Data analysis of car collisions

By Sourja Paul

Introduction

- With increase in the number of cars on road there has been an increase in traffic which has led to increase in car collisions
- 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 car collisions
- To solve this business problem we need to have previous data set about car collisions having specification on various attributes
- This will help us in determining the key factors leading to these collisions

Data Collection

❑ Seattle Police Department (SPD) has been collecting detailed data about cars collisions

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

❑ The data has 38 variables and 194673 records

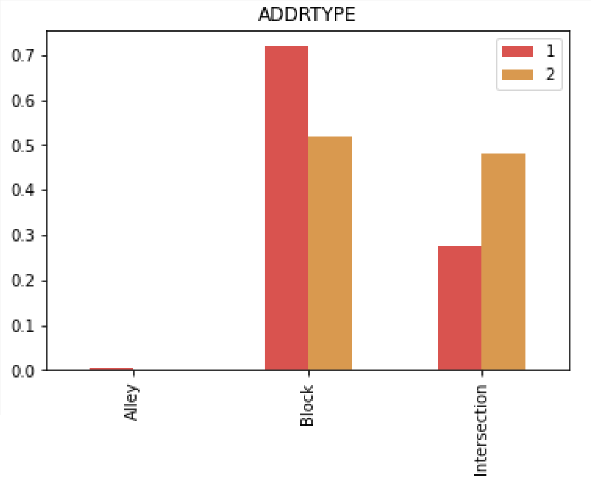
❑ The dependent/target variable for this dataset is SEVERITYCODE

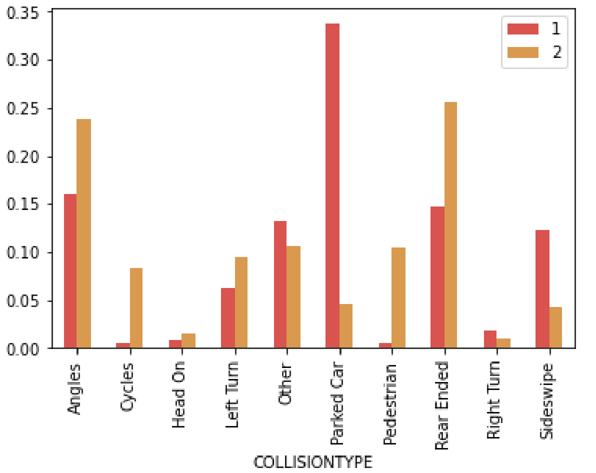
❑ From the list of remaining 37 variables, 8 variables are chosen for independent variables

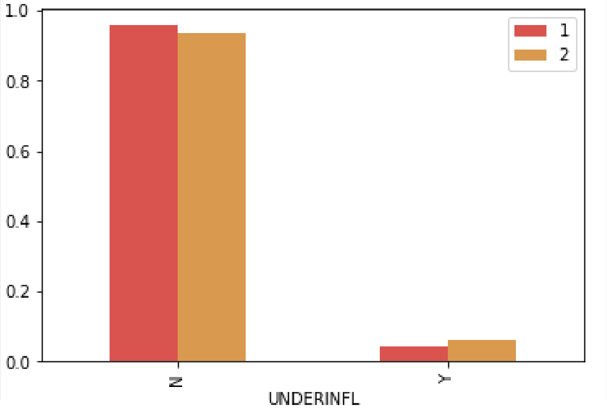
SEVERITYCODE	SEVERITYDESC
1	Property Damage Only Collision
2	Injury Collision

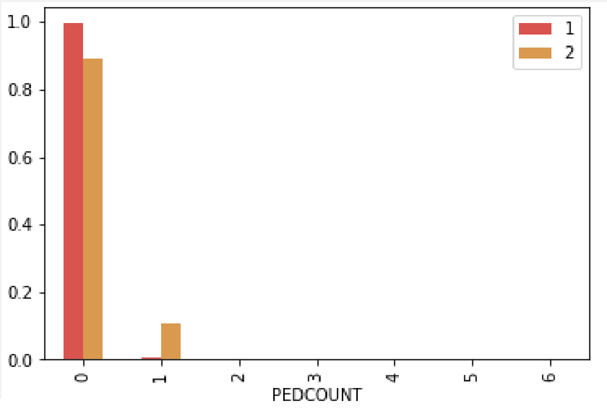
Variable Name	Variable Description
ADDRTYPE	Address type
COLLISIONTYPE	Type of collision
UNDERINFL	Whether the driver was under the influence of alcohol
PEDCOUNT	Number of pedestrians injured
INCDATE	Date of collision
WEATHER	Weather during collision
ROADCOND	Road condition during collision
LIGHTCOND	Light condition during collision

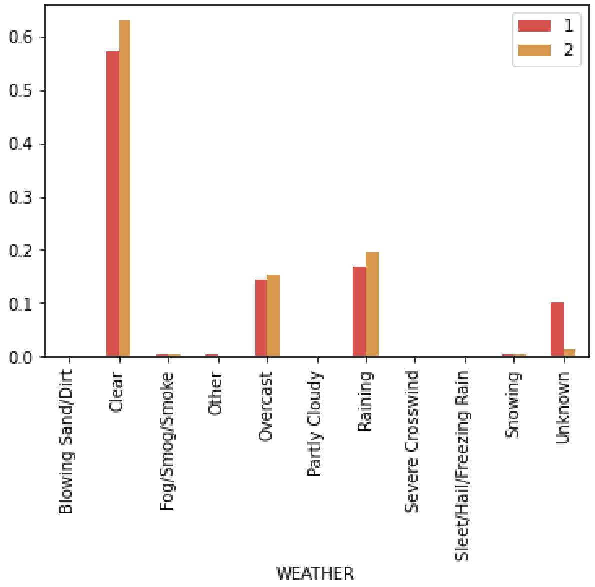
Data Visualization

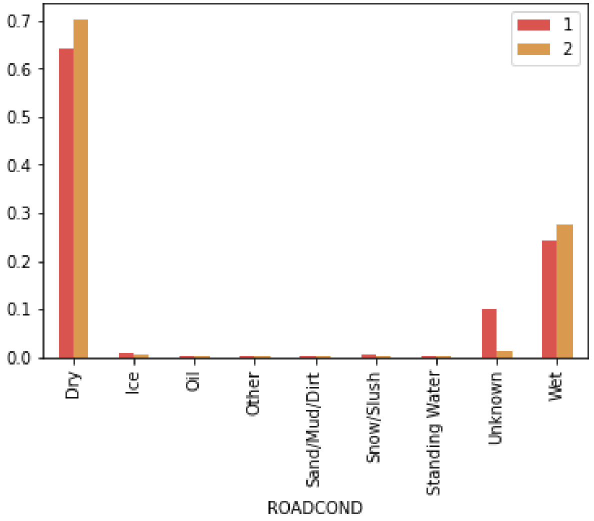


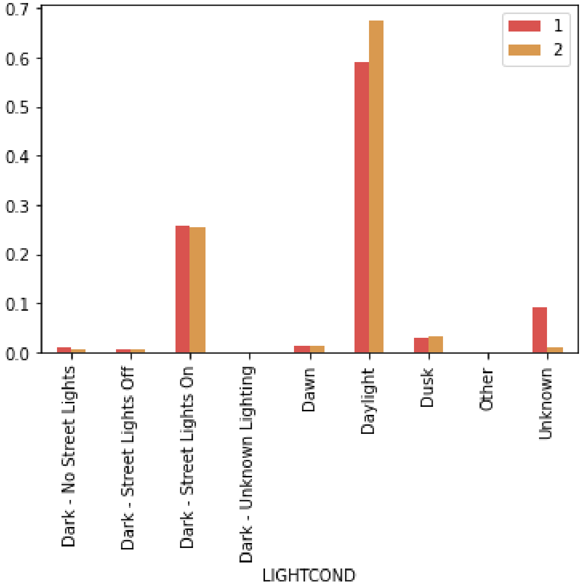


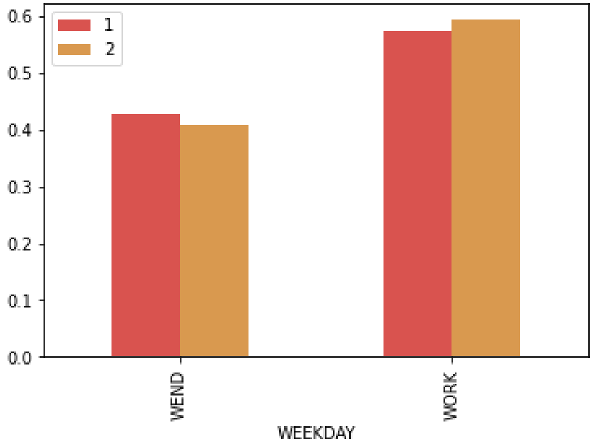








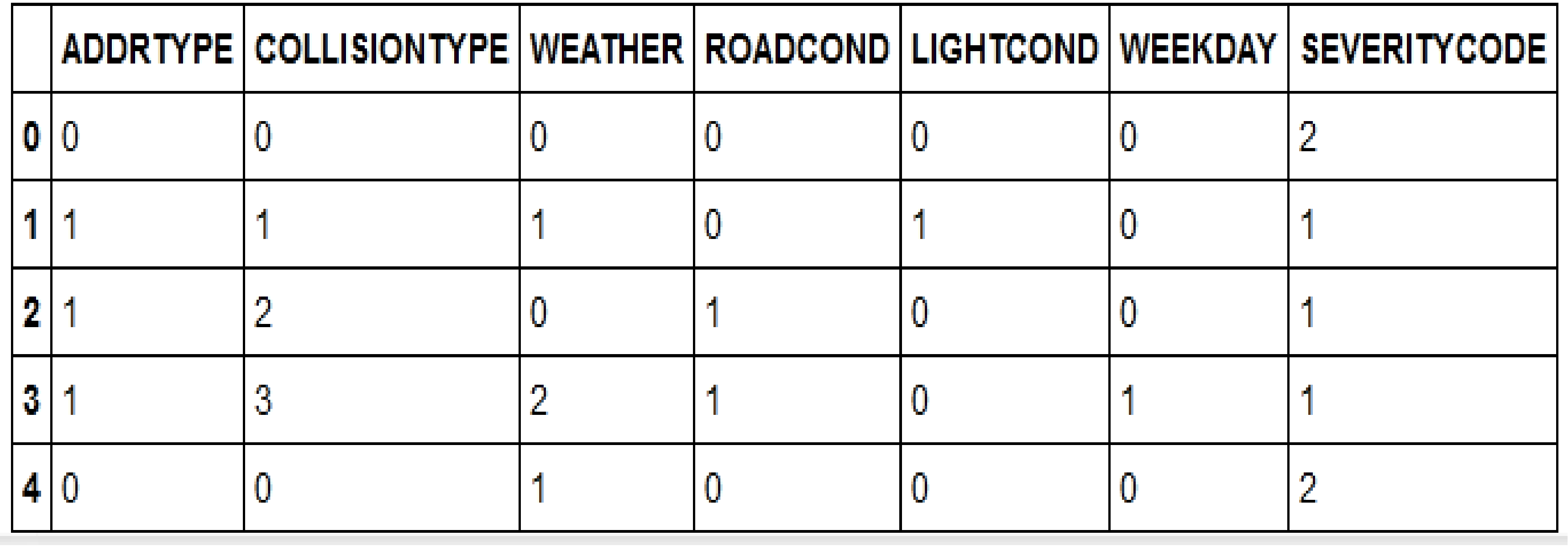




Predictive Modelling

❑ From the bar charts we can deduce that UNDERINFL and PEDCOUNT doesn’t have much impact on SEVERITYCODE

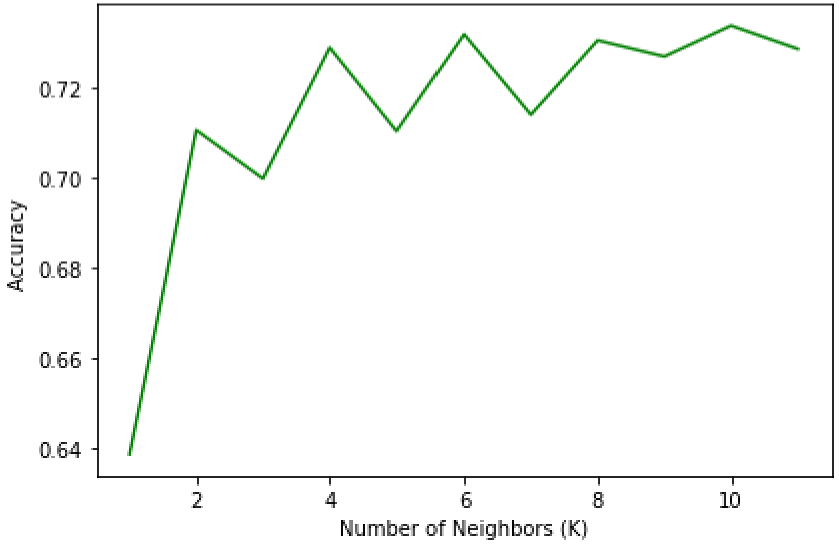
❑ We select the other remaining six variables for our predictive modelling but first we factorize our data for easy modelling



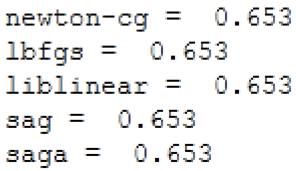
	ADDRTYPE	COLLISIONTYPE	WEATHER	ROADCOND	LIGHTCOND	WEEKDAY	SEVERITYCODE
0	0	0	0	0	0	0	2
1	1	1	1	0	1	0	1
2	1	2	0	1	0	0	1
3	1	3	2	1	0	1	1
4	0	0	1	0	0	0	2

Predictive Modelling

❑ The objective of the model is to predict the severity code of the collision based on the available data
❑ Two classification model were used

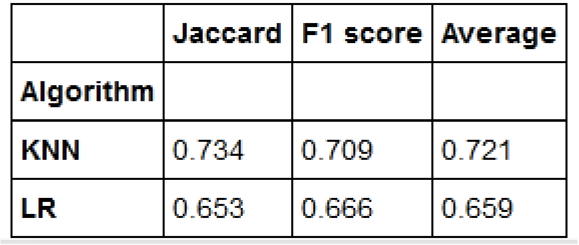


Based on the graph k=10 was chosen for best accuracy



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

Statistical values for Logistic Regression



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

➢ Data Summary of the two Classification Algorithm
➢ based on the average value KNN Algorithm will be the preferred choice of model

Analysis

❑ The KNN algorithm thus confirms that the independent variables 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

❑ Majority collisions happened on work days during the day

❑ 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