Data Science Capstone

Prediction of Accident Severity using Machine Learning Models

IBM – Coursera Project

By : Parthiv Lakshman

Introduction and Business understanding

- Each year traffic increases on roads
- Leads to more accidents
- Need for a system to predict accident severity based on external factors such as weather, lighting, etc.
- Machine learning models can help to accurately predict severity

Understanding the data

- Dataset of accident severity taken from SDOT Traffic management division in Seattle
- Classify accident severity in two levels –
  level 1 : property damage
  level 2 : Injury
- Goal : predict severity levels based on predictor variables such as road conditions, weather conditions, etc.

Original dataset

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGH
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Dayli
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark · Lights
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Dayli
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Dayli
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	...	Wet	Dayli

194673 rows, 38 columns

Data Preparation

- Cleaning of dataset
- Data wrangling
- Removing rows having no significant information
- Removing columns that do not influence severity
- Reducing dataset, renaming columns
- One-hot encoding : converting categorical values to numerical values
- Removing predictors that have less than 1% influence on outcome

Final dataset

	SC	Persons	Vehicles	Dark	Dawn	Daylight	Dusk	Dry	Ice	Snow/Slush	Wet	Clear	Fog/Smog/Smoke	Overcast	Raining	Snowing
0	2	2	2	0	0	1	0	0	0	0	1	0	0	1	0	0
1	1	2	2	1	0	0	0	0	0	0	1	0	0	0	1	0
2	1	4	3	0	0	1	0	1	0	0	0	0	0	1	0	0
3	1	3	3	0	0	1	0	1	0	0	0	1	0	0	0	0
4	2	2	2	0	0	1	0	0	0	0	1	0	0	0	1	0

Influence of predictors on Severity

```
SEVERITYCODE  LIGHTCOND
1             Daylight                      0.586028
              Dark - Street Lights On       0.257030
              Unknown                       0.097187
              Dusk                          0.029893
              Dawn                          0.012673
              Dark - No Street Lights       0.009086
              Dark - Street Lights Off      0.006669
              Other                         0.001382
              Dark - Unknown Lighting       0.000053
2             Daylight                      0.675050
              Dark - Street Lights On       0.253512
              Dusk                          0.034047
              Dawn                          0.014431
              Unknown                       0.010596
              Dark - No Street Lights       0.005850
              Dark - Street Lights Off      0.005534
              Other                         0.000911
              Dark - Unknown Lighting       0.000070
Name: LIGHTCOND, dtype: float64
```

Modeling – Classification Algorithms

Decision Tree Classifier :

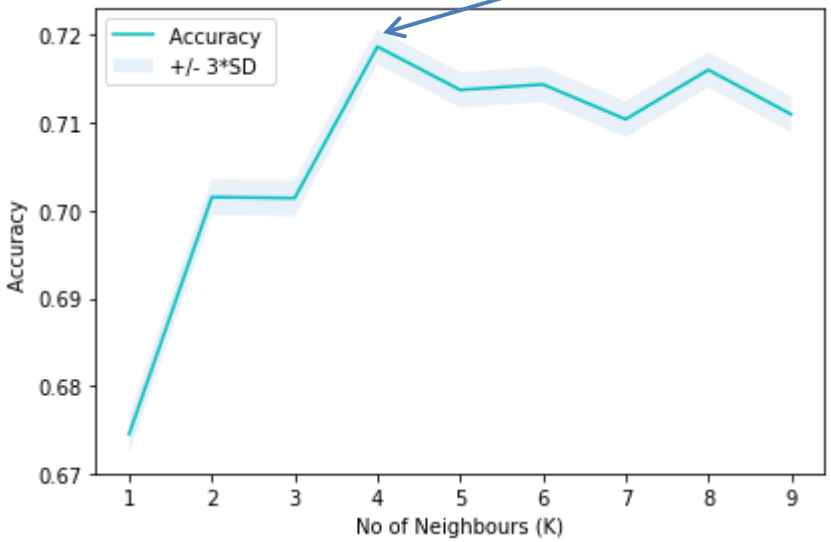
Built by splitting dataset into distinct nodes, where the nodes contain either categories or information regarding categories.

```
Initial classification :    [1 1 1 1 1]
Predicted classification :  [1 1 1 1 1]
Training set accuracy :  0.7219023443081811
Test set accuracy :  0.7273083862869695
```

K-Nearest Neighbors Classifier:

Takes a set of labeled points and uses the information to label nearby points.

First, the best value of k is determined to be 4, and then is used in the model.




```
Initial classification :    [1 1 1 1 1]
Predicted classification :  [1 2 1 2 1]
Training set accuracy :  0.715354420059007
Test set accuracy :  0.7186004550090217
```

Support Vector Machine (SVM) Classifier:

Classifies points based on a separator. Has the highest computation time.

```
Initial classification :    [1 1 1 1 1]
Predicted classification :  [1 1 1 1 1]
Training set accuracy :  0.7216417722263783
Test set accuracy :  0.7256413273711462
```

Logistic Regression Classifier:

Is a statistical machine learning algorithm, for classifying records of a dataset based on values of input field.

```
Initial classification :    [1 1 1 1 1]
Predicted classification :  [1 1 1 1 1]
Training set accuracy :  0.6740075145626172
Test set accuracy :  0.6801992625715855
```

Model Evaluation

Models evaluated based on Jaccard score, F1 score and Log Loss (for logistic regression). Comparison shown in table below.

Method	Jaccard Index	F1 Score	Log loss
Decision Tree	0.73	0.67	X
KNN	0.72	0.63	X
SVM	0.73	0.58	X
Logistic regression	0.68	0.67	23

As per Jaccard score – Decision tree and SVM are best
As per F1 score – Decision tree and logistic regression are best
Overall is Decision tree the best algorithm for this dataset.