Seattle Traffic Accident Study

Geoffrey Longtin

1. Introduction

1.1 Background

In Seattle every year countless accidents occur on the city roads which are linked to driver external factors. Current there are many apps that improve the driving experience by helping the drivers navigate the roads more efficiently by mapping the faster routes, cheapest routes or avoiding traffic jams. However, there are no features to help predict the risks of accidents due to external factors while driving.

1.2 Problem

The ultimate goal of this study is to use the data at hand and help built a predictive model which would help the driver identify the risks of accidents based on the current external factors such as light, road visibility and weather.

1.3 Interest

The entire Seattle community would greatly benefit from such an approach since drivers would be able to avoid bad driving conditions, or become more aware of bad conditions before hand as to avoid to be caught by surprise. Overall, road safety would improve overall and also reduce costs related to accidents.

2. Data Acquisition and Cleaning

2.1 Data Source

The Seattle police department has recorded in their Traffic Records all traffic accidents since 2004. This data can be found using the following link: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

The data contains 194 673 rows and 38 features. The feature SEVERITYCODE will be used as the target variable since it will help predict accidents. Moreover, out of the list of possible accidents, the data only contains severity codes of 1 and 2 which are respectively Property Damage and Injury types of collisions.

2.2 Data Cleaning and Feature Selection

The first step in the data cleaning operation is to remove any accident cases where the driver was at fault. Therefore, we dropped all the data which had Y (Yes) in the columns INATTENTIONIND (Inattention while driving), UNDERINFL (Driver was under the influence of drugs or alcohol), and SPEEDING (Speeding was a factor in the accident). This chain of operation resulted in data containing 152 293 rows.

The second step is to keep only the three features we desire for this analysis being: 1) WEATHER, 2) ROADCOND (Road Condition) and 3) LIGHTCOND (Light Condition). These feature have been selected since they are the only features which relate to external conditions during driving, and thus would help build the most useful models to help drivers decide accident risks based on factors driving conditions.

Therefore, all other columns were dropped resulting data of 4 features (which includes the target variable) as shown in the image below

	SEVERITYCODE	WEATHER	ROADCOND	LIGHTCOND
0	2	Overcast	Wet	Daylight
1	1	Raining	Wet	Dark - Street Lights On
2	1	Overcast	Dry	Daylight
3	1	Clear	Dry	Daylight
4	2	Raining	Wet	Daylight

The third operation required to delete all data with labels “Other” and “Unknown” as they are not helping in the modeling.

The fourth operation for the models to be more realistic, we need to balance the cases for each type of Severity Code. Therefore, it was found that the data at the previous stage had 88 259 rows for Code 1, and 40 857 rows for Code 2. We therefore randomly selected 40 857 rows from the majority cases Code 1, and keeping all the remaining Code 1 cases. This leaves use with a database of 81 714 rows and 4 features.

And lastly, for the models to be able to predict the column data will also need to be translated into numbers with the corresponding categorization:

For Weather:

0: Blowing Sand/Dirt

1: Clear

2: Fog/Smog/Smoke

3: Overcast

4: Partly Cloudy

5: Raining

6: Severe Crosswind

7: Sleet/Hail/Freezing Rain

8: Snowing

For Road Condition:

0: Dry

1: Ice

2: Oil

3: Sand/Mud/Dirt

4: Snow/Slush

5: Standing Water

6: Wet

For Light Condition: (Note: All 'dark' attributes are put into a single group '0')

0: Dark - No Street Lights

0: Dark - Street Lights Off

0: Dark - Street Lights On

0: Dark - Unknown Lighting

1: Dawn

2: Daylight

3: Dusk

Thus the final database is made of 81 714 rows and 7 features (including the 3 new categorization columns) as shown below.

	SEVERITYCODE	WEATHER	ROADCOND	LIGHTCOND	WEATHER_CAT	ROADCOND_CAT	LIGHTCOND_CAT
193258	1	Raining	Wet	Daylight	5	6	2
53471	1	Raining	Wet	Daylight	5	6	2
115753	1	Raining	Dry	Daylight	5	0	2
170248	1	Clear	Wet	Daylight	1	6	2
29902	1	Clear	Dry	Dark - Street Lights On	1	0	0

3. Exploratory Data Analysis

We first start by looking at count for each categories for all 3 variables and see which are the more predominant.

```
df_downsampled['WEATHER_CAT'].value_counts()
```

```
1    52719
5    15471
3    12857
8      333
2      248
7       48
0       25
6        9
4        4
Name: WEATHER_CAT, dtype: int64
```

From the Weather data it is show that the top 3 categories are as follows: 1) Dry, 2) Raining, 3) Overcast

```
df_downsampled['ROADCOND_CAT'].value_counts()
```

```
0    58994
6    21894
1      398
4      316
2       41
3       36
5       35
Name: ROADCOND_CAT, dtype: int64
```

From the Road Condition data it is show that the top 3 categories are as follows: 1) Dry, 2) Wet, 3) Sand/Mud/Dirt

```
df_downsampled['LIGHTCOND_CAT'].value_counts()
```

```
2    55646
0    22130
3     2786
1     1152
Name: LIGHTCOND_CAT, dtype: int64
```

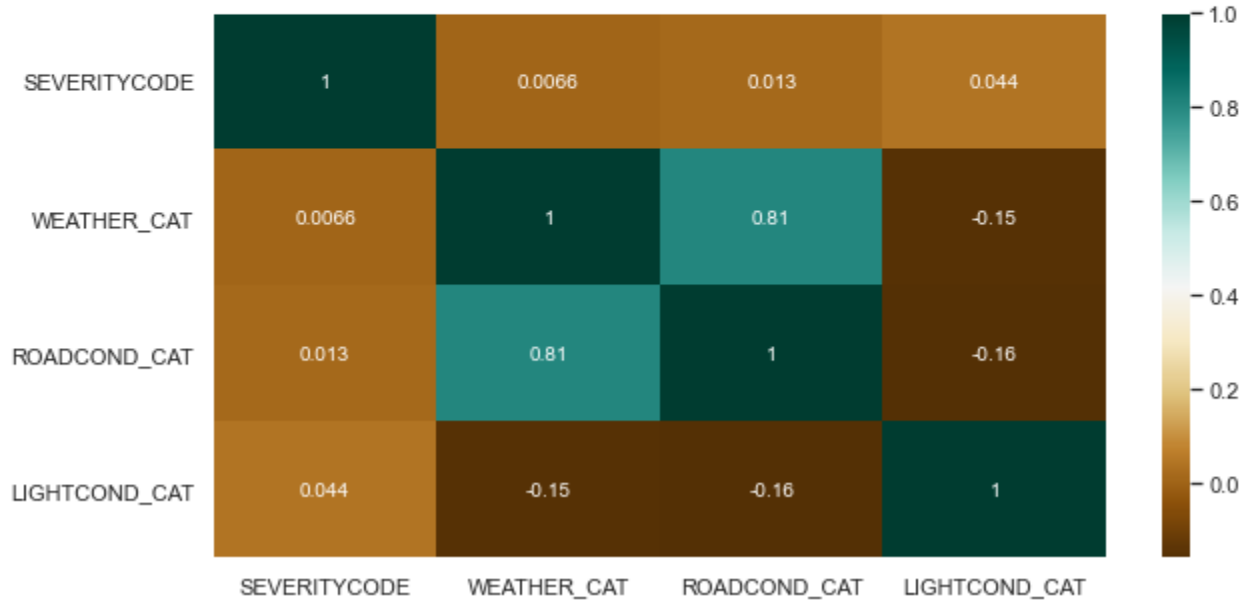
From the Light Condition data it is show that the top 3 categories are as follows: 1) Daylight, 2) Dark, 3) Dusk

From the three set of observation above, the majority of accidents would occur in Dry, and Daylight Conditions, this might be contradictory to our purpose of finding conditions leading to accidents. These results can be attributed to the fact that the majority of traveling is done during those conditions and therefore skews the data.

When look at the correlation between the target variable and the 3 variables, we notice that all 3 variable are mildly correlating to the Severity Code, which the highest correlation being with the Light Condition.

As for the inter variable correlation, we see somewhat a strong correlation between Weather and Road Condition, which is to be expected since the weather directly affects road conditions.

	SEVERITYCODE	WEATHER_CAT	ROADCOND_CAT	LIGHTCOND_CAT
SEVERITYCODE	1.000000	0.006581	0.013403	0.043564
WEATHER_CAT	0.006581	1.000000	0.809683	-0.150318
ROADCOND_CAT	0.013403	0.809683	1.000000	-0.159022
LIGHTCOND_CAT	0.043564	-0.150318	-0.159022	1.000000

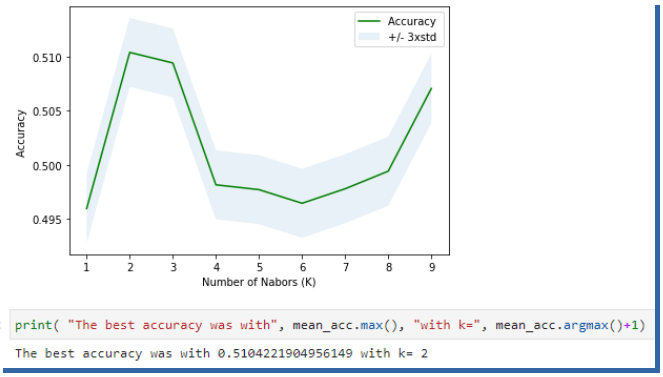


4. Predictive Modeling

All predictive and evaluation calculations are done using a distributed data of 70% for training and 30% for testing.

4.1 K-Nearest Neighbor (KNN) Model

When building a KNN model, we first need to assess which is the best K value needed. In order to do this we drafted the following graph to evaluate this value.



From this graph, we find that K equal to 2 is the best fit for this model. In doing so, we built a KNN model and found that the accuracy was as follows:

- Jaccard score is 0.3496

- F1 score is 0.5103

4.2 Decision Tree Model

The Decision Tree Model was built using a entropy criterion and a maximum depth of 7.

The model evaluation show that

- Jaccard score is 0.1912

- F1 score is 0.4748

4.3 Logistic Regression Model

The Logistic Regression Model was built using a C value of 0.1, and liblinear solver.

The model evaluation show that

- Jaccard score is 0.2759

- F1 score is 0.5062

- Logloss score is 0.6923

5. Conclusions

Based on the 3 models built, the most accurate model uses a Logistic Regression approach. Despite being the most accurate, we can clearly see that with a Logloss score of 0.6923 the is room for improvement for the model.

6. Recommendations

As discussed in the conclusion, the Logistic Regression could take advantage of better accuracy.

When doing the data exploration, we noticed that many of the data for accident were happening during the dry and daylight conditions, which could be attributed to the fact that most driving occurs in these conditions. Therefore the data could benefit by being balanced to reflect more balanced data by types of variable categories.

Moreover, some other driver data which is not covered in the original dataset could be gather to further see correlation with the Severity Code such as the age, sex, physical disabilities, type of car, etc. This would allow to have a better customized model for each type of drivers and help better predict the accident risks based on external driving conditions.