Credit Card Default Prediction

INTRODUCTION

American Express is a fully integrated payments firm that gives people access to goods, information, and experiences that improve quality of life and help businesses succeed. Thanks to the efforts of their staff members around the world, the organization makes every attempt to give customers the finest experience possible.

In order to improve the lives of the company's globally linked digital consumers and strengthen their bond with it, American Express develops digital goods and services. Deep comprehension and understanding of what customers need and expect form the foundation of all digital goods developed. The company's focus on four key areas—Data & Technology, Platforms & Capabilities, Engaging and Intentional Experiences, and a Deep Partner Ecosystem—allows the company to provide memorable experiences to our clients at every touch point.

The credit card industry has grown greatly overall. This sector is seeing tremendous advancements, including various technological developments. These recent advancements provide easy onboarding processes, unique card products, tailored offers, rewards, and superior portable applications, which have proven to be incredibly helpful to current customers as well as attracting new customers. Credit card issuing companies are also working to advance this industry's awareness and innovation.

PROBLEM STATEMENT

A credit card default happens when the customer fails to make any payment towards the credit card outstanding bill for a long period of time. Customer application and bureau data with the default tagging i.e., if a customer has missed a cumulative of 3 payments across all open trades, his default indicator is 1 else 0. Data consists of independent variables at the time T0 and the actual performance of the individual (Default/ Non Default) after 12 months.

So, the problem statement is to correctly predict whether the applicant will be going default in next 12 months from the details of their new credit card application, previous credit history and other relevant features.

Training Dataset with 83000 entries has been given with around 50 features that can be used to build and train prediction models. But the dataset is full of missing values and non-compatible values that cannot be fed to prediction models, as it is and hence, requires extensive data cleaning and pre-processing, in order to predict whether an applicant will default or not in the next 12 months, as required in the problem statement above.

MOTIVATION

A default is bad for any cardholder and the company providing the credit. It can profoundly affect both the stakeholders involved financially. So, the motivation for carrying out the project has been listed below:

- Account sent to collections - Credit card issuers can either close the customer's account and transfer the debt to a collection agency.
- Legal action - Some creditors are more aggressive. They may file for a local court judgment against the default customer if they're seeking an immediate resolution, which can lead to a paycheck lien, requiring the employer to send a portion of the income directly to your creditors.
- Decrease on credit score - A late payment is the biggest factor on the credit report. A failure to make timely payments for six months can lead to a drop on credit score by hundreds of points, which can take up to years to recover.
- Increase in interest rates - If it is 60 days past due on a payment, the interest rate will probably go up drastically. Since the customers are missing payments and continuing to carry a balance, the interest payments will only accrue too, making it all the more harder to repay the debts.
- Decrease in credit limit - Defaulting on a credit card makes the defaulter look especially risky to creditors. They may lower their own risk by limiting the amount of credit to which the customer has access to. So, a default on a credit card, may lead to a credit limit decrease on other cards.

On the other hand, default in credit payment affects the profitability of the credit providers. So, to maintain the profit margins they will be forced to increase the interest rates which will again affect the economy. Hence, there is a need to predict which customers had the highest probability of defaulting so it may be prevented.

METHODOLOGY

- Pre-processing of collected data for checking data quality and data cleaning by fixing or removing incorrect and corrupted data within the data set
- Transformation of data through suitable statistical and mathematical algos to make the data more suitable for predicting the results for the problem at hand and making the data more compatible for robust prediction by models.
- Building various training models using common statistical and algorithmic models like regression, decision trees and artificial neural networks
- Training the various models built and then testing their performance
- Evaluation of trained models and comparison with each other to identify the best model for the given dataset.
- Further Tuning of the best model identified to create the finest and most optimal model for predicting the results.
- Cleaning of given test dataset and feeding it to the model to get the predicted results of the given problem at hand.

DATA PREPROCESSING

In order to convert the raw data into a more usable format with all the errors eliminated and incompatible data removed, the following pre-processing steps have been done.

- Check for Missing Values - Data has been tested for missing values. And found that there are several types of missing values exists in the dataset i.e, 'N/A', 'na', 'missing', '#VALUE!', 'NULL', None, 'Nan', 'nan', float('NaN'), '-1', np.nan.

- Dropping Columns with High Missing Data - Columns that have more than 50% of missing data have been removed from the training set, as generalizing the missing values by replacement through mean might skew the results.

- Replacing Missing Values - Missing values were replaced by corresponding mean of the particular column.

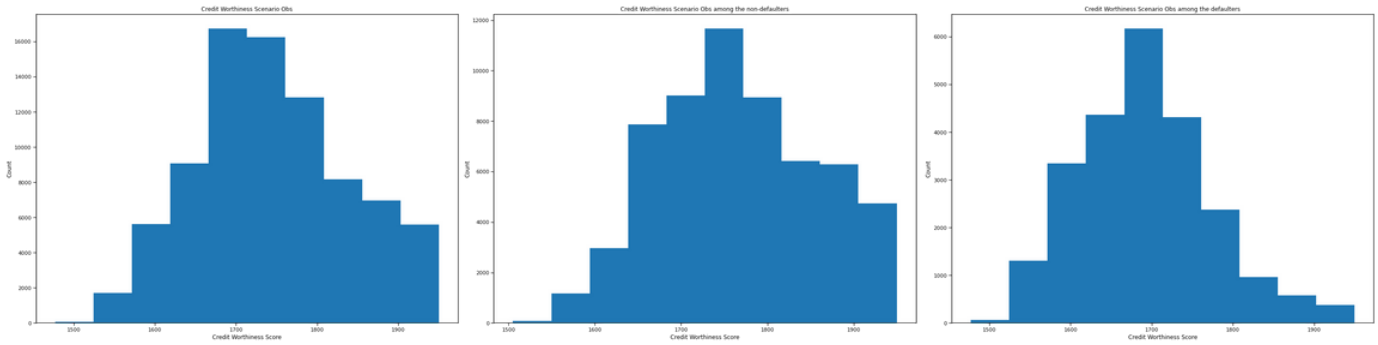
- One Hot Encoding - Most Machine Learning algorithms cannot work with categorical data and hence the data needs to be converted into numerical format. One-hot encoding has been used to quantify the categorical data. Using this method, one of the columns, which contains categorical data has been encoded into two new separate columns of binary data based on the two original classes in the parent column.

EXPLORATORY DATA ANALYSIS

Exploratory data analysis (EDA) is a method used by analysts to better understand and summarize a dataset. It is an important step in the data analysis process that helps identify trends, patterns, and relationships within a dataset. By performing EDA, analysts can gain a better understanding of the data and create more accurate models and predictions. Another reason why EDA is important is that it helps analysts understand the underlying structure of the data. This is particularly useful when working with large and complex datasets, where it can be difficult to see patterns and trends without some form of summarization or visualization.

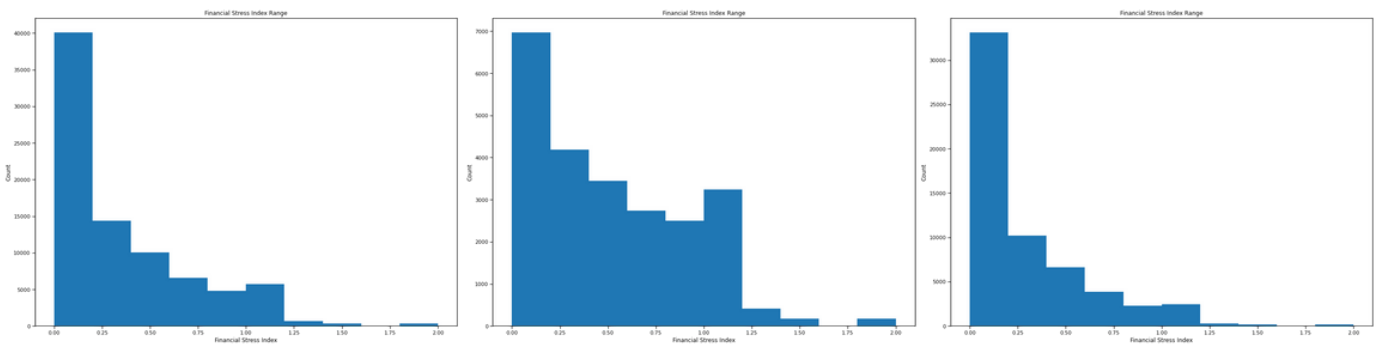
EXPLORATORY DATA ANALYSIS

The figures below show the credit worthiness score of all applicants, in comparison with that of defaulters and non-defaulters. The figure on the left represents the credit worthiness of the applicants where most of the applicants credit worthiness score is normally distributed i.e., between the range of 1700-1800, whereas the one in the middle, which represents the credit worthiness of non-defaulters, is negatively skewed i.e., between the range of 1800-1900 and the one on the right, which represents the credit worthiness of the defaulters, is positively skewed i.e., between the range of 1500-1700.



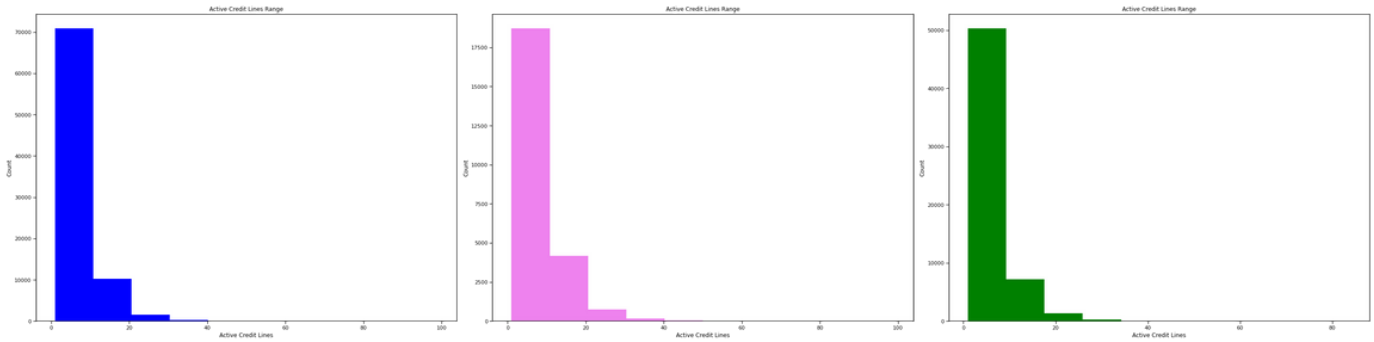
The figures below show the financial stress index of all applicants on the left, in comparison with that of the defaulters in the middle and non-defaulters on the right separately.

The degree of financial stress of an individual is a function of collection trades, bankruptcies files, tax liens invoked, etc. The graph shows that non-defaulters are less stressed financially, as compared to defaulters, which is obviously to be expected of defaulters.

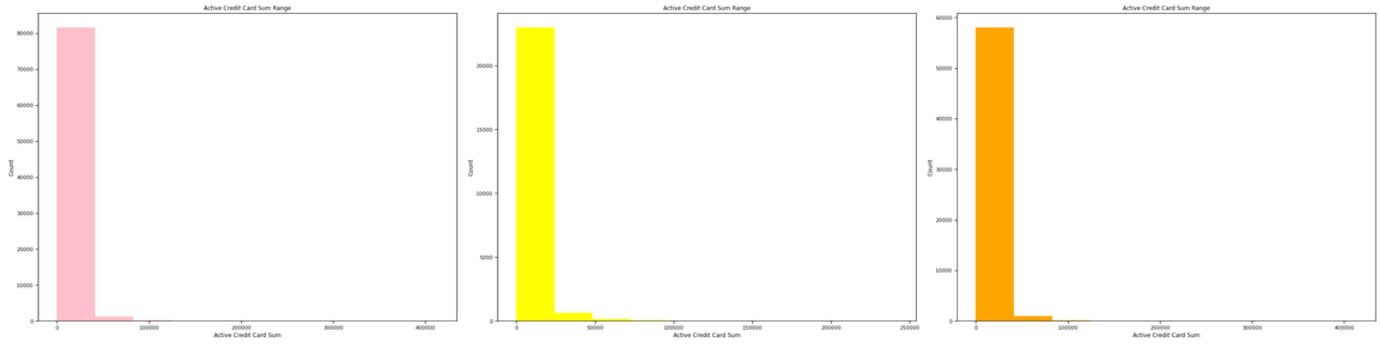


EXPLORATORY DATA ANALYSIS

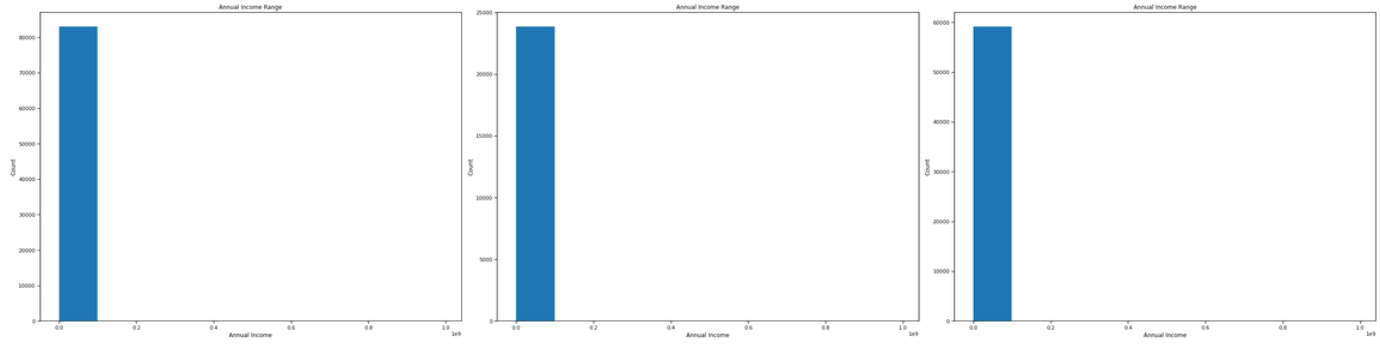
The figures below show the active credit lines range of all applicants on the left, in comparison with that of the defaulters in the middle and non-defaulters on the right separately. Interestingly, more number of non-defaulters have active credit lines than defaulters, probably because non-defaulters are more financially capable to prevent defaults.



The figures below show the active credit card sum range of all applicants on the left, in comparison with that of the defaulters in the middle and non-defaulters on the right separately. All the figures show that the data is positively skewed, with most values in the range of 0-80000.

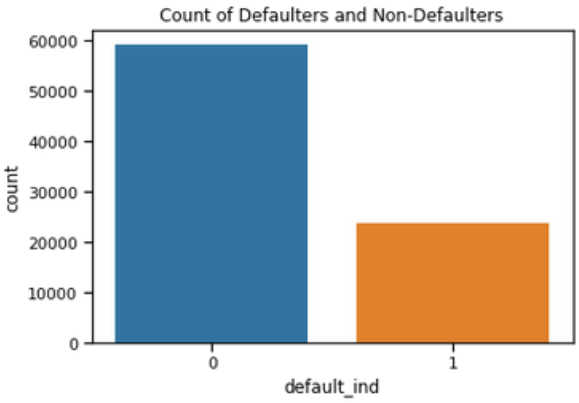


The figures below show the annual income of all applicants on the left, in comparison with that of the defaulters in the middle and non-defaulters on the right separately. As is to be expected, non-defaulters have higher annual income than defaulters.

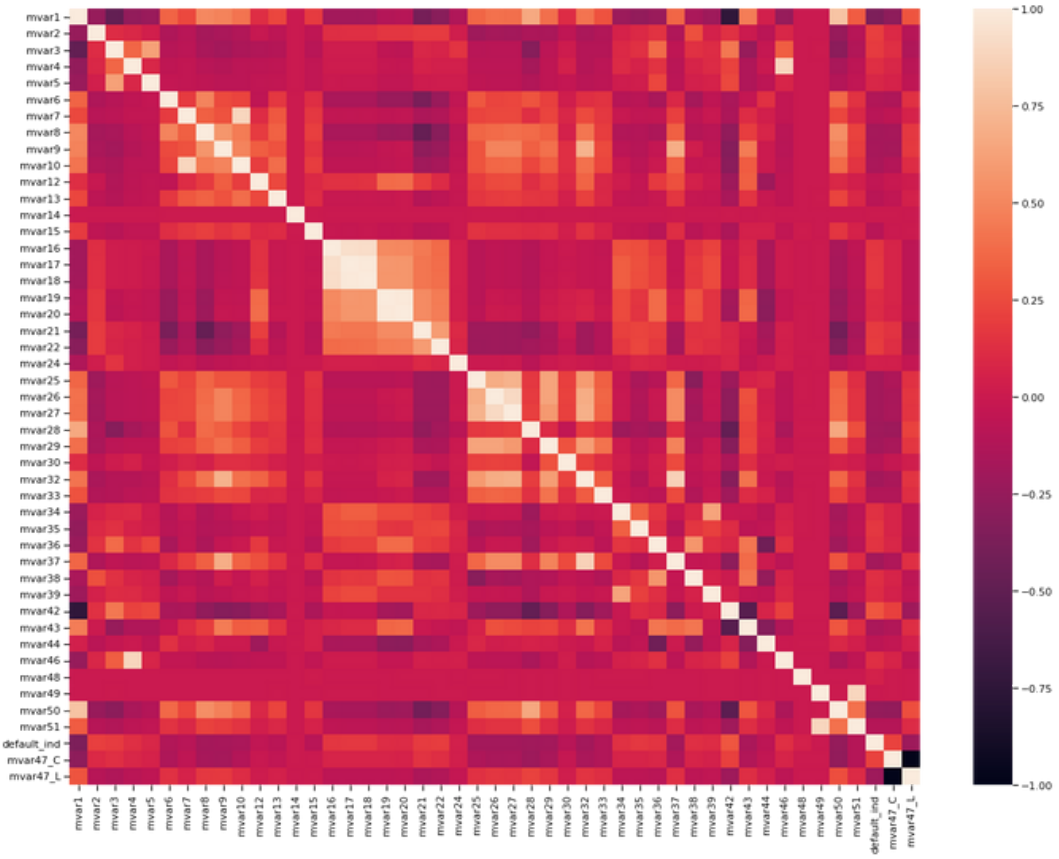


EXPLORATORY DATA ANALYSIS

The figure on the right shows the distribution of the target variable that has to be predicted in the training set. The count of non-defaulters is higher than that of defaulters in the given training sample, which indicates a class imbalanced dataset and can create bias and other problems during model training and predictions.



The heat map shown below, represents the correlation between any two given variables in the form of colors, with darker colors implying higher negative correlation and lighter colors signifying higher positive correlation. As can be seen from the plot, the variable sets of (16,17,18), (19,20), (26,27) and (49,51) exhibit high positive correlation, whereas most of the other variables exhibit zero or negative correlation with each other. The high correlations among the independent variables can affect the model's ability to accurately estimate the coefficients of the independent variables. PCA is one of the techniques that can be used to identify and remove these correlated variables, which can help to improve the performance of the regression model.

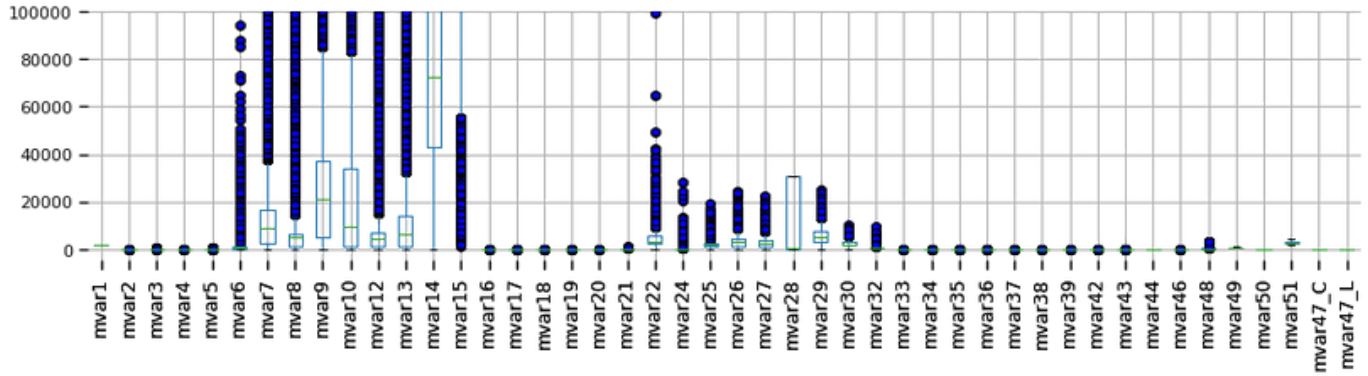


EXPLORATORY DATA ANALYSIS

Box plots are used to show distributions of numeric data values, especially when you want to compare them between multiple groups. They are built to provide high-level information at a glance, offering general information about a group of data's symmetry, skew, variance, and outliers.

The box plot below shows the distribution of each variable in the training dataset, showing that many variables are on a different scale to each other, which can significantly affect the predictability of the model and hence, all these variables need to be standardized.

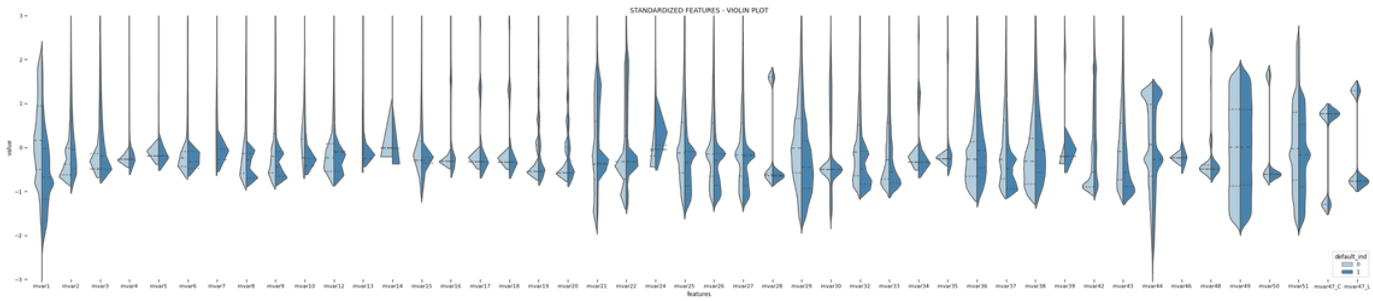
The box plot also shows the presence of a huge number of outliers in some of the variables, which can affect predictability of the models, but these outliers have been decided to be included in the training set, as removal of all the fields corresponding to the outlier values in each of the variables can significantly reduce the size of the training dataset that is needed to build the models.



STANDARDIZATION

Data standardization is the process of converting data into a common format that allows it to be more easily compared and analyzed. This is often necessary because different organizations and systems may use different standards for storing and representing data, which can make it difficult to combine and analyze data from multiple sources. Data standardization can help improve the efficiency of data analysis, as it allows analysts to use common tools and methods to analyze data from multiple sources. This can save time and effort, as analysts do not need to develop separate analysis methods for each individual source of data.

Since, data is of different formats across the columns of the dataset, it has been standardized in order to reduce the influence of the variables that have highly different scales from each other. It helps ensure data accuracy and consistency, allows data from multiple sources to be combined and analyzed, and can improve the efficiency of data analysis.



The figure above shows the distribution of data in each column in the format of a violin plot. As can be observed, the data across different variables, now appears to be more comparable and is distributed in a much better way that allows for more efficient data analysis and prediction.

PCA

Principal component analysis (PCA) is a statistical technique used to reduce the dimensionality of a dataset. It is a type of linear transformation that identifies the underlying structure of the data and projects it onto a lower-dimensional space, while retaining as much of the original information as possible.

PCA is necessary because it helps overcome the limitations of traditional data analysis methods, which can become ineffective when dealing with high-dimensional datasets. These methods often rely on visualizing the data in two or three dimensions, which can be difficult or impossible when the data has many more dimensions. In these cases, PCA can be used to reduce the dimensionality of the data, making it more manageable and allowing it to be visualized and analyzed more easily.

Another reason why PCA is important is that it can help identify the most important features or variables in a dataset. By identifying and projecting the data onto a lower-dimensional space, PCA can highlight the variables that contribute most to the overall variance in the data. This can be useful for selecting a subset of variables for further analysis, or for identifying patterns and trends in the data that may not be apparent when looking at all the variables together.

As the given dataset has a large number of variables in the independent set, we are using PCA to further reduce the dataset dimension, in order to help models to be better fitted and improve prediction results.

OVERSAMPLING

Oversampling is a technique used to address the problem of imbalanced classes in a dataset. Imbalanced classes occur when there is a disproportionate ratio of observations in different classes, such as when one class has significantly more observations than another. This can be problematic for classification algorithms, as they may be biased towards the majority class and have difficulty accurately predicting the minority class.

Oversampling is necessary in these situations because it helps balance the classes in the dataset, making it easier for classification algorithms to accurately predict all classes. This is done by generating additional synthetic observations for the minority class, which increases the overall number of observations in that class and makes the distribution of classes more even.

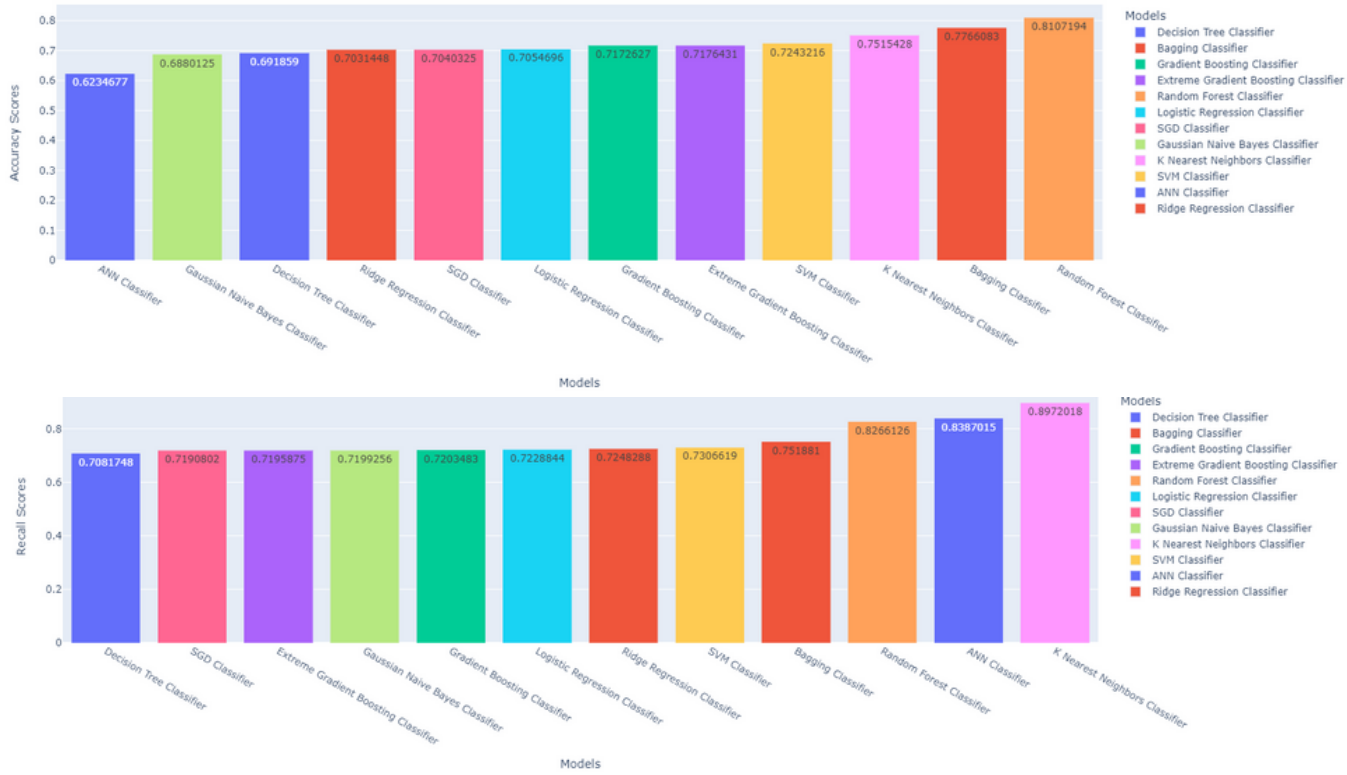
There are several benefits to using oversampling in classification tasks with imbalanced classes. First, it can improve the performance of the classification algorithm, as the increased number of observations in the minority class allows the algorithm to better learn the characteristics of that class. This can lead to more accurate predictions and higher overall performance.

Second, oversampling can help prevent bias in the classification algorithm, as the increased number of observations in the minority class reduces the likelihood that the algorithm will be biased towards the majority class. This can help ensure that the algorithm is able to accurately predict all classes, rather than just the majority class.

As shown previously, since the default indicators were imbalanced towards 0, sampling is required to increase or decrease the dataset size and subsequently improve model performance. Since, undersampling reduces dataset size and resulted in a loss of accuracy when models were trained, we have chosen to employ oversampling via SMOTE, which involves creating additional observations that are similar to the existing observations in the minority class, but are not identical. This can help the classification algorithm learn the characteristics of the minority class more effectively, and can lead to more accurate predictions. By interpolating between existing observations rather than simply copying them, SMOTE can create new observations that are similar to the original data, but that have some variation. This can help prevent the classification algorithm from overfitting to the synthetic observations and improve its ability to generalize to new data.

MODEL SELECTION

The dataset was trained using default parameters of multiple models initially, in order to check which model would be a good fit for the given dataset, sort of in a brute force approach. Of the 12 models that were trained and tested on the given dataset, Random Forest Classifier showed the best accuracy values, with k-Nearest Neighbors Classifier giving the highest recall score, as shown in the below figures.



RANDOM FOREST CLASSIFIER

A random forest classifier is a type of machine learning algorithm that uses a collection, or "forest," of decision trees to make predictions. A decision tree is a flowchart-like structure in which an internal node represents a feature or attribute, the branches represent the possible values of that attribute, and the leaves represent the class labels. In a random forest classifier, each tree in the forest is trained on a random subset of the data, and the final prediction is made by averaging the predictions of all the individual trees. This process is known as "ensemble learning," and it can often produce more accurate results than a single decision tree.

WHY RANDOM FORESTS?

One of the reasons that we went with Random Forest Classifier was it gave the best results on accuracy of predictions, when multiple models were evaluated on the training dataset and hence, we chose to further tune this model for improving its accuracy. Random forests are a good choice for binary data classification with high-dimensional data because they are able to handle large numbers of features and they are relatively robust to overfitting.

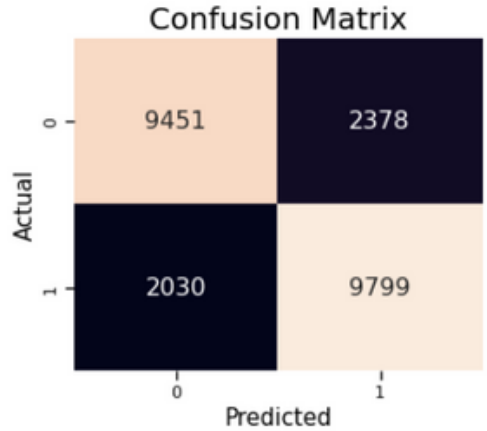
Also, they can handle missing values in the data. This is often a problem in real-world data, where it is common to have missing or incomplete information. A random forest classifier can handle missing values by using a method known as "imputation," where the missing values are replaced with estimates based on the available data. This allows the model to use all the available data to make predictions, even if some values are missing.

One advantage of a random forest classifier is that it can handle high-dimensional data, meaning data with a large number of features or attributes. Because each decision tree in the forest is trained on a random subset of the data, the overall model is able to capture the complexity of the data without overfitting, which is the problem of a model becoming too closely matched to the training data and losing its ability to generalize to new, unseen examples. Another advantage of the random forest classifier is that it can provide estimates of the relative importance of each feature in the data. This is useful in understanding the data and identifying the most important features that the model is using to make predictions. In addition, the random forest classifier is relatively easy to use and can be trained efficiently, even on large datasets. It is also robust to overfitting, meaning that it generally produces good results even when trained on noisy or incomplete data.

Another possible disadvantage of the random forest classifier is that it is sensitive to the hyperparameters used to train the model. These are the settings that control the overall structure and behavior of the model, such as the number of decision trees in the forest or the maximum depth of each tree. If the hyperparameters are not set properly, the model may not perform well, and finding the optimal values can require careful experimentation and tuning. This can be time-consuming and may require expert knowledge to get right. However, once the hyperparameters have been set properly, the random forest classifier can produce highly accurate and reliable results.

RESULTS

The Random Forest Model was built with multiple different parameters and the best one was identified using k-Fold Cross validation with k as 5. The results of the trained model on the test set (a random 20% set of fields from the pre-processed dataset) are shown below:



From the confusion matrix shown above, we can observe the following:

- 9451 predictions about an applicant would default was correct.
- 9799 predictions about an applicant would not default was correct.
- 2378 predictions about an applicant would default was incorrect. This is type 1 error.
- 2030 predictions about an applicant would not default was incorrect. This is type 2 error.

The interpretation of class-wise test results, shown on the right are given below:

- An accuracy score of 0.81 indicates that the majority of the predictions by the model whether an applicant would default or not were correct.
- The number of wrong predictions is measured by error. Error of 0.186 indicates that low proportion of predictions made by the model, whether an applicant would default or not were wrong.
- The ratio of correct predictions by the model to total number of actual correct predictions is sensitivity. Sensitivity of 0.79 and 0.83 indicates that 0.72 and 0.83 proportion of applicants defaulting out of total number of defaulting applicants were predicted correctly by the model.
- The ratio of failure predictions by the model to total number of failure predictions is specificity. Specificity of 0.82 and 0.79 indicates that 0.82 and 0.72 proportion of applicants not defaulting out of total applicants not defaulting were predicted correctly by the model.

```
Accuracy:
          0
0  0.813678
1  0.813678

Error:
          0
0  0.186322
1  0.186322

Sensitivity:
          0
0  0.798969
1  0.828388

Specificity:
          0
0  0.828388
1  0.798969
```

RESULTS

	precision	recall	f1-score	support
0	0.82	0.80	0.81	11829
1	0.80	0.83	0.82	11829
accuracy			0.81	23658
macro avg	0.81	0.81	0.81	23658
weighted avg	0.81	0.81	0.81	23658

The accuracy of predictions made by a classification algorithm is evaluated using a classification report. The interpretation of the classification report shown in the above figure, of the trained random forest model is given below:

- Precision is calculated by dividing the total number of true positives by the total number of positive forecasts. To put it another way, precision determines what percentage of forecasted positives actually are positive. Precision of 0.82 and 0.80 indicates there are very low false positives compared to true positives.

- Recall is true positive divided by true positive and false negative. It evaluates how well a model can foretell positive outcomes. Recall of 0.8 and 0.83 indicates the major predictions were made correct by model. It also indicates low false negatives compared to true positives.

- F1 score is the harmonic mean of precision and recall. F1 score closer to 1 is preferable as it indicates better precision and recall. F1 score of 0.81 and 0.82 indicates that the model have achieved high precision and recall.

- Weighted average and macro average for precision, recall and f1 score are observed to be similar.

SUMMARY

In summary, credit card default prediction is important because it can help credit card issuers manage their risk and support customers who are at high risk of defaulting on their payments. This can help protect the issuer's financial health and support customers in managing their debt.

In this project, data has been pre-processed post and models have been built, which have then been trained, to identify the best model, which has then been further tuned and cross-validated to better fit the given training data and make accurate prediction on the test data.

LEADERBOARD RESULTS

The test set was also pre-processed in the same way as the training dataset and was fed into the random forest built earlier. The summary of the results of the predicted outcomes from the model is shown below, along with the leaderboard rank of **38**, with a score of 50.05%.

```
print(y_pred.value_counts())
```

```
0    34764
1    12236
dtype: int64
```

Rank	Team Name	Score	Entries	Best Submission
38	Sixers Jump to me	50.05%	4	03:22 pm | December 15, 2022 History

SUMMARY

In summary, credit card default prediction is important because it can help credit card issuers manage their risk and support customers who are at high risk of defaulting on their payments. This can help protect the issuer's financial health and support customers in managing their debt.

In this project, data has been pre-processed post and models have been built, which have then been trained, to identify the best model, which has then been further tuned and cross-validated to better fit the given training data and make accurate prediction on the test data.