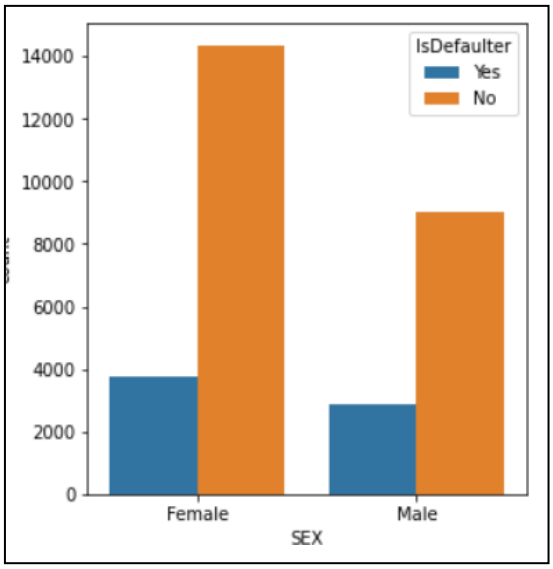


From the graph above, both the classes are not in proportion. Which means that dataset imbalanced. Data balancing is required.

Gender wise distribution can be seen from count plot of feature SEX.

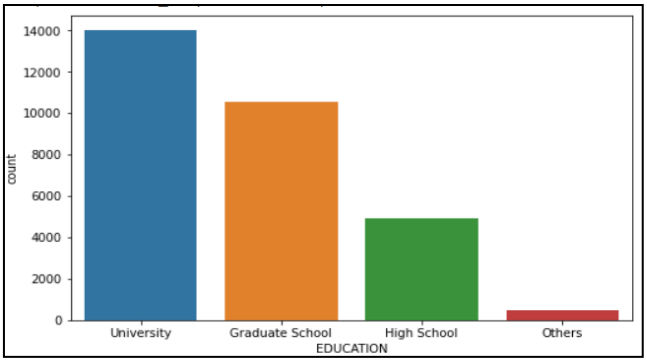


Female have larger credit cards than male.

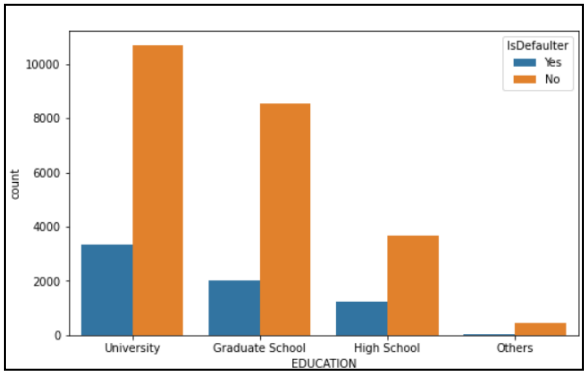


As the number female credit card holder is larger than male, their credit card defaults are also higher than male.

Following are the values in each category of Education feature.

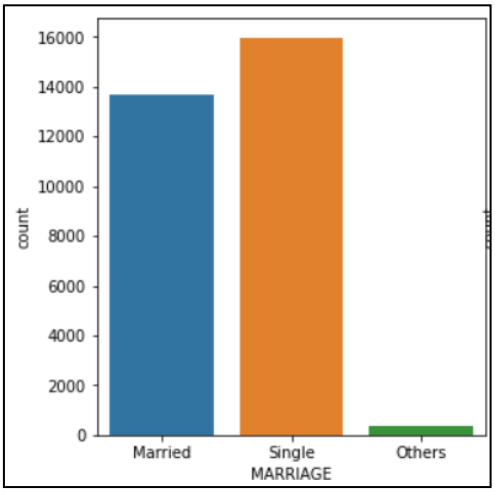


University has maximum credit card holder customers.
Let's look into education wise credit card default values.

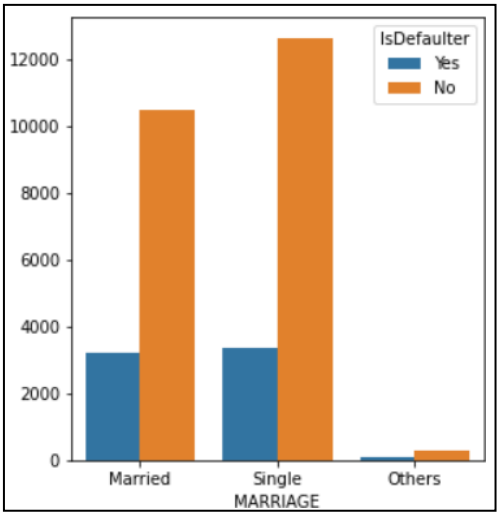


As the number university is higher credit card default are also higher in this case.

Categorical values in marital status.

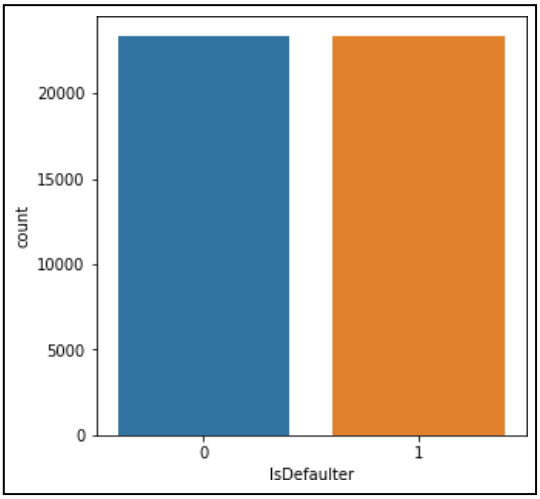


Singles have more credit cards.
Let's look into marital status wise credit default values.



Though the number of credit card holder is maximum in singles, but credit card defaults are almost same in case of single and married people.

- **Handling Class Imbalance:**
  As we have seen that data is imbalanced and we need to balance it. SMOTE (Synthetic Minority Oversampling Technique) is the technique to make data class balanced.



From the above graph, data class is balanced now.

- **Standardization of features**
  Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it.
  The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

- **Splitting Data:**

  Data splits into training dataset and testing dataset. Training dataset is for making algorithm learn and train model. And test dataset is for testing the performance of train model. Here 80% of data taken as training dataset and remaining 20% of dataset used for testing purpose.

- **Fitting Different Models:**

  For modelling we tried various algorithms like:

  - Logistic Regression
  - Decision Tree Classifier
  - Random Forest Classifier
  - Support Vector Machine
  - Gradient Boosting
  - XG Boosting

- **Cross Validation & Hyperparameter Tunning:**

  Cross Validation is a very useful technique for assessing the effectiveness of your model, particularly in cases where you need to mitigate overfitting.

  Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting.

  Cross validation and hyperparameter tunning performed on all above machine learning model.

- **Comparison of Models:**

  Comparison of all the above six model on the basis od several evaluation metrics. Most important accuracy and f1 score.

- **Combined ROC Curve:**

  Combined plot of ROC curve for each of the above model.

- **Feature Importance:**

  Feature importance to note down highly related features in prediction of credit card default.

4. Algorithms:

4.1 Logistic Regression:

Logistic regression is a machine learning algorithm for classification problem. In this algorithm, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function. It is most useful for understanding the influence of several independent variables on a single outcome variable.

Following are the evaluation metrics after fitting data into logistic regression model:

LOGISTIC REGRESSION					
Accuracy		Precision	Recall	F1	AUC
Train	Test				
0.828	0.831	0.795	0.857	0.825	0.833

Following are the evaluation metrics after fitting data into cross validated and hyperparameter tunned logistic regression model:

Tunning on Logistic Regression					
Accuracy		Precision	Recall	F1	AUC
Train	Test				
0.827	0.832	0.799	0.855	0.826	0.833

4.2 Decision Tree Classifier:

Given a data of attributes together with its classes, a decision tree produces a sequence of rules that can be used to classify the data. Decision Tree is simple to understand and visualize, requires little data preparation, and can handle both numerical and categorical data.

Following are the evaluation metrics after fitting data into decision tree classifier:

Decision Tree Classifier					
Accuracy		Precision	Recall	F1	AUC
Train	Test				
1	0.791	0.810	0.781	0.795	0.792

Following are the evaluation metrics after fitting data into cross validated and hyperparameter tunned decision tree classifier:

Tunning on Decision Tree Classifier					
Accuracy		Precision	Recall	F1	AUC
Train	Test				
0.837	0.824	0.779	0.857	0.816	0.827

4.3Random Forest Classifier:

Random Forest classifier is a meta-estimator that fits a number ofdecision trees on various sub-samplesof datasets and uses average toimprove the predictive accuracy ofthe model and controls over-fitting.The sub-sample size is always thesame as the original input sample sizebut the samples are drawn withreplacement.

Reduction in over-fitting and random forest classifier is more accurate than decision trees in most cases.

Following are the evaluation metrics after fitting data into random forest classifier:

Random Forest Classifier					
Accuracy		Precision	Recall	F1	AUC
Train	Test				
1	0.863	0.821	0.897	0.857	0.866

Following are the evaluation metrics after fitting data into cross validated and hyperparameter tunned random forest classifier:

Tunning on Random Forest Classifier					
Accuracy		Precision	Recall	F1	AUC
Train	Test				
0.844	0.833	0.794	0.860	0.826	0.835

4.4Support Vector Machine:

Support vector machine is a representation of the training data as points in space separated into categories by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

Effective in high dimensional spaces and uses a subset of training points in the decision function so it is also memory efficient.

Following are the evaluation metrics after fitting data into support vector classifier:

Support Vector Machine					
Accuracy		Precision	Recall	F1	AUC
Train	Test				
0.848	0.840	0.765	0.900	0.827	0.848

Following are the evaluation metrics after fitting data into cross validated and hyperparameter tunned support vector classifier:

Tunning on Support Vector Machine					
Accuracy		Precision	Recall	F1	AUC
Train	Test				
0.846	0.841	0.768	0.900	0.829	0.849

4.5 Gradient Boosting:

It is a technique of producing an additive predictive model by combining various weak predictors, typically Decision Trees. Due to this sequential connection, boosting algorithms are usually slow to learn, but also highly accurate. The final model aggregates the result of each step and thus a strong learner is achieved.

It is a generalized algorithm which works for any differentiable loss function.

Following are the evaluation metrics after fitting data into gradient boosting:

Gradient Boosting					
Accuracy		Precision	Recall	F1	AUC
Train	Test				
0.847	0.843	0.801	0.875	0.836	0.846

Following are the evaluation metrics after fitting data into cross validated and hyperparameter tunned gradient boosting:

Tunning on Gradient Boosting					
Accuracy		Precision	Recall	F1	AUC
Train	Test				
0.951	0.866	0.824	0.899	0.860	0.868

4.6 XG Boosting:

XG Boost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework.
It is a perfect combination of software and hardware optimization techniques to yield superior results using fewer computing resources in theshortest amount of time.

Following are the evaluation metrics after fitting data into XG boosting:

XG Boosting					
Accuracy		Precision	Recall	F1	AUC
Train	Test				
0.847	0.843	0.799	0.877	0.836	0.846

Following are the evaluation metrics after fitting data into cross validated and hyperparameter tunned XG boosting:

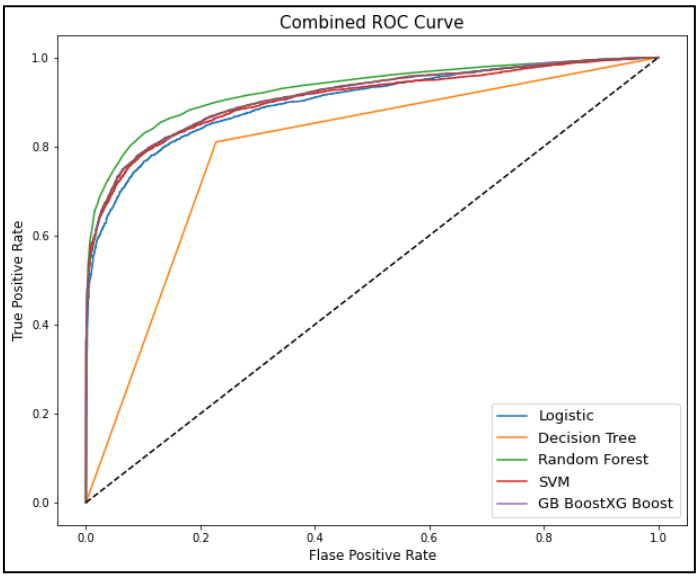
Tunning on XG Boosting					
Accuracy		Precision	Recall	F1	AUC
Train	Test				
0.995	0.871	0.831	0.904	0.866	0.874

5. Combined ROC Curve

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds.

- True Positive Rate (TPR)
- False Positive Rate (FPR)

An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives.

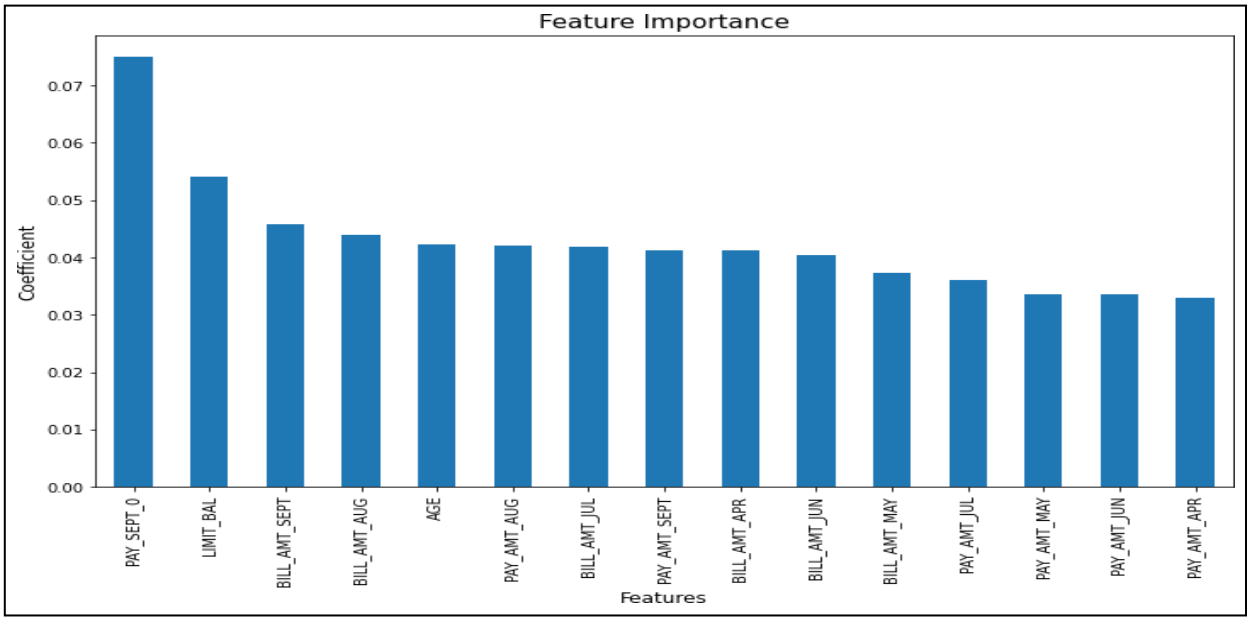


6. Feature Importance

Feature selection is the process of reducing the number of input variables when developing a predictive model.

It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model.

Feature selection techniques are often used in domains where there are many features and comparatively few samples



7. Comparison of Models

In machine learning, classification problem means training a model to specify which category an entry belongs to. There are so many classification algorithms in machine learning.

All the above machine learning model sorted descending by their AUC.

	Classifier	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	AUC
11	Optimal XG Boosting	0.995	0.871	0.831	0.904	0.866	0.874
2	Random Forest	0.999	0.867	0.832	0.895	0.862	0.869
10	Optimal Gradient Boosting	0.951	0.866	0.824	0.899	0.860	0.868
3	SVM	0.846	0.841	0.768	0.900	0.829	0.849
9	Optimal SVM	0.846	0.841	0.768	0.900	0.829	0.849
4	Gradient Boosting	0.845	0.845	0.801	0.878	0.838	0.848
5	XG Boosting	0.847	0.844	0.801	0.877	0.837	0.847
8	Optimal Random Forest	0.844	0.833	0.794	0.860	0.826	0.835
0	Logistic Regression	0.827	0.832	0.796	0.857	0.826	0.834
6	Optimal Logistic Regression	0.826	0.832	0.797	0.857	0.826	0.834
7	Optimal Decision Tree	0.841	0.825	0.779	0.858	0.817	0.828
1	Decision Tree	1.000	0.802	0.814	0.795	0.804	0.802

8. Conclusion:

1. From all baseline model, Random Forest classifier shows highest test accuracy and F1 score and AUC.

2. Baseline model of Random Forest and decision tree shows huge difference in train and test accuracy which shows overfitting.

3. After cross validation and hyperparameter tunning, XG Boost shows highest test accuracy score of 87% and AUC is 0.874.

4. Cross validation and hyperparameter tunning certainly reduce chances of overfitting and also increases performance of model.

References-

1. Stack Overflow
2. Analytics Vidhya