Capstone Project – Car Accident Severity

Prediction Model Presentation

By: Prasad Jaywant

Sept 25th 2020

Background

Seattle Department of Transportation (SDOT) is on a mission to deliver a transportation system that provides safe and affordable access to the places. The council's goal is to create safe transportation environments and eliminate serious and fatal crashes in Seattle. Making sure people can get around the growing city safely is the council's top priority.

It becomes a growing need if there is something in place that could warn road commuters, given the weather and the road conditions about the possibility of getting into a car accident and how severe it could be. Based on such alerts, people could drive more carefully or even change their travel if they are able to.

"Capstone Project aims to predict severity of a car accident reliably, so as to help citizens reach places safely and timely."

Stakeholders & beneficiaries

The severity impact prediction model (which is scope of this project) could be published as a REST API or web service (future scope of work) for the Seattle Department of Transportation (SDOT). The SDOT may have options to own or to subscribe to this service. By inputting necessary data to the service it could receive predictions regarding severity of accidents. This would help SDOT formulate traffic routing decisions or alerts in the geography under its monitoring.

Daily commuters and road travelers would find it much convenient to know about live traffic information, traffic diversion alerts and notifications when they tune with the SDOT broadcast channels. It would help save everyone's precious time, hectic travels and help avert mishaps or accidents due to such forewarnings.

Methodology

We introduce here the research methods and data source used for the analysis. We would discuss in key highlights in below sections about the data, choice of variables, modelling methods and how they would help answer the problem statement. The methodology steps essentially are as follows;

- ✓ Data collection and understanding
    - Data source
    - Data understanding
- ✓ Data preparation
    - Basic insight of dataset
    - Feature selection
    - Data cleansing
    - Data transforming
    - Test of correlation and significance
    - Conclusion of important variables
- ✓ Model development
    - Algorithms and empirical findings
    - Results summary
- ✓ Discussion
- ✓ Conclusion
- ✓ References
- ✓ Acknowledgement

Data Understanding

We have used shared data of Seattle city as basis to deal with the accidents data

(source: http://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0.csv).

At first glance at the CSV file, we could see what type of data we have with us. The label for the data set is Severity, which describes the fatality of an accident. The remaining columns have different types of attributes. Also noticed that the data had some unbalanced attributes which need to be normalized during next steps.

We also used the collisions meta data available for about 16 years to understand the nature of all attributes. Having about 2.21L data observations, we could notice that a split of these could be used to train and test the prospective model.

Data Preparation (Data Fields)

In the given dataset, SeverityCode is identified as the target variable (labelled or dependent) while rest of the fields are construed as independent variables or the attributes.

The case objective with the given data, does qualify it as a classification problem of the supervised machine learning.

All columns that could influence the cause and impact of an accident need to be selected for training and testing the model.

```
[6]:  1 df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 221389 entries, 0 to 221388
Data columns (total 40 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   X                213918 non-null  float64
 1   Y                213918 non-null  float64
 2   OBJECTID         221389 non-null  int64
 3   INCKEY           221389 non-null  int64
 4   COLDETKEY        221389 non-null  int64
 5   REPORTNO         221389 non-null  object
 6   STATUS           221389 non-null  object
 7   ADDRTYPE         217677 non-null  object
 8   INTKEY           71884 non-null   float64
 9   LOCATION         216801 non-null  object
 10  EXCEPTRSNCODE    100986 non-null  object
 11  EXCEPTRSNDESC    11779 non-null   object
 12  SEVERITYCODE     221388 non-null  object
 13  SEVERITYDESC     221389 non-null  object
 14  COLLISIONTYPE    195159 non-null  object
 15  PERSONCOUNT      221389 non-null  int64
 16  PEDCOUNT         221389 non-null  int64
 17  PEDCYLCOUNT      221389 non-null  int64
 18  VEHCOUNT         221389 non-null  int64
 19  INJURIES         221389 non-null  int64
 20  SERIOUSINJURIES  221389 non-null  int64
 21  FATALITIES       221389 non-null  int64
 22  INCDATE          221389 non-null  object
 23  INCDTTM          221389 non-null  object
 24  JUNCTIONTYPE     209417 non-null  object
 25  SDOT_COLCODE     221388 non-null  float64
 26  SDOT_COLDESC     221388 non-null  object
 27  INATTENTIONIND   30188 non-null   object
 28  UNDERINFL        195179 non-null  object
 29  WEATHER          194969 non-null  object
 30  ROADCOND         195050 non-null  object
 31  LIGHTCOND        194880 non-null  object
 32  PEDROWNOTGRNT    5192 non-null    object
 33  SDOTCOLNUM       127205 non-null  float64
 34  SPEEDING         9928 non-null    object
```


```
 33  SDOTCOLNUM       127205 non-null  float64
 34  SPEEDING         9928 non-null    object
```


```
 35  ST_COLCODE       211976 non-null  object
 36  ST_COLDESC       195159 non-null  object
 37  SEGLANEKEY       221389 non-null  int64
 38  CROSSWALKKEY     221389 non-null  int64
 39  HITPARKEDCAR     221389 non-null  object
dtypes: float64(5), int64(12), object(23)
memory usage: 67.6+ MB
```

1 Python | Idle   Saving completed   Mode: Command   Ln 1, Col 1   Capstone Project_Prasad Jaywant_Submission.ipynb

Data Preparation (Rationale)

Sr. No.	Attribute	Data type, length	Description	Wrangling Method	Rationale
1	OBJECTID	OBJECTID	ESRI unique identifier	Dropped	Insignificance
2	X	Longitude	ESRI geometry field	Dropped	Insignificance
3	Y	Latitude	ESRI geometry field	Dropped	Insignificance
4	ADDRTYPE	Text, 12	Collision address type: Alley, Block, Intersection	Retained	2% missing data, replaced by max. frequency
5	INTKEY	Double	Key that corresponds to the intersection associated with a collision	Dropped	Insignificance
6	LOCATION	Text, 255	Description of the general location of the collision	Dropped	Insignificance
7	EXCEPTRSNCODE	Text, 10		Dropped	Insignificance
8	EXCEPTRSNDESC	Text, 300		Dropped	Insignificance
9	SEVERITYCODE	Text, 100	A code that corresponds to the severity of the collision: 3—fatality, 2b—serious injury, 2—injury, 1—prop damage, 0—unknown	Retained	Target variable
10	SEVERITYDESC	Text	A detailed description of the severity of the collision	Retained	Target variable
11	COLLISIONTYPE	Text, 300	Collision type	Retained	12% missing data, replaced by max. frequency
12	PERSONCOUNT	Double	The total number of people involved in the collision	Retained	
13	PEDCOUNT	Double	The number of pedestrians involved in the collision. This is entered by the state.	Retained	
14	PEDCYLCOUNT	Double	The number of bicycles involved in the collision. This is entered by the state.	Retained	
15	VEHCOUNT	Double	The number of vehicles involved in the collision. This is entered by the state.	Retained	
16	INJURIES	Double	The number of total injuries involved in the collision. This is entered by the state.	Retained	
17	SERIOUSINJURIES	Double	The number of serious injuries involved in the collision. This is entered by the state.	Retained	
18	FATALITIES	Double	The number of fatalities involved in the collision. This is entered by the state.	Retained	
19	INCDATE	Date	The date of the incident.	Dropped	Insignificance
20	INCDTTM	Text, 30	The date and time of the incident.	Dropped	Insignificance
21	JUNCTIONTYPE	Text, 300	Category of junction at which collision took place	Retained	5.5% missing data, replaced by max. frequency
22	SDOT_COLCODE	Text, 10	A code given to the collision by SDOT.	Dropped	Insignificance
23	SDOT_COLDESC	Text, 300	A description of the collision corresponding to the collision code.		
24	INATTENTIONIND	Text, 1	Whether or not collision was due to inattention (Y/N).	Dropped	86% data is missing
25	UNDERINFL	Text, 10	Whether or not a driver involved was under the influence of drugs or alcohol.	Dropped	only 4.5% observations are influencing
26	WEATHER	Text, 300	A description of the weather conditions during the time of the collision.	Retained	12% missing data, replaced by max. frequency
27	ROADCOND	Text, 300	The condition of the road during the collision.	Retained	12% missing data, replaced by max. frequency
28	LIGHTCOND	Text, 300	The light conditions during the collision.	Retained	12% missing data, replaced by max. frequency
29	PEDROWNOTGRNT	Text, 1	Whether or not the pedestrian right of way was not granted. (Y/N)	Dropped	97.7% data missing
30	SDOTCOLNUM	Text, 10	A number given to the collision by SDOT.	Dropped	Insignificance
31	SPEEDING	Text, 1	Whether or not speeding was a factor in the collision. (Y/N)	Dropped	only 4.5% observations are influencing, rest data unavailable
32	ST_COLCODE	Text, 10	A code provided by the state that describes the collision. For more information about these codes, please see the State Collision Code Dictionary.	Dropped	Insignificance
33	ST_COLDESC	Text, 300	A description that corresponds to the state's coding designation.	Retained	12% missing data, replaced by max. frequency
34	SEGLANEKEY	Long	A key for the lane segment in which the collision occurred.	Dropped	Insignificance
35	CROSSWALKKEY	Long	A key for the crosswalk at which the collision occurred.	Dropped	Insignificance
36	HITPARKEDCAR	Text, 1	Whether or not the collision involved hitting a parked car. (Y/N)	Retained	
37	STATUS	Text, 10	Matched, Unmatched	Dropped	Insignificance
38	REPORTNO	Long	Sr. No. of report for internal purposes	Dropped	Insignificance
39	COLDETKEY	Long	Secondary key for the incident	Dropped	Insignificance
40	INCKEY	Long	A unique key for the incident	Dropped	Insignificance

Data Preparation (Cleansing)

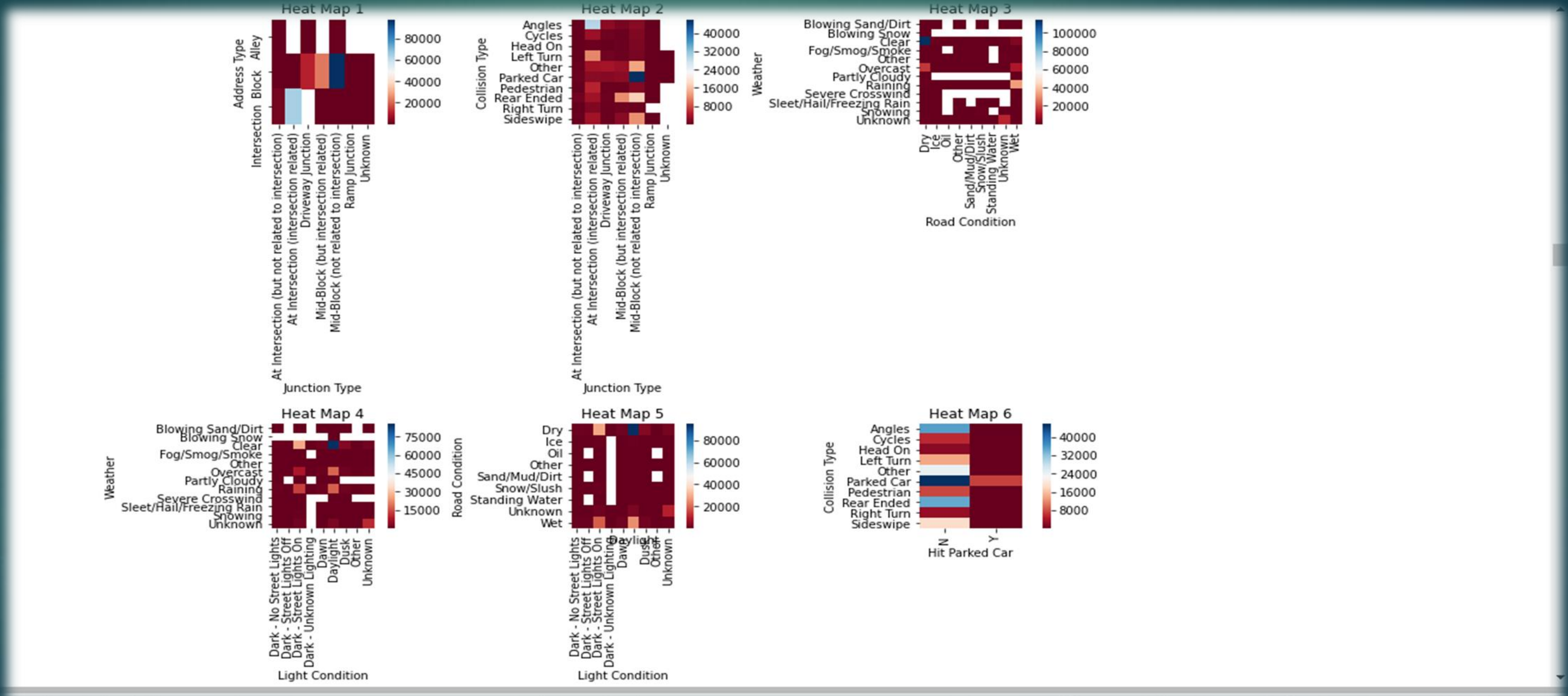
```
[11]:  1 df.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 199794 entries, 0 to 221388
Data columns (total 20 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   ADDRTYPE         199794 non-null  object
 1   SEVERITYCODE     199794 non-null  int64
 2   SEVERITYDESC     199794 non-null  object
 3   COLLISIONTYPE    199794 non-null  object
 4   PERSONCOUNT      199794 non-null  int64
 5   PEDCOUNT         199794 non-null  int64
 6   PEDCYLCOUNT      199794 non-null  int64
 7   VEHCOUNT         199794 non-null  int64
 8   INJURIES         199794 non-null  int64
 9   SERIOUSINJURIES  199794 non-null  int64
 10  FATALITIES       199794 non-null  int64
 11  JUNCTIONTYPE     199794 non-null  object
 12  SDOT_COLCODE     199794 non-null  float64
 13  SDOT_COLDESC     199794 non-null  object
 14  WEATHER          199794 non-null  object
 15  ROADCOND         199794 non-null  object
 16  LIGHTCOND        199794 non-null  object
 17  ST_COLCODE       199794 non-null  object
 18  ST_COLDESC       195157 non-null  object
 19  HITPARKEDCAR     199794 non-null  object
dtypes: float64(1), int64(8), object(11)
memory usage: 32.0+ MB
```

The next step in data cleansing would be to check and make sure that all data is in the correct format (int, float, text or other). To use categorical variables for regression analysis, indicator variables (or dummy variable) were used for transforming categorical variables into binary (0s and 1s) or numeric values.

Data Preparation (Correlations)

To get a better measure of the important characteristics, we looked at the correlation of attributes vis-a-vis target variable i.e. Accident Severity. The correlations are depicted by constructing heat maps between pairs of the variables.



Data Preparation (Important Variables)

Here we have a better idea of what our data looks like and which variables are important for consideration while predicting the 'Severity' class.

[19]:

	index	Pearson Correlation Coefficient	P-value
6	INJURIES	0.735846	0
7	SERIOUSINJURIES	0.468547	0

[20]:

```
df_corr.info(max_cols=130)
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 199794 entries, 0 to 221388
Data columns (total 123 columns):
 #   Column                                                    Non-Null Count   Dtype
---  ------                                                    --------------   -----
 0   SEVERITYCODE                                              199794 non-null  int64
 1   PERSONCOUNT                                               199794 non-null  int64
 2   PEDCOUNT                                                  199794 non-null  int64
 3   PEDCYLCOUNT                                               199794 non-null  int64
 4   VEHCOUNT                                                  199794 non-null  int64
 5   INJURIES                                                  199794 non-null  int64
 6   SERIOUSINJURIES                                           199794 non-null  int64
 7   FATALITIES                                                199794 non-null  int64
 8   ADDRTYPE_Alley                                            199794 non-null  uint8
 9   ADDRTYPE_Block                                            199794 non-null  uint8
 10  ADDRTYPE_Intersection                                     199794 non-null  uint8
 11  COLLISIONTYPE_Angles                                      199794 non-null  uint8
 12  COLLISIONTYPE_Cycles                                      199794 non-null  uint8
 13  COLLISIONTYPE_Head On                                     199794 non-null  uint8
 14  COLLISIONTYPE_Left Turn                                   199794 non-null  uint8
 15  COLLISIONTYPE_Other                                       199794 non-null  uint8
 16  COLLISIONTYPE_Parked Car                                  199794 non-null  uint8
 17  COLLISIONTYPE_Pedestrian                                  199794 non-null  uint8
 18  COLLISIONTYPE_Rear Ended                                  199794 non-null  uint8
 19  COLLISIONTYPE_Right Turn                                  199794 non-null  uint8
 111 ST_COLCODE_73                                             199794 non-null  uint8
 112 ST_COLCODE_74                                             199794 non-null  uint8
 113 ST_COLCODE_8                                              199794 non-null  uint8
 114 ST_COLCODE_81                                             199794 non-null  uint8
 115 ST_COLCODE_82                                             199794 non-null  uint8
 116 ST_COLCODE_83                                             199794 non-null  uint8
 117 ST_COLCODE_84                                             199794 non-null  uint8
 118 ST_COLCODE_85                                             199794 non-null  uint8
 119 ST_COLCODE_87                                             199794 non-null  uint8
 120 ST_COLCODE_88                                             199794 non-null  uint8
 121 HITPARKEDCAR_N                                            199794 non-null  uint8
 122 HITPARKEDCAR_Y                                            199794 non-null  uint8
dtypes: int64(8), uint8(115)
memory usage: 35.6 MB
```

Model Development

A Model would help us understand the exact relationship between different variables and how these variables are used to predict the result. We developed Classification model based on following algorithms that would predict the severity of an accident using the variables or features.

Logistic Regression:

It produces a formula that predicts the probability of a class label as function of the independent variables. Logistic regression fits a special s-shaped curve by transforming the numeric estimate into a probability with the sigmoid function σ.

K-Nearest Neighbors (KNN):

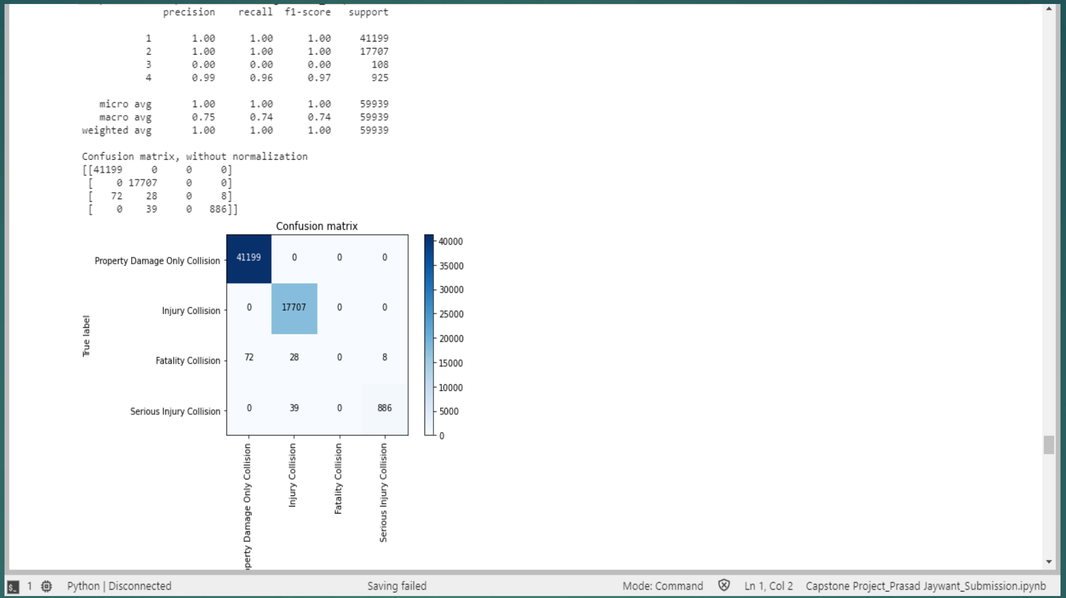
K-Nearest Neighbors is an algorithm for supervised learning, where the data is 'trained' with data points corresponding to their classification. Once a point is to be predicted, it considers the 'K' nearest points to it to determine its classification.

Decision Trees:

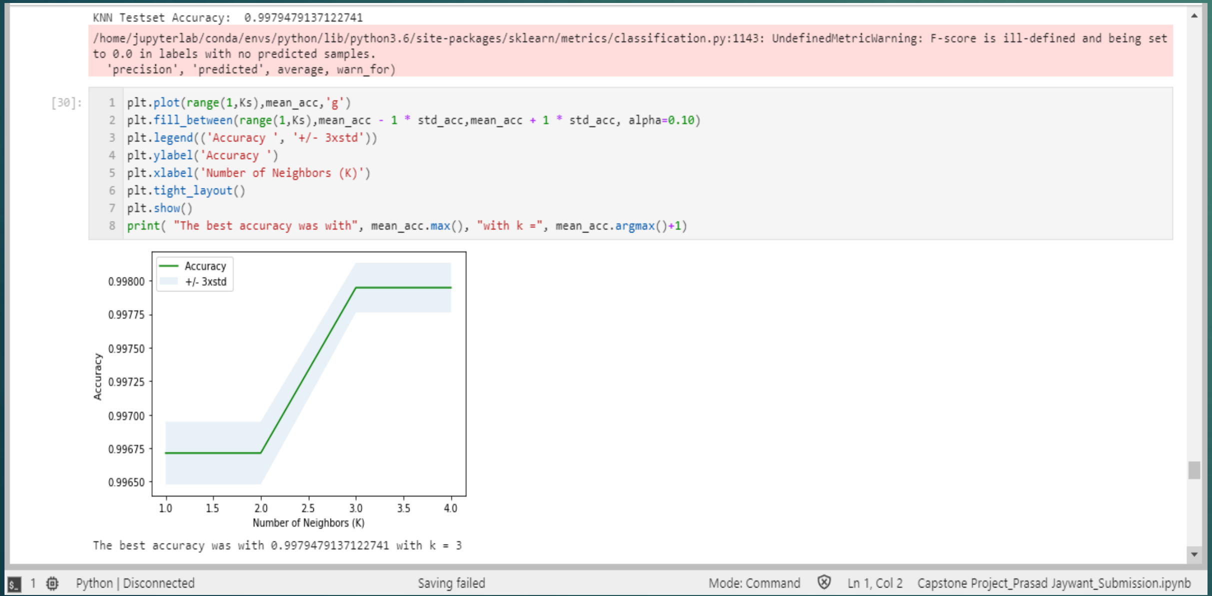
Based on the 'minimizing entropy (degree of randomness)' and 'maximising information gain (level of certainty)' criteria.

Model Development (Evaluation)

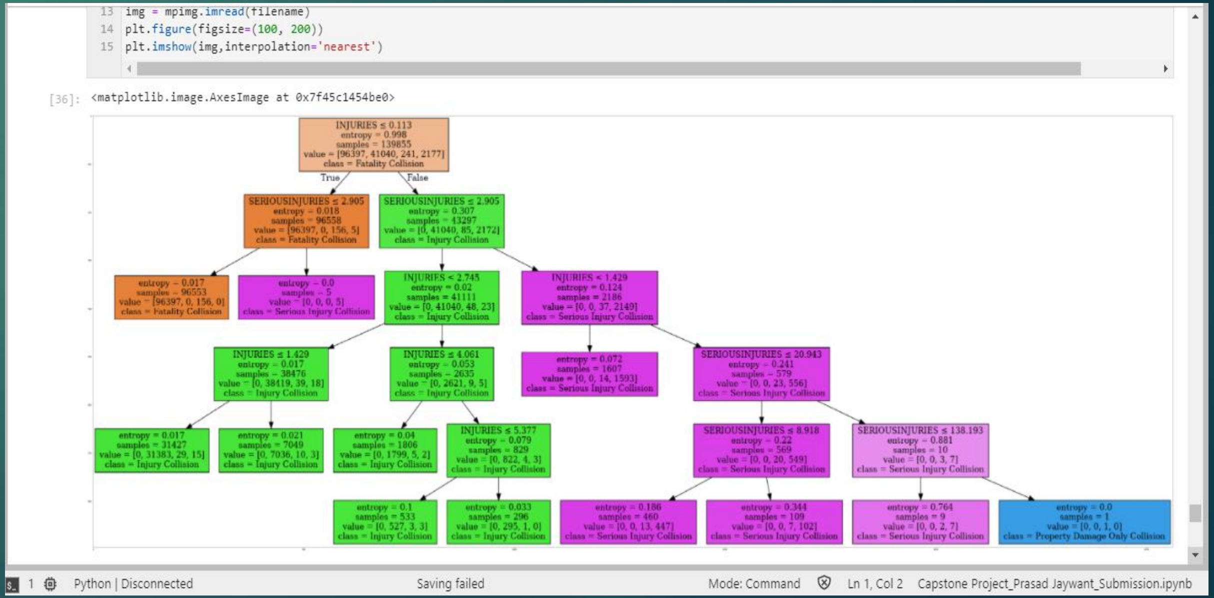
Logistic Regression



K-Nearest Neighbors



Decision Trees



Model Development (Results Summary)

The accuracy of the models built using different evaluation metrics can be summarized as follows;

Report

Here we report accuracy of the built model using different evaluation metrics:

[37]:

```
Report={'Algorithm':['KNN','Decision Tree','Logistic Regression'],'Jaccard':[knn_jac,dtree_jac,lr_jac],
       'F1-score':[knn_f1_score,dtree_f1_score,lr_f1_score],
       'LogLoss':['NA','NA',lr_log_loss]}
report_frame=pd.DataFrame(Report)
report_frame.set_index('Algorithm',inplace=True)
report_frame
```

[37]:

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.997948	0.997047	NA
Decision Tree	0.997948	0.997047	NA
Logistic Regression	0.997548	0.996641	0.0307091

1 Python | Disconnected Saving failed Mode: Command Ln 1, Col 2 Capstone Project_Prasad Jaywant_Submission.ipynb

Discussion

The data set is well structured and offers good number of useful observations (about 2L). The data wrangling was mostly accomplished by substituting the values with maximum frequency of the available data.

The correlation method shortlisted some variables such as injuries, which are related to the impact of accident, contributed moderately to the severity. Though the influence of causal factors such as weather, road/light conditions on accident severity was expected, they seemed not significant in contribution as was suggested by the low values (<0.4) of Pearson coefficients. Correlation of address type and junction type to severity was also not significantly evident.

We had split given data set into 70:30 ratio for training/testing the model. Model's prediction accuracy seems acceptable due to high Jaccard and F1-score and near-zero Log loss values.

Conclusion

The model has fairly taken care of the missing values which are of common occurrence in the real data gathering scenarios. The selected algorithms are in sync with the prediction accuracy, thereby poses high confidence in predicting the real cases. As envisioned in section 1.4, the model seems capable of implementing it at the client site. Also poses high potential for extending it to more city councils having similar data sources.

In the roadmap ahead, the model could be enriched with deeper analysis of causation factors, although the focus at present was more on the correlations within given data. The model deployment and integration with client systems could be the next steps of project implementation. With study of advanced Python capabilities, statistical/probabilistic algorithms and graphical visualizations, it could provide opportunity for iterative improvements in the model.

References

Preparation of this report must cite help of valuable references as follows;

- ✓ IBM Data Science Professional Certificate Course – All 9 modules, labs, tutorials and links therein: https://www.coursera.org/professional-certificates/ibm-data-science
- ✓ IBM Watson Studio Resources: https://cloud.ibm.com/resources
- ✓ Github Repository: https://github.com/
- ✓ Pandas open source literature: http://pandas.pydata.org
- ✓ Scikit Learn open source content: https://scikit-learn.org/
- ✓ Technology community sites: https://stackoverflow.com/ and many more from Google Search

Acknowledgement

I must thank all the mentor team members; Alex Aklson, Polong Lin, Romeo Kienzler, Svetlana Levitan, Joseph Santarcangelo, Hima Vasudevan, Rav Ahuja, Saeed Aghabozorgi for their valued guidance, exciting videos and structuring comprehensive labs and the tutorials. The illustrations, case studies showcased in these sessions helped me understand the concepts and practical applications of Data Science tools and techniques.

I also take this opportunity to thank my peers who took out their time to review my graded submissions and provided the valued feedbacks. I owe my special thanks to the contributors who are active in posting replies to various queries or issues being raised in the discussion forums. Their efforts in helping people to come out of the stuck up or no-clue situations are much appreciated.

Equally important, please convey my best regards to the Coursera organizing team for designing this unique course for the benefit of hundreds of thousands of such Data Science enthusiasts globally.

For any queries/suggestions, please revert to

ppjaywant@gmail.com

Thank You!