Capstone Project

Accident Severity Predictions

Prepared by:
Pradiv Gnanaraj

Analysis | Machine Learning

Agenda

- Business Understanding
- Data Understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment

What could be the problem?

Case Study : Predict the Severity of Accident

- Prevention is better than cure.
  A Model to alert the travelers.
- Predicting accident severity.
  Based on the previous data collected from the accident report
  To alert the traveler prior travel

Various factors :

- Weather and Light Conditions
- Road type and Speed

Data Understanding

Content

Dataset has been fetched from UK open data and the files have been merged and cleaned to reach the final data attached.

Primarily Captures Road Accidents in UK between 1979 and 2015 and has 70 features/columns and about 250K rows.

Also attached with it is an excel file with Multiple Tabs that can help one to understand the Data.

Acknowledgements and Data Source

Data has been fetched from Open Data Platform UK and is being shared under Open Government Licence.

For more details refer to Open Data UK

The data set was uploaded to Kaggle.

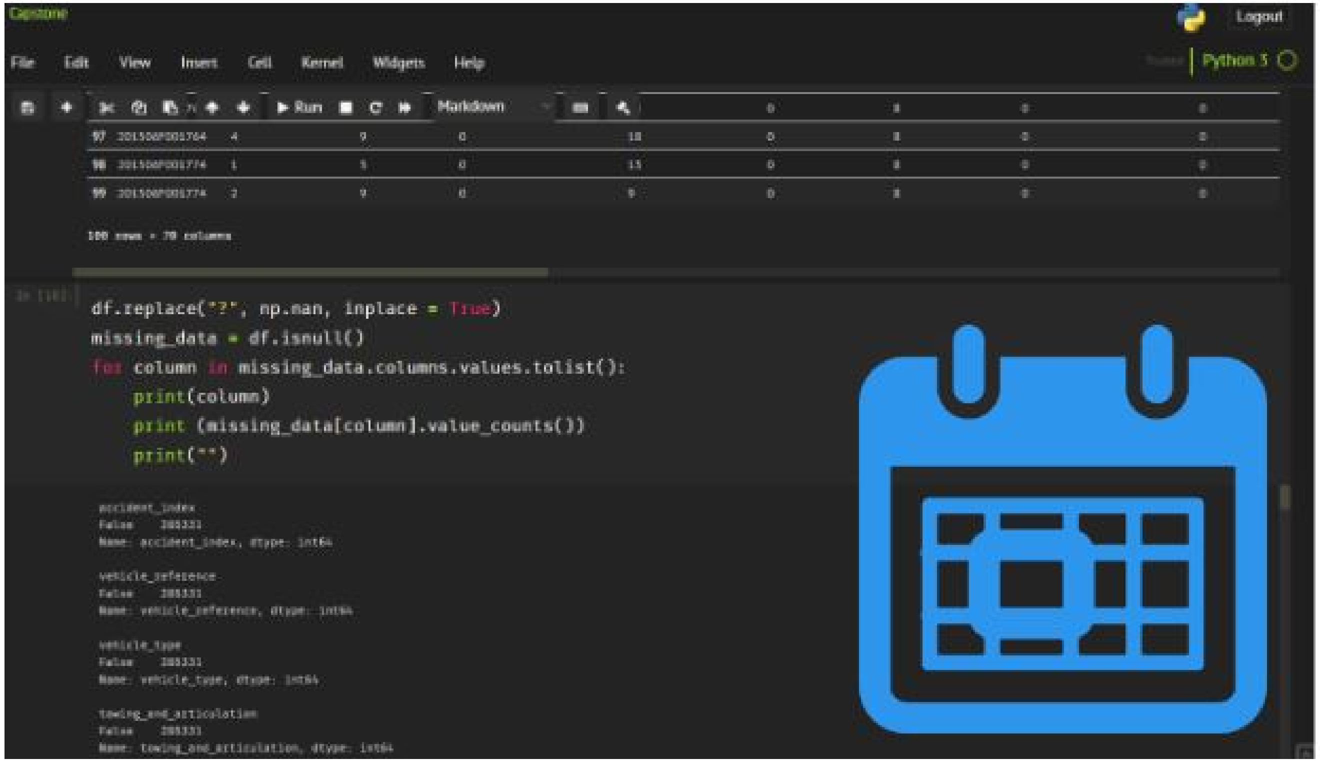
<link will be shared during the final submission>

Data Preparation

Data preparation can be performed multiple times and it includes balancing the labeled data, transformation, filling missing data, and cleaning the dataset.

Python Loop "True" represents a missing value, "False" means the value is present in the dataset.

Each column has 285331 rows of data with 22 columns containing missing data.



```
97  201506GF001764  4   9   0   10   0   8   0   0
98  201506GF001774  1   5   0   15   0   8   0   0
99  201506GF001774  2   9   0   9    0   8   0   0

100 rows × 70 columns
```

```
df.replace("?", np.man, inplace = True)
missing_data = df.isnull()
for column in missing_data.columns.values.tolist():
    print(column)
    print (missing_data[column].value_counts())
    print("")
```

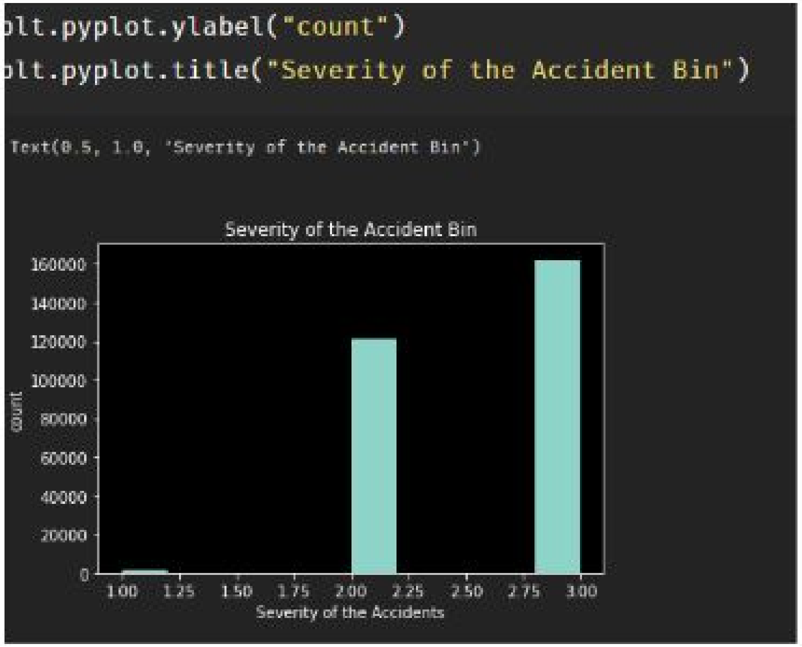
```
accident_index
False    285331
Name: accident_index, dtype: int64

vehicle_reference
False    285331
Name: vehicle_reference, dtype: int64

vehicle_type
False    285331
Name: vehicle_type, dtype: int64

towing_and_articulation
False    285331
Name: towing_and_articulation, dtype: int64
```

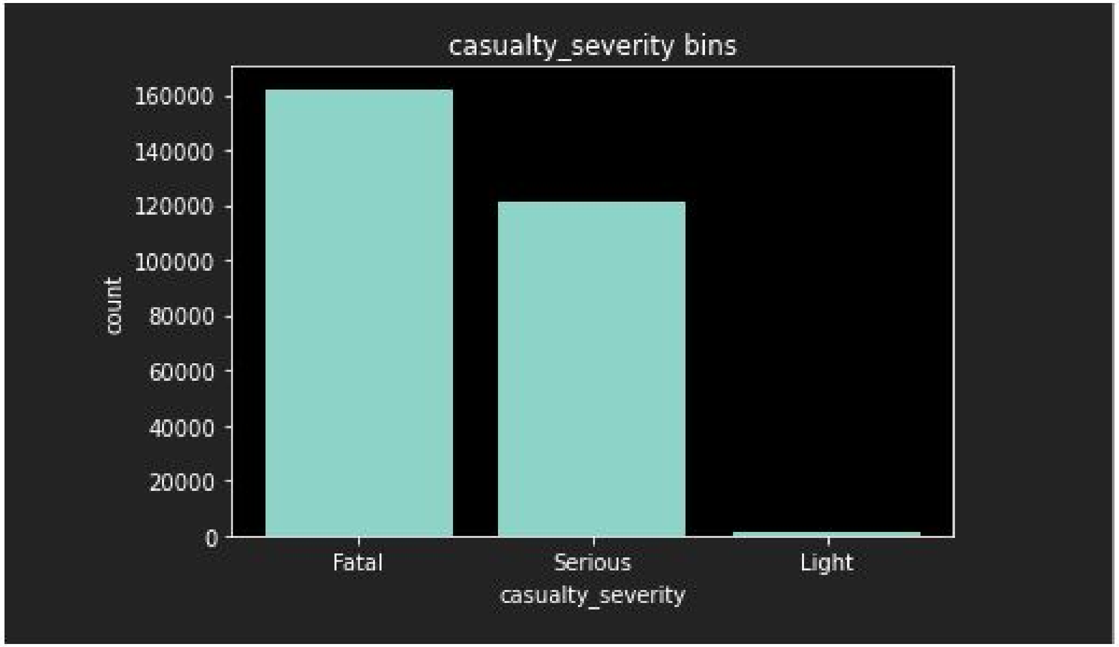
Clean Data



```
plt.pyplot.ylabel("count")
plt.pyplot.title("Severity of the Accident Bin")
```

```
Text(0.5, 1.0, 'Severity of the Accident Bin')
```

All the missing data was removed and replaced with values, which makes our data set a clean data.

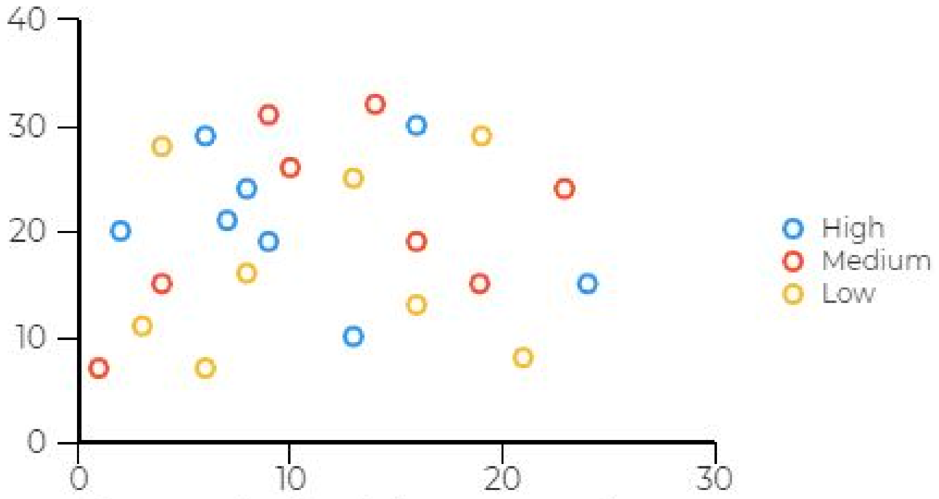


Modeling and Evaluation

The correlation between weather_conditions, light_conditions, speed_limit, road_surface_conditions, special_conditions_at_site, day_of_week and casualty_severity attributes are sigificant.

The linear distribution are weak for the following:

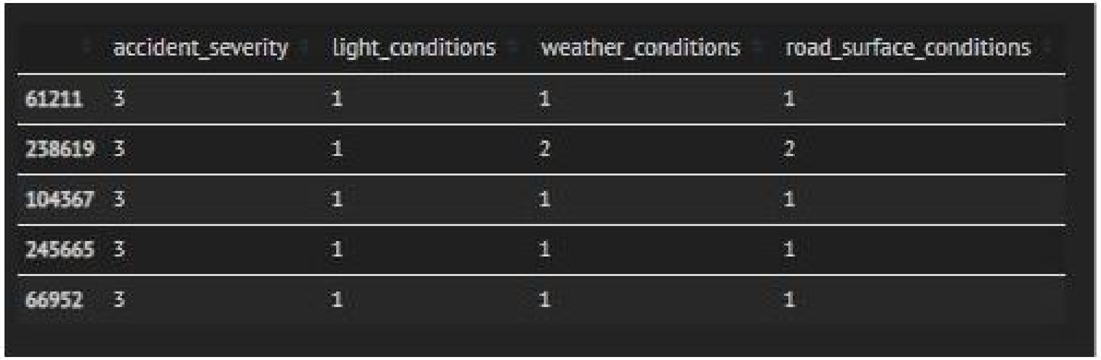
weather_conditions
light_conditions
speed_limit
road_surface_conditions
special_conditions_at_site
day_of_week



However, a weak correlation can be statistically significant, as the sample size is large enough.
Gauss Markov assumptions
- Collinear variables change at the same time, and therefore it is difficult to assess each variables distinct effect on the outcome variable.

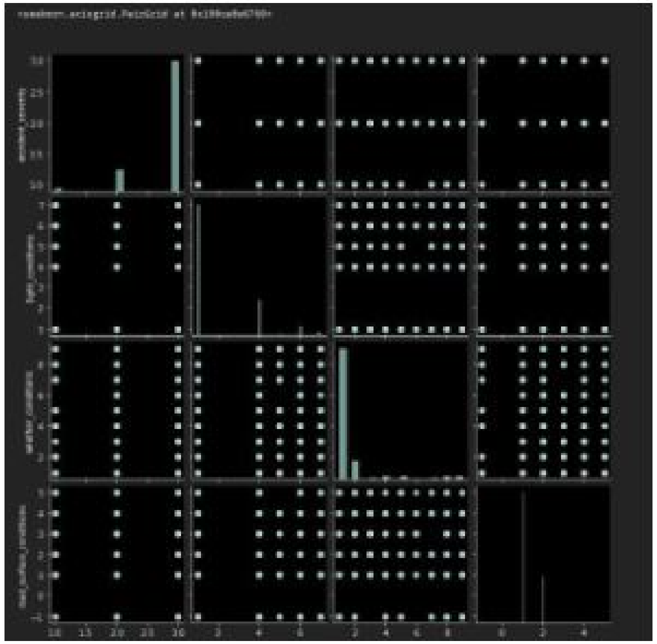
We will feed the model with cleaned data set. And Procced with Modelling, Evaluation and Deployment.

Selected Data for prediction



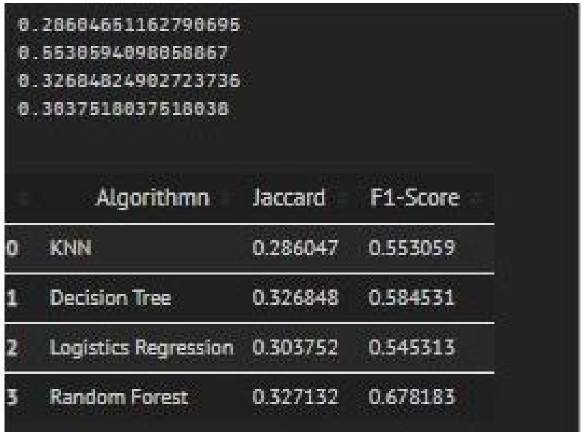
	accident_severity	light_conditions	weather_conditions	road_surface_conditions
61211	3	1	1	1
238619	3	1	2	2
104367	3	1	1	1
245665	3	1	1	1
66952	3	1	1	1

Multivariate plots to better understand the relationships between attributes.



In this phase, various algorithms and methods can be selected and applied to build the model including supervised machine learning techniques. I have selected decision tree, KNN , Linear Regression and . At this phase, stepping back to the data preparation phase was often required.

Results



```
0.28604651162790695
0.5530594098058867
0.32684824902723736
0.3037518037518038
```

	Algorithmn	Jaccard	F1-Score
0	KNN	0.286047	0.553059
1	Decision Tree	0.326848	0.584531
2	Logistics Regression	0.303752	0.545313
3	Random Forest	0.327132	0.678183

K-Nearest Neighbor (KNN)
KNN will help us predict the severity code of an outcome by finding the most similar to data point within k distance.

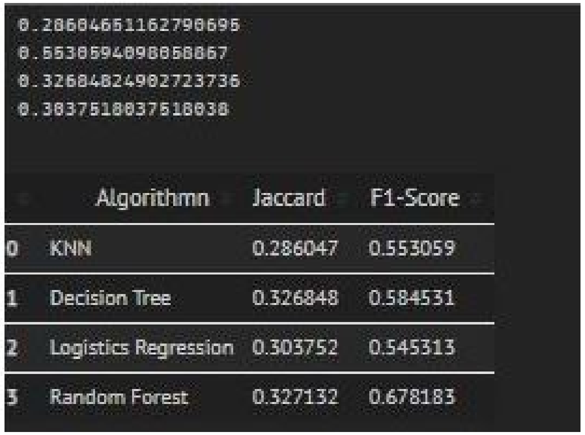
Decision Tree
A decision tree model gives us a layout of all possible outcomes so we can fully analyze the consequences of a decision. It context, the decision tree observes all possible outcomes of different weather conditions.

Logistic Regression
Because our dataset only provides us with three severity code outcomes, our model will only predict one of those two classes. This makes our data binary, which is perfect to use with logistic regression.

Conclusion

After building and comparing various classification models to predict the severity of the accident. I analyzed the relationship between severity of an accident and characteristics which involved conditions like road, light and weather to see the gravity of the accident.

Severity of an accident can be predicted in real time by using above data when an accident is reported and from there measures can be taken.



```
0.28604651162790695
0.5530594098058867
0.32684824902723736
0.3037518037518038
```

	Algorithmn	Jaccard	F1-Score
0	KNN	0.286047	0.553059
1	Decision Tree	0.326848	0.584531
2	Logistics Regression	0.303752	0.545313
3	Random Forest	0.327132	0.678183