Car Collision Severity Modeling in Seattle, WA, USA

Grady Weinheimer
9 September 2020

INTRODUCTION

The consequences of car collisions vary greatly from incident to incident. Since 2004, over 86,000 total injuries and nearly 400 deaths have resulted from car collisions in Seattle, WA, USA[(1)]. There is no easy cure to totally prevent car collisions in the future, but predicting the expected severity of future car collisions can help Seattle's emergency services become better prepared. Having this prediction capability would allow the Seattle local government to make an optimized budget for equipment and operations relating to car collision emergencies. To make these insightful predictions, a machine learning model can be created to accurately predict car collision severity by the use of variables that represent driving conditions and driver behavior. The ultimate goal of this project was to create a model that can accurately predict car collision severity based on collision data. That way, a model can take various estimated conditions, such as weather or amount of drunk drivers, and predict the severity of car collisions for future situations.

DATA

The model for predicting car collision severity depended on historical collision data from the Seattle government's open data portal. The data set is a comma-separated values (.csv) file called 'Collisions'. It contains collision data from thousands of car collisions that occurred in Seattle from 2004 to the present. Each incident, or row, has a labeled severity code which allows for supervised machine learning.

A code that corresponds to the severity of the collision:

- **3**—fatality
- **2b**—serious injury
- **2**—injury
- **1**—prop damage
- **0**—unknown

Figure 1: Collision Severity Values and their Meaning

Since the "unknown" severity code cases do not provide value in predicting a collision's severity, incidents from that category were excluded from analysis.

A significant amount of data preparation was required. In this dataset, many of the variables are categorical. For variables with only two value options (such as yes or no), one-hot encoding was used to make the value numerical. One example of this was the variable of driver drunkenness. After applying one-hot encoding, a value of '1' meant the driver was under the influence, and a value of '0' meant the driver was not under the influence. For weather, the values were first classified as hazardous or non-hazardous before one-hot encoding. In a similar fashion, the days of the week were classified as weekend or weekday before one-hot encoding. For variables with three or more different values, dummy values were created. One example of this was the time of the incident variable. Regarding the time of the incident, the times were each classified as one of the four major times of day: morning, daytime, evening, and nighttime. Then dummy values were applied to numerically indicate what time of day each incident occurred. This was also done with the collision type, and time of year (season) variables. After making the variables numerical and easier to use for machine learning models, there was a little more data clean-up required. One major issue with the unaltered dataset is that the amount of incidents per severity code is quite unbalanced. As a result the model's predictions would be naturally biased towards the majority case if the data set was left unbalanced. To balance the data set, 10,000 samples from each severity code, 1 to 3, were taken for analysis. Two of the severity codes had more than 10,000 samples, so they were downsampled. The other two codes were opposite and needed upsampling. The last major issue was the presence of missing values for some of the entries and some of the variables. To rectify this, the average value of a variable for a given severity code was inserted to replace the missing value.

The goal of the project was to take variables describing the driving conditions and driver behavior at each incident and make a model that uses that data to predict accident severity. The variables deemed the best to represent driving conditions and driver behavior included the collision type, weather, time of day, season, driver drunkenness, and day of the week. These variables were selected for the featured data set for modeling.

METHODOLOGY

Data importing, exploration, analysis, modeling, and evaluation were all conducted in the IBM Cloud Watson Studio platform. A project was created and the historical collision data was uploaded there. A Jupyter Notebook was created to conduct the entire programming workflow in Python, and the data was imported into it. Numpy, Pandas, and sci-kit learn were the main Python programming packages used. For data exploration, the goal was to see which variables relating to driving conditions and driver behavior were the most relevant. Since much of the data was categorical, one-hot encoding and dummies were used for many variables. Once the featured set of variables were selected and refined, the Seaborn software package was used to visually compare the differences in values for the featured variables regarding the collision code. For example, the total counts of daytime collisions for each collision code were totaled up

and then graphed against each other. See the below figures for visuals explaining the key differences between the severity codes of values 1, 2, 2b, and 3.

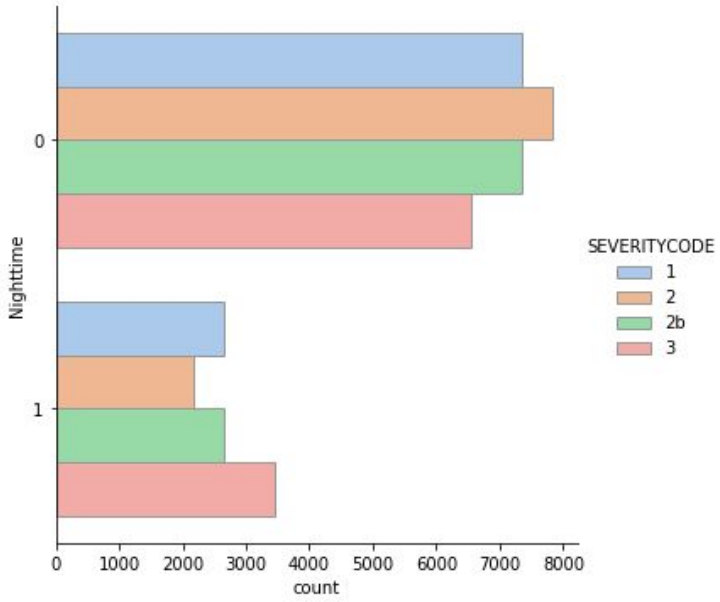


Figure 2: For incidents occurring at nighttime, their 'Nighttime' values were scored as '1'. All other incidents were scored as '0' for this variable. Based on the graph, more serious collisions (Severity Codes: 2b and 3) were more likely to occur at night than collisions of lower severity.

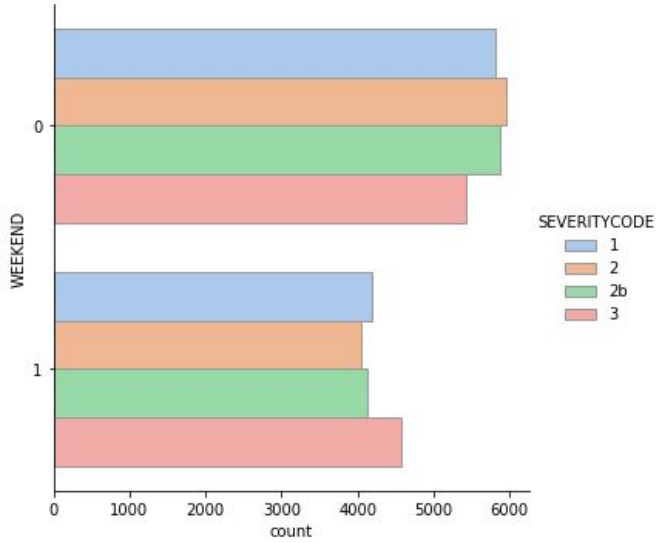
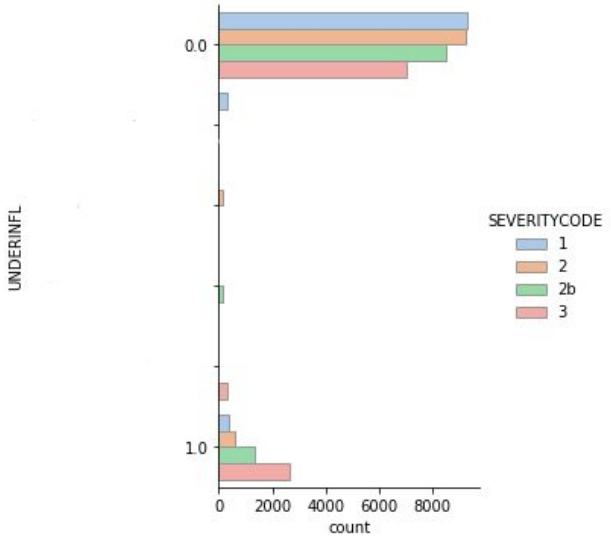


Figure 3: For incidents occurring on the weekend, their 'Weekend' values were scored as '1'. All other incidents were scored as '0' for this variable. Based on the graph, more serious collisions (Severity Code: 3) were more likely to occur on the weekend than collisions of lower severity.



<u>Figure 4</u>: For incidents involving a drunk driver, their 'UNDERINFL' values were scored as '1'. All other incidents were scored as '0' for this variable. Based on the graph, more serious collisions (Severity Code: 3) were more likely to occur on the weekend than collisions of lower severity. Note - the in-between values were average values used for removing null values.

Figures 2-4 represent three of the main variables that differentiated the severity codes. Additional visuals showing characteristics of the different severity codes can be found in the supplement coding notebook.

Once the critical variables were determined for modeling, a feature set was created. Null values in each severity code were replaced with average values and the amount of samples from each severity code was equalized to make a balanced dataset. The feature set was split into a training set, which took 80% of the samples, and a test set, which had 20% of the samples. Four machine learning models were created including K Nearest Neighbors (KNN), Decision Tree (DT), Support Vector Machine (SVM), and Logistic Regression (LR). Due to the large size of the data set, machine learning model parameters were only tuned on K Nearest Neighbors and Decision Tree. The computing power available in the IBM Watson Studio Lite Plan could not process the code for tuning SVM and LR in a reasonable amount of time. The final models were created by fitting the training data. After that, the Jaccard similarity and F1-score for each model were calculated by predicting the outputs of the model from the test data and comparing it to the true test data labels. Lastly, those results were entered in a table to see which model had the highest accuracy.

RESULTS

The project goal of creating a machine learning model to predict car collision severity from historical collision data was successfully completed in a Jupyter Notebook. Four different

machine learning models were created, and it is clearly evident that the Decision Tree model is the best performing one with a Jaccard similarity and F1-score of 0.5706 and 0.5628 respectively. The below table shows a summary of the model accuracy scored for each model algorithm.

Algorithm	Jaccard	F1-score
KNN	0.5310	0.5262
Decision Tree	0.5706	0.5628
SVM	0.4848	0.4719
LogisticRegression	0.4999	0.4862

Figure 5: A table showing results from evaluating the accuracy of the Machine Learning algorithms.

The model is capable of taking newly found characteristics of driving conditions and driver behavior to make predictions on car collision severity. It will not predict the likelihood of a car collision, but rather the severity of consequences, which was the main goal.

DISCUSSION

The overall goal of creating a model to predict car collision severity based on historical data regarding driving conditions and driver behavior was successful. However, the best model found has fairly poor accuracy with a Jaccard similarity and F1-score of 0.5706 and 0.5628 respectively. Car collisions are caused by a wide variety of factors and can happen at any given time a car is on the road. As a result, the data randomness makes it difficult to differentiate why some collisions end up being more severe than others when looking at data from collisions. Another challenge to overcome was the fact that many critical factors that affect the severity of a car collision were not available in the data set. These include the wearing of seatbelts, the safety rating of the car, the volume of traffic of the roads driven, and many more. Having this additional data would surely make the severity code categories more distinctive. For example, the National Highway Traffic and Safety Administration in the USA estimates that 14,955 lives were saved by seat belts in 2017.[(2)] Knowing this, the differences between the severity codes could have been further accentuated if seatbelt data was also present in the original data set. The initial imbalance of severity codes per incident made balancing the data very difficult. While over 100,000 cases were classified as “property damage”, less than 400 had death(s) involved. In addition to data randomness and insufficient data availability, there was room for improvement in parameter tuning for the models. As mentioned in the Methodology section, the computing power available in the IBM Watson Studio Lite Plan could not process the code for tuning the SVM and LR models in a reasonable amount of time. Despite some data shortcomings and opportunity for model improvement, this model can still be useful for the

Seattle Government to predict car collision severity. It may also be good to supplement this model with a model that predicts the probability or total number of car collisions to occur in a year. While it is good to know what variables contribute to collision severity, it is also good to be able to estimate how many collisions will happen in a year. Combined, those two models could truly help the Seattle Government with its emergency services budget and planning. In the meantime, the Seattle Government will have this model available for use, and the Decision Tree model can receive new data to make predictions as needed.

CONCLUSION

To reiterate, the overall project goal was to create a machine learning model that can predict car collision severity based on historical data regarding driving conditions and driver behavior. Using collision type, weather, dates, times, and driver drunkenness proved useful in predicting the severity code defined by the Seattle Government. While the task was accomplished, there is room for improvement in both the data used and the model tuning parameters. The Seattle Government now has a Decision Tree model with a Jaccard similarity and F1-score of 0.5706 and 0.5628, respectively, available for use. This will help the government predict expected relative amounts of severe car collisions versus minor ones which will ultimately help the city's budgeting and planning for emergency services.

REFERENCES

[1] "Collisions." City of Seattle, 14 Mar. 2018, https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf.

[2] "Seat Belts." National Highway Traffic Safety Administration, https://www.nhtsa.gov/risky-driving/seat-belts#:~:text=Of%20the%2037%2C133%20people%20killed,had%20been%20wearing%20seat%20belts.