**Model Building**

After performing basic preprocessing, visualization of important features, cleaning the data and correlation between features, we get our final dataset which consists of 24 features. Now, we need machine learning models to predict loan-status feature which consists of 2 categories i.e. Fully Paid and Charged Off. We encoded the categorical columns to assign integers to all different types of categories.

First, we check the class balance for the column loan-status. We see that the class is imbalanced. To solve this problem of imbalanced class, we can use many techniques like Assigning weight class, use ensembled algorithms with cross-validation and upsample the minority class or downsample the majority class. Here, we upsample the minority class.

Now, we define a function which determines the accuracy for the models which we use for loan-status prediction.

First, we split the dataset into training and testing parts where 25% data is used for testing. Now, we standardize the data so that the mean is zero and standard deviation is one. This optimizes the models to run faster. Here, we use Scikit-learn StandardScaler method.

Now after preparing our data, we can build models Here, the feature loan-status is binary. So, we use binary classification models to train and test our data.

**Logistic Regression:** We use one the most basic model for binary classification. Logistic Regression is a classification algorithm that is used where the target variable is categorical. The notion of LR is to find relationship between features and probability of particular outcome. We usually have binary classification in LR, but we also have Multinomial Logistic Regression where the response can have three or more values.

For our dataset, we use Logistic Regression from Scikit-learn. After implementing this model, we can see that we have a decent accuracy of almost 65%. But since it is one of the most basic classification models, we use try more complex models.

**Random Forest:** This model consists of a large number of individual decision trees that operate as an ensemble. Each individual decision tree in the RF gives a class prediction and the class with most votes becomes the model’s prediction.

Here, after applying the RF model we see that the accuracy jumps to 75% from 64% in the LR model. This proves that RF is more precise and accurate. The random forest model works so well because a large number of relatively uncorrelated trees operating as a group can outperform any other individual constituent models. The main reason for this is that the trees protect each other from their individual errors. Some trees may be wrong, many other trees will be right, so as group the trees are able to move in correct direction.

**Need for Generalization of Models:** Generalization refers tp model’s ability to adapt properly to new, previous unseen data which is drawn from the same distribution as the one which we used for the creation of model.

So, now we try a gradient boosted tree-based algorithm to classify.

**LightGBM Model:** LightGBM is one of the most successful and powerful machine learning algorithms. In simple words, LightGBM is a gradient boosting framework that uses tree-based learning algorithm. It grows tree vertically while other models grow trees horizontally. It always chooses the leaf nodes with max delta loss to grow. While growing same leaf nodes, leaf-wise algorithms can reduce more loss than a level-wise algorithm. The size of data is increasing day by day and it is quite difficult for traditional data science algorithms to give faster results. LightGBM is called as ‘Light’ because of its high speed. It can handle the large size of data and it takes lower memory to run. It focuses on accuracy of results plus it has GPU learning also. This model is best for huge datasets where the number of rows is usually more than 10000 because it is sensitive to overfitting and can easily overfit small data. Since, our dataset consists of almost 1.3 million rows, this model is best for our use.

After implementing this model, we can see that our accuracy is now at almost 89% which is lot better than our previous models.

**CONCLUSION:** After pre-processing and cleaning the dataset, visualization and correlation of all the important features, we built the machine learning models on our finalized dataset so that we can finally predict loans. The accuracy of the models jumped to almost 89% which is highly accurate. Some important conclusions are:

1. We got correlations between features that truly match with practical examples. For example, interest-rate & grade OR loan-amount and grade.
2. Through visualization of interest-rate column, we can conclude that people who pay back fully usually have low interest rate compared to those who are charged off. This conclusion relates to real world also.
3. Through visualization of grade column, we can conclude that people who have good credit grade tend to pay back the loan fully than those who have bad credit grade.
4. We got lot of deep-insights through visualization like what was the range of loan-amounts, where do most people who borrow loan live.
5. We also concluded that Tree-based models perform better for classification of data with more categorical columns.