Prediction Collisions Severity

Hubert Aliaga

August 27, 2020

Contents

1 Business Understanding 3
1.1 Background . . . 3
1.2 Problem . . . 3
1.3 Interest . . . 3

2 Data Understanding 4
2.1 Data sources . . . 4

3 Data Preparation 5
3.1 Data cleaning . . . 5
3.2 Feature selection . . . 5
3.3 Calculation of target variable . . . 7
3.4 Relationship between the collision and the days of the week . . . 7
3.5 Relationship between the collision and the hours of a day . . . 8

4 Modeling 10
4.1 Regression Model . . . 10
4.2 Classification Models . . . 10

5 Evaluation 11

6 Conclusions 12

Chapter 1

Business Understanding

1.1 Background

Say you are driving to another city for work or to visit some friends. It is rainy and windy, and on the way, you come across a terrible traffic jam on the other side of the highway.
Long lines of cars barely moving. As you keep driving, police car start appearing from afar shutting down the highway.
Oh, it is an accident and there's a helicopter transporting the ones involved in the crash to the nearest hospital. They must be in critical condition for all of this to be happening.
Now, wouldn't it be great if there is something in place that could warn you, given the weather and the road conditions about the possibility of you getting into a car accident and how severe it would be, so that you would drive more carefully or even change your travel if you are able to.

1.2 Problem

Data that might contribute to determining Collisions and his severity might include Collision Address type,Location, Collision type, Road Condition, wheather and number of objects or people involved in the collision. This project aims to predict severity type of any accidents.

1.3 Interest

Clearly, those who work long distances or frequent certain places are interested in finding routes where they are not delayed or even postponed due to sudden accidents.
In addition, the state may also be interested to be on the lookout for any collisions or accidents at certain locations.

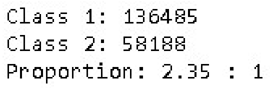
Chapter 2

Data Understanding

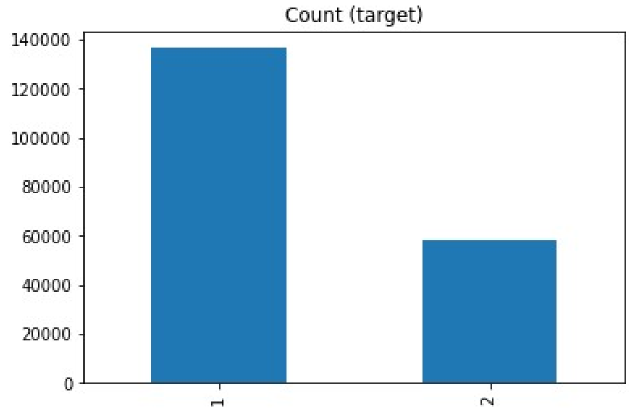
2.1 Data sources

Let’s use our shared data for Seattle city as an example of how to deal with the accidents data. The label (target) for the data set is SEVERITY, which describes the fatality of an accident.

You will notice that the shared data has unbalanced labels (approximately 70% Class 1 and 30% Class 2).
So, balance data with Random Over-Sampling (might be Random Under-Sampling).



```
Class 1: 136485
Class 2: 58188
Proportion: 2.35 : 1
```



Chapter 3

Data Preparation

3.1 Data cleaning

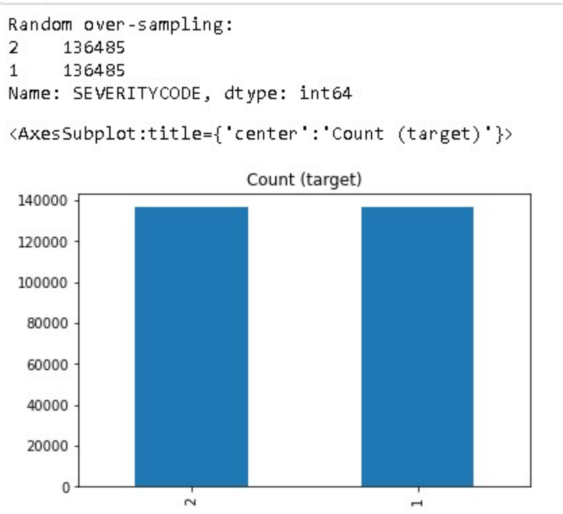
For the cleaning of the data it was structured as follows:
First, we look for missing values in the quantitative variables, that is where we find the *INTKEY* variable which is a variable of type int and it only gives us the index when an accident was at an intersection. Which is not necessary for our model.

Second, we focus on redundant information, for example we have the *REPORTNO* variable which is the report number of the accident or collision. There are also variables such as *SEVERITYDESC* and *SDOTCOLDESC* which already exist in their number formats in the data and it only gives us a brief description of the variables and that is why it is not considered. In the Notebook there is a detailed list of why each variable was not considered one by one.

3.2 Feature selection

To assemble the features we first have to balance our label (objective) because the data as it is at 70% to 30% as shown in the previous figure.

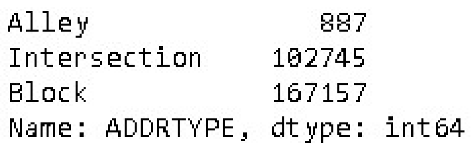
To do this we have two methods, Over-Sampling and Under-Sampling. As a convenience, Over-Sampling was used and the data was obtained as follows.



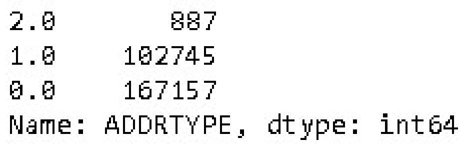
```
Random over-sampling:
2    136485
1    136485
Name: SEVERITYCODE, dtype: int64

<AxesSubplot:title={'center':'Count (target)'}>
```

After balancing the data, now we have to recognize which categorical variables to use and also know which ones have relevant information and for this we will use two methods to quantify these categorical variables:

1. Binarization, for example the variable ADDRTYPE has as labels Block, Intersection and Alley, so we change these variables to 0,1 and 2 respectively.

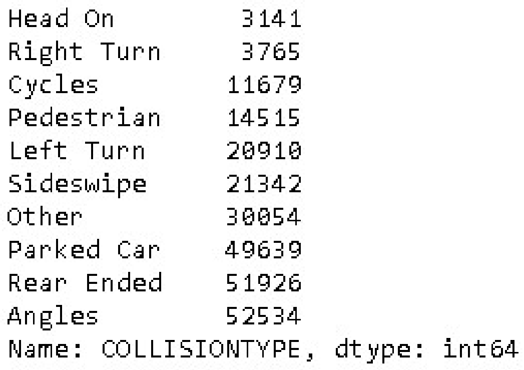


```
Alley              887
Intersection    102745
Block           167157
Name: ADDRTYPE, dtype: int64
```



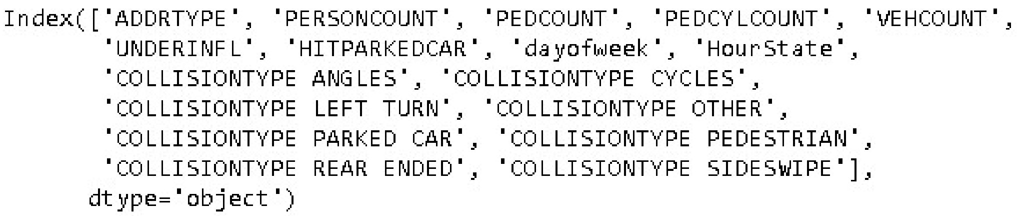
```
2.0       887
1.0    102745
0.0    167157
Name: ADDRTYPE, dtype: int64
```

2. One hot encoding, this method is about converting the information in a column to columns as many times as only variables appear and filling the new columns with 1 or 0. For example, the COLLISIONTYPE variable has 10 values as shown in the following figure:



```
Head On         3141
Right Turn      3765
Cycles         11679
Pedestrian     14515
Left Turn      20910
Sideswipe      21342
Other          30054
Parked Car     49639
Rear Ended     51926
Angles         52534
Name: COLLISIONTYPE, dtype: int64
```

Then we see that the types of collisions Head On and Right Turn is unsignificative information, so we are going to do one hot encoding without considering these labels.



```
Index(['ADDRTYPE', 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT',
       'UNDERINFL', 'HITPARKEDCAR', 'dayofweek', 'HourState',
       'COLLISIONTYPE ANGLES', 'COLLISIONTYPE CYCLES',
       'COLLISIONTYPE LEFT TURN', 'COLLISIONTYPE OTHER',
       'COLLISIONTYPE PARKED CAR', 'COLLISIONTYPE PEDESTRIAN',
       'COLLISIONTYPE REAR ENDED', 'COLLISIONTYPE SIDESWIPE'],
      dtype='object')
```

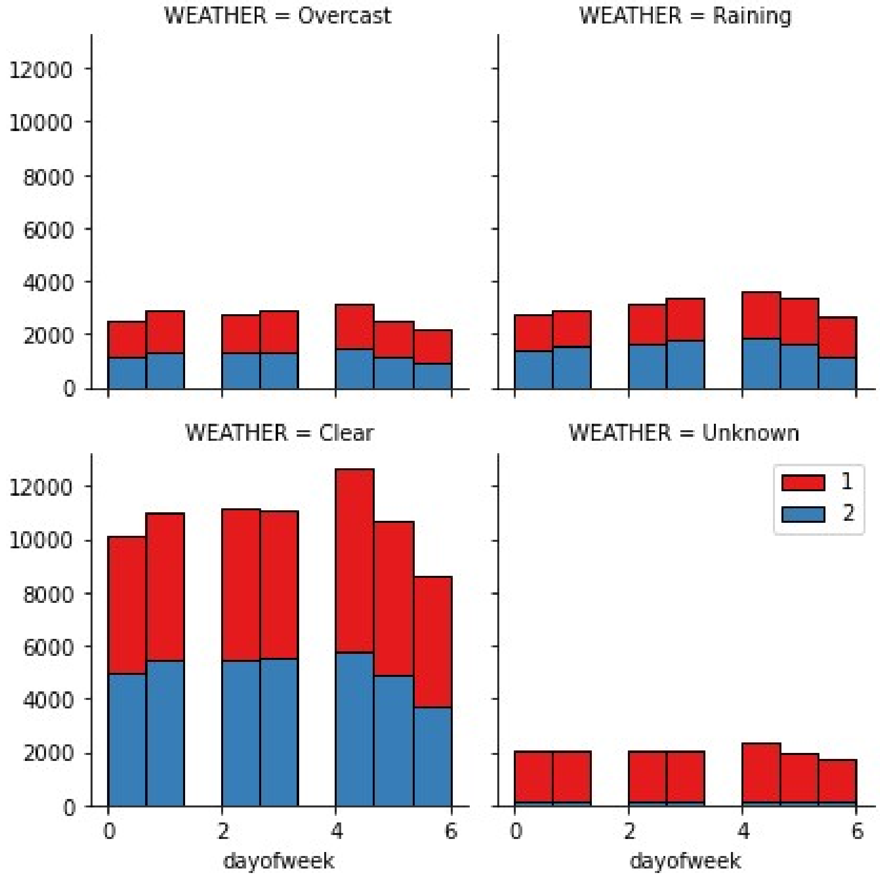
Then we see that the information of all the COLLISIONTYPE labels appears except Head On and Right Turn.

3.3 Calculation of target variable

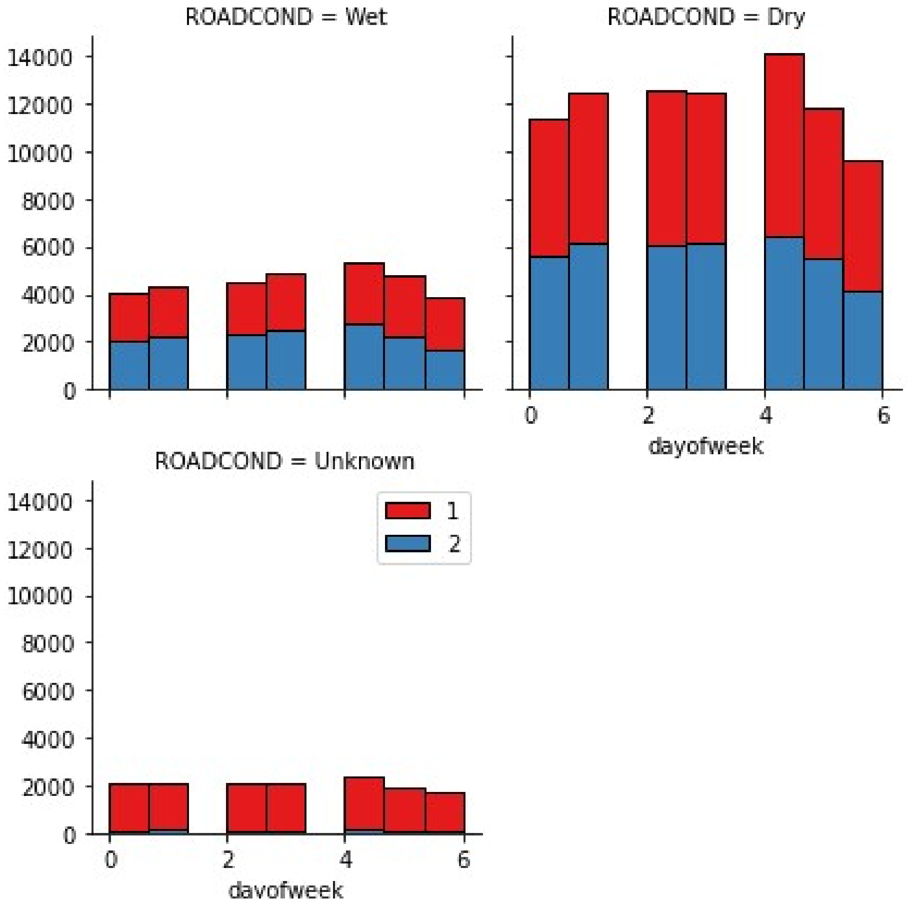
The target variable is the SEVERITYCODE column and it is a categorical binary variable, so this leads us to perform an exploratory analysis regarding this variable.

By having both the weather variable and the road conditions in a collision, we can use this to compare it with the time variables, as they are at the level of days or hours.

3.4 Relationship between the collision and the days of the week



As we can see, we do not have a particular day a week where we can say the type of fatality of a collision. Even though the graph shows us different climates.

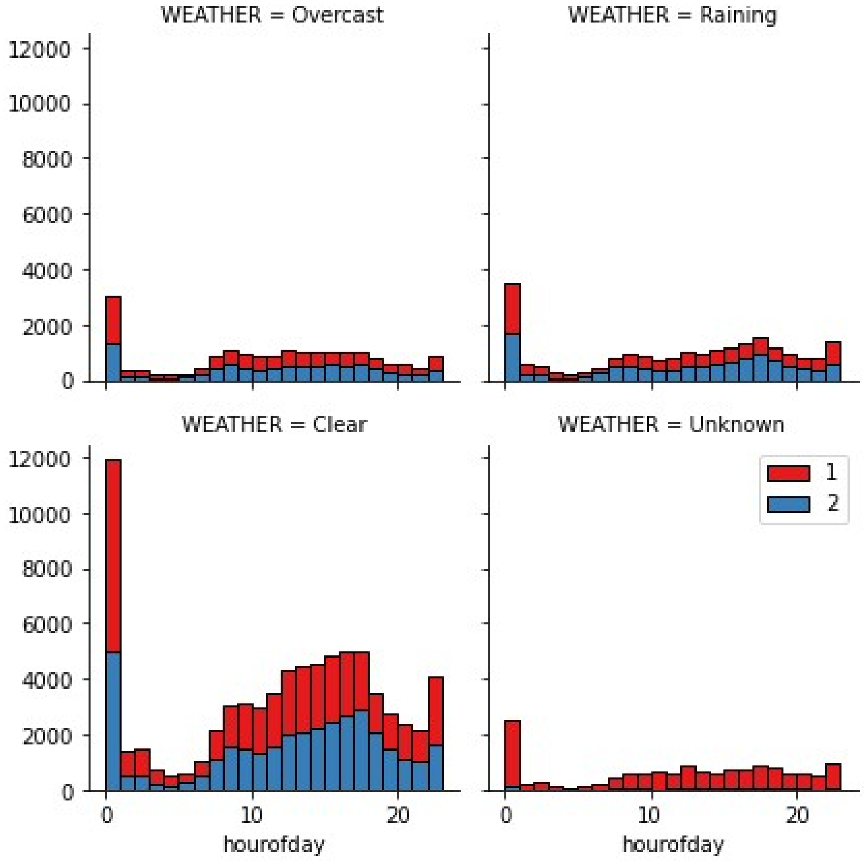


Likewise, considering now the variable of road conditions, we do not see anything in particular that tells us what day or days there is a type of fatality or severity of a collision.

The following comparison is something that will take us further.

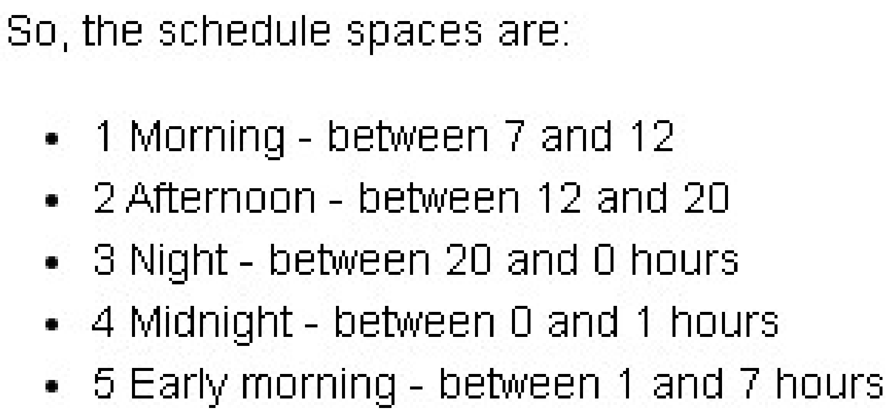
3.5 Relationship between the collision and the hours of a day

We will use the comparison between climates, as road conditions will show familiar behavior.



As we clearly see how it is divided into groups where more of one type of collision is concentrated than another. One visible case is collisions that occurred at midnight.

So in order not to have 24 columns with one hot encoding that would be 24 hours a day, we are going to separate them into 5 large groups, as follows:



So, the schedule spaces are:

- 1 Morning - between 7 and 12
- 2 Afternoon - between 12 and 20
- 3 Night - between 20 and 0 hours
- 4 Midnight - between 0 and 1 hours
- 5 Early morning - between 1 and 7 hours

Therefore, we are going to create the variable HOURSTATE that will have information from 1 to 5, which will represent to which time group a collision occurred with its respective fatality or severity.

Chapter 4

Modeling

4.1 Regression Model

When we talk about the severity of a collision, we cannot say. A certain collision was very serious and a half, or less serious and a quarter. To answer this we must say on a certain scale, precisely this problem is given as a label.

1. Prop damage
2. Injury

Therefore, our model must be trained to be able to classify more optimally what type of collision severity will occur and with the location data to be able to know exactly if this will lead to a change of route in a certain route.

To solve this, we will concentrate on the Classification models.

4.2 Classification Models

To make this model we will do it in several traditional steps:

1. Verify that both the characteristics and the target variable have the same dimensions.
2. Normalize the data
3. Separate training and validation set
4. Train the model with the training data

To carry out the mentioned steps, I will concentrate on the Logistic Regression model, the other models, KNN, SVM and Decission Tree will be left as optional in the Notebook code.

Chapter 5

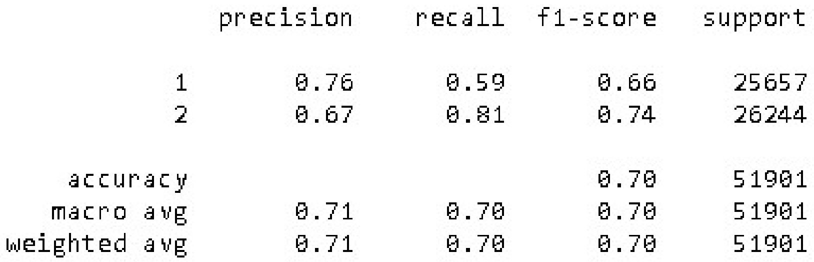
Evaluation

To evaluate the Logistic Regression model, we will focus on three specific measures:

1. Jaccard Score

2. F1 Score

3. Log Loss

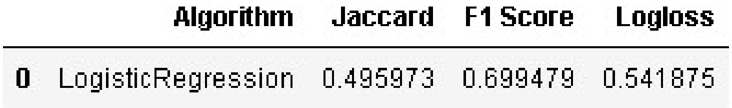


```
              precision    recall  f1-score   support

           1       0.76      0.59      0.66     25657
           2       0.67      0.81      0.74     26244

    accuracy                           0.70     51901
   macro avg       0.71      0.70      0.70     51901
weighted avg       0.71      0.70      0.70     51901
```

Chapter 6

Conclusions

To finish let's show the results table:



	Algorithm	Jaccard	F1 Score	Logloss
0	LogisticRegression	0.495973	0.699479	0.541875

As we see our model works well, not excellent but well. Since if we had considered some insignificant information or redundant columns, we could get overfitting. Which if we should review more carefully.

We can say that when they provide us with information about a weather, the type of street or knowing the location of the accident. We are 70 % confident in saying what kind of collision it is and thus taking other alternatives if necessary. Remember that an important variable is what time this collision occurs, as we saw that it influences a lot.