ROAD TRAFFIC PREDICTION AND FORECASTING

DHRUVA G
PES1UG20CS131
DEPARTMENT OF COMPUTER SCIENCE
PES UNIVERSITY,BANGALORE
dhruva.pesu@gmail.com

DISHA SUNIL NIKAM
PES1UG20CS133
DEPARTMENT OF COMPUTER SCIENCE
PES UNIVERSITY,BANGALORE
sup.disha@gmail.com

SANIKA M RANGAYYAN
PES1UG20CS901
DEPARTMENT OF COMPUTER SCIENCE
PES UNIVERSITY,BANGALORE
sanrangayyan@gmail.com

ABSTRACT-This project seeks to leverage historically collected data from over 140 urban zones to forecast and predict traffic conditions in the future. It is an attempt to establish correlation between features such as weather and temperature with road traffic conditions. We also sought to forecast how the traffic flow will fluctuate in the future using time-series forecasting to identify traffic flow trends in future time periods.

KEYWORDS- forecasting,prediction,time series,traffic,junctions.

I . INTRODUCTION

Transportation is an important aspect of a sustainable city and society. Urban road transportation networks, as the carrier of human activities in the city, have been studied in terms of the structural characteristics and dynamics for decades. Most of the academic studies, however, are in a fragmented state. In the fields of geography and urban planning, physics and related domains, academic scholars have paid more attention to the structure of urban street networks, and the transport dynamics is the traditional research content of transportation scholars. Determining the network characteristics of structure and function is the key challenge of current research, and the ultimate goal of the network research is to better understand the behaviours of the transport systems

However, the actual data acquisition of the network level is always difficult, with most of the results obtained via traffic simulation. Currently, big data provides opportunities to gain insight into the relationship between transport dynamics and network intrinsic properties. The source and quality of big data are the main constraints for most researchers. There are many studies based on trajectory data or smart cards from different location devices (GPS, phone, etc.), however, the traffic detector data is managed by the government and traffic police. Related work is seldom reported. Even worse, data sources from different modes or samples are likely to show diverse vehicular patterns & is unreliable to accurately predict traffic.

Additionally, forecasting of traffic over different dates of the week using time-series forecasting is difficult, yet can lead to several crucial insights – as different days of the week have varying levels of traffic flow and congestion in various zones, making it a pivotal focus of our project, setting our project apart from other research conducted in this domain.

II. RELATED WORK

In the research paper titled “BAYESIAN ANALYSIS OF TRAFFIC FLOW ON INTERSTATE I-55: THE LWR MODEL” by : Nicholas Polson et al., the data is measured by loop-detector sensors installed on inter- state highways. Loop-detector is a simple presence sensor that measures when a vehicle is present and generates an on/off. Traffic flow parameters: The primary variable of interest is traffic density, which is a macroscopic characteristic of traffic flow and the control variable of interest in transportation. We were able to gain deeper insight into the modelling of traffic flow as a function of other physical attributes. The traffic flow is considered as a function of location x and time t. The flow-density relation, which is called the fundamental diagram, allowed them to calculate flow via density.

A QUEUEING MODEL FOR ROAD TRAFFIC FLOW proposed that on roads which are uninterrupted by traffic signals, intersections, etc., vehicles should be considered as travelling in random queues. Criteria for determining the queues in actual traffic are found. A crude model is then used to study the formation of these queues in an attempt to derive the Borel-Tanner distribution of queue lengths. The random queues model is then used to study waiting times for pedestrians (or vehicles) wishing to cross one lane of traffic.

In urban road networks, intersection usually constitute major bottlenecks, due to conflicting interactions between traffic streams in different directions. Intersections are the most critical points from capacity, congestion and safety viewpoints for the operation of an urban road network. The paper titled PREDICTIVE MODELLING OF TRAFFIC FLOW IN AKURE, NIGERIA: UNSIGNALIZED INTERSECTIONS IN FOCUS and the models developed models in this research provided insight into the combined effect of speed, density, headway and delay as well as the roadway geometric characteristics on traffic flow

III. METHODOLOGY

ABOUT THE DATASET

The two datasets used are the open source and collected over a period of time. The project uses two datasets-one from the UCI dataset and the other from open source projects. The first dataset consists of encoded features that denote day of the week,zone,weather and a numeric column for temperature. The dependent variable is the traffic.The traffic column values range from 1 to 5. 1 denotes number of vehicles passing is the specific zone for a given day of the week with temperature and weather lying in the range 0 to 5 vehicles. Similarly 2 denotes 5 to 20 vehicles, 3 for 20-45 vehicles and so on. The second dataset consists of the junction, date and time. The dependent column was the count of vehicles passing through the given junction at a particular date and time. The second dataset required implementing time series analysis.

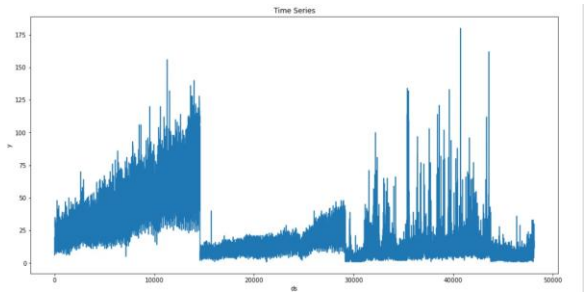
PREPROCESSING

The dataset was built after collecting the data over a period of time and did not have any null or NaN values. The datasets used for the project had most of its features encoded, hence required data cleaning. The techniques used were Label encoding and One hot encoding. Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form.

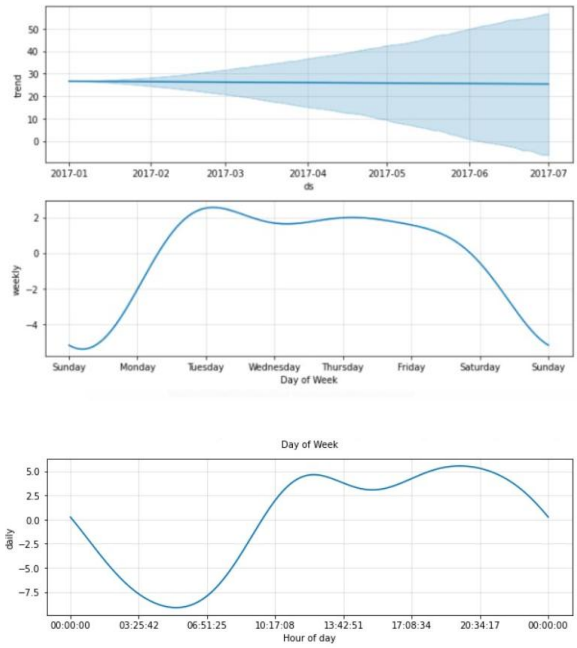
Machine learning algorithms can then decide in a better way how those labels must be operated. Most Machine Learning algorithms cannot work with categorical data and needs to be converted into numerical data.

EXPLORATORY DATA ANALYSIS

The below graphs visualizes the count of vehicles crossing a junction for a given duration of time in a time series way.



The below graphs visualizes the weekly,daily trend in the vehicles movement across a junction.The first graph shows that trend followed in the time series data on a daily basis.



Hence it is conclude that there is least traffic on Sundays with respect to a week. There is less traffic from 3am to 7am and remains relatively high in all other times of the day.

MODELS USED

The accuracy measure to compare the performances of the models is:

```
df_error=(y_predicted - y_test)/y_test
df_error=round(df_new.mean()*100,2)
accuracy=100-df_error
```

For the first dataset,regression or classification algorithms are used due to the fact that the ouput of the model must be a number ranging from 1 to 5.Hence below models are used.

1) SVM

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems.The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.SVM model has a good ability for traffic flow prediction. The model is interacted using a streamlit front-end.

2) RANDOM FOREST

The random forest algorithm is an extension of the bagging method as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees. Random forest algorithms have three main hyperparameters, which need to be set before training. These include node size, the number of trees, and the number of features sampled. From there, the random forest classifier can be used to solve for regression or classification problems.The random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn from a training set with replacement, called the bootstrap sample. The model is interacted using a streamlit front-end.

3) DECISION TREE

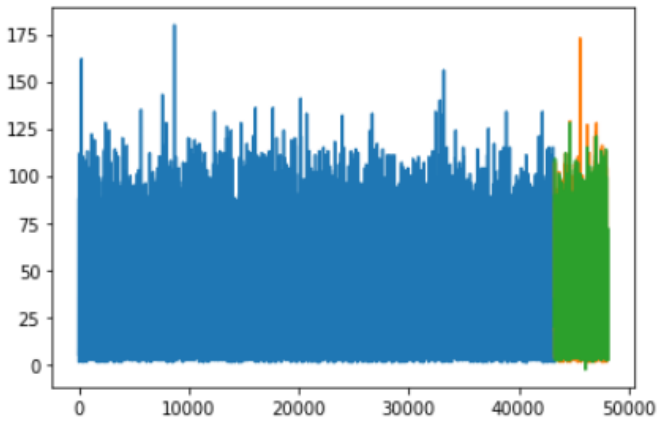
The actual flavour of machine learning is demonstracted in this algorithm as the whole algorithm is coded from scratch in this project.

A tree can be “learned” by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. Gini Index is a score that evaluates how accurate a split is among the classified groups. Gini index evaluates a score in the range between 0 and 1, where 0 is when all observations belong to one class, and 1 is a random distribution of the elements within classes.

For second dataset , time series forecasting algorithms are used to predict future count of vehicles for a given junction.The models used are:

1) BAGGING

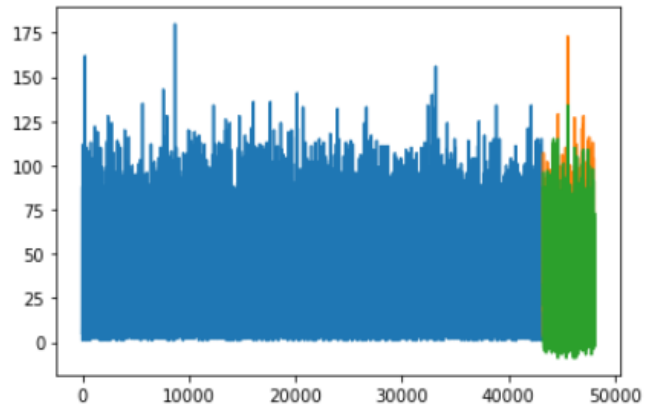
Bagging, also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce variance within a noisy dataset. In bagging, a random sample of data in a training set is selected with replacement—meaning that the individual data points can be chosen more than once. After several data samples are generated, these weak models are then trained independently, and depending on the type of task—regression or classification, for example—the average or majority of those predictions yield a more accurate estimate. The project uses gradient regression as the backend model to train the data.



The blue part of the graph is the training set, green is the validation and orange is the predicted part. The graph shows that model has a good overlap between precicted and validation set.

2) ADABOOST

AdaBoost also called Adaptive Boosting is a technique in Machine Learning used as an Ensemble Method. The most common algorithm used with AdaBoost is decision trees with one level that means with Decision trees with only 1 split. These trees are also called Decision Stumps.What this algorithm does is that it builds a model and gives equal weights to all the data points. It then assigns higher weights to points that are wrongly classified. Now all the points which have higher weights are given more importance in the next model. It will keep training models until and unless a lowe error is received.The project uses gradient regression as the backend model for training.



The blue part of the graph is the training set, green is the validation and orange is the predicted part. The graph shows that model has a good overlap between precicted and validation set.

3) FB-PROPHET

Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.

IV. CONCLUSION

The project successfully demonstrated traffic prediction for the first dataset using random forest and SVM with the Streamlit front-end which runs the model and predicts traffic value for the given values of independent variable. The Decision tree algorithm was coded from scratch using the first dataset. The time series forecasting was done on the second dataset using Bagging and Ada boost which computed Gradient regression in backend and FBProphet module was imported to study the time series in depth. The project implemented the above algorithms with good accuracy value.

V. REFERENCES

1) BAYESIAN ANALYSIS OF TRAFFIC FLOW ON INTERSTATE I-55: THE LWR MODEL Author(s): Nicholas Polson and Vadim Sokolov

2) A QUEUEING MODEL FOR ROAD TRAFFIC FLOW Author(s): Alan J. Miller

3) PREDICTIVE MODELLING OF TRAFFIC FLOW IN AKURE, NIGERIA: UNSIGNALIZED INTERSECTIONS IN FOCUS Author(s): Adebayo O. Owolabi, Olugbenga J. Oyedepo and Enobong E. Okoko