Predicting the accidents on road due to bad weather

and

bad road conditions

Saurabh Parmar

September 07, 2020

1. Introduction

1.1 Background

While travelling from one city to another city by road, sometimes people face terrible traffic jam on road due to accident happens due to bad weather condition or bad road condition. And due to that lots of people suffer in traffic jam and needs to wait till the traffic jam gets clear.

If people can predict the accident on road before start travelling, based on bad weather (windy, rainy, etc) or bad road condition, people can drive more carefully or change the travel plan or change the travel route.

1.2 Problem

People get stuck longer hours in traffic jam and due to that people face below problems:

- Can't reach at destination place on right time, due to that they may miss some task to finish. e.g.: fail to attend interview on time, miss the doctor consultation, miss the business meetings, etc.
- Lots of fuel gets waste due to slow vehicle movement.
- People gets irritate while waiting in long queue, due to that they get tired physical and mentally.
- Sometimes, people may feel physical problems (due to hunger, stopping urination, etc).
- Patients may face severe health problem, if gets stuck for longer hours.

1.3 Interest

This model can be useful for the people who travels from one city to another city, in bad weather condition or in bad road condition.

2. Data acquisition and cleaning

2.1 Data sources

Govt of Seattle has created the dataset and it is available here. Which contains all types of collisions. Timeframe from 2004 to Present. Please find the details of each columns here.
This dataset contains around 38 attributes. In which severity code is a target variable. And other 37 attributes we can use as input features (independent variables).
It has features like address type, location, severity description, pedestrian count, vehicles count, injuries, date, time, fatalities, weather, road condition, light condition, speeding and many more.

2.2 Data cleaning

There are total 136485 samples available into this dataset. In which most of the samples having Nan values. And out of 37 features, most of the features are not having co-relation with severity of accident. Few attributes are having numerical values and some attributes are having categorial strings.

2.3 Feature Selection

SEVERITYCODE and SEVERITYCODE.1 both are having same data. So, only considered SEVERITYCode as target variable.

Based on the correlation matric, it's been found that there is no relationship between SEVERITYCODE and X, Y, OBJECTID, INCKEY, COLDETKEY, INTKEY, SDOTCOLNUM. So, better to exclude these attributes.

There is very less relationship between SEVERITYCODE and PERSONCOUNT, SEGLANEKEY. So, removed it as well.

As there is some categorial features are also available in dataset, and it doesn't have any correlation with severity code, so removed it as well. Such as status, ADDRTYPE, ST_COLDESC, etc.

Initially checked the correlation of all the numeric features and selected the features which is having more relationship with severity code. After that looked more deeply in each of the features, converted important categorial features into numeric values, such as weather and road condition.

Weather features having categories like Raining, clear, overcast, snowing, Fog/smog/smoke, Hail/Sleet, severe crosswind, partly cloudy.

Road condition feature is having categories like wet, Dry, snow/slush, Ice, Standing water, oil, sand/Mud/Dirt.

	SEVERITYCODE	WEATHER	ROADCOND
0	2	Overcast	Wet
1	1	Raining	Wet
2	1	Overcast	Dry
3	1	Clear	Dry
4	2	Raining	Wet

The aim is to create model to predict the accident due to weather and road condition, Weather condition and Road condition are the best features to select for model creation.

In weather and roadcond features, there are most of the samples are unwanted and can decrease the accuracy of the model, such as "Clear", "Unknown", "Others", etc. So, removed such samples from these two features to get better accuracy of model.

3. Exploratory Data Analysis

3.1 Calculation of target variable

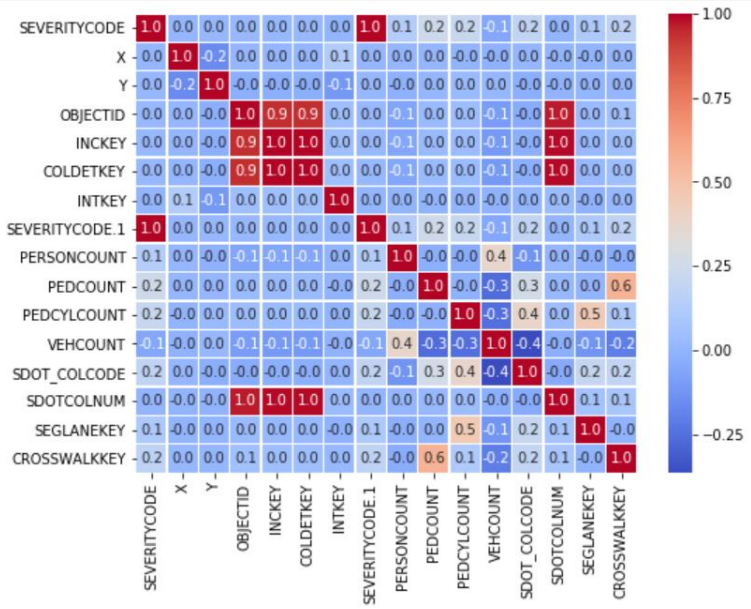
Here target variable is a severity code. It has only two types of severity code available with this dataset. It's 1 and 2. There are total 1,36,485 samples available of severity 1, and 58188 samples available of severity 2.

```
df['SEVERITYCODE'].value_counts()
```

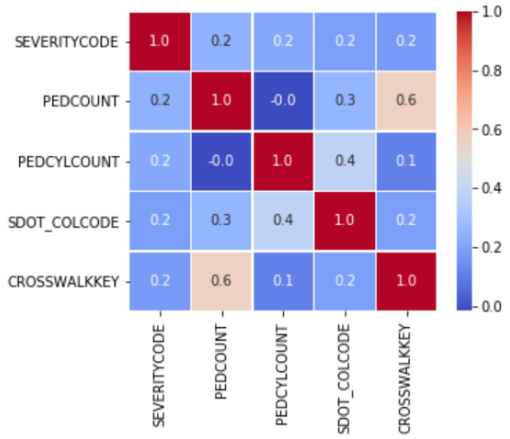
```
]: 1    136485
   2     58188
```

3.2 Correlation between numeric features

Based on the correlation matrix of all the numeric shown in below heatmap, it seems that there are few



Direct relations are there between PEDCOUNT and CROSSWALKKEY, OBJECTID and INCKEY, OBJECTID and COLDETKEY, SEGLANEKEY and PEDCYLCOUNT, etc.



But it seems that these all the features are not having direct relationship with severity code by which we can predict the accident on road.

4. Predictive Modeling

4.1 Classification models

Here, values of severity code are only two (1 and 2), so regression algorithm is not suitable to train this model. For this kind of feature and target variable, classification algorithms are best suitable.
Used below algorithms to train the model using training data. And then evaluate all four models using test data.

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

4.2 Performances of different models

Based on the four algorithms, it seems that F1 score of SVM and Logistic regression is more than KNN and Decision Tree algorithm. And, Jaccard score is similar for all the algorithms. It would be good to choose the SVM or Logistic Regression model for better prediction of target value. Log loss(Logarithmic loss) measures the performance of a classifier where the predicted output is a probability value between 0 and 1. It's value is 0.63 in Logistic Regression model.

	Algorithm	Jaccard	F1-score	LogLoss
0	KNN	0.67679	0.546484	NA
1	Desicion Tree	0.67679	0.546484	NA
2	SVM	0.67679	0.807245	NA
3	LogisticRegression	0.67679	0.807245	0.627976

5. Conclusions

Purpose of this project was to predict the accident and traffic jam because of bad weather or bad road condition. To create this model, generic approach is applied where prediction of accident is found based on current weather condition and bad road condition features. There are other 37 features were available in dataset, but those are not relevant to predict accident on road and not having direct relationship with severity code. So, after checking the correlation between severity code and other features, it observed that there was some relation among independent features, but not having direct relation with target label. So, by going deep and understanding each feature and based on correlation & heatmap method, excluded number of features from creating final model.

As, target values are not continuous, used classification algorithm to predict the accident. First divided dataset into 80-20 ratio. then, used four classification algorithms and trained models one by one using training dataset comprises of 80% of total final dataset. and evaluate it on remaining 20% testing dataset.

Based on the result it seems that SVM and logistic regression algorithms gives better accuracy compared to KNN and Decision tree algorithms. As this model is generic, location is not included to predict the bad weather and road condition between two locations/cities.

6. Future directions

For future, we can get the weather and road condition details between source and destination places/cities and considering that location information to predict the severity and probability of traffic jam on road to give more accurate result.