Assignment 3: Scale Up

In this assignment, you will practice how to prepare and analyze larger and more complex datasets in a scalable manner using Apache Spark. In this assignment, you will perform analysis tasks on a UK Road Safety dataset utilizing both Spark SQL and Python, and train a predictive model using MLlib.

This homework is due on **April 19th, at 11:59 PM EST**.

Please upload the submission as a single .zip file to CMS. It should include:

1. A report that includes the figures you generated, the descriptions you think are necessary and your answers to the questions that require your respond.
2. The IPython notebook of your experiments with all the figures and results.

While you can run Spark locally, using DataBricks Community Edition[1] is our recommended way of carrying out this assignment. It saves your time to set up the environments and it should be adequately fast for this assignment.

Introduction

The UK government collects and publishes (usually on an annual basis) detailed information about traffic accidents across the country. This information includes, but is not limited to, geographical locations, weather conditions, type of vehicles, number of casualties and vehicle manoeuvres.

The dataset comprises of two files:

- **Accident_Information.csv**: every line in the file represents a unique traffic accident (identified by the Accident_Index column), featuring various properties related to the accident as columns.
- **Vehicle_Information.csv**: every line in the file represents the involvement of a unique vehicle in a unique traffic accident, featuring various vehicle and passenger properties as columns.

The two above-mentioned files/datasets can be linked through the unique traffic accident identifier (Accident_Index column).

More information about the dataset can be found here[2].

1 https://community.cloud.databricks.com/

2 https://www.kaggle.com/tsiaras/uk-road-safety-accidents-and-vehicles

Data Analysis

Use SQL queries to perform the following analysis. Report the results of the query in a table and add a visualization (of your choice) that best portrays the data. Ignore any rows that lack relevant fields.

1. The average road speed limits at the accident site, under different lighting conditions. Give a reasonable explanation of the result.
2. A breakdown of the number of accidents according to accident type (“Serious” versus “Slight” accidents) and driver age.
3. The serious accident ratio of cars from top 50 most common manufactures.
4. [BONUS] Generate a 2D heat map that shows the accident rate at different locations. Overlay the heat map on a UK map.

Severeness Prediction

Based on MLlib, you will build a model that predict the severeness of an accident given other information. For the choice of models, you are recommended to choose between random forest and logistic regression model.

For the choice of features, you are free to choose from whatever feature you think are appropriate. However, be sure to include at least one categorical feature and one numerical feature.

Report the ROC and AUC of your model. Perform a feature importance analysis and briefly describe the result in your report.

Additional Resources

Data Uploading

You will need to upload and load the data CSVs into spark dataframe before any processing. It is recommended to upload the zipped dataset and unzip it on the cloud, especially if you have a slow connection. Here is a sample script that unzips the data on DBFS.

```
%sh
unzip /dbfs/FileStore/tables/Accident_Information_csv-84de4.zip -d /dbfs/FileStore/tables/
unzip /dbfs/FileStore/tables/Vehicle_Information_csv-d57fc.zip -d /dbfs/FileStore/tables/
```

```
Archive:  /dbfs/FileStore/tables/Accident_Information_csv-84de4.zip
  inflating: /dbfs/FileStore/tables/Accident_Information.csv
Archive:  /dbfs/FileStore/tables/Vehicle_Information_csv-d57fc.zip
  inflating: /dbfs/FileStore/tables/Vehicle_Information.csv
```

Databricks Tutorial

Databricks Quickstart Tutorial contains most of the information on how to run Python and SQL queries on its cloud platform. Be sure to check it out if you are not familiar with their notebook system and their file system.

Limitations of Databricks Community Edition

One of the biggest limitations of community edition is that an idling cluster will be shut down automatically after 2 hours of inactive. After that you will have to create a new cluster and attach your notebook to the new cluster as the cluster being shut down is not allowed to be started again. While this is usually not an issue, it is always a good idea to keep it in mind.