**Loan Default Prediction**

**Problem Statement:**

To classify if the borrower will default the loan using borrower’s finance history. That means, given a set of new predictor variables, we need to predict the target variable as 1: Defaulter or 0: Non-Defaulter.

The metric we use to choose the best model is ‘False Negative Rate’.

**Data engineering:**

For this project we downloaded the data from Kaggle. The dataset consists of 360,000 observations and 145 features. Out of these features many of them were empty. We have removed those. Also, those features which were not relevant were removed too.

String values have been formatted to integers

Categorical values have been transformed to numerical.

Redundant variables have been dropped.

Filled NAN values with mean values of corresponding columns.

All the numerical values have been scaled to a range between -1 and 1.

Using correlation matrix of the dataset we chose 20 related features to our objective. Then the dataset has been reduced to 11000,10.

**Predictor Variables:**

On these 20 features we have implemented Recursive Feature Elimination (REF) using logistic Regression model to get the best 10 features:

|  |  |
| --- | --- |
| **Predictor Variables** | **Description** |
| funded\_amnt | The total amount committed to the loan |
| grade | Employment lengths in years |
| annual\_inc | The annual Income |
| last\_pymnt\_amnt | Last total payment amount received |
| bc\_open\_to\_buy | Total open to buy on revolving bankcards. |
| int\_rate | Interest Rate on the loan |
| mo\_sin\_old\_rev\_tl\_open | Months since oldest revolving account opened |
| mo\_sin\_rcnt\_rev\_tl\_open | Months since most recent revolving account opened |
| acc\_open\_past\_24mths | Number of trades open in past 24 months |
| Loan\_status | Current status of the loan |

**Target variable:**

The target variable here is ‘Loan\_status’ which shows the status of the loan.

It has 3 different values:

**Charged Off:** Loan for which there is no longer a reasonable expectation of further payments

**Fully Paid:** Loan has been fully repaid

**Default:** Loan has not been current for 121 days or more

Since the goal is to predict whether a borrower will default the loan, we are considering only the observations where loan\_status is either Fully paid or charged off. We changed Fully paid as 0 and Charged off as 1 where 1 indicates the borrower as a defaulter.

**Models Applied:**

**Random Forest classification:**

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. Random forest select a subset of features in each of its decision trees thereby reducing the bias of the model.

**Multi-layer Perceptron:**

MLP utilizes backpropagation for training. Its multiple layers and non-linear activation function helps us distinguish data that is not linearly separable.

**Support Vector Machine:**

We choose to build SVM classifier because once a hyperplane is found, most of the data other than the support vectors become redundant. This means that small changes to data cannot greatly affect the hyperplane and hence the SVM. So support vector machine tends to generalize very well.

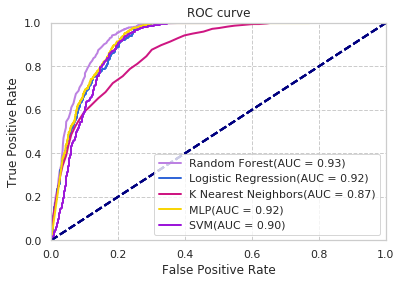
**Logistic Regression:**

Logistic regression, despite its name, is a linear model for classification rather than regression. With logistic regression outputs have a nice probabilistic interpretation and the algorithm can be regularized to avoid overfitting.

**K-Neighbors**

Neighbors-based classification is a type of instance-based learning or non-generalizing learning: it does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the nearest neighbors of each point: a query point is assigned the data class which has the most representatives within the nearest neighbors of the point.

|  |  |
| --- | --- |
| **Random Forest** | **Logistic Regression** |
| **K-Neighbors** | **MLP** |
| **SVM** | |



**Mode Accuracies:**

Random Forest accuracy is 87.81

Logistic regression accuracy is 85.27

SVM accuracy is 85.81

KNN accuracy is 78.04

MLP accuracy is 86.27