Money plays an important role in any human being’s life. While one might argue that money cannot buy happiness, the fact that it is required in every sphere of life to buy goods or services means that it should be accessible to all. So even if money cannot buy happiness, it sure can provide peace of mind. However, due to many social as well as other factors, not everyone have equal access to enough money to live a financially secure life. Apart from this, sometimes people are in need of a large sum of money which they might not have access to immediately. Even for people with strong finances, having immediate access to money maybe difficult as their money may be tied to some other assets (like stocks, or property) and liquidating them takes time. In order to deal with these problems, one popular solution that people turn to is loans.

There are various types of banking as well as non-banking financial institutions that give out loans to people and business. The way these institutions earn from these loans are by charging the borrower interests on the loan. There are several types of loans which are tailored to suit various different needs. For example, there is housing loans which are given to help purchase real estate. There are education loans which are given to students to pursue academics and unlike other types of loan, the repayment of these loans starts only after the student have obtained their degrees.

Taking a loan can be beneficial to not only the financial institution sanctioning the loan but also to the one obtaining the loan. Firstly, it helps them meet their immediate financial obligation which he or she may not have been able to if there wasn’t anyone to give them a loan. Secondly, if the borrower stick to their repayment schedule, follows the good practises as well as makes their repayments on time, it also helps them build a good credit history which would enable them to obtain larger loans at lower rate of interest in the future.

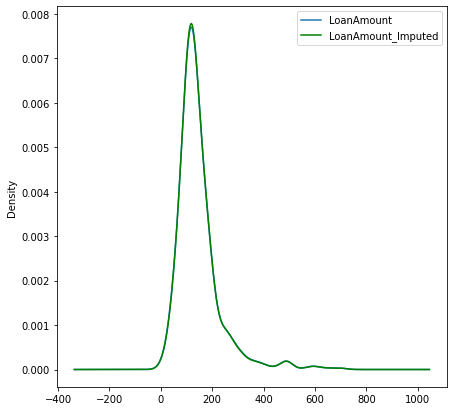
However, from the financial institution’s point of view, they are taking a lot of risk by giving out loans then the borrower. There is always a chance that the borrower might not repay the money they owe to them. This is called defaulting a loan. Loan defaults are not just the problem of the financial institution issuing it. If the loan defaults get large enough, it will result in recession. This was the case that led to the 2007 financial crisis which is now famously (or infamously!) known as ‘The Subprime Mortgage Crisis’ which lasted for 3 years and caused a recession in not only the United States but also to the entire world. This crisis was the result of banks giving out loans to bad customers who were not in a position to pay back the loans later. These loans were called ‘sub-primes’. Even though the financial crisis lasted for three years, its effects are still felt in the economy. Therefore, it is crucial that banks and financial institutions give out loans only to those who are more likely to repay the loan.

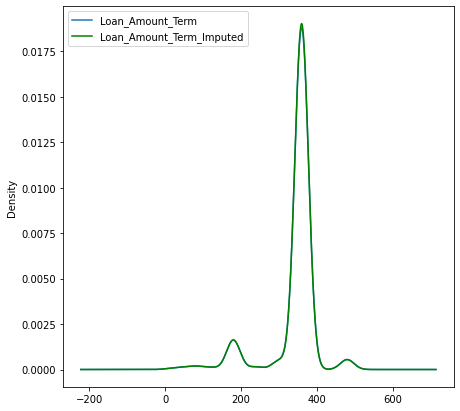
In order to do so, it is necessary to have a thorough background check of the loan applicant. Traditionally, this was done with the help of a lot of paperwork which the borrower was asked to provide. This made taking loan a tedious and inefficient process. This also affect the overall customer experience in a negative way as the borrowers would have to wait until the banks have finished their due diligence before giving out a loan. In case the loan was not sanctioned, the customer would have waited to no avail which might leave them frustrated and with a bitter taste for the organisation. Therefore, while it is important for loan giving organisations to do their due diligence in order to mitigate their risks, it is also equally important for them to do it as promptly as possible in order to deliver as quicker and smoother customer experiences. This can be achieved in many ways. While most of those ways are out of the scope of this discussion, one way is, in fact, right down our alley which is using data to solve problems. For this project, we will use data to predict if a loan will be sanctioned or rejected for an individual.

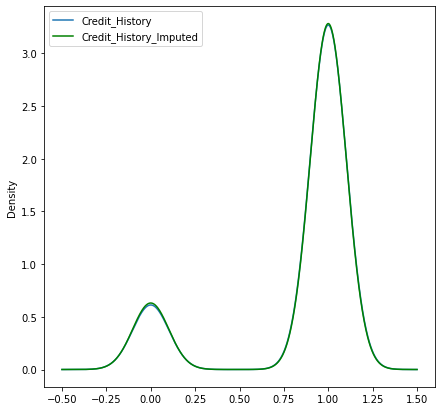
However, before we start, it is important to point out that deciding to accept or reject a loan application requires a lot of dimensions to be considered, something only a human is capable of (at least for now). Some of these may be incomprehensible for a computer algorithm. This may be due to the lack of data points for the algorithm to consider, lack of computing power even when the data is available or simply because the right algorithm just doesn’t exist. Therefore, relying only on computers to judge a fellow human’s credibility raises a few ethical concerns. In this regard, machine learning and artificial intelligence should be used to filter applicants so that the important ones get priority first and not to completely reject applications without any human reviewing them first.

Having addressed the problem statement and the ethical concerns, we are finally ready to dive into the project. The task we have at hand is that of a supervised classification problem. The data set contains twelve columns and 614 rows. Out of these 12 columns, the target column is ‘Loan\_Status’ with the labels ‘Y’ for those cases where a loan was sanctioned and ‘N’ for those that were rejected. The other columns are Loan\_ID, Gender, Married, Dependents, Education, Self\_Employed, ApplicantUncome, CoapplicantIncome, LoanAmount, Loan\_Amount\_Term, CreditHistory, and Property\_Area. The column names are self-explanatory and does not require further explanation.

Out of the 12 columns, 7 of them have missing values. Since that data set has only 614 rows, dropping the rows with missing values will result in significant data loss. Therefore, instead of dropping them, we will impute the missing values. For the numeric columns, we use random sample imputing to replace the missing values. First we make a copy of the original column and then impute the missing values in the new column .This technique is chosen as it does not affect the existing data distribution (and variance) a lot. This is evident from the plots below.



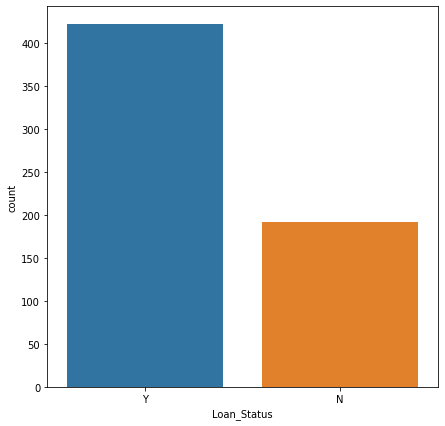




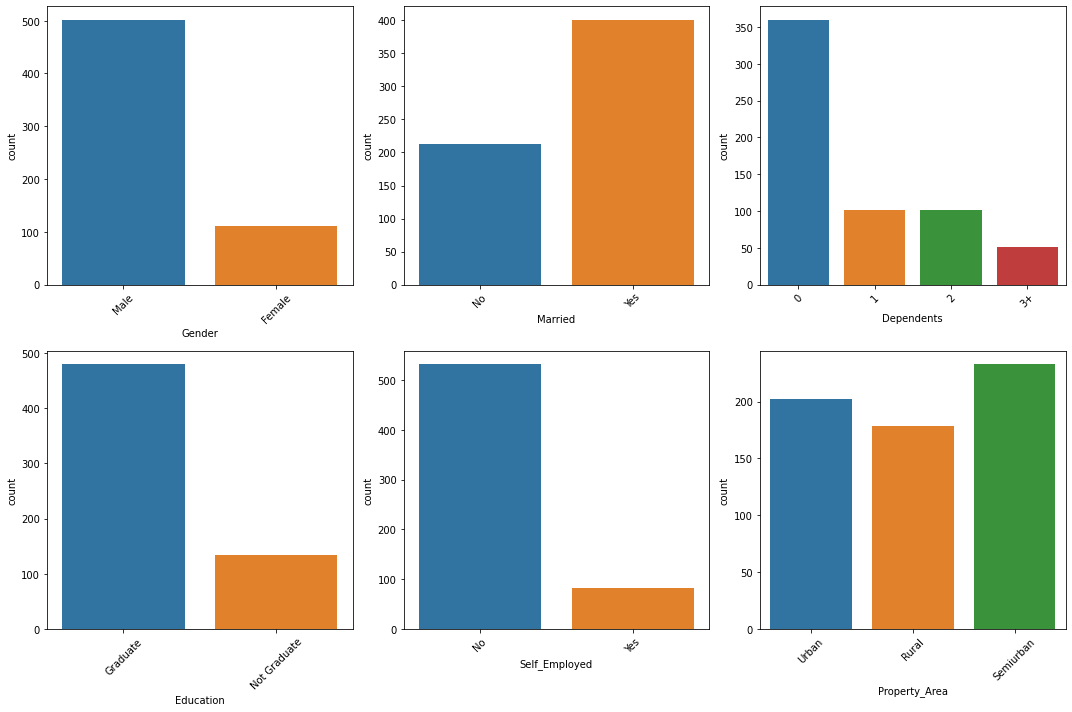
As it can be seen clearly, by using the random sample imputation technique, the imputed values match the distribution of the original column so well that that the distribution curve almost appear as a single one.

For the categorical columns containing missing values, we replace the missing values with the mode of the column. We also create a new column for each categorical column with missing values and fill it with 0 for the rows without missing values in the original column and 1 for the rows with a missing value in the original column. This is done to capture any relation the missing values might be having with the target column.

Before building our model, it is important to have a better understanding of the data. So, we will dive into some exploratory data analysis using a combination of visualization and hypothesis testing.



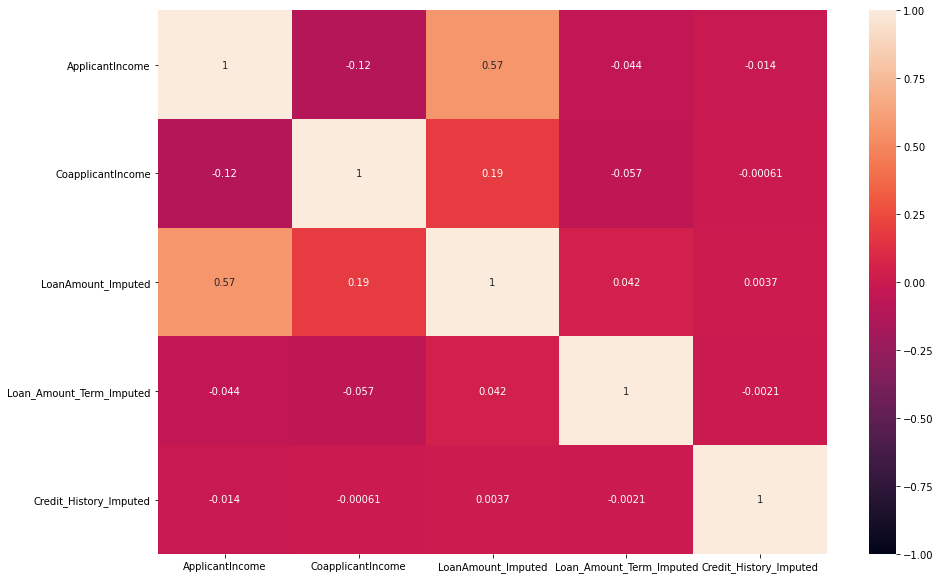
We begin by looking at the count plot of our target column as seen above. The label ‘Y’ has a count about 450 while the ‘N’ label has a count of about 200. Since the ‘Y’ label is more then the ‘N’, the final model will better at predicting the Y(s) then the N(s). To solve this, we need to balance the dataset before fitting a model. Now we can turn our attention to the feature columns in the dataset.



The figure above shows the count plots for all the categorical columns. We see that most applicants are male. Approximately 500 applicants are male while only about 100 applicants are female. In the column Married, majority of the applicants are married. Most have 0 dependants, and out of all the applicants, those with more than three dependents are the least. In terms of education and employment, majority of the applicants are graduate and not self-employed. However, in terms of residing/property area, there is a more or less balance between applicants from Urban, Rural or Semi-urban areas.

Since our target is a categorical variable, in order to see if it has a statistically significant relation with other categorical columns in the data set. This can be achieved by performing Chi-Squared test of Independence. Performing this test reveals that Married, Education, and Property\_Area has a statistically significant relationship with the target column at 95% significance level. We will only consider these categorical columns for model building.

Next, we turn our attention to the numeric columns in the data set. We start by inspecting the correlation heat map.



Inspecting the heat map above, we see that no columns are very highly correlated to each other. Therefore, we can rule out the possibility of multicollinearity. All the independent numeric columns had skewness ranging from -1.85 to as high as 7.49. Therefore, all these columns were transformed using either log transformation, cube root transformation or boxcox transformation. After this transformation, the final skewness ranged between -0.19 to 0.49.

Once we have removed/reduced the skewness, we can go forward with separating the target and the feature variables. After this, we perform encoding on the target variable as it is originally a categorical column. This is done by replacing all ‘N’ labels with 0 and all ‘Y’ labels as 1. Then we move on to standardising the numeric columns and one hot encoding the independent categorical columns.

Finally, we perform oversampling using random oversampling technique to balance the dataset. Once the pre-processing stage is out of the way, we fit an arbitrarily chosen model multiple times with different random states during train-test split to find the best random state where accuracy is the highest. In our case, that random state was 199. We then use this random state to build all our models after train-test split.

After train-test split, we fit multiple models classification models namely, Logistic Regression, Random Forest Classifier, Decision Tree Classifier, Support Vector Classifier, KNeighbors Classifier and AdaBoost classifier. In order to select the final model, we looked in to the accua

Now, before we choose a final model, we need to ask ourselves what are we trying to achieve with this model. The evaluation metrics we select will depend on this. Therefore, the question that we are trying to answer is, what is more important for us? Is it more important to identify which applications will be rejected or is it important to correctly identify which application will be approved? While the answer to this question would generally require some domain knowledge, for the purpose of this project, our priority would be to correctly identify those which are likely to get accepted so that their applications can be reviewed first and those likely to get rejected are reviewed later. In technical terms, we are aiming to select a model with high sensitivity (also called recall) for the label ‘Y’ (encoded as ‘1’) which would tell us ‘How many of the application which were accepted were correctly identified?’. Therefore, going by this metrics, we see that support vector classifier (or SVC) has the highest recall for the label ‘Y’ of 0.92. Once SVC is selected as the final model, we perform hyper parameter tuning on this algorithm to further optimise it. After hyper parameter tuning, the recall for label ‘Y’ increased by 0.06 to 0.98. Finally, we serialise the model for later using the library called joblib. This will enable us to make predictions in the future without re-training the model every time we need to make predictions.

While this was a fun short project to work on, there are still room for improvement here. Firstly, the oversampling technique used was random oversampling which randomly takes existing rows in the data sets and duplicates them to balance the dataset. Here, a better and more sophisticated oversampling maybe be used to improve the model. Secondly, the AUC-ROC curve was not inspected to further optimise the model. This should also be done to make the model more accurate and robust.