**Abstract**

The demand for loans from banks is constantly rising as a result of people's growing requirements, especially due to the current economic climate and rising cost of living. After determining an applicant's eligibility, which is a challenging and time-consuming process, banks normally handle the applicant's loan as to which they hope they are paid off with a short repayment time. With the implementation of machine learning tools that employ classification algorithms to anticipate the favoured loan applicants, the machine learning techniques are perfect for minimising human effort and facilitating effective decision-making in determining which applicant has a ‘paid off’ loan status type. The models applied in this project are *K-Nearest Neighbours* (KNN), *Logistic Regression* (LR*), Decision Trees* (DT) and *Random Forest Classification* (RFC). After the experimental analysis on the ‘*Loan Data 2016’* it was concluded that the best performing model was the *LR model* with an 89% accuracy, 85% R2 score (co-efficient of determination) and F1-scores of 84%-94% for each of the three target variable categories – further validated by a plot of the actual vs predicted loan status.

**Introduction**

Loans account for a large portion of bank profits. Despite the fact that many people are looking for loans – especially students - finding a legitimate applicant who will return the loan, and on time, is difficult. Choosing the right applicant is time consuming if the process is done manually, additionally many misconceptions may happen to select the genuine applicant. Therefore, the aim of this project is to create machine learning-based loan status prediction systems given the ‘Loan Data 2016’ dataset, which will automatically choose the ideal applicants – resulting in a significant reduction in the loan sanctioning period time for banks. Furthermore, this study is to assess the capabilities of several machine learning algorithms such as KNN, LR, DT, RFC to forecast the type of loan status, such as paid off, collection, and collection paid off – evaluating the best model by quantifying the classification report metrics such as accuracy, F1 and R2 scores.

The project's loan data can be accessible on Kaggle under the heading "Loan Data 2016." The dataset contains 500 rows and 11 columns with feature headers for paid off time, age, gender, education level, due date, and more.

**Objectives**

* Extensive EDA to explore insights on the existing dataset using Seaborn and PowerBI
* Predict the loan status type using a range of machine learning models on the 2016 ‘Loan Data’ dataset
* Use metrics such as accuracy,F1 and R2  to determine the best performing model

**Methodology**

As with all data, pre-processing is the first step. The collected data may contain missing values resulting in discrepancies. Pre-processing is necessary to improve outcomes and the algorithm's efficiency, for example, outliers should be eliminated, unnecessary columns dropped, formatting columns and nulls evaluated and replaced.

The steps taken for pre-processing in the Loan Data were as follows:

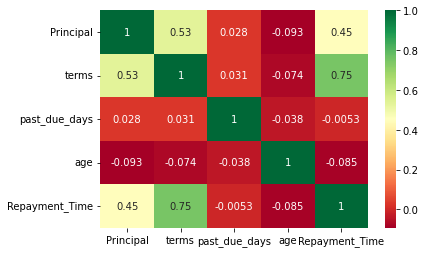
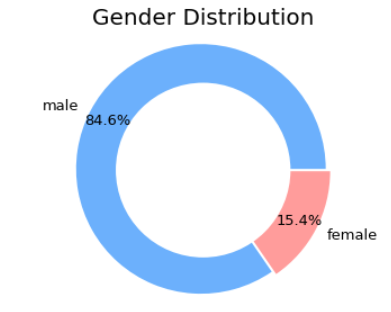
* Parsing through the paid off time column in excel – separating the date from the time, forming two new separate columns
* Importing the data into python by reading the file as a csv and using the Pandas library for data cleaning
* Dropping the Loan ID column as each value is unique and therefore has no impact as a feature column when using machine learning models
* Viewing the unique outputs of each column
* Formatting columns with dates using to\_datetime function so that the column was useable.
* Filling and replacing the nulls in the past due days columns with zero
* Creating new columns with the dates separated into months and weekday names
* Dropping the original columns used to create the new columns to have a tidier data frame

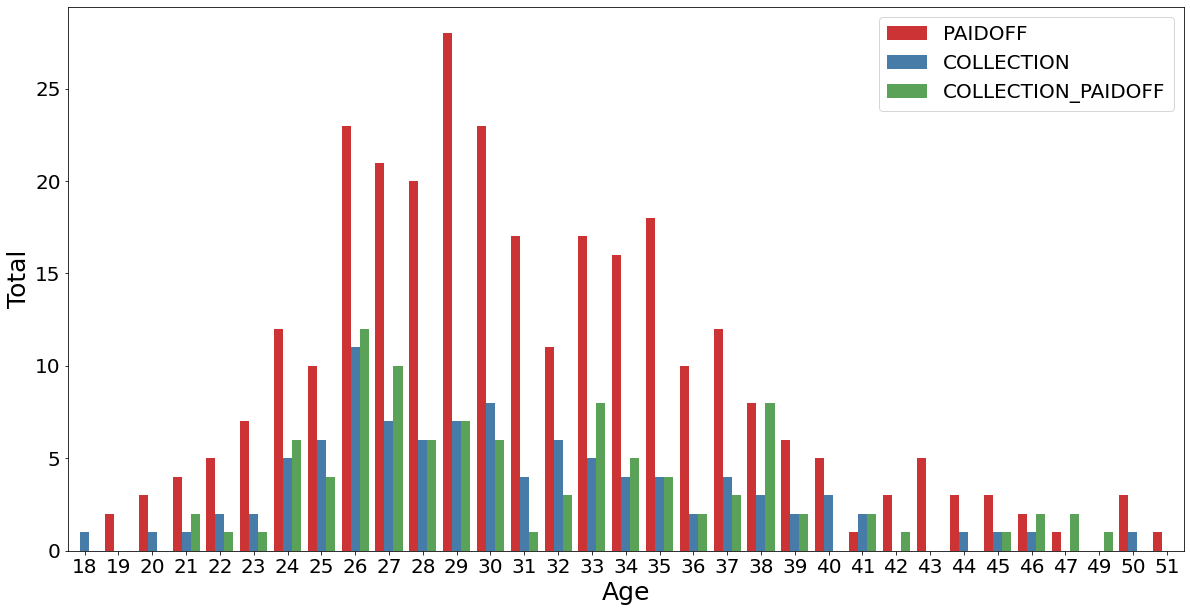
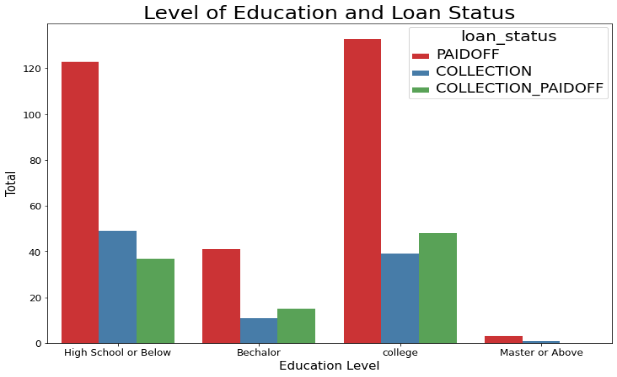
**Exploratory Data Analysis**

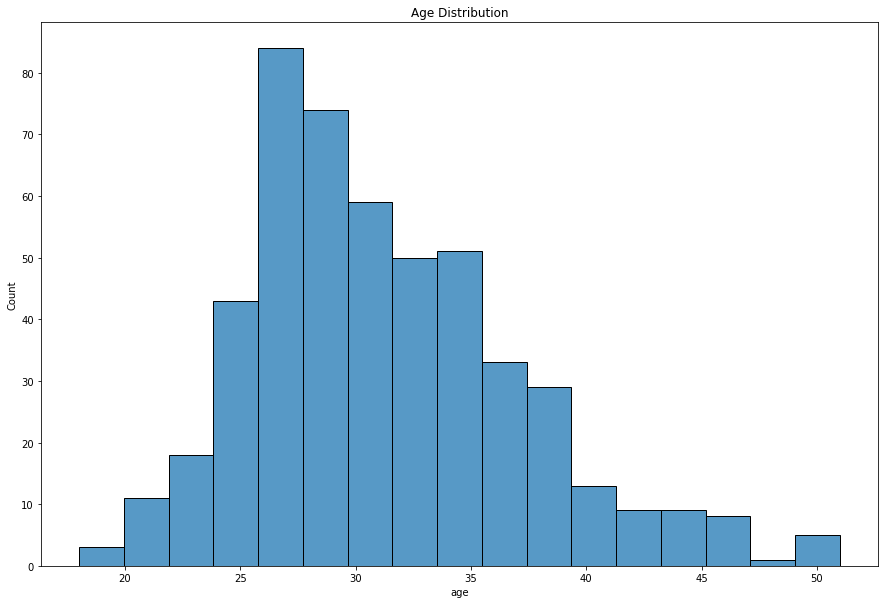
* 500 rows and 16 columns (after feature engineering and cleaning)
* Gender split – 85% Male & 15% Female
* Mean age was 31 years old
* Most common age to take out a loan was 26 years old
* Minimum age 18 and Maximum age 51 years old
* 29-year-olds most likely to have ‘Paid off’ loan status type
* Most common repayment time allocated by provider is 29 days
* Loans are always due in the months of Sept, Oct, Nov – most commonly October
* Loans always start in the month of September
* Given a month time frame to pay off the loan
* Loan most likely paid off on a Monday
* Principle amount / amount given from loan provider ranges from 300-1000

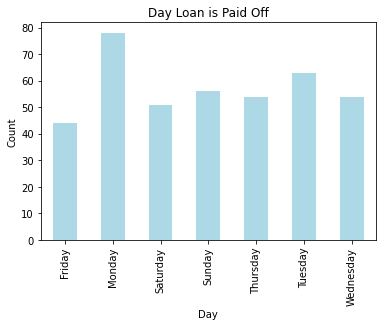
**Machine Learning**

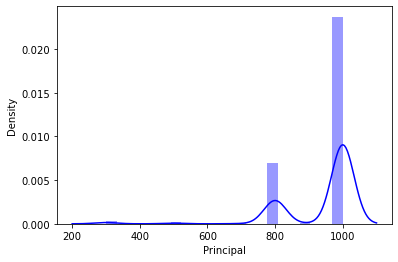
* Label encoding all the categorical columns in the dataset into a numeric form to allow labels to be machine-readable
* Train-Test-Split (80/20)
* Standardizing / Normalizing the dataset if required
* Balancing target variable using SMOTE
* Applying Machine Learning Algorithms (KNN, LR, DT, RFC)
* Quantifying ML metrics using classifications reports (determining accuracy and F1-scores)
* Plotting actual against predicted for the various ML techniques used to predict loan status type

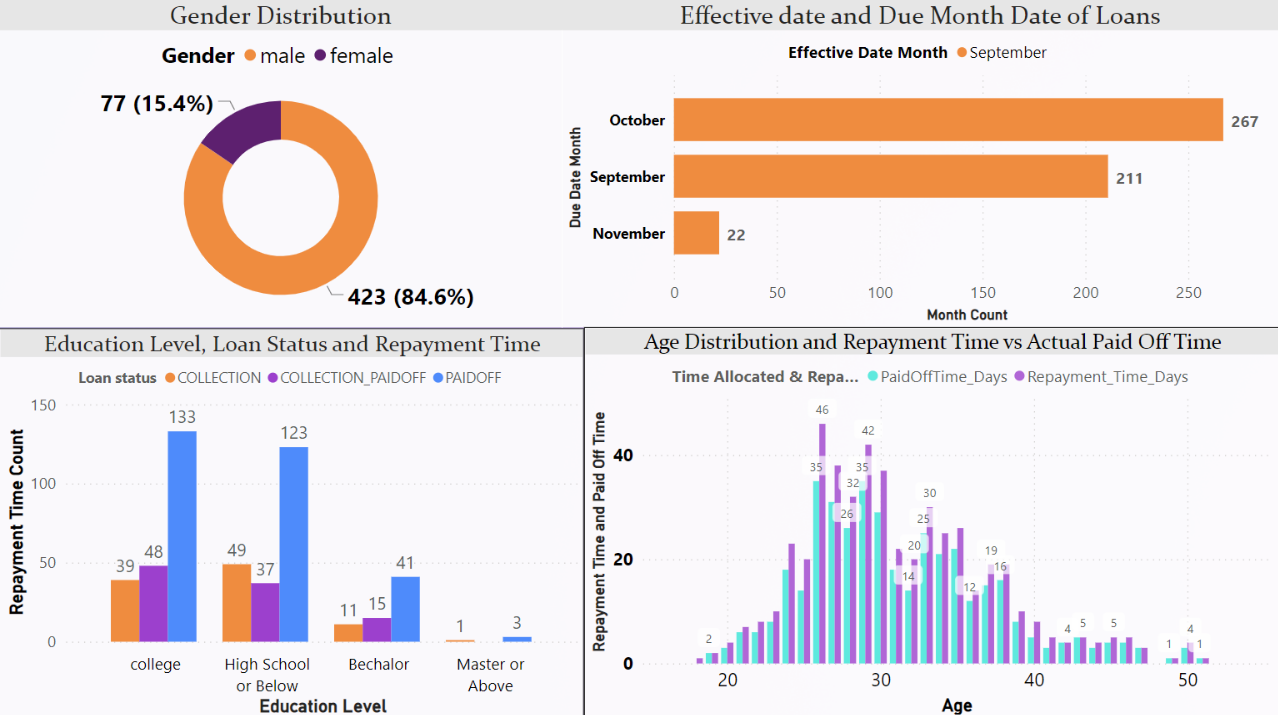
**EDA - Visuals**







****

**PowerBI Dashboard**

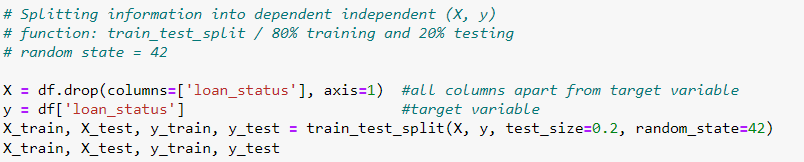
The PowerBI dashboard above displays four separate graphs, including one on the gender distribution and another for the age distribution with the repayment vs. actual paid-off times , as well as data on the effective and due date months and finally the education level against the loan status type.

**Machine Learning**

Machine learning is a field within computer science focused on giving computers the capabilities to learn. The goal of machine learning is to create algorithms that can learn and make predictions based on data and feedback. An important characteristic of machine learning is that it is not explicitly programmed to follow certain decision rules to create results. Instead, it has the capability of creating those rules based on data and feedback.

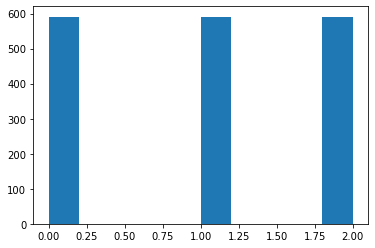
The most crucial phase of data analytics is feature engineering. It deals with formulating predictions and choosing the attributes that are used in training as well as aiding in modelling the dataset and comprehending it. An inaccurate or subpar prediction model may result from inadequate feature selection. The selection of the appropriate characteristics affects both accuracy and predictive power. It eliminates all the extraneous or pointless elements, if wrong features are selected, then even the good algorithm may produce the bad predictions. Therefore, feature engineering acts like a backbone in building an accurate predictive model.

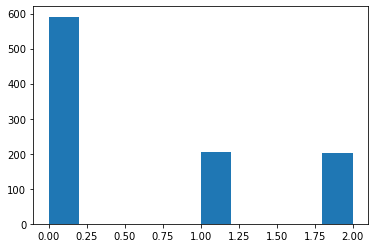
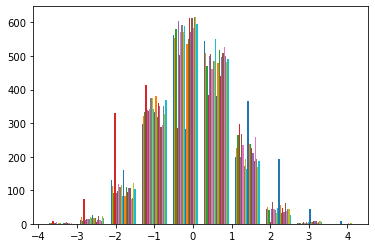
A method for assessing a machine learning algorithm's performance is the train-test split. It can be applied to issues involving classification or regression as well as any supervised learning algorithm. The process entails splitting the dataset into two subsets – training data and testing data. In this project a split of 80:20 was used, meaning 80% training and 20% testing. Below is an example code of splitting the data:

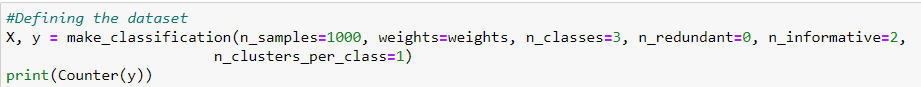


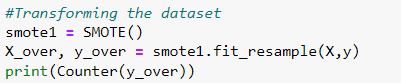
One common issue with all machine learning algorithms is over and underfitting. For Decision Tree, it means growing too large tree (with strong bias, small variation) so it loses its ability to generalise the data and to predict the output - therefore resampling is required.

**Resampling**

The data sets used in this research project were imbalanced, particularly the target variable (‘loan status’). To remove the imbalance the SMOTE (Synthetic Minority Oversampling Technique) method had been implemented. SMOTE works by generating new instances from existing minority cases that can be supplied as an input. After using SMOTE on the data set, it is concluded that resampling can have a positive effect, but it should not be applied without reservation. The visuals below depict the imbalance in the target variable and how the imbalance is rectified using SMOTE resampling:



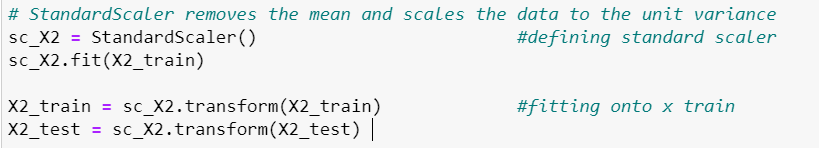
SMOTE resampling codes used:



**Standardisation / Normalization**

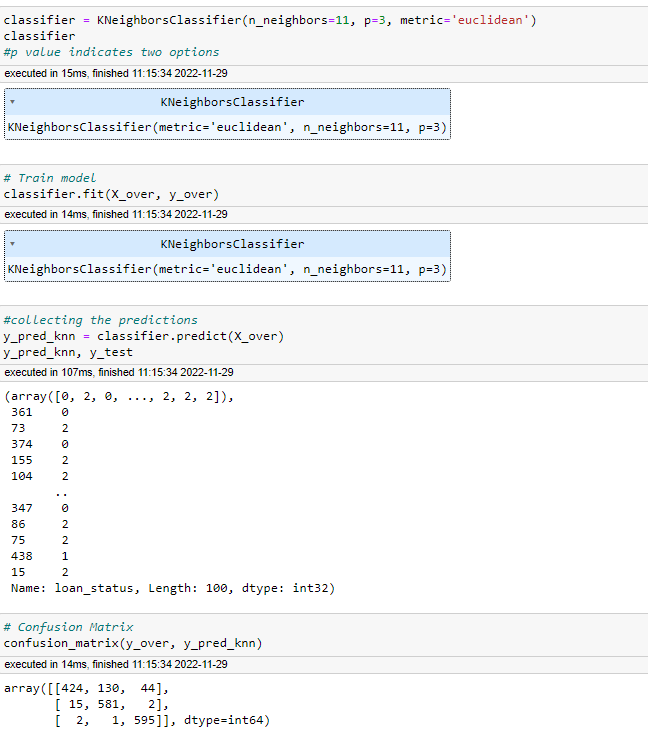
For proper data analysis, standardised data is necessary. A Standard Scalerensures that for each feature in the dataset the mean is 0 and variance is 1 - bringing all features to the same magnitude. This scaling doesn’t ensure any minimum and maximum values for the features. Any machine learning method that determines the separation between data points must use aspect Scaling. Similar to scaling, the objective of normalisation is to convert the values of the dataset's numeric columns to a standard scale without distorting variations in the value ranges. Data must be scaled to meet a conventional normal distribution as part of standardisation. A distribution with a mean of 0 and a standard deviation of 1 is known as a standard normal distribution.

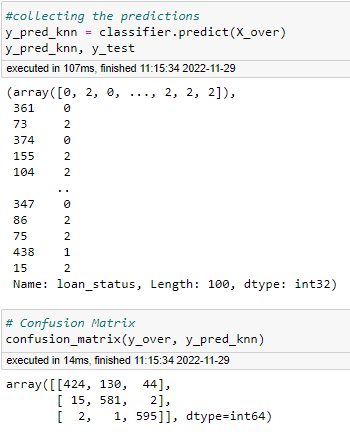
The code below shows the process of standardisation:

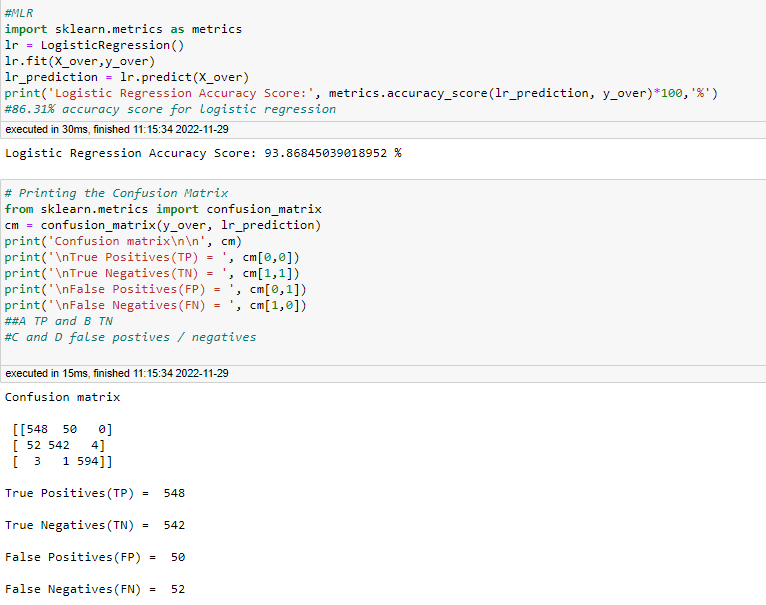


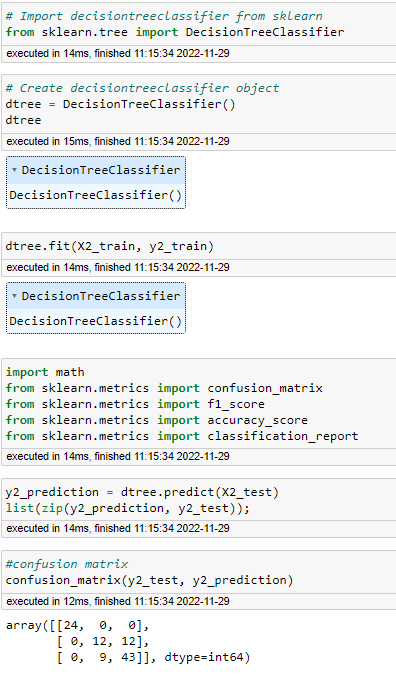
**Modelling**

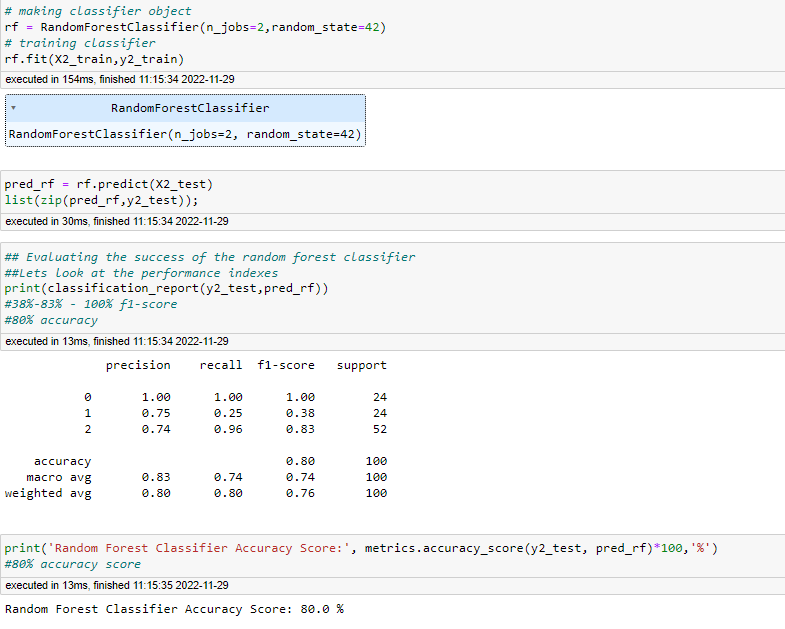
For the machine learning data modelling, the Logistic Regression algorithm will be utilised, along with Decision Tree, Random Forest Classification and K-Nearest Neighbours. Below are the codes for the various models:

KNN



Logistic Regression

Decision Trees

Random Forest Classifier

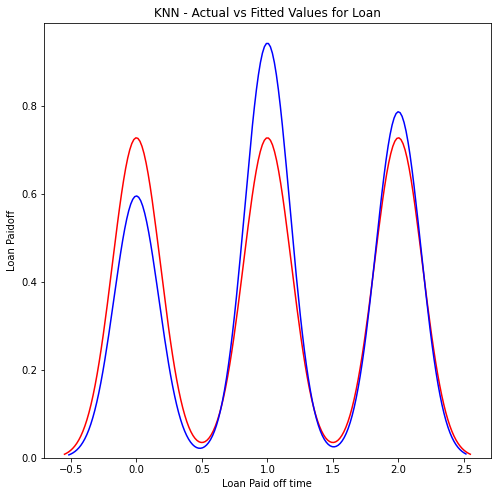
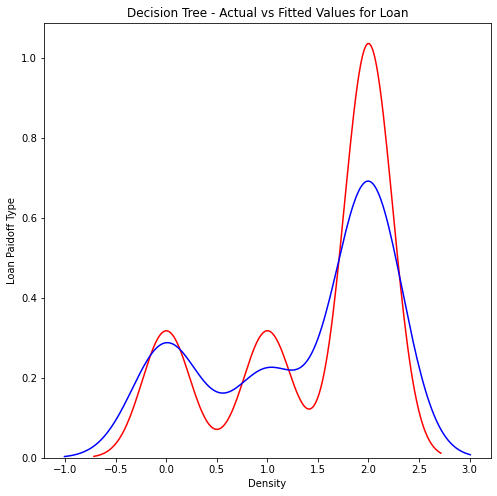
**Results**

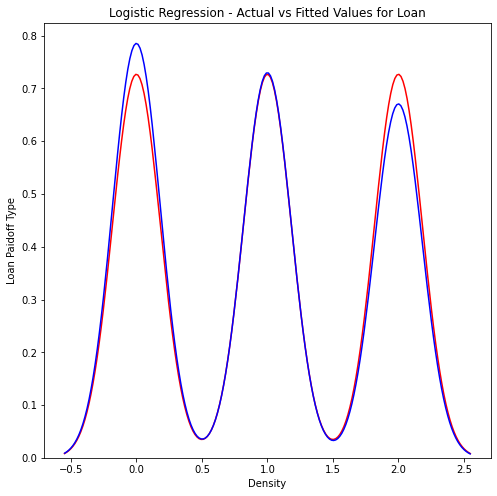
Visualisations representing Machine Learning results, including table with classification report metrics, R2 scores and a plot of actual vs predicted loan status type using KNN, LR and DT:

* 0 = Paid Off
* 1 = Collection
* 2 = Collection-Paid Off

|  |  |  |
| --- | --- | --- |
| Machine Learning Model | F1-Scores | Accuracy |
| KNN | 85% | 85% |
| Logistic Regression | 84%, 84%, 94% | 89% |
| Decision Tree | 100%, 52%, 82% | 80% |
| Random Forest Classifier | 100%, 38%, 82% | 80% |

|  |  |
| --- | --- |
| Machine Learning Technique | R2  Value |
| Logistic Regression | 85% |
| KNN | 64% |
| Decision Tree | 70.65% |





**Analysis & Conclusion**

The dataset is ambiguous regarding the type of loan data however it is reasonable to assume that this dataset targets student / young adults considering the heavy focus on the education level as well as the maximum principal amount being 1000. Additionally, all loan payments start in the month of September and are short term loans which could indicate the start of a semester – further reinforcing the possibility that the loan provider is a student loan provider. The target variable in this data was the ‘loan status’ column in which the three categories were – ‘paid off’ , ‘collection’ and ‘collection paid off’ these categories were encoded as ‘0’, ‘1, and ‘2’ respectively. Having applied 4 different ML algorithms, the LR model had the highest accuracy and F1-scores of 89% and 84-94% - this is further validated with a graph indicating the actual vs predicted values of loan status type with the plots closely overlapping each other as well as the highest co-efficient of determination (R2) score of 85%.

The machine learning strategies implemented in this project, particularly Logistic Regression models, can save the banking industry and its staff a significant amount of time when processing loan eligibility and requests from students / young professionals. This study has accurately shown how to forecast loan status types using machine learning algorithms on a very difficult dataset. It has been demonstrated that data pre-processing, a careful selection of dataset balancing strategies, and classification algorithms are all crucial for achieving the highest outcome.

To further enhance model performance on this crucial predication task, in future works, exploring more machine learning techniques such as Naïve Bayes Gradient and Gradient / AdaBoosting algorithms should be implemented which can also solve the problem which may even lead to more robust techniques with higher classification metrics.

**Reflection**

Ideally using a broader dataset that is not target towards students, with more feature columns that have a bigger significance on loan approvals such as credit scores, work employment history and dependencies should be used. This in turn will likely have a bigger impact in predicting the loan status and eligibility of an applicant – in which both the applicant and the bank staff will benefit from this. Since the data is old (2016) it can be assumed additional and more relevant columns have been included in the dataset as well as spelling errors and types in the outputs such as ‘Bechalor’ for the education level which should be ‘Bachelor’. If this project were to be repeated, the column names should be renamed as well as any typos in the categorical row outputs. Additionally, the principal amount currency units should be determined as no unit value was provided in the dataset – leading to ambiguity regarding loan amount. Finally, machine learning results should be validated with previous works found in articles to prove repeatability and accuracy in the experimental result findings.

**Appendix**

Dataset: <https://www.kaggle.com/datasets/zhijinzhai/loandata>

Jupyter Notebook Workbook - Attached as PDF in email

PowerBI – Available upon request

PowerPoint Slide Decks – Available upon request

Seaborn Viz Examples – Attached as PDF