San Jose State University

Data Mining - CMPE 255

Professor Carlos Rojas

MILESTONE 2 Report

Team Members

Raymond Hong - (009957389)-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

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.

Introduction & Background

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 cau se the likelihood of an accident to occur. How are those statistics proven? Over a course of a given timeframe, researchers compile data to find out what specific problems are to be blamed for accidents. Our dataset has plenty of information to use, specifically such as the date, weather condition, traffic objects currently present and a lot more.

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. The goal is not to only find new information, but it also verify information that we know is true. For example, holiday travels cause more accidents or more accidents occur on rainy days, because if we can verify those facts, we can trust the new information we will find.

While many data scientists have performed some data visualization techniques and different machine learning models, we plan to do the same but using different representations. We also plan to find frequent items consisting of traffic objects using the apriori algorithm.

Methods

The most common methods they use 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 level 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.

Example Analysis

The dataset consists of car accidents from 49 states of the United States and is collected from February 2016 through December 2019. The data provides us with critical information captured during an accident. For example; time of the accident, location, weather, severity of the accident and other additional information. By performing data analysis on this data set, we will be able to extract important information such as the impact of weather on the severity of the accident or frequency of accidents in certain locations.

Github Link :
https://github.com/amirradman/cmpe255
Dataset :
https://www.kaggle.com/sobhanmoosavi/us-accidents/#US_Accidents_Dec19.csv

Milestone 1 - Proposal Responses

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. We used deep learning techniques that are not used before as most of the methods used before are normal machine learning techniques. As the data that is provided is categorical data we used the labelencoder for converting categorical data to numerical data.

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.

4. 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, and also during the working hours (As everyone goes to their respective work space in the morning and return during the evening.) thus meaning more chances for accidents to occur. We will strive to prove these assumptions in our data.

5. What 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 and label encoder in the dataset and then compare the results of both the algorithms.

6. Do you see any local trends that differ from the global dataset?
We did fetch the data of accidents in India and the UK and visualized the data using the same algorithms and found trends that were different for different countries. There are several other factors that contribute to the rate of accidents like human factors, vehicle factors, and road environment factors. 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 • Report
Amir	• Pre-Processing and Cleaning Up the multiple datasets • Data Visualization • Decision Tree • Report
Amrita	• Implement One Hot encoding • Pre-Processing and Cleaning Up the multiple datasets • Data Visualization • Neural and compare the results • Report

Sum Mohan	• Implement Label Encoder • Pre-Processing and Cleaning Up the multiple datasets • Data Visualization • Neural Networks and compare the results • Report

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.