Assignment

Alex Doyle

11/20/2020

Predicting Arrest numbers at NFL stadiums

Table of contents

1.0 Background

1.1 Introduction

For this analysis, the brief set out was to analyze a real data and ask some interesting questions. After searching through various websites this data set jumped out. It is a record of the number of people arrested at National Football League games starting in 2011 up until the 2015 season. Also included are 11 other factors such as the home score such as the visiting team, scores for both teams, the time the game was played etc.

- It is interesting and unique.

- Although the data is a number of years old there has not been an in depth extraction of the data to answer any questions other than "What city had the most arrests over this period of time?"

From these records I hope to be able to explain what factors have the biggest impact on the number of data. Is it the case that some NFL fanbases are are more prone to cause trouble, or is it the case the intense pressure of close game results in a spike in the numbers. Meaning I aim to identify what factors lead to data at games, are these the factors that will be used in the prediction model to accurately predict the number of people arrested at a given game. Personally I expect to see the a rise in the number of people arrested at close games but I do not believe there will be a massive difference across teams as surely higher arrests leads to a reactive security surge.

```
## [1]
"C:/Users/alexd/Documents/C7801_Assignment/data_archive/C7081_Assignment"
```

names	var_type	var_description
season	Factor	year of data recorded 2011 - 2015
week_num	Factor	week of game 1 – 17
day_of_week	Factor	day game was
played		
gametime_local	Factor	time of game(loacl timezone
home_team	Factor	Home team
away_team	Factor	Away team
home_score	Numeric	Score of home team
away_score	Numeric	Score of away team
OT_flag	Factor	indicator of extra time played at the end of the game if
game is tied at the end of regulation time		
arrests	Numeric	number of people arrested
division_game	Factor	game between two teams in the same division (32 teams divided between 8
divisions		

From the first couple rows of data we can get a sense of the variables. We can see that there are 1006 observations of 11 variables. At a glance there are also a number of teams that didn’t release any information about how many arrests occurred. These teams were Cleveland, New Orleans, Buffalo, Miami and Oakland.

2.0 Method

2.1 Feature Data clenseing

From reading the original README from kaggle as mentioned above it does not include numbers from various team but there is are also variables that require editing.

Detroit

Detroit has no information provided relating to the number of arrests over the 5 year period. It would be difficult make a prediction of the average number of arrests for the entire team as we will see later on in the analysis that the home team is an important predictor. This means that all inputs for the detroit are removed. We can see that there are 40 NA values in arrests variable.

Overtime The variable OT_flag which signify extra time having to be played at the end of regulation time due to the game finishing a tie is currently set up as 1 (there was overtime) and NA (indicating no overtime).
To make the Overtime data useful it was changed to the NA inputs to zero and the “OT” inputs to 1 to create a factor variable with 2 levels that could be used for analysis.

Abbreviations It is worth changing the names of the teams to make it easier to read decision trees further along the analysis, the various team names are shortened with the abbreviation function.
Before being shortened the names are long and rather unwieldy particularly for the decision trees plots further along.

```
##  [1] "Arizona"         "Baltimore"       "Carolina"        "Chicago"
##  [5] "Cincinnati"      "Dallas"          "Denver"          "Green Bay"
##  [9] "Houston"         "Indianapolis"    "Jacksonville"    "Kansas City"
## [13] "Miami"           "New England"     "New York Giants" "New York Jets"
## [17] "Oakland"         "Philadelphia"    "Pittsburgh"      "San Diego"
## [21] "San Francisco"   "Seattle"         "Tampa Bay"       "Tennessee"
## [25] "Washington"

##  [1] "Arzn" "Bltm" "Chcg" "Cncn" "Crln" "Dlls" "Dnvr" "GrnB" "Hstn" "Indn"
## [11] "Jcks" "KnsC" "Miam" "NwEn" "NwYG" "NwYJ" "Okln" "Phld" "Ptts" "SnDg"
## [21] "SnFr" "Sttl" "TmpB" "Tnns" "Wshn"
```

The shortened names are 4 letter and much more manageable.

Character vaiables

Looking at the data again we can see that there some variables are recorded as `characters` instead of factors which is required for the analysis.using the `lapply`function all of the charcther variables are converted to the class `Factor`.

Score Difference To explore the features further I decided to try decipher some more variables score difference was the main one. I was hoping to see if there was any relationship between a close game and rising arrest rates. As tight game comes to an end I would imagine that tempers flare under such citcumstances.

Outliers The arrests data has a huge right skew to it with a very long tail. Looking at, it has the appearance of a Poisson distribution to it but when a test of the distribution is carried out the results say that because of the long tail it is unlikely that the distibution fits a Poisson model. It is important to note that the distribution is not Gaussian either.

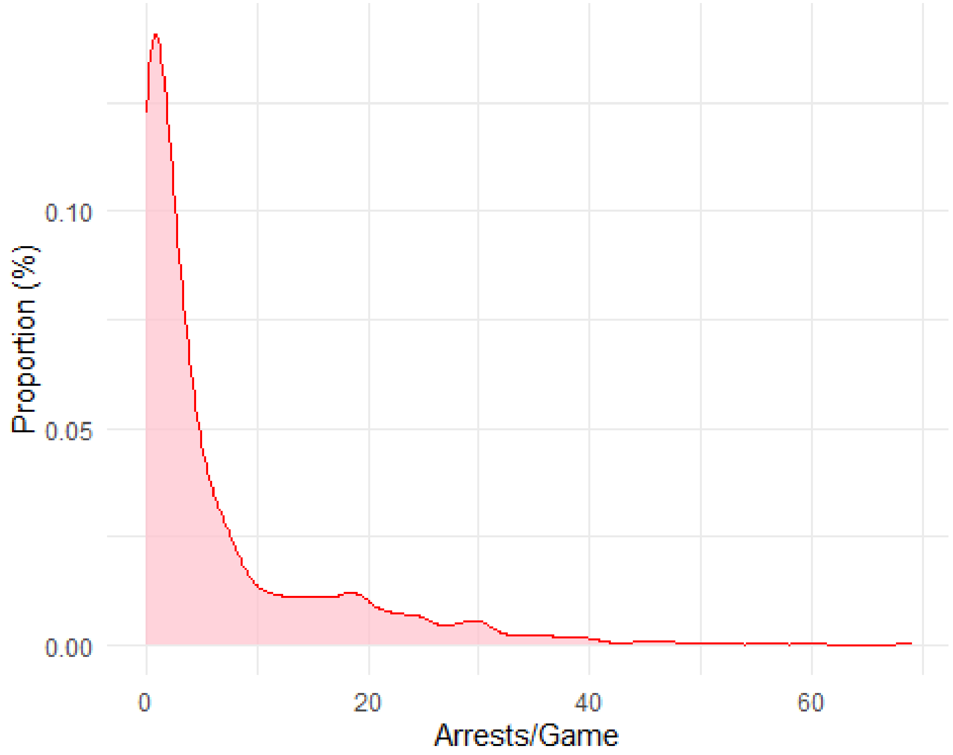


Fig. 2.1. Frequency of arrests/game

```
## Dispersion test of count data:
## 966 data points.
## Mean: 6.57
## Variance: 90.48
## Probability of being drawn from Poisson distribution: 0
```

It was tempting to remove the top one percent of the variables as outliers but on further inspection 5 out of the top 10 points are from the same team (San Diego Chargers), removing some of these may adversely affect the quality of the model. For this analysis no actions being taken as regards to removing outliers.

```
##  [1] SnDg SnDg Ptts Ptts Okln SnDg NwYG SnDg NwYJ SnDg
## 25 Levels: Arzn Bltm Chcg Cncn Crln Dlls Dnvr GrnB Hstn Indn Jcks KnsC ...
Wshn
```

2.2 Model Selection

The data is divided randomly into 2 sets; the training set(80%) and a test set(20%). The training set will be used to create the models and the test set then can be used to test the quality of the fit of the model. It will produce the test set error. Due to the score difference variable being created from the 2 other numeric variables it creates an error when running these functions so it is removed for until the decision tree methods.

```
## [1] 772  12

## [1] 194  12
```

2.3.1 Subset Selection

The 3 main methods of the subset selection are

Best Subset Selection: where all combinations of of each number of predictors are used.

Forward Stepwise Selection where starting at 0 predictors, the predoictor with the most additional improvement is added up until all predictors are included.

Backwards Stepwise Selection opposite to forward stepwise it starts with all predictors and removes the least useful each time.

The motivation for using stepwise selection is that while it doesn't compared every possible model which can be huge 2^n for n predictors. Looking at Fig. 2.2. which plots the Adjusted R^2 (the percentage of the response variable that is explained by the model) to the number of variables of the model all 3 models use the same predictors to create in the models with 8 identical predictor. Due to the number of factor variables creates a huge number of dummy variables makes it difficult to decipher the key factors but does show what dummy variables are important.

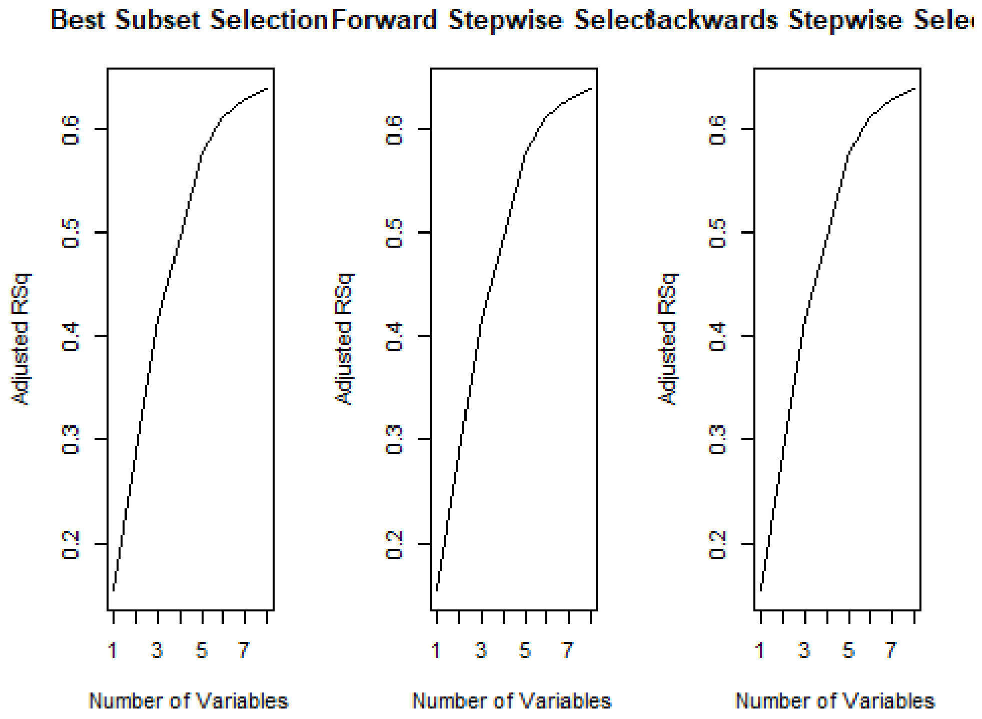


Fig. 2.2.

Using the `coef` function it identifies the coefficient for the model, looking at the model with 8 variables as it has the max R^2 for the forward subset selection.

2.3.2 The Validation Set

To compare the models that are developed, the root mean square error(RMSE) of a test set will be used as an indicator of the quality of fit of the model. Starting with the simplest, the validation set approach which will use the training and test set created earlier to build a model and the test set is used to the predict arrests which then can be compared to the actual arrests.

Using a training set and a test set to check the predictive performance of the models with differing numbers of predictors After testing the various models we can see that the 7th model had the lowest MSE of 34.97

2.3.4 k-fold Cross Validation

To try and imporve accuracy of the prediction the observations can be divided into k groups, or folds. The first fold is used to as a validation set for the remaining k - 1 folds which are used as a “training set”. This preoces is repeated with a different fold used as a test set until all folds have been used. The resulting MSE is calculated using the MSE from the average of MSE_1 to MSE_k. This gives an overall more accurate than the previous validation set due to the bias-variance trade off.

```
## Reordering variables and trying again:
```

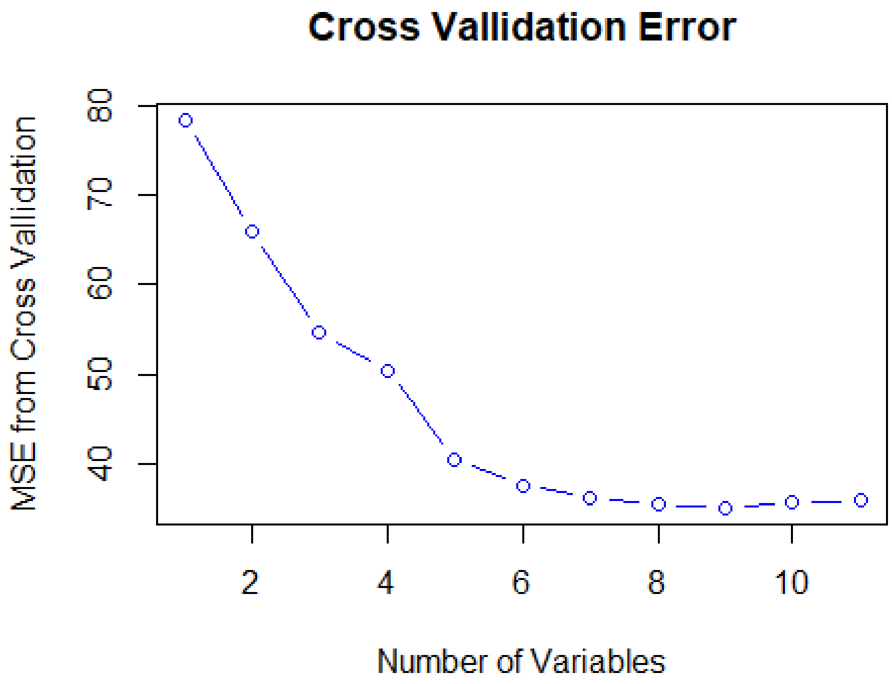


Fig. 2.3. Plots the number of varibales in the model against the average MSE. After cross validation the best model used 9 variables. All of these aside from the time of games were the dummy variables. Looking at the variables selected the majority are `home_team` variables with some `away_teamand` `gametime_local` included.

```
coef(reg.best, 9)

##    (Intercept) gametime_local   home_teamGrnB   home_teamNwYG   home_teamNwYJ
##      -3.768417      10.112168        4.545440       19.780062       19.247333
##  home_teamOkln   home_teamPtts   home_teamSnDg   home_teamSnFr   away_teamClvl
##      15.914983      14.419543       22.429560        9.095714       -3.363916
```

2.3 Ridge Regression

While similar to the least square models used previously the ridge regression it uses a shrinkage penalty to shrink coefficients towards zero, this shrinkage coefficient is set by the tuning parameter (λ) for the best accuracy. The observations again have to be split up into the training set and test set but in the form of a matrix for the ridge regression to be carried out. A large λ creates a penalty with a large effect and small λ creates a small shrinkage penalty.

A grid of 100 values ranging from 0.135 up to 22026 (plotted on the logarithmic scale) is created this will test model with 100 different λ and give the MSE for 100 different models enabling us to circle in on the best λ for the model. If we set the model up from just the intercept, the MSE would be {r intercept}, where as the average across all λ was {r mean.rr}. Fig. 7. graphs the log of λ on the x-axis with the resulting MSE on the y-axis.

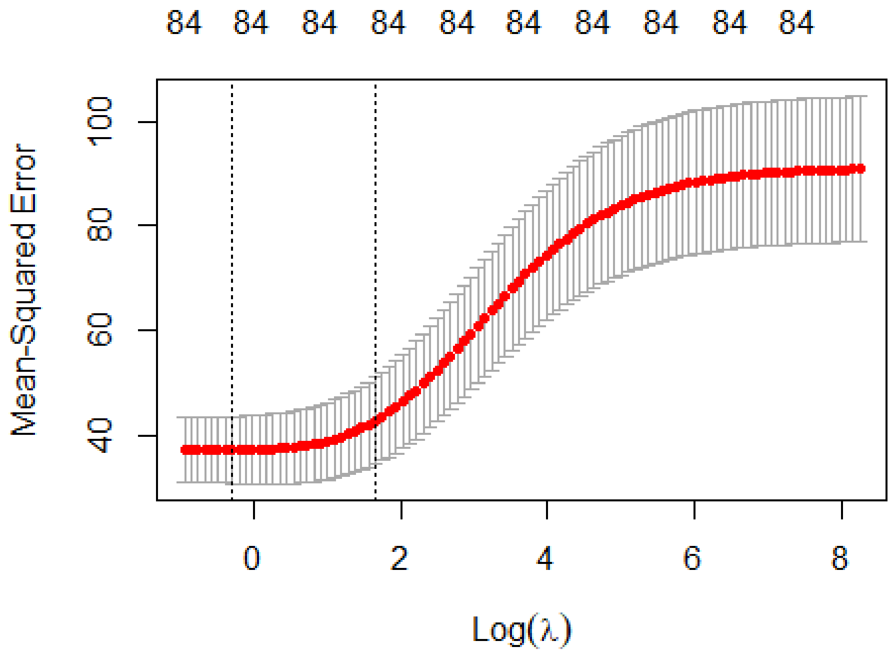


Fig. 2.4 Is a plot of the MSE at the different tuning parameter(λ) for this model it decreased with a smaller λ with the best being achieved at 0.74.

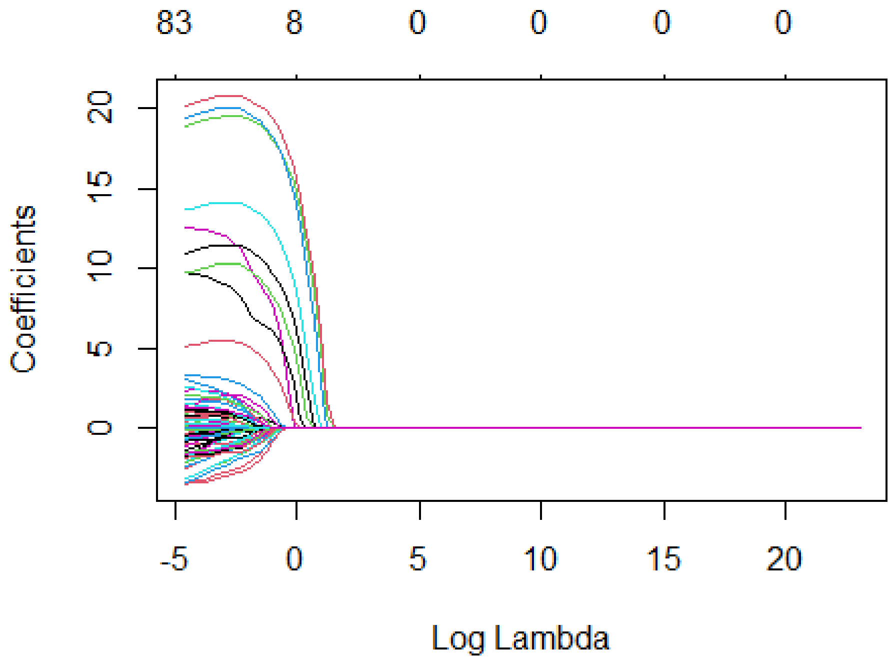


Fig.2.5. Each line in the plot is a predictor in the model as it approaches 0

2.4 The Lasso

The lasso method improves on ridge regression by shrinking coefficient estimators to 0 compared to ridge regression which will have all predictors included. That is still higher than the MSE from the ridge regression model, but it is still smaller than the null model

2.5 Principal Component Regression

Due to the majority of the variables in the `NFL_data`are factors not numeric it caused the PCR to not work well

2.6 Decision Tree

Stepping away from the linear models and using a decision tree to approach the prediction problem. “Decision trees consist of a series of splitting rules that segments the predictors in n regions” (James, G. *et al*, 2013). “Regression tree leaves contain constant values as predictions for the class value”(Richardson, B. *et al*, 2017). Decision trees can be easily interpreted which is a massive advantage but decision trees can be thought of as a black box (Breiman, 2002) meaning the between the input and the output there it can be hard to interpret what is variables are important.

The variable for score difference can be reintroduced as it can not cause a linear discrepancy in a decision tree. While adding another does not allow for a direct comparison of the tree based methods to linear models, it may offer superior prediction which is more important for this analysis.

2.7.1 Decision tree

```
## 
## Regression tree:
## tree(formula = arrests ~ ., data = NFL_data, subset = train)
## Variables actually used in tree construction:
## [1] "home_team"  "away_team"  "week_num"   "score_diff"
## Number of terminal nodes:  10 
## Residual mean deviance:  18.42 = 8714 / 473 
## Distribution of residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## -17.830  -2.077  -0.625   0.000   1.923  16.380
```

.

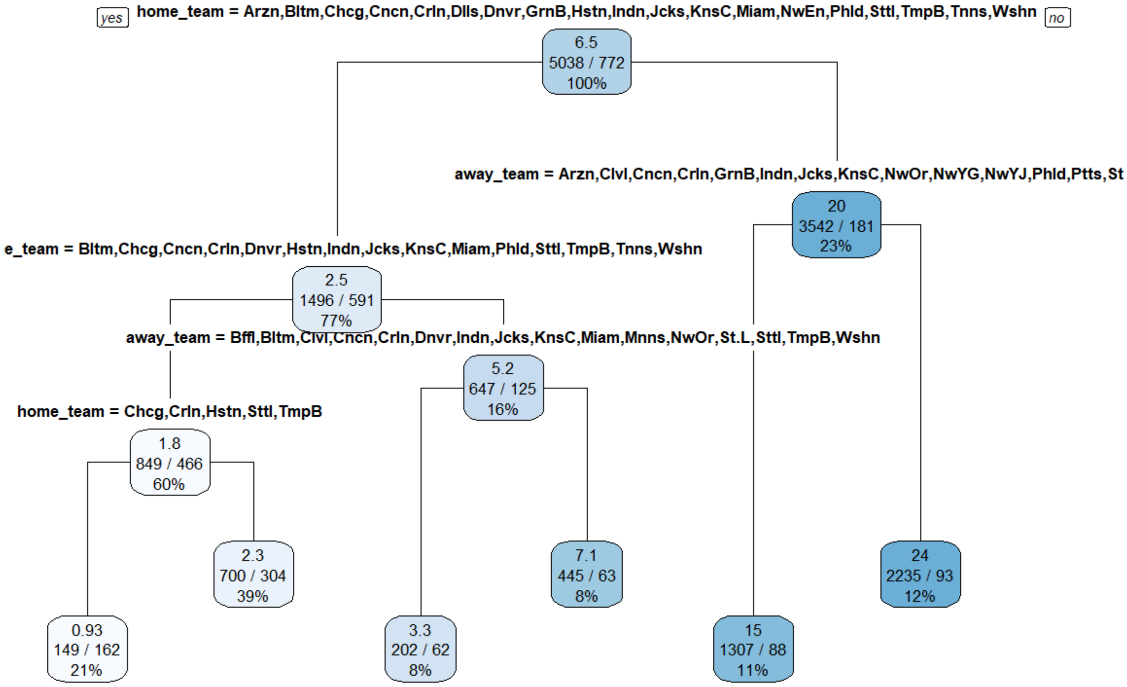

yes
home_team = Arzn,Bltm,Chcg,Cncn,Crln,Dlls,Dnvr,GrnB,Hstn,Indn,Jcks,KnsC,Miam,NwEn,Phld,Sttl,TmpB,Tnns,Wshn
no
6.5
5038 / 772
100%
away_team = Arzn,Clvl,Cncn,Crln,GrnB,Indn,Jcks,KnsC,NwOr,NwYG,NwYJ,Phld,Ptts,St
20
3542 / 181
23%
e_team = Bltm,Chcg,Cncn,Crln,Dnvr,Hstn,Indn,Jcks,KnsC,Miam,Phld,Sttl,TmpB,Tnns,Wshn
2.5
1496 / 591
77%
away_team = Bffl,Bltm,Clvl,Cncn,Crln,Dnvr,Indn,Jcks,KnsC,Miam,Mnns,NwOr,St.L,Sttl,TmpB,Wshn
5.2
647 / 125
16%
home_team = Chcg,Crln,Hstn,Sttl,TmpB
1.8
849 / 466
60%
2.3
700 / 304
39%
7.1
445 / 63
8%
24
2235 / 93
12%
0.93
149 / 162
21%
3.3
202 / 62
8%
15
1307 / 88
11%

Fig. 2.6 Is a single decison tree using 5 different variables in the prediction resulting in 11 terminal nodes. The inputs shift left or right depending on a boolean output

2.7.2 Bagging, Random Forest and Boosting

There are a number of methods to try and improve the accuracy of the decision trees through methods like

Bagging: This involved creating a distinct data set by sampling the original multiple times repplacing values. These data sets are used to create prediction models and these models are averaged to decrease the varience.

Random Forest: Works by only allowing a subset of predictors to create a decision tree over multiple(500) trees. This decorrelates the tree and makes the average of the resultuing trees less variable.

Boosting: Builds on the bagging method by growing trees sequentially using information from the previous trees

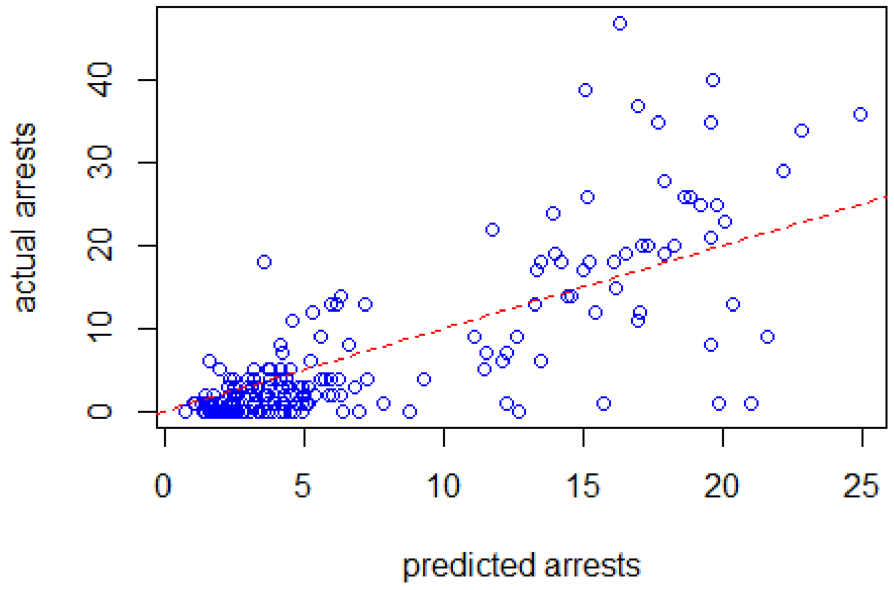
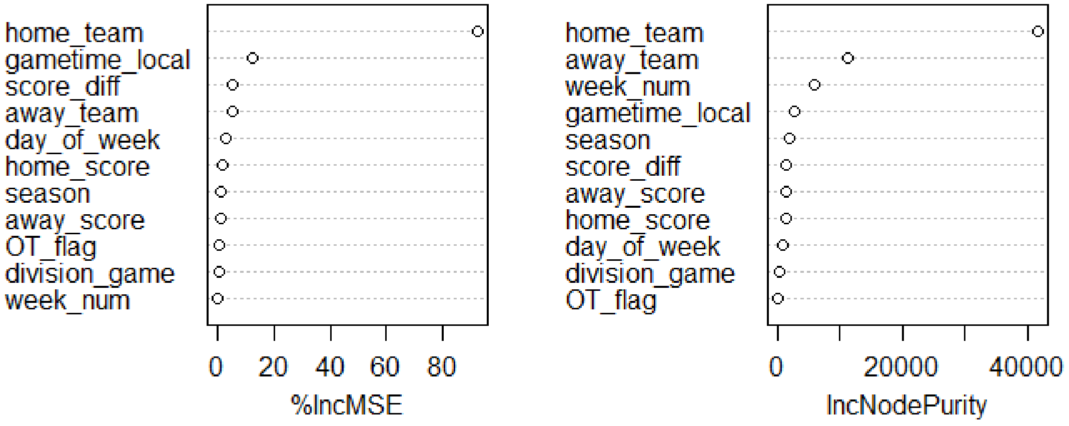


Fig. 2.7 is a a plot of the predicted number of arrests from a `Random Forest` on the x axis and the actual number of arrests on the y axis, the dashed line would be the correct prediction.

Fig. 2.8 Importance of each variable in the Random Froest(n=500)



2.7 Generalized Linear Models

Generalized Liner Models (GLMs) are a large class of statistical models that consist of three components

- A random component specifying the distribution of the response variable
- A linear predictor
- A link function

With these components GLMs attempt to accomodate a range of distribution types (Gaussian, Poisson etc.)(Venables *et al*, 2004). This is useful for this analysis as the arrests variable appears to have a poisson distrubtion or quasi-poisson, which we can include in the model.

Four different models are created to incorporate the factors that have consistently been included in models for both model selection and the decision tree methods. The root meean square error is calculated on each of the models to compare to previous models.

GLM 1: poisson distribtion with all variables

GLM 2: poisson distribution with the top 4 predictors from decision tree

GLM 3: quasipoisson distrubtion with the top 4 predictors from the decision tree

GLM 4: gaussian distribution with the top 4 predictors from decision tree

It is imporatant to note that 3 of the 4 predictors were consistanly in the model selection methods

3.0 Results

Looking back to the main objectives starting out this analysis the 2 main aims were to

1. Identify the factors that have the biggest impact on the nunmber of people arrested at an NFL game
2. Create a model that can predict the number of people arrested at an NFL game.

After applying a number of methods to the data along with a exploration into the data which can be found here https://github.com/alexdoyle115/C7081_Assignment/blob/main/C7081_assignment_EDA.rmd.

3.1 Key factors

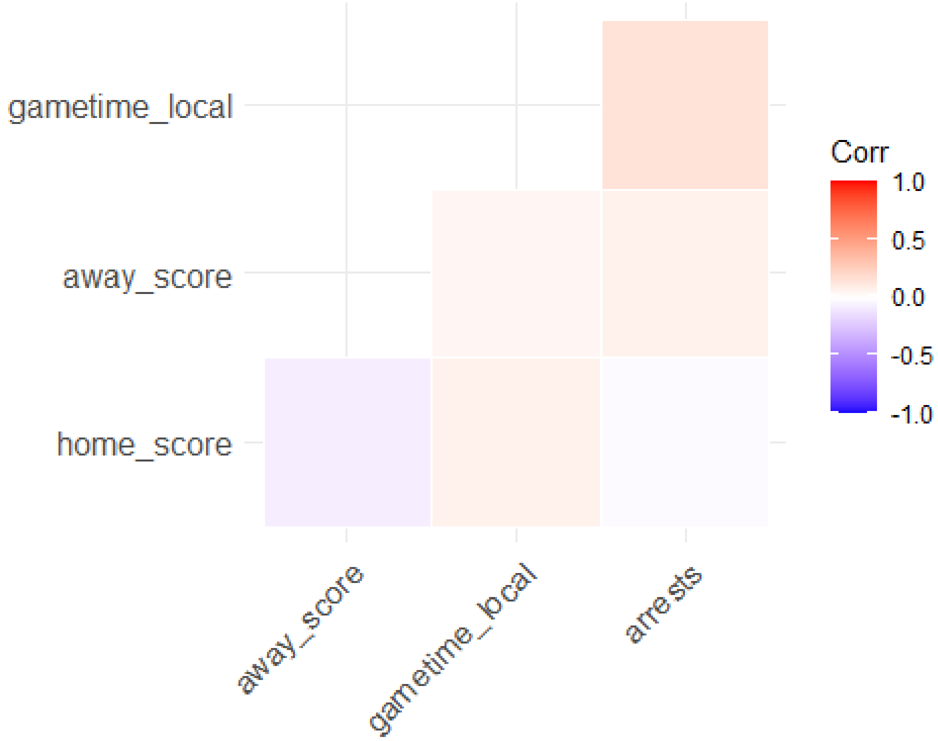


Fig. 3.1 Correlation plot

The methods used in this analysis are to create prediction models but the coefficients offer insight into what variables are heavily weighted. Looking at Fig. 3.1 it highlights that there is little correlation between the number of arrests and the other numeric variables. In this analysis there are 2 methods that can be used for selecting the main variables for importance are the model selection and the decision tree. Using model selection to pick out the important variables was problematic as mentioned due to the creating of dummy variables. Linear regression creates dummy variables to incorporate qualitative variables into to model. This method works well for small numbers of predictors but as in this case with 25 predictors for the `home_team` column it create a lot of dummy models which is difficult to decipher. However it does indicate the important teams and when the dummy variables are from the same predictor it does offer some insight.

The decision tree method for selecting variables has drawbacks also. Methods like `Random Forest` do not try to connect predictions to the attributes and even removing important predictors can have little impact on a models prediction accuracy (Efron, 2020).

Despite this there is a consistency across all of the methods. We can see from the dummy variables from subset selection that the `home_team`variable is very important for prediction but there is not much else that can be explained. In Fig. 2.11 it indicates the impact of each predictor on the Mean Square Error across the 500 iteration of the `Random Forest`, overall the this model accounts for 61% of the variance. Comparing this to Fig. 2.2. we see that the R^2 is just above 60% also. As we mentioned that these graphs are not a fool proof indicator but for the small number of predictors in this model and it being backed up by the GLM model for prediction the main predictors are

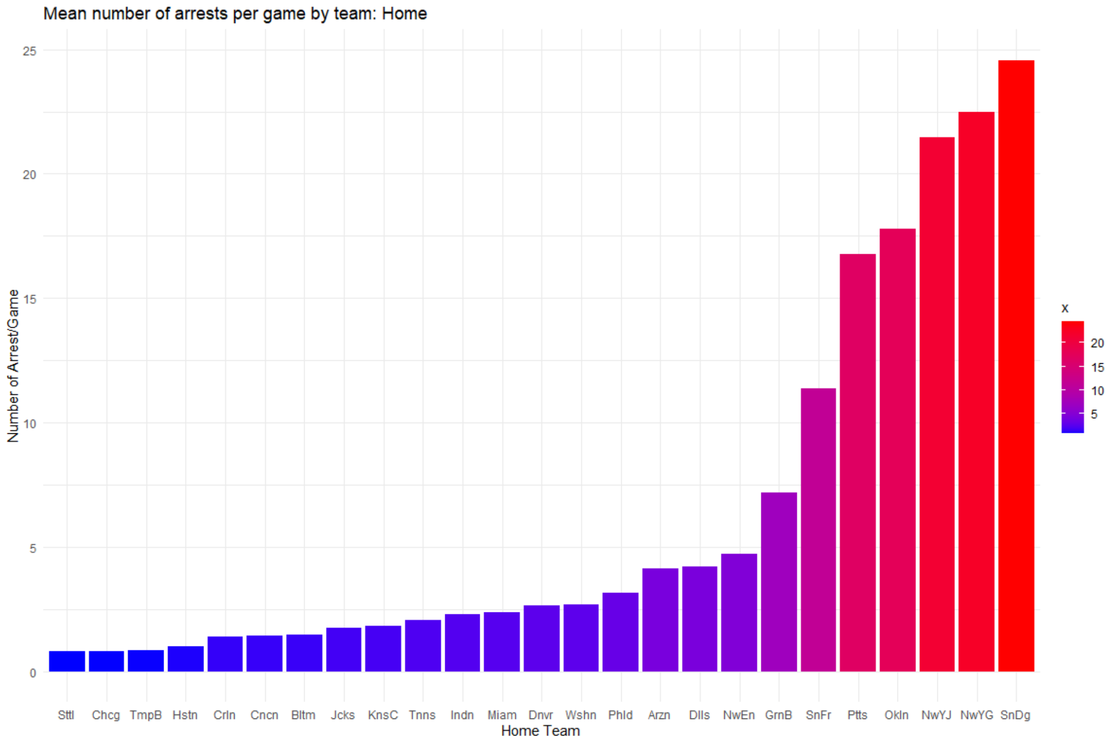


Fig. 3.2. Does show there is some variance across team when it comes to the number of arrests again wheater that due over zealous security or rowdy supporters can not be deciphered but it is the strongest predictor.

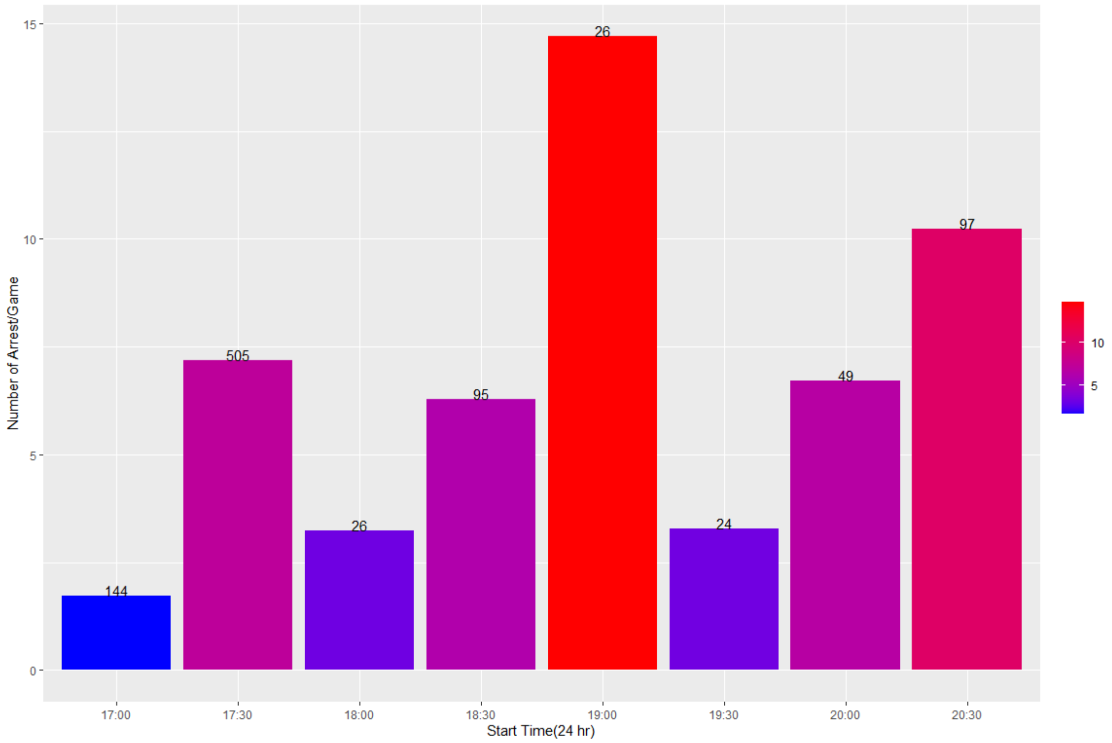


Fig. 3.3 Mean number of arrests of game time(Number of Kickoffs at this time).

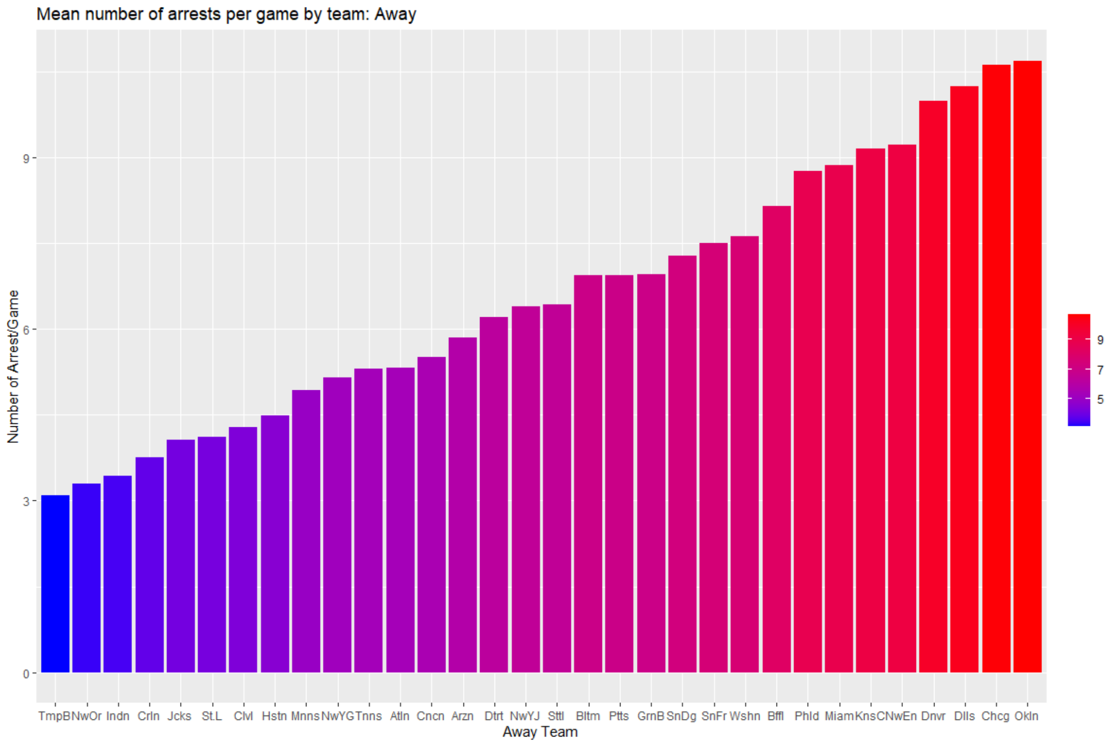


Fig. 3.4.Relationship between away teams and arrests

3.2 Predictions

The prediction aspect of the analysis was simpler to gain a quantitative result with the root mean square error (RMSE) of the predictions being used to select the most accurate model. Although the RMSE can be influenced by any outliers. Looking at Fig. 3.5 it displays the differing RMSE for each model. The main point to take away is the superior predictive ability of the deciision tree method (aside from `boosting`) The other important point is the GLM models built from the most important factors above have simliar very little change in RMSE which is unsurprising as the arrests did not fit any of the distributions. The GLMs also were much simpler to interpret.

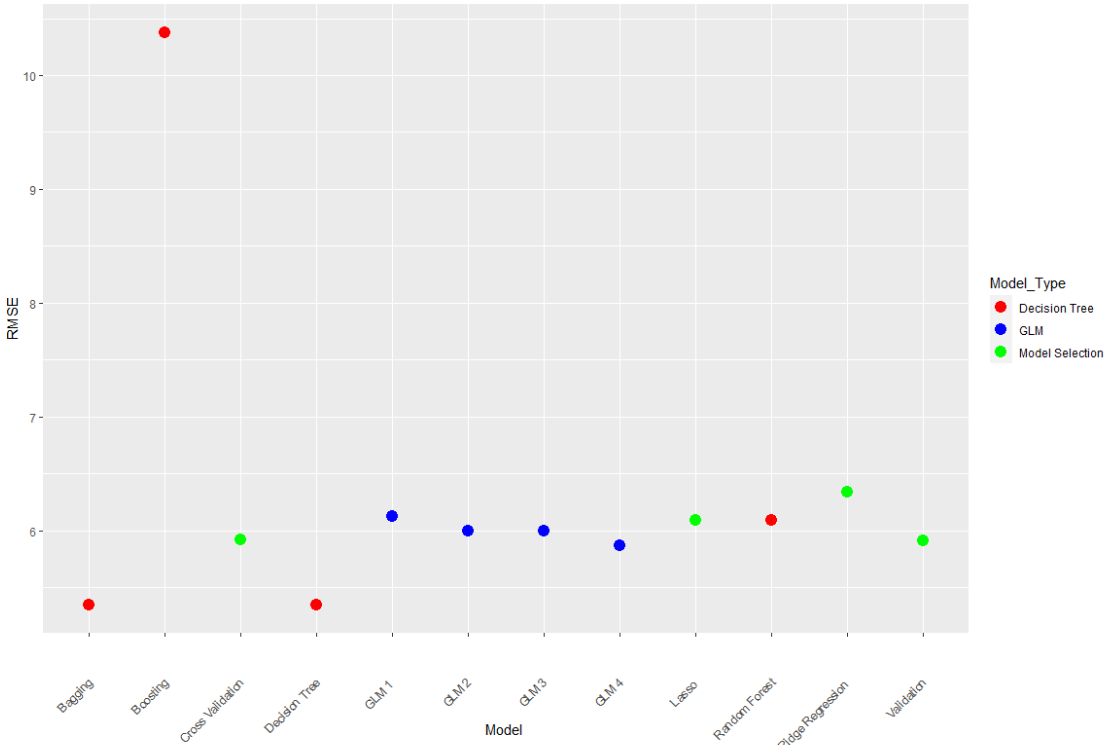


Fig. 3.5. Comparison of the different RMSE of the different models

4.0 Conclusion

The aims to identify the important factors and create a model to predict the number of arrests overall was reasonably successful. The home team was the most dominant factor by quiet some margin which is understandable when looking at the Fig. 3.1 and makes sense that, but this doesn't give any information into what actually causes this. The big down side of having a character variable as the main one it makes the linear model difficult to interpret and visualize. The GLMs while consistently good are out competed by the prediction quality of the decision tree based methods. Their simplicity and interpretability make up for this giving in my opinion the overall best model for the requirements. Approaching the problem again it would be worth trying to get a complete data set including all 32 teams and trying to examine the causation of the higher number of arrests. Unfortunately there was no new information regarding the results available to use as a testing set or create an updated model.

5.0 Litriture Cited

Richardson, B., Fuller-Tyszkiewicz, M., O'Donnell, R., Ling, M. and Staiger, P.K., 2017. *"Regression tree analysis of ecological momentary assessment data."* Health Psychology Review, 11(3), pp.235-241.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. *"An introduction to statistical learning"* (Vol. 112, p. 18). New York: springer.

Venables, W.N. and Dichmont, C.M., 2004. *"GLMs, GAMs and GLMMs: an overview of theory for applications in fisheries research."*, Fisheries research, 70(2-3), pp.319-337. article:

Bradley Efron (2020) *"Prediction, Estimation, and Attribution, Journal of the American Statistical Association"*, 115:530, 636-655, DOI: 10.1080/01621459.2020.1762613