BA/MIS – 749
GROUP PROJECT

Travel Insurance Prediction

HIMANI KESKAR (MIS, RED ID: 825026807)

SINDHUJA THURAKA (MIS, RED ID: 827151137)

TAMANNA GAUR (MIS, RED ID: 824677588)

ANUJA GAGANGRAS (BA, RED ID: 826040885)

KRUTTIKA BHAGWAT (BA, RED ID: 827151735)

AIM

- OUR PROJECT – " TRAVEL INSURANCE PREDICTIONS" MAINLY FOCUSES ON THE ANALYSIS AND INFERENCE OF WHETHER OR NOT A CUSTOMER IS LIKELY TO PURCHASE TRAVEL INSURANCE , BASED ON VARIOUS SOCIAL, PERSONAL AND FINANCIAL FACTORS – WHICH ACT AS OUR PREDICTORS/KEY INDICATORS IN THE MODEL.

- THE DATASET IS BASED ON THE INSURANCE PACKAGES BEING OFFERED TO THE CUSTOMERS BY A TOUR & TRAVEL COMPANY. THE DATA OF CUSTOMERS PROVIDED IS REQUIRED TO BUILD A MODEL THAT CAN INFER IF THE CUSTOMER WILL BE INTERESTED TO BUY THE TRAVEL INSURANCE PACKAGE BASED ON VARIOUS PARAMETERS THAT ARE RELATED TO THE SAME.

- CONTAINS BOTH QUALITATIVE AND QUANTITATIVE VARIABLES.SINCE THIS DATASET IS PRIMARILY USEFUL FOR DETERMINATION OF A DECISION, IT IS A INFERENCE RELATED PROBLEM.

- OUT OF ALL THE MACHINE LEARNING ALGORITHMS, AS WE KNOW THE OUTPUT OF THE RESPONSE VARIABLE, WE ARE USING SUPERVISED LEARNING, LOGISTIC REGRESSION WHICH WILL GIVE THE OUTPUT AS 0 OR 1 (NO= 0 AND YES = 1)

Our dataset contains 9 variables with 1987 rows. The data is clean and is based on various factors that count towards the decision of whether the customer will be interested to buy the travel insurance or not.

- **Age** – In numbers (Quantitative)
- **Employment Type**- (Government /Private / Self-employed)
- **Graduate Or Not**- Graduation status (Yes/No)
- **Annual Income**- The Yearly Income of The Customer in Indian Rupees (rounded to 50k INR, Quantitative)
- **Family Members**- Number of Members in Customer's Family (Quantitative)
- **Chronic Disease**- If the Customer Suffers from Any Major Disease or Conditions Like Diabetes/High BP or Asthma etc.(Yes/No)
- **Frequent Flier**-Customer’s booking history-based data, flying frequency.(Yes/No)
- **Ever Traveled Abroad** - Yes or no
- **Travel Insurance**- Whether the customer bought travel insurance or not?? (Yes/No)

RESPONSE VARIABLE

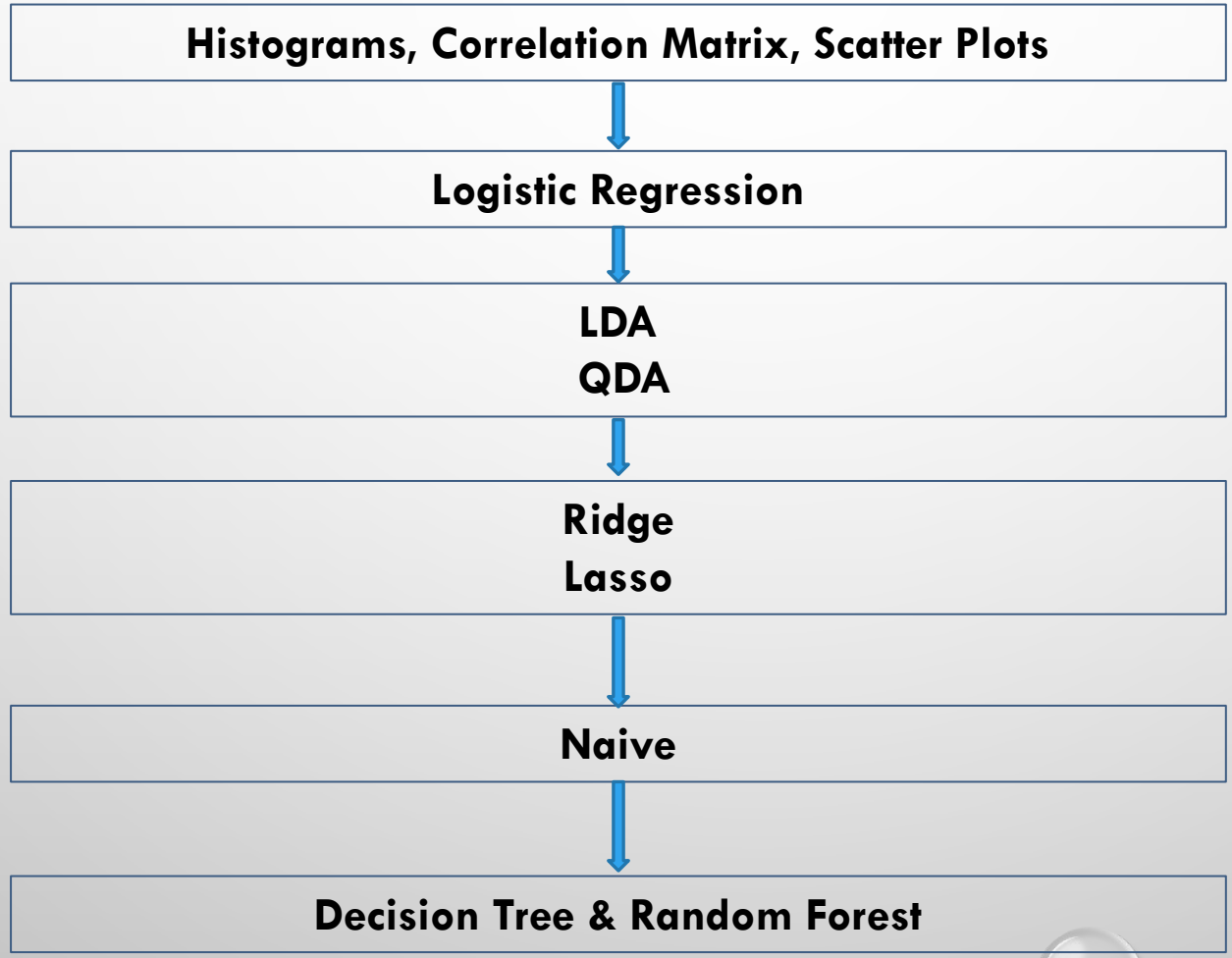
SUMMARY STATISTICS

- RESPONSE VARIABLE : TRAVEL INSURANCE (YES/NO)
- PREDICTOR VARIABLES : AGE, EMPLOYMENT TYPE, GRADUATION STATUS, ANNUAL INCOME, FAMILY COUNT, CHRONIC DISEASE , FLIER STATUS , TRAVEL ABROAD STATUS.
- STATISTICS – MEAN, MEDIAN, MIN, MAX AND STANDARD DEVIATION FOR QUANTITATIVE VARIABLES.

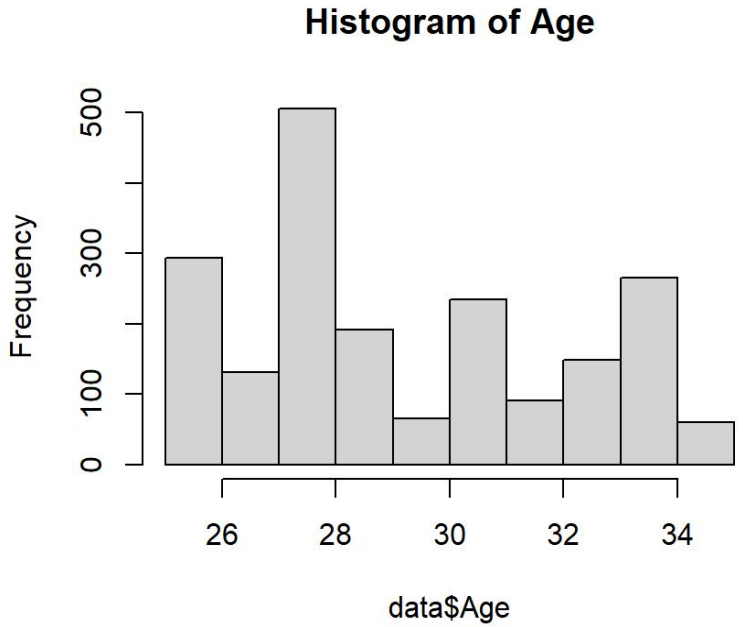
Variable	Mean	Median	Minimum	Maximum	Standard deviation
age	29.65	29	25	35	2.913308
AnnualIncome	932763	900000	300000	1800000	376855.7
FamilyMembers	4.753	5	2	9	1.60965

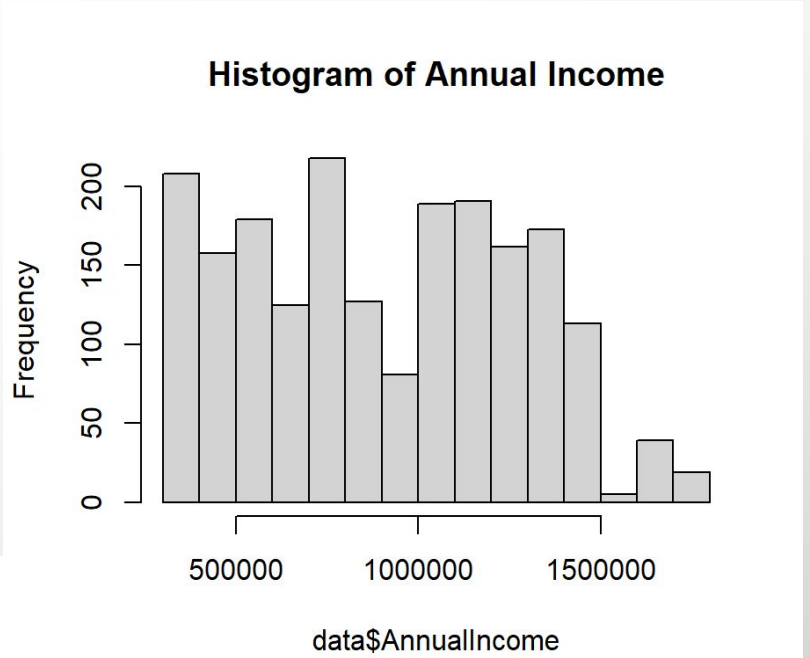
CLASSIFICATION

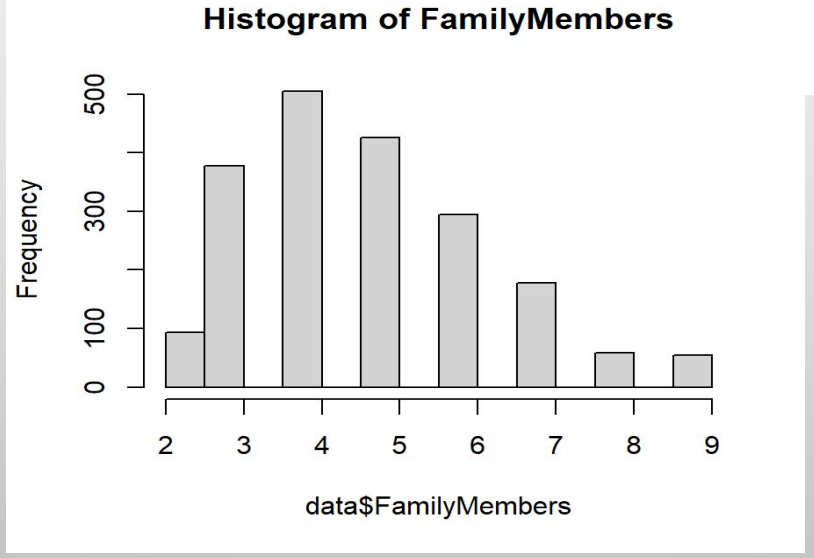
(Supervised Learning)



HISTOGRAMS


Histogram of Age
Frequency
0
100
300
500
26
28
30
32
34
data$Age


Histogram of Annual Income
Frequency
0
50
100
150
200
500000
1000000
1500000
data$AnnualIncome

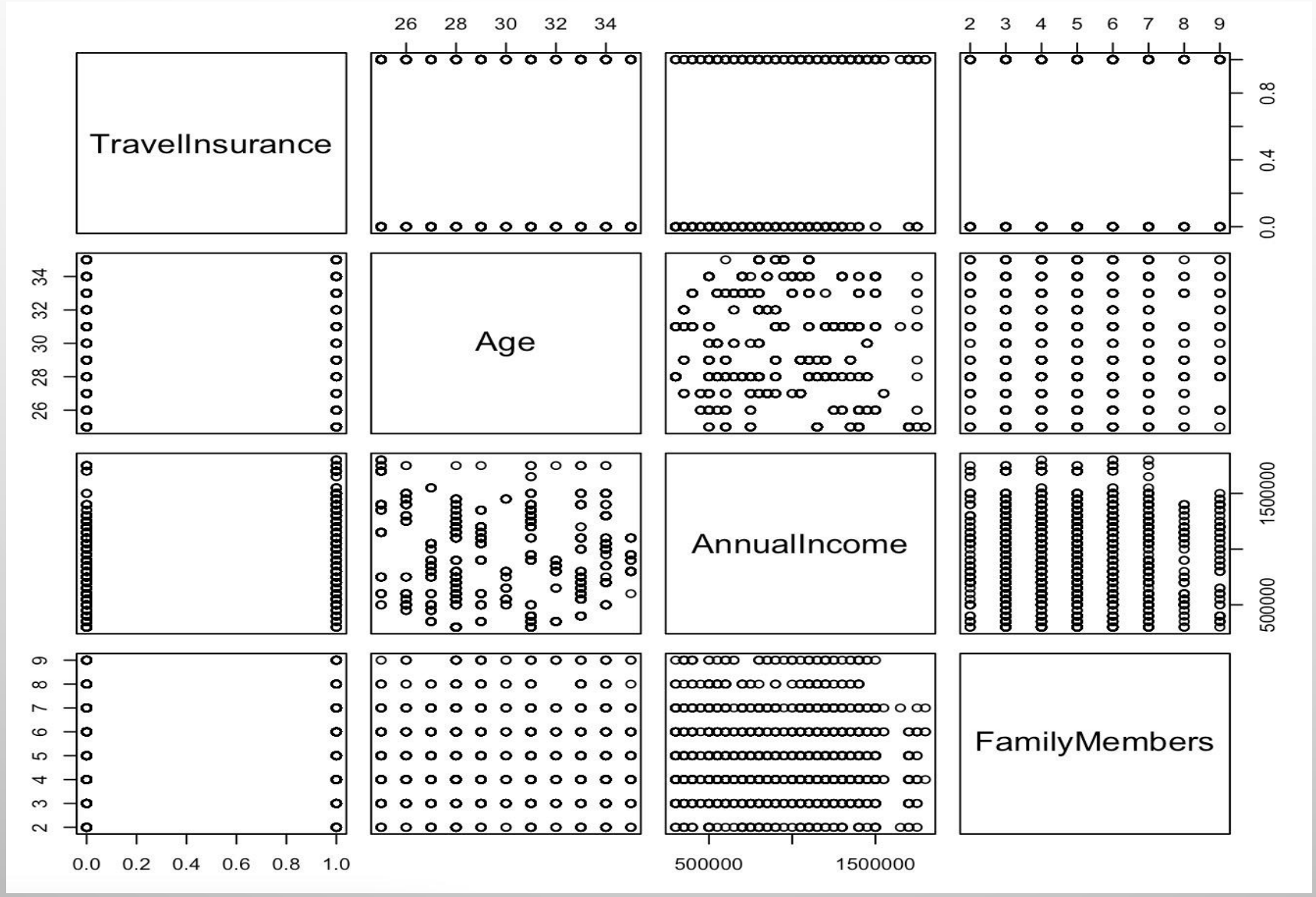

Histogram of FamilyMembers
Frequency
0
100
300
500
2
3
4
5
6
7
8
9
data$FamilyMembers

STEPS

- CURATED SCATTER PLOT AND CORRELATION MATRIX FOR RELATIONSHIP BETWEEN VARIABLES
- BUILT AND RAN VARIOUS MODELS FOR OUR DATA, MOST IMPORTANTLY LOGISTIC REGRESSION , LASSO AND RANDOM FOREST.
- CALCULATED AND COMPARED THEIR MSE, ACCURACY AND AUC CURVE TO FIND THE BEST POSSIBLE MODEL FOR THE PREDICTION OF TRAVEL INSURANCE.

```
> #Correlation Matrix
> cor(data[10], data[5:6])
                AnnualIncome FamilyMembers
TravelInsurance    0.3967632    0.07990901
>
```

RELATIONSHIP BETWEEN VARIABLES



LOGISTIC REGRESSION

- LOGISTIC REGRESSION IS ONE OF THE MOST IMPORTANT MODELS FOR OUR TRAVEL INSURANCE PREDICTION DATASET, AS IT GIVES RESULTS IN THE FORM OF “0” AND “1” OR “YES” AND “NO” , WHICH WILL HELP THE COMPANY DETERMINE WHETHER THE INSURANCE WILL BE PURCHASED OR NOT.

- WE HAVE DIVIDED THE DATA INTO 70% TRAINING SET TO TRAIN THE MODEL AND INTO 30% TESTING DATA TO GET THE ACCURACY OF THE MODEL.

- WE HAVE CONSIDERED CROSS VALIDATION APPROACH TO DIVIDE THE DATA AS IT DIVIDES THE DATA RANDOMLY TO GET BETTER ACCURACY AND TO MINIMISE THE TEST ERROR INSTEAD OF STANDARD DIVISION APPROACH.

MODEL COMPARISON

Logistic regression is very popular for classification, especially when K = 2.

In case of binary outcome (0 and 1), results from Linear Discriminant Analysis are same
As logistic regression. So LDA is more appropriate for multiclass classification.

Naive Bayes is useful when p (no of predictors) is very large.

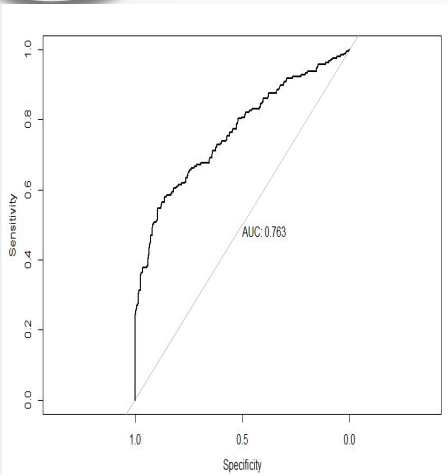
Accuracy (overall fraction of correction prediction) of different models are calculated by confusion matrix.

Test data is used to get the accuracy.

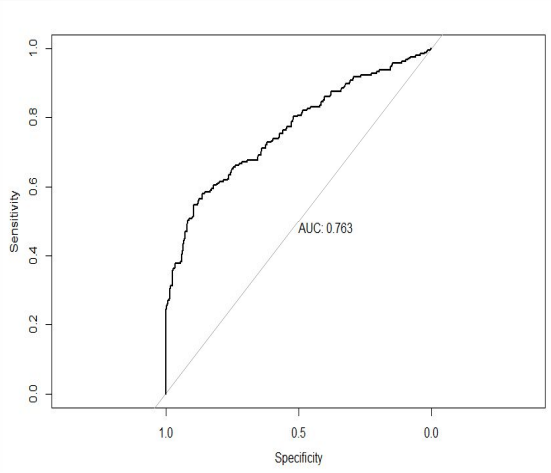
Confusion matrix also gives the False positive (Type 1 error) and False negative (Type 2 error) error rate.

ROC curve gives the area under the curve. Higher AUC gives better performance of the model.

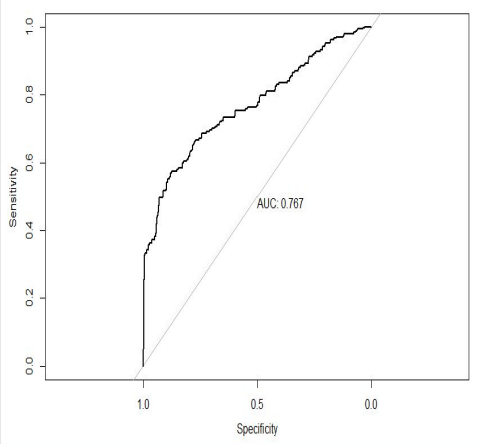
Model Comparison based on R output



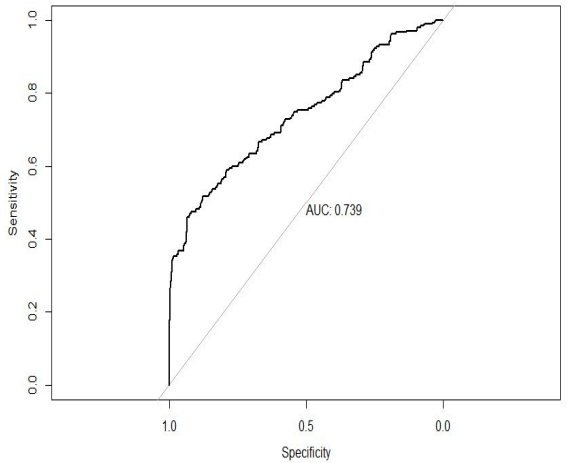
AUC – Logistic Regression



AUC – LDA



AUC – QDA



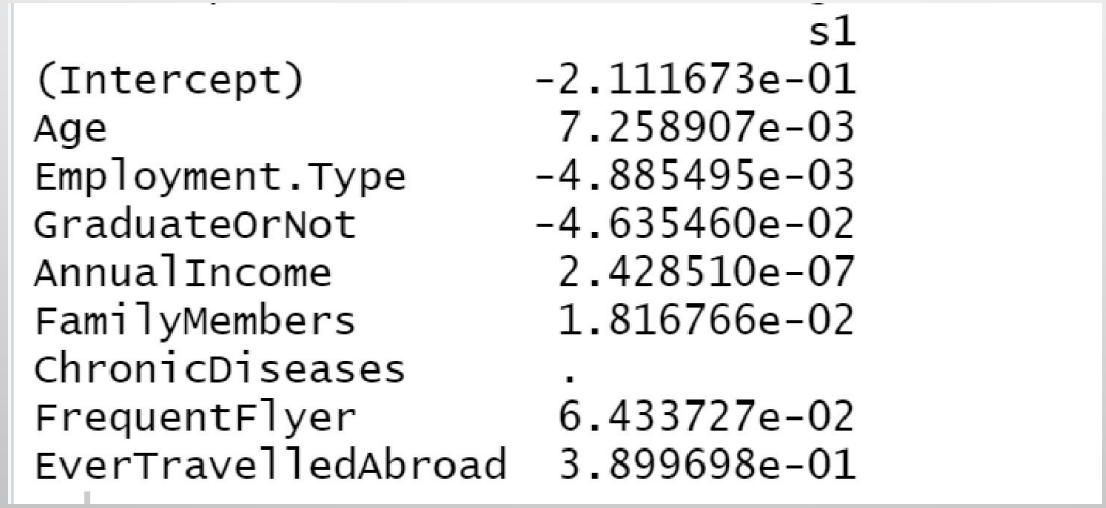
AUC – Naïve Bayes

Model Name	ACCURACY	AUC (Area Under Curve)
Logistic regression	75.82%	0.7631
LDA	75.82%	0.7631
QDA	77.8%	0.7672
Naïve Bayes	75.82%	0.7392

By comparing Accuracy (overall fraction of correction predictions) and AUC (Area under the curve) , QDA works better.

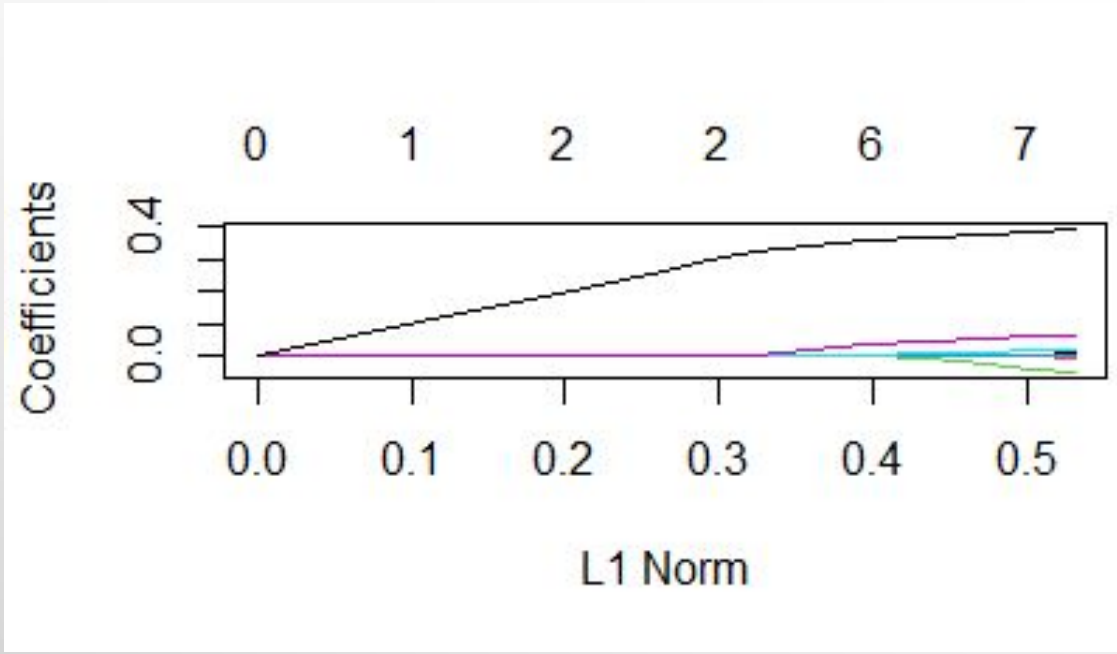
LASSO

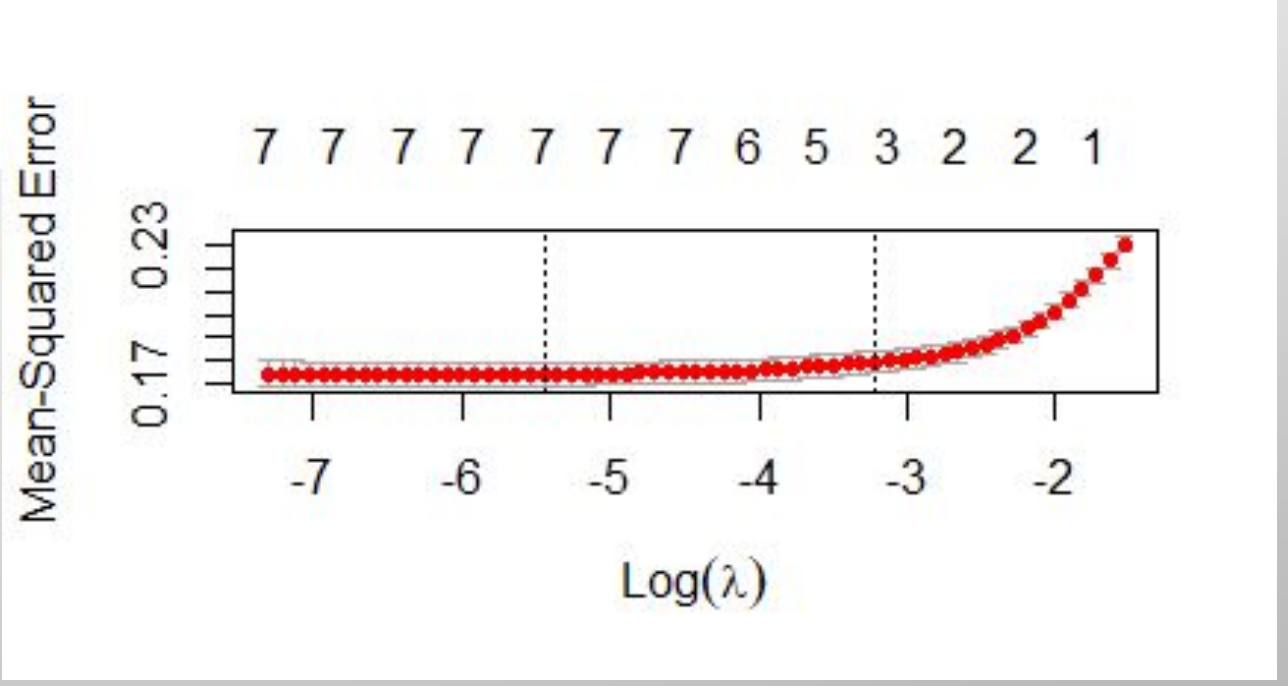
- IMPORTANT TO ANALYSE DIFFERENT MODELS FOR OUR DATA SO THAT WE CAN FIND THE BEST AND MOST OPTIMAL MODEL WHICH CAN ACCURATELY PREDICT WHETHER THE CUSTOMER OF THE TRAVEL COMPANY WOULD PURCHASE THEIR INSURANCE OR NOT.
- LASSO IS ONE OF THE POPULAR MODELS WHICH ALSO FITS CLOSELY WITH OUR DATASET.
- DATA MATRIX FOR LASSO :



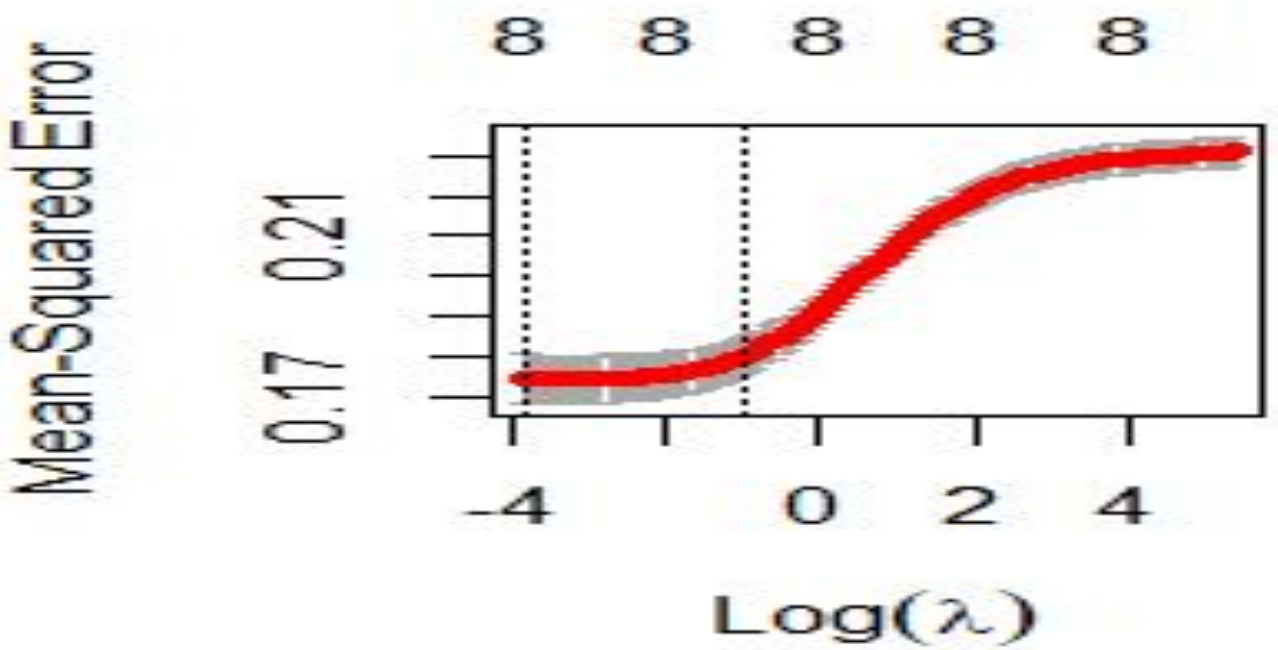
	s1
(Intercept)	-2.111673e-01
Age	7.258907e-03
Employment.Type	-4.885495e-03
GraduateOrNot	-4.635460e-02
AnnualIncome	2.428510e-07
FamilyMembers	1.816766e-02
ChronicDiseases	.
FrequentFlyer	6.433727e-02
EverTravelledAbroad	3.899698e-01

LASSO MODEL


0 1 2 2 6 7
Coefficients
0.0 0.4
0.0 0.1 0.2 0.3 0.4 0.5
L1 Norm


7 7 7 7 7 7 7 6 5 3 2 2 1
Mean-Squared Error
0.17 0.23
-7 -6 -5 -4 -3 -2
Log(λ)

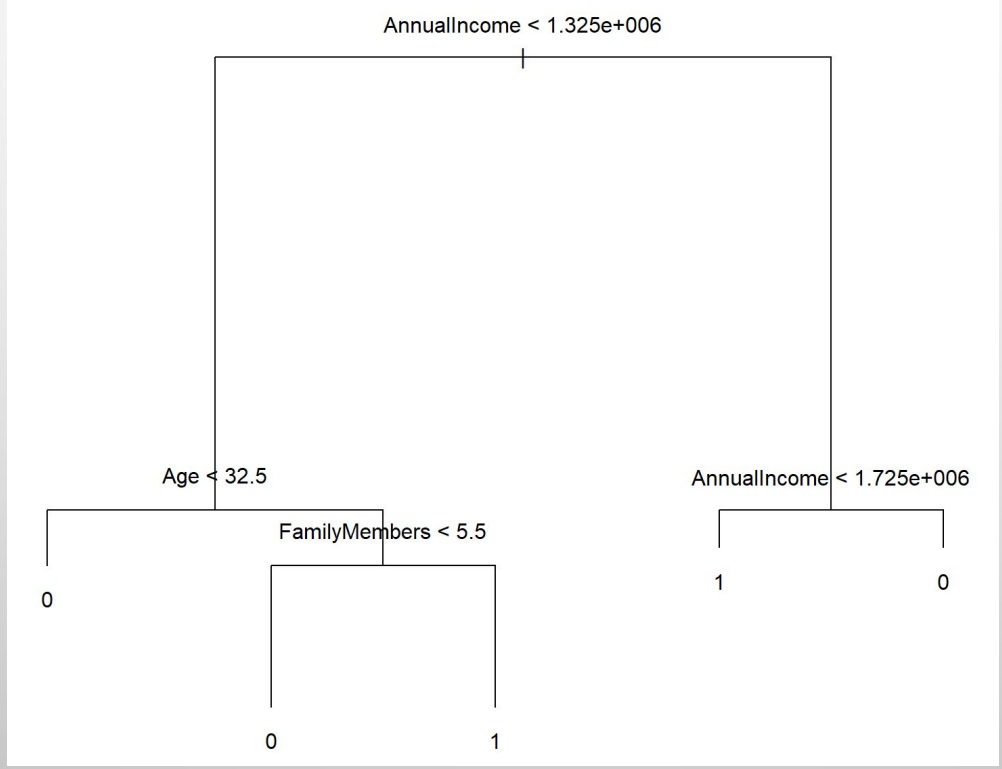
Ridge Regression




```
predict(out, type = "coefficients", s = bestlam)[1:9, ]
      (Intercept)                       Age     Employment.Type      GraduateOrNot
    -4.167198e-01              1.121801e-02       -1.321650e-02       -3.031305e-02
    AnnualIncome              FamilyMembers     ChronicDiseases      FrequentFlyer
     2.783141e-07              2.548614e-02        1.329201e-02        8.493530e-02
verTravelledAbroad
     3.596385e-01
```

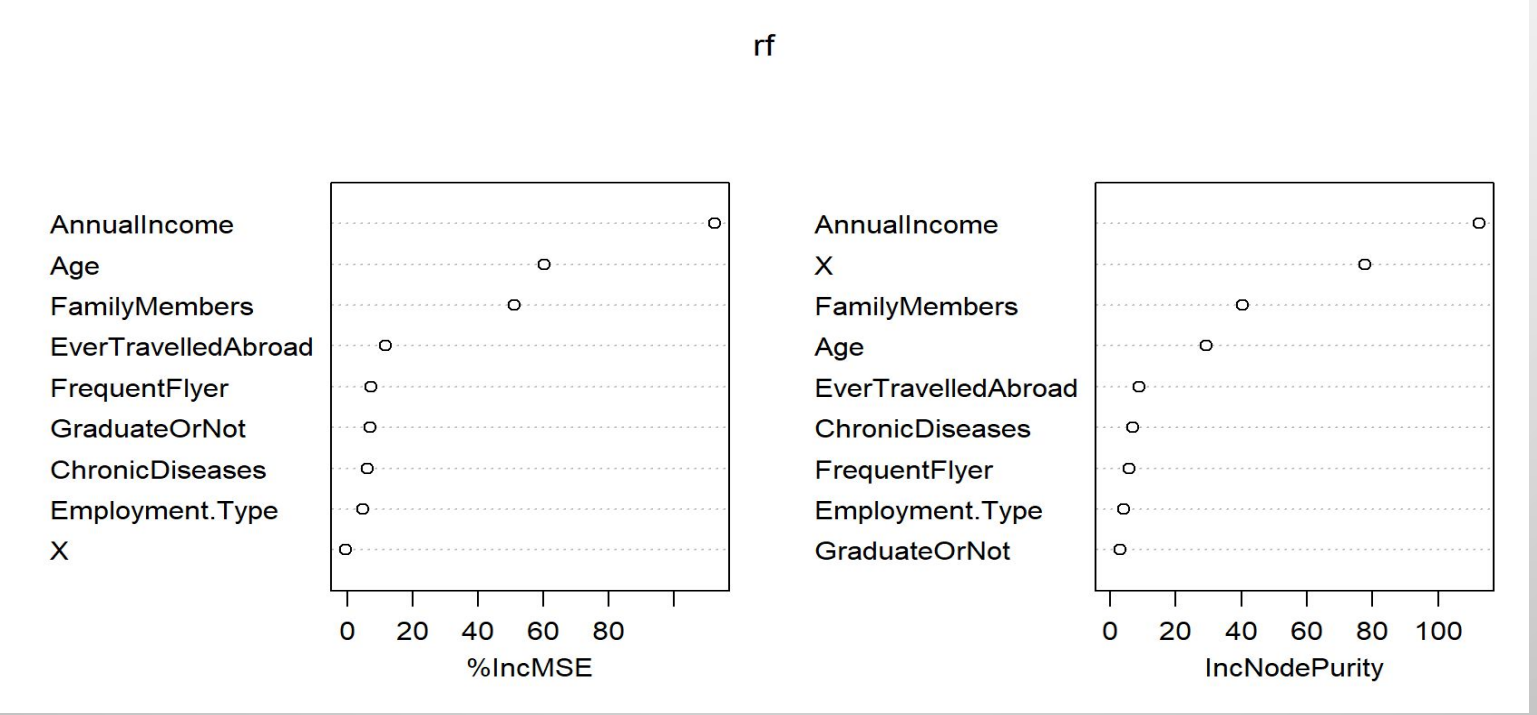
Decision Tree

- Categorial Variable decision tree
- Easy to prepare, easy to understand and less data cleaning required



RANDOM FOREST

- Random Forest model has higher accuracy over the decision tree.
- It handles both classification and regression tasks, and fits large datasets well.
- The data matrix for our dataset :



Conclusion

- Based on the results, among the Logistic Regression, LDA, QDA and Naïve Bayes:
  QDA best model (Higher AUC – 0.7672 , Accuracy - 77.8%)
- Decision Tree and Random Forest :
  Age, Annual Income, Family Members to the significant
- Ridge Regression – MSE 1.797014
  Lasso Regression – MSE 0.1726044
  Lasso is a better model as the MSE value is lesser.

Questions/Answers