Data Mining Project Proposal - Group 5

Abdulaziz Gebril, Jenny Tsai, & Mojahid Osman

April 15, 2020

- **What problem did you select and why did you select it?**
  Tens of thousands of lives are lost every year in severe car accidents in the U.S. For example, the National Highway Traffic Safety Administration estimated that 36,750 people were killed in the U.S. in traffic crashes in 2018. Therefore, it is important to find factors that can affect the severity of car accidents, based on which we can craft strategies to lower its occurrence, especially those of high severity. In this project, we will look at weather and road conditions, and identify the most important ones that can impact the severity of the car accidents.

- **What database/dataset will you use? Does it need to be cleaned?**
  We will use the U.S. accident dataset created by Moosavi et. all (2019). The dataset was compiled from various sources, which contains about three million instances of traffic accidents that took place within the contiguous United States over the last three years. Each accident record consists of a variety of intrinsic and contextual attributes such as location, time, weather, and points-of-interest. Though the preliminary merging and cleaning was done, the dataset still has a lot of missing values and attributes that are not relevant to our problem statement. Therefore, a significant amount of efforts will be devoted into data pre-processing.

- **What data mining algorithm will you use? Will it be a standard form, or will you have to customize it?**
  We will use algorithms of random forest, logistic regression, and possibly SVM. The algorithm will be of a standard form as our dataset and problem statement are pretty straightforward. The only customizations that we will make are the parameters in the models.

- **What softwares will you use to implement the network? Why?**
  - Python
    It wasn't that difficult for us to decide to go with Python because Python is a great and efficient tool for data preprocessing, especially when it comes to large dataset. It's also the most common tool used for Machine Learning as Python has libraries (e.g., scikit-learn) that are designed to do. Python is also compatible with a wide range of programs in the market (e.g, QT), which saves our time on code conversion.

  - PyQt5 - QT designer
    In addition to python, we chose to use QT to build the application interface to give the users the ability to interact with the application we are building. One of the main reasons we chose Qt is the time to market and its full support python and it can run on any OS like (Windows, Linux, and Mac) with no issue.

- **What reference materials will you use to obtain sufficient background on applying the chosen network to the specific problem that you selected?**

Since we use the binary variable *Car Accident Severity* (0 = low, 1 = high) as our target, this is categorized as a supervised learning - classification problem (García et al., 2015). Specifically, we will adopt the *filter model* as our feature selector (García et al., 2015), using the evaluation measures (see next section) to obtain the best feature subset for training. To do that, we will build a random forest and select features based on importance. Prior related research has implemented logistic regression model to solve problems of similar type (e.g., Moosavi et al., 2019), therefore, we will also build a logistic regression model to compare with our network. Finally, if possible, we will try more models (e.g., SVM) and use the ensemble method to predict the outcome, which in turn, increases the reliability of our estimation.

- **How will you judge the performance of your results? What metrics will you use?**
  For logistic regression, we intend to build a confusion matrix from which we can obtain the relevant performance metrics (e.g., accuracy, recall, precision, F1 score, etc). We may also check other performance metrics such as McFadden and AIC to see if they are consistent with the findings from the confusion matrix. For random forest, we will check entrophy/gini and information gain to evaluate the model's performance.

- **Provide a rough schedule for completing the project.**

Week	Date	Progress	Deliverables
Week 1	3/22 - 3/28	- Find a research topic and dataset	
Week 2	3/29 - 4/4	- EDA on the dataset, finalize problem statement and solution approach - Data-cleaning/pre-processing - Set up github repo	
Week 3	4/5 - 4/11	- Complete research proposal - Data-cleaning/pre-processing - Modeling - Logistic regression & Decision Tree - Explore QT designer	
Week 4	4/12 - 4/18	- Streamline the findings - Visualize findings using QT designer - Write the summary paper	4/14 - Exam 2 4/15 - Proposal due
Week 5	4/19 - 4/25	- Work with Professor on issues - Building slides & Presentation dry-run	4/21 - Q&A session
Week 6	4/26 - 4/28	- Finalize presentation & paper	4/28 - Presentation & Paper due

References

García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining*. Switzerland: Springer.

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, & Rajiv Ramnath. (2019). *A Countrywide traffic accident dataset.*

Moosavi, S., Samavatian, M. H., Parthasarathy, S., Teodorescu, R., & Ramnath, R. (2019). *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.