IBM Data Science Professional Certificate Capstone Project – Car Accident Severity

Dave

August 2020

Introduction

- The Capstone Project of the IBM Data Science Specialization covers all the topics taught in the courses
- The project is based on a provided dataset of car accidents occurred in Seattle since 2004
- The data was recorded by Traffic Records and collected by the Seattle Police Department
- The dataset can be found here (link to metadata)
- The dataset includes attributes such as:
  - Severity
  - Location
  - Collision type
  - Number of injuries
  - Weather, road and light conditions, etc.

The Business Problem

- Not all accidents can be predicted
- Many uncontrollable factors exist in every accident:
  - Weather
  - Location
  - Time, etc.

The Business Problem

- However, manageable recorded factors may include:
  - Lighting – if many accidents occur in dark areas ➲ proper lighting should be installed
  - Pedestrians – if many pedestrians involved in some areas ➲ noticeable crosses are needed
  - Cyclists – if many cyclists involved in some areas ➲ bicycle lanes should be paved
  - Parked cars – if many accidents involve parked cars ➲ proper parking is needed
- The impact of one such factor can be huge and save lives
- The local authority may benefit a lot from the analysis
- A safer living space for the citizens can be provided

The Data

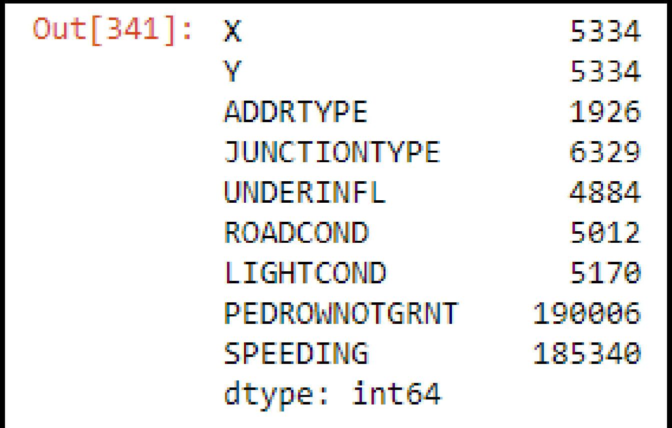
- The dataset gathers all collision events in Seattle since 2004
- The attributes in the dataset include, among others:
  - Severity of collision (damage level)
  - Collision type (head on, involved pedestrians or cyclists)
  - Time of accident date and time
  - Affected persons (also if cyclists, pedestrians or passengers were involved)
  - Involved parked cars
  - Address (alleys, blocks or intersections)
  - Weather, road and light conditions
- Overall there are 194673 accidents recorded in 38 attributes
- As mentioned before, this report will focus on the manageable factors

Data Preparation

- The first step is removing the irrelevant/uncontrollable attributes:
  - OBJECTID
  - INCKEY
  - COLDETKEY
  - REPORTNO
  - STATUS
  - INTKEY
  - LOCATION
  - EXCEPTRSNCODE
  - EXCEPTRSNDESC
  - SEVERITYCODE.1
  - SEVERITYDESC
  - COLLISIONTYPE
  - INCDTTM
  - SDOT_COLCODE
  - SDOT_COLDESC
  - INATTENTIONIND
  - WEATHER
  - SDOTCOLNUM
  - ST_COLCODE
  - ST_COLDESC
  - SEGLANEKEY
  - CROSSWALKKEY
- Now the dataset has 194673 accidents and 16 attributes

Data Preparation

- The second step is locating NaN cells:



```
Out[341]: X                   5334
          Y                   5334
          ADDRTYPE            1926
          JUNCTIONTYPE        6329
          UNDERINFL           4884
          ROADCOND            5012
          LIGHTCOND           5170
          PEDROWNOTGRNT     190006
          SPEEDING          185340
          dtype: int64
```

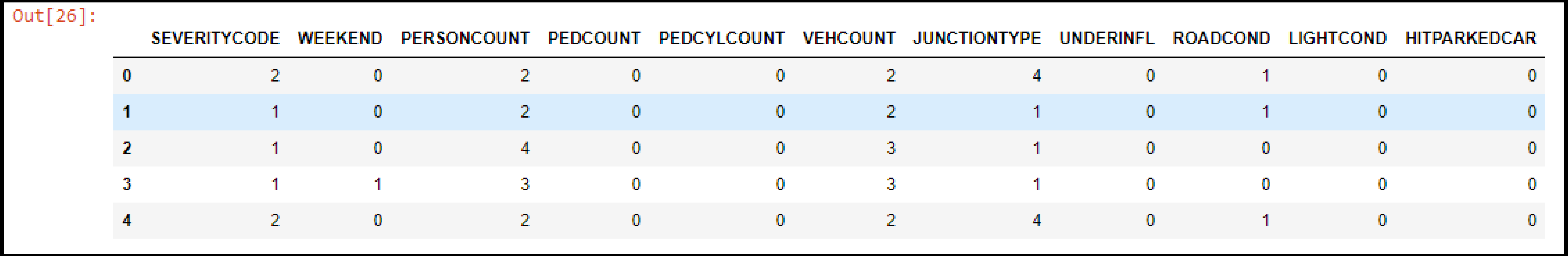
- PEDROWNOTGRNT and SPEEDING attributes are almost full with NaN cells, so they are dropped
- Also all rows with NaN cells in the following attributes are dropped:
    - ADDRTYPE
    - JUNCTIONTYPE
    - UNDERINFL
    - ROADCOND
    - LIGHTCOND

Data Preparation

- The third step is converting the data to numeric in the following attributes:
  - HITPARKEDCAR
  - UNDERINFL
  - ROADCOND (dry=0, all others=1)
  - LIGHTCOND (daylight=0, dark with street lights=0, all others=1)
  - JUNCTIONTYPE ("unknown" and "ramp" values are dropped, all others=1 to 5)
- After this step, ADDRTYPE attribute looked too similar to JUNCTIONTYP, therefore dropped

Data Preparation

- The fourth step is converting the INCDATE attribute data to binary
  - By asking if the day of the week is weekend or not
- And finally attributes X,Y and INCDATE are removed
- Now the dataset contains 168500 accidents in 11 attributes:

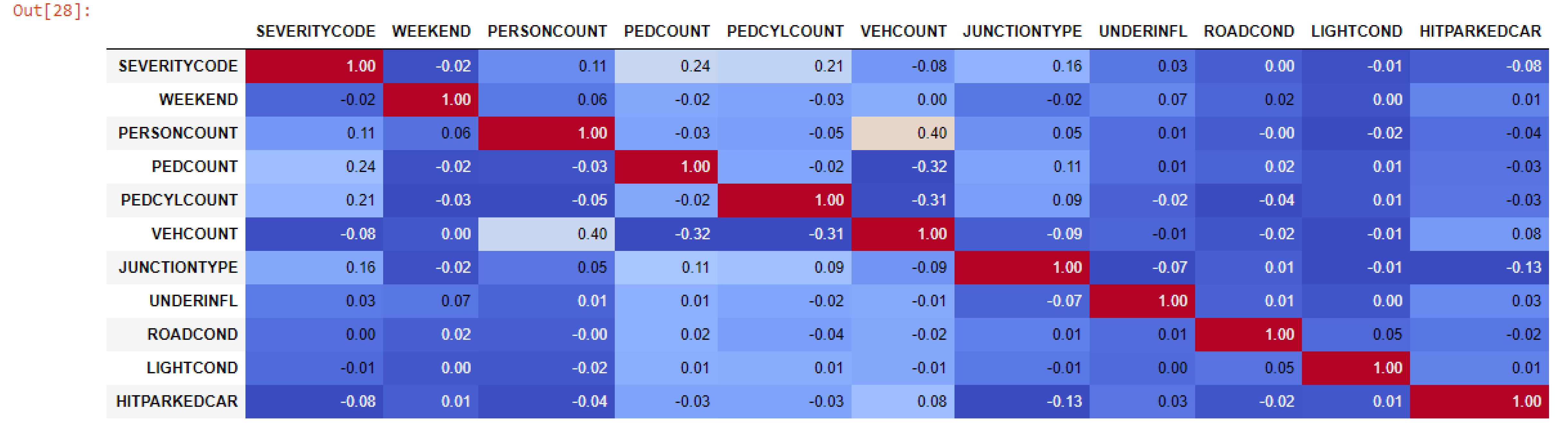


Out[26]:

	SEVERITYCODE	WEEKEND	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	JUNCTIONTYPE	UNDERINFL	ROADCOND	LIGHTCOND	HITPARKEDCAR
0	2	0	2	0	0	2	4	0	1	0	0
1	1	0	2	0	0	2	1	0	1	0	0
2	1	0	4	0	0	3	1	0	0	0	0
3	1	1	3	0	0	3	1	0	0	0	0
4	2	0	2	0	0	2	4	0	1	0	0

Initial Exploration

- An initial correlation map was built:



Out[28]:

	SEVERITYCODE	WEEKEND	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	JUNCTIONTYPE	UNDERINFL	ROADCOND	LIGHTCOND	HITPARKEDCAR
SEVERITYCODE	1.00	-0.02	0.11	0.24	0.21	-0.08	0.16	0.03	0.00	-0.01	-0.08
WEEKEND	-0.02	1.00	0.06	-0.02	-0.03	0.00	-0.02	0.07	0.02	0.00	0.01
PERSONCOUNT	0.11	0.06	1.00	-0.03	-0.05	0.40	0.05	0.01	-0.00	-0.02	-0.04
PEDCOUNT	0.24	-0.02	-0.03	1.00	-0.02	-0.32	0.11	0.01	0.02	0.01	-0.03
PEDCYLCOUNT	0.21	-0.03	-0.05	-0.02	1.00	-0.31	0.09	-0.02	-0.04	0.01	-0.03
VEHCOUNT	-0.08	0.00	0.40	-0.32	-0.31	1.00	-0.09	-0.01	-0.02	-0.01	0.08
JUNCTIONTYPE	0.16	-0.02	0.05	0.11	0.09	-0.09	1.00	-0.07	0.01	-0.01	-0.13
UNDERINFL	0.03	0.07	0.01	0.01	-0.02	-0.01	-0.07	1.00	0.01	0.00	0.03
ROADCOND	0.00	0.02	-0.00	0.02	-0.04	-0.02	0.01	0.01	1.00	0.05	-0.02
LIGHTCOND	-0.01	0.00	-0.02	0.01	0.01	-0.01	-0.01	0.00	0.05	1.00	0.01
HITPARKEDCAR	-0.08	0.01	-0.04	-0.03	-0.03	0.08	-0.13	0.03	-0.02	0.01	1.00

Initial Exploration

- The map shows very small correlation between severity and manageable attributes:
    - WEEKEND
    - JUNCTIONTYPE
    - UNDERINFL
    - ROADCOND
    - LIGHTCOND
    - HITPARKEDCAR

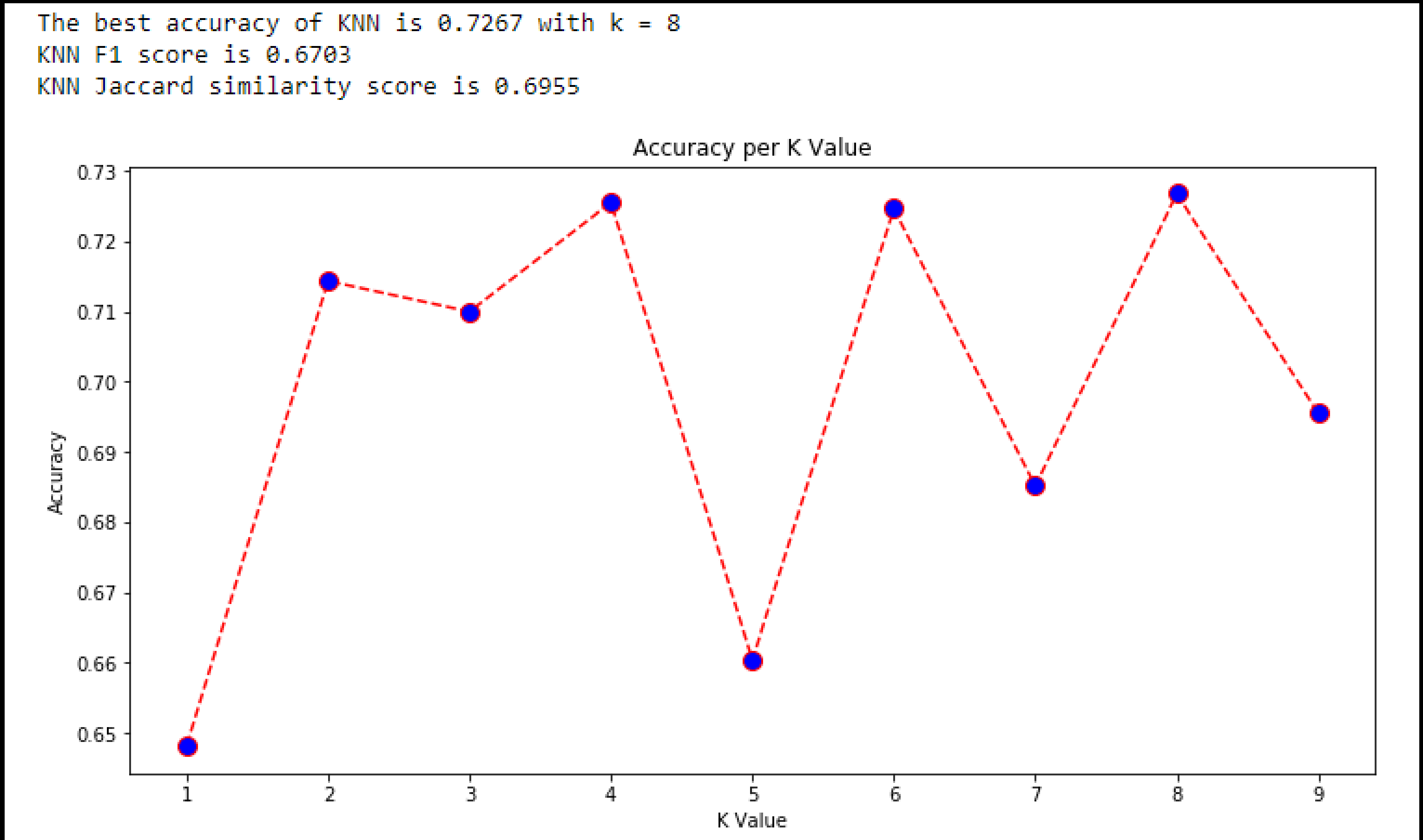
Further Exploration

- The following attributes were chosen for further exploration:
  - SEVERITYCODE
  - PEDCOUNT
  - PEDCYLCOUNT
  - JUNCTIONTYPE
  - VEHCOUNT
  - PERSONCOUNT
- The correlation in these attributes is > 0.15, slightly higher than others
- The further exploration will be performed using Machine Learning algorithms
  - KNN
  - Decision Tree
  - Logistic Regression

Further Exploration

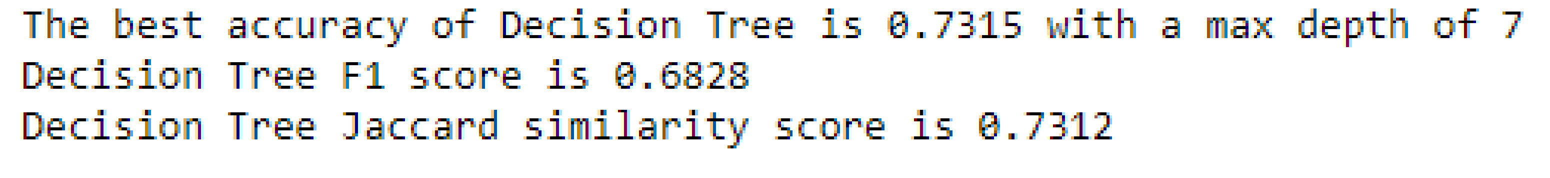
- The following attributes were chosen for further exploration:
  - PEDCOUNT
  - PEDCYLCOUNT
  - JUNCTIONTYPE
  - VEHCOUNT
  - PERSONCOUNT
- The correlation in these attributes is > 0.15, slightly higher than others
- The further exploration will be performed using Machine Learning algorithms
  - KNN
  - Decision Tree
  - Logistic Regression

KNN



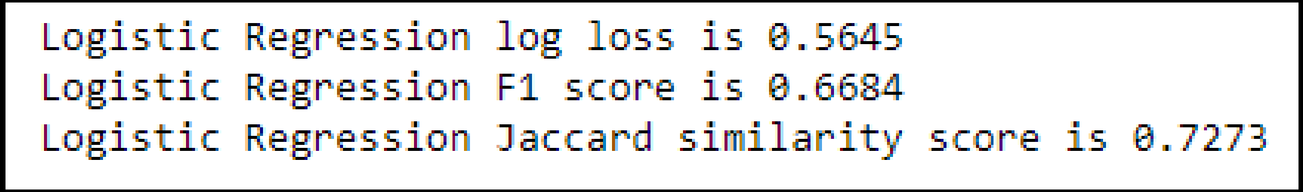
```
The best accuracy of KNN is 0.7267 with k = 8
KNN F1 score is 0.6703
KNN Jaccard similarity score is 0.6955
```

Decision Tree



```
The best accuracy of Decision Tree is 0.7315 with a max depth of 7
Decision Tree F1 score is 0.6828
Decision Tree Jaccard similarity score is 0.7312
```

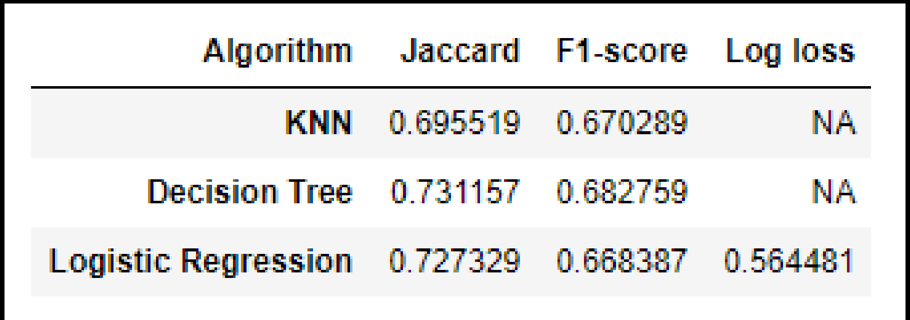
Logistic Regression



```
Logistic Regression log loss is 0.5645
Logistic Regression F1 score is 0.6684
Logistic Regression Jaccard similarity score is 0.7273
```

Results

- The Decision Tree algorithm is the most accurate in this case:



Algorithm	Jaccard	F1-score	Log loss
KNN	0.695519	0.670289	NA
Decision Tree	0.731157	0.682759	NA
Logistic Regression	0.727329	0.668387	0.564481

Discussion

- In contrast to the initial goal, the dataset shows that there is no correlation between severity and manageable attributes:
  - WEEKEND
  - UNDERINFL
  - ROADCOND
  - LIGHTCOND
  - HITPARKEDCAR
- However, it is highly correlated with logical attributes, such as:
  - PEDCOUNT
  - PEDCYLCOUNT
  - JUNCTIONTYPE
  - VEHCOUNT
  - PERSONCOUNT

Conclusions

- Despite the large amount of collected data, the Seattle car accident dataset analysis can’t predict accidents based on authority-manageable features
- However, observing the dataset, some steps can be taken to reduce the accident rate:
    - Water draining on the roads
    - Deicing
    - Installing street lights
    - Performing more alcohol tests