IBM Data Science Certificate

- Capstone Project –

Predicting Severity of Car Accidents

Introduction

- The aim of the project is to predict the severity of car accidents by using data of former car accidents in the area of seattle
- The labeled data has been obtained from 2004 to present and is updated one a weekly basis
- Might be useful for emergency services such as paramedics or firefighters or insurances

The Data

- The data comprises of 194673 total entries and the number of categories or parameters is 37

- The severity of the car accidents is measured as a categorical number, which means a supervised classification model is used

- The categories include several information about the accident, like the location, persons and cars involved, injuries of persons, the type of collisions, the road, wheather or light conditions and time and dates

The Data – Used Models

- K-Nearest Neighbours
- Decision Tree
- Support-Vector Machine
- Logistic Regression

The Methodology

- Select the parameters for the model
- Fill missing entries
- One-hot-Encoding
- Final Feature Selection and Modeling

Parameter Selection

- Features = ['ADDRTYPE','SEVERITYDESC','COLLISIONTYPE','PERSONCOUNT','PEDCOUNT','PEDCYLCOUNT','VEHCOUNT','INCDTTM','JUNCTIONTYPE','INATTENTIONIND','UNDERINFL','WEATHER', 'LIGHTCOND', 'ROADCOND','PEDROWNOTGRNT','SPEEDING','SEGLANEKEY','CROSSWALKKEY','HITPARKEDCAR‘]
- Target = [‘SEVERITY‘]

See: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf.

Fill Missing Entries

- There are several features with a significant amount of missing values.

- Those empty entries are filled with illed with 'No', 'Unknown' and 'Other' values (depending on the feature). Unknown and Other are values are combined in each feature.

```
Features.isnull().sum()
```

```
ADDRTYPE           1926
SEVERITYDESC          0
COLLISIONTYPE      4904
PERSONCOUNT           0
PEDCOUNT              0
PEDCYLCOUNT           0
VEHCOUNT              0
JUNCTIONTYPE       6329
INATTENTIONIND    164868
UNDERINFL          4884
WEATHER            5081
LIGHTCOND          5170
ROADCOND           5012
PEDROWNOTGRNT    190006
SPEEDING         185340
SEGLANEKEY            0
CROSSWALKKEY          0
HITPARKEDCAR          0
WEEKDAY               0
HOURDAY               0
dtype: int64
```

One-Hot-Encoding

- Features are then one-hot-encoded, which results in a features array with 194673 rows and 55 features in total.

Feature Selection

- To reduce the number of features (preventing overfitting and reducing computation time), feature selection is applied using the Chi2-Score of each feature

- 20 remaining features

	Specs	Score
8	CROSSWALKKEY	3.186417e+09
7	SEGLANEKEY	8.638549e+07
15	Injury Collision	1.364850e+05
20	Parked Car	1.356930e+04
1	PEDCOUNT	1.248884e+04
21	Pedestrian	1.132928e+04
2	PEDCYLCOUNT	8.818806e+03
17	Cycles	8.608345e+03
5	PEDROWNOTGRNT	8.085270e+03
26	At Intersection (intersection related)	5.360734e+03
14	Intersection	5.137017e+03
29	Mid-Block (not related to intersection)	2.939348e+03
22	Rear Ended	2.811283e+03
0	PERSONCOUNT	2.473877e+03
24	Sideswipe	2.395238e+03
13	Block	2.312244e+03
9	HITPARKEDCAR	1.931147e+03
16	Angles	1.462505e+03
11	HOURDAY	9.587739e+02
46	Daylight	6.028858e+02

Model building

	Algorithm	Accuracy	Jaccard	F1-score
0	KNN	0.998973	0.998973	0.998973
1	Decision Tree	1.000000	1.000000	1.000000
2	SVM	0.999486	0.999486	0.999487
3	LogisticRegression	1.000000	1.000000	1.000000

- High Accuracy for each model

- Decision Tree and LogisticRegression perform best

Results, Discussion and Conclusion

- If the accident happend at a crosswalk, if parking cars were involved, if pedestrians or cyclist have been involved, the number of involved persons and injuries play an important role for the prediction models
- The trained models showed an excellent performance and can therefore be used for the desired applications
- However, one might get a similar result using less parameter and therefore less computation time
- Therefore, one should try to further reduce the parameterset.