# Final Project Report

**Traffic in the Lehigh Valley**

# Prepared By: Group #5

# Joel Camacho

# Mohamed Mazarul

# Bradley John

# 

# 

# Executive Summary

According to the Lehigh Valley Insider, “the Lehigh Valley ranks among Pennsylvania’s fastest growing regions”. The 2020 U.S. Census reported that the Lehigh Valley has grown over 6% the past ten years, which accounts to over 40,000 people. As the population increases the number of accidents has increased. A combination of more users and aging infrastructure has led the government to make key development decisions about where to spend money on road safety projects. The data from the National Highway Safety Transportation Administration FARS database would suggest that corresponding projects selected during the years up to 2014 did not meet objectives in terms of eliminating the most severe of roadway accidents at key locations in the area. We theorized that reviewing the road, temporal, and weather factors associated with accidents during the data period would lead to better decision making regarding future investment in road safety projects leading to reduced accident counts and less time impact on traffic as a result.

After performing analysis on the FARS data through multiple models we found that road characteristics should play a key part in the decision-making process. However, local knowledge of the areas involved must be considered in concert with the analysis in order to produce the optimum decision impact.

# Motivation

# We set to model the factors contributing to accident severity based on weather and traffic patterns as well as the frequency of accidents on certain roadways in the Lehigh Valley area. Having data regarding time of accident, locations, weather conditions, and how traffic was affected afterward will assist in making predictions for high impact locations. This will make for more informed budgeting decisions for departments of transportation in making taxpayer funded roadway expansion decisions and in the placement of preventative measures. It will also assist Insurance companies in pricing in that specific market. If roads are safer, then accidents are reduced, thus making auto insurance pricing more affordable and overall lowers taxpayer funded costs.

# Introduction

As mentioned in the Executive Summary, the Lehigh Valley Insider, “the Lehigh Valley ranks among Pennsylvania’s fastest growing regions”. The 2020 U.S. Census reported that the Lehigh Valley has grown over 6% the past ten years, which accounts to over 40,000 people. This does not account for the new distribution warehouses that have moved into the area, the growth of the trucking industry and workforce traveling to and from those facilities; yet the major roadways in use for ten years remain mostly unchanged today. As everyone shares the roadways, fellow drivers are also concerned with car accidents as they cause traffic and significant time delay. Insurance companies and departments of transportation pay close attention to the frequency and nature of car accidents, especially those considered a higher severity.

The objective of this project is to make Lehigh Valley roadways safer overall. The Lehigh Valley Planning Commission (LVPC) has accumulated data indicating that the five-year period between 2010 and 2014 shows an average of 56 fatalities per year. Their overall safety traffic plan mission is “while these fatalities have declined, despite growth in population and registered vehicles, we must continue to strive toward a zero injury and death goal”.

We are applying various models to identify any trends in the traffic accident data to offer a probabilistic determination of accidents occurring in the future. We want to determine if the past data has any hidden trends, whether accidents continue to or will occur in certain locations (i.e., highways, roadways, exits), weather conditions, traffic patterns, and traffic in general can be determined with a high degree of confidence. We are hoping to provide valuable information to the Pennsylvania Department of Transportation (PennDOT) on how to allocate taxpayer funds in future engineering and construction projects such as the construction and new exit at 22 and Front St, reflective markers along Route 22 and 378, and the replacement bridge over the Lehigh River connecting Coplay to Northampton. The total of these three projects alone is over $92 million. As with any government agency, they are limited by budget constraints and we are hoping to alleviate some budget pressure and create safer roadways.

We are reviewing six years of data (2016-2021) to determine if properly supported predictions for 2022 can be performed. The results will then compare our predictions to the National Highway and Safety 2022 quarter one accident data to assist in validating if the models are progressing as expected.

# Methods

There were several accident-related datasets available across a variety of publishers. After reviewing each one, we selected a dataset from the National Highway Transportation and Safety Administration via Kaggle.com. One of the major reasons for selection is that the dataset disaggregated the data by county, ZIP code, and city. It provided the opportunity to satisfy our requirement for the predictive model to be beneficial and improve our lives. We are able to select the areas in Pennsylvania that we wanted to perform our analysis, which were the three counties that make up the Lehigh Valley (Lehigh, Carbon, and Northampton).

It contained several variables including various weather conditions, traffic patterns, description of accidents, and geo-locations that we thought would offer diversity for our predictive model. It also contains variables such as start and end times of traffic which we could then create additional variables such as total traffic times. We also decided to keep those years that were impacted due to COVID so we could scale against the average daily vehicle counts.

After the selection of the dataset, we examined and cleaned the data utilizing several tools and options. The initial dataset was over 2.8 million rows with 49 variables, thus making it unreadable in its entirety in excel and SAS as both have limits of 1 million rows. We converted the dataset into a Microsoft Access database to filter the three counties worth of data which came to 3,226 rows. There is a column with accident start times which contains the corresponding date and time. Using Access, we separated the start time data into Years, Months, Days, and Time. We then converted the filtered data back into excel.

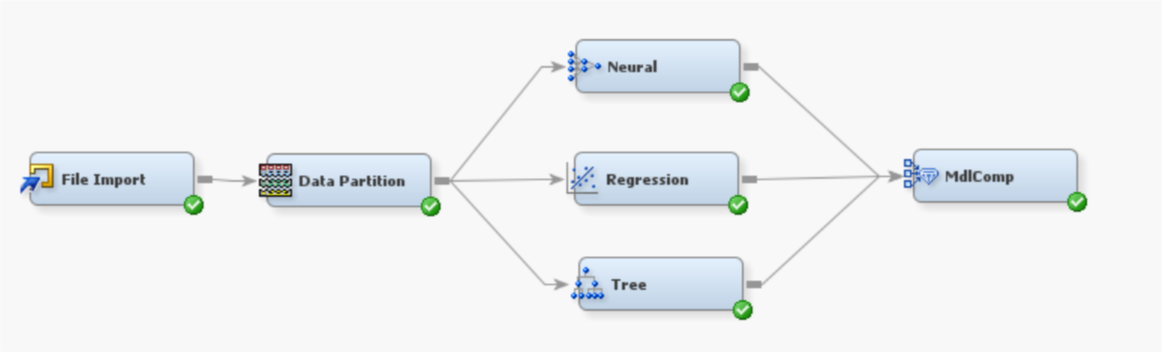
Excel gave us the ability to calculate an additional variable by subtracting the traffic start time from the traffic end time. The dataset was then entered into Tableau Prep for further cleaning by performing the following:

* Removed data that was not determined to be beneficial such as three different dusk types, state (as all three counties are in the state of PA), and airport code (weather station that collected weather data)
* Removed any columns (weather and traffic pattern related) that did not contain any values due to filtering in the three counties
* Replaced or combined any null values in remaining binary data columns with 0
* Combined any duplicative data in wind (N and North), street (I-78E and I-78 East), and ZIP code (anything after the normal 5 digits gets combined into original 5-digit code)
* Properly classified the type of data in each column (text, numbers, calendar date, time, etc.)

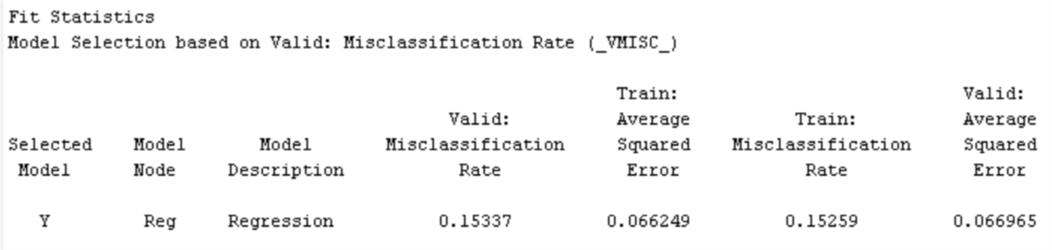
The dataset was then cleaned through Tableau Prep and given in an excel file that was entered into SAS. The final cleaned dataset contained 3,226 rows and 38 columns.

After cleaning the data, we exported it to SAS Enterprise Miner Workstation 15.1 for analysis. We applied several models to the data to look for key indicators. Regression, decision tree and neural network analysis was used to identify what road characteristics have the greatest impact on accident severity. Text Mining was used on accident descriptions to determine if common themes exist. Where do certain directions, construction, or intersections appear frequently? We used the Data Partition node to split the data into a training data set (40%, 1,290 observations), a validation data set (30%, 968 observations), and a test set (30%, 967 observations) via simple random sampling. We then compared three models; decision tree, logistic regression, and neural network for the target variable. We have several categories of predictive variables that we could choose from including weather, geographic, temporal, and traffic features. We choose to start with traffic features since they would intuitively seem to have the most direct impact on the severity. Additionally, the traffic feature variables all happened to be binary variables (either they existed at the scene of the crash, or they didn’t) where multiple features could exist but would not necessarily be conditional upon each other in determining the severity of the accident.

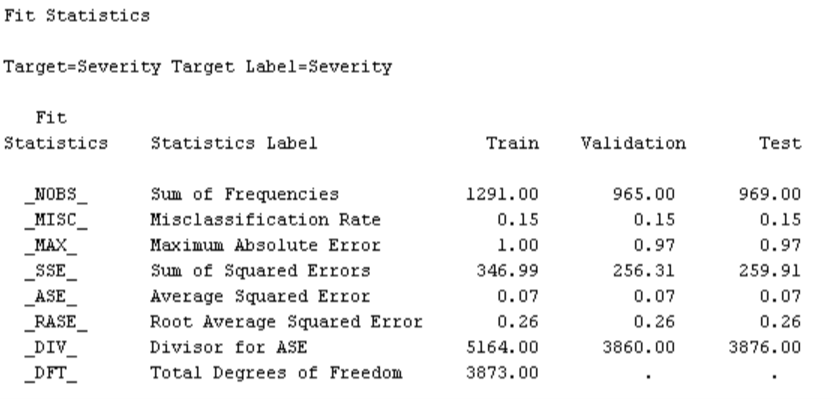
# Results



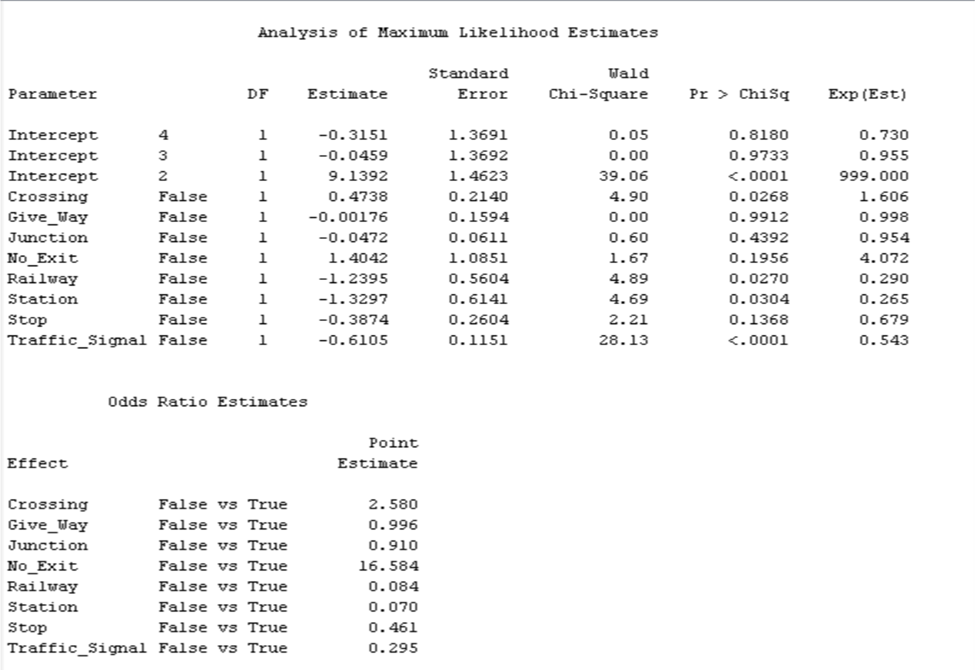
After running a model comparison on the models mentioned above, the logistic regression was selected for the best model to determine the severity level based upon the binary road characteristics data. The key factor was misclassification rate on the test data. Interestingly the misclassification rate was the same between the decision tree and the logistics regression.



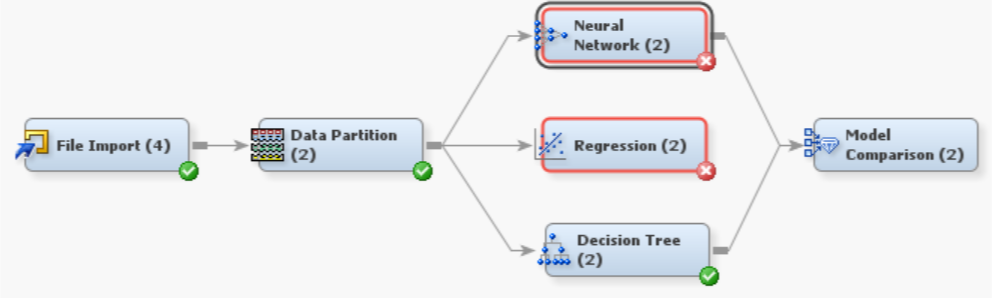
We choose a logistics regression since our target variable is a categorical ordinal target. Additionally, the fit statistics for the regression were generally more positive than the decision tree.



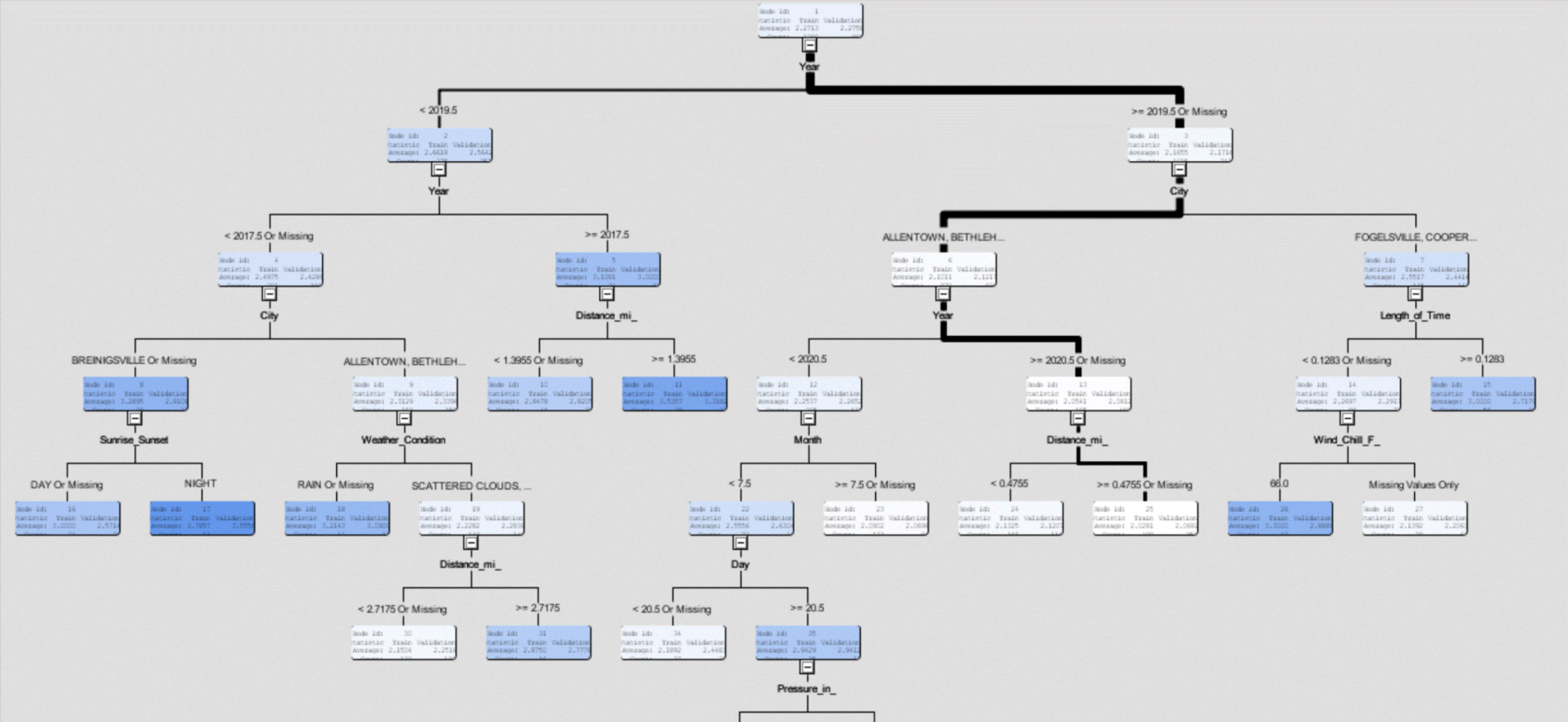
An interesting feature of the binary traffic variables was that some were assumed to have a positive impact on the target variable but others would be assumed to have a negative impact (higher degree of severity). That is exactly what our initial analysis found. The features that met our p-value criteria (.05) included; whether it was a street intersection, whether a railway was a factor in the crash, the presence of a gas station, and the presence of a traffic signal. Conversely there were a couple surprises. For example, the presence of a stop sign did not have the weight that we would have expected.

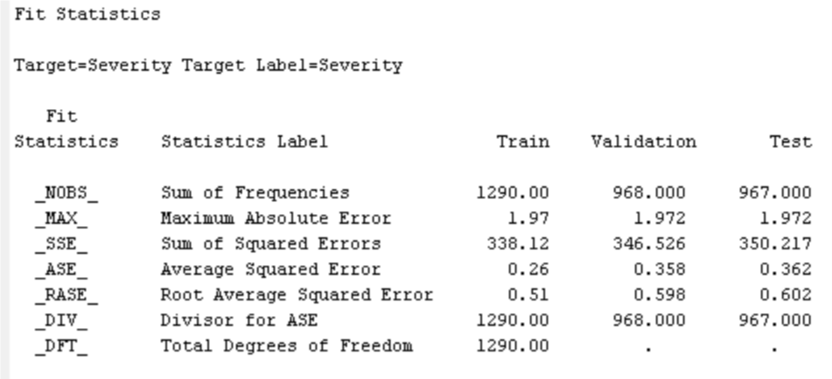


The misclassification rate was still higher than we would have liked so we expanded the data inputs to include weather conditions and other categorical variables from the original data set. When we expanded the variables the neural network and logistics regression errored out as the number of unique records was exceeded.

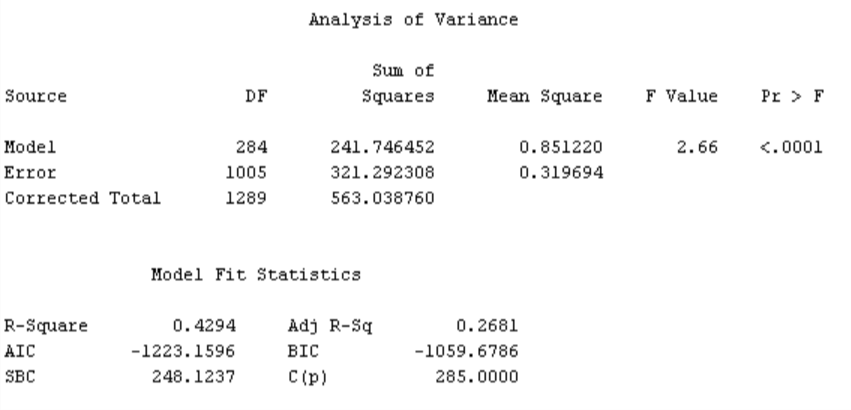


Even after scaling back the number of records the other models errored as Enterprise Miner has a cap of 512 unique records. So, we were left with the decision tree. The fit statistics for the tree deteriorated compared to the logistics regression which left us to wonder - did this mean that the regression model was better but for the Enterprise Miner restriction or was it that the new attributes introduced were less impactful from a classification perspective?

****

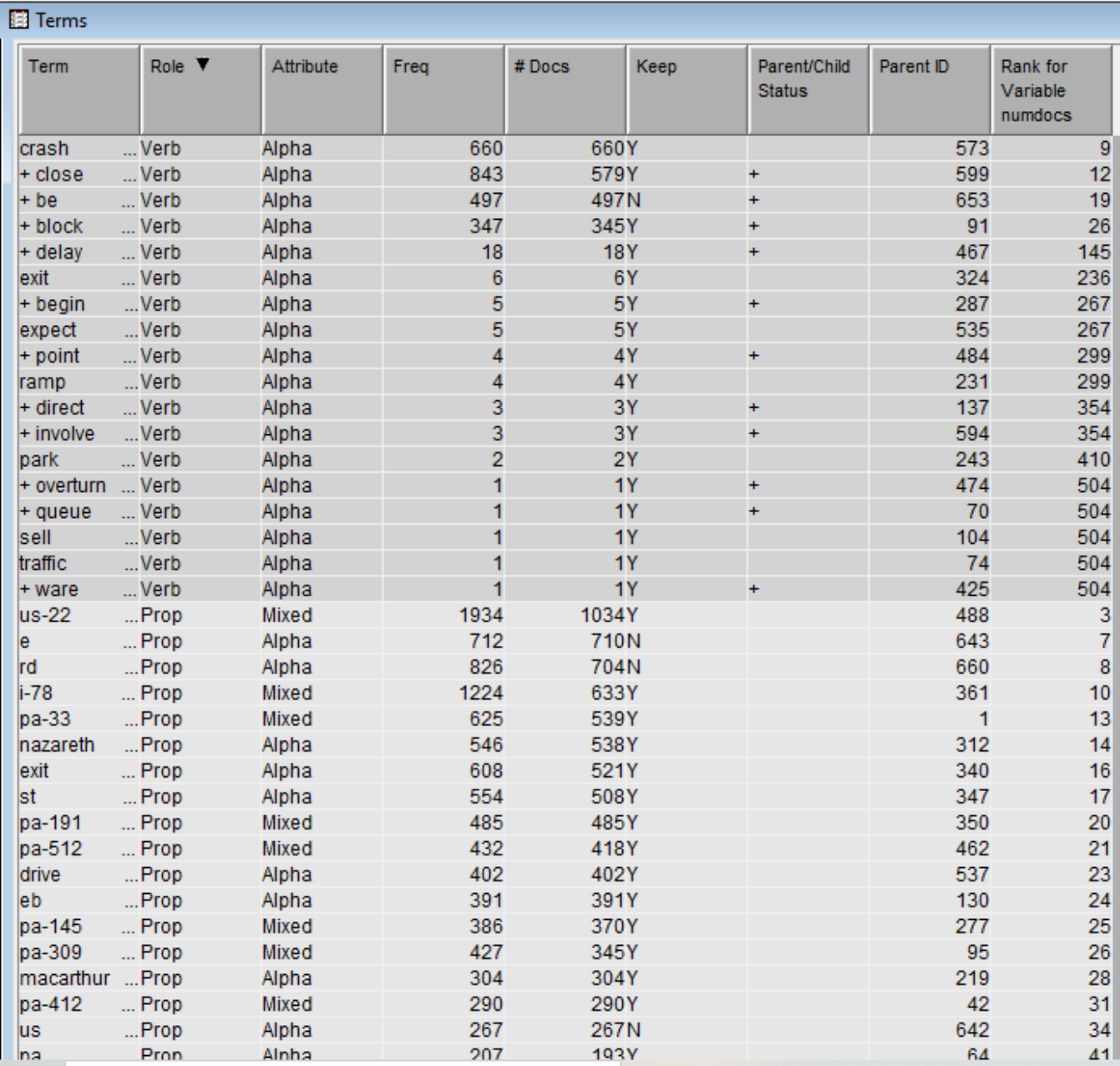


To answer this question, we re-ran the regression model using solely the newly introduced variables. The challenge with this was that we still ran into the enterprise miner cap. So, to mitigate this impact we eliminated the time and distance measures since they were continuous and unique to every record. This allowed us to produce fit statistics for the non-road attributes. The results demonstrated that to our surprise the road characteristics had more of an impact than the weather or temporal attributes in the aggregate. That said, there were of course individual attributes that had a much stronger impact which included; precipitation, wind speed, wind direction, and city.



This analysis would suggest generally speaking that road characteristics have a greater impact on the accident severity that the weather and temporal attributes. Road characteristics are items that are with human control whereas the weather and temporal attributes generally speaking are not. So this implies that sound decision making on the part of governmental authorities can have a meaningful impact on accident severity.

We turned to text mining to identify specific locations and other characteristics that drove accident severity that were as a result of features not identified in the FARS data. The text mining node of Text Parsing indicated that construction (while not active) and commuter hours were the additional drivers of accident severity. Additionally, to validate the data in Excel, we created a macro to partition the description and then performed a query to identify the number of counts for each key term. The counts validated the findings from the text mining shown below.



# Recommendations and Conclusion

* New road projects for safety should focus on those that have characteristics noted above combined with a high rate of frequency identified in the text mining. Locations that met those criteria included sites on I78 and Route 22, specifically at PA 191 and 512.
* Text mining also revealed locations that had high frequency and were negatively impacted by the lack of such road characteristics. For example, certain stretches of main highways or interstates. Those roads are not going to have a stop sign and placing a stop sign at those locations is not an option. Analysis of those locations revealed an indirect factor in the severity data - human factors. An example is 22E at route 191. When we originally reviewed the data, we thought that there must be an error as the frequency of accidents was showing on the eastbound lanes rather than the westbound lanes. Intuitively we thought the opposite should be the case given the shorter on ramp onto the highway. In reviewing the descriptions of the accidents identified in the text mining we realized that drivers aware of the short onramp on the westbound lanes tend to move over to the passing lane knowing a hazard is ahead whereas in the eastbound lanes where the accidents occurred people didn’t feel the need to get over and tended to “race” down the onramp leading to more accidents. This would suggest that when dollars for road improvement are not available an alternative would be to announce a hazard or high incident area and suggest an action such as moving over for the next mile or something similar.
* Having individuals involved in the decision-making process that have personal knowledge of the geographic areas in this scenario is critical. Our finding above is a great example of why this needs to be the case. If we were not familiar with the area where accidents that seemed counterintuitive were taking place, we would have just assumed that there were errors in the data rather than being able to determine some of the logical reasons behind the data.
* Other preventative measures funded by taxpayers that may decrease accidents in these high-volume areas are electronic speed limit posting that the department of transportation or analytical algorithm can update on a real time basis. This will notify drivers of potential hazards such as construction or large volume of vehicles by slowing down all traffic prior to the point of impact. These can replace current speed limit postings along Route 22 and Interstate 78, specifically in the miles east and west of the junctions of PA 191 and PA 512.
* Due to the data analysis concluding that roadway construction areas have a large number of accidents, another preventative measure are electronic speed indicators that enforce the speed limits by capturing photos of the vehicle. These devices are used in surrounding states such as Maryland in their construction zones, are easily moved, and simple to calibrate. A camera is mounted facing the speed indicator and once a certain maximum speed is indicated by the radar, a photo is taken of the speed and license plate of the vehicle in question. A traffic violation is then mailed to the registered owner of the vehicle. By providing a means of paying a potential large fine, traffic speeds should stay at or below the posted speed limit. These devices should be in use in all active major roadway construction sites.
* Larger project recommendations include re-designing the on and off ramps in question above. This would involve a large-scale construction project, significant traffic due to construction (implementation of previously mentioned recommendations to reduce likelihood of accidents), and long-time frame. This is costly and time consuming but may eliminate the volume traffic accidents in those regions. It can address the impact of large trucks navigating traffic patterns that were not designed for them.

# References

* <https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents>
* <https://lvpc.org/data-lv-transportation.html>
* <https://lehighvalley.org/census-2020-lehigh-valley-ranks-pa-fastest-growing-regions/>
* <https://www.penndot.pa.gov/RegionalOffices/district-5/Pages/default.aspx>

# 

# Appendix

Data attributes from National Highway Safety Transportation Administration

