University of Connecticut
MS in Business Analytics and Project Management

OPIM 5512 Data Science using Python

Road Accident Severity Prediction

Professor: Dave Wanik

Team 9

Prem Shah

Yixuan Lu

Guan-Yi Li

Lei Cao

Table of Contents

Executive Summary

Machine learning has become more ubiquitous, or significant in certain fields, the Department for Transport of the United Kingdom is constantly collecting data and publishing the statistics on their website. Due to the urbanization process around the globe, traffic accidents have been on a tremendous rise, causing significant life and property losses. Predicting traffic accidents is critically important in improving transportation, public safety and as well as safe routing.

Our objective was to focus on analyzing the factors which could help predict the severity of the accident. The results of analysis could be utilized by the Department of transport for safe route planning, Emergency Vehicle allocation, Roadway design and utilized by Google maps for better route recommendation to the users.

We started our analysis with exploratory data analysis to discern the dataset. Machine learning algorithms were used to explore the complex interactions among roadways, traffic, environmental elements and predicting accident severity. Since most of the predictor variables in the dataset were categorical, we recoded categorical variables. 11 models were built, evaluated for complexity and accuracy, and compared to conclude which model is the best fit for predicting accident severity.

Spot Checking technique was used to fit the 11 models to determine which models would predict the accident severity with the highest accuracy. We also performed feature engineering to enrich our dataset Hyperparameter tuning and pipelining the best performing model helped to improve the performance of the model by making accurate predictions. Gradient Boosting performed well

with the accuracy of 78.69% and which were further improved by doing permutation testing for feature importance which played an important role in predictions.

Introduction

Traffic accidents are extremely common, and it is more frequent in the sprawling metropolis. Because of their frequency, traffic accidents are a major cause of death globally, cutting short millions of lives per year. By 2017, the total number of cars in the UK is 31,200,182, indicating that per 1000 people have 471 motor vehicles, rank 35th in 160 countries. However, road accidents happen every day. In 2018, there were a total of 160,597 casualties of all severities in road traffic crashes.

The potential to predict the accident severity (e.g. What will cause accidents? What will make accidents worse?) is therefore useful not only to public safety stakeholders but also transportation administrators and individual travelers. A system that can predict the cause of traffic accidents and predict the accident severity can potentially save lives and aid in developing better roadway designs. The detailed data, public transportation information, and motor vehicle crash reports could provide us valuable insights for traffic accident analysis.

However, this problem is very taxing due to several issues such as class imbalance and low significance of predictor variables depending upon the location meaning too much randomness in the data. Since we are creating a binary target variable between slight accidents (0) and serious or Fatal accidents (1) the class will be severely imbalanced. Simple linear models might not be useful for predicting the accident severity due to non-linearity.

We will use various analysis techniques and build models to predict accident severity, including Logistic Regression, Random Forest, Linear SVC, KNN, Decision Tree. By using predictive analysis and comparing the models, we can have a better understanding of different variables like what kind of road condition, vehicle condition and what factors are involved that contribute to road accidents.

Data Description

The dataset on road traffic accident analysis has about 2 million rows and 63 columns. The predictor variables in the dataset had information geographical locations, weather conditions, type of vehicle, number of casualties and vehicle maneuvers. The dataset had missing values which were dropped.

During the exploratory data analysis part, we found some compelling details like the age group between 26 years to 35 years was more frequent with accidents than any other age group. We also explored that most of the accidents happened on Friday. We were surprised to see that weather conditions did not play much role at the time of accident. Most of the accidents happened during the afternoon rush that is between 3.00 pm to 7.00 pm. This data exploration gave us clear insights that road conditions, rush hours, weekdays, age group, age of vehicle, and junction details played a major role in road accidents.

In order to increase the practicability of the dataset, extensive feature engineering for all the predictor variables was performed. We recoded all the categorical variables and we also performed interaction terms between two predictor variables i.e. age of predictor variable and speed limit. Following this, we applied Z -score standardization for all the numeric variables so that the scale of any numeric variable does not influence the predictions during the modeling.

We classified our multiclass target variable into 2 classes i.e. slight as ’0’ and Fatal or severe as ‘1’. The original dataset had only 13% rows for severe or Fatal accident severity, this was a highly unbalanced dataset. In order to balance it for the modeling, we used a sampling technique to balance out the Slight and Severe or Fatal accident severity. After under-sampling technique, we had a ratio of 1:1 for slight and severe or Fatal accident severity which was used for modeling and training the data.

Model Description

Our final modeling dataset of 381K rows and 62 columns was split into training data with 70% and test data with 30%. We implemented Spot-checking technique to determine the machine learning model which was best suited to predict the accident severity. Machine learning models which we used are Logistic Regression, Random Forest Classifier, K Neighbors Classifier, Gaussian NB, Perceptron, SGD Classifier, Decision Tree Classifier, Gradient Boosting Classifier, Linear Discriminant Analysis, Extra Trees Classifier, and Bagging Classifier.

Based on the accuracy, from 11 different models we narrowed down to 6 best models whose accuracy is more than 60%. The dataset was too huge to do the hyperparameter tuning and gridsearchcv, so we subset the data with the accidents that happened in 2016. We used hyperparameter tuning to modify parameters in all the 6 models which are Logistic Regression, Random Forest classifier, Gradient Boosting Classifier, Linear Discriminant analysis, Extra Trees classifier and Bagging Classifier to improve the performance of the models.

We further implemented permutation testing for feature importance for our best model i.e. Gradient Boosting. Here we saw that the features like Engine capacity, Number of vehicles, speed limit, age group of drivers are some of the features which were believed to be important in making predictions. While we also saw features like pedestrian crossing human control, vehicle location restricted lane, road surface condition code, weather condition code, were comparatively less important features.

We further tried to improve our final model i.e. Gradient Boosting using the results of permutation testing and we eliminated all the less important variables and it resulted in far better accuracy that is 86.71%.

Results

We used spot checking with 11 different models to find out which all models are the best suitable for predicting the results. Refer to figure 2.1, (Appendix). Our criteria to consider the best 6 models was the total accuracy and we considered the best 6 models whose accuracy was above 60% for the hyperparameter tuning and gridsearchcv. The logistic regression had the total accuracy of 60%, random forest's accuracy was 63%, Gradient Boosting's accuracy was 64%, Linear Discriminant analysis's analysis accuracy was 60%, Extra Trees classifier's accuracy was 62% and Bagging classifier's accuracy was 60%.

After this step, we did the hyper parameter tuning for all the 6 models. With this step we got an insight about the best parameters for each algorithm. Refer to figure 2.2, (Appendix). Test set accuracy score for best params of Logistic Regression is 57%, Test set accuracy score for best params of Random Forest classifier is 63%, Test set accuracy score for best params of Gradient Boosting Classifier is 62%, Test set accuracy score for best params of Linear discriminant analysis is 57%, Test set accuracy score for best params of Extra Trees classifier is 58% and Test set accuracy score for best params of Bagging classifier is 57%.

From the above step, we set the best parameters for each model and trained the dataset. Refer to figure 2.3, (Appendix). The parameters that were changed are minimum samples leaf and minimum samples split along with the maximum split. After using hyperparameter tuning and based on the accuracy score, we decided that the best model was Gradient Boosting Classifier Subsequently we also implemented permutation testing for feature importance and we identified features that were significant for the model predictions and also the features which were less significant for the model predictions. Refer to figure 2.4 (Appendix).

We did this permutation testing using a Random Forest Classifier algorithm. since Random Forest Classifier tells us the clear picture about which features are important and which features are less important. After permutation testing, we polished our best models with the variables that were considered significant. We fit our models again by including only important features for predicting the results. We saw that accuracy of Logistic regression was 56%, accuracy of random forest was 60%, accuracy of gradient boosting was 86%, accuracy of linear discriminant analysis was 55%, accuracy of extra trees classifier was 58% and accuracy of bagging trees was 55%. We saw that gradient boosting performs well with the accuracy of 86%. It is safe to say that dropping all the less significant variables did help in terms of accuracy of the model for Gradient Boosting Classifier. Please refer to figure 2.5 and figure 2.6 (Appendix).

Discussion

Road accidents are a serious problem in our societies across the globe. The world health organization estimated that 1.25 million deaths were related to road traffic injuries in the year 2010. Transport authorities worldwide have been striving to implement strategies to minimize the road traffic accidents by introducing safety regulations. There were many strategies implemented for road traffic accidents reduction but has proven to be an elusive goal as these measures have hitherto failed to make a considerable reduction in the frequency of the road traffic accidents. Accidents are influenced by many measurable factors such as driving speed, road condition, weather condition, light condition and so on. Therefore, many researchers have come together to understand the dynamics of road traffic accidents.

Katannya Kapeli and Meraldo Antonio (2019) researched that weather conditions had no role in severe or fatal accidents. This is quite evident because there were multiple factors that affected the road traffic accidents, they considered junction, time, origin and destination played a vital role in predicting the road traffic accidents. They used a negative sampling technique using the several hundreds of accident hot spots and they classified their target variable with accident and no accident. They used classification machine learning models and their best model was random forest with only numerical predictors.

Other researchers aim to classify the target variable into binary classes {accident, no accident}. They compared the performance of artificial Neural Network with the negative binomial regression models. Artificial Neural Network achieved 64% and 61% accuracy for training and test data. They also applied a decision tree, random forest, k nearest neighbor to predict road accidents. The best accuracy achieved by them was 70%. Their dataset was quite old therefore comparing it directly is unfair. They have not mentioned anything about the imbalanced class problem.

Our work, by contrast, incorporates no negative sampling and classified our target variable into slight accident severity and Fatal or severe accident severity. Predicting the accident severity makes sense rather than predicting about accident and no accident because if we predict whether the accident is going to happen or not it would not show us how severe the accident can be and how much casualties can happen at a time. Whereas predicting the accident severity gives us more insights about how severe the accident can be given the condition or set of parameters. Based on this information the transportation authorities can take immediate actions about the conditions of the road, designing of the road, light conditions and various other factors that affect the accident severity.

Conclusions and Next Steps

We investigated the problem of traffic accident severity prediction using Road traffic accident analysis data. This was a challenging problem due to class imbalanced as the majority of the accident severity belonged to Slight accident severity. It was surprising to see that weather and light conditions did not have much impact on the fatal and severe accident severity. We saw that during the permutation testing, features like engine capacity, number of vehicles, speed limit, age group of drivers played an important role for improving the accuracy of the Gradient boosting classifier model.

For the future work, we want to see the results while working with the whole dataset and not have computing restrictions. Since the scope of the course was about learning the data science concept using machine learning algorithms, we hope to see what we can achieve using deep learning models for such a heavy dataset which has many categorical predictor variables.

Looking at the results, we believe that we should have done negative sampling and tried building the models. We also learned that with more features such as real time traffic information, construction, important events, higher resolution weather could have helped the model to improve significantly. We also want to see how our model performs by assigning weights to each class of our target variable.

References

https://www.datacamp.com/projects/462

https://towardsdatascience.com/live-prediction-of-traffic-accident-risks-using-machine-learning-and-google-maps-d2eeffb9389e https://www.gov.uk/government/collections/road-accidents-and-safety-statistics

https://parsers.me/deep-learning-machine-learning-whats-the-difference/

https://www.sciencedirect.com/science/article/pii/S2352146516304033

file:///D:/UCONN%20Acads/semester%202/python%20learnings/39x.pdf

https://www.hindawi.com/journals/jat/2018/3869106/#abstract

https://towardsdatascience.com/time-series-forecasting-for-road-accidents-in-uk-f940e5970988

Appendix

1. Exploratory data analysis

Figure 1.1

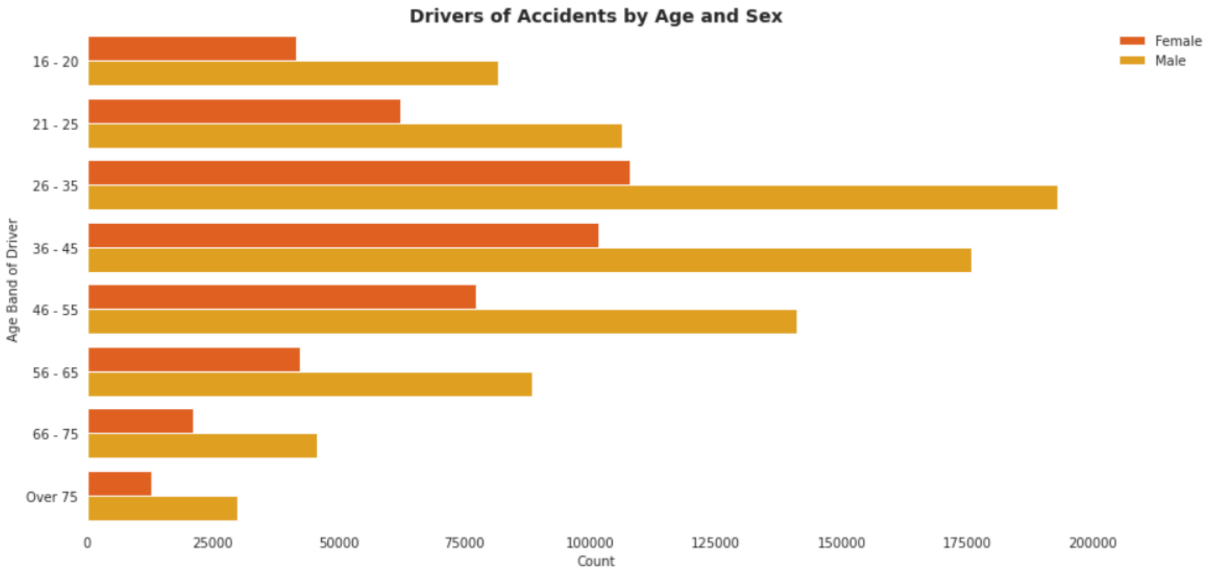


Figure 1.2

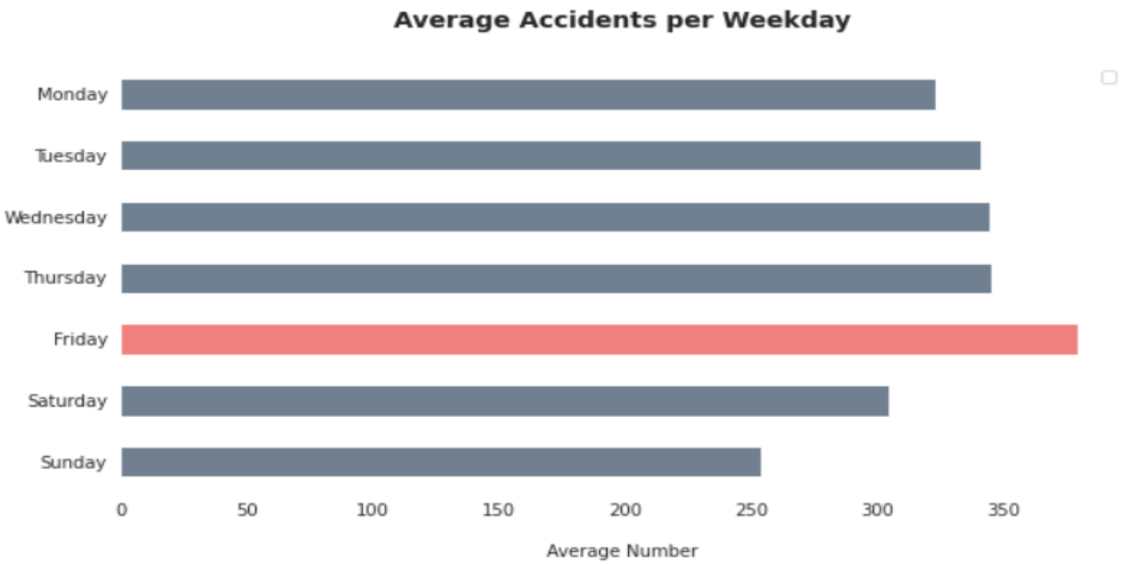


Figure 1.3

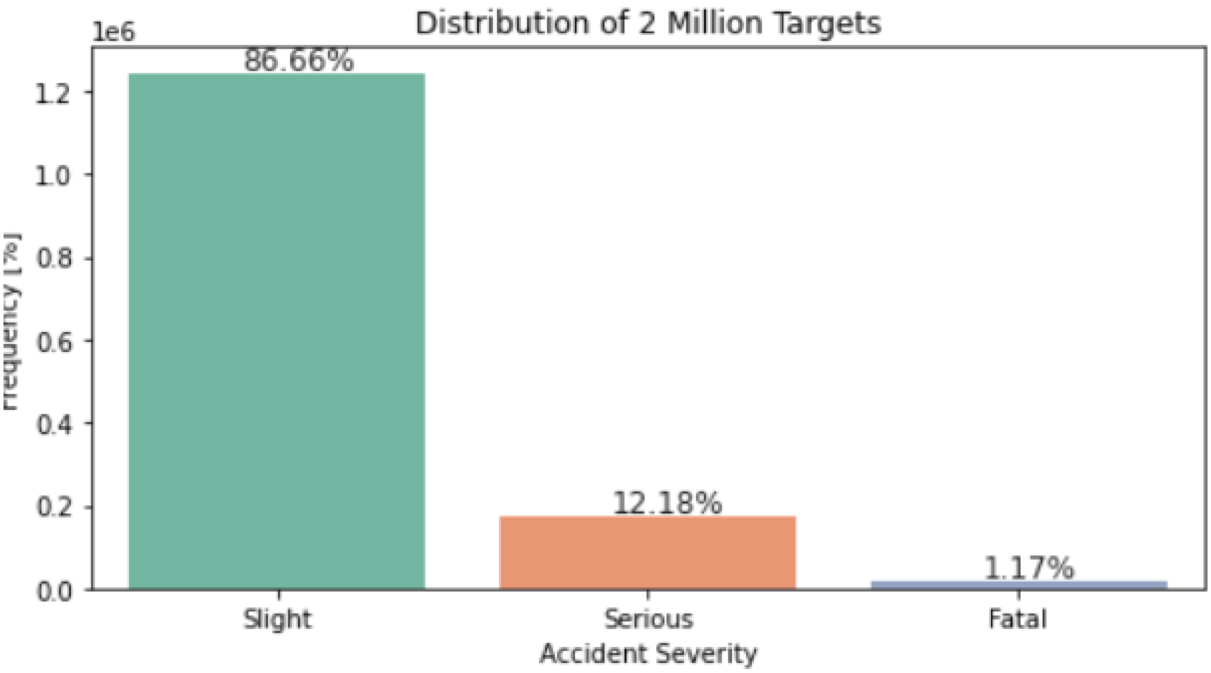


Figure 1.4

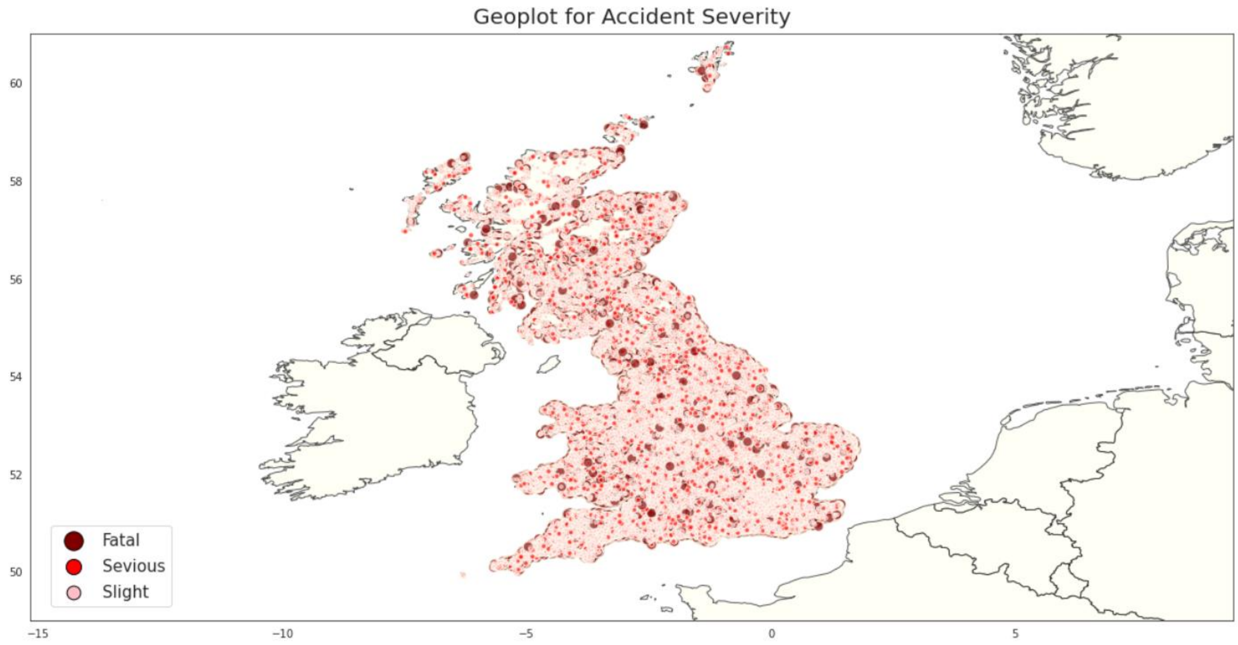
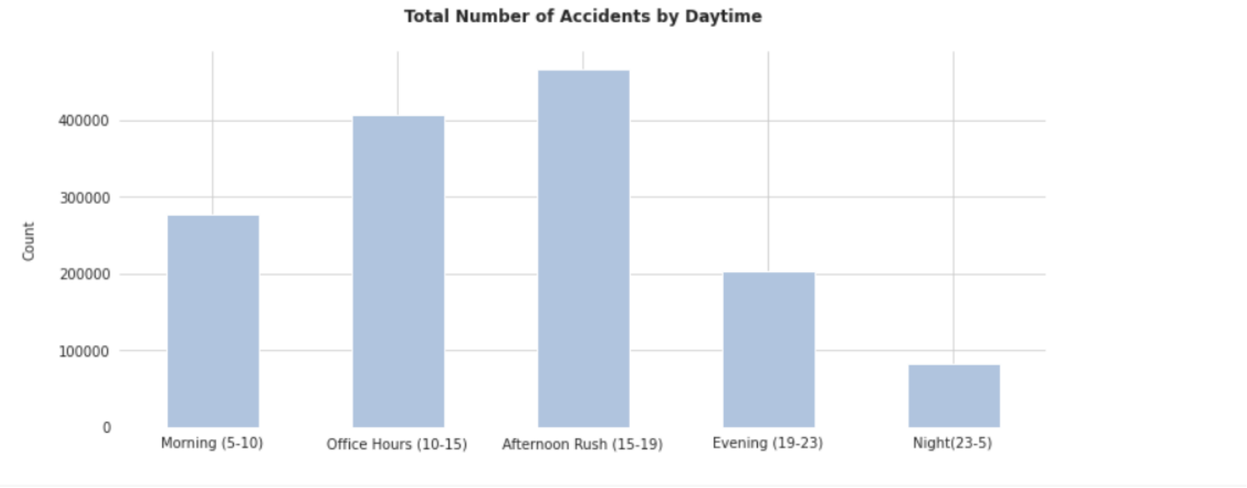
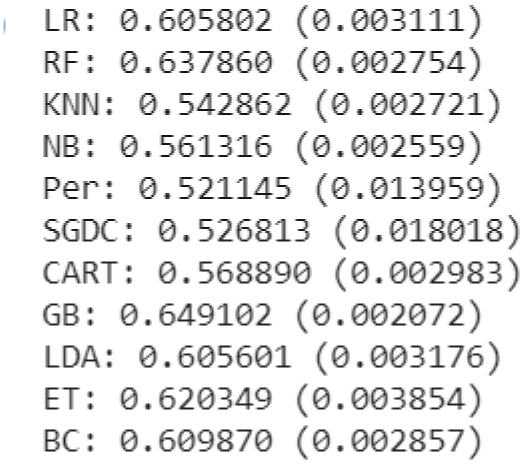


Figure 1.5



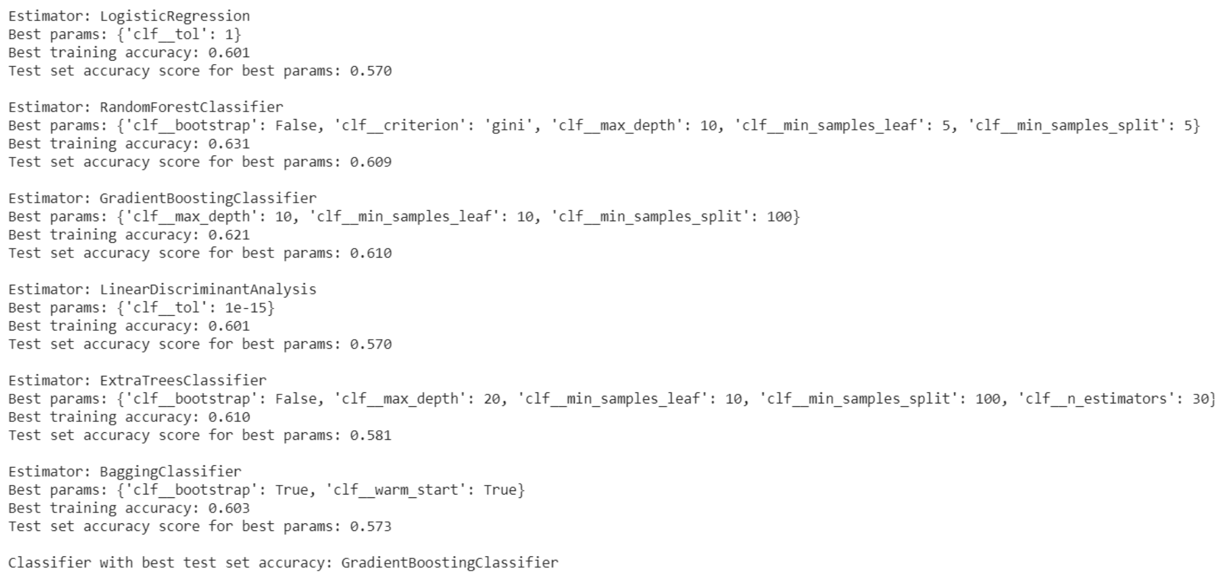
2. Modelling

Figure 2.1



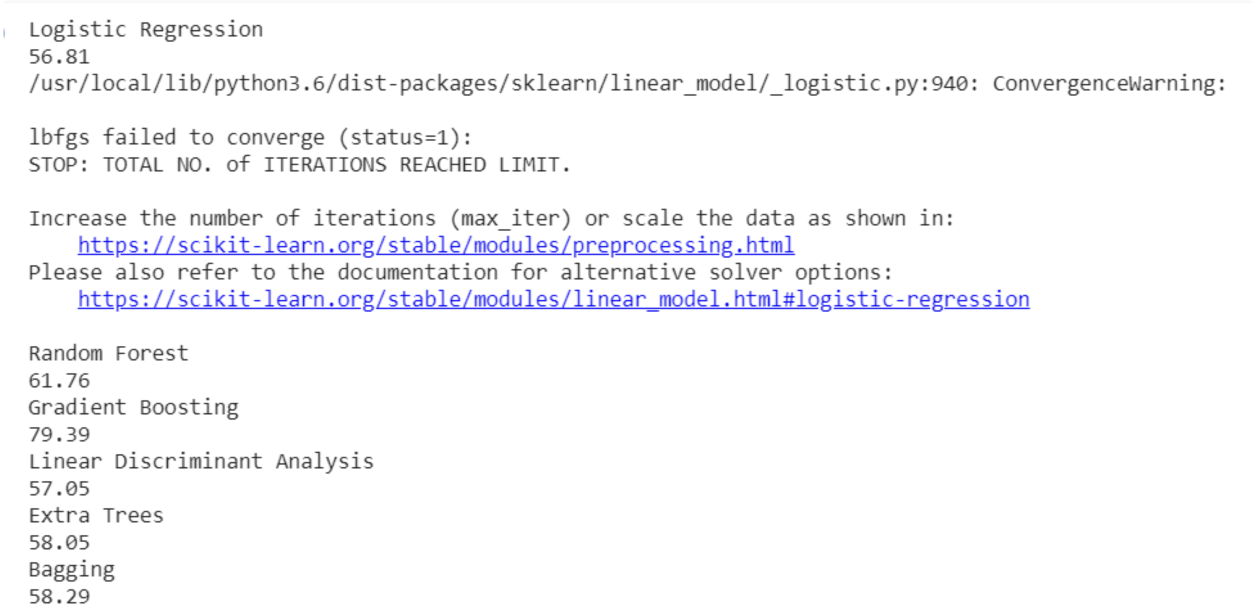
```
LR: 0.605802 (0.003111)
RF: 0.637860 (0.002754)
KNN: 0.542862 (0.002721)
NB: 0.561316 (0.002559)
Per: 0.521145 (0.013959)
SGDC: 0.526813 (0.018018)
CART: 0.568890 (0.002983)
GB: 0.649102 (0.002072)
LDA: 0.605601 (0.003176)
ET: 0.620349 (0.003854)
BC: 0.609870 (0.002857)
```

Figure 2.2



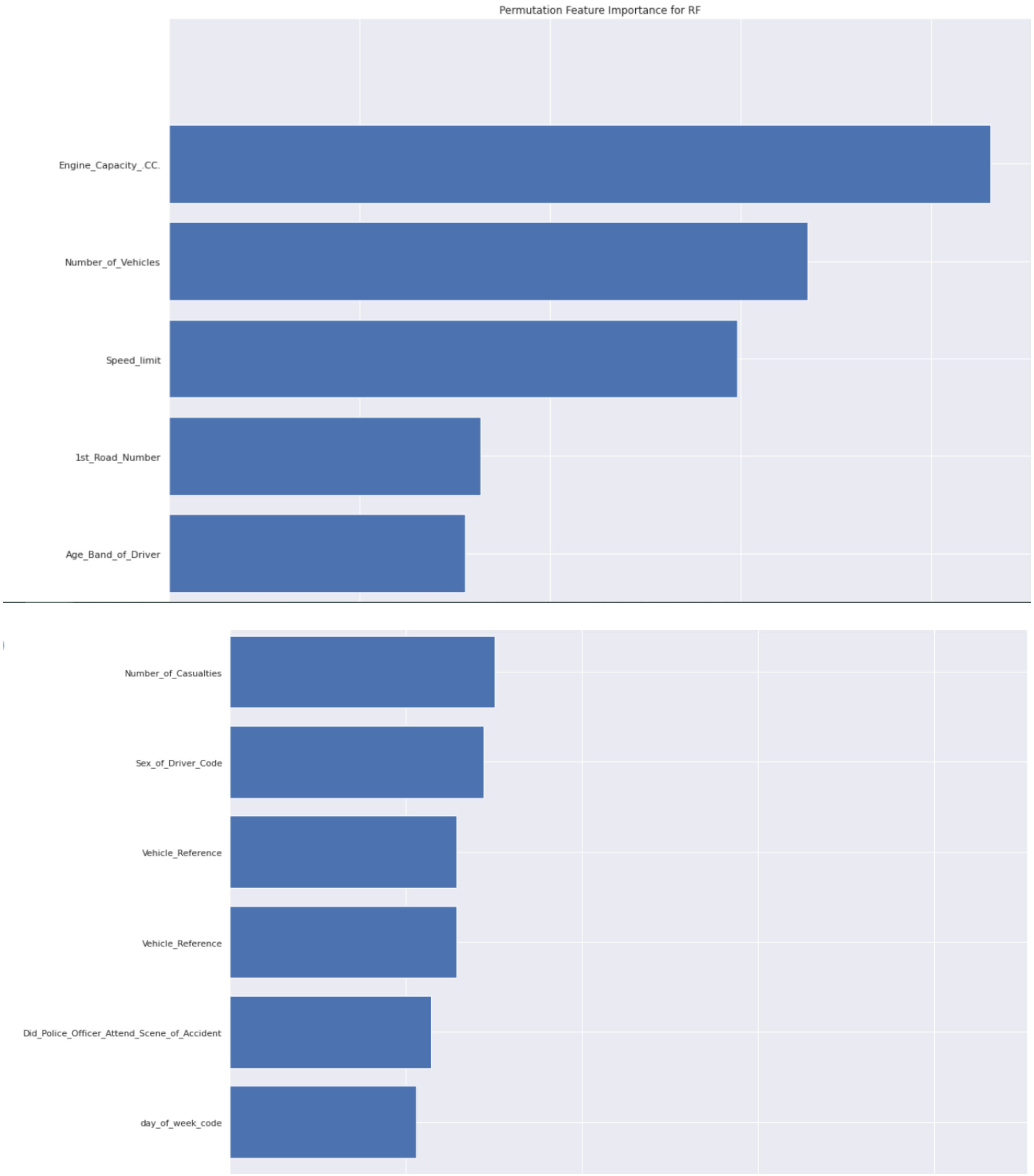
```
Estimator: LogisticRegression
Best params: {'clf__tol': 1}
Best training accuracy: 0.601
Test set accuracy score for best params: 0.570

Estimator: RandomForestClassifier
Best params: {'clf__bootstrap': False, 'clf__criterion': 'gini', 'clf__max_depth': 10, 'clf__min_samples_leaf': 5, 'clf__min_samples_split': 5}
Best training accuracy: 0.631
Test set accuracy score for best params: 0.609

Estimator: GradientBoostingClassifier
Best params: {'clf__max_depth': 10, 'clf__min_samples_leaf': 10, 'clf__min_samples_split': 100}
Best training accuracy: 0.621
Test set accuracy score for best params: 0.610

Estimator: LinearDiscriminantAnalysis
Best params: {'clf__tol': 1e-15}
Best training accuracy: 0.601
Test set accuracy score for best params: 0.570

Estimator: ExtraTreesClassifier
Best params: {'clf__bootstrap': False, 'clf__max_depth': 20, 'clf__min_samples_leaf': 10, 'clf__min_samples_split': 100, 'clf__n_estimators': 30}
Best training accuracy: 0.610
Test set accuracy score for best params: 0.581

Estimator: BaggingClassifier
Best params: {'clf__bootstrap': True, 'clf__warm_start': True}
Best training accuracy: 0.603
Test set accuracy score for best params: 0.573

Classifier with best test set accuracy: GradientBoostingClassifier
```

Figure 2.3



```
Logistic Regression
56.81
/usr/local/lib/python3.6/dist-packages/sklearn/linear_model/_logistic.py:940: ConvergenceWarning:

lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

Random Forest
61.76
Gradient Boosting
79.39
Linear Discriminant Analysis
57.05
Extra Trees
58.05
Bagging
58.29
```

Figure 2.4


Permutation Feature Importance for RF
Engine_Capacity_.CC.
Number_of_Vehicles
Speed_limit
1st_Road_Number
Age_Band_of_Driver
Number_of_Casualties
Sex_of_Driver_Code
Vehicle_Reference
Vehicle_Reference
Did_Police_Officer_Attend_Scene_of_Accident
day_of_week_code

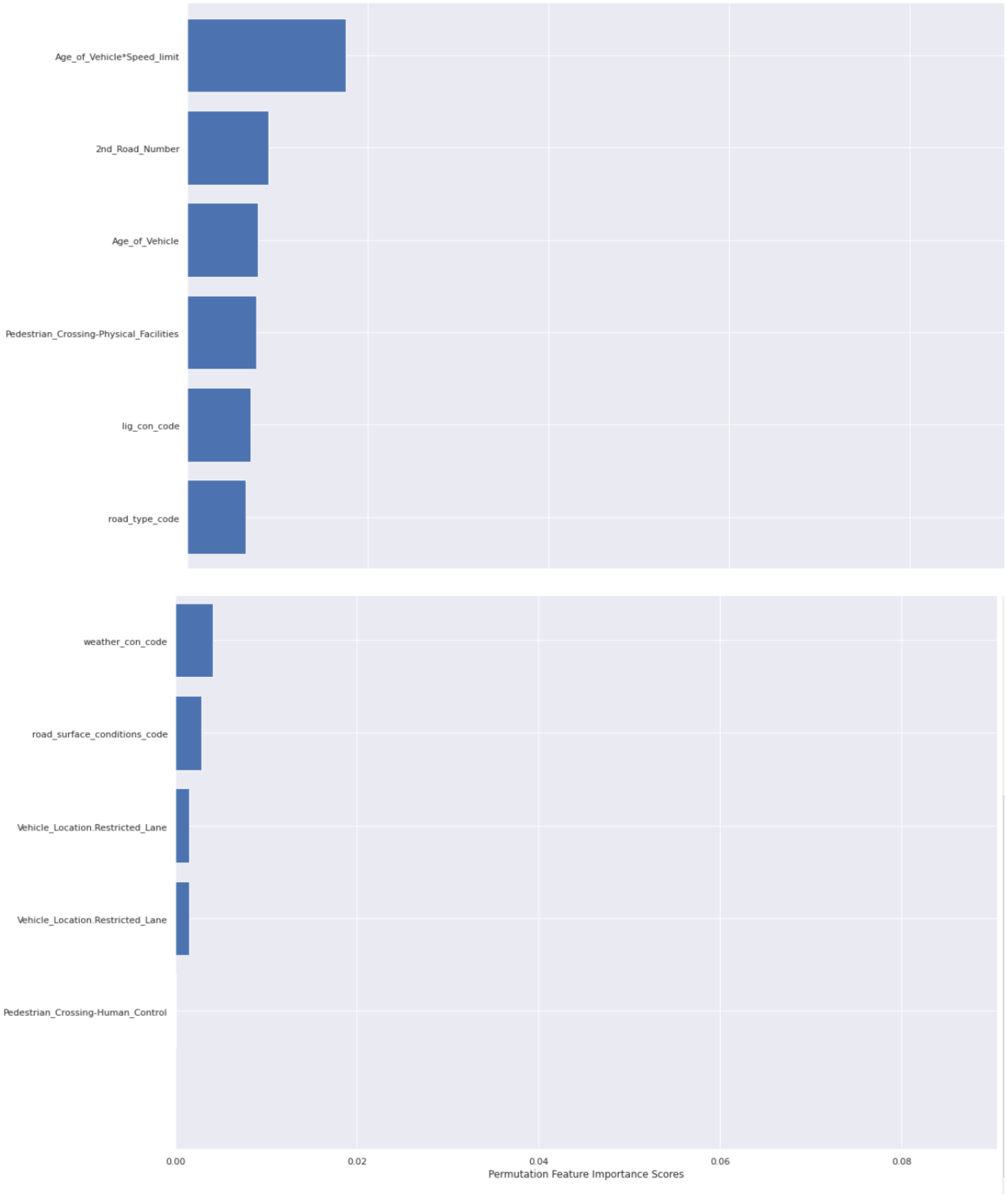
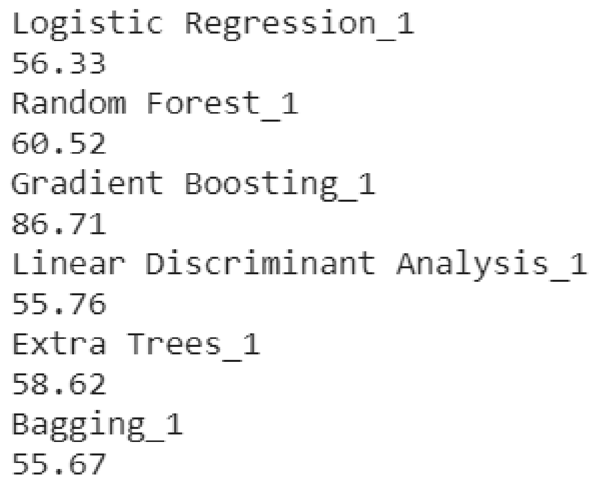
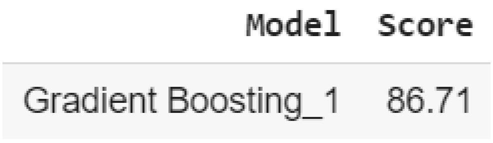

Age_of_Vehicle*Speed_limit
2nd_Road_Number
Age_of_Vehicle
Pedestrian_Crossing-Physical_Facilities
lig_con_code
road_type_code
weather_con_code
road_surface_conditions_code
Vehicle_Location.Restricted_Lane
Vehicle_Location.Restricted_Lane
Pedestrian_Crossing-Human_Control
0.00
0.02
0.04
0.06
0.08
Permutation Feature Importance Scores

Figure 2.5



```
Logistic Regression_1
56.33
Random Forest_1
60.52
Gradient Boosting_1
86.71
Linear Discriminant Analysis_1
55.76
Extra Trees_1
58.62
Bagging_1
55.67
```

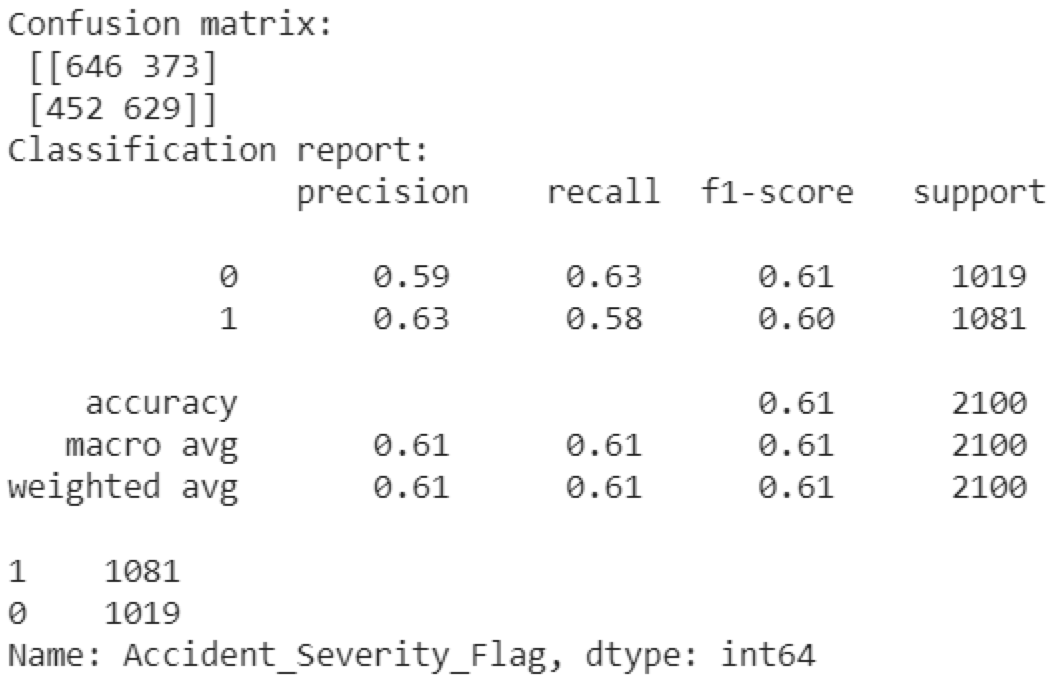
Figure 2.6



Model	Score
Gradient Boosting_1	86.71

Figure 2.7

Results before permutation testing for gradient boosting



```
Confusion matrix:
 [[646 373]
 [452 629]]
Classification report:
              precision    recall  f1-score   support

           0       0.59      0.63      0.61      1019
           1       0.63      0.58      0.60      1081

    accuracy                           0.61      2100
   macro avg       0.61      0.61      0.61      2100
weighted avg       0.61      0.61      0.61      2100

1    1081
0    1019
Name: Accident_Severity_Flag, dtype: int64
```