Default Prediction for a Peer To Peer Lending Platform

Kago Ronald Thipe

May 28, 2024

1 Introduction

In this task, We will be using data from a random sample of loans obtained from a peer-to-peer lending platform with the aim of helping the platform identify risky loan applicants more accurately.We will be using different Machine Learning models with different parameters and then finally choose the best model.

We have been provided with 3 csv datasets:

- trainData: This is the dataset on which we will train all our models
- testData: This is the dataset on which we will evaluate our model's fit.
- varDescription: This contains a description of the features

Our approach to this assignment involved:

- **Data Preprossessing:** Some features(mths-since-last-delinq) had more than 50% missing values so we removed it. Other columns that had missing values, we imputed the median. We chose median imputation as our features were not normally distributed. We used one-hot encoding to convert the categorical features ['grade','home ownership', 'application type] to numerical features. We also converted 'emp length' to a numerical. This was because we were to use were regression and thus required numerical features. We also scaled the features before modelling.
- **Exploratory Data Analysis:** To gain interesting insights such as correlation between features and the target variable, distribution of important features and outlier detection.
- **Model Selection:** We used regression techniques, as well as Random Forest, and Neural Network. Parameters we chose had to be computationally effective as well as provide the best improvement to our models.
- **Performance Metrics:** We used Mean Squared Error on both Training and Test Sets to measure performance as well choose the best parameters. We also used ROC AUC scores to compare the different models and choose the best one.

2 Data Description

Number of Numerical Features: 27
Number of Categorical Features: 4

The table below shows summary statistics for numerical features. From this summary, we can see that some features have got missing values, that is those features with a count of less than 226067. We can also get an idea of the distribution of the features by looking at the range of values and identifying the presence of outliers by noticing the margin between mean and median of each feature.

The variable 'y' which is going to be our target variable takes values 1 for 'loan status' that is charged off, and 0 for 'loan status' that is either 'fully paid', 'current' or any other status that shows that the applicant did not default on their payments. We can also see in the table that the mean for variable 'y' is lower, thus meaning in the original dataset, Most applicants are Non-defaulters, hence there is an imbalanced distribution of our target variable.

Table 1: Summary Statistics

Variable	Count	Mean	Std	Min	50%	Max
loan_amnt	226067	15000	9000	500	13000	40000
int_rate	226067	13	5	5	13	31
installment	226067	450	270	20	380	1700
annual_inc	226067	78000	82000	0	65000	9300000
dti	225887	19	14	0	18	1000
delinq_2yrs	226062	0	1	0	0	30
inq_last_6mths	226062	1	1	0	0	20
mths_since_last_delinq	110227	34	22	0	16	160
open_acc	226062	12	6	0	11	90
pub_rec	226062	0	1	0	0	50
revol_bal	226067	17000	24000	0	11000	2560000
revol_util	225877	50	25	0	50	160
total_acc	226062	24	12	1	22	150
total_pymnt	226067	12000	10000	0	9100	63000
total_pymnt_inv	226067	12000	9900	0	9100	63000
total_rec_prncp	226067	9300	8300	0	6800	40000
total_rec_int	226067	2400	2700	0	1500	28000
total_rec_late_fee	226067	1	10	-0.005	0	770
recoveries	226067	140	740	0	0	35000
collection_recovery_fee	226067	20	130	0	0	5200
last_pymnt_amnt	226067	3400	6000	0	590	41000
collections_12_mths_ex_med	226051	0	0.2	0	0	8
acc_now_delinq	226062	0	0.07	0	0	5
tot_coll_amt	218960	260	20000	0	0	9150000
tot_cur_bal	218960	143000	161000	0	80000	9970000
total_rev_hi_lim	218960	35000	35000	0	15000	1880000
y	226067	0.12	0.32	0	0	1

Furthermore we can plot a bar plot showing distribution of loan distribution. These information shows that our dataset contains our classes of loan distribution are not equally distributed, and there is a higher probability of our models predicting an applicant' s outcome as non defaulting. This also supports information we got from the summary statistics of variable 'y'

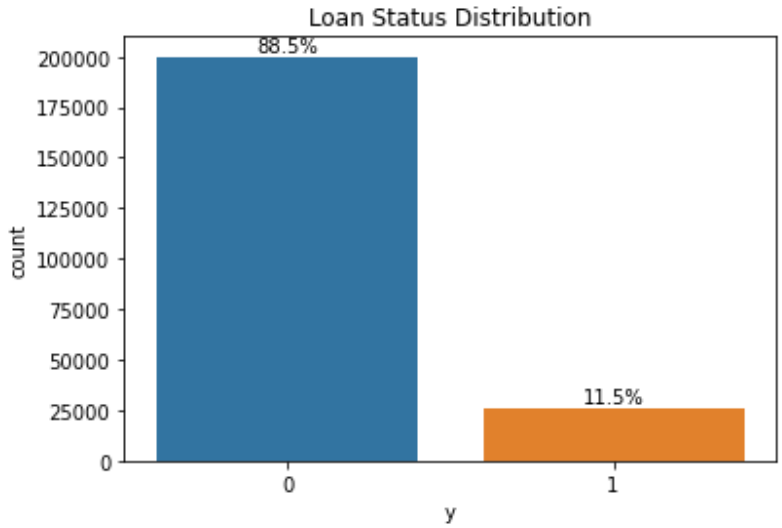


Figure 1: 1 Represents Charged Off/Defaulted, 0 Represents Otherwise

3 Model Comparison

3.1 Linear Regression

1. Mean Squared Error (Training - Linear Regression): 0.0677961069164743
2. Mean Squared Error (Testing - Linear Regression): 0.06860668637158215

For Ridge and Lasso Regression there is parameter λ which it's purpose is to regularize Linear Regression. Intention is to reduce Generalisation error while keeping training error low. We need to ensure that we make a reasonable choice of λ because if λ an is too low, overfitting might occur, and if it is too high, the fit might be too general d thus for our models we had to choose λ between 0.01 and 100.

3.2 Ridge Regression

1. Best lambda (Ridge): 99.53
2. Mean Squared Error (Training - Ridge): 0.06824535482911973
3. Mean Squared Error (Testing - Ridge): 0.06861122101595804

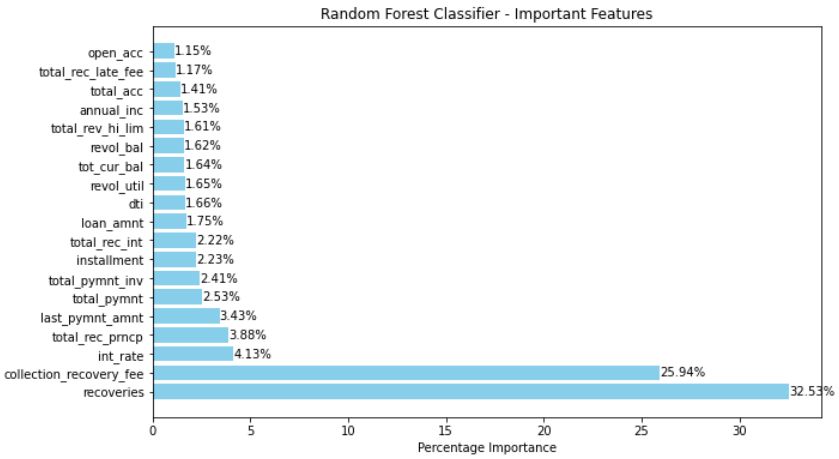
3.3 Lasso Regression

1. Best lambda (Lasso): 54.61
2. Mean Squared Error (Training - Lasso): 0.10213034189740981
3. Mean Squared Error (Testing - Lasso): 0.10247702603422762

3.4 Random Forest

1. Best Parameters: n-estimators: 200, max-depth: 50
2. Accuracy (Training - Random Forest Classifier): 1.0
3. Accuracy (Testing - Random Forest Classifier): 0.9655898472576714
4. Mean Squared Error (Training - Random Forest): 0.0
5. Mean Squared Error (Testing - Random Forest): 0.0344101527423286

The most important variables in predicting default focusing on those that the percentage of importance was greater than 1 percent.



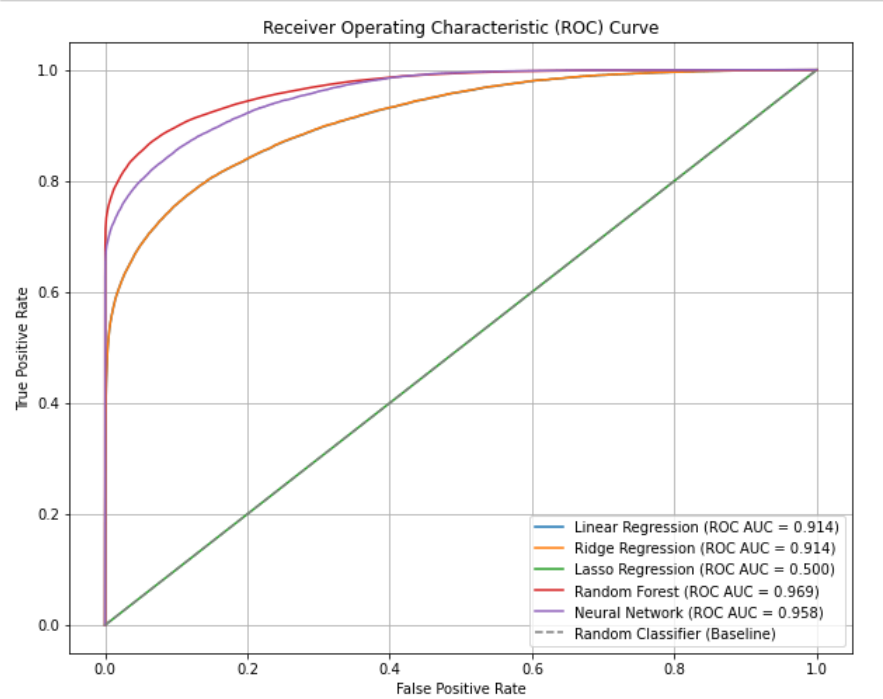
3.5 Neural Network

We chose MLP classifier as it proved to be computationally effective for our purposes.MLP classify can also handle large datasets, and can learn from high dimensional spaces, thus it was suitable for our classification task.

1. Best Parameters (MLP Classifier): {'solver': 'adam', 'learning-rate': 'constant', 'hidden-layer-sizes': 100, 'alpha': 0.01, 'activation': 'relu'}
2. Accuracy (Training - MLP Classifier): 0.9636390981434707
3. Accuracy (Testing - MLP Classifier): 0.960516130173798

3.6 Model Evaluation

The Mean Squared Error and Accuracy measurements acquired from all the models are relatively good, and the difference between the train scores and the test scores are also low. On top of considering the Mean Squared Error metrics and Accuracy Tests carried out on both the Training and Test Sets, We will use the **Area Under the ROC Curve(AUC)** on the testing dataset. Higher AUC values indicate better predictive power.



We can therefore conclude that the Random Forest Model is a better model predictor to determine if an applicant will default or not. This is because Random Forest performs better on the Test Set when compared to there models. However the Random Forest gave an accuracy of 100 percent and MSE of 0.0 on the Training Dataset which might give issues of Over fitting.

4 Insights

As seen in the Random Forest plot, some features are important in predicting the default outcome. We also found which features are mostly correlated to the 'y' variable and those that are least correlated. For example, we can see that 'Recoveries' are highly correlated with default status, this might inform the Peer To Peer Platform about the effectiveness of their post-default strategies. We can also see that applicants with higher interest rates tend to default, this information might help the business in determining if the risk of giving applicants higher interest rates is worth it. It can also assist with customer targeting approaches.

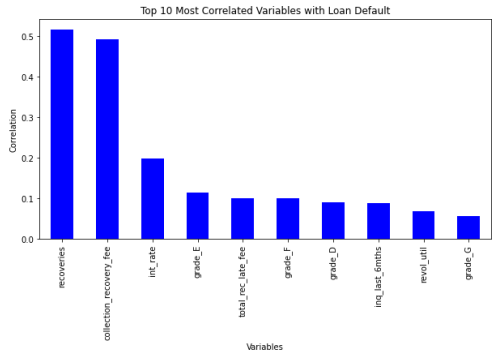


Figure 2: Top 10 Most Frequent Words

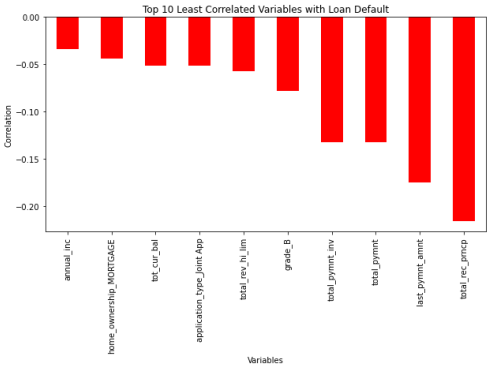
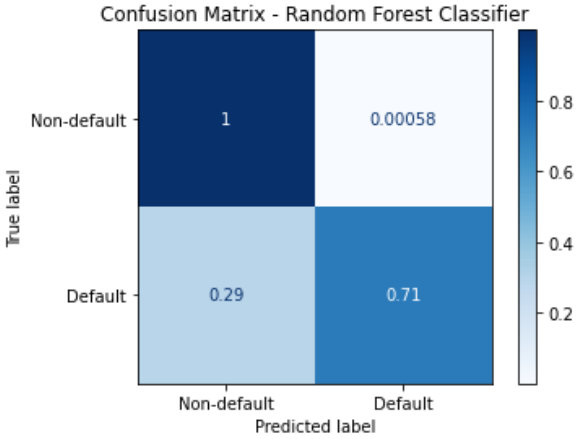


Figure 3: Top 10 Least Frequent Words

Figure 4: Correlation of features with 'y' variable

Since we have given random forest as the best model,it is necessary for us to know whether our model is accurate in predicting 'defaulters' or 'non-defaulters'. The goal of this assignment was to build a model that could predict if an applicant will default as such present a great risk to the business.



It can be seen that the model 100% can predict defaulters. This might be because in our dataset that was already an imbalance of non-defaulters - 88.5% to defaulters - 11.5%. However the model can predict that 71% will default which is also relatively good.

5 Conclusion

We have built 4 models that can predict if an applicant will default on their loan or not, and have chosen Random Forest as the Best Performing model. We have however identified that the unequal distribution between the 2 Classes {1:'default', 0:'non-default'} posed problems which might have led to over fitting. The 29% that has been predicted as Non-default, but rather Defaulted poses a risk of lost revenue, cost of recoveries etc. As such, an improvement in the Random Forest model needs to be considered.We might consider using more data, hopefully with more data showing behaviour of defaulters, additionally we can also consider other modelling techniques such as Logistic Regression. Another challenge encountered was tuning parameters for each model. This was a challenge as the models took too long to run, we have used randomized search to go through, this improved the time it takes to run a model.