

**MISM: 6203 Final Project Report**

**By**

**Zhi Ji and Spencer Young**

**Dec 15, 2020**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Project Report**

**Business Analytics Methods**

### **D'Amore-McKim School of Business**

**Northeastern University**

**Table of Contents**

**Part 1: Introduction**

1.1 Topic Introduction & Social Understanding

1.2 Data Introduction

1.3 Research Target

**Part 2: Data Processing**

2.1 Data Collection & Wrangling

**Part 3: Data Analyzing & Visualization**

3.1 Visualizations & Findings

3.1.1 Geographical Views

3.1.2 Demographical Views

3.1.3 Time-related Views

3.1.4 Environmental Elements Views

3.2 Predictive Modeling

3.2.1 Support Vactor Machine Model

3.2.2 K-nearest Neighbor Model

3.2.3. Logistic Regression Model

**Part 4: Conclusion**

4.1 Findings

4.2 Management Recommendation

4.3 Societal Implications

**Part 1: Introduction**

**1.1 Topic Introduction and Social Understanding**

In this project, publicly available car accident data is analyzed. As the data suggests, an unfortunate truth in the United States is that car accidents are common and a huge public safety issue. As with most problems or issues in America, businesses attempt to profit or capitalize financially based on the respective problem’s occurrence. Car insurance companies focus on the occurrence and threat of car accidents. Car insurance companies also use large amounts of data to gain a better understanding of how and why car accidents occur in order to assess risk. The problem that must be addressed for insurance companies is how data can be used to reliably predict car accidents and what factors best indicate their occurrence. From a societal perspective, the problem is using such data most effectively to make driving a safer activity for all of those that are involved.

**1.2 Data Introduction**

The data that was used in our research was obtained from Kaggle.com, a popular online community of data scientists and machine learning practitioners that share publicly available datasets. The original car accident data set included 3.5 million observations and 49 variables. The observations that were included in the dataset occurred in the United States from February 2016 to June 2020. The dataset was generated using application programing interfaces (API’s), which provide streaming traffic incident data. The dataset has a large number of variables, with each variable offering potentially valuable information for identifying factors that are predictive of a car accident. The dataset contains a series of variables that were recorded at the time of each accident, which can be organized into several subgroups as shown below in Table 1:

Table 1: Variables in Car Accident Dataset

|  |  |
| --- | --- |
| Type of Variable | Variables |
| Locational Related Variables | Longitude, Latitude, Street, Side of the roads, City, County, State, and Zip code. |
| Time Related Variables | Start time, Time zone, Time of the day based on: sunrise or sunset, civil twilight, nautical twilight, and astronomical twilight. |
| Weather Related Variables | Temperature (in °F), humidity, Pressure, Visibility (in mile), Wind direction, Wind speed (in mph), Precipitation (in inch), and Weather condition. |
| Environmental Related Variables | Amenity (presence of amenity in a nearby location, the rest are the same logic), Bump, Crossing, Give way, Junction, No exit, Railway, Roundabout, Station, Stop, Traffic calming, Traffic signal, Turning loop. |

(2018 whole year data not included)

U.S. Traffic Accident Dataset:

<https://www.kaggle.com/sobhanmoosavi/us-accidents>

State Population Totals Dataset:

<https://www.census.gov/data/datasets/time-series/demo/popest/2010s-state-total.html>

U.S. Holiday Dates 2010-2020

<https://data.world/jennifer-v/us-holiday-dates-2010-2020>

**1.3 Research Target**

The first research target of interest is identifying factors associated with the severity of car accidents. Severity is of paramount interest as car accidents are responsible for on average 3,000,000 nonfatal injuries and 38,000 deaths annually. We believe that better preventive measures can be developed by understanding the factors distinguishing nonfatal accidents from fatal accidents. The second research target is identifying factors that are associated with the occurrence of car accidents. Identifying such factors will enhance our general understanding of how to reduce the occurrence of accidents.

**Part 2: Data Processing**

**2.1 Data Collecting & Wrangling**

Our data wrangling can be broken down into several parts. First, the original U.S. Accident dataset contained over 3.5 million observations. This number of observations was far too big for the efficiencies of the analysis. Accordingly, our first step was to randomly separatea 500,000-observation size sample dataset for the analysis.

Secondly, we read through the data dictionary to get a basic understating of the context of our dataset. Reading through the data dictionary also helped us to eliminate some irrelevant variables that wouldn’t be used for the analysis. Then, we addressed some of the issues that arose from formatting the dataset such as dates and the name of states that were abbreviated. We also found that some observations had missing values for some variables. In such cases, we replaced them with the mean values for the appropriate variable.

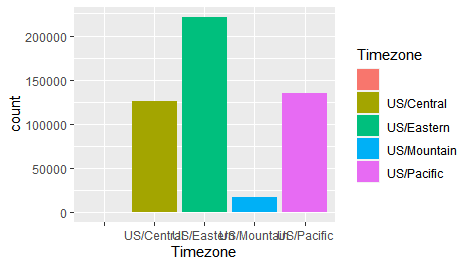
The third step entailed constructing additional variables that could be helpful for the analysis. Since we were dealing with a 4-year time series analysis, we assumed that it would be useful to break the time-related variables into a series of periods and time nodes. Thus, we broke down the date variable and formed several new variables such as day-of-week and holidays. We also did the same for the locational related variables and thus generated more dimensions to the dataset.

**Part 3: Data Analyzing**

**3.1 Visualization & Findings**

**3.1.1 Geographical Views**

Figure 1: Car Accidents by Time Zone

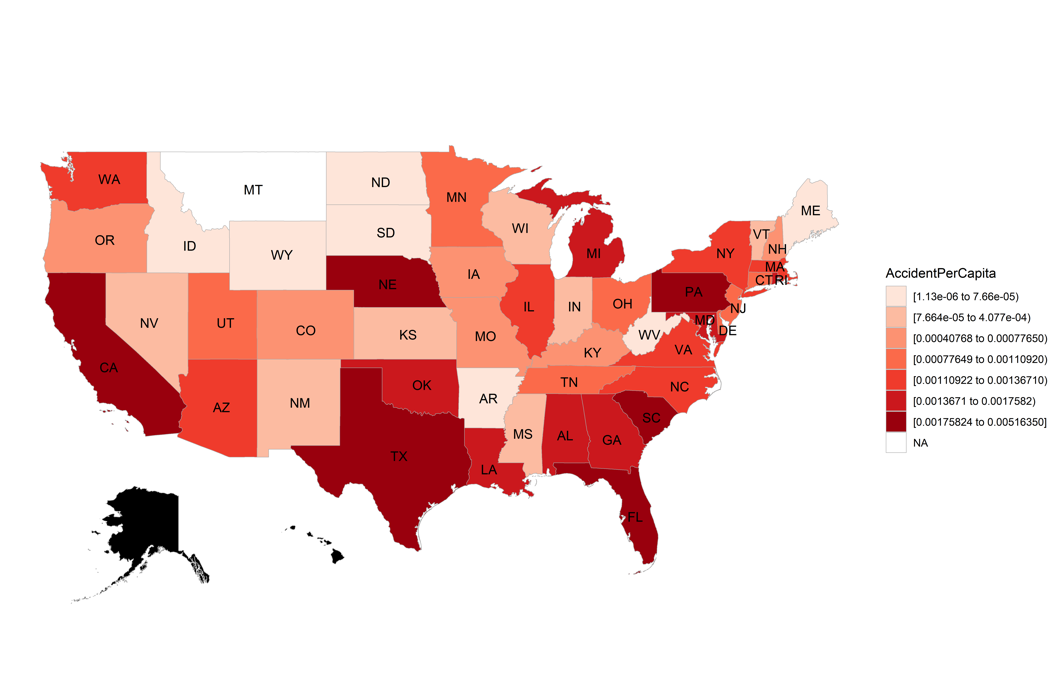


Our analysis included a geographical perspective. The variable “Timezone” helped to divide the U.S. into different geographical sections. Figure 1 presents an overview of the total number of accidents that occurred in each part of the U.S based on the respective time zone where they occurred.

As illustrated in the bar chart, the Eastern Time Zone had the highest number of car accidents by a relatively large margin. We believe this result can be attributed to two different factors. First, the Eastern Time Zone accounts for nearly 50% of the population in the United States as of 2020. Second, the Eastern Time Zone has, on average, more inclement weather than any other time zone in the United States. Conversely, the Mountain Time Zone has the least number of car accidents by a relatively large margin. We believe this may be attributed to the Mountain Time Zone having the lowest population density among the time zones.

**3.1.2 Demographical Views**

Figure 2: National Wide Traffic Accidents Heatmap

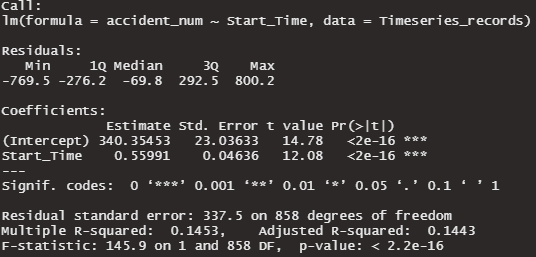


As a second step, we focused on demographical factors that accounted for the different population base in each state. By combining our U.S. Accident dataset with our State Population dataset, we were be able to eliminate the effect of the states’ population. The accident per capita value uncovers the truth of “holding each state’s population constant; that is, which states are more likely to have more accidents than others.”

Figure 2 presents the national wide traffic accidents heatmap that we generated from our dataset. States in dark red had the highest accident per capita -- California, Texas, Nebraska, Florida, South Carolina, and Pennsylvania. Because our dataset doesn’t have any records from Alaska and Hawaii, these states were marked as black.

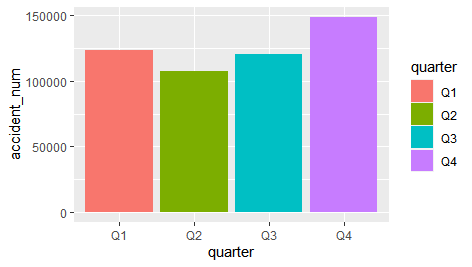
**3.1.3 Time-related Views**

Figure 3: Timeseries Linear Regression



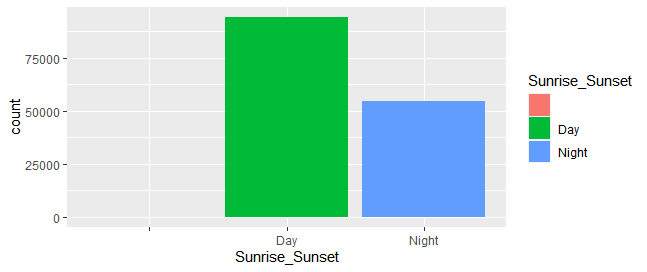
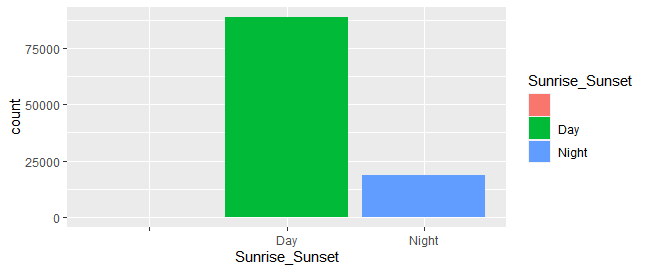
Our next focal point is examining possible seasonal and other time-related factors underlying traffic accident volume. We calculated the total volume of accident for each day. Next, we analyzed a timeseries trend by applying a linear regression model (Figure 3). From the total of 860 individual days covered in the whole dataset, we observed that time and number of accidents is highly correlated. Specifically, starting from February 8th 2016, daily traffic accidents volume, on average, increased by 0.56 per day.

Figure 4: Car Accidents by Quarter of the Year



Next, we grouped the dates into different quarters of a year, then calculated the total number of traffic accidents by quarter (Figure 4). The results show that the 4th quarter of a year has the highest number of accidents, which accounted for nearly 30% of the annual volume. The 2nd quarter has the lowest number of accidents, which accounts for only 21.47% of the annual volume. Thus, in the U.S., the 4th quarter has 459.26 more traffic accidents per day than the 2nd quarter, on average. Given these findings, we kept digging deeper into the data and tried to find out what underlies these patterns.

Figure 5: Day-Night accident in 2nd quarter Figure 6: Day-Night accident in 4th quarter

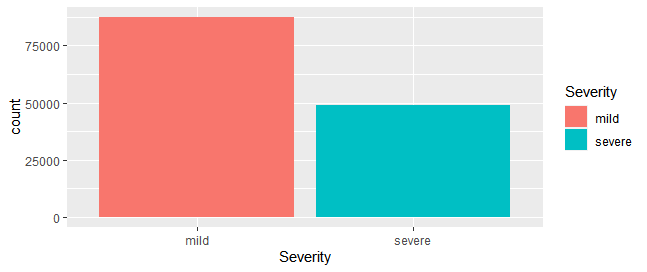
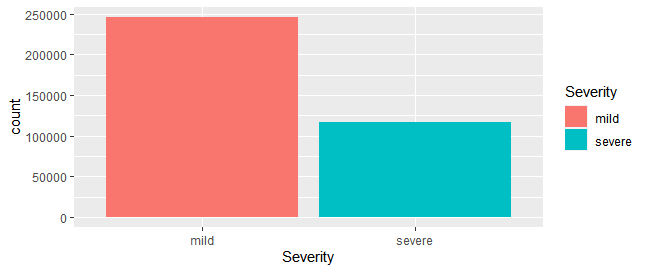


A possible reason for why the 4th quarter has more car accidents is the relative duration of daytime and nighttime. The northern hemisphere has longer nights and shorter days in the 4th quarter of a year, especially in the winter. People tend to have less activities in the nighttime than they do in the daytime.

Figures 5 and 6 above show differences in the percentage of traffic accidents that occur during the day and night. Figure 5 presents the 2nd quarter and Figure 6 presents the 4th quarter. During the 2nd quarter 88,852 accidents occurred the day and 18,471 accidents occurred during the night whereas during the 4th quarter 94,166 accidents occurred during the day and 54,484 accidents occurred during the night. Clearly, more accidents occur during the daytime when people are more likely to be driving.

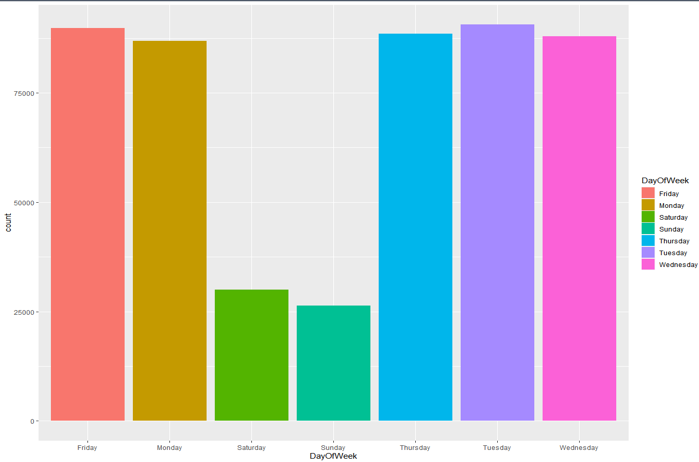
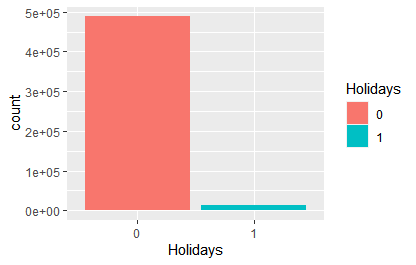
This day/night differential in accidents helps explain the differences observed between in accidents between the 2nd and 4th quarter. From the 2nd quarter to the 4th quarter, the number of accidents during the day increased by 5.98% and the number of accidents during the night increased by a shockingly 194.97%. Thus, the daily number of accidents is highly positively correlated to the length of nighttime.

Figure 7: Severity During Daytime Figure 8: Severity During Nighttime



In the next step, we continued to group the dataset by daytime and nighttime, then examined the composition of each group. Interestingly, we found out that of the total 363,737 accidents that occurred during the day, 32.29% were severe accidents. By comparison, of the total 36,242 accidents that occurred during the night, 35.77% were severe accidents. Thus, although the total traffic accidents that occurred during the nighttime were less than in the daytime, there is a 3.48% higher chance that the accident is going to be severe in the nighttime than in the daytime. Accident severity appears to be somewhat associated with time of day, specifically daytime vs nighttime.

Figure 9: Accident VS Holiday Figure 10: Accident VS Day of Week



Accordingly, we have observed that different time periods during the day are associated with the total number and severity of traffic accidents. As mentioned, in the Data Wrangling section, we generated two variables based on dates. As a next step, we evaluated the effects of “Holiday” and “Day of Week”.

By looking at Figure 9, one can see that holidays are only a very small portion of car accidents in the three-year period (excluding year 2018). In detail, there were 24 holidays out of the 860 individual dates, which accounts for 2.79% of the dates, but there were 12,081 accidents out of 500,000 total accidents that occurred during those holidays, which accounts for 2.42% of the accidents. Thus, the chance of accidents occurring during holidays was lower during normal days.

The other factor is “Day of Week,” presented in Figure 10. The two bars that are significantly less than the others are “Saturday” and “Sunday”, showing that the number of traffic accidents that occurred during the weekends was significantly less than the number on weekdays.

We can come to the conclusion that the number of accidents that happened during the weekends and holidays were less than on normal days and weekdays; this could be explained by people tending to stay at home and spend time with their families on the weekend, however, we can’t definitively say the specific reasons due to the limited information on hand.

**3.1.4 Environmental Elements Views**

Figure 11: Traffic Signs, Environments VS Total Accidents

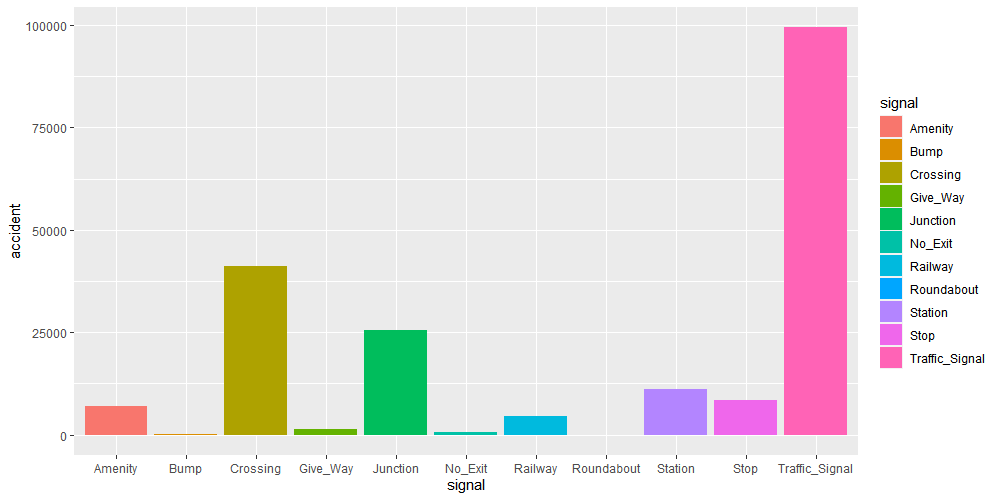
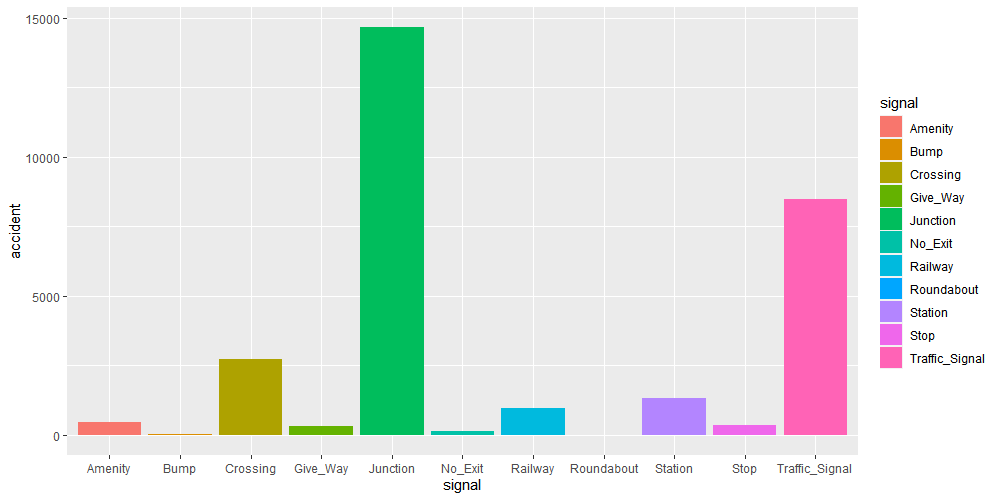


Figure 11 shows the number of total accidents that occurred near each type of traffic sign. We observed that the factor “Traffic\_Signal” has the most records of traffic accidents and “Crossing” has the second most records. The factor “Junction” is in the 3rd position.

Figure 12: Traffic Signs, Environments VS Severe Accidents



The results of previously discussed analysis inspired us to examine if severe accidents have the same pattern as overall accidents. Figure 12 shows only the number of severe accidents that happened near each type of traffic sign or specific type of environment. The results were different from the previous analysis. The factor “Junction” has the greatest number of records, indicating that severe accidents were more likely to occur near junctions, “Traffic\_Signal” and “Crossing” were the second and third in accidents, respectively.

**3.2 Predictive Molding**

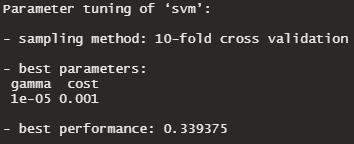
From the previous visualizations and interpretations, we considered some basic ideas of the key drivers of the daily accident numbers as well as accident severities. In the next step, we decided to build predictive models for the purpose of finding hidden patterns in the data.

**3.2.1 Support Vector Machine Model**

In the SVM modeling, the first step was to identify the type of problem that we were trying to solve. In this case, we were dealing with a classification problem that analyzed “under what conditions, is the accident more likely to be a severe one.” Then, the original target variable that had 4 levels of severity was merged and formed into two groups: Severe and Mild.

Then we formed a training set that contained 80% of observations and a testing set that had the remaining 20% of observations. We then tuned the model by using the tune.svm function. The results showed that the applying of nonlinear kernels by use of the Radial Basis Function kernel, the cost and gamma was equal to 0.001 and 1e-05 provided the best accuracy for the model.

Figure 13: Parameter Tuning of SVM Figure 14: SVM Model 1

Above is the confusion matrix for the “best parameters”, we realized that the model was simply predicting all the records to be mild, and the accuracy was 0.6635. In our case, however, we can’t tolerate false negative cases since that would be much more costly than other types of mistakes.

Figure 15: SVM Model 2

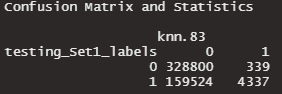


After trying several parameters, we found that the best combination was when cost=50 and gamma=0.07 as model had less false negative cases. After assessing the result, we concluded that the SVM is not suitable for this dataset and so attempted other models.

**3.2.2 K-nearest Neighbor Model**

The second model we built was K-nearest Neighbor model. The optimal K value for the training set was the square root the total number of observations. In this case, the optimal K value fell between 83 and 84 and through the follow up calculations, we found that K=83 did slightly better than K=84.

Figure 16: Confusion Matrix of K model

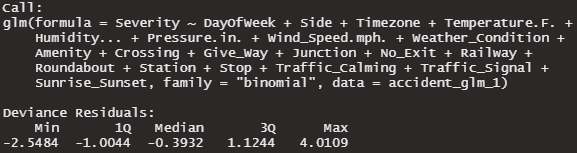


The confusion matrix above (Figure 16) shows that the performance of the K-NN model did a similar job as the SVM model. The overall accuracy (0.6757) was slightly higher than SVM, however, it did much better on precision. That is, when then model predicted a severe traffic accident, there was 0.9275 chance that it was correct.

**3.2.3 Logistic Regression Model**

The logistic regression model entails a binary dependent variable, so the first step was to assign severity level 1 and 2 to “0”, and level 3 and 4 to "1”. Next, we dumped all possible variables into the model to run it first. Then applied step() function with likelihood ratio test to eliminate unnecessary variables from the modeling step. By doing so, we were be able to reduce the 24 factors down to 20.

Figure 17: Logistic Regression Formula



We evaluated the logistic regression model by looking at its Akaike Information Criterion (AIC), and we chose the model that gave us the smallest AIC, in this case, 527925; also, the overall accuracy of this model is 69.21%. The likelihood ratio test helped us to tune the model and the formula of the final logistic regression was shown above in Figure 17.

**Part 4: Conclusion**

**4.1 Findings**

Car accidents result from a combination of several different factors. In this report, these factors were isolated into subsets (geographical, demographical, time-related and environmental-related) to determine which factors have the utmost predictive value in regard to car accidents.

Indeed, our findings suggest that certain factors and attributes contribute significantly to the occurrence of car accidents in the United States. Our finding was as follows:

* The Eastern Time Zone recorded the highest number of car accidents during this time frame
* Accident per capita values were highest in the states California, Texas, Nebraska, Florida, South Carolina, and Pennsylvania.
* The highest total of car accidents occurred at a traffic sign by a large margin; however, accidents were found to be most severe at the junction
* Most car accidents occurred during the daytime, but the proportion of severe accidents were much higher during the nighttime than in the daytime
* Significantly less car accidents occurred on weekends and holidays relative to the number of accidents on a standard weekday
* The 4th quarter of the year has highest total number of car accidents, on average, while the 2nd quarter has the lowest number of accidents

**4.2 Management Recommendation**

We believe our findings offer valuable information and insights for car insurance companies. Car insurance companies can use this data analysis to more reliably predict the occurrence of car accidents throughout the country based on a wide range of attributes. Specifically, they can better assess the risk by building driving profiles of their customers. The insurance companies can also educate their members by informing them of these findings. For example, they can recommend preceding with extra caution during the nighttime and at junctions as this timeframe and location has been found to be associated with highest numbers of severe accidents. Educating members is mutually beneficial as less accidents is better for both the company and client.

**4.3 Societal Implications**

Our analyses make a small but meaningful step towards making the roads in the United States safer for all of those involved. The government can use this information to better understand the factors that contribute to unsafe driving conditions. Once the appropriate government agencies fully understand the risk factors for accidents, they can implement safety regulations and measures that will hopefully incrementally reduce the car accident rate in the United States. Certain initiatives could be an increased presence of police officers at junctions due to the high number of severe accidents occurring there. Additionally, automobile makers can use this information to improve safety features in their cars.