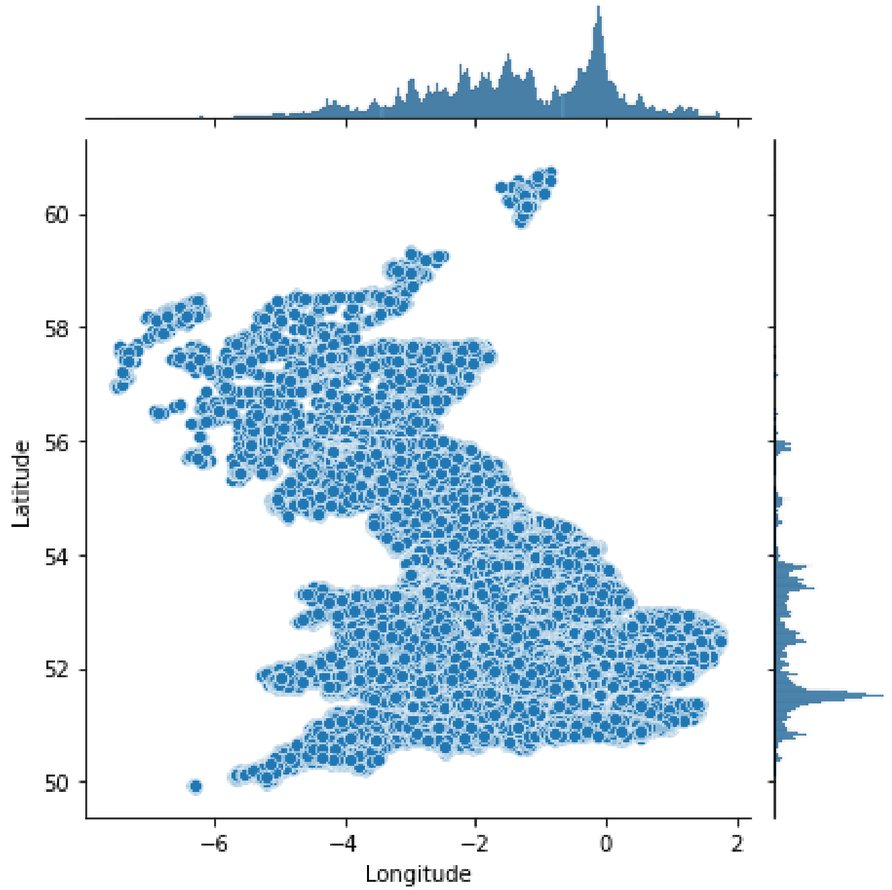
Data Analytics project
2021

UK Road Accidents Analysis



Utkarsh Gupta - PES1UG19CS549
Varun Mk - PES1UG19CS558
Harshita Vidapanakal - PES1UG19CS185

UK Car Accident Data

Traffic accidents are the leading cause of death for the age group 15-29 year. Data regarding these accidents can drive analysis and shed light on this issue

With the advent of technology in surveillance, data regarding circumstances of road accidents and consequential causalities was able to be consolidated

This has provided us with vast amounts of data to assist in accident prevention techniques

Objectives

Identify factors affecting the frequency and scale of accidents to perform time series prediction of number of crashes/year

Predict severity for any accident given the conditions it occurs in

Detect "Hot spots" which are places identified by a high accident occurence

Dataset

- Comprises of 2 .csv files
- AccidentInformation.csv - 34 attributes
- Vehicle_Information.csv - 24 attributes
- Date Range: 2005-2017 and 2004-2016 respectively
- The 2 files can be linked through the unique traffic accident identifier : Accident_Index column
- Kaggle Link : https://www.kaggle.com/silicon99/dft-accident-data

Approach

1

ML Models to predict the severity of an accident as 'Fatal', 'Serious' or 'Slight'

2

- Naive Bayes Classifier
- Logistic Regression
- Adaboost
- XGBoost
- Random Forest

3

Quartic weighed KDE plots are used to analyse accident hotspots over the years.

4

Time Series Analysis using ARIMA and SARIMA models

Approach

1

ML Models to predict the severity of an accident as 'Fatal', 'Serious' or 'Slight'

2

- Naive Bayes Classifier
- Logistic Regression
- Adaboost
- XGBoost
- Random Forest

3

Quartic weighed KDE plots are used to analyse accident hotspots over the years.

4

Time Series Analysis using ARIMA and SARIMA models

Evaluation

1 Accuracy, Precision, Recall and F1-Score calculated for every machine learning model

2 ADF Statistic and p-value to determine if data is stationary. Log-likelihood and AIC to determine ARIMA model

3 MSE, MAE, RMSE and MAPE statistics for ARIMA and SARIMA Models

Observations and Insights

We select Random Forest classifier as our optimal model with an accuracy of 92%. We then check for feature importance and explore the relevant parameters.

Through the KDE plots, we explore the East Sussex and Hampshire locations and observe the evolution of hostpots.

We used an ARIMA and SARIMA model for time series analysis and forecasting the number of accidents in Liverpool.

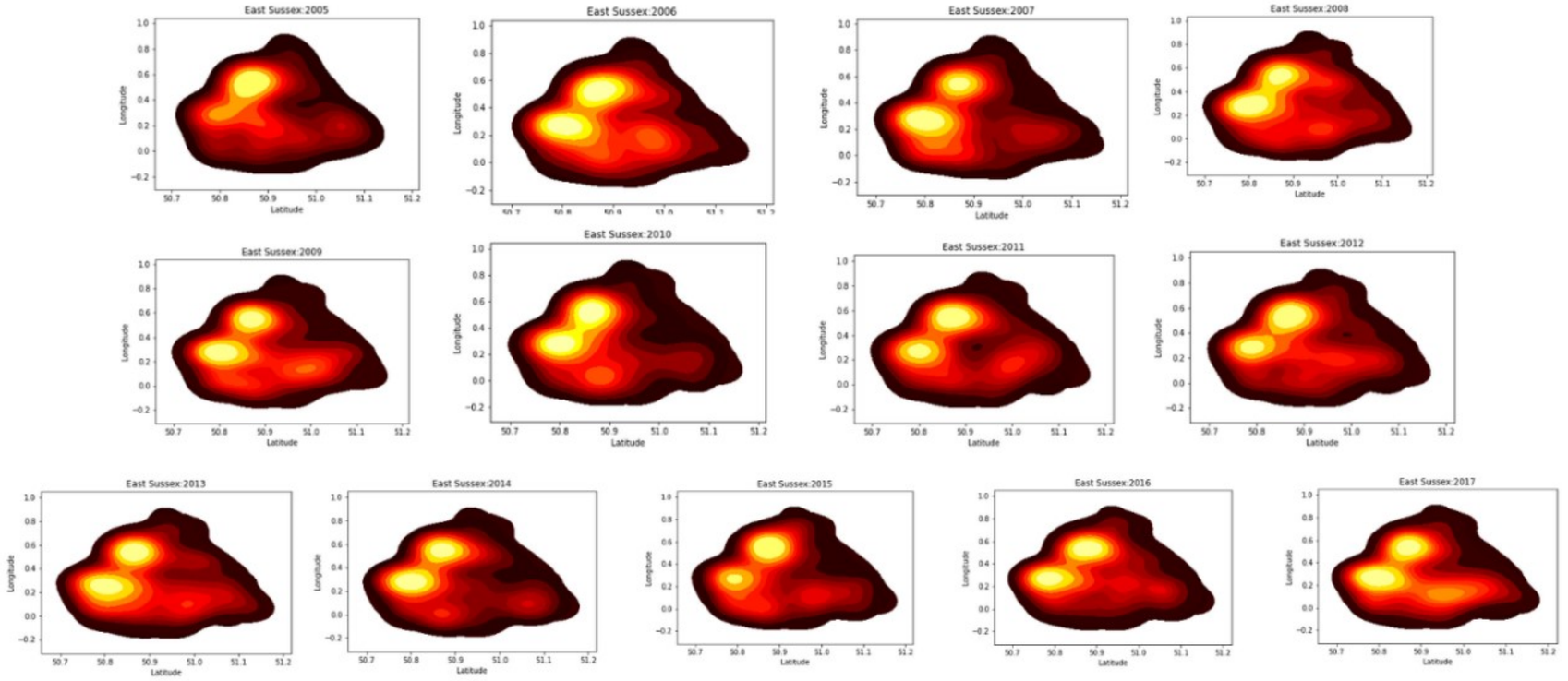
Observations and Insights

Considering the imbalance of the dataset, we synthesize new samples using SMOTE for the minority classes to help create a less biased model.

We check this against the performance of models trained on the original dataset and it outperforms by a good margin.

On seasonal decomposition, we can observe a repeating seasonal pattern. Our SARIMA model performed better with a MAPE value of 0.077.

KDE Plot for East Sussex



Contributions

- Utkarsh aided the data pre-processing tasks and trained classification machine learning models on the dataset to predict accident severity. In addition to this, he worked on the hotspot analysis pipeline with KDE and inferred results for the same
- Varun aided the time series analysis tasks and trained xgboost machine learning model on the dataset to predict severity. He also worked on feature importance and plotting correlation of severity with other important attributes to draw inferences.
- Harshita aided with the EDA and literature survey which helped us gauge what to work on and form our goals for this project, hotspot detection using KDE. She, in addition to this, contributed to some of the machine learning based models used for severity prediction


Thank you!