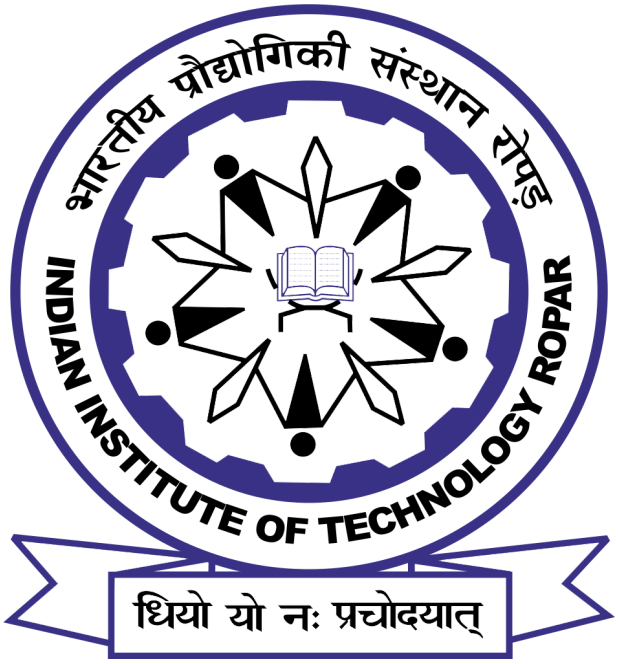
MA515 Project Report



Indian Institute of Technology, Ropar

Ropar – 140001

Submitted To - Dr. Arun Kumar

Date - 30th November 2021

Submitted By -

Pranjali Bajpai - 2018EEB1243

Table of Contents

1. Problem Statement

Do exploratory data analysis on the data. Use logistic regression and LDA to predict whether a customer will buy a Caravan Insurance policy. Compare the findings from different methods

2. About Dataset

2.1. The dataset consists of 5822 data points It has total 86 columns out of which the last column corresponds to output whether customer will purchase Caravan Insurance policy or not.

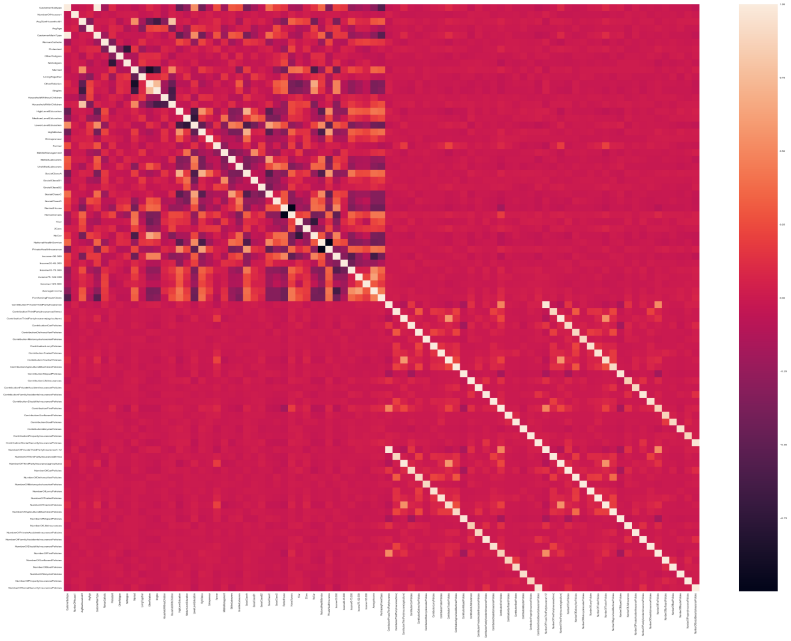
2.2. There are no Null values in any column of the dataset

2.3. There are no NA values in any column of the dataset

2.4. To conclude the dataset does not consist of any missing value

2.5. The predictors of the dataset are divided into two categories. Column codes beginning with M (MOSTYPE CustomerSubtype to MKOOPKLA PurchasingPowerClass) refer to demographic statistics of the postal code, while Column codes beginning with P and A refer to product ownership and insurance statistics in the postal code.

3. EDA



Correlation Plot

Following are the observations from the Correlation Plot:

3.1. There is high positive correlation between Customer Sub Type and Customer Main Type

3.2. Average Size Household is positively correlated with Household With Children

3.3. Married and No Car are negatively correlated

3.4. National Health Service and Private Health Insurance are negatively correlated

3.5. Social Class A is positively correlated with High level Education

3.6. Social Class A is positively correlated with Private Health Insurance

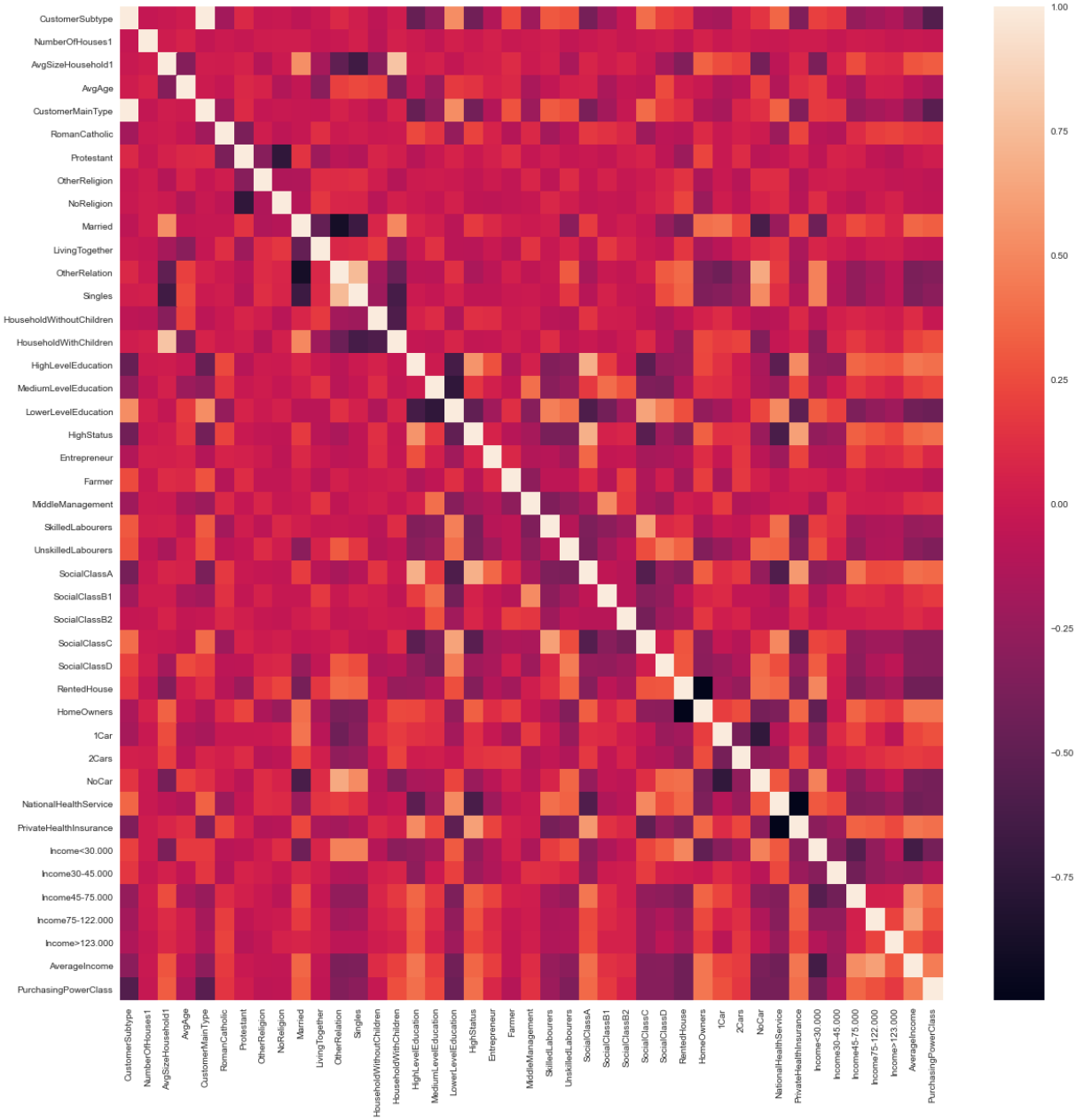
3.7. Social Class A is positively correlated with High Status

3.8. Social Class C is positively correlated with Lower Level Education

3.9. Social Class C is positively correlated with Private Health Insurance

3.10. National Health Service and Social Class A are negatively correlated

3.11. Skilled Laboureres and Social Class C are highly positively correlated



Correlation Plot of Demographic Statistics Variable

Histogram

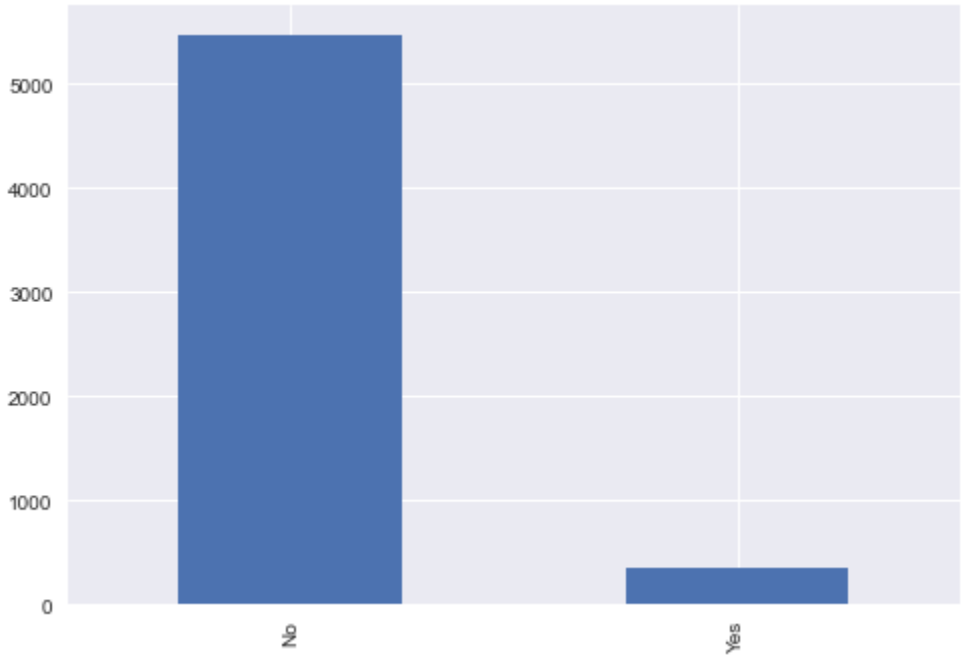
Following are the observations from the Histogram Plot:

3.12. The count of other policy buyers are very low in the dataset except Car Policy and Fire Policy

3.13. Similarly Contribution Car and Fire Policy is comparatively greater than other policy's contribution

3.14. Count of National Health Service is low as compared to Private Health Insurance

Analyzing Response Variable



Bar Plot of Response Variable

3.15. From the above plot we observe that the dataset is highly unbalanced as total percentage of "Yes" is approxiamately 5.97% and percentage of "No" is 94.02%

4. Data Preprocessing

4.1. Separate X and y.

4.2. Get the value of columns from the numerical label.

Get the value of AvgAge according to numerical label

- The AvgAge is a categorical column with value in range [1, 6] and with following meaning
  - 1: 20-30 years
  - 2: 30-40 years
  - 3: 40-50 years
  - 4: 50-60 years
  - 5: 60-70 years
  - 6: 70-80 years Set the value of the variable as median of the given range

In [1796]:
```
# Get the value of AvgAge according to numerical label
for i in range(rows):
    ageLabel=X[i, 3]
    if(ageLabel==1):
        X[i, 3]=25
    if(ageLabel==2):
        X[i, 3]=35
    if(ageLabel==3):
        X[i, 3]=45
    if(ageLabel==4):
        X[i, 3]=55
    if(ageLabel==5):
        X[i, 3]=65
    if(ageLabel==6):
        X[i, 3]=75
```

4.3. For later columns set the median value of the range as the value of variable.

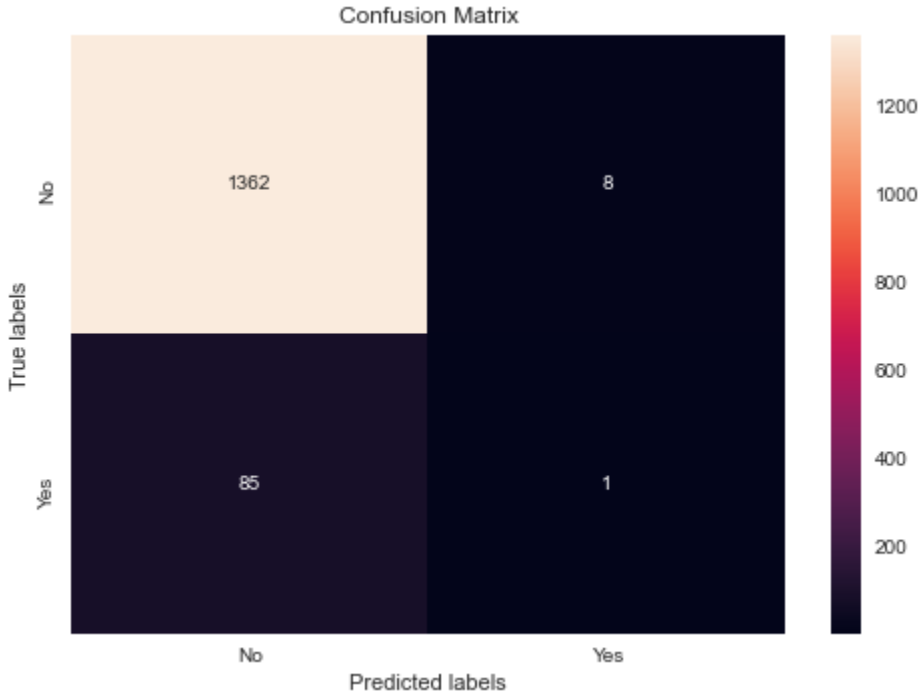
4.4. Use LabelEncoder to encode response variable labels to numerical values

4.5. Use OneHotEncoder for the categorical column CustomerSubType and CustomerMainType

4.6. After splitting the data into test set and training set and performing feature scaling, apply logistic regression and LDA to fit a model on training data.

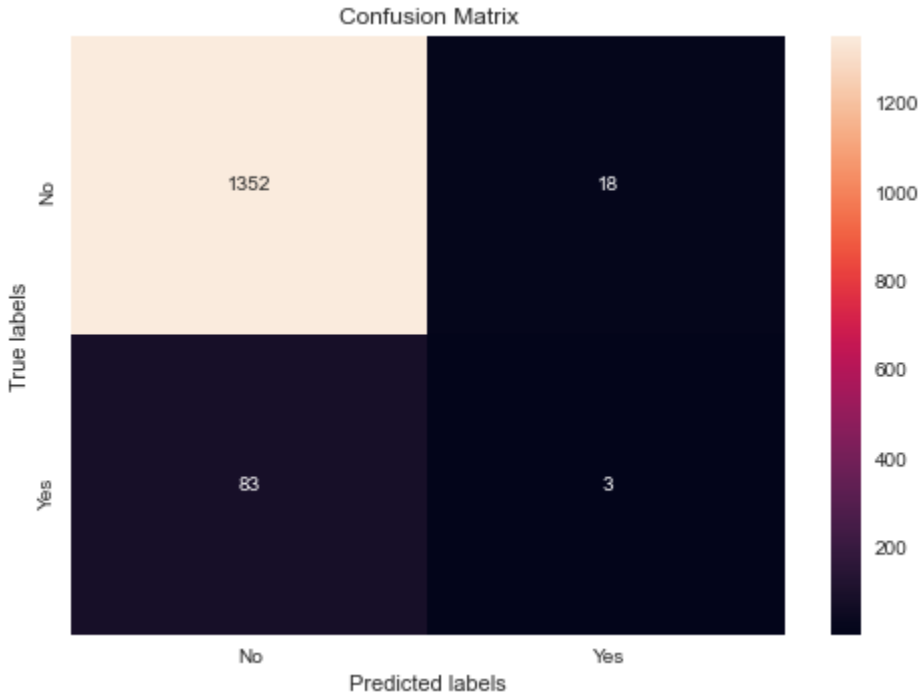
**** For better understanding of data preprocessing please see the comments in the code***

5. Results



Confusion Matrix for Logistic Regression

```
Test Accuracy from logistic regression: 0.9361263736263736
Train accuracy from logistic regression: 0.94
```



Confusion Matrix for LDA

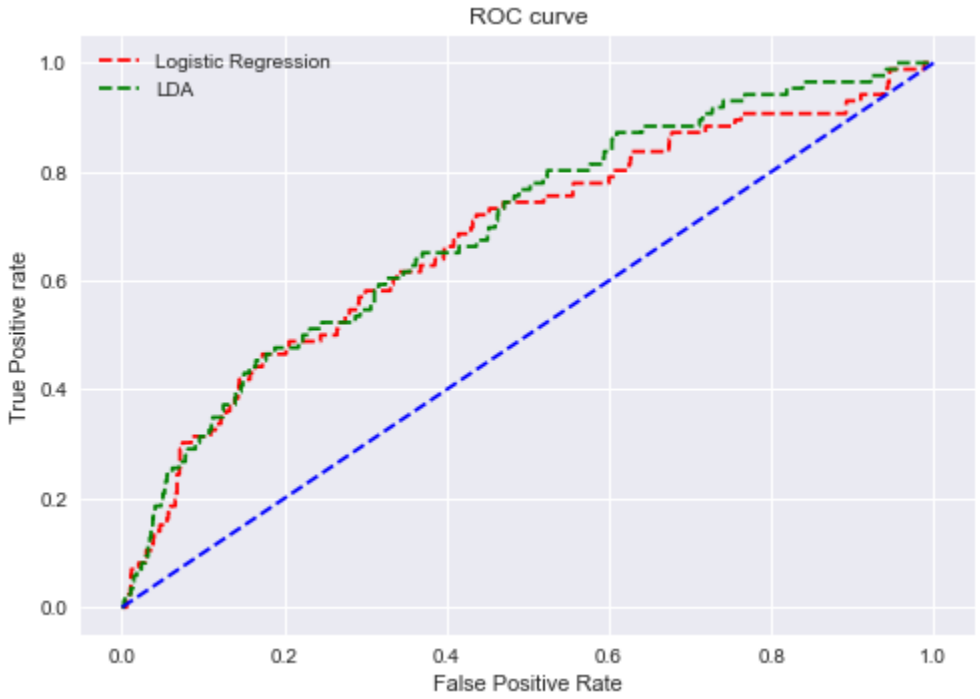
```
Test Accuracy from LDA: 0.9306318681318682
Train accuracy from LDA: 0.94
```

Since the data is highly unbalanced, it is very important to choose the correct evaluation metrics.

In this case, the accuracy will be not be a correct measure for evaluation. As the count of "No" is very high, even though the classifier will misclassify the "Yes" values, the dominating term will be count of "No" and hence accuracy will be a misleading metric. Hence it will lead to wrong conclusion.

Here finding AUC score for both the algorithms would give better insights about the performance of both the classifiers.

```
AUC Socre for Logistic Regression: 0.6773086063486674
AUC Socre for LDA: 0.6962272958750636
```



ROC Curve

From the AUC score, we can say that the LDA performs better than Logistic regression on the given dataset as AUC score for LDA is 69.62% and AUC Score for Logistic Regression is 67.73%. We know that higher the AUC score better is the classifier.

6. References

6.1. https://www.kaggle.com/uciml/caravan-insurance-challenge