Data Analytics Tools & Techniques Final Report

Accident Data Analysis

Team 6

Chelsea Nowlin - voi967

Felicia Villela - fmg659

Katherine Perritano - oqt134

Liani Castillo - ncs947

Rubina Saya - dxa185

May 12th, 2020

Outline

1. Problem Statement
2. Data Descriptions of Variables
3. Data Cleaning and Pre-Processing
4. Visualizations
5. Predictive Modeling
    a. Data Splitting – training/validation (2013 through 2017)
6. Interpretation of Model Results
7. Generalizations and Recommendations
8. References
9. Appendix

Problem Statement

The main goal of this predictive modeling was to identify the leading cause of fatal accidents in Texas. Our initial hypothesis is that drunk driving has a significant effect on the number of fatalities in an accident. For purposes of this study, we used the publicly available Fatality Analysis Reporting System (FARS) published by the National Highway Traffic Safety Administration (NHTSA). The data set contains information about fatal injuries incurred in the United States.

Data Descriptions of Variables

The descriptions of the variables used and explored in the dataset are located in the Appendix section of this report. Variables and their respective descriptions were derived from the FARS Analytical User's Manual.

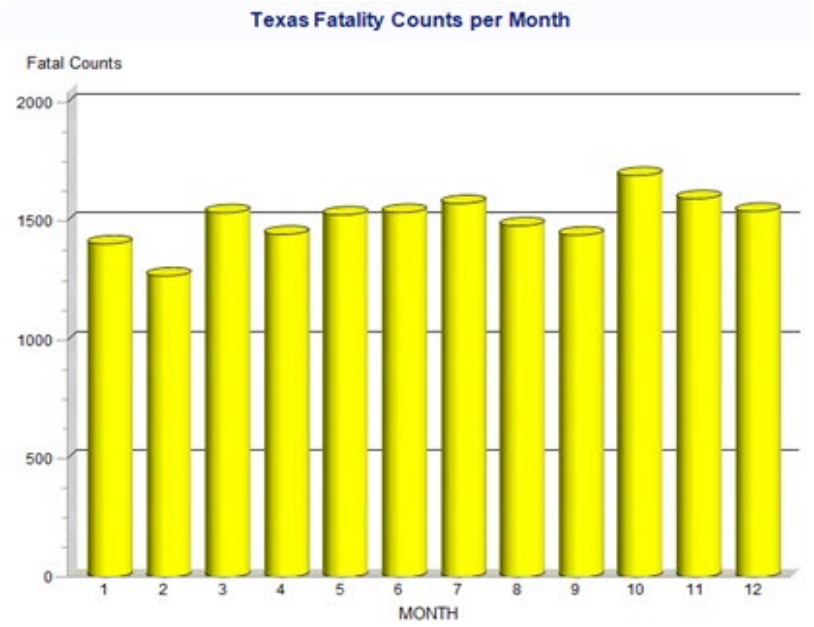
Data Cleaning Pre-Processing

Our accident datasets were available in comma separated value formats, and we utilized excel to complete the data cleaning and pre-processing prior to bringing the data into SAS Enterprise Guide and SAS Enterprise Miner. We did not have any missing values in our dataset, and we removed all observations except Texas to narrow the scope of our analysis. Additionally, we eliminated variables that were either non-pertinent or would not bring value to the problem we chose to evaluate. We discovered the accident dataset reuses the State Case IDs for each individual year. Since we wanted to train our model based on the five years of data available, we had to make a unique key with each observation by combining the Year and the State Case. Once we had the unique identifier, we were able to import the data into Enterprise Guide, and then append all of the years into a single dataset. The total number of observations for Texas was 16,206.

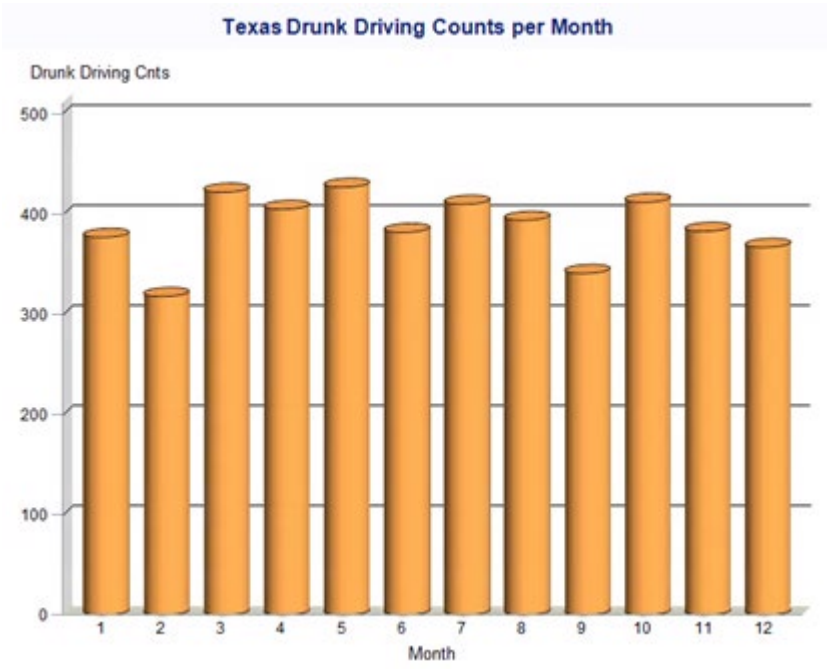
Visualizations

Using the combined Accident datasets from 2013 – 2017, we completed various visualizations to identify outliers and determine if we had cross correlated attributes. We considered our business problem regarding the occurrence of drunk drivers in an accident where a fatality occurred. We filtered our dataset to conduct analysis on Bexar county, however we found that the smaller number of observations did not provide as much insight. Therefore, we based our visualizations in this report on the 16,206 Texas observations and are including some of the insightful graphs in the following section:

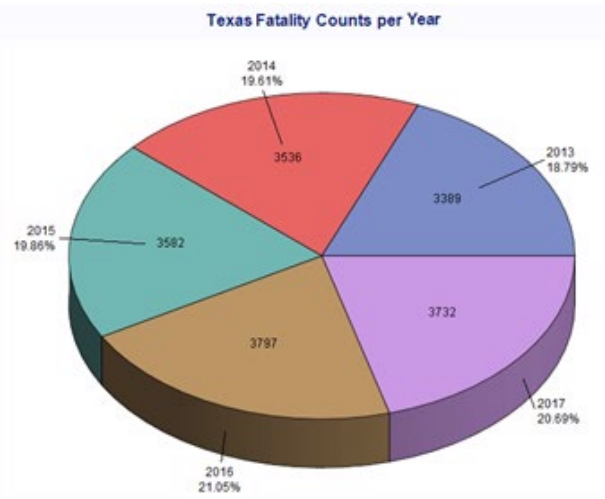
For the analysis of Texas Fatality Counts per month, we found some variation in the total counts by month with October being the highest count, and February being the lowest. However, with February naturally being a shorter month, (in regards to the amount of days) this could be an indication as to why the amount of fatalities was low. Overall, the distribution of fatality counts from January to December straddles around 1500 each month.

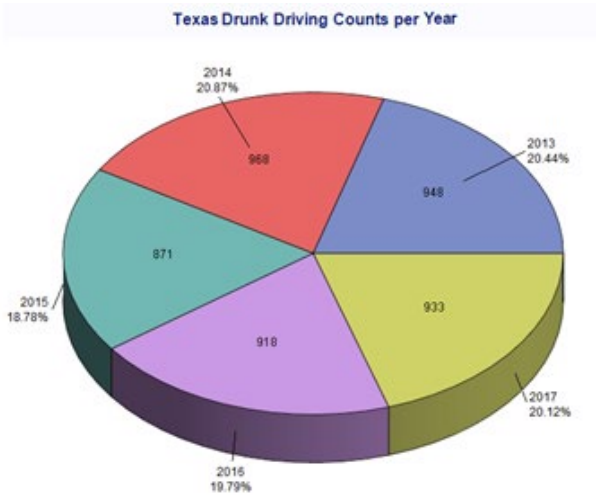


Similarly, we created a visualization to see Texas Drunk Driving accident counts per month. We found a very similar distribution pattern with some variation in the total counts by month. May was the highest count, though only slightly higher than March and October. February had the lowest count of drunk drivers (the shorter month note mentioned above was still taken into consideration). We found that the total number of drunk drivers per month ranged from 300 to slightly above 400. We compared these totals to the fatality counts each month, and found that only May and June seem to differ in the spread of high and low counts each month. We could not determine if there was a correlation between drunk drivers and fatalities in an accident from the monthly count totals.

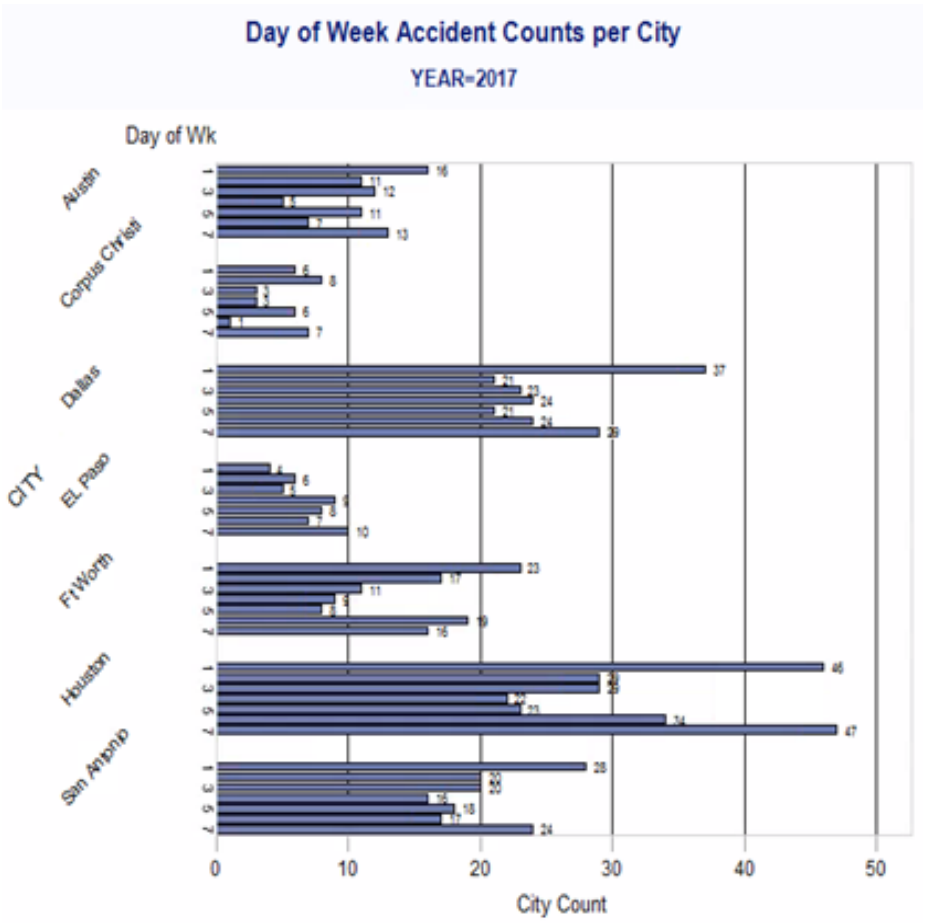


We decided to complete the same count analysis by year since our dataset included 2013 – 2017 values to see how the proportion of total counts by year differed for fatalities and drunk drivers separately. The below pie charts indicated that each slice size is almost exactly split by 20%, meaning the total by year is staying about the same for fatalities and drunk drivers. At this point, it seems that driving drunk on its own has its own effect on accidents but it is not an indication of a fatality occurring since the attributes are appearing similarly distributed.

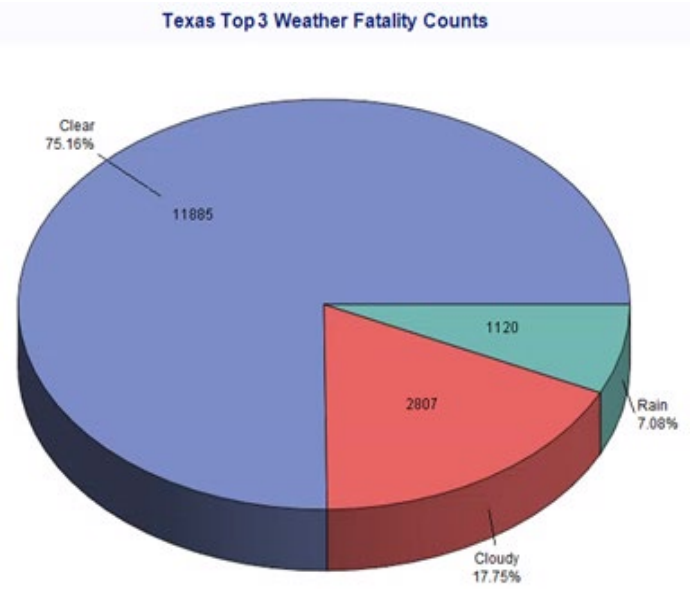




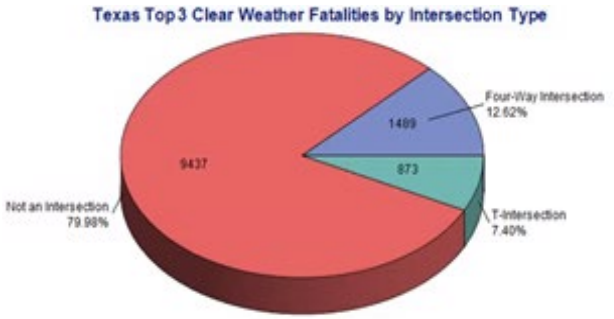
We completed further analysis on other variables to determine if there was a day of the week that would stand out among cities based on the amount of accidents. The following visualization shows the amount of accidents that happen each day of the week across a selection of seven cities in our dataset. Since the variation was not significantly different across each year, we are only including the 2017 graph in this report. Sunday and Saturday had the most accidents for Austin, Dallas, Houston, and San Antonio. However, we recognize the increase in accidents on a weekend could be due to population density of these larger cities compared to weekdays. There was not a clear indication of fatalities increasing on these days of the week.

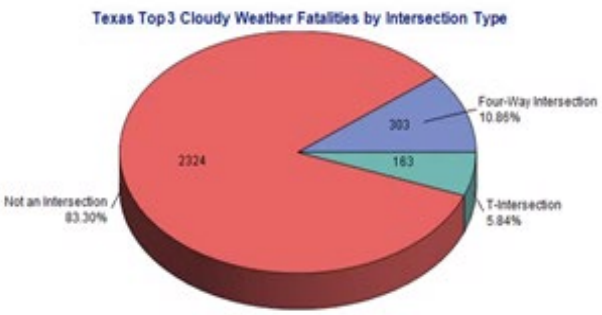


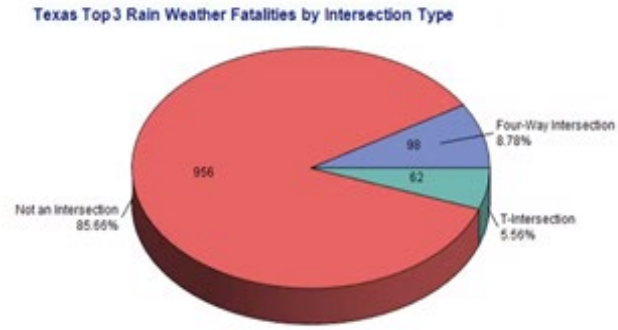
Weather was another attribute in our dataset, and we decided to analyze the frequencies of fatalities based on the weather conditions when the accident occurred. Our original thought was that rainy or cloudy conditions would affect the occurrence of a fatality in an accident. The visualization determined that clear conditions were where most fatalities occurred, followed by cloudy and rainy respectively.



Due to our incorrect assumption of weather's correlation to fatality counts, we decided to dig deeper into those categories and determine which intersection types influenced the counts. It was determined the majority (~80-86%) of fatal accidents occurred where there was not an intersection, with four-way intersections and T-intersections following as the next highest categories.

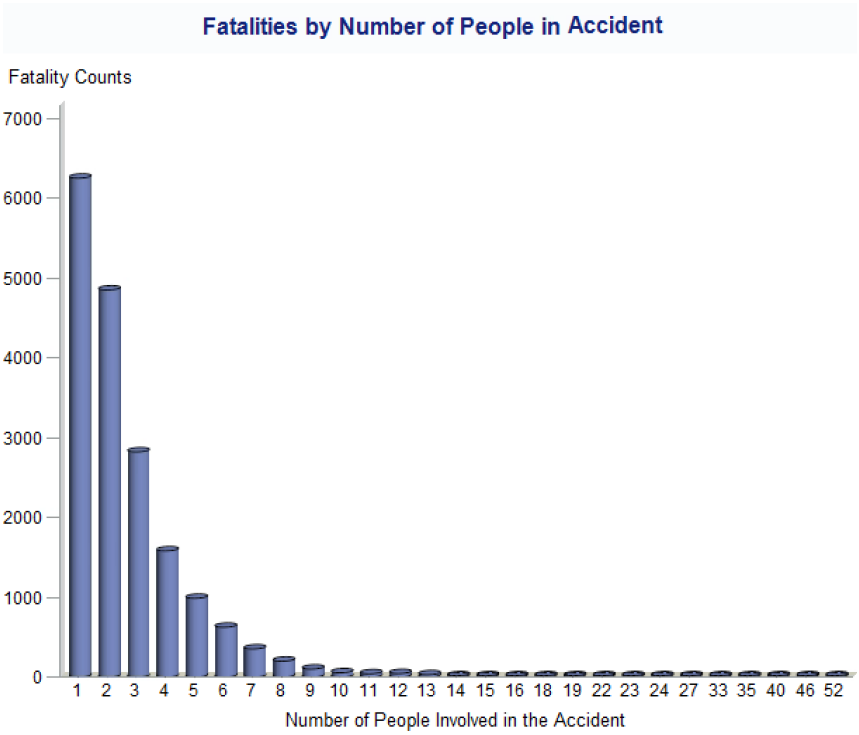






After conducting analysis on several accident factors, such as driving drunk, day of the week, month of the year, year, weather conditions, and intersection type – we found that more fatalities were occurring in accidents where drivers were not drunk, in clear conditions,

and not in an intersection. The fatal accident could occur on any day, and any month of the year. This led us to consider that the likeliness of fatality may be more impacted by the actual driver, than the surrounding conditions. We used the person variable to complete additional visualizations for this impact on fatality counts. The breakdown of how many people were involved in accidents is shown in the graph below, and showed there were more fatalities in accidents that involved less people. When a single person was involved, we had the highest number of fatalities. This could be an indication of the driver having more distractions alone, or potentially being less cautious rather than if they had additional passengers in the vehicle.

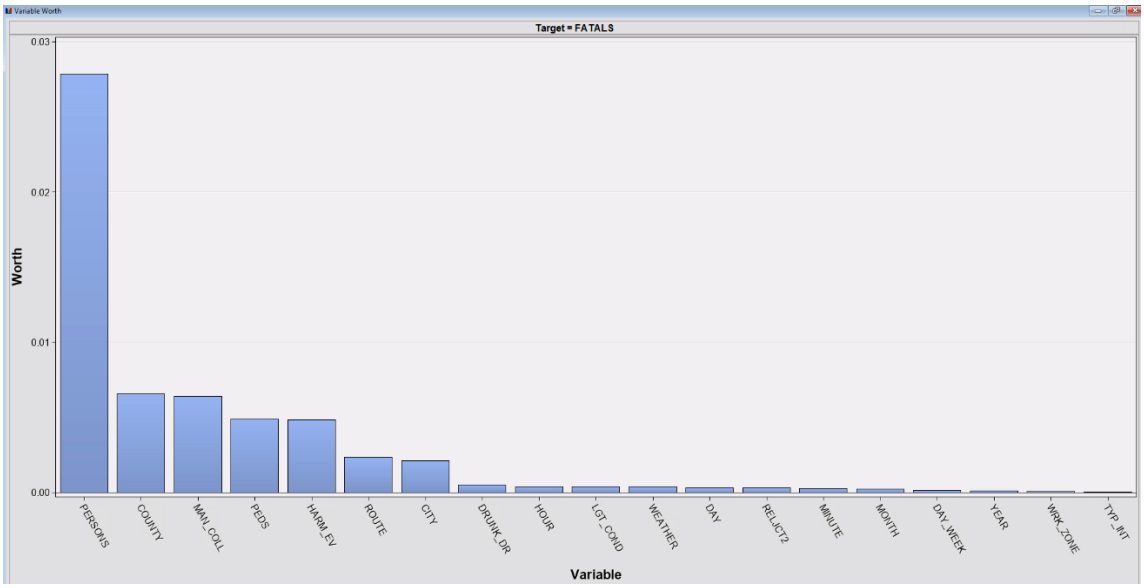


Predictive Modeling

The appended accident data for 2013 to 2017 was imported into SAS Miner for building the final model. One of the critical steps in building a model in SAS Miner is to determine the variable Role and Level. This step has a significant impact on the final model as changes to variable Role and Level result in entirely different final models. The variables with quantitative characteristics were coded as Interval and non-quantitative data were coded as Nominal. We began by running some preliminary analysis to understand our data in order to identify any modifications that might be needed. We started out with exploring our dataset to view our target variable Fatal in more detail. Once this was completed, we ran both a StatExplore and a Regression line so that we could view the various distributions and relationships between our variables to best prepare our data to build the final model.

Name	Role	Level
CITY	Input	Nominal
COUNTY	Input	Nominal
DAY	Input	Nominal
DAY_WEEK	Input	Nominal
DRUNK_DR	Input	Interval
FATALS	Target	Interval
HARM_EV	Input	Nominal
HOUR	Input	Nominal
LATITUDE	Rejected	Interval
LGT_COND	Input	Nominal
LONGITUD	Rejected	Interval
MAN_COLL	Input	Nominal
MINUTE	Input	Nominal
MONTH	Input	Nominal
PEDS	Input	Interval
PERSONS	Input	Interval
RELJCT2	Input	Interval
REL_ROAD	Rejected	Interval
ROUTE	Input	Nominal
STATE	Rejected	Interval
ST_CASE	Rejected	Interval
TWAY_ID	Rejected	Nominal
TWAY_ID2	Rejected	Nominal
TYP_INT	Input	Interval
VE_TOTAL	Rejected	Interval
WEATHER	Input	Nominal
WRK_ZONE	Input	Nominal
YEAR	Input	Nominal
YEAR_ST_CASE	ID	Nominal

Once we had a better understanding of our data, we added the Data Partition node and set the Training Method as Simple Random partitioning. For the breakdown of our Training and Validation allocation, we chose to use 80 percent of the data for training and 20 percent for validation. Since the variable with unknowns were coded either 98 or 99 and the number of unknowns were not significant in each given variable, we determined that it was not nesscessary for us to address or impute missing values. Finally, we moved forward with our model building.



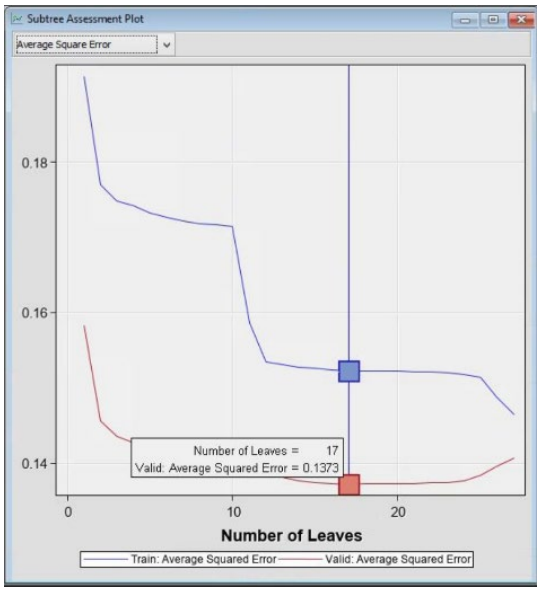
Fit Statistics

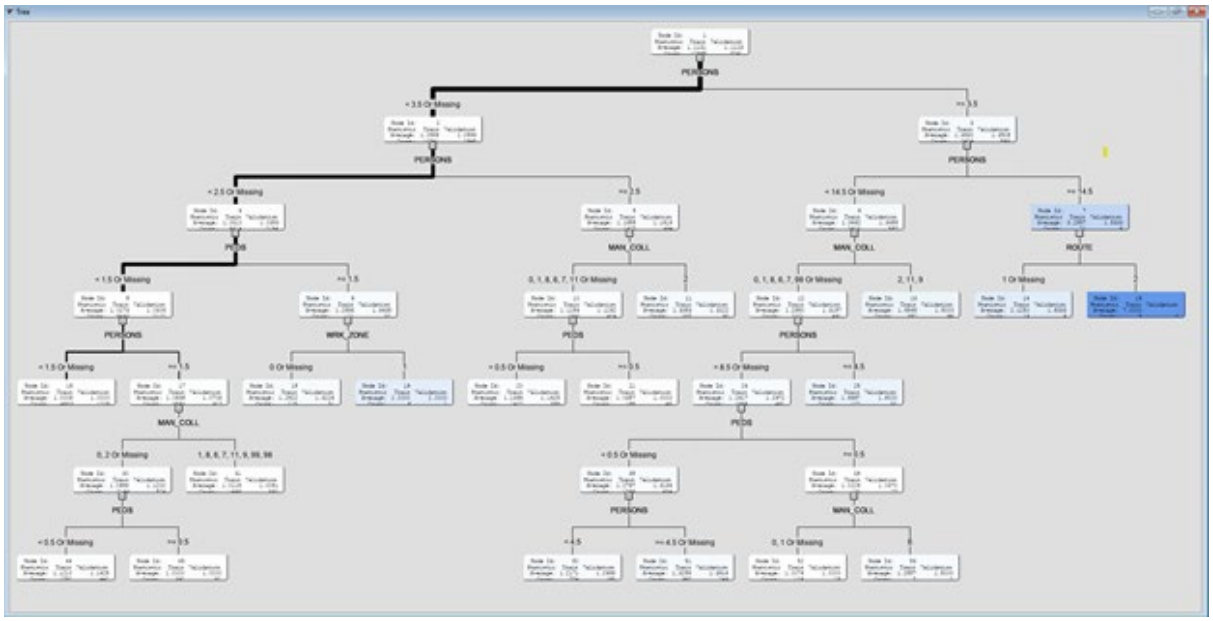
Target	Target Label ▲	Fit Statistics	Statistics Label	Train
FATALS		_AIC_	Akaike's Information Criterion	-29940.5
FATALS		_ASE_	Average Squared Error	0.151416
FATALS		_AVERR_	Average Error Function	0.151416
FATALS		_DFE_	Degrees of Freedom for Error	15880
FATALS		_DFM_	Model Degrees of Freedom	326
FATALS		_DFT_	Total Degrees of Freedom	16206
FATALS		_DIV_	Divisor for ASE	16206
FATALS		_ERR_	Error Function	2453.846
FATALS		_FPE_	Final Prediction Error	0.157633
FATALS		_MAX_	Maximum Absolute Error	10.1776
FATALS		_MSE_	Mean Square Error	0.154524
FATALS		_NOBS_	Sum of Frequencies	16206
FATALS		_NW_	Number of Estimate Weights	326
FATALS		_RASE_	Root Average Sum of Squares	0.389122
FATALS		_RFPE_	Root Final Prediction Error	0.39703
FATALS		_RMSE_	Root Mean Squared Error	0.393096
FATALS		_SBC_	Schwarz's Bayesian Criterion	-27432.5
FATALS		_SSE_	Sum of Squared Errors	2453.846
FATALS		_SUMW_	Sum of Case Weights Times Freq	16206

The models we decided to run were Decision Tree, Linear Regression, Polynomial Regression, and Neural Network. Once each node was configured for our dataset, we reviewed the results of each model to assess its viability. Finally, we entered the Model Comparison node so that we could determine which of the four models had the lowest Average Square Error (ASE) for both training and validation data.

Interpretation of Model Results

The first model we ran was the Decision Tree Model. The initial split assigns the observations Persons involved in the accident by splitting the variables based on the decision boundries less than 3.5 or missing and greater than or equal to 3.5 and then again by 2.5. The group is then further divided by Pedestrians, Work_Zone, Man_Coll respectively. Next, we used the Subtree Assessment Plot to find the tree with the best validation rate had 17 leaves.



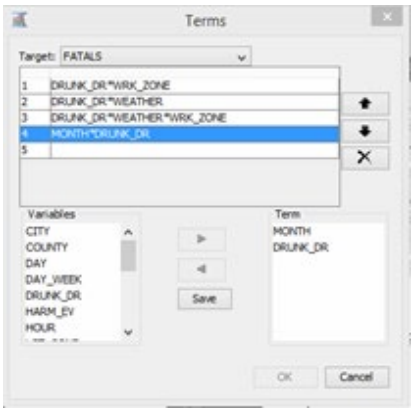


The second model we used to analyze our data was Linear Regression. We ran the model as Stepwise Selection and reviewed the results, which once again showed Persons to be added to the model followed by Manner of Collision and County. We also see the variables Drunk Driver, Pedestrians, Harmful Event, and Route, which also have p-value of less than .01 making these variables significant to the model.

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	F Value	Pr > F	Validation Error Rate
1	PERSONS	1	1	1890.95	<.0001	465.8
2	MAN_COLL	10	2	32.20	<.0001	457.0
3	COUNTY	252	3	2.13	<.0001	479.5
4	DRUNK_DR	1	4	25.22	<.0001	479.9
5	PEDS	1	5	21.79	<.0001	480.5
6	HARM_EV	48	6	2.89	<.0001	480.6
7	ROUTE	8	7	4.66	<.0001	481.5

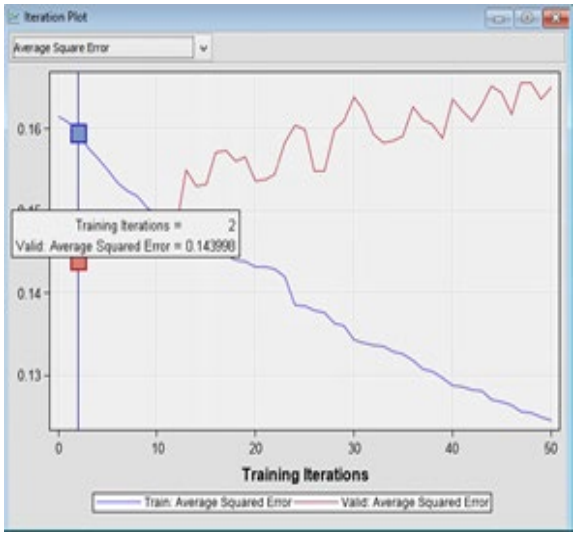
The third model we chose to use was a Polynomial Regression. This model had the same variables as the linear regression. Additionally, we added interactions effects we wanted to test, but only Drunk Driving*Weather*Work_Zone was the interaction effect that made it into the final model and showed any significance. We can see that the interaction effect between Drunk Driver and Work Zone was entered in step four of the Stepwise analysis, but was later removed in step nine as the value no longer had significance to the model. We do see that the Weather variable was also added into this model. The final model contained the variables that have a significant effect on the model with p-values less than .05.



Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	F Value	Pr > F	Validation Error Rate
1	PERSONS		1	1	1890.95	<.0001	465.8
2	MAN_COLL		10	2	32.20	<.0001	457.0
3	COUNTY		252	3	2.13	<.0001	479.5
4	DRUNK_DR*WRK_ZONE		4	4	11.30	<.0001	479.6
5	PEDS		1	5	21.05	<.0001	480.2
6	HARM_EV		48	6	3.03	<.0001	480.5
7	ROUTE		8	7	4.64	<.0001	481.4
8	DRUNK_DR*WEATHER*WRK_ZONE		15	8	1.75	0.0353	482.2
9		DRUNK_DR*WRK_ZONE	4	7	1.56	0.1818	482.2
10	WEATHER		12	8	2.49	0.0029	482.7

The last model we used was the Neural Network due to its ability to compare multiple elements at a given time. This would help us make sure that every possible combination was considered in our model. Even though the performance of the training sample becomes better with additional iterations, the validation performance is the best at iteration 2 as indicated by the plot, resulting in a Validation ASE of 0.143998.

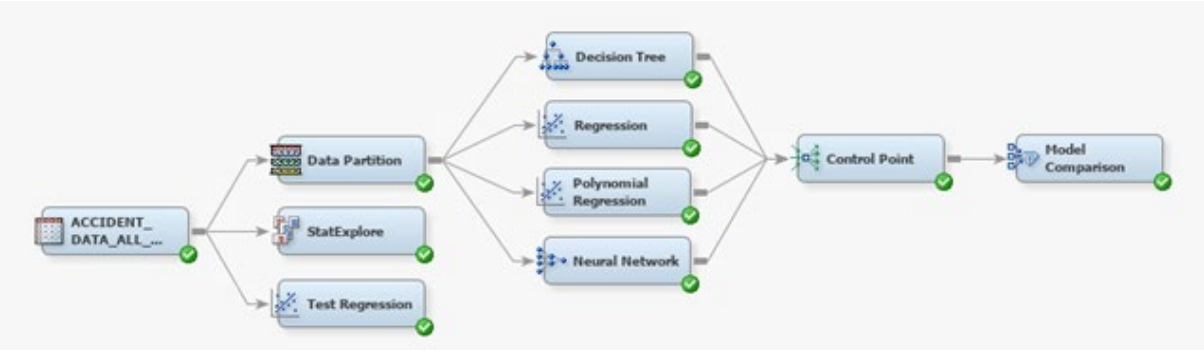


After all of our models were made, we used the Model Comparison node to find the lowest Validation ASE. According to the model comparison, the Decision Tree was the best fit for our data set with a Validation ASE of 0.13725.

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Train: Average Squared Error	Valid: Average Squared Error
Y	Tree	Tree	Decision Tree (2)	FATALS	0.152329	0.13725
	Neural	Neural	Neural Network	FATALS	0.159499	0.143998
	Reg	Reg	Regression	FATALS	0.163031	0.140995
	Reg3	Reg3	Polynomial Regression	FATALS	0.163031	0.140995

Below is a snapshot of our final model:



Generalizations and Recommendations

Based on our visualizations and model results, Texas remained consistent in the number of fatal accidents year over year. Overall, it is apparent that the Accident dataset in and of itself was not sufficient in making a predictive model on the probability of fatality in an accident. As can be seen in each model, Person variable was of significance in the final results of the model. However, since the Accident dataset only reports accidents that resulted in fatality, our final model cannot be used on an unknown dataset to predict fatality. When we apply this predictive model to general accidents that include both fatal and non-fatal accidents, the chances are that the model will predict that most accidents will result in a fatality. A critical flaw in this dataset is that it is not representative of the entire population of vehicle accidents, i.e. fatal as well as non-fatal accidents; therefore, this will result in an inaccurate predictive model.

Given the limitation of the dataset, we still assumed that accidents involving a drunk driver would result in more fatalities than non-drunk driving accidents; however, only 27.6% of the fatal accidents involved drunk driving. Additionally, we assumed that weather conditions would have an impact on fatalities, but our results showed otherwise. This was also the case, for days of the week, and month of the year. We found that single-occupant drivers were more likely to be in a fatal accident than vehicles with multiple occupants. This can potentially be attributed to drivers being more cautious about the safety of the passengers in the vehicle.

In conclusion, our analysis provided us with valuable information that can be used to raise awareness surrounding drinking and driving. We would recommend the Texas Department of Transportation (TxDOT) continue to increase awareness on the high number of fatalities in an effort to lower the accident fatality rate. We also recommend to continue publishing the statistics from their data analysis and research. As drivers ourselves, we are aware of the TxDOT electronic signs providing numbers and facts about accidents and fatalities, which is already increasing awareness.

References

NHTSA. (n.d.). NHTSA FTP. Retrieved from https://www.nhtsa.gov/node/97996/251

NHTSA. (n.d). NHTSA FARS Analytical User's Manual. Retrieved from https://www.nhtsa.gov/node/97996/119141

Appendix
Accident and Vehicle Data Attribute Column Guide

Accident Data Set

- **STATE –** Identifies state crash occurred using GSA Geographic Location Codes (GLC)

  Attribute Codes

  1. Alabama
  2. Alaska
  3. (Blank)
  4. Arizona
  5. Arkansas
  6. California
  7. (Blank)
  8. Colorado
  9. Connecticut
  10. Delaware
  11. District of Columbia
  12. Florida
  13. Georgia
  14. (Blank)
  15. Hawaii
  16. Idaho
  17. Illinois
  18. Indiana
  19. Iowa
  20. Kansas
  21. Kentucky
  22. Louisiana
  23. Maine
  24. Maryland
  25. Massachusetts
  26. Michigan
  27. Minnesota
  28. Mississippi
  29. Missouri
  30. Montana
  31. Nebraska
  32. Nevada
  33. New Hampshire
  34. New Jersey
  35. New Mexico
  36. New York
  37. North Carolina
  38. North Dakota
  39. Ohio
  40. Oklahoma
  41. Oregon
  42. Pennsylvania
  43. Puerto Rico
  44. Rhode Island
  45. South Carolina
  46. South Dakota
  47. Tennessee
  48. Texas
  49. Utah
  50. Vermont
  51. Virginia
  52. Virgin Islands (since 2004)
  53. Washington
  54. West Virginia
  55. Wisconsin
  56. Wyoming

- **YR_ST_CASE – Unique ID**
- **ST_CASE -** Unique case number assigned to each crash
  - Two characters for State Code followed by four characters for Case Number xxxxxx
- **VE_TOTAL –** Number of vehicles involved in crash (includes parked cars if applicable)
  - 1-999 Number of vehicles in crash
- **PEDS –** Number of case forms submitted for persons (non-occupants of any vehicle) involved in crash
  - 0-99 Number of persons not in motor vehicles
- **PERSONS –** Counts number of occupants in vehicles in crash ** In hit and run cases where driver and occupants are not known, coded as unknown
  - 0-999 Number of Person Forms
- **COUNTY –** County where crash occurred using GLC codes
  - 0 - Not Applicable
  - 1-996 - Use GSA Geographical Codes

    - 997 - Other
    - 998 - Not Reported
    - 999 - Unknown
- **CITY –** City where crash occurred using GLC codes
    - 0 - Not Applicable
    - 1-9996 - GSA Geographical Codes
    - 9997 - Other
    - 9898 - Not Reported
    - 9999 - Unknown
- **DAY –** Day of crash
    - 1-31 - Day of the month of the crash
    - -- Unknown
- **MONTH –** Month of crash
    - 1 - January
    - 2 - February
    - 3 - March
    - 4 - April
    - 5 - May
    - 6 - June
    - 7 - July
    - 8 - August
    - 9 - September
    - 10 - October
    - 11 - November
    - 12 - December
    - -- Unknown
- **YEAR –** Year of crash
- **DAY_WEEK –** Day of the week of crash
    - 1 - Sunday
    - 2 - Monday
    - 3 - Tuesday
    - 4 - Wednesday
    - 5 - Thursday
    - 6 - Friday
    - 7 - Saturday
    - -- Unknown
- **HOUR –** Hour (TIME) crash occurred
    - 0-23 - Hour
    - -- Not Applicable or Not Notified
    - 99 - Unknown
- **MINUTE –** Minute (TIME) crash occurred
    - 0 -59 - Minute
    - -- Not Applicable or Not Notified
    - 99 - Unknown
- **ROUTE -** Identifies the route signing of the trafficway on which the crash occurred
    - 1 - Interstate
    - 2 - U.S. Highway
    - 3 - State highway
    - 4 - County Road
    - 5 - Local Street - Township
    - 6 - Local Street - Municipality
    - 7 - Local Street - Frontage Road
    - 8 - Other
    - 9 - Unknown
- **TWAY_ID –** Trafficway on which crash occurred (actual posted number, assigned number, or common name)
- **TWAY_ID2 –** Trafficway on which crash occurred; added beginning 2004 when to accommodate intersection related crashes where officer provides identifier for second trafficway
- **LATITUDE –** Latitude position of crash location using Global Position coordinates
- **LONGITUDE –** Longitude position of crash location using Global Position coordinates

- **HARM_EV –** Describes the first injury or damage producing the event of the crash. First Harmful Event applies to the crash not the vehicle and is based on best judgement of FARS analyst (1-99)
- **MAN_COLL -** Describes the orientation of two motor vehicles in-transport when they are involved in the "First Harmful Event" of a collision crash. If the "First Harmful Event" is not a collision between two motor vehicles in-transport it is classified as such.
    - 0 - Not collision with motor vehicle in transport
    - 1 - Front to rear
    - 2 - Front to front
    - 6 - Angle
    - 7 - Sideswipe - Same direction
    - 8 - Sideswipe - Opposite direction
    - 9 - Rear to side
    - 10 - Rear to rear
    - 11 - Other (End swipes and others)
    - 98 - Not reported
    - -- - Unknown
    - 99 - Reported as unknown
- **RELJCT2 –** Identifies location of crash with respect to junction or interchange area
    - 1 - Non junction
    - 2 - Intersection
    - 3 - Intersection related
    - 4 - Driveway access
    - 5 - Entrance/ Exit ramp related
    - 6 - Railway grade crossing
    - 7 - Crossover related
    - 8 - Driveway access related
    - 16 - Shared use path crossing
    - 17 - Acceleration/deceleration lane
    - 18 - Through roadway
    - 19 - Other location within interchange area
    - 20 - Entrance/exit ramp
    - 98 - Not reported
    - -- - Unknown
    - 99 - Reported as unknown
- **TYP_INT –** Type of intersection
    - 1 - Not an intersection
    - 2 - Four way intersection
    - 3 - T -intersection
    - 4 - Y intersection
    - 5 - Traffic circle
    - 6 - Roundabout
    - 7 - Five point, or more
    - 10 - L - intersection
    - 98 - Not reported
    - 99 - Unknown
- **WRK_ZONE –** Identifies if crash occurred in a work zone area. If crash is identified as a "Work Zone Accident" the type of work activity is identified
    - 0 - None
    - 1 - Construction
    - 2 - Maintenance
    - 3 - Utility
    - 4 - Work zone, Type unknown
    - -- Not reported
- **REL_ROAD -** Identifies the location of the crash as it relates to its position within or outside the trafficway based on the "First Harmful Event."
    - 1 - On roadway
    - 2 - On shoulder
    - 3 - On median
    - 4 - On roadside
    - 5 - Outside trafficway
    - 6 - Off roadway - location unknown
    - 7 - In parking lane/zone
    - 8 - Gore
    - 10 - Separator
    - 11 - Continuous left turn lane
    - 98 - Not reported
    - 99 - Unknown
- **LGT_COND –** Reports the type and level of light that existed at the time of the crash

    - 1 - Daylight
    - 2 - Dark, not lighted
    - 3 - Dark, lighted
    - 4 - Dawn
    - 5 - Dusk
    - 6 - Dark, unknown lighting
    - 7 - Other
    - 8 - Not reported
    - 9 - Unknown
- **WEATHER –** Prevailing atmospheric conditions that existed at the time of the crash
    - 0 - No additional atmospheric conditions
    - 1 - Clear
    - 2 - Rain
    - 3 - Sleet, hail
    - 4 - Snow
    - 5 - Fog, smoke, smog
    - 6 - Severe crosswinds
    - 7 - Blowing sand, soil, dirt
    - 8 - Other
    - 10 - Cloudy
    - 11 - Blowing snow
    - 12 - Freezing rain or drizzle
    - 98 - Not reported
    - 99 - Unknown
- **FATALS –** Number of fatalities that occurred in the crash
    - 1-99 - Number of fatalities that occurred in the crash
- **DRUNK_DR –** Identified number of drinking drivers in accident. Driver is included as drinking if tested positive for alcohol presence; not only those whose BAC tests over legal limit. ANYONE with alcohol presence (drivers only) is counted. (0-99)
    - 0 - No drinking
    - 1 - Drinking
    - -- Unknown

    ** The change to a three-digit BAC in 2015 means that a BAC of .001 or greater qualifies as a drinking driver whereas prior to 2015 a BAC of .01 or greater qualified as a drinking driver. This may have ramifications for trend analyses.