Capstone Project - Car accident severity¶

A. Introduction

The effective treatment of road accidents and thus the enhancement of road safety is a major concern to societies, due to the losses in human lives and the economic and social cost.

The Seattle government try find the best solution to prevent avoidable accident by warn drivers to be more careful in critical situation by given weather, 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, we will handle with many attribute it could help us to avoid this accident, There could be some tendencies of driving habit with day of week that could help us know about factors impacting the collisions better. (For Example, people tend to drink on Friday or Saturday night, drunk people loves to speeding, speeding lead to collisions).

The main purpose for the project to provide advices for target audience such as police, rescue group, insurance companies and drivers themselves to make insightful decision it could help to reducing accident and accident severity.

B. Data Description

B.1.Data acquisition

The Data provided by Seattle Police Department and Traffic Records department from 2004 to Present, in total there are 37 attribute (independent variables) , The dependent variable (SEVERITYCODE) contains numbers of the correspond to different level of severity caused by accident from 0 to 4 , are as follows:

0: Little to no Probability (Clear Condition)\

1: very low probability (Chance or property damage)\

2: low probability (Chance of injury)\

3: Mid probability (Chance of serious injury)\

4: High probability (Chance of fatality)\

Some attribute have missing data, have numerical and categorical types of data need for preprocessing before any future processing.

B.2.Data Preprocessing

The dataset consist 194,674 rows and 37 columns (attributes), we also noticed the original dataset is not ready for data analysis, first we will check data type of every column, then we will drop columns which is not required, in my case I will choose four attributes to solve the problem which is (SEVERITYCODE, WEATHER, ROADCOND, LIGHTCOND), and also we have categorical data type we have to convert it to numerical data type and numerical type.
Also we noticed that data unbalanced so we will use some statistical method to balance it.

C. Methodology

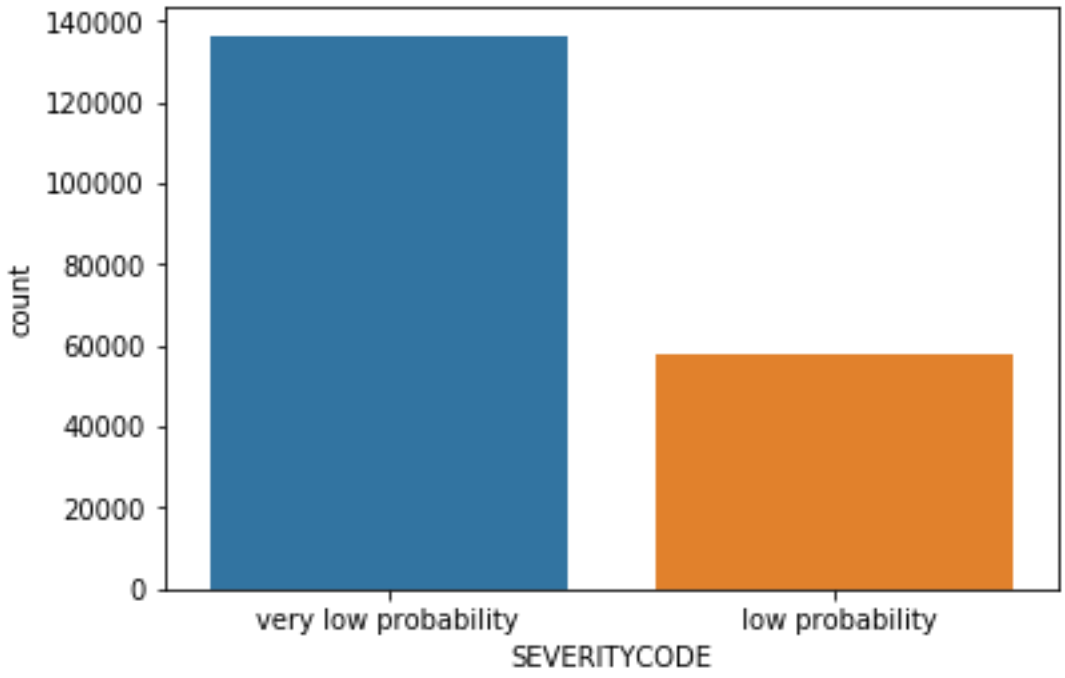
In this project our direct our effort to find proper way to provide advices for target audience and help them to make insightful decision that could help them to reducing accident severity.

So first I chose my target variables is severity code which describes the fatality of an accident. Then I choose three main conditions that could causes accident to train our model and then we will find out the relation between these conditions and severity code.

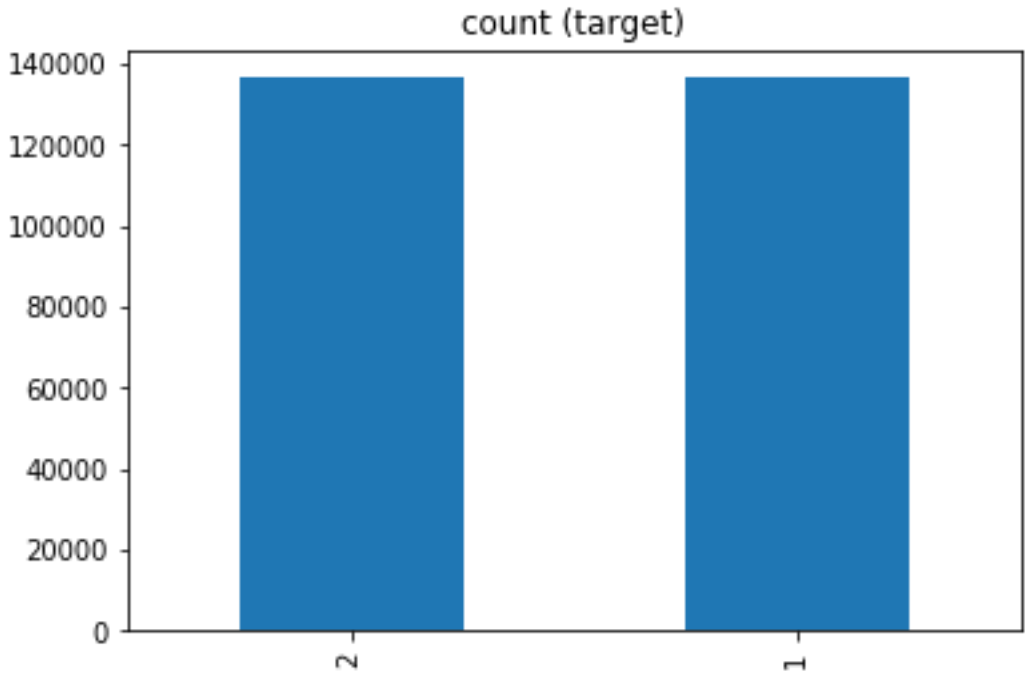
As we noticed before that we have categorical data that we have to transform it to numeric data so I convert the columns to a category, then I used those category values for my label encoding.

Also we have unbalanced data, we should balance it to improve the predictability of your model so I applied Random Over sampling by adding more copies to the minority class.

Unbalanced data



Balanced data

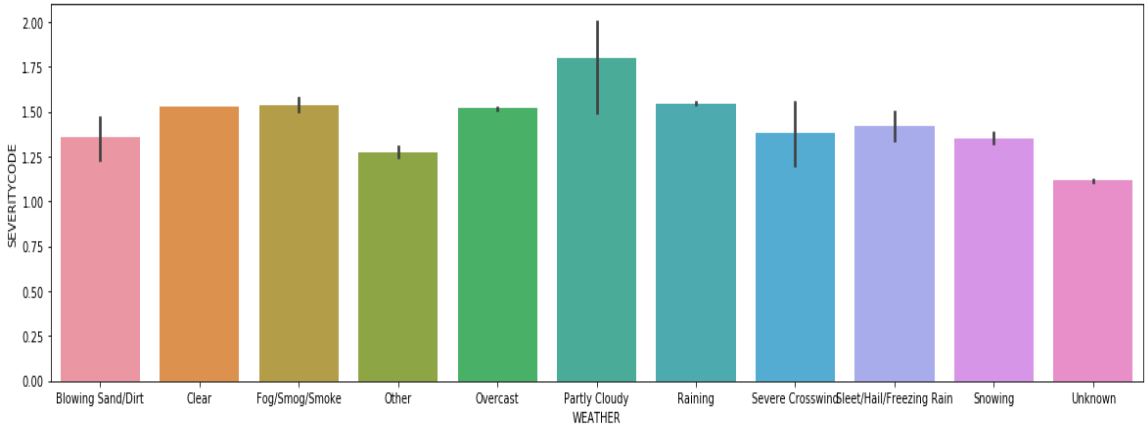


D. Analysis

D.1.Relation between weather and accident severity

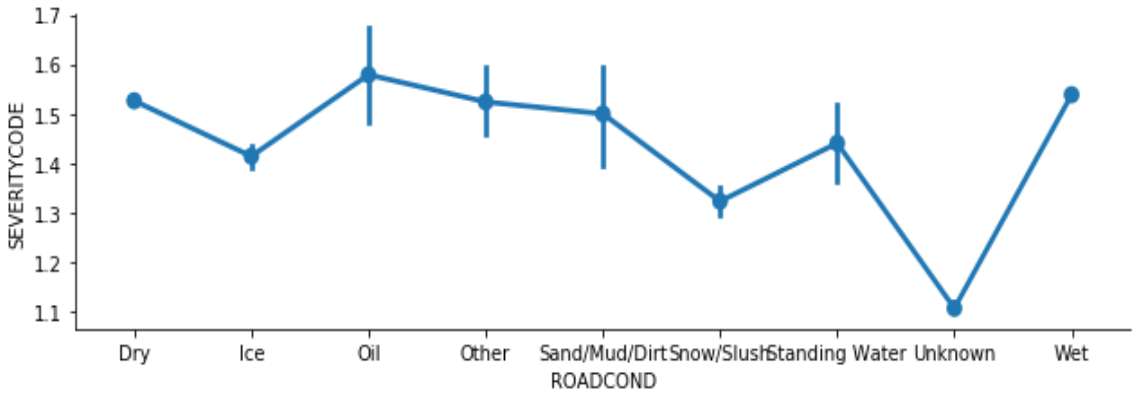
Road accident severity may be influenced by a number of factors.
The weather information recorded was taken as the prevailing weather at the time of the accident. At the local authority level, accident severity for the various adverse weather categories of rain, fog, and high winds is compared with the nonhazardous condition of fine weather. Severity ratios are then

calculated. Findings establish that accident severity accident severity increase in partly cloudy weather compared with raining.



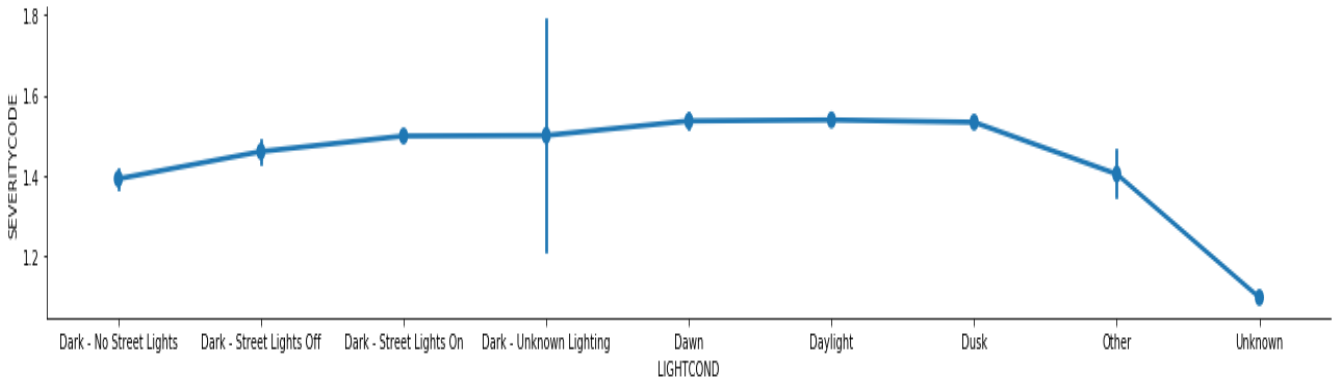
D.2. Relation between road conditions and accident severity

For every crash, information collected from police reports on reported environmental and road conditions. These conditions are not subsequently cross-referenced with observed meteorological data. Each record indicates an environmental condition around the time of the crash: clear, cloudy, snow, rain, sleet, or fog. For road conditions, each record indicates dry, wet, snow/slush, ice, or sand/dirt/oil. The adverse road conditions of wet, snow/slush, and ice were defined to have an environmental precursor our criteria to identify adverse environmental and road conditions. And as we noticed in our table when we have oil on the road the accident severity increased.



D.3.Relation between Light condition and accident severity

Accident reports also refer to road light condition, which indicate to important attribute that could cause accident severity and as table shown the severity increase when it dusk and dark _ unknown lighting.



E. Predictive Modeling

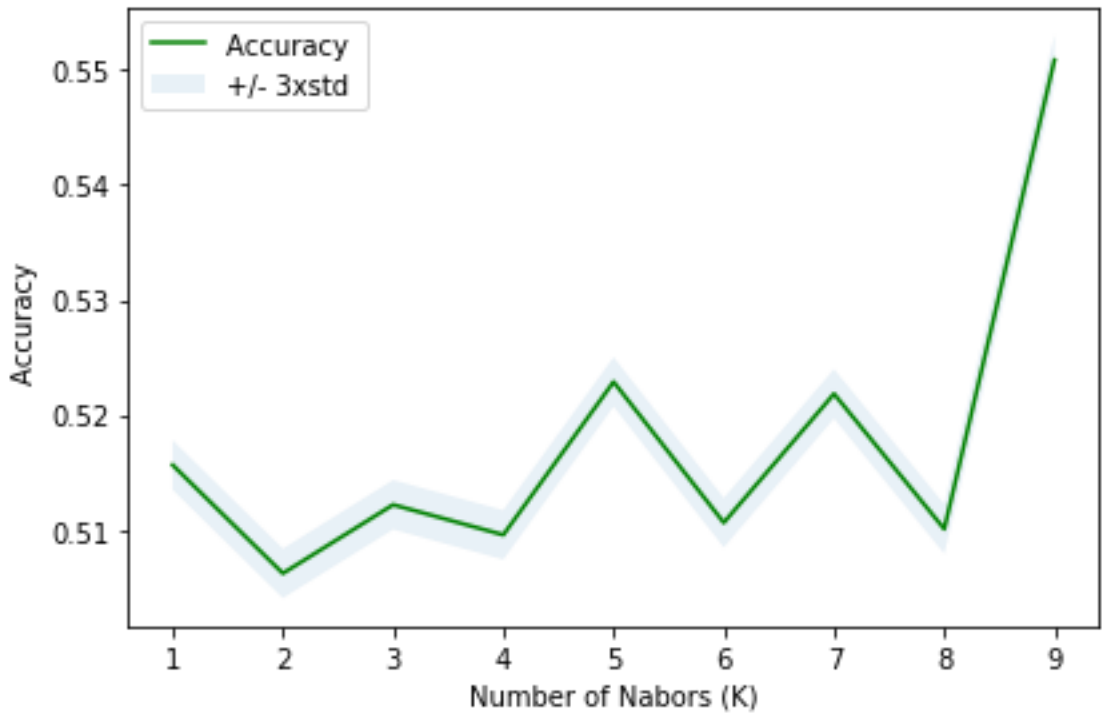
We are looking to predict accident and how serve it could be, therefore I will use Classification model, because classification models focus on the probabilities, I will apply K nearest neighbor (KNN), support vector machines (SVM) , logistic linear regression .

E.1. Classification model

E.1.1. K nearest neighbor

I split our dataset into train and test set, then I train with the training set and test with the testing set.

This will provide a more accurate evaluation on out-of-sample accuracy because the testing dataset is not part of the dataset that have been used to train the data.And I calculate the accuracy of KNN for different Ks , Then I use the model to predict the test set.



The best accuracy was with 0.5508847126057809 with k= 9

Model evaluation:

Train set Accuracy 0.5488286258563212

Test set Accuracy 0.5508847126057809

jaccard_similarity_score 0.5508847126057809

f1_score 0.5270148932750356

E.1.2. support vector machines (SVM)

First of all I split the data to train and test set, and I use RBF (Radial Basis Function) as kernel function for this project, after I fitted the data set to the model, I use it to predict new values

Finally,

Model evaluation:

jaccard_similarity_score 0.5642195113016082

f1_score 0.5433004303284255

E.1.3. Logistic Regression

Logistic Regression is a variation of Linear Regression, useful when the observed dependent variable, y, is categorical. It produces a formula that predicts the probability of the class label as a function of the independent variables. To apply this model first I normalize the data set, and split it into train and test set.

To build our model I use numerical optimizers to find parameters which is (liblinear) .

After fitting, I use our model to predict using our test set.

Finally,

Model evaluation:

jaccard_similarity_score 0.5299117119097336

F1_score 0.5160896282263393

Log_loss 0.6840821331431094

F. Classification models

Performance of classification models.

	Algorithm	Jaccard	F1-score	Logistic Regression
0	**KNN**	**0.550885**	**0.527015**	**Nan**
1	**SVM**	**0.564220**	**0.543300**	**Nan**
2	**Logistic Regression**	**0.529912**	**0.516090**	**0.684082**

G. Conclusions

The proposed model was simulated to predict the likelihood of the accident severity caused by diverse road factors so that some effective preventive measures can be taken to reduce such accidents. The projected model may pave the way to the relevant authorities to initiate preventive measures which is inexpensive rather than undertake costly corrective measures in long run depending on the unfavorable road conditions that cause such serious accident severity.