A report on data analysis of car collisions

By

Sourja Paul

Business Problem

With increase in the number of cars on road there has been an increase in traffic which has led to increase in road accidents. Mishaps due to car accident are a major concern for both the law enforcers and the riders behind the wheels.

As data analyst we can address this problem by designing a predictive model to be used by traffic regulators and common people to understand what conditions are the major reasons for the road accidents.

To solve this business problem we need to have previous data set about car collisions having specification on various attributes like the road condition, the weather and the car condition. This will help us in determining the key factors leading to these accidents. The data will help us in designing a predictive model that will help in providing recommendations on the basis of prevailing factors and minimize the quantity of car collisions.

Data collection and wrangling

Seattle Police Department (SPD) has been collecting detailed data about cars collisions. It will be used for developing model that allows to determine locations, weather conditions, days of week, time of day and other factors that are helpful for car collisions prediction.

The dataset is available at https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

File with metadata is available athttps://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

The data has 38 variables and 194673 records. The dependent/target variable for this dataset is SEVERITYCODE. It has only two values and our analysis and model design will revolve around this variable only.

SEVERITYCODE	SEVERITYDESC
1	Property Damage Only Collision
2	Injury Collision

From the list of remaining 37 variables, 8 variables are chosen for independent variables.

Variable Name	Variable Description
ADDRTYPE	Address type
COLLISIONTYPE	Type of collision
UNDERINFL	Whether the driver was under the influence of alcohol
PEDCOUNT	Number of pedestrians injured
INCDATE	Date of collision

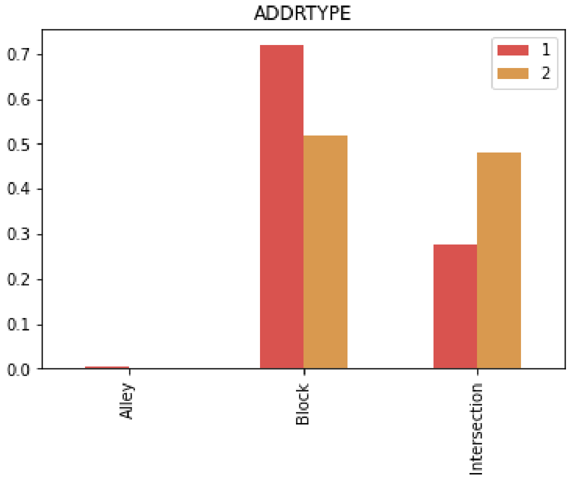
WEATHER	Weather during collision
ROADCOND	Road condition during collision
LIGHTCOND	Light condition during collision

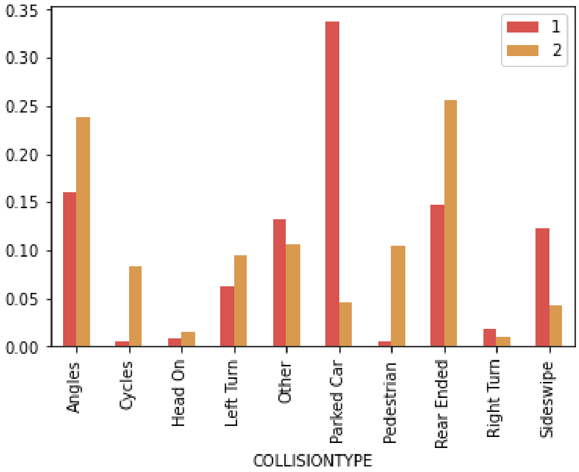
We will do data visualization with these variables to find out how much impact they have on SEVERITYCODE and then design our predictive model based on the selected variables among this.

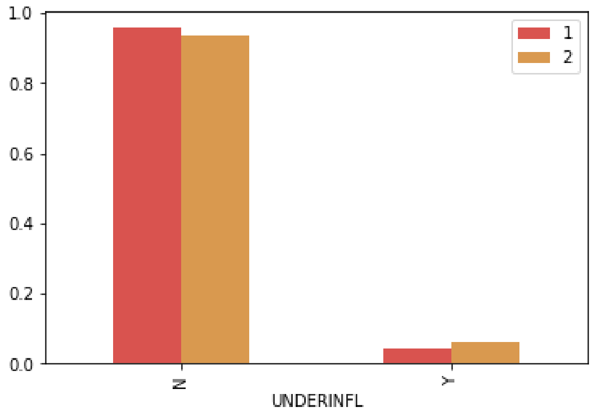
For data wrangling we first filter out the missing data and delete those data. The revised dataset have 187504 records. The data for UNDERINFL was amended to make it in sync with either N for No or Y for Yes. From INCDATE we filtered out the days and clubbed Monday-Thursday as WORK (workday) and Friday-Sunday as WEND(weekend).

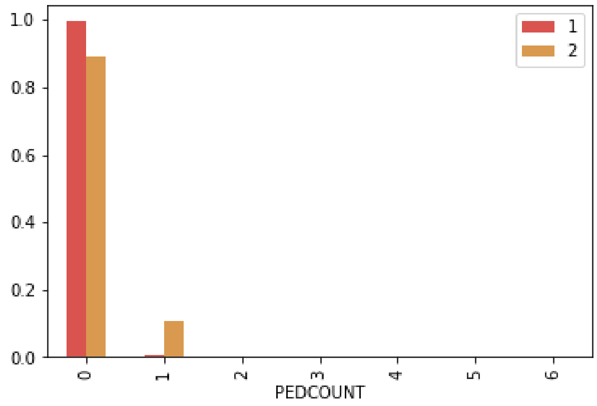
Data Visualization

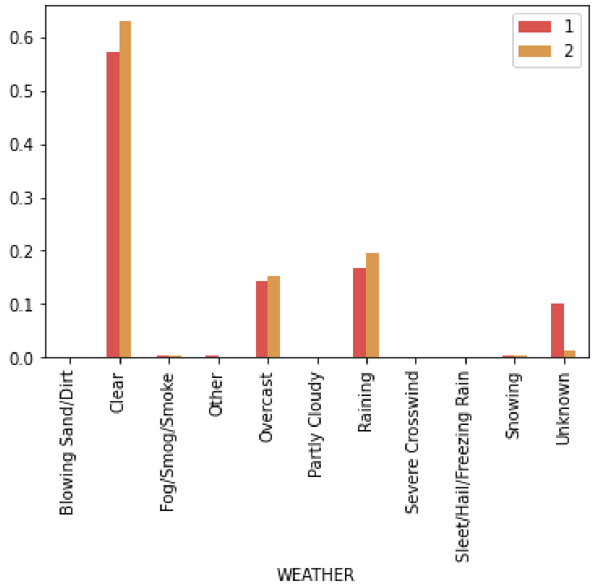
Bar charts for normalized independent variables were plotted to identify those variables that varies for different SEVERITYCODE.

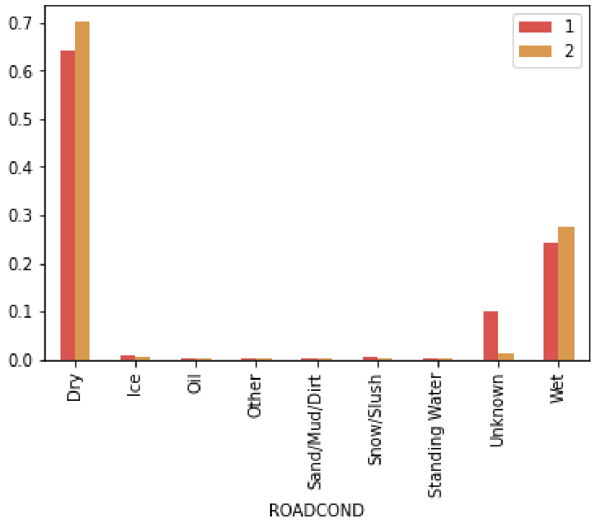


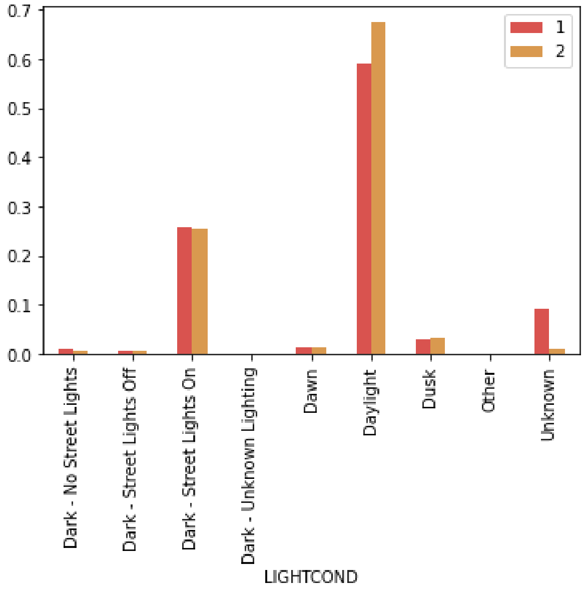


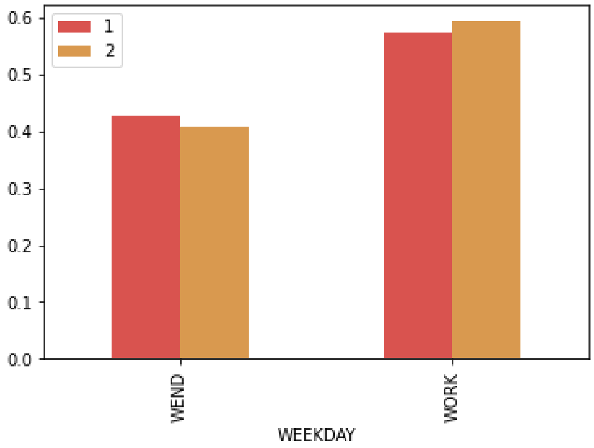

1.0
0.8
0.6
0.4
0.2
0.0
1
2
N
Y
UNDERINFL


1.0
0.8
0.6
0.4
0.2
0.0
1
2
0
1
2
3
4
5
6
PEDCOUNT


0.6
0.5
0.4
0.3
0.2
0.1
0.0
1
2
Blowing Sand/Dirt
Clear
Fog/Smog/Smoke
Other
Overcast
Partly Cloudy
Raining
Severe Crosswind
Sleet/Hail/Freezing Rain
Snowing
Unknown
WEATHER


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1
2
Dry
Ice
Oil
Other
Sand/Mud/Dirt
Snow/Slush
Standing Water
Unknown
Wet
ROADCOND

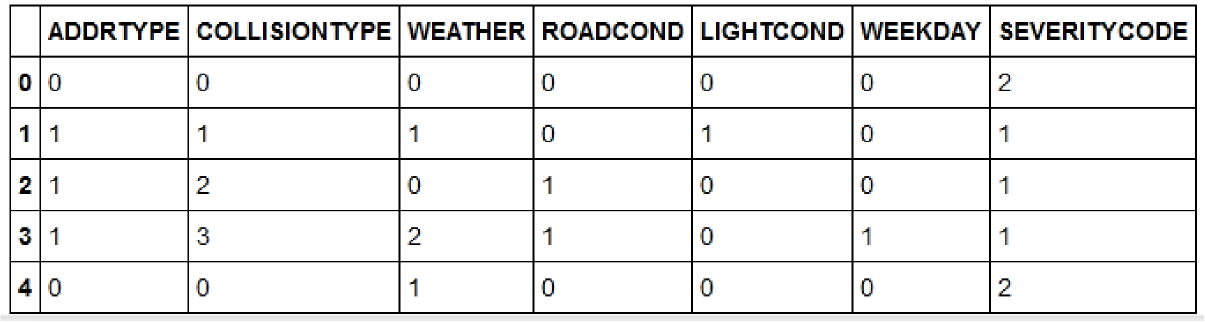

0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1
2
Dark - No Street Lights
Dark - Street Lights Off
Dark - Street Lights On
Dark - Unknown Lighting
Dawn
Daylight
Dusk
Other
Unknown
LIGHTCOND


0.6
0.5
0.4
0.3
0.2
0.1
0.0
1
2
WEND
WORK
WEEKDAY

Predictive Modelling

From the above bar charts we can deduce that UNDERINFL and PEDCOUNT doesn’t have much impact on SEVERITYCODE, a little less than 100% were not UNDERINFL and 0 number of Pedestrians injured. We select the other remaining six variables for our predictive modelling but first we factorize our data for easy modelling.

A snapshot of the factorised data

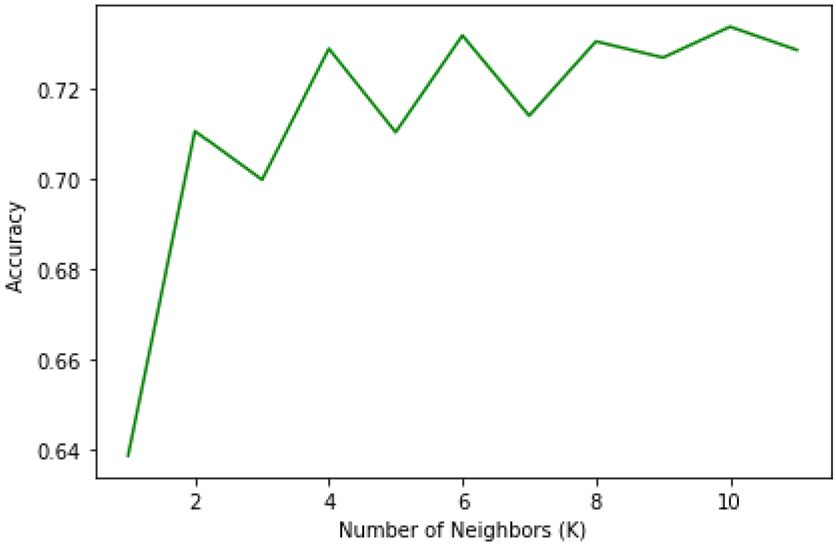


	ADDRTYPE	COLLISIONTYPE	WEATHER	ROADCOND	LIGHTCOND	WEEKDAY	SEVERITYCODE
0	0	0	0	0	0	0	2
1	1	1	1	0	1	0	1
2	1	2	0	1	0	0	1
3	1	3	2	1	0	1	1
4	0	0	1	0	0	0	2

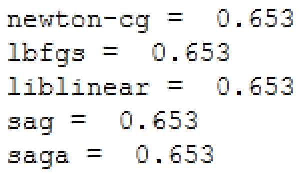
The objective of the model is to predict the severity code of the collision based on the available data. Two classification model were used and the one with better accuracy will be selected for this dataset.

Dataset was divided into two sets, training (80%) and test (20%).

KNN Algorithm was run first and we found k =10 has the best accuracy.

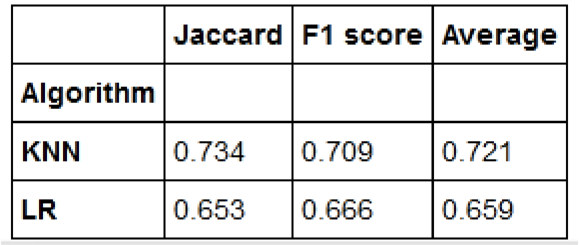


Then we implemented Logistic Regression.



```
newton-cg =  0.653
lbfgs =  0.653
liblinear =  0.653
sag =  0.653
saga =  0.653
```

We compared the Jaccard score , F1 score and the average of the two values.



	Jaccard	F1 score	Average
Algorithm			
KNN	0.734	0.709	0.721
LR	0.653	0.666	0.659

Based on the average value KNN algorithm model will be our best choice.

Discussion

The KNN algorithm thus confirms that the independent variables ADDRTYPE, COLLISIONTYPE, INCDATE, WEATHER, ROADCOND and LIGHTCOND does have an uimpact on the severity of car collision. Most collisions take place at intersection and at block. Parked cars are the major type of collisions and majority collisions happened on work days during the day. So a better traffic management plans on car parked at roads near intersection and blocks with more focus during office hours when the volume of cars on road is at large might lead to decrease of car collision.

Conclusion

In this study the emphasis was on to identify the variables that impact car collision and filter out the unwanted data. The variables were then put to a data visualization test from where we selected the variables really impacted the severity based on bar graphs. The data was factorized and put under two classification algorithm KNN and Logistic Regression and based on the average score KNN was selected.