**[Introduction](https://github.com/Kalpesh-Rajput/Classification_Credit_Card_Default_Prediction" \l "introduction)**

Default credit cards are an important issue that brings negative consequences to both sides, i.e, banks and customers. If a customer does not pay his obligations, banks lose money, the customer will lose credibility in future payments, collection calls start to be made and in last resort, the case may go into court. In order to avoid all of that trouble, effective methods that are able to predict the default of credit cards are needed. Therefore, default credit card prediction is an important, challenging, and useful task that should be addressed.

In recent years, the credit card issuers in Taiwan faced the cash and credit card debt crisis and the delinquency is expected to peak in the third quarter of 2006 (Chou,2006). In order to increase market share, card-issuing banks in Taiwan over-issued cash and credit cards to unqualified applicants. At the same time, most cardholders, irrespective of their repayment ability, overused credit card for consumption and accumulated heavy credit and cash–card debts. The crisis caused a blow to consumer finance confidence and it is a big challenge for both banks and cardholders.

[**Problem Statement**](https://github.com/Kalpesh-Rajput/Classification_Credit_Card_Default_Prediction#problem-statement)

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. We can use the K-S chart to evaluate which customers will default on their credit card payments

Data Description Attribute Information: 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:

* X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
* X2: Gender (1 = male; 2 = female).
* X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
* X4: Marital status (1 = married; 2 = single; 3 = others).
* X5: Age (year).
* X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. 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.
* X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.
* X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

[**Steps involved:**](https://github.com/Kalpesh-Rajput/Classification_Credit_Card_Default_Prediction#steps-involved)

[**Exploratory Data Analysis**](https://github.com/Kalpesh-Rajput/Classification_Credit_Card_Default_Prediction#exploratory-data-analysis)

After loading the dataset we compared our target variable that is “Default payment next month” with other independent variables. This process helped us figure out various aspects and relationships among the dependent and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the dependent variable.

[**Null values Treatment**](https://github.com/Kalpesh-Rajput/Classification_Credit_Card_Default_Prediction#null-values-treatment)

Our data set didn’t have any null values to be treated. Outliers treatment Isolation Forest Algorithm: Isolation forest is a tree-based algorithm that is very effective for both outlier and novelty detection in high-dimensional data. Encoding of categorical columns We used One Hot Encoding(converting to dummy variables) to produce binary integers of 0 and 1 to encode our categorical features because categorical features that are in string format cannot be understood by the machine and needs to be converted to the numerical format.

[**Standardization of features**](https://github.com/Kalpesh-Rajput/Classification_Credit_Card_Default_Prediction#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. Fitting different models For modeling, we tried various classification algorithms like:

* Logistic Regression
* SVM
* Decision Tree
* Random Forest Classification
* XGBoost Classification
* CatBoost Classification
* LightGBM Classification

[**Tuning the hyperparameters for better recall**](https://github.com/Kalpesh-Rajput/Classification_Credit_Card_Default_Prediction#tuning-the-hyperparameters-for-better-recall)

Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting in the case of tree-based models like Random Forest Classifier and XGBoost classifier.

[**Features Explainability**](https://github.com/Kalpesh-Rajput/Classification_Credit_Card_Default_Prediction#features-explainability)

We have applied SHAP on the XGBoost and CatBoost model to determine the features that were most important while predicting an instance And also build a feature importance graph to find out which features were important and which were redundant in a model

[**performance matrix**](https://github.com/Kalpesh-Rajput/Classification_Credit_Card_Default_Prediction#performance-matrix)

[**Conclusion**](https://github.com/Kalpesh-Rajput/Classification_Credit_Card_Default_Prediction#conclusion)

From the project we can conclude that the default rate is higher for males, increases as the education increases, and also increases as the age of a person increases. i.e clients whose age over 60 was higher than mid-age and young people. In all of these models, our recall revolves in the range of 76 to 84%.with the best fit model as random forest