Coursera Capstone

IBM Applied Data Science Capstone

Car Accident Severity in Seattle

By:Shalini S

September 2020

1.Introduction

1.1.Background

Seattle is a seaport city located in the west coast of the United States. It is the seat of King County, Washington. Seattle is the largest city in both the state of Washington and the Pacific Northwest region of North America. Seattle is the northernmost U.S. city with at least 500,000 people, farther north than Canadian cities such as Toronto, Ottawa, and Montreal, and at about the same latitude as Salzburg, Austria. Regarding the transportation, the city has started moving away from the automobile and towards mass transit. From 2004 to 2009, the annual number of unlinked public transportation trips increased by approximately 21%.Even with the high numbers for public transportation—and a typical 58 hours per year searching for parking among Seattle drivers—a 29.7 percent of downtown commutes still happen alone in a vehicle. The survey notes that this includes people who drove with children, so that could be a factor in the commute choice. Also Seattleites own a lot of cars. It has 637 cars for every 1000 residents.

1.2.Business Problem

But the road accident curve is not so cool in Seattle. They have been reporting a high number of car accidents due to varied reasons. So the government of Seattle is in its pursuit of reducing the car collisions that is happening on its roads. The best way to approach this would be to analyse the data that contains details of the previously occurred such accidents. There is a high need for an algorithm or system that could prevent or initially, at least flatten the curve of accidents taking place, through its predictions so that required force is sent for their rescue. Beyond several factors relating to the physical, mental and driving conditions of the people who are on the driver seat, the status of the roads being travelled through, the weather and climatic conditions and other common commodities are also to be taken into account.

1.3.Target Audience

The target audience of the project are the Seattle government, the police department, the rescue groups and the car insurance companies too. This project and its results might open up some statistics for the target audience and make insightful decisions for reducing the number of accidents and injuries on its roads.

2.Data acquisition and cleaning

2.1.Data sources

The data we are considering for this is the unbalanced dataset that is provided by the Seattle Department of Transportation Traffic Management Division. The dataset that is used here consists of 194673 entries of accidents with 37 attributes to be considered for each. Each row is given a severity code which is dependent on other factors/attributes. It covers accidents from January 2004 to May 2020.

Our main variable "SEVERITY CODE" contains values from 0 to 4.

It corresponds to the severity of the collision.

3->fatality

2b->serious injury

2->injury

1->prop damage

- 0->unknown as given in metadata.

The severity code in the data set considered is 1 and 2.

1- 136485

2- 58188

Other important variables include:

1. ADDRTYPE: Collision address type: Alley, Block, Intersection

2. LOCATION: Description of the general location of the collision

3. PERSONCOUNT: The total number of people involved in the collision helps identify severity level

4. PEDCOUNT: The number of pedestrians involved in the collision helps identify severity level

5. PEDCYLCOUNT: The number of bicycles involved in the collision helps identify severity level

6. VEHCOUNT: The number of vehicles involved in the collision identifies severity level

7. JUNCTIONTYPE: Category of junction at which collision took place helps identify where most collisions occur

8. WEATHER: A description of the weather conditions during the time of the collision

9. ROADCOND: The condition of the road during the collision

10. LIGHTCOND: The light conditions during the collision

11. SPEEDING: Whether or not speeding was a factor in the collision (Y/N)

12. SEGLANEKEY: A key for the lane segment in which the collision occurred

13. CROSSWALKKEY: A key for the crosswalk at which the collision occurred

14. HITPARKEDCAR: Whether or not the collision involved hitting a parked car

15. INJURIES: The number of total injuries in the collision.

16. SERIOUSINJURIES: The number of serious injuries in the collision.

17. FATALITIES: The number of fatalities in the collision.

18. INATTENTIONIND : Whether or not collision was due to inattention.

19. UNDERINFL: Whether or not a driver involved was under the influence of drugs or alcohol.

20. INCDTTM : The date and time of the incident.

2.2.Data cleaning

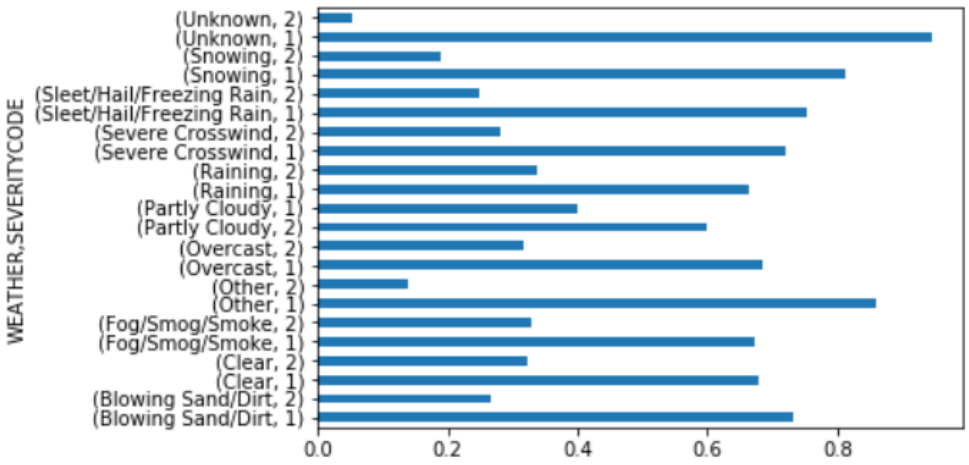
For cleaning the data, we are considering the columns with varied entries to be equalized into correct format. Certain columns from which certain details are needed are formatted accordingly. For example, the incident date is present. For that, need day of week for analysis. So such pre-processing works are carried out.

2.3.Feature selection

For analysis, as severity code is the target variable, certain important independent variables are considered. WEATHER, LIGHTCOND, ROADCOND, ADDRTYPE, DAYOFWEEK (extracted from incident date) are considered.

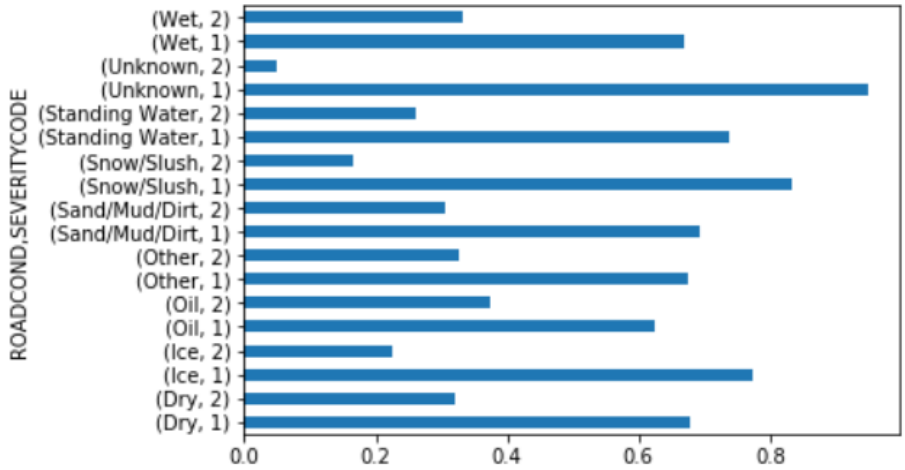
3.Exploratory data analysis and visualization

3.1.Relationship between severity code and weather



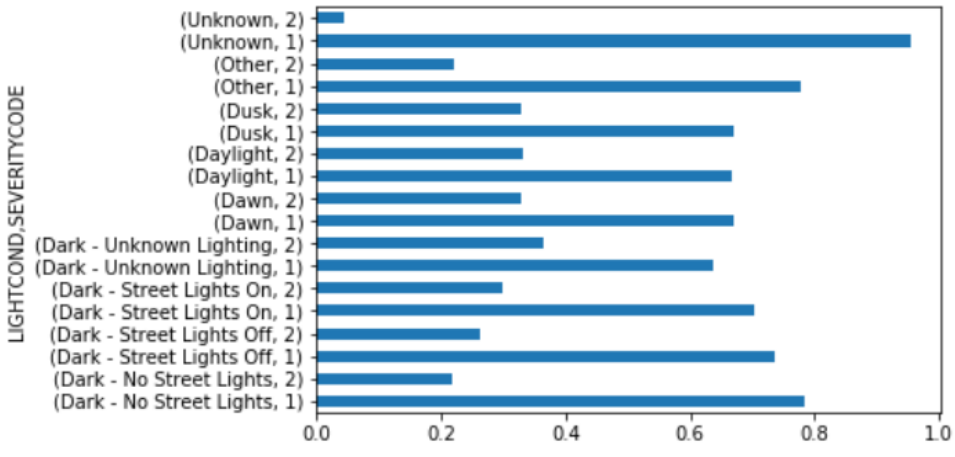
Shows the relationship between severity code and weather. Taking severity code as 1 and 2 and all weather types.

3.2.Relationship between severity code and road conditions



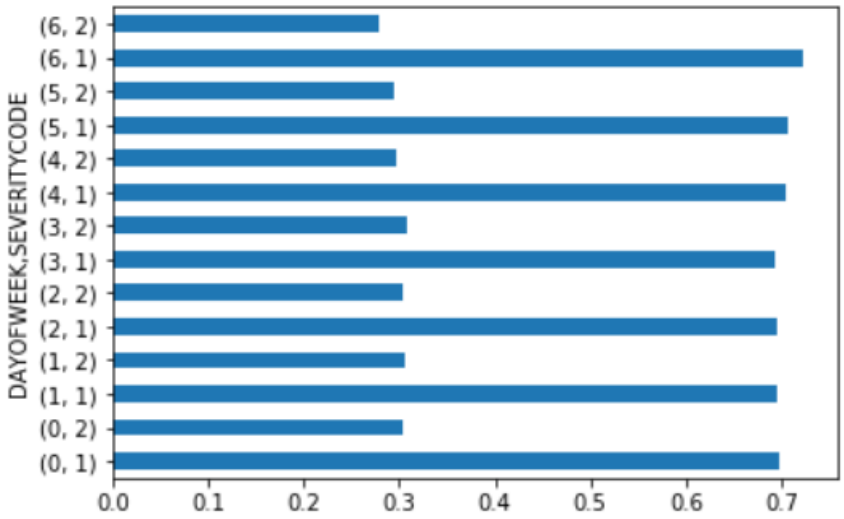
Shows the relationship between severity code and road conditions. Taking severity code as 1 and 2 and all road conditions.

3.3.Relationship between severity code and light conditions



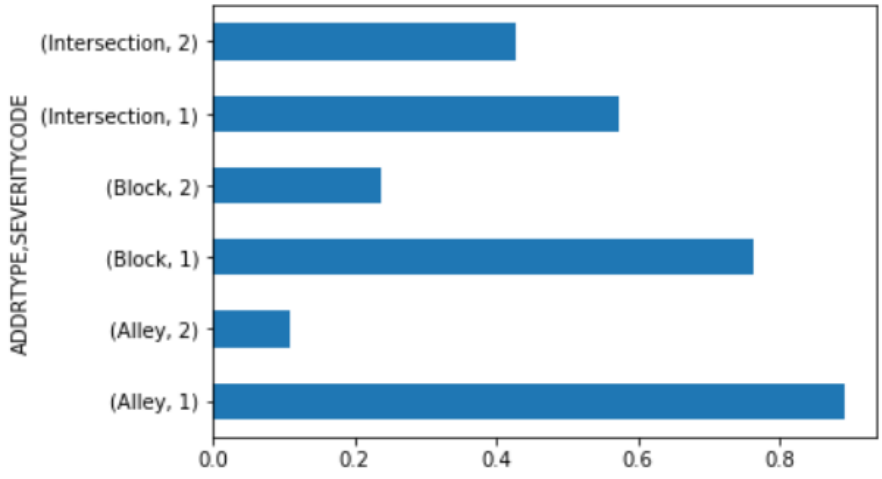
Shows the relationship between severity code and light conditions. Taking severity code as 1 and 2 and all light conditions.

3.4.Relationship between severity code and day of week



Shows the relationship between severity code and day of week. Taking severity code as 1 and 2 and day of week extracted from incident date.

3.5.Relationship between severity code and address type



Shows the relationship between severity code and address type. Taking severity code as 1 and 2 and address type.

4.Machine Learning Models

4.1.K-nearest Neighbor

It can be used for both classification and regression problems. However, it is more widely used in classification problems in the industry. K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors. The case being assigned to the class is most common amongst its K nearest neighbors measured by a distance function.

These distance functions can be Euclidean, Manhattan, Minkowski and Hamming distance. First three functions are used for continuous function and fourth one (Hamming) for categorical variables. If K = 1, then the case is simply assigned to the class of its nearest neighbor. At times, choosing K turns out to be a challenge while performing kNN modeling.

4.2.Decision Tree Algorithm

It is a type of supervised learning algorithm that is mostly used for classification problems. Surprisingly, it works for both categorical and continuous dependent variables. In this algorithm, we split the population into two or more homogeneous sets. This is done based on most significant attributes/ independent variables to make as distinct groups as possible.

4.3.Logistic Regression

It is a classification not a regression algorithm. It is used to estimate discrete values (Binary values like 0/1, yes/no, true/false) based on given set of independent variable(s). In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function. Hence, it is also known as logit regression. Since, it predicts the probability, its output values lies between 0 and 1 (as expected).

Let’s say your friend gives you a puzzle to solve. There are only 2 outcome scenarios – either you solve it or you don’t. Now imagine, that you are being given wide range of puzzles / quizzes in an attempt to understand which subjects you are good at. The outcome to this study would be something like this – if you are given a trigonometry

based tenth grade problem, you are 70% likely to solve it. On the other hand, if it is grade fifth history question, the probability of getting an answer is only 30%.

The dataset is split into train and test sets for modelling as X_train, X_test, y_train and y_test sets. These are applied to the machine learning models and the best one obtained for prediction.

5.Evalaution

There are certain metrics that are considered for evaluation of the models.

5.1.Jaccard score

The Jaccard index [1], or Jaccard similarity coefficient, defined as the size of the intersection divided by the size of the union of two label sets, is used to compare set of predicted labels for a sample to the corresponding set of labels in y_true.

This jaccard_score may be a poor metric if there are no positives for some samples or classes. Jaccard is undefined if there are no true or predicted labels, and our implementation will return a score of 0 with a warning.

5.2.F1-score

Computing the F1 score, also known as balanced F-score or F-measure.

The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal.

F1 = 2 * (precision * recall) / (precision + recall)

When true positive + false positive == 0, precision is undefined; When true positive + false negative == 0, recall is undefined. In such cases, by default the metric will be set to 0, as will f-score, and UndefinedMetricWarning will be raised. This behavior can be modified with zero_division.

5.3.Accuracy

In multilabel classification, this function computes subset accuracy: the set of labels predicted for a sample must *exactly* match the corresponding set of labels in y_true.

In binary and multiclass classification, this function is equal to the jaccard_score function.

6.Results

The models of k Nearest Neighbor, Decision Tree and logistic regression were able to provide classification and prediction.

Algorithm	Jaccard	F1-score	Accuracy
KNN	0.69	0.48	0.69
Decision Tree	0.69	0.41	0.69
LogisticRegression	0.70	0.41	0.70

7.Discussion

This shows the evaluation results of the machine learning process. The logistic regression shows a better accuracy based on the dataset considered.

So considering these the prediction of accidents could be done using WEATHER, ROADCOND, LIGHTCOND, ADDRTYPE and DAYOFWEEK as these play a major role apart from physical and driving conditions of the driver.

8.Conclusion and future directions

So this is the analysis on the car accident severity and the prediction chance of these. With these, the accidents could be predicted based on day to day conditions and appropriate actions could be taken before hand to avoid unbearable losses and problems. All the necessary stakeholders could be alerted so that there is ease in travelling.