Car accident severity

Nikhil Veeramalli

1. Introduction

The Seattle government is going to prevent avoidable car accidents by employing methods that alert drivers, health system, and police to remind them to be more careful in critical situations.

In most cases, not paying enough attention during driving, abusing drugs and alcohol or driving at very high speed are the main causes of occurring accidents that can be prevented by enacting harsher regulations.

Besides the aforementioned reasons, weather, visibility, or road conditions are the major uncontrollable factors that can be prevented by revealing hidden patterns in the data and announcing warning to the local government, police and drivers on the targeted roads.

The model and its results are going to provide some advice for the target audience to make insightful decisions for reducing the number of accidents and injuries for the city.

The data we have consists of includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. Timeframe: 2004 to Present.

2. Data

The data used for this study is given by the Applied Data Science Capstone course on Coursera.org via the following link https://s3.us.cloud-object-storage.appdomain.cloud/cf-coursesdata/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv.

The dataset has information gathered on the road traffic accidents of Seattle City. Python packages will be used to conduct this study. The dataset will be cleaned according to the requirements of this project.

We chose the unbalanced dataset provided by the Seattle Department of Transportation Traffic Management Division with 194673 rows (accidents) and 37 columns (features) where each accident is given a severity code. It covers accidents from January 2004 to May 2020. Some of the features in this dataset include and are not limited to Severity code, Location/Address of accident, Weather condition at the incident site, Driver state (whether under influence or not), collision type. Hence we think its a good generalized dataset which will help us in creating an accurate predictive model. The unbalance with respect to the severity code in the dataset is as follows.

Dependent variable/ target: “SEVERITYCODE” (0 to 2 levels)

1 - Very Low Probability - Chance or Property Damage

2 - Low Probability - Chance of Injury

The initial dataset consists of 38 columns (features/attributes) and 194673 rows. Using pandas drop function we drop all columns expect the desired ones. Columns with the same information but in either categorical or numerical is dropped and only one is chosen.

3. **Methodology**

Feature selection is the process of reducing the number of input variables when developing a predictive model. It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model

Here, In the Data, we are considering only few features. They are WEATHER,ROADCOND,JUNCTIONTYPE,VEHCOUNT,PERSONCOUNT,SEVERITYDESC,ADDRTYPE,SDOT_COLDESC,LIGHTCOND.

/* Description of Variables*/

#SEVERITYDESC : A description of code that corresponds to the severity of the collision

#WEATHER : A description of the weather conditions during the time of the collision.

#ROADCOND : The condition of the road during the collision.

#JUNCTIONTYPE : Category of junction at which collision took place

#VEHCOUNT : The number of vehicles involved in the collision.This is entered by the state.

#PERSONCOUNT : The total number of people involved in the collision

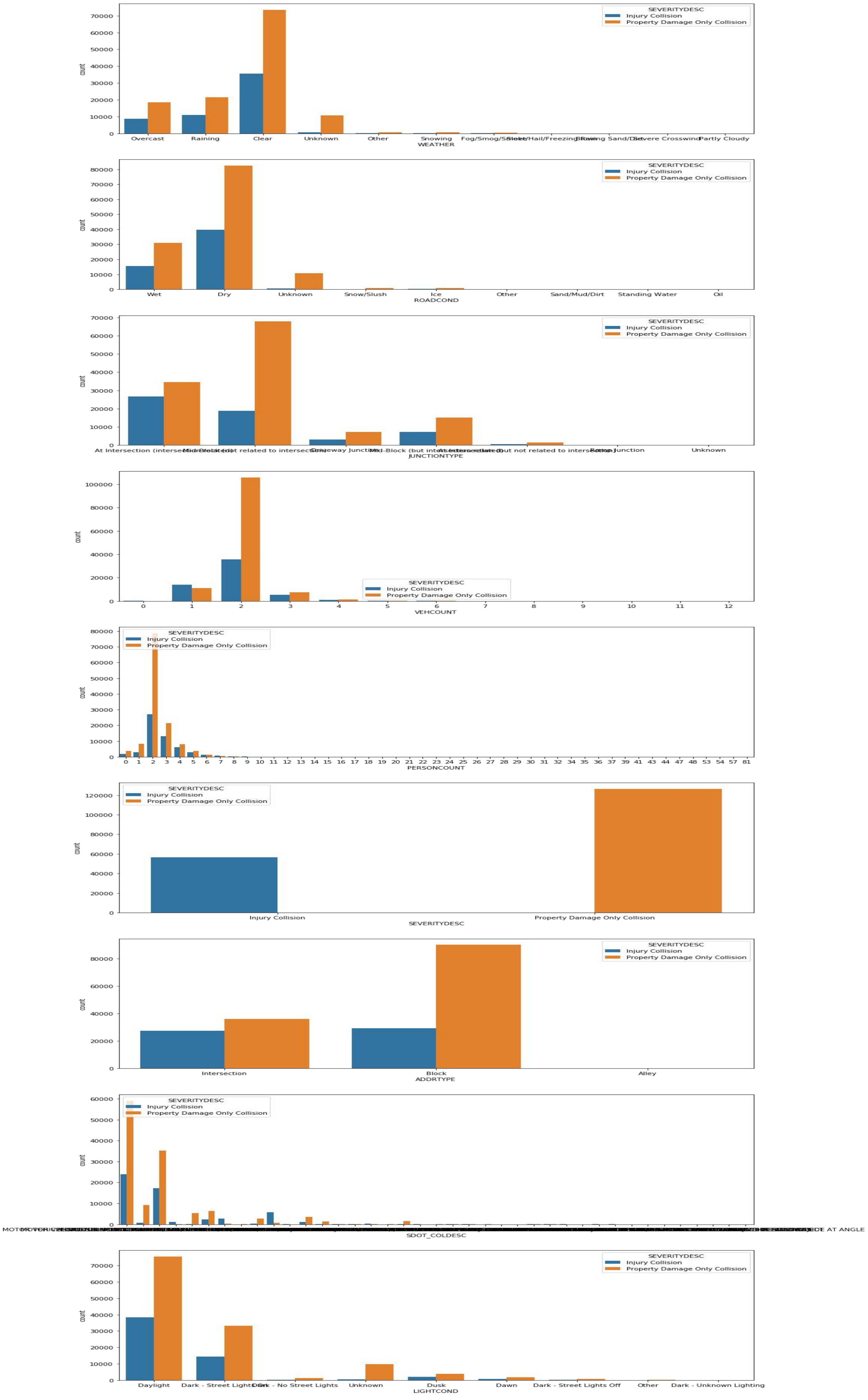
#ADDRTYPE : Collision address type-- Alley,Block,Intersection

#SDOT_COLDESC : A description of the collision corresponding to the collision code.

#LIGHTCOND : The light conditions during the collision.

Checked every attributes and Removed the missing values by removing whole rows I.e only few values are missing.

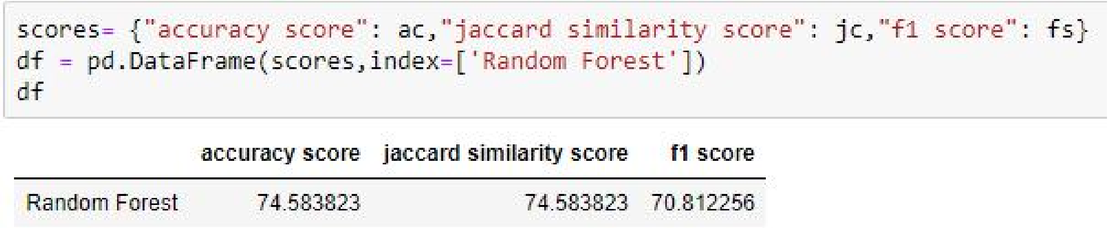
plot of Input variables w.r.t Target variable "SEVERITYDESC"


SEVERITYDESC
Injury Collision
Property Damage Only Collision
count
0
10000
20000
30000
40000
50000
60000
70000
Overcast
Raining
Clear
Unknown
Other
Snowing
Fog/Smog/Smoke
Sleet/Hail/Freezing Rain
Blowing Sand/Dirt
Severe Crosswind
Partly Cloudy
WEATHER
80000
Wet
Dry
Unknown
Snow/Slush
Ice
Other
Sand/Mud/Dirt
Standing Water
Oil
ROADCOND
At Intersection (intersection related)
Mid-Block (not related to intersection)
Driveway Junction
Mid-Block (but intersection related)
At Intersection (but not related to intersection)
Ramp Junction
Unknown
JUNCTIONTYPE
100000
0
1
2
3
4
5
6
7
8
9
10
11
12
VEHCOUNT
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 34 35 36 37 39 41 43 44 47 48 53 54 57 81
PERSONCOUNT
120000
Injury Collision
Property Damage Only Collision
SEVERITYDESC
Intersection
Block
Alley
ADDRTYPE
MOTOR VEHICLE
SIDE AT ANGLE
SDOT_COLDESC
Daylight
Dark - Street Lights On
Dark - No Street Lights
Unknown
Dusk
Dawn
Dark - Street Lights Off
Other
Dark - Unknown Lighting
LIGHTCOND

4. Model Development and Evaluation

The attributes for the data have categorical variables for their elements and the suitable model type for such data is a classification model. We will use Random Forest model for our model. Random Forest is a supervised learning technique that predicted the response by learning decision rules derived from features. It shows the possible outcomes and can be used to map out an algorithm that predicts the best choice mathematically. We will use Accuracy score , Jaccard Index and F1-score to evaluate our model. Jaccard Index, it is used to compare set of predicted values to their true values. The higher the score the better the accuracy. From 0 to 1, 1 being the best score. F1-score, it is the weighted average of the precision and recall. The score is between 0 and 1, best score is 1 and worst score at 0.

5. Results



```
scores= {"accuracy score": ac,"jaccard similarity score": jc,"f1 score": fs}
df = pd.DataFrame(scores,index=['Random Forest'])
df
```

	accuracy score	jaccard similarity score	f1 score
Random Forest	74.583823	74.583823	70.812256

6. Discussion and Conclusion

This project aimed to study the relationship between severity level of road traffic collisions in junction types. From our data analysis illustrates that there are different levels of collision count in junction and address types. Such as the difference in severity level in junction types in not so much in all junction expect Mid-block (not related to intersection) junction type. As for address types, at block type we can see the same thing for Property Damage Collison Only but for Injury Collison between Alley and Block the difference is so much different. To conclude, this project aimed at exploring the data to provide insight in severity levels for road collisions at junctions. The predictive model would be useful to help local authorities decide on whether to implement new safety measures in certain areas