1. **Abstract**

New York City has intense traffic problems. We intend to use statistical methods to identify some New York City traffic pattern issues and suggest practical alternatives to solve these problems. The peak taxi pick-ups and drop-offs exists in particular timings of 4pm to 7pm. Our objective here is to write an algorithm so that I can find the time taken for a ride in New York City Limits. I have used the Streamlit platform for model development after several machine learning train test algorithms have been tested. The one with the highest accuracy has been used in the final Streamlit model. Random Forest Classifier, Ada-boost, and X.G. boost showed better results than other models. Random Forest Model showed the highest accuracy. I have used Random Forest Classifier for my final ensemble deployment to estimate travel time between origination and destination points. The Pick-up and Drop-off points have been plotted on the base map of the USA so that the concentration points can be established and we can estimate in which areas there is higher traffic flow. At these points, vehicles will have to stop for pick-up and drop-offs, which slower the traffic flow in the city. The areas of transit, railways, industries, and the Central Business Districts such as Manhattan have shown higher concentration points on the geographical Map.

1. **Introduction**

New York City has traffic issues. These issues are pestering the city, making the city earn less because of traffic snarls in the city. Traffic delays cost more than $115 billion to Americans in wasted fuel each year (Texas Transport Institute, 2015). People and officials have many problems thanks to New York's overwhelming traffic. The latter usually involves building extra roads and intersections to alleviate congestion. However as more roads and parking lots are made available to motorists, traffic congestion worsens. In the 1970s and 1980s, traffic congestion increased due to infrastructure spending in Central Europe.

This project aims to identify concentration points in the city based on *the NYC Taxi and Limousine Commission* data. The data has pick-up and drop-off points at different locations in the city of New York (after concatenating the two datasheets). We can hence identify the time of the day with peak travel time and estimate the travel time from origination to destination points.

Since New York's traffic woes are well known across the globe, we have chosen New York City traffic patterns because a study on this city can be of much influence on other cities with similar populations and metropolitan characteristics. A better planning model for an already advanced city like New York can help address issues in, say, New Delhi or Beijing.

1. **Literature Review**

**Traffic-Related Case Studies of Different Cities**

The related case studies for other cities cannot be contrasted entirely. The contingent environmental conditions may not be similar for different countries, but studying different countries and their traffic conditions expands our horizons and prepares us to identify common characteristics

1. By separating the route into autonomous fragments and predicting the origin-destination-based traffic flow behaviour of the fragments, Jain et al. (2017) were able to anticipate congestion on the urban highway and sub-arterial roads in a city with a diverse traffic set-up. The predicted trip time for a journey is modelled based on partial traffic attributes (flow and speed) at the origin and destination of road fragments, and roadway and segment traffic factors, including divergence routes, are also explored to account for travel time (Jain et al., 2017). The authors limited observation locations along the path to a few selected nodes rather than the whole distance. Then, using multiple linear regression, they developed a model to forecast travel time on a given fragment.
2. The issue of traffic congestion in the Rio Grande Valley typically extends beyond the difficulties it causes (Clarkson, 2014). It might have a substantial effect on the economy of the Rio Grande Valley. Psychological and Personal relationship issues start to appear in individuals who tend to spend higher time in city travel time.
3. Another case study is in the city of Cairo; A study by the world bank on Cairo's problem found that it led to the loss of about 50 billion Egyptian pounds (4% of the GDP). (Clarkson,2014). The economic impact of higher travel time has been discussed here.
4. Another Case Study in Jakarta, Indonesia Traffic Problems cause a loss of 0.6% of GDP annually for Indonesia (Clarkson,2014). Economic impact of higher travel time in the city of Jakarta has been discussed here.
5. An increase in traffic congestion is the fixed cost of owning and buying a car, which includes depreciation, car tax, motor insurance, and the price (Economics Online, 2017). Car sales and insurance companies have made the process simpler and hence causing single driver passenger car units to increase causing traffic congestion.

Data analysis of Transportation data – Case Studies

1. The relationship between taxi operations and weather conditions, including precipitation, snow depth, and snowfall, is well known. An extensive data analysis of taxi operations in New York City was conducted in a study titled 'Big Data Analytics of Taxi Operations in New York City' (scirp.org). Here, geospatial analysis has been performed using Geopandas.
2. Whong examined 170 million taxi journeys in NYC in 2013 and gathered data on the tip, total fare, number of passengers, trip start point, and trip finish point for each [Whong]. He then visualized these statistics to show how each component of taxi operation has altered over time.

[Big Data Analytics of Taxi Operations in New York City (scirp.org)](https://www.scirp.org/journal/paperinformation.aspx?paperid=94087)

1. **Data Description**

The data set was obtained from New York City Taxi and Limousine Commission. The Yellow Pages data is our primary data based on which the trip duration has been calculated. It is an open-source database type, i.e. available for the public to use and analyze to improve its services.

The link to the source data is below

Source of Data: [TLC Trip Record Data - TLC (nyc.gov)](https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page).

Source of Data: [NYC Taxi Zones | NYC Open Data (cityofnewyork.us)](https://data.cityofnewyork.us/Transportation/NYC-Taxi-Zones/d3c5-ddgc)

Yellow Taxi Trip Data Sheet

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| VendorID | A code indicating the TPEP provider that provided the record. 1= Creative Mobile Technologies, LLC; 2= VeriFone Inc. |
| tpep\_pickup\_datetime | The date and time when the meter was engaged. |
| Passenger\_count | The number of passengers in the vehicle. |
| Trip\_distance | The elapsed trip distance in miles reported by the taximeter. |
| PULocationID | TLC Taxi Zone, in which the taximeter was engaged. |
| DULocationID | TLC Taxi Zone, in which taximeter was disengaged. |
| RateCodeID | The final rate code in effect at the end of the Trip.  1= Standard rate  2=JFK  3=Newark  4=Nassau or Westchester  5=Negotiated fare  6=Group ride |
| Store\_and\_fwd\_flag | This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka "store and forward," because the vehicle did not have a connection to the server.  Y= store and forward trip N= not a store and forward Trip |
| Payment Type | A numeric code signifying how the passenger paid for the Trip.  1= Credit card  2= Cash  3= No charge  4= Dispute  5= Unknown  6= Voided trip |
| Fare\_amount | The time-and-distance fare calculated by the meter. |
| Extra | $0.50 MTA tax that is automatically triggered based on the metered rate in use. |
| Improvement\_surcharge | $0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015. |
| Tip\_amount | Tip amount – This field is automatically populated for credit card tips. Cash tips are not included. |
| Tolls\_amount | Total amount of all tolls paid in Trip |
| Total\_amount | The total amount charged to passengers. Does not include cash tips. |
| Congestion\_Surcharge | Total amount collected in Trip for NYS congestion surcharge. |
| Airport\_fee | $1.25 for pick up only at LaGuardia and John F. Kennedy Airports |

Taxi Zones Data Sheet

|  |  |
| --- | --- |
| Attribute | Description |
| Object ID | A Unique Identification number |
| Shape\_Leng | Perimeter of the census tract, in meters (computed after we erase the coastal water from census tracts) |
| The\_Geom | New York Geometry data |
| Shape Area | The area of the census tract, in square meters (computed after we erase the coastal water from census tracts) |
| Zone | Different boroughs in New York City have been further subdivided into different zones |
| Location ID | Unique Identification for each zone |
| Borough | NYC has five boroughs  1.the Bronx  2.Brooklyn  3.Manhattan  4.Queens  5.Staten Island |

Map

Description automatically generated

5.3 Example- Borough of Manhattan subdivided into different Zones

The data sheets have been obtained from the New York Taxi and Limousine Website. (Link mentioned above). The data sheets from January to June 2022 have been used for the analysis.

1. **Methodology**

The project can be divided into four parts.

1. Data collection.
2. Data pre-processing and Exploratory data analysis.
3. Machine Learning Testing and Training
4. Streamlit Final Deployment.

**Data Collection:** This dataset is collected and provided by the NYC Taxi and Limousine Commission (TLC), an open-source data type, i.e., available for the public to use and analyse to improve its services.

In our case, we have two data sheets concatenated based on the Location Id on one data sheet and the pick-up and drop-off locations on the other. Since these attributes are the common attributes in both the data sheets, such concatenation has been done.

The concatenation of the January to June data sheets has been done. The data on the TLC website is in parquet format; we converted it to csv format to analyse the dataset.

I have picked 700000 rows (around 4800000 after concatenation) randomly so that the I can reduce the computation time. After adding the trip duration column, I have randomly chosen three hundred thousand selections from the seven hundred thousand data for further analysis to reduce the computation time. I also performed the same analysis randomly, selecting only thirty thousand selections. The results are different in some ways concerning the three hundred thousand selections in terms of the Passenger count.

**Data Pre-processing and EDA:**  In data pre-processing, we understand the attributes of the dataset and come up with a consensus as to what variables might affect the choice of the Response variable. In Exploratory Data Analysis, by the end of it, we make sure our dataset is suitable for modelling by changing object columns to integer columns, dropping the columns which are not crucial for the prediction, and so on.

I have dropped the following columns which I believe do not have much influence on the trip duration-*VendorID','RatecodeID','store\_and\_fwd\_flag,'PULocationID', 'DOLocationID, payment\_type', 'extra', 'mta\_tax', 'tip\_amount', 'tolls\_amount','total\_amount, 'improvement\_surcharge','congestion\_surcharge.'*

**Trip Duration**

The pick-up and drop-off times of the cabs are recorded in the datasheet. I have used these times to calculate the duration of each trip. I then used the trip duration as my response variable to determine the time taken for a trip for different locations based on other attributes. The other pertinent attributes have been used as predictor variables for the determination of trip duration between two sets of geographical points.

Graphs have been drawn between the trip duration and different predictor variables using the Matplotlib and Seaborn packages. Boxplots have been plotted using the Seaborn package. Graphviz package has been used for Decision Tree visualization.

**Model Train and Test:** Once the EDA and Feature Engineering are done, we will divide our dataset into train and test then feed the training dataset to the model and predict with Test dataset.

I have used the following models

1. Regularisation models- Ridge and Lasso
2. KNN Regression
3. Support Vector Regression
4. Polynomial Regression
5. Hyper Parameter tuning for KNeighbours Regressor
6. Decision Tree
7. Random Forest
8. Ada boost Regressor
9. Bayesian Ridge

All these machine learning models for Train and Test prediction and regression models have been used to develop the best fit model for analysis. In addition, the random Forest Model has been used for the Streamlit deployment model as it has the highest accuracy among the existing models.

**Random Forest Machine Learning Model**

Random Forest is a classifier that uses many decision trees on different subsets of the input dataset and averages the results to increase the dataset's predicted accuracy.It is a combination of different predictions done from various Decision Trees. My Random Forest model shows almost similar results to Bayesian Ridge model and Adaboost Regressor. I have chosen the Random Forest in my final model for the following reasons:

* Random Forest can complete problems including classification and regression.
* It is able to handle big datasets with lots of dimensions.
* It improves the model's accuracy and avoids the overfitting problem.

RANDOM FOREST CLASSFIER

DECISION TREE

DATASET

MAJORITY VOTE TAKEN

FINAL PROJECTION

PREDICTION

PREDICTION

PREDICTION

Fig 5.1 Random Forest Classifier Flow Chart

**Deployment:** We download the pickle file and deploy the file to predict in real-time scenarios.

**About 'Pickle'**: It is a python module used to serialize a python object into binary format and deserialize it back to the python object. We use pickle to save the final data and train it with multiple models, or we can keep the model and test it on various data without training the model again.

We give the inputs like pick-up latitude, pick-up longitude, drop-off latitude, drop-off longitude, date and time of the journey, and the number of passengers. Once all the required inputs are given, we can predict the time required for that ride.

Streamlit model has been used for the final ensemble deployment. (GeeksforGeeks)

When a function is marked with Streamlit's cache annotation, it instructs Streamlit to do the following three checks every time the function is called:

* The input parameters that were used to invoke the function,
* the actual bytecode that makes up its body,
* the files, variables, and other pieces of code it depends on.

1. **Results and Discussion**

**Trip Duration Estimation with Streamlit Model**

Trip pick-ups and Drop-offs have been plotted at different times of the day

Chart, bar chart, histogram

Description automatically generated

Fig 6.1 Pick-ups at different times of the day

Chart, bar chart, histogram

Description automatically generated

Fig 6.2 Drop-offs at different times of the day

Trip Counts on different days of the week

Chart, bar chart

Description automatically generated

Fig. 6.3 Pick-ups at different days of the week

Chart, bar chart

Description automatically generated

Fig.6.4 Drop-offs at different days of the week

Passenger Count influence on the Trip Duration

Analysis was performed on 2 sets of data. One is on thirty thousand randomly selected variables and the following graph has been obtained between *Trip Duration* and *Passenger Count.* Chart, bar chart

Description automatically generated

Fig 6.5 Trip Duration v/s Passenger Count

When the analysis was performed on three hundred thousand data, the graph between Trip Duration and Passenger Count is as below.

Chart

Description automatically generated

Fig 6.6 Trip Duration v/s Passenger Count

Another iteration of three hundred thousand data gives me another graph,

Chart, bar chart

Description automatically generated

Fig 6.7 Trip Duration v/s Passenger Count (three hundred thousand random data)

From the graphs plotted above, there seems to be no definite pattern between Trip Duration and Passenger Count. The exact pattern may be developed if we consider the entire dataset for the analysis.

Several Regression and machine learning models have been used to find the right fit so that an apt prediction model based on our dataset can be developed. The above-said models (mentioned in methodology section) have been used here for fitting a regression so that the best fit model can be used for the final ensemble.

The results comparison for different models has been presented below in a tabular column.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | R2 | MAE (Minutes) | RMSE (Minutes) |
| Linear Regression | 0.0085 | 5.3075 | 6.7482 |
| Lasso | 5.3042 | 5.4032 | 6.7792 |
| Ridge | 0.0089 | 5.3064 | 6.7468 |
| Polynomial Regression | -5.3714 | 22576436.8156 | 1570709275.178456 |
| KNN Regression | -5.3714 | 1562724787.4333 | 70455179209.9072 |
| Decision Tree | 0.1234 | 5.3304 | 6.8866 |
| Bayesian Ridge | 0.743 | 2.8596 | 4.1288 |
| **Random Forest** | **0.743** | **2.8596** | **4.1288** |
| AdaBoost Regressor | 0.743 | 2.8596 | 4.1288 |

Below is the demonstration of a real-time trip duration prediction in deployment.

A picture containing text, road, screenshot, several

Description automatically generated

A picture containing text, road, screenshot, monitor

Description automatically generated

While other models predict with high errors, the Random Forest outperforms other different models and forecasts with a Mean Absolute Error (MAE) of 2.859 minutes and Root Mean Squared Error(RMSE) of 4.1288 minutes, the results is better compared most other models.

1. **Conclusion and Recommendations**
2. Estimating population increase has to be established more thoroughly so we can model our planning infrastructure accordingly. There is a gap between a city's demand and investment in city expansion and modelling. Planning deficiencies must be redressed, so Central Business Districts do not get overcrowded.
3. Car share and Rideshare practices need to be more widely followed by people in metropolitan cities to relieve the pressure of single-driver vehicles from the city.
4. In my home country city of New Delhi, a population of around 170 million people, traffic jams are a daily occurrence. Therefore, the odd vehicle number plate vehicles are allowed to ply on one day and even-numbered cars on the next day. This model is to reduce traffic congestion and encourage the usage of public transport.
5. Timings for Congestion in the city are during the day.The office timings need not be the same for all to accommodate traffic problems. For example, in summer, some offices in New York City can begin at 6 am and end at 2 pm.
6. The Manhattan area in New York has the highest daily Pick-ups and Drop-offs, as confirmed by the concentration maps. Times Square is an international destination for Tourists across the globe. In addition, New York City has a thriving stock market in downtown Manhattan. Due to these reasons, Traffic congestion is high mainly in this.
7. Traffic Engineers can use the Folium maps generated in the analysis to understand the points of heavy cab pick-ups and drop-offs so that a dedicated lane may be created in the city where such activities are disproportionately higher.
8. Traffic Engineers can utilize the Trip Duration between Origination and Destination points to understand which route is causing a delayed travel time. Then, the identified problems on that particular route can be resolved.

**References**

1. Bharadwaj, S., Ballare, S., & Chandel, M. K. (2017). Impact of Congestion on greenhouse gasemissions for road transport in Mumbai metropolitan region. Transportation Research Procedia, 25, 3538–3551. <https://doi.org/10.1016/j.trpro.2017.05.282>
2. Bouyahia, Z., Haddad, H., Derrode, S., & Pieczynski, W. (2022). Traffic state prediction using conditionally Gaussian observed Markov fuzzy switching model. Journal of Intelligent Transportation Systems, 1-20.
3. Chang, Y. S., Lee, Y. J., & Choi, S. S. B. (2017). Is there more traffic congestion in larger cities?-Scaling analysis of the 101 largest U.S. urban centers? Transport Policy, pp. 59, 54–63.
4. Clarkson, N. (2014). What is The Economic Impact of Traffic? Retrieved on 17 Apr. 2018 from

https://www.virgin.com/disruptors/what-economic-impact-traffic

1. Das, S., & Nayyar, A. (2019). Innovative ideas to manage urban traffic congestion in cognitive cities. In Driving the development, management, and sustainability of cognitive cities (pp. 139-162). IGI Global.
2. Di, X., Xiao, Y., Zhu, C., Deng, Y., Zhao, Q., & Rao, W. (2019, June). Traffic congestion prediction by spatiotemporal propagation patterns. In 2019 20th IEEE International Conference on Mobile Management (MDM) (pp Data. 298–303). IEEE
3. Emami, A., Sarvi, M., & Asadi Bagloee, S. (2019). Using Kalman filter algorithm for short-term traffic flow prediction in a connected vehicle environment. Journal of Modern Transportation, 27(3), 222–232.
4. Ewing, R., Tian, G., & Lyons, T. (2018). Does compact development increase or reduce traffic congestion? Cities, pp. 72, 94–101.
5. Faghih-Imani, A., Anowar, S., Miller, E. J., & Eluru, N. (2017). Hail a cab or ride a bike? A travel time comparison of taxi and bicycle-sharing systems in New York City. TransportationResearch Part A: Policy and Practice, 101, 11-21.
6. Fiedler, D., Čáp, M., & Čertický, M. (2017, October). Impact of mobility-on-demand on traffic congestion: Simulation-based study. In 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC) (pp. 1-6). IEEE.FHWA. (2022). Sources for FHWA Reports.
7. FHWA. Highways.dot.gov. <https://highways.dot.gov/research/resources/research-library/sources-fhwa-reports>
8. Geeks for Geeks blog

[Deploy a Machine Learning Model using Streamlit Library - GeeksforGeeks](https://www.geeksforgeeks.org/deploy-a-machine-learning-model-using-streamlit-library/)

1. García-Ramírez, Y. (2020). Developing a Traffic Congestion Model based on Google Traffic Data: A Case Study in Ecuador. In VEHITS (pp. 137-144).Jain, S., Jain, S. S., & Jain, G. (2017). Traffic congestion modeling based on origin and destination. Procedia Engineering, 187, 442–450.

<https://doi.org/10.1016/j.proeng.2017.04.398>

1. Jayasooriya, S. A. C. S., & Bandara, Y. M. M. S. (2017, May). Measuring the Economic costs of traffic congestion. In 2017 Moratuwa Engineering Research Conference (Mercon) (pp. 141-146). IEEE.
2. Kohan, M., & Ale, J. M. (2020). Discovering traffic congestion through traffic flow patterns generated by moving object trajectories. Computers, Environment and Urban Systems, 80, 101426.
3. Koźlak, A., & Wach, D. (2018). Causes of traffic congestion in urban areas. Case of Poland. In SHS Web of Conferences (Vol. 57, p. 01019). EDP Sciences.Liu, Y.D.
4. Liu, C. and Zheng, Z.L. (2020) Traffic Congestion and Duration Prediction Model Based on Regression Analysis and Survival Analysis. Open Journal of Business and Management, 8, 943-959. <https://doi.org/10.4236/ojbm.2020.82059>.
5. Lomax, T., Schrank, D., Eisele, B., & Albert, L. (2021). Urban Mobility Report — Urban Mobility Information. Mobility.tamu.edu. <https://mobility.tamu.edu/umr/>
6. Lu, J., Li, B., Li, H., & Al-Barakani, A. (2021). Expansion of city scale, traffic modes, traffic congestion, and air pollution. Cities, 108, 102974.Lyu, D., & Lin, Y. (2022). Impact Estimation of Traffic Accident Duration Based on SurvivalAnalysis by Using Field Urban Traffic Condition. In International Conference on Green Building, Civil Engineering and Smart City (pp. 1247-1254). Springer, Singapore.
7. Marshall, W. E., & Dumbaugh, E. (2018). Revisiting the relationship between traffic congestion and the economy: A longitudinal examination of U.S. metropolitan areas. Transportation, pp. 1–40. <https://doi.org/10.1007/s11116-018-9884-5>
8. Metz, D. (2018). Tackling urban traffic congestion: The experience of London, Stockholm, and Singapore. Case Studies on Transport Policy, 6(4), 494–498.
9. Nagy, A. M., & Simon, V. (2021). Traffic congestion propagation identification method in smart cities. Infocommunications Journal, 13(1), 45-57
10. Schneider, T. (2016) Taxi, Uber, and Lyft Usage in New York City.  
    <http://toddwschneider.com/posts/taxi-uber-lyft-usage-new-york-city/>
11. Schneider, T. (2015) Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance.
12. Whong, C. (2013) NYC Taxis: A Day in the Life. [http://chriswhong.github.io/nyctaxi/#](http://chriswhong.github.io/nyctaxi/)