**Credit card payment default analysis**

**Project Description:**

**Context**

This project is about predicting the Credit Card Bill Payment Defaulters using Logistic Regression, K-Nearest Neighbors, Random Forest Models using the “Default of Credit Card Clients” Dataset from UCI Machine Learning Repository. This dataset contains 30,000 records of data on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

**Problem Statement:**

In general, Banks issue credit cards based on a person's credit history and score. A banking institution faces issues related to distinguishing between borrowers that are likely to repay their credit card debt and borrowers that are likely to default on their debt. A credit card holder “defaults” on their debt when they make purchases with their credit card that they fail to repay or fail to pay before a given window (usually 30-45 days). They must identify high-risk credit card default users effectively as there is no set mechanism for identifying defaulters because any client can default at any time. Although banks make money by selling the assets of defaulters in extreme circumstances, this usually comes at high litigation costs and takes a long time. As a result, banks must lend responsibly to reduce defaults.

**Data Sources:**

**Citations:** Lichman, M. (2013). UCI Machine Learning Repository [ <https://archive.ics.uci.edu/ml/index.php> ]. Irvine, CA: University of California, School of Information and Computer Science.

**Link to dataset**: <https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>

Our goal is to perform exploratory data analysis and machine learning models for predicting the default of credit card clients using python.

**Data Preparation:**

This dataset contains 30,000 records of data on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005, and twenty-four input variables explaining almost every aspect of the default.

**Identifying the input variables and output variables**:

Our target variable (Output variable) is Default Payment Next Month, the dependent variable for prediction. We have twenty-four features (Input Variables) like sex, education, age, pay the amount, bill amount, Limit balance, etc.

**Data description:**

Graphical user interface, application, table

Description automatically generated

# **Variable description:**

1. ID: ID of each client
2. LIMIT\_BAL: Amount of given credit in New Taiwan dollars (includes individual and family/supplementary credit
3. SEX: Gender (1=male, 2=female)
4. EDUCATION: (1=graduate school, 2=university, 3=high school, 0=others, 4=others, 5=unknown, 6=unknown)
5. MARRIAGE: Marital status (1 = married; 2 = single; 3 = divorce; 0=others)
6. AGE: Age in years
7. PAY\_0: Repayment status in September 2005 (-2: No consumption; -1: Paid in full; 0: The use of revolving credit; 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.)
8. PAY\_2: Repayment status in August 2005 (scale same as above)
9. PAY\_3: Repayment status in July 2005 (scale same as above)
10. PAY\_4: Repayment status in June 2005 (scale same as above)
11. PAY\_5: Repayment status in May 2005 (scale same as above)
12. PAY\_6: Repayment status in April 2005 (scale same as above)
13. BILL\_AMT1: Amount of bill statement in September 2005 (NT dollar)
14. BILL\_AMT2: Amount of bill statement in August 2005 (NT dollar)
15. BILL\_AMT3: Amount of bill statement in July 2005 (NT dollar)
16. BILL\_AMT4: Amount of bill statement in June 2005 (NT dollar)
17. BILL\_AMT5: Amount of bill statement in May 2005 (NT dollar)
18. BILL\_AMT6: Amount of bill statement in April 2005 (NT dollar)
19. PAY\_AMT1: Amount of previous payment in September 2005 (NT dollar)
20. PAY\_AMT2: Amount of previous payment in August 2005 (NT dollar)
21. PAY\_AMT3: Amount of previous payment in July 2005 (NT dollar)
22. PAY\_AMT4: Amount of previous payment in June 2005 (NT dollar)
23. PAY\_AMT5: Amount of previous payment in May 2005 (NT dollar)
24. PAY\_AMT6: Amount of previous payment in April 2005 (NT dollar)
25. default.payment.next.month: Default payment (1=yes, 0=no)

**Data Cleaning & Validation:**

* Checking for Null/NAN values.
* Dropped the ID column (1st column) as it is irrelevant to the target variable
* Renaming the columns
  + Pay\_0 to Pay\_1, as the dataset has no PAY\_1 between PAY\_0 and PAY\_2
  + DEFAULT\_PAYMENT\_NEXT\_MONTH to DEFAULT\_PAYMENT
* From the dataset, values for Education column are 1 = graduate school; 2 = university; 3 = high school; 0, 4, 5, 6 = others. So, changing 0,5,6 to 4 in order to bring it under one category.

Graphical user interface, table

Description automatically generated

**Visualization:**

We plotted some preliminary visualization charts and graphs and explored our data to get an idea of its overall shape and the distribution of various demographic categories.

**Payment defaulters by Ratio:**

* The pie chart gives the information of the number of clients defaulting on the payments.
* We can infer from the chart that 22.12% (6636 clients) of the total clients are defaulting on their next month’s payment.

Chart, pie chart

Description automatically generated

**Correlation Matrix:**

* There is no correlation for most of the predictor variables except BILL\_AMT1 to BILL\_AMT6, PAY\_1 to PAY\_6.
* Correlation decreases with the distance between months. The lowest correlations are between Sept & April.

Chart, treemap chart

Description automatically generated

**Clients by Age:**

* We plotted the bar graph between the age of clients and the number of clients.
* Analyzing the client data according to clients’ age, we got to know that the number of defaulters is more in the age group of 20 to 40. Also, most of the clients are from the same age group.

Chart, bar chart

Description automatically generated

**Education level of Clients:**

Icon

Description automatically generated

From dataset: 1=graduate school, 2=university, 3=high school, 4=others; we can see from the above graph that most clients have an education of university and graduate school.

**Clients and their credit limit:**

* Analyzing the defaulter's data according to their credit limit, we learned that the number of defaulters is decreasing with an increase in their credit limit.
* The below trend is because, in general, the credit limit balance will be more for the clients with good credit scores and no previous default history.

Chart, bar chart

Description automatically generated

**Previous Bills and their Payments of 6 months:**

* This scatter plot gives the relationship between the payment made by the clients against their actual bill amount of each month from September 2005 to April 2005.

A picture containing text, computer, computer, screenshot

Description automatically generated

**Repayment Status:**

* The box plot gives the range of the client data according to their repayment status, i.e. (-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)

Chart

Description automatically generated

**Data Partition:**

We divided the dataset into training and test data in an 80%: 20% ratio.

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Description** | **Records** |
| Training Data | Consist of 80% of total records and features that would be used in overall training and modeling. | 24000 |
| Testing Data | Consist of 20% of total records and features that would be used during the final prediction of the Credit default. | 6000 |

**Modeling:**

We have implemented the below models:

* Logistic Regression
* Classification using K-Nearest Neighbors
* Random Forest classification

## Logistic Regression:

## Logistic regression model is used to predict a binary outcome, such as True or False, based on prior observations of a data set. A logistic regression model predicts a dependent variable by analyzing the relationship between existing independent variables.

We used Confusion Matrix to evaluate the accuracy of test data (20% of the dataset, which is 6000 records).

CM Results: True Negatives:4680, False Positives:1315, False Negatives:3, True Positives: 2

The Accuracy of the testing data using the Logistic regression model is 78.03%.

Graphical user interface, text

Description automatically generated

**Classification using K-Nearest Neighbors:**

K-Nearest Neighbors is often used as a benchmark for more complex classifiers. We have used 23 independent features for KNN implementation. The Accuracy of the testing data is 77.66%.

We have implemented KNN with different optimal weights by changing k values, and the highest accuracy we noticed is 78.58% for k=20.

Text

Description automatically generated

**Classification using Random Forest:**

Random forest model is used widely in Classification and Regression problems. Random forests are very good for classification problems but are slightly less good at regression problems. It builds decision trees on different samples and takes their majority vote for classification. The Accuracy of the testing data is 82.05%.

**Text

Description automatically generated**

## Evaluation:

## Accuracy for all the three models implemented in the project is as shown below:

Chart

Description automatically generated with low confidence

From all the three models discussed, we can see that the Random Forest model has the highest accuracy of 82.05%, followed by Logistic Regression and KNN with an accuracy of 78.03% and 77.66%, respectively.

# **Recommendations:**

* According to the demographics, female clients, better educated, single, and between 30 and 40, are more likely to pay on time.
* Clients aged 20 to 40 are likely to default the payment, so we recommend banks lend them responsibly.