2020

Predicting the Severity of Road Accidents

FINAL REPORT FOR COURSERA IBM CAPSTONE PROJECT

ALFRED APPIAH

1.0 Introduction/Business Problem

Road accidents continue to kill many people every day in the world. In the UK for example, 1,784 people died in 2018 from road accidents in addition to over 25,000 people that sustained serious injuries (UK Department for Transport, 2018). All road accidents differ in severity based on a multiplicity of factors. Understanding the impact of road characteristics, weather conditions, car and driver characteristics on the severity of road accidents continue to a key policy discussion in road safety.

The intent of this project is to build a machine learning classification model to predict the severity of a road accident based on road characteristics, weather conditions, car and driver characteristics. Understanding those relationship would help drivers determine when to take a trip given their characteristics, the weather, the road conditions or the car they are driving. The model will also contribute to ongoing government efforts and campaign messages. For instance, if certain age groups driving certain types of cars are found to be at risk of severe accidents then campaign messages could be tailored to those groups appropriately. Also, if certain types of road under certain types of weather conditions are more likely to result in severe road accidents, then appropriate investments and warnings can be developed as a response.

2.0 Data

The data to be used for this analysis comes from Kaggle.com (https://www.kaggle.com/tsiaras/uk-road-safety-accidents-and-vehicles). The original data is from the UK's Department for Transport. There are two data sources available:

- **Accident information,** which contains information on distinct traffic accidents from 2005 to 2017. This dataset contains attributes, include a target variable for accident severity, that can be used in machine learning models. This dataset contains 2,047,256 rows and 34 columns.
- **Vehicle information,** which contains information on the vehicles involved in the accident and passenger information from 2004 to 2016. This dataset contains 2,177,205 rows and 24 columns.

The features and target variable available in the datasets are shown below.

Feature selection

I joined the two datasets using a distinct ID (Accident_Index). The target variable was *Accident_Severity.* Features for training the model will include the following:

- *Day_of_Week*
- *Junction_Detail*
- *Junction_Control*
- *Road_Surface_Conditions*
- *Road_Type*
- *Speed-Limit*
- *Urna_or_Rural_Area*
- *Weather_Conditions*
- *Age_of_Vehicle*

- *Engine_Capacity*
- *Sex_of_Driver*
- *Age_of_Driver*
- *Was_Vehicle_Left_Hand_Drive*
- *Vehicle_Type*
- *Light_Conditions*

Exploring the target variable

The plot below shows the distribution of classes in the target variable. The distribution shows that the data is highly unbalanced, with the vast majority of accidents (86%) categorized as slightly severe. That poses issues for modelling and therefore was addressed appropriately. In addition, given how small the proportion of fatal cases are, I grouped them together with the serious category, making this a binary classification problem.

Table 1 Distribution of target variable

Class	**Proportion**
Slight	85.8%
Serious	12.9%
Fatal	1.3%

3.0 Methodology

This section outlines the methodology used in the study. The section begins with a discussion of exploratory data analysis that was performed to understand the data, then discusses how the steps undertaken in feature engineering to make the data ready for modelling. Classification models developed for this project are then discussed, including the evaluation metrics used

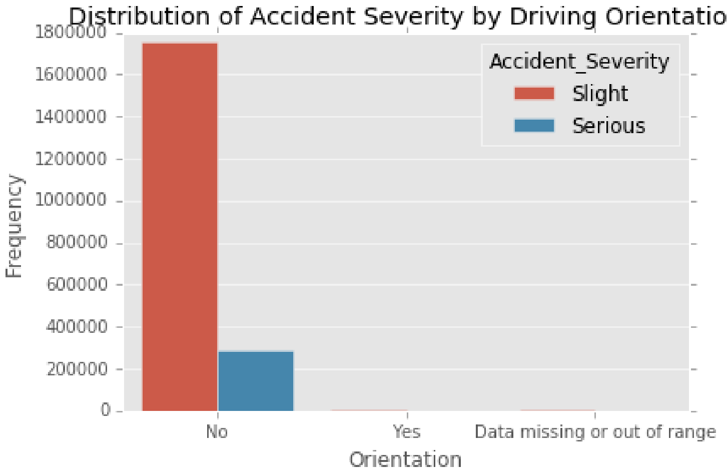
3.1 Exploratory data analysis

3.1.1 Vehicle features

The first vehicle feature explored was the age of the vehicle. Because this was a numeric feature, measures of central tendency like mean and median were used. This showed that the mean and median age of vehicles were similar (7.14 and 7 respectively). Next, we explored the engine capacity of the vehicles and the average engine capacity was 2,028 cubic centimeters with a median of 1,598 CC.

The vast majority of vehicles were right hand driven as shown in Figure 1.

Figure 1 Distribution of accident severity by vehicles driving orientation



3.1.2 Driver features

Figure 2 shows the distribution of accident severity by driver's age group. We observe that the vast majority of drivers were between 26 and 55 years of age

Figure 2 Distribution of accident severity by age group of drivers

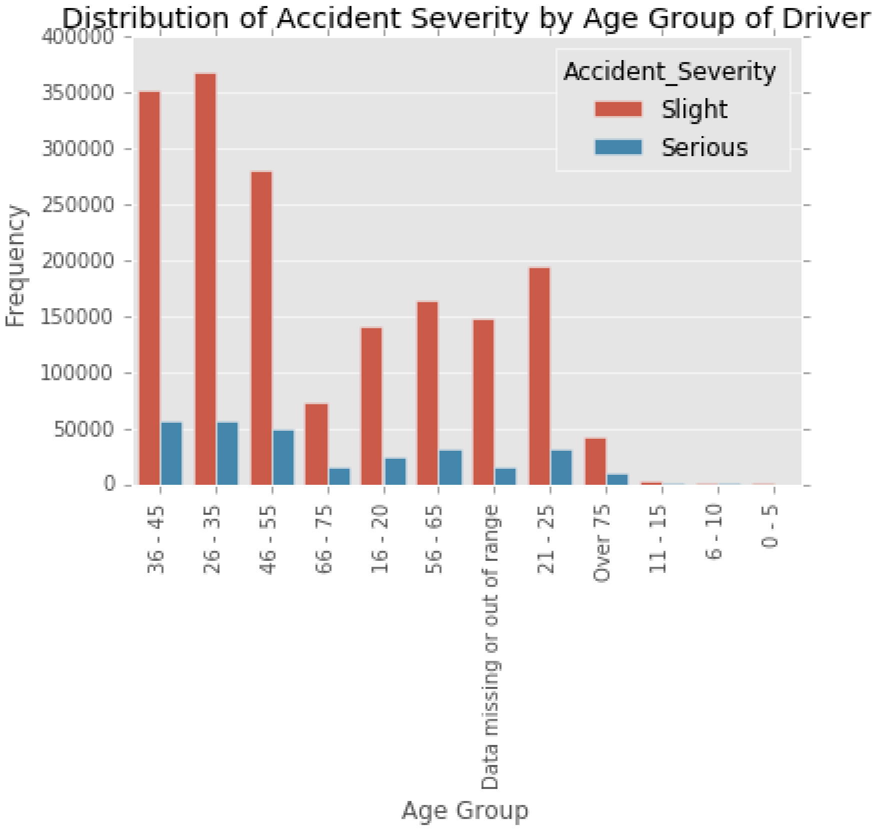
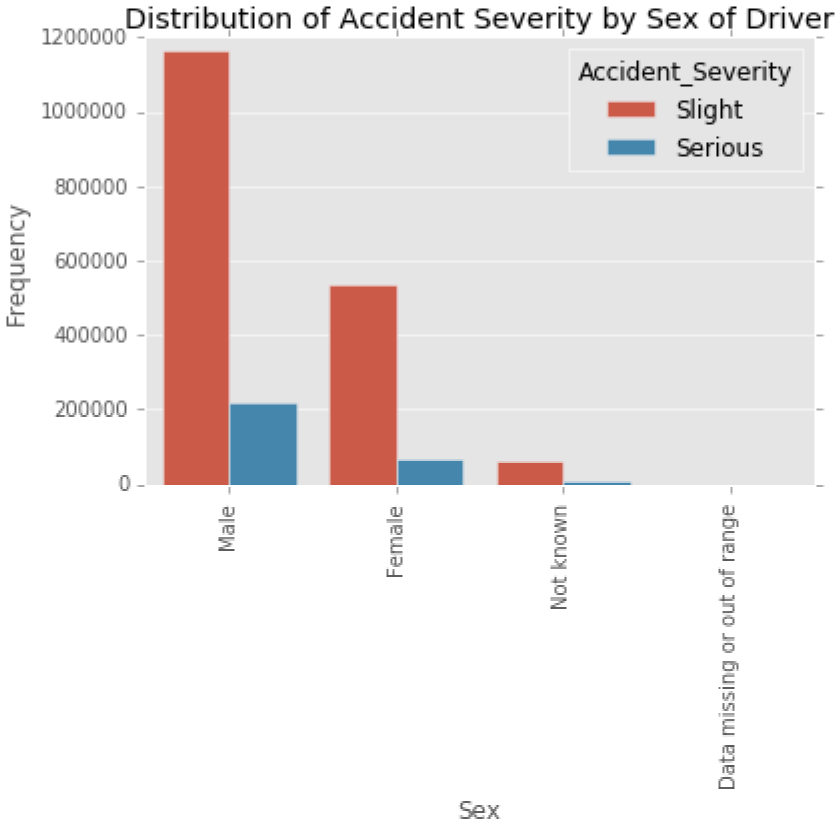


Figure 3 shows the distribution of accident severity by sex of the driver. We observe that the vast majority of drivers were male.

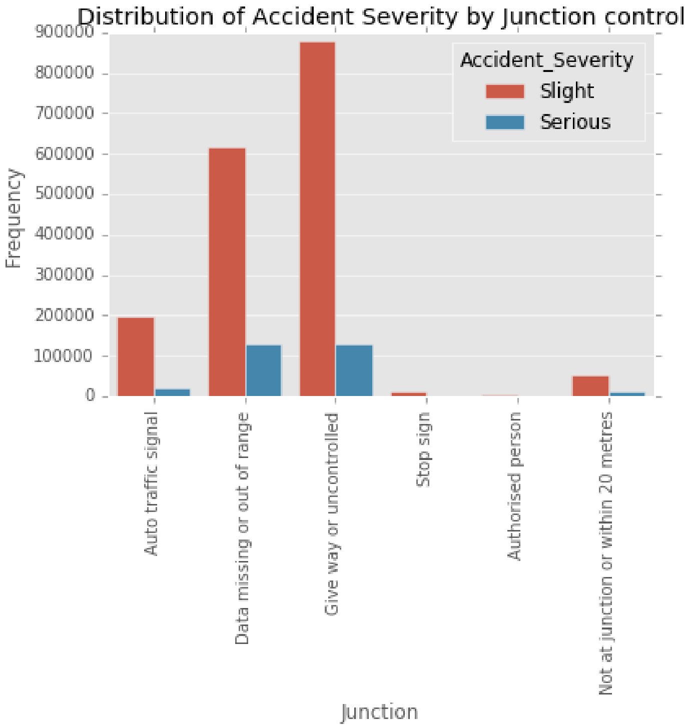
Figure 3 Distribution of accident severity by sex of driver



3.1.3 Road and weather features

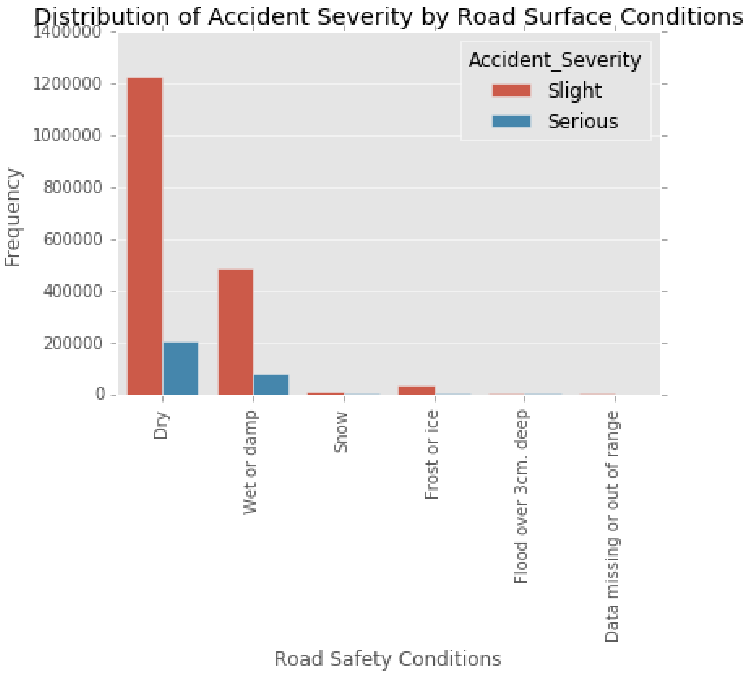
From Figure 4 we see that the majority of the accidents occurred at *Give way or uncontrolled* junctions. In terms of the details of the junctions, most were either not at junction or at T or staggered junctions.

Figure 4 Distribution of accident severity by junction control



We also see that most accidents happened on dry, wet or dump road surfaces as shown in Figure 5.

Figure 5 Distribution of accident severity by road surface conditions



The final road feature I looked at was whether the road was in an urban or rural area. About two-thirds of the roads were in urban areas.

In terms of weather conditions, the majority of accidents happened on days when the weather was fine with no high winds.

3.2 Feature engineering

3.2.1 Dealing with missing data

From the exploratory data analysis, we observed that there some missing values were coded as *data out of range or missing.* The first step I did in feature engineering was to recode all those values to missing.

Given the volume of data available for modelling, I dropped all rows with missing values instead of using imputation techniques to replace missing values.

3.2.2 Numeric features

Numeric features were normalized as part of feature engineering.

3.2.3 Time features

The dataset had a time variable to indicate what time the accident happened. To make it useful in modelling, I divided the time into 5 groups to reflect various activities of the day. This included early morning (5 am to 9:59 am), business hours (10 am to 2:59 pm), rush hour (3pm to 6:59pm), evening (7pm to 10:59pm). The remaining time period was coded as night.

3.2.4 Categorical features

Processing of categorical features started with dropping rows that has missing information on some selected features, for example, road and weather conditions. Dummy variables were created for categorical features through one hot encoding.

3.3 Modelling

3.3.1 Splitting the data

The first step in modelling was to split the data into training and testing sets. As indicated in the data section, the data from 2005 to 2013 was used as the training set and the data from 2014 to 2016 was used as the testing set. That meant I used approximately 71% of the dataset for training the models and 29% for testing.

3.3.2 Handling imbalanced classes

From the exploratory data analysis, we observed that the data was very unbalanced. I applied a synthetic minority oversampling technique to address the data imbalance. SMOTE was only applied to the training dataset. Table 2 shows the distribution of the target variable before and after SMOTE. Applying SMOTE increased the minority serious accident class as shown in Table 2.

Table 2 Distribution of target variable before and after SMOTE

Class	Before SMOTE	After SMOTE

Serious	89,014	614,213
Slight	614,213	614,213

3.4 Model Estimation and Evaluation

Given that the target variable I was trying to predict was categorical, classification models were estimated. Three models were estimated on the resampled training dataset. The models were decision tree, random forest, and logistic regression. The fitted models were then used to make predictions using the test dataset. The weighted f1-score and confusion matrix were example of metrics used to evaluate the models. In addition, I identified which variables were the best predictors of the severity of accidents to inform policy recommendations.

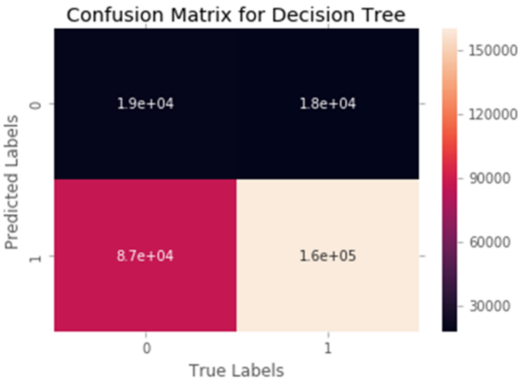
4.0 Results and discussion

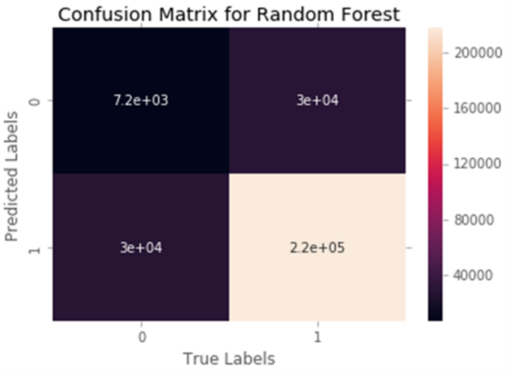
Based on the evaluation metrics, the best performing model was the random forest as shown in Table 3.

Table 3 Classification report for selected models

Model	Weighted average recall	Weighted average f1-score
Decision Tree	0.63	0.69
Random Forest	0.79	0.79
Logistic Regression	0.67	0.72

Confusion matrices for all three models are shown below.





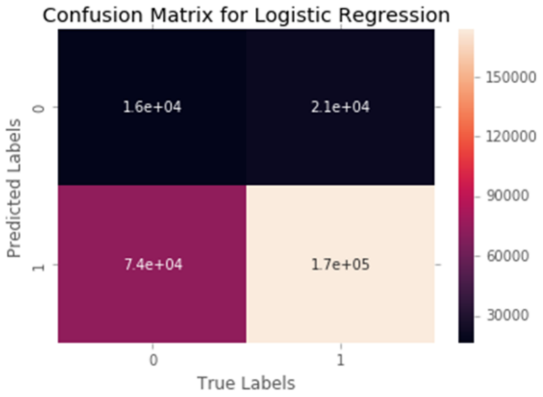
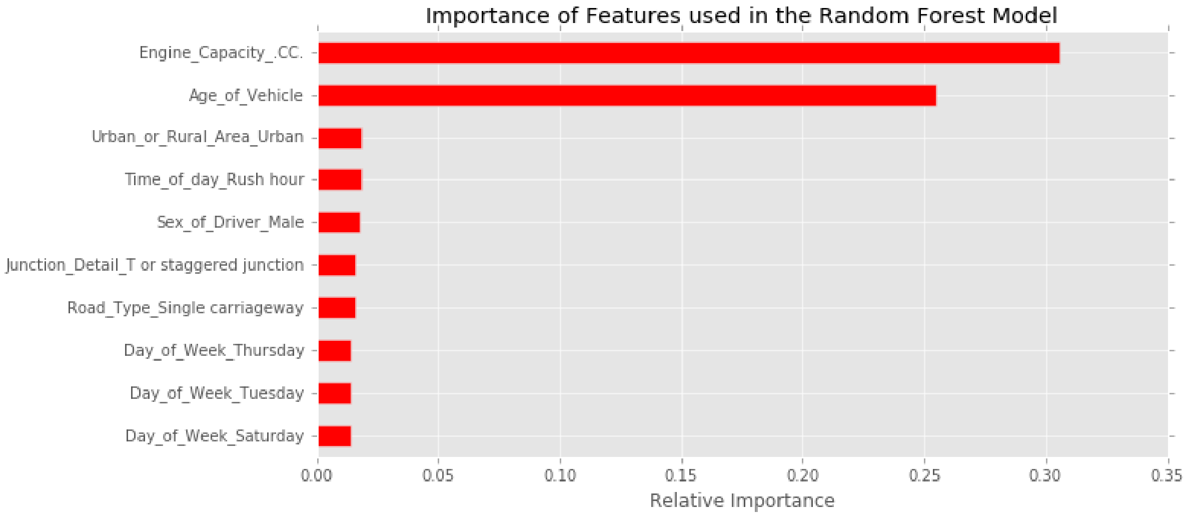


Figure 6 shows the top 10 best predictors in the random forest model. Engine capacity and age of the vehicle had the top 2 highest relative importance in the model signalling a need to focus on vehicle characteristics. Urban road was the next important feature. None of the weather related features showed on the top 10 best predictors of accident severity.

Figure 6 Relative importance of features (top 10)



5.0 Conclusions

This project has developed a machine learning classification model for predicting the severity of road accidents using datasets from the UK. The model has shown that features that are within the control of public transportation and safety authorities like vehicle, road and driver characteristics are the best predictors of the severity of road accidents. Features that cannot be controlled by public transportation authorities such as weather did not show up in the top 10 predictors. The implication for policy is that, transportation policies should focus on improving vehicle, driver and road characteristics to ensure that accident severity is reduced.