MIS 620 Group Project Proposal

Project Title:

Accident Rate Analysis

Team Members:

Shraddha Masuti

Lee Parker

Rashmi Sawant

Deepika Siriah

Contents

1. Team Information

1.1 Team Name

Accident Rate Analysis

1.2 Team Members and Expertise

Shraddha Masuti – Data Analyst
Lee Parker – Project Manager
Rashmi Sawant – Data Analyst
Deepika Siriah – Business Analyst

2. Data Analytics Problem

2.1 Executive Summary

The UK government amassed accidents data from 2012 - 2014, on over 1.6 million accidents in England. This project dataset contains 3 years of data for every recorded major accident in England from 2012 to 2014, including detailed information about the accidents.

This project will analyze and process the data to discover the statistical relevance of the various factors on accidents in England. As there is very large amount of data, the data consistency must be checked. Also, the data consists of codes for which master data must be prepared and mapped to the raw data. This is the most time consuming and critical task as it provides the base data for further modeling and analysis. The objective of the project is to discover and describe trends in major road accidents, and to create a model that can accurately predict rate of road accidents. The purpose of creating this model is to provide information to the government that can be used to take appropriate preventative measures.

2.2 Data Sources

The data is collected from the Kaggle website. The original dataset, available as a 'csv' file, is observed as structured. The raw data consists of variables like location of the road accident, number of causalities, date, time, road type, road surface condition etc. Once the data is processed and cleaned it will be suitable for use in R programming.

2.3 Technical Resources

Data Preparation:

For analyzing the data set, the data will be stored in a csv file. The data processing and cleaning will be performed using R or SQL. Further, R will be used for predictive analysis.

Model Planning:

Out of the 3 years of data, exploration will be performed to learn about the relationships between variables, and to select the appropriate data segments. Furthermore, we can select the most suitable models using R.

Model Building:

We plan to use R to develop data sets for testing, training, and production purposes. The model building and execution will also be done using R.

2.4 Initial Hypothesis

The purpose of the project is to analyze the data to find the relationships between various factors such as location of the road accident, number of causalities, date, time, road type, road surface condition, weather conditions etc. The interrelationships in the data will then be studied to determine whether the data is sufficient for trend analysis that will enable a model to be built that can accurately predict road accident rates.

H_1: There is no correlation within the dataset between the factors and road accidents.

H_2: The data analysis does not provide the trends to enable a model to predict rate of road accidents greater than chance (i.e. <= 50%).

3. Management

3.1 Challenges

- Insufficient data. There is insufficient data about the causalities involved like gender, age, which should be a factor in predicting rate of road accident. The model will be built without this factor included until the data becomes available and can be incorporated.
- Overfitting/oversampled data. Due to the volume of data available, there is a risk of overfitting data to the model. To mitigate, the data will be segregated into smaller samples.
- Incomplete data. Some columns have many rows for which there is no data. Variables that have many missing values will be removed from the dataset and not used.

3.2 Objectives

- The objective of the project is to demonstrate the relationship of the factors to the road accidents rate, determine trends, and build a model that can predict road accidents at an accuracy rate greater than chance. The model would be able to demonstrate trends to aid in the creation of preventative measures.
- Success criteria for this project is to determine if the hypotheses can be rejected, and if so, to create a model that can predict road accidents at an accuracy rate greater than chance.
- Failure criteria for this project is to not be able to reach a conclusion about if the hypothesis can be rejected.