# A Project REPORT on

**Loan application screening and aspect of Fair lending practice**

SUBMITTED BY

**Kishor Lagad**

M.Sc. [Semester IV]



# Dr. D.Y. Patil Unitech Society’s

# Dr. D. Y. Patil Arts, Commerce and Science College Pimpri, Pune-18

SUBMITTED TO

**SAVITRIBAI PHULE PUNE UNIVERSITY**

ACADEMIC YEAR

**2021-2022**

**DECLARATION**

Mr. **Kishor Lagad** of **Dr. D. Y. Patil Arts, Commerce and Science College** of S.Y.M.Sc. [Semester IV] hereby declare that I have completed my project, titled ‘**Loan application screening and aspect of Fair lending practice’** in the academic year 2021 – 2022. The information submitted hereby declare that all facts and figures are true to my knowledge.

**Signature of Student**

**Kishor Lagad**

**Certificate**

This is to certify that Kishor Lagad (S.Y. Msc statistics) of the department of statistics of this institute has satisfactorily completed the project. The project entitled “**Loan application screening and aspect of Fair lending practice**” was carried out under my direct supervision.

Ms Komal Kothawade Mrs Deepali Akolkar Project Guide Head of Department

Internal Examiner External Examiner

|  |  |  |
| --- | --- | --- |
| Sr.no | Title | Page no. |
| 1 | Acknowledgement |  |
| 2 | Motivation |  |
| 3 | Abstract |  |
| 4 | Introduction |  |
| 5 | Terminologies involved |  |
| 6 | Objectives |  |
| 7 | Explanatory Analysis |  |
| 8 | Model building |  |
| 9 | Conclusion |  |
| 10 | References |  |

**Index**

**ACKNOWLEDGEMENT**

First of all I would like to take this opportunity to thank the Pune University for having projects as a part of the M.S. curriculum.

I wish to express my heartfelt gratitude to the following individuals who have played a crucial role in the research for this project. Without their active cooperation the preparation of this project could not have been completed within the specified time limit.

The first person I would like to acknowledge is my project `guide **Asst. Prof. Komal Kothawade** who supported me throughout this project with utmost cooperation and patience. I am very much thankful to them for sparing their precious time for me and for helping me in doing this project

Finally, I would like to thank project partners and all my friends who have helped in all possible ways in making this project presentable.

Last but not least I would like to thank the almighty for always helping me.

**MOTIVATION**

One of the most significant financial items is the loan. All banks are attempting to devise successful business tactics for convincing clients to apply for loans.

However, some consumers act in an unfavourable manner once their application has been granted. To avoid this issue, banks must develop systems for predicting client behaviour.

Machine learning techniques, which are commonly employed in banking, work well for this purpose. We will be working on loan behaviour prediction using machine learning models in this project.

**Abstract**

The cost of assets is rising every day, and the amount of money needed to buy a complete asset is quite significant. As a result, you won't be able to buy it with your funds. Applying for a loan is the simplest approach to obtain the finances needed. However, obtaining a loan is a lengthy procedure. The application must go through a number of processes before being granted, and approval is not guaranteed.

Many loan prediction models have been devised to reduce the time it takes for a loan to be approved and to reduce the risk associated with the loan. The goal of this research was to analyse multiple Loan Prediction Models and determine which one had the least amount of error and could be used by banks in the real world to forecast whether a loan should be authorised or denied while considering the risk factor.

After comparing and analysing the models, it was discovered that the prediction model was the most accurate and fitting of all. This can save time and money by lowering the amount of time and labour needed to approve loans and weeding out the best applicants for lending.

**Introduction**

To forecast a result, a Prediction Model employs data mining, statistics, and probability. Predictors are factors in any model that are expected to impact future outcomes. After gathering data from multiple sources, a statistical model is created. It can utilise a basic linear equation or a sophisticated neural network that has been mapped with the help of complicated software.

As additional data becomes available, the model becomes more sophisticated, and the error lowers, allowing it to forecast with the least amount of risk and in the shortest amount of time possible. The Prediction Model is beneficial to the Banks benefit by reducing the risk involved with the loan approval processes, while applicants benefit by reducing the time it takes to complete the process. The project’s main goal is to evaluate Loan Prediction Models developed using multiple algorithms and select the best one that may reduce loan approval time and risk. It is done by forecasting whether or not the loan may be granted to that person based on a variety of factors such as credit score, income, age, marital status, gender, and so on. The prediction model benefits both the applicant and the bank by lowering risk and lowering the number of defaulters.

In the present scenario, a loan needs to be approved manually by a representative of the bank which means that person will be responsible for whether the person is eligible for the loan or not and also calculating the risk associated with it. As it is done by a human it is a time consuming process and is susceptible to errors. If the loan is not repaid, then it accounts as a loss to the bank and banks earn most of their profits by the interest paid to them. If the banks lose too much money, then it will result in a banking crisis. These banking crisis affects the economy of the country. So it is very important that the loan should be approved with the least amount of error in risk calculation while taking up as the least time possible. Loan prediction model is required that can predict quickly whether the loan can be passed or not with the least amount of risk possible.

**Context**

For a financial organization in the business of lending money, understanding risk profile of a customer is central to lending money in the form of a structured product. Improper understanding of the risk profile of a customer carries two types of risks i.e. risk of losing money by lending it to an unworthy customer (who either intentionally or due to circumstances can’t pay it back) and second is opportunity cost of declining credit to a worthy customer.

While understanding the risk profile is important, it is critical that lenders don’t base their decisions which may come across as biased and unfair on account of a customer’s race, religion, nation of origin, gender, and age etc. In US, the [Fair Housing Act](https://www.hud.gov/program_offices/fair_housing_equal_opp) (FHA) and the [Equal Credit Opportunity Act](http://www.ftc.gov/bcp/edu/pubs/consumer/credit/cre15.shtm) (ECOA) protect consumers by prohibiting unfair and discriminatory practices while lending and ensure the same lending opportunity to everyone.

Risk profile should therefore be focused on relevant qualitative and quantitative aspects of a customer such as past credit behaviour, credit worthiness and loan related attributes for any past or present lending relationship with other lenders

**Data Content:**

Inside the following dataset, there are around 42,000 loan accounts, booked in between 2007 to 2011, with some demographic and credit bureau attributes along with current loan status. Credit bureau is an agency, which collects credit performance related information for every customer from banks. This information can be summarized into attributes which can tell about any customer’s historical credit behaviour. Loan status tells about how the loan is performing currently, i.e. if the loan is delinquent or the customer has missed any payment or if the loan is already fully paid or charged off or the loan is doing well and paying on time.

**Terminologies**

### **Logistic Regression**

This is a classification algorithm which uses a logistic function to predict binary outcome (True/False, 0/1, Yes/No) given an independent variable. The aim of this model is to find a relationship between features and probability of particular outcome.

### **Decision Trees**

This is a supervised machine learning algorithm mostly used for classification problems. All features should be discretized in this model, so that the population can be split into two or more homogeneous sets or subsets. This model uses a different algorithm to split a node into two or more sub-nodes. With the creation of more sub-nodes, homogeneity and purity of the nodes increases with respect to the dependent variable.

### **Random Forest**

This is a tree based ensemble model which helps in improving the accuracy of the model . It combines a large number of Decision trees to build a powerful predicting model. It takes a random sample of rows and features of each individual tree to prepare a decision tree model. Final prediction class is either the mode of all the predictors or the mean of all the predictors.

### **XGBoost**

This algorithm only works with the quantitative variable. It is a gradient boosting algorithm which forms strong rules for the model by boosting weak learners to a strong learner. It is a fast and efficient algorithm which recently dominated machine learning because of its high performance and speed.

### **Pandas**

Pandas is a Python package to work with structured and time series data.The data from various file formats such as csv, json, sql etc can be imported using Pandas. It is a powerful open source tool used for data analysis and data manipulation operations such as data cleaning, merging, selecting as well wrangling

### **Sklearn**

This python library is helpful for building machine learning and statistical models such as clustering, classification, regression etc. Though it can be used for reading, manipulating and summarizing the data as well, better libraries are there to perform these functions

**Ensemble Classifier**

Ensemble learning helps improve machine learning results by combining several models. This approach allows the production of better predictive performance compared to a single model. Basic idea is to learn a set of classifiers (experts) and to allow them to vote.

**Underfitting**

A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data.

**Overfitting**

A statistical model is said to be overfitted when we train it with a lot of data (just like fitting ourselves in oversized pants!).

**Feature**

A feature is an attribute that has an impact on a problem or is useful for the problem, and choosing the important features for the model is known as feature selection.

**Id**

An unique identifier for each loan

**member\_id**

An unique identifier for each borrower

**loan\_amnt**

The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.

**funded\_amnt**

The total amount committed to that loan at that point in time.

**funded\_amnt\_inv**

The total amount committed by investors for that loan at that point in time.

**term**

Term of the loan. Values are in months and can be either 36 or 60.

**installment**

The monthly payment owed by the borrower if the loan originates.

**emp\_title**

The job title supplied by the Borrower when applying for the loan.

**emp\_length**

Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.

**home\_ownership**

The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.

**annual\_inc**

The self-reported annual income provided by the borrower during registration.

**verification\_status**

Indicates if income was verified, not verified, or if the income source was verified

**issue\_d**

The month which the loan was funded

**loan\_status**

Current status of the loan

**desc**

Loan description provided by the borrower

**purpose**

A category provided by the borrower for the loan request.

**title**

The loan title provided by the borrower

**dti**

A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested loan, divided by the borrower's self-reported monthly income.

**open\_acc**

The number of open credit lines in the borrower's credit file.

**revol\_bal**

Total credit revolving balance

**total\_acc**

The total number of credit lines currently in the borrower's credit file

**out\_prncp\_inv**

Remaining outstanding principal for portion of total amount funded by investors

**total\_pymnt**

Payments received to date for total amount funded

**total\_pymnt\_inv**

Payments received to date for portion of total amount funded by investors

**total\_rec\_int**

Interest received to date

**revol\_utilization**

Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.

**number\_bc\_gt\_75**

number of bankcard accounts > 75% of limit.

**fico\_score**

Origination Fico score or the fico score at the point of loan origination

**iti**

Loan value to income ratio

**race\_name**

Race of the borrower

**gender**

Gender of the borrower

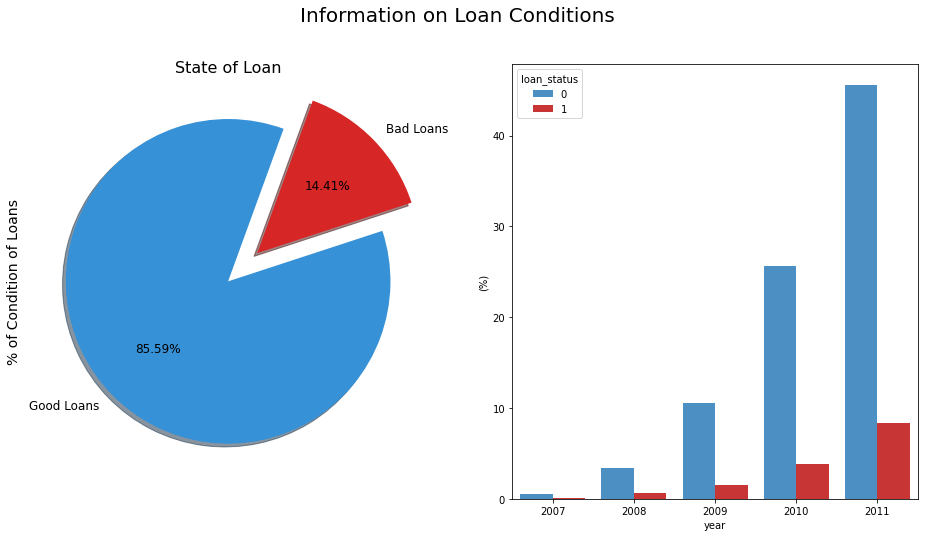
**Objectives**

1.As a decision maker, looking into these given booked accounts, how would you want to take a decision to approve/ decline the same applications again in future? What would be your screening criterion(s), if any?

2.Please provide data driven rationale to substantiate the appropriateness of the characteristics you have identified in the screening logic

3. Given that you have a decision engine, which helps you to take a call based on the given data, how would you identify if your process has an unintentional bias and discriminate your customers based on some of the sensitive attributes, available in this dataset?

**Explanatory analysis**



For analysis purpose only the customers whose loan status is default or fully paid were taken into consideration and other were skipped.

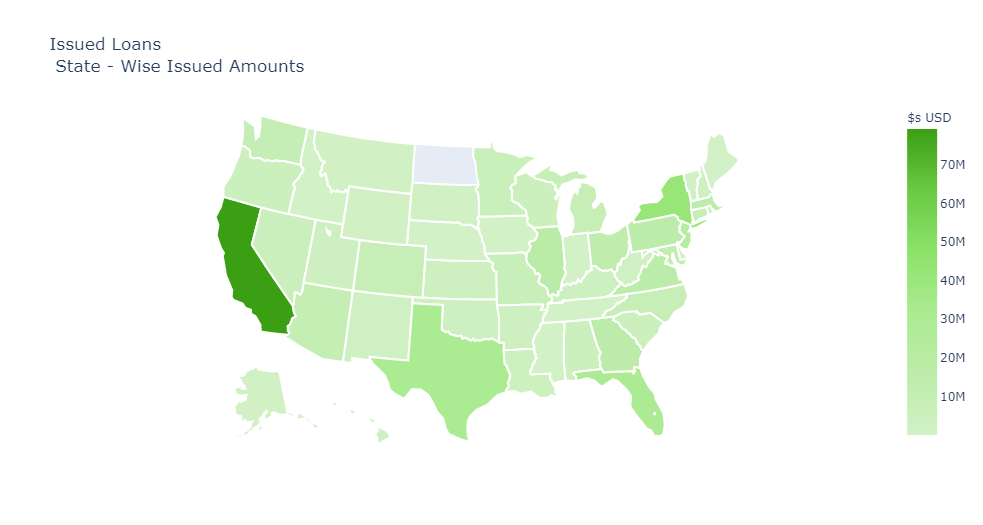
Out of total customers in considerations, approximately 14.41% were defaulters and this kind of loan were named as bad loan for convenience purpose and other 85.59% who paid fully were named as good loan (non defaulters)

It is essential to focus on what factors makes customer as defaulter of loan for that purpose we will first understand the distribution of Loan amount and default rates state wise in USA

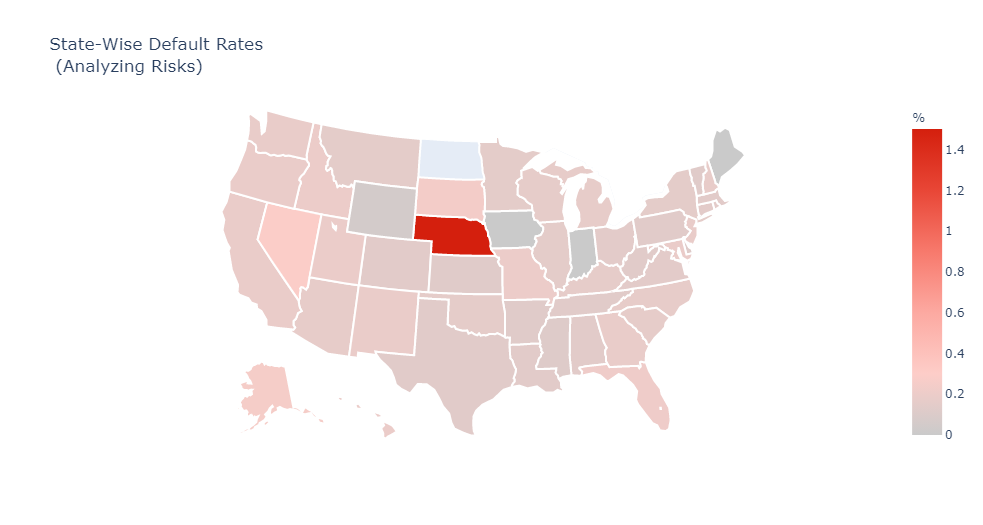
**Distribution of Loan amount and default rates state wise in USA**

California is the state with highest issued loan amount followed by Texas, New York, Florida. North Dakota has lowest issued loan amount among all the states in USA

Issued Loans (State wise issued amounts)



State wise default rates (Analysing risks)

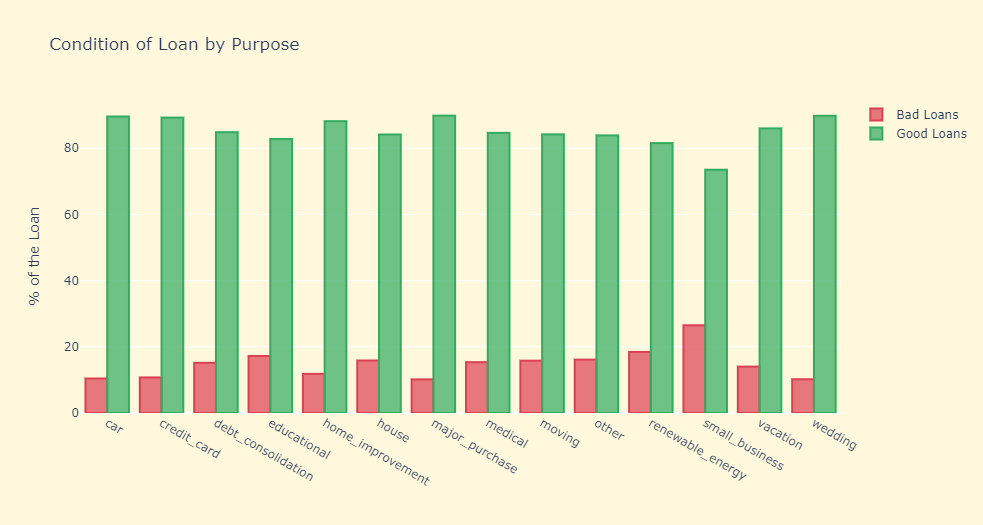


In case of default rate Nebraska is the state with highest default rate where North Dakota has lowest default rate (this can be due to lowest issued loan amount)

California the state with highest issued Loan amount has average default rate between 0.2 to 0.4

**Purpose of loan and default rate**

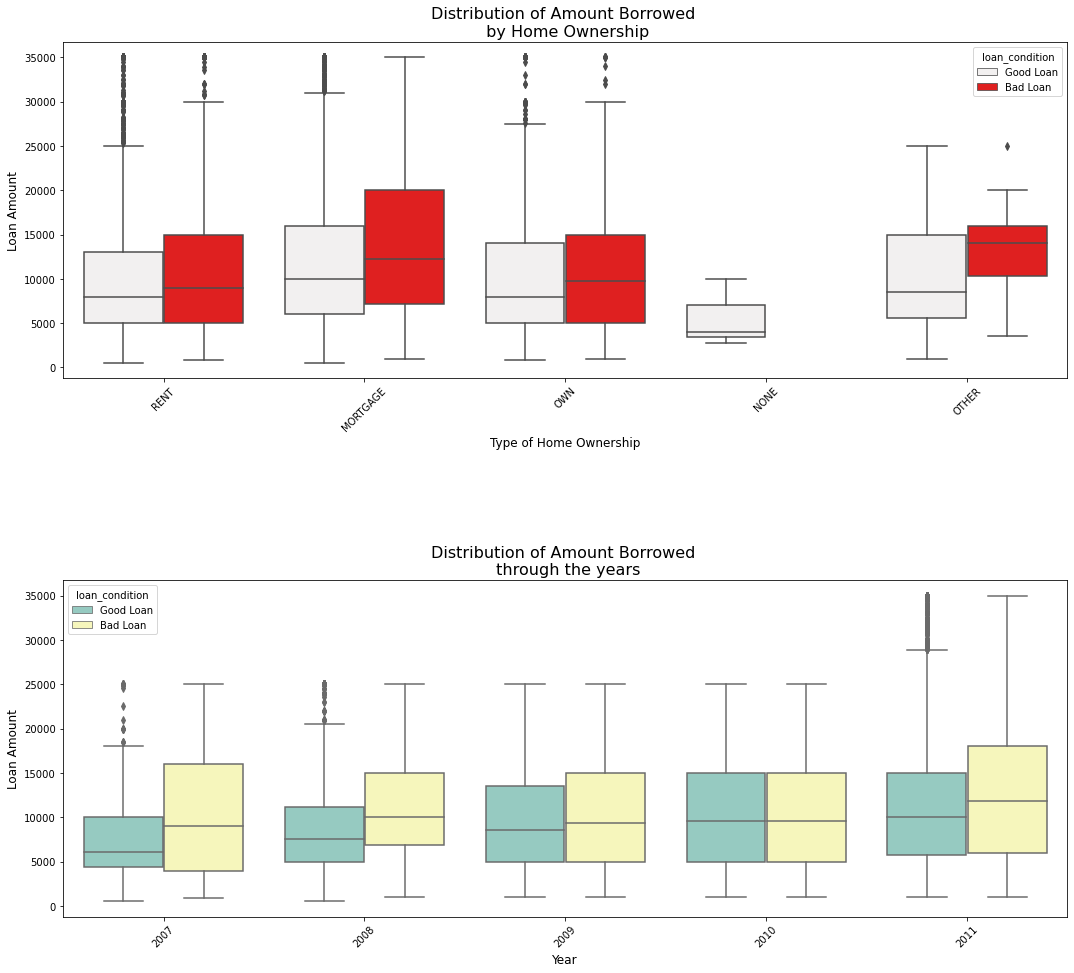
To understand Loan defaulters, we also need to focus on if the purpose of loan and Loan being defaulted has any relation or not.



Loan for small business has the highest bad loans as compared to other loan purposes, followed by renewable energy, educational, debt consolidation, house, medical, moving and so on.

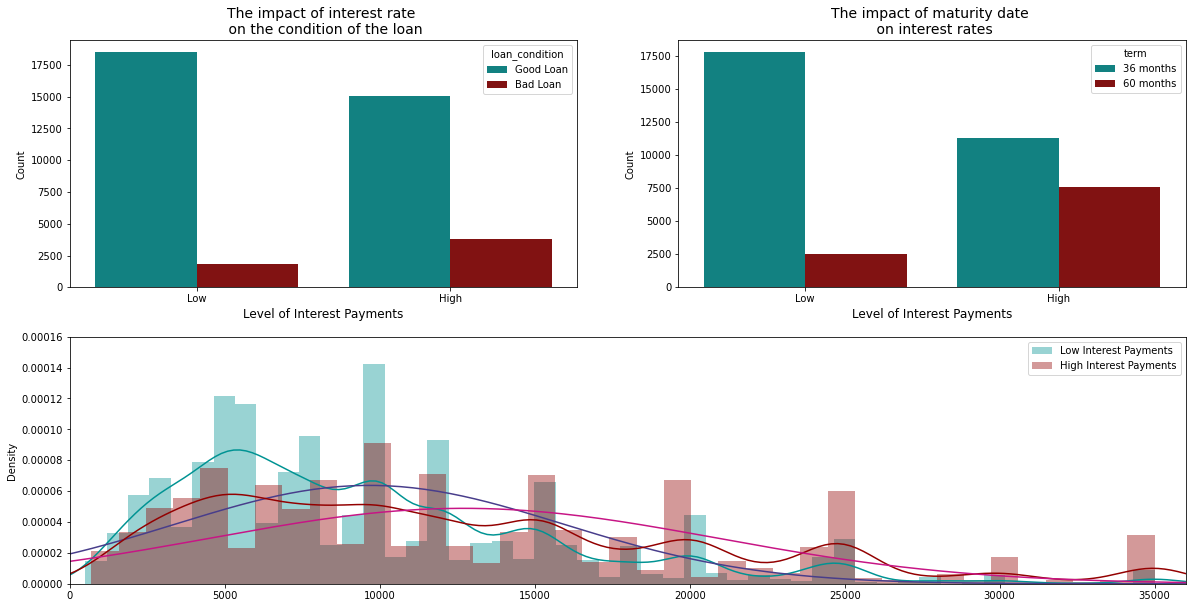
Lowest default rate is observed in case of car, credit card, major purpose, wedding etc.

**Home ownership and default rate**



In case of home ownership, no matter what is the status of home ownership, **bad loan has always high loan amount than good loan** also observed and this case is also observed in case of distribution of Amount borrowed through the years.

**Interest rate and maturity date**



This concludes, higher the Loan amount, higher the chances of Loan being defaulted.

Next factor in consideration is impact of Interest rate and maturity date on condition of loan. Higher the Interest rate higher the chances of Loan default as well as if the maturity date on interest rates is high (here in this case 60 months) has highest chances of loan being defaulted than the tow less maturity term (36 months)

**Model Building**

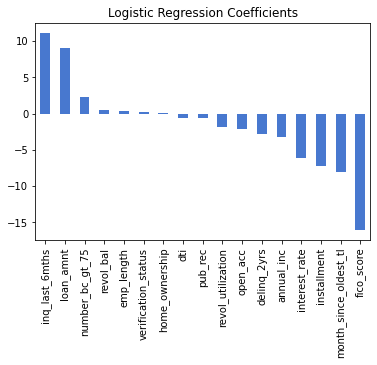
Data is being processed so that it can be determined how many values are missing from each column. The count of missing values present in the non-numerical attributes are processed and computed using the statistics.

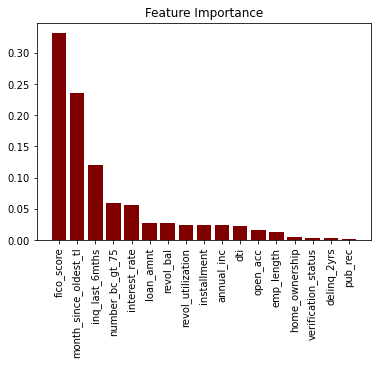
A generic function is defined whose input is a model that helps determine the Xgboost scores and accuracy using ensemble learning method

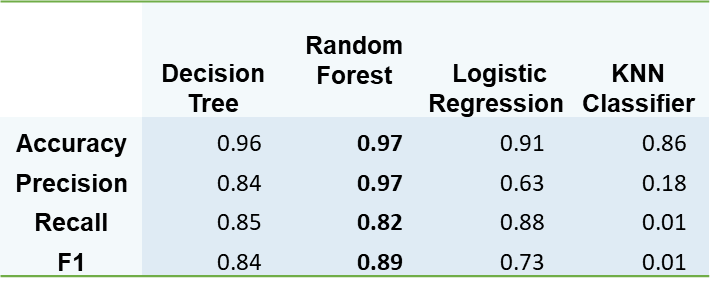
Model based on Random forest has an advantage that it can find the most important features that greatly affect the accuracy of the result among all the features using feature importance matrix.

Since 82% accuracy was observed in the previous results, the most important features are taken from the feature importance matrix now to avoid overfitting

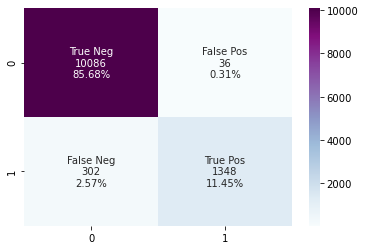
The predictive models based on Logistic Regression, Decision Tree and Random Forest, give the accuracy as 80.945%, 93.648% and 83.388% whereas the cross-validation is found to be 80.945%, 72.213% and 80.130% respectively. This shows that for the given dataset, the accuracy of model based on decision tree is highest but random forest is better at generalization even though it’s cross validation is not much higher than logistic regression.

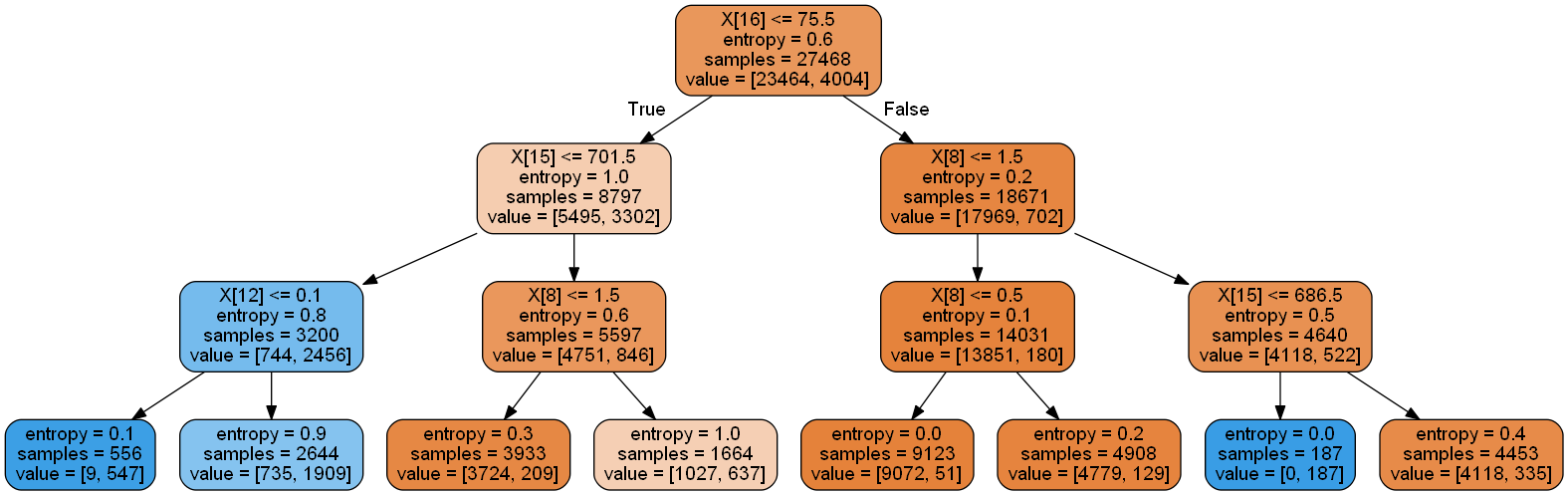
****

****

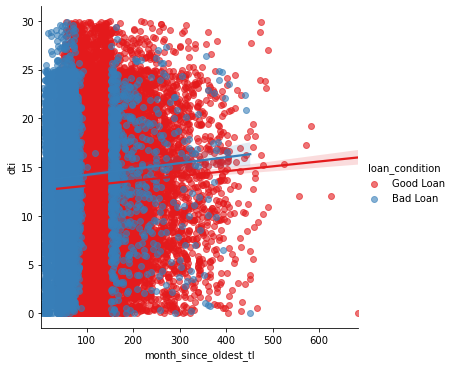
****

**The random forest algorithm outperforms all other classification models, with over 97% accuracy**

****

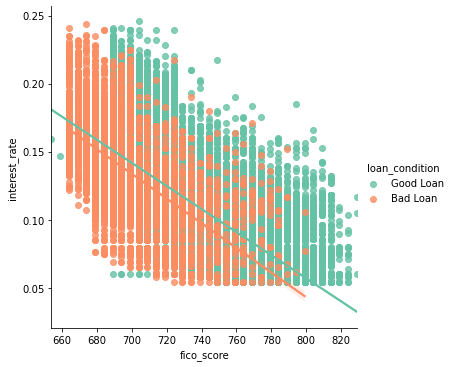
**Business plan**

* **We use the results from our random forest and logistic models to justify the criteria for loan approval.**
* **The most important determinants – FICO score, months since last trade line, inquiries in last 6 months, interest rates, and number of bank account cards > 75% of the limit**

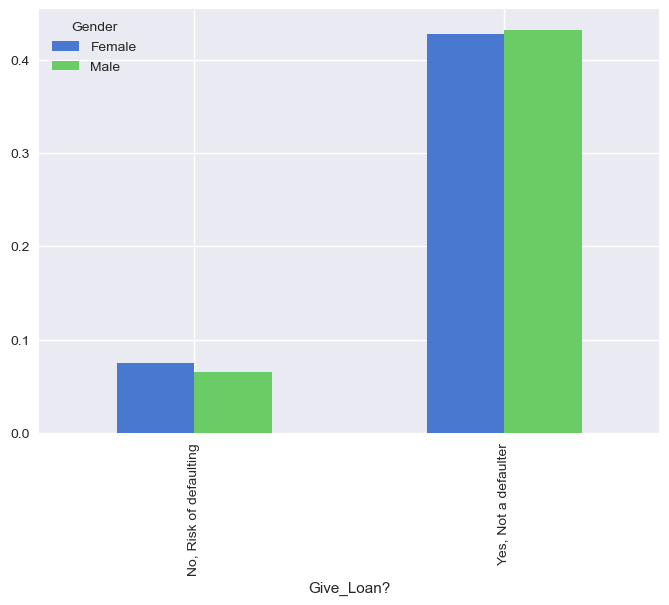
****

**Identifying a good loan:**

* + **Maintain a FICO score greater than 700**
  + **Keep interest rates low, below 15%**
  + **Ensure a DTI ratio below 10**
  + **Over 100 months since last tradeline**

****

**This also shows that some factors are relatively unimportant in determining the default risk on a loan, i.e. the length of employment, income verification status, and home ownership status, among others**

****

* **Two sensitive categories present in the data – gender and race**
* **Does our model unintentionally discriminate?**
* **Loans are distributed proportionately between different races and genders.**

****