Coursera - IBM Data Science Capstone

Predicting Accident Severity

Car accident severity (Week 2)

Rousseau, Antoine
10-21-2020

Contents

Problem 2
Methodology 2
Data Understanding 2
Data Preparation 3
Modelling 4
Results 5
Discussion 5
Conclusion 6

Problem

Traffic accidents are an endemic problem in the US. In 2018 alone, 12 million vehicles were involved in a car crash[1]. Individuals take the road, often out of necessity, not thinking twice about the conditions they will face, and how much they will put themselves and others at risk. **What if we could predict the severity of an accident, if one were to occur, based on time, date and environmental data, and thereby provide information that could reduce their frequency?**

As a result, authorities could optimize the allocation of rapid-response services and signal the risk of taking the road to the population in real-time through billboards. Car manufacturers could signal this information in the navigation system of the cars. Individuals could have access to an app that would suggest a higher risk of a severe accident if they were to take the road.

Methodology

Our analysis is built upon the Seattle Collisions dataset. It contains, in tabular form, 31 attributes detailing each of the 194 673 collisions that took place in the Seattle from 2004 to today. Each entry is labelled according to severity.

1- We will start by describing the dataset, gauging the quality of the data, and assessing the relevance to the problem of the available attributes.
2- We will then transform it, selecting the useful attributes, addressing missing or unknown entries, and encoding it.
3- We will then run several classification algorithms and evaluate their performance.
4- Finally, we will discuss the findings and assess how the model answers the problem statement.

Data Understanding

The first step in analysing the data is to get an understanding of the attributes. Reading the Metadata pdf, we can allocate the attributes in the following buckets:

- **ID**: OBJECTID, INCKEY, COLDETKEY, REPORTNO, STATUS
- **Location**: X, Y, ADDRTYPE, INTKEY, LOCATION, JUNCTIONTYPE, SEGLANEKEY, CROSSWALKKEY, HITPARKEDCAR
- **Unknown**: EXCEPTRSNCODE, EXCEPTRSNDESC
- **Accident description**: SEVERITYDESC, COLLISIONTYPE, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, SDOT_COLCODE, SDOT_COLDESC, SDOTCOLNUM, ST_COLCODE, ST_COLDESC
- **Time/Date**: INCDATE, INCDTTM
- **Driver Behavior/State**: INATTENTIONIND, UNDERINFL, PEDROWNOTGRNT, SPEEDING
- **Environmental**: WEATHER, ROADCOND, LIGHTCOND

To answer our problem, we are interested in the predictive ability of the Time/Date and Environmental attributes and will therefore discard all other attributes in the data preparation stage.

[1] https://www.statista.com/topics/3708/road-accidents-in-the-us/

We print the table to see how the data frame looks (Figure 1).

Figure 1

	SEVERITYCODE	INCDATE	INCDTTM	WEATHER	ROADCOND	LIGHTCOND
0	2	2013/03/27 00:00:00+00	3/27/2013 2:54:00 PM	Overcast	Wet	Daylight
1	1	2006/12/20 00:00:00+00	12/20/2006 6:55:00 PM	Raining	Wet	Dark - Street Lights On
2	1	2004/11/18 00:00:00+00	11/18/2004 10:20:00 AM	Overcast	Dry	Daylight
3	1	2013/03/29 00:00:00+00	3/29/2013 9:26:00 AM	Clear	Dry	Daylight
4	2	2004/01/28 00:00:00+00	1/28/2004 8:04:00 AM	Raining	Wet	Daylight

We describe the data (Figure 2).

Figure 2

```
RangeIndex: 194673 entries, 0 to 194672
Data columns (total 6 columns):
SEVERITYCODE    194673 non-null int64
INCDATE         194673 non-null object
INCDTTM         194673 non-null object
WEATHER         189592 non-null object
ROADCOND        189661 non-null object
LIGHTCOND       189503 non-null object
```

As we can see, the two Time/Date attributes are not of the “datetime” type nor the Environmental attributes of the “category” type. This will have to be addressed in the data preparation stage. We can also see that several rows from the Environmental attributes are missing (there are less than 194673 values for WEATHER, ROADCOND and LIGHTCOND)

The next step is checking the quality of the data starting with missing values. Checking the number of rows with at least one null reveals that less than 3% of them contain missing values. Additionally, after inspection of the unique categorical values of WEATHER, ROADCOND and LIGHTCOND, there appears to be the entries Unknown and Other. We decide to categorise them as missing values as well. This results in 13% of the rows containing missing values. We will fill the missing data with the most common entry for each attribute in the data preparation phase and will investigate if discarding the rows instead improves the model at a later stage.

Data Preparation

Before we ingest the data frame into the different models, we perform the following operations:

- Change the types of all the attributes from Object to:
    - datetime for the Time/Date attributes
    - category for the Environmental attributes
- Change INCDTTM to the hour of the day (0-23)
- Change INCDATE to the day of the week (0-6)
- Replace Unknown/Other entries by n/a’s
- Fill n/a’s with most frequent values

In order to end up with numerical variables, we now encode the environmental values. For each, we decide a value for each category according to its potential detrimental effect on driver experience. The encoding choice can be seen in Figure 3. The resulting data frame can be seen in Figure 4.

Figure 3

```
weather_dic ={"Clear":0,
              "Partly Cloudy":0.5,
              "Overcast":1,
              "Blowing Sand/Dirt":1.5,
              "Severe Crosswind":2,
              "Raining":2,
              "Sleet/Hail/Freezing Rain":2.5,
              "Snowing":3,
              "Fog/Smog/Smoke":3,
             }
```

```
road_dic ={"Dry":0,
           "Wet":1,
           "Sand/Mud/Dirt":1,
           "Snow/Slush":1.5,
           "Oil":1.5,
           "Standing Water":2,
           "Ice":3,
           }
```

```
light_dic ={"Daylight":0,
           "Dark - Street Lights On":1.5,
           "Dawn":1,
           "Dusk":1,
           "Dark - Unknown Lighting":2,
           "Dark - Street Lights Off":3,
           "Dark - No Street Lights":3,
           }
```

Figure 4

	SEVERITYCODE	INCDATE	INCDTTM	WEATHER	ROADCOND	LIGHTCOND
0	1	2	14	1	1	0
1	0	2	18	2	1	1
2	0	3	10	1	0	0
3	0	4	9	0	0	0
4	1	2	8	2	1	0

Before we move to the modelling phase, it is worth noting that the data set is quite imbalanced. In fact, 30% of the labelled entries are Severe accidents – a stretch from a 50/50 dataset. We will use the balancing features of the upcoming algorithms to account for this.

Modelling

As this is a classification problem, we will compare the performance of some common Machine learning algorithms: Logistic Regression (LR), KNN, and Decision Tree (DT). We train-test split our data, leaving 10% for testing. We also add a Dummy Classifier which will return the most frequent class.

The parameter choices are as follows:

- KNN: neighbours = 3
- LR: class-weight = balanced
- DT: max depth = 5, class-weight = balanced

Results

To evaluate the performance of the models, we use accuracy, the F1-score and the Jaccard-score. Figure 5 presents the scores. We can see that all the models perform worse than the dummy classifier when it comes to accuracy, but all others beat it in terms of F1 and Jaccard score.

Figure 5

	Accuracy	F1	Jaccard
Dummy Classifier	0.706030	0.000000	0.000000
Nearest Neighbors	0.618708	0.273039	0.158104
Logistic Regression	0.528354	0.395444	0.246451
Decision Tree	0.496302	0.436566	0.279236

The scores obtained for all three models are low. This indicates that further iterations are needed to improve the performance and conclude if our problem can be resolved with the current data set.

Alternative A: Removing rows that contain n/a, Unknown or Other data instead of filling them with the most frequent category worsens the performance.

	Accuracy	F1	Jaccard
Dummy Classifier	0.670628	0.000000	0.000000
Nearest Neighbors	0.599965	0.270572	0.156452
Logistic Regression	0.471287	0.438445	0.280775
Decision Tree	0.549012	0.386063	0.239206

Alternative B: One-hot encoding features (n/a's still removed). Once again, this does not dramatically improve model performance.

	Accuracy	F1	Jaccard
Dummy Classifier	0.670628	0.000000	0.000000
Nearest Neighbors	0.582784	0.290687	0.170061
Logistic Regression	0.520181	0.423716	0.268807
Decision Tree	0.446517	0.465845	0.303649

Discussion

At this stage it is worth asking whether the scores can be improved with the features selected. Alternatives A and B did not show improvements. We should go back to the original problem and reshape it, in a perhaps less ambitious way, using a wider set of features.

Conclusion

Our original question, “What if we could predict the severity of an accident, if one were to occur, based on time, date and environmental data, and thereby provide information that could reduce their frequency?” was not answered in this study.

We were however able to identify that taking date, time and environmental data available in the Seattle dataset was insufficient to build a reliable model.