UK Car Accidents

CS 210 Group Project

Welcome!

Welcome to our project blog! This site is created for CS 210 Group Project and the details about our Group Project will be given via this site.

About our Project

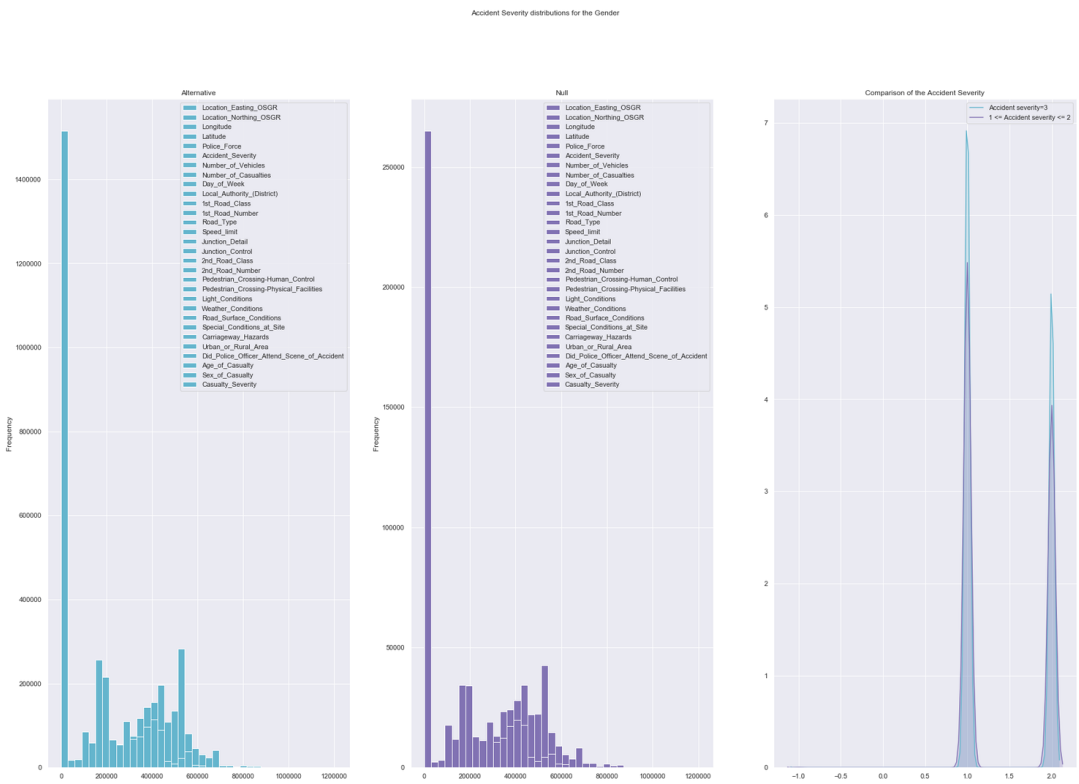
Problem selection is very significant since we are trying to select large data sets with many set points. Therefore, we choose a data set which is called UK Car Accidents between 2005 and 2015. Our data set consists of two different data sources which is called Accidents and Casualties. The casualties file presents a detailed information about the situations of an accident from 2005 to 2015 such as casualty type, sex, casualty severity, casualty age, casualty social class etc. and the accident file presents a detailed information about the situations of an accident from 2005 to 2015 such as location's latitude and longitude, area type, road type, date, time, day of week etc. The motivation for selecting this data set is that the examination of road accidents is important to understand the factors involved and their impact. Accidents usually include multiple variables such as accidents severity, the gender of driver, the age of driver that's why it is difficult to analyze the data. As we can see from the graph with variables of gender of driver and accident's severity, it would come up with an interesting result because usually the expected result has a tendency of females cause more accidents than males, but the result is opposite. In the following visualizations, it is attempted to demonstrate how the variables interact with each other.

Various visualizations can be built in order to describe the data sets. You can reach our visualizations regarding to our datasets from the following link: https://nbviewer.jupyter.org/gist/ftalay/1f64aaf055ea616ae03d32f2f9fde3d8

Hypothesis Testing

Hypothesis Testing is applied whether the attributes have a relationship between each other. Hypothesis tests are based on a statement called null hypothesis which presumes nothing interesting is going on between whatever variables you are testing. Two cases are tested with our data. Firstly, we wanted to find out whether the gender affect the accident severity. Secondly, the relationship between weather conditions affect casualty.

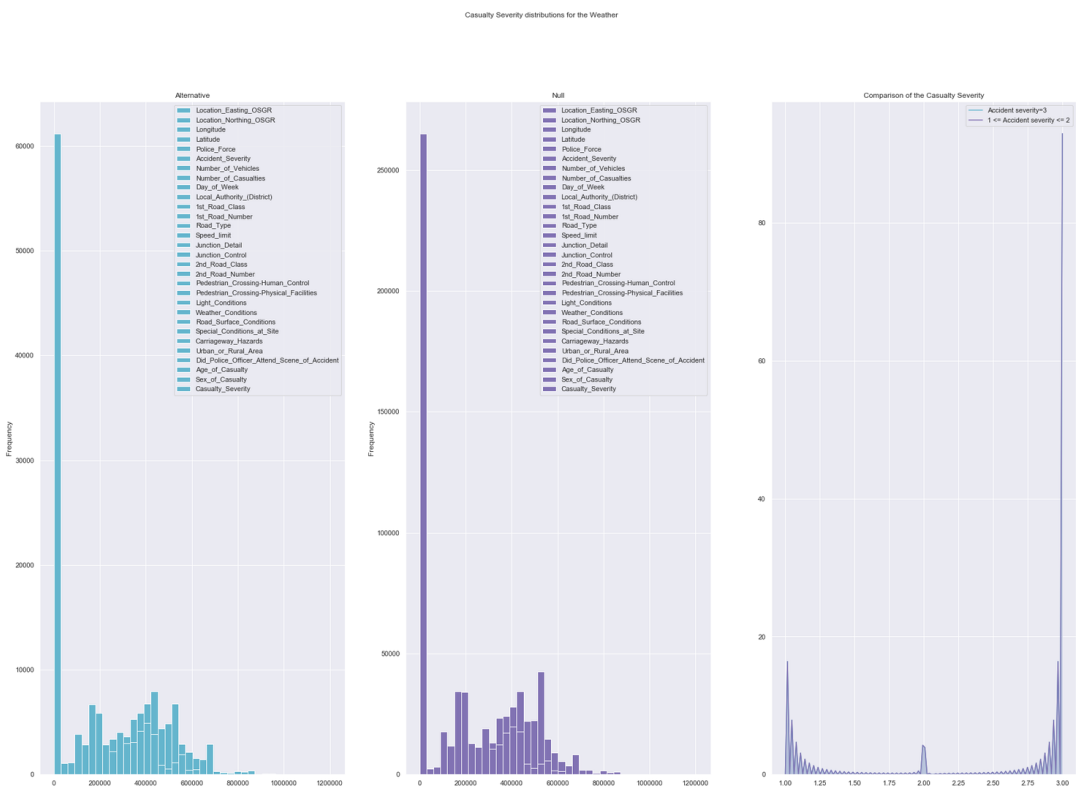
Effect of gender on accident severity is shown in the following figure. The accident severity values are measured on average.



Accident Severity Distribution for the Gende

```
The resulted p-value is 0.28778083775536534. Since it is greater th
```

```
In the second example, the effect of weather conditions on casualty
```



Casualty Severity Distributions for the Weather

```
The resulted p-value is 0.5877458515943549. Since it is greater tha
```

Linear Regression

Linear regression is generally applied to quantify the relationship between two or more variables. We will try to determine which measured effects most cause the accidents in UK between 2005 and 2015 by using linear regression. In this section, MSE (mean squared error) are going to be discussed. MSE means that it can be defined as measure of the average of the squares of the errors—that is, the average squared difference between the estimated values and what is estimated.

The following effects are measured in order to see the relationship between the accident severity: “Number_of_Vehicles”, “Light_Conditions”, “Weather_Conditions”, “Urban_or_Rural_Area”, “Age_of_Casualty”, “Sex_of_Casualty”, “Casualty_Severity”.

The mean squared error (MSE) is shown as follows:

```
0.1576933590880602
```

The same effects are also measured in order to see the relationship between casualty severity. The effects are as following: “Number_of_Vehicles”, “Light_Conditions”, “Weather_Conditions”, “Urban_or_Rural_Area” , “Age_of_Casualty”, “Sex_of_Casualty”, “Accident_Severity”.

The mean squared error (MSE) is shown as follows:

```
0.1341912704479353
```

The mean squared error tells you how far is a set of points to regression line. In these two cases second one is more close to 0 so it can be said that the set of points is more close to the regression line than in the first case.

The completed version of the code can be seen from the following link:

https://nbviewer.jupyter.org/gist/ftalay/8363d77411c3fb7a18ce76bdb21c8391

Building Two Models with a ML Technique

Decision Tree and Random Forest are used as two machine learning technique. Decision tree is a learning method used for classification and regression.

Decision tree constructs regression or classification models in the shape of tree structure. It splits up a data set into smaller subsets while simultaneously an associated decision tree is progressively developed. The final outcome is a tree with leaf nodes and decision nodes.

Decision Tree is a subset of Random Forest and Random Forest grows many classification trees. Basically, we can say that Random Forest made of many decision trees. Instead of averaging the prediction of trees, this model applies two essential concepts that gives it the name random:

1. Random sampling of training data points when building trees
2. Random subsets of features considered when splitting nodes

Verifying our Models

Cross-validation is used for two of our machine learning techniques. For the cross validation of decision tree: The cross-validation score for decision tree regressor is negative but it does not mean that model is not successful, it is about the library.

Describing the Performance of the Machine Learning Techniques

Random Forest Model performs better since its accuracy score is better than decision tree model which is 0.7323346002821433. By Random Forest, we train more than one decision tree on a random feature subset with random points. On the prediction part, predict from each decision tree and use techniques such as, majority voting etc. to use their common ideas. Therefore, Random Forest model performs better.

As we could observe from Feature Importance Rankings; as much as we increase ‘Age_of_Casualty’ column, RandomForest accuracy rate increases. Therefore, ‘Age_of_Casualty’ feature works best. Since, decision tree is a subset of RandomForest whatever we conclude for RandomForest we could directly mention also for decision tree. Thus, ‘Age_of_Casualty’ also works best for decision tree.

Edit