TECH4DEV

LOAN PREDICTION MODEL

Group 2 mini project models an analysis of the borrower's eligibility/chance of being granted a loan

JUNE 20, 2021

GROUP 2 MINI REPORT

TECH4DEV WOMEN TECHSTERS FELLOWSHIP 2021
Learning Track - Data Science & AI

Group 2 - Judith Malepe, Caroline Ayieko, Omobolanle Adeyemi, Chege Jacinta, Purity Chepkurui, Oluwatosin Ehindero, Alozie Ijeoma, Nkechi Ijeoma, Clemence Ruhi

Table of Contents

List of Figures

1. Introduction

The current economic downturn has resulted with an increased need for consumer credit as people find themselves turning more to financial institutions for loans. One of the key factors which financial institutions use to assess the borrower's eligibility to qualify for credit is by creating and designing prediction models to provide an automatic assessment upon input of certain variables to model the borrower's chances of qualifying for the loan. This method is a less time-consuming process and quickens the loan shopping process. The main objective of the model is to provide an automation of the loan processing system for financial institutions. The model provides a binary response of the target variable, i.e., loan approval status. In this project, the borrowers' details are loaded from a dataset and a model generated to predict the loan approval status.

2. Methodology

The following figure represents the data science lifecycle flow diagram which was adopted to approach the project plan to realize the solution.

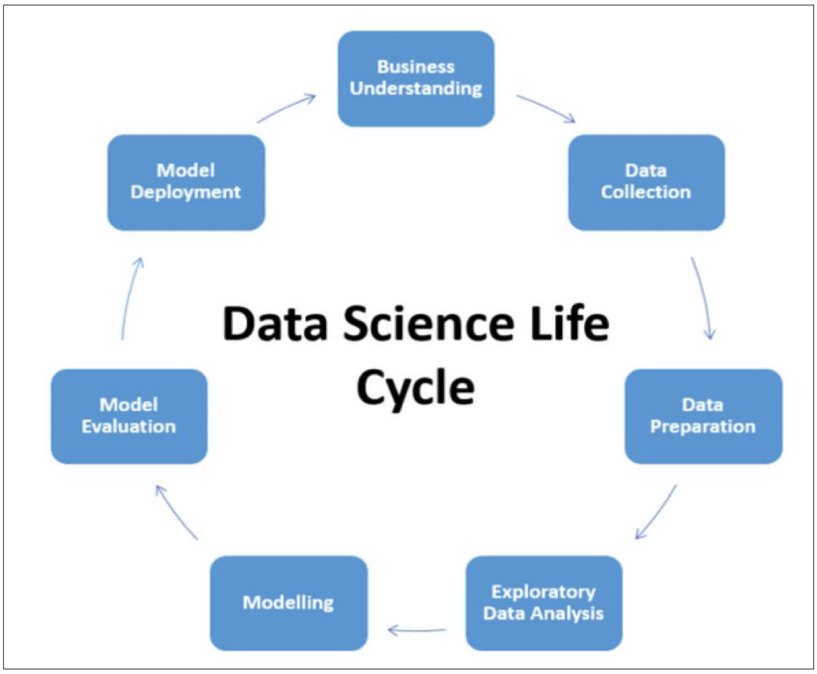


Figure 1: Data Science project Lifecycle [1]

2.1. Business Understanding

According to the Financial Institutions Centre, one of the bank's core business is enabling firms and households to cope with economic uncertainties. This can be achieved in different ways, amongst which, the flow of funds from the ultimate lenders to the ultimate borrowers is facilitated. The following diagram indicates the general business process used to apply for a loan.

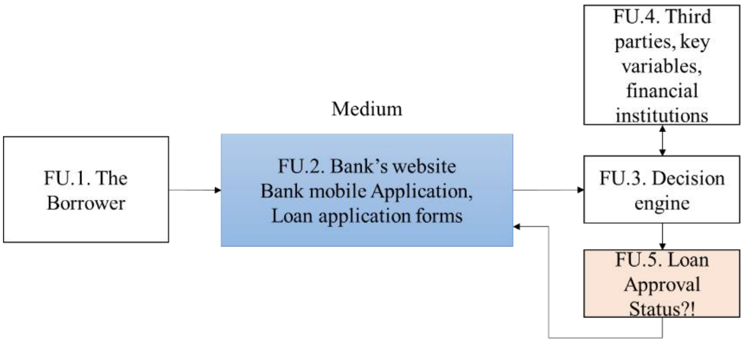

FU.4. Third parties, key variables, financial institutions
Medium
FU.1. The Borrower
FU.2. Bank's website Bank mobile Application, Loan application forms
FU.3. Decision engine
FU.5. Loan Approval Status?!

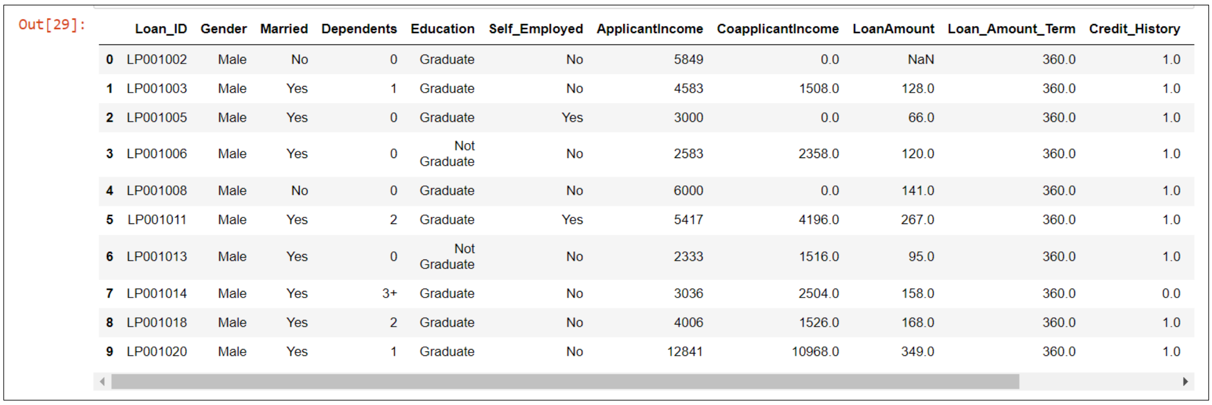
Figure 2: Business application process

Functional Unit (FU) 1 represents the customer who applied for credit/a loan. FU.2 represents the different mediums of application for a loan provided by the financial institution. FU.2 data is sent to FU.3, the control center where key variables are processed, and other third-party entities are consulted to obtain other customer data such as credit history. A decision is made, and a loan approval status determined.

As a result of the economic downturn, there is an increase in the number of credit applications which may slow down the decision-making process of financial institutions to carefully review the borrower's application. More information from the customer might be required by loan credit analysts to review the customer. This process can be time consuming and introduce a problem to the bank in accordance with the response time specifications stated. The bank requires a reliable and quick system to analyze the customers eligibility. The loan processing system must be accurate to reduce the risk of lending to a customer who is likely to default on their loan payments.

2.2. Data Acquisition

A dataset which contains the borrower's application details was obtained from Kaggle and used to design the model for automating the loan approval status. The dataset consisted of the training dataset (for generating the model) and testing dataset (for testing and assessing the model). The two datasets are csv files and were loaded onto the Jupiter notebook which is the analytic environment we will be using for building the loan prediction model (see figure below).


Out[29]:

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	1.0
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0
4	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0
5	LP001011	Male	Yes	2	Graduate	Yes	5417	4196.0	267.0	360.0	1.0
6	LP001013	Male	Yes	0	Not Graduate	No	2333	1516.0	95.0	360.0	1.0
7	LP001014	Male	Yes	3+	Graduate	No	3036	2504.0	158.0	360.0	0.0
8	LP001018	Male	Yes	2	Graduate	No	4006	1526.0	168.0	360.0	1.0
9	LP001020	Male	Yes	1	Graduate	No	12841	10968.0	349.0	360.0	1.0

Figure 3: Dataset loaded on Jupiter notebook environment

2.2.1. Data preparation

The dataset was audited by cleaning the data to produce a higher quality of information and removes errors from the data. The cleaned data is used as an input to the model. This process control method is carried out to eliminate errors; it contributes proportionally to the accuracy of the data, thereby ensuring accurate insights are captured.

2.3. Exploratory Data Analysis

The following figure presents data visualization for the married, self-employed, gender and credit history variables.

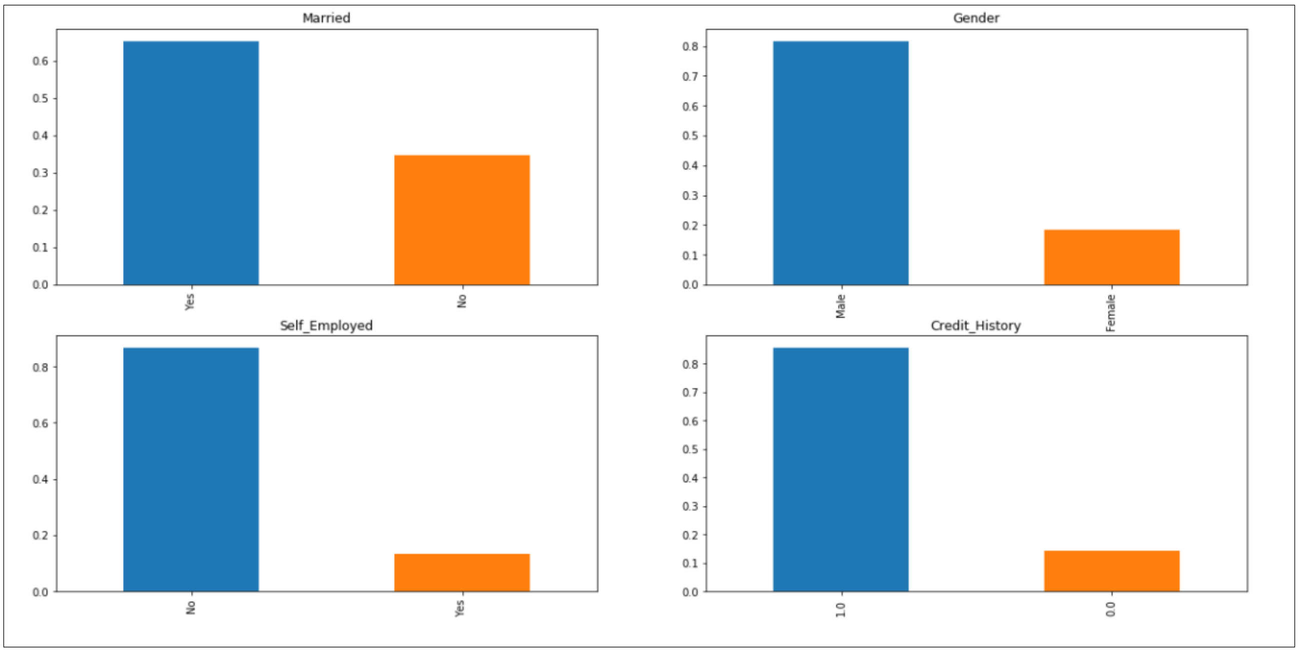


Figure 4: Training data composition for married, self-employed, gender and credit history variables

The following graphs represents the data composition of people with dependents, percentage of those who are graduates or not, and the representation of the property area where the borrowers come from.

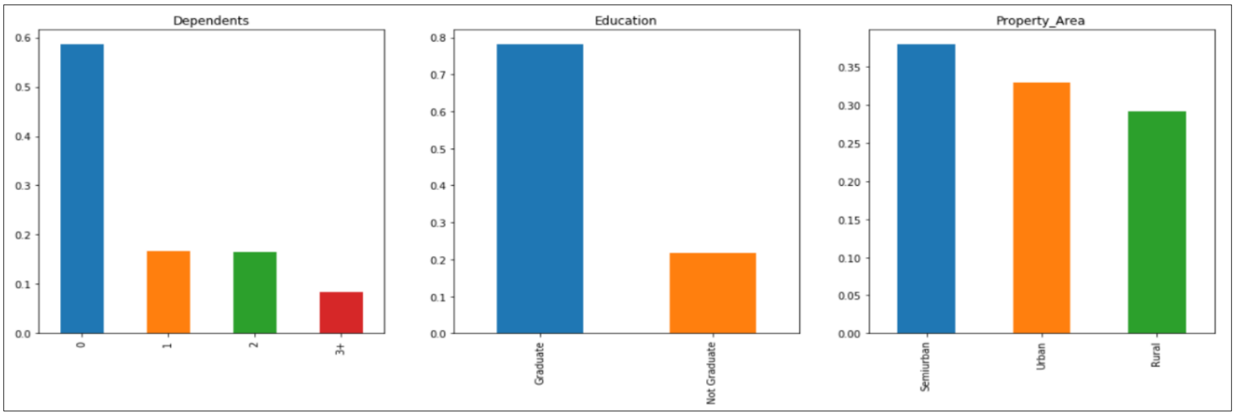


Figure 5: Training data composition based on dependants, education, and property area variables

The following figure represents the loan approval status for the married, gender and dependents variables.

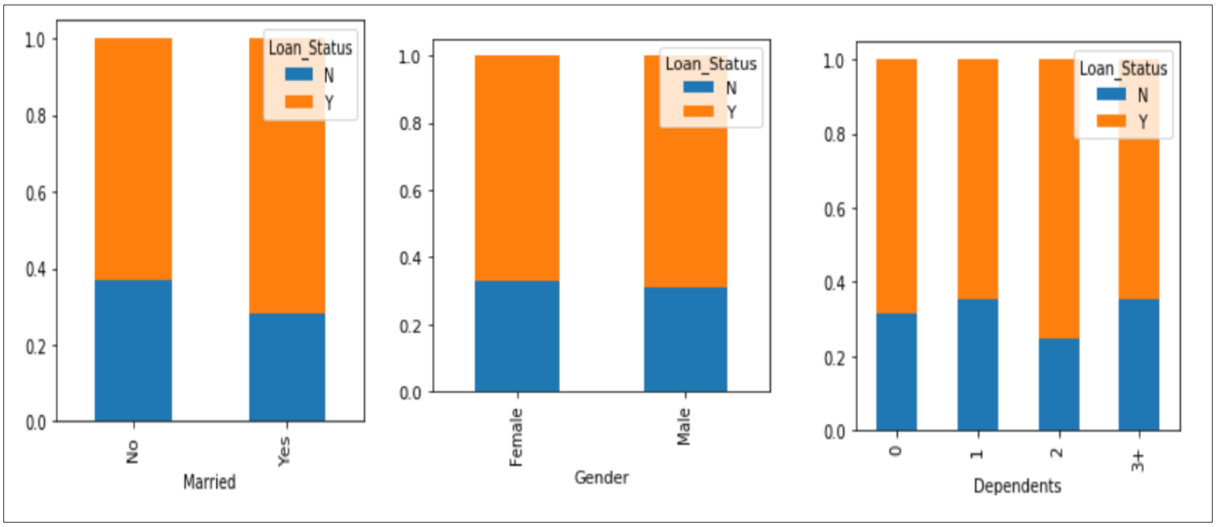


Figure 6: loan approval status using the married, gender and dependents variables

Married: The leftmost graph indicated in the above figure shows that a higher percentage of married people are more likely to qualify for a loan than those who are not. There is 62% chance of qualifying if not married, and 75% chance if married.

Gender: Males and females have an almost equal chance of qualifying for a loan. Females have 65% chance while males have 70% chance over females.

Dependents: People with higher number of dependents are less likely to qualify for a loan while those with less dependents and/or no dependents have higher chances of being eligible for the loan. The graph indicates 70% chance for those with no dependents, 65% chance for those with 1 dependent and 3+ dependents.

The distribution analysis of the borrower’s income and the loan amount applied for is presented below.

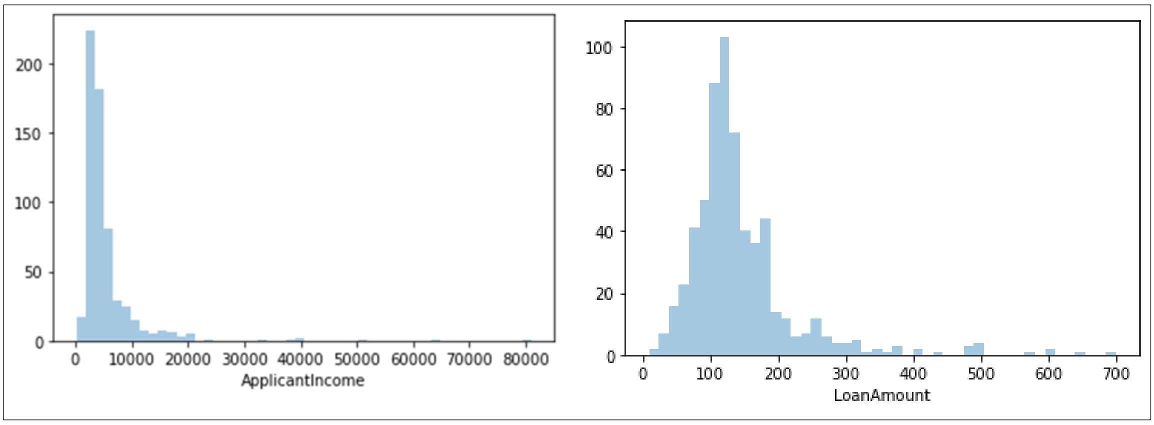


Figure 7: Distribution graphs for the applicant income and loan amount applied

The loan amount application and applicant’s income are a positively skewed distribution graph. This means the mean and median of the dataset is not equal [2]. Thus, statistical techniques cannot be applied to the data, but rather, logarithms and quantile regression techniques should be used to make a more accurate prediction model.

The below figure shows a comparison of the applicant's income between the graduates and non-graduates.

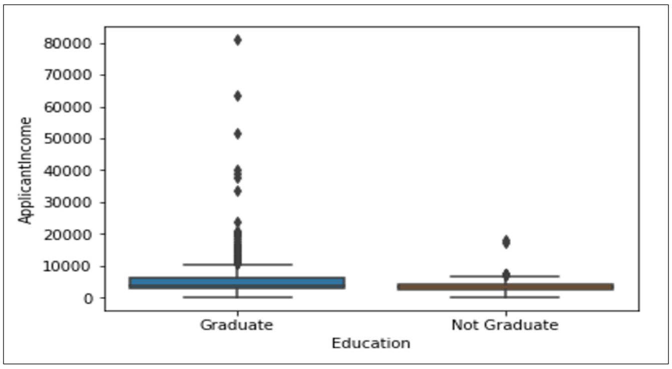


Figure 8: Outliers for the education variable

Graduates have more outliers, hence people with more income are most likely to be graduates and educated.

2.4. Data modelling

The data modelling process indicated below deploys machine learning approach, and the process is indicated below. This approach was used to build and generate the loan prediction model.

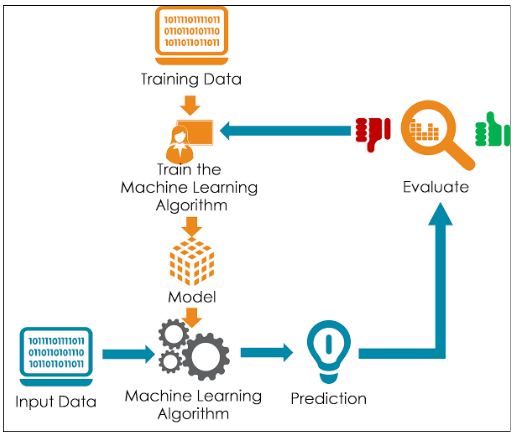

Training Data
Train the Machine Learning Algorithm
Evaluate
Model
Input Data
Machine Learning Algorithm
Prediction

Figure 9: Modelling using Machine learning approach [3]

2.4.1. Machine learning methods

There are five machine learning algorithms which can be used to learn the target function that maps the input variables into output variables. These are:

- Linear Regression
- Logistic Regression
- Linear Discriminant Analysis
- Decision Trees
- Naïve Bayes
- Random Forest

Linear regression uses statistics to model the prediction. It uses an equation of a straight line that best fits the linear relationship of the input variables and output variables.

Logistic regression is the method commonly used for binary classification problems, i.e., problems with two class values. The method uses a nonlinear function called the logistic function to transform the output.

Linear discriminant analysis is commonly limited to two-class classification problems. The representation of the linear discriminant analysis consists of the data statistics for each class. The technique is best suited for removing outliers.

Decision tree method is represented by a binary tree. Each node represents a single input variable and a split point on that variable (assuming the variable is numeric).

Naive Bayes model comprises of the probability for each class and the conditional probability of each class given the input value.

Random forest uses decision trees to create suboptimal splits by introducing randomness.

2.4.2. Model development

As indicated by our positively skewed distribution graphs for loan amount applied and applicant income when exploring data, logarithms and quantile regressions are better suited for our dataset. The logistic regression method was chosen for assembling the model.

2.4.3. Feature engineering

Feature importance was used for getting rid of variables which are unnecessary to improve model accuracy by selecting features with a greater impact.

The bar graph below indicates the feature importance of the model.

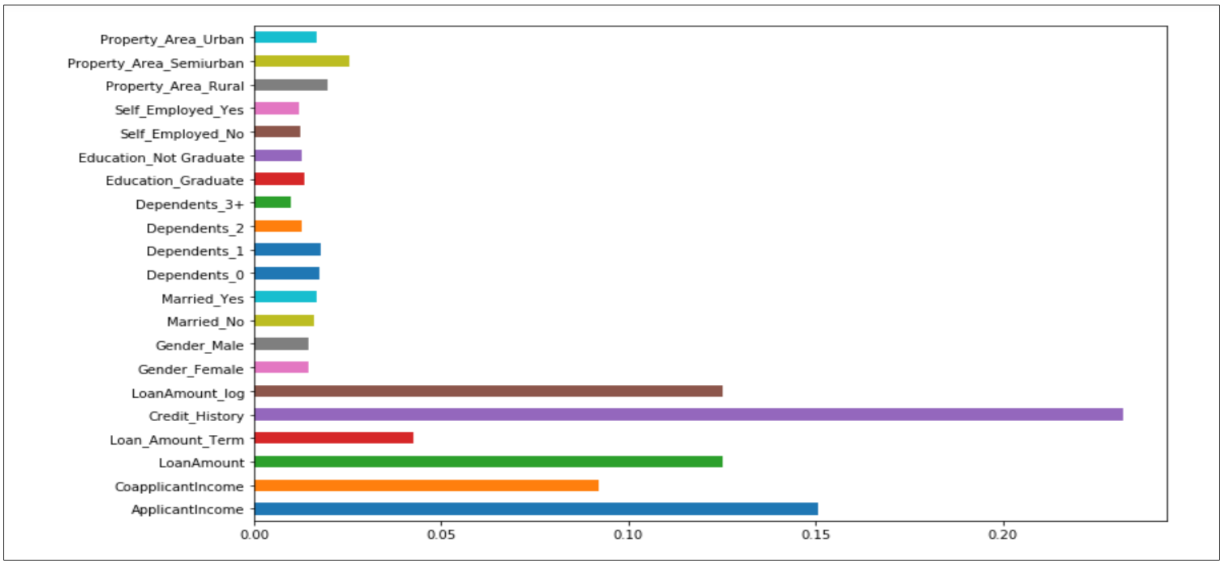


Figure 10: Feature importance

From the figure above, the long bars indicate greater significance than small bars.

Feature importance variables for our model:

- Loan amount
- Credit history
- Co-applicant income
- Applicant income
- Loan term

2.4.4. Model audit/validation

A logistical regression method to determine the accuracy of the model with another method (random forest), was used to validate the accuracy of the model.

- When using logistic regression model, the accuracy obtained was 78% and using random forest, the accuracy was 77%
- Upon deployment of important features, the model was retrained, and the accuracy improved to 81% from 77% for random forest and for logistic regression, the accuracy improved from 78% to 83%.

3. Model deployment: Results & findings

The graph below represents the loan approval chances against loan amount term.

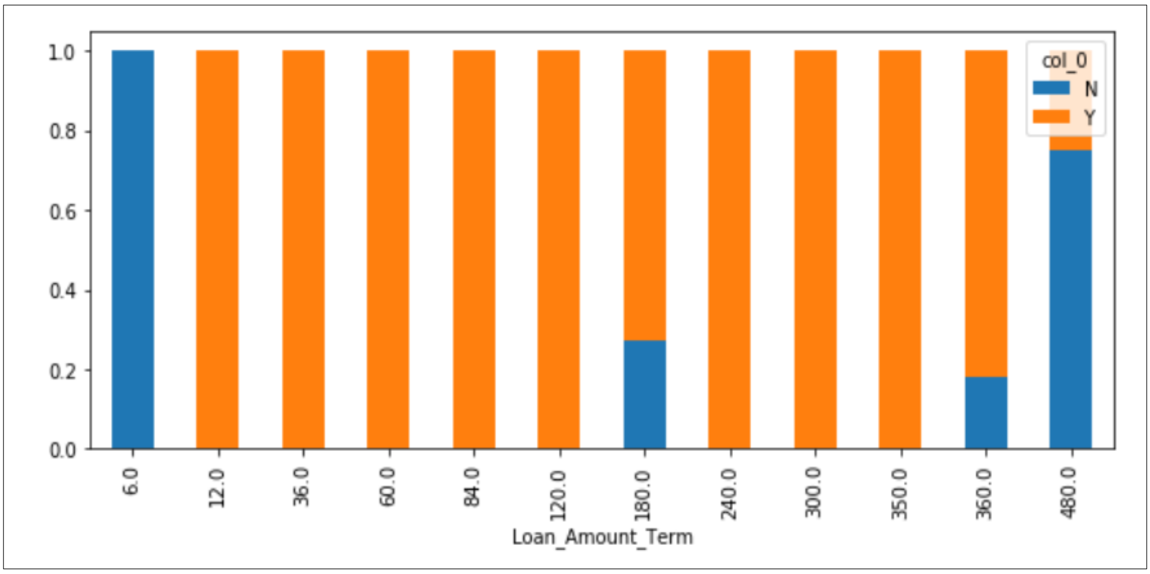


Figure 11: Loan approval results for different loan periods

From the above figure, a loan term between 12 - 120 months and 240 – 360 months has a 100% chance of a qualifying applicant than for a period less than a year, or more than 480 months.

The following graph is a loan approval status vs credit history bar graph.

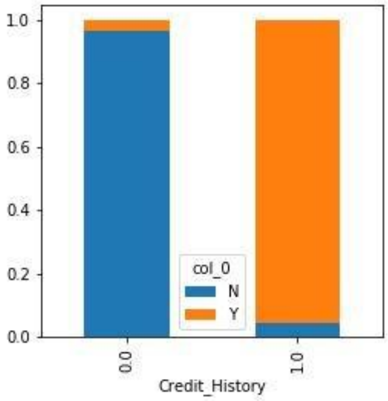
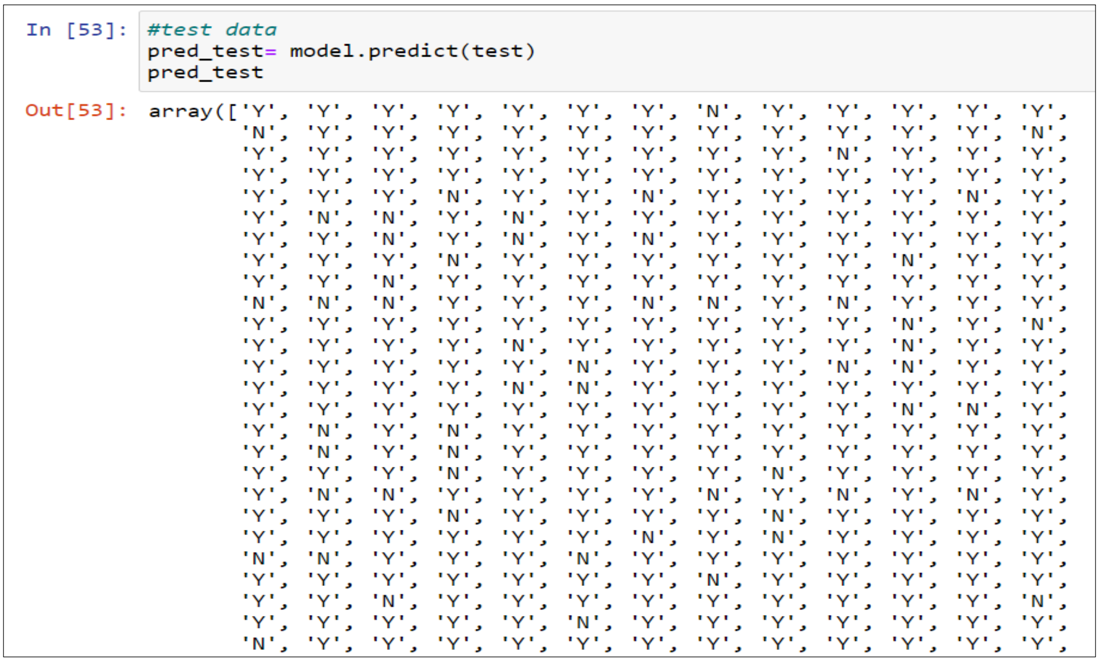


Figure 12: Loan approval status vs credit history graph

It can be observed from the above graph that people with no credit history are 95% guaranteed a decline in qualifying for a loan compared to those who possess a credit history.

The following figure deploys the model to predict the applicant approval status using the test data.

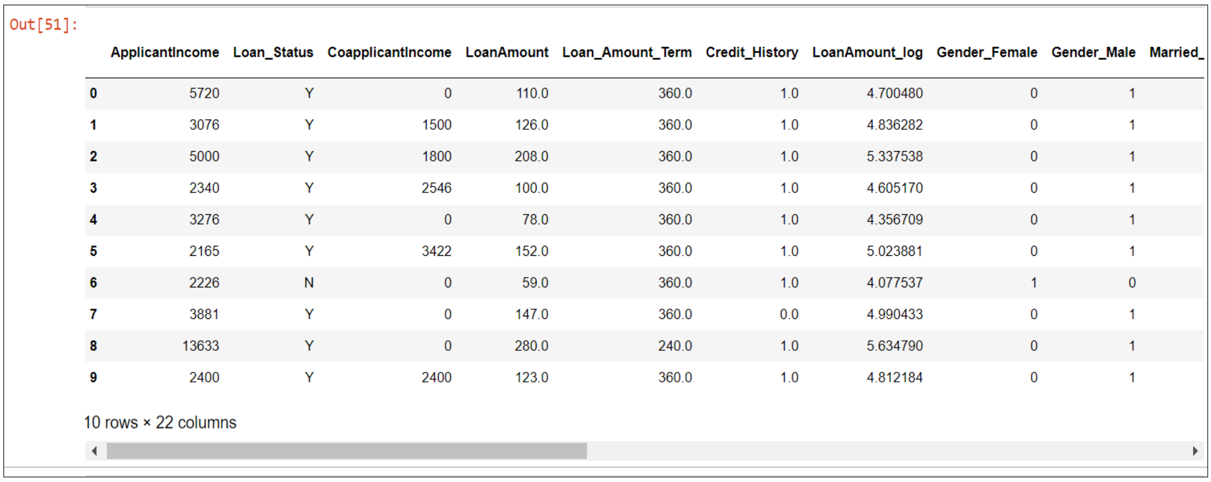


```
In [53]: #test data
         pred_test= model.predict(test)
         pred_test

Out[53]: array(['Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y',
                'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'N',
                'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y',
                'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y',
                'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'N', 'Y',
                'Y', 'N', 'N', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y',
                'Y', 'Y', 'N', 'Y', 'N', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y',
                'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y',
                'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y',
                'N', 'N', 'N', 'Y', 'Y', 'Y', 'N', 'N', 'Y', 'N', 'Y', 'Y', 'Y',
                'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'N',
                'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y',
                'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'N', 'N', 'Y', 'Y',
                'Y', 'Y', 'Y', 'Y', 'N', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y',
                'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'N', 'Y',
                'Y', 'N', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y',
                'Y', 'N', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y',
                'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y',
                'Y', 'N', 'N', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'N', 'Y', 'N', 'Y',
                'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y',
                'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'N', 'Y', 'Y', 'Y', 'Y',
                'N', 'N', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y',
                'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y',
                'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'N',
                'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y',
                'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y',
```

Figure 13: Model Output using test data

The test dataset snapshot of the model output is shown below with an entry for loan approval status presented. It can be noted that the declined loan applicant has no credit history.



Out[51]:

	ApplicantIncome	Loan_Status	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	LoanAmount_log	Gender_Female	Gender_Male	Married_
0	5720	Y	0	110.0	360.0	1.0	4.700480	0	1	
1	3076	Y	1500	126.0	360.0	1.0	4.836282	0	1	
2	5000	Y	1800	208.0	360.0	1.0	5.337538	0	1	
3	2340	Y	2546	100.0	360.0	1.0	4.605170	0	1	
4	3276	Y	0	78.0	360.0	1.0	4.356709	0	1	
5	2165	Y	3422	152.0	360.0	1.0	5.023881	0	1	
6	2226	N	0	59.0	360.0	1.0	4.077537	1	0	
7	3881	Y	0	147.0	360.0	0.0	4.990433	0	1	
8	13633	Y	0	280.0	240.0	1.0	5.634790	0	1	
9	2400	Y	2400	123.0	360.0	1.0	4.812184	0	1	

10 rows × 22 columns

Figure 14: Model output data

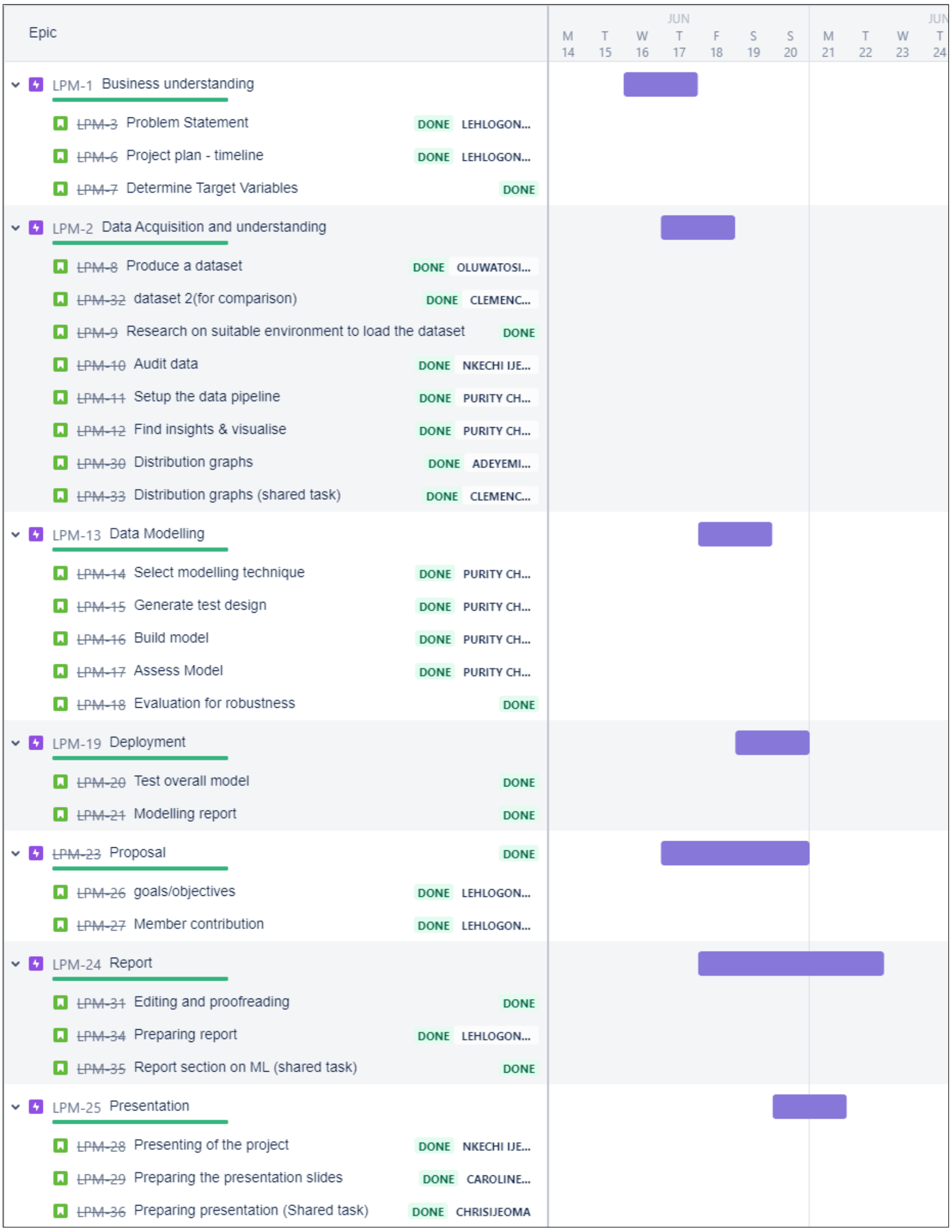
4. Conclusion

An automated loan approval prediction model was created and used to check the eligibility of the borrower's chances of qualifying for a loan.

The target variable is the loan approval status. Two methods were used in verifying the accuracy of the model and for benchmarking. The model accuracy was 78% using the logistic regression method and 77% using the random forest method. After variable selection was performed, the data accuracy significantly improved. The accuracy further improved to 83% for the logistic regression method and 81% for the random forest method. From the validation test dataset results, it can be noted that every person must at least have a credit history to become eligible for loan approval.

Recommendations: during the project presentation with the facilitator, it was realised that prediction models require continuous improvement and testing to improve model performance. Other factors which can further improve the model are deploying model assembly fitting, fine tuning (parameter and hyper-parameter tuning) the model to produce a more reliable prediction model which will ensure that business objectives are met while reducing the risk of borrowers defaulting on their payments.

A summary of the roadmap which was utilized to execute the tasks is shown below



Epic			
LPM-1	Business understanding		
~~LPM-3~~	Problem Statement	DONE	LEHLOGON...
~~LPM-6~~	Project plan - timeline	DONE	LEHLOGON...
~~LPM-7~~	Determine Target Variables	DONE	
LPM-2	Data Acquisition and understanding		
~~LPM-8~~	Produce a dataset	DONE	OLUWATOSI...
~~LPM-32~~	dataset 2(for comparison)	DONE	CLEMENC...
~~LPM-9~~	Research on suitable environment to load the dataset	DONE	
~~LPM-10~~	Audit data	DONE	NKECHI IJE...
~~LPM-11~~	Setup the data pipeline	DONE	PURITY CH...
~~LPM-12~~	Find insights & visualise	DONE	PURITY CH...
~~LPM-30~~	Distribution graphs	DONE	ADEYEMI...
~~LPM-33~~	Distribution graphs (shared task)	DONE	CLEMENC...
LPM-13	Data Modelling		
~~LPM-14~~	Select modelling technique	DONE	PURITY CH...
~~LPM-15~~	Generate test design	DONE	PURITY CH...
~~LPM-16~~	Build model	DONE	PURITY CH...
~~LPM-17~~	Assess Model	DONE	PURITY CH...
~~LPM-18~~	Evaluation for robustness	DONE	
LPM-19	Deployment		
~~LPM-20~~	Test overall model	DONE	
~~LPM-21~~	Modelling report	DONE	
~~LPM-23~~	Proposal	DONE	
~~LPM-26~~	goals/objectives	DONE	LEHLOGON...
~~LPM-27~~	Member contribution	DONE	LEHLOGON...
LPM-24	Report		
~~LPM-31~~	Editing and proofreading	DONE	
~~LPM-34~~	Preparing report	DONE	LEHLOGON...
~~LPM-35~~	Report section on ML (shared task)	DONE	
LPM-25	Presentation		
~~LPM-28~~	Presenting of the project	DONE	NKECHI IJE...
~~LPM-29~~	Preparing the presentation slides	DONE	CAROLINE...
~~LPM-36~~	Preparing presentation (Shared task)	DONE	CHRISIJEOMA

Figure 15: Project timeline/Roadmap

5. References

[1]. https://medium.com/swlh/complete-life-cycle-of-a-data-science-machine-learning-project-13df81bbd8eb

[2]. Emeto W, Statistical modelling notes, ‘*Tech4Dev-DataScience Learning track*’, May 2021.06.23

[3]. https://intellipaat.com/blog/tutorial/data-science-tutorial/modeling-the-data/

[4]. Analytic environment - Jupiter Notebook

[5]. Testing and training data - https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset