# **Abstract**

Loans are the core business of banks. The main profit comes directly from the loan’s interest. The loan companies grant a loan after an intensive process of verification and validation. In our industry, banks have several products to sell however main supply of financial gain of any banks is on its credit line. in order that they will earn from interest of these loans which they credits. A bank’s profit or a loss depends to an oversized extent on loans i.e. whether or not the shoppers square measure return the loan or defaulting. By predicting the loan defaulters, the bank

can cut back its Non-performing Assets. This makes the study of this development vital. Previous analysis during this era has shown that there square measure such a lot of ways to check the problem of dominant loan default. however because the right predictions square measure vital for the maximization of profits, it is essential to check the character of the various ways and their comparison. An awfully vital approach in prognostic analytics is employed to check the matter of predicting loan defaulters (i) assortment of information, (ii) information cleansing and (iii) Performance analysis. Experimental tests found that XGBoost classifier works better in comparison to other machine learning algorithms in terms of loan prediction.

**Key Words**: Machine Learning, XGBoost, Logistic Regression, Loan Prediction.

# **Table of Contents**

1. Abstract
2. List of Figures
3. Introduction to the Problem Statement
4. Exploratory Data Analysis
5. Data Preprocessing
6. Feature Selection and PCA
7. Results and Conclusion

# **List of Figures**

* Figure 1: Null Values Percentage
* Figure 2: Countplots of Categorical Variables
* Figure 3: Normalized countplots of Ordinal features
* Figure 4: Distplot and Boxplot of Applicant Income
* Figure 5: Segregating Applicant Income by education levels.
* Figure 6: Co-Applicant income distribution
* Figure 7: Stacked Bar plot between Gender and Loan Status
* Figure 8: Stacked Bar plots of different categorical variable and target variables
* Figure 9: Stacked Bar plot between credit history and property area vs Loan status
* Figure 10: Mean of income values for target status
* Figure 11: Binning of income values for more detailed analysis
* Figure 12: Stacked Bar plot of Co-Applicant income vs Loan Status
* Figure 13: Correlation Heatmap
* Figure 14: Feature Importance Plot using Decision Tree classifier attributes

# **Introduction to Problem Statement**

Loan Prediction is very helpful for employees of banks as well as for the applicant also. The aim of this report is to provide a quick, immediate, and easy way to choose deserving applicants. Dream housing Finance Company deals in all loans. They have a presence across all urban, semi-urban, and rural areas. Customer first applies for a loan after that company or bank validates the customer eligibility for the loan. Company or bank wants to automate the loan eligibility process (real-time) based on customer details provided while filling the application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History, and others. This project has taken the data of previous customers of various banks to whom on a set of parameters loans were approved. So the machine learning model is trained on that record to get accurate results. Our main objective of this project is to predict the safety of loan, which in turn will help protect the bank from loan defaulters.

# **Exploratory Data Analysis**

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

The goal during EDA is to develop an understanding of your data. The easiest way to do this is to use questions as tools to guide your investigation. When you ask a question, the question focuses your attention on a specific part of your dataset and helps you decide which graphs, models, or transformations to make. Generate questions about your data. Search for answers by visualising, transforming, and modelling your data. Use what you learn to refine your questions and/or generate new questions. EDA is not a formal process with a strict set of rules.

More than anything, EDA is a state of mind.

During the initial phases of EDA you should feel free to investigate every idea that occurs to you. Some of these ideas will pan out, and some will be dead ends. As your exploration continues, you will home in on a few particularly productive areas that you’ll eventually write up and communicate to others.

**Dataset Columns**

The following points give a brief information about the information contained in the columns of the dataset used:

* Loan\_ID --------------> Unique Loan ID.
* Gender --------------> Male/ Female
* Married --------------> Applicant married (Y/N)
* Dependents ------------> Number of dependents
* Education -------------> Applicant Education (Graduate/ Under Graduate)
* Self\_Employed ---------> Self-employed (Y/N)
* ApplicantIncome -------> Applicant income
* CoapplicantIncome -----> Coapplicant income
* LoanAmount -----------> Loan amount in thousands
* Loan\_Amount\_Term ------> Term of a loan in months
* Credit\_History --------> Credit history meets guidelines
* Property\_Area ---------> Urban/ Semi-Urban/ Rural
* Loan\_Status -----------> Loan approved (Y/N)

**Null Values Analysis**

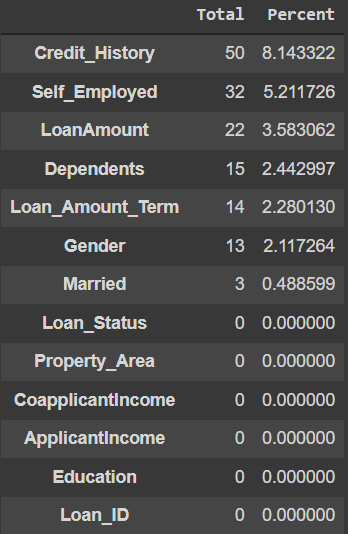
The image below shows the percentage of null values present in each feature of the dataset:

Fig. 1: Null Values Percentage

The table screenshot on the left implies we need to perform data imputation which will explain the different methods used for filling the null values of the features.

We can see there are three formats of data types:

* **object**: Object format means variables are categorical. Categorical variables in our dataset are Loan\_ID, Gender, Married, Dependents, Education, Self\_Employed, Property\_Area, Loan\_Status.
* **int64**: It represents the integer variables. ApplicantIncome is of this format.
* **float64**: It represents the variable that has some decimal values involved. They are also numerical

Now, let's visualize each variable separately. Different types of variables are Categorical, ordinal, and numerical.

* **Categorical features**: These features have categories (Gender, Married, Self\_Employed, Credit\_History, Loan\_Status)
* **Ordinal features**: Variables in categorical features having some order involved (Dependents, Education, Property\_Area)
* **Numerical features**: These features have numerical values (ApplicantIncome, Co-applicant income, LoanAmount, Loan\_Amount\_Term)

**Data Analysis of Independent Variable (Categorical)**

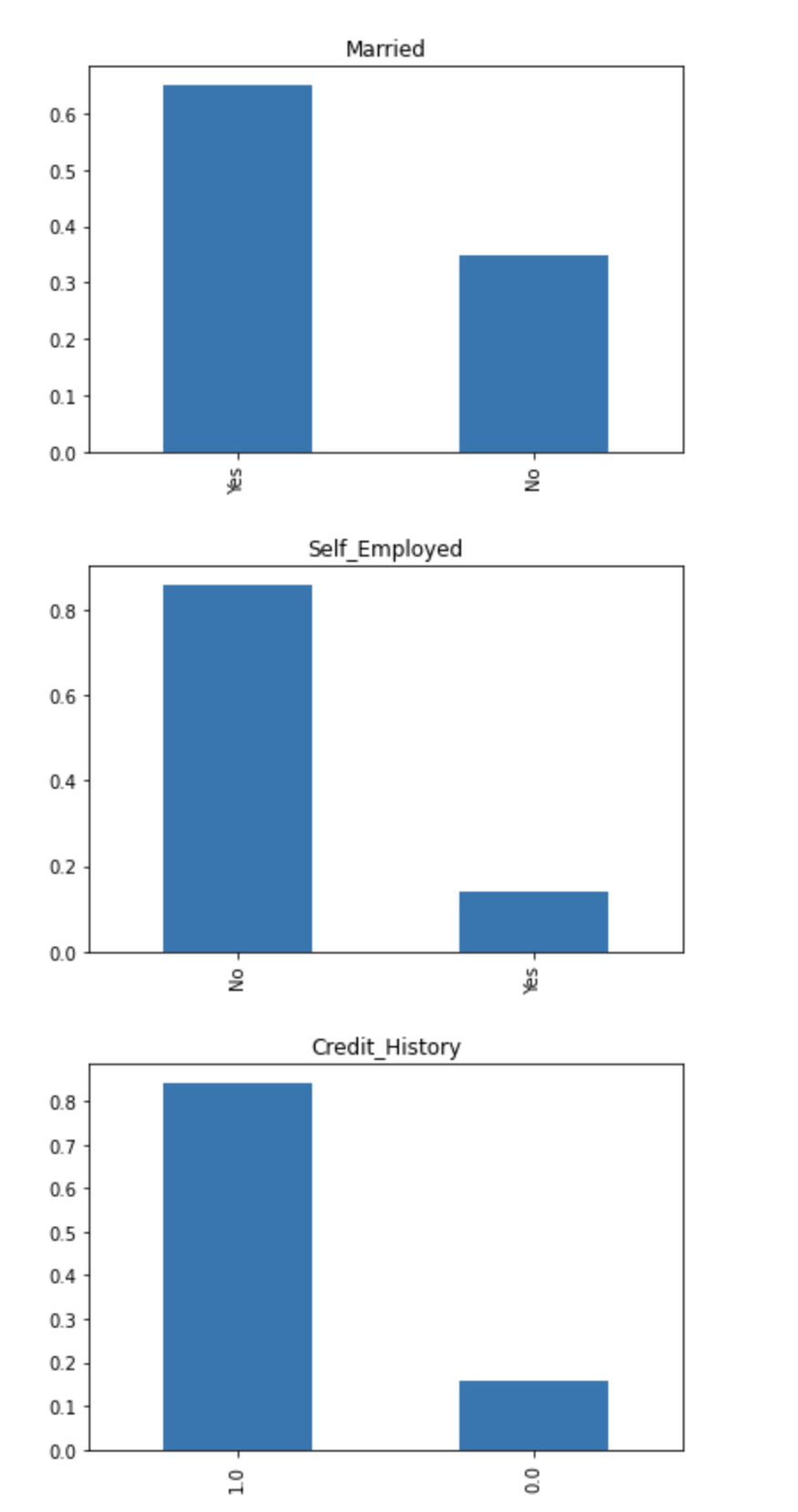


Fig. 2: Countplots of Categorical Variables

It can be inferred from the above bar plots that:

* 80% of applicants in the dataset are male.
* Around 65% of the applicants in the dataset are married.
* Around 15% of applicants in the dataset are self-employed.
* Around 85% of applicants have repaid their debts.

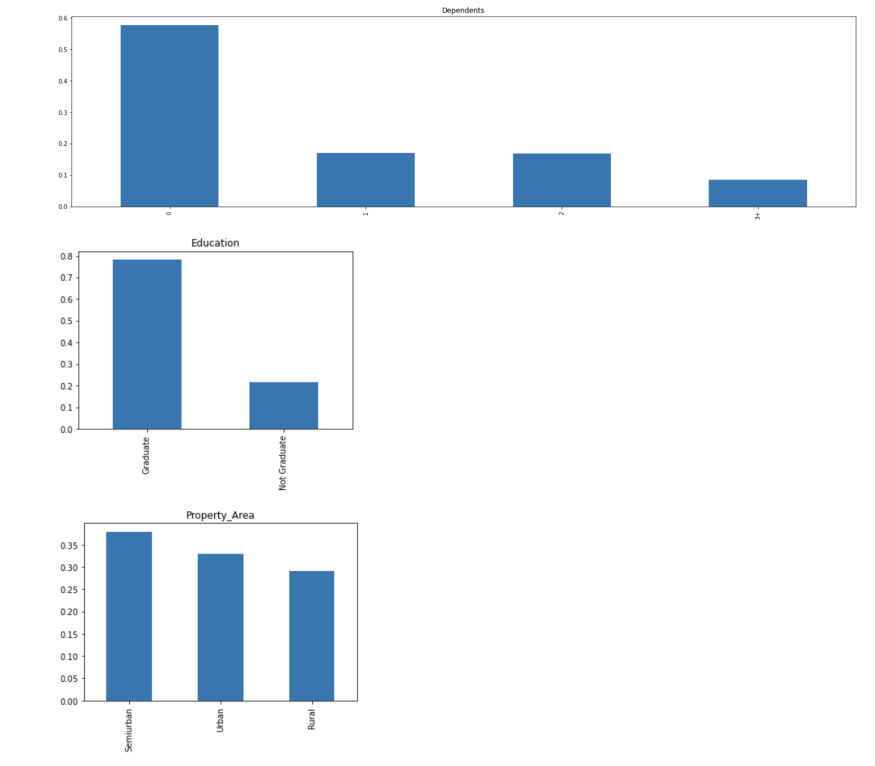
**Data Analysis of Independent Variable (Ordinal)**

Fig. 3: Normalized countplots of Ordinal features

The following inferences can be made from the above bar plots:

* Most of the applicants don't have any dependents.
* Around 80% of the applicants are graduates.
* Most of the applicants are from semi-urban area.

**Data Analysis of Independent Variable (Numerical)**

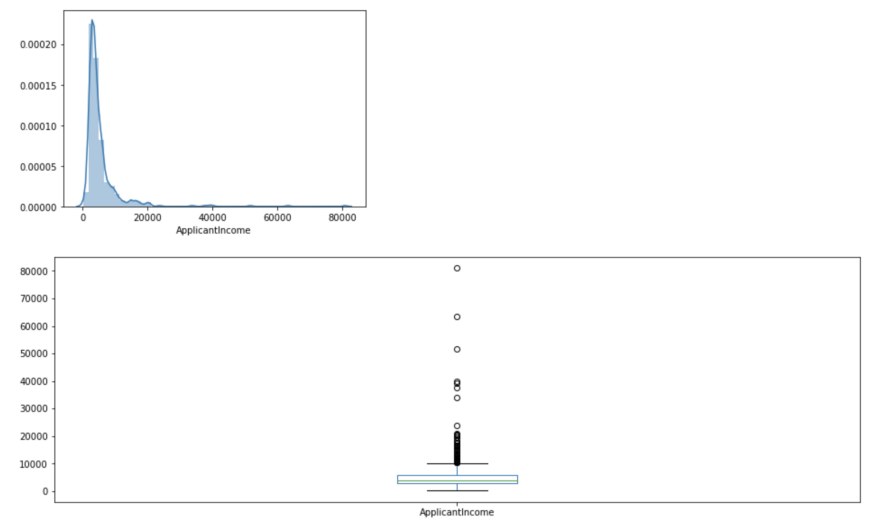


Fig. 4: Distplot and Boxplot of Applicant Income

It can be inferred that most of the data in the distribution of applicant income are towards the left which means it is not normally distributed. We will try to make it normal in later sections as algorithms work better if the data is normally distributed.

The boxplot confirms the presence of a lot of outliers/extreme values. This can be attributed to the income disparity in the society. Part of this can be driven by the fact that we are looking at people with different education levels. Let us segregate them by Education.

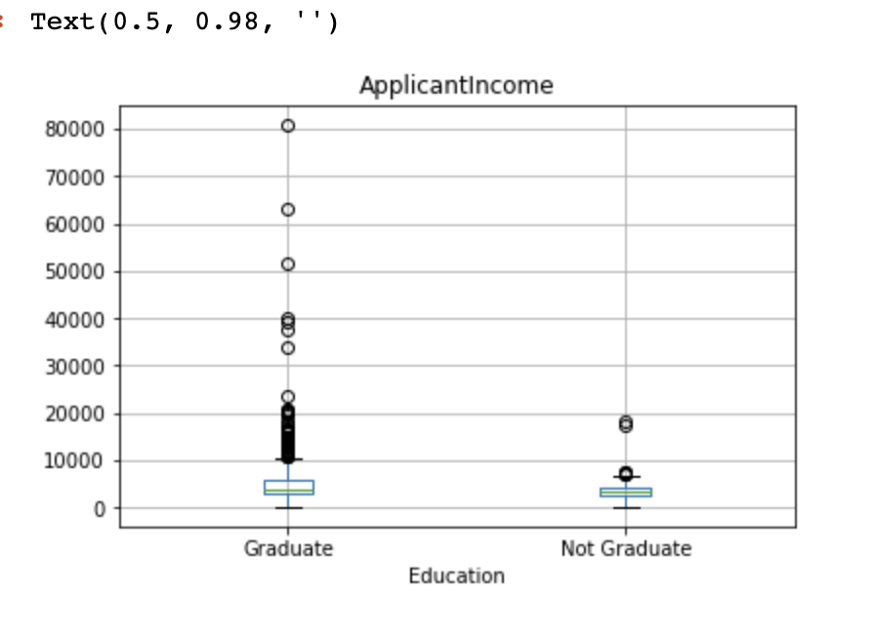


Fig. 5: Segregating Applicant Income by education levels.

We can see that there are a higher number of graduates with very high incomes, which are appearing to be outliers.

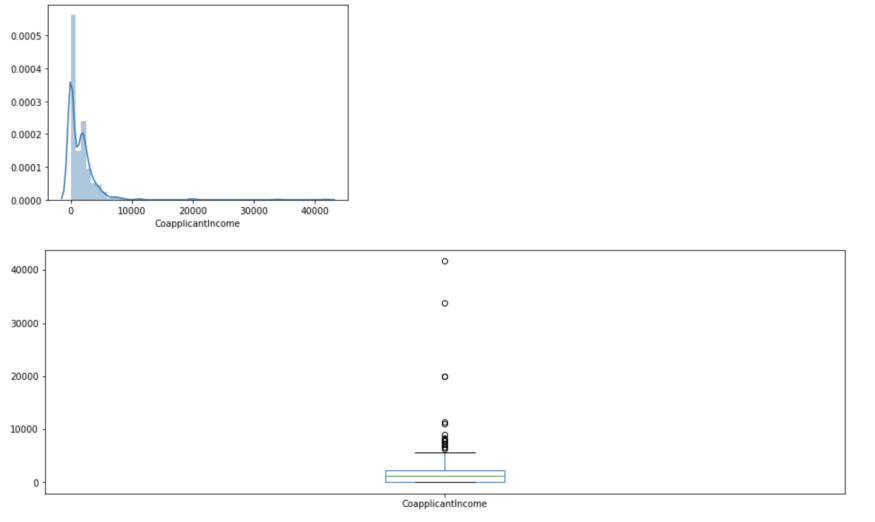
**Co-Applicant income Distribution**

Fig. 6: Co-Applicant income distribution

We see a similar distribution as that of the applicant's income. The majority of co-applicants income ranges from 0 to 5000. We also see a lot of outliers in the applicant's income and it is not normally distributed.

**Bivariate Analysis**

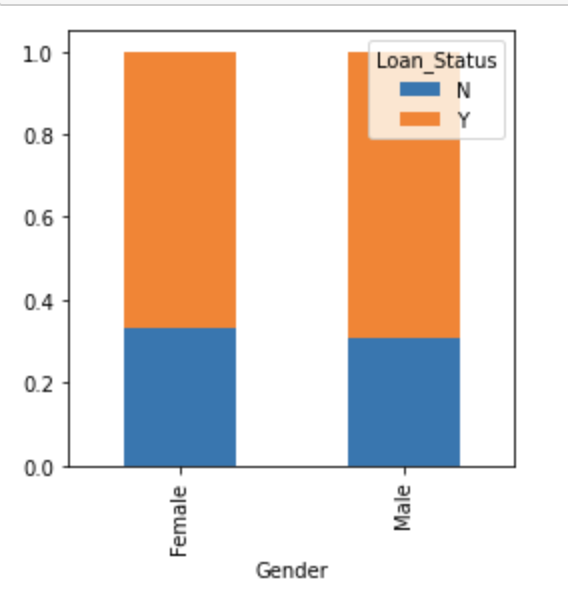
**Categorical Independent Variable vs Target Variable**

Fig. 7: Stacked Bar plot between Gender and Loan Status

It can be inferred that the proportion of male and female applicants is more or less the same for both approved and unapproved loans.

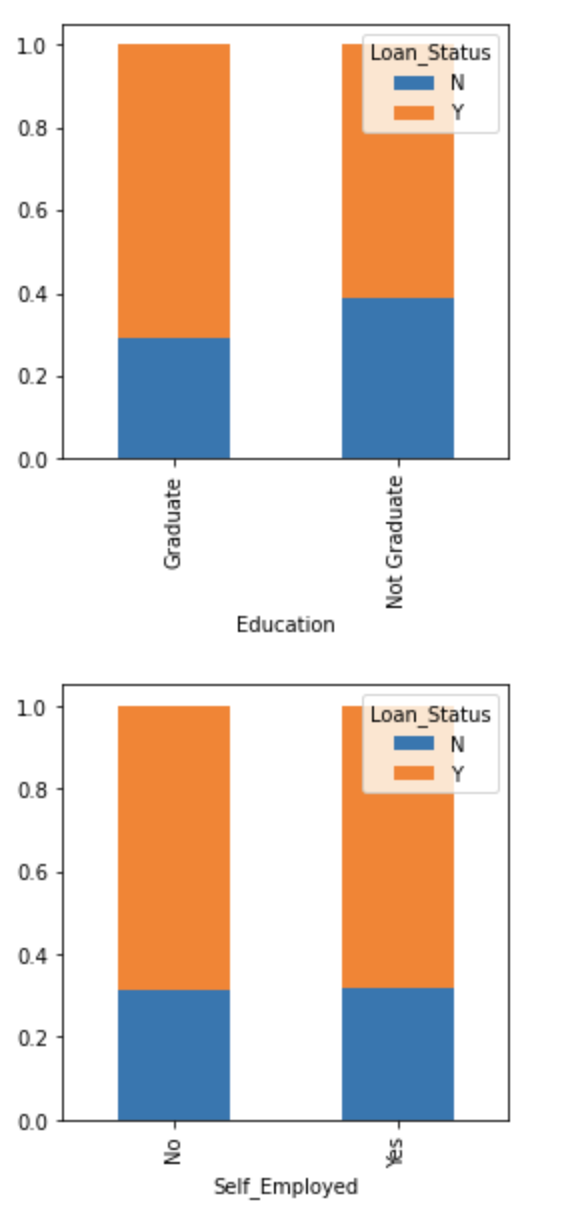
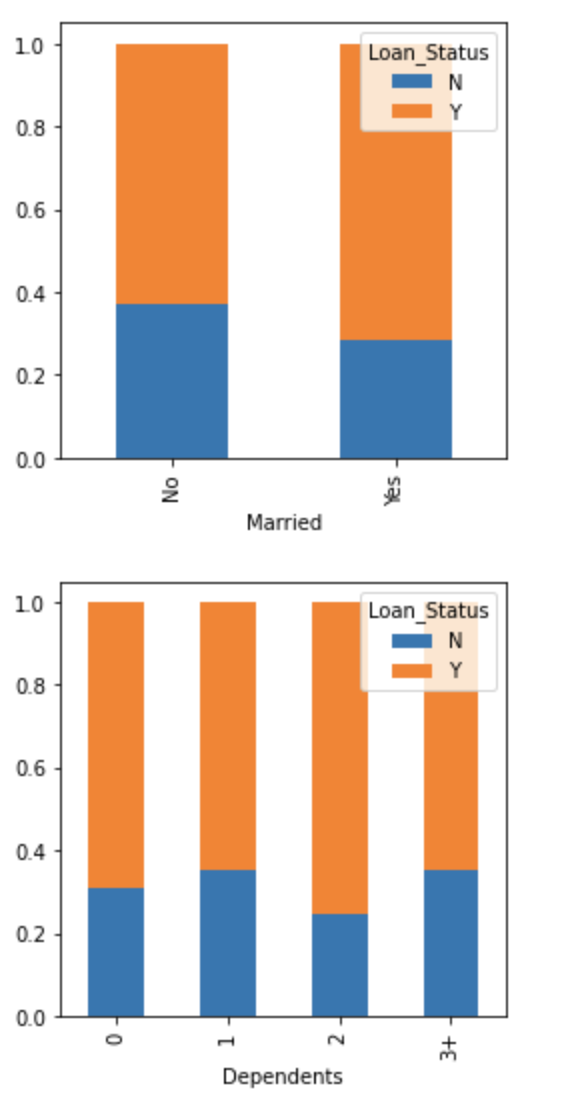


Fig. 8: Stacked Bar plots of different Categorical variable and target variables

Some of the observations from the above figures are:

* The proportion of married applicants is higher for approved loans.
* The distribution of applicants with 1 or 3+ dependents is similar across both the categories of Loan\_Status.
* There is nothing significant we can infer from Self\_Employed vs Loan\_Status plot.

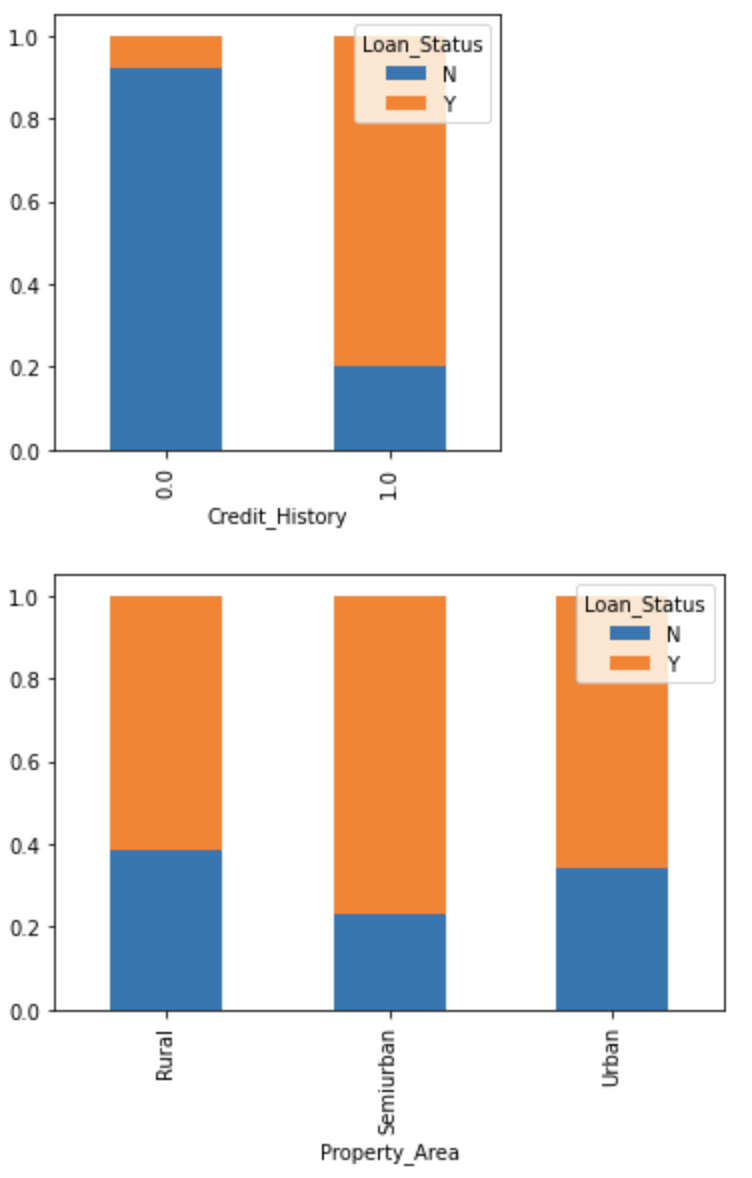
Now we will look at the relationship between remaining categorical independent variables and Loan\_Status.

Fig. 9: Stacked Bar plot between credit history and property area vs Loan status

* It seems people with a credit history as 1 are more likely to get their loans approved.
* The proportion of loans getting approved in the semi-urban area is higher as compared to that in rural or urban areas.

**Numerical Independent Variable vs Target Variable**

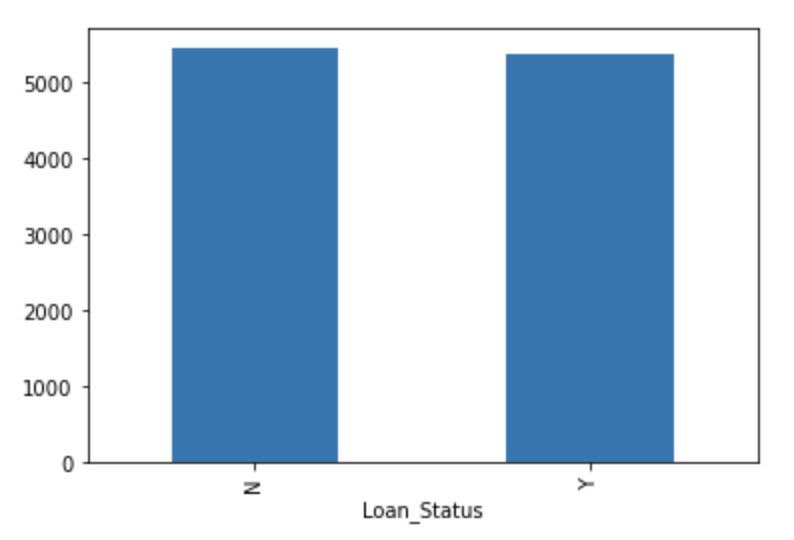
We will try to find the mean income of people for which the loan has been approved vs the mean income of people for which the loan has not been approved.

Fig. 10: Mean of income values for target status

Here the y-axis represents the mean applicant income. We don’t see any change in the mean income. So, let’s make bins for the applicant income variable based on the values in it and analyze the corresponding loan status for each bin.

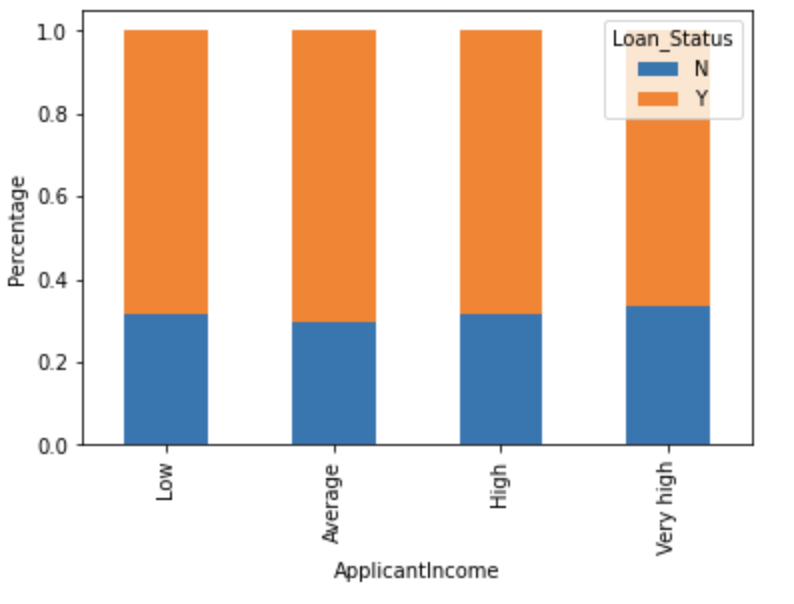


Fig. 11: Binning of income values for more detailed analysis

It can be inferred that Applicant's income does not affect the chances of loan approval which contradicts our hypothesis in which we assumed that if the applicant's income is high the chances of loan approval will also be high.

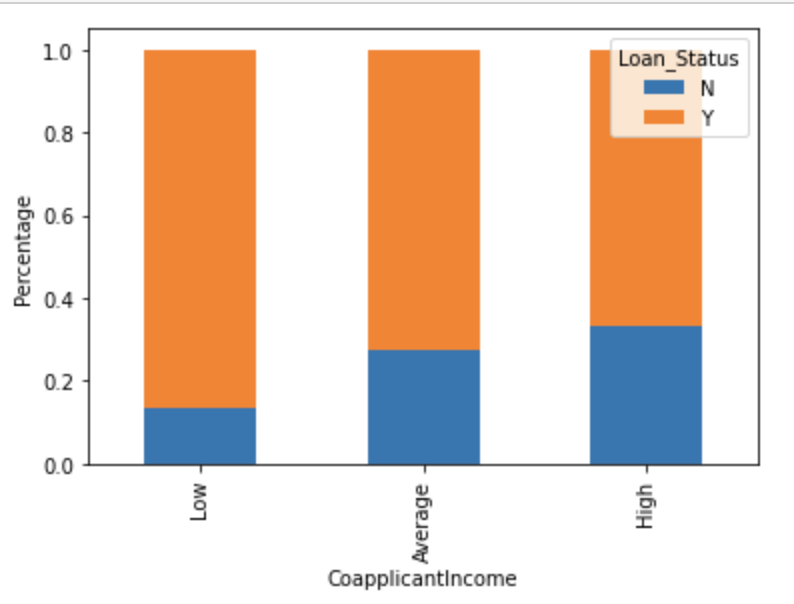
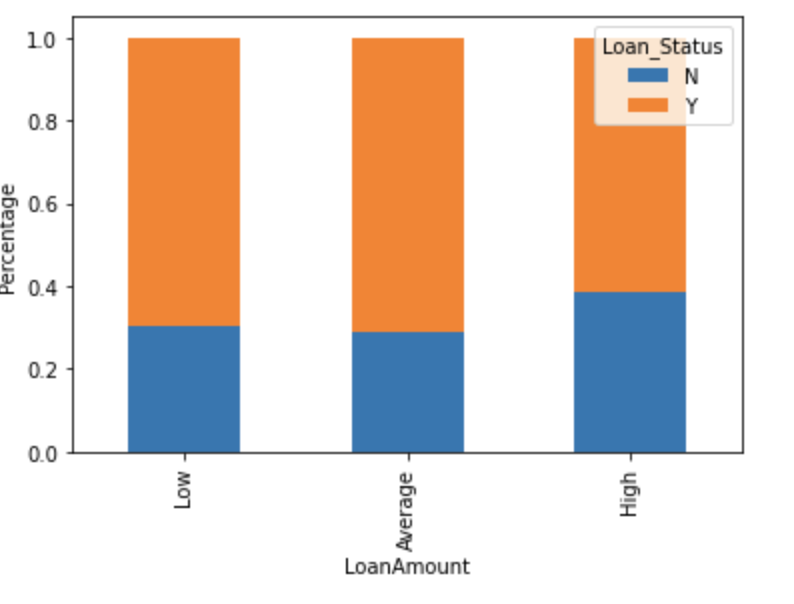
**Co-applicant income and loan amount variable**

Fig. 12: Stacked Bar plot of Co-Applicant income vs Loan Status

It shows that if co-applicants income is less the chances of loan approval are high. But this does not look right. The possible reason behind this may be that most of the applicants don’t have any co-applicant so the co-applicant income for such applicants is 0 and hence the loan approval is not dependent on it. So, we can make a new variable in which we will combine the applicant’s and co-applicants income to visualize the combined effect of income on loan approval.

**Let’s visualize the Loan Amount variable.**



It can be seen that the proportion of approved loans is higher for Low and Average Loan Amount as compared to that of High Loan Amount which supports our hypothesis in which we considered that the chances of loan approval will be high when the loan amount is less.

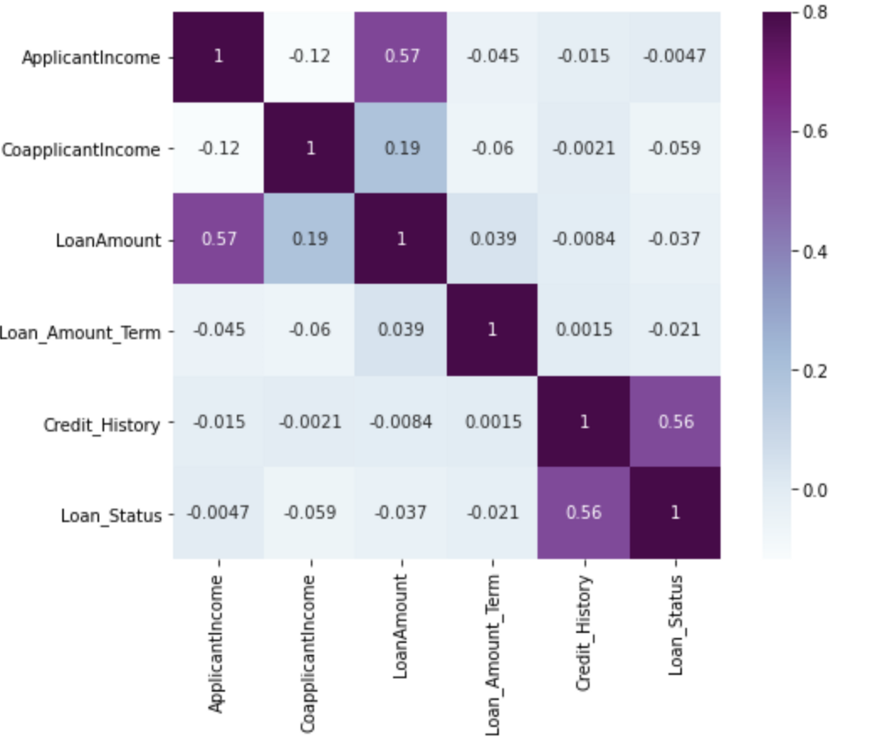
**Correlation Plot**

Fig. 13: Correlation Heatmap

We see that the most correlated variables are (ApplicantIncome — LoanAmount) and (Credit\_History — Loan\_Status). LoanAmount is also correlated with CoapplicantIncome.

# **Data Preprocessing**

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format. The first step in this process is Data Cleaning which refers to filling of the missing values, which is followed by Data Normalization, which will be followed by PCA and feature selection which will be explained in details in further sections

**Filling Missing Data**

* There are missing values in Gender, Married, Dependents, Self\_Employed, LoanAmount, Loan\_Amount\_Term, and Credit\_History features.
* We will treat the missing values in all the features one by one.
* There are very few missing values in Gender, Married, Dependents, Credit\_History, and Self\_Employed features so we can fill them using the **mode** value of the features.
* **Deterministic Regression Imputation -** The null values of the two features “LoanAmount” and “Loan\_Amount\_Term” were filled using the values predicted by regression model, which is also called as regression based null values imputation.

**Feature Encoding**

The categorical features such as Gender, Married, Self\_Employed, Credit\_History, and Loan\_Status were OneHot Encoded, and the ordinal categorical features such as Dependents, Education and Property area were Label Encoded. I have experimented with different types of feature encoding techniques in order to get the highest performing set of features.

# **Feature Selection and PCA**

Feature selection is one of the crucial processes in any Machine Learning project, in which we reduce the number of input features to the predictive model. This method has various different advantages.

The top reasons to use feature selection are:

* It enables the machine learning algorithm to train faster.
* It reduces the complexity of a model and makes it easier to interpret.
* It improves the accuracy of a model if the right subset is chosen.

I have plotted a barplot showing feature importance of features using the attribute “feature\_importances\_” of decision tree classifier, in order to perform feature selection.

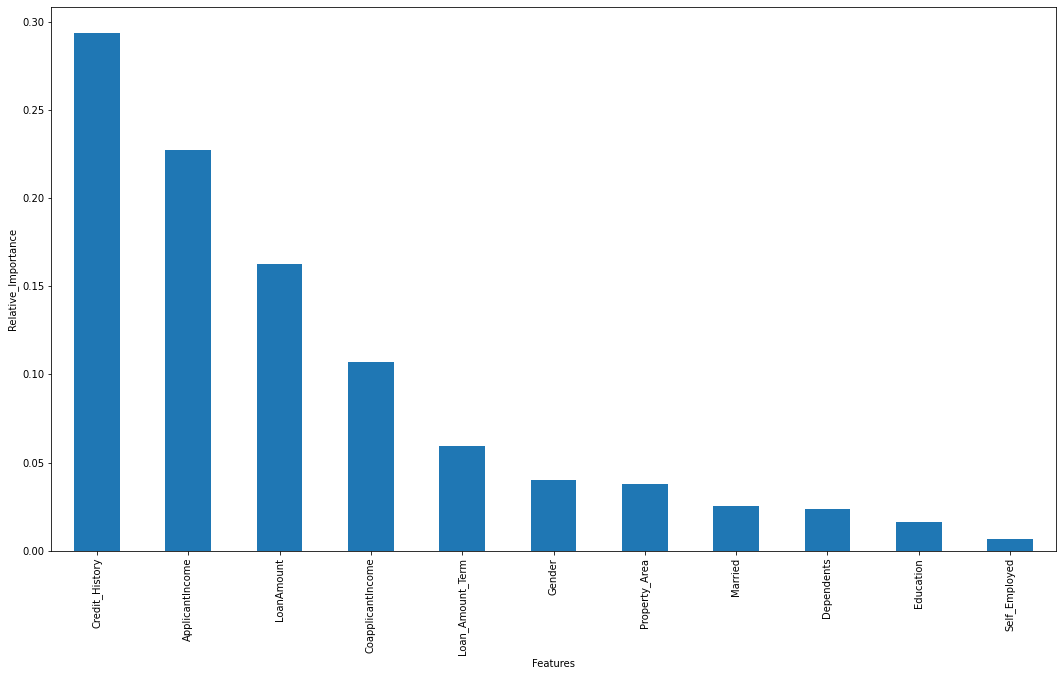


Fig. 14: Feature Importance Plot using Decision Tree classifier attributes

I have selected the top eight features out of the eleven features by the help of this feature importance plot. The number of features used here are varied from a range of six to ten and the best performing set of features were selected for the final model training.

Following this, Principal Component Analysis (PCA) were performed on the features of the dataset in order to perform feature reduction. Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance.

# **Results and Conclusion**

We implemented various machine learning classifiers for baseline creation and finally selected XGBoost Classifier as the best model to be further used for hyper-parameter tuning. XGBoost is a scalable and accurate implementation of gradient boosting machines and it has proven to push the limits of computing power for boosted trees algorithms as it was built and developed for the sole purpose of model performance and computational speed.

We were finally able to achieve 82.073% accuracy in the final week of coding using XGBoost classifier. The following table gives a detailed results summary of different classifiers used in the baseline as well as the final model results.

| **Machine Learning Classifier** | **Accuracy** |
| --- | --- |
| Logistic Regression | 80.48 |
| Support Vector Machine | 80.49 |
| K Nearest Neighbour Classifier | 65.85 |
| Decision Tree Classifier | 77.234 |
| LightGBM | 78.86 |
| AdaBoost (Tuned) | 81.46 |
| **XGBoost (Tuned best model)** | **82.07** |

**Table** **1**: Classifier Results Summary

This application is working properly and meeting to all Banker requirements. This component can be easily plugged in many other systems. It works correctly and fulfills all requirements of bankers and can be connected to many other systems. There were multiple malfunctions in the computers, content errors and fixing of weight in computerized prediction systems.In the near term, the banking software could be more reliable, accurate, and dynamic in nature and can be fit in with an automated processing unit.