Coursera Capstone Project

Accident Severity Prediction using Machine Learning

Arabinda Behera

11th September, 2020

Introduction and Business Problem

- Lot of lives are lost in car accidents every year all over the world
- Huge economic losses are also incurred
- Several factors like car conditions, weather conditions, traffic, driver's state etc can affect the car accidents' numbers and severity
- Design of machine learning models for predicting severity of car accidents can be useful for drivers, police, insurance agencies etc.
- It can also lead to the development of advanced navigation and warning softwares to provide with information and warnings to the drivers so that they can plan ahead of their journey.

Data Source and Data Reduction

- Data is taken from the open data website of the UK government published by the Department of Transport
- The two .csv files were loaded and merged to give a dataframe of size (2058408,57)
- Dataset was reduced to select only 30% of data for fast analysis
- 25 most important features were selected to construct a new dataframe
- The final dataframe has size - (617522,25)
- The first 5 rows are shown below

Out[48]:

	Accident_Index	1st_Road_Class	Day_of_Week	Junction_Detail	Light_Conditions	Number_of_Casualties	Number_of_Vehicles	Road_Surface_Conditions
892589	2011120033766	A	Wednesday	Not at junction or within 20 metres	Daylight	1	2	Wet or damp
1118119	2012350212112	A	Saturday	Not at junction or within 20 metres	Daylight	5	3	Dry
784164	2010070171227	C	Wednesday	Not at junction or within 20 metres	Daylight	1	2	Dry
1822070	201601BS70346	A	Monday	Crossroads	Daylight	1	2	Dry
286989	2007230795809	A	Monday	Private drive or entrance	Daylight	5	3	Dry

5 rows × 25 columns

Data Cleaning and Wrangling

- The target column Accident_Severity (Slight, Severe and Fatal) was converted from multi-class to two class by merging Severe and Fatal into one class Severe

- Pipelines were created to do the following

    i) Address null values in features

    ii) Feature transformation

    iii) One-Hot Encodings

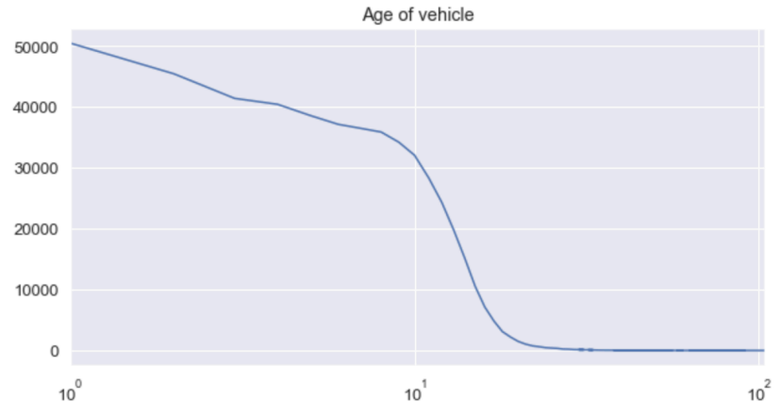
- The null values in features were replaced using SimpleImputer() function by the most frequent or median values

- Feature transformation were applied on some features explained in the report

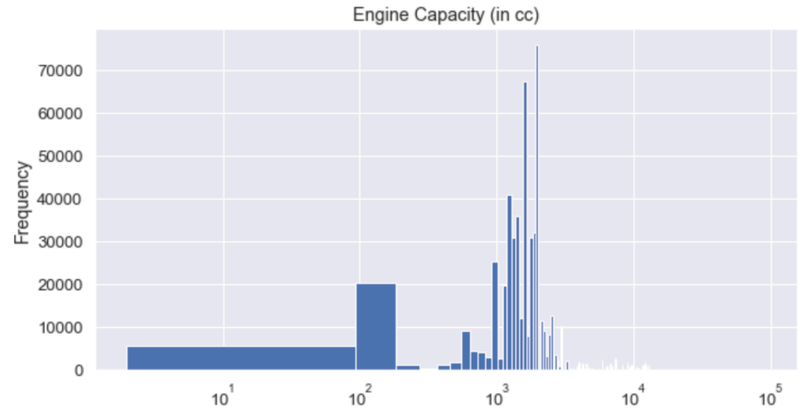
- One-Hot encodings were applied on the categorical features using OneHotEncoder() function

Data Visualisation

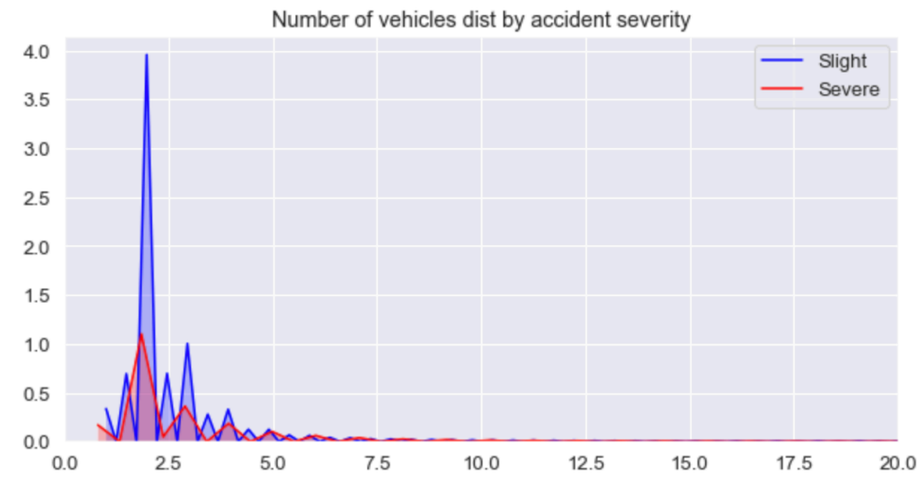
- Some important plots in the data frame are shown below

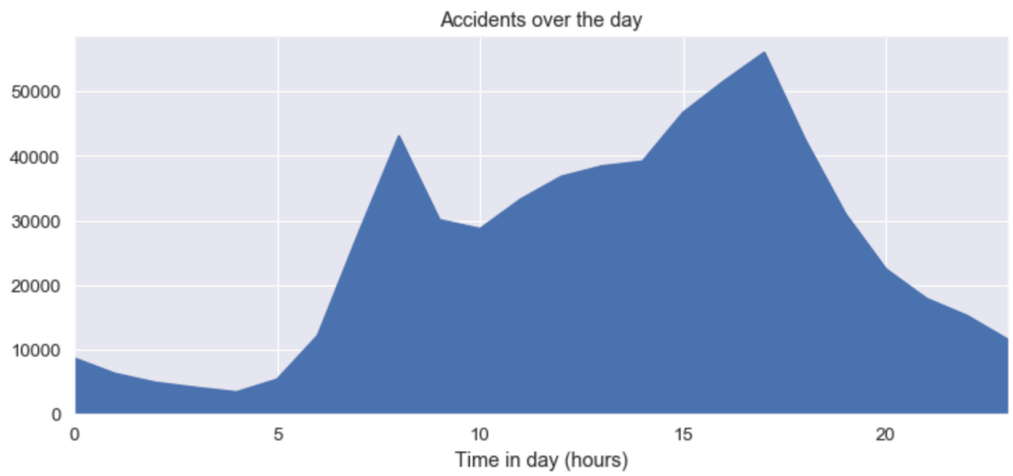


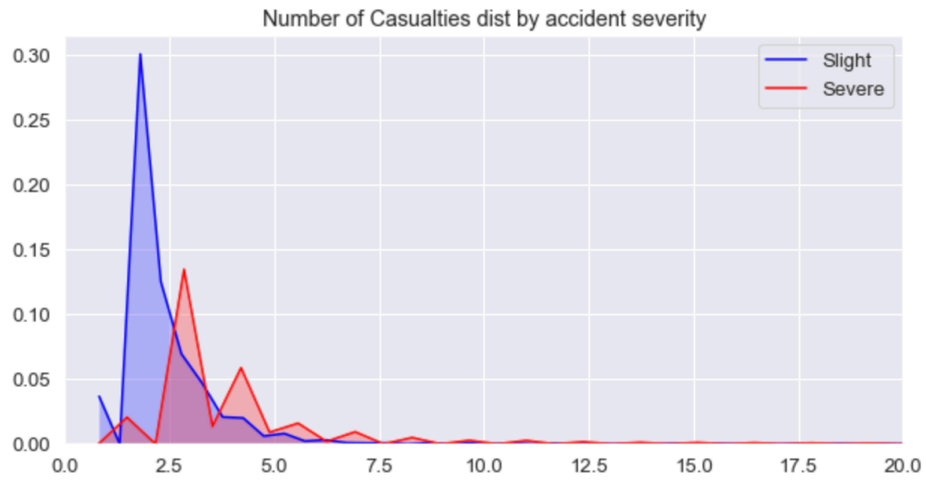

```
Min: 1.0
Max: 104.0
Median: 7.0
```




```
Min: 2.0
Max: 91000.0
Median: 1598.0
```







- Median age of cars in accidents is 7 years and median engine capacity is 1598 c.c
- In accidents over day plot - maximum no. of accidents around 5pm and minimum around 12 am to 5 am

Modeling

- The dataset is split into training and test sets in the ratio 3:1 - (X_train,y_train) and (X_test,y_test)
- The feature sets X_train and X_test were passed through pipelines to apply the transformations
- The problem is of classification so supervised learning models will be used
- Three models are used for classification of accidents into two severity classes - Slight or Severe

    1. Logistic Regression Classifier
    2. Gradient Boosting Classifier
    3. Random Forest Classifier

- All three models were trained on the training set (X_train,y_train)

Results

- The trained models were then applied on X_test to get the predictions - y_pred
- The probabilities of the classes for the prediction (y_pred_proba) were also obtained
- The models were evaluated using three evaluation metrics :

1. Jaccard Similarity Score
2. F1 Score
3. ROC_AUC Score

- Jaccard score and F1 score were calculated using y_test and y_pred
- ROC_AUC score was calculated using y_test and y_pred_proba

Results

- The scores are shown in the table below

Metric	Logistic Regression Classifier	Gradient Boosting Classifier	Random Forest Classifier
Jaccard Similarity Score	0.66	0.67	0.80
F1 Score	0.70	0.76	0.84
ROC_AUC Score	0.71	0.71	0.86

- Random Forest performs the best and Logistic Regression performs the worst among the three models

- Random Forest gives the highest prediction accuracy of 86%

Discussion

- The models perform quite well and the accuracy of 86% for the Random Forest model is pretty good
- Only a small amount of data ~ 600,000 is used in the project. The full dataset which has ~2,000,000 rows can further improve the performance
- A more detailed hyperparameter search for the models using cross validation can further improve the performance
- More advanced models or deep learning models like Artificial Neural Networks can significantly boost the performance

Conclusion

- Machine learning models were developed for classification of accident severity using UK government data
- Several data cleaning and feature engineering steps were performed to prepare the data for the models
- Three classification models were trained on the training sample of the data - Logistic Regression, Gradient Boosting and Random Forest
- The model prediction on the test sample were evaluated using three metrics - Jaccard Similarity Score, F1 score and ROC_AUC score
- Random Forest Classifier performed the best with accuracy of 86 %
- There is scope of improvement and other advanced deep learning models like artificial neural networks can improve the accuracy further

Thank You