**Fatality Analysis of Motor Vehicle Crashes in U.S. States [2010 – 2015]**

Logo

Description automatically generated

Team 4

Phimonkae Wilairat

Vidhya Lakshmi Vaithilingam Palanimurugan

ALY 6015: Intermediate Analytics

Instructor: Dr. Roy Wada

August 21, 2022

**INTRODUCTION**

The project is interested in studying factors that affect the number of fatalities in a crash and using related variables to create a model that could accurately predict the number of fatally injured persons. This project uses EDA to catch hold of as many insights from the data set. By investigating data, we could see data trends or patterns and the relationship between variables with the help of summary statistical testing and graphical representations. We expect to discern the remarkable trend of several important sub-category factors, which are consistent with the changes in fatalities. Ultimately, studying related variables leads to choosing appropriate predictors for the predictive model estimating the number of fatalities. Our group applied analytic methods with FARS data set to understand the given dataset, draw meaningful conclusions from the data by asking various research questions and create a prediction model to answer the questions.

**DATASET DESCRIPTION**

The source of FARS data is U.S. Department of Transportation and the National Highway Traffic Safety Administration. Every crash recorded in this data set is defined using 58 different features out of which 34 are quantitative variables and the remaining 24 are categorical variables. A total of 185,426 observations are found in this dataset where 153,260 observations were made during the years 2010 to 2014 and 32,166 observations in the year 2015. The data set contains only numeric data and does not include any text data.

**RESEARCH QUESTIONS**

Below are the questions that can be asked to draw meaningful conclusions from the FARS dataset:

1. How have the fatalities number changed over 5 years (2010 to 2015)?
2. What factors affect the traffic fatalities the most that the number of fatalities showed substantial change?

2.1 Environment condition (areas: state, road conditions: light conditions, and weather)

2.2 Driver and other causes (Impaired driving, speeding, etc.)

1. Are genders related to high fatalities number?
2. Does the total number of occupants involved in the crash affect the number of fatalities?
3. Which variables move in relation to the fatalities number?
4. Is there a difference in mean of the number of drunk drivers, number of drivers with invalid licenses, and number of drivers under age 16?

**DESCRIPTIVE STATISTICS**

The overview of the data set is obtained by summarizing the basic descriptive statistics of all the variables. Furthermore, the dependent variable *‘fatals’* is reviewed by year in which the crash occurred.

*Table 1. Descriptive statistics of all variables in the dataset*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Mean | Std. Deviation | N missing | Completion Rate |
| state | 27.89 | 16.22 | 0 | 1.00 |
| st\_case | 279,508.54 | 162,204.77 | 0 | 1.00 |
| **ve\_total** | **1.53** | **0.82** | **1,833** | **0.99** |
| **ve\_forms** | **1.49** | **0.79** | **0** | **1.00** |
| county | 91.47 | 95.38 | 1,840 | 0.99 |
| city | 1,194.31 | 1,890.69 | 1,901 | 0.99 |
| day | 15.65 | 8.84 | 1,833 | 0.99 |
| month | 6.76 | 3.35 | 1,833 | 0.99 |
| year | 2,012.52 | 1.72 | 1,833 | 0.99 |
| date\_SIF | 19,374.25 | 635.27 | 1,833 | 0.99 |
| day\_week | 4.14 | 2.10 | 1,833 | 0.99 |
| hour | 12.66 | 6.93 | 3,189 | 0.98 |
| minute | 28.40 | 17.47 | 3,213 | 0.98 |
| nhs | 0.33 | 0.47 | 3,688 | 0.98 |
| latitude | 36.70 | 5.17 | 3,688 | 0.98 |
| longitud | -91.94 | 14.78 | 3,688 | 0.98 |
| lgt\_cond | 1.83 | 1.02 | 2,647 | 0.99 |
| weather | 2.62 | 3.29 | 3,395 | 0.98 |
| **fatals** | **1.09** | **0.36** | **1,833** | **0.99** |
| **drunk\_dr** | **0.31** | **0.48** | **1,833** | **0.99** |
| **total\_numoccs** | **2.27** | **1.96** | **1,006** | **0.99** |
| total\_hit\_run | 0.05 | 0.21 | 837 | 1.00 |
| **total\_registered\_owner** | **0.84** | **0.75** | **4,690** | **0.97** |
| **total\_not\_registered** | **0.43** | **0.58** | **4,690** | **0.97** |
| **total\_other\_owner** | **0.22** | **0.49** | **4,690** | **0.97** |
| total\_fire\_exp | 0.04 | 0.20 | 694 | 1.00 |
| **total\_valid\_license** | **1.30** | **0.85** | **5,039** | **0.97** |
| **total\_invalid\_license** | **0.18** | **0.40** | **694** | **1.00** |
| **no\_prev\_acc** | **1.27** | **0.80** | **18,658** | **0.90** |
| **one\_prev\_acc** | **0.16** | **0.39** | **18,658** | **0.90** |
| **two\_prev\_acc** | **0.08** | **0.29** | **18,658** | **0.90** |
| **no\_prev\_sus** | **1.27** | **0.84** | **6,016** | **0.97** |
| **one\_prev\_sus** | **0.10** | **0.31** | **6,016** | **0.97** |
| **two\_prev\_sus** | **0.14** | **0.36** | **6,016** | **0.97** |
| **no\_prev\_dwi** | **1.44** | **0.80** | **6,016** | **0.97** |
| **one\_prev\_dwi** | **0.04** | **0.19** | **6,016** | **0.97** |
| **no\_prev\_spd** | **1.22** | **0.80** | **6,016** | **0.97** |
| **one\_prev\_spd** | **0.19** | **0.42** | **6,016** | **0.97** |
| **no\_prev\_oth** | **1.22** | **0.81** | **6,016** | **0.97** |
| **one\_prev\_oth** | **0.18** | **0.41** | **6,016** | **0.97** |
| speed\_related | 0.23 | 0.42 | 7,945 | 0.96 |
| **dr\_age\_lower16** | **0.00** | **0.07** | **5,598** | **0.97** |
| **dr\_age\_lower18** | **0.03** | **0.19** | **5,598** | **0.97** |
| **dr\_age\_lower21** | **0.10** | **0.31** | **5,598** | **0.97** |
| **dr\_age\_lower30** | **0.32** | **0.52** | **5,598** | **0.97** |
| **dr\_age\_lower65** | **0.83** | **0.80** | **5,598** | **0.97** |
| **dr\_age\_65over** | **0.20** | **0.43** | **5,598** | **0.97** |
| **dr\_male** | **1.10** | **0.76** | **5,388** | **0.97** |
| **dr\_female** | **0.39** | **0.59** | **5,388** | **0.97** |
| drugs\_inv | 0.25 | 0.43 | 120,991 | 0.35 |
| dr\_alcohol\_drug\_med | 0.26 | 0.44 | 50,925 | 0.73 |
| dr\_other\_impair | 0.08 | 0.27 | 50,925 | 0.73 |
| **total\_moving\_violations** | **0.15** | **0.36** | **3,364** | **0.98** |
| nm\_alcohol\_drug\_med | 87.47 | 31.72 | 13,779 | 0.93 |
| nm\_other\_impair | 87.45 | 31.78 | 13,779 | 0.93 |
| nm\_involved | 0.18 | 0.39 | 0 | 1.00 |
| **pernotmvit** | **0.21** | **0.49** | **32,129** | **0.83** |
| **permvit** | **2.26** | **1.86** | **32,129** | **0.83** |

From table 1, bold data represents quantitative variables, while non-bold information represents categorical variables. The number of fatalities (fatals) shows its average value at 1.09 i.e, mostly one person died in a car crash. Our group is interested in studying the traffic crash cases that will cause high death tolls. Therefore, all crash cases are divided into group of cases with a normal number of deaths and group of cases with a high number of fatalities. It is assumed that cases of high fatalities have more than one person died in an accident. By dividing the data into two groups, we could more clearly understand the effect of other variables on different fatality numbers.

Beside mean of all variables, we can observe that most of the data have missing value and each variable has a different amount of missing data. The dataset is pre-processed before any further experiments on carried out on the dataset.

**DATA PRE-PROCESSING**

It is observed that only 20% of the records are complete in the entire dataset. Since we can’t make a good prediction model with only 20% of the data, we clean the dataset based on our requirements to utilise maximum data for better prediction. As our goal is to predict the number of fatal in a crash, we reject rows which doesn’t have information on this criterion. An object ‘*FARS\_DS\_1*’ is created filtering the dataset by dropping the rows with NA values in ‘*fatals*’ column which removes 1833 rows of data. Next, all the columns in the dataset are tested for duplication. It is found that all the columns are unique, and no two columns has same set of values. The feature “*drugs\_inv*” is observed to have 35% of data and remaining 65% to be missing. Imputing this will lead to incorrect prediction and thus this feature is omitted from the prediction model. An object “*FARS\_final*” is created by filtering only the complete cases with 87,754 observations over 57 variables.

Furthermore, another object “*FARS\_imputed*” is created by omitting 1,910 rows with missing values in the features “*latitude*” and “*city*” and the remaining data is imputed with mean and mode for the quantitative and categorical variables respectively. The object has 181,683 observations over 57 features.

**EXPLANATORY DATA ANALYSIS**

**Q1: How have the fatalities number changed over 5 years (2010 to 2015)?**

*Table 2.* Descriptive statistics of variable ‘fatal’ grouped by year

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **year** | **n** | **mean** | **median** | **sd** |
| 2010 | 30,296 | 1.089220 | 1 | 0.3584198 |
| 2011 | 29,867 | 1.087454 | 1 | 0.3555516 |
| 2012 | 31,006 | 1.089531 | 1 | 0.3703867 |
| 2013 | 30,202 | 1.089100 | 1 | 0.3579330 |
| 2014 | 30,056 | 1.089433 | 1 | 0.3649784 |
| 2015 | 32,166 | 1.090966 | 1 | 0.3657935 |

*Table 3.* Number of total crash cases and crash cases with high number of fatalities

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Fatals/year** | **2010** | **2011** | **2012** | **2013** | **2014** | **2015** | **Total** |
| crash case | 14734 | 14514 | 15310 | 15102 | 15303 | 16245 | 91208 |
| high number of fatalities (>1) | 1050 | 1019 | 1066 | 1092 | 1037 | 1174 | 6438 |
| percentage | 7.13% | 7.02% | 6.96% | 7.23% | 6.78% | 7.23% | 7.06% |

From table 2 and 3, The number of fatalities almost did not change over the past 6 years, as can be observed from the mean in the table 2. Cases with a high number of deaths has a fixed proportion to the total number of cases at around 7% and it has remained constant for six years. From the data collected for 6 years, it can imply that when a car crash occurs, there is about one fatality per case and there is 7% of serious car crash that kills more than one person in an accident.

**Q2: What factors affect the traffic fatalities the most that the number of fatalities showed substantial change?**

1. **Environment conditions**: (areas: state, road conditions: light conditions, and weather)

*Figure 1.* Bar graph shows the number of fatalities in each state

*Figure 2.* Bar graph shows the highest percentage high number of fatalities cases

From figure 1, we can observe from the data that Texas has the highest number of high death cases. However, the total case of Texas is also high that means higher number of death cases compared to other states may be due to the high number of total crashes. Therefore, we compared each state by percentage of high death case to total cases as we can see from the figure above (figure 2.). The graph shows North Dakota has the highest percentage at 10.63%. We can imply from the result that when a car crash occurs, an accident in North Dakota state has the greatest chance of causing death. Figures 1 and 2 show that the likelihood of fatalities from a car crash in small cities is higher by comparing the proportion of cases with more than one death in the incident.

*Table 4.* The fatalities number with weather condition

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Fatals/weather conditions** | **Clear** | **Cloudy** | **Rain** | **Total** |
| crash case | 68976 (76%) | 13302 (15%) | 6618 (7%) | 91208 (100%) |
| high number of fatalities (>1) | 4812 (75%) | 946 (15%) | 495 (8%) | 6438 (100%) |

*Table 5.* The fatalities number with light condition

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Fatals/light conditions** | **Daylight** | **Dark-Not lighted** | **Dark-Lighted** | **Total** |
| crash case | 42147 (46%) | 25454 (28%) | 19613 (22%) | 91208 (100%) |
| high number of fatalities (>1) | 3033 (47%) | 1968 (31%) | 1174 (18%) | 6438 (100%) |

From tables 4 and 5, we found that car crash incidents usually occur when the weather is clear and during the daylight whether it be the number of total crash case or cases with high number of fatalities.

1. **Driver errors and other causes**

Our group wanted to study the factors affecting the number of fatalities; therefore, car crash cases are recorded whether these factors are related or not to the incidents. We are interested in the factors as follow:

* Speed related
* Alcohol, drugs, or medication influence on the driver
* A driver who has other physical impairments
* A crash that occurred on a traffic way that is part of the National Highway System
* A crash that a contact vehicle did not stop to render aid
* A crash that a fire occurred in the vehicle

These factor variables are recorded as binary: yes = 1 and no = 0, so we categorized the data into groups recorded as 1 and 0.

For example, when we want to test the speed variable to see if it affects the number of fatalities or not, we grouped the fatalities data into groups with speed involved and the group without speed involved. Then, we do a hypothesis test, examining whether there was a difference in the average number of fatalities between the two groups. If there was, it shows that speed affects the number of fatally injured persons in a car crash. Our group used the Kruskal-Wallis test to see if there was a difference in means at α = 0.05.

1. **Speed related**

**Hypotheses**:

Null Hypothesis, H0: There is no difference in means of number of fatalities between speed related group and unrelated group

Alternate Hypothesis, H1: There is a difference in means of number of fatalities between speed related group and unrelated group(claim)

**Testing the Hypothesis:**

*Table 6. Result of the Kruskal Wallis test*

|  |  |  |  |
| --- | --- | --- | --- |
| **Kruskal Wallis test at α = 0.05** | | | |
| **factors** | **Chi - squared** | **p-value** | **Result** |
| Speed related | 1.1733 | 0.2787 | fail to reject the null |
| **Al\_Drugs\_MED** | **167.59** | **< 2.2e-16** | **reject the null** |
| Other\_impair | 0.17882 | 0.6724 | fail to reject the null |
| **Highway** | **265.42** | **< 2.2e-16** | **reject the null** |
| **Hit and run** | **16.806** | **4.141e-05** | **reject the null** |
| **Fire\_exp** | **426.28** | **< 2.2e-16** | **reject the null** |

The test result shows that there is difference in means of number of fatalities between binary group of alcohol, drugs, or medication influence on the driver, a crash that occurred on a highway, hit and run case, and fire occurrence case. Therefore, these factors affect the number of fatalities.

**Summarize the results**

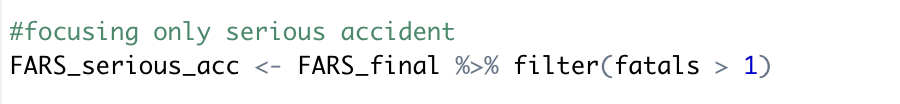
The test result shows that the statistical test value at 1.1733 and the p-value is 0.2787. Therefore, we failed to reject the null hypothesis as p-value is greater than alpha. It means there is not enough evidence to conclude that There is difference in means of number of fatalities between speed related group and unrelated group. In other word, speed did not affect the number of fatalities. We also tested the hypothesis in this way to examine whether there is a difference between two binary groups with other factors mentioned above, and the result shows as follow:

**Recommendation**

As an organization involved in the surveillance of road accidents, the state government and involved officers should have strict preventive measures and surveillance on drivers affected by alcohol, drug, and medication. They should pay more attention to hit-and-run incidents or a fire that occurs when the car crashes with preparation for rescuers that can go in to help promptly. In addition, drivers need to be more careful when drunk or intoxicated by medical treatment and when driving on the highway.

**Q3: Are genders related to high fatalities number?**

Several men and women drivers in a car crash was counted for each crash incident and recorded into "dr\_male" and "dr\_women" variables. Our group is interested in examining whether there was a difference in the number of drivers between males and females. We focus only on the crash case with a high number of fatalities (more than one person died). We tested this claim using the Wilcoxon rank sum test to see if there was a difference in means at α = 0.05.



**Hypotheses**:

Null Hypothesis, H0: There is no difference in the number of drivers by each gender

Alternate Hypothesis, H1: There is difference in the number of drivers by each gender(claim)

**Test Results**

Wilcoxon Ran Sum Test is used to test the hypothesis. The test result shows that the statistical test value (W) at 32511126 and the p-value is <2.2e-16. Therefore, we rejected the null hypothesis as p-value is less than alpha. It means there is enough evidence to conclude that there is difference in means of number of drivers between male and female. In other word, gender of driver is related to high number of fatalities.

**Q4: Does the total number of occupants involved in the crash affect the number of fatalities?**

The total number of occupants involved in the crash are recorded for each car crash. To examine that this number influent the number of fatalities, our group divided the fatalities data into two groups by using the total number of occupants; therefore, we got the group with high number of occupants and the group with low number of occupants. We used the average vehicle occupancy value at 1.50 (based on data from the National Household Travel Survey) as a cutoff point to divide data into two groups.

Text

Description automatically generated

Then, we do a hypothesis test, examining whether there was a difference in the average number of fatalities between group with high number of occupants and the group with normal number of occupants. If there was, it shows that occupants number affects the number of fatally injured persons in a car crash. Our group used the Kruskal-Wallis test to see if there was a difference in means at α = 0.05.

**Total number of occupants**

**Hypotheses**:

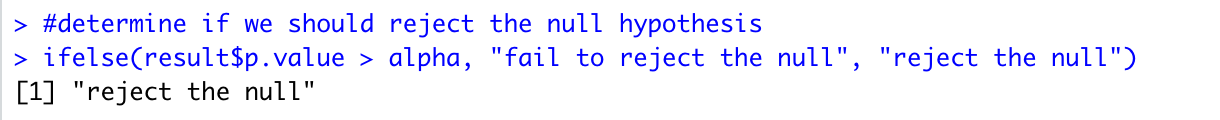
Null Hypothesis, H0: There is no difference in means of number of fatalities between the high occupant group and low occupant group

Alternate Hypothesis, H1: There is difference in means of number of fatalities between the high occupant group and low occupant group (claim)

Chart, text

Description automatically generated with medium confidence

**Make the decision**



**Summarize the results**

The test result shows that the chi-squared value at 2382.3 and the p-value is < 2.2e-16. Therefore, we reject the null hypothesis as p-value is less than alpha. It means there is enough evidence to conclude that There is difference in means of number of fatalities between the high occupant group and low occupant group. In other word, the total number of involved in the crash affect the number of fatalities.

**Q5: Which variables move in relation to the fatalities number?**

The correlation matrix is created to see the correlation between quantitative variables. We use it to discover the relationship related to the number of fatalities. The number of fatalities is correlated with the total number of occupants and number of people in motor vehicles in transport. The correlation coefficient is about 0.5, which means they’re positively correlated. More number of people involved in the crash is moderately related to higher number of fatalities. Figure 3 shows the correlation between various numeric variables in the dataset.

Chart, scatter chart

Description automatically generated

*Figure 3.* Correlation matrix of quantitative variables from FARS dataset

**Q6: Is there a difference in mean of the number of drunk drivers, number of drivers with invalid licenses, and number of drivers under age 16?**

**One-way ANOVA**

Is there sufficient evidence to conclude that there is a difference in mean of the number of drunk drivers, number of drivers with invalid licenses, and number of drivers under age 16? Assume the 0.05 level of significance.

**Hypotheses**

Null hypothesis, Ho: There's no difference in mean number of drivers among three types of drivers (µ1=µ2=µ3).

Alternative hypothesis, Ha: There's a difference in mean number of drivers among three types of drivers (claim).

**Test Results**

ANOVA’s result show that p-value is much smaller than the alpha value (α = 0.05); therefore, we reject the null hypothesis that the means are all equal. In other words, there's difference in mean number of drivers among drunk drivers, drivers with invalid licenses, and drivers under age 16.

By using Tukey's ‘Honest Significant Difference’ method, the result show that the p-value after adjustment for the multiple comparisons (p adj) of each pair of driver type is lower than significant (alpha level of 0.05). Therefore, there is sufficient evidence to prove that a significant difference in mean number of drivers exists among drunk drivers, drivers with invalid licenses, and drivers under age 16.

**PREDICTION MODELS**

With the filtered data in “*FARS\_final*” containing 87,754 observations over 57 variables, the training (70,204 observations) and testing (17,550 observations) sets are created by splitting the dataset in a ratio of 80:20. Five different prediction models are created using the principles are Generalized Linear Model. The AIC, BIC and RMSE values for both training and testing data are calculated to compare between the models.

**Model #1: Linear Regression with all variables as predictors**

To being with, the simplest model is built with all variables as predictors using linear regression method. The chosen response variable “*fatals*” is predicted using all the remaining 56 variable used as predictors. The model is trained using the training data. The summary of the model shows that there are 19 significant variables and 13 are highly significant. The variables identified as highly significant are *ve\_forms, county, city, hour, nhs, drunk\_dr, total\_fire\_exp, no\_prev\_dwi, one\_prev\_dwi, dr\_alcohol\_drug\_med, total\_moving\_violations, pernotmvit* and *permvit*. Looking at the coefficients of all the predictor variables, it is observed that 22 variables are found to have negative coefficients. Some of them with significant coefficients includes *nm\_involved, year, ve\_forms, one\_prev\_dwi, total\_moving\_violations,* and *no\_prev\_dwi*.

**Model #2: Linear Regression with highly significant variables as predictors**

This model is created based on the highly significant variables (0.001 significance level) identified in the previous model. The variables *ve\_forms, county, city,* and *hour* are omitted and the remaining 9 variables are used to build the prediction model. All the variables are found to be highly significant, and the equation of the model is given by:

|  |
| --- |
| **Y =** 0.957 + 0.016(nhs) + 0.034(drunk\_dr) + 0.137(total\_fire\_exp) **-0.034**(no\_prev\_dwi)  **- 0.047**(one\_prev\_dwi) + 0.021(dr\_alcohol\_drug\_med) **-0.062**(total\_moving\_violations) + 0.022(pernotmvit) + 0.069(permvit) |

It is evident that the variables *no\_prev\_dwi, no\_prev\_dwi* and *total\_moving\_violations* seem to have negative impact on number of fatal in a crash. Also, the occurrence of fire in the crash has high impact on the fatality.

**Model #3: Linear Regression with selected 10 variables as predictors**

The top 10 reasonable predictor variables are chosen from the 56 variables and used to build this model. The predictor variables used in this model are *total\_moving\_violation, drunk\_dr, total\_not\_registered, total\_invalid\_license, two\_prev\_sus, one\_prev\_oth, speed\_related, dr\_age\_lower65, dr\_other\_impair,* and *dr\_alcohol\_drug\_med*. The summary of the model shows all of them are highly significant except *two\_prev\_sus* and *speed\_related*. The equation of the model is given by:

|  |
| --- |
| **Y** = 1.042 **- 0.016**(total\_moving\_violations) + 0.019(drunk\_dr) + 0.037(total\_not\_registered) +  0.013(total\_invalid\_license) **-0.0003(two\_prev\_sus)** + 0.015(one\_prev\_oth) + 0.001(speed\_related) + 0.023(dr\_age\_lower65) + 0.014(dr\_other\_impair) + 0.018(dr\_alcohol\_drug\_med) |

From the above equation, it is evident that the variables *two\_prev\_sus* and *total\_moving\_violations* have negative impact on number of fatal in a crash. Also, the fatality is impacted by the factors like drunk driver, unregistered vehicles, and drivers’ age lower than 65. Looks like the speed related violations doesn’t affect the fatality in the crash.

**Model #4: Linear Regression with stepwise selection method**

Too many characteristics in a model might cause issues. Hence, we use stepwise selection method to select the optimal predictors for our model. The direction of the step is set to “both” which used both forward and backward direction stepwise to find the best set of predictors. It is found that this method resulted in using 36 predictors and has lowest AIC and BIC values among all the models.

Text

Description automatically generated

*Figure 4.* Stepwise selection method model summary

**Model #5: LASSO Regression Model**

The LASSO regularization uses the tuning parameter (λ) to avoid overfitting. The function *cv.glmnet()* from the "*glmnet*" library is used to get the lambda's ideal value. For the generalized linear model, this function does k-fold cross validation and returns lambda values for various mean square error values that may be graphed. Figure 5 demonstrates the output of *cv.glmnet()* for cross validation of the training dataset.

The summary of this function shows that it tried 86 lambda values, and the minimum lambda value is **0.00020** and the largest value lambdasuch that error is within 1 standard error of the minimum is **0.01313**. The non-zero parameters for these two lambda values are 47 and 5 respectively. Lambda minimum value is used to create a prediction model with 48 predictor variables.

Histogram

Description automatically generated

*Figure 5.* Mean-square error for different lambda values during the cross validation for GLM model with LASSO regularization

From the summary of the model, we can see that the predictors *total\_moving\_violations, ve\_forms, one\_prev\_dwi, total\_hit\_run,* and *no\_prev\_dwi* affects the fatality negatively. Predictors such as *total\_fire\_exp, pernotmvit, permvit, drunk\_dr, nhs, dr\_alcohol\_drug\_med, dr\_other\_impair,* along with variable containing the details of vehicleownership, license status of driver and previous accident history have high positive impact on the fatality.

**Model Comparison**

*Table 7.* Comparing all the five prediction models

| **Prediction Model** | **AIC** | **BIC** | **RMSE of**  **Training Data** | **RMSE of**  **Testing Data** |
| --- | --- | --- | --- | --- |
| LRM with all variables | 51104.99 | 51627.06 | 0.3479 | 0.3442 |
| LRM with sig. variables | 51801.69 | 51902.44 | 0.3499 | 0.3463 |
| LRM with 10 variables | 59152.84 | 59262.75 | 0.3687 | 0.3592 |
| LRM with stepwise selection | 51076.14 | 51415.03 | 0.348 | 0.3442 |
| LASSO Regression Model | - | - | 0.348 | 0.344 |

The models are compared using the metrics such as AIC, BIC and RMSE. Among the first four linear regression models, it is observed that Linear Regression model using stepwise selection method has the lowest AIC and BIC values among all the linear models. It’s RMSE value is similar to have the Linear model with all 56 predictor variables. However, comparing only the RMSE values, we can conclude that LASSO regression model performs well for our dataset.

**LASSO REGRESSION MODEL**

To further optimize the model, we use two datasets namely *‘FARS\_final’* and *‘FARS\_imputed’* from the data pre-processing step. Also, three different sample sizes are taken and experimented on the training and testing dataset. The sample sizes to be used are 20000, 40000 and 85000 with 80% of the data for training and 20% for testing.

*Table 8.* Comparing the LASSO model for different sample sizes and different dataset

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dataset used:** | **FARS\_final** | | | **FARS\_imputed** | | |
| **LASSO Model** | **No. of predictors** | **RMSE of** | **RMSE of** | **No. of predictors** | **RMSE of** | **RMSE of** |
| **Training Data** | **Testing Data** | **Training Data** | **Testing Data** |
| Sample #1  (16000, 4000) | 32 | 0.3496 | 0.3475 | 34 | 0.3354 | 0.3869 |
| Sample #2  (32000, 8000) | 37 | 0.3503 | 0.3468 | 36 | 0.3358 | 0.3539 |
| Sample #3  (68000, 17000) | 42 | 0.348 | 0.3352 | 49 | 0.3401 | 0.3386 |

From table 8, we can observe that the RMSE values of the testing dataset is gradually decreasing when the size of the sample is increased in case of both the dataset. Also, we can see that when the imputed dataset is used, the RMSE value for the training set may appear less but the when the model is used on testing dataset, the RMSE value is higher and appears to be overfitting. Using the unimputed original dataset, we can see the RMSE values are similar for both training and testing datasets and the RMSE values constantly comes down by increasing the sample size. On a closer look at the number of predictors used in the models in different scenarios, we can see that the imputed dataset uses larger number of predictors than the original one. To conclude, having a larger sample size improves the performance of the model and it is recommended to use the unimputed dataset.

**CONCLUSION**

In this project, we have used LASSO Regression model to create a prediction model to forecast the fatality rate in a crash. It is observed that when the number of occupants involved in a crash is directly proportional to the number of fatally injured people involved in the crash. From the results of the model, it is evident that driving under the Influence of Alcohol, Drugs, or Medication leads to an increased risk of serious injury in an accident. In crash cases occurred on highways, high fatalities numbers are observed when driver of the contact vehicle did not stop to render aid, and crash involves fire. To conclude, the number of people involved in the crash, occurrence of fire in the crash, violation history of drivers, their drunk status, Influence of Drugs, or Medication are the top factors (positively) that contributed to predicting fatalities.

**REFERENCES**

Bluman, A. G. (2017). Elementary statistics: A step by step approach. New York: McGraw-Hill Higher Education.

Cran.r.(2022). *Using Skimr*. <https://cran.project.org/web/packages/skimr/vignettes/skimr.html>

Kabacoff, R. I. (2015). R in Action (2nd ed.). Manning Publications.

Narasimhan, K. A., (2021). *Why Linear Regression is not suitable for classification?.* Analytics Vidhya. <https://medium.com/analytics-vidhya/why-linear-regression-is-not-suitable-for-classification-cd724dd61cb8>

Statistics how to. (n.d.). *Kruskal Wallis H Test: Definition, Examples, Assumptions, SPSS.* <https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/kruskal-wallis/>

**APPENDIX**

The R file named *ALY6015\_FinalProject\_Team4.R* is attached along with the word document.