**Predicting Fatal Traffic Accidents Involving a Distracted Driver**

**Springboard Capstone Project Final Report**



**INTRODUCTION**

Traffic accident related deaths unexpectedly increased by 7.2% in 2015, following a 20 year declining trend. In August 2016, the National Highway Traffic Safety Administration along with the Whitehouse put out a call to action to investigate 2015 US traffic fatalities data.

This project focuses on distracted driving related traffic fatalities. Smartphone-related distracted driving is one potential contributing factor for increased traffic fatalities in 2015 and car accidents in general, as smartphone ownership has quadrupled since 2010. More than 207 million Americans now own a smartphone. Although it is illegal to use a cell phone while driving in many states, drivers continue to text and talk and drive.

Moreover, recent advancements in vehicle technology combined with machine learning algorithms and access to large datasets are revolutionizing the car insurance industry. These advances allow for innovative analyses that could predict certain types of traffic fatalities in addition to what factors are associated with these specific types accidents. Furthermore, car insurance companies are always competing for new customers, and with these advancements in big data and vehicle technology – namely the ability to collect data directly from “smart” cars and to access national data on traffic accidents - these companies are innovating and changing their policies. Now that these companies can use data to design and offer different insurance plans based on driving habits, they can reward safe drivers with lower premiums and charge more to drivers who exhibit less safe driving habits. Moreover, with access to all sorts of metrics such as mileage, fuel consumption, breaking patterns, speeding, time of day, weather conditions, and more, analysts are practically able to recreate the scenario of a car crash, and in turn, to assess liability in accidents. Data science is and will continue to change the car insurance industry.

**PROJECT OBJECTIVES**

The project aims to predict

(1) factors associated with fatal accidents involving distracted drivers and

(2) the actual incidence/occurrence of fatal accidents involving distracted drivers.

**DATASET**

The primary dataset used for this project is the 2015 US Traffic Fatalities dataset, which can be downloaded here:

<https://www.kaggle.com/anokas/2015-us-traffic-fatalities>

The Traffic Fatalities data released by the National Highway Traffic Safety Administration in conjunction with the U.S. Department of Transportation each year is comprised of three main datasets: *Accidents, Persons,* and *Vehicles*. This project focuses on the *Accidents* dataset.

The *Accidents* dataset has 38 columns (variables) and 32,167 rows, each representing a traffic accident that involved at least one fatality. The dataset includes a case number for each accident as well as information about the accident. The data is accessible in CSV format.

**ACC\_AUX.csv - Dataset of 2015 accidents (38 variables,  >32,000 rows of data)**

Limitations of the datasets

Following the initial exploration of each dataset, it was determined that there are multiple rows per case in both the *Persons* and *Vehicles* datasets that make it somewhat difficult to link them to the Accidents dataset cleanly. Unfortunately, this means that demographic information (apart from age) or vehicle information cannot be linked to the distracted driver fatality events while keeping one row of data per case/accident.

**INITIAL DATA CLEANING AND WRANGLING**

Since the data were released by the National Highway Traffic Safety Administration and the U.S. Department of Transportation, they were very tidy and clean from the outset. The data were already arranged in columns, with cases as rows.  All of the data in the *Accidents* dataset were discrete, apart from the number of fatalities involved in the accident. In many cases the variables included an “unknown” category. Because this impacted quite a number of variables, I decided to include the “unknown” categories in all analyses.

Nearly all of the variables in the *Accidents* dataset were discrete integers (apart from the number of fatalities). All integers were recoded into factor variables for the initial data exploration.

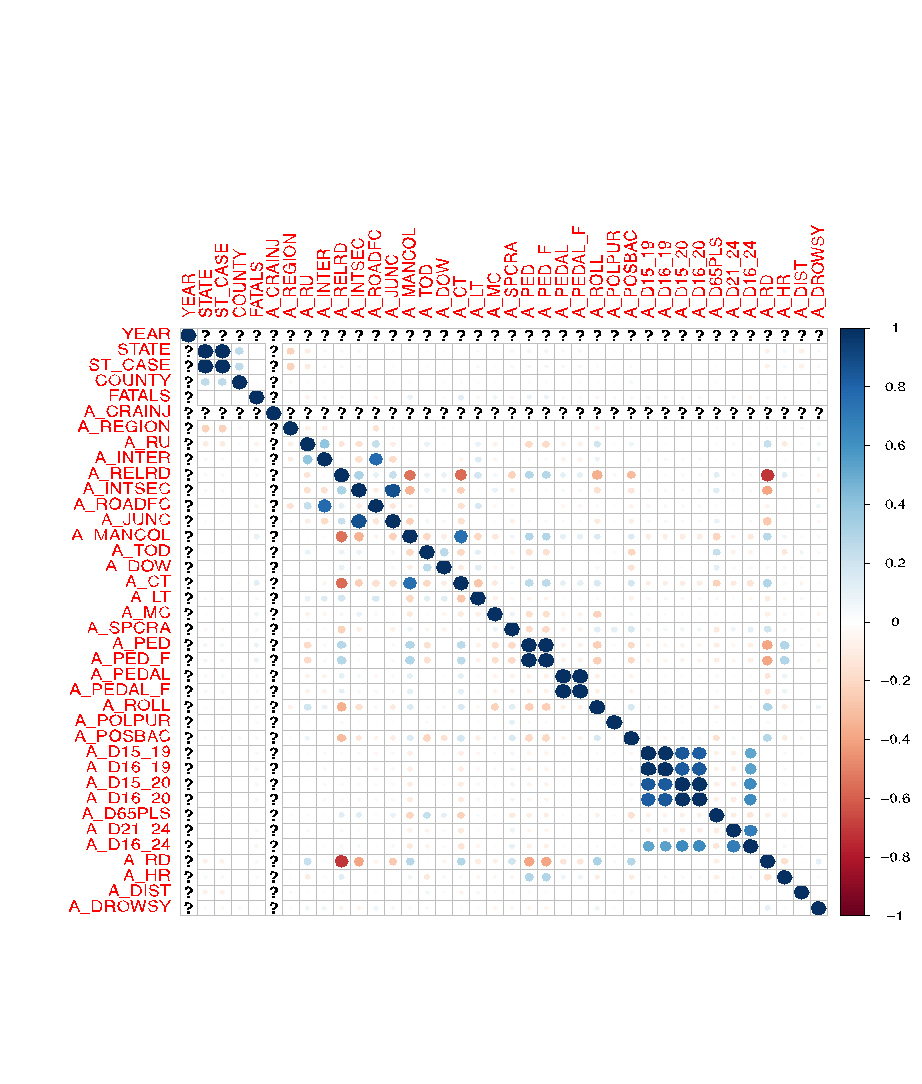
Even though rigorous data cleaning was not necessary for this particular dataset, there are some noteworthy techniques that could be use in the future for data cleaning/tidying needs. For example, missing data/null values in a continuous variable could be replaced with the mean or median of that variable. Categories without many cases from a discrete variable could be grouped into fewer categories. If the data was not available in a column format, it could be reshape. New variables could be made by uniting or separating variables; many other things can be done to clean and tidy the data with the *dplyr* and *tidyr* packages in R.

**EXPLORATORY DATA ANALYSIS**

Correlations

Correlations were performed on all the variables in each of the three datasets using the *corrplot* package in R.  Since I was interested in exploring specific uncommon outcomes like distracted driving and drowsy driving, for example, the correlation plots provided me with initial ideas about which variables I might want to include as part of other initial data exploration, such as bar plots, and eventually modelling.

Correlations from the *Accidents* dataset



From this initial correlation matrix it is clear that while not many variables are strongly correlated with distracted driving., there are still a few correlations. State is positively correlated with distracted driving fatalities while time of day is negatively correlated with distracted driving fatalities. There is at least some variation in distracted driving fatalities that can be explored further.

Histograms

Histograms allow us to see the distribution of each categorical variable in the dataset. A selected group of histograms is presented here, beginning with our outcome variable.

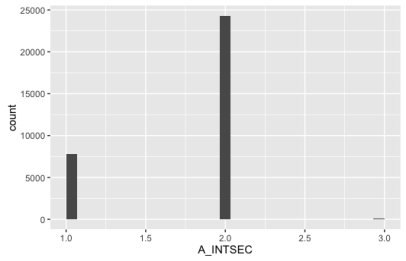
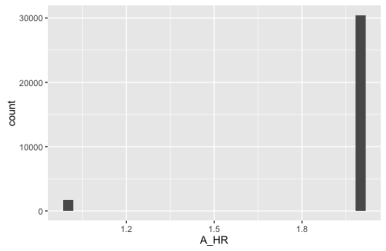
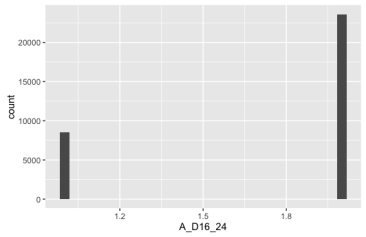
Distracted Driving

|  |  |  |
| --- | --- | --- |
| Distracted Driver | N | % |
| Yes | 3,196 | 9.9% |
| No | 2,8970 | 90.1% |



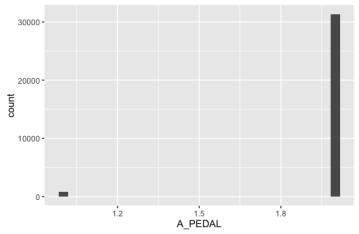
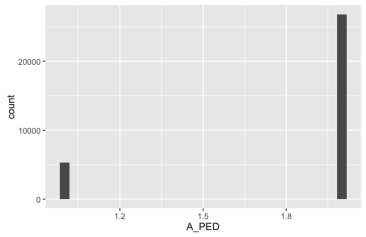
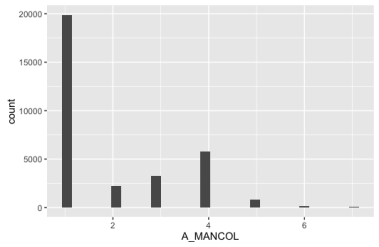
It is important to note here that the distribution of the outcome variable, traffic fatalities involving a distracted driver, is highly unbalanced in this dataset, with an overall incidence of less than 10%. However, there is still a good number of observations; there were 3,196 traffic fatalities where distracted driving was reported.

Age under 25 Hit and Run Intersection



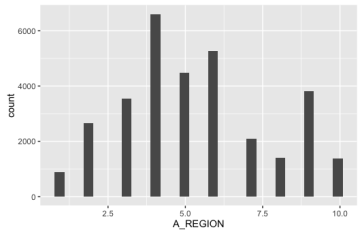
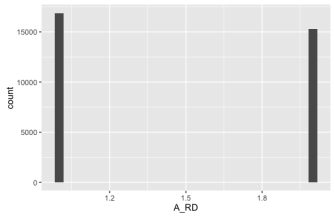
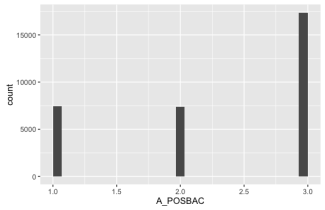
From the three histograms shown above, it clear that a relatively large proportion of fatal accidents in 2015 involve drivers under 25 years of age, as well as intersections. From this I want to investigate whether distracted driving fatalities were more common in younger drivers. I would also like to see whether distracted driving fatalities occur more often at intersections.

Manor of Collison Involving a Pedestrian Involving a Cyclist



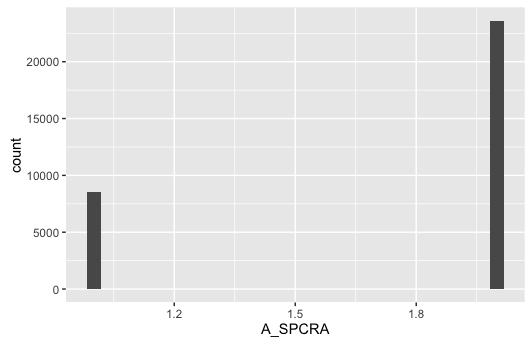
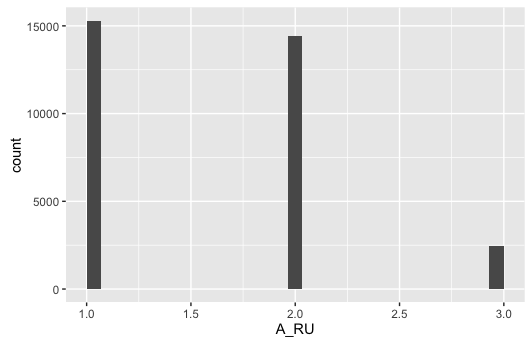
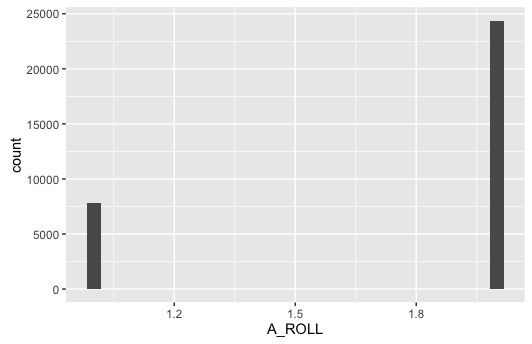
Manor of collision is a very interesting variable. I was somewhat surprised to note that most fatal accidents do not involve another vehicle, that is, they are single vehicle accidents. Despite this, rear-end, head-on, angle, and sideswipe collisions accounted for thousands of fatal accidents. I suspect distracted driving collisions are included here. I am also interested to explore whether distracted drivers are more likely to hit and kill pedestrians and cyclists.

Involved alcohol Roadway departure     Region



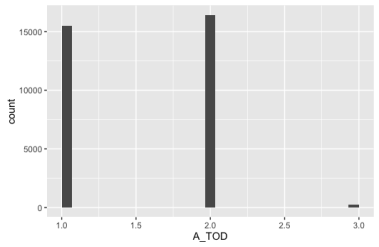
I am surprised by the number of fatal accidents that involve alcohol. I am also surprised at the number of accidents where the influence of alcohol is unknown. Roadway departure is also very common in fatal accidents, while fatal accidents appear to vary across region, with the most fatal accidents occurring in the southern states of Alabama, Florida, Georgia, South Carolina, and Tennessee.

Rollover Rural or Urban Involving speeding



Vehicle rollover and speeding also appear to be involved in a significant number of fatal accidents in 2015. Fatal traffic accidents occurred in urban and rural areas at approximately the same rate. The same is true for daytime and nighttime fatal accidents (see below).

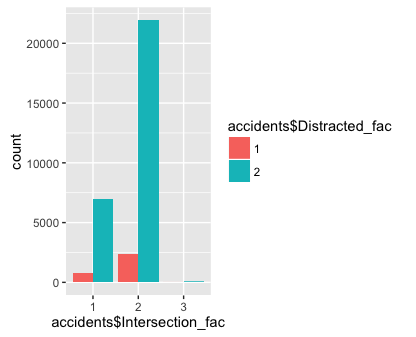
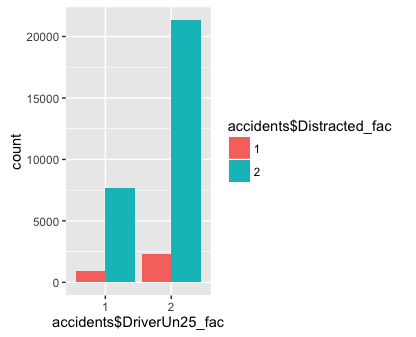
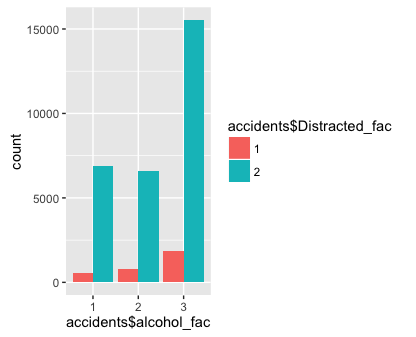
Time of Day



Bar plots with 2 variables

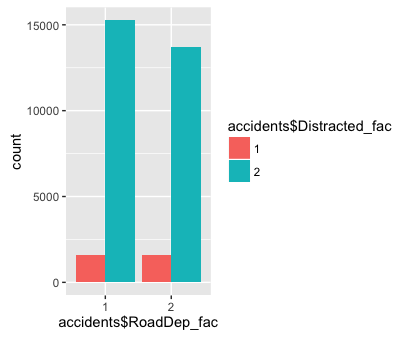
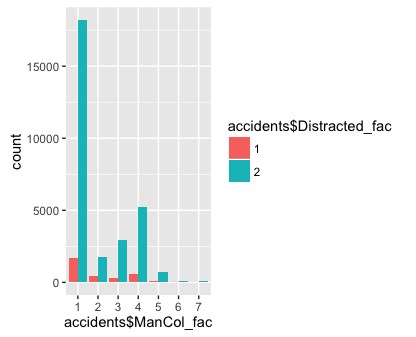
Bar plots allow us to examine the relationship between key variables and potential outcome variables. Here are some examples of bar plots with the distracted driver variable.

Alcohol Age under 25 Intersection



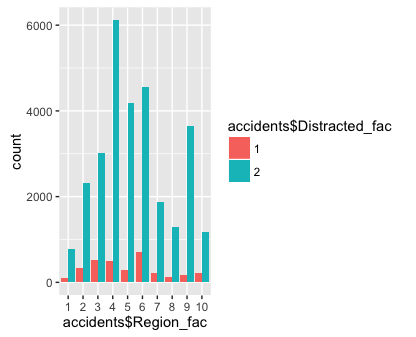
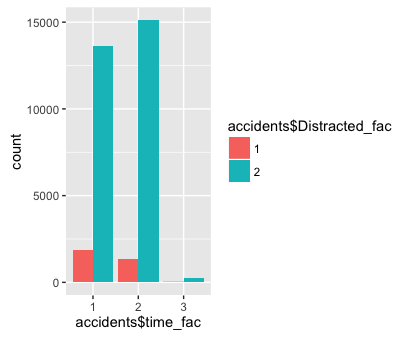
From these bar plots we begin to see some interesting relationships. Based solely on raw counts, distracted driving related traffic fatalities were less likely to involve alcohol or an intersection, and more likely to involve drivers age 25 and older. This was counter-intuitive to my initial thought that perhaps younger people were texting and driving. Perhaps it’s actually older drivers.

Manor or collision Roadway Departure



Distracted driver fatalities happened across all manors of collision and equally involved a roadway departure vs not.

Time of Day Region



Finally, it is interesting to note that more distracted driver fatalities occurred during the daytime as opposed to nighttime. Distracted driving fatalities varied widely across region, with the most taking place in region category 6, which includes Louisiana, Mississippi, New Mexico, Oklahoma, and Texas.

**MODEL BUILDING PROCESS**

Approach to Modeling

The exploratory data analysis confirmed that distracted driving deaths vary by different factors. The data are non-linear, as most of the variables are categorical and the outcome is also a discrete variable with two categories (fatal traffic accident that (1) involved a distracted driver vs. (2) did not involve a distracted driver), the data are non-linear in nature. Therefore, a logistic regression model and a random forest model are the most appropriate models for predicting a fatal accident involving a distracted driver. Furthermore, producing both models allows me to assess which is the most accurate model for prediction.

Training and Testing Datasets

For both the logistic regression and the random forest final models, training and testing datasets were randomly selected using a 70% to 30% ratio respectively. Each model was trained on a 70% sample of the dataset and subsequently tested on a 30% sample of the dataset.

**LOGISTIC REGRESSION MODEL**

This section describes the process to build a logistic regression model that predicts the likelihood of a traffic fatality involving a distracted driver, and that only included variables that had a statistically significant relationship with distracted driving fatalities. To begin I produced a logistic regression model that included 20 independent variables to predict the likelihood of a fatal accident involving a distracted driver (yes, no). These 20 variables and their categories are listed here:

* Alcohol involved (yes, no, unknown)
* Driver under age 25 (yes, no)
* Day of week (weekday, weekend, unknown)
* Hit and Run (yes, no)
* Interstate (yes, no, unknown)
* Intersection (yes, no, unknown)
* Large Truck (yes, no)
* Manor of collision (Not collision with another motor vehicle, rear-end, head-on, angle, sideswipe, other, unknown)
* Motorcycle involved (yes, no)
* Pedestrian involved (yes, no)
* Bicyclist involved (yes, no)
* Roadway departure (yes, no)
* Region (10 categories)
* Rollover (yes, no)
* Rural/Urban (rural, urban, unknown)
* Speeding (yes, no, unknown)
* Time of day (daytime, nighttime, unknown)
* Driver over age 65 (yes, no)
* Drowsy driver (yes, no)
* Road class (interstate, freeway and expressway, principal arterial, minor arterial, collector, local, unknown)

Choosing Variables for Inclusion & Checking for Collinearity

There were two main steps to reduce the number of variables in the model. First, all the variables that did not have a statistically significant relationship with distracted driving fatalities were removed from model. This reduced the number of independent variables from 20 to 9. Second, Variance Inflation Factor tests were performed to ensure that none of the remaining variables were collinear, and thus operating to destabilize the model.

In order to run Variance Inflation Factor tests all the remaining variables had to be recoded into numeric dummy variables; otherwise they would not run in the tests. This also allowed for all the remaining variables to be simplified by eliminating categories that were not significant in the first model. For example, initially the “region” variable had 10 categories, but only 6 of them were statistically significant predictors of distracted driving in the first model. Therefore, the statistically significant region categories were coded into six dummy variables, while four others were eliminated altogether.

Finally, using the *car* package in R, the recoded dummy variables were evaluated for collinearity. None of my remaining variables were collinear; therefore all these remaining recoded variables were included in the logistic regression model.

**Variance Inflation Factor Testing: Coding and Output**

Independent Variable Coding for Variance Inflation Factor tests

ManCol\_NoOthCar

motorcycle\_Yes

RoadDep\_Yes

Roll\_Yes

speeding\_Yes

DriverUn25\_Yes

Alcohol\_Yes

daytime

ManCol\_RearEnd

Region\_midAtlantic

Region\_SouthEast

Region\_NorMidwest

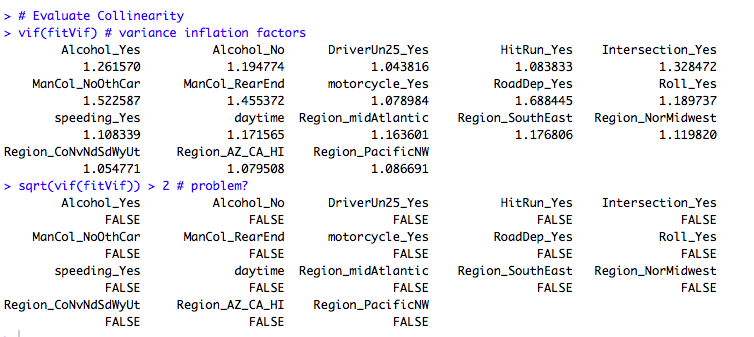
Region\_CoNvNdSdWyUt

Region\_AZ\_CA\_HI

Region\_PacificNW

Variance Inflation Factor Testing Output

This shows that none of the remaining variables were overly collinear. This particular test recommends excluding a variable if the square root of the model fit of the variable is larger than 2.



Feature Engineering

Feature engineering is a technique to make features out of variables that are more flexible in order to simplify the model and improve its predictive accuracy. Interactions are a type of feature engineering that is possible with categorical variables. Four or five interactions were hypothesized and tested, and one, between daytime and rear end collisions, was found to slightly improve the McFadden R-squared value of the overall model, and was therefore included.

**FINAL LOGISTIC REGRESSION MODEL AND RESULTS**

Based on process of refinement described previously, a logistic regression model was built and trained as follows:

Distracted\_driver ~ ManCol\_NoOthCar+motorcycle\_Yes+RoadDep\_Yes+

Roll\_Yes+speeding\_Yes+DriverUn25\_Yes+Alcohol\_Yes+

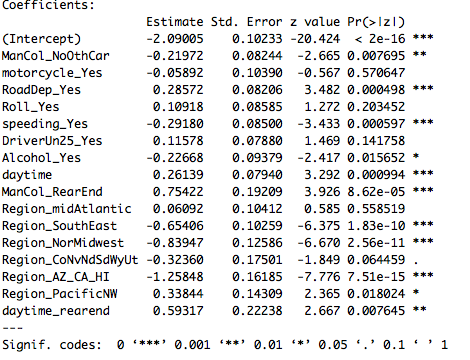
Daytime+ManCol\_RearEnd+daytime\_rearend + Region\_midAtlantic+Region\_SouthEast+Region\_NorMidwest+

Region\_CoNvNdSdWyUt+Region\_AZ\_CA\_HI+Region\_PacificNW

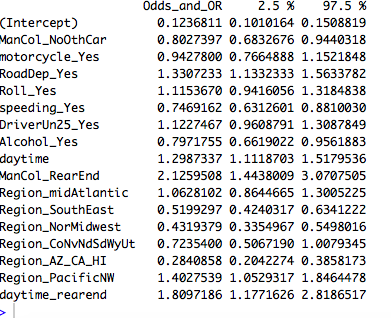
**Test Data Results**

It is clear from the logistic regression results that the odds of distracted driving fatalities vary by different factors. The test data confirm that traffic accident fatalities have higher odds of being due to distracted driving in the Pacific Northwest, and lower odds in the Southeast and North Midwest states as well as Arizona, California, and Hawaii. Moreover, distracted driver fatalities are more likely to involve another vehicle, and less likely to result in a roadway departure, involve speeding or alcohol. Distracted driving related deaths are also more likely to happen in the daytime and to involve a rear end collision.

Testing Data Model Coefficients Summary



Testing Data Model Odds Ratio Summary



**Logistic Regression Model Accuracy**

The subsequent McFadden pseudo R-Squared values on both the testing and training dataset were as follows:

Final Logistic Regression Model McFadden pseudo R-squared

(training dataset)

***0.04681***

(testing dataset)

***0.054700***

The final logistic regression model was only able to explain approximately 4 to 5% of the variation in distracted driving fatalities. This is a fairly poor model.

Moreover, a confusion matrix was used to test the model’s accuracy in predicting distracted driving fatalities. While the model was very accurate at predicting non-distracted driving, it had a 0% accuracy in predicting distracted driving fatalities.

Therefore, building a random forest model was necessary to test whether predicting distracted driver fatal accidents with better accuracy was possible.

**RANDOM FOREST MODEL**

A random forest model was developed using the same dummy variables (apart from the interaction) that were included in the logistic regression model, as using logistic regression is regression is one method to choose variables for inclusion in a random forest model. The random forest model was modelled as follows;

Distracted\_driver ~ ManCol\_NoOthCar+motorcycle\_Yes+RoadDep\_Yes+

Roll\_Yes+speeding\_Yes+DriverUn25\_Yes+Alcohol\_Yes+

Daytime+ManCol\_RearEnd+Region\_midAtlantic+

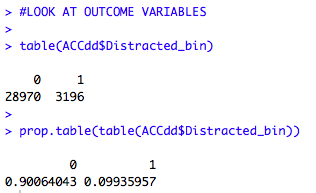
Region\_SouthEast+Region\_NorMidwest+

Region\_CoNvNdSdWyUt+Region\_AZ\_CA\_HI+Region\_PacificNW

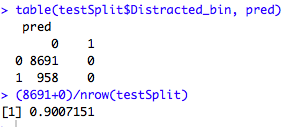
**Model Accuracy and the Accuracy Paradox**

Once again the model was rained and tested on 70/30 split of the sample and once again the model seemingly performed very well on the test dataset. Unfortunately, this was yet another case of accuracy paradox, whereby the model accurately predicts the underlying unbalanced distribution of distracted driving cases by ***never*** predicting a traffic fatality involving a distracted driver. As can be seen below, less than 10% of fatal traffic accidents involved a distracted driver. This is a very skewed and unbalanced dataset. The model learns that it can produce a nearly 90% accuracy rate by always predicting that an accident does not involve a distracted driver. With zero true positive predictions, it was necessary to try a different strategy; hence, an R package called *smote* was implemented.

Actual Distracted Driver Events from the Dataset



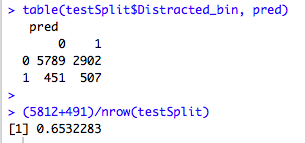
Predicted Distracted Driver Events and Model Accuracy without *smote*



Note that while it has a 90% accuracy rate, the model predicts all negative outcomes. This is not very useful

Predicted Distracted Driver Events and Model Accuracy using *smote*

The R package, *smote*, over samples rare events in a dataset that occur less than 15% of the time. The distribution above indicates that distracted drivers were involved in less than 10% of traffic fatalities in the 2015 dataset.



After resampling the data using *smote*, the model was rerun using on the training and testing data. After utilizing *smote* the model’s accuracy for predicting traffic fatalities involving a distracted driver was tested again with a confusion matrix. While the model now only has a 65% overall accuracy rate, more importantly, it predicts distracted driving fatalities. This is much better than making a random guess, where one could only hope to achieve an accuracy rate of 50% at best. Smote worked and this is a big improvement.

**65% OVERALL ACCURACY WITH SMOTE**

|  |  |  |
| --- | --- | --- |
|  | **Predicted** | |
| **Actual** | 0 | 1 |
| 0 | 5789 | 2902 |
| 1 | 451 | 507 |

Taking an even closer look at how the model performed on the testing dataset shows how accurate it is by quadrant.

|  |  |  |
| --- | --- | --- |
| **Actual** | **Predicted** | **Rate** |
|  | True positive | 1 predict 1 |
| 958 | 507 | 53% |
|  | False positive | 0 predict 1 |
| 8691 | 2902 | 33% |
|  | True negative | 0 predict 0 |
| 8691 | 5789 | 67% |
|  | False negative | 1 predict 0 |
| 958 | 451 | 47% |

While the 65% overall accuracy rate is a good start, there are some concerns with the model, namely, falsely predicting that 2,902 traffic fatalities involved a distracted driver.

**CHOOSING A MODEL**

Given the fact that both the logistic regression model and the non-smote sampled random forest model both displayed severe accuracy paradox whereby they predicted no true positives whatsoever, the *smote* sampled random forest model that predicts distracted driving fatalities with 65% overall accuracy is the best model. This model should be recommended to car insurance companies.

**RECOMMENDATIONS AND INSIGHTS**

From this preliminary study predicting distracted driving fatalities, car insurance companies could use this model

1. *to predict whether a fatal crash* ***did not*** *involve a distracted driver.*

Reasoning: Evidence like this could help reduce the burden of liability and potentially save the driver and insurance company money in a settlement.

1. *to develop more innovative, individualized, and competitive insurance premiums*

Reasoning: Earlier findings from the study identified some subgroups of drivers as well as other conditions that may be associated with a lower likelihood of distracted driving accidents. These targeted premiums, which need to be further informed by more models with other outcomes, should incentivize safe driving (through lower premiums), which could also help reduce traffic fatalities. While insurance companies once justified higher premium rates for young drivers based on research demonstrating a higher risk for accidents, targeted premiums are the next innovation in car insurance industry pricing.

It is important to note here that an insurance company would not want to risk falsely categorizing a fatal accident as involving a distracted driver (false positive), as this is a serious allegation and sensitive topic. As the model currently stands, unfortunately it too risky for an insurance company to predict a distracted driving fatalities with adequate confidence, as the model had a false positive rate of 33%. This brings us to future study and research that could improve the model.

**FURTHER MODEL DEVELOPMENT AND FUTURE RESEARCH**

There are a number of ways the model and the overall project could be improved:

1. **Combine multiple years** **of data** (e.g. 2014 and 2015) and build, train and test the model on this larger dataset. This could strengthen the model and improve its accuracy.
2. **Include more variables.** This model’s level of accuracy (65%) was achieved using only information from the *Accidents* dataset. It would be worth exploring whether the variables from the *Persons* and *Vehicles* datasets could inform and improve this model. One way to begin investigating this would be to join distracted driving cases into the *Persons* and *Vehicles* datasets to do some exploratory data analysis examining how factors in those dataset that are not available in the *Accidents* dataset, such as vehicle model and year as well as demographic information, relate to distracted driving fatalities. A limitation is that if an accident involved more than one vehicle, there is not sufficient information in the dataset to know which person in the accident was the distracted driver.
3. **Produce more models predicting other outcomes** (e.g. Age under 25, age over 65, alcohol involved, specific states, drowsy drivers, involving motorcycles, etc… An insurance company could use the information from all of these separate models to offer more competitive premiums broken down by all sorts of factors and subgroups, such as: age, state, etc.
4. **Develop more feature engineered variables**, as this can really improve a model’s accuracy.