**Car Crash Analysis**

Name: Saurabh Dharmadhikari

PGP-DSBA Online January’ 21

Date: 18/10/2022

Table of Contents:

Executive Summary……………………...................................................................................................................................3

Introduction ...................................................................................................................................................................3

Data Description ............................................................................................................................................................3

Sample of the dataset ...................................................................................................................................................4

Exploratory Data Analysis..............................................................................................................................................5

Let us check the types of variables in the data frame. .................................................................................................5

Check for missing values in the dataset .......................................................................................................................5

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. .........................................7

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). ......................................................................................12

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. …………………………………………………………………………………………………………...................15

2.4) Inference: Based on these predictions, what are the insights and recommendations. .....................................18

Conclusion & Insights ..................................................................................................................................................18

THE END!......................................................................................................................................................................18

Executive Summary

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.

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. The data consists of 11217 entries of accidents. We are also provided with different attributes concerning these accidents mentioned below in the Data Description. This data set provides us with information on number of incidents where passengers survived. Our mission is to build a model that can predict incidents where passengers survive in case of an accident. For this we will be developing a logistic regression model and linear discriminant analysis model.

Data Description

1. dvcat: factor with levels (estimated impact speeds) 1-9km/h, 10-24, 25-39, 40-54, 55+

2. weight: Observation weights, albeit of uncertain accuracy, designed to account for varying

sampling probabilities. (The inverse probability weighting estimator can be used to demonstrate

causality when the researcher cannot conduct a controlled experiment but has observed data to

model)

3. Survived: factor with levels Survived or not\_survived

4. airbag: a factor with levels none or airbag

5. seatbelt: a factor with levels none or belted

6. frontal: a numeric vector; 0 = non-frontal, 1=frontal impact

7. sex: a factor with levels f: Female or m: Male

8. ageOFocc: age of occupant in years

9. yearacc: year of accident

10. yearVeh: Year of model of vehicle; a numeric vector

11. abcat: Did one or more (driver or passenger) airbag(s) deploy? This factor has levels deploy,

nodeploy and unavail

12. occRole: a factor with levels driver or pass: passenger

13. deploy: a numeric vector: 0 if an airbag was unavailable or did not deploy; 1 if one or more

bags deployed.

14. injSeverity: a numeric vector; 0: none, 1: possible injury, 2: no incapacity, 3: incapacity, 4:

killed; 5: unknown, 6: prior death

15. caseid: character, created by pasting together the populations sampling unit, the case

number, and the vehicle number. Within each year, use this to uniquely identify the vehicle.

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.

Sample of the data set

|  | **Unnamed: 0** | **dvcat** | **weight** | **Survived** | **airbag** | **seatbelt** | **frontal** | **sex** | **ageOFocc** | **yearacc** | **yearVeh** | **abcat** | **occRole** | **deploy** | **injSeverity** | **caseid** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 0 | 55+ | 27.078 | Not\_Survived | none | none | 1 | m | 32 | 1997 | 1987.0 | unavail | driver | 0 | 4.0 | 2:13:2 |
| **1** | 1 | 25-39 | 89.627 | Not\_Survived | airbag | belted | 0 | f | 54 | 1997 | 1994.0 | nodeploy | driver | 0 | 4.0 | 2:17:1 |
| **2** | 2 | 55+ | 27.078 | Not\_Survived | none | belted | 1 | m | 67 | 1997 | 1992.0 | unavail | driver | 0 | 4.0 | 2:79:1 |
| **3** | 3 | 55+ | 27.078 | Not\_Survived | none | belted | 1 | f | 64 | 1997 | 1992.0 | unavail | pass | 0 | 4.0 | 2:79:1 |
| **4** | 4 | 55+ | 13.374 | Not\_Survived | none | none | 1 | m | 23 | 1997 | 1986.0 | unavail | driver | 0 | 4.0 | 4:58:1 |

We can see the initial 5 rows of our data set in the above sample. Let us study more about this data set.

Exploratory data analysis

RangeIndex: 11217 entries, 0 to 11216

Data columns (total 16 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Unnamed: 0 11217 non-null int64

1 dvcat 11217 non-null object

2 weight 11217 non-null float64

3 Survived 11217 non-null object

4 airbag 11217 non-null object

5 seatbelt 11217 non-null object

6 frontal 11217 non-null int64

7 sex 11217 non-null object

8 ageOFocc 11217 non-null int64

9 yearacc 11217 non-null int64

10 yearVeh 11217 non-null float64

11 abcat 11217 non-null object

12 occRole 11217 non-null object

13 deploy 11217 non-null int64

14 injSeverity 11140 non-null float64

15 caseid 11217 non-null object

dtypes: float64(3), int64(5), object(8)

We have a total of 11217 rows and 16 columns. We will not be needing Unnamed: 0 and caseid to build our model as Unnamed: 0 is like index numbering and a unique case id can not predict the outcome of an accident. These two columns will be removed latter.

So we are left with 7 object type and 7 integer or float type features. All these features could be good predictors for an outcome of an accident.

Check for null values

Unnamed: 0 0

dvcat 0

weight 0

Survived 0

airbag 0

seatbelt 0

frontal 0

sex 0

ageOFocc 0

yearacc 0

yearVeh 0

abcat 0

occRole 0

deploy 0

injSeverity 77

caseid 0

dtype: int64

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.

injSeverity has 77 missing values. These null values will be filled with mode value of injSeverity. This will help us keep the data in the other columns and also help us fill the missing values without making much difference to data in injSeverity.

Unnamed: 0 0

dvcat 0

weight 0

Survived 0

airbag 0

seatbelt 0

frontal 0

sex 0

ageOFocc 0

yearacc 0

yearVeh 0

abcat 0

occRole 0

deploy 0

injSeverity 0

caseid 0

dtype: int64

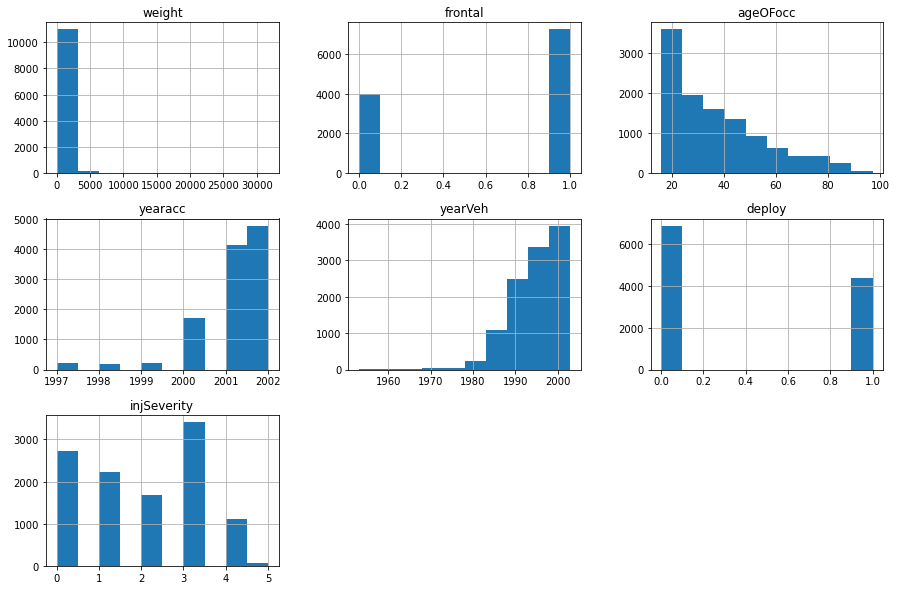
We can see that we have been able to fill the missing values in our data frame.

We have also removed the Unnamed: 0 and caseid columns.

|  | **count** | **unique** | **top** | **freq** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **dvcat** | 11217 | 5 | 10-24 | 5414 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **weight** | 11217.0 | NaN | NaN | NaN | 431.405309 | 1406.202941 | 0.0 | 28.292 | 82.195 | 324.056 | 31694.04 |
| **Survived** | 11217 | 2 | survived | 10037 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **airbag** | 11217 | 2 | airbag | 7064 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **seatbelt** | 11217 | 2 | belted | 7849 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **frontal** | 11217.0 | NaN | NaN | NaN | 0.644022 | 0.47883 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 |
| **sex** | 11217 | 2 | m | 6048 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **ageOFocc** | 11217.0 | NaN | NaN | NaN | 37.427654 | 18.192429 | 16.0 | 22.0 | 33.0 | 48.0 | 97.0 |
| **yearacc** | 11217.0 | NaN | NaN | NaN | 2001.103236 | 1.056805 | 1997.0 | 2001.0 | 2001.0 | 2002.0 | 2002.0 |
| **yearVeh** | 11217.0 | NaN | NaN | NaN | 1994.177944 | 5.658704 | 1953.0 | 1991.0 | 1995.0 | 1999.0 | 2003.0 |
| **abcat** | 11217 | 3 | deploy | 4365 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **occRole** | 11217 | 2 | driver | 8786 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **deploy** | 11217.0 | NaN | NaN | NaN | 0.389141 | 0.487577 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 |
| **injSeverity** | 11217.0 | NaN | NaN | NaN | 1.833645 | 1.377214 | 0.0 | 1.0 | 2.0 | 3.0 | 5.0 |

In the above table we can see the counts, mean, standard deviation, minimum value and maximum value. For object type features we can also see the unique values, most occurring category, and its frequency of occurrence.

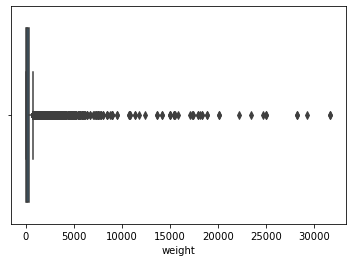
Univariate Analysis



In the above histograms we can see that most of the data we have is for front collisions, which is about 50% of the times. Records for accidents for age 16 to 20 is the highest. Also, records from 2021 and 2022 are the highest. It could be because of better record keeping and also because of increase in the number of cars. We see that cars manufactured in the year 2000 show more in our data. Most accidents end up with incapacity for the passenger according to the records.

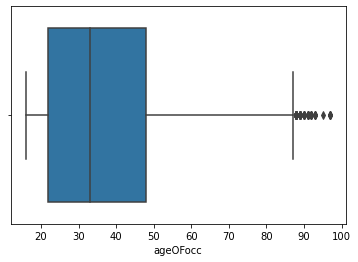
Let us see box plots of these features and also check for outliers.

Weight



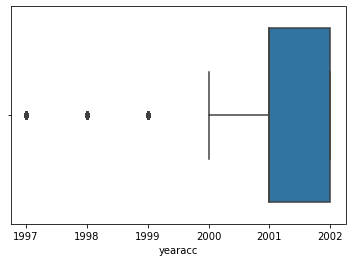
There are many outliers.

Age



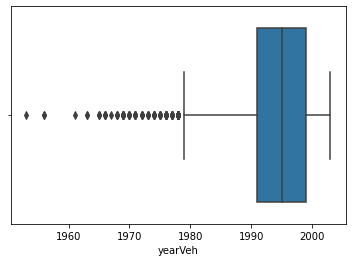
There are only a few outliers in this feature for, about 88 to 96.

Year of accident

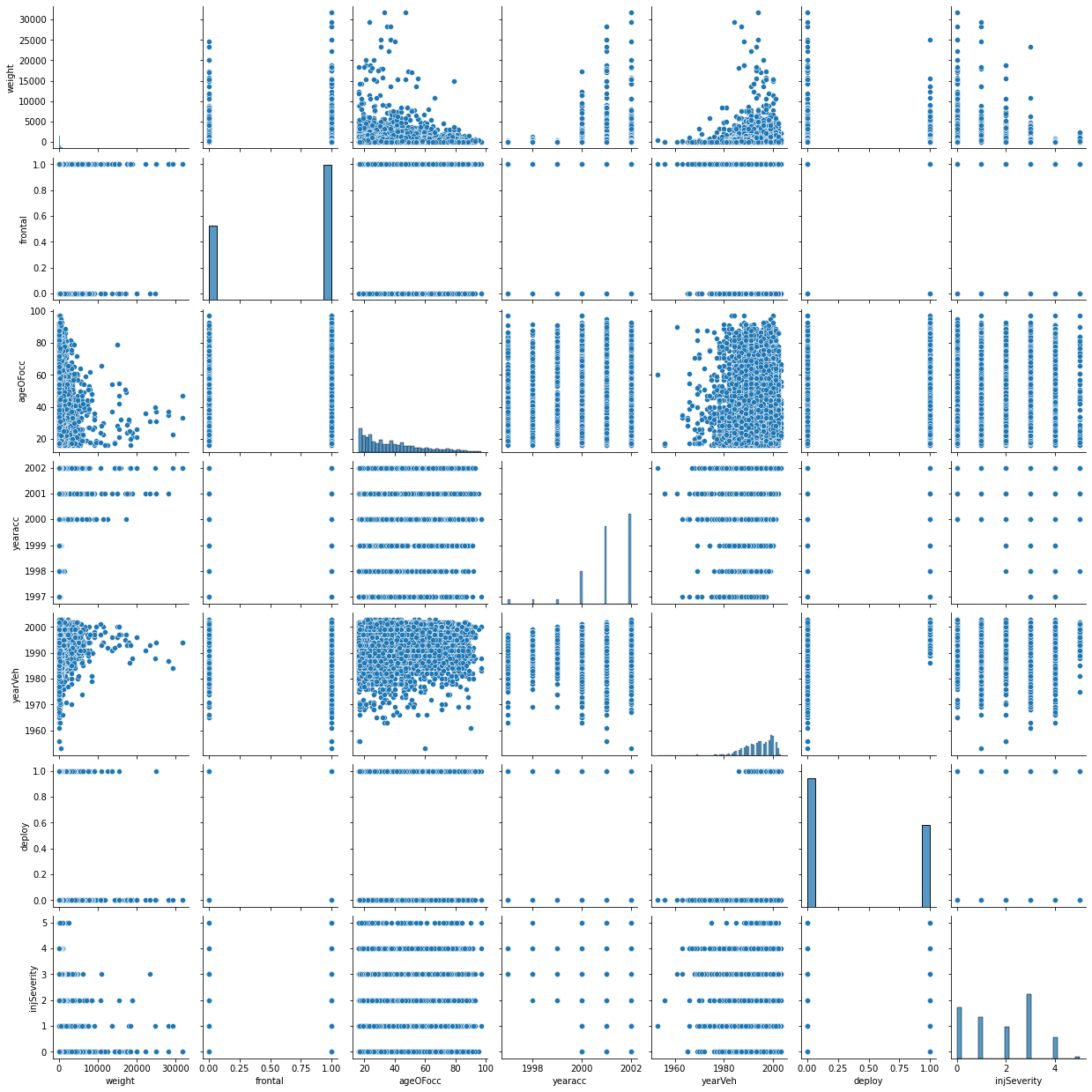


There are only 3 outliers.

Year vehicle was manufactured



Pair plot



We do not find a linear relationship between features, this could help us in our predictive model.

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).

We need to convert categorical values to numeric values for ease of building a model.

Hence, we will label values in these features.

|  | **dvcat** | **weight** | **Survived** | **airbag** | **seatbelt** | **frontal** | **sex** | **ageOFocc** | **yearacc** | **yearVeh** | **abcat** | **occRole** | **deploy** | **injSeverity** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 4 | 27.078 | 0 | 1 | 1 | 1 | 1 | 32 | 1997 | 1987.0 | 2 | 1 | 0 | 4.0 |
| **1** | 2 | 89.627 | 0 | 0 | 0 | 0 | 0 | 54 | 1997 | 1994.0 | 1 | 1 | 0 | 4.0 |
| **2** | 4 | 27.078 | 0 | 1 | 0 | 1 | 1 | 67 | 1997 | 1992.0 | 2 | 1 | 0 | 4.0 |
| **3** | 4 | 27.078 | 0 | 1 | 0 | 1 | 0 | 64 | 1997 | 1992.0 | 2 | 0 | 0 | 4.0 |
| **4** | 4 | 13.374 | 0 | 1 | 1 | 1 | 1 | 23 | 1997 | 1986.0 | 2 | 1 | 0 | 4.0 |

We can observe in the above table we have labelled as follows

Survived - 'survived': 1, ‘Not\_Survived': 0

airbag - 'none': 1, 'airbag': 0

seatbelt - 'none': 1, 'belted': 0

sex - 'm': 1, 'f': 0

occRole - 'driver': 1, 'pass': 0

dvcat - '55+': 4, '25-39': 2, '10-24': 1, '40-54': 3, '1-9km/h: 0'

RangeIndex: 11217 entries, 0 to 11216

Data columns (total 14 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 dvcat 11217 non-null int32

1 weight 11217 non-null float64

2 Survived 11217 non-null int64

3 airbag 11217 non-null int64

4 seatbelt 11217 non-null int64

5 frontal 11217 non-null int64

6 sex 11217 non-null int64

7 ageOFocc 11217 non-null int64

8 yearacc 11217 non-null int64

9 yearVeh 11217 non-null float64

10 abcat 11217 non-null int32

11 occRole 11217 non-null int64

12 deploy 11217 non-null int64

13 injSeverity 11217 non-null float64

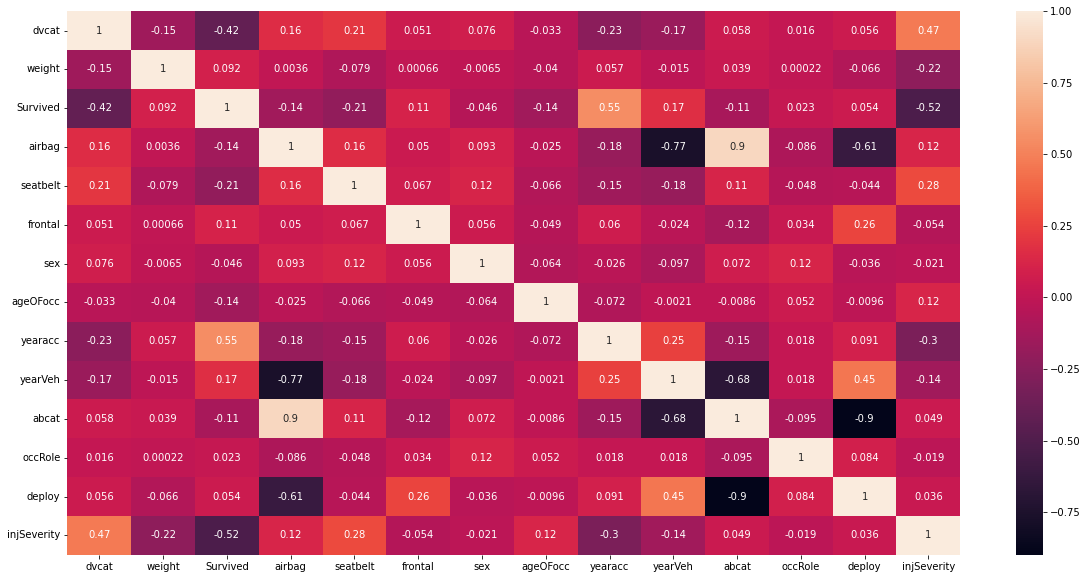
dtypes: float64(3), int32(2), int64(9)

Now we have all the features as integer type or float type.

Below is a correlation plot.

|  | **dvcat** | **weight** | **Survived** | **airbag** | **seatbelt** | **frontal** | **sex** | **ageOFocc** | **yearacc** | **yearVeh** | **abcat** | **occRole** | **deploy** | **injSeverity** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **dvcat** | 1.000000 | -0.145590 | -0.415593 | 0.160533 | 0.205410 | 0.051341 | 0.075768 | -0.033358 | -0.233494 | -0.166158 | 0.057847 | 0.015879 | 0.055622 | 0.469143 |
| **weight** | -0.145590 | 1.000000 | 0.091640 | 0.003574 | -0.078739 | 0.000659 | -0.006471 | -0.040111 | 0.056892 | -0.015226 | 0.038795 | 0.000219 | -0.065783 | -0.220659 |
| **Survived** | -0.415593 | 0.091640 | 1.000000 | -0.139679 | -0.206467 | 0.107990 | -0.046499 | -0.135473 | 0.549885 | 0.165096 | -0.107829 | 0.023460 | 0.054346 | -0.517637 |
| **airbag** | 0.160533 | 0.003574 | -0.139679 | 1.000000 | 0.157501 | 0.050272 | 0.092886 | -0.025109 | -0.181478 | -0.766181 | 0.896724 | -0.086011 | -0.611983 | 0.124394 |
| **seatbelt** | 0.205410 | -0.078739 | -0.206467 | 0.157501 | 1.000000 | 0.066590 | 0.117071 | -0.066066 | -0.149208 | -0.180534 | 0.111991 | -0.047712 | -0.044132 | 0.283063 |
| **frontal** | 0.051341 | 0.000659 | 0.107990 | 0.050272 | 0.066590 | 1.000000 | 0.055639 | -0.048856 | 0.059768 | -0.024267 | -0.117856 | 0.033721 | 0.260388 | -0.053709 |
| **sex** | 0.075768 | -0.006471 | -0.046499 | 0.092886 | 0.117071 | 0.055639 | 1.000000 | -0.063575 | -0.025957 | -0.097390 | 0.071708 | 0.116228 | -0.036143 | -0.021284 |
| **ageOFocc** | -0.033358 | -0.040111 | -0.135473 | -0.025109 | -0.066066 | -0.048856 | -0.063575 | 1.000000 | -0.072271 | -0.002070 | -0.008569 | 0.052485 | -0.009556 | 0.123495 |
| **yearacc** | -0.233494 | 0.056892 | 0.549885 | -0.181478 | -0.149208 | 0.059768 | -0.025957 | -0.072271 | 1.000000 | 0.247743 | -0.151650 | 0.018217 | 0.091252 | -0.300495 |
| **yearVeh** | -0.166158 | -0.015226 | 0.165096 | -0.766181 | -0.180534 | -0.024267 | -0.097390 | -0.002070 | 0.247743 | 1.000000 | -0.677852 | 0.018416 | 0.452448 | -0.138475 |
| **abcat** | 0.057847 | 0.038795 | -0.107829 | 0.896724 | 0.111991 | -0.117856 | 0.071708 | -0.008569 | -0.151650 | -0.677852 | 1.000000 | -0.094860 | -0.898811 | 0.048724 |
| **occRole** | 0.015879 | 0.000219 | 0.023460 | -0.086011 | -0.047712 | 0.033721 | 0.116228 | 0.052485 | 0.018217 | 0.018416 | -0.094860 | 1.000000 | 0.084323 | -0.018918 |
| **deploy** | 0.055622 | -0.065783 | 0.054346 | -0.611983 | -0.044132 | 0.260388 | -0.036143 | -0.009556 | 0.091252 | 0.452448 | -0.898811 | 0.084323 | 1.000000 | 0.036133 |
| **injSeverity** | 0.469143 | -0.220659 | -0.517637 | 0.124394 | 0.283063 | -0.053709 | -0.021284 | 0.123495 | -0.300495 | -0.138475 | 0.048724 | -0.018918 | 0.036133 | 1.000000 |

Heatmap



We do not have a very strong linear relationships between features. Except a few like abcat and airbag have a positive relationship. And yearVeh and airbag also abcat and deploy have a prominent negative relationship.

The data now looks clear and we are ready to build our prediction model.

We have taken a test size of 30% and rest is our train set.

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.

Logistic regression

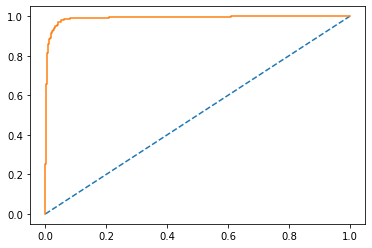
A total accuracy score for test set is 0.9824717765894236.

This is a good score for our prediction.

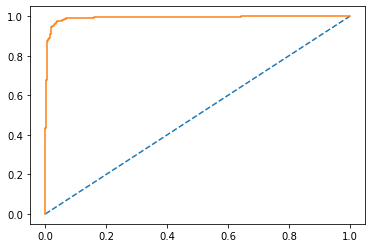
A total accuracy score for train set in 0.9801299197554452.

This is also a good score for our prediction.

We observe that our model is able to generalize well as we have good and a balanced accuracy scores for train set and test set.



We can see AUC and ROC in the above graph for train set.



We can see AUC and ROC in the above graph for test set. They both have a high accuracy score.

Following is the report of train set used and its confusion matrix.

precision recall f1-score support

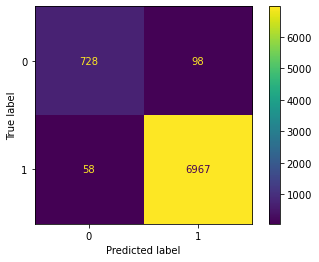
0 0.93 0.88 0.90 826

1 0.99 0.99 0.99 7025

accuracy 0.98 7851

macro avg 0.96 0.94 0.95 7851

weighted avg 0.98 0.98 0.98 7851



Following is the report of test set used and its confusion matrix.

precision recall f1-score support

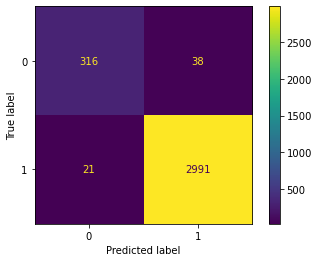
0 0.94 0.89 0.91 354

1 0.99 0.99 0.99 3012

accuracy 0.98 3366

macro avg 0.96 0.94 0.95 3366

weighted avg 0.98 0.98 0.98 3366



We can see in the confusion matrix that our model was able to predict 316 plus 2991 times right but it did not predict 21 plus 38 right.

This model seems very capable as the accuracy is very high.

Linear Discriminant Analysis

We have again taken the same data sets of train and test to build our model.

The following is the report after building and running our LDA model.

precision recall f1-score support

0 0.90 0.70 0.79 1180

1 0.97 0.99 0.98 10037

accuracy 0.96 11217

macro avg 0.93 0.85 0.88 11217

weighted avg 0.96 0.96 0.96 11217

We observe that the accuracy score for this model is about 96%.

This score is much less than the logistic regression model.

Thus, we can say that in this case study a logistic regression model performs far better than a linear discriminant analysis model.

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

Using logistic regression model, we can say

For {Passengers who did not survive (Label 0)}:

Precision (90%) – 90% of passengers who did not survive are correctly predicted, out of all passengers who did not survive that are predicted.

Recall (70%) – Out of all the passengers who actually did not survive, 70% of passengers who did not survive have been predicted correctly.

For {Passengers who did survive (Label 1)}:

Precision (97%) – 97% of Passengers who did survive are correctly predicted, out of all passengers who had accident that are predicted.

Recall (99%) – Out of all the passengers who actually did survive, 99% of Customers who did Churn have been correctly predicted.

Accuracy, AUC, Precision and Recall for test data is almost in line with training data. This proves no overfitting or underfitting has happened, and overall, the model is a good model for classification.