Business Process Analytics Project Report

Loan Analysis: Understanding the Client and Business

EXECUTIVE SUMMARY

When loan company or the banking team are trying to decide whether the user is eligible for loan or not, a lot of loan company paying too much attention or even only focus on the parameter like credit history. They often ignore the other parameter or information like gender, self-employed etc. All these have become the root cause of uncertain credit risk. Failed to really understand the actual situation of the borrower, even mistrusting the credit information fabricated by the customer, lays hidden risks for the substantial credit risk. Therefore, several aspects (variables) are taken into account while approving a loan. Predicting whether a borrower will default on loan is a significant concern of most financial institutes. If the lender is too strict, fewer loans get approved, which means there's less interest to collect. But if they're too lax, they end up approving loans that default.

Among all industries, insurance domain has the largest use of analytics & data science methods. This dataset is about the company that wants to automate the loan eligibility process based on customer details provided while filling online application. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers.

We have used second-hand data from Kaggle. The data has 614 rows and 13 columns. Data set details are Loan ID, Gender, Marital Status, Education, Self-employed, Number of Dependents, Applicant Income, Coapplicant Income, Loan Amount, Loan Amount Term, Credit History, Property Area, Loan status.
In this study, loan behavior was analyzed using logistic regression machine learning model. Machine learning model was trained to predict for loan repayment. Machine learning model was evaluated via K-fold classification technique, confusion matrix and correlation matrix. Feature Importance was performed to identify the most useful variables in the dataset.

PROBLEM STATEMENT

As the loan company spend too much time looking into the variables like credit history, and at the same time ignore other parameter's importance, many companies failed to really understand borrower's actual situation and have to take the uncertain credit risk. Our dataset includes parameter Gender, Marital Status, Education, Self-employed, Number of Dependents, Applicant Income and Credit History. However, we want to drop the parameter credit history as it plays really important role in the model. As a result, we will be able to use those model's output to understand others parameter's importance. The loan company or the banking team will then realize and start to pay some attention to other variable which is also important to the loan decision making process. Therefore, we have provided details below regarding our problem-solving approach, explaining the output of our analysis.

SUMMARY OF PROBLEM-SOLVING APPROACH

We have used logistic regression models, Descriptive Analysis, Basic Statistics, Missing Value Handling, Train Test Split, Model Evaluation, Model Analysis, K-Fold Cross Validation, Confusion Matrix, Classification Report, Correlation Matrix and Feature Importance to do some experimental among those loan process parameters. By comparing those different models' output, we can determine which parameter is important in the loan approval decision making process. The solution is that banks would give loans to only those customers who are eligible so that they can be assured of getting their money back.

MAJOR PROJECT RESULTS AND RECOMMENDATIONS

Recommendations:
To determine whether a customer qualifies for the loan, can be a complicated preposition for banks. Banks need to determine whether a customer qualifies for the loan based on the collected information.

With a high debt-to-income ratio, make the application more appealing by considering the addition of collateral or a cosigner. Collateral should be something of significant value that is at least comparable to the size of the loan. Collateral in our dataset is Property Area whose values are urban, semi-urban and rural.
The technical limitations of legacy lending systems reduce the lender's ability to replace manual steps with automated decisions. They also make it impossible to integrate alternative data sources that enable lenders to make more informed, accurate lending decisions.

1.0 IMPROVEMENT OPPORTUNITY: DEFINE PHASE

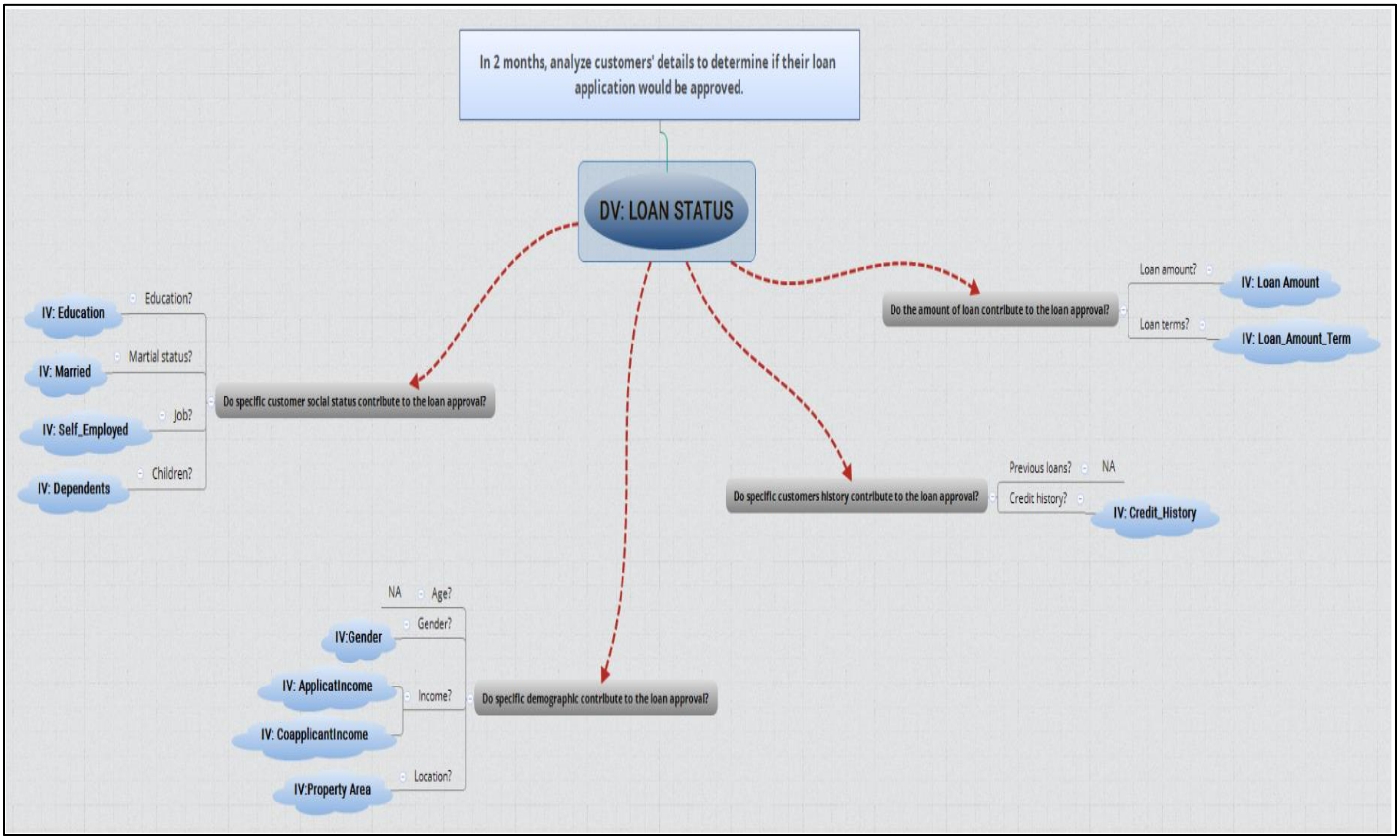
1.1 Problem Statement/Discussion of the process being examined:

The company deals with distribution of loans. Customers first apply for loan and after that the company validates the customer eligibility for loan. However, doing so manually is time consuming. Hence it wants to automate the loan eligibility process (real-time) based on customer information.

So, the key is to identify the factors or customer segments that are eligible for taking loan.
The immediate concern is how will the company benefit if we give customer segments. The solution is that banks would give loans to only those customers who are eligible so that they can be assured of getting their money back. Hence, more the accurate is the prediction of eligible customers, more beneficial it would be for the company.

1.2 Type of problem:

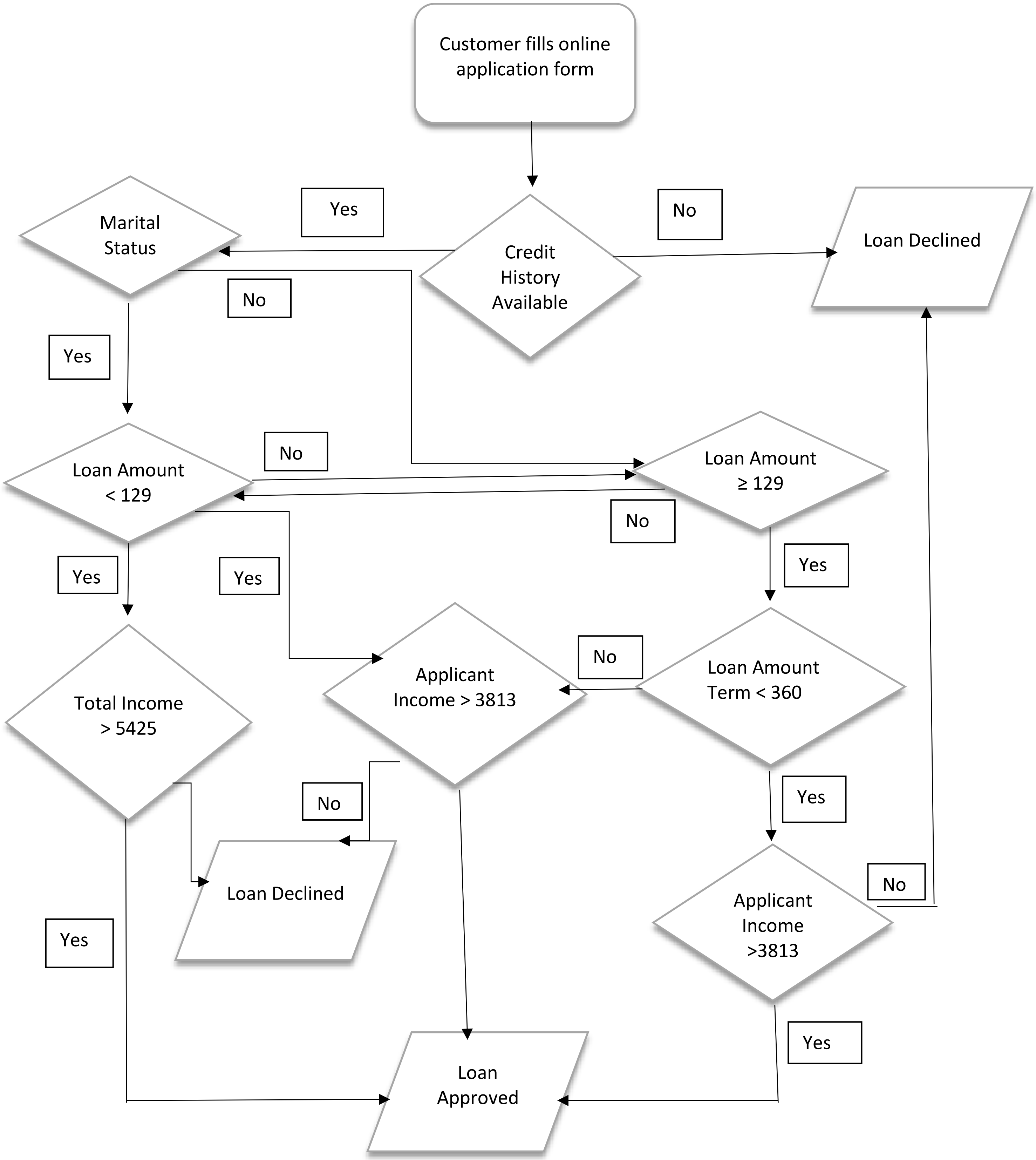
This is a clear classification problem as we need to classify whether the loan status is yes or no. This was resolved using a classification technique like logistic regression.



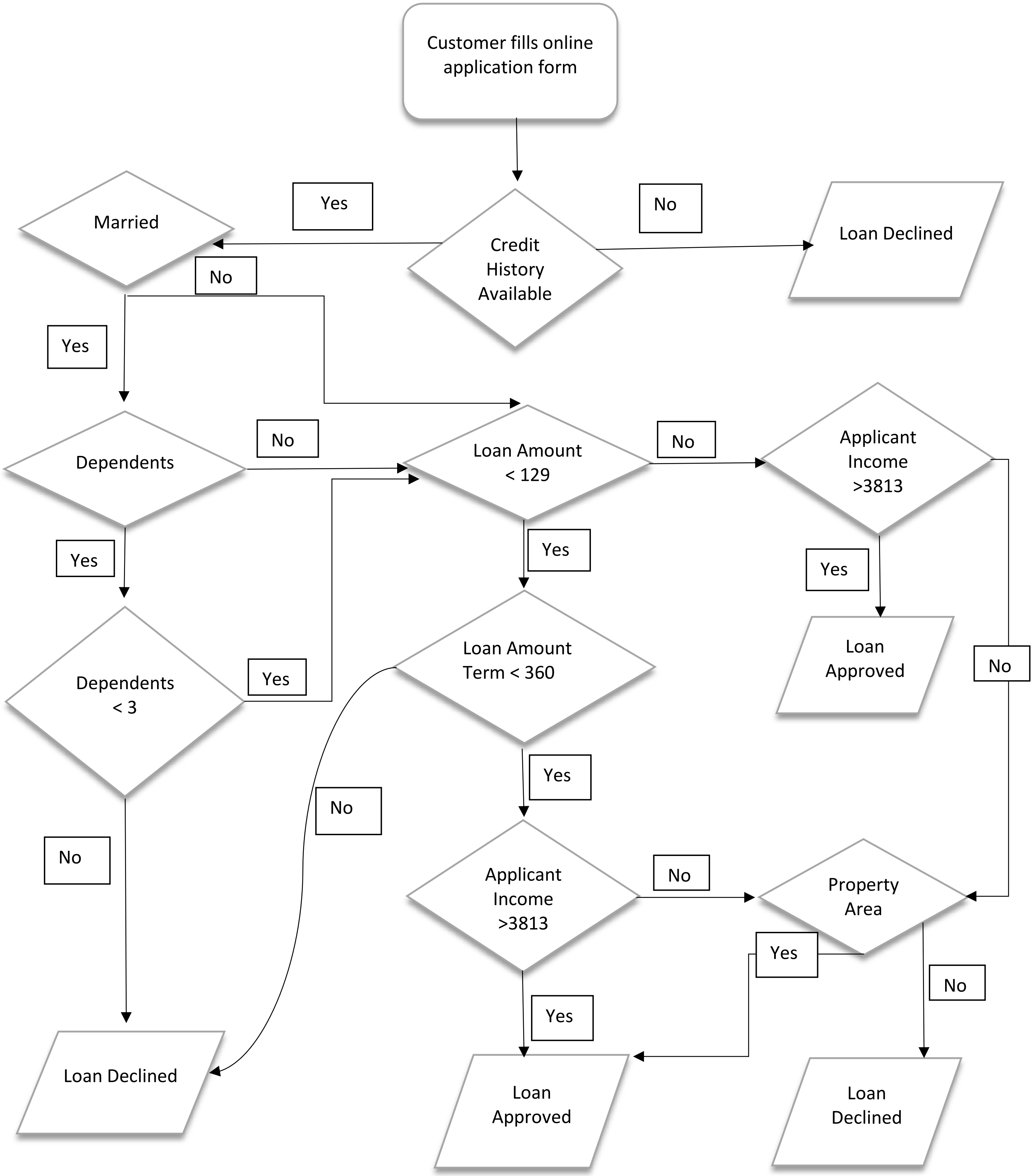
In 2 months, analyze customers' details to determine if their loan application would be approved.
DV: LOAN STATUS
Do specific customer social status contribute to the loan approval?
Education?
IV: Education
Martial status?
IV: Married
Job?
IV: Self_Employed
Children?
IV: Dependents
Do the amount of loan contribute to the loan approval?
Loan amount?
IV: Loan Amount
Loan terms?
IV: Loan_Amount_Term
Do specific customers history contribute to the loan approval?
Previous loans?
NA
Credit history?
IV: Credit_History
Do specific demographic contribute to the loan approval?
NA
Age?
Gender?
IV:Gender
Income?
IV: ApplicatIncome
IV: CoapplicantIncome
Location?
IV:Property Area

Source: http://databoosting.com/loan-prediction-problem/

Process Map for Model – 1:



Process Map for Model – 2:



Identification of key measures used to evaluate the success of your project:

- Loan_Status is the response variable and rest all are the variables /factors that decide the approval of the loan or not. Loan status (loan_status), binary, yes (yes = 1) or no (no = 0).
- There are 614 observations in our dataset and 13 columns altogether.
- Noise variables – Loan_Id and Education were dropped from our dataset for analysis purpose.

Qualitative data (Nominal):

- Gender, Married, Self_Employed, Dependents, Property_Area.

Quantitative data (Ordinal):

- LoanAmount, Loan_Amount_Term, ApplicantIncome, CoapplicantIncome, Credit_History.

1.3 Discussion of project scope:

- Look for correlation between dependent variable (Loan_Status) and independent variables in the dataset.
- The bank adds weightage for parameters collected from sources, such as customers and credit rating agencies, to understand their impact in the final decision-making process. Banks use this information to create rules to check if an applicant is eligible for the loan.
- Using Python, perform logistic regression, k-fold cross validation, confusion matrix, correlation matrix and feature importance.

2.0 CURRENT STATE OF THE PROCESS: MEASURE PHASE

2.1 Current Performance Level

We have used second-hand data from Kaggle. The data has 614 rows and 13 columns. Data set details are Gender, Marital Status, Education, Self-employed, Number of Dependents, Income, Loan Amount, Loan Amount Term, Credit History, Property Area, Loan status.

https://www.kaggle.com/ninzaami/loan-predication

Data Preprocessing is presented as follows:

We have used Python 3.8 for dataset cleaning, out of 13 columns, we have removed loan id and Education columns as these two columns are not necessary for our study. There were total 7 columns with missing values. Categorical variables such as Gender, Married, Dependents, Self-employed, Property Area are converted into numeric variable by encoding the categories and then imputed missing values with mode. For continuous variable such as Loan amount and Loan term amount, we imputed the values with their respective medians. Median was used instead of meaning as outliers are present in the dataset.

Creation of New Variables:

We have created one new variable as ‘Total income’.

- Total_Income = ApplicantIncome + CoapplicantIncome

To determine the loan eligibility criteria the bank adds weightage for parameters collected from sources, such as customers and credit rating agencies, to understand their impact in the final decision-making process. Banks use this information to create rules to check if an applicant is eligible for the loan. To get better the best result we will be creating two weightage list and will develop models based on each of them.

For Model 1 and Model 2 following table lists some parameters that a bank collects, in their order of priority, along with the granted weightage:

Parameters	**Weightage factor for Model -1**	**Weightage factor for Model - 2**
Loan Amount	10	10
Loan Amount Term (Number of months to pay back the loan)	8	8
Applicant Income	8	8
Property_Area (Net Assets)	8	8
Total Income (applicant income + co-applicant income)	8	8
Credit_History	6	6
Marital Status	7	5
Self employed	7	5
Gender	7	5
Number of dependents	4	4

These articles were used as reference for deciding the weightage factors of gender and marital status for Model 1 and Model 2.

https://www.sciencedirect.com/science/article/abs/pii/S0378426610002669

https://libproxy.library.unt.edu:6995/global/article/GALE%7CA164595268?u=txshracd2679&sid=summon

https://www.fdic.gov/regulations/compliance/manual/4/iv-1.1.pdf

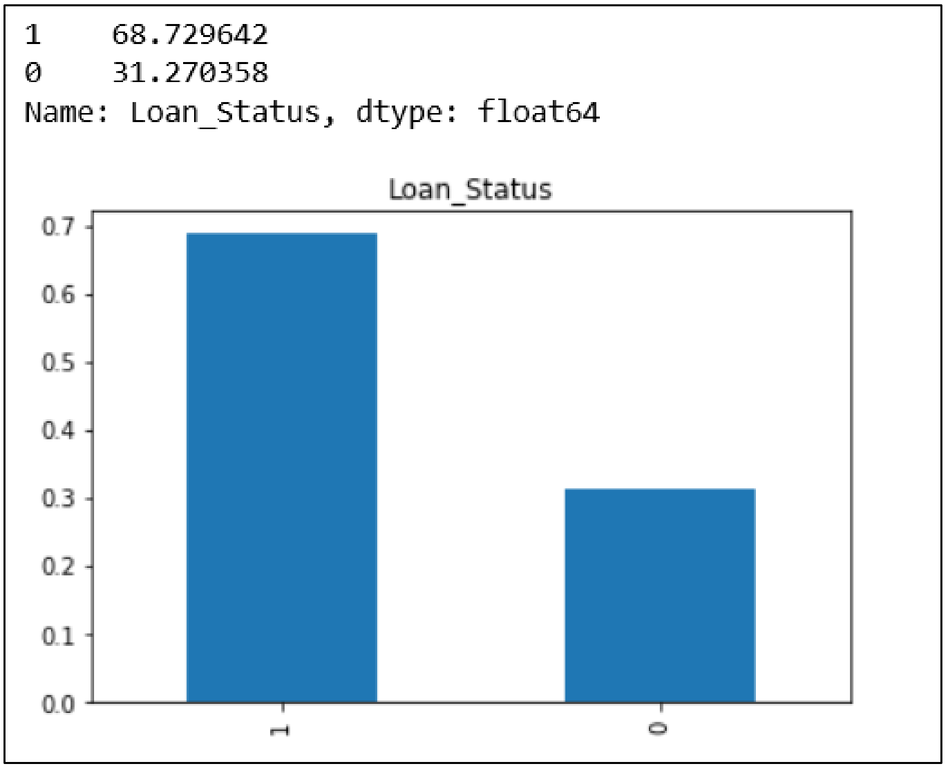
Using the above weightage parameter factors, we have created new variables for our analysis for Model 1 and Model 2.

The following are the new variables:

Model 1	**Model 2**
Gender_Max_Points_7	Gender_Max_Points_5
MaritalStatus_Max_Points_7	MaritalStatus_Max_Points_5
Dependent_Max_Points_4	Dependent_Max_Points_4
Self_Employed_Max_Points_7	Self_Employed_Max_Points_5
ApplicantIncome_Max_Points_8	ApplicantIncome_Max_Points_8
LoanAmount_Max_Points_10	LoanAmount_Max_Points_10
Loan_Amount_Term_Max_Points_8	Loan_Amount_Term_Max_Points_8
Credit_history_max_points_6	Credit_history_max_points_6
Property_Area_max_points_8	Property_Area_max_points_8
Total_income_max_points_8	Total_income_max_points_8
Loan_Status	Loan_Status

Current State of Key Y Outputs

Target variable of our process is Loan_Status, which indicates the denial or approval of the loan. By looking at below plot we can say that among all the data 68.72% (422 out of 614) of the people got loan approval. Here, Value 1 states Loan status approved, and Value 0 states Loan status denied.

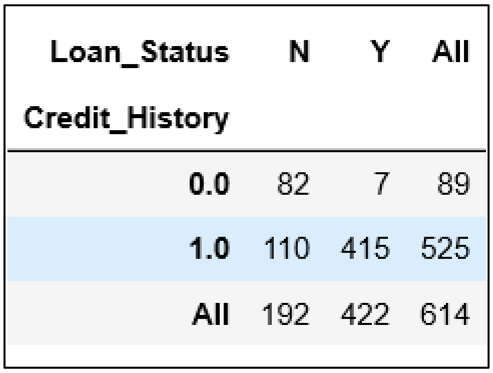


```
1    68.729642
0    31.270358
Name: Loan_Status, dtype: float64
```

Distribution/Data Patterns of Key Y Outputs.

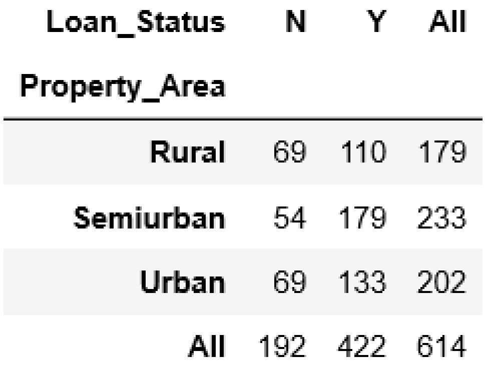
When customer apply for a mortgage, checking the credit score is one of the first things most lenders do. Credit score is determined based on your past payment history and borrowing behavior. The higher the credit score, the more likely it is customer will be approved for a mortgage and the better the interest rate will be. Hence, we have estimated the distribution between credit history and loan status.

We can observe (using cross tabulation) the distribution of credit history with load status. There are total 79.04 % of the applicants whose loans were approved and have credit history equals to 1.

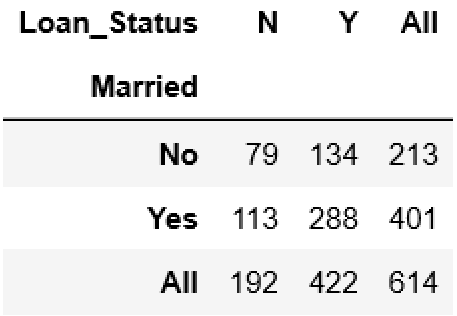


Loan_Status	N	Y	All
Credit_History			
0.0	82	7	89
1.0	110	415	525
All	192	422	614

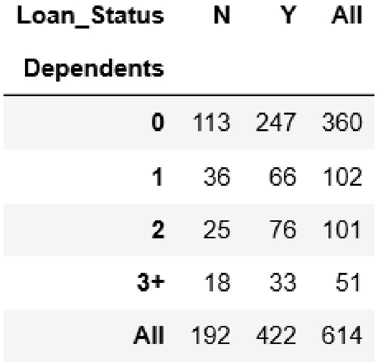
Likewise, we can observe the distribution of original variables such as gender, marital status, property area and number of dependents with loan status as shown below.



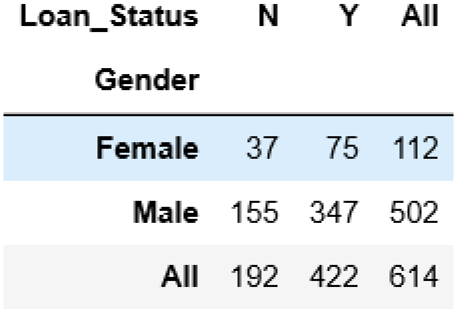
Loan_Status Property_Area	N	Y	All
Rural	69	110	179
Semiurban	54	179	233
Urban	69	133	202
All	192	422	614



Loan_Status Married	N	Y	All
No	79	134	213
Yes	113	288	401
All	192	422	614



Loan_Status Dependents	N	Y	All
0	113	247	360
1	36	66	102
2	25	76	101
3+	18	33	51
All	192	422	614



Loan_Status Gender	N	Y	All
Female	37	75	112
Male	155	347	502
All	192	422	614

2.2 Identification of Key Variables

Several variables can be considered when evaluating the loan prediction. Variables can be defined with a qualifier of input, output. Input variables have their values passed into a process or service when the process is initiated. Output variables have their values returned from a process when it completes. Below are the key variables.

- **Key Output Variable:**

Loan_Status: This is our target variable, binary 1/0 or string “yes” and “No” – tells whether loan is approved or not.

- **Key Input Variables:**

Credit_History: Credit History (0/1)
Marital status: Married (Yes)/(No)
Property Area: Semi urban, Urban and Rural
Applicant_Income: Applicant Income
Self_Employed: self Employement income
Dependents: Number of persons depending on the client
ApplicantIncome: Applicant Income

2.3 Identification of Target Performance Levels or Project Goals

The primary goal of this project is to build a model to predict the likely loan approvals by using classification data algorithms. The historical data of the customers such as their income, loan amount, credit etc. will be used

to do the analysis. Also, the analysis will also be done to find the most relevant attributes, i.e., the factors that affect the prediction result the most.
The entire loan approval process can have several parts, including getting pre-approved, getting the home appraised, and getting the actual loan. In a normal market, this process takes about 30 days on average. During high-volume months, it can take longer—an average of 45 to 60 days, depending on the lender. If the lender uncovers any financial issues in the record (e.g., a low credit score, previous foreclosure, or overwhelming debt), getting a mortgage can become a slower and more complicated process.

https://www.realtor.com/advice/finance/how-long-does-it-take-to-get-a-mortgage/

3.0 ANALYSIS AND FINDINGS

The loan is one of the most important products of any financial institute. Distribution of loans is the core business part of almost every banks. Effective business strategies for loan application process is an important aspect taken into consideration by financial institutes. Therefore, several aspects (variables) are taken into account while approving a loan. Predicting whether a borrower will default on loan is a significant concern of most financial institutes. If the lender is too strict, fewer loans get approved, which means there's less interest to collect. But if they're too lax, they end up approving loans that default.

In this study, loan behavior is analyzed using logistic regression machine learning model. Machine learning model was trained to predict for the loan repayment. Machine learning model was evaluated via K-fold classification technique, confusion matrix and correlation matrix. Feature importance was performed to identify the most useful variables in the dataset.

3.1 Analysis Variables:

- **For Model – 1:** Gender_Max_Points_7, MaritalStatus_Max_Points_7, Dependent_Max_Points_4, Self_Employed_Max_Points_7, ApplicantIncome_Max_Points_8, LoanAmount_Max_Points_10, Loan_Amount_Term_Max_Points_8, Credit_history_max_points_6, Property_Area_max_points_8, Total_income_max_points_8.

- **For Model – 2:** Gender_Max_Points_5, MaritalStatus_Max_Points_5, Dependent_Max_Points_4, Self_Employed_Max_Points_5, ApplicantIncome_Max_Points_8, LoanAmount_Max_Points_10, Loan_Amount_Term_Max_Points_8, Credit_history_max_points_6, Property_Area_max_points_8, Total_income_max_points_8.

These variables were created by adding weightage factors to parameters and used for analysis purpose.

3.2 Model Analysis:

Logistic Regression was used for model analysis. It is one of the simplest and commonly used machine learning algorithms for two-class classification. It is easy to implement and can be used as the baseline for any binary classification problems. It describes and estimates the relationship between one binary variable (loan_status) and independent variables.

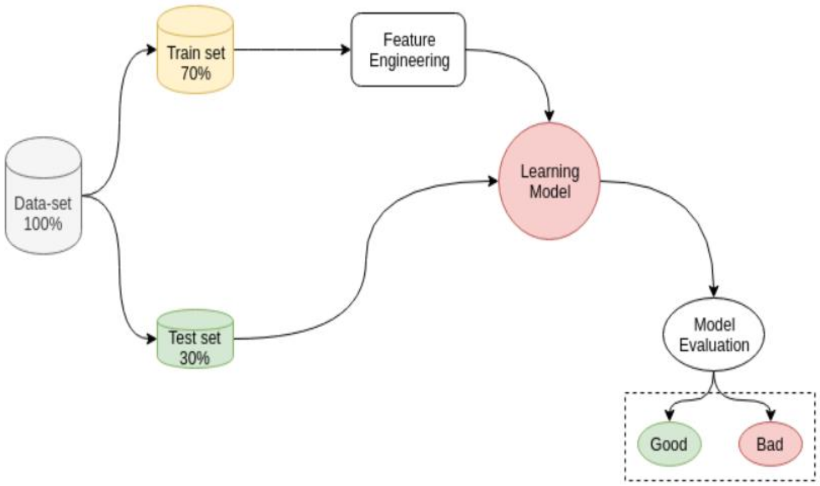
Logistic regression was used to fit into the loan data.

Step 1: Logistic regression model with regularization was constructed to avoid overfitting and conducted a confusion matrix to see how the model performed on the loan prediction dataset.

Step 2: We tested different parameters in order to see how accuracy changes.

3.3 Model Evaluation:

Entire data was split into two sets, one is used to train the model upon and the other is kept as a holdout set which is used to check how the model behaves with completely unseen data. The figure below summarizes the entire idea of performing the split.

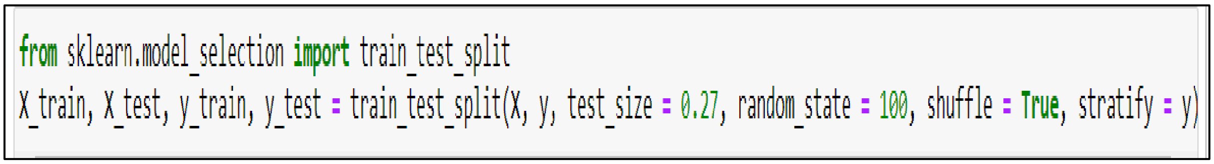


A very high-level view of the model building and evaluation pipeline

Source: https://towardsdatascience.com/cross-validation-430d9a5fee22

Splitting the dataset:

The data was split into training set (75%), and test set (25%). Training set was used to fit the model, and test set was used to evaluate the best model to get an estimation of generalization error.



```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.27, random_state = 100, shuffle = True, stratify = y)
```

X is the original entire set of features and y is the entire set of corresponding true labels. The above function splits the entire set into train and test set with a ratio of 0.25 assigned for the test set. The parameter shuffle is set to true, thus the data set will be randomly shuffled before the split. The parameter stratify is recently added to Sci-kit Learn from v0.17, it makes a split so that the proportion of values in the sample produced will be the same as the proportion of values provided to the parameter stratify. For example, if the variable y is a binary categorical variable with values 0 and 1 and there are 10% of zeros and 90% of ones, stratify = y will make sure that your random split has 10% of 0's and 90% of 1's.

3.4 K-Fold Cross Validation:

K-fold cross validation was applied to try to reduce over-fitting. The idea was to estimate the predictive performance of the model for data that is not yet known. Cross-validation is a robust form, repeats the experiment several times, using several different parts of the training set as validation sets. This provides a more accurate analysis of how the model generalizes to previously unknown data.

In order to run the K-Fold, the dataset was split into a number of folds, in this case 5. So then, each fold was divided into a train and test subset each with its own features and targets. Each classifier was trained using the training features and targets, then they predict the result using the fold test features. Finally, the performance of the classifier was measured using acc: accuracy

In K-fold cross validation repeats the model evaluation process multiple times (instead of one time) and calculates the mean skill. The mean estimate of any parameter is less biased than a one-shot estimate. There is still some bias though.

The mean accuracy scores of the four models obtained using K – fold cross validation was:

	Model-1		Model-2	
	Model1_full data (using credit_history = 1 and credit_history = 0)	Model1_data (using credit_history = 1 only)	Model2_full data (using credit_history = 1 and credit_history = 0)	Model2_data (using credit_history = 1 only)
Mean Accuracy Score using K-fold cross validation	0.80946 (80.946 %)	0.78666 (78.666 %)	0.80624 (80.624%)	0.79047 (79.047%)

3.5 Confusion Matrix:

It provides a summary of the predictive results in a classification problem. Correct and Incorrect predictions are summarized in the table with their values and broken down by each class.

- **True Positive:** Both predicted and actual value is true. The predicted and actual value of loan_status is 1 (1 or Yes).

- **True Negative:** Both the predicted and actual value is false. The predicted value of loan_status not 1 (i.e. it is 0 or No) and it actually is 0

- **False Positive (Type I Error):** The predicted value is positive and it’s false. The predicted value of loan_status is 1 (Yes) but its actual value is not (it is 0 or No).

- **False Negative (Type II Error):** The predicted value is negative and it’s false. The predicted value of loan_status is not 1 but its actual value is 1.

Confusion Matrix:

Predictive Values

Actual Values

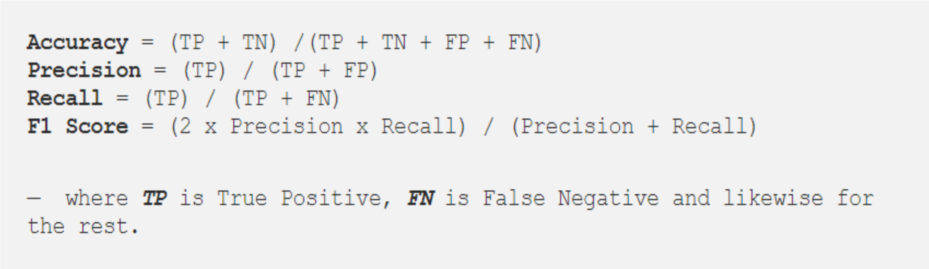
	Loan_status = 1(or Yes)	Loan_status = 0 (or No)
Loan_status = 1(or Yes)	True Positive (loan_status = 1 or Yes)	False Negative (Type II Error) (loan_status = 0 or No i.e. loan_status ǂ 1 or Yes)
Loan_status = 0 (or No)	False Positive (Type I Error) (loan_status ǂ 0 or No i.e. loan_status = 1 or Yes)	True Negative (loan_status ǂ 1 or Yes i.e. loan_status = 0 or No)

Confusion Matrix Values (Figure 1, 3, 5, 7):

	Model-1		Model-2	
	Model1_full data (using credit_history = 1 and credit history = 0)	Model1_data (using credit history = 1 only)	Model2_full data (using credit_history = 1 and credit history = 0)	Model2_data (using credit_history = 1 only)
True Positive	19	1	19	1
True Negative	105	103	105	104
False Positive (Type I Error)	1	1	1	0
False Negative (Type II Error)	29	27	29	27

3.6 Classification Report:

There are several metrics that can be deduced from the confusion matrix, such as –



```
Accuracy = (TP + TN) /(TP + TN + FP + FN)
Precision = (TP) / (TP + FP)
Recall = (TP) / (TP + FN)
F1 Score = (2 x Precision x Recall) / (Precision + Recall)


—  where TP is True Positive, FN is False Negative and likewise for
the rest.
```

- **Precision:**

Precision is defined as the ratio of the total number of correctly classified positive classes divided by the total number of predicted positive classes. Or, out of all the predictive positive classes, how much we predicted correctly. Precision should be high.

- **Recall or Sensitivity:**

Recall is defined as the ratio of the total number of correctly classified positive classes divide by the total number of positive classes. Or, out of all the positive classes, how much we have predicted correctly. Recall should be high.

- **F-score or F-1 score:**

It is difficult to compare two models with different Precision and Recall. So, to make them comparable, we use F-Score. It is the Harmonic Mean of Precision and Recall. As compared to Arithmetic Mean, Harmonic Mean punishes the extreme values more. F-score should be high.

Classification Report:

	Model-1				Model-2			
	Model1_full data (using credit_history = 1 and credit_history = 0)		Model1_data (using credit_history = 1 only)		Model2_full data (using credit_history = 1 and credit_history = 0)		Model2_data (using credit_history = 1 only)	
Loan_status	0	1	0	1	0	1	0	1
Accuracy	0.81 (81%)		0.79 (79%)		0.81 (81%)		0.80 (80%)	
Recall	0.40	0.99	0.04	0.99	0.40	0.99	0.04	1.00
Precision	0.95	0.78	0.50	0.79	0.95	0.78	1.00	0.79
F1-score	0.56	0.88	0.07	0.88	0.56	0.88	0.07	0.89
Support	48	106	28	104	48	106	28	104

3.7 Correlation Matrix:

Correlation Matrix describes the correlation between different features in the dataset. To understand the relationship between multiple variables and attributes in the dataset, the first step is to check the correlation relationship between each variable as correlation is used as a basic quantity for many modelling techniques, as it can also help in predicting one attribute from another.

Correlation values range between -1 and 1. There are two key components of a correlation value:

- Magnitude – the larger the magnitude (closer to 1 or -1), the stronger the relation.
- Sign – If negative, there is inverse correlation. If positive, there is regular correlation.

(Figure 1, 3, 5, 7) - In our dataset, there is a strong positive correlation (approximately 0.84) in both the models (Model-1 and Model-2) between:

- Self-Employed_Max_Points and Applicant_Income_Max_Points variables.
- Applicant_Income_Max_Points and Total_income_Max_Points.

3.8 Feature Importance:

Feature selection is the process of finding and selecting the most useful features in a dataset. Unnecessary features decrease training speed, the model interpretability and the generalization performance on the test set. Estimating the influence of a given feature to a model prediction is important mainly in large datasets for performance gain by selecting only the most relevant ones.

Feature Importance: (Figure 2,4,6,8) – The following were the top two features:

	Model-1		Model-2	
	Model1_full data (using credit_history = 1 and credit_history = 0)	Model1_data (using credit_history = 1 only)	Model2_full data (using credit_history = 1 and credit_history = 0)	Model2_data (using credit_history = 1 only)
Useful Features	Credit_history_max_points_6 and Marital_status_max_points_7	Loan_Amount_Term_Max_Points_8 and Marital_status_max_points_7	Credit_history_max_points_6 and Loan_Amount_Term_max_points_8	Loan_Amount_Term_max_points_8 and Property_Area_Max_Points_8

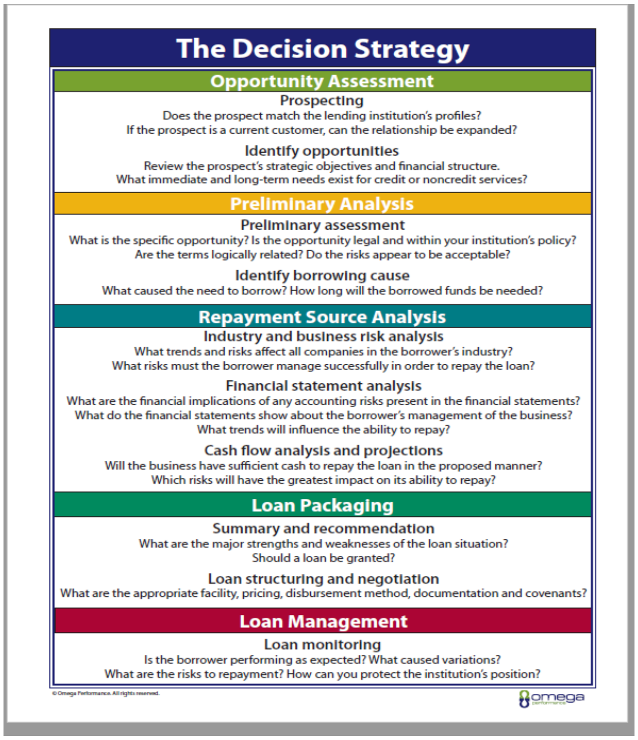
4.0 RECOMMENDATIONS: THE IMPROVE PHASE

- To determine whether a customer qualifies for the loan, can be a complicated preposition for banks. Banks need to determine whether a customer qualifies for the loan based on the collected information. In our collected data from the Kaggle, important parameters such as ‘Criminal History’, ‘Frequency of borrowing’, and ‘Number of years in job’ are not defined. Theses parameters must be included as these variables plays an important role in making decisions. For example, in case of criminal history – ‘description of the specific charges’ (e.g. assault, forgery, robbery), ‘description of the level of individual charges’ (e.g. misdemeanor, felony), ‘date of offense including month, day (if possible) and year’, ‘City and state, or county and state where offense/s occurred’ plays an vital role in loan approval process. Also, Failure to provide appropriate information while handling the data or managing the process can put some concerns that affect the evaluation process.

- In our Loan process, FICO or ‘Fair Isaac credit scores’ of the customer must be included and checked before approving the loan. FICO scores are used by 90 of top lenders to make decision about credits approval. It makes decision faster, consistent, and fair.

- The technical limitations of legacy lending systems reduce the lender’s ability to replace manual steps with automated decisions. They also make it impossible to integrate alternative data sources that enable lenders to make more informed, accurate lending decisions. Via the cloud, however, modern loan origination systems offer pre-integrated access to data sources. This data can be automatically accessed without a manual login to verify applicant information using decision rules. Results of verification, in combination with decision rules, can also help provide a better assessment of creditworthiness.

- Streamlining the decision process is often the primary difference between completion and abandonment. Driving change in how lending decision are made take a serious level of introspection and internal advocacy on the financial institutions part. Decision makers have to advocate an internal process review. They need to emphasize on automation and work with compliance officers to determine what elements of the lending process they can eliminate or consolidate and what must be retained. Below is the snapshot of Omega Performance Decision Strategy which serves as a foundation for all their credit training programs.



The Decision Strategy

Opportunity Assessment

Prospecting
Does the prospect match the lending institution's profiles?
If the prospect is a current customer, can the relationship be expanded?

Identify opportunities
Review the prospect's strategic objectives and financial structure.
What immediate and long-term needs exist for credit or noncredit services?

Preliminary Analysis

Preliminary assessment
What is the specific opportunity? Is the opportunity legal and within your institution's policy?
Are the terms logically related? Do the risks appear to be acceptable?

Identify borrowing cause
What caused the need to borrow? How long will the borrowed funds be needed?

Repayment Source Analysis

Industry and business risk analysis
What trends and risks affect all companies in the borrower's industry?
What risks must the borrower manage successfully in order to repay the loan?

Financial statement analysis
What are the financial implications of any accounting risks present in the financial statements?
What do the financial statements show about the borrower's management of the business?
What trends will influence the ability to repay?

Cash flow analysis and projections
Will the business have sufficient cash to repay the loan in the proposed manner?
Which risks will have the greatest impact on its ability to repay?

Loan Packaging

Summary and recommendation
What are the major strengths and weaknesses of the loan situation?
Should a loan be granted?

Loan structuring and negotiation
What are the appropriate facility, pricing, disbursement method, documentation and covenants?

Loan Management

Loan monitoring
Is the borrower performing as expected? What caused variations?
What are the risks to repayment? How can you protect the institution's position?

© Omega Performance. All rights reserved.

omega performance

Source: https://www.omega-performance.com/omega-decision-strategy/

5.0 POTENTIAL BENEFITS OF MAKING RECOMMENDATIONS:

Below are the benefits of recommendations made above (point 4):

- Customers who are frequent borrowers establish a reputation which directly impacts on their ability to secure debt at advantageous terms. Thus, frequency of borrowing is an important variable for lenders in determining the loan status. Also, Employers in the financial services industry such as banks, credit unions and broker-dealers, are subject to various background investigation and screening requirements. Criminal history enables lenders to gain honesty information about the applicant which also becomes a key constraint in the loan approval process.

- The benefit of using FICO scores rather than any other credit score gives customer a better understanding of their credit and more confidence while filing the application. FICO Scores are trusted to be a fair and reliable measure of whether a person will pay back their loan on time. By consistently using FICO Scores, lenders take on less risk, and you get faster and fairer access to the credit you need and can manage.

- Modern loan origination software overcomes limitations of legacy systems. Using cloud technologies and digital documentation application processing time and application cost can be reduced. Integration of alternative data sources, as well as the application of decision rules, result in better quality lending decisions.

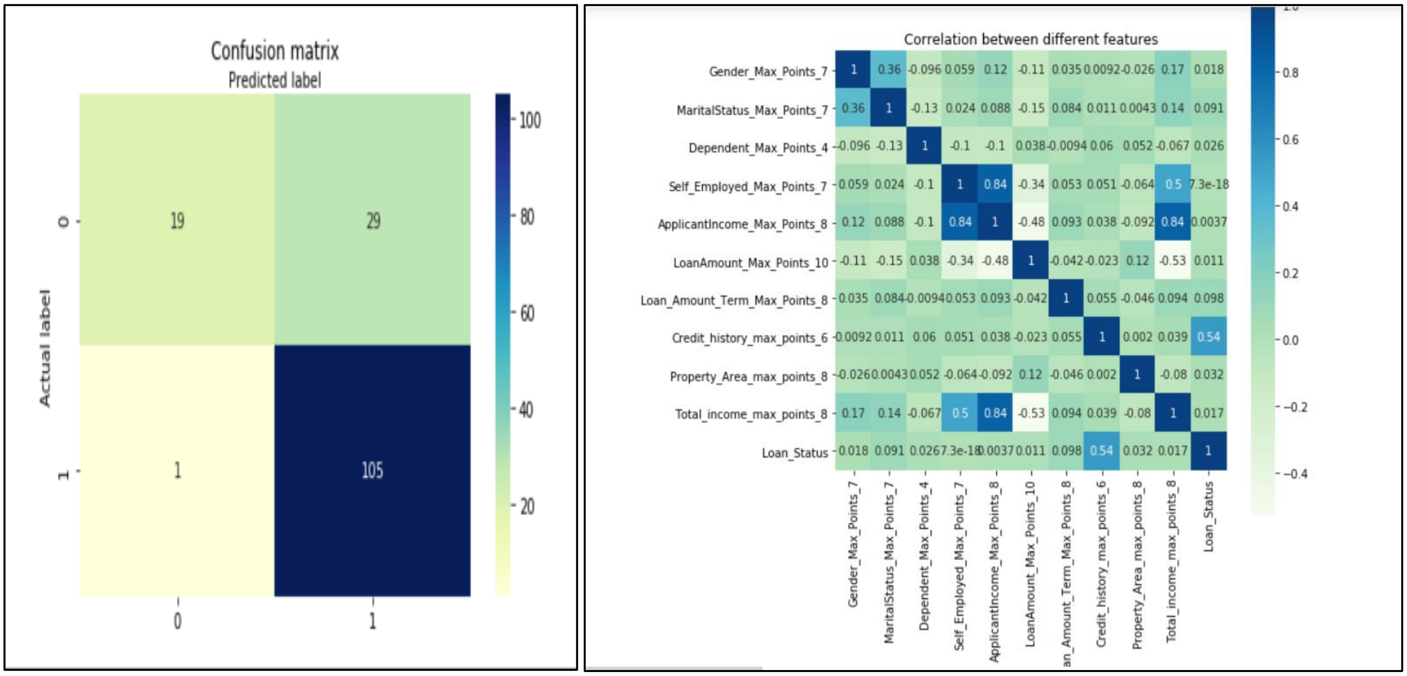
- Decision Strategy ensures that lenders capture all the information necessary to assess the risk in lending to a business/customer. This results in a consistent and reliable approach to credit decision making which is the foundation of a solid credit culture and aids in quicker loan applications and decisions.

6.0 SUMMARY/CONCLUSION

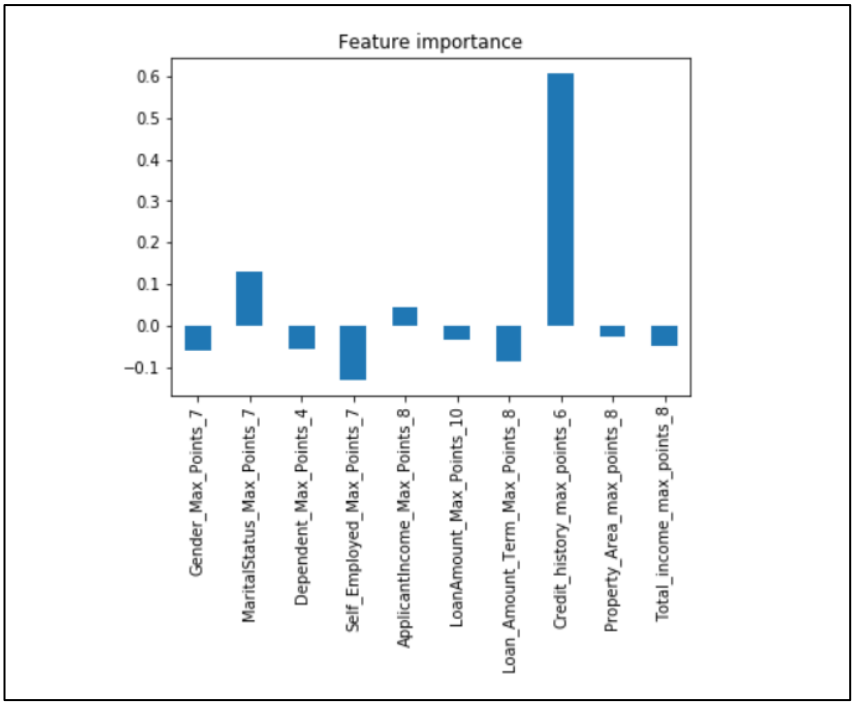
- Loan_Status is the dependent (Y) variable in our dataset.
- Loan behavior was assessed using logistic regression machine learning model. Model was evaluated by splitting the dataset in the ratio of 75:25.
- K-Fold Classification was performed on Model -1 and Model -2 to understand the accuracy of the trained dataset. The following was the mean accuracy of the models:
  Model1_fulldata (using credit_history = 1 and credit_history = 0) – 80.946%
  Model1_data (using credit_history = 1 only) – 78.67%
  Model2_full data (using credit_history = 1 and credit_history = 0) – 80.624%
  Model2_data (using credit_history = 1 only) – 79.047%
  From the mean accuracy score obtained via K-Fold Classification of the models did not have any significant difference.
- A strong positive correlation was found between the analysis variables:
  Self-Employed_Max_Points and Applicant_Income_Max_Points variables.
  Applicant_Income_Max_Points and Total_income_Max_Points.
- With the help of feature importance technique, we can deduce that:
  For Model-1: Credit_history_max_points_6, Marital_Status_Max_Points_7 and Loan_Amount_Term_Max_Points_8 were the most features.
  For Model-2: Credit_history_max_points_6, Loan_Amount_Term_Max_Points_8 and Property_Area_Max_Points_8 was found to be the most useful features. Therefore, these features(variables) should be taken into account while approving loan.
- Moreover, irrespective of the weightage factor to gender and self-employment parameters, these features are not highly significant in loan decision making process.
- The decision whether marital status might affect the loan approval process lies at the discretion of the lending officer.

APPENDICES

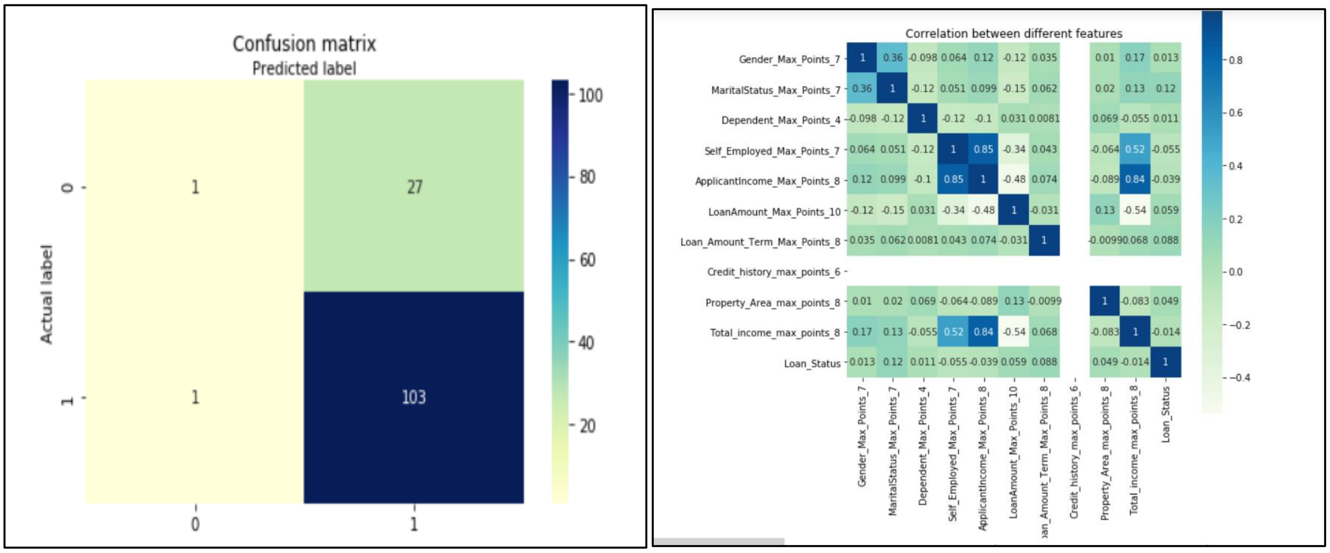
- Figure -1- Confusion Matrix & Correlation Matrix: Model1_full data (using credit_history = 1 and credit_history = 0)



- Figure -2 - Feature Importance: Model1_full data (using credit_history = 1 and credit_history = 0)



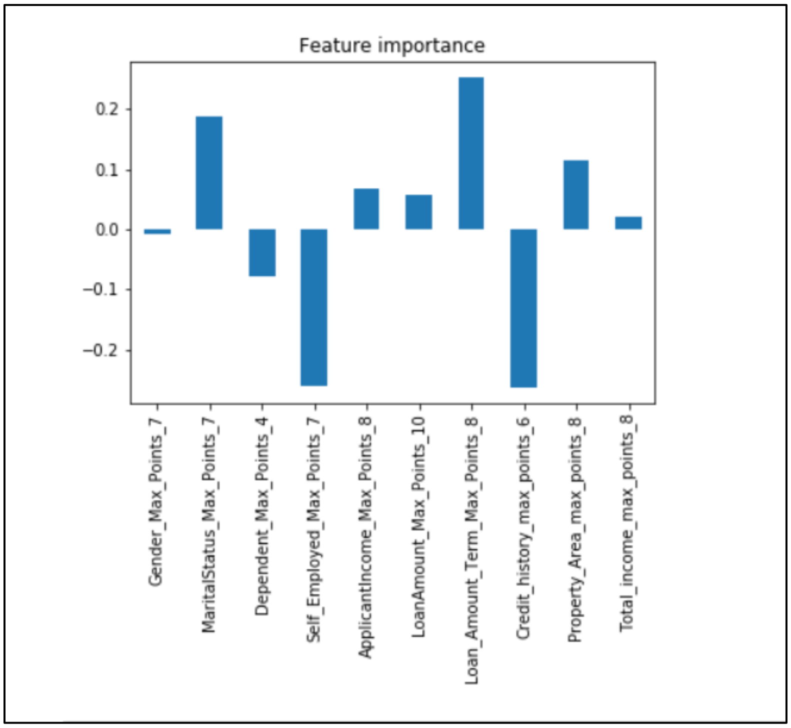
- Figure -3 - Confusion Matrix & Correlation Matrix: Model1_data (using credit_history = 1 only)



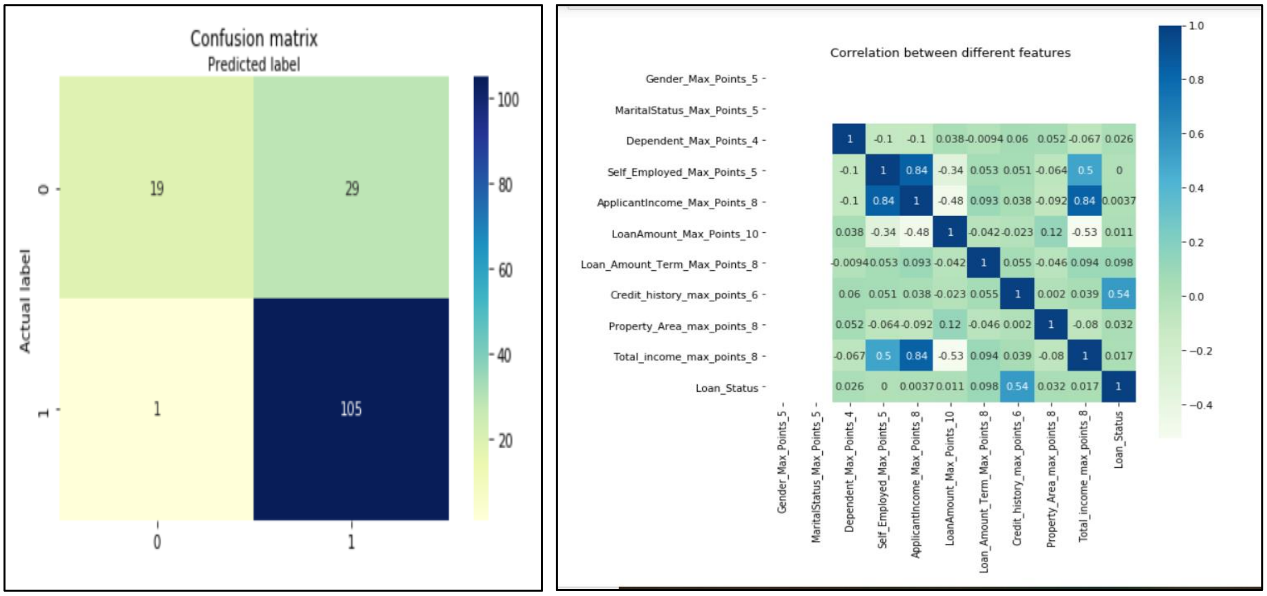
Correlation between different features

	Gender_Max_Points_7	MaritalStatus_Max_Points_7	Dependent_Max_Points_4	Self_Employed_Max_Points_7	ApplicantIncome_Max_Points_8	LoanAmount_Max_Points_10	an_Amount_Term_Max_Points_8	Credit_history_max_points_6	Property_Area_max_points_8	Total_income_max_points_8	Loan_Status
Gender_Max_Points_7	1	0.36	-0.098	0.064	0.12	-0.12	0.035		0.01	0.17	0.013
MaritalStatus_Max_Points_7	0.36	1	-0.12	0.051	0.099	-0.15	0.062		0.02	0.13	0.12
Dependent_Max_Points_4	-0.098	-0.12	1	-0.12	-0.1	0.031	0.0081		0.069	-0.055	0.011
Self_Employed_Max_Points_7	0.064	0.051	-0.12	1	0.85	-0.34	0.043		-0.064	0.52	-0.055
ApplicantIncome_Max_Points_8	0.12	0.099	-0.1	0.85	1	-0.48	0.074		-0.089	0.84	-0.039
LoanAmount_Max_Points_10	-0.12	-0.15	0.031	-0.34	-0.48	1	-0.031		0.13	-0.54	0.059
Loan_Amount_Term_Max_Points_8	0.035	0.062	0.0081	0.043	0.074	-0.031	1		-0.0099	0.068	0.088
Credit_history_max_points_6											
Property_Area_max_points_8	0.01	0.02	0.069	-0.064	-0.089	0.13	-0.0099		1	-0.083	0.049
Total_income_max_points_8	0.17	0.13	-0.055	0.52	0.84	-0.54	0.068		-0.083	1	-0.014
Loan_Status	0.013	0.12	0.011	-0.055	-0.039	0.059	0.088		0.049	-0.014	1

- Figure - 4 - Feature Importance: Model1_data (using credit_history = 1 only)

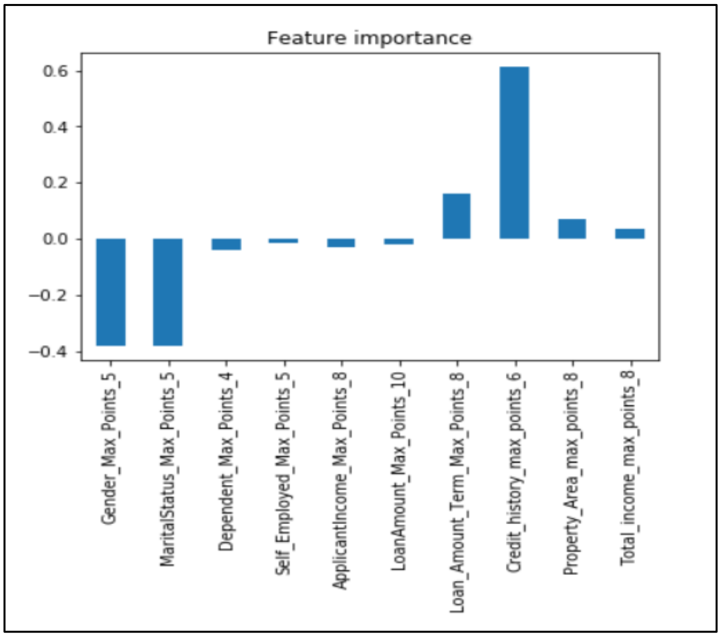


- Figure – 5 – Confusion Matrix & Correlation Matrix: Model2_full data (using credit_history = 1 and credit_history = 0)

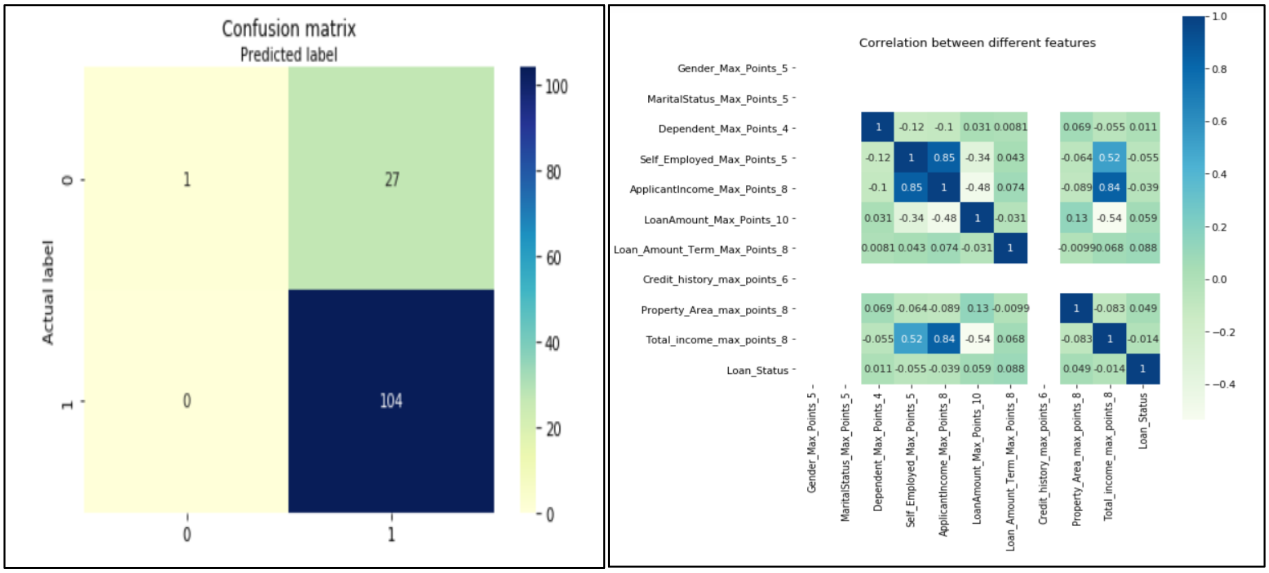


	Gender_Max_Points_5	MaritalStatus_Max_Points_5	Dependent_Max_Points_4	Self_Employed_Max_Points_5	ApplicantIncome_Max_Points_8	LoanAmount_Max_Points_10	Loan_Amount_Term_Max_Points_8	Credit_history_max_points_6	Property_Area_max_points_8	Total_income_max_points_8	Loan_Status
Gender_Max_Points_5											
MaritalStatus_Max_Points_5											
Dependent_Max_Points_4			1	-0.1	-0.1	0.038	-0.0094	0.06	0.052	-0.067	0.026
Self_Employed_Max_Points_5			-0.1	1	0.84	-0.34	0.053	0.051	-0.064	0.5	0
ApplicantIncome_Max_Points_8			-0.1	0.84	1	-0.48	0.093	0.038	-0.092	0.84	0.0037
LoanAmount_Max_Points_10			0.038	-0.34	-0.48	1	-0.042	-0.023	0.12	-0.53	0.011
Loan_Amount_Term_Max_Points_8			-0.0094	0.053	0.093	-0.042	1	0.055	-0.046	0.094	0.098
Credit_history_max_points_6			0.06	0.051	0.038	-0.023	0.055	1	0.002	0.039	0.54
Property_Area_max_points_8			0.052	-0.064	-0.092	0.12	-0.046	0.002	1	-0.08	0.032
Total_income_max_points_8			-0.067	0.5	0.84	-0.53	0.094	0.039	-0.08	1	0.017
Loan_Status			0.026	0	0.0037	0.011	0.098	0.54	0.032	0.017	1

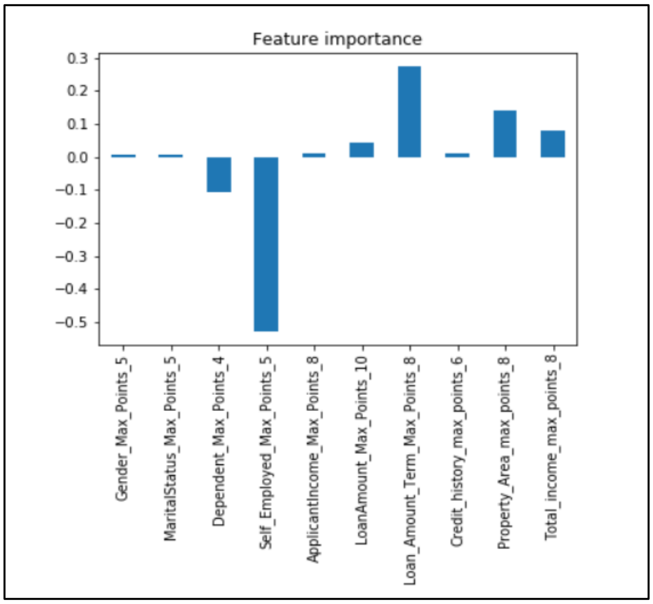
- Figure – 6 – Feature Importance: Model2_full data (using credit_history = 1 and credit_history = 0)



- Figure – 7 – Confusion Matrix & Correlation Matrix: Model2_data (using credit_history = 1 only)



- Figure – 8 – Feature Importance: Model2_data (using credit_history = 1 only)



Statistical Analysis Plan
for

Pi -Oneering Consulting Group

Version: 2

CONTENTS

1.INTRODUCTION TO YOUR PROJECT WITH BASIC BACKGROUND

Among all industries, insurance domain has the largest use of analytics & data science methods. This dataset is about the company that wants to automate the loan eligibility process based on customer details provided while filling online application. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers.

2. DATA SOURCE

We have used second-hand data from Kaggle. The data has 614 rows and 13 columns. Data set details are Gender, Marital Status, Education, Self-employed, Number of Dependents, Income, Loan Amount, Loan Amount Term, Credit History, Property Area, Loan status.

https://www.kaggle.com/ninzaami/loan-predication

3. ANALYSIS OBJECTIVES

In this study, loan behavior was analyzed using logistic regression machine learning model. Machine learning model was trained to predict for loan repayment. Machine learning model was evaluated via K-fold classification technique, confusion matrix and correlation matrix. Feature Importance was performed to identify the most useful variables in the dataset.

4. ANALYSIS SETS/ POPULATIONS/SUBGROUPS

By using customer's information to help banking team to decide whether to look at some specific parameter (e.g. gender, self-employed) or not.

5. ENDPOINTS AND COVARIATES

Covariates are variables that covary with the dependent variable but are not focus of the study. Here in our study, Loan amount term and property area are covariates as focus of the study is to measure the effect of gender, marital status, self-employed, application income on the dependent variable which is loan status.

In our study, an endpoint is an outcome used to judge the approval of the loan; it is a precisely defined variable intended to reflect an outcome of interest that is statistically analyzed to address a research question. Primary endpoint of our model is to check the loan status.

6. HANDLING OF MISSING VALUES, OUTLIERS AND OTHER DATA CONVENTIONS

Out of 13 columns, we have removed Loan id and Education columns as these two columns are not necessary for our study. There were total 7 columns with missing values. Categorical variables such as Gender, Married, Dependents, Self-employed, Property Area are converted into numeric variable by encoding the categories and then imputed missing values with mode. For continuous variable such as Loan amount and Loan term amount, we imputed the values with their respective medians.

7. STATISTICAL METHODOLOGY

7.1 STATISTICAL PROCEDURES

With the help of machine learning model like logistic regression, K-fold cross validation technique, confusion matrix, correlation matrix and feature importance were performed within our research industry/topic to improve business decision making and achieve desired goal(s).

With the help of Rule Point informatica document, we created weighted parameter factors for our data. We created two models having different weighted parameters for selected variables. The two models were compared based on the scoring metrics.
All statistical Analysis will be performed using Python3.8.

Logistic Regression:
http://utstat.toronto.edu/~ali/papers/creditworthinessProject.pdf

https://www.researchgate.net/publication/290994615_Predicting_the_Probability_of_Loan-Default_An_Application_of_Binary_Logistic_Regression

https://machinelearningmastery.com/k-fold-cross-validation/

Rule Point Informatica 6.2:
https://docs.informatica.com/complex-event-processing/rulepoint/6-2/business-process-management-use-case-example/business-process-management-use-case.html

7.2 MEASURES TO ADJUST FOR MULTIPLICITY, CONFOUNDS, HETEROGENEITY, ETC.

This is not applicable to our dataset.

8. SENSITIVITY ANALYSES

For sensitivity analysis and to conduct a what-if analysis we have used python where we studied how our independent variables affect the loan status of applicant which is also the dependent variable.

9. RATIONALE FOR ANY DEVIATION FROM PRE-SPECIFIED ANALYSIS PLAN PERFORMED

Instead of dimensionality reduction technique like principal component analysis, K-fold cross validation technique was used to estimate the skill of machine learning model. This technique provides a more accurate analysis of how the model generalizes to previously unknown data. Moreover, in machine learning, “dimensionality” refers to the number of features (i.e. input variables) in the dataset. Dimensionality reduction technique like principal component analysis is performed when the number of features is very large to the number of observations in the dataset such that certain algorithms struggle to train effective models.

The dataset used is relatively small with 614 observations and 13 columns. This technique was dropped later and thus the deviation.

The analysis plan is not considered to be a contract, but rather a set of goals and guidelines for the analysis of research outputs. Nevertheless, deviations of the plan should be documented so that they may be taken into consideration when statistical analyses and modeling are carried out. Possible circumstances under which the plan might be changed include:

- Changes in the objectives of the research team not predicated by data snooping
- Loss of data, or failure to acquire intended data
- Excessive amount of missing data in a pre-specified analysis variable, without an opportunity to improve data completeness
- Additional data that is made available, which was not presumed to be available during the formulation of the analysis plan
- The inability of acquired data to satisfy the assumptions anticipated in the analysis plan
- Results suggest research phenomenon is more/less general than had been anticipated
- Results suggest estimates will be insufficiently precise to meet objectives
- Results suggest planned statistical approach will be inappropriate (e.g. more complicated model required, perhaps with additional covariates) While changes due to these (and other) ventualities are sometimes inevitable, failing to document such changes can result in invalid statistical inference in some circumstances.

http://biostat.mc.vanderbilt.edu/wiki/Main/PreStatAnalysisPlan

10. PLANS TO ENSURE QUALITY AND ETHICS

A. Professional Integrity and Accountability

The ethical statistician uses methodology and data that are relevant and appropriate; without favoritism or prejudice; and in a manner intended to produce valid, interpretable, and reproducible results. The ethical statistician does not knowingly accept work for which he/she is not sufficiently qualified, is honest with the client about any limitation of expertise, and consults other statisticians when necessary or in doubt. It is essential that statisticians treat others with respect.

The ethical statistician:

Identifies and mitigates any preferences on the part of the investigators or data providers that might predetermine or influence the analyses/results. Employs selection or sampling methods and analytic approaches appropriate and valid for the specific question to be addressed, so that results extend beyond the sample to a population relevant to the objectives with minimal error under reasonable assumptions. Respects and acknowledges the contributions and intellectual property of others. When establishing authorship order for posters, papers, and other scholarship, strives to make clear the basis for this order, if determined on grounds other than intellectual contribution. Discloses conflicts of interest, financial and otherwise, and manages or resolves them according to established (institutional/regional/local) rules and laws.

Accepts full responsibility for his/her professional performance. Provides only expert testimony, written work, and oral presentations that he/she would be willing to have peer reviewed.

Exhibits respect for others and, thus, neither engages in nor condones discrimination based on personal characteristics; bullying; unwelcome physical, including sexual, contact; or other forms of harassment or intimidation, and takes appropriate action when aware of such unethical practices by others.

B. Integrity of data and methods

The ethical statistician is candid about any known or suspected limitations, defects, or biases in the data that may affect the integrity or reliability of the statistical analysis. Objective and valid interpretation of the results requires that the underlying analysis recognizes and acknowledges the degree of reliability and integrity of the data.

The ethical statistician:

Acknowledges statistical and substantive assumptions made in the execution and interpretation of any analysis. When reporting on the validity of data used, acknowledges data editing procedures, including any imputation and missing data mechanisms. Reports the limitations of statistical inference and possible sources of error. In publications, reports, or testimony, identifies who is responsible for the statistical work if it would not otherwise be apparent. Reports the sources and assessed adequacy of the data, accounts for all data considered in a study, and explains the sample(s) actually used. Clearly and fully reports the steps taken to preserve data integrity and valid results. Where appropriate, addresses potential confounding variables not included in the study. In publications and reports, conveys the findings in ways that are both honest and meaningful to the user/reader. This includes tables, models, and graphics. In publications or testimony, identifies the ultimate financial sponsor of the study, the stated purpose, and the intended use of the study results. When reporting analyses of volunteer data or other data that may not be representative of a defined population, includes appropriate disclaimers and, if used, appropriate weighting. To aid peer review and replication, shares the data used in the analyses whenever possible/allowable and exercises due caution to protect proprietary and confidential data, including all data that might inappropriately reveal respondent identities. Strives to promptly correct any errors discovered while producing the final report or after publication. As appropriate, disseminates the correction publicly or to others relying on the results.

C. Responsibilities to Science/Public/Funder/Client

The ethical statistician supports valid inferences, transparency, and good science in general, keeping the interests of the public, funder, client, or customer in mind (as well as professional colleagues, patients, the public, and the scientific community).
The ethical statistician:
To the extent possible, presents a client or employer with choices among valid alternative statistical approaches that may vary in scope, cost, or precision. Strives to explain any expected adverse consequences of failure to follow through on an agreed-upon sampling or analytic plan. Applies statistical sampling and analysis procedures scientifically, without predetermining the outcome. Strives to make new statistical knowledge widely available to provide benefits to society at large and beyond his/her own scope of applications. Understands and conforms to confidentiality requirements of data collection, release, and dissemination and any restrictions on its use established by the data provider (to the extent legally required), protecting use and disclosure of data accordingly. Guards privileged information of the employer, client, or funder.

D. Responsibilities to Research Subjects

The ethical statistician protects and respects the rights and interests of human and animal subjects at all stages of their involvement in a project. This includes respondents to the census or to surveys, those whose data are contained in administrative records, and subjects of physically or psychologically invasive research.

The ethical statistician:

Keeps informed about and adheres to applicable rules, approvals, and guidelines for the protection and welfare of human and animal subjects.
Strives to avoid the use of excessive or inadequate numbers of research subjects—and excessive risk to research subjects (in terms of health, welfare, privacy, and ownership of their own data)—by making informed recommendations for study size. Protects the privacy and confidentiality of research subjects and data concerning them, whether obtained from the subjects directly, other persons, or existing records. Anticipates and solicits approval for secondary and indirect uses of the data, including linkage to other data sets, when obtaining approvals from research subjects and obtains approvals appropriate to allow for peer review and independent replication of analyses.
Knows the legal limitations on privacy and confidentiality assurances and does not over-promise or assume legal privacy and confidentiality protections where they may not apply.
Considers whether appropriate research-subject approvals were obtained before participating in a study involving human beings or organizations before analyzing data from such a study and while reviewing manuscripts for publication or internal use. The statistician considers the treatment of research subjects (e.g., confidentiality agreements, expectations of privacy, notification, consent, etc.) when evaluating the appropriateness of the data source(s). In contemplating whether to participate in an analysis of data from a particular source, refuses to do so if participating in the analysis could reasonably be interpreted by individuals who provided information as sanctioning a violation of their rights.

Recognizes any statistical descriptions of groups may carry risks of stereotypes and stigmatization. Statisticians should contemplate, and be sensitive to, the manner in which information is framed to avoid disproportionate harm to vulnerable groups.

E. Responsibilities to Research Team Colleagues

Science and statistical practice are often conducted in teams made up of professionals with different professional standards. The statistician must know how to work ethically in this environment.

The ethical statistician:

Recognizes other professions have standards and obligations, research practices and standards can differ across disciplines, and statisticians do not have obligations to standards of other professions that conflict with these guidelines.
Ensures all discussion and reporting of statistical design and analysis is consistent with these guidelines.
Avoids compromising scientific validity for expediency.
Strives to promote transparency in design, execution, and reporting or presenting of all analyses.

F. Responsibilities to Other Statisticians or Statistics Practitioners

The practice of statistics requires consideration of the entire range of possible explanations for observed phenomena, and distinct observers drawing on their own unique sets of experiences can arrive at different and potentially diverging judgments about the plausibility of different explanations. Even in adversarial settings, discourse tends to be most successful when statisticians treat one another with mutual respect and focus on scientific principles, methodology, and the substance of data interpretations.
Out of respect for fellow statistical practitioners, the ethical statistician:
Promotes sharing of data and methods as much as possible and as appropriate without compromising propriety. Makes documentation suitable for replicate analyses, metadata studies, and other research by qualified investigators.
Helps strengthen the work of others through appropriate peer review; in peer review, respects differences of opinion and assesses methods, not individuals. Strives to complete review assignments thoroughly, thoughtfully, and promptly.
Instills in students and non-statisticians an appreciation for the practical value of the concepts and methods they are learning or using.
Uses professional qualifications and contributions as the basis for decisions regarding statistical practitioners’ hiring, firing, promotion, work assignments, publications and presentations, candidacy for offices and awards, funding or approval of research, and other professional matters.

G. Responsibilities Regarding Allegations of Misconduct

The ethical statistician understands the differences between questionable statistical, scientific, or professional practices and practices that constitute misconduct. The ethical statistician avoids all of the above and knows how each should be handled.

The ethical statistician:

Avoids condoning or appearing to condone statistical, scientific, or professional misconduct.
Recognizes that differences of opinion and honest error do not constitute misconduct; they warrant discussion, but not accusation.
Knows the definitions of, and procedures relating to, misconduct. If involved in a misconduct investigation, follows prescribed procedures.
Maintains confidentiality during an investigation, but discloses the investigation results honestly to appropriate parties and stakeholders once they are available.
Following an investigation of misconduct, supports the appropriate efforts of all involved—including those reporting the possible scientific error or misconduct—to resume their careers in as normal a manner as possible.
Avoids, and acts to discourage, retaliation against or damage to the employability of those who responsibly call attention to possible scientific error or to scientific or other professional misconduct.

H. Responsibilities of Employers, Including Organizations, Individuals, Attorneys, or Other Clients Employing Statistical Practitioners

Those employing any person to analyze data are implicitly relying on the profession's reputation for objectivity. However, this creates an obligation on the part of the employer to understand and respect statisticians' obligation of objectivity.

Those employing statisticians are expected to:

Recognize that the ethical guidelines exist and were instituted for the protection and support of the statistician and the consumer alike.
Maintain a working environment free from intimidation, including discrimination based on personal characteristics; bullying; coercion; unwelcome physical (including sexual) contact; and other forms of harassment.

Recognize that valid findings result from competent work in a moral environment. Employers, funders, or those who commission statistical analysis have an obligation to rely on the expertise and judgment of qualified statisticians for any data analysis. This obligation may be especially relevant in analyses known or anticipated to have tangible physical, financial, or psychological effects.

Recognize the results of valid statistical studies cannot be guaranteed to conform to the expectations or desires of those commissioning the study or the statistical practitioner(s).

Recognize it is contrary to these guidelines to report or follow only those results that conform to expectations without explicitly acknowledging competing findings and the basis for choices regarding which results to report, use, and/or cite.

Recognize the inclusion of statistical practitioners as authors or acknowledgement of their contributions to projects or publications requires their explicit permission because it implies endorsement of the work.

Support sound statistical analysis and expose incompetent or corrupt statistical practice.

Strive to protect the professional freedom and responsibility of statistical practitioners who comply with these guidelines.

11. PROGRAMMING PLANS (USE OF PYTHON, R, etc.)

Python 3.8 was used for dataset analysis.

Step 1 – Input: Loading Source data file (loan_prediction)

Step 2 – Data preparation and preprocessing:

Clean Missing Data –
Categorical variables that have missing values were replaced with mode
Numeric variables that have missing values were replaced with median.
Median was used instead of mean as outliers are present in the dataset.

Removal of Noise Variables –
Variables like Loan_ID and Education were dropped from the dataset.

Creation of New Variables –
Total Income variable was created
Total_Income = ApplicantIncome + CoapplicantIncome

Defined weightage factors for Model-1 and Model-2 –

Parameters	Weightage factor for Model - 1	Weightage factor for Model - 2
Loan Amount	10	10
Loan Amount Term (Number of months to pay back the loan)	8	8
Applicant Income	8	8
Property_Area (Net Assets)	8	8
Total Income (applicant income + co-applicant income)	8	8
Credit_History	6	6
Marital Status	7	5
Self employed	7	5
Gender	7	5
Number of dependents	4	4

By adding these weightage factors to these parameters, new variables were created for analysis. The following analysis variables were created for model 1 and model 2:

These articles were used as reference for deciding the weightage factors of gender and marital status for Model-1 and Model-2.

https://www.sciencedirect.com/science/article/abs/pii/S0378426610002669

https://libproxy.library.unt.edu:6995/global/article/GALE%7CA164595268?u=txshracd2679&sid=summon

https://www.fdic.gov/regulations/compliance/manual/4/iv-1.1.pdf

Step 3 – Analysis
Model - 1 and Model - 2 were created and used for analysis based on different weightage factors
Split the data into train and test in the ratio 75 -25.
Logistic Regression, K-Fold Cross Validation, Confusion Matrix, Correlation Matrix and Feature Importance were performed.

12. REFERENCES USED IN THE COURSE OF YOUR ANALYSES CAN BE FOUND IN APPENDICES UNDER "APPENDIX B"

13. APPENDICES

To create your appendices, do the following: 1) add a blank page after this page here and label it "APPENDIX A" and then after that page add the pdf document from Canvas called "Editable Ethical Statement for Each Firm"; and, 2) add a blank page after the ethical statement page and label it "APPENDIX B" and then after that page add a page with the heading "REFERENCE" that will contain all of the references (statistics textbooks, journal articles, etc. that you use to create your SAP and to do your analyses. When you do your SAP for projects 1, 2, & 3, you will likely have additional references that you didn't have for your general SAP that you turn in February 12.

Appendix – A

OUR FIRM'S ETHICAL CODE

We pledge in writing to abide by the American Statistical Association's (ASA) and INFORMS' Codes of Ethics. Our adherence to these Codes signifies voluntary assumption of self-discipline. As the professional associations for our firm in the United States, the ASA and INFORMS requires adherence to their Codes of Ethics as a condition of membership. The standards of conduct set forth in these Codes provide basic principles in the ethical practice of data analysis consulting. The purpose of these Codes is to help us maintain our professionalism and adhere to high ethical standards in the conduct of providing services to clients and in our dealings with our colleagues and the public. Our individual judgment requires we apply these principles. We are liable to disciplinary action under the ASA's and INFORMS' Rules of Procedure for Enforcement of this Code if our conduct is found by the ASA's or INFORMS' respective Ethics Committees to be in violation of their respective Codes or to bring discredit to the profession or to ASA and INFORMS.

Our Commitment to Our Clients

1) We will serve our clients with integrity, competence, independence, objectivity, and professionalism.

2) We will mutually establish with our client's realistic expectations of the benefits and results of our services.

3) We will only accept assignments for which we possess the requisite experience and competence to perform and will only assign staff or engage colleagues with the knowledge and expertise needed to serve our clients effectively.

4) Before accepting any engagement, we will ensure that we have worked with our clients to establish a mutual understanding of the objectives, scope, work plan, and fee arrangements.

5) We will treat appropriately all confidential client information that is not public knowledge, take reasonable steps to prevent it from access by unauthorized people, and will not take advantage of proprietary or privileged information, either for use by ourselves, the client's firm, or another client, without the client's permission.

6) We will avoid conflicts of interest or the appearance of such and will immediately disclose to the client circumstances or interests that we believe may influence my judgment or objectivity.

7) We will offer to withdraw from a consulting assignment when we believe my objectivity or integrity may be impaired.

8) We will refrain from inviting an employee of an active or inactive client to consider alternative employment without prior discussion with the client.

Our Commitment to Fiscal Integrity

9) We will agree in advance with a client on the basis for fees and expenses and will charge fees that are reasonable and commensurate with the services delivered and the responsibility accepted.

10) We will not accept commissions, remuneration, or other benefits from a third party in connection with the recommendations to a client without that client's prior knowledge and consent, and will disclose in advance any financial interests in goods or services that form part of such recommendations.

Our Commitment to the Public and the Profession

11) If within the scope of my engagement, we will report to appropriate authorities within or external to the client organization any occurrences of malfeasance, dangerous behavior, or illegal activities.

12) We will respect the rights of consulting colleagues and consulting firms and will not use their proprietary information or methodologies without permission.

13) We will represent the profession with integrity and professionalism in my relations with our clients, colleagues, and the general public.

14) We will not advertise our services in a deceptive manner nor misrepresent or denigrate individual consulting practitioners, consulting firms, or the consulting profession.

15) If we perceive a violation of the Code, we will report it to the APA and INFORMS and will promote adherence to the Code by other member consultants working on our behalf.

APPENDIX – B

Logistic Regression:
http://utstat.toronto.edu/~ali/papers/creditworthinessProject.pdf

https://www.researchgate.net/publication/290994615_Predicting_the_Probability_of_Loan-Default_An_Application_of_Binary_Logistic_Regression

https://machinelearningmastery.com/k-fold-cross-validation/

Rule Point Informatica 6.2:
https://docs.informatica.com/complex-event-processing/rulepoint/6-2/business-process-management-use-case-example/business-process-management-use-case.html

http://biostat.mc.vanderbilt.edu/wiki/Main/PreStatAnalysisPlan

https://www.sciencedirect.com/science/article/abs/pii/S0378426610002669

https://libproxy.library.unt.edu:6995/global/article/GALE%7CA164595268?u=txshracd2679&sid=summon

https://www.fdic.gov/regulations/compliance/manual/4/iv-1.1.pdf

http://databoosting.com/loan-prediction-problem/

https://www.kaggle.com/ninzaami/loan-predication

https://www.realtor.com/advice/finance/how-long-does-it-take-to-get-a-mortgage/

https://towardsdatascience.com/cross-validation-430d9a5fee22

https://towardsdatascience.com/cross-validation-430d9a5fee22

https://towardsdatascience.com/cross-validation-430d9a5fee22

http://databoosting.com/loan-prediction-problem/

https://towardsdatascience.com/financial-data-analysis-51e7275d0ae

https://www.myfico.com/credit-education/fico-scores-vs-credit-scores https://www.badcredit.org/how-to/5-tips-improve-chances-loan-approval/

https://defisolutions.com/defi-insight/2018/05/23/process-improvement-ideas-in-banking/

https://www.inc.com/encyclopedia/credit-evaluation-and-approval.html

https://defisolutions.com/defi-insight/2018/05/23/process-improvement-ideas-in-banking/

https://www.omega-performance.com/omega-decision-strategy/