Car Accident Severity Prediction Report

IBM Applied Data Science Capstone

Qing Yang (Tony) Xie

October 3, 2020

Table of Contents

Introduction/Business Problem 3

Data 3

Data Understanding 3

Analysis Platform 5

Data Cleaning 5

Methodology 6

Data Exploration 6

Classification 10

Results 11

Discussion 12

Conclusion 12

Introduction/Business Problem

In the modern society, travelling by motor vehicle is essential for a large part of the population to reach their destinations efficiently. Daily commute by car has become an important aspect of many people's everyday life to get to work or school. Due to various type of circumstances, car accidents can occur. Car accidents can inflict damage to the vehicle and harm to the people involved. Specifically, it can cause economic loss, injuries and even fatalities. This study focuses on analyzing several main factors that could lead to car accidents and developing a suitable model to predict the severity of accident under specified conditions. The target audiences are motor vehicle drivers. The goal of this study is to produce a model that improves driver's awareness of a dangerous condition and reduce the likelihood of a car accident.

Data

Data Understanding

The data set used in this study is originated from Seattle Department of Transportation, Traffic Records group. The original data set contains records from 2004 to present and consist of 37 attributes. Each attribute has a maximum of 194673 records. The target variable or dependent variable is "SEVERITYCODE", which describes the severity of the collision. Here is a detailed description of the type of severity codes in the dataset:

- 3 – fatality
- 2b – serious injury
- 2 – injury
- 1 – prop damage

The original dataset needs to be cleaned and pre-processed due to the following reasons:

1. Unbalanced data: There are 136485 records for severity code 1 and 58188 records for severity code 2.
2. Null entries: 17 independent variables in the dataset contains null entries
3. Duplicated attribute: "SEVERITYCODE" appeared twice in the dataset
4. Text records: some text records need to be converted to numerical type for analysis

5. Key Variable Selection: Several less relevant attributes need to be dropped out before carrying out the analysis

The key variables selected for this study are collision address type (ADDRTYPE), weather conditions (WEATHER), road conditions (ROADCOND) and light conditions (LIGHTCOND). These attributes are selected because they are indicative of driver's surrounding geographical and atmospheric conditions. Unlike other attributes in the original data set such as UNDERINFL (if or not a driver involved was under the influence of drugs or alcohol) which can be fully controlled and prevented by the driver before attempting to start the road trip or HITPARKEDCAR (Whether or not the collision involved hitting a parked car) which describes one of the outcomes of the collision, the selected attributes focus more on the naturally occurring conditions that the driver is likely to encounter, actively or passively. Below is a more detailed description of the selected key variables extracted from the Metadata Form as a part of the original data set package.

Table 1. Detailed summary of the selected key attribute

Attribute	**Data type**	**Description**
ADDRTYPE	Text	Collision address type: • Alley • Block • Intersection
WEATHER	Text	A description of the weather conditions during the time of the collision.
ROADCOND	Text	The condition of the road during the collision.
LIGHTCOND	Text	The light conditions during the collision.

If the selected key attributes have a significant impact on the severity of car accidents, an alert system can be developed to assess the current day's atmospheric conditions and the type of junctions and road conditions that the driver would encounter based on the chosen start location and destination. It can then notify the driver about the suitability of driving under these conditions.

Analysis Platform

All the analysis was conducted using Jupyter Notebook on IBM Cloud's Watson Studio platform. The programming language used for the analysis is Python (version 3.6).

Data Cleaning

The original dataset was cleaned and preprocessed to make it adequate for data analysis. The target variable "SEVERITYCODE" appeared twice in the dataset. The redundant "SEVERITYCODE" was removed to address this issue.

Since the selected key variables for this study are collision address type (ADDRTYPE), weather conditions (WEATHER), road conditions (ROADCOND) and light conditions (LIGHTCOND) as discussed in the "Data Understanding" section, all other attributes except for the target variable were removed from the dataset.

An aggregation of null values for each attribute was populated in Table 2. Note that the total number of rows/records for the entire data set is 194673. The attribute with the greatest number of null values is LIGHTCOND, with only 5170 null records. Since the number of null values only account for a small portion of the entire dataset, they are removed prior to analysis.

Table 2. Number of null values for each attribute in the original dataset

Attribute	**Number of Null Values**
SEVERITYCODE	0
ADDRTYPE	1926
WEATHER	5081
ROADCOND	5012
LIGHTCOND	5170

Since severity code is the target variable, it was examined more closely to prevent imbalanced data. As shown in Table 3, the number of values under severity code 1 is 2.35 times greater than the number of values under severity code 1. This kind of uneven class distribution may lead to skewed or biased results. Therefore, the dataset was balanced by down sampling the records in severity code 1 to match the number of records in severity code 2. After the process of down sampling, the resulting dataset contained equal amount of records for severity code 1 and severity code 2, which is 56883.

Table 3. Number of values under each severity code in the dataset after filtering out null values

Severity Code	Number of Values
1	136485
2	58188

The final step of preprocessing involved checking the data type for each attribute and converting them to the proper form. It was determined that the severity code has a data type of "int64", whereas the selected attributes all have the same "object" data type. The data types of selected attributes are converted to categorical values and subsequently assigned with numerical labels in preparation for analysis.

Methodology

Data Exploration

The preprocessed data set is first examined using series of count plots to determine the counts in each categorical bin for the selected attributes. Figure 1, Figure 2, Figure 3 and Figure 4 shows the count plot for collision address type, weather conditions, road conditions and light conditions, respectively.

From these figures, it is apparent that some categories in each attribute occurs considerably more frequently than the other categories. For instance, there are substantially more car accident records from "Block" and "Intersection" categories than the "Alley" category from "collision address type" attribute. This is expected as there tends to be more traffic at an intersection or a mid-block than an alley, which can lead to higher number of accidents.

Note that the highest counts of car accident records for weather conditions, road conditions and light conditions attributes are "Clear", "Dry" and "Daylight", respectively. This is somewhat less expected as they are commonly considered as safe conditions when driving. One possible explanation is that there might be more vehicles on the road under these conditions, which can in turn increase the probability of collision.

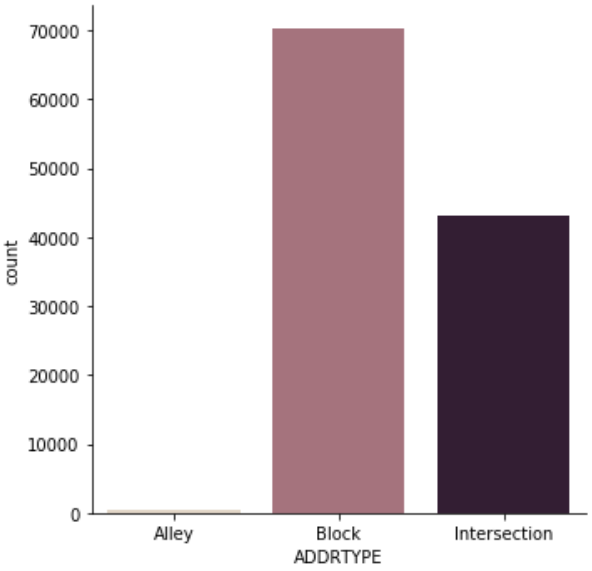

Figure 1. Count plot for the collision address type attribute

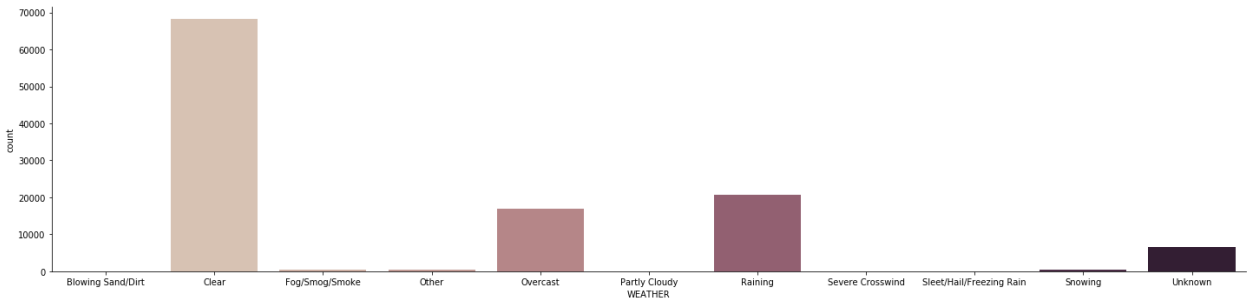

Figure 2. Count plot for the weather condition attribute

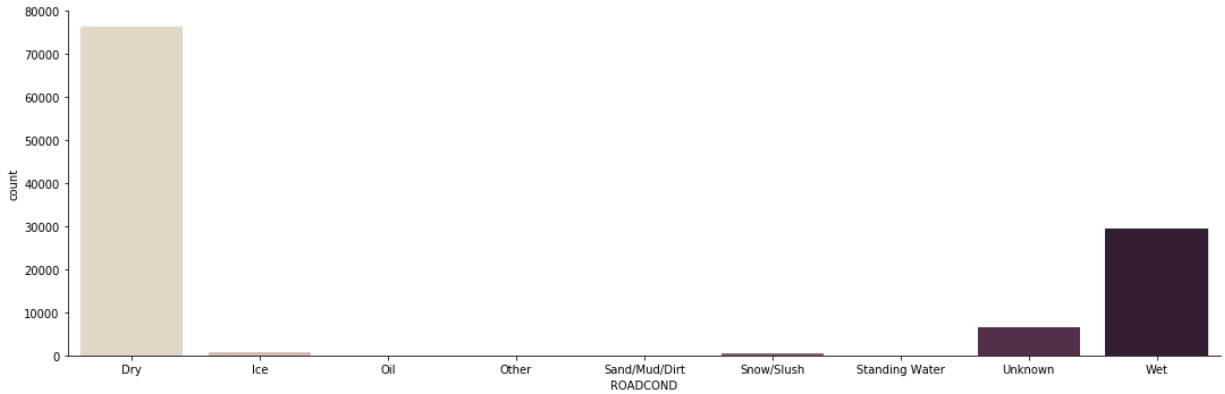

Figure 3. Count plot for the road condition attribute

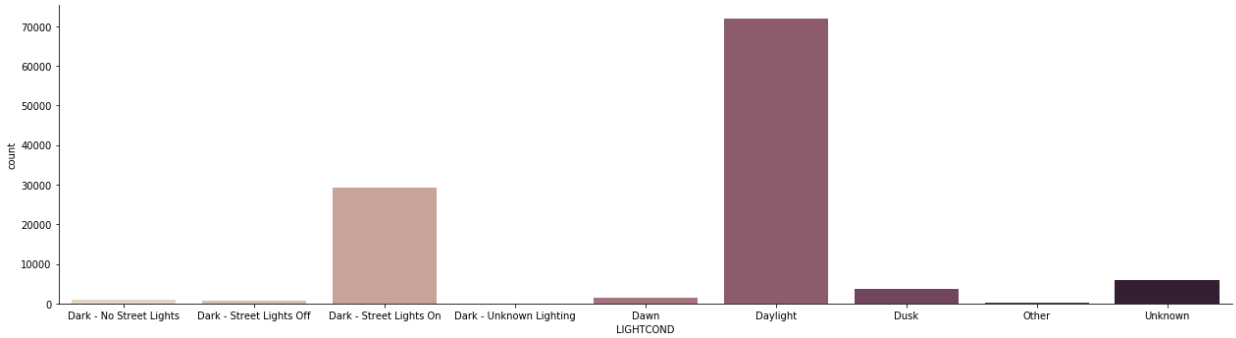

Figure 4. Count plot for the light condition attribute

The preprocessed data is then further examined by grouping each categorical bin in each attribute by severity code. Figure 5, Figure 6, Figure 7 and Figure 8 shows the count plots for each attribute after grouping by severity code.

From these plots, it is evident that there are several categories under each attribute that have noticeable differences between the count of severity 1 accidents and count of severity 2 accidents. Specifically, in Figure 5, there are thousands more severity 1 accidents than severity 2 accidents from the block. In contrast, there are thousands more severity 2 accidents than severity 1 accidents from the intersection. These types of discrepancies indicate that the level of severity might be related to some of the conditions in the selected attributes.

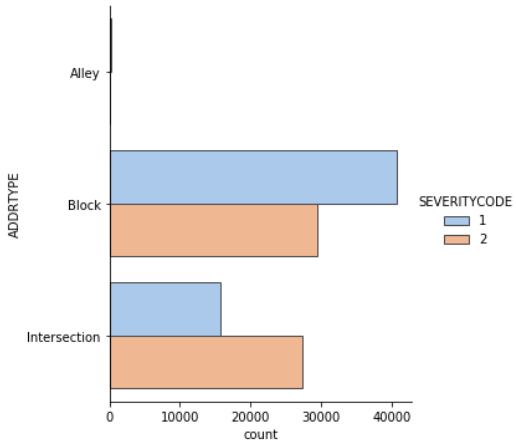

Figure 5. Counts grouped by severity code for collision address type attribute

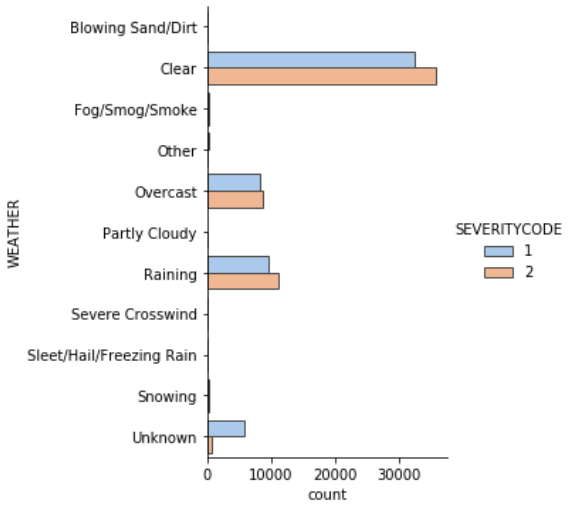

Figure 6. Counts grouped by severity code for weather conditions attribute

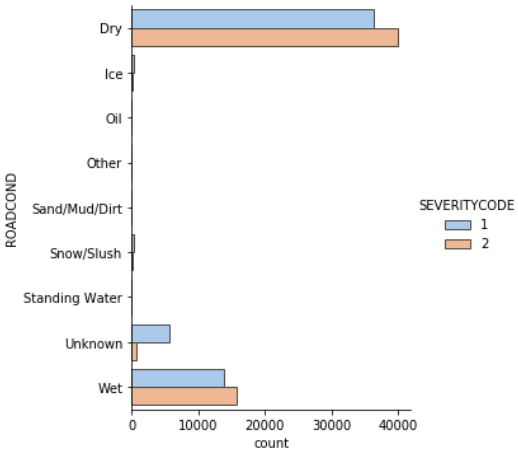

Figure 7. Counts grouped by severity code for road conditions attribute

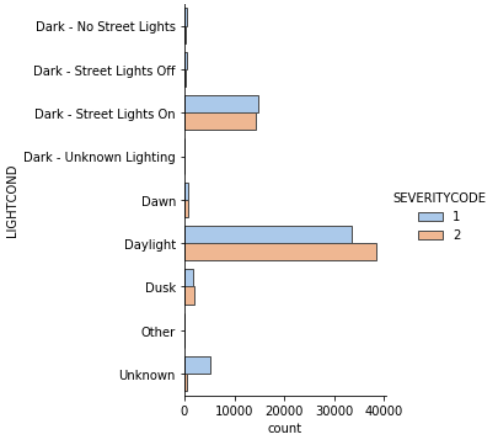

Figure 8. Counts grouped by severity code for light conditions attribute

Classification

Since the goal of this study is to determine the outcome (severity) of car accident under different conditions, the following machine learning algorithms were used to build the model for this study:

- K Nearest Neighbor (KNN)
- Decision Tree
- Support Vector Machine
- Logistic Regression

Before applying these algorithms, data was first normalized to achieve 0 mean and unit variance. Then, the dataset was split into 2 parts, where 70% of the data was used for training the model and 30% of the data was used for testing the accuracy of the models.

K Nearest Neighbor is a non-parametric algorithm used to predict the outcome of events based on 'K' nearest points. For this study, K Nearest Neighbor was applied iteratively for k values between 1 and 15. Figure 9 shows the accuracy of K Nearest Neighbor with different k values. Based on the plot, the best accuracy was attained when k equals 11. Therefore, the model was then re-constructed using an optimal k value of 11.

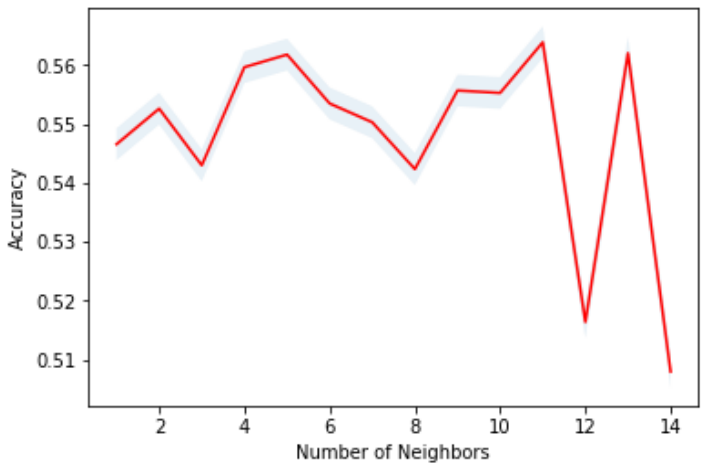

Figure 9. Accuracy of K Nearest Neighbor with different k values

Decision Tree is an algorithm that predicts event outcomes using a tree-like structure. In this study, the default settings were applied to most of the properties of decision tree classifier to

build the model. The criterion used for the split was “entropy”. A max depth of 5 was used to prevent the nodes from over expanding.

Support Vector Machine is a classification method that categorizes data by mapping points to high-dimensional feature space so that they can be separated by hyperplanes. The kernel function used for this model was Radial Basis Function (RBF). A default C value of 1 was used as this algorithm is quite computationally intensive for the number of data points that need to be processed. It would require time and better computational resources to iterate through different C values effectively.

Logistic Regression is a variation of linear regression that predicts the probability of a class label with respect to the independent variables. In this case, the class label would be severity and the independent variables would be the selected attributes. The solver used for this model was “liblinear” and a C value of 0.01 was used to achieve strong regularization.

Results

Table 1 shows the accuracy of the models used in this study. The accuracy of each model was evaluated using the following metrics: Jaccard similarity coefficient score, F1-score and log loss. Jaccard similarity coefficient score calculates the size of intersection between the predicted data set and actual test data set over their total size. F1-score determines test accuracy based on precision and recall. Log loss is a loss function that penalizes inaccurate prediction, which should be relatively low when the model is highly accurate. Among the applied algorithms, log loss is only suitable for evaluating logistic regression. Hence to compare these models, Jaccard similarity coefficient score and F1-score are used instead. As shown in Table 4, the average of Jaccard score and F1-score for Support Vector Machine model is slightly higher than the other models used in this study.

Table 4. Accuracy of KNN, Decision Tree, Support Vector Machine and Logistic Regression

Algorithm	Jaccard	F1-score	LogLoss	Average (Jaccard & F1-score)
KNN	0.51641	0.51207	NA	0.51424
Decision Tree	0.60149	0.59513	NA	0.59831
SVM	0.60152	0.59531	NA	0.59842
LogisticRegression	0.59815	0.59337	0.66458	0.59576

Discussion

Based on the results from each evaluation metrics, Support Vector Machine model is slightly more accurate than the other three models. Therefore, when evaluating severity of car accidents through collision address type, weather conditions, road conditions and light conditions, Support Vector Machine model should be used. However, since the average of Jaccard and F1-score is close to 0.6, the model is only moderately accurate. Note that other models have relatively similar accuracy results, with less than 1 percent difference between decision tree and Support Vector Machine. Thus, for any future studies, it is recommended to explore other factors that may contribute to car accidents within or outside of the original dataset. It is also recommended to test different maximum depth for decision tree and C values for Support Vector Machine and Logistic Regression to increase the accuracy of existing models.

Conclusion

According to the results of this study, Support Vector Machine only has moderate accuracy to predict the severity of car accident when using collision address type, weather conditions, road conditions and light conditions as the key attributes. Although existing models can be optimized by tuning max depth for decision tree and C values for Support Vector Machine and Logistic Regression, other car accident related attributes can also be explored in future studies to build models with higher accuracy.