San Jose State University
Data Mining - CMPE 255
Professor Carlos Rojas

MILESTONE 2 Report

Team Members
Raymond Hong - ()-raymond.hong01@sjsu.edu
Amrita Kasaundhan-(013854204)-amrita.kasaundhan@sjsu.edu
Sum Mohan Reddy - (014545453)-summohanreddy.mallannagari@sjsu.edu
Amir Radman(010208367) - amirhossein.radman@sjsu.edu

Introduction & Background

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. 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.

Researchers and data scientists are already at work trying to see what information that can retrieve in order to reduce the number of accidents. For this topic, we will be using the accident dataset supplied from Kaggle, where different users have already implemented their solution to this problem. While many of them have performed some data visualization techniques and different machine learning models, we plan to do the same but using different representations.

Additional Datasets

Obviously, we know to expect trends such as cell phone usage, holiday travel, bad weather conditions to have some sort of effect on accidents. We want to prove these by finding them in the data.

In order to obtain additional information about accidents, we will use other datasets to extract them. One dataset from the NHTSA provides roughly 330,000 fatal accidents that have occurred in 2018. We've already begun exploring this data to find trends like holiday travel fatalities.

Methods

Our approach in data visualization would be to implement more methods aside from your normal bar graph. We want to apply the apriori algorithm on the obstacles present during times of accidents, to see which obstacles are associated with others. We want to find out the correlation between different types of weather, and traffic obstacles as well as using correlation matrices, paragraphs, etc. The dataset is large so we should be able to implement a number of graphical representations to convey trends.

Can you please answer the following:

1. Summarize what methods others have applied.

The majority of the kernels that exist on Kaggle try to find useful information out of a dataset. Through data visualization, clean up, machine learning, the goal is to extract information that hasn't been found or to apply different methods and compare the results.

Most machine learning related work uses SKlearn to predict severity levels of all accidents after performing some preprocessing of data. They use various algorithms such as KNN, Decision Trees, Random Forests to prove which algorithms are the best suited for this application. These indirectly tell us what features are important to the learning.

There is also a kernel that focuses on what causes the accidents. The program using numpy,panda and plotpy for cleaning up the data and visualizing the results. By doing the analysis, They are able to extract important information such as :

- Number of accidents in the weekends compared to weekdays
- Identifying the country with the most number of accidents.
- Analysis of the GPS location data that could provide us with information about frequency of accidents in certain junctions

2. How does your approach differ?

1. An interesting characteristic of this dataset is that there are some features that provide the existence of various traffic obstacles at the presence of an accident. This gives us a chance to perform association analysis. We can perform the Apriori Algorithm on this to determine what obstacles tend to be with other obstacles when accidents occur.

2. Because of the large amount of data in this dataset (~1Gb), we plan to use PySpark and it's various methods and compares it to non-distributed methods.

3. Did you find anything new?

Have you considered using multiple datasets to analyze the same problem? Check out https://data.world/cdc/impaired-driving-death-all Or something as simple as a days of holidays and how they correlate with accidents.

We will use other datasets to see if we can extract more useful information. Including the dataset you provided, we are also looking into NHTSA Fatalities dataset to obtain meaningful information. Using that dataset, we are able to find the amount of fatal accidents occurring on different days.

Also, consider the following questions:
1. What are some obvious results that you expect? For example, bad weather usually correlates with higher accident rates.

We know that bad weather contributes to higher accident rates and we know already that holidays tend to generate traffic, thus meaning more chances for accidents to occur. We will strive to prove these assumptions in our data.

2. What's happens if you reduce the number of variables to answer questions about your dataset? Can you outperform more sophisticated algorithms (deep learning models) by feature hacking the dataset.

Answers: It depends upon whether the variable has significance in machine learning models. It does have an impact on ML models then the underfitting of data models occurs. Underfitting of data means the data model is so simplistic which means low variance in the prediction and high bias which leads to the wrong prediction.
Yes, feature selection and feature extraction is a deep learning technique to reduce the dimension of the dataset to a trade-off between bias and variance. Applying these may result in outperforming the ML models as we have applied the Apriori Algorithm in our dataset and found that the application of this algorithm in our dataset gives a better result. Along with that, we also have the plan to apply one-hot encoding in the dataset and then compare the results of both the algorithms.

3. Do you see any local trends that differ from the global dataset?

We did fetch the data of accidents in India and UK and visualize the data using the same algorithms and found trends were different for different countries. There are several other factors that contributes to the rate of accidents like human factors, vehicle factors, and road environment factor. The rate of accidents at night is lower than that of day, which is the same for all of the countries' data set that we picked for the making prediction.

what will be the workload distribution among your teammates? Write down the tasks that each person will do.

Raymond	• Implement Apriori Algorithm to obtain frequent itemsets. • Pre-Processing and Cleaning Up the multiple datasets • Data Visualization • Decision Tree
Amir	• Pre-Processing and Cleaning Up the multiple datasets • Data Visualization • Decision Tree
Amrita	• Pre-Processing and Cleaning Up the multiple datasets • Data Visualization • KNN and compare the results
Sum Mohan	