**Predicting the Traffic Collision Severities in Seattle City**

For Applied Data Science Capstone Project at Coursera

Yuying Ren

August 30, 2020

**Executive Summary**

The project presents collision severity prediction in Seattle city aimed to help travelers assess their traveling needs against risk factors that bear an impact on how severe collision might be. A comparison among four machine learning algorithms, namely K Nearest Neighbor, Decision Tree, Support Vector Machine, and Logistic Regression, are presented. Thirty eight attributes are investigated, and the major contributing factors and their effects are identified. The results indicate that the goodness of fit of Decision Tree is higher than that of the rest, though the performances among models are similar. Logistic Regression model is not a good fit. Project results can help improve the accuracy in collision severity prediction, which is one of the essential steps in modern day accident management process. It presents value-add to both travelers and to first responders. By recognizing the key influences, this research also provides insights that could feed into suggestions for government to take effective measures to reduce accident impacts and improve traffic safety.

**Table of Contents**

1. **Introduction**
2. **Literature Review**
3. **Data Acquisition and Cleaning**
4. **Predictive Modeling**
5. **Results and Model Evaluation**
6. **Discussions**
7. **Conclusion and Future Directions**
8. **Acknowledgement**
9. **References**
10. **Introduction**
    1. **Background**

As massive amount of data that involves almost every single aspect of our lives as well as more powerful analytical tools become available, the use of data in traffic and transportation started to serve as the foundation for making informed decisions on nearly work at all major cities around the world, from safety improvements to repaving to grant applications.

Traffic accidents are the 8th cause of mortality in different countries and are expected to rise to the 3rd rank by 2020[[1]](#footnote-1). There is no surprise that increased resources have been directed at accident severity prediction, which has become one of the most interesting areas of studies that attract attentions from academic researchers to government agencies. By understanding when, where, how and why severe collisions happen, cities across the globe could not only stop millions of disabilities and deaths, but also avoid or at least ease the cities’ finance and improve public health.

* 1. **Problem**

Numerous factors, such as environmental, vehicle-related, host (driver, pedestrian), and their type of interaction affect the characteristics of these accidents. In addition, to accurately identify these factors, different solutions, such as evaluation of public health, risk factors, and Haddon matrix, are developed, which further adds to the collection of different features that may have an impact on collision severities. On top of that, area- or city-specific factors, such as traffic law, weather conditions, and human behavior----both of the drivers and of the pedestrians----could influence collision severities.

The narrow it down, this project aims to predict whether a collision in Seattle city will cause personal injury or not by looking at a set of features provided by Seattle Department of Transportation (SDOT).

* 1. **Interest**

Being one of the major steps of accident management, accident severity prediction can provide crucial information for travelers to assess their traveling needs and reduce their risks of being involved in collisions. It will also help emergency responders to evaluate the severity level of accidents, estimate the potential impacts, and implement efficient accident management procedures.

1. **Literature Review**

Many previous studies provided insights into collision severity prediction. However, most studies only used after-math factors and the traditional statistical models. For example, Zong, Xu, and Zhang looked at fatalities, number of injuries and property damage using Bayesian network and Regression model. The results indicate that the goodness of fit of Bayesian network is higher than that of Regression models in accident severity modeling. This finding facilitates the improvement of accuracy for accident severity prediction[[2]](#footnote-2).

Later on, Zong, Zhang, Xu, Zhu, and Wang examined both severity and duration and presented a model system to predict severity and duration by employing Ordered Probit model and Hazard model, respectively. With the developed models, they found that the goodness-of-fit of Ordered Probit model is higher than that of SVC model in severity modeling. In addition, accident severity is proven to be an important determinant of duration; that is, more fatalities and injuries in the accident lead to longer duration[[3]](#footnote-3).

Although previous works presented the effectiveness of numerous models in accident severity modeling, there is no contribution that conducts a quantitative comparison of different classification models, especially machine learning algorithms. Moreover, previous studies tend to take a first-responders’ point of view, instead of that of a traveler. Therefore, this project will add values in helping travels predict and assess their traveling needs against potential collision risks utilizing machine learning algorithms.

1. **Data Acquisition and Cleaning**
   1. **Data Source**

Seattle Department of Transportation (SDOT)’s mission is to deliver a transportation system that provides safe and affordable access to places and opportunities. Traffic collisions data is part of their annual Traffic Report released to the public.

The *All collisions* data provided by SDOT includes all types of collisions and associated information, as such weather condition, road condition, and light condition from 2004 to Present. The breadth and depth of the data collected allows objective discussion of project merits and results. A list of attribute information can be found [here](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf).

* 1. **Data Cleaning**

The first thing I did was to locate the target variable. Since this project aims to predict collision severity, it is clear that the target variable is expressed as numerical numbers and stored in the “SEVERITYCODE” column, where “1” denotes property damage only and “2” denotes injuries.

Data downloaded has missing information, shown as blank in the CSV file. The metadata notes stated that these are due to lack of information recorded by first responders or people who filed collision report. If I were to remove all rows that include NAN value, more than 120,000 rows will be removed as most rows contain at least one NAN value. Thus, I decided to keep all rows until feature selection stage.

There are several problems with the dataset. First, incident dates are not displayed in the same format. I converted the “INCDATE” column to Datetime object in pandas library.

Second, there are some inconsistencies in how information is displayed. For example, the “UNDERINFL” column, a field that flags whether or not a driver involved were under the influence of drugs or alcohol, has four values: “Y”, “N”, “0”, and “1”. The metadata description clarified that “0” denotes “N” and “1” denotes “Y”. I converted “0” and “1” into “N” and “Y” to make the field values consistent.

Third, only 751 out of 194,673 records are collisions that happened at alleys, compared to that happened at blocks and intersect. This is intuitive and understandable. However, to avoid the possibility of bias, I decided to drop the 751 “Alley” records.

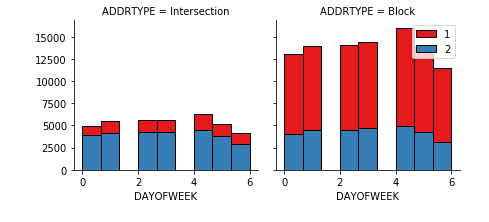
* 1. **Feature Selection**

After data cleaning, there were 193,922 samples and 38 columns in the data. Upon examining the meaning of each column, it was clear that some fields are not meaningful for collision prediction. For example, four columns ---- “OBJECTID” (ESRI unique identifier), “INCKEY” (A unique key for the incident), “COLDETKEY” (Secondary key for the incident), and “REPORTNO” (Report Number) ---- are associated with unique identifiers and are not informative in predicting collision severity. In addition, “SDOT\_COLCODE” and “ST\_COLCODE”, along with their descriptions, are codes given by SDOT and by the state, respectively. These are not features that may impact collision severity.

In addition, several fields that may bear an impact on collision severity have significant amount of missing data. For example, “SPEEDING”, a flag on whether or not speeding was a factor in the collision, has 104,857 records with “blank” values. Even though it is intuitive that speeding may influence collision severity, due to missing data, I decided to exclude it from the modeling.

After discarding these inappropriate fields, I used plots to further examine the rest of the fields. For example, Figure 1 displays the severity code with day of the week people travel and indicates that during weekend, more collisions were identified as type 1 collision ---- the collisions with property damage only. It seems that day of the week people travel is a reasonable indicator.

**Figure 1**



Similarly, I then examined address type (i.e. whether a collision incurred at blocks, or intersects), whether the driver involved is under influence, weather condition, road condition, and light condition and decided these are appropriate features for the predictive model.

* 1. **Preprocessing**

Since weather condition, road condition, and light condition are categorical fields, I use one hot encoding technique to convert them to binary variables and appended them to the feature data frame. All three features have “Unknown” and “Other” categories. Since these two categories are not informative to predict collision severity, I decided to drop these two categories from the feature data frame.

Next I revisited the NAN values in the dataset. There are more than 6,000 rows with NAN values in the feature data frame and are removed. Now I have a clean dataset ready for modeling with 187,223 sample records. The clean dataset consists of a collection of features, namely day of week people travel, address type, whether a driver is under influence, weather condition, road condition, and light condition, and the target variable----severity code, where “1” denotes property damage only and “2” denotes injuries.

The final step is to normalize data, which is straightforward suing StandardScaler in preprocessing in Sklearn.

1. **Predictive Modeling**

Two types of models, regression and classification, can be used in a predictive model. Since the target value is binary, the application of classification models is much more straightforward. In this project, I decided to apply four machine learning algorithms, namely the K Nearest Neighbor (KNN), Decision Tree, Support Vector Machine (SVM), and Logistic Regression, to the dataset, using Jaccard Index, F1 Score and Log Loss as the evaluation metrics. I also plotted confusion matrix to examine type 1 and type 2 errors more closely.

* 1. **Train-Test Split and Model Application**

I first divided the samples into two datasets, one for training and one for testing. I used the rule of thumb, which is 80/20 split in this case, where the training set consists of 149,778 records and the testing set 37,445 records.

Since the four models of my choice already exist in Sklearn, I simply imported KNN, Decision Tree, SVM and Logistic Regression from Sklearn and trained them using the training dataset. I then made the prediction and evaluated model performance using the test dataset.

* 1. **Performances Evaluation**

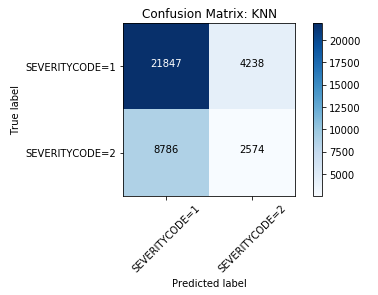
Among the four models, the Decision Tree and SVM models have very similar performance and they performed the best based on Jaccard Index (~70%), though that of the Logistic Regression is very close. KNN model has the highest F1 score (~62%). It is interesting to note that looking at Jaccard Index and F1 score, the differences among models are small.

**Table 1**

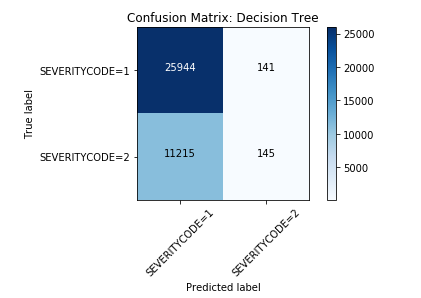
|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithms** | **Jaccard Index** | **F1 Score** | **Log Loss** |
| KNN | 0.6522 | 0.6226 | N/A |
| Decision Tree | 0.6967 | 0.5791 | N/A |
| SVM | 0.6967 | 0.5791 | N/A |
| Logistic Regression | 0.6964 | 0.5821 | 0.9610 |

I also evaluated the models using confusion matrix. In this particular problem, lower false negative count is more important. In other words, it is more important to be sure that a traveler or a first responder will correctly predict a severe collision where injuries are involved, rather than correctly predict a less sever collision that involves property damage only, as these predictions may impact people’s behavior. A false negative prediction may cause the traveler to be more risk-seeking and the first responder to be less attentive. In the confusion matrix with low false-positive cases, the Decision Tree model had slightly lower false negative count than other models (Figure 2-5).

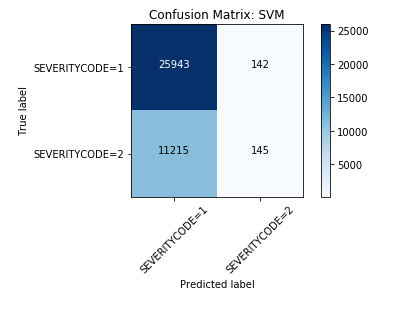
**Figure 2**

****

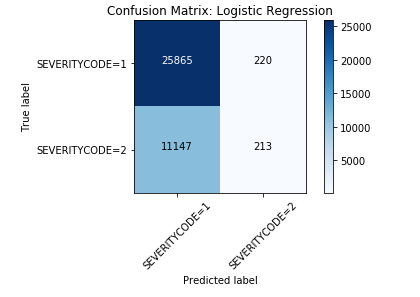
**Figure 3**



**Figure 4**

****

**Figure 5**



1. **Discussions**

The first thing to point out is the dataset studied is not a balance dataset. The raw dataset has 136,485 records associated with type 1 collision and only 58,188 records with type 2 collision, roughly 70/30 spilt. It is expected. The vast majority of the collisions only caused property damage, not person injury.

To combat imbalanced data, Accuracy is not the metric to use. It can be misleading. There are metrics that have been designed to tell a more truthful story when working with imbalanced classes. In this particular case, F1 Score, a weighted average of precision and recall, is more appropriate.

Second, the Log Loss of the Logistic Regression model indicates that the model has high uncertainty. Recall that the Log Loss metric takes into account the probabilities underlying the model, and not only the final output of the classification. A low Log Loss means a low uncertainty or entropy. In this particular case where the target variable is binary, non-informative value is roughly 0.6. This is obtained by setting p (i) = prevalence (i), in this case, 30/70. My Log Loss of 0.96 indicates that the Logistic Regression model has higher entropy than that of a “dumb” prediction, which is randomly predicting 30% type 2 collisions and 70% type 1 collisions. I concluded that the Logistic Regression model is not a good fit.

1. **Conclusion and Future Directions**

In this project, I analyzed the relationship between collision severity and thirty eight associated information categories provided by SDOT. I identified six related features, namely week of day people travel, whether the driver is under influence, weather condition, road condition, and light condition, among the most important features that affect collision severity.

Four machine learning algorithms, that is, K Nearest Neighbor, Decision Tree, Support Vector Machines, and Logistic Regression, are considered. The goodness of fit of the four methods is compared according to the test results. The results suggest that Decision Tree is more suitable for collision severity prediction in Seattle city, though the performances of the four models are very similar.

Project results can be applied to predicting collision severity in Seattle and identifying the key effects of contributing factors. By comparing the four models, it also makes a methodological contribution in enhancing prediction accuracy of severity estimation.

One limitation of current work is that some factors, such as driver characteristics and whether speeding was observed, which have potential effects on collision severity, are not considered because of the large amount of missing data. Proper data source should be identified and further study should be conducted to examine the impacts of these factors.

1. **Acknowledgements**

Thanks to everyone who made the IBM Data Science Professional Certificate available, especially everyone on the Instructors team who helped me get here. I also want to thank all my peers who kindly reviewed my assignments. Without you, I could not be able to do any of this.

1. **References**
2. Kambiz Masoumi, Arash Forouzan, Hassan Barzegari, Ali Asgari Darian, Fakher Rahim, Behzad Zohrevandi, Somayeh Nabi, “Effective Factors in Severity of Traffic Accident-Related Traumas; an Epidemiologic Study Based on the Haddon Matrix”*, Emerg (Tehran),* 2016 Spring; 4(2): 78–82, PMCID: PMC4893755
3. Fang Zong, Hongguo Xu, Huiyong Zhang, "Prediction for Traffic Accident Severity: Comparing the Bayesian Network and Regression Models", *Mathematical Problems in Engineering*, vol. 2013, Article ID 475194, 9 pages, 2013, <https://doi.org/10.1155/2013/475194>
4. Fang Zong, Huiyong Zhang, Hongguo Xu, Xiumei Zhu, and Lu Wang, “Application of Discrete Mathematics in Urban Transportation System Analysis”, *Research Article*, Volume 2013, Article ID 547904, <https://doi.org/10.1155/2013/547904>
5. 2019 TRAFFIC REPORT, Seattle Department of Transportation, <https://www.seattle.gov/Documents/Departments/SDOT/VisionZero/2019_Traffic_Report.pdf>
6. [Seok-Lyong Lee](https://www.researchgate.net/profile/Seok-Lyong_Lee), “Predicting Traffic Accident Severity Using Classification Techniques”, October 2015, *Advanced Science Letters* 21(10):3128-3131, DOI: 10.1166/asl.2015.6446

1. Kambiz Masoumi, Arash Forouzan, Hassan Barzegari, Ali Asgari Darian, Fakher Rahim, Behzad Zohrevandi, Somayeh Nabi, “Effective Factors in Severity of Traffic Accident-Related Traumas; an Epidemiologic Study Based on the Haddon Matrix”*, Emerg (Tehran),* 2016 Spring; 4(2): 78–82, PMCID: PMC4893755 [↑](#footnote-ref-1)
2. Fang Zong, Hongguo Xu, Huiyong Zhang, "Prediction for Traffic Accident Severity: Comparing the Bayesian Network and Regression Models", *Mathematical Problems in Engineering*, vol. 2013, Article ID 475194, 9 pages, 2013. https://doi.org/10.1155/2013/475194 [↑](#footnote-ref-2)
3. Fang Zong, Huiyong Zhang, Hongguo Xu, Xiumei Zhu, and Lu Wang, “Application of Discrete Mathematics in Urban Transportation System Analysis”, *Research Article*, Volume 2013, Article ID 547904, https://doi.org/10.1155/2013/547904 [↑](#footnote-ref-3)