CSE303: Statistics for Data Science [Spring 2021]

Project Report

Submitted by:

Student ID	Student Name	Contribution Percentage
2018-2-60-033	Bijoy Basak	**29%**
2018-2-60-036	Simonta Chakraborty	**32%**
2018-2-60-111	Farhan Ahmed	**39%**

1. Introduction

From the given dataset we had to analyse and then implement as per our understandings. We found many categorical attributes in both our train and test dataset. In our project for developing models, we used machine learning model such as Logistic Regression, Linear Support Vector Classifier, Ridge Regression and Lasso Regression. We used data processing for checking null and unique value. After that we encoded categorical values from given dataset. For encoding we used Label Encoder. Then we tried to reduce dimension of the dataset. For that we used Standard Scaler. After all of this we fitted our dataset into above discussed models. Then we focused on evaluating the models. We got accuracy, precision, recall, f1 score and ROC area under the curve score for each of our models. We used matplotlib.pyplot and seaborn API for visualization.

2. Exploratory Data Analysis

After exploratory data analysis, these are our findings:

None of the train or test datasets had any null values. Pandas built in nunique() function was used to check for unique data in the dataset. Pandas drop_duplicates() was used to drop duplicate. But after using the function to remove duplicates, the dataset size was same as before. Thus, we can say there was no duplicate values in datasets as well.

Now, for the features in dataset there are 19 categorical and 11 numerical variables. A few plots of the train dataset shown below.

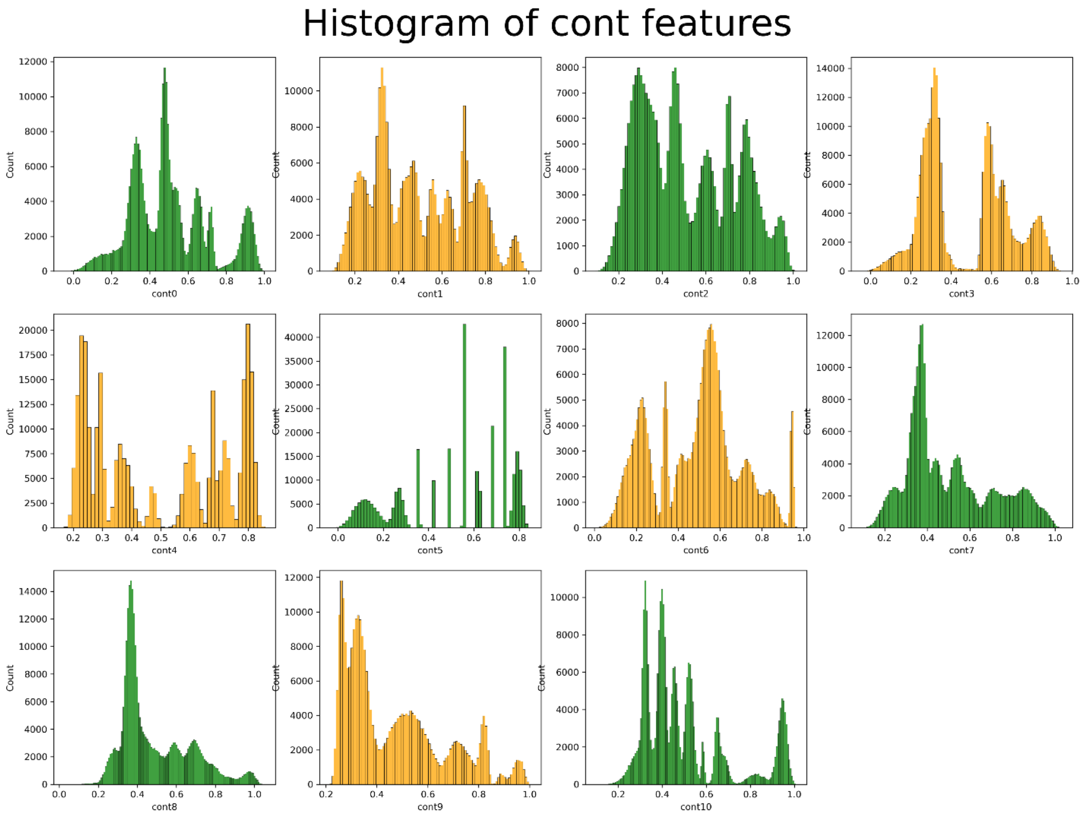


Fig 1: Histogram of cont features

From the above plot, we can see that none of the features are Normally Distributed as there is no plot which has bell shaped curve.

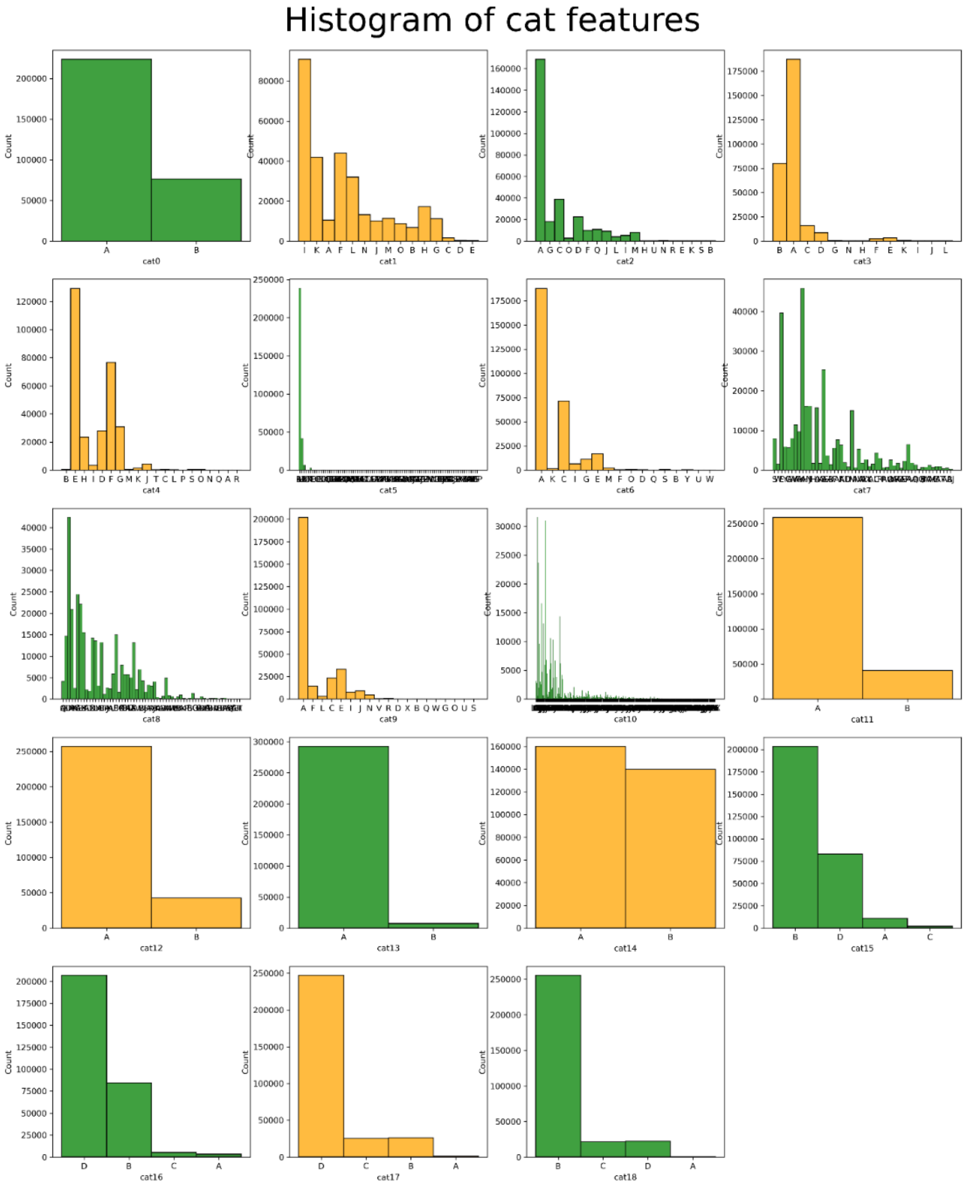


Fig 2: Histogram of cat features

From the above figure, we see that most of the categorical attributes are binary. None of them are Normally distributed as well.

In the following figure we have the Correlation Heatmap of the numerical attributes of the dataset. If we observe, we can see that column ‘Cont0’ has quite high correlation with column ‘Cont10’ and ‘Cont7’. Also, ‘Cont7’ and ‘Cont10’ have a significant correlation.

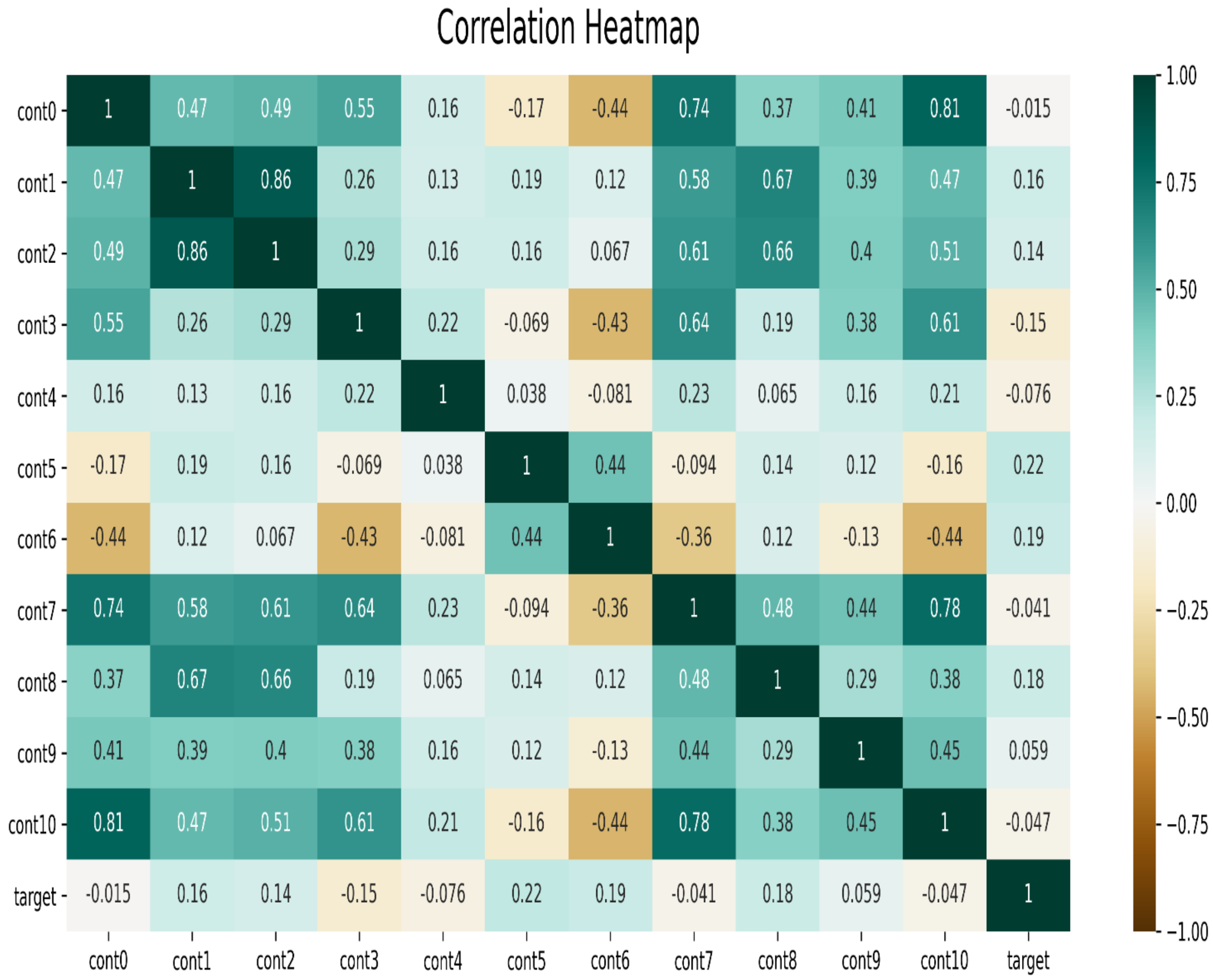


Fig 3: Correlation Heatmap

Lastly in figure 4, we see the ‘target’ column’s correlation heatmap with other columns. The column has the highest correlation with ‘cont6’ and lowest with ‘cont3’.

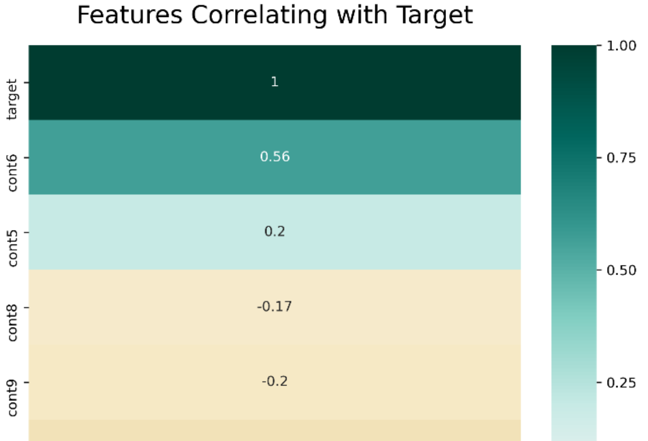


Fig 4: Correlation with Target

3. Machine Learning Models

Logistic Regression

Logistic regression is a supervised classification algorithm. It is the alternation regression analysis when the dependent variable has a binary solution. Logistic regression gives the result which range is 0 to 1, because it uses sigmoid function. There are three types of logistic regression such as, Binomial, Multinomial and Ordinal. In this dataset we used binomial logistic regression. Logistic regression nicely fits our dataset as the values are between 0 to 1 and the decision boundary is 0.5. Logistic regression model curve is S-shape.

Linear Support Vector Classifier

Linear support vector classifier is similar to support vector classifier when the kernel value is 'linear'. For penalizing it has more flexibility than support vector classifiers and performs better in terms of a large number of datasets.

Ridge Regression

Ridge regression is a regularization model and a type of linear regression model which avoids over or underfitting of the model by implementing a penalty. It is also known as L2 regularization.The ridge regression does not fit the training set as well as linear or polynomial regression and as output of this is a more accurate and better model. By providing penalty value it makes some bias to the model and the amount of bias we get are in small numbers. For that small amount of bias the ridge line is fit to the dataset and it returns for that small amount of bias, variance can be dropped. Because of the penalties, the coefficients tend to go close to zero but never become absolute zero. We can say ridge regression can predict well.

Lasso Regression

Lasso regression is also known as L1 regularization. This is also a model that uses shrinkage. By using lasso, we can convert high to low variance. It can avoid models from overfitting. For multiplying penalty values with the sum of weights, the coefficients can become zero. Lasso penalizes less important features. Lasso sometimes struggles with some types of data and selects one of the collinear variables randomly.

4. Data Pre-processing

First, we checked null and unique values for both test and train dataset. Then we dropped duplicate values from our dataset. For replacing categorical values, we did encode. Because machine learning models can only deal with numerical values. Finally, we tried to reduce dimension of our given dataset.

5. Different Models

Model Name	Parameter	Description
Logistic Regression	solver='liblinear'	For high dimension dataset it gives better result than others.
	C =1	For handling missed classified data and strengthen regularization.
	penalty = l1	For limiting the size of coefficient.
LinearSVC	C =1	For handling missed classified data and strengthen regularization.
Ridge Regression	alpha = 1	For decreasing variance.
	solver = 'sag'	Fast converges on columns that are on the same scale.
Lasso Regression	alpha=0.001	Best cross validation for lasso regression.

6. Performance Evaluation

On the given dataset we used four machine learning models. They are Logistic Regression, LinearSVC, Ridge Regression and Lasso Regression.

Accuracy of the Models:
The best accuracy we got is from the LinearSVC Model. Accuracy Score of LinearSVC is Almost 83.8%. The worst Accuracy Score is from Ridge Regression Model, the score is less than 83.7%. Logistic and Lasso Regression have almost the same accuracy score of ~83.73%.

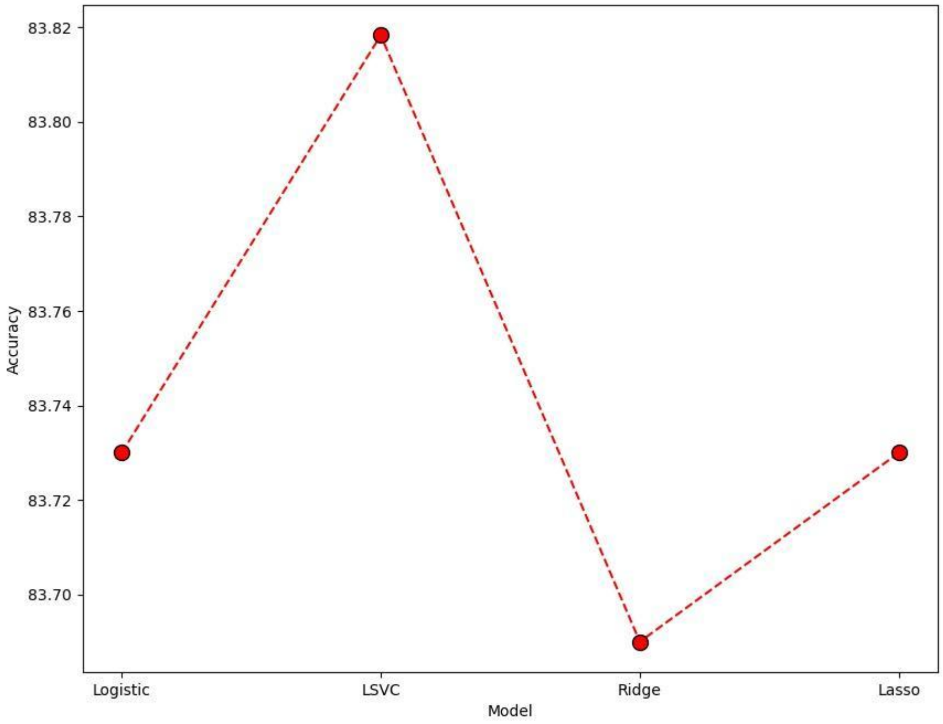


Figure 5: Accuracy of Various Models

Precision and Recall:

In the below figure, we can see Precision of the models on the left subplot and Recall of the Models on the right subplot. The Precisions and recalls are almost equal for each model. But, if compared to other models similarly as the Accuracy Score, highest precision and recall is for LinearSVC and Lowest is Ridge while other two are almost similar.

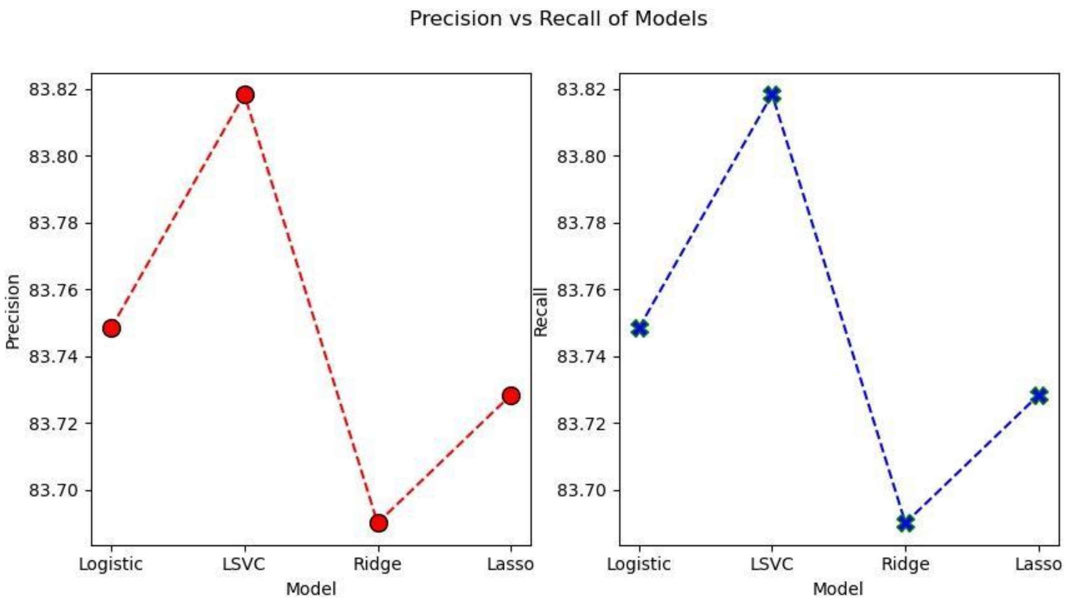


Figure 6: Precision vs Recall

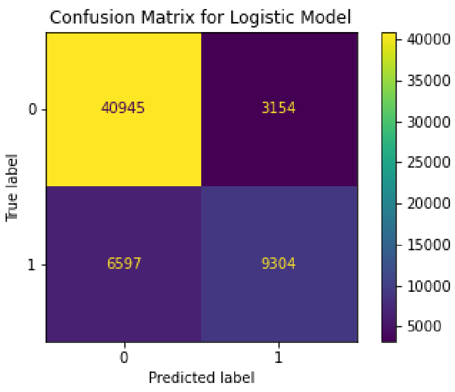
F1 Score and ROC Auc Score:

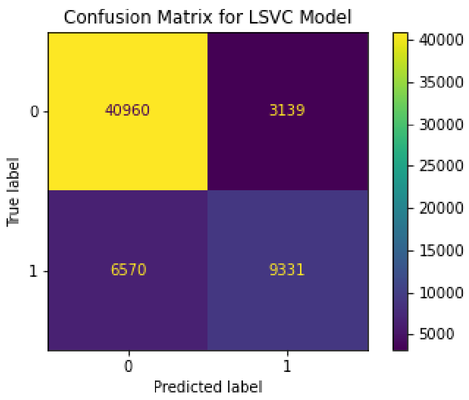
F1 and Auc Scores of the models are following.

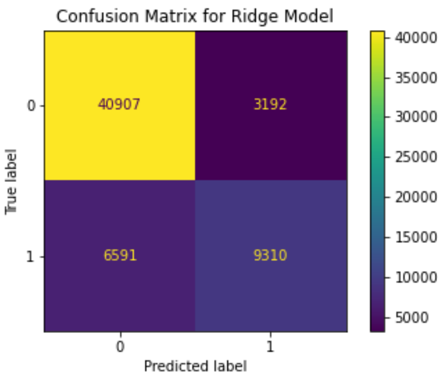
Model	F1 Score	ROC Auc Score
Logistic Model	65.61585387354984	75.67997682558722
LinearSVC	65.77843572662225	75.78188433469337
Ridge Regression	65.54467995211604	75.648335961115
Lasso Regression	65.57111118947702	75.65229656853217

Confusion Matrix:

We can see the confusion matrix of the Models below. We can see that they have highest number in True Positive.







7. Discussion

Our dataset had a lot categorical and continuous values. To use such dataset, we had to pre-process the data as well. Finally, the output had be 'target' column with binary values. Not every Model is good with binomial values. This, we have seen in the Performance Evaluation seen already. LinearSVC model outperforms all the other models. On the other hand, ridge regression had the lowest accuracy.

LinearSVC is Linear Support Vector Classification which works great with large sized data. It is also very flexible in terms of penalty and for multiclass data. The LinearSVC reduces variance greatly. Thus, it increases the overall performance of the Model.

Logistic Regression is also great with binomial valued datasets. It has the concept of odds which is greatly beneficial. But as it creates a S shape curve the accuracy can be less than linear models.

Lastly, for Lasso and Ridge regression models, the lasso regression model performed close to logistic regression but ridge regression performed the worst among all the models. Lasso and Ridge both are regularization methods. But the penalization of lasso regression is more aggressive than ridge. Lasso regression also has feature selection which ridge lacks. As the dataset is quite large and has a lot of categorical and continuous value the lack of aggressive penalization and feature selection the ridge could not perform well. But with more processing of the data performance can be improved, not for just Ridge regression model but may favour all four of them.

ROC AUC Plots:

If we see the ROC AUC Plots, we see that Lasso Regression has the highest ROC AUC value among all of the graphs.

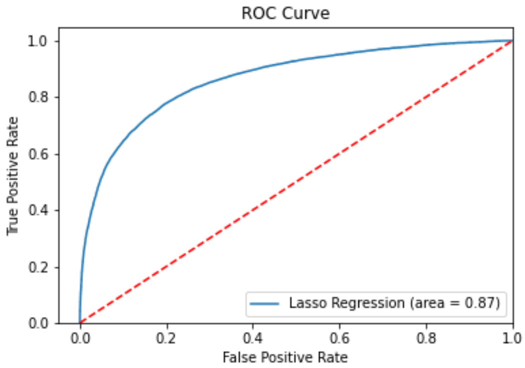


Figure 7.1: ROC AUC for Lasso Reg.

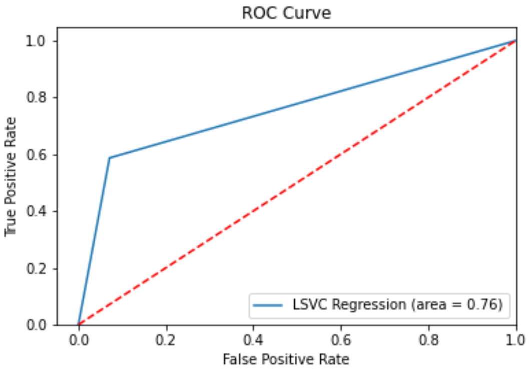


Figure 7.2: ROC AUC for LSVC

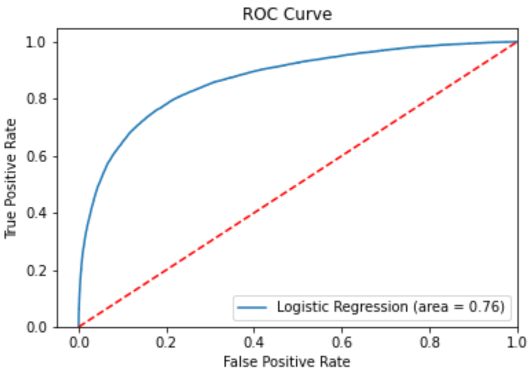


Figure 7.3: ROC AUC for Logistic Reg.
Reg.

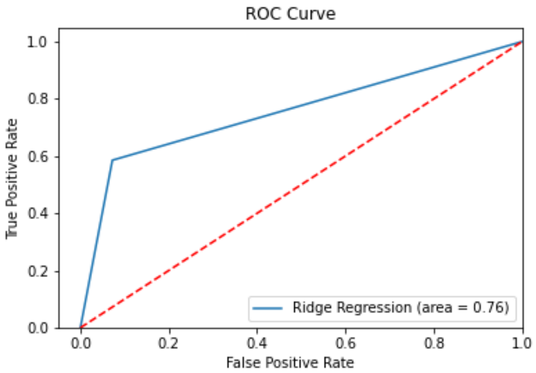


Figure 7.4: ROC AUC for Ridge

8. Conclusion

After finishing this project, we feel like it was roller coaster ride. We faced many difficulties. The main difficult thing was to improve accuracy of the models. Even after working so hard, we can hardly change the accuracy. Although there were many difficulties, we learnt new concepts like null value detection, duplicate value checking, encoding, scaling etc. We tried to apply all our knowledge that achieved from this course. But we are very disappointed that we could not change the accuracy significantly. It was challenging but interesting project for us, we learnt many new things and improved our teamwork. Also, we got a glimpse of future that, we may have to spend a lot of sleepless nights if we can become data scientists as we have spent for this project.