Team 233 Project Proposal

With our project, we intend to predict the severity (in terms of impact on traffic) of an automobile accident given a set of environmental, geographic and other features1. This information could be used for a variety of purposes, including staffing and response decisions for emergency services, identification of accident-prone locations (as well as areas prone to *not* having accidents, where *this* information could be used to remediate the aforementioned accident-prone locations), and estimates of delays for GPS routing, among others.

Today, first responders are called to the scene of an accident based on the description provided by the 911 caller. A complete picture is not always possible, especially when the caller is distressed and distracted, limiting the information the operator is able to gather2. Our approach will provide a predicted severity based on immediately available information including time of day, weather conditions, and similar accidents in the area, as well as using features unique to the location itself (e.g., whether there is an amenity, railway, roundabout or even speed bump nearby)3.

The impact of our study should provide benefits to first responders, roadway designers and GPS developers and users alike, to mention a few4. For first responders, success could be measured by changes in fatality rates, response time, and site cleanup times. Success for a roadway designer would be measured through relative accident frequencies post construction. For GPS developers and users, accuracy of time estimates can be measured, with improvement expected by leveraging our algorithm5.

Our project risks primarily consist of inability to discover new findings that would bring benefits to our stakeholders6. The project team has identified free dataset(s) and will be able to conduct the analysis with minimal or no cost via leveraging team members computing resources and open source software packages7.

The project timeline will run over six weeks between 28 February 2020 and 17 April 2020 with milestones expected to be completed as follows8. It is expected that there may be instances of rework or shifts in the analysis trajectory based on findings during the analysis. Success at each of the milestones will be measured by completion of each interim deliverable, with results shared with all team members9. It is expected that responsibility for all deliverables will be shared equally by team members, but ownership of specific items may be assigned as the project progresses.

* 28 February - Project Proposal submitted.
* 13 March - Exploratory Data Analysis completed and results reviewed by team members ; analysis methodology chosen and assigned.
* 27 March - Analysis completed and results reviewed by team members; Final Report sections assigned.
* 17 April - Final Report completed and submitted.

Moghaddam et al [2] trained artificial neural networks to understand drivers of severities of accidents occurring on the highways of Tehran, Iran. Their study had a high degree of accuracy but noted many driver-specific characteristics and actions that led to higher severities. These data points would not be available for first responders, a focus of our study.

Beshah and Hill [3] studied accident severities in Ethiopia, using a variety of classification models including decision trees, naive Bayes, and KNN. Their findings dialed in on road design, but seemed very limited in terms of data they were able to source. It likely will not be useful for our project, which should be able to deliver a more meaningful output given the additional data available to us.

Garrido et al [4] looked at factors impacting the severity of injury in accidents in Portugal. Their study is useful for our project and notes some predictors and a methodology that we can start with in our analysis. The main shortcoming of their study was a limited dataset of just 4,528 records, which could easily lead to overfitting issues and unjustified extrapolation. Our dataset includes millions of records, providing a significantly more credible base.

Schnyder et al [5] dug into drivers of PTSD symptoms following severe injuries sustained in accidents. It is not useful for our project, as it did not include any information on outside factors that resulted in the accident that caused these injuries. While accident-related injuries are related to our study, our focus is on using readily available information to predict traffic delays from accidents.

Zong et al [6] used an automobile accident dataset to determine which model worked better for predicting severities, the Bayesian network or regression models, ultimately finding the former to be superior. This finding is useful for our study and can help guide our selection of models to use. Our study will improve upon this by using a far bigger dataset, as well as identifying drivers of accident severity, as opposed to only looking at which modeling approach produces better results.

Ma et al [7] took a novel approach to modeling accident frequencies, which led to recommendations for safety treatment and design of highways. Knowledge of the types of road design associated with higher accident rates will benefit our analysis, and we intend to take this a step further to understand the impact these have on severities, as well as uncovering additional drivers.

Al-Ghamdi [8] took a binary view to studying accident severities, limiting his analysis to comparing fatal versus non-fatal accidents using logistic regression. His approach is not going to be useful for our project, as we intend to consider more than two possible outcomes. Further, he had access to just 9 independent variables to consider, limiting his scope considerably compared to what we should be able to achieve.

Wang et al [9] combined accident frequency and severity models to identify sections of roadway that were prone to accidents, ranking their results by monetary cost of accidents by location. This study will be beneficial to our project, as the general approach and variables considered should also be valuable and predictive in our analysis. Its scope was fairly limited, looking at 12,254 accidents in the London vicinity. Our study will be much larger in scale.

Alkheder et al [10] used an artificial neural network to predict accident severities in Abu Dhabi using 4 severity groupings: Minor, Moderate, Severe, and Death. While their approach and variables considered could be leverage for our project, their model accuracy fell off substantially as accident severity increased. The additional data available for our study should lead to much higher levels of accuracy.

Ratanavaraha and Suangka [11] modeled severities of accidents occurring on highways in Thailand. Their main finding was that an increase in speed led to an increase in accident severity. Speed should be a valuable predictor in our model as well, but we expect to find more than just one variable associated with accident severities.

Sameen and Pradhan [12] improved on previous artificial neural network models for predicting accident severity by leveraging recurrent neural networks, which are more effective when your data is sequential. This result could be useful for our project, insofar as providing a model framework to explore. The focus of this paper was more about finding the right model for predicting severities as opposed to the predictions themselves. We expect our model will provide more valuable insight and potential social benefit with a focus on identifying drivers that could be mitigated.

Abbas [13] studied causes of accidents occurring in Egypt, with a goal of recommending improvements to lower their overall frequency. Abbas specifically tested whether his data could be combined, or if each roadway needed to be studied separately. He found the former to be the case, which is useful for our project. We will improve on his findings by going beyond accident counts to prediction of severities, giving our results more practical applications.

Heilmeier Question Reference Footnotes

1 What are you trying to do? Articulate your objectives using absolutely no jargon.

2 How is it done today; what are the limits of current practice?

3 What’s new in your approach? Why will it be successful?

4 Who cares?

5 If you’re successful, what difference and impact will it make, and how do you measure them (e.g., via user studies, experiments, ground truth data, etc.)?

6 What are the risks and payoffs?

7 How much will it cost?

8 How long will it take?

9 What are the midterm and final “exams” to check for success? How will progress be measured?

Team Members

Robert King

Anna Lan

Kareem Naguib

Nathan Rugge

Notes

Project will be completed using the following tools:

1. Python version: 3.x.x
2. Hadoop and/or other cloud-computing software utility
3. Tableau, D3, and/or other visualization software(s)

References

1. US Accidents: A Countrywide Traffic Accident Dataset (2016-2019)
   1. <https://www.kaggle.com/sobhanmoosavi/us-accidents>
   2. Additional citations for dataset:
      1. Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019. (<https://arxiv.org/abs/1906.05409>)
      2. 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. (<https://arxiv.org/abs/1909.09638>)
2. Rezaie Moghaddam F, Afandizadeh S, Ziyadi M. “Prediction of accident severity using artificial neural networks.” IJCE. 2011; 9 (1) :41-48 (<http://ijce.iust.ac.ir/article-1-544-en.html>)
3. Tibebe Beshah, Shawndra Hill. “Mining Road Traffic Accident Data to Improve Safety: Role of Road-Related Factors on Accident Severity in Ethiopia.” AAAI Spring Symposium Series (2010) (<https://www.aaai.org/ocs/index.php/SSS/SSS10/paper/viewPaper/1173>)
4. Rui Garrido, Ana Bastos, Ana de Almeida, Jose Paulo Elvas. “Prediction of road accident severity using the ordered probit model.” (<https://www.sciencedirect.com/science/article/pii/S2352146514002701>)
5. Ulrich Schnyder, M.D.; Hanspeter Moergeli, M.A., M.Sc.; Richard Klaghofer, Ph.D.; and Claus Buddeberg, M.D. “Incidence and Prediction of Posttraumatic Stress Disorder Symptoms in Severely Injured Accident Victims.” (<https://ajp.psychiatryonline.org/doi/full/10.1176/appi.ajp.158.4.594>)
6. Fang Zong, Hongguo Xu, and Huiyong Zhang. “Prediction for Traffic Accident Severity: Comparing the Bayesian Network and Regression Models.” (<https://www.hindawi.com/journals/mpe/2013/475194/>)
7. Jianming Ma, Kara Kockelman, Paul Damien. “A Multivariate Poisson-Lognormal Regression Model for Prediction of Crash Counts by Severity, using Bayesian Methods.” (<https://www.caee.utexas.edu/prof/kockelman/public_html/TRB07MVPLN.pdf>)
8. Ali S Al-Ghamdi. “Using logistic regression to estimate the influence of accident factors on accident severity.” (<https://www.sciencedirect.com/science/article/abs/pii/S0001457501000732>)
9. Chao Wang, Mohammed A Quddus, Stephen G Ison. “Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model.” (<https://www.sciencedirect.com/science/article/abs/pii/S0001457511001357>)
10. Sharaf Alkheder, Madhar Taamneh, Salah Taamneh. “Severity Prediction of Traffic Accident Using an Artificial Neural Network.” (<https://onlinelibrary.wiley.com/doi/abs/10.1002/for.2425>)
11. Vatanavongs Ratanavaraha, Sonnarong Suangka. “Impacts of accident severity factors and loss values of crashes on expressways in Thailand.” (<https://www.sciencedirect.com/science/article/pii/S038611121300023X>)
12. Maher Ibrahim Sameen, Biswajeet Pradhan. “Severity Prediction of Traffic Accidents with Recurrent Neural Networks.” (<https://www.mdpi.com/2076-3417/7/6/476>)
13. Khaled A Abbas. “Traffic safety assessment and development of predictive models for accidents on rural roads in Egypt.” (<https://www.sciencedirect.com/science/article/abs/pii/S0001457502001458>)

Project Proposal Notes

* Interactive visualization:
  + Team to explore different models and choose the best for the user.
  + Visualization will have parameters that the user can select and this will change the display to build an inherent understanding of how each factor/feature in the dataset affects the model.
* Project title: Understanding the Severity of Accidents Through Predictive Modeling.
  + Alt Option: Understanding the factors for Accident Severity through Predictive Modeling.
* Modeling:
  + Need to find the appropriate platform to train the model.
* Collaboration & Division of Labor
  + Github for repository
  + Robert to send out his start at the EDA notebook for other members to begin their investigation.
* Montgomery County (MD) Fire & Rescue Services sends the following units for reports of vehicle collisions:
  + Limited Access Highways - 2 Engines, 2 Ambulances, 1 Rescue Squad
  + High Speed (40mph+) road - 1 Engine, 1 Ambulance, 1 Rescue Squad
  + Low Speed (35mph or less) - 1 Engine, 1 Ambulance
  + Non-Highway reported people trapped/pinned - 1 Engine, 1 Ambulance, 1 Rescue Squad
  + HazMat (fuel spills) adds 1 HazMat unit, 1 additional Engine
  + Additional special units might be added for additional odd circumstances such as cars off the highway into the woods or down an embankment.

Team 233 Project Presentation

**Introduction**

Today, first responders are called to the scene of an accident based on the description provided by the 911 caller. After classifying the incident, first responders are dispatched and the scene is managed to restore the roadways as safely and quickly as possible. The operator faces the challenges of inaccurate reports and overload from multiple callers – determining the traffic severity of the accident typically requires additional verification [ref best practices]. In addition to the problem the operator faces, transportation agencies have different criteria for determining the accident severity and its categorization - there isn’t a common classification scheme[ref IMPM]. Traffic severity is a key component used across agencies for determining the classification of the incident [ref IMPM] - this impacts the response, traffic control planning, resource optimization, and operational improvements [ref Analysis, Modeling, and Simulation].

Our project intends to provide a tool that helps people understand and explore the factors of accident severity on traffic. The tool will provide interactive visualization and predict severity based on immediately available information including time of day, weather conditions, and similar accidents in the area, as well as using features unique to the location itself (e.g., whether there is an amenity, railway, roundabout or even speed bump nearby). The impact of our study should provide benefits to first responders, transportation agencies, roadway designers and GPS developers and users alike, to mention a few.

* **First responders:** improve response decisions in presence of missing information or overload of information from multiple callers. Success can be measured through response time and site cleanup times.
* **Transportation agencies:** Improve planning and resource optimization. The gains can be measured by changes in fatality rates, response time, and site cleanup times.
* **Roadway designers:** Identification of accident-prone locations (as well as areas prone to *not* having accidents, where *this* information could be used to remediate the aforementioned accident-prone locations). Success for a roadway designer would be measured through relative accident frequencies post construction.
* **GPS developers:** Improve estimating delays for routing - the change in accuracy of time estimates can be a measure of success.

Our project risks primarily consist of inability to discover new findings that would bring benefits to our stakeholders6. The project team has identified free dataset(s) and will be able to conduct the analysis with minimal or no cost via leveraging team members computing resources and open source software packages7.

The project timeline will run over six weeks between 28 February 2020 and 17 April 2020 with milestones expected to be completed as follows8. It is expected that there may be instances of rework or shifts in the analysis trajectory based on findings during the analysis. Success at each of the milestones will be measured by completion of each interim deliverable, with results shared with all team members9. It is expected that responsibility for all deliverables will be shared equally by team members, but ownership of specific items may be assigned as the project progresses.

* 28 February - Project Proposal submitted.
* 13 March - Exploratory Data Analysis completed and results reviewed by team members; analysis methodology chosen and assigned.
* 27 March - Analysis completed and results reviewed by team members; Final Report sections assigned.
* 17 April - Final Report completed and submitted.

**Relevant Work**

There is extensive research conducted on predicting severity of accidents in terms of injury and traffic to understand the impact of human factors and road design on severity, feature and model selection, and model performance comparison. We will be leveraging general approach and exploring model frameworks as well as findings, but there are some limitations due to the dataset used and the information would not be available to the end-uses of our project.

**Human Factors**

**Roadway Design**

**Feature Selection and Model Performance**

We will be leveraging the general approach and feature selection from these studies as well as exploring model frameworks. Useful findings such as roadway designs associated with higher accident rates, and each roadway needs to be studied separately as opposed to combining datasets will be taken into consideration when developing our model.

Moghaddam et al [2] trained artificial neural networks to understand drivers of severities of accidents occurring on the highways of Tehran, Iran. Their study had a high degree of accuracy but noted many driver-specific characteristics and actions that led to higher severities. These data points would not be available for first responders, a focus of our study.

Beshah and Hill [3] studied accident severities in Ethiopia, using a variety of classification models including decision trees, naive Bayes, and KNN. Their findings dialed in on road design, but seemed very limited in terms of data they were able to source. It likely will not be useful for our project, which should be able to deliver a more meaningful output given the additional data available to us.

Garrido et al [4] looked at factors impacting the severity of injury in accidents in Portugal. Their study is useful for our project and notes some predictors and a methodology that we can start with in our analysis. The main shortcoming of their study was a limited dataset of just 4,528 records, which could easily lead to overfitting issues and unjustified extrapolation. Our dataset includes millions of records, providing a significantly more credible base.

Schnyder et al [5] dug into drivers of PTSD symptoms following severe injuries sustained in accidents. It is not useful for our project, as it did not include any information on outside factors that resulted in the accident that caused these injuries. While accident-related injuries are related to our study, our focus is on using readily available information to predict traffic delays from accidents.

Zong et al [6] used an automobile accident dataset to determine which model worked better for predicting severities, the Bayesian network or regression models, ultimately finding the former to be superior. This finding is useful for our study and can help guide our selection of models to use. Our study will improve upon this by using a far bigger dataset, as well as identifying drivers of accident severity, as opposed to only looking at which modeling approach produces better results.

Ma et al [7] took a novel approach to modeling accident frequencies, which led to recommendations for safety treatment and design of highways. Knowledge of the types of road design associated with higher accident rates will benefit our analysis, and we intend to take this a step further to understand the impact these have on severities, as well as uncovering additional drivers.

Al-Ghamdi [8] took a binary view to studying accident severities, limiting his analysis to comparing fatal versus non-fatal accidents using logistic regression. His approach is not going to be useful for our project, as we intend to consider more than two possible outcomes. Further, he had access to just 9 independent variables to consider, limiting his scope considerably compared to what we should be able to achieve.

Wang et al [9] combined accident frequency and severity models to identify sections of roadway that were prone to accidents, ranking their results by monetary cost of accidents by location. This study will be beneficial to our project, as the general approach and variables considered should also be valuable and predictive in our analysis. Its scope was fairly limited, looking at 12,254 accidents in the London vicinity. Our study will be much larger in scale.

Alkheder et al [10] used an artificial neural network to predict accident severities in Abu Dhabi using 4 severity groupings: Minor, Moderate, Severe, and Death. While their approach and variables considered could be leverage for our project, their model accuracy fell off substantially as accident severity increased. The additional data available for our study should lead to much higher levels of accuracy.

Ratanavaraha and Suangka [11] modeled severities of accidents occurring on highways in Thailand. Their main finding was that an increase in speed led to an increase in accident severity. Speed should be a valuable predictor in our model as well, but we expect to find more than just one variable associated with accident severities.

Sameen and Pradhan [12] improved on previous artificial neural network models for predicting accident severity by leveraging recurrent neural networks, which are more effective when your data is sequential. This result could be useful for our project, insofar as providing a model framework to explore. The focus of this paper was more about finding the right model for predicting severities as opposed to the predictions themselves. We expect our model will provide more valuable insight and potential social benefit with a focus on identifying drivers that could be mitigated.

Abbas [13] studied causes of accidents occurring in Egypt, with a goal of recommending improvements to lower their overall frequency. Abbas specifically tested whether his data could be combined, or if each roadway needed to be studied separately. He found the former to be the case, which is useful for our project. We will improve on his findings by going beyond accident counts to prediction of severities, giving our results more practical applications.

Notes:

For example, NY DOT determines the different levels of threshold based on anticipated duration, time of day and severity of the accident, with the strategy of “throwing more at it” the more severe the accident is classified. While Arizona DOT has three levels of threshold ranging from minimal traffic impact, traffic flow restriction, and severe (fatality, HAZMAT, homicide, etc.) with each level of severity having a difference in response in notification and prioritization [IMPM].