Capstone Project

Accident Severity Predictions

Prepared by:

Pradiv Gnanaraj

Analysis | Machine Learning

Agenda

- Business Understanding
- Data Understanding
- Data Preparation
- Conclusion and Future Directions
- Review

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.

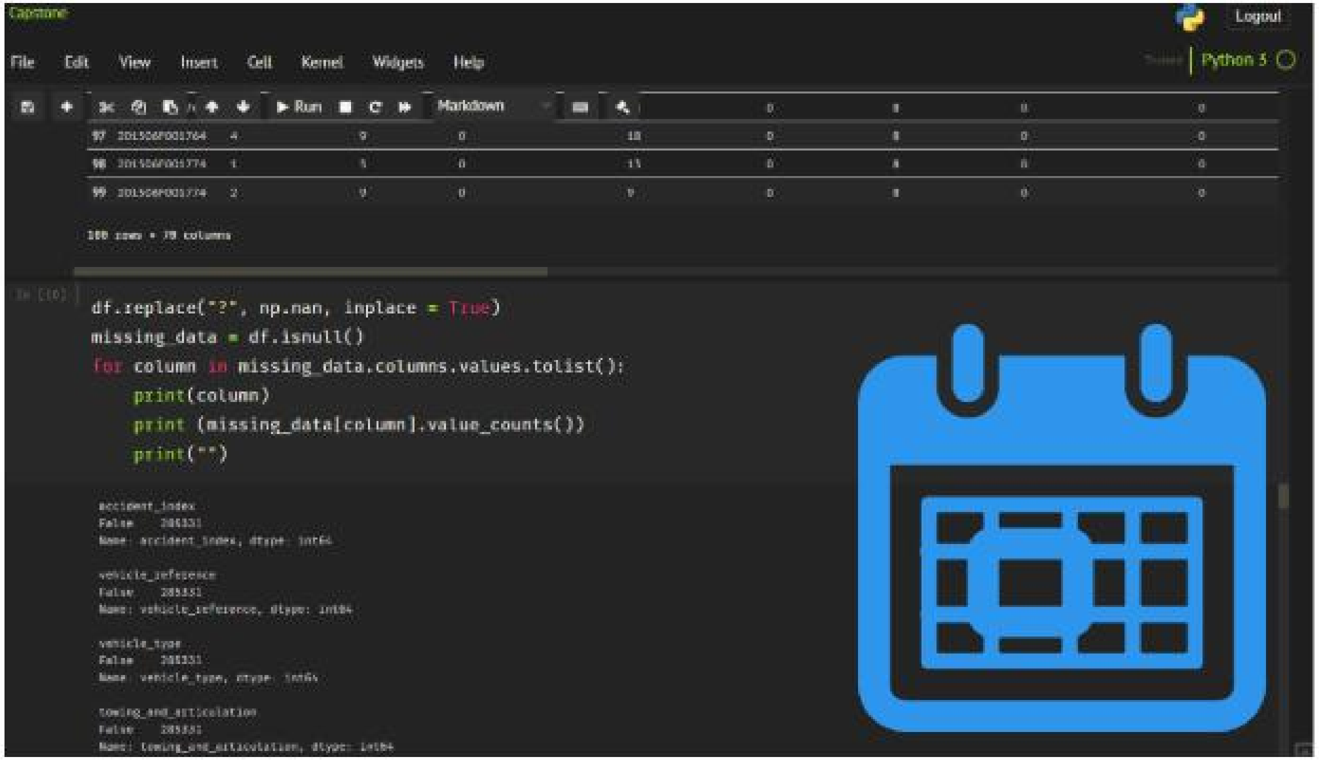
<link will be shared during the final submission>

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.



```
                                                    0      8      0      0
97  2015O6F001764  4       9      0        10       0      8      0      0
98  2015O6F001774  1       5      0        15       0      8      0      0
99  2015O6F001774  2       9      0         9       0      8      0      0

100 rows × 78 columns
```

```
In [16]:
df.replace("?", np.nan, inplace = True)
missing_data = df.isnull()
for column in missing_data.columns.values.tolist():
    print(column)
    print (missing_data[column].value_counts())
    print("")
```

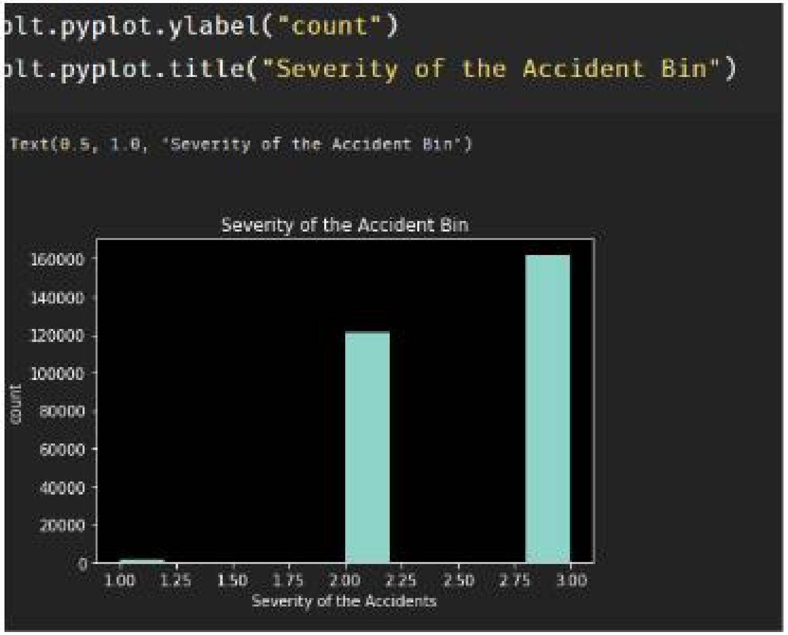
```
accident_index
False    285331
Name: accident_index, dtype: int64

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
```

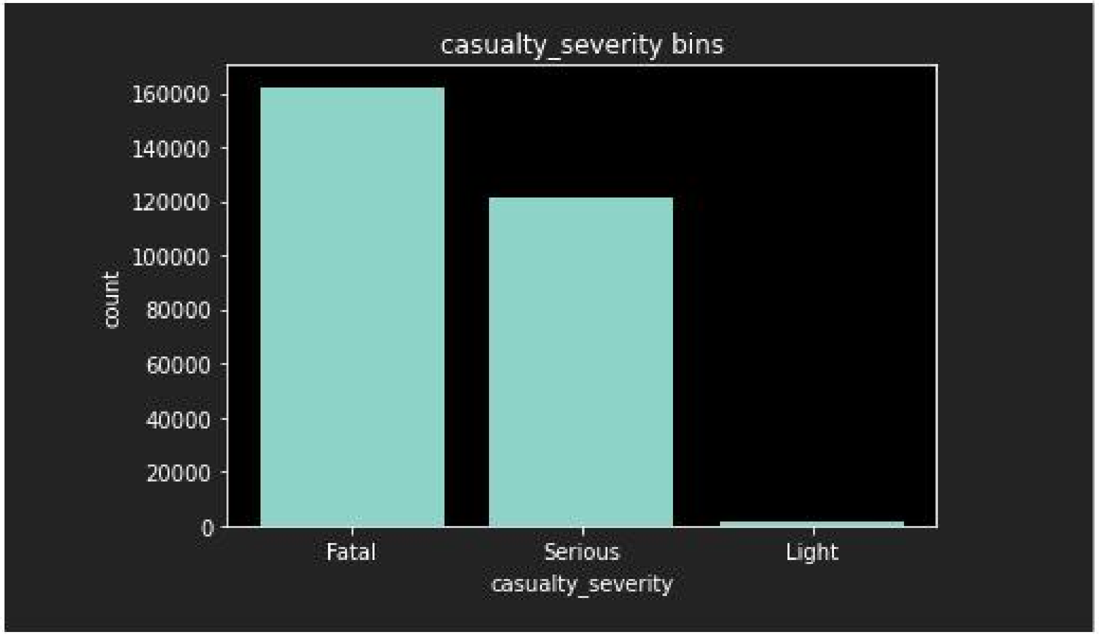
Clean Data



```
lt.pyplot.ylabel("count")
lt.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.

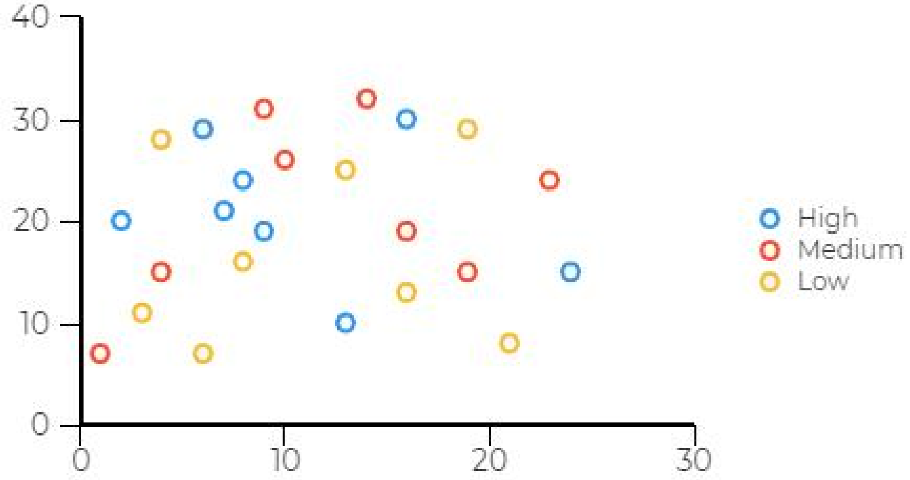


Conclusion and Future Directions

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.