ACTL3142 Assignment

Evan Xiong z5421149

## Executive Summary

The report analyses the design and effectiveness of multiple predictive models. These models aim to understand what factors influence fatal accidents in hopes of reducing fatalities on the roads in Victoria. The models are evaluated with the AUC criterion and stem from classification tree methods to shrinkage methods. Through analysing each model; this report utilises the boosting model to provide a description in regards to which characteristics of the drivers would result in a fatal accident.

## Exploratory Data Analysis

Figure 1 displays the proportion of fatal accidents that occurred at a particular speed zone and road geometry. This stacked column chart shows that a major proportion of fatal accidents occurred at speeds of around 100 to 110km/h, as under these speed zones around 40% of accidents were fatal. By moving along the x-axis, we can see that the faster speeds at which the drivers were allowed to move have led to a shift to where fatal accidents occur. At lower speeds, fatal accidents mainly occur not at intersections, whereas at higher speeds fatal accidents tend to occur regardless of the road geometry. Therefore, a key factor that drives the fatality rate is the speed at which the person was driving. As regardless of how the accident has occurred, moving at faster speeds results in more fatal accidents.

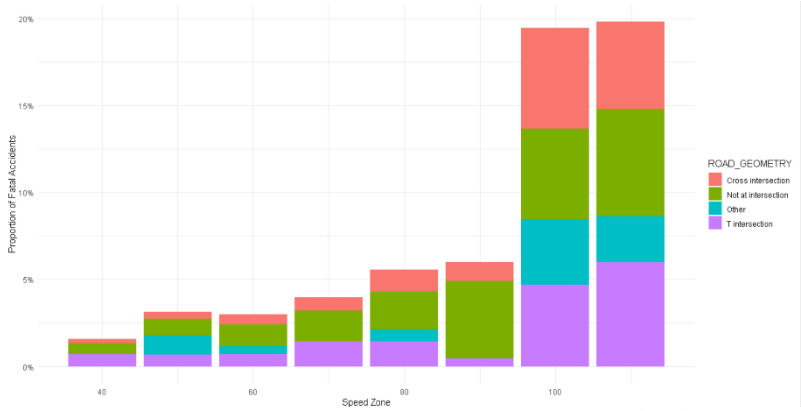



*Figure 1 The Relationship Road Geometry and Speed Zone have on the Proportion of Fatal Accidents*

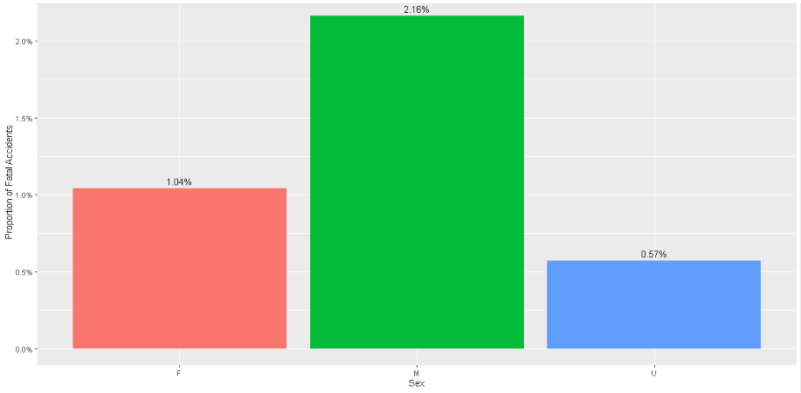



*Figure 2 Proportion of Fatal Accidents by Sex*

Figure 2 displays the proportion of fatal accidents by sex. This column chart shows that 2.16% of accidents involving males are fatal. Comparatively, 1.61% of fatal accidents involve other genders. This high fatality rate of 2.16% is a result of males being risk-takers. Due to this, males tend to adopt a more reckless driving style which corresponds to more fatal accidents. Thus, the gender of the driver and more specifically whether the driver was male or not is crucial in determining the fatality rate.

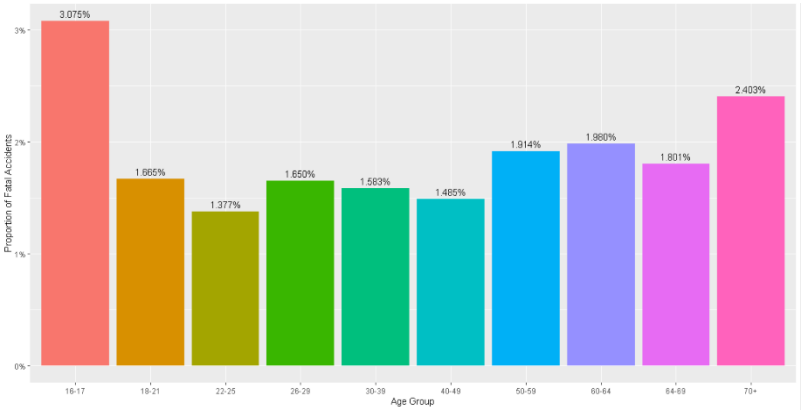



*Figure 3 Proportion of Fatal Accidents by Age Group*

Figure 3 displays the proportion of fatal accidents by age group. The modes of this bimodal chart lie in age groups 16-17 and 70+. From the graph we can see that, 3.075% of accidents involving drivers in the age group 16-17 were fatal whereas 2.403% of accidents involving

drivers in the age group 70+ were fatal. These high figures were due to the reckless driving of young adults and the fragile bodies of the elderly. Thereby age, more specifically, determining whether people are older than 70 or not is crucial in driving the fatality rate. We disregard the other modal peak as people aged 16-17 are under-represented within the data and hence are not significant predictors.

Figure 4 displays the proportion of fatal accidents by the manufacturing year of the vehicle. This line graph illustrates a downward trend in the fatality rate. The decrease in the fatality rate over time is due to the implementation of various safety features in newer vehicle models like airbags and brake assist. However, surprisingly cars manufactured in 2020 have the highest fatality rate. This abnormality in the data is maybe caused by the advent of Covid-19. By disregarding this abnormality, we can see that the manufacturing year of the vehicle is an integral predictor that drives the fatality rate.

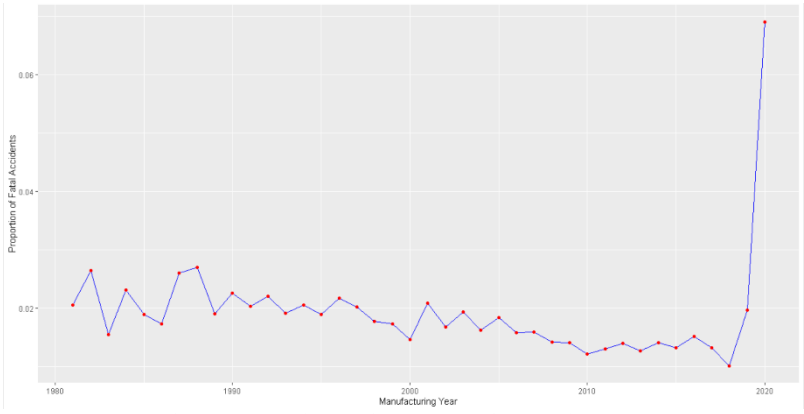



*Figure 4 Proportion of Fatal Accidents by Manufacturing Year of the Vehicle*

Figure 5 displays the proportion of fatal accidents by time of day. This multiple-line graph has a positive skew which indicates that most fatal accidents happen in the early hours of the day. The higher proportion of accidents around these times is caused by the reckless driving of intoxicated drivers. We can also see that most of the fatal accidents occur on Sunday and Monday. This corresponds with our previous findings as during the weekend, people take their minds off work by being under the influence of alcohol. Therefore, the hour of the day is significant in driving the fatality rate.

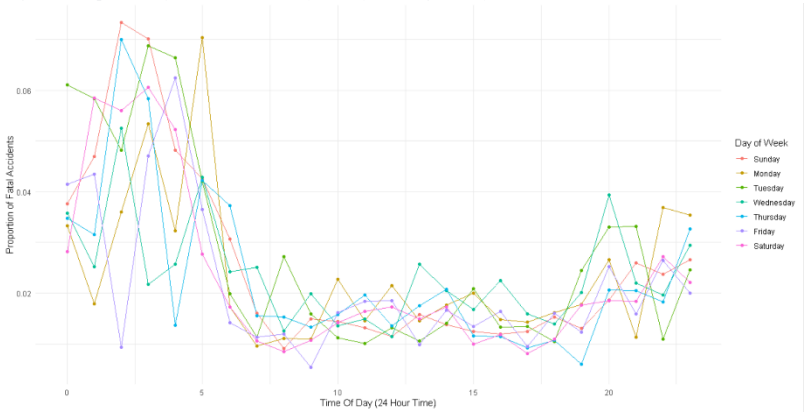



*Figure 5 Proportion of Fatal Accidents by Time of Day*

Figure 6 displays the proportion of fatal accidents by vehicle colour. This column graph shows that 2.240% of accidents involved a white-coloured vehicle. This higher proportion contradicts the studies found online (Youi Insurance 2017) as it is known that white vehicles are the safest as they do not blend in with the road. Therefore, we can deduce that this high figure is a result of a larger proportion of white cars on the road and other factors like the type of vehicles involved in the accident, the type of accident, and the surface conditions when the accident occurred. Thus, the vehicle colour is insignificant in driving the fatality rate.

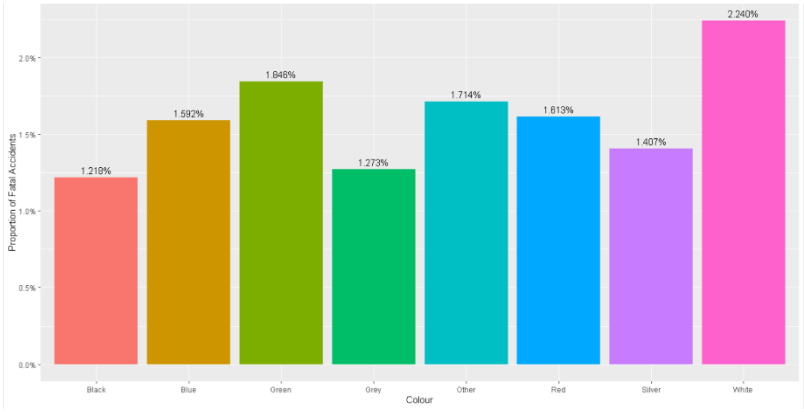



*Figure 6 Proportion of Fatal Accidents by Vehicle Colour*

| Coefficient | Predictors ($X_k$) | Estimate | Type | P-value | Significance |
|---|---|---|---|---|---|
| $\beta_1$ | Speed_Zone | 0.739468 | Numeric | <2e-16 | TRUE |
| $\beta_2$ | ACCIDENT_TYPEcollision with some other object | -1.615810 | Binary | 5.30e-07 | TRUE |
| $\beta_3$ | ACCIDENT_TYPECollision with vehicle | -0.497199 | Binary | <2e-16 | TRUE |
| $\beta_4$ | ACCIDENT_TYPEFall from or in moving vehicle | -0.009957 | Binary | 0.9749 | FALSE |
| $\beta_5$ | ACCIDENT_TYPENo collision and no object struck | -1.133724 | Binary | 1.90e-05 | TRUE |
| $\beta_6$ | ACCIDENT_TYPEStruck animal | -1.445423 | Binary | 9.63e-07 | TRUE |
| $\beta_7$ | ACCIDENT_TYPEStruck Pedestrian | 1.427956 | Binary | < 2e-16 | TRUE |
| $\beta_8$ | ACCIDENT_TYPEVehicle overturned (no collision) | -1.042139 | Binary | < 2e-16 | TRUE |
| $\beta_9$ | Belt_not_worn | 0.189787 | Binary | < 2e-16 | TRUE |
| $\beta_{10}$ | VEHICLE_TYPEHeavy Vehicle (Rigid) > 4.5 Tonnes | 1.217020 | Binary | < 2e-16 | TRUE |
| $\beta_{11}$ | VEHICLE_TYPEOther | 0.494869 | Binary | 6.60e-10 | TRUE |
| $\beta_{12}$ | VEHICLE_TYPEPanel Van | 0.108928 | Binary | 0.3398 | FALSE |
| $\beta_{13}$ | VEHICLE_TYPEPrime Mover - Single Trailer | 1.074797 | Binary | < 2e-16 | TRUE |
| $\beta_{14}$ | VEHICLE_TYPEStation Wagon | 0.089443 | Binary | 0.0688 | FALSE |
| $\beta_{15}$ | VEHICLE_TYPETaxi | -0.098670 | Binary | 0.6078 | FALSE |
| $\beta_{16}$ | VEHICLE_TYPEUtility | 0.295819 | Binary | 1.69e-07 | TRUE |
| $\beta_{17}$ | Male_SEX | 0.210176 | Binary | < 2e-16 | TRUE |
| $\beta_{18}$ | AGE_Over_70 | 0.148396 | Binary | < 2e-16 | TRUE |
| $\beta_{19}$ | TOTAL_NO_OCCUPANTS | 0.098722 | Numeric | < 2e-16 | TRUE |
| $\beta_{20}$ | SURFACE_CONDOther | -1.279828 | Binary | 3.02e-09 | TRUE |
| $\beta_{21}$ | SURFACE_CONDWet | -0.264656 | Binary | 1.34e-07 | TRUE |
| $\beta_{22}$ | Not_At_Inter | 0.152186 | Binary | 9.84e-14 | TRUE |
| $\beta_{23}$ | VEHICLE_YEAR_MANUF | -0.100319 | Numeric | 1.29e-08 | TRUE |

Building a Logistic Regression with 23 Predictors

We model the data with a logistic regression since the response $Y$ is binary. We let $Y = 1$ denote that the accident was fatal whereas we let $Y = 0$ denote that the accident was not fatal.

$$\ln\left(\frac{Pr(Y=1|X)}{1-Pr(Y=1|X)}\right) = -4.29 + 0.74X_1 - 1.62X_2 - 0.50X_3 - 0.01X_4 - 1.13X_5 - 1.45X_6 + 1.43X_7 - 1.04X_8 + 0.19X_9 + 1.22X_{10} + 0.49X_{11} + 0.11X_{12} + 1.07X_{13} + 0.09X_{14} - 0.10X_{15} + 0.30X_{16} + 0.21X_{17} + 0.15X_{18} + 0.10X_{19} - 1.28X_{20} - 0.26X_{21} + 0.15X_{22} - 0.1X_{23}$$

The above table of predictors summarises the impact of each of the predictors on the probability of an accident being fatal. Under logistic regression, the response and predictors have a multiplicative relationship. Therefore, by holding all other factors constant, a unit increase in $X_k$ corresponds to the odds changing by a factor of $e^{\beta_k}$. If the $\beta$'s are negative, this translates to odds that are less than one which thereby indicates that Pr $(Y = 1|X)$ decreases as the covariates increases. Through this, we can deduce that if the covariates with positive coefficient increase, the fatality probability will also increase whereas if the covariates with negative coefficient decrease, the fatality probability will also decrease. This relationship between the predictors and the response is supported by our exploratory data analysis.

Figure 1 indicates that driving in areas with higher speed zones corresponds to a higher fatality proportion. From the table, we can see that the coefficient for the "SPEED_ZONE" variable is 0.74, this thereby indicates that an increase in "SPEED_ZONE" corresponds to an increase in the fatality probability. Therefore, this aligns with our findings in figure 1, as a higher fatality proportion is a result of higher fatality probabilities. This can be further extended to other variables.

By considering figure 2, we deduce that being male corresponds to an increase in the fatality rate, this aligns with the positive coefficient of Male_SEX. Additionally, by considering figure 3, we deduced that individuals being aged over 70 corresponds to an increase in the fatality rate, this aligns with the positive coefficient of "AGE_Over_70". Lastly, by considering figure 4, we can support the relationship "VEHICLE_YEAR_MANUF" has with fatal accident probability. In figure 4, we deduced that owning newer manufactured cars would result in a decrease in the fatality rate, this aligns with the negative coefficient of "VEHICLE_YEAR_MANUF". Therefore, the nature of the covariates in this model can be explained by exploratory data analysis.

Since many of the variables in the dataset are categorical, most of the covariates we consider here take on binary values. This means that either $X_k = 0$ or $X_k = 1$. These covariates take on binary values because they are levels of a categorical variable.

From the table, we can see that out of the 23 predictors, 4 of them were deemed insignificant. We reached this conclusion by performing hypothesis testing. By testing the null $H_0: \beta_k = 0$ against the alternative $H_1: \beta_k \neq 0$ we can determine which predictors appear to be significant. By considering the p-values, we can see that the p-values of $\beta_4, \beta_{12}, \beta_{14}, \beta_{15}$ are larger than the 5% significant level and as a result these predictors are deemed insignificant. Thus, the regression model has 19 significant predictors.

Selection of Predictors

| | Df | Deviance | Resid. Df | Resid. Dev | Pr(>Chi) | |
|---|---|---|---|---|---|---|
| NULL | | | 199999 | 34343 | | |
| SPEED_ZONE | 1 | 2201.93 | 199998 | 32141 | < 2.2e-16 | *** |
| ACCIDENT_TYPE | 7 | 1069.38 | 199991 | 31072 | < 2.2e-16 | *** |
| Belt_not_worn | 1 | 331.38 | 199990 | 30741 | < 2.2e-16 | *** |
| VEHICLE_TYPE | 7 | 345.64 | 199983 | 30395 | < 2.2e-16 | *** |
| Male_SEX | 1 | 126.00 | 199982 | 30269 | < 2.2e-16 | *** |
| AGE_Over_70 | 1 | 84.16 | 199981 | 30185 | < 2.2e-16 | *** |
| TOTAL_NO_OCCUPANTS | 1 | 72.45 | 199980 | 30112 | < 2.2e-16 | *** |
| SURFACE_COND | 2 | 78.21 | 199978 | 30034 | < 2.2e-16 | *** |
| Not_At_Inter | 1 | 54.53 | 199977 | 29980 | 1.530e-13 | *** |
| VEHICLE_YEAR_MANUF | 1 | 32.07 | 199976 | 29947 | 1.487e-08 | *** |

After cleaning the data, there were 10 variables left to perform logistic regression with. To ensure that the subset selection process was correct, we performed a confirmatory test. By plotting and analysing a deviance table, we can determine whether there was significant improvement in the model when increasing the model complexity via the addition of each predictor. We achieve this by comparing two models $M_1$ and $M_2$, whereby $M_1$ is the initial model whereas $M_2$ is the model after adding a predictor. A general rule of thumb in determining whether the improvement in the model was significant is if $D_1 - D_2 > 2(p - q)$, where $D_1 - D_2$ is the difference in the deviances of the models and $p - q$ is the increase in degrees of freedom when incorporating a predictor into the model. Therefore, we can conclude that our model requires all these predictors as the above inequality is satisfied for each variable added into the model.

Imbalanced Data

When dealing with imbalanced data, instead of random sampling, we utilised stratified sampling, as doing so ensures that the proportion of fatal accidents is kept the same in the training and test set. With this, a more accurate prediction can be achieved as the number of fatal accidents will not be over/under-represented in the training and test sets. Additionally, we standardised all the non-categorical variables to ensure the data is of the same scale. Through this, we can directly compare the predictors.

Best Predictive Model Analysis

Out of all the models we tested, the best-performing model is the boosting model. By using this model, we were able to obtain the most accurate predictions regarding whether an accident was fatal or not. The boosting model is an extension of the classification trees. It involves combining a large number of decision trees. Unlike classification trees, each tree grows sequentially utilising the information from previously grown trees, these trees then focus on areas the previous trees did not perform well in. This model was chosen as it has the highest AUC of 0.807. I chose the best model based on this criteria as finding the optimal threshold of the confusion matrix for each model is too dependent on various factors. This is illustrated in the ROC curves below in figure 7. From the plot, we can see the ROC curve of the boost model hugs the top left corner considerably more than other tree-based models.

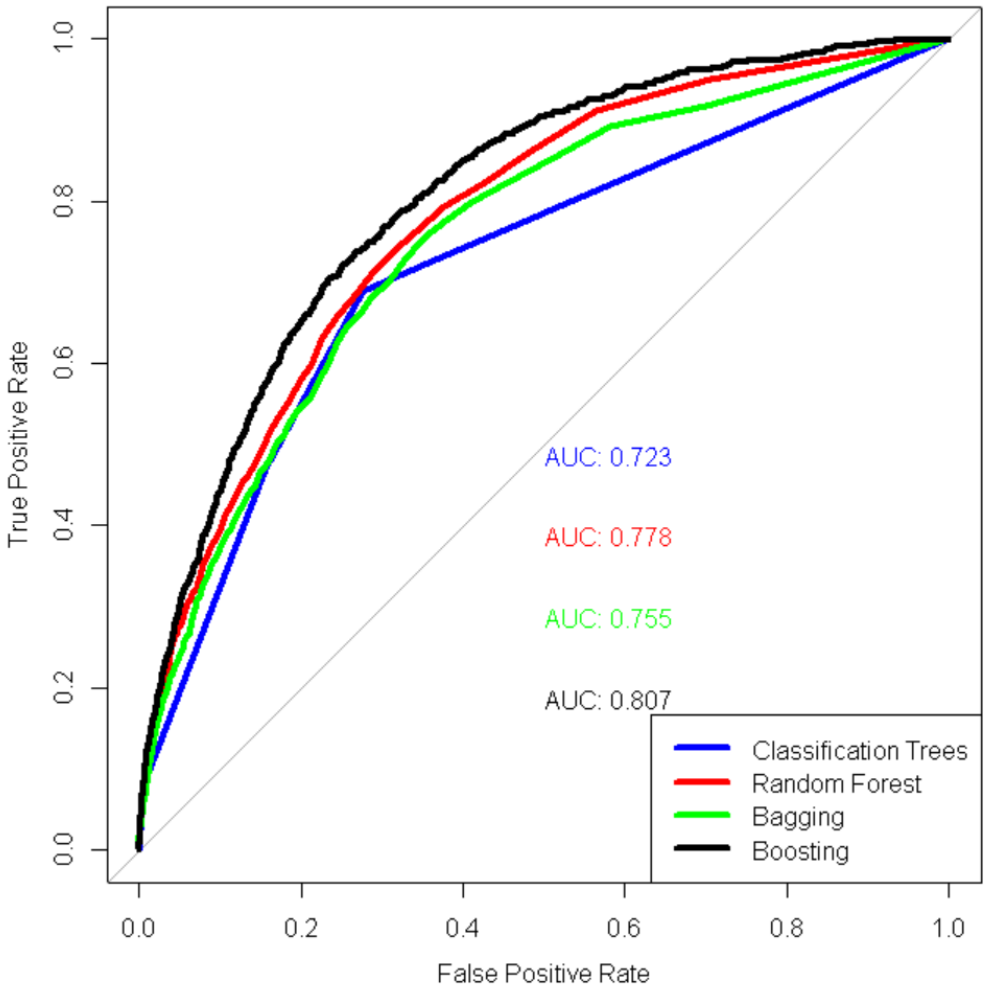



*Figure 7 ROC Curves for Tree-Based Models*

To obtain more inferences on the boost model, we can calculate the true positive rate. Under the 1.7% threshold, we obtain the confusion matrix below. This threshold was chosen such that it aligns with the probability that an accident is fatal in the training set. Furthermore, this particular threshold was also chosen in order to remain consistent with the thresholds of the confusion matrices for the other models (see in Technical Appendix).

| | Predicted | |
|---|---|---|
| Actual | 0 | 1 |
| 0 | 37778 | 11396 |
| 1 | 248 | 578 |

Since, we are interested in predicting whether an accident is fatal or not, calculating the true positive rate becomes more meaningful than the test error rate. The $TPR = \frac{578}{11974} \approx 0.048$. This rate is relatively higher than the $TPR$ of other models (see in Technical Appendix).

The boosting model consists of three tuning parameters, namely the number of trees, the number of splits in each tree and the shrinkage. Due to computational reasons and a lack of strong hardware, the number of trees we chose was 500. The other two parameters were left as their default setting, as running the boost model was already computationally intensive.

Recommendation

To assist in the development of a prevention campaign, we are providing a non-technical description of the characteristics of the drivers that are mostly likely to be involved in fatal accidents. With the use of the variable importance table, we can identify the most important predictors in determining whether a driver would be involved in a fatal accident or not. The table illustrates the relative influence of each predictor. Variables with high relative influence are considered strong predictors of the model. Through this, with a relative influence of 26.09, 22.60, and 19.13 we can deduce that the most important predictors are “AGE”, “VEHICLE_YEAR_MANUF” and “VEHICLE_TYPE” respectively. These results are consistent with the findings in our exploratory data analysis and construction of the logistic regression model. Figure 3 and 4 highlights the importance of age and the manufacturing year of the vehicle whilst the fitted logistic regression model highlights the significance of all of these predictors. Based on these previous findings, individuals aged over 70 who drive a vehicle that was manufactured before the 2000s and heavier than 4.5 tonnes are mostly likely to be involved in a fatal accident. A list of the 2500 drivers who are most likely to engage in a fatal accident out of 10000 drivers in the evaluation dataset is provided in the file “FATAL_DRIVERS_EVAL.csv”. This list was generated by tabulating the predictions of the best predictive model, the boost model. These prediction probabilities are then arranged in descending order such that the 2500 most vulnerable drivers are selected. We recommend that these 2500 drivers be investigated as they most likely exhibit the characteristic of drivers that would be involved in a fatal accident as describe above.

| | var | rel.inf |
|---|---|---|
| AGE | AGE | 26.091333 |
| VEHICLE_YEAR_MANUF | VEHICLE_YEAR_MANUF | 22.602058 |
| VEHICLE_TYPE | VEHICLE_TYPE | 19.136479 |
| HELMET_BELT_WORN | HELMET_BELT_WORN | 13.321391 |
| TOTAL_NO_OCCUPANTS | TOTAL_NO_OCCUPANTS | 9.557358 |
| FUEL_TYPE | FUEL_TYPE | 6.286745 |
| SEX | SEX | 3.004636 |

Technical Appendix

Construction of Our Logistic Regression Model

For this analysis, we have assumed that the dataset contains accurate and reliable data. This is not always true as there may be discrepancies within the data as well as the existence of unexplained predictors. Additionally, we assumed that the predictors are independent of each other. This assumption isn't as important as lasso regression aims to reduce the multicollinearity of the predictors.

Prior to the construction of the model the dataset should be cleaned to reduce the computational time of the code. To start we removed 14 variables from the dataset. These variables "DRIVER_ID", "VEHICLE_ID", "OWNER_POSTCODE", "LICENCE_STATE", "ACCIDENT_NO", "FUEL_TYPE", "ACCIDENTTIME", "AGE", "ACCIDENTDATE", "VEHICLE_MAKE", "VEHICLE_BODY_STYLE", "ATMOSPH_COND", "ROAD_SURFACE_TYPE" and "LIGHT_CONDITION" were excluded as they were irrelevant or encompassed by another variable. For example, "ACCIDENTTIME" and "ACCIDENTDATE" are encompassed by a variable we've made "HOUR_OF_DAY". Additionally, the response parameter "fatal" was converted from a "LOGICAL" type to a "NUMERIC" type. This would allow us to ultimately perform logistic regression with "fatal".

In order to obtain the most parsimonious model, we used lasso regression to perform another round of variable selection as currently, each variable has numerous levels which in turn results in a highly uninterpretable model. Firstly, we conduct stratified sampling to create a 75 to 25 split of the training and test set.

```
> print(prob)
[1] 0.01699333
> print(pro)
[1] 0.016935
> print(proba)
[1] 0.01676
```

With this, we can see that the fatality proportion is around 1.7% for the dataset, training set and test set. The slight discrepancy in the proportions is caused by random sampling.

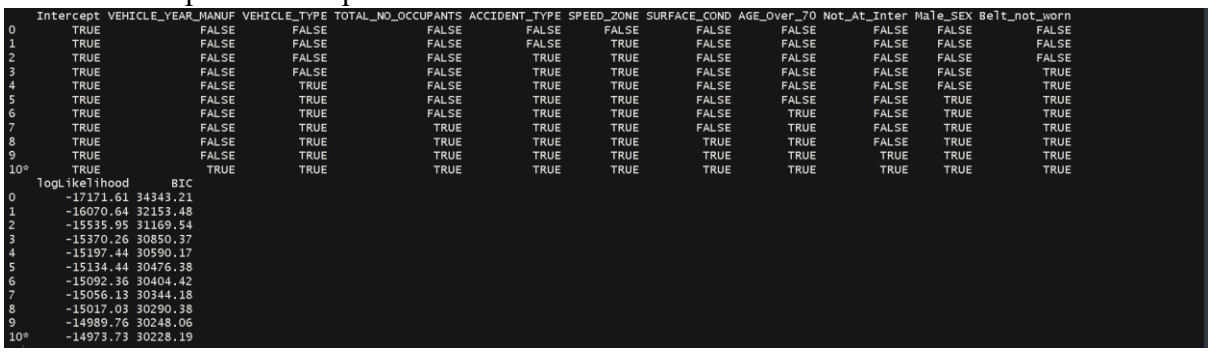
By performing lasso regression and conducting cross validation on the training set, we obtain the best lambda within one standard error. With this, we obtain 19 predictors that were not minimised to 0.

```
 [1] "(Intercept)"                                  "SEXM"
 [3] "Age.Group70+"                                 "HELMET_BELT_WORNSeatbelt not worn"
 [5] "VEHICLE_YEAR_MANUF"                           "VEHICLE_TYPEHeavy Vehicle (Rigid) > 4.5 Tonnes"
 [7] "VEHICLE_TYPEOther"                            "VEHICLE_TYPEPrime Mover - Single Trailer"
 [9] "VEHICLE_TYPEUtility"                          "TOTAL_NO_OCCUPANTS"
[11] "ACCIDENT_TYPEcollision with some other object" "ACCIDENT_TYPECollision with vehicle"
[13] "ACCIDENT_TYPEStruck animal"                   "ACCIDENT_TYPEStruck Pedestrian"
[15] "ACCIDENT_TYPEVehicle overturned (no collision)" "ROAD_GEOMETRYNot at intersection"
[17] "SPEED_ZONE"                                   "SURFACE_CONDOther"
[19] "SURFACE_CONDWet"
```

Knowing this, we further cleaned the data by adding variables "Helmet_not_worn", "Age_over_70", "Not_at_intersection" and "Male_SEX" and removed the predictors that were minimised to 0 by lasso regression. These include "HOUR_OF_DAY", "VEHICLE_COLOUR", "AGE", "SEX" and "ROAD_GEOMETRY". We added specific levels of the categorical variables in order to reduce the computational time of subset

selection. “VEHICLE_TYPE” and “ACCIDENT_TYPE” were not manipulated and transformed further as most of its level were significant.

Using this new cleaned dataset. We perform forward stepwise subset selection for logistic regression. Through this, we obtain a model which requires the use of 10 variables. The information criterion used when performing subset selection was the Bayesian Information Criterion (BIC). By utilising this criterion for subset selection, we can penalise more heavily on model complexity compared to Akaike Information Criterion (AIC). This ensures we obtain the simplest model possible.



| | Intercept | VEHICLE_YEAR_MANUF | VEHICLE_TYPE | TOTAL_NO_OCCUPANTS | ACCIDENT_TYPE | SPEED_ZONE | SURFACE_COND | AGE_Over_70 | Not_At_Inter | Male_SEX | Belt_not_worn |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE |
| 1 | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE |
| 2 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE |
| 3 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE |
| 4 | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE |
| 5 | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| 6 | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE |
| 7 | TRUE | FALSE | TRUE | TRUE | TRUE | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE |
| 8 | TRUE | FALSE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | FALSE | TRUE | TRUE |
| 9 | TRUE | FALSE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE |
| 10* | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE |

| | logLikelihood | BIC |
|---|---|---|
| 0 | -17171.61 | 34343.21 |
| 1 | -16070.64 | 32153.48 |
| 2 | -15535.95 | 31169.54 |
| 3 | -15370.26 | 30850.37 |
| 4 | -15197.44 | 30590.17 |
| 5 | -15134.44 | 30476.38 |
| 6 | -15092.36 | 30404.42 |
| 7 | -15056.13 | 30344.18 |
| 8 | -15017.03 | 30290.38 |
| 9 | -14989.76 | 30248.06 |
| 10* | -14973.73 | 30228.19 |

Before fitting our model, we standardise the non-categorical variables such that the numeric variables can be directly compared against each other. With this, a logistic regression is fitted using these 10 variables.

Predictive Models

Logistic Regression

Logistic Regression measures the changes in the log-odds as the values of the predictors change. Under the 1.71% threshold, we obtain the confusion matrix below. This threshold is chosen based on the proportion of fatal accidents in the dataset.

| | 0 | 1 |
|---|---|---|
| 0 | 36385 | 12789 |
| 1 | 237 | 589 |

Given that we are interested in what drives fatality, we should be interested in the true positive rate.

$$TPR = \frac{589}{13378} \approx 0.044$$

Since there are too many factors when deciding the optimal threshold for plotting a confusion matrix. A better measure for the model’s performance is AUC. The AUC for this model is 0.795 which is surprisingly accurate given it is extremely simple to compute.

KNN

The K-Nearest-Neigbours approach classifies a data point based on its nearest K neighbours. For this KNN model, we chose K = 1, as by reducing the K the AUC increased by a minuscule amount. Under the 50% threshold, we obtain the confusion matrix below.

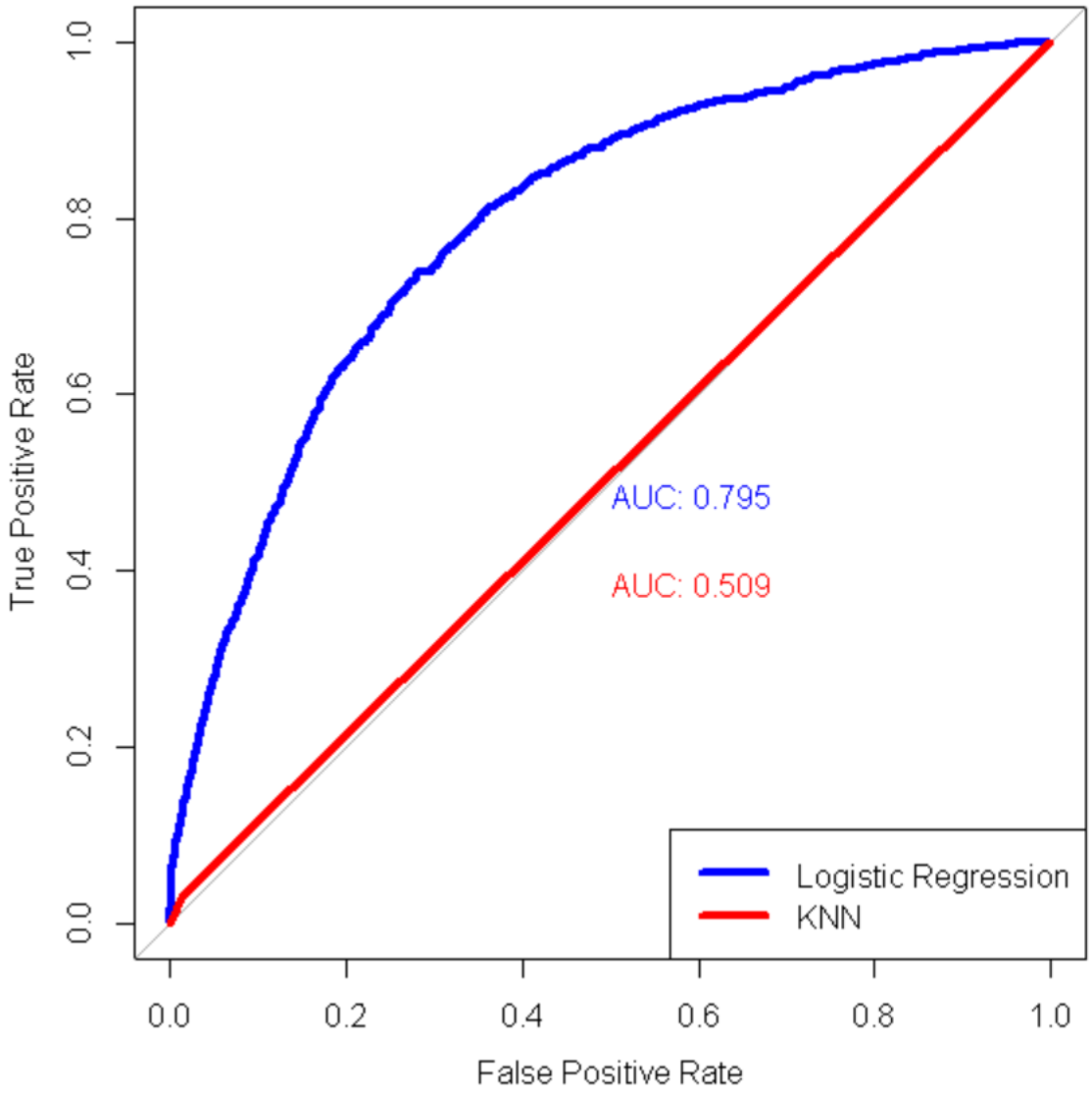
| | 0 | 1 |
|---|---|---|
| 0 | 48841 | 808 |
| 1 | 333 | 18 |

The true positive rate for this model is

$$TPR = \frac{18}{826} \approx 0.022.$$

The AUC for KNN was 0.509 which was the lowest out of all the models available. This is due to KNN's inability to handle large datasets.

ROC curves for these Methods




Shrinkage Methods

Lasso Regression

Lasso regression is a shrinkage method that pushes the estimates of $\beta_k$ towards 0, this allows it to perform feature/variable selection. The tuning parameter $\lambda$ was chosen using cross validation. Since we wanted to make the most accurate predictions, we choose the lowest $\lambda$ possible to fit the model. Under the 1.7% threshold, we obtain the confusion matrix below.

| | 0 | 1 |
|---|---|---|
| 0 | 36087 | 13075 |
| 1 | 237 | 601 |

The true positive rate for this model is

$$TPR = \frac{601}{13676} \approx 0.044\%.$$

The AUC for lasso regression is also 0.795 which is surprising as we assume that lasso regression would perform better than logistic regression.

Ridge Regression

Ridge regression is a shrinkage method extremely similar to lasso regression. The only difference is that the estimates of $\beta_k$ are unable to be minimised to 0. The tuning parameter $\lambda$ was chosen using a similar manner using cross validation. Under the 1.7% threshold, we obtain the confusion matrix below.

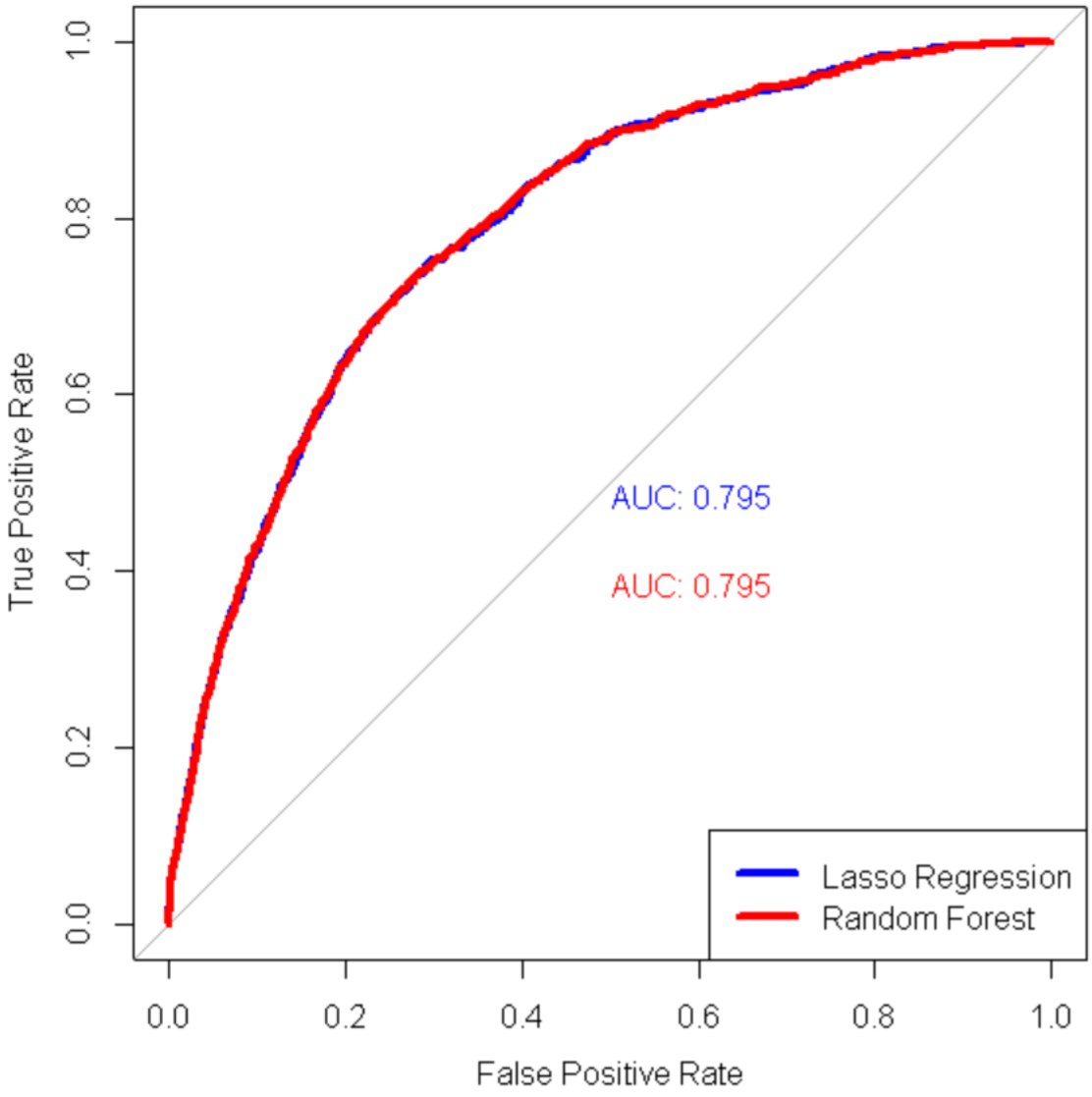
| | 0 | 1 |
|---|---|---|
| 0 | 35841 | 13321 |
| 1 | 231 | 607 |

The true positive rate for this model is

$$TPR = \frac{607}{13928} \approx 0.044\%.$$

The AUC for ridge regression is also 0.795 which indicates there isn't much of a difference between using lasso and ridge regression.

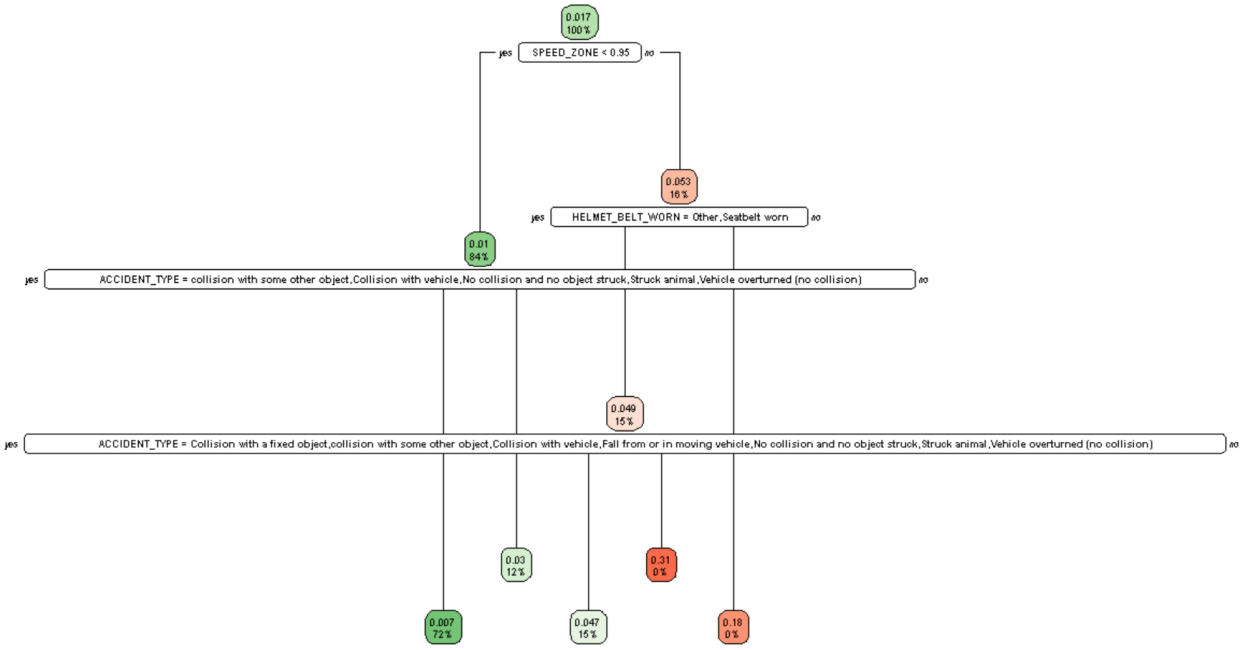
ROC curves for Skrinkage Methods




Tree-Based Methods

Classification Tree

The classification tree is an interpretable approach for predicting qualitative responses. It stems from a root node continually branching out. The number of branches and terminal nodes of the tree can be adjusted to strike a balance between detail and interpretability. The complexity parameter used for the tree below is 0.002 as any complexity parameter above 0.003 only had two terminal nodes and hence didn't provide enough information.




The plot above is a classification tree that represents the data. Although, it is very simple unlike KNN, it can compete with other classification models as the AUC for this model is 0.723.

Random Forest

Random forest is an extension of classification trees. It builds on these bootstrapped training sample such that a random sample of $m \approx \sqrt{p}$ predictors is chosen as split candidates from the full set of $p$ predictors. This model aims to reduce the variance of classification trees. The AUC for this model is 0.778.

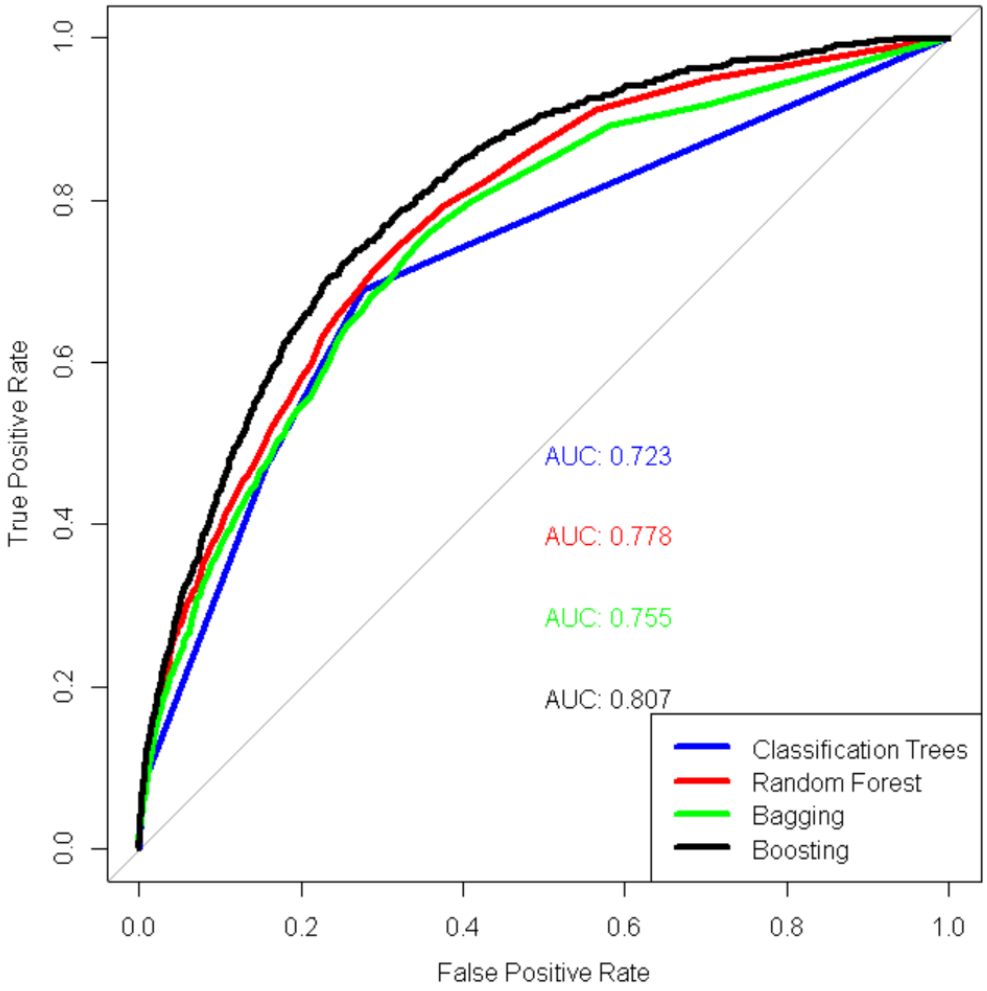
Bagging

Bagging is a special case of a random forest with $m = p$. It usually returns a result worse than random forests. The AUC for this model is 0.755.

Boosting

- Analysis done in main report

ROC curves for Tree-Based Methods



Generative AI Appendix

I have used ChatGPT for mainly debugging and cleaning my code as I don't trust its ability to explain the intricacies about the theory such as subset selection, information criterion, etc.

Some of the key prompts used during the project were:

- [Inputting error messages]
- Pasting my code to GPT (as a sanity check)
- Pasting a packages name into GPT (obtain some elementary data about the package)
- Descriptions of desired outcomes
- Explaining the code on websites

Some Examples

- knn <- knn(train = train, test = test, cl = factor(train$fatal), k = 5) Error in knn(train = train, test = test, cl = factor(train$fatal), k = 5) : NA/NaN/Inf in foreign function call (arg 6) In addition: Warning messages: 1: In knn(train = train, test = test, cl = factor(train$fatal), k = 5) : NAs introduced by coercion 2: In knn(train = train, test = test, cl = factor(train$fatal), k = 5) : NAs introduced by coercion
- ROCR package roc plot
- > plot(roc_score) > roc_score = roc(test_set$fatal, response) Setting levels: control = 0, case = 1 Setting direction: controls < cases Warning message: In roc.default(test_set$fatal, response) : Deprecated use a matrix as predictor. Unexpected results may be produced, please pass a numeric vector. > plot(roc_score)
- step(glm, direction = "forward", IC = "BIC")
- splittools vs createdatapattion stratified sampling r
- predict_col <- subset(new_Vic_crashes_clean, -c(fatal, VEHICLE_TYPE, + ACCIDENT_TYPE)) Error in subset.data.frame(new_Vic_crashes_clean, -c(fatal, VEHICLE_TYPE, : 'subset' must be logical

References


- Forest 2016, *Why Decision tree is outperforming Random Forest in this simple case?*, Cross Validated, viewed 30 July 2023, <https://stats.stackexchange.com/questions/241062/why-decision-tree-is-outperforming-random-forest-in-this-simple-case>.

- Gagan 2017, *Getting probability as 0 or 1 in KNN (predict_proba)*, Stack Overflow, viewed 28 July 2023, <https://stackoverflow.com/questions/41956049/getting-probability-as-0-or-1-in-knn-predict-proba>.

- Price, A 2017, *How can I draw a ROC curve for a randomForest model with three classes in R?*, Stack Overflow, viewed 28 July 2023, <https://stackoverflow.com/questions/46124424/how-can-i-draw-a-roc-curve-for-a-randomforest-model-with-three-classes-in-r>.

- *Study: What Are The Safest Car Colours and Which Are The Most Popular?* 2017, Youi, viewed 30 July 2023, <https://www.youi.com.au/youi-news/study-which-car-colours-are-most-popular-and-which-are-the-safest>.

- Starmer, J 2018, *ROC and AUC in R, YouTube*, viewed 28 July 2023, <https://www.youtube.com/watch?v=qcvAqAH60Yw>.

- in 2015, *Can I use binary variables in logistic glm in R?*, Cross Validated, viewed 28 July 2023, <https://stats.stackexchange.com/questions/183150/can-i-use-binary-variables-in-logistic-glm-in-r#:~:text=The%20short%20answer%20is%20yes,with%20one%20binary%20predictor%20variable.&text=What%20if%20you%20skip%20the,get%20a%20perfect%20separating%20predictor.>.

- user35577 2013, *stratified splitting the data*, Stack Overflow, viewed 30 July 2023, <https://stackoverflow.com/questions/20776887/stratified-splitting-the-data>.