Data Mining and Prediction of US Automobile Accidents

Raymond Hong
Computer Engineering Department
San Jose State University
San Jose, USA
raym.hong@gmail.com

Sum Mohan Reddy Mallannagari
Computer Engineering Department
San Jose State University
San Jose, USA
summohanreddy.mallannagari@sjsu.edu

Amir Hossein Radman
Computer Engineering Department
San Jose State University
San Jose, USA
amirhossein.radman@sjsu.edu

Amrita Kasaundhan
Computer Engineering Department
San Jose State University)
San Jose, USA
amrita.kasaundhan@sjsu.edu

***Abstract*—Automobile accidents are a frequent occurrence in our everyday life. With a responsible driver who knows the rules of driving, it is important to understand the risks that come into play when certain situations or obstacles occur. Given a dataset that has accident information along with its severity, and can visualize trends in the data, and create models that can help us predict in what scenarios severe accidents would occur. The models we will implement will be decision trees and neural networks.**

***Index Terms*—accidents, fatalities, United States, automobiles, cars, impaired, deaths**

I. Introduction

You might hear that you are 23 times more likely to get into an accident while on the phone, or that a rainy day can cause the likelihood of an accident to occur. How are those statistics proven? Data is far more useful than just being a record of historical data. Not only does one use this data to get reference instances that happened in the past, but they also perform analysis on the data. Analyzing data extracts important information that one might not recognize from a local scale. One can perform many operations on the data, to make information much more obvious and prevalent. One example is the use of visualizations. Of course, looking at rows and columns of data does not make trends obvious. If we can represent these in a more pictorial visual, we can make it obvious to the audience what is clearly going on with the data. One example of this is plotting the number of accidents that occur on weekdays. Expressing this in a bar graph would given the reader instantaneous understanding of what the trend is.

Another example is computing, or generating information from the given data. One might need to transform the data into another representation through the use of operations or conversions. This is where feature extracting is a little more difficult to do because often one must fully understand what the data means, and know how to manipulate the data into another form that can be interpreted. It's like squeezing all the information one can get out of the data. These can be done with operations such as principle component analysis, added columns together, or changing continuous values into categorical values, which you will see later in this paper that it has been done. In order to implement some of the machine learning algorithms, some data like barometric pressure, needed to be converted from a continuous value to categorical, for the method of decision trees to work. It's also easier to understand and extract meaning from that form.

Machine learning algorithms are also a huge determining factor of extracting new information. Creating a machine learning model creates something where we can simulates various conditions, and have it predict what the likely outcome would be. Without this technique, we would be burdened with the task of really digging into what factors contribute to what manually. For example, having the machine learn and understand what are the factors that contribute to an accident at severity 4 vs severity 1 saves us as the researchers a lot of burden to finding it out. It extracts the patterns for us, and from there, can report these behaviors to the public in order to reduce severe accidents. It also lets us go into unknown territory, as we can simulate a fake accident with these created conditions, and have the model predict the severity. There are endless possibilities for how much information we are able to extract.

Using these techniques to analyze accident data, we feel that implementing these methods would find new information that could potentially saves lives in the long run. Many people have already implemented their own approach, such as using random forests to predict severity, or regression models. We will try to find new and more information by implementing new methods such as decision trees, and neural networks. We will also use association algorithms to find what traffic obstacles are frequent with each other.

II. Methods

To visualize the data is the clearest way, we performed various methods of clean up, so that any attempts at visualization will not be distorted. Some of the methods are using pandas build-in functions such as dropping "nan"'s, as shown in Figure 1, unrelated columns, or converting data columns into necessary representations. After we've obtain multiple visualizations that we can infer from the data, we begin running machine learning methods. From other kernels that others have produced on Kaggle, most people implement the usage of random forests to drive their predictions. For this paper, we implement decision trees and neural networks to run our predictions.

The most common methods used on Kaggle are KNN and Random Forests. They have achieved around 90% accuracy using these methods. In terms of data visualization, they have implemented bar graphs to see which weather conditions were the most prominent during accidents, which states have the most accidents, and the number of accidents from each of severity. Some interesting data we plan to implement is to extract additional information. We also plan to utilize decision trees and neural networks for this application.

III. Experiment and Analysis

A. Data Visualization

We will be begin by addressing some aspects of the data set, so anyone who decides to use this dataset will understand the work required in order to prepossess this data. First off, we check for any null values that can be eliminated from the row. According to 1, there is quite a lot of null values. We eliminate them to increase dataframe efficiency and processing.

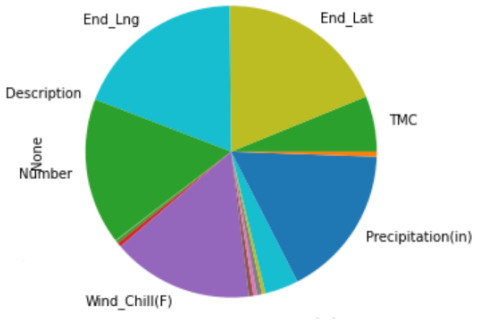

Fig. 1. Pie chart of the null count for each column.

This dataset has classified the dataset with a severity between 1 and 4, where 1 is the accident is small, and 4 where the accident is large. The get an idea of whether this dataset is balanced or not, fig 2 was produce to understand if the data has any bias to it.

As you can see in figure 2 and in table I, severity 2 has the highest number of records, followed by 3, 4 and 1. This might make the data a little challenging to produce accurate results. The Data has also been visualize based on various frequency namely daily, weekly and monthly and yearly basis.

TABLE I
Severity Count

Severity	Count
1	968
2	1993410
3	887620
4	92337

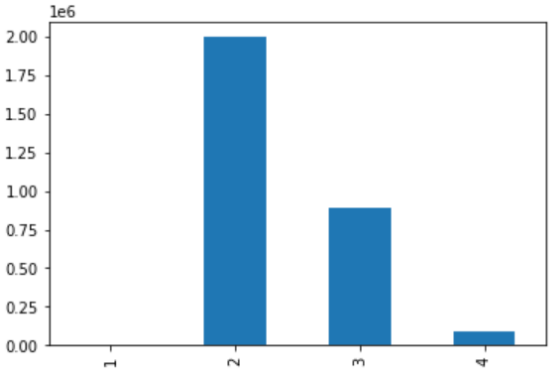

Fig. 2. Bar graph of the severity count.

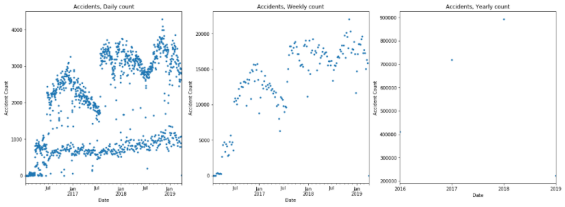

Fig. 3. Displays the accidents count on daily, weekly, monthly basis.

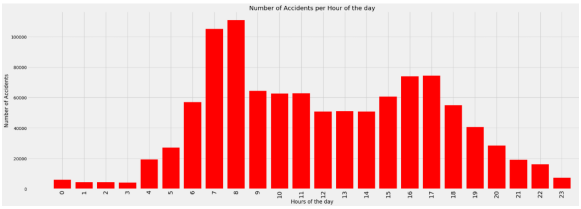

Fig. 4. Displays the accidents count occured by time.

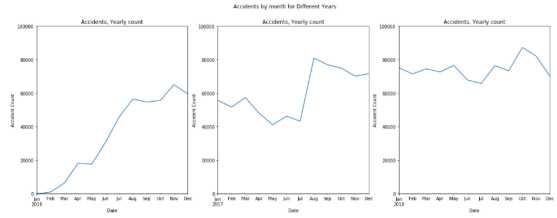

Fig. 5. Displays the accidents count in various years.

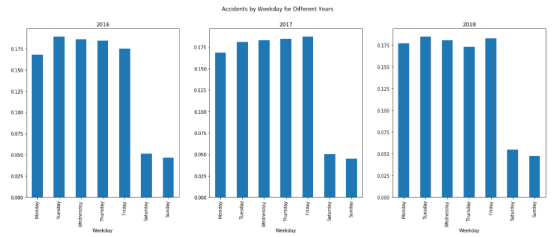

Fig. 6. Displays the accidents count on weekdays in different years.

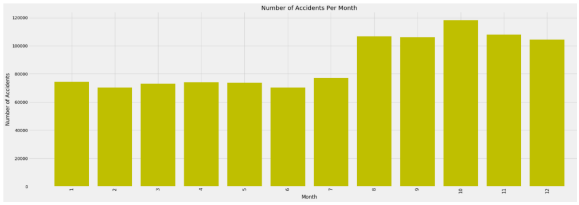

Fig. 7. Displays the accidents count by months.

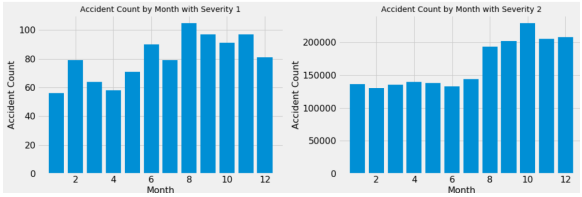

Fig. 8. Displays the accidents severity by months.

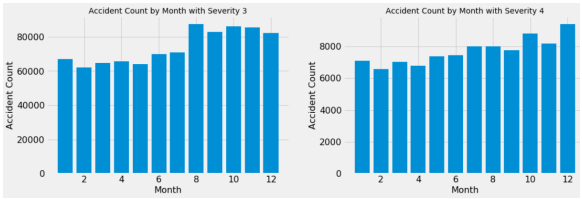

Fig. 9. Displays the accidents severity by months.

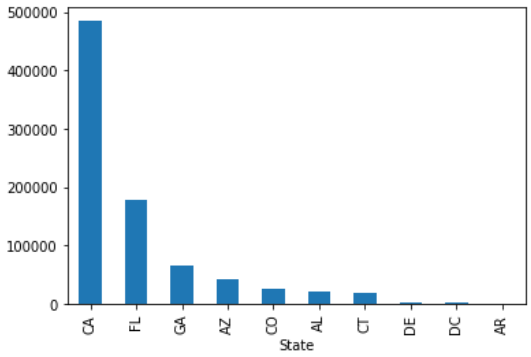

Fig. 10. Displays the accidents count in different states.

B. Feature Selection

Feature selection is process of selecting feature that contribute to the prediction variable where the features are selected based on various factors like its importance and correlation observed while visualizing from exploratory data analysis.

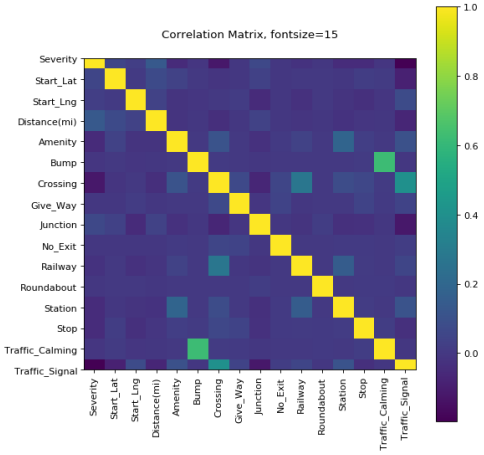

Fig. 11. Correlation Matrix of features.

These selected features majorly contribute to determine the effect of road conditions and weather conditions on predicting accidents risk. Features that are related to road conditions include Amenity, Bump, Crossing, GiveWay, Junction, NoExit, Railway, Roundabout, Station, Stop, TrafficCalming, TrafficSignal, TurningLoop. While features that are related to weather conditions include Temperature(F), WindChill(F), Humidity, Pressure(in), Visibility(mi), WindDirection, WindSpeed(mph), Precipitation(in), WeatherCondition, SunriseSunset.

C. Data Pre-processing

In the Data Preprocessing step, data cleaning techniques like removing null valued rows and columns, filling the null values with median and mode are performed. Data scaling using standard scaler on train data and normalization are performed. Data transformation using Label encoder, One Hot encoder are performed. As the label of the dataset is very skewed, the label of the data contains very few elements of the 0 class and very high elements of 4 class. So, the data of 0 class is skewed and need to do sampling of the dataset. We performed under sampling of the data by combining both classes of 0 and 1 to 8Dimensionality reduction is performed using Principal Component Analysis (PCA). Parameter tuning using Grid Search CV is performed.

D. Machine Learning

Our first model we used were decision trees. In order to implement decision trees, we had to figure out what we could do with the columns that contained continuous values. For example, humidity, windspeed, and temperature were among the few that had to undergo categorization. To do this, we set thresholds for each , varying the number of depths to an

appropriate number depending on the column. We feel that doing this could lead into several issues as we do not really know what is the true appropriateness of levels. For example, we set temperature to four levels: freezing, cold, normal and very hot. There could be less levels that would be far more ideal that 4. We will continue to explain this issue in the upcoming section.

After categorizing the columns with continuous values, we proceeded to implement decision trees with 2, 4, 6, and 8 layers of depth and compare the results. An example of a 4 layer decision tree can be found in Figure 12

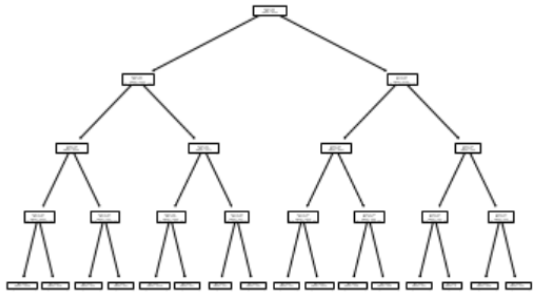

Fig. 12. Image of a four layered decision tree.

Various accidents models are used to minimize the accidents by monitoring the effectiveness of road safety policies that have been introduced. Second model is based on the deep learning methods, we used Neural Network on our data set using MLP classifer. Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns. Neural Networks(NN) is one of the novel computational approaches includes modelling tools that can be of great utility in understanding and predicting the accidents.The ability of NN's to work with complex, instructive processing characteristics such as parallelism, noise tolerance, nonlinearity, learning and generalization capabilities, provide superior predictive power in non-linear and complex relationships which suits their application to accident data analysis. They interpret sensory data through a kind of machine perception, labeling or clustering raw input.It takes multiple input to product one single out. Figure 13 displays the mathematical representation of Neural Network.

$$\text{output} = \begin{cases} 0 & \text{if } \sum_j w_j x_j \leq \text{ threshold} \\ 1 & \text{if } \sum_j w_j x_j > \text{ threshold} \end{cases}$$

Fig. 13. Mathematical representation of Neural Network.

To understand the architecture of Neural Network, we took Figure 14 for reference. The diagram shows the network of four-layer network has two hidden layers. The layers are made up of many neurons, which are highly interconnected by weighted links through which information exchange happens.The leftmost layer in this network is called the input layer. The rightmost or output layer contains the output while the middle layer is called a hidden layer, since the neurons in this layer are neither inputs nor outputs. Confusion Matrix is plotted to visualize the number of correct predictions in the test data.

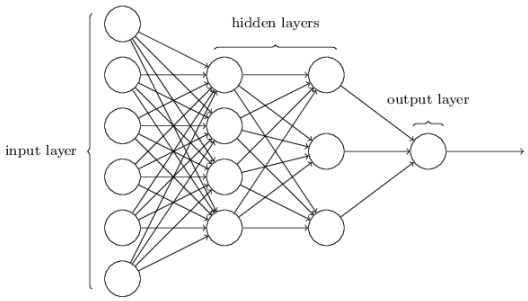

Fig. 14. Neural Network Architecture.

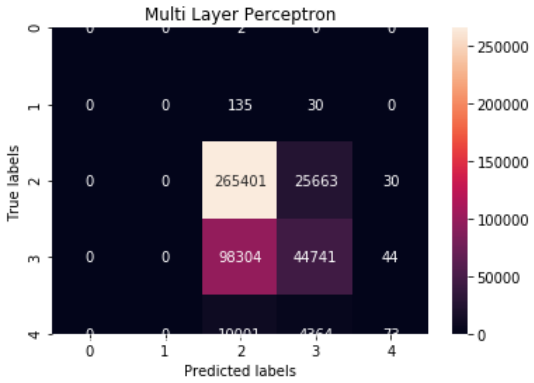

Fig. 15. Neural Network Confusion Matrix

IV. Comparisons

As a team, we focused on ensuring our analysis is different from other people's analysis on the same dataset. Some of the analysis that had been previously done on this dataset includes:

- Analysis of types of cars depending on the number of accidents in different states.
- Cause of accidents.
- Predicting accidents severity.
- Time of accidents.

A. Decision Tree

Four decision trees were created in order to see how the s vary the accuracy. using the SKlearn library, the accuracy from each were pretty much the same, as shown in Table III. We feel the reason for this is due to the categorization of some of the columns. Wind Speed is an example of this. We do not know the ranges in which wind speed should be classified based on the danger. For this project, our categories in this section we based on our opinion on what they should be. Thus, if these are off, we feel very confident that the depth behaves this way. Figure 16 shows this as a graph.

B. Frequent Traffic Obstacles

In figure 11, you can see the correlation between traffic obstacles. We further this exploration by applying Apriori Algorithm on this to find frequent item sets in these accidents. Table II shows the frequent itemset with a support threshold of 0.025. The reason for this low threshold is that there were no itemsets above this number. We can say that there are practically no frequent items in this dataset for any high threshold.

TABLE II
FREQUENT ITEMSETS

Itemset	Support
crossing	0.069
Junction	0.080
Traffic Signal	0.162
Crossing, Traffic Signal	0.053

TABLE III
DEPTH ACCURACY

Depth	Accuracy
2	0.7170925646862676
4	0.7171310397896694
6	0.7171310397896694
8	0.7171887524447722

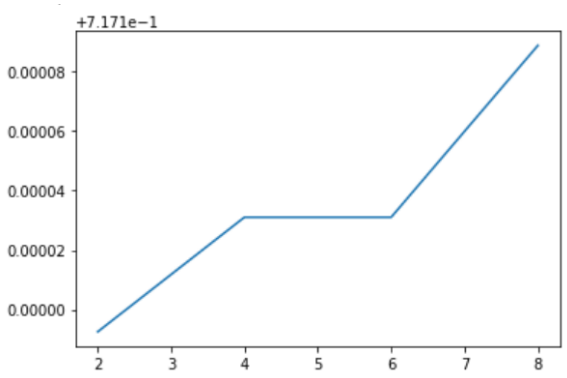

Fig. 16. Graph of the accuracy over depth of Decision Tree.

C. Neural Networks

The results have been obtains using below parameters with different values. It is observed that confusion matrix at hidden layer size of(30, 30, 30) and (20, 20, 20). Diagram 17 displays the best parameters which contributes to the more accurate accident prediction.

- solver
- alpha
- hidden layer size
- max iteration
- learning rate

```
Best parameters found:
 {'hidden_layer_sizes': (20, 20, 20), 'max_iter': 200}
0.675 (+/-0.007) for {'hidden_layer_sizes': (10, 10, 10), 'max_iter': 200}
0.680 (+/-0.002) for {'hidden_layer_sizes': (10, 10, 10), 'max_iter': 300}
0.691 (+/-0.002) for {'hidden_layer_sizes': (20, 20, 20), 'max_iter': 200}
0.691 (+/-0.007) for {'hidden_layer_sizes': (20, 20, 20), 'max_iter': 300}
```

Fig. 17. Best Parameters.

V. CONCLUSIONS

In conclusion, prior to performing the analysis, our team used various techniques to ensure our data was clean and not include any unnecessary records that could affect the accuracy of our work. Additionally, by implementing decision tree, neural network and association analysis algorithm, we extracted new and critical information about the dataset that by comparison, had not been done previously. We also focused on providing various visualization charts and figures for the audience to better understand the data and compare the data visually. For example, by performing the analysis we discovered majority of accidents occur in state of California and towards the last quarter of the year, which could indicate the raise in people's traveling during holiday season and cause more accidents in congested roads.

REFERENCES

[1] L. Li, S. Shrestha and G. Hu, "Analysis of road traffic fatal accidents using data mining techniques," 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), London, 2017, pp. 363-370, doi: 10.1109/SERA.2017.7965753.

[2] Sanjay Misra" An Artificial Neural Network Model for Road Accident Prediction: A Case Study of a Developing Country"

[3] M. F. Labib, A. S. Rifat, M. M. Hossain, A. K. Das and F. Nawrine, "Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh," 2019 7th International Conference on Smart Computing Communications (ICSCC), Sarawak, Malaysia, Malaysia, 2019, pp. 1-5, doi: 10.1109/ICSCC.2019.8843640.

[4] Lipovac, Krsto Tešić, Milan Maric, Bojan erić, Miroslav. (2015). Accident Analysis and Prevention. Accident; analysis and prevention. 84. 74-82. 10.1016/j.aap.2015.08.010.

[5] S. Sonal and S. Suman, "A Framework for Analysis of Road Accidents," 2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research (ICETIETR), Ernakulam, 2018, pp. 1-5, doi: 10.1109/ICETIETR.2018.8529088.