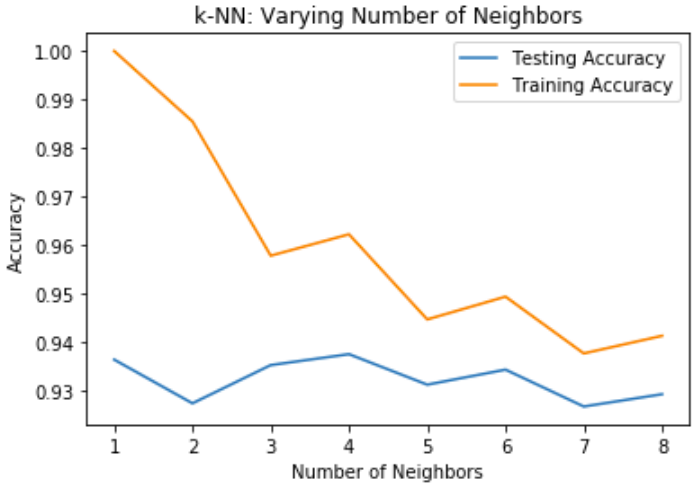
In Depth Analysis-ML

Capstone-1

We built our first model:

1)**Knn Classifier:** We got an accuracy of 0.93 . As our model has good loans dominating the sample we may need to over sample our data with bad sample.

We then now checked the effect on the accuracy by changing n in the model and the graph is plotted below:



Due to imbalanced data out score results showed 1.0 as 85% of the loans were good(target variable). Hence we used

2) **imblearn** for oversampling and **RandomForestClassifier** :

We see that the recall score and precision score comes out to be 1, which does not seen to be right. it can also be because of the dominating class of good loans or some other dominating features.

We then checked the important features in the model and observed that few features such as: loan_status,loan_condition_int,recoveries,collection_recovery_fee,last_credit_pull_d,total_rec_late_fee,total_rec_int,installment,total_pymnt,last_pymnt_d,total_pymnt_inv are to be dropped due to there insignificance/unavailability before the approval or similar terms to the loan_condition(target variable).

Hence we drop the above columns from the new data set and train our model again.

and got the following result:

Validation Results

0.8879022147931467

0.9444308145240432

Test Results

0.8820985332831892

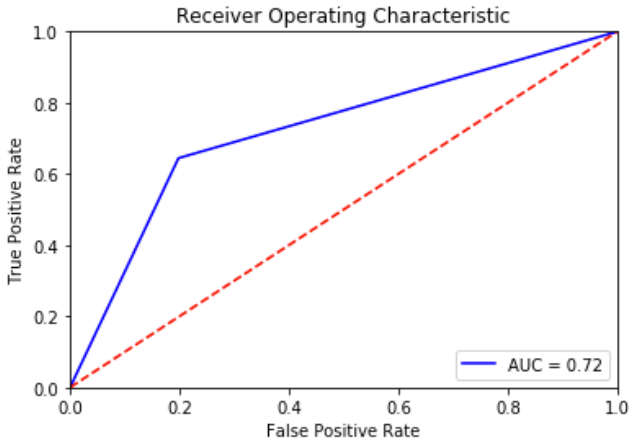
0.9388568896765618

Next, we did manual oversampling:It shows that out of 43000 datasets 6431 data represents the bad loans . So to make our dataset balanced we will create a new dataset "balanced" comprising both the type in 3:2 ratio.

We used two models separately to plot the ROC curve using the manually over sampled data.

3) KNN Classifier : Accuracy: 0.6489321843387036

4)Logistic Regression: precision score of :0.72



We will be proceeding for an Ensembled Learning Model to give much better result and hence we will now observe the performance of Six models together using a FOR loop.
The six models used below are:

1)LogisticRegression	**Cv_score(mean)=0.72**	**recall_score=0.73**	**precision_score=0.74**
2)DecisionTreeClassifier	0.88	0.88	0.89
3)LinearDiscriminantAnalysis	0.74	0.73	0.74
4)SVC	0.60	0.5	0.3
5)KNeighborsClassifier	0.64	0.62	0.63
6)MultinomialNB	0.53	0.53	0.56

It seems like **logistic regressio, Decision Tree and LDA models** agree to each other.Hence we will finally use this models into our Ensemble Techniques as our main aim is to increase the Precision and accuracy(decrease the False Positives which are more dangerous in our case).

Ensemble Model

This method combines the decisions from multiple models to improve the overall performance. This can be achieved in various ways,the first method which we would use is **Max Voting.**

a)Max Voting:

The accuracy score of the test model using the Max Voting Ensemble Method is:

0.77

b)Advanced Ensemble Method of Bagging meta-estimator

Bagging meta-estimator is an ensembling algorithm that can be used for both classification (BaggingClassifier) and regression (BaggingRegressor) problems. It follows the typical bagging technique to make predictions. Following are the steps for the bagging meta-estimator algorithm:

The score using the Decision Tree bagging Classifier is:

0.91

The ROC Curve using above model is as follows: