CAR INSURANCE CLAIM PREDICTION

GROUP - 12

Jahnani Nagarajan Sivakumar

Kirthika Kulandaivel Senthilkumar

Krishna Apurva

Manusha Medarametla

PROJECT OVERVIEW

Introduction → Dataset Description → Exploratory Data Analysis → Decision Tree Analysis → Random Forest Analysis → Logistic Regression Analysis → Neural Network Analysis → Model Evaluation → Conclusion

INTRODUCTION

- Aims to predict if policyholders will file a claim in the next six months by analyzing a comprehensive dataset.
- Helps insurance companies refine their risk assessment and pricing strategies.
- Revolutionize managing risk in the car insurance industry using Advanced analytics and Machine learning.

Goal:

- The project aims to develop an accurate predictive model using policyholder attributes for data-informed decision-making.

DATASET DESCRIPTION

Data Source: Kaggle

97656 instances and 44 attributes.

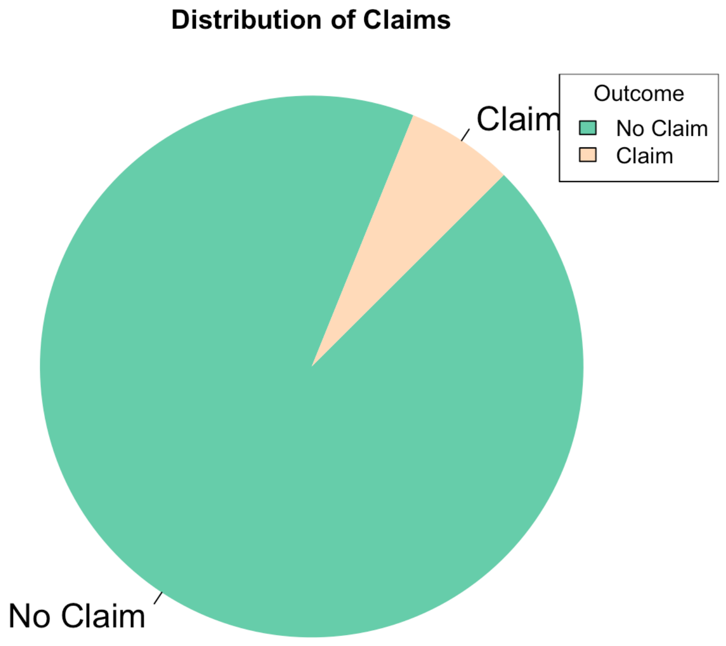
Describes policyholder's details like policy tenure, age of the car, age of the car owner, the population density of the city, make and model of the vehicle, power, engine type, etc.,

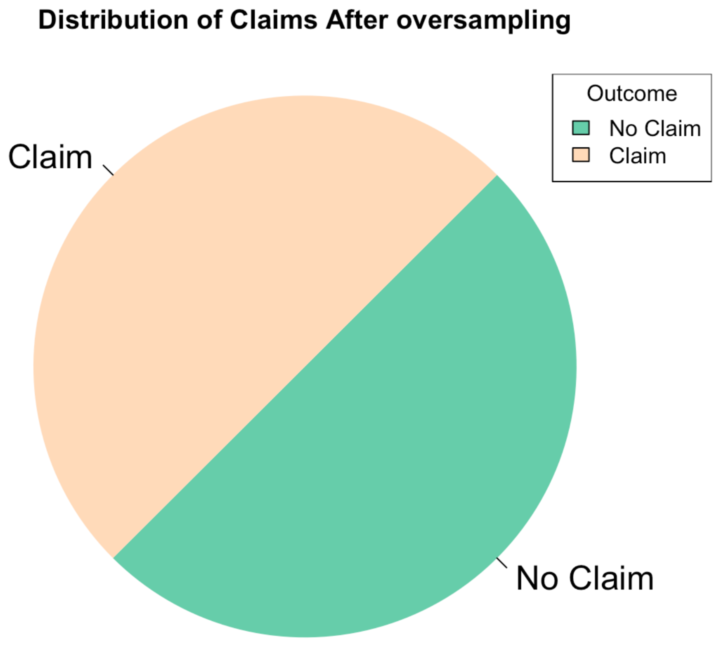
Target variable indicating whether the policyholder files a claim in the next six months.

Split for the training and testing with a ratio of 60:40

EXPLORATORY DATA ANALYSIS

- An imbalance in the target data distribution occurs
- Oversampled to balance data





PRE-PROCESSING PROCEDURES

Identification of null values present, if any

Dropping attributes that are not required for Classification Analysis

Mutated '0' and '1' instead of 'No' and 'Yes' in the dataset

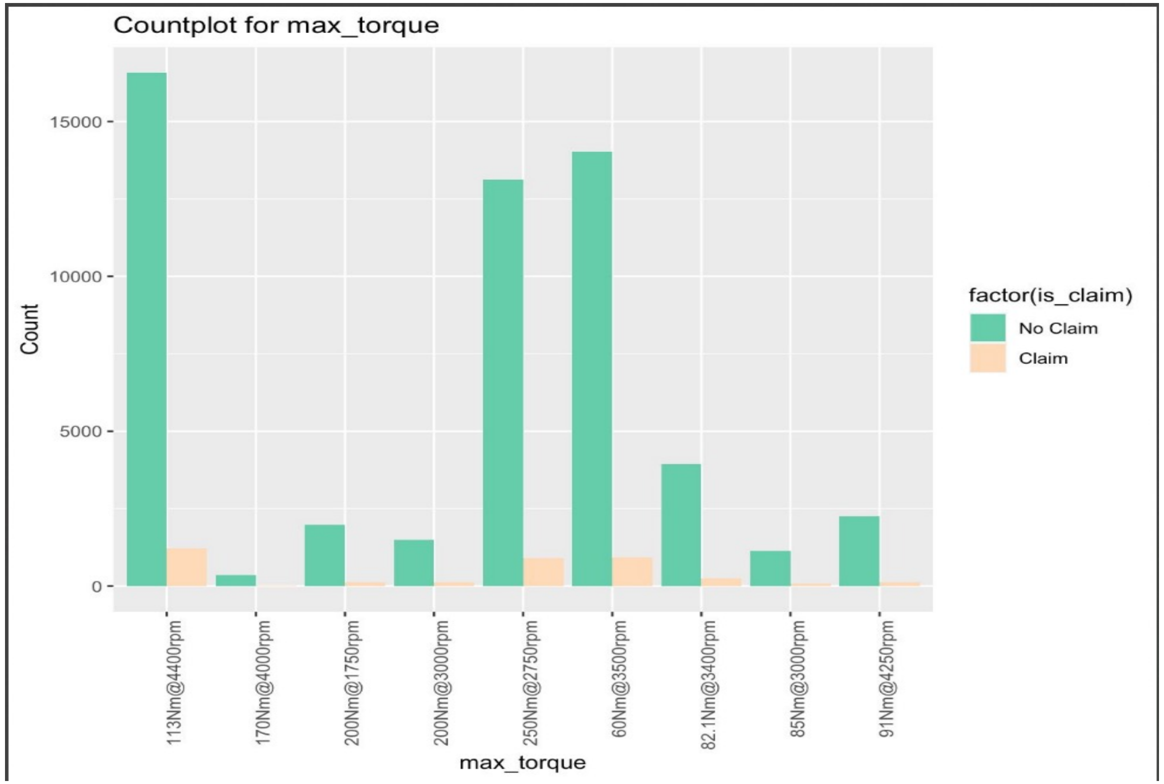
Grouped categorical and numerical attributes

Extracted numerical values from Max torque and Max power to estimate ratio with RPM values

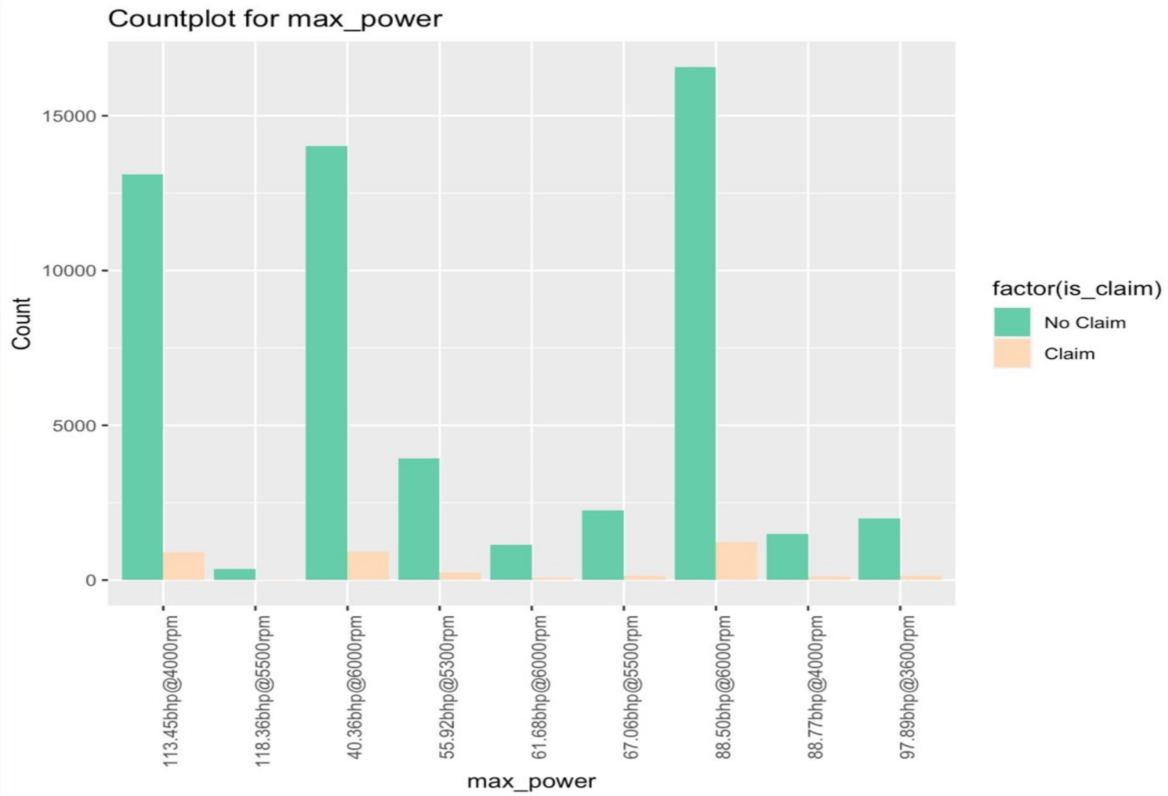
Added new columns to the dataset by splitting the categorical data

DATA DISTRIBUTION ANALYSIS

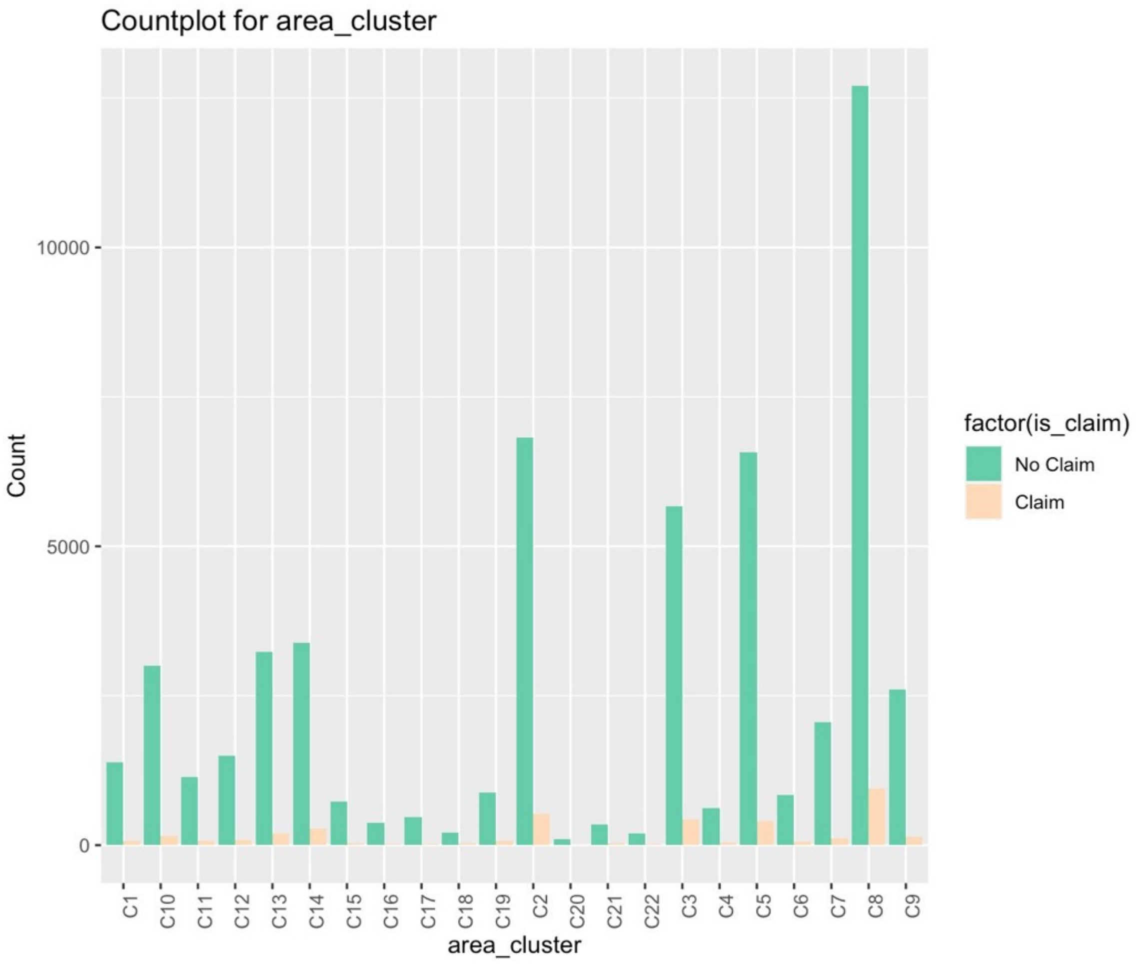
Max-torque

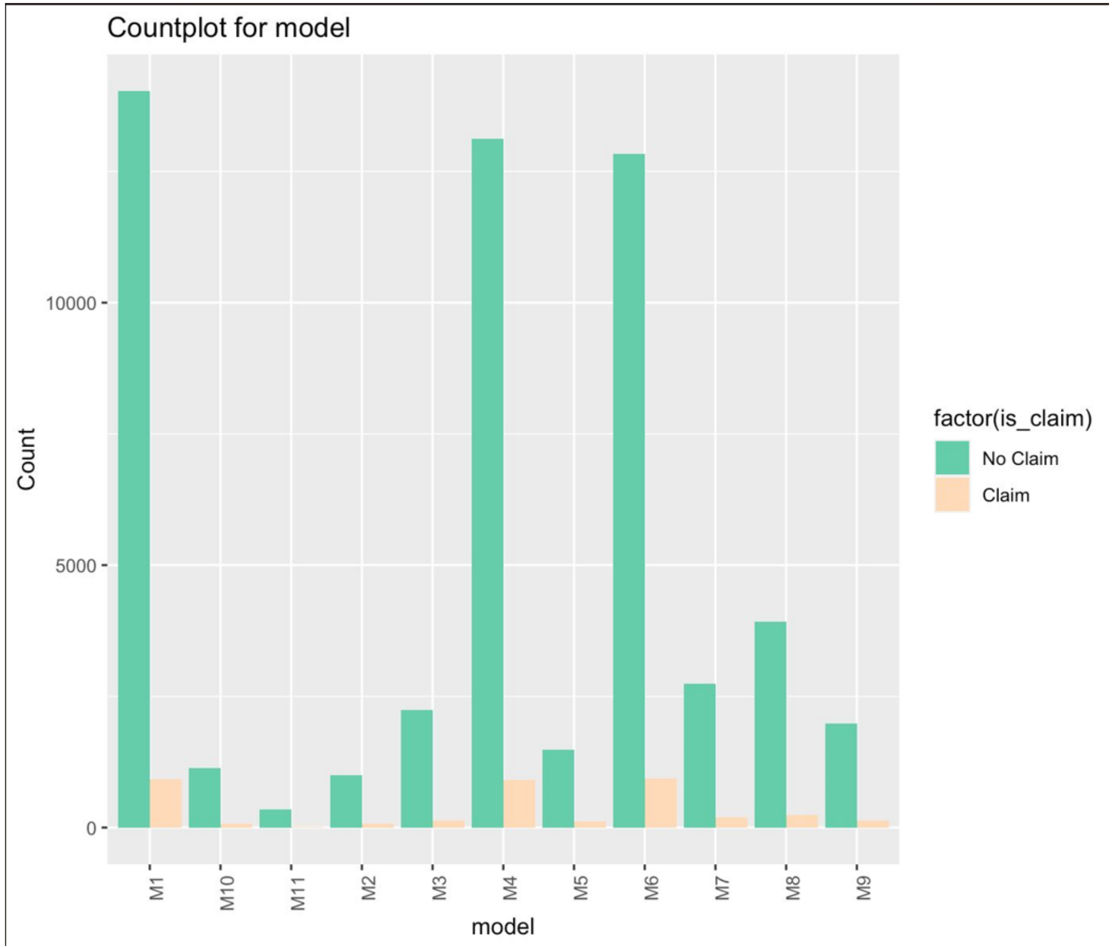


Max-power

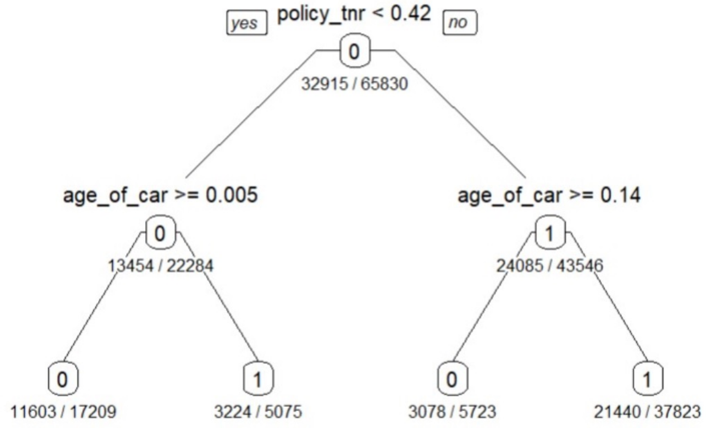


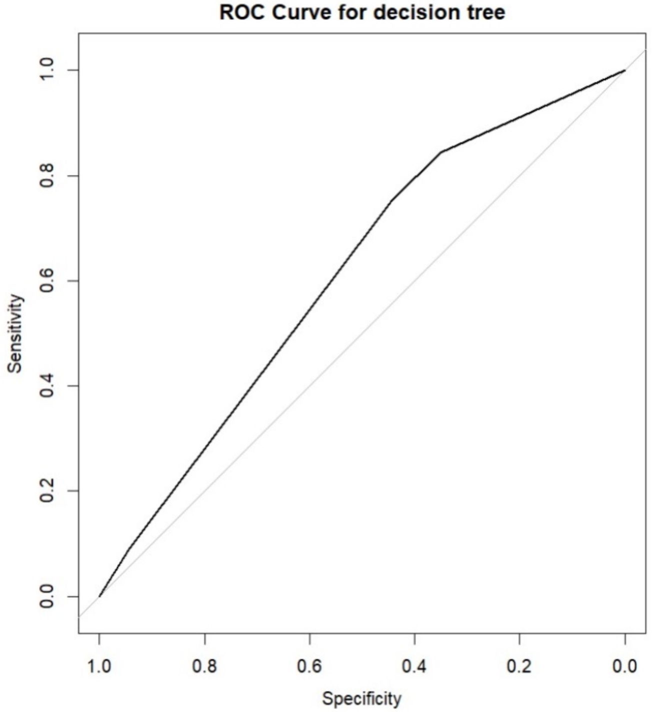
AREA AND MODEL VS CLAIM





DECISION TREE ANALYSIS






```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0  9730   371
         1 12199  1137

               Accuracy : 0.4637
                 95% CI : (0.4573, 0.4701)
    No Information Rate : 0.9357
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0425

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.44370
            Specificity : 0.75398
         Pos Pred Value : 0.96327
         Neg Pred Value : 0.08526
             Prevalence : 0.93566
         Detection Rate : 0.41516
   Detection Prevalence : 0.43099
      Balanced Accuracy : 0.59884

       'Positive' Class : 0
```

RANDOM FOREST ANALYSIS

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 13512   657
         1  8417   851

               Accuracy : 0.6128
                 95% CI : (0.6066, 0.6191)
    No Information Rate : 0.9357
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0531

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.61617
            Specificity : 0.56432
         Pos Pred Value : 0.95363
         Neg Pred Value : 0.09182
             Prevalence : 0.93566
         Detection Rate : 0.57652
   Detection Prevalence : 0.60456
      Balanced Accuracy : 0.59025

       'Positive' Class : 0
```

10 TREES CONFUSION MATRIX

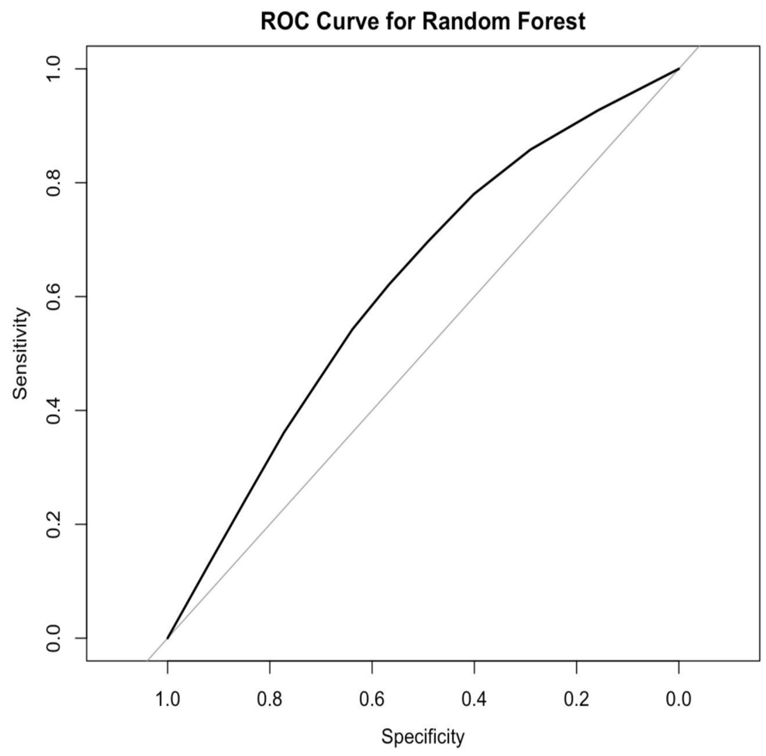
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 13610   646
         1  8319   862

               Accuracy : 0.6175
                 95% CI : (0.6112, 0.6237)
    No Information Rate : 0.9357
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0571

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.62064
            Specificity : 0.57162
         Pos Pred Value : 0.95469
         Neg Pred Value : 0.09389
             Prevalence : 0.93566
         Detection Rate : 0.58071
   Detection Prevalence : 0.60827
      Balanced Accuracy : 0.59613

       'Positive' Class : 0

>
```

100 TREES CONFUSION MATRIX

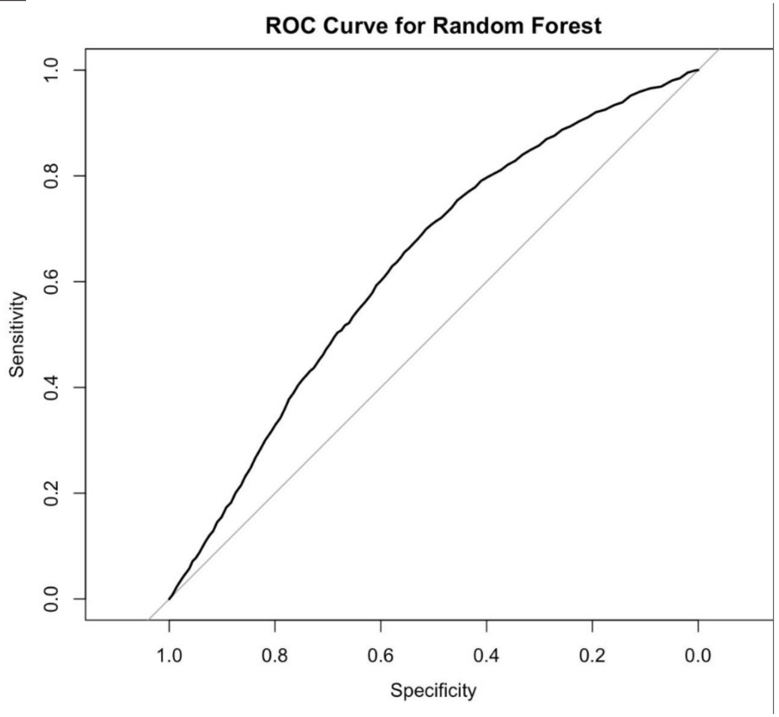
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 13567   635
         1  8362   873

               Accuracy : 0.6161
                 95% CI : (0.6099, 0.6224)
    No Information Rate : 0.9357
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0584

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.61868
            Specificity : 0.57891
         Pos Pred Value : 0.95529
         Neg Pred Value : 0.09453
             Prevalence : 0.93566
         Detection Rate : 0.57887
   Detection Prevalence : 0.60596
      Balanced Accuracy : 0.59880

       'Positive' Class : 0
```

500 TREES CONFUSION MATRIX

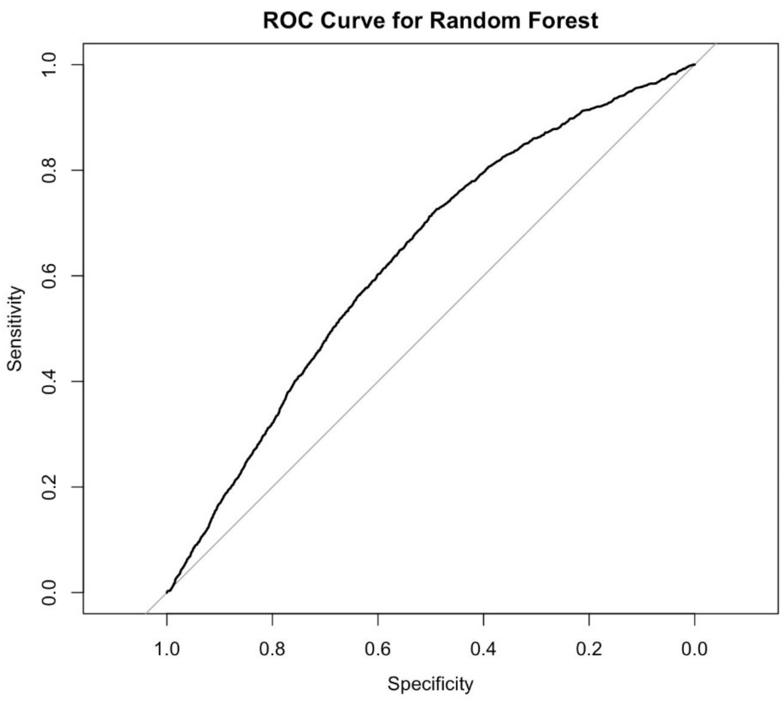
RANDOM FOREST ANALYSIS (CONT..)



10 TREES ROC CURVE



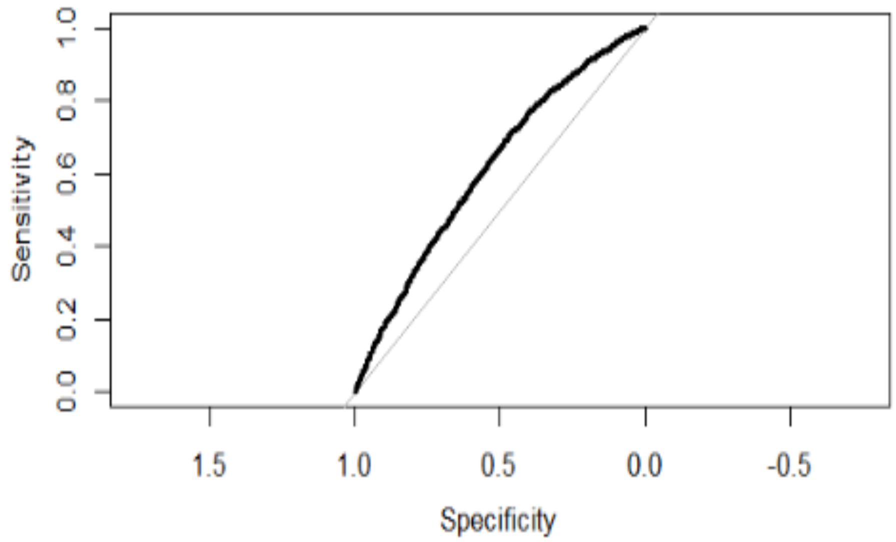
100 TREES ROC CURVE



500 TREES ROC CURVE

LOGISTIC REGRESSION ANALYSIS

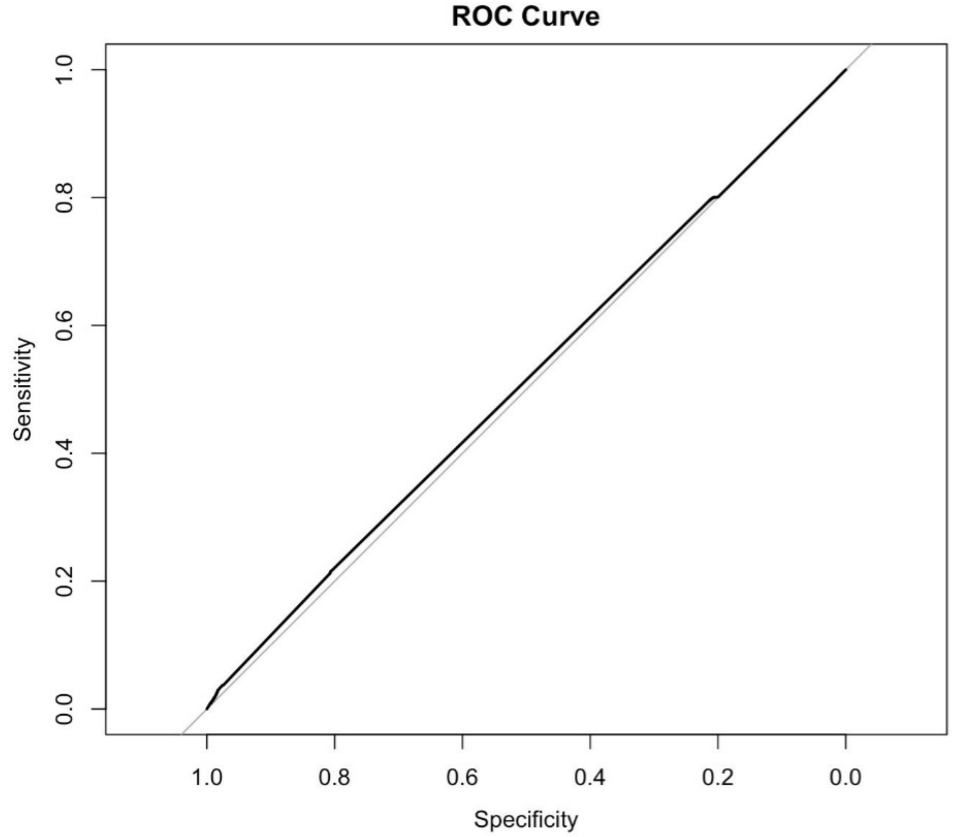
	actual	predicted
3	0	0.5804978
5	0	0.3813829
6	0	0.5156848
9	0	0.3699288
11	0	0.5901706



Area under the curve: 0.6128

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 18930  1176
         1  2999   332

               Accuracy : 0.8219
                 95% CI : (0.8169, 0.8267)
    No Information Rate : 0.9357
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0534

 Mcnemar's Test P-Value : <0.0000000000000002

            Sensitivity : 0.86324
            Specificity : 0.22016
         Pos Pred Value : 0.94151
         Neg Pred Value : 0.09967
             Prevalence : 0.93566
         Detection Rate : 0.80770
   Detection Prevalence : 0.85787
      Balanced Accuracy : 0.54170

       'Positive' Class : 0
```

NEURAL NETWORK ANALYSIS




```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 17652  1182
         1  4277   326

               Accuracy : 0.7671
                 95% CI : (0.7616, 0.7725)
    No Information Rate : 0.9357
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0108

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.80496
            Specificity : 0.21618
         Pos Pred Value : 0.93724
         Neg Pred Value : 0.07082
             Prevalence : 0.93566
         Detection Rate : 0.75317
   Detection Prevalence : 0.80360
      Balanced Accuracy : 0.51057

       'Positive' Class : 0
```

OVERVIEW

	Accuracy	Area Under Curve (ROC)
Decision Tree	46.37%	0.6109
Random Forest	61.62%	0.6316
Logistic Regression	82.19%	0.6128
Neural Network	76.71%	0.5118

SUMMARY

- Logistic regression has the highest accuracy among the models.
- Low AUC may suggest that its ability to discriminate between classes is not as strong as Random Forest
- Random Forest accuracy is lower than logistic regression
- Higher AUC indicates better discrimination between positive and negative cases.
- Random forests are less interpretable than logistic regression.
- Decision trees are prone to overfitting and might have struggled with generalization to new data

CONCLUSION

Based on the business objective, interpretability is crucial to determine what factors influence the policyholder to claim. Therefore, logistic regression clearly explains the attributes that can make a policyholder claim a policy. For example, policy tenure, age of the car, and area have a higher influence on claim prediction than any other attribute.

The insurance company can use this prediction model to charge the policyholder with a premium or higher-cost policy.

Acknowledgement:
Prof. Zhang's R scripts and Notes

THANK YOU