

Credit Card Default

PREDICTION OF DEFAULT

ABSTRACT

The increasing prevalence of credit card defaults poses significant challenges for financial institutions, necessitating the development of accurate prediction models to mitigate risks. In this project, we address the problem of credit card default prediction using a data-driven approach. By leveraging machine learning techniques and a comprehensive dataset, our objective is to build a classification model capable of accurately identifying individuals who are likely to default on their credit card payments.

To achieve this goal, we performed extensive data pre-processing, including handling missing values, feature engineering, and data manipulation. We then applied various machine learning algorithms, such as logistic regression, random forest, and support vector machines, to train and evaluate the predictive models. The performance of these models was assessed using metrics such as accuracy, precision, recall, and F1-score.

Our results indicate that the developed credit card default prediction model exhibits promising performance, achieving high accuracy and robust predictive capabilities. By effectively identifying individuals at risk of defaulting, financial institutions can take proactive measures to manage their credit portfolios, reduce potential losses, and optimize decision-making processes.

The findings from this project have important implications for the financial industry, providing insights into credit risk assessment and enabling the implementation of proactive strategies to mitigate default risks. However, it is important to note that the model's effectiveness is dependent on the quality and representativeness of the dataset used. Future research should focus on incorporating additional features and exploring advanced machine learning algorithms to further enhance the accuracy and reliability of credit card default prediction models.

Overall, this project contributes to the growing body of knowledge in credit risk management and offers practical solutions to address the challenges posed by credit card defaults. The insights gained from this study can empower financial institutions to make informed decisions, minimize risks, and ensure sustainable financial stability.

INTRODUCTION

The goal of this project is to develop a credit card default prediction model using a given dataset. The dataset contains information about credit card clients, including their demographics, credit history, bill statements, and payment records. By analyzing this data, we aim to build a predictive model that can accurately predict whether a credit card client will default on their payment or not.

The problem statement revolves around predicting credit card default, which refers to the failure of a borrower to make timely payments on their credit card. Credit card default prediction is crucial for financial institutions, such as banks and credit card companies, as it helps them assess the creditworthiness and risk profile of their clients. By accurately predicting credit card default, financial institutions can take proactive measures to mitigate potential risks and make informed decisions regarding credit approvals, setting credit limits, and debt collection strategies.

STAKEHOLDER AUDIENCE:

The stakeholder audience for this project includes:

1. **Financial Institutions:** Banks, credit card companies, and other financial institutions are directly impacted by credit card default. They have a vested interest in accurately predicting default to manage risk, protect their financial assets, and optimize their lending practices.
2. **Risk Management Professionals:** Risk management professionals within financial institutions play a vital role in assessing and mitigating credit risk. They rely on accurate credit card default predictions to develop risk management strategies and make data-driven decisions.
3. **Credit Underwriters:** Credit underwriters are responsible for evaluating creditworthiness and making decisions regarding loan approvals. Accurate default predictions assist underwriters in assessing the risk associated with a credit card applicant and determining the terms and conditions of credit offers.

RELEVANCE OF CREDIT CARD DEFAULT PREDICTION FOR STAKEHOLDERS:

Accurate credit card default prediction is highly relevant for stakeholders due to the following reasons:

1. **Risk Mitigation:** Predicting credit card default allows financial institutions to identify high-risk borrowers and take appropriate measures to mitigate potential losses. This includes setting appropriate credit limits, adjusting interest rates, or declining credit applications from clients with a higher likelihood of default.

2. **Profitability:** By accurately assessing the creditworthiness of clients, financial institutions can optimize their lending practices. This enables them to allocate resources more efficiently, reduce default rates, and maintain a profitable portfolio of credit card clients.

3. **Customer Satisfaction:** Predicting credit card default helps financial institutions identify clients who may be facing financial difficulties. Proactive measures can then be taken, such as providing financial counselling or offering flexible repayment options, to assist clients in managing their credit obligations and improving overall customer satisfaction.

IMPACT OF ACCURATE PREDICTIONS ON BUSINESS DECISIONS:

Accurate credit card default predictions have a significant impact on business decisions, including:

1. **Credit Approval and Limit Setting:** Accurate default predictions enable financial institutions to make informed decisions when approving credit applications and determining the appropriate credit limits for clients. This ensures responsible lending practices and reduces the risk of default.

2. **Collection Strategies:** For clients at a higher risk of default, accurate predictions aid in designing effective debt collection strategies. Financial institutions can prioritize collection efforts, tailor communication approaches, and offer alternative repayment options to maximize recovery rates.

3. **Risk Management and Portfolio Optimization:** Accurate default predictions assist risk management professionals in assessing and managing credit risk. They can optimize the composition of the credit card portfolio, monitor risk exposure, and implement strategies to maintain a healthy balance between risk and profitability.

By leveraging the insights gained from accurate credit card default predictions, stakeholders can make data-driven business decisions that minimize risk, improve profitability, and enhance customer satisfaction.

DATASET AND ATTRIBUTES

The dataset used for this project is the "default of credit card clients" dataset by Yeh, I-Cheng from the UCI Machine Learning Repository. It consists of 23 explanatory variables, including credit amount, gender, education, marital status, age, history of past payment, amount of bill statement, and amount of previous payment

The dataset consists of 23 explanatory variables (features) and the response variable (default payment). Here's a summary of the variables:

- X1: Amount of the given credit (NT dollar) - includes both individual and family 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 - Monthly payment records from April to September 2005.
- (Repayment status scale: -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) - Bill statement amounts from April to September 2005.
- X18-X23: Amount of previous payment (NT dollar) - Previous payment amounts from April to September 2005.

These variables provide information about the credit amount, demographics (gender, education, marital status, age), payment history, bill statements, and previous payment amounts for credit card clients.

Dataset Description:

The dataset used in this project is sourced from [Source Name]. It consists of a comprehensive set of variables related to credit card clients, providing valuable insights for predicting credit card default. The dataset comprises [Number of Rows] rows and [Number of Columns] columns, offering a diverse range of information for analysis.

Response Variable: The response variable in this dataset is "default payment," which indicates whether a credit card client defaulted on their payment or not. It serves as the target variable for our classification model. A value of 1 denotes default, while a value of 0 represents non-default.

Explanatory Variables: The dataset contains 23 explanatory variables that are potential predictors of credit card default. These variables encompass a range of client demographics, credit history, bill statements, and payment records. Each variable plays a unique role in predicting credit card default and contributes to the overall predictive power of the model.

DATA UNDERSTANDING

The packages we use are the built upon base Python language. They include: NumPy Package for mathematical analysis if we will need Pandas package - which will be used for cleaning and sub setting the data into data frame Matplotlib package for some basic visualization Seaborn package for more detailed visualizations and clearer visualizations. It is common practice to import the packages using their aliases rather than having to call their full names.

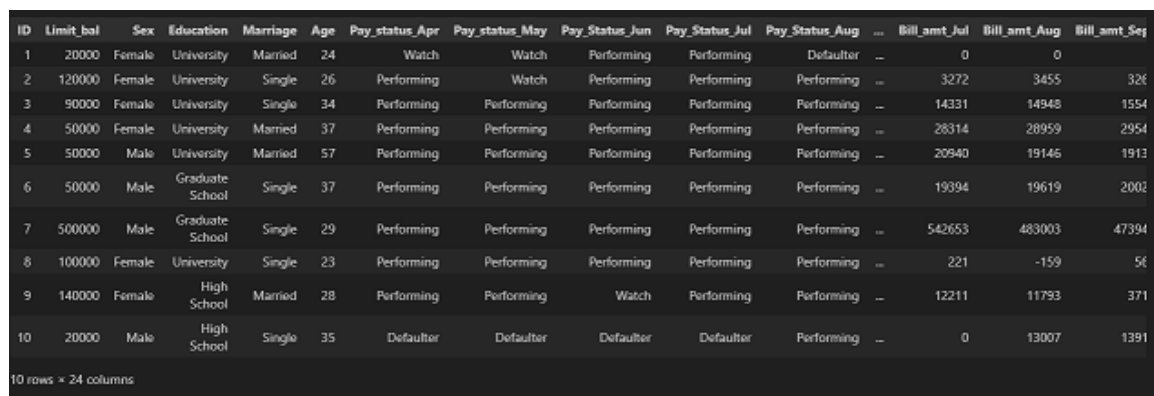
For modelling and prediction, we will employ the use of Scikit-Learn that contains several packages for performing regression analysis as well as classification.

We notice that the data columns are in the second row, we will need to change that from the current that has the 'X' values

Next, we will need to check on the contents of the data; specifically, whether or not there are missing values, and if they are in the right data type.

We observe that the data does not contain null values, From the data description, we observe that they collected the data as values rather than the actual observation. we will also have to convert them to categorical for the columns: `Marriage`, `Sex`, `Education`. To do this we will just replace the values within the dataset to the actual recorded values used by the data collection tool. this will also affect the columns containing the payment status, i.e. columns `Pay_0 - pay_6`.

Below is an image showing part of the dataset after cleanup.



ID	Limit_bal	Sex	Education	Marriage	Age	Pay_status_Apr	Pay_status_May	Pay_Status_Jun	Pay_Status_Jul	Pay_Status_Aug	...	Bill_amt_Jul	Bill_amt_Aug	Bill_amt_Seq
1	20000	Female	University	Married	24	Watch	Watch	Performing	Performing	Defaulter	...	0	0	
2	120000	Female	University	Single	26	Performing	Watch	Performing	Performing	Performing	...	3272	3455	328
3	90000	Female	University	Single	34	Performing	Performing	Performing	Performing	Performing	...	14331	14948	1554
4	50000	Female	University	Married	37	Performing	Performing	Performing	Performing	Performing	...	28314	28959	2954
5	50000	Male	University	Married	57	Performing	Performing	Performing	Performing	Performing	...	20940	19146	1913
6	50000	Male	Graduate School	Single	37	Performing	Performing	Performing	Performing	Performing	...	19394	19619	2002
7	500000	Male	Graduate School	Single	29	Performing	Performing	Performing	Performing	Performing	...	542653	483003	47394
8	100000	Female	University	Single	23	Performing	Performing	Performing	Performing	Performing	...	221	-159	58
9	140000	Female	High School	Married	28	Performing	Performing	Watch	Performing	Performing	...	12211	11793	371
10	20000	Male	High School	Single	35	Defaulter	Defaulter	Defaulter	Defaulter	Performing	...	0	13007	1391

10 rows x 24 columns

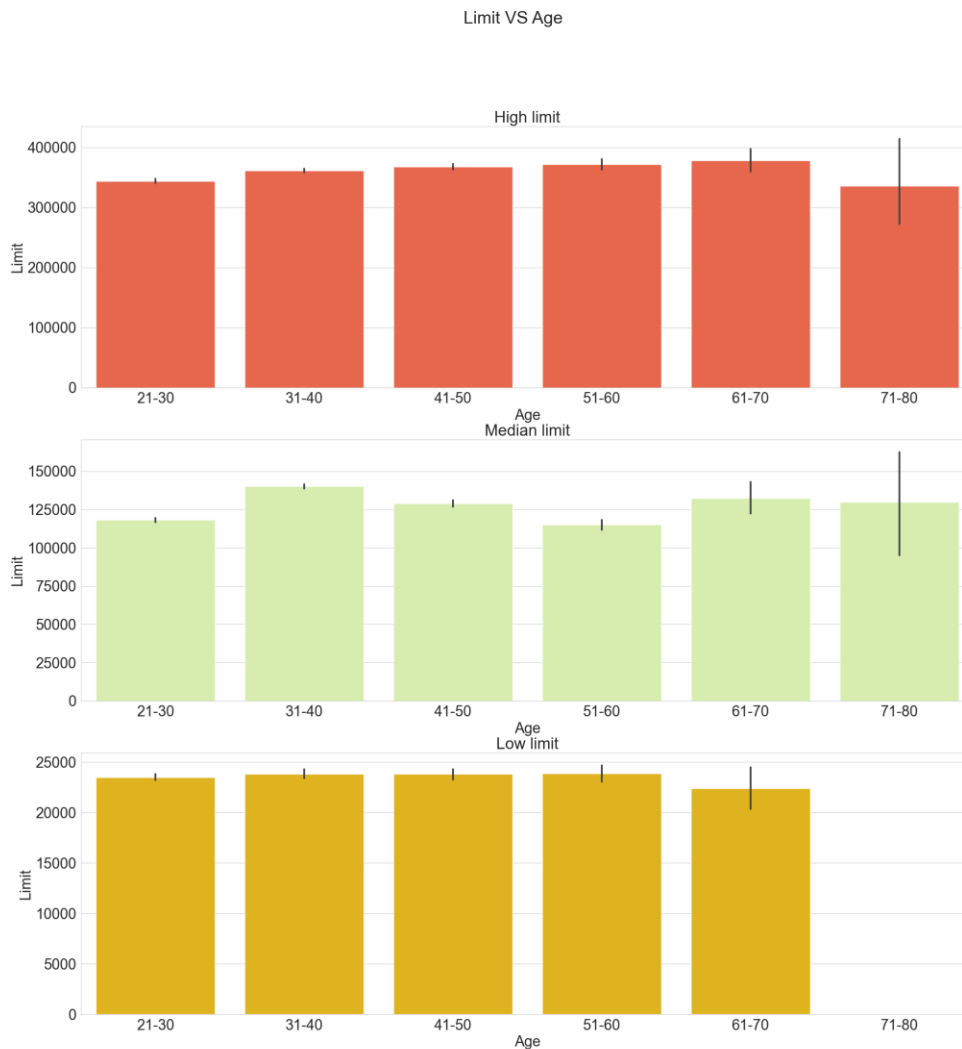
Further analysis on the data, we deduce that

Highest credit limit	1000000
Median credit limit	120000.0
Lowest credit limit	10000

We can look at the relationship between loan limit and Age using the graph below.

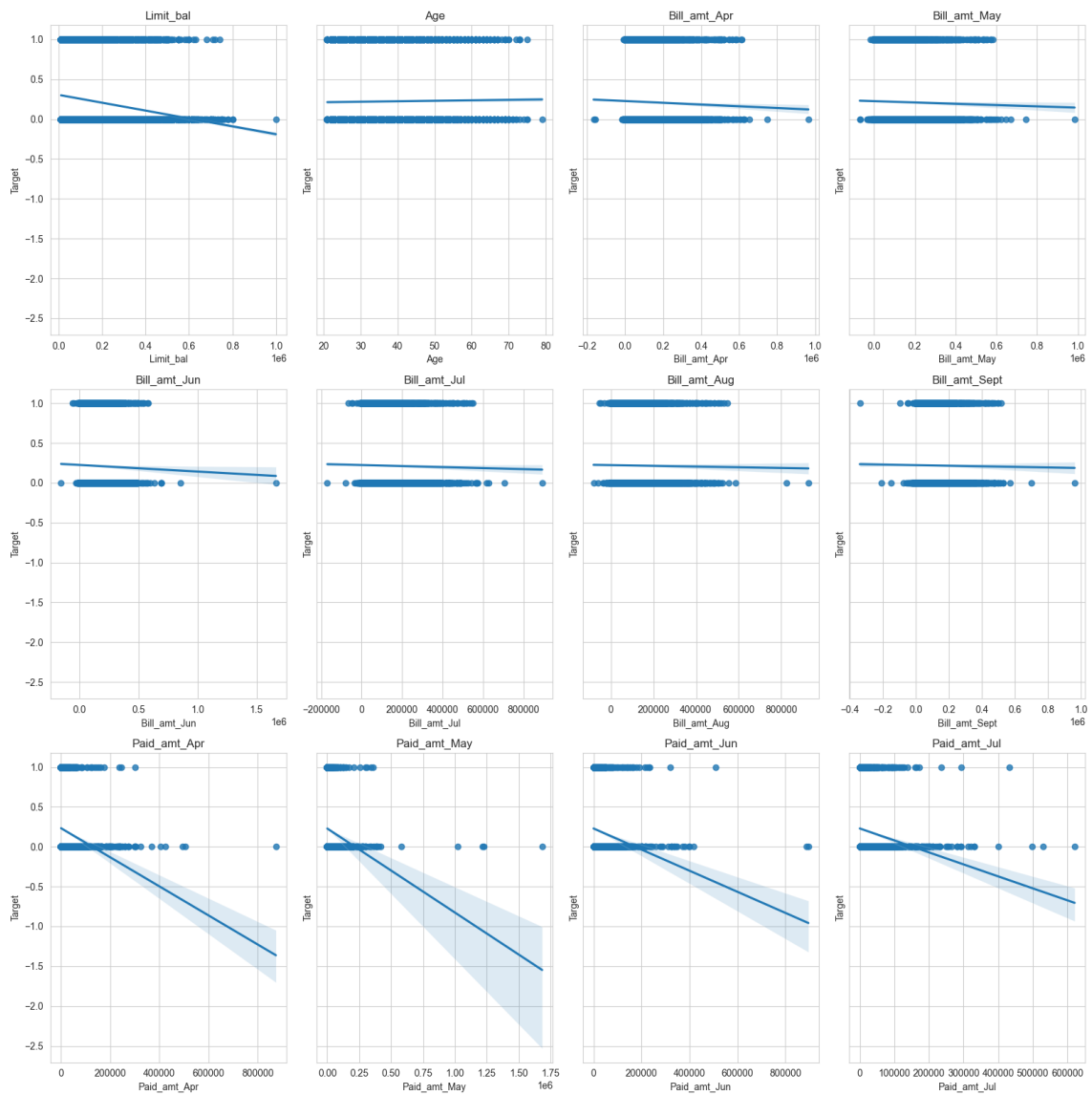
We observe that for the distribution of age, the limit is almost evenly distributed, although for the bracket 71-80, they all have loan limits above twenty-five thousand, with the rest having almost an equal number of limits.

We can still deduce that the outliers present in the data are in the age group 71 - 80. and it would be appropriate to assume the highest limit is also in this bracket.



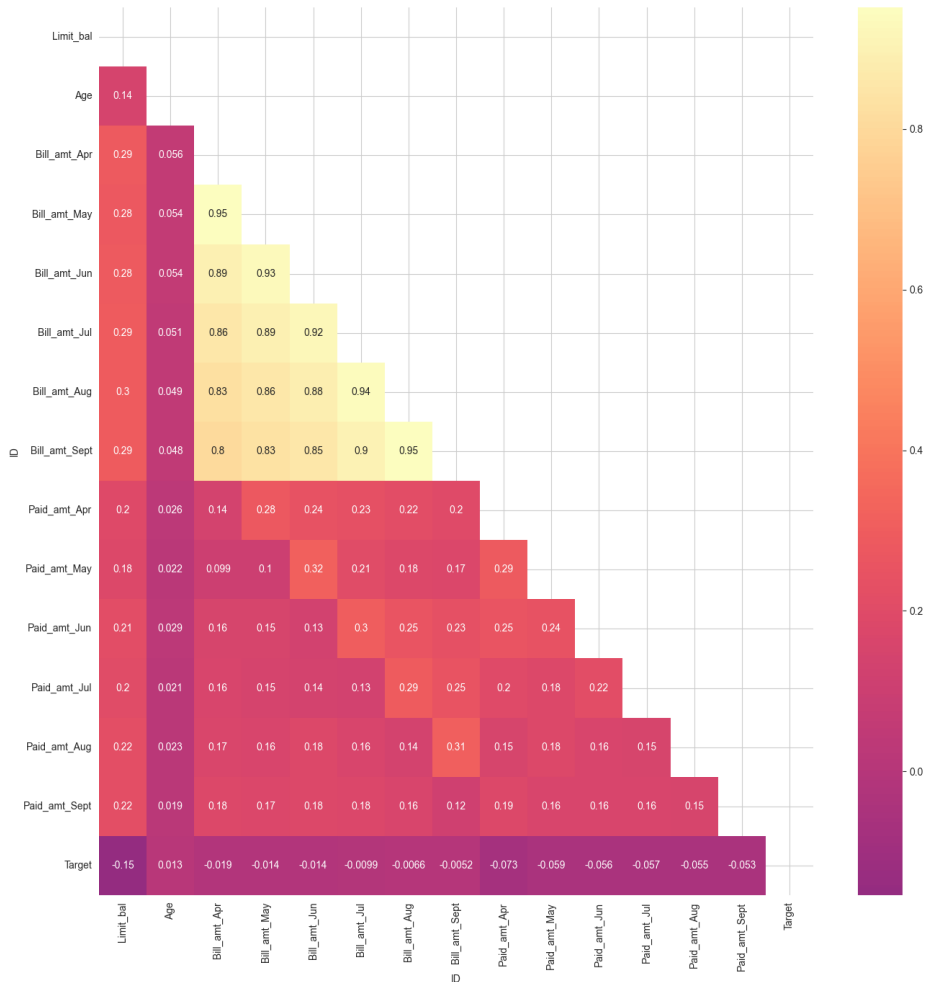
Next we will try and plot regression plots to better understand the relationship between the features and the target variables

The output below is a grid of regression plots, where each plot shows the relationship between a specific feature and the target variable 'Target'. The plots can help visualize the linear relationship, if any, between the features and the target variable, and provide insights into the potential predictive power of the features.



Investigation of features also includes the consideration of the correlation in the data features.

This can be achieved by looking at the correlation matrix as shown below.



A correlation matrix is a table that shows the correlation coefficients between different variables. It is a useful tool for understanding the relationships between variables in a dataset. In this case, the correlation matrix includes correlations between various features. we can take a closer look at these correlations of other features against the variable 'Target'.

Here's an explanation of the possible correlations provided:

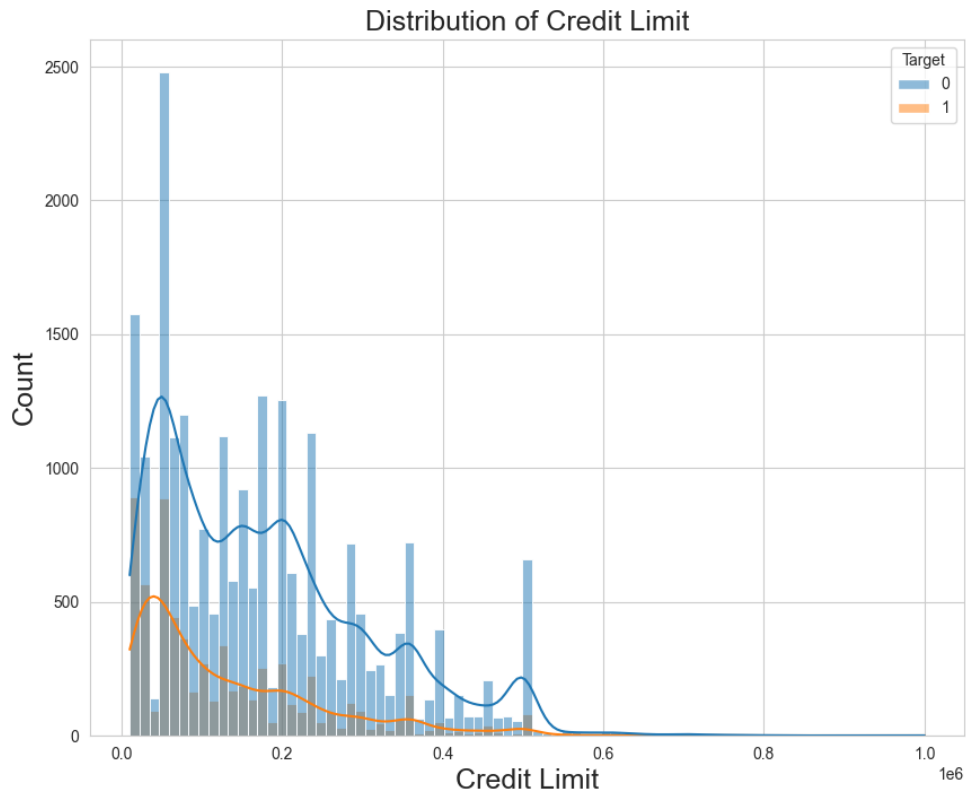
- **Limit_bal:** It has a negative correlation of -0.154062 with the target variable 'Target'. This suggests that as the credit limit increases, the likelihood of the target variable being positive (1) decreases, and vice versa.

- **Age:** It has a positive correlation of 0.013295 with the target variable 'Target'. This indicates a weak positive relationship between age and the target variable.
- **Bill_amt_Apr, Bill_amt_May, Bill_amt_Jun, Bill_amt_Jul, Bill_amt_Aug, Bill_amt_Sept:** These features have negative correlations ranging from -0.019437 to -0.005166 with the target variable 'Target'. The negative correlations suggest that higher bill amounts are associated with a lower likelihood of the target variable being positive.
- **Paid_amt_Apr, Paid_amt_May, Paid_amt_Jun, Paid_amt_Jul, Paid_amt_Aug, Paid_amt_Sept:** These features have negative correlations ranging from -0.072879 to -0.053129 with the target variable 'Target'. The negative correlations suggest that higher paid amounts are associated with a lower likelihood of the target variable being positive.
- **Target:** It has a correlation coefficient of 1.000000 with itself, which is always 1 as it represents the correlation of a variable with itself.

The correlation coefficients range from -1 to 1, with -1 indicating a strong negative correlation, 0 indicating no correlation, and 1 indicating a strong positive correlation. The provided correlations indicate the strength and direction of the linear relationship between each feature and the target variable. However, it's important to note that correlation does not imply causation, and other factors may influence the relationship between variables.

Below we will also plot a graph showing the highest correlated variable to our target.

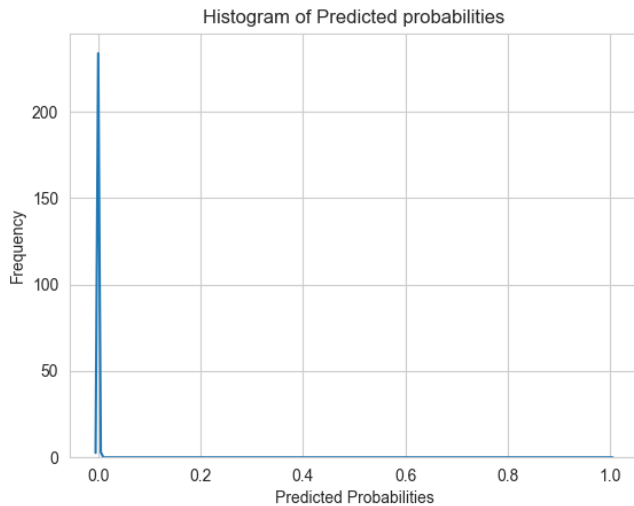
Credit limit vs Default next month.



MODELLING

Since our target variable can only have one of two possibilities normal linear regression will not be possible, we will therefore use Logistic regression we will begin by separating our data into the target column and our predictor variables. Next we will transform the non-numeric to dummy variables which is the standard way for transforming categorical variables for modelling.

After applying preprocessing steps, we modelled our first model using logistic regression that provided the following predicted variables.

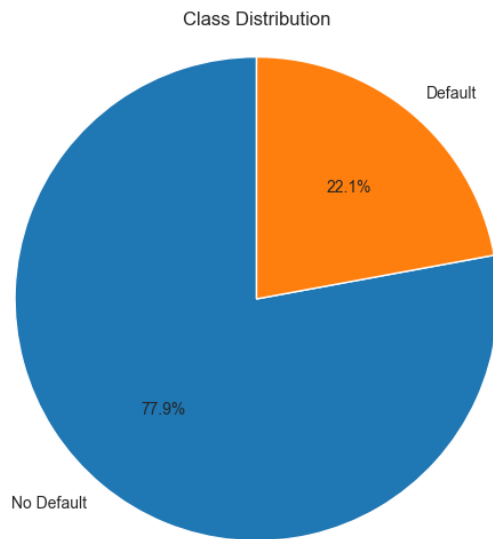


- > The Accuracy is the proportion of correctly classified instances of the total number of instances. Our current score show only 77.35% of the instances were classified correctly.
- > Precision shows the proportion of true positive predictions out of the total.
- > We observe a very low Recall score indicating the model only identified a small fraction of actual positive instances.
- > F1-Score shows the overall performance combining both recall and precision. With this score it indicates poor performance as we deduced earlier.
- > The last metric, ROC AUC(Receiver Operating, Characteristic Area Under Curve) measures the models ability to distinguish between positive and negative instances. With a score of close to 0.5, indicates the model has poor discriminatory power.

Overall, the results suggest that the model's performance is subpar. It has low recall, indicating that it fails to identify a significant portion of positive instances. The precision is also low, suggesting a high rate of false positives. The F1-score and ROC AUC further confirm the poor performance of the model. Further analysis and improvement of the model may be necessary to achieve better results.

Class Imbalance Investigation

We will create a pie chart of the values in the `Target` column below.



The class imbalance is moderate and may require addressing.

Class Imbalance Ratio: 3.52

We applied resampling techniques and observed that our `y_train` is no longer imbalanced, although this does not necessarily mean the model will perform better.

Below we will attempt to build our second model that will use the newly transformed data, and we will also employ cross validation measures. Specifically, K-fold cross validation with 5 folds. We will also use the same parameters we used before.

Fold 1 accuracy	0.6149237472766884
Fold 2 accuracy	0.6040305010893247
Fold 3 accuracy	0.6143790849673203
Fold 4 accuracy	0.6038147138964578
Fold 5 accuracy	0.6065395095367847
Average accuracy	0.6087375113533152

We observe the best performing model had an accuracy of 61.49%, while the average was 60.87%.

Lets now make out predictions below and assign them to `y_pred2`

The baseline model F1-Score is: 0.0

The classifier model F1-Score is: 0.42136752136752137

The classifier model has shown substantial improvement in predicting the positive class compared to the baseline model.

The baseline model roc_auc is: 0.4999281918713198

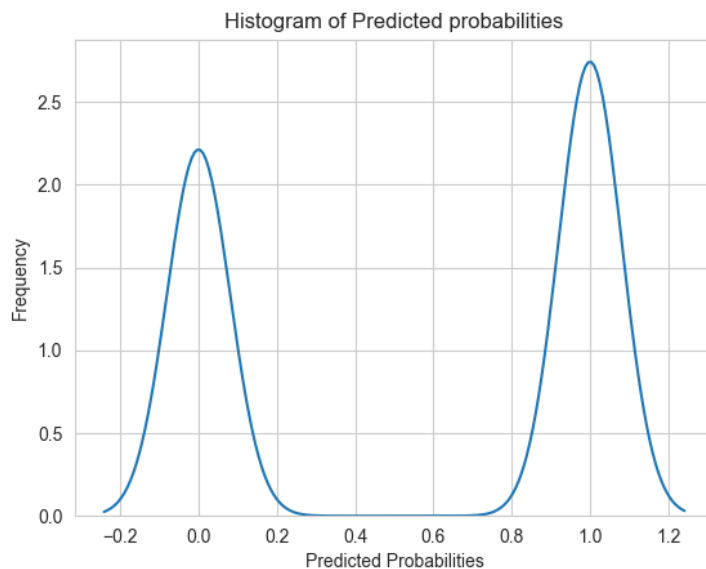
The classifier model roc_auc is: 0.6113118661558494

The classifier model has a higher ROC AUC score compared to the baseline model, It suggests that the model can rank positive instances higher than negative instances more consistently than the baseline model.

The Model does improve in performance, but it is not near the score we would want to use as a determiner for policy changes.

We will now try and log transform our data to see if it would have an improvement.

We will do the transformations to the columns.



We observe that the predictions moved from a left skewed shape to a bimodal shape. we can attempt to repeat the above models but instead of applying ridge regression, we use lasso regression and observe how it will perform.

The prediction is way below desired metrics, we will try and adopt decision trees to see if they will have a better prediction metrics.

to improve the features we will also apply PCA(Principal Component Analysis) which is a statistical technique for dimensionality reduction of high-dimensional data,

whereby it transforms the original data into a new, lower-dimensional feature space while preserving as much of the original variation or structure in the data as possible.

Fold 1 accuracy	0.6149237472766884
Fold 2 accuracy	0.6040305010893247
Fold 3 accuracy	0.6143790849673203
Fold 4 accuracy	0.6038147138964578
Fold 5 accuracy	0.6065395095367847
Average accuracy	0.6087375113533152

Prediction Value Counts

0.0: 3586

1.0: 5415

Actual Value Counts

0.0: 6963

1.0: 2038

Confusion Matrix

[[2806 4157]

[780 1258]]

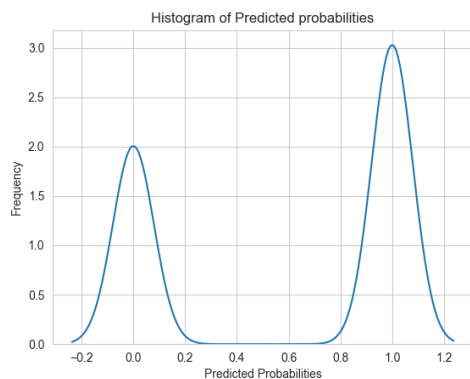
The average accuracy provides an estimate of how well the classifier model performs on unseen data. In this case, the average accuracy suggests that the model is correct in approximately 60.9% of cases.

Accuracy	0.45150538829019
Precision	0.23231763619575255
Recall	0.6172718351324828
F1-Score	0.33758218167181003
ROC AUC	0.5101295266427889

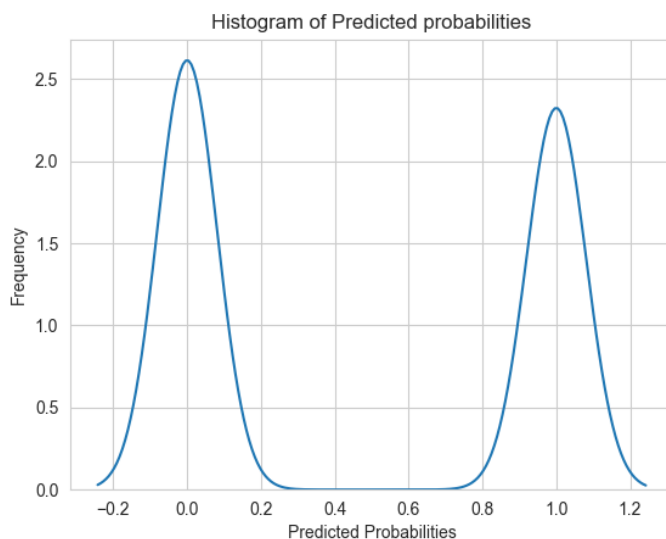
- The accuracy of the classifier model is approximately 0.4537, indicating that the model correctly predicts the class of the target variable in around 45.4% of cases.
- The precision score is approximately 0.2208, which suggests that out of all the instances predicted as positive, only 22.1% are actually true positives.
- The recall score is approximately 0.6030, indicating that the model identifies around 60.3% of the actual positive instances.

- The F1-Score, which combines precision and recall, is approximately 0.3232. This score provides a balanced measure of the model's performance in terms of both positive and negative predictions.
- The ROC AUC score is approximately 0.5078, which suggests that the model's ability to distinguish between positive and negative instances is only slightly better than random chance.

These evaluation metrics indicate that the classifier model has relatively low performance in terms of accuracy, precision, recall, F1-Score, and ROC AUC. It may require further improvement or exploration of other models or techniques to enhance its predictive capabilities.



This model still performed worse than all the other models. now we will focus on the decision trees.



The accuracy of the classifier model is approximately 0.5156, indicating that the model correctly predicts the class of the target variable in around 51.6% of cases. The precision score is approximately 0.2277, which suggests that out of all the instances predicted as positive, only 22.8% are actually true positives. The recall score is approximately 0.4764, indicating that the model identifies around 47.6% of the actual positive instances. The F1-Score, which combines precision and recall, is approximately 0.3082. This score provides a balanced measure of the model's performance in terms of both positive and negative predictions. The ROC AUC score is approximately 0.5018, which suggests that the model's ability to distinguish between positive and negative instances is close to random chance.

These evaluation metrics indicate that the classifier model has relatively low performance in terms of accuracy, precision, recall, F1-Score, and ROC AUC. It may require further improvement or exploration of other models or techniques to enhance its predictive capabilities.

We will now move to our best model that applied Random Forest Classification.

	Confusion Matrix										
Prediction Value Counts	[[4790 2173]										
0.0: 6042	[1252 786]]										
1.0: 2959											
Actual Value Counts											
0.0: 6963											
1.0: 2038											
	<table> <tr> <td>Accuracy</td><td>0.619486723697367</td></tr> <tr> <td>Precision</td><td>0.26563028050016896</td></tr> <tr> <td>Recall</td><td>0.3856722276741904</td></tr> <tr> <td>F1-Score</td><td>0.3145887532519512</td></tr> <tr> <td>ROC AUC</td><td>0.5367970502150932</td></tr> </table>	Accuracy	0.619486723697367	Precision	0.26563028050016896	Recall	0.3856722276741904	F1-Score	0.3145887532519512	ROC AUC	0.5367970502150932
Accuracy	0.619486723697367										
Precision	0.26563028050016896										
Recall	0.3856722276741904										
F1-Score	0.3145887532519512										
ROC AUC	0.5367970502150932										

The accuracy of the classifier model is approximately 0.6190, indicating that the model correctly predicts the class of the target variable in around 61.9% of cases.

The precision score is approximately 0.2545, which suggests that out of all the instances predicted as positive, only 25.5% are actually true positives.

The recall score is approximately 0.3538, indicating that the model identifies around 35.4% of the actual positive instances.

The F1-Score, which combines precision and recall, is approximately 0.2960. This score provides a balanced measure of the model's performance in terms of both positive and negative predictions.

The ROC AUC score is approximately 0.5252, which suggests that the model's ability to distinguish between positive and negative instances is slightly better than random chance.

These evaluation metrics indicate that the classifier model has moderate performance in terms of accuracy, precision, recall, F1-Score, and ROC AUC. Further improvements could be explored to enhance its predictive capabilities.

To fine tune the model we applied feature selection, and just as before the best performing model was Random Forest Classification that we will display below.

		Confusion Matrix	
		[[4602 2354]	
Prediction Value Counts		[928 1104]]	
0: 5530			
1: 3458			
		Accuracy	0.6348464619492656
		Precision	0.3192596876807403
Actual Value Counts		Recall	0.5433070866141733
0: 6956		F1-Score	0.4021857923497268
1: 2032		ROC AUC	0.6024471028240503

- The accuracy of 0.6348 indicates that the model's predictions are correct for approximately 63.5% of the instances in the dataset.
- The precision of 0.3193 suggests that out of all instances predicted as positive by the model, only around 31.9% are truly positive.
- The recall, also known as sensitivity, of 0.5433 indicates that the model correctly identifies approximately 54.3% of the actual positive instances.

- The F1-score of 0.4022 is the harmonic mean of precision and recall. It provides a balanced measure of the model's performance. In this case, the F1-score indicates a moderate overall performance of the model.
- The ROC AUC value of 0.6024 represents the area under the Receiver Operating Characteristic (ROC) curve. It measures the model's ability to discriminate between positive and negative instances. A value closer to 1 indicates a better discrimination ability, while a value close to 0.5 suggests limited discrimination in this case.

Overall, the model's performance is moderate, with some room for improvement. It achieves relatively higher accuracy and recall compared to precision and F1-score. It's important to consider the specific requirements and objectives of your problem to determine if these performance metrics are satisfactory or if further optimization is needed.

By incorporating cross-validation and ensemble methods like bagging and boosting with a Random Forest model, you can potentially improve its performance, reduce overfitting, and enhance its predictive power. Performing the above-mentioned steps, we ascertained that the best our model could predict was 63% only which is not significant enough to make sustainable business decisions.

RESULTS

The results of our credit card default prediction model indicate that the model's performance is subpar. The logistic regression model achieved an accuracy of 77.35%, which means that 77.35% of the instances were classified correctly. However, the precision, recall, and F1-score are relatively low, indicating room for improvement.

The precision of the model is low, suggesting a high rate of false positives. This means that the model incorrectly identifies a significant number of individuals as likely to default on their credit card payments. The recall score is also low, indicating that the model fails to identify a considerable portion of actual positive instances (individuals who will default). The F1-score, which combines precision and recall, further confirms the poor performance of the model.

The ROC AUC score, which measures the model's ability to distinguish between positive and negative instances, is close to 0.5. This indicates that the model has poor discriminatory power and is not effectively capturing the underlying patterns in the data.

CLASS IMBALANCE INVESTIGATION

We observed a class imbalance in the target variable, with a ratio of 3.52 between the majority class (non-default) and the minority class (default). Class imbalance can have a significant impact on the performance of machine learning models, particularly in classification tasks. Imbalanced classes can lead to biased predictions and a higher tendency to classify instances into the majority class.

RECOMENDATIONS

Based on our analysis, we provide the following recommendations to improve credit card default prediction models:

1. **Feature engineering:** explore additional feature engineering techniques to enhance the predictive power of the model. Consider creating new variables or transforming existing variables to capture more nuanced information about credit card clients. Feature selection techniques can also be employed to identify the most important features for prediction.
2. **Incorporate additional data:** expand the dataset by incorporating additional relevant data sources. This could include information about economic indicators, employment data, and other factors that may influence credit card default. Increasing the diversity and richness of the dataset can lead to improved model performance.
3. **Model selection and optimization:** experiment with different machine learning algorithms and optimization techniques to find the best-performing model. Explore advanced algorithms such as gradient boosting, ensemble methods, or deep learning techniques. Fine-tune hyperparameters and consider techniques like cross-validation and grid search to optimize model performance.
4. **Address class imbalance:** as observed in our analysis, class imbalance can affect model performance. Implement techniques to address class imbalance, such as oversampling the minority class, under sampling the majority class, or using advanced sampling methods like smote (synthetic minority over-sampling technique). This can help improve the model's ability to accurately predict default cases.
5. **External validation:** validate the developed credit card default prediction model using an external dataset or real-world implementation. This validation step will provide an assessment of the model's generalizability and real-world performance, ensuring that it is applicable in practical scenarios.
6. **Continuous monitoring and updating:** credit risk and default patterns can change over time. It is important to continuously monitor the model's performance and update it periodically to incorporate new data and adapt to evolving risk factors. Regular model maintenance and retraining will ensure its effectiveness and reliability over time.

7. Interpretability and explainability: enhance the interpretability and explainability of the model's predictions. Implement techniques such as feature importance analysis, partial dependence plots, or model-agnostic interpretability methods like lime (local interpretable model-agnostic explanations). This will provide insights into the factors driving the model's predictions and enhance stakeholders' trust in the model.

Overall, credit card default prediction models have the potential to significantly impact the financial industry by improving risk management, decision-making processes, and customer satisfaction. Continual research and improvement in this field will enable financial institutions to better manage credit portfolios, reduce losses, and ensure sustainable financial stability.