Coursera

IBM Applied Data Science Capstone Project

Seattle Accident Analysis and Classification

Hossein Faghih

Sep 2020

# Introduction

There are almost 200 million cars in United States that are being used every day. City of Seattle, as a sample, has reported nearly 200 thousand reports of car accidents during 2005-2020. An accident can lead to property damage, injury, and loss of life.

If we can predict the chances of accidents in different situations, we may be able to stop them at their origin or create other safety measures to reduce the number and severity of crashes happening every day.

For this aim, we are trying to use the accident log of the Seattle police department for the period of 2005-2020 to create a model that predicts the chance of a severe accident in different weather and road conditions. We will also get insights into how several human factors will affect crashes which might lead to policies to limit those incidents.

# Data

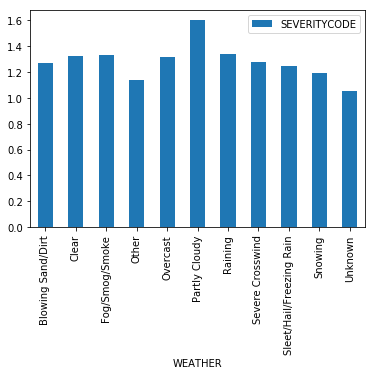
We have 194673 reported accidents with 38 columns of descriptive, administrative, and contextual data. As we go through the data, assuming that the context of an accident has a significant role on it, we try to choose the features of the model.

We have data on weather, road and lighting conditions, type of the address, whether a parked car was hit and several human factors, e.g. whether the situation happened because of being under influence, speeding, inattention.

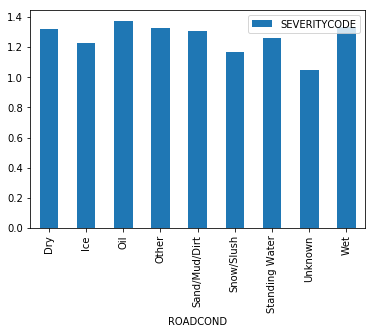
|  |  |  |
| --- | --- | --- |
| Variable | Description | Details |
| SEVERITYCODE | A code that corresponds to the severity of the collision | 3—fatality  2b—serious injury  2—injury  1—prop damage  0—unknown |
| ADDRTYPE | Collision Address Type | Alley  Block  Intersection |
| WEATHER | A description of the weather conditions during the time of the collision. | "Clear"  "Raining"  "Overcast"  "Snowing"  "Fog/Smog/Smoke"  "Sleet/Hail/Freezing Rain"  "Blowing Sand/Dirt"  "Severe Crosswind"  "Partly Cloudy"  "Other"  "Unknown" |
| LIGHTCOND | The light conditions during the collision. | "Daylight"  "Dark - Street Lights On"  "Dark - No Street Lights"  "Dusk"  "Dawn"  "Dark - Street Lights Off"  "Dark - Unknown Lighting"  "Other"  "Unknown" |
| ROADCOND | The condition of the road during the collision. | "Dry"  "Wet"  "Ice"  "Snow/Slush"  "Other"  "Standing Water"  "Sand/Mud/Dirt"  "Oil"  "Unknown" |
| SPEEDING | Whether or not speeding was a factor in the collision. | Y  N |
| UNDERINFL | Whether or not a driver involved was under the influence of drugs or alcohol. | Y  N |
| INATTENTIONIND | Whether or not collision was due to inattention. | Y  N |

We evaluated each variable regarding the relations with severity of accident:

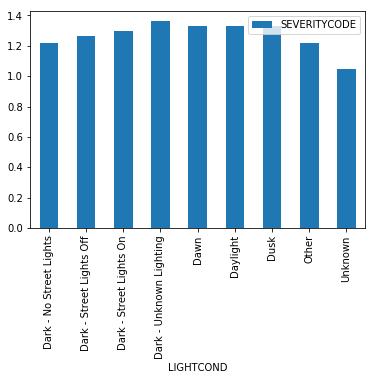
Weather:



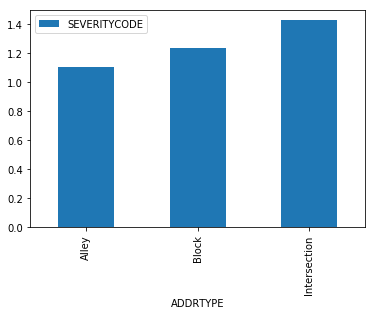
Road conditions:



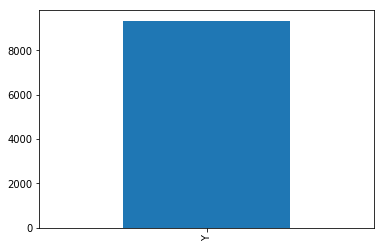
Light conditions:



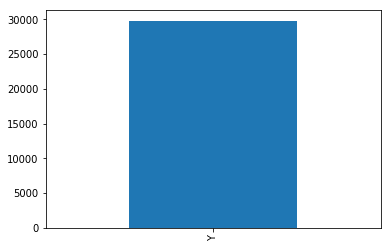
Address Type:



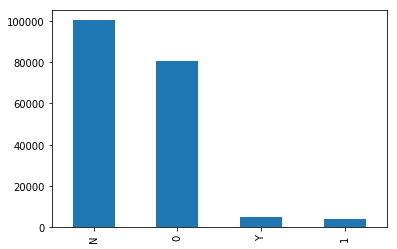
Speeding:



Inattention:



Under influence:



Then we cleansed the data with the following standard and replaced the NaN with median of each column.

# Methodology

We built two models based on our data.

1. Linear Regression Model
2. Decision Tree model

Comparing the results, Decision Tree is slightly working better and is more explainable due to the nature of the model and the problem we had at hand. Details can be seen in the notebook.

We tried to create a Support Vector model as well, however, the computations seemed to be outside the available computing power.

# Results

For Logistic Regression model the results are:

precision recall f1-score support

1 0.71 0.97 0.82 27425

2 0.44 0.06 0.10 11510

micro avg 0.70 0.70 0.70 38935

macro avg 0.58 0.51 0.46 38935

weighted avg 0.63 0.70 0.61 38935

for decision tree model the results are:

precision recall f1-score support

1 0.71 0.99 0.83 27425

2 0.53 0.01 0.03 11510

micro avg 0.70 0.70 0.70 38935

macro avg 0.62 0.50 0.43 38935

weighted avg 0.66 0.70 0.59 38935

# Conclusion

By looking at the accuracy scores of the models we see that both model are working at the same level of 71% on predicting a property damage index in severity score while the decision tree model is working slightly better (53% vs. 44%) in terms of the injuries index of the severity score.

It is astonishing that each element of the model is very weakly correlated with the outcome of the model, however, the model is producing an accuracy of 70% for the results.