TRAFFIC PREDICTION

A Project by Team SciPy

Traffic congestion has been a menace in most areas across the world and has generated a lot of hiccups across other sectors including healthcare, transportation, and logistics.

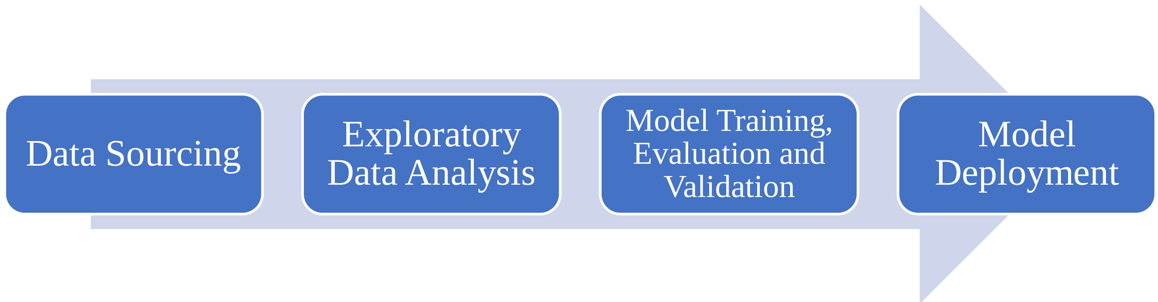
Observations from research have shown that traffic in most areas has taken a pattern, too many people moving on a particular road at the same time will most likely lead to congestion. Various factors contribute to traffic, but increased vehicular activities at a junction are the leading cause of congestion.

Several efforts in the past and recent times have been made to solve the problem of traffic congestion to no avail. Machine learning models that predict the number of vehicles per time at a particular junction will largely help people to know when to avoid a junction and or take an alternative route and in return reduce traffic congestion drastically.

Aims and Objectives

This project aims to deploy a machine-learning model that predicts the likelihood of traffic at a particular time at four different junctions in a city. Past data will be collected from the junctions and processed into usable data using data science techniques, the cleaned data will be used to train several machine learning models and the best model with the highest accuracy will be deployed.

Flow Process



Data Sourcing

This is a process of collecting data from internal, external sources, or a combination of both. For this project, the near-perfect dataset was sourced from Kaggle.

https://www.kaggle.com/datasets/fedesoriano/traffic-prediction-dataset

Exploratory Data Analysis

This includes cleaning the dataset, and analyzing and creating visuals to find insights.

The dataset gotten from Kaggle contained 48,120 rows and 4 columns;

DateTime: this contains the time at which sensors collect traffic data at the different junctions. The sensor collected data every hour.

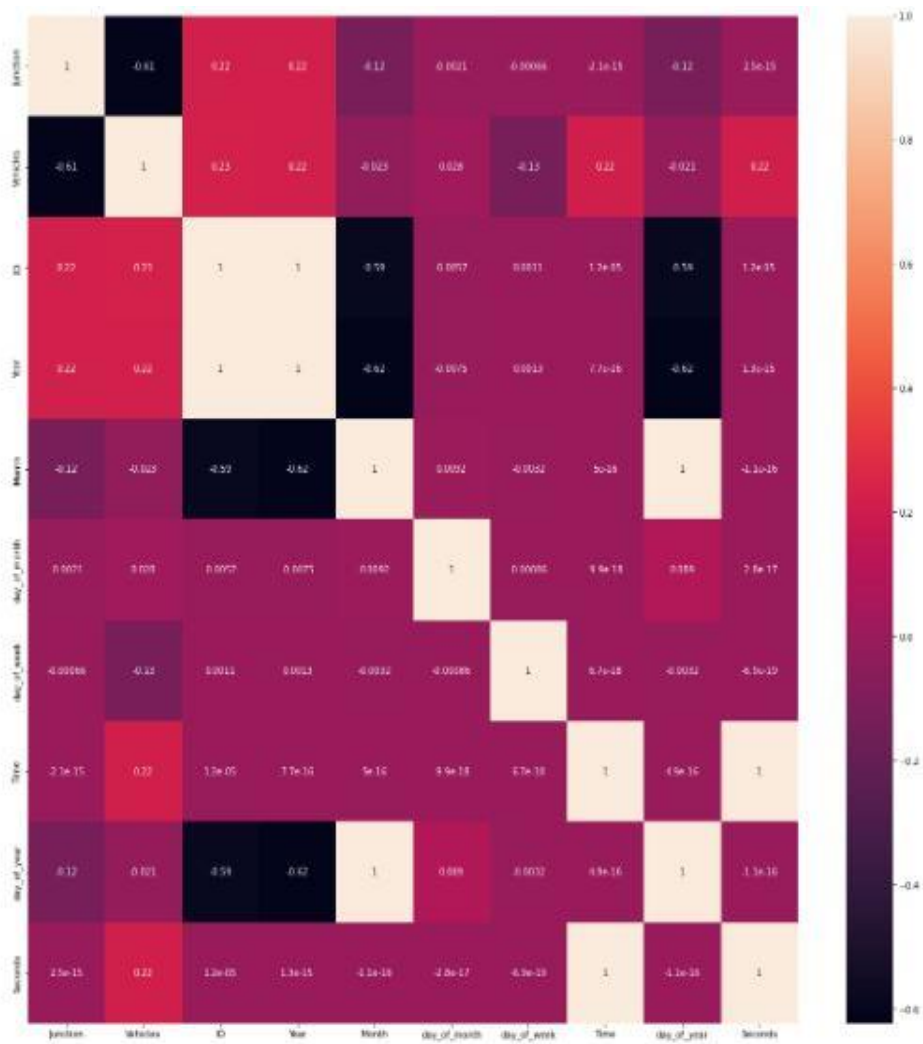
Junction: this represents the four different junctions from which the traffic data was collected.

Vehicles: this represents the number of vehicles at the time the sensors capture the traffic data.

ID: this is the unique ID of the sensors.

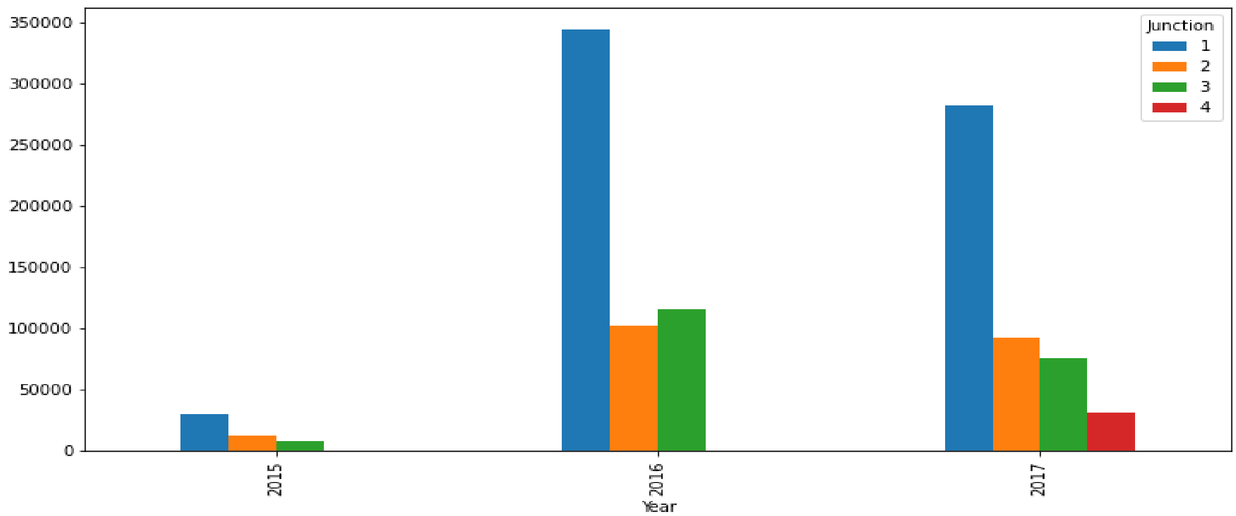
This project involves a time series analysis, so, more columns were engineered for quality analysis. The new columns include Year, Month, Day_of_Month, Day_of_Week, Day_of_Year, Date, Time, and Seconds.

The number of vehicles at a particular time is the most important factor affecting traffic. A heatmap showing the correlation between the number of vehicles and other features is shown below. "Junction 1" had the highest correlation to the number of Vehicles while Day_of_year and Month had the lowest correlation to the target.



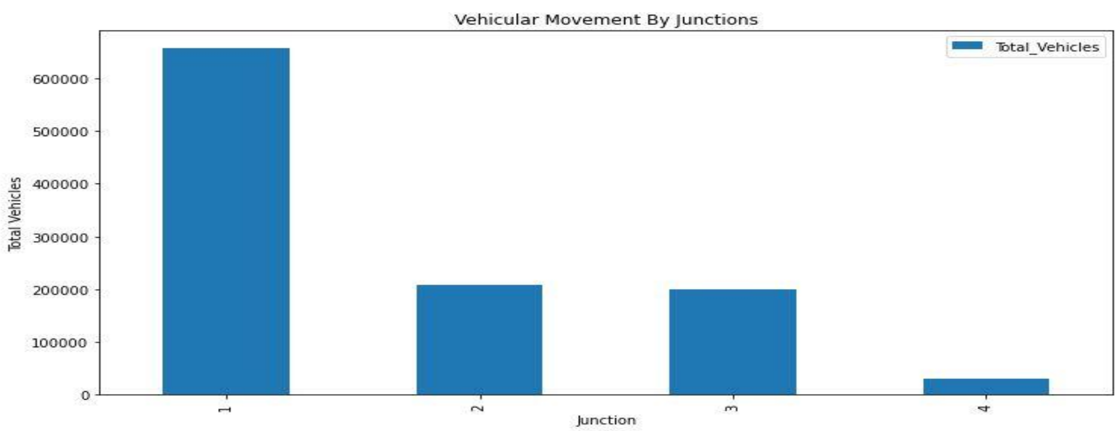
Heatmap showing the correlation between variables

Below are the charts showing the relationship between the Number of Vehicles at different times by junction.

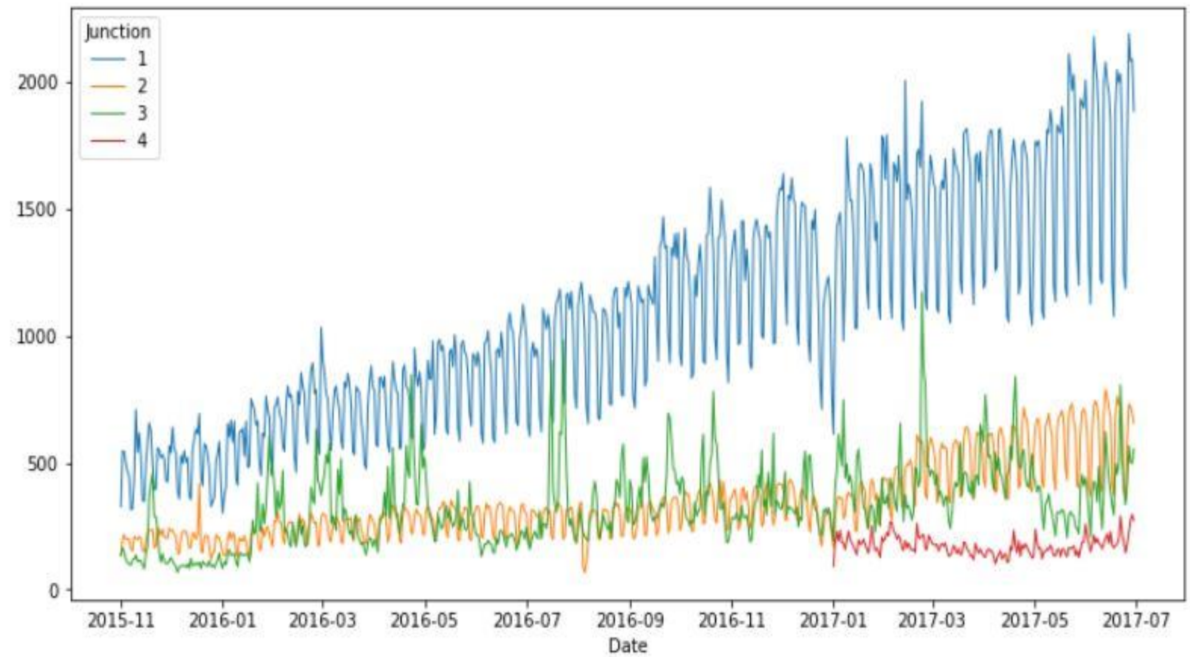


The Year Bar Plot above showed a sharp increase in Traffic Movement for Junction 1 through 3 for 2016 as compared to 2015 which dropped for 2017 with the addition of junction 4.

Sdqa

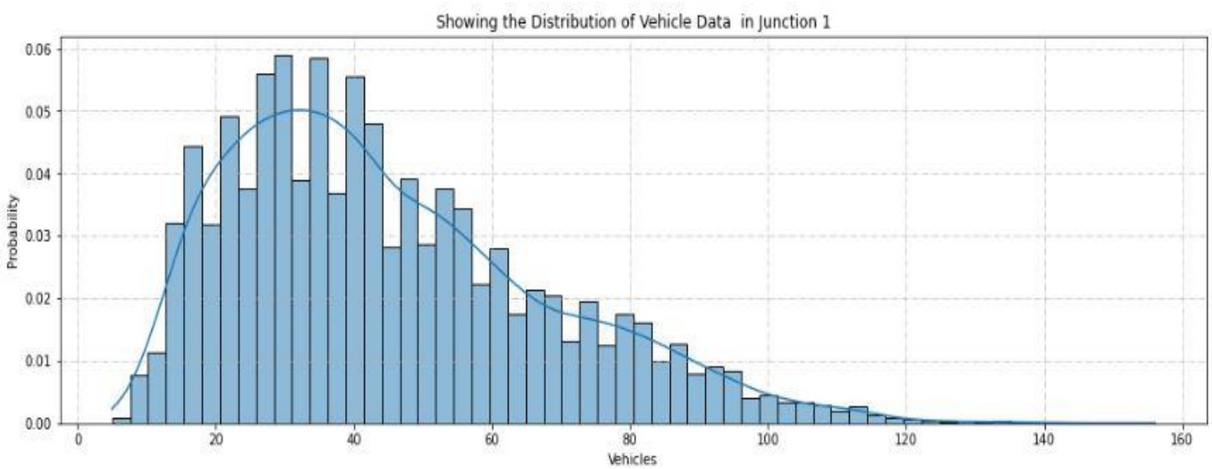


A bar chart showing the vehicular movement at different junctions

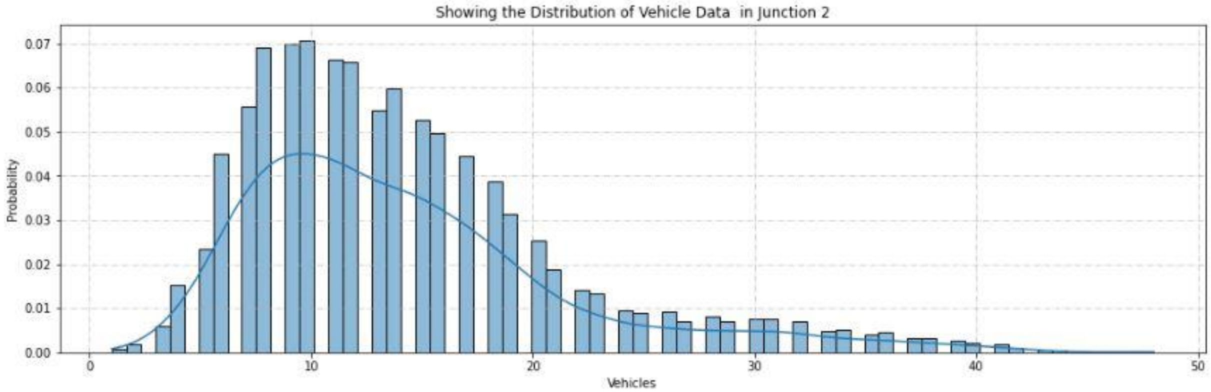


Vehicular movement at junctions between 2015-11 and 2017-07

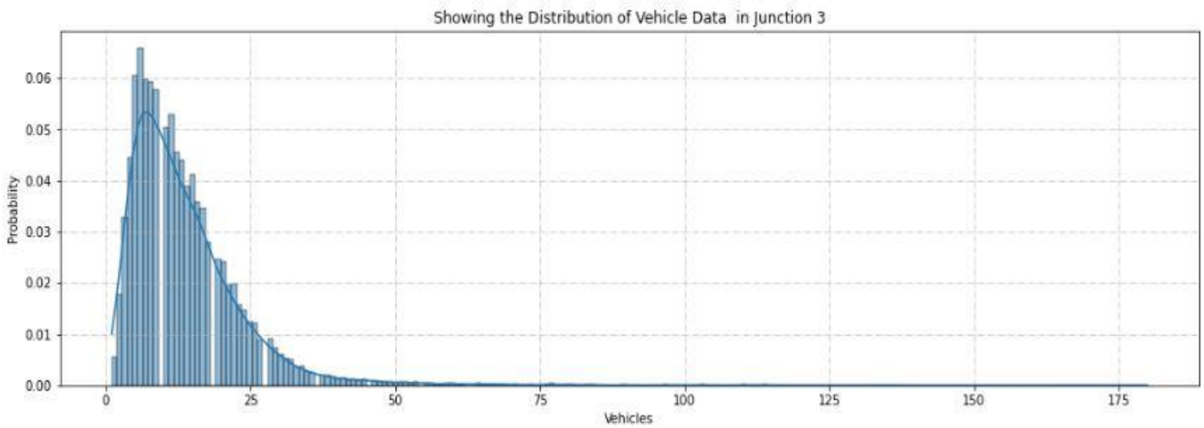
Charts showing the distribution of Vehicles at different junctions



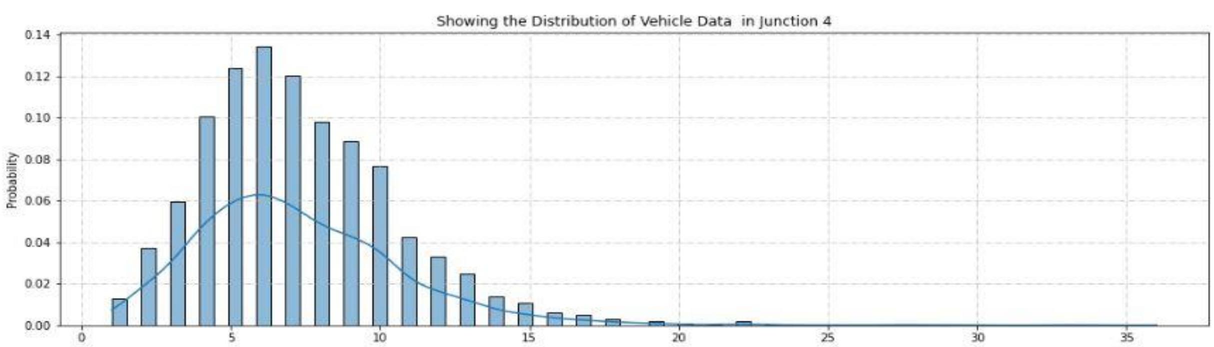
Junction 1



Junction 2

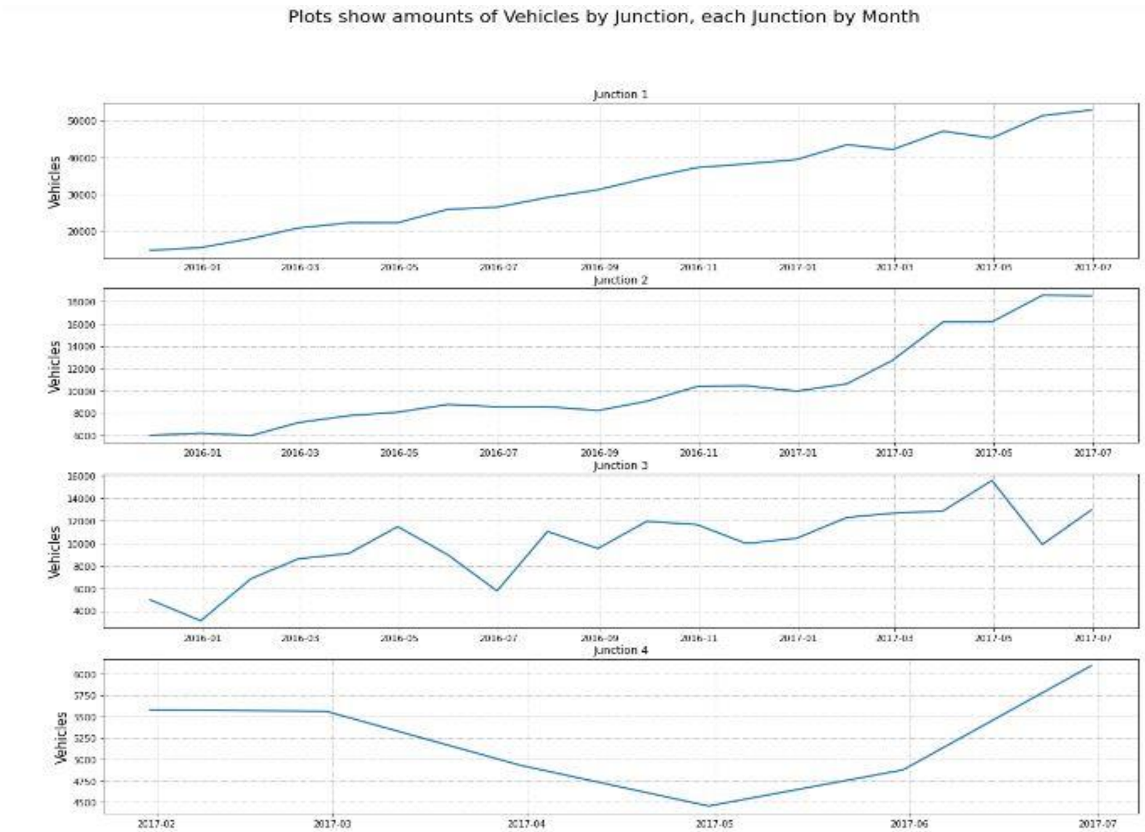


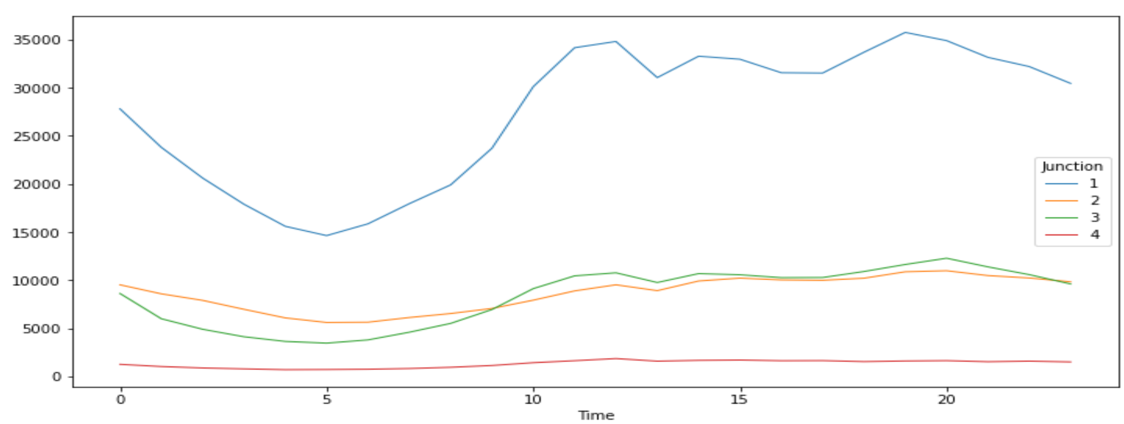
Junction 3



Junction 4

Plots showing the number of vehicles at each junction per month





Chat showing traffic at different junctions per hour

The Hourly Analysis Shows the Least Traffic situation during the early period of the day, with peak traffic experienced at the later hour of the day

Model Training and Validation

Using target encoding, more features with aggregate functions (STD, Max, Min, Mean, and Median) were engineered and the dataset was split to represent the four junctions differently.

Baseline models were built for each junction. The tables below show the baseline models and their Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Average Score.

JUNCTION 1

Model	RMSE	MAPE	Average Score
RandomForest	7.010026	7.689260	7.349643
LGBM	7.026064	7.906487	7.466276
GradientBostimg	7.481111	8.651360	8.066235
XGBoost	8.201345	8.987725	8.594535
DecisionTree	9.052603	9.769647	9.411125
AdaBoost	10.104497	13.934631	12.019564
Prophet	13.790623	14.182331	13.986477
LinearRegression	12.854838	15.230175	14.042507
Ridge	12.695813	15.475677	14.085745
Lasso	12.790993	17.673074	15.232033
CatBoost	17.109869	17.215516	17.162693
ARIMAX	18.234103	28.908571	23.571337
SVR	37.676057	38.914948	38.295502
LinearSVR	71.207159	100.000000	85.603580

JUNCTION 2

Model	RMSE	MAPE	Average
LGBM	4.583478	13.839091	9.211284
RandomForest	4.932235	15.287121	10.109678
XGBoost	4.632021	15.688251	10.160136
GradientBoosting	5.061104	15.636265	10.348685
DecisionTree	5.617521	17.554998	11.586259
AdaBoost	5.839895	19.045331	12.442613
CatBoost	7.177207	19.705083	13.441145
LinearRegression	7.627751	21.368942	14.498346
Ridge	7.931942	22.272597	15.102269
Lasso	8.612488	25.109018	16.860753
Prophet	6.579840	29.841092	18.210466
ARIMAX	7.391952	33.855458	20.623705
SVR	13.920349	41.996867	27.958608
LinearSVR	25.088684	100.000000	62.544342

JUNCTION 3

Model	RMSE	MAPE	Average Score
LGBM	9.437156	39.265525	24.351340
LinearRegresssion	8.418616	41.727789	25.073203
Ridge	8.5294755	44.654804	26.592139
CatBoost	10.028169	45.982760	28.005465
SVR	12.075476	44.539007	28.307241
GradientBoosting	9.925743	46.990505	28.458124
Lasso	8.581689	49.206843	28.894266
RandomForest	11.028847	52.278331	31.653589
XGBoost	10.644478	57.295947	33.970212

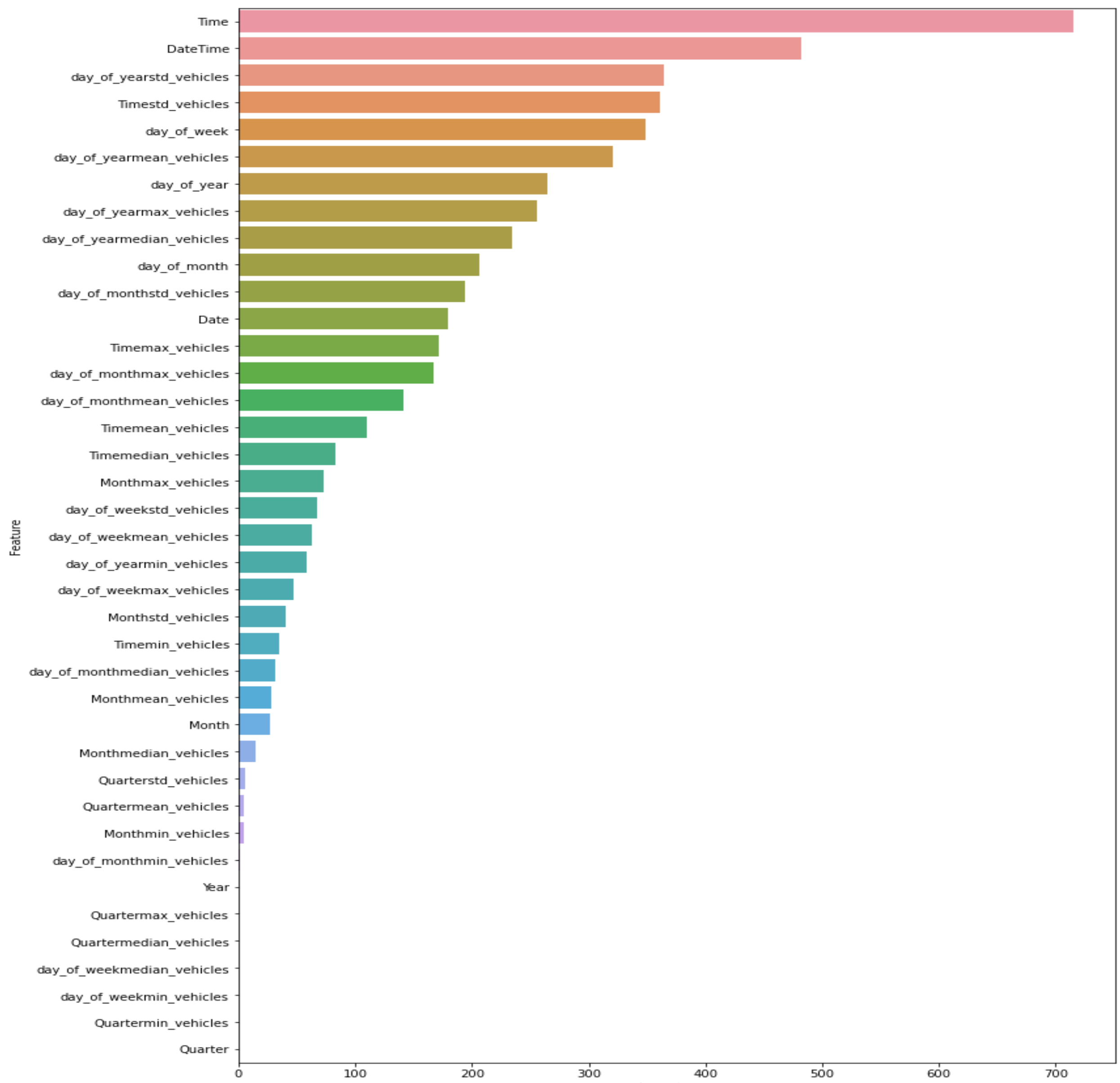
ARIMAX	10.275442	70.132817	40.204130
DecisionTree	14.463622	67.261362	40.862492
AdaBoost	12.682227	76.035432	44.358829
LinearSVR	20.413881	100.00000	60.206940
SVR	15.943360	122.272906	69.108133

JUNCTION 4

Model	RMSE	MAPE	Average Score
LGBM	3.279228	28.993053	16.136140
GradientBoosting	3.298566	29.364646	16.331606
RandomForest	3.303135	30.246825	16.774980
Ridge	3.646825	31.389219	17.518022
LinearRegression	3.646906	31.399350	17.523128
CatBoost	3.547152	31.749286	17.648219
Lasso	3.691968	31.641518	17.666743
ARIMAX	3.421977	33.179423	18.300700
XGBoost	3.330306	33.964769	18.647538
AdaBoost	3.366164	40.539051	21.952608
DecisionTree	4.777540	42.351461	23.564500
Prophet	3.403392	46.238255	24.820824
SVR	4.428277	47.955974	26.192125
LinearSVR	9.197400	100.000000	54.598700

LGBM has the best baseline performance for Junctions 2, 3, and 4, so it was chosen and tuned for each junction. RandomForest had the best performance for Junction 1, followed by LGBM, but LGBM outperformed RandomForest after tuning both models and was chosen and optimized to improve performance.

We checked the importance of all the features to see which ones added noise to our models and we noticed the features with no importance to the models and dropped them.



A bar chart showing features and their importance

Model Deployment

The Model was deployed using the Streamlit library in python on the Streamlit cloud to enable users to make live predictions. See the link below:

https://team-scipy-traffic-predictor-ap.streamlitapp.com/

Result

From the analysis, the increased number of vehicles is the leading cause of traffic congestion. More vehicles were present in Junction 1 while Junction 4 had the least number of vehicles, there has been an upward trend of vehicles yearly in all four junctions with junction 1 having the highest upward trend

We notice a daily increase in Vehicular movement in all the junctions except "Junction 4" which started recording data in January 2017.

Traffic flow was observed to be steady across all junctions until the Fourth day (Thursday) where there is a sharp drop in movement till the rest of the week Except junction 4, We notice the data increasing during the morning time, around 6 am, staying steady throughout the afternoon, and decreasing during the evening time around 8 pm.

We also notice that we have less traffic during the weekend and steady traffic during the weekdays.

Junction 4 was created to reduce the overall traffic situation on the axis which seemed to work.

Conclusion and Recommendation

The analysis shows that junction 1 has the highest chance of traffic congestion, it is advised that the deployed traffic prediction app should be utilized to know the state of the road, especially during traffic peak periods, and alternative routes (junction 4) should be plied by motorists

Team Members

1. Chisom Promise Nnamani
2. Pearse Jim
3. Victoria Udoh
4. Babatunde Raji
5. Oguamanam Chinyere
6. Zainab Mohammed
7. Bernard Boateng
8. Lakshay Arora
9. Bissalla Daniel
10. Subair Hussein
11. Samuel Nnamani
12. Gozie Ibekwe

13. Pragati Thakur

14. Fidel Imasuen

15. Odion Sonny-Egbeahie

16. Kevin Outta

17. Djardo Isaac

18. Emekobong Udoh

19. Omotayo Waheed

20. Joshua Obikunle

21. Bertram Okonkwo