

PREDICTIVE MODELLING

PROJECT

NAME – DEEPAK SINGH

BATCH – PGDSBA SEPT D

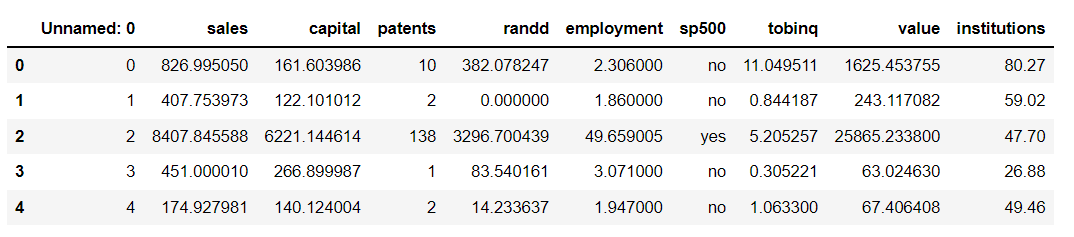
Problem 1: Linear Regression.

You are a part of an investment firm and your work is to do research about these 759 firms. You are provided with the dataset containing the sales and other attributes of these 759 firms. Predict the sales of these firms on the bases of the details given in the dataset so as to help your company in investing consciously. Also, provide them with 5 attributes that are most important.

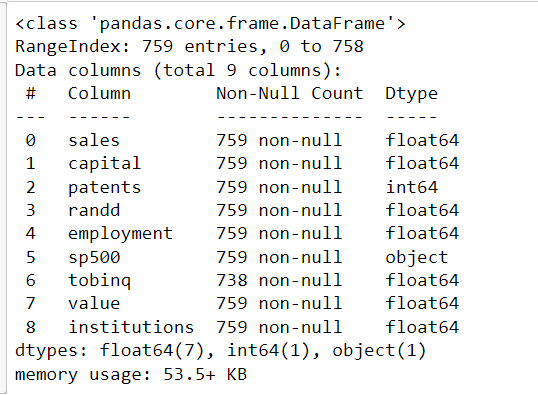
Questions for Problem 1:

1.1) Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, data types, shape, EDA). Perform Univariate and Bivariate Analysis. (8 marks)

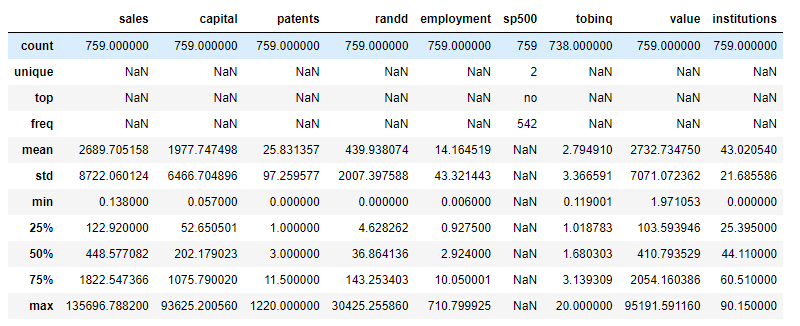
Read and check head and tail of the data



Checking shape, information and summary statistics There are 759 observations for 10 variables We can remove the column 'Unnamed: 0

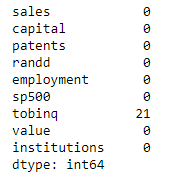


There are Integer, Float and Object Data Type.



This data set continuous, discrete and categorical data 'sp500' has categorical data whereas 'patents' has discrete data and all other variables have continuous data for most of the variables except 'institutions’, the mean is far greater than median indicating positive skewness in the data Standard deviation is also high for all numerical variables.

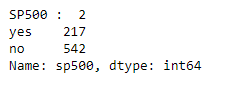
Check for Nulls values:



There are null values for the variable 'tobinq', the ratio between a physical asset's market value and its replacement value. In the variable 'tobinq', we have 2.77% of null values.

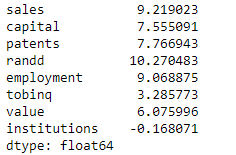
Check for duplicates: There are no duplicates in the data set.

Checking unique values for categorical variable 'sp500’:

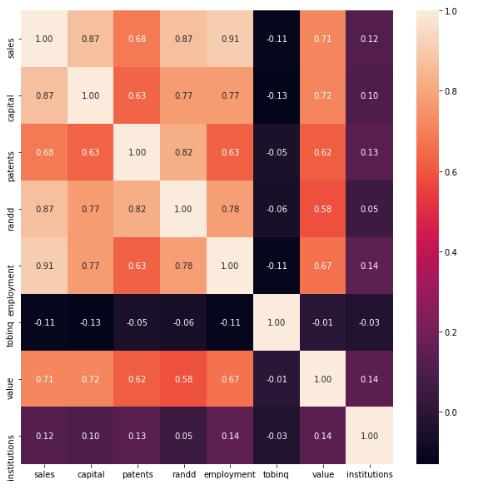


217 firms of the total firms in the dataset are members of S&P 500 Index.

Conduct Univariate Analysis As per the analysis, we see that all the continuous variables - sales, capital, patents, randd, employment, tobinq, value except institutions have skewed data.



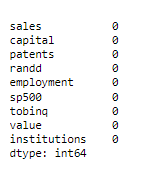
**\*Please refer to input 15 of the jupyter file – Linear Regression Project final Submitted.**



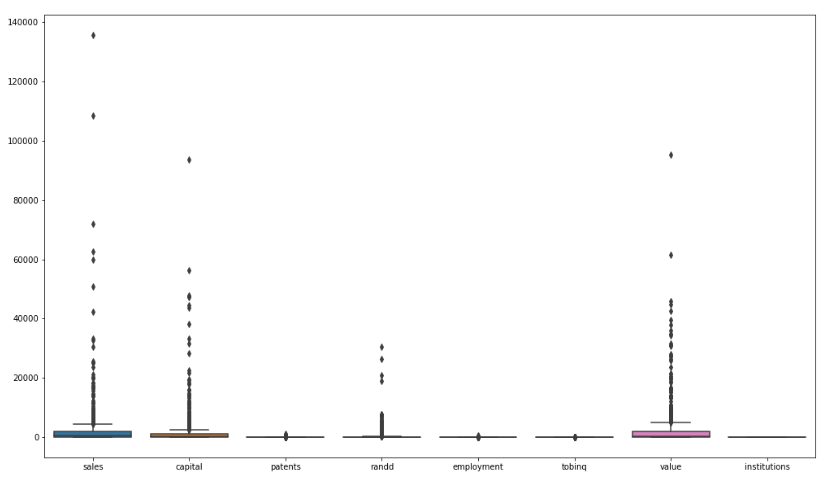
Pair plot and correlation matrix shows that there is correlation between capital and randd and capital and employment. It also shows strong correlation between dependent variable sales with capital, randd, employment and value

1.2) Impute null values if present? Do you think scaling is necessary in this case? (8 marks)

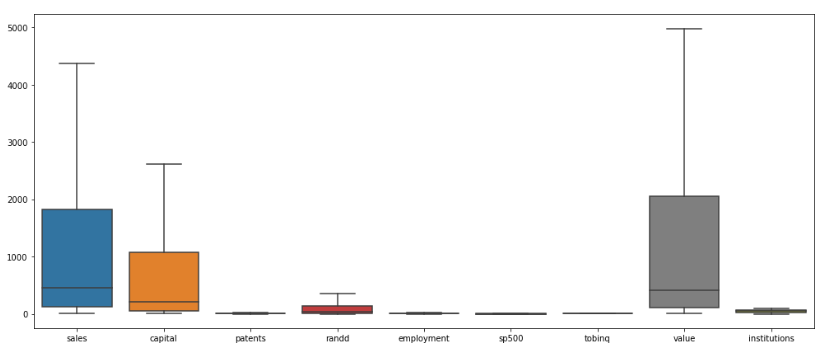
We have Null values in tobinq, since tobinq being continuous variable and a ratio null values are imputed using median.



There are outliers, the boxplots with outliers.

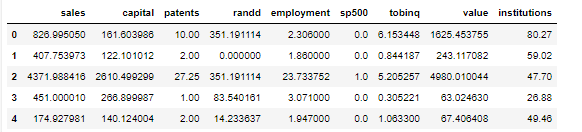


After outlier treatment.

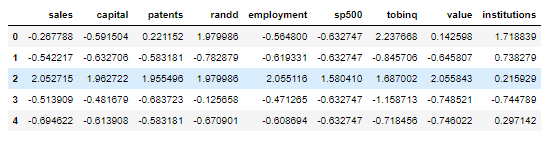


The different variables are in different scale or magnitude, patents are whole numbers , Tobinq is a ratio converted in decimal numbers and similarly all other variables are in different scale, hence scaling will be required.

Before scaling:

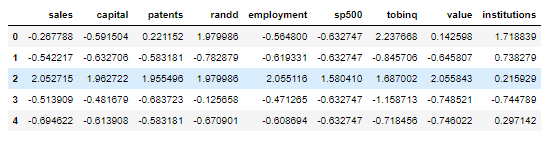


After Scaling:



1.3) Encode the data (having string values) for Modelling. Data Split: Split the data into test and train (30:70). Apply Linear regression. Performance Metrics: Check the performance of Predictions on Train and Test sets using R-square, RMSE. (8 marks).

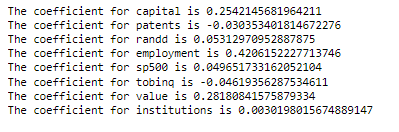
One feature 'sp500' with string value is encoded to create dummies



Data set is split into train and test data in 70:30 ratio-

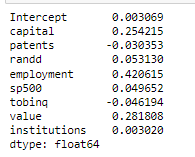


Linear Regression model is applied:

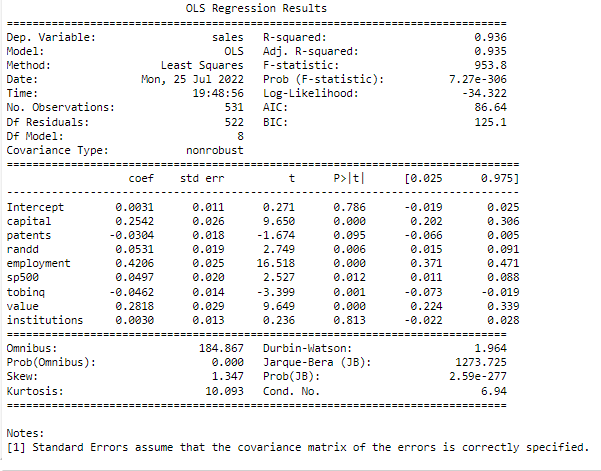


1. The intercept for our model is -0.02834809299281312
2. R square on the train data 0.9359702538559448
3. 93% of the variation in the sales is explained by the predictors in the model for train set
4. R square on the test data 0.9240311293641786
5. RMSE train 0.2581275829531501
6. RMSE test 0.2618357790172932

Linear Regression using stats models:

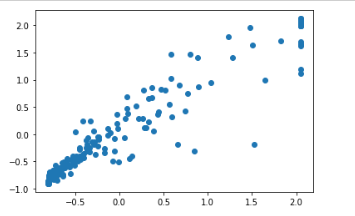


Inferential Statistics:



1. **mean\_sq\_error train - 0.2581275829531501**
2. **mean-sq-error test - 0.2618357790172932**

Prediction Scatter plot



(0.0) \* Intercept + (0.25) \* capital + (-0.03) \* patents + (0.05) \* randd + (0.42) \* employment + (0.05) \* sp500 + (-0.05) \* tobinq + (0.28) \* value + (0.0) \* institutions +

1.4) Inference: Based on these predictions, what are the business insights and recommendations. (6 marks)

1. As per the model output below features are significant
2. Capital has positive impact on the sales
3. Number of patents and tobinq have slight negative impact on the sales
4. whereas randd has positive impact on the sales increase and employment has the highest impact on the sales performance
5. With low beta coefficient the current the above attribute should be able to predict the sales performance significantly in the future.

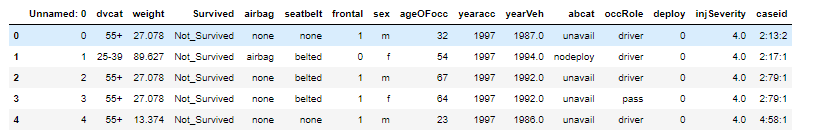
Problem 2: Logistic Regression and Linear Discriminant Analysis

You are hired by the Government to do an analysis of car crashes. You are provided details of car crashes, among which some people survived and some didn't. You have to help the government in predicting whether a person will survive or not on the basis of the information given in the data set so as to provide insights that will help the government to make stronger laws for car manufacturers to ensure safety measures. Also, find out the important factors on the basis of which you made your predictions.

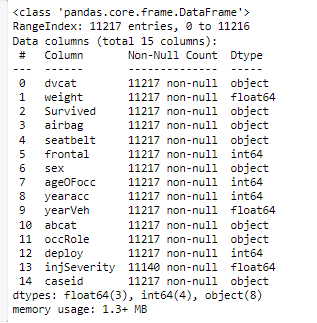
Questions for Problem 2:

2.1) Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis. (8 marks)

Survived is the dependent variable. All other are variables are independent.



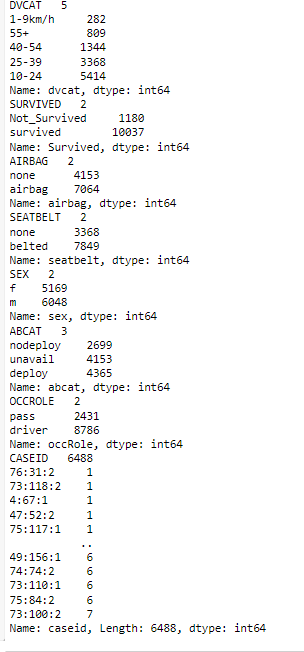
1. 8 object data variables ()-dvcat, Survived, airbag, seatbelt, sex, abcat, occRole, caseid.5 are integer or float variables.
2. You can drop caseid as it has no relevance when it comes to analyzing the data set
3. 77 values in in Severity are having null values. Basically, these variable states the extent of the injury.
4. we would need to see how to impute the missing data. Since it has distinct values
5. Let’s check how many duplicate values are there are 7 distinct values in the in Severity variable, it would be best to impute the 77 null values with mode of the variable



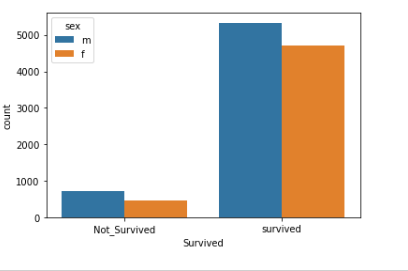
no duplicate values found in the data set:

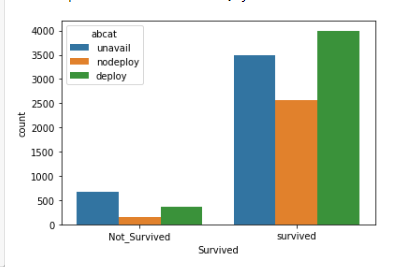


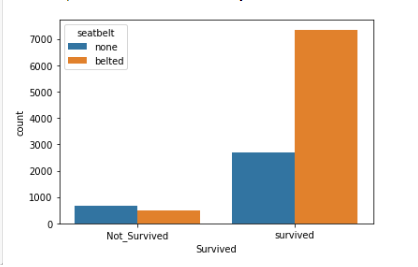
Getting counts of object variable:

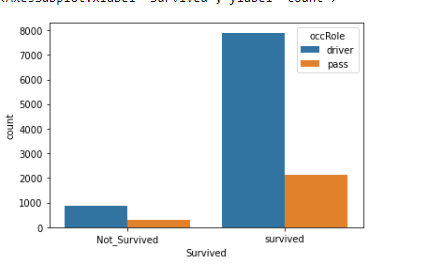


Below is the pictorial representation of all the DVcat, Survived, airbag, Seatbelt, Sex, Abcat, Occ Role variables:



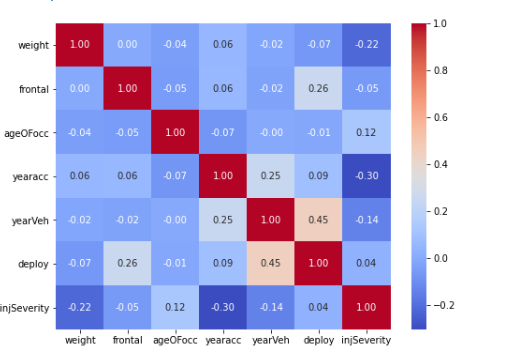




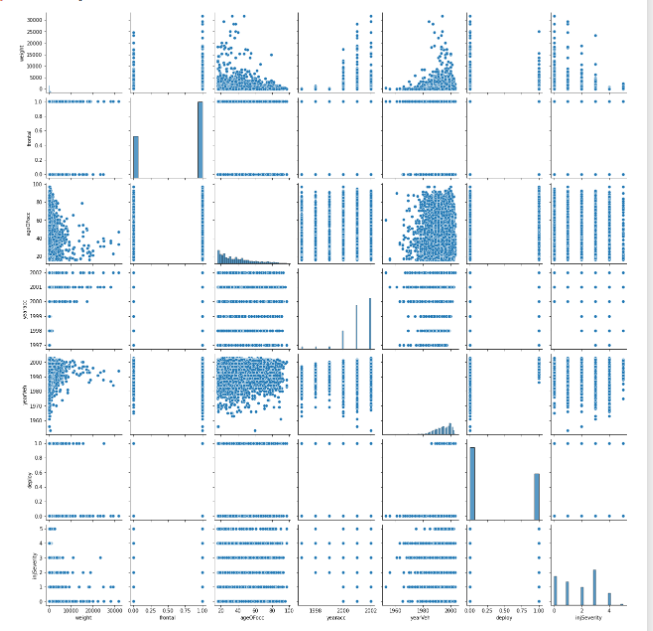


Checking for Outliers I have checked Weight and AgeOfOcc variable and both have outliers:

Correlation Matrix: no correlation ship exists between independent variables:

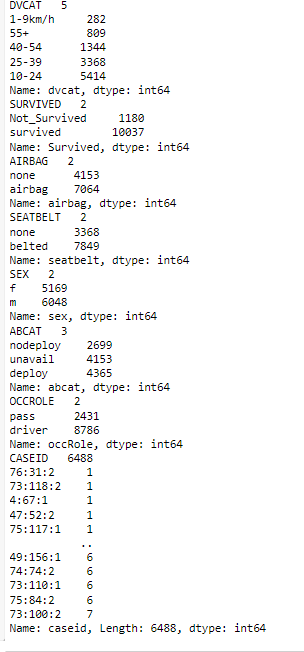


There is not much correlation ship between any independent variables

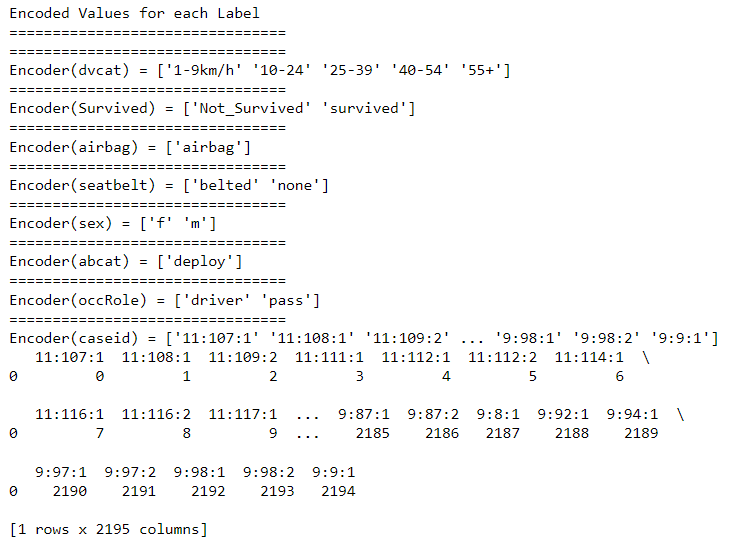


2.2) Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis). (8 marks)

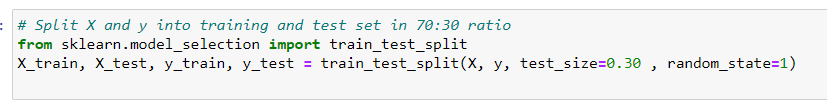
Since Machine level algorithms require us to convert String values to Numeric Values, we need to convert the same using Label encoding:



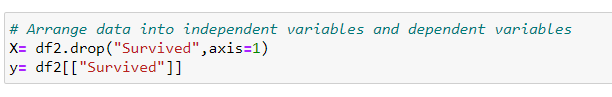
Post Label encoding: After converting the String values to Machine readable Values, below are the counts and corresponding values for the same:



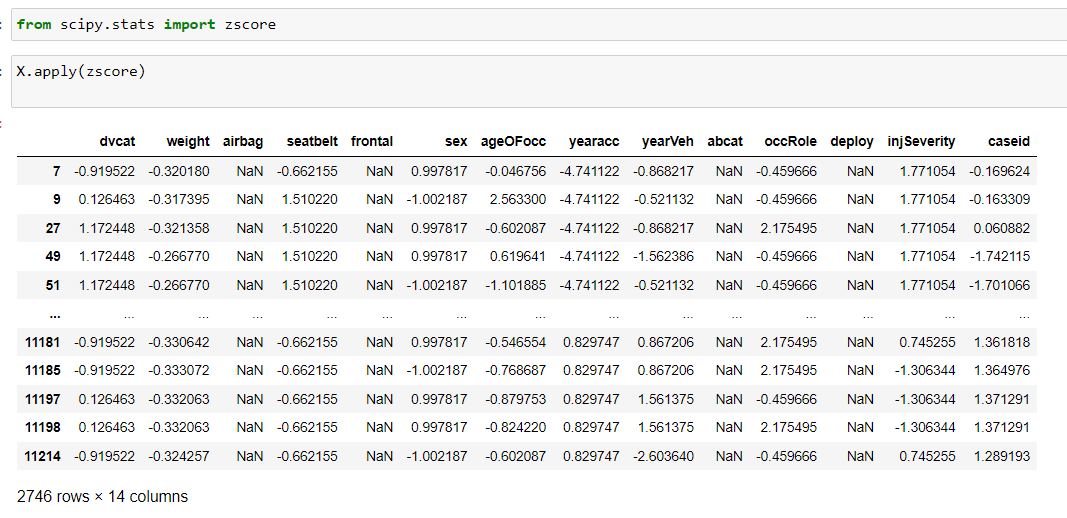
Splitting data into training and test data set: I have split the data post removal of caseid columns as this variable is not useful. I have split the data in 70%(training Data Set) and 30%(testing data set):



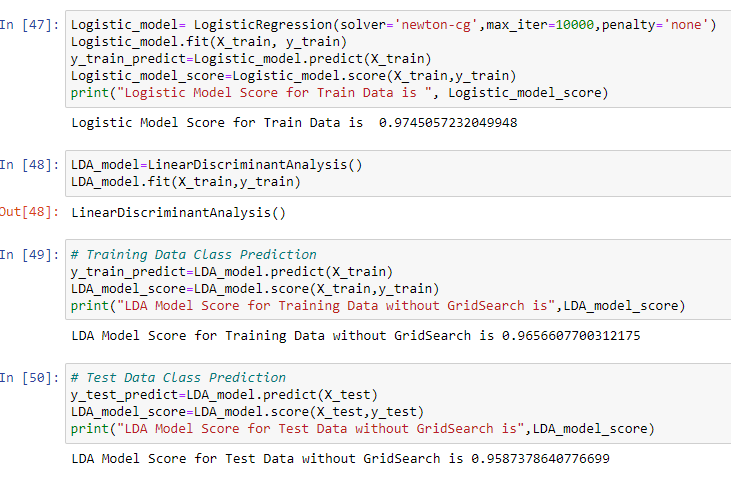
Target variable taken is Survived:



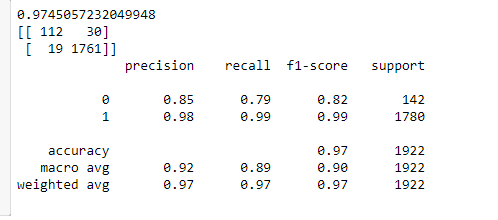
Scaling is applied as the data for Logistic Regression and LDA is very sensitive and the categorical count in this data set very high.

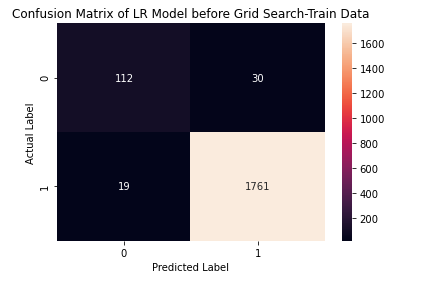


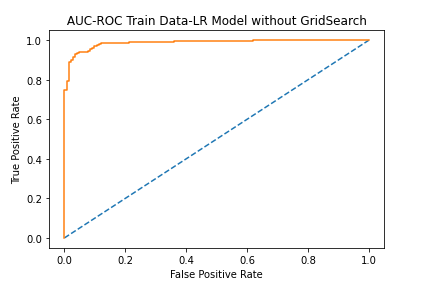
Applied Logistic Regression and LDA (linear discriminant analysis):



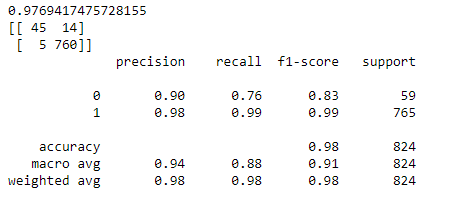
Metrics for Logistic Regression Train data:

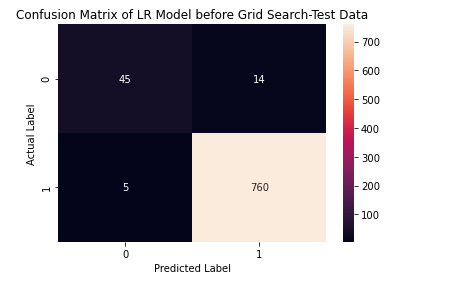


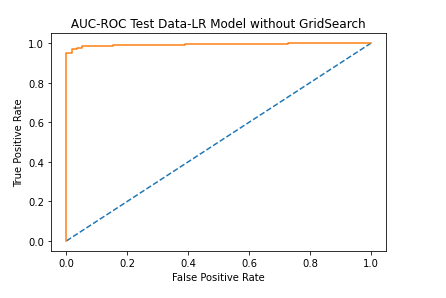




Testing Data:







Inference:

By using the performance metrics such as confusion matrix, the True Positive, False Positive, False Negative, and True Negative values can be extracted which will aid in the calculation of the accuracy score, precision score, recall score, and f1score. Listing below model performance metrics before fine tuning the mode For train data: tp-112 tn-1761 fp-68 fn- 19 Auc- 0.987 Precision-.98 Recall-0.99 Accuracy- 97 For test data: tp-45 tn- 760 fp-5 fn-14 Auc-99 Precision-0.98 Recall-0.99 Accuracy-98 Here, both Type I Error (False Positives) and Type II Error (False Negatives) are low indicating high Sensitivity/Recall, Precision, Specificity and F1 Score. Train and Test data scores are mostly in line and the overall performance of model looks good.

Best/optimized Model From the performance metrics we can come to a conclusion that the best optimizes model is Logistic Regression.

* 1. Inference:

Based on these predictions, what are the insights and recommendations. Based on the insights Government can issue the following features as a mandatory inclusion while manufacturing a car :

1. Airbags A car’s airbags will deploy in a crash and cushion both the driver and passengers from common impact points, like the steering wheel, dash, and sides of the vehicle. Also, check for safety features that will turn off the airbags if a child is placed in the front seat, as airbags are quite dangerous when children are not riding in the back.

2. Antilock Brakes Antilock brakes prevent the wheels from locking when you hit the brakes quickly. Locked wheels can cause the car to spin out on a slippery service, making it impossible for you to steer. An ABS system will use sensors on each wheel to pulse the braking system to each wheel during hard braking in order to prevent locking the wheels. This safety feature will help keep you in control, even when you have to hit the brakes hard.

3. Electronic Stability Control These sensors detect things like wheel speed, sideways motion, steering angel and rotation. If the car moves out of the driver’s intended path, the system will apply brakes to one or more of the wheels to help the driver regain stability.

4. Adaptive Headlights Headlights improve vision after the sun sets, and adaptive headlights help take visibility a step farther. These systems use sensors that monitor the car’s elevation to illuminate the road more effectively while also reducing glare for oncoming drivers. By improving visibility and making your car easier for other driver’s to see, adaptive headlights are an important safety feature.

5. Traction Control Finally, when considering car safety features, consider traction control. This is an electronic control system that limits how much the wheels can spin when you are accelerating. This means that the wheels will have maximum traction, even in wet or slippery conditions

2.3) Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model. Compare both the models and write inferences, which model is best/optimized. (8 marks)

2.4) Inference: Based on these predictions, what are the insights and recommendations. (6 marks