**Predictive Analysis of traffic violations**

R. Bikmetov

UNC Charlotte  
9201 University City Blvd,  
Charlotte, NC 28223, USA  
Tel (980) 277-1027

[rbikmeto@uncc.edu](mailto:rbikmeto@uncc.edu)

A.Raghuvanshi

UNC Charlotte  
9201 University City Blvd,  
Charlotte, NC 28223, USA  
Tel (609) 251-3541

[araghuva@uncc.edu](mailto:araghuva@uncc.edu)

S.Pothina

UNC Charlotte  
9201 University City Blvd,  
Charlotte, NC 28223, USA  
Tel (704) 493-7555

[spothina@uncc.edu](mailto:spothina@uncc.edu)

S.Narasaraj

UNC Charlotte  
9201 University City Blvd,  
Charlotte, NC 28223, USA  
Tel (704) 794-4485

[snarasar@uncc.edu](mailto:snarasar@uncc.edu)

D.Mehta

UNC Charlotte  
9201 University City Blvd,  
Charlotte, NC 28223, USA  
Tel (704) 236-2921

dmehta9@uncc.edu

J. Park

UNC Charlotte  
9201 University City Blvd,  
Charlotte, NC 28223, USA  
Tel (980) 202-9341

jpark97@uncc.edu

**ABSTRACT**

**This paper produced results based on traffic violation data that was updated daily in Montgomery County in the USA. Using the data set, we analyzed the effect of a traffic violation on traffic accidents by using various big data analysis techniques. In particular, three modeling hypotheses have been developed based on initial data understanding and the performed 5 visualizations. The required data preprocessing has been performed to address these hypotheses using 3 models with different algorithms.**

**Several classifications, clustering and regression algorithms have been considered in our analysis, such as Single Tree, Random Trees, K-Means clustering, Multiple regression, etc. Due to an importance of high sensitivity and a high F1-score, Random Trees and Single Trees have been considered as the best algorithms for the first and second model according to our business case. K-means clustering and Single Tree classification have been considered as the best algorithms for model 3 being able to provide numbers of different types of violations along with the number of injuries for various clusters.**

**Based on the obtained results, 3 strategic recommendations have been provided to increase the awareness of traffic regulating authorities and improve traffic safety.**

**Keywords**

Knowledge Discovery, Data Mining, Traffic Violations Analysis, Predict Traffic Accident

# INTRODUCTION

During the last decade, different data analytics and mining techniques have become useful in various research and industrial sectors [1-5]. Extending the range of applications for these techniques is ever evolving task challenging modern data scientists. Besides, investigating domain specific problems often requires enhancements to current data analytics and mining practices. Application of data mining techniques for traffic control and regulation is one of the common trends [2-5].

It is known that road accidents present a big worldwide threat that continues to cause losses, injuries, and fatalities on road ways resulting in a huge impact at the economic and social levels. Specifically, accidents cause about 1.27 million deaths worldwide per year and up to 50 million injuries (de Ona, Lopez, & Abellan, 2013). Hence, such global problem requires more attention for reduction of frequency and severity of accident occurrence.

The historical data about previous traffic violations represents an arduous opportunity for researchers to recognize the most significant factors in such violations. The main challenges in extracting knowledge from this data are its huge size and high dimensions [2-4]. Recently, a number of data-mining techniques have been efficiently utilized for extraction of useful information from big data sets about traffic accidents [3, 5, 6]. Classification methods were commonly used techniques in data-mining road traffic accidents (and other violations). The main purpose of these methods was constructing classifiers for predicting of new accidents and their severity. It was shown that the accuracy of such classifiers strongly depends on the collected data characteristic [3, 5, 6]. This aspect makes it practically impossible to recommend a certain classifier for certain type of problems. Therefore, it is a common practice to test multiple classifiers before deciding to use one or the other for a particular problem and corresponding data set [2-7].

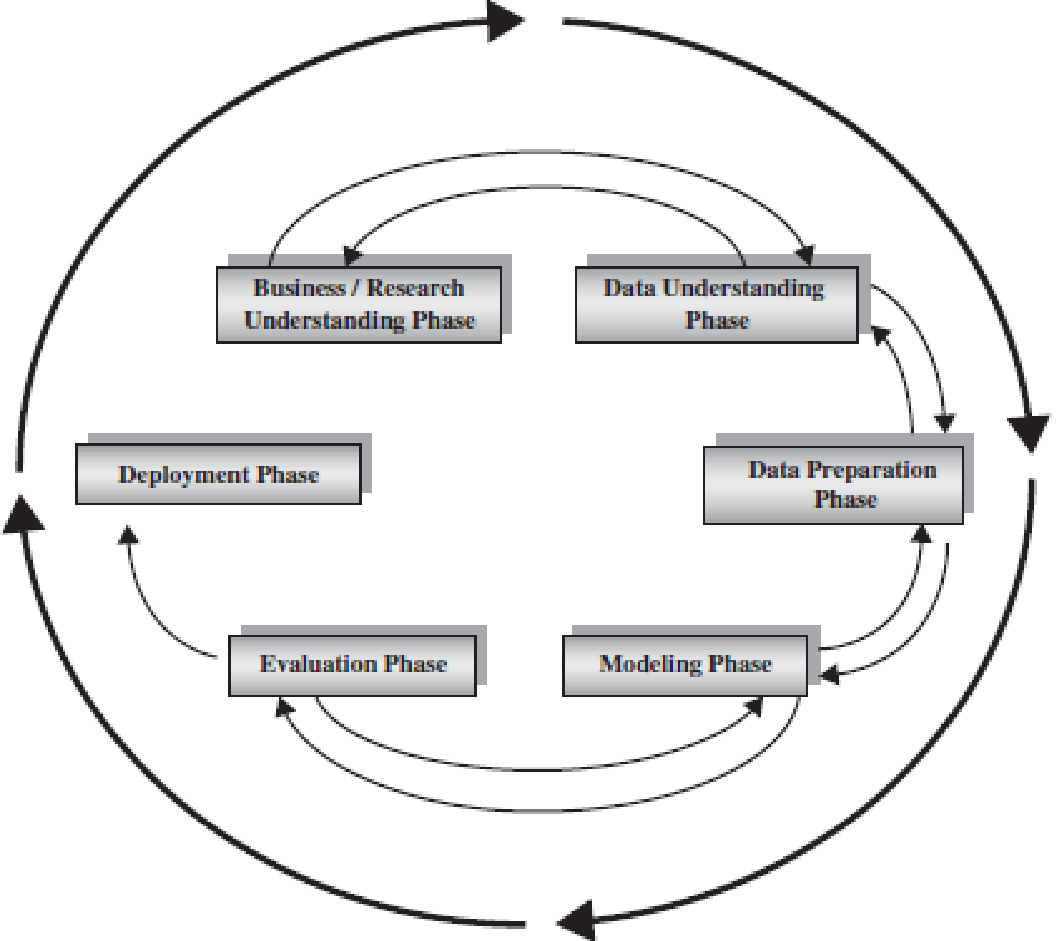


Fig. 1 – Six phases of iterative and adaptive CRISP-DM [1]

In this work, we perform our analysis of chosen dataset “Traffic Violations” following well-known Cross-Industry Standard Process for Data Mining (CRISP-DM) (Fig. 1), which allows fitting data mining problem into the general problem-solving strategy of a research unit [1]. The project research tasks were formulated in the project assignment description […]. The main tasks of it are selecting an available on web dataset, study its data and draw preliminary conclusions from it, formulate the initial hypotheses based on data analysis, and develop the business case. The other important tasks of the project are exploring dataset and producing 5 visualizations, review the hypotheses based on visualizations, produce dataset(s) satisfying the produced hypotheses. At the final tasks are to present 3 modeling techniques for developed hypotheses, implement 3 algorithms for each model, and provide strategic recommendations based on modeling results.

Following CRISP-DM framework and project research tasks, Research Understanding, Data Understanding and Visualization have been performed in sections 2, 3, and 4 respectively. After that, Data Preparation and Modeling Phases have been completed in sections 5 and 6 respectively. According to the assigned research task, Modeling phase has been completed by application of different well-known data-mining techniques. Besides, the investigation of the performance of several data-mining algorithms for the initially formulated hypotheses has been performed. Finally, Validation and Deployment phases are accomplished in accordance with the CRISP-DM framework and the project assignment in sections 6 and 7 respectively. Besides, the strategic recommendations have been developed in section 8.

# RESEARCH UNDERSTANDING

The objective for the project is to analyze a dataset on Traffic Violations of Maryland State to improve awareness of Maryland Capitol Police (MCP) and to help MCP in reducing the number of violations based on their location, injury and fatality rates, number of property damages, etc. At the same time, the information about the frequency of violations caused by out-of-state drivers can be used by MCP for sharing with the police departments of the other states for their considerations. The data is created by the MCP and hosted through data.gov. MCP is focused on ensuring proper safety measures to decrease the number of violations happening in its state by tightening various traffic rules at places where violation frequency is high. We believe that it is important to advise Maryland Capitol Police about the number of violations happening at various locations based on the data provided and frequency of particular violations over the time. Moreover, we can predict the likelihood of violations turning out into an accident, which involved injury or property damage. We also can estimate seasonal violations by bringing out conclusions, such as the season with the highest rate of violations or the season with the highest rate of violations involved injuries or property damages. Finally, we can conduct analysis on a portion of violations involved commercial vehicles. This information can give suggestions for tightening certain traffic rules and regulations for commercial vehicles only. As the overall result, provided information and insights will be helpful to MCP to ascertain about proper safety measures and decrease the number of violations in the state by tightening certain traffic rules at places where violation frequency is high. Maryland Capitol Police (MCP) has asked us to use our skills with data to help to reduce the number of violations based on their location, injury and fatality rates, number of property damages, and other factors. MCP is willing to redistribute their available police personnel and take several extra measures to improve traffic safety and reducing the number violations, injuries, and property damages.

# DATA UNDERSTANDING

The chosen dataset “Traffic Violations” was obtained from <https://catalog.data.gov/> which was recently updated on January 17, 2018. This dataset consists of traffic violation data from 2012 to 2018 from all the electronic traffic violations issued in Montgomery County of Maryland state. This data provides information about all the traffic violations in that county, their impact, and the place at which they happened. Moreover, information related to the vehicle and its type as well as its driver is also provided. The description of non-descriptive variables is given in table 1.

# In order to find the variables that have the greatest effect on the automobile accident, we made three search questions by analyzing the characteristics of the data set shown in the data search and created a model for verifying the question. Before making the model, we tried to find out whether the use of safety belts, alcohol consumption, use of the phone while driving, and type of automobile would affect the accident through data search. We designed models to verify the search questions using the algorithms, which are frequently used in the big data techniques.

Table 1. Non-descriptive variables

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| Accident | If traffic violation involved an accident. |
| Belts | If traffic violation involved a seat belt violation. |
| Personal Injury | If traffic violation involved Personal Injury. |
| Property Damage | If traffic violation involved Property Damage. |
| Fatal | If traffic violation involved a fatality. |
| HAZMAT | If the traffic violation involved hazardous materials. |
| Vehicle Type | Type of vehicle (Examples: Automobile, Station Wagon, Heavy Duty Truck, etc.) |
| Violation Type | Violation type. (Examples: Warning, Citation, SERO) |
| Charge | Numeric code for the specific charge. |
| Geolocation | Geo-coded location information. |

We derived the following initial research questions/hypotheses through the dataset

● Using the information about traffic violations collected over 6 years, the future violations can be predicted using attributes such as the likelihood of accidents, and geo-coordinates of the location, etc. This prediction will be useful to aware transportation authorities, who can modify and reform the existing laws of traffic regulation to improve the traffic safety.

● The chosen dataset can be analyzed to obtain the information about the estimated number of future violations that can lead to personal or fatal injury and property damage. Such information will be helpful for insurance companies, who provides their quotes based on the historical statistical data.

● The obtained information about the frequency of different types of violations and corresponding charges will help police and other traffic regulation authorities to optimize their operation of traffic control.

After data visualization and data preparation tasks are completed, these initial hypotheses will be further updated. The obtained updated hypotheses will be used to implement various modeling techniques with the application of several algorithms.

# DATA VISUALIZATION

# Based on the requirement, several visualizations have been developed to increase the understanding of the initial dataset, communicate insights and interesting findings from it. The presented visualization has been performed in Tableau software.

# Visualization of violations based on time is shown in Fig. 6 in the appendix. A number of traffic violations observed from the dataset, clearly explains the number of violations during the hours between 22:00 (10 pm) and 00:00 (12 am) is very large as compared to the number of violations during other hours. Blue color represents the number of violations in the daytime, and orange color represents the number of violations in the nighttime. During daytime, we can observe that the number of violations is highest during morning peak hours 08:00 - 09:00 am. Thus, the visualization clearly explains that MCP has to put extra precautions (e.g., extra personnel, security cameras, etc.) at night hours and early morning peak hours.

Visualizations of vehicles more prone to the violation are shown in Fig. 7 in the appendix.The visualization clearly shows; the larger the circle, the more the number of violations occurred for that vehicle type. Based on the chart, we can clearly state that violations by automobile are the highest and light truck vehicles share the second highest percentage of a number of violations in Maryland.

Visualizations of violations based on the seasonare shown in Fig. 8 in the appendix.This visualization is performed to explain the number of violations based on four seasons throughout the year as a *Heatmap.* Darker colors show higher numbers in a heatmap. It can be noticed that the average number of violations doesn’t have big change between different season. At the same time, it can be noticed that the number of violations is the highest during the fall month and second highest during winter months. Both of these seasons are shown in darker blue colors. Seasons with the lower numbers of violations are shown in light colors.

Visualizations of violations involved child based on year and month are shown in Fig. 9 in the appendix. This graph shows the number of violations related to children (failure to secure the child with a belt, transporting child without a proper child seat, etc) for each year and month. We found that the highest number of violations is in May for each year. Additionally, the number of these violations in the indicated period was drastically reduced starting from 2016 and on. Such reduction can be the consequence of accident prevention activities that were probably started from 2016.

Visualizations of phone violations with year and month are shown in Fig. 10 in the appendix.The timeline graph shows violations by mobile device usage for the corresponding year and month. It can be seen from the graph that the average number of violations is several times increased starting from Dec 2013. Besides, the graph shows that there is a peak of cell phone usage violations in April each year. It can also be seen that this peak was decreasing up to Apr 2016 and stays the same after that. This information can give awareness and insights for the future actions of the MCP.

Visualization of personal injury based on seat belt violations is shown in Fig. 11 in the appendix. The visualization below depicts, the number of violations involved not wearing the seat belt and leading (or not leading) to personal injury. The number of violations where the driver did not wear the seat belt and was injured is shown in yellow color, the area covered for the case is small, because the dataset is huge with violations based on a number of other features. Even though the highest number of violations didn’t cause any injuries with no seat belt, it can be seen that the number of injuries for people wearing seat belt is about 5 times lower than for the people not wearing it. MCP can consider the seat belt enforcement, which can save 100’s of people.

Based on the performed visualizations, the following two research questions/ hypotheses can be presented:

1. The number of violations involved phone usage is still high in the month of April and stays relatively constant during the last two years. What locations contribute the most to phone usage violations? The awareness of MCP can be increased to these locations.

2. The number of violations is very high during 10 pm - 12 am and relatively high at 8 am - 9 am. What are the main factors that contribute to violations during this time? The awareness of MCP and traffic authorities can be increased to these types of violations.

After analyzing and combining the initial hypotheses formulated after data understanding and data visualization tasks, the following three final hypotheses have been formulated:

1. For a given combination of factors (i.e., values of predictor variables), predict whether a violation occurred contributed to an accident or not.
2. For a given combination of factors (i.e., values of predictor variables), predict whether the violation that led to an accident was fatal or not.
3. For a given combination of factors (i.e., values of predictor variables), predict the geolocation, at which a certain violation is likely to occur, also showing the locations with the highest number of certain violations.

To perform the modeling of these hypotheses, a data preparation task should be completed. This task allows to define and introduce the required target variables, properly utilize the required predictors, and avoid any bias that can present in the initial dataset.

# DATA PREPARATION

## Handling inconsistent and missing data

As a part of dataset cleansing and preprocessing, it has been observed and that two attributes “Agency” and “Accident”. The first variable shows the agency created a warning or citation notice. The second variable indicates accident occurrence at the moment of the stop. Since all dataset was taken from MCP, “Agency” variable has only one value. According to it, description (see metadata) “Accident” variable should have Boolean values “yes or no” indicating of occurrence or non-occurrence of an accident at the time of the stop. However, the initial dataset has “no” values for all the records, which does not provide any practical use from data analytics point of view. Hence, both, “Agency” and “Accident” variables have been removed as a single-valued attribute.

## Missing Data

Handling missing data and blank fields is an important part of the data wrangling process. The initial dataset has been processed using XL-miners transformation for missing data handling for null values. Records with null values can be handled based on specific filters available in XLMiner. Since the number of these records in the original dataset was less than 1% of the total number of records, the records with null values were removed.

## Normalization/Standardization

The used dataset has not been normalized and standardized because our analysis and research hypothesis didn’t require the stable convergence of weight for attributes. Besides, the absolute majority of variables in our dataset are categorical. Hence, all our further analysis has been performed without normalization and standardization.

## New Attributes

According to our data understanding based on the performed visualizations and data preparation, the dataset has been updated by the introduction of the following new attributes defined by new variables “Phone usage”, “Contributed to the accident”, and “Fatal”. All of these new variables are binary. Values of “1” for all records of variable “Phone usage” contain “phone” statement in the “description” attribute included in the original data. If the “description” attribute of the original data doesn’t contain the “phone” statement, “Phone usage” variable has its corresponding values as “0”. “Contributed to an accident” and “Fatal” variables can be described in a similar manner. However, the purpose of creating these variables is different from it is for “Phone usage”.

“Contributed to an accident” and “Fatal” have been created as target variables. Whereas, “Phone usage” variable has been used for prediction of a chosen target variable, e.g., “Contributed to an accident” and “Fatal”. In our opinion, this choice can provide a valuable insight to MCP about this type of traffic violation fulfilling the corresponding research requirements.

## Outliers detection

Several outliers have been detected in vehicles’ “Year” variable. These outliers were presented by non-standard values (irregularities) and further replaced by imputing or deleted. Imputing have been performed in the following manner: all values of “Year” variable below 1953 have been replaced with the random value of 2000. Whereas, all values of “Year” variable above 2018 have been deleted since there are no currently existing vehicles that have been produced after 2018.

## Dataset filtering

To perform a proper prediction, the dataset should have an equal weightage of target variable values, i.e., the dataset should be unbiased. To achieve this goal, filtering can be applied. R script has been used to filter the dataset and transform it into an unbiased one for further modeling.

The first dataset filtering has been performed to approach research hypothesis predicting “Contributed to an Accident” target variable. In this filtering, all the records that have target variable as true have been selected. In addition to that, the equal number of records with the value of target variable as false has been selected. The same dataset has been used for modeling of hypothesis 3. For the second hypothesis, this process was repeated by considering fatal as our target variable.

## Binning and creating dummies

Binning and creating dummies are the important steps of data analysis process that includes classification tasks [1]. Several algorithms prefer binned categorical rather than continuous numerical variables [1]. According to this principle, variable “Year” has been binned into 10 equal intervals based on its original value, thus, transformed from numerical into a categorical variable. The results of binning are presented in the appendix. In order to obtain binary categorical variables from the corresponding initial data, dummies have been created for attributes “Belts”, “HAZMAT”, “Commercial Vehicle”, “Alcohol”, “Phone Usage”, and the previously binned variable “Year”. All these variables received values “1” and “0” in accordance with their description (see metadata). The results of the binning procedure are shown in Fig. 13 and Fig. 14 in the appendix.

# MODELING TECHNIQUES

After the final hypotheses have been developed and corresponding data preparation has been completed, the development of three modeling techniques has been performed according to the project requirements. These modeling techniques have been developed investigating the performance of several classification algorithms in predicting the corresponding hypotheses. The work focuses on classification algorithms due to and initial research described in section 1.

## Model 1

In the first model, prediction of contribution to an accident for a given scenario is performed (see hypothesis 1) by classification technique using three well-known algorithms: Single Tree, Random Trees, and Naive Bayes. In this model, “contribution to an accident”, a new variable created during data preparation, has been utilized as a target variable. Categorized variables “Belts”, “Alcohol”, and “Phone usage” as well as binned variables “Vehicle Type” and “Year” have been used as predictors. The obtained best-pruned tree is shown in Fig. 2. The corresponding full tree is shown in Fig. 12 in the appendix.



Fig. 2 – Best-pruned tree of classification method based on Single Tree algorithm. The method is used to predict “contribution to an accident” target variable.

It can be seen that the top predictors for this model are “Belts” and “Phone usage” variables. The performance of the developed model can be assessed using parameters shown in Table 2. It can be seen from the Table 2 that Single Tree algorithm has the highest precision. Whereas, Random Trees algorithms have the highest sensitivity and F1-score. In order to assess the performance of the above algorithms correctly, it is important to recall the meaning of the above parameters (precision, sensitivity, specificity, and F1-score) in accordance with standard evaluation metrics [5-9]. Additionally, depending on particular prediction requirements, different performance parameters can be considered for model assessment and evaluation of the results obtained from it. The highest sensitivity of Random Trees algorithm shows the highest proportion of contribution to accident cases that are correctly identified. Since sensitivity is an important parameter for the developed business case and hypothesis 1, Random Trees algorithm will be preferable for this hypothesis. Specifically, the awareness of accident contribution is the highest priority for prediction task to be modeled. The highest F1-score of Random Trees algorithm gives the optimal combination of sensitivity and precision showing the proportion of contribution to accident cases that were correctly identified. The highest precision of the Single Tree algorithm becomes more useful for the business case, which requires a high general prediction performance.

Table 2 – Performance parameters of Model 1

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Single Tree** | **Random Trees** | **Naïve Bayes** |
| Precision | 0.626 | 0.503 | 0.621 |
| Sensitivity | 0.275 | 0.996 | 0.266 |
| Specificity | 0.836 | 0.023 | 0.838 |
| F1-Score | 0.383 | 0.669 | 0.373 |

## Model 2

In the second model, prediction of accident fatality for a given scenario (see hypothesis 2) is performed by classification technique using the same algorithms as in model 1. In model 2, “fatal”, a new variable created during data preparation, has been utilized as a target. In the beginning, the predictor variables have been chosen the same as in model 1, besides “Vehicle Type”. The performance of the developed model can be assessed using parameters shown in Table 3. Similar to the analysis described in model 1, different performance parameters can be considered for model assessment based on the particular prediction requirements. It can be noticed that Single Tree algorithm has the highest sensitivity and F1-score parameters for this model. Whereas, Random Trees algorithm has the highest precision.

Table 3 – Performance parameters of the initial Model 2

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Single Tree** | **Random Trees** | **Naïve Bayes** |
| Precision | 0.579 | 0.639 | 0.579 |
| Sensitivity | 0.79 | 0.383 | 0.543 |
| Specificity | 0.404 | 0.776 | 0.59 |
| F1-Score | 0.668 | 0.478 | 0.561 |

The best-pruned tree obtained for model 2 is shown in Fig. 3. The corresponding full tree is shown in Fig. 15 in the appendix. It can be seen that the top predictors for this model are “Binned\_Year\_9”, “Binned\_Year\_10”, and “Alcohol” variables (see appendix for binned variables description). The best-pruned tree shows “Binned\_Year\_4” as an additional top predictor. These results have been possibly obtained due to a large number of dataset records, which belong to 9th and 10th binned interval. Based on our hypothesis 2, the obtained classification results are not very useful, because they don’t provide good insights and awareness about certain combinations of factors that led to an accident that would be fatal or not. In other words, it is required to identify the various types and combinations of violations that would lead to an accident. In the current case, the years of vehicles, defined by “Binned\_Year” and being the main predictor variable, cannot properly address hypothesis 2. Hence, the results of this modeling cannot be used by MCP to take certain measures to reduce the number of fatal accidents.

In order to improve the predictive performance of model 2 according to the developed business case, variable “Year” has been removed from the prediction analysis. The obtained best-pruned tree had a commercial vehicle as the main predictor of fatalities, which occurred due to the bias of data. Since vehicle type dependence was not among the goals for our hypothesis 2, this variable has been removed and the modeling process has been repeated. The obtained best-pruned tree had only single node showing bias in data towards the variable “Year”. However, the corresponding full tree shown in Fig. 4 indicate that the top predictors for the updated model are “Alcohol”, “Phone” and “HAZMAT” variables. These results can be useful for our developed prediction tasks because they show the particular violations and situations that can be controlled by MCP, which becomes more aware of them based on our analysis. The performance of the updated model can be assessed similarly to model 1 using parameters shown in Table 4. Prediction of all possible combination of factors that would lead to fatal accidents (see hypothesis 2) is the highest priority of model 2. Therefore, Single Tree algorithm is the best choice for model 2 due to the same reasons that have been presented for the assessment of model 1 performance.

Table 4 – Performance parameters of the updated Model 2

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Single Tree** | **Random Trees** | **Naïve Bayes** |
| Precision | 0.509 | 0.519 | 0.5 |
| Sensitivity | 1 | 0.98 | 0.07 |
| Specificity | 0 | 0.057 | 0.92 |
| F1-Score | 0.675 | 0.67 | 0.12 |

Fig. 3 – Best-pruned tree of classification method based on Single Tree algorithm. The method is used to predict “fatal” target variable predicting fatality of an accident.



Fig. 4 – Full tree of updated classification method based on Single Tree algorithm. The method is used to predict “fatal” target variable predicting fatality of an accident.

## Model 3

In the model 3, the geolocation with a high likelihood of an accident is predicted (see hypothesis 3). This model has been started with clustering of the dataset using K-means and hierarchical clustering algorithms. After that, various classification and regression prediction techniques have been used. K-Means clustering algorithm has been chosen as the best for the first part of the modeling due to better in-cluster and inter-cluster distances and ability to variate the number of clusters using the available computational performance. Considering the area of Maryland state and the number of corresponding police departments, the optimal number of clusters has been chosen to be ten due to an approximately equal weight of overall records. In this model, cluster ID, a new variable created during data preparation, has been utilized as a target variable showing the most probable cluster for a given set of predictors.

After completing clustering, the selection of predictors to address hypothesis 3 has been performed, which is performed using the same classification algorithms as in model 1. The prediction has been performed using the following classification algorithms: Single Tree, Random Trees, and Naive Bayes. Besides, multiple linear and logistic regression algorithms have been applied to satisfy project requirements. However, the developed prediction model using multiple linear and logistic regression algorithms had a very low adjusted R-squared value and a number of errors respectively. Therefore, the results of modeling using regressions haven’t been considered in the evaluation and presented at Fig. 16 in the appendix for general consideration. It has been noticed that the prediction results from Random Trees and Naive Bayes algorithms became biased for developed clustering model. Whereas, Single Tree algorithm has shown sustained results being able to perform classification to the formed clusters. Hence, Single Tree has been selected as the best prediction algorithm for model 3.

The best-pruned tree obtained for Single Tree classification algorithm of model 3 contains a single node due to a higher weight of records from cluster 9. The full tree is shown in Fig. 5 indicates classification score performance based on the developed model. From this performance, it can be seen, for example, that there are several alcohol violations along with personal injuries occurred in cluster 5. At the same time, it can be observed that there are multiple belt violations along with personal injuries occurred in cluster 9. This results can increase the awareness of MCP to clusters 5 and 9.



Fig. 5 – Full tree classification method based on Single Tree algorithm performed after K-means clustering.

# VERIFICATION

The obtained results can be verified using multiple factors: performance parameters metrics (sensitivity, F1-score, etc.), business case, and utilization of different data-mining software (e.g., SAS, Python and Matlab libraries, etc.), which gives an opportunity to utilize a higher number of records than XLMiner and perform a different data preprocessing. The utilization of different data-mining tools is the goal of the future research. The initial verification based on the parameter metrics and the business case is provided below.

For model 1, it can be seen from Table 2 that Random Trees implementation has the most noteworthy affectability, which demonstrates the extent of commitment to accident cases that are accurately recognized. Along these lines, affectability is an essential parameter for business situations when the familiarity with accident commitment is the most astounding need for forecast assignment to be displayed, which is the case in our business scenario. Adjacent to, Random Trees calculation has the most astounding F1-score, which gives the ideal mix of affectability and accuracy demonstrating the extent of commitment to accident cases that were effectively recognized. In the meantime, Single Tree calculation has the most noteworthy exactness, which turns out to be more helpful for general case forecast execution.

The acquired best-pruned tree had just single node demonstrating inclination in information towards the variable "Year". Nonetheless, the relating full tree appeared in Fig. 4 show that the best indicators for the refreshed model are "Alcohol", "Phone" and "HAZMAT" factors. These outcomes model 2 showing can be valuable for our created forecast undertakings since they demonstrate the specific infringement and circumstances that can be controlled by MCP, which turns out to be more mindful of them in view of our examination. The execution of the refreshed model can be surveyed correspondingly to model 1 utilizing parameters appeared in Table 4.

Single Tree algorithm has shown sustained results for model 3 being able to perform classification to the formed clusters. Hence, it has been selected as the best prediction algorithm for model 3. The best-pruned tree obtained for Single Tree classification algorithm of model 3 contains a single node. From this algorithm performance, it can be seen that there are several alcohol violations along with personal injuries occurred in cluster 5. At the same time, it can be observed that there are multiple belt violations along with personal injuries occurred in cluster 9.

# STRATEGIC RECOMMENDATIONS

Based on the hypotheses developed in section 4 and corresponding modeling performed in section 6, the following recommendations can be given to MCP and traffic law enforcement authorities to increase their awareness, modify and reform the existing laws of traffic regulation, and improve the traffic safety:

1. Increase the overall attention and enforcement of rules for belts and phone usage as the main reason for accident contribution based on modeling for hypothesis 1.
2. Pay an extreme attention and take extra measures to drunk drivers and phone usage when driving, since these factors are the top ones for fatal accidents.
3. Alert MCP about the major areas where violations are likely to be caused:

cluster 5: alcohol violations with personal injuries;  
 cluster 9: multiple belt violations with injuries.  
  
 Similar recommendations can be given to the insurance companies to consider for their policies for drivers in analyzed areas.

# DISCUSSION

Our overall business use case is to minimize as many as possible traffic violations, which can lead to corresponding accidents, injuries, and fatalities. The 5 performed visualizations helped us obtain insights about hourly based violations, vehicle type majorly affected the violations, seasonal violations, child involvement as well as insights about how the phone was a major concern in the traffic violations. We produced datasets, which satisfy our main objective, we pre-processed data and remove error and inconsistencies, and etc. The selected dataset for analysis belongs to data.gov website. During the analysis, 2 new columns phone\_usage and child\_involved variables have been created. To address prediction requirements, 3 research hypotheses have been developed. The new target variables that we have derived are “Contribution to an accident”, “Fatal”, and “Cluster ID”.

All of these hypotheses have been aimed to develop the strategic recommendations to increase the awareness of transportation authorities, who can modify and reform the existing laws of traffic regulation to improve the traffic safety. Besides, the developed questions can provide the information about the estimated number of future violations and to optimize the operation of traffic control.

Several classifications, clustering and regression models have been considered in our analysis, such as Single Tree, Random Trees, Naive Bayes, K-Means clustering, Multiple regression, etc. Due to an importance of high sensitivity and a high F1-score, Random Trees and Single Trees have been considered as the best algorithms for the first and second model according to our business case. K-means clustering and Single Tree classification have been considered as the best algorithms for model 3 being able to provide numbers of different types of violations along with the number of injuries for various clusters.

The overall project timeline is shown in Fig. 17 as per requirements.

## FUTURE WORK

Several directions can be taken for future work: extended results validation using other data analytics software, e.g., Python and Matlab libraries, SAS, etc.; data refining, and combined analysis.

From data refining point of view, the “charge” attribute of the initial dataset can be properly documented and synchronized to the particular statements of traffic law, “contribution to an accident” target variable can be specified to a specific traffic violation citation. This measure can help further categorize violations by the specific type and increase the awareness to specific violations.

Combined analysis of various data sets related to traffic will enable more sophisticated data analysis. For example, in order to increase the accuracy of “contribution to an accident” and “fatal” prediction, the particular weather and road conditions should be considered among the predictors.

# ACKNOWLEDGEMENT

We are thankful to our instructor, Dr. Tsay, and our teaching assistants, Ved and Bhanu, for help and guidance through the completion of our project.

# APPENDIX

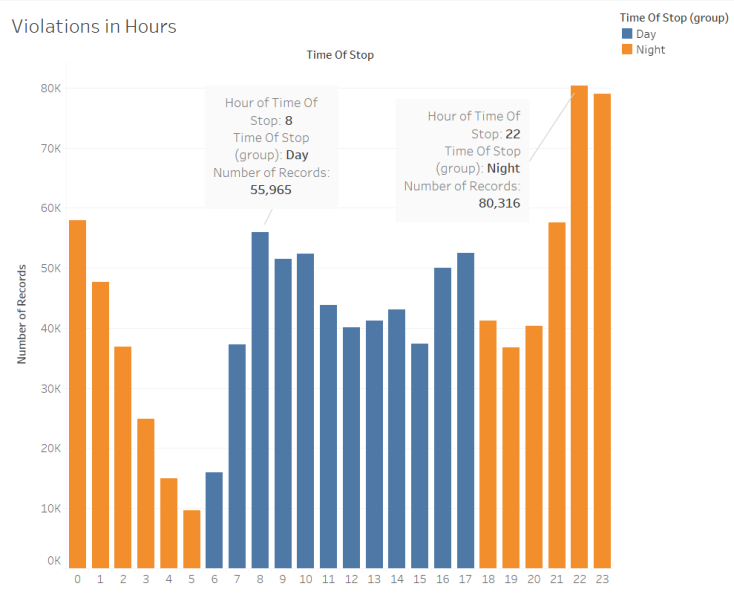


Fig. 6 – Violations in hours

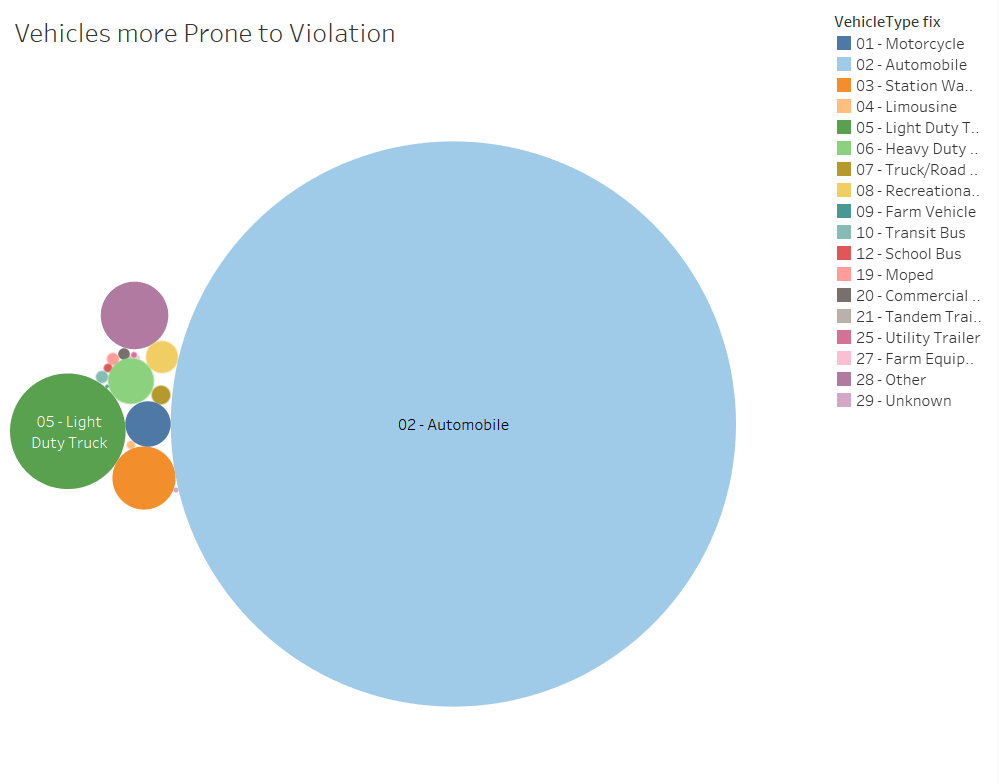


Fig. 7 – Vehicles more prone to violations

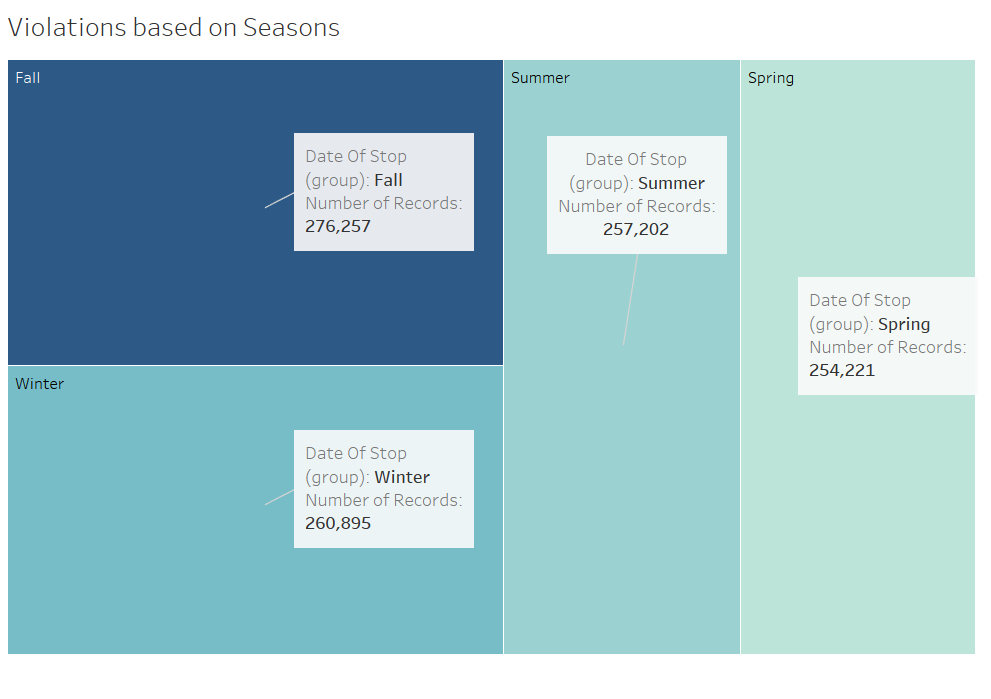


Fig. 8 – Violations based on seasons

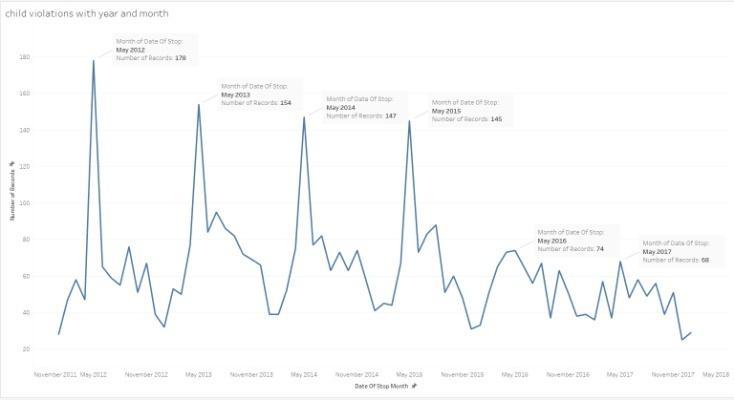


Fig. 9 – Child violations with year and month

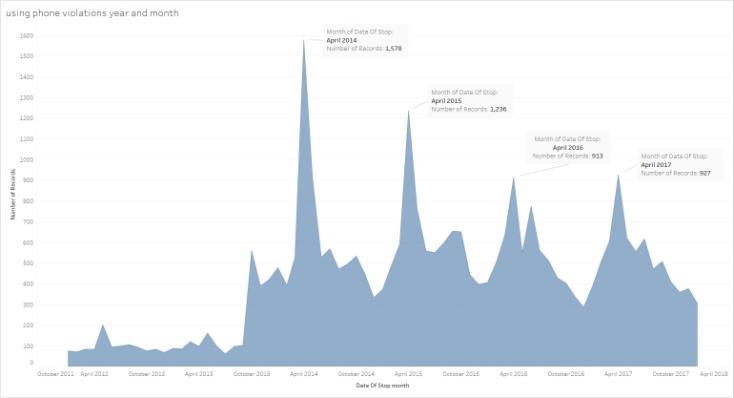


Fig. 10 – Phone violation year and month

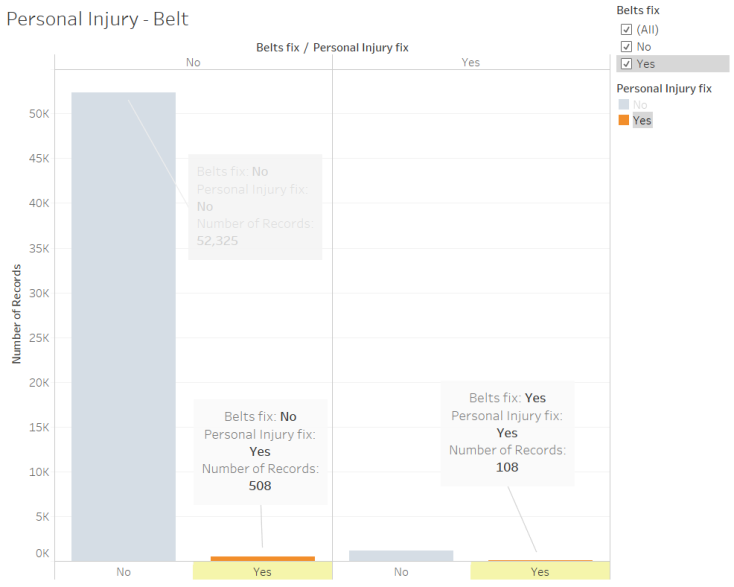


Fig. 11 – Personal Injury due to Belt

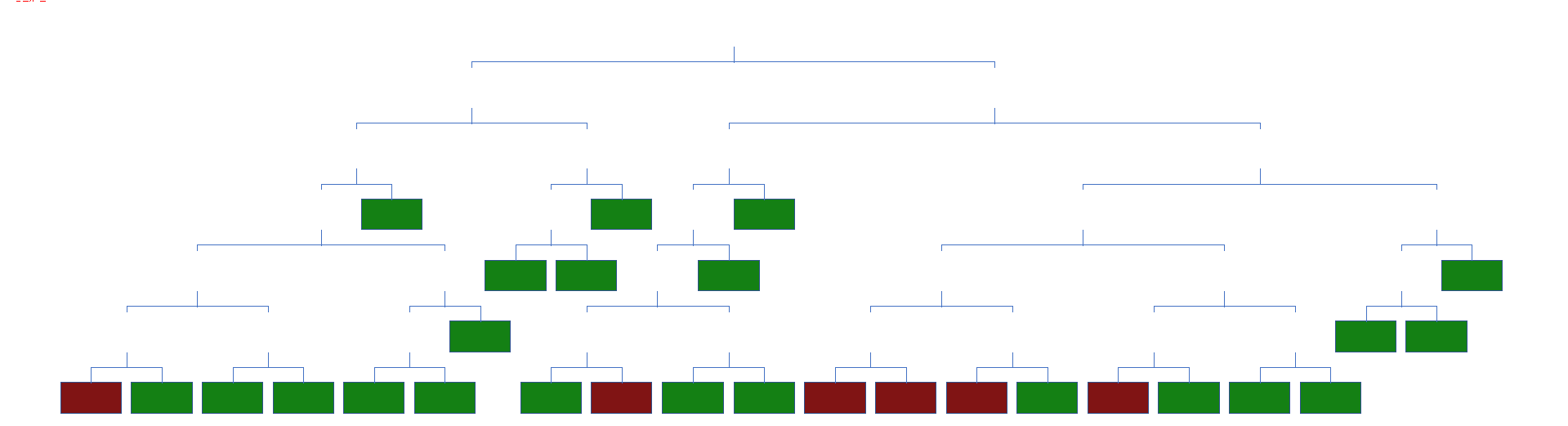


Fig. 12 – Full tree for model 1

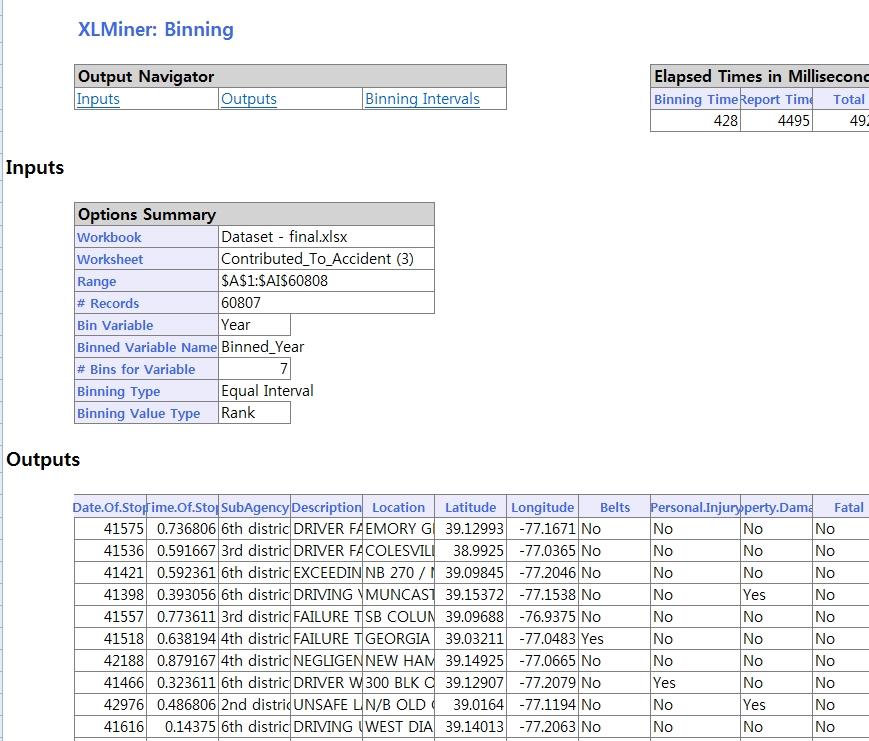


Fig. 13 – Binning and creating dummies

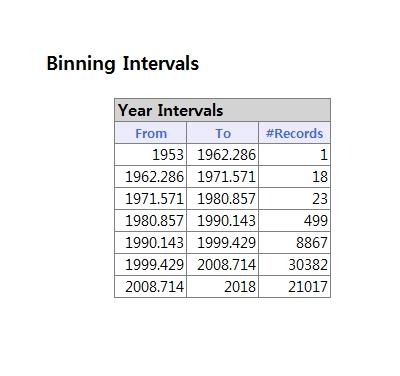
****

Fig. 14 – Binning and creating dummies

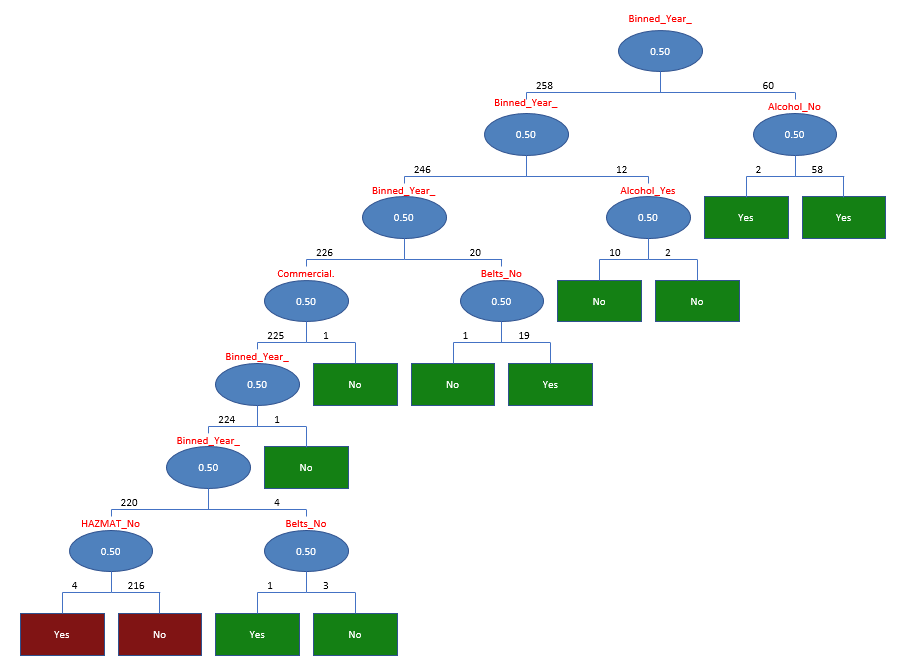


Fig. 15 – Full tree for model 2.

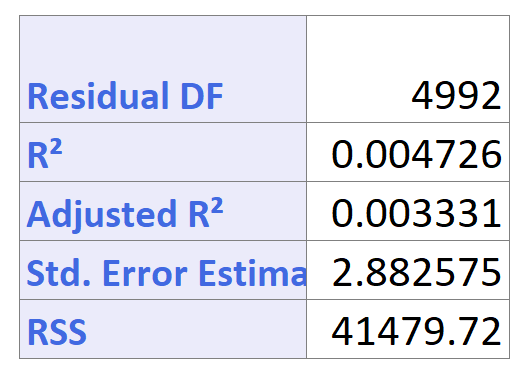


Fig. 16 - Results evaluation of regression prediction for model 3.

# REFERENCES

1. Discovering Knowledge in Data: An Introduction to Data Mining, Daniel T. Larose and Chantal D. Larose, Wiley, 2nd edition: Wiley ISBN 978-0-470-90874-7
2. Abellan, J., Lopez, G., & De O~na, J. (2013). Analysis of traffic accident severity using Decision Rules via Decision Trees. Expert Systems with Applications, 40, 6047–6054.
3. Chang, L.-Y., & Chien, J.-T. (2015). Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model. Safety Science, 51(1), 17–22.
4. Chen, W. H., & Jovanis, P. P. (2012). Method for identifying factors contributing to driver injury severity in traffic crashes. Transportation Research Record, 1717, 1–9.
5. Kashani, A. T., Rabieyan, R., & Besharati, M. M. (2014). A data mining approach to investigate the factors influencing the crash severity of motorcycle pillion passengers. Journal of Safety Research, 51, 93–98.
6. Kwon, O. H., Rhee, W., & Yoon, Y. (2015). Application of classification algorithms for analysis of road safety risk factor dependencies. Accident Analysis and Prevention, 75, 1–15.
7. Xie, Y., Zhang, Y., & Liang, F. (2009). Crash injury severity analysis using Bayesian ordered probit models. Journal of Transportation Engineering ASCE, 135(1), 18–25.
8. Mujalli, M. O., & de O~na, J. (2011). A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks. Journal of Safety Research, 42, 317–326.
9. De O~na, J., Lopez, G., & Abellan, J. (2013). Extracting decision rules from police accident reports through decision trees. Accident Analysis & Prevention, 50, 1151–1160.

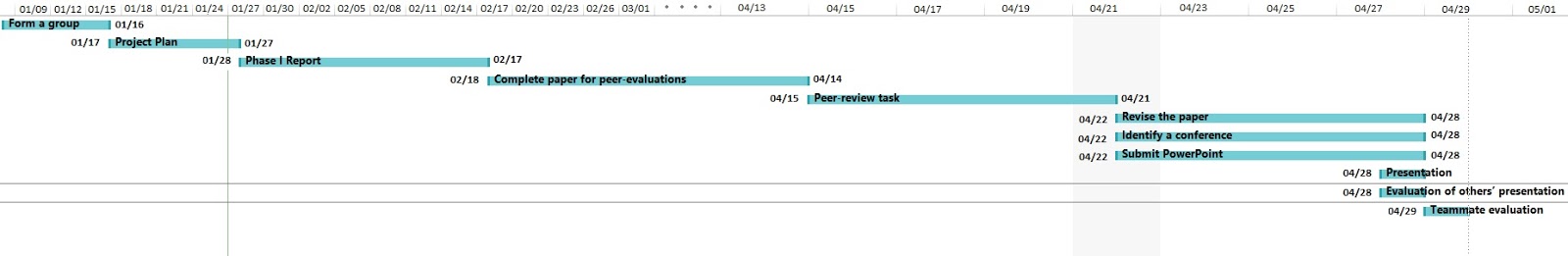


Fig. 17 – Completion timeline

NOTE:

All modeling techniques have been included in the previous submission.