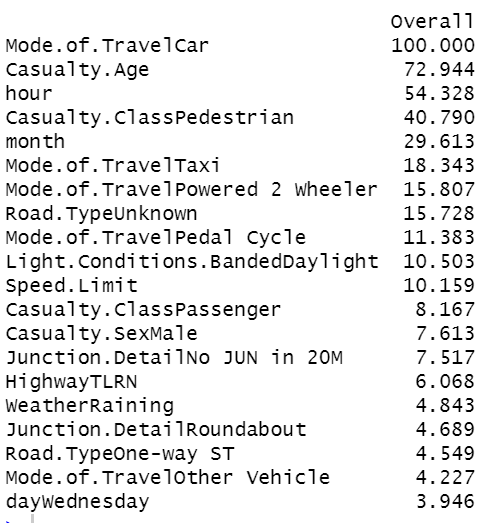
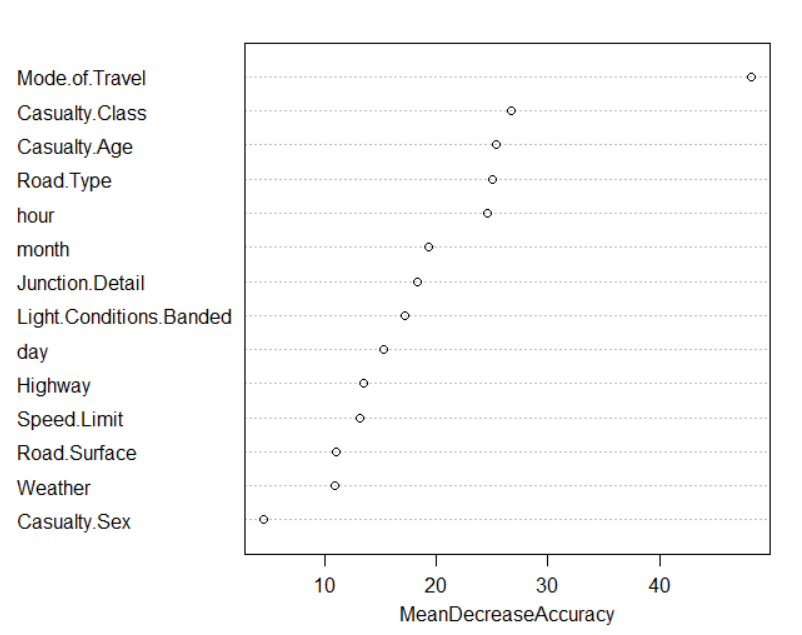
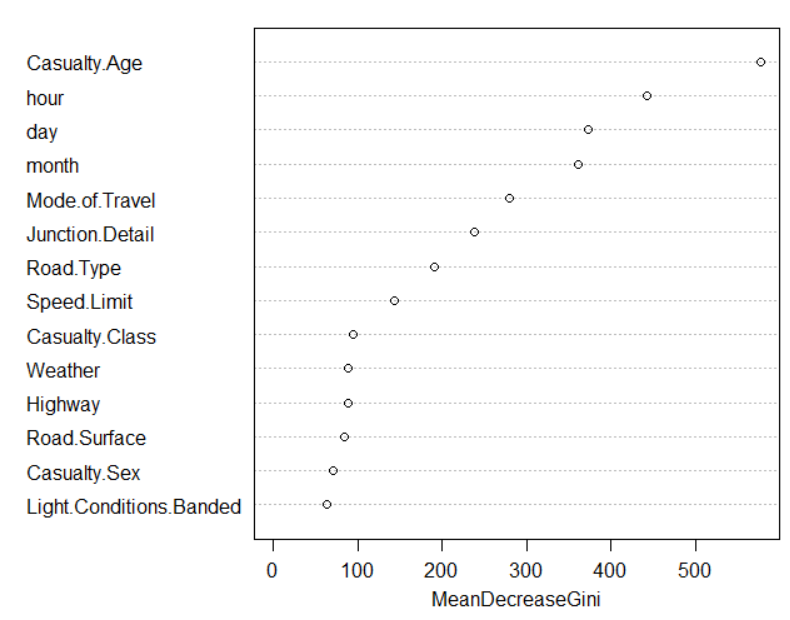
|  |  |  |
| --- | --- | --- |
| Original Dataset (n=10,032) | | |
| Class | Serious | Slight |
| Number of samples | 4,032 | 6,000 |
| Percentage | 40.21% | 59.79% |

|  |  |  |
| --- | --- | --- |
| 10% of Original Dataset (n=1,003) | | |
| Class | Serious | Slight |
| Number of samples | 403 | 600 |
| Percentage | 40.18% | 59.82% |

|  |  |  |
| --- | --- | --- |
| 90% of Original Dataset (n=9,029) | | |
| Class | Serious | Slight |
| Number of samples | 3,620 | 5,400 |
| Percentage | 40.19% | 59.81% |

****

****

****

**1. Why you chose the data**

-Road accidents constitute a significant proportion of the number of serious injuries reported every year. Yet, it is often challenging to determine which specific conditions lead to such events, making it more difficult for local law enforcement to address the number and severity of road accidents. We all know that some characteristics of vehicles and the surroundings play a key role (engine capacity, condition of the road, etc.). However, many questions are still open. Which of these factors are the leading ones? How much are the external factors to blame, compared to the driver skills?

-we analysed situational information (such as road type, weather conditions, etc.) to estimate the severity of an accident. Such insights would help governments to better understand the sources of accidents and act to reduce them.

**2. What the data are about**

-Road Safety Data about the circumstances of personal injury road accidents in United Kingdom from 1979, the types of vehicles involved and the consequential casualties. The statistics relate only to personal injury accidents on public roads that are reported to the police.

-I use subset of 2019 dataset.

**3. An outline of the analysis process**

**4. Description of all the relevant steps**

**5. Graphics/tables or whatever you need to explain what decisions you made**

3.1 Data preprocessing step

-Full dataset is highly unbalanced which have 3 classes: Fatal,Serious,Slight (of 0.5%/14%/84.5%) like 500 out of 40K so I drop that class and turn the problem into binary classification

-merge 3 files of attendant, vehicle, and environment details based on ID

-preprocess dataset by EDA such as seeing matrix of missing value

-drop features that has more than 90% percent of same values

-drop one of pair features that have correlation more than 95% percent

-drop column with text that need nlp technique to extract information

-process datetime column to hour, day, and month

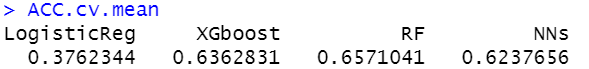
-collapse some values of feature to smaller class so we have smaller dummy encoding ex. 7 class to 3 class + 1 as other

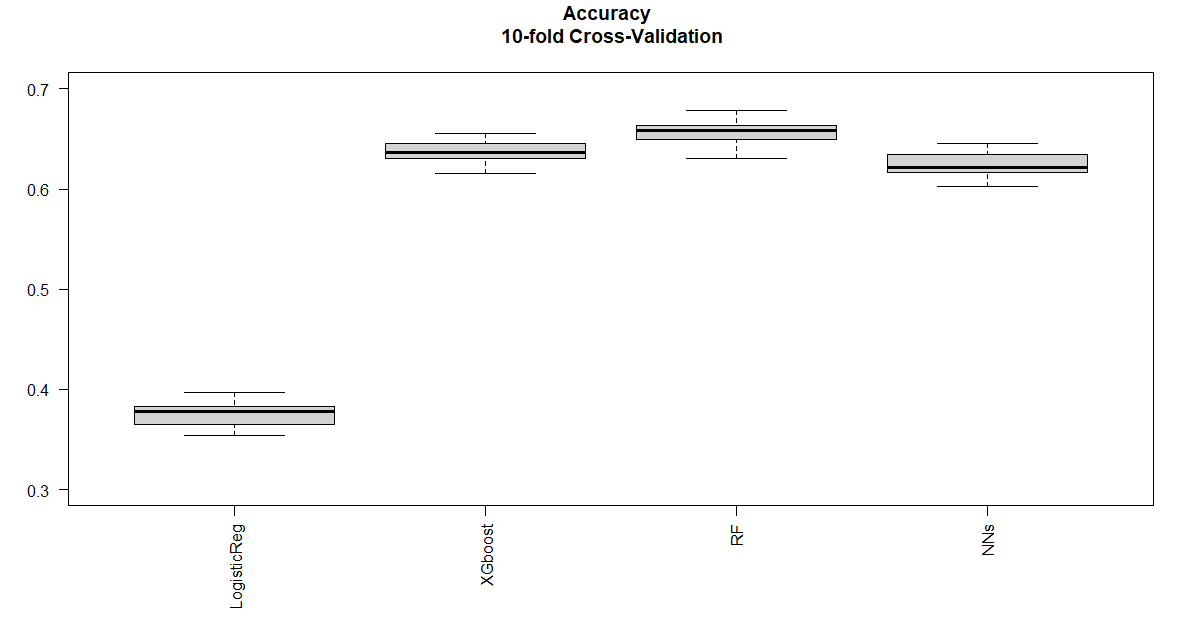
3.2 Data splitting

-split 90/10 into training and test dataset

3.3 Explore models

-using 10-fold cross validation

****

****

- the metric I use to measure these models is accuracy which is 1-misclassification rate

-surprising result is that logistic regression has far lower accuracy and RF beats XGboost.

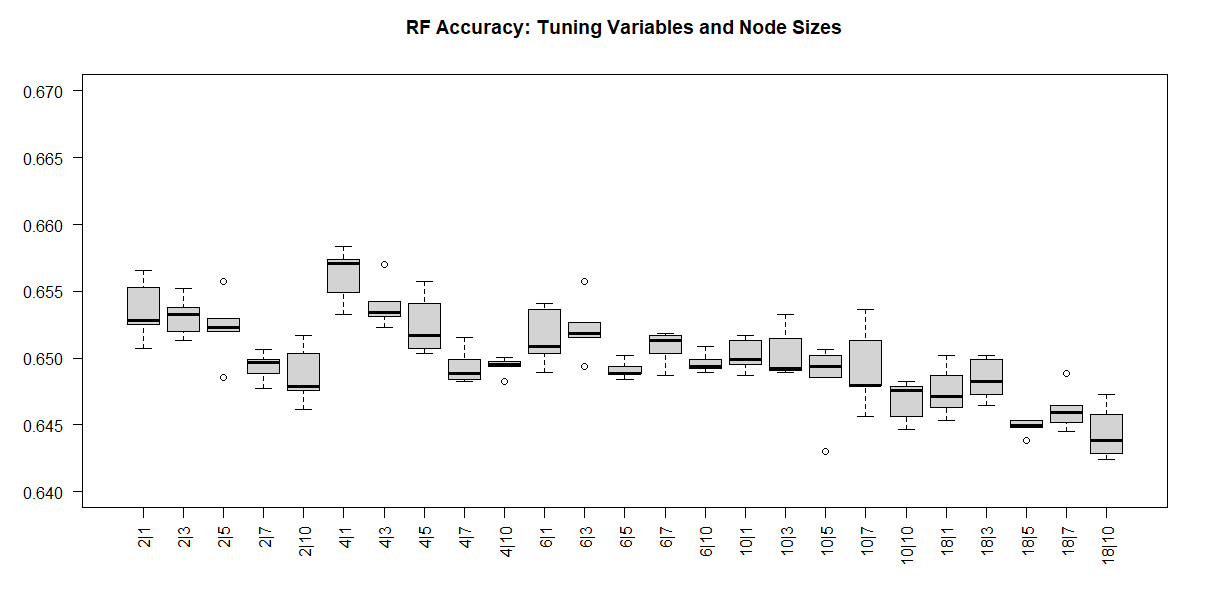
3.3 Tuning RF models

-rerun the code: split that training data into 80/20 training, validation dataset

-using that training data (80) to try 2 methods:

1. using repeatedcv, 10-fold cv with 5 repeats

2. using OOB error Repeated 5 times with 2 hyperparameter (mtry and nodesize)

****

-get the best model of each method

-then using that model to make prediction on validation dataset

-get very similar result but the first method takes longer time to train

**6. How you did your final predictions**

**7. A measure of uncertainty on the predictions, and how it compares to**

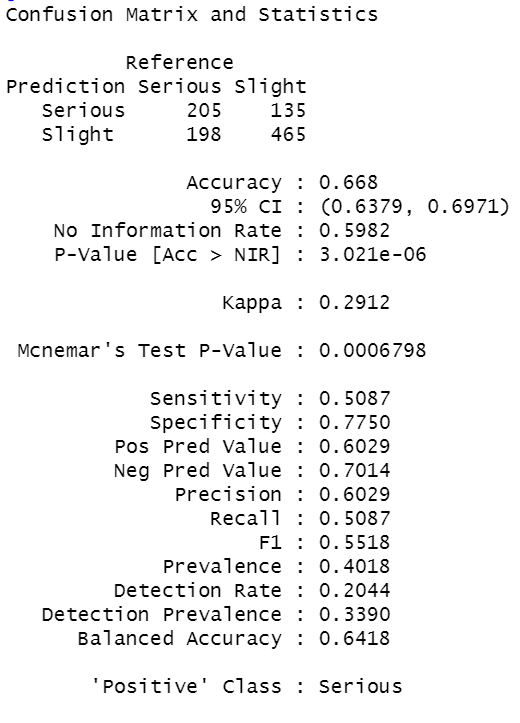
**(a) Confusion matrix and misclassification rate**

-use the sets of parameters from that training dataset to train full model

-make prediction on holdout test dataset

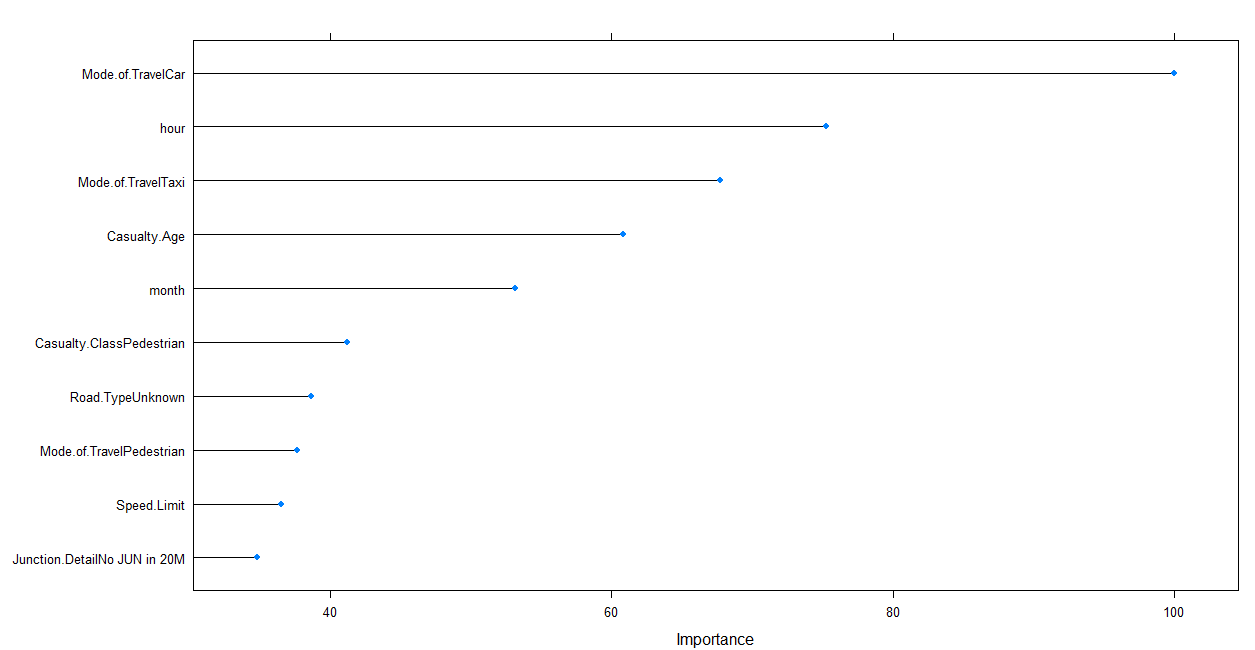
|  |  |  |
| --- | --- | --- |
|  | Reference | |
| Prediction | Serious | Slight |
| Serious | 205 | 135 |
| Slight | 198 | 465 |

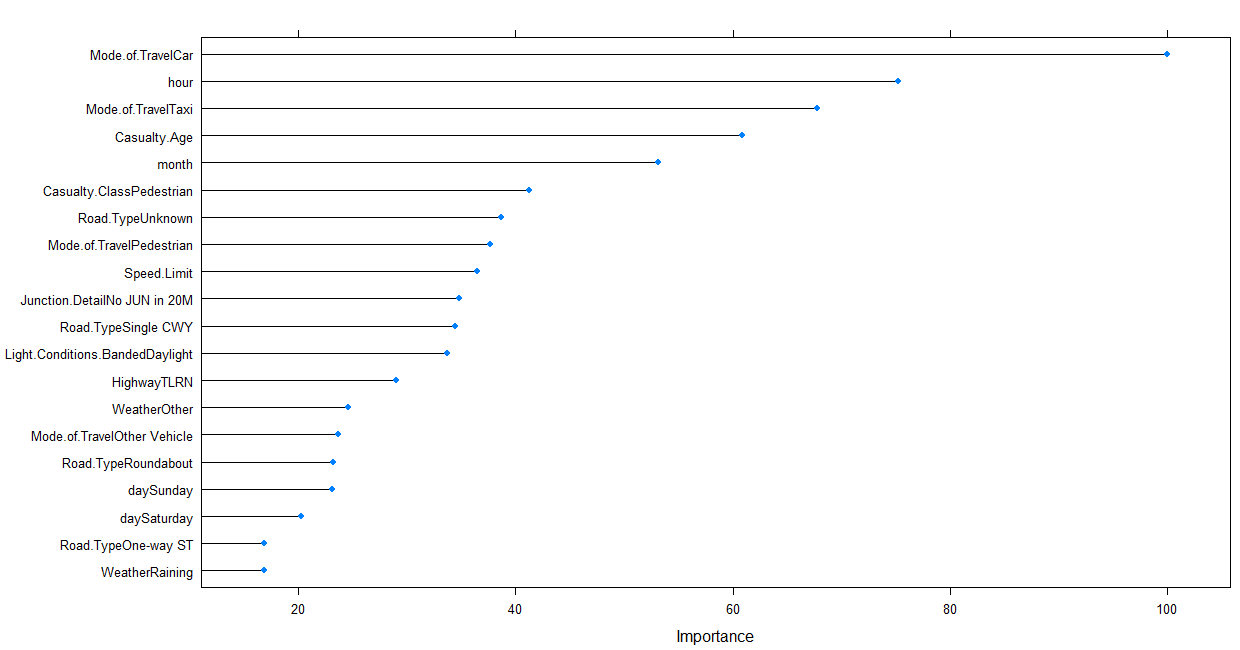
|  |  |
| --- | --- |
| Accuracy | 0.668 |
| Precision | 0.603 |
| Recall | 0.508 |
| F1 | 0.552 |

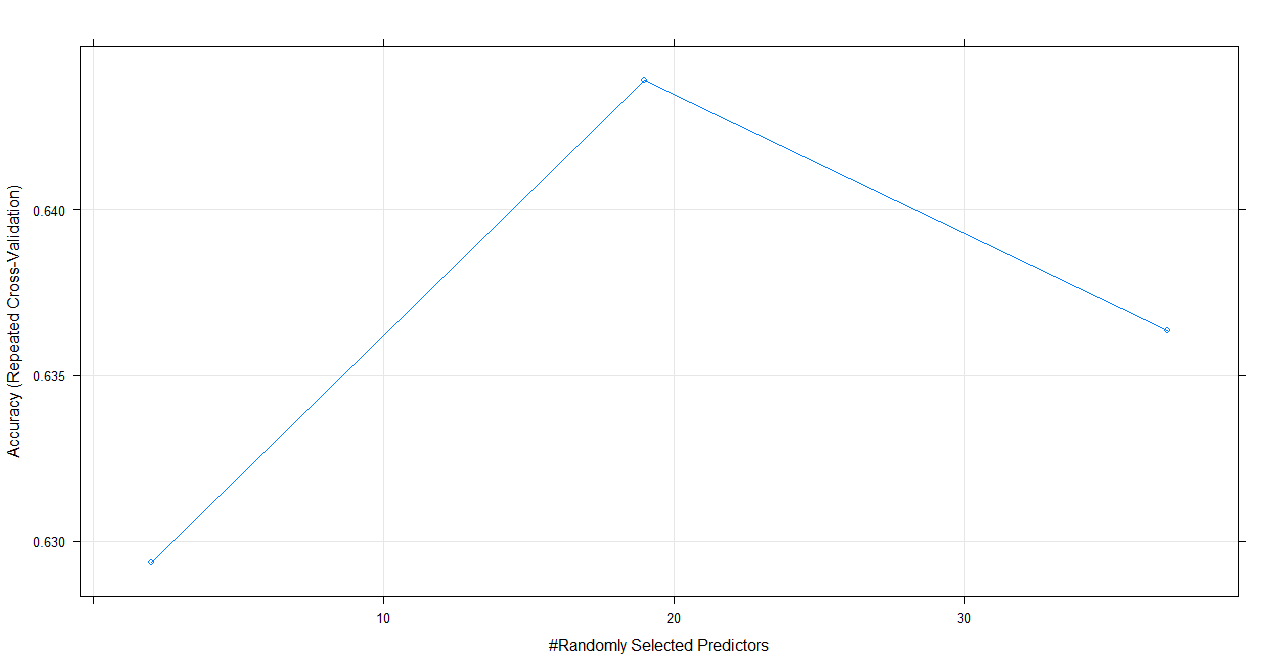


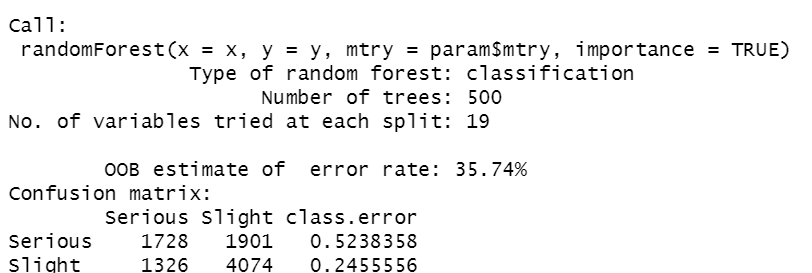
-even though the accuracy is around 70% but the F1 and Recall is quite low compared to Precision; therefore, this model has problem with False Negative rate

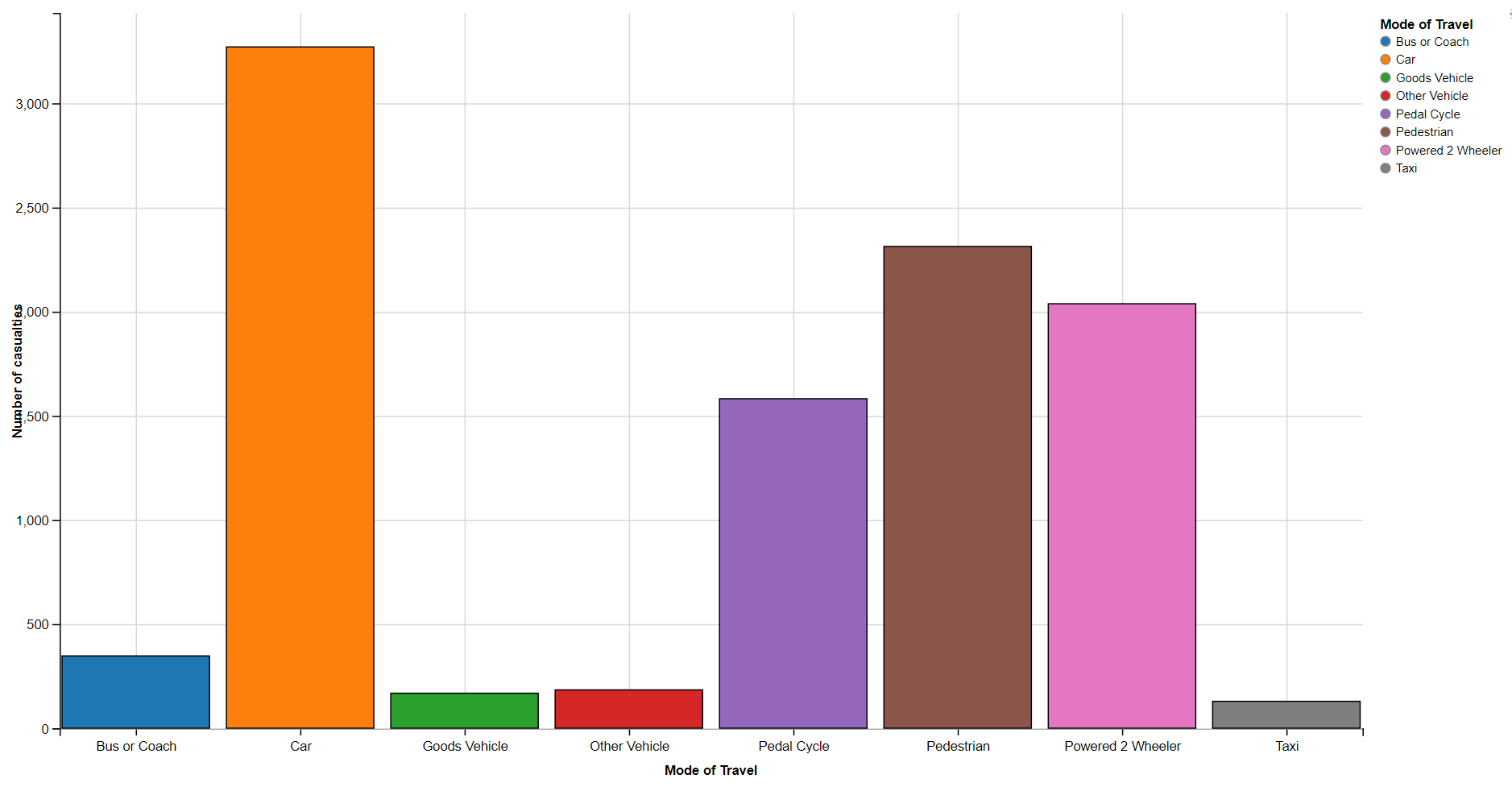
**8. Any relevant summary or conclusions.**

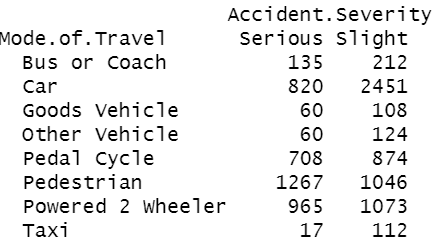


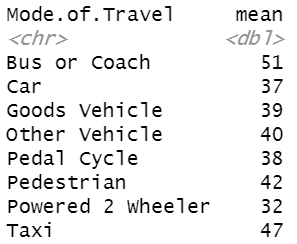


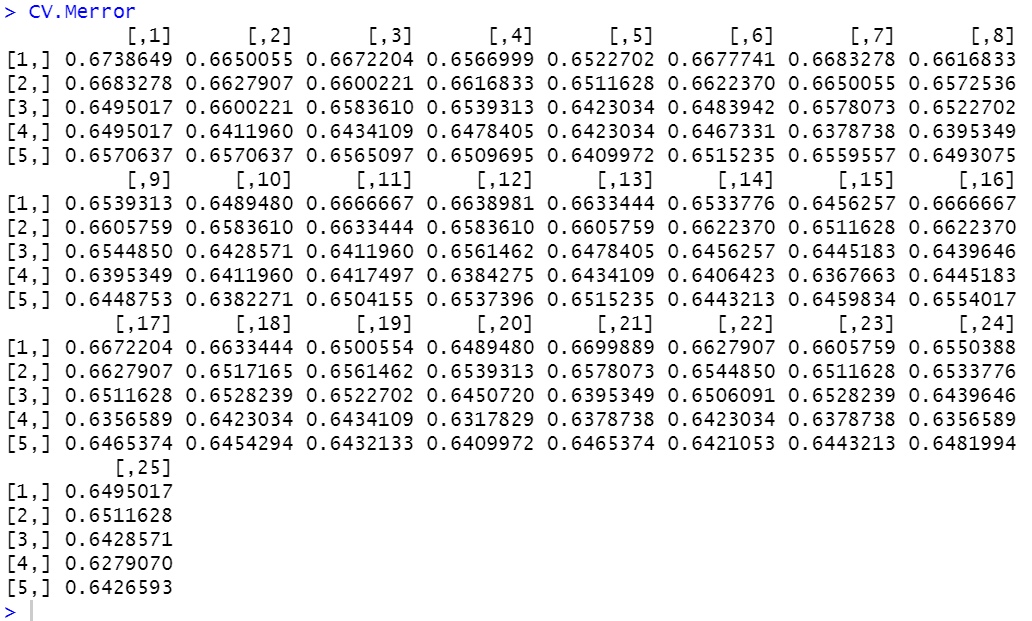


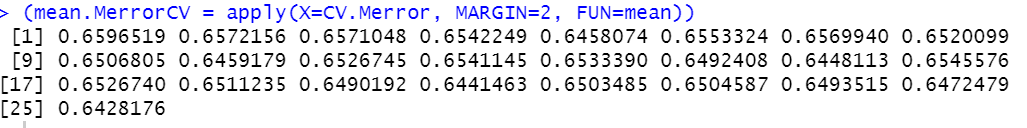


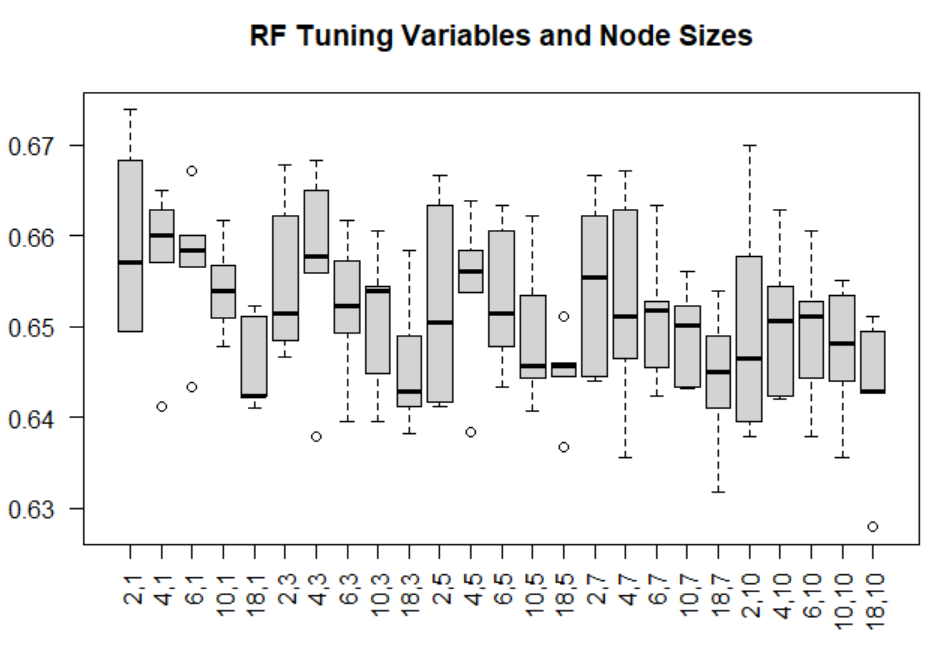


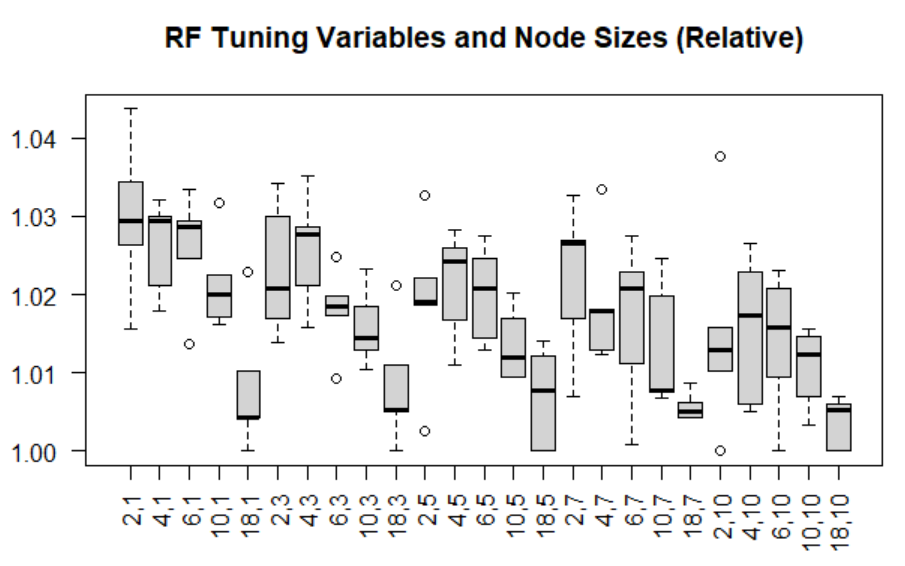












2.

-Road accidents are major causes of deaths and injuries reported every year.

-According CDC, Every day, almost 3,700 people are killed globally in crashes

-It is estimated to be the ninth leading cause of death globally for all age groups. More people now die in crashes than from HIV/AIDS.

-it is estimated to cost world economy around 1.8 trillion dollar.

-However, it is often challenging to determine which conditions lead to these events, making it more difficult for us to address the severity of road accidents.

-We know that some characteristics of vehicles and the surroundings play a key role (engine capacity, condition of the road, etc.).

-However, many questions are still open. Which of these factors are the leading ones?

(- How much are the external factors to blame, compared to the driver skills?)

3. -In this project, I use 2019 UK Road Safety Data.

-the casualties in road accidents,

-the types of vehicles involved and the circumstances.

-However, I use a subset of 2019 dataset.

-since Full original dataset is highly unbalanced which have 3 classes: Fatal, Serious, Slight (of 0.5%/14%/84.5%)

-But Fatal is just 0.5 percent, so I drop that class and turn the problem into binary classification

4.

Data Preprocessing + Exploratory Data Analysis

Model Exploration

Parameter Tuning

Prediction and Analysis

5.

-merge 3 files of attendant, vehicle, and environment details based on ID

-drop features that has more than 90% percent of same values

-drop one of pair features that have correlation more than 95% percent

-drop column with text that need nlp technique to extract information

-process datetime column to hour, day, and month

-Then split dataset into train/test set as shown in the slide

-I try to split dataset into same proportion between 2 class as seen in 40/60% in all table

-the test dataset is kept into separate csv file for using in the prediction steps (last step)

6.

-preprocess dataset by EDA such as seeing matrix of missing value

-collapse some values of feature to smaller class so we have smaller dummy encoding ex. 7 class to 3 class + 1 as other

7.

-in model exploration, I pick 4 different models and use 10-fold cross validation

-the left one is I did PCA before doing modelling

- the metric I use to measure these models is accuracy which is 1-misclassification rate

-surprising result is that logistic regression has far lower accuracy and

RF beats XGboost.

8.

-I also did feature scaling, feature engineering like datetime

-and this is some variable importance I get from random forest.

-surprising thing to see is hour, day, month, and age are very important.

9.

-I tune the random forest.

-using that training data to try 2 methods:

1. using 5-fold cv with 5 repetition of each pair model in each fold

meaning that in each fold, I run every set of parameter 5 times to calculated its average accuracy

2. using OOB error with 5 repetition for each pair of hyperparameter (mtry and nodesize)

-get the best model of each method

-use the sets of parameters from this tuning to train model all full training dataset

the OOB rate from random forest is computed while the model is being build whereas CV uses holdout samples that are predicted after the random forest model is computed.

10.

-make prediction on holdout test dataset that is in separate csv file

-we can see in all the models we have has accuracy below 70%

-the best model in term of accuracy is RF with 2 mtry and 2 nodes

-but we can see trade-off between 2 and 4 mtry: 2 mtry get lower Recall but higher precision

-while 4 mtry is opposite.

-and F1 of 4 mtry is higher

(-even though the accuracy is around 70% but the F1 and Recall is quite low compared to Precision; therefore, this model has problem with False Negative rate )

11.

-Road accident is still a challenging problem as we can see our best accuracy is still below 70%

Especially for real world dataset

-parameter tuning is important as we can see slight improvement

-For future work, we can try better feature engineering

Explore underlying parameters inside RF

Add more data (using multiple years)