Data Science Course Capstone Project - Report

The present Jupyter Notebook implements assignments of the course "Applied Data Science Capstone", the last of the nine courses of the IBM Professional Certification "Data Science".

Car Accident Severity Prediction Based on Contextual Conditions

Table of contents

Introduction

Business Understanding

The goal of the project is to build a predictor capable of predicting the severity of a road accident given traffic, weather and other environmental conditions. The severity is to be predicted in terms of property damage only or some type of bodily injury event in case of accident. The purpose of the predictor is to help travelers to judge if the conditions they are currently encountering during their trip are a known factor relevant for serious consequences in case of accident or not.

The idea for reaching the goal is to use the accident dataset maintained and made publicly available by the Transportation Department of the City of Seattle. This accident dataset can be retrieved in CSV format at the address https://data-seattlecitygis.opendata.arcgis.com/datasets/collisions. Such dataset will be use to build and validate a predictor using the Machine Learning techniques presented in the courses of the "IBM Data Science Professional Certificate" series on Coursera.

Data

Data Understanding

Following a first analysis of the accident dataset provided by the City of Seattle Transportation Department, the following columns are found to be representative of the conditions that are felt to be most relevant to affect the severity of an accident:

- LOCATION - Location of the accident
- JUNCTIONTYPE - Type of of road junction (mid-block with or without junction, intersection, driveway, ramp, etc.)
- WEATHER - Weather condition (Overcast, Raining, Clear, Snowing, etc.)

- ROADCOND - Road condition (Wet, Dry, Snow/Slush, Ice, etc.)
- LIGHTCOND - Lighting conditions (Daylight, Dark - Street Lights On, Dark - No Street Lights, etc.)
- SPEEDING - Speeding vehicles (Yes or No)
- VEHCOUNT - Number of vehicles involved
- PERSONCOUNT - Number of persons involved

The target field that the predictor will need to predict is "SEVERITYCODE", with the values 2 and 1 for 'Injury Collision' and 'Property Damage Only Collision'.

Based on the above column selection, a first analysis of the raw data provides the following results:

```
Dataset shape: (194673, 38)
SEVERITYCODE unique values: [2 1]
SEVERITYDESC unique values: ['Injury Collision' 'Property Damage Only
Collision']
JUNCTIONTYPE unique values: ['At Intersection (intersection related)'
 'Mid-Block (not related to intersection)' 'Driveway Junction'
 'Mid-Block (but intersection related)'
 'At Intersection (but not related to intersection)' nan 'Unknown'
 'Ramp Junction']
WEATHER unique values: ['Overcast' 'Raining' 'Clear' nan 'Unknown' 'Other'
'Snowing'
 'Fog/Smog/Smoke' 'Sleet/Hail/Freezing Rain' 'Blowing Sand/Dirt'
 'Severe Crosswind' 'Partly Cloudy']
ROADCOND unique values: ['Wet' 'Dry' nan 'Unknown' 'Snow/Slush' 'Ice' 'Other'
'Sand/Mud/Dirt'
 'Standing Water' 'Oil']
LIGHTCOND unique values: ['Daylight' 'Dark - Street Lights On' 'Dark - No Street
Lights' nan
 'Unknown' 'Dusk' 'Dawn' 'Dark - Street Lights Off' 'Other'
 'Dark - Unknown Lighting']
SPEEDING unique values: [nan 'Y']

Number of damage-only incidents: 136485
Number of bodily injury incidents: 58188

Number of unique LOCATION: 24103

Most typical locations for damage-only incidents: LOCATION
BATTERY ST TUNNEL NB BETWEEN ALASKAN WY VI NB AND AURORA AVE N
198
BATTERY ST TUNNEL SB BETWEEN AURORA AVE N AND ALASKAN WY VI SB
185
ALASKAN WY VI NB BETWEEN S ROYAL BROUGHAM WAY ON RP AND SENECA ST OFF RP
175
N NORTHGATE WAY BETWEEN MERIDIAN AVE N AND CORLISS AVE N
171
AURORA AVE N BETWEEN N 130TH ST AND N 135TH ST
151
6TH AVE AND JAMES ST
145
ALASKAN WY VI SB BETWEEN COLUMBIA ST ON RP AND ALASKAN WY VI SB EFR OFF RP
144
1ST AVE BETWEEN BLANCHARD ST AND BELL ST
140
RAINIER AVE S BETWEEN S BAYVIEW ST AND S MCCLELLAN ST
137
WEST SEATTLE BR EB BETWEEN ALASKAN WY VI NB ON RP AND DELRIDGE-W SEATTLE BR EB
ON RP    134
```

```
AURORA AVE N BETWEEN N 117TH PL AND N 125TH ST
134
AURORA BR BETWEEN RAYE ST AND BRIDGE WAY N
129
ALASKAN WY VI NB BETWEEN SENECA ST OFF RP AND WESTERN AV OFF RP
109
1ST AVE BETWEEN UNION ST AND PIKE ST
108
RAINIER AVE S BETWEEN S DEARBORN ST AND S CHARLES N ST
102
5TH AVE AND VIRGINIA ST
101
RAINIER AVE S BETWEEN S HENDERSON ST AND S DIRECTOR N ST
101
dtype: int64

Most typical locations for bodily injury incidents: LOCATION
AURORA AVE N BETWEEN N 117TH PL AND N 125TH ST    120
6TH AVE AND JAMES ST                              107
dtype: int64

Number of incidents by JUNCTIONTYPE value: JUNCTIONTYPE
Mid-Block (not related to intersection)              89800
At Intersection (intersection related)               62810
Mid-Block (but intersection related)                 22790
Driveway Junction                                    10671
At Intersection (but not related to intersection)     2098
Ramp Junction                                          166
Unknown                                                  9
dtype: int64

Number of incidents by WEATHER value: WEATHER
Clear                       111135
Raining                      33145
Overcast                     27714
Unknown                      15091
Snowing                        907
Other                          832
Fog/Smog/Smoke                 569
Sleet/Hail/Freezing Rain       113
Blowing Sand/Dirt               56
Severe Crosswind                25
Partly Cloudy                    5
dtype: int64

Number of incidents by ROADCOND value: ROADCOND
Dry               124510
Wet                47474
Unknown            15078
Ice                 1209
Snow/Slush          1004
Other                132
Standing Water       115
Sand/Mud/Dirt         75
Oil                   64
dtype: int64

Number of incidents by LIGHTCOND value: LIGHTCOND
Daylight                    116137
Dark - Street Lights On      48507
Unknown                      13473
Dusk                          5902
Dawn                          2502
```

```
Dark - No Street Lights       1537
Dark - Street Lights Off      1199
Other                          235
Dark - Unknown Lighting         11
dtype: int64

Number of incidents by SPEEDING value: SPEEDING
Y    9333
dtype: int64
```

Data Preparation and Cleaning

The above results call for a clean-up of the dataset consisting of a number of adjustments as follows:

- Transformation of SPEEDING column "NaN" values into "0" for "No" and of "Y" values into "1" for "Yes"
- Elimination of all records containing "NaN" values or other undefined column values like "Unknown"
- Substitution of the LOCATION identifier column with the probability of a bodily injury in case of an incident at that location based on the content of the dataset, as it is felt that such indication can help the predictor in identifying the severity of the incident at a certain location much more than a categorization of the very numerous locations (over 2000) mentioned in the dataset.

The dataset resulting from the above adjustments is as follows:

	SEVERITYCODE	JUNCTIONTYPE	WEATHER	ROADCOND	LIGHTCOND	SPEEDING	VEHCOUNT	PERSONCOUNT	LOCINJRISK
0	2	At Intersection (intersection related)	Overcast	Wet	Daylight	0	2	2	0.483871
1	1	At Intersection (intersection related)	Overcast	Dry	Dark - Street Lights On	0	3	4	0.483871
2	2	At Intersection (intersection related)	Clear	Dry	Daylight	0	3	5	0.483871
3	2	At Intersection (intersection related)	Overcast	Wet	Daylight	0	2	2	0.483871
4	1	At Intersection (intersection related)	Overcast	Wet	Daylight	0	2	2	0.483871
...	...	...	...	...	...	...	...	...	...
169282	2	Mid-Block (not related to	Clear	Dry	Dusk	0	2	2	0.000000

	SEVERITYCODE	JUNCTIONTYPE	WEATHER	ROADCOND	LIGHTCOND	SPEEDING	VEHCOUNT	PERSONCOUNT	LOCINJRISK
		intersection)							
169283	1	Mid-Block (not related to intersection)	Raining	Wet	Dark - Street Lights On	0	2	2	0.000000
169284	2	Driveway Junction	Clear	Dry	Daylight	0	2	2	0.000000
169285	1	At Intersection (intersection related)	Clear	Dry	Daylight	0	2	2	0.000000
169286	1	Mid-Block (not related to intersection)	Raining	Wet	Dark - Street Lights On	0	1	1	0.000000

169287 rows × 10 columns

Data Pre-Processing

A final adjustment before building the predictor is to transform the categorical features selected for predicting the incident severity into numerical values that can be used by the Machine Learning techniques that will be used to develop the predictor.

```
JUNCTIONTYPE unique values: ['At Intersection (intersection related)'
 'At Intersection (but not related to intersection)'
 'Mid-Block (not related to intersection)'
 'Mid-Block (but intersection related)' 'Driveway Junction'
 'Ramp Junction']
WEATHER unique values: ['Overcast' 'Clear' 'Raining' 'Blowing Sand/Dirt'
'Snowing'
 'Fog/Smog/Smoke' 'Sleet/Hail/Freezing Rain' 'Severe Crosswind'
 'Partly Cloudy']
ROADCOND unique values: ['Wet' 'Dry' 'Ice' 'Snow/Slush' 'Standing Water' 'Oil'
'Sand/Mud/Dirt']
LIGHTCOND unique values: ['Daylight' 'Dark - Street Lights On' 'Dusk' 'Unknown'
'Dawn'
 'Dark - Street Lights Off' 'Dark - No Street Lights'
 'Dark - Unknown Lighting']
```

Following the above categorization the data set results to be as follows:

```
X, the Feature Matrix (data): [[1 3 6 ... 2 2 0.4838709677419355]
 [1 3 0 ... 3 4 0.4838709677419355]
 [1 1 0 ... 3 5 0.4838709677419355]
 ...
 [2 1 0 ... 2 2 0.0]
 [1 1 0 ... 2 2 0.0]
 [4 5 6 ... 1 1 0.0]]

y, the response vector (target): 0         2
1         1
2         2
3         2
4         1
         ..
169282    2
```

```
169283    1
169284    2
169285    1
169286    1
Name: SEVERITYCODE, Length: 169287, dtype: int64
```

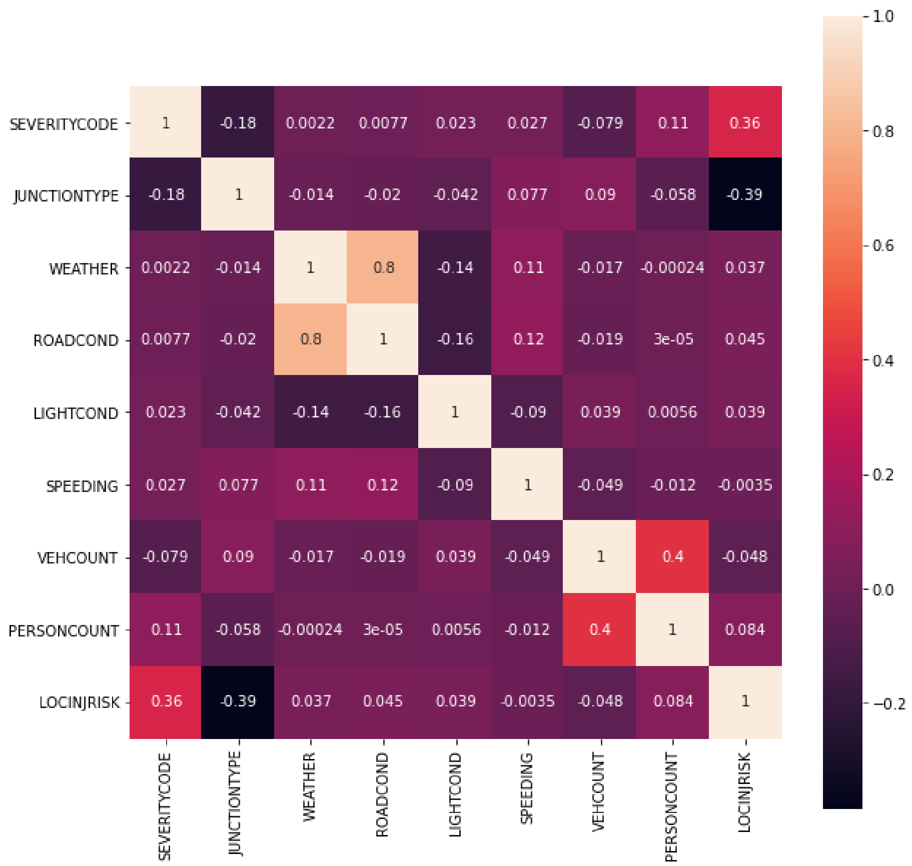
The final step of the pre-processing is the separation of the data set into a train set for the development of the predictor and a test set for its validation.

```
Train set: (135429, 8) (135429,)
Test set: (33858, 8) (33858,)
```

Methodology

In the present section various predictors are built using the train data set according to the Machine Learning techniques presented during the specialization courses series. The predictors are then validated using the test set previously prepared and a table is finally presented to summarize the performance of the various predictors and indicate the one that provides the best accuracy.

Before building and evaluating the various predictors, it is useful to evaluate the level of correlation in the available dataset to assess the chances that we will have to build solid predictors for identifying the incident severity in terms of damage-only or bodily injury in case of incident given the incident features identified. Such an assessment is possible by building a "heat-map" to represent the mutual correlation between the features and the target. Analyzing the prepared data the heat-map is as follows:



It can be seen that in general the level of mutual correlation between the various features and the target is very poor. Most of the correlations are very close to zero with only few unsurprising exceptions:

- The strongest correlation (value 0.8) appears to be between the weather (WEATHER) and the road condition (ROADCOND) features. This is typically related to the fact that a rainy weather is connected to a wet road, clear weather is connected to a dry road, etc.
- A second significant correlation (value 0.8) is present between the number of vehicles (VEHCOUNT) and the number of people (PERSONCOUNT) involved in the incident. Also this correlation is unsurprising as it is to be expected that with more vehicles are involved also the number of people will inevitably grow and vice-versa.
- The third significant correlation (value 0.36) is between the target feature severity code (SEVERITYCODE) and the newly added feature representing the likeliness of a bodily injury vs. a damage-only at the incident location (LOCINJRISK), which was calculated according to the severity of the incidents recorded in the dataset for the various locations.

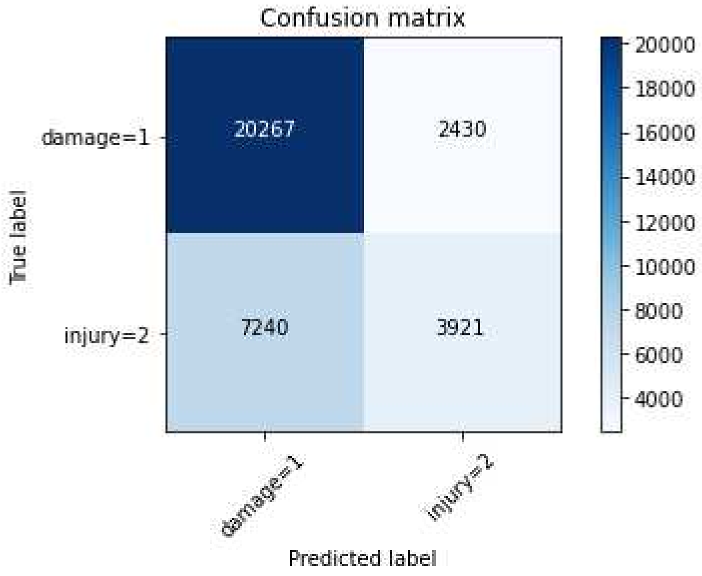
Being connected to the target feature, the third correlation appears to be the most useful one for helping the prediction of the incident severity. Due to the absence of further helpful correlations it is to be expected that it will be difficult to build predictors with a good level of accuracy. Nevertheless let us build and compare the performance of possible predictors built using the following Machine Learning techniques:

- Logistic Regression
- Support Vector Machine (SVM)
- K-Nearest Neighbor (KNN)
- Decision Tree

Using each single of the above techniques and the Python library "Scikit Learn", a predictor will be built using the previously prepared train data set and the accuracy will be assessed using the previously prepared test data set. The resulting prediction accuracies will be then gathered in a single table, which will be commented in the "Results" section of the present report.

Modeling with Logistic Regression

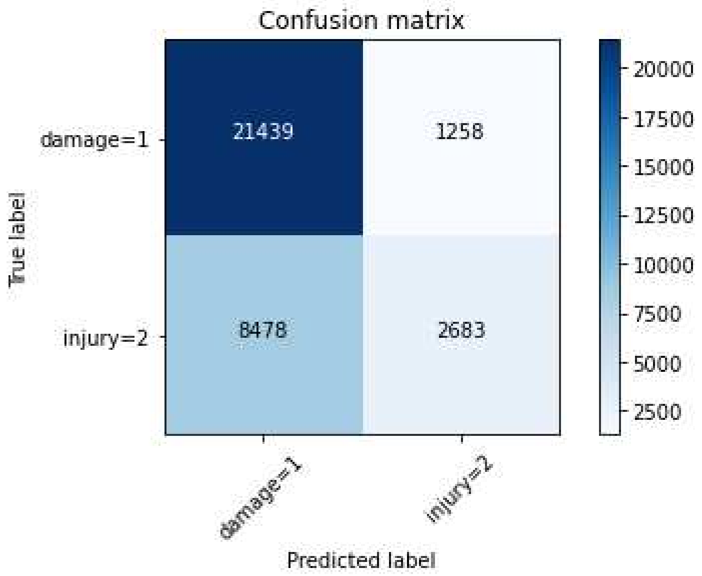
```
LogisticRegression(C=0.01, solver='liblinear')
```




```
              precision    recall  f1-score   support

           1       0.74      0.89      0.81     22697
           2       0.62      0.35      0.45     11161

    accuracy                           0.71     33858
   macro avg       0.68      0.62      0.63     33858
weighted avg       0.70      0.71      0.69     33858

Jaccard score: 0.6769883421852557
F1-score: 0.8073858656680741
Log-loss score: 0.5546087301960763
```

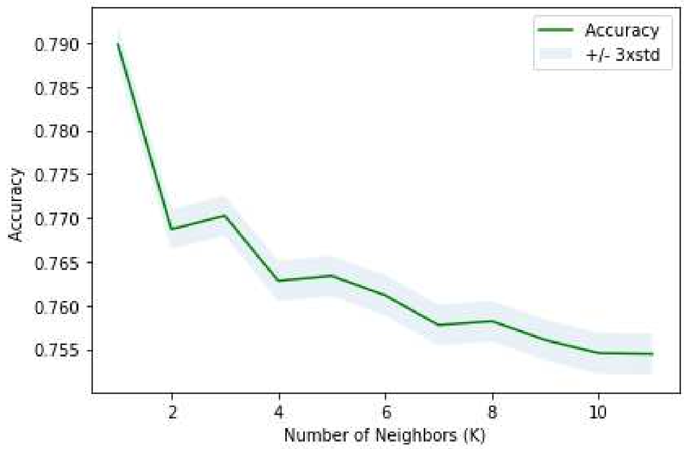
Modeling with SVM

```
SVC(kernel='linear', probability=True)
```



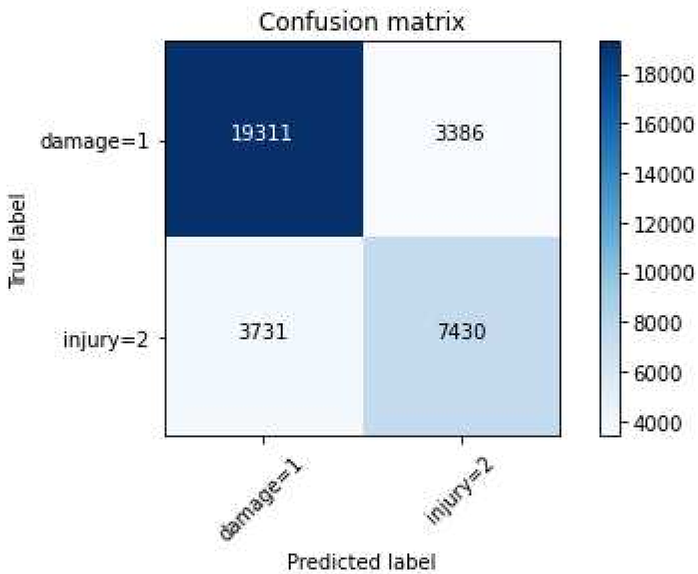

```
              precision    recall  f1-score   support

           1       0.72      0.94      0.81     22697
           2       0.68      0.24      0.36     11161

    accuracy                           0.71     33858
   macro avg       0.70      0.59      0.59     33858
weighted avg       0.70      0.71      0.66     33858

Jaccard score: 0.6876984763432238
F1-score: 0.8149541947010301
Log-loss score: 0.5613624392049108
```

Modeling with KNN




```
Best Accuracy is 0.7897985705003249 for k = 1
```

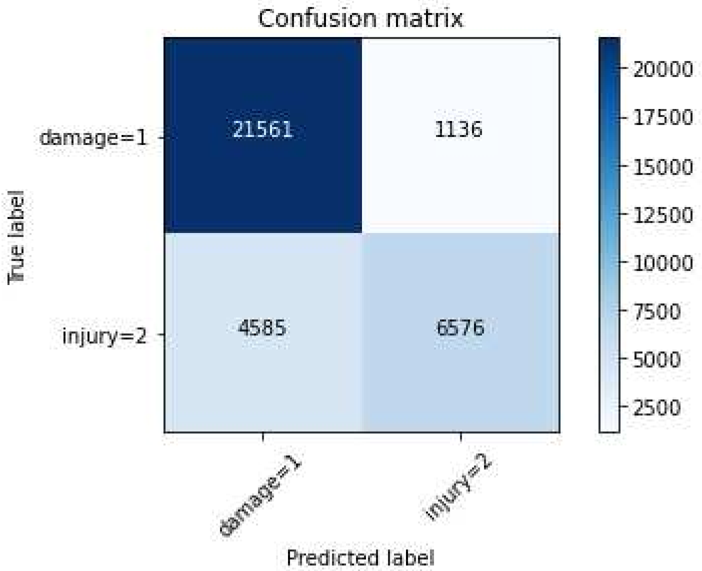
```
KNeighborsClassifier(n_neighbors=1)
```




```
              precision    recall  f1-score   support

           1       0.84      0.85      0.84     22697
           2       0.69      0.67      0.68     11161

    accuracy                           0.79     33858
   macro avg       0.76      0.76      0.76     33858
weighted avg       0.79      0.79      0.79     33858

Jaccard score: 0.7307022854548206
F1-score: 0.8443997463871095
Log-loss score: 7.260100171379862
```

Modeling with Decision Tree

```
DecisionTreeClassifier(criterion='entropy', max_depth=30)
```




```
              precision    recall  f1-score   support

           1       0.82      0.95      0.88     22697
           2       0.85      0.59      0.70     11161

    accuracy                           0.83     33858
   macro avg       0.84      0.77      0.79     33858
weighted avg       0.83      0.83      0.82     33858

Jaccard score: 0.7057160587464663
F1-score: 0.8274719055703776
Log-loss score: 0.5018186555575692
```

Results

The following table summarizes the weighed accuracies on precision and recall plus F1 scores achieved by the various predictors on the test data set:

Predictor Type	Weighed Avg. Precision	Weighed Avg. Recall	Weighed Avg. F1 Score	Notes
Logistic Regression	0.70	0.71	0.69	
Support Vector Regression	0.70	0.71	0.66	
K-Nearest Neighbor	0.79	0.79	0.79	
Decision Tree	**0.83**	**0.83**	**0.82**	**Best performer!**

Discussion

Evaluation

It must be noted that the performances achieved by the various predictors were in general surprisingly good despite the heat-map of the correlations between the features showed quite poor and unpromising values. The Decision Tree predictor clearly proved to be the most accurate in this case. This result was obtained building a quite deep tree, which in any case resulted in a quite short training time, actually much shorter than the ones that were required for training the SVM and KNN predictors.

Conclusion

The project consisted in building a successful predictor to predict the severity in terms of "damage-only" or "bodily injury" of the outcome of a car incident occurring in the Seattle city area given a number of environmental conditions consisting of location, road condition, weather condition, road junction type, speeding, number of people involved, light conditions, number of vehicles involved.

The data available for building the predictor consisted in a CSV file maintained by the Department of Traffic of Seattle consisting of 169287 records. Such data set was divided into two distinct sets of 135429 records for training candidate predictors and 33858 records for testing them.

In order to find a good solution, multiple predictors have been built using the different Machine Learning techniques suitable for binary predictions presented during the specialization courses. In particular four predictors have been built based on Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Decision Tree.

The comparison of the performances of the four predictors on the test set indicated that the Decision Tree predictor was the best performer in terms of accuracy.