CREDIT CARD DEFAULT PREDICTION

TECHNICAL DOCUMENTATION

Vineeta Singh

Abstract:

This project 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**.** 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 more easily compared with cash transactions or bank statements. One downside that has been witnessed over the past few years of this increasing digital phenomenon is the rise of fraud on credit cards. Global fraud has increased by almost three times, from $9.84 billion to $32.39 billion in less than a decade (2011 to 2020).

Keywords: Exploratory Data Analysis, Train -Test split Classification, Machine learning model, (LR, , RF, KNN, SVM)

PROBLEM STATEMENT

After understanding the gravity of the fraud situation worldwide, particularly in the United States and some of the major European countries, the next automatic 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. let us discuss some of the challenges that we face while dealing with credit card fraud as below.

Data imbalance: The fraud and non-fraud data are generally much skewed. To give an example in the sample open-source dataset that we will be dealing with here, we have **6636** frauds out of a total of 30000 transactions. This is roughly only **21%** 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.

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. However, we will soon see that 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.

Real-Time Detection: For most of the fraud detection models in practice they have to work under very stringent timing conditions. We can take an example of a transaction-level fraud detection model. 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.

1.OBJECTIVE

Objective of our project is to predict which customer might default in upcoming months. Before goingany further let's have a quick look on definition of what actually meant by Credit Card Default. Creditcard default happens when somebody becomes severely delinquent (usually a young person who regularly performs illegal or immoral acts) on their credit card payments. Missing credit card payments once or twice does not count as a default. A payment default occurs when you fail to pay the Minimum Amount Due on the credit card for a few consecutive months.

Our goal is to look at the past data and learn from it the characteristics and behavior of past defaulters so as to make accurate predictions in the future. This will help the banks and credit card companies filter candidate applicants for credit cards more accurately so as to minimize the cases of default in the future.

2.INTRODUCTION

Credit risk has traditionally been the greatest risk among all the risks that the banking and credit card industry are facing, and it is usually the one requiring the most capital. This can beproven by industry business reports and statistical data. For example, “The Federal Reserve Bank of New York measures credit card delinquencies based on the percent of balances that are at least 90 days late. For the third quarter of 2019, that rate was about 8%, about the same level as in the previous quarter.” Thus, assessing, detecting and managing default risk is the key factor in 2 generating revenue and reducing loss for the banking and credit card industry.

Despite machine learning and big data have been adopted by the banking industry, the current applications are mainly focused on credit score predicting. The disadvantage of heavily relying on credit score is banks would miss valuable customers who come from countries that are traditionally under banked with no credit history or new immigrants who have repaying power but lack credit history. According to a literature review report on analysing credit risk using machine and deep learning models, “credit risk management problems researched have been around credit scoring; it would go a long way to research how machine learning can be applied to quantitative areas for better computations of credit risk exposure by predicting probabilities of default.”

The purpose of this project is to conduct quantitative analysis on credit card default risk by using interpretable machine learning models with accessible customer data, instead of credit score or credit history, with the goal of assisting and speeding up the human decision-making process.

DATA DESCRIPTION

The suggested system uses the original UCI repository report. There are 25 factors and 30,000 documents for customers. This dataset contains information on the credit card clients, regular charges, demographic factors, credit records, payment. 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 23 variables as explanatory variables:

- **ID**: ID of each client
- **LIMIT_BAL**: Amount of given credit in NT dollars (includes individual and family/supplementary credit
- **SEX**: Gender (1=male, 2=female)
- **EDUCATION**: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
- **MARRIAGE**: Marital status (1=married, 2=single, 3=others)
- **AGE**: Age in years
- **PAY_0**: Repayment status in September, 2005 (-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 ninemonths and above)
- **PAY_2**: Repayment status in August, 2005 (scale same as above)
- **PAY_3**: Repayment status in July, 2005 (scale same as above)
- **PAY_4**: Repayment status in June, 2005 (scale same as above)
- **PAY_5**: Repayment status in May, 2005 (scale same as above)
- **PAY_6**: Repayment status in April, 2005 (scale same as above)
- **BILL_AMT1**: Amount of bill statement in September, 2005 (NT dollar)
- **BILL_AMT2**: Amount of bill statement in August, 2005 (NT dollar)
- **BILL_AMT3**: Amount of bill statement in July, 2005 (NT dollar)
- **BILL_AMT4**: Amount of bill statement in June, 2005 (NT dollar)
- **BILL_AMT5**: Amount of bill statement in May, 2005 (NT dollar)
- **BILL_AMT6**: Amount of bill statement in April, 2005 (NT dollar)
- **PAY_AMT1**: Amount of previous payment in September, 2005 (NT dollar)
- **PAY_AMT2**: Amount of previous payment in August, 2005 (NT dollar)
- **PAY_AMT3**: Amount of previous payment in July, 2005 (NT dollar)
- **PAY_AMT4**: Amount of previous payment in June, 2005 (NT dollar)
- **PAY_AMT5**: Amount of previous payment in May, 2005 (NT dollar)
- **PAY_AMT6**: Amount of previous payment in April, 2005 (NT dollar)
- **default. payment. next. month**: Default payment (1=yes, 0=no)

I. EXPLORATORY DATA ANALYSIS

This dataset contains information on default payments, demographic factors, credit limit, history of payments, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. It includes 30,000 rows and 25 columns, and there is no credit score or credit history information.

Overall, the dataset is very clean, but there are several undocumented column values. As a result, most of the data wrangling effort was spent on searching information and interpreting the columns.

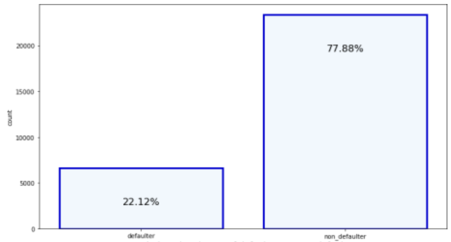
More details about the data cleaning can be found in this Colab Notebook. The purpose of exploratory data analysis is to identify the variables that impact payment default likelihood and the correlations between them. We use graphical and statistical data exploratory analysis tools to check every categorical variable. Each starts with a visualization and is followed by a statistical test to verify the findings.

1. CATEGORICAL VARIABLES

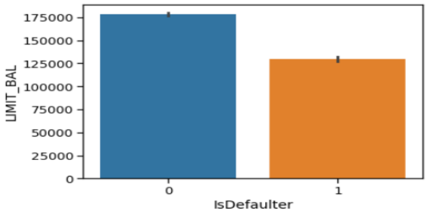
We have several categorical variables in our data frame. We will examine these values and the relationship these features have with our dependent variable.

- Looking at the sex column from the graph below, we can see that there are more female credit card holders than male. And while looking at the relationship with our dependent variable, we see that there are more female defaulters. Now this makes sense as most purchases are made by women today so they will have more credit cards and therefore are more likely to default.
- We also see that single people have more credit cards than couples and more no. of defaults too. As can be seen from the pie graph below, 53% credit card holders are single and 45% married.
- We can also see that highly educated people have way more number of credit cards than lower educated or uneducated people.

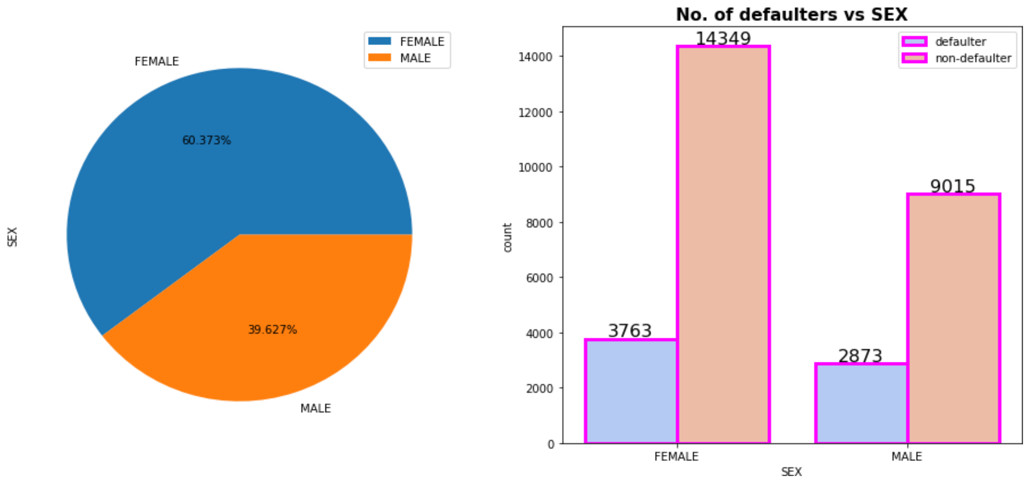
DEPENDENT VARIABLE - We can clearly see from the below graph that we have around 22 % of **i**nstances labeled as defaulter.



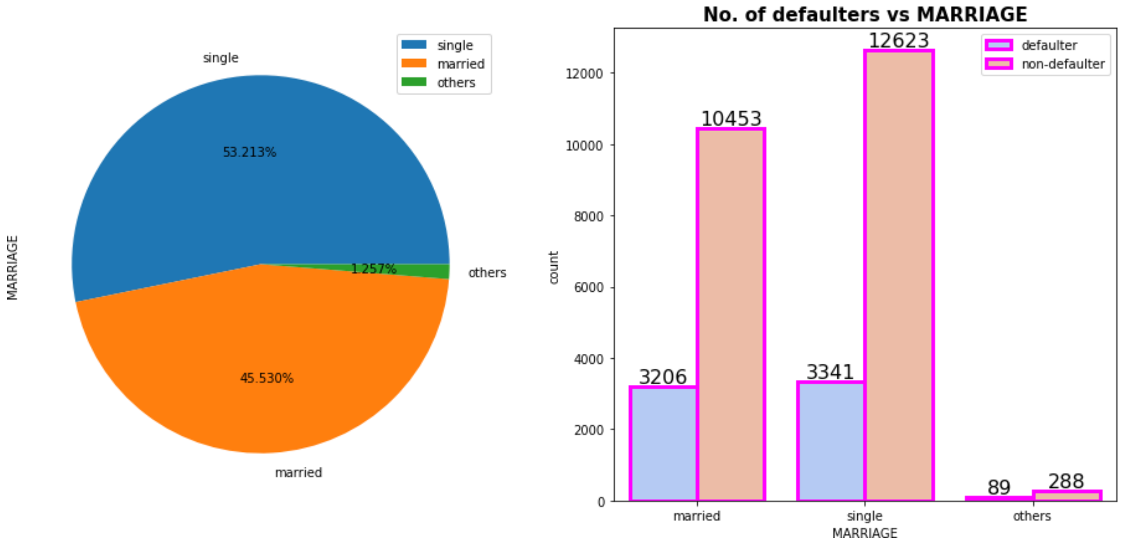
Graph denoting the no. of defaulters vs non-defaulters



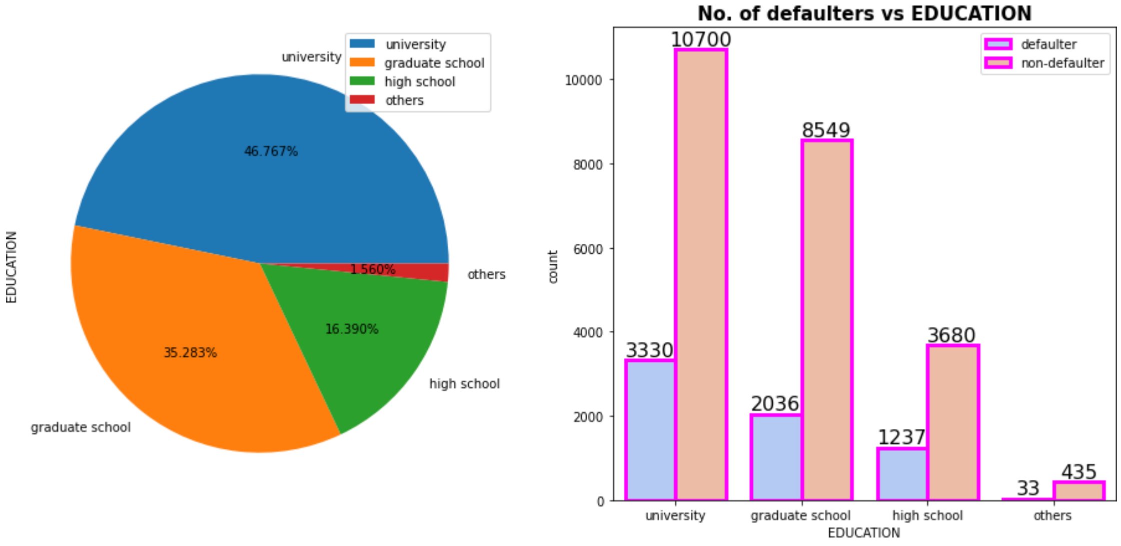
After different categorical feature SMOTE algorithms and feature engineering, we transform our data for reading purposes for a better understanding of the data set.



Dependent variable vs SEX

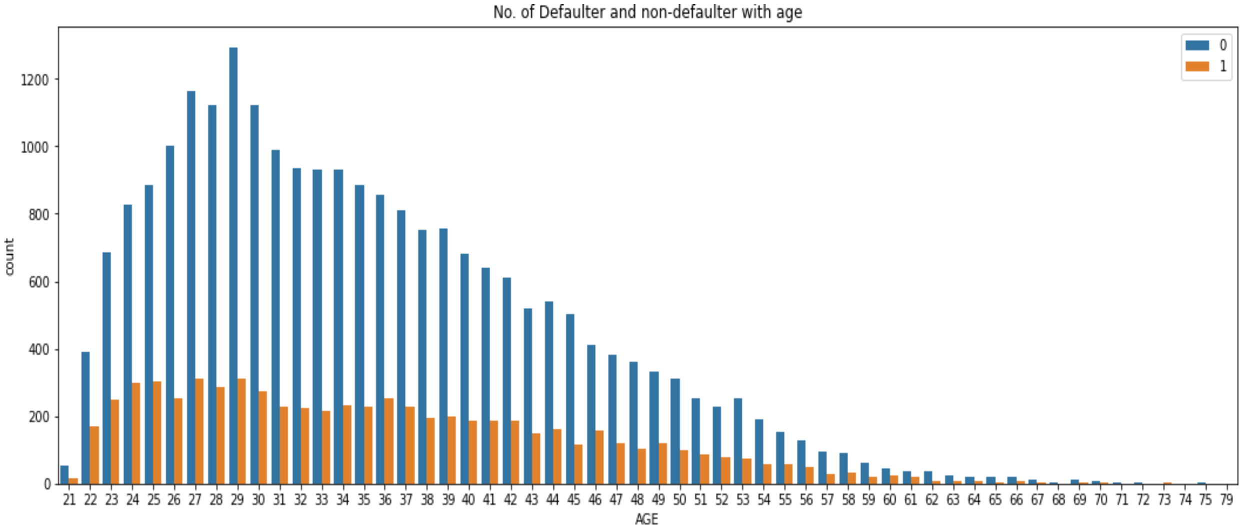


Dependent variable vs MARRIAGE



Dependent variable vs EDUCATION

2. NUMERICAL FEATURES

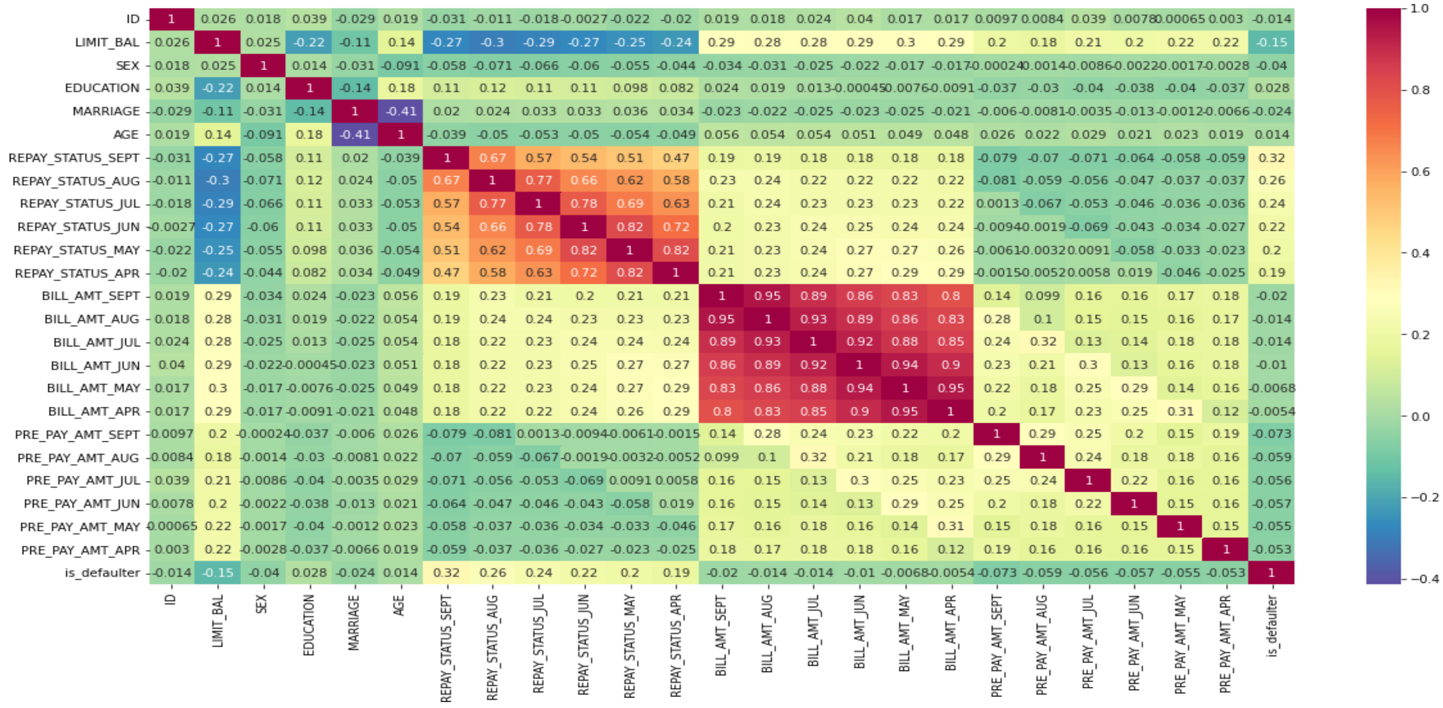


Dependent variable vs AGE

From the above graph we can clearly see that middle-aged customers have the highest no. of credit cards. The count plot shows that default probability increases for customers younger than 30 and older than 70. Customers aged between 30 and 50 have the lowest delayed payment rate, while younger groups (20-30) and older groups (50-70) all have higher delayed payment rates. This aligns with social reality that customers aged 30-50 typically have the strongest earning power. We also notice that there are not a lot of older credit card users. This is because of two reasons – low consumption and technological barrier.

3. CORRELATION ANALYSIS: -

Correlation heatmaps are a type of plot that visualize the strength of relationships between numerical variables. Correlation plots are used to understand which variables are related to each other and the strength of this relationship. The values in the cells indicate the strength of the relationship, with positive values indicating a positive relationship and negative values indicating a negative relationship.



	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	REPAY_STATUS_SEPT	REPAY_STATUS_AUG	REPAY_STATUS_JUL	REPAY_STATUS_JUN	REPAY_STATUS_MAY	REPAY_STATUS_APR	BILL_AMT_SEPT	BILL_AMT_AUG	BILL_AMT_JUL	BILL_AMT_JUN	BILL_AMT_MAY	BILL_AMT_APR	PRE_PAY_AMT_SEPT	PRE_PAY_AMT_AUG	PRE_PAY_AMT_JUL	PRE_PAY_AMT_JUN	PRE_PAY_AMT_MAY	PRE_PAY_AMT_APR	is_defaulter
ID	1	0.026	0.018	0.039	-0.029	0.019	-0.031	-0.011	-0.018	-0.0027	-0.022	-0.02	0.019	0.018	0.024	0.04	0.017	0.017	0.0097	0.0084	0.039	0.0078	0.00065	0.003	-0.014
LIMIT_BAL	0.026	1	0.025	-0.22	-0.11	0.14	-0.27	-0.3	-0.29	-0.27	-0.25	-0.24	0.29	0.28	0.28	0.29	0.3	0.29	0.2	0.18	0.21	0.2	0.22	0.22	-0.15
SEX	0.018	0.025	1	0.014	-0.031	-0.091	-0.058	-0.071	-0.066	-0.06	-0.055	-0.044	-0.034	-0.031	-0.025	-0.022	-0.017	-0.017	-0.00024	-0.0014	-0.0086	-0.0022	-0.0017	-0.0028	-0.04
EDUCATION	0.039	-0.22	0.014	1	-0.14	0.18	0.11	0.12	0.11	0.11	0.098	0.082	0.024	0.019	0.013	-0.00045	-0.0076	-0.0091	-0.037	-0.03	-0.04	-0.038	-0.04	-0.037	0.028
MARRIAGE	-0.029	-0.11	-0.031	-0.14	1	-0.41	0.02	0.024	0.033	0.033	0.036	0.034	-0.023	-0.022	-0.025	-0.023	-0.025	-0.021	-0.006	-0.0081	-0.0035	-0.013	-0.0012	-0.0066	-0.024
AGE	0.019	0.14	-0.091	0.18	-0.41	1	-0.039	-0.05	-0.053	-0.05	-0.054	-0.049	0.056	0.054	0.054	0.051	0.049	0.048	0.026	0.022	0.029	0.021	0.023	0.019	0.014
REPAY_STATUS_SEPT	-0.031	-0.27	-0.058	0.11	0.02	-0.039	1	0.67	0.57	0.54	0.51	0.47	0.19	0.19	0.18	0.18	0.18	0.18	-0.079	-0.07	-0.071	-0.064	-0.058	-0.059	0.32
REPAY_STATUS_AUG	-0.011	-0.3	-0.071	0.12	0.024	-0.05	0.67	1	0.77	0.66	0.62	0.58	0.23	0.24	0.22	0.22	0.22	0.22	-0.081	-0.059	-0.056	-0.047	-0.037	-0.037	0.26
REPAY_STATUS_JUL	-0.018	-0.29	-0.066	0.11	0.033	-0.053	0.57	0.77	1	0.78	0.69	0.63	0.21	0.24	0.23	0.23	0.23	0.22	0.0013	-0.067	-0.053	-0.046	-0.036	-0.036	0.24
REPAY_STATUS_JUN	-0.0027	-0.27	-0.06	0.11	0.033	-0.05	0.54	0.66	0.78	1	0.82	0.72	0.2	0.23	0.24	0.25	0.24	0.24	-0.0094	-0.0019	-0.069	-0.043	-0.034	-0.027	0.22
REPAY_STATUS_MAY	-0.022	-0.25	-0.055	0.098	0.036	-0.054	0.51	0.62	0.69	0.82	1	0.82	0.21	0.23	0.24	0.27	0.27	0.26	-0.0061	-0.0032	0.0091	-0.058	-0.033	-0.023	0.2
REPAY_STATUS_APR	-0.02	-0.24	-0.044	0.082	0.034	-0.049	0.47	0.58	0.63	0.72	0.82	1	0.21	0.23	0.24	0.27	0.29	0.29	-0.0015	-0.0052	0.0058	0.019	-0.046	-0.025	0.19
BILL_AMT_SEPT	0.019	0.29	-0.034	0.024	-0.023	0.056	0.19	0.23	0.21	0.2	0.21	0.21	1	0.95	0.89	0.86	0.83	0.8	0.14	0.099	0.16	0.16	0.17	0.18	-0.02
BILL_AMT_AUG	0.018	0.28	-0.031	0.019	-0.022	0.054	0.19	0.24	0.24	0.23	0.23	0.23	0.95	1	0.93	0.89	0.86	0.83	0.28	0.1	0.15	0.15	0.16	0.17	-0.014
BILL_AMT_JUL	0.024	0.28	-0.025	0.013	-0.025	0.054	0.18	0.22	0.23	0.24	0.24	0.24	0.89	0.93	1	0.92	0.88	0.85	0.24	0.32	0.13	0.14	0.18	0.18	-0.014
BILL_AMT_JUN	0.04	0.29	-0.022	-0.00045	-0.023	0.051	0.18	0.22	0.23	0.25	0.27	0.27	0.86	0.89	0.92	1	0.94	0.9	0.23	0.21	0.3	0.13	0.16	0.18	-0.01
BILL_AMT_MAY	0.017	0.3	-0.017	-0.0076	-0.025	0.049	0.18	0.22	0.23	0.24	0.27	0.29	0.83	0.86	0.88	0.94	1	0.95	0.22	0.18	0.25	0.29	0.14	0.16	-0.0068
BILL_AMT_APR	0.017	0.29	-0.017	-0.0091	-0.021	0.048	0.18	0.22	0.22	0.24	0.26	0.29	0.8	0.83	0.85	0.9	0.95	1	0.2	0.17	0.23	0.25	0.31	0.12	-0.0054
PRE_PAY_AMT_SEPT	0.0097	0.2	-0.00024	-0.037	-0.006	0.026	-0.079	-0.081	0.0013	-0.0094	-0.0061	-0.0015	0.14	0.28	0.24	0.23	0.22	0.2	1	0.29	0.25	0.2	0.15	0.19	-0.073
PRE_PAY_AMT_AUG	0.0084	0.18	-0.0014	-0.03	-0.0081	0.022	-0.07	-0.059	-0.067	-0.0019	-0.0032	-0.0052	0.099	0.1	0.32	0.21	0.18	0.17	0.29	1	0.24	0.18	0.18	0.16	-0.059
PRE_PAY_AMT_JUL	0.039	0.21	-0.0086	-0.04	-0.0035	0.029	-0.071	-0.056	-0.053	-0.069	0.0091	0.0058	0.16	0.15	0.13	0.3	0.25	0.23	0.25	0.24	1	0.22	0.16	0.16	-0.056
PRE_PAY_AMT_JUN	0.0078	0.2	-0.0022	-0.038	-0.013	0.021	-0.064	-0.047	-0.046	-0.043	-0.058	0.019	0.16	0.15	0.14	0.13	0.29	0.25	0.2	0.18	0.22	1	0.15	0.16	-0.057
PRE_PAY_AMT_MAY	0.00065	0.22	-0.0017	-0.04	-0.0012	0.023	-0.058	-0.037	-0.036	-0.034	-0.033	-0.046	0.17	0.16	0.18	0.16	0.14	0.31	0.15	0.18	0.16	0.15	1	0.15	-0.055
PRE_PAY_AMT_APR	0.003	0.22	-0.0028	-0.037	-0.0066	0.019	-0.059	-0.037	-0.036	-0.027	-0.023	-0.025	0.18	0.17	0.18	0.18	0.16	0.12	0.19	0.16	0.16	0.16	0.15	1	-0.053
is_defaulter	-0.014	-0.15	-0.04	0.028	-0.024	0.014	0.32	0.26	0.24	0.22	0.2	0.19	-0.02	-0.014	-0.014	-0.01	-0.0068	-0.0054	-0.073	-0.059	-0.056	-0.057	-0.055	-0.053	1

Multicollinearity

We can clearly see from the above heatmap that there are numerous columns where there is high correlation for example in bill amount columns (darker shade in heatmap). This can be a problem for some of our models.

The variance inflation factor (VIF) identifies correlation between independent variables and the strength of that correlation. Using Variance Inflation Factor- VIF- we can determine if two independent variables are collinear with each other. Multicollinearity is when there's correlation between predictors (i.e., independent variables) in a model; it's presence can adversely affect your results.

4. DATA PREPARATION

Data preparation includes feature engineering and feature selection. In feature engineering we converted categorical features such as Marriage, Education and Sex and into numerical data.

Feature selection is a way of selecting the subset of the most relevant features from the original features set by removing the redundant, irrelevant, or noisy features.

- One hot encoding is a process by which categorical variables are converted into a numerical variable that could be provided to ML algorithms to do a better job in prediction.
- Here we perform one hot encoding on several categorical features.
- Label encoding can be done for features having few categories.
- get dummies converts categorical data into dummy or indicator variables.

MODEL IMPLEMENTATION

Classification is a technique for determining which class the dependent belongs to based on one or more independent variables.

For modeling we tried various classification models such as-

1) Logistic Regression

2) K-Neighbors Classifier

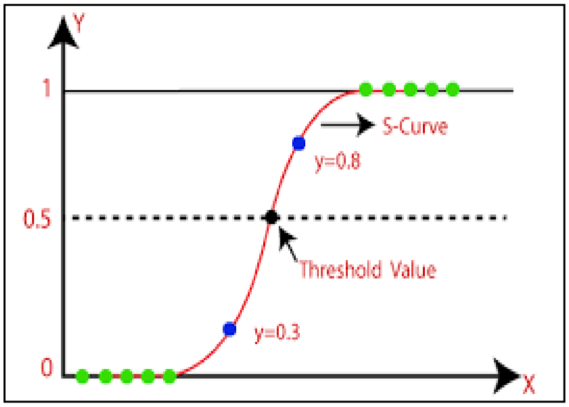
3) Support Vector Classifier

4) Random Forest Classifier

1. ALGORITHMS-

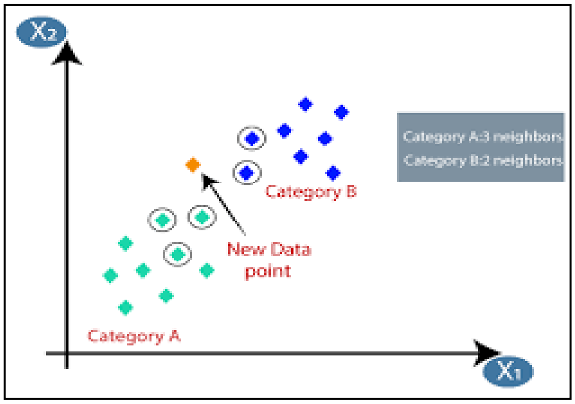
1. LOGISTIC REGRESSION:

Logistic regression is kind of like linear regression, but is used when the dependent variable is not categorical. It's called regression but it actually performs classification based on the regression equation and it classifies the dependent variable into classes. Logistic Regression is used when the dependent variable(target) is categorical.



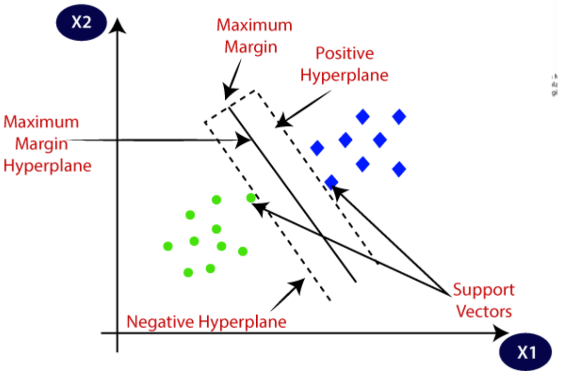
2. K-NEIGHBORS CLASSIFIER:

The k-nearest neighbors algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. KNN is a non- parametric algorithm. It works by finding k similar datapoints that are close in values and take mean/mode of their labels.



3. SUPPORT VECTOR CLASSIFIER:

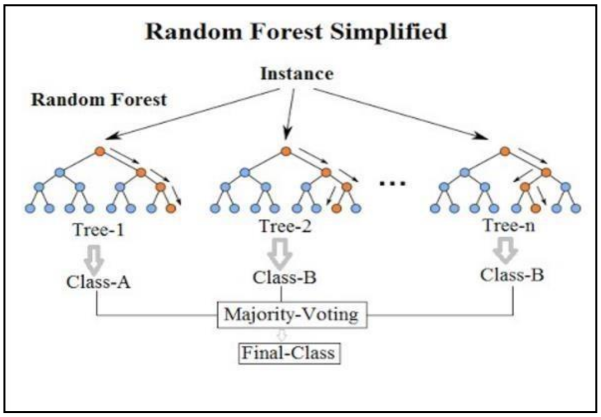
Support vector is used for both regression and classification. It is based on the concept of decision planes that define decision boundaries. A decision plane (hyperplane) is one that separates between a set of objects having different class memberships. It performs classification by finding the hyperplane that maximizes the margin between two classes with the help of Support Vectors.



4. RANDOM FOREST CLASSIFIER:

Random forest classifier is an ensemble algorithm based on bagging i.e bootstrap aggregation. Ensemble methods combine more than one algorithm of the same or differentkind for classifying objects.

Random forest, like its name implies, consists of a large number of individual decision treesthat operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes become our model’s prediction.



2. MODEL PERFORMANCE-

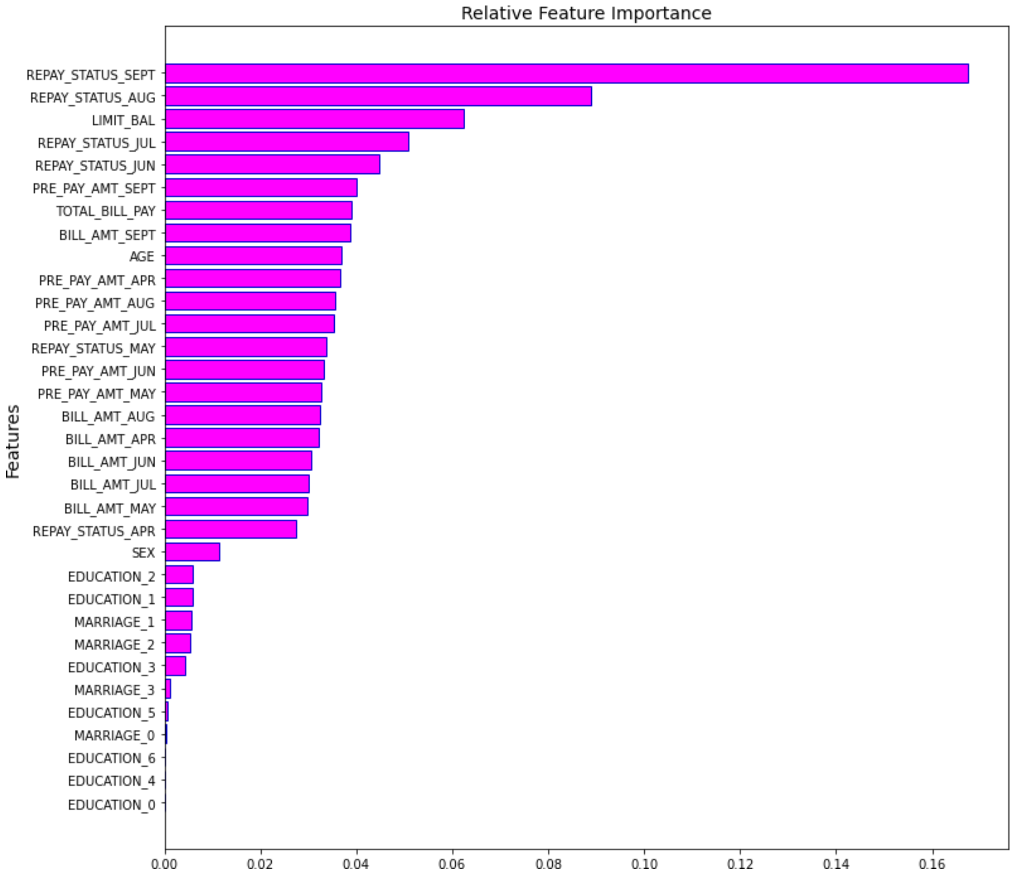
Evaluation Metrics:

There are different metrics for supervised algorithms (classification and regression) and unsupervised algorithms. Let's start exploring various Evaluation metrics we used.

- **Accuracy** of a classifier is calculated as the ratio of the total number of correctly predicted samples by the total number of samples. Accuracy metric should not be used when the data set is imbalanced.
- **Precision** is a good metric to use when the costs of false positives (FP) is high.
- Precision = TP / (TP + FP)
- **Recall** is a good metric to use when the cost associated with false negatives (FN) is high.
- Recall = TP / (TP + FN)
- **F1 score** is a weighted average of precision and recall. Thus, it considers FP and FN. This metric is very useful when we have uneven class distribution, as it seeks a balance between precision and recall.
- F1 score = 2 (precision recall) / (precision + recall)
- **Confusion matrix** is an N dimensional square matrix, where N represents the total number of target classes or categories. Confusion matrix can be used toevaluate a classifier whenever the data set is imbalanced.
- **AUC-ROC Curve** is a performance metric that is used to measure the performance for theclassification model at different threshold values. ROC is Receiver Operating Characteristic Curve and AUC is Area Under Curve. The higher the value of AUC (Area under the curve), the better is our classifier in predicting the classes.

FEATURE IMPORTANCE:

Feature importance refers to techniques that assign a score to input features based on how useful they are at predicting a target variable. Feature importance scores play an important role in a predictive modeling project, including providing insight into the data, insight into the model, and the basis for dimensionality reduction and feature selection that can improve the efficiency and effectiveness of a predictive model on the problem.



We can see in the above plot, our features are ranked top to bottom in descending order of importance.

CONCLUSIONS DRAWN

After conducting this thorough exercise, we found that :

* Most of the credit card users are Female and have higher number of defaults.
* Most of the credit card users are highly educated.
* single users have more no. of credit cards.
* The number of credit card users goes down with increase in age as old people have less
* consumption and may not be able to use credit cards and their purchases are usually made by younger family members.

* Using a Logistic Regression classifier, we can predict an accuracy of 67.7% and ROC_AUC score of 0.663

* Using Random Forest Classifier, we can predict an accuracy of around 87.6% and ROC_AUC score of 0.837

* Using K-Neighbor Classifier, we can predict an accuracy of 85.69% and ROC_AUC score of 0.858
* Using Support Vector Machine Classifier, we can predict an accuracy of 76.66% and ROC_AUC score of 0.723

* Random Forest Classifier performs best among all models.
* Logistic Regression is not giving good precision score

Our best models are Random Forest, K-Neighbor Classifier and Support Vector Machine that score really well on Precision, Recall, ROC_AUC and F1 score with K NEAREST NEIGHBOR CLASSIFIER & RANDOM FOREST being the best performers as they score the best on nearly every metric out there.

References:

1] Z. Feng and M. Feng, "Research on credit card scoring model based on AHP," *Finance Theory and Practice*, vol. 1, pp. 74–77, 2016.View at: Google Scholar

[2] R. Mei, Y. Xu, and G. Wang, "Study on analysis and influence factors of credit card default prediction model," *Statistics and Applications*, vol. 5, no. 3, pp. 263–275, 2016.Viewat: Publishe site | Google Scholar

[3]M. Zan, G. Yanrong, and F. Guanlong, "Credit card fraud classification based on GAN-AdaBoost-DT imbalance classification algorithm," *Journal of Computer Applications*, vol. 39, no. 2, pp. 314–318, 2019.View at: Google Scholar

[4]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

[5]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