Car Accident Severity Analysis using Machine Learning Algorithms

1.Introduction & Business Understanding

Road accidents are one of the major causes of death and disability all over the world. Reason for road accidents can be environmental conditions such as weather, traffic on road, type of road, speed and light conditions. This paper addresses the in-depth analysis that identifies as the contributory factors behind the road accidents and the quantification of the factors that affect the frequency and

severity of accidents based on the crash data available. The severity of each accident can be predicted quite accurately with various classification machine learning algorithms. This can ultimately help government, traffic police, medical institutions, individual drivers and the insurance companies by getting useful insights of the accident severity regarding the causes and consequences of the accidents. The Machine Learning 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 model will predict the accident severity with various supervised machine learning algorithms i.e. * Algorithm A. Logistic regression * Algorithm B. The K-Nearest Neighbours (KNN) algorithm * Algorithm C. Decision Tree * Algorithm D. Random Forest And finally, the accuracy score versus algorithm will be plotted to check which algorithm performs better.

2. Data Understanding

The data used for this project was collected by the SDOT traffic management Division and Seattle Traffic Records Group from 2004 to present. It was downloaded from the link shared in the IBM Applied Data Science Capstone course. The data consists of 38 independent variables and 194,673 rows. The dependent variable, "SEVERITYCODE", contains numbers that correspond to different levels of severity caused by an accident from 1 to 2. Severity codes are as follows: 1: Property Damage Only Collision 2: Injury Collision Furthermore, as there are null values in some records, the data needs to be pre-processed before any further processing.

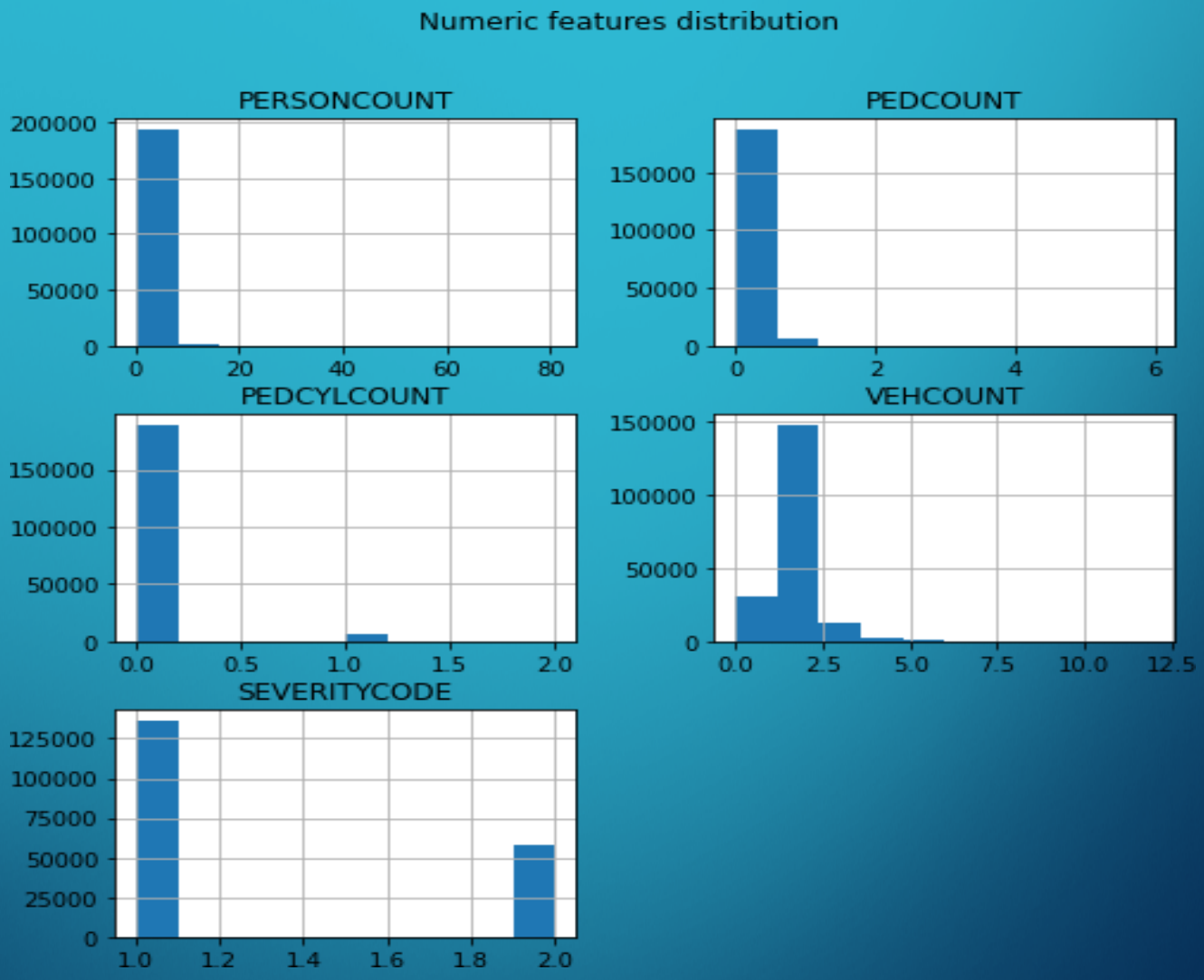
Reading the CSV Data

<class 'pandas.core.frame.DataFrame'> RangeIndex: 194673 entries, 0 to 194672 Data columns (total 37 columns):

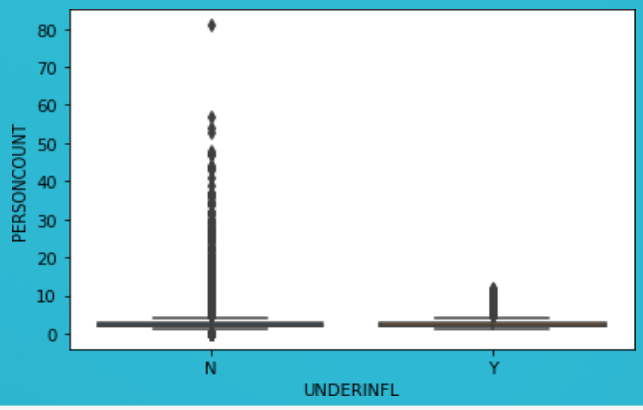
#	Column	Non-Null Count	Dtype
0	SEVERITYCODE	194673 non-null	int64
1	longitude	189339 non-null	float64
2	latitude	189339 non-null	float64
3	OBJECTID	194673 non-null	int64
4	INCKEY	194673 non-null	int64
5	COLDETKEY	194673 non-null	int64
6	REPORTNO	194673 non-null	object
7	STATUS	194673 non-null	object
8	ADDRTYPE	192747 non-null	object
9	INTKEY	65070 non-null	float64
10	LOCATION	191996 non-null	object
11	EXCEPTRSNCODE	84811 non-null	object
12	EXCEPTRSNDESC	5638 non-null	object
13	SEVERITYDESC	194673 non-null	object
14	COLLISIONTYPE	189769 non-null	object
15	PERSONCOUNT	194673 non-null	int64
16	PEDCOUNT	194673 non-null	int64
17	PEDCYLCOUNT	194673 non-null	int64
18	VEHCOUNT	194673 non-null	int64
19	INCDATE	194673 non-null	object
20	INCDTTM	194673 non-null	object
21	JUNCTIONTYPE	188344 non-null	object
22	SDOT_COLCODE	194673 non-null	int64
23	SDOT_COLDESC	194673 non-null	object
24	INATTENTIONIND	29805 non-null	object
25	UNDERINFL	189789 non-null	object
26	WEATHER	189592 non-null	object
27	ROADCOND	189661 non-null	object
28	LIGHTCOND	189503 non-null	object
29	PEDROWNOTGRNT	4667 non-null	object
30	SDOTCOLNUM	114936 non-null	float64
31	SPEEDING	9333 non-null	object
32	ST_COLCODE	194655 non-null	object
33	ST_COLDESC	189769 non-null	object
34	SEGLANEKEY	194673 non-null	int64
35	CROSSWALKKEY	194673 non-null	int64
36	HITPARKEDCAR	194673 non-null	object

Description of the Numeric Features

	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	SEVERITYCODE
count	194673.000000	194673.000000	194673.000000	194673.000000	194673.000000
mean	2.444427	0.037139	0.028391	1.920780	1.298901
std	1.345929	0.198150	0.167413	0.631047	0.457778
min	0.000000	0.000000	0.000000	0.000000	1.000000
25%	2.000000	0.000000	0.000000	2.000000	1.000000
50%	2.000000	0.000000	0.000000	2.000000	1.000000
75%	3.000000	0.000000	0.000000	2.000000	2.000000
max	81.000000	6.000000	2.000000	12.000000	2.000000



Plotting the under Influence of alcohol along with the person count



3. Data Preparation

Categorical Features percentage (%) of samples from the selected Data

1."COLLISIONTYPE"

	counts	Percent
Parked Car	47987	25.3%
Angles	34674	18.3%
Rear Ended	34090	18.0%
Other	23703	12.5%
Sideswipe	18609	9.8%
Left Turn	13703	7.2%
Pedestrian	6608	3.5%
Cycles	5415	2.9%
Right Turn	2956	1.6%
Head On	2024	1.1%

2. "LIGHTCOND"

	counts	Percent
Daylight	116137	61.30%
Dark - Street Lights On	48507	25.60%
Unknown	13473	7.10%
Dusk	5902	3.10%
Dawn	2502	1.30%
Dark - No Street Lights	1537	0.80%
Dark - Street Lights Off	1199	0.60%
Other	235	0.10%
Dark - Unknown Lighting	11	0.00%

3."ROADCOND"

	counts	Percent
Dry	124510	65.60%
Wet	47474	25.00%
Unknown	15078	7.90%
Ice	1209	0.60%
Snow/Slush	1004	0.50%
Other	132	0.10%
Standing Water	115	0.10%
Sand/Mud/Dirt	75	0.00%
Oil	64	0.00%

4. "SDOT_COLDESC"

	counts	Percent
MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END A...	85209	43.80%
MOTOR VEHICLE STRUCK MOTOR VEHICLE, REAR END	54299	27.90%
MOTOR VEHICLE STRUCK MOTOR VEHICLE, LEFT SIDE S...	9928	5.10%
NOT ENOUGH INFORMATION / NOT APPLICABLE	9787	5.00%
MOTOR VEHICLE RAN OFF ROAD - HIT FIXED OBJECT	8856	4.50%
MOTOR VEHCILE STRUCK PEDESTRIAN	6518	3.30%
MOTOR VEHICLE STRUCK MOTOR VEHICLE, LEFT SIDE A...	5852	3.00%
MOTOR VEHICLE STRUCK OBJECT IN ROAD	4741	2.40%
MOTOR VEHICLE STRUCK PEDALCYCLIST, FRONT END AT...	3104	1.60%
MOTOR VEHICLE STRUCK MOTOR VEHICLE, RIGHT SIDE ...	1604	0.80%
MOTOR VEHICLE STRUCK MOTOR VEHICLE, RIGHT SIDE ...	1440	0.70%
PEDALCYCLIST STRUCK MOTOR VEHICLE FRONT END AT ...	1312	0.70%
MOTOR VEHICLE OVERTURNED IN ROAD	479	0.20%
MOTOR VEHICLE STRUCK PEDALCYCLIST, REAR END	181	0.10%
PEDALCYCLIST STRUCK MOTOR VEHICLE LEFT SIDE SID...	180	0.10%
MOTOR VEHICLE RAN OFF ROAD - NO COLLISION	166	0.10%
PEDALCYCLIST STRUCK MOTOR VEHICLE REAR END	139	0.10%
MOTOR VEHICLE STRUCK PEDALCYCLIST, LEFT SIDE SI...	124	0.10%
DRIVERLESS VEHICLE RAN OFF ROAD - HIT FIXED OBJECT	107	0.10%
DRIVERLESS VEHICLE STRUCK MOTOR VEHICLE FRONT E...	104	0.10%
MOTOR VEHICLE STRUCK TRAIN	102	0.10%
DRIVERLESS VEHICLE STRUCK MOTOR VEHICLE REAR END	93	0.00%
PEDALCYCLIST STRUCK PEDESTRIAN	75	0.00%
PEDALCYCLIST OVERTURNED IN ROAD	69	0.00%
DRIVERLESS VEHICLE STRUCK MOTOR VEHICLE LEFT SI...	53	0.00%
PEDALCYCLIST STRUCK MOTOR VEHICLE RIGHT SIDE SI...	50	0.00%
PEDALCYCLIST STRUCK OBJECT IN ROAD	23	0.00%
MOTOR VEHICLE STRUCK PEDALCYCLIST, RIGHT SIDE S...	17	0.00%
DRIVERLESS VEHICLE STRUCK MOTOR VEHICLE RIGHT S...	12	0.00%
PEDALCYCLIST STRUCK MOTOR VEHICLE LEFT SIDE AT ...	9	0.00%
DRIVERLESS VEHICLE STRUCK PEDESTRIAN	8	0.00%
PEDALCYCLIST STRUCK PEDALCYCLIST REAR END	7	0.00%
DRIVERLESS VEHICLE STRUCK MOTOR VEHICLE RIGHT S...	6	0.00%
PEDALCYCLIST STRUCK PEDALCYCLIST FRONT END AT A...	5	0.00%
PEDALCYCLIST RAN OFF ROAD - HIT FIXED OBJECT	4	0.00%
DRIVERLESS VEHICLE STRUCK MOTOR VEHICLE LEFT SI...	4	0.00%
DRIVERLESS VEHICLE STRUCK OBJECT IN ROADWAY	3	0.00%
PEDALCYCLIST STRUCK MOTOR VEHICLE RIGHT SIDE AT...	2	0.00%
DRIVERLESS VEHICLE RAN OFF ROAD - NO COLLISION	1	0.00%

5. "WEATHER"

	counts	Percent
Clear	111135	58.60%
Raining	33145	17.50%
Overcast	27714	14.60%
Unknown	15091	8.00%
Snowing	907	0.50%
Other	832	0.40%
Fog/Smog /Smoke	569	0.30%
Sleet/Hail /Freezing Rain	113	0.10%
Blowing Sand/Dirt	56	0.00%
Severe Crosswind	25	0.00%
Partly Cloudy	5	0.00%

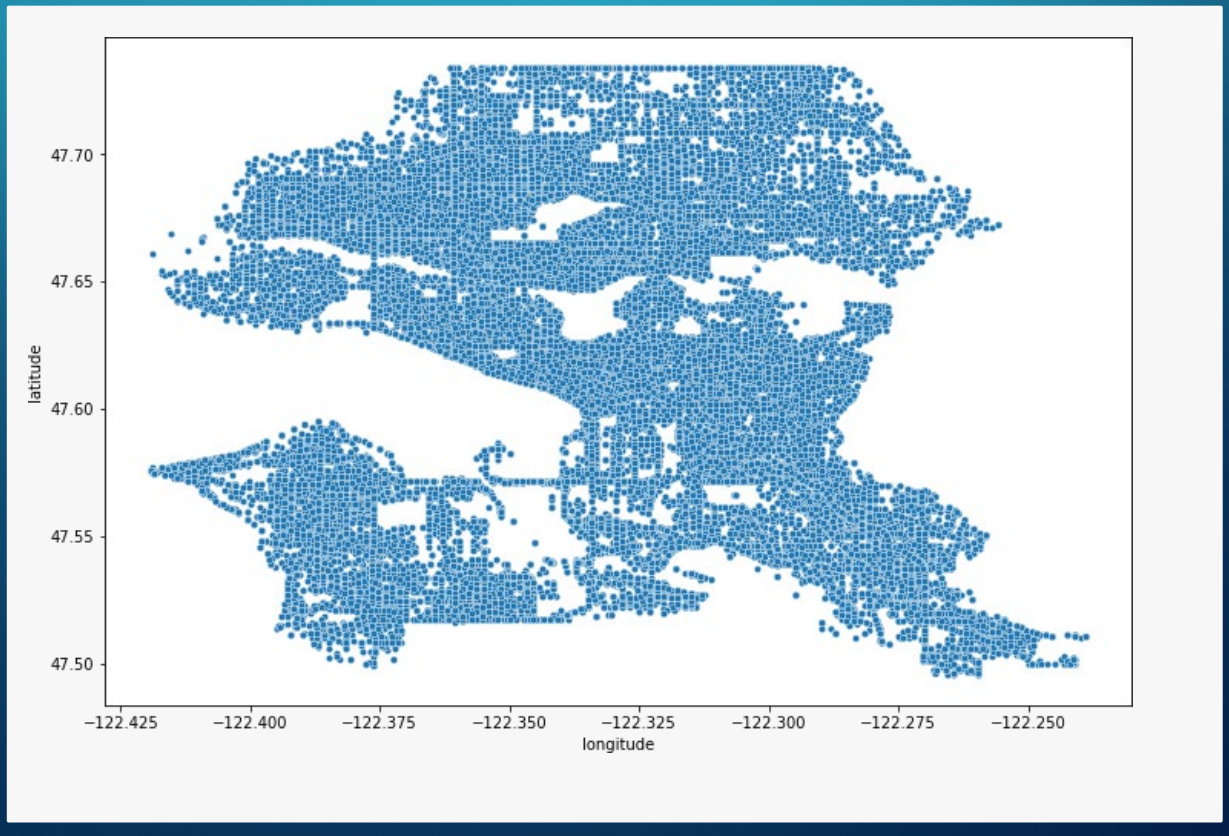
6. "JUNCTIONTYPE"

	counts	Percent
Mid-Block (not related to intersection)	89800	47.70%
At Intersection related to intersection	62810	33.30%
Mid-Block (but intersection related)	22790	12.10%
Driveway Junction	10671	5.70%
At Intersection_not related to intersection	2098	1.10%
Ramp Junction	166	0.10%
Unknown	9	0.00%

7. "UNDERINFL"

	counts	Percent
N	180668	95.20%
Y	9121	4.80%

Scatter plot of the accident coordinates



Selecting and finalizing the features for Machine Learning Model

Selected Features = ["SEVERITYCODE", "longitude", "latitude", "PERSONCOUNT", "PEDCOUNT", "PEDCYLCOUNT", "VEHCOUNT", "ADDRTYPE", "COLLISIONTYPE", "WEATHER", "ROADCOND", "LIGHTCOND", "SDOT_COLDESC", "HITPARKEDCAR", "Hour"]

4. Modeling and Evaluation

1.Logistic Regression

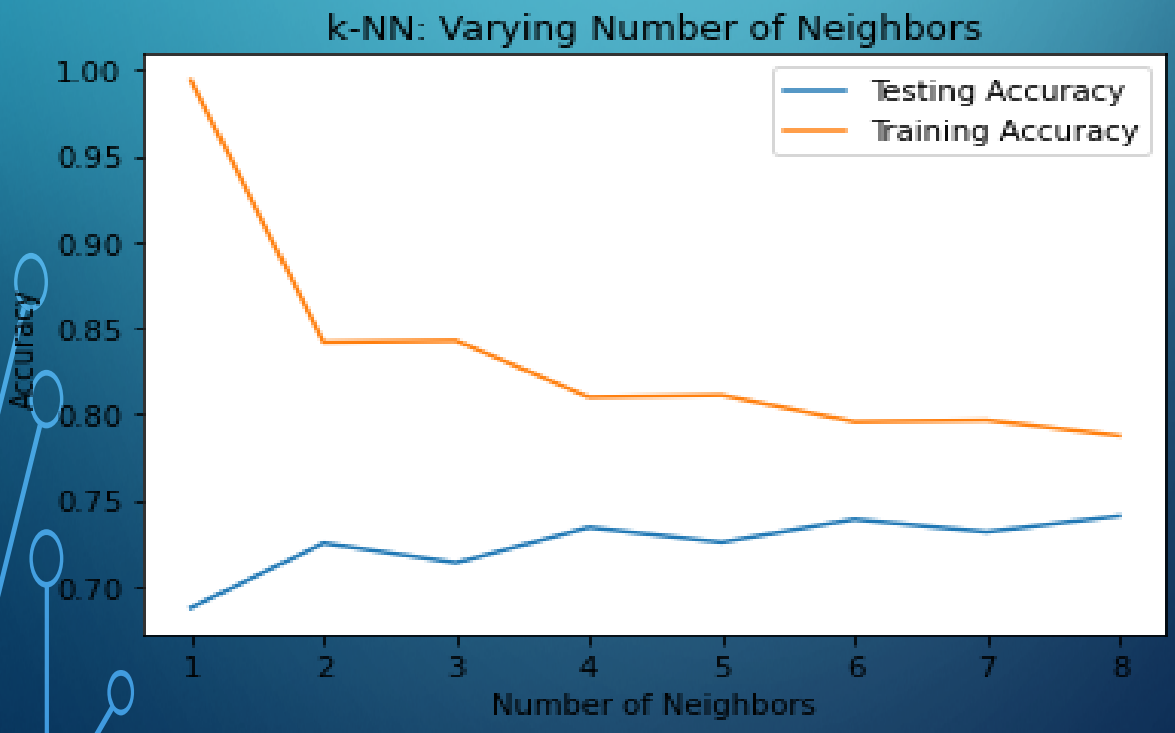
[Logistic regression algorithm] accuracy score: 0.756.

2.K-NN Neighbors

[K-Neighbors Classifier(n_neighbors=6)

[K-Nearest Neighbors (KNN)] knn.score: 0.739.
[K-Nearest Neighbors (KNN)] accuracy_score: 0.739.

Generating a plot for K-NN with varying number of Neighbours



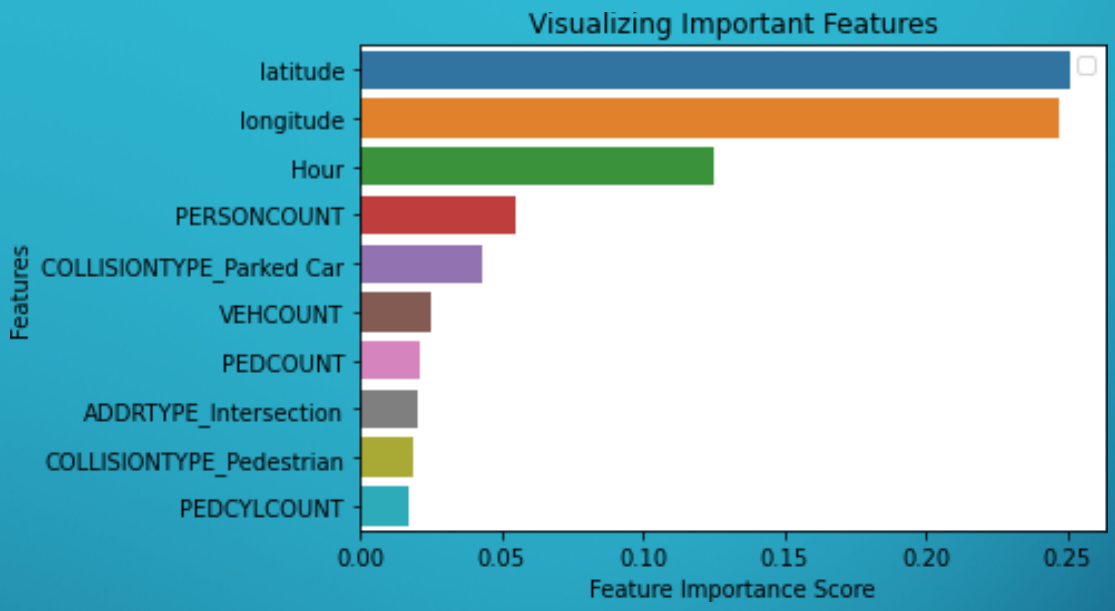
3.Decision Tree Algorithm

[Decision Tree -- entropy] accuracy score: 0.754.
[Decision Tree -- Gini] accuracy score: 0.754.

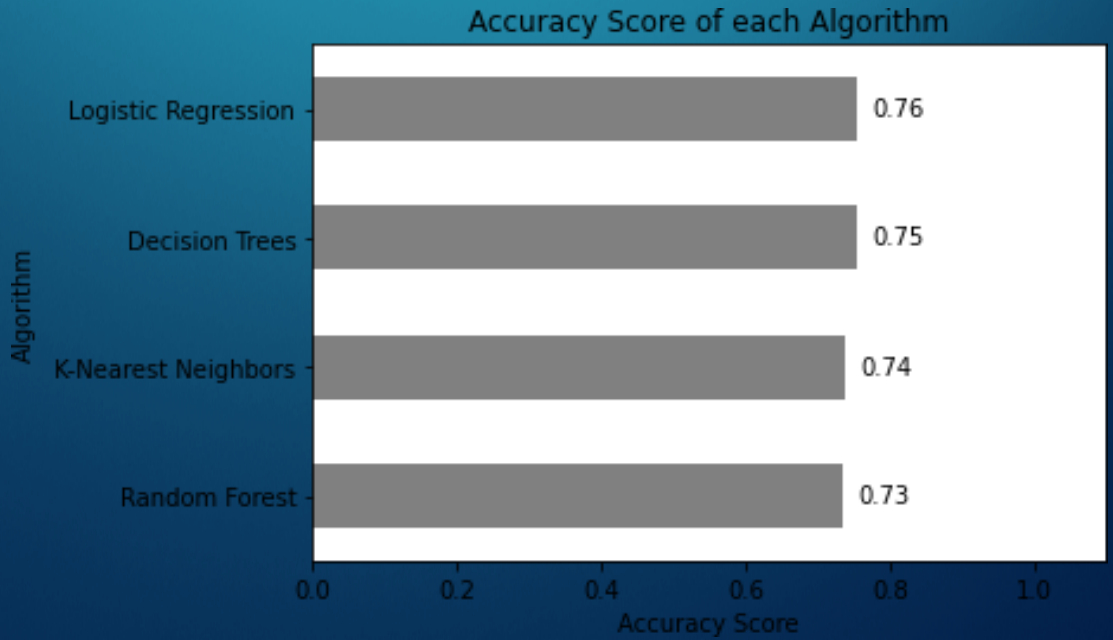
4. Random Forest Algorithm

[Random forest algorithm] accuracy score: 0.735.

Visualizing the important features



Accuracy score of considered Algorithms



5. Deployment

For the deployment phase as it can vary from project to project a simple pdf report has been generated.

6. Summary

- Seattle road accidents data has been analyzed to get useful insights.
- The data contains multiple attributes e.g. accident severity, collision type, coordinates of the incident, date and time of the incident, weather and road conditions, address types, no of persons injured and property damage and many other attributes.
- There are two accident severity types i.e.
  - Property damage only collision(1)
  - Injury collision(2)
- After the data preparation understanding phase, data preparation phase was carried out by selecting the right features for the machine learning model.
- In the Modeling phase, 4 algorithms were selected where the target class was "accident severity".
- Based on the predictions, "Logistic Regression" relatively performed better among the others with the percentage of 76%.

7. Conclusion

Based on the selected dataset(features) for this capstone that include mainly, coordinates, hour, person count and the collision type, it can be concluded that these particular classes have a somewhat impact on whether or not travelling along the Seattle roads could result in property damage (class 1) or injury (class 2). In this study, the technique of association rules with a large set of accident data to identify the reasons of road accidents were used.The results show that this model could provide good predictions against traffic accident with 76% correct rate.It should be noted that due to the constraints of data and research condition, there are still some factors, such as engine capacity, traffic flows, gender, age of the driver, attaining the missing data etc. that are not considered in this model and can be taken into account for future study. The results of this study can be used in vehicle safety assistance driving and provide early warnings and proposals for safe driving and hence help in reducing the number of accidents.