Project for Data Analysis and Visualization

Project Title: Predicting the severity of Accidents based on various conditions.

Members' Names: Ishaan Shrivastava, Sajal Saini, Rajat Bansal, Amber Singh, Rohit Raj

Q1. Business/Research/Social Objective of the project:

Road accidents are among the top causes of death across the globe. The consequences of road accidents may vary from minor injuries to death. Our project will attempt to capture the severity of accidents in terms of various possible factors which will include weather conditions, time, location, visibility, state of the driver, description of the accident, etc. With driverless cars under development, our project can help to implement an adaptive driving strategy for such cars depending on the expected level of severity of the accident.

Q2. Sources of data, method of data collection, and an estimate of data cleaning, transformation effort.

We are using a dataset of 3.5 million accidents in the US from February 2016 to June 2020 which is available on Kaggle. This dataset is itself derived from various traffic APIs of different states of the US.

The dataset is based on the following papers (which are required to be cited if we take the aforementioned dataset in use):

● Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath."A Countrywide Traffic Accident Dataset.", arXiv preprint arXiv:1906.05409 (2019).

● Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu,
and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In Proceedings of the 27th ACM SIGSPATIAL International Conference on
Advances in Geographic Information Systems, ACM, 2019.

To scale latitudes and longitudes on the 0-1 range for incremental models we used z-score to transform them.

We also identified the data which was collected post-accident and we deleted those columns.

Also, based on some visualizations we deemed some of the columns as irrelevant.

Weather conditions were converted into boolean with each unique condition having its column. Since these columns were interdependent, we rejected Bayesian Modelling as it assumes each feature to be independent of the other.

Q3. Data analysis, modeling, and model validation/testing.

To model the data, we implemented Random Forest, Decision Tree, Linear SVM, and Logistic Regression. The latter two were implemented as incremental algorithms (batch-wise learning) but they both yielded very bad results and hence we concluded that they work badly on our skewed data.

Random Forest yielded decent accuracy whereas Decision Tree's accuracy was phenomenal (around 80%). Another observation was that the skewed classes with fewer occurrences had a bad recall even in the best models. (refer to modeling section of the slides for more)

Q4. How do you intend to visualize your raw data, intermediate and final results?
On raw data, we have performed Exploratory Data Analysis and have highlighted some trends according to our data set. The visualizations can be found in the notebook we have shared, in the slides, and some older ones in the mid-term project report.
In modeling, we have generated our validation curve for random forest and decision tree. Besides, we have also generated the most dominant features affecting our severity prediction according to each of the mentioned models.
Accuracy of models is also compared and learning curves of SVM and Logistic Regression are also generated.

Q5. Problems faced and Where we would have headed if we had more time:

Frequent RAM blow-ups during training: For this, we ensured that nothing unimportant is in the memory. Only the training set and validation set are occupying the memory in addition to the model. This helped us in getting results that were initially blowing up.

Very high training time is SVM, and non-incremental Logistic Regression and SVM due to which validation was not feasible. Also even though these models were left for overnight training they didn't yield any good results.

Batch Learning Algorithms like SGD for SVM and Logistic Regression didn't prove helpful, despite training on large data size. Our main motivation to use incremental algorithms was to save training time and still get decent accuracy. We even changed the number of iterations and batch sizes to get better results but those efforts went in vain as beyond a limit the training time which was our motivation to implement these algorithms has become pretty long.

Where we were planning to head?
1.) Refining the results of incremental algorithms.
2.) Planning to explore 'Undersampling' to get better Recall for severity 4 and severity 1.

Initially, we planned to implement Neural Networks but since we had ~2 million records to train we dropped the idea as Neural Networks are more suitable for fewer records.

Q6. Member Contributions:

Ishaan Shrivastava: Majorly involved in modeling data and analyzing post modeling data, have some contributions in EDA and data cleaning

Amber Singh: Major involvement in modeling data and analyzing post modeling data, have some contributions in EDA and data cleaning

Sajal Saini: Moderate contributions in modeling and major role in EDA and Data Cleaning. Crucial roles in pre-modeling feature selection.

Rajat Bansal: Moderate contributions in modeling and major role in EDA and Data Cleaning. Crucial roles in pre-modeling feature selection.

Rohit Raj: Contributions in EDA and Data Cleaning.

Link to Dataset: https://www.kaggle.com/sobhanmoosavi/us-accidents