Business Problem

Car accidents are estimated to cost the United States approximately $900 billion dollars per year and kill nearly 40,000 individuals yearly. With this large of an impact on every day life it is important to understand what factors influence car accidents and inform the public on how to reduce their risks.

The goal of this project is to gain the ability to predict the severity of an auto accident based on factors such as the current weather conditions, time of day and/or day of the week.

Data

The US Accident June 20 data set from Kaggle is used for this research. It contains an accident severity indicator ranging from 1 to 4. 1 indicates the a lesser impact on traffic and 4 indicates a significant impact on traffic. The data set contains accident data from February 2016 through June 2020 for 49 of the 50 states within the US (excluding Hawaii). It contains approximately 3.5 million (3,513,617) observations across 42 variables (Data Dictionary). The data was collected from three sources: Bing 29.45%, Map Quest 68.71% and Map Quest – Bing 1.84%. The independent variables for this report are the start date (timestamp), temperature (in Fahrenheit), visibility (in miles), wind speed (in miles per hour), and precipitation (in inches)

Data Cleaning

To gain an understanding of this large data set, I subset the data by state to South Carolina (my current state of residence). This left me with a more manageable 173,277 observation. I also removed and altered some of the variables within the data set to better fit the business plan. I kept the Severity indicator, split the time stamp of start_time into its date, day and hour parts (start_date, start_day and start_hour). I also retained Temperature, Visibility, Wind Speed and Precipitation.

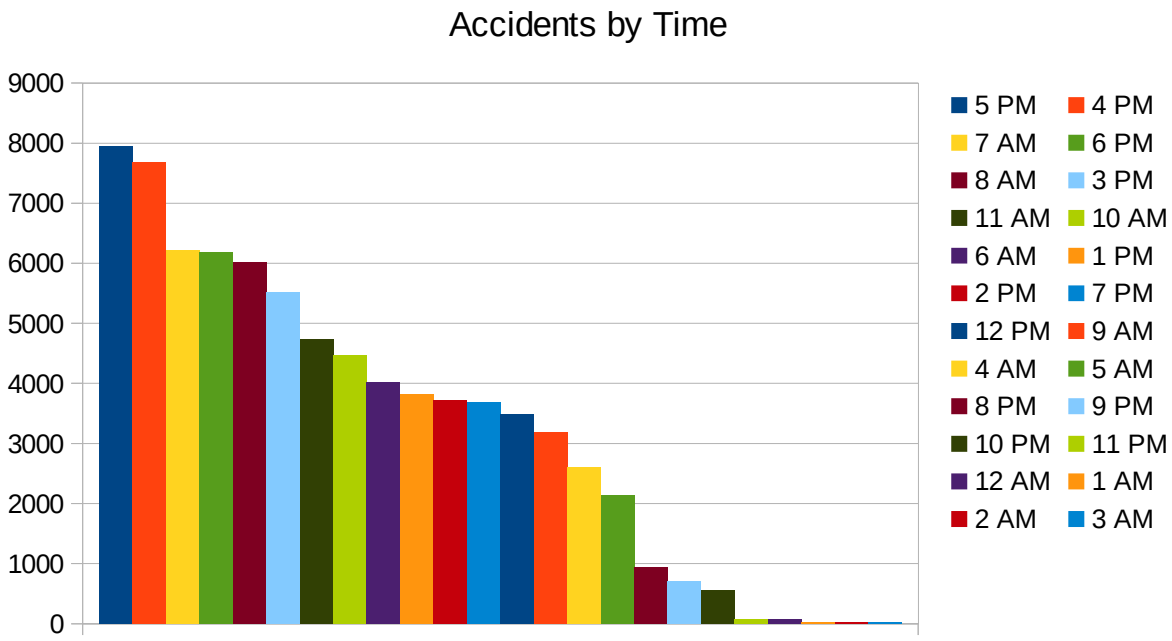
Data Application

I used DB2 (SQL) to subset and reduce the number of variables for this data and created a new CVS file to load into python3 with Jupyter Notebooks. To provider further insight into the business problem, regression and a decision tree will be used to further understand what the day of the weak, time of day and weather conditions play on the severity of the accident. The data will also be standardize to mean zero standard deviation one.

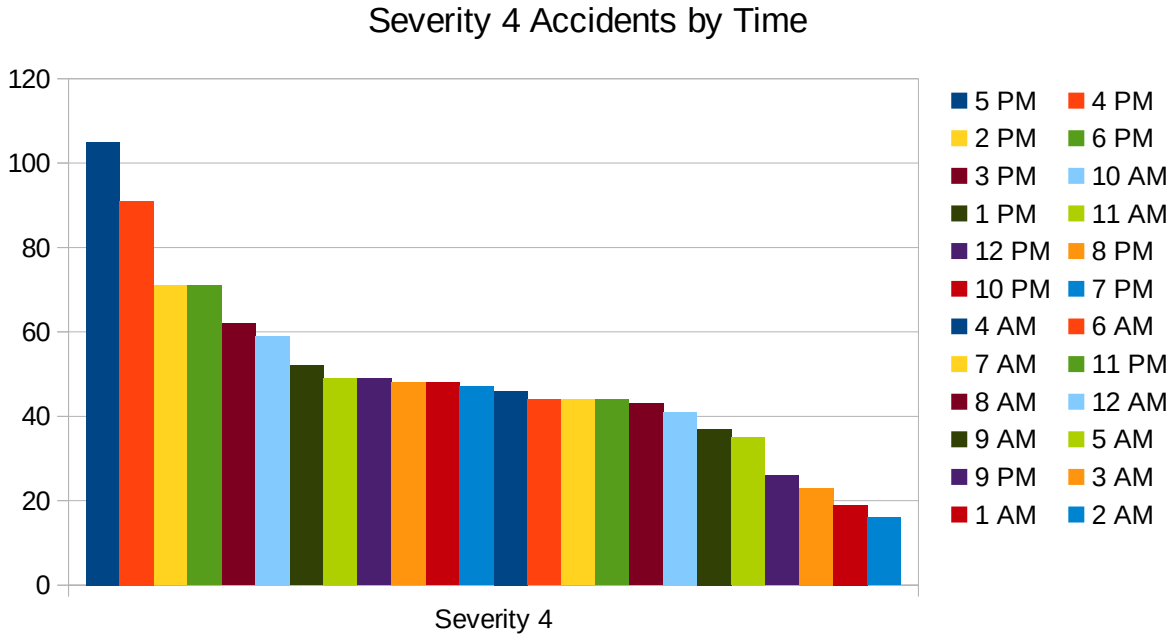
Exploration:

Severity	Count	Percent of Total
1	116	0.07%
2	137,371	79.28%
3	34,620	19.98%
4	1170	0.68%

The first thing I did was examine the counts for the dependent variable, severity. It is clear to see that most of the car accidents in South Carolina are classified as Severity 2 accident (79.28%).

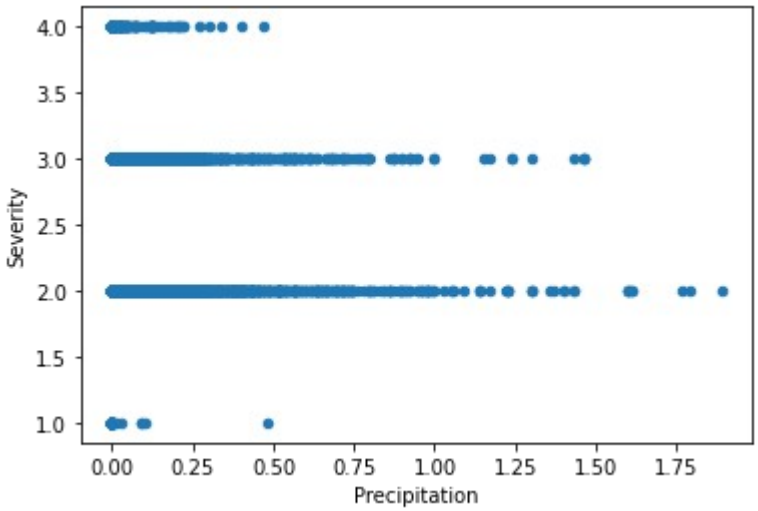


Looking at accidents by time of day, it is interesting to see that the high accident times are the major commuting times for leaving and going to work (5pm, 4pm, 7am, 6pm). It is also interesting to note that there is low accident count past 10 pm.



With this interesting finding I then examined the severity 4 accidents by time. Again, the counts followed the patter of high counts for work commuting times. This may be explained by the definition of the severity indicator. A 4 indicate long delays in traffic. This would lead me to believe that traffic volumes and the accident severity indicator or positively correlated. In other words, the more cars on the road when an accident occurs, the greater the traffic delay.

I then looked at a scatter plot of severity by precipitation. I initially assumed that I would see a positive correctional but was shocked to see very little correlation.



This also noticed with temperature, visibility and wind speed. This partially because severity is not continuous variable, but also because of the lac of variability in severity. As noted above nearly 80% of the observations are of severity 2. This is not what I had expected.

Modeling

To understand to effects the hour of day, day of week, year, temperature (in Fahrenheit), visibility (in miles), wind speed (in miles per hour), and precipitation (in inches) had on the severity of an accident I created a liner regression model. To classify the severity of an accident based on the same variables above I created a decision tree.

Linear Regression

The regression algorithm produced the model

Severity = 2.184 – 0.047 (Year) – 0.014 (Month) + 0.006 (Day) – 0.012 (Hour) + 0.013 (Temp) – 0.015 (Visibility) + 0.004 (Wind Speed) + 0.013 (Precipitation)

This model was not significant. This has to do with two things. First, severity is not a continuous numeric variable. Second, The variance within severity was very small. This can been seen in the first table where nearly 80% of the observation are severity 2. Very little insight was gained by running a linear regression in this manner.

Decision Tree

Before running the final decision tree, I first split the data set into training and test. I used these to data sets to tune the depth parameter. The optimal depth was set at 9.

I then crated my final tree with the full data set and depth set to 9. I then predicted my y hats with this newly trained try. The final decision tree was able to classify correctly 83% of the time. This is just better then always classifying severity as 2.

The more interest outcome came from examining the branches of the tree. The first branches of the tree was split on year. Paired with the information gained by the linear regression, I am able to conclude that every additional year we are able to decrease the severity of accident. I find this interest and would like to research if this decrease in severity due to time.

Conclusion

With the tools I gained from this classed, I was able to determine that an accident is more likely to occur during the common commute times for going and returning from work. This is also the time that the more sever accidents happen. It was also determined that as we continue through time we are lowering the severity of auto accidents. This could be due to better designs of our road instructor, the design of automobiles, and/or our response time to accidents.

Further Research

Further research is needed in the measurement and accuracy of the response variable of severity. With the vast majority of the accidents being classified as 2, it makes using machine learning algorithms difficult because of the lack of variation.

It is also important to see if the patterns identified in this report are consistent accords the other 47 states included in the data set. It is very likely that states can be clustered into groups with similar accident severity.

Further research will also need to be conducted on identifying the cause in the reduction of severity with time. It is very interesting to further understand what factors are driving this downward trend and how we can support them.