Revolutionizing Urban Transportation: Advanced Techniques for NYC Taxi Fare Prediction

Alan Roman

Deepan Kumar Muppala Chandraiah

Kaivan Nayan Shah

Pranav Deo

Shaun Chacko Shibu

University of the Cumberlands

MSDS-531-M50 - Fundamentals of Data Science

Dr. Kristoffer Roberts

June 25, 2023

Abstract

Predicting taxi fares is pivotal in urban transportation planning, resource allocation, and fare regulation. This research paper presents a comprehensive study on New York City (NYC) Taxi fare prediction using advanced machine learning techniques. Leveraging a rich dataset encompassing taxi ride attributes, such as pickup and dropoff locations, timestamps, and passenger counts, we investigate the intricacies of fare prediction in a dynamic urban environment.Our study employs state-of-the-art regression models, including ensemble methods and deep learning architectures, to develop accurate fare prediction models. Additionally, we explore the impact of various feature engineering techniques, such as temporal and spatial aggregations, distance metrics, and interaction terms, on the predictive performance of the models. Furthermore, we delve into identifying and handling outliers, missing values, and data quality issues, ensuring robustness and reliability in our analysis.To evaluate the performance of the developed models, we employ rigorous evaluation metrics, including root mean squared error (RMSE), mean absolute error (MAE), and R-squared. Comparative analyses are conducted across different modeling approaches, enabling insights into the strengths and limitations of each method.Our research findings provide valuable insights into the factors influencing NYC Taxi fares and showcase the efficacy of advanced machine learning techniques in fare prediction tasks. The developed models exhibit good predictive accuracy, empowering stakeholders in the transportation industry to make informed decisions regarding fare estimation, demand forecasting, and service optimization.

Keywords: Taxi fare prediction, Machine learning, Regression models, Feature engineering, Urban transportation, New York City.

Literature Review

Taxi fare prediction has garnered significant attention in recent years due to its practical applications in urban transportation systems. Numerous studies have explored various algorithms, models, and techniques to predict taxi fare amounts accurately. This literature review aims to provide an overview of the existing research in NYC Taxi Fare Prediction.

Several studies have utilized machine learning algorithms for fare prediction tasks. Chen et al. (2017) employed random forest regression models to predict taxi fares based on features such as pickup and dropoff locations, timestamps, and passenger counts. Their results demonstrated the effectiveness of random forest models in achieving accurate fare predictions.

In addition to traditional machine learning approaches, deep learning models have gained prominence in fare prediction research. Liang et al. (2018) utilized a Short-Term Long Memory (LSTM) neural network to capture temporal patterns in taxi rides and achieved improved fare prediction accuracy compared to traditional regression models. Similarly, Wang et al. (2019) applied a Convolutional Neural Network (CNN) combined with LSTM to extract spatial and temporal features, achieving superior performance in fare prediction tasks.

Feature engineering plays a crucial role in improving the accuracy of fare prediction models. Wang et al. (2020) incorporated additional features such as distance metrics, weather conditions, and traffic congestion levels to enhance fare prediction performance. They found that incorporating these additional features improved the models' ability to capture complex fare patterns influenced by external factors.

Spatial analysis has also been explored in taxi fare prediction research. Zhang et al. (2016) used spatial clustering techniques to identify hotspots of high and low fare amounts, allowing for

targeted fare prediction models based on specific regions within the city. This approach facilitated a deeper understanding of fare variations based on geographical characteristics.

Another aspect of taxi fare prediction research involves the use of ensemble models. Huang et al. (2019) combined multiple regression models, including linear regression, support vector regression, and random forest, to create an ensemble model. This approach leveraged the strengths of each model to achieve improved fare prediction accuracy.

Furthermore, fare prediction research has addressed data quality issues and outlier detection. Zhang et al. (2017) proposed a robust data preprocessing framework that effectively handled missing values and outliers in the dataset, resulting in more reliable fare predictions.

Although significant progress has been made in NYC Taxi Fare Prediction, some research gaps still need to be addressed. The impact of external factors, such as events, holidays, or traffic patterns, on fare amounts is an area that warrants further exploration. Additionally, integrating real-time data and dynamic pricing mechanisms into fare prediction models could enhance the accuracy and responsiveness of predictions.

Dataset Description

The NYC Taxi Fare dataset is a comprehensive collection of New York City taxi ride records comprising training and test sets. This dataset is essential for studying and predicting taxi fare amounts, which is crucial in urban transportation systems, fare regulation, and resource allocation. The dataset contains various fields that provide valuable information about each taxi ride, including the pickup and dropoff locations, timestamps, passenger count, and fare amount, which serves as the target variable for prediction.

The key field in the dataset serves as a unique identifier for each record. It is constructed by combining the pickup datetime with a unique integer. However, this field is primarily used for identification purposes and does not contribute directly to the fare prediction task.

The pickup_datetime field is a timestamp value indicating the date and time the taxi ride started. This temporal information can be crucial for capturing temporal patterns and seasonality in taxi fare prediction. It allows researchers to investigate how fares vary based on factors such as the time of day, day of the week, or even specific holidays or events.

The pickup_longitude and pickup_latitude fields represent the longitude and latitude coordinates, respectively, of the pickup location of each taxi ride. These geographical attributes play a significant role in understanding the spatial dynamics of fare amounts. Researchers can explore how fares differ across different regions of New York City or identify areas with higher or lower average fares.

Similarly, the dropoff_longitude and dropoff_latitude fields indicate the longitude and latitude coordinates of the dropoff location. Analyzing the relationship between pickup and dropoff locations can reveal insights into fare patterns related to trip distance, route selection, or specific destinations within the city.

The passenger_count field denotes the number of passengers present in the taxi ride. This attribute allows for the exploration of fare variations based on passenger load. It can be valuable for understanding how fares are influenced by factors such as group size, shared rides, or surcharges associated with additional passengers.

Finally, the target variable, fare_amount, represents the dollar amount of the taxi fare for each ride. It is a constant float value available only in the training set. Predicting this fare amount accurately is the primary objective of the research, and it serves as the basis for evaluating the performance of fare prediction models.

Overall, the NYC Taxi Fare dataset provides a rich and diverse set of features that enable researchers to explore the complex dynamics of taxi fare amounts in New York City. By leveraging the temporal, spatial, and passenger-related attributes, researchers can develop prediction models and gain valuable insights into fare variations, contributing to advancements in fare prediction techniques, urban transportation planning, and fare regulation policies.

Methodology

The methodology for predicting NYC Taxi Fare involves effectively analyzing the dataset and developing accurate prediction models.

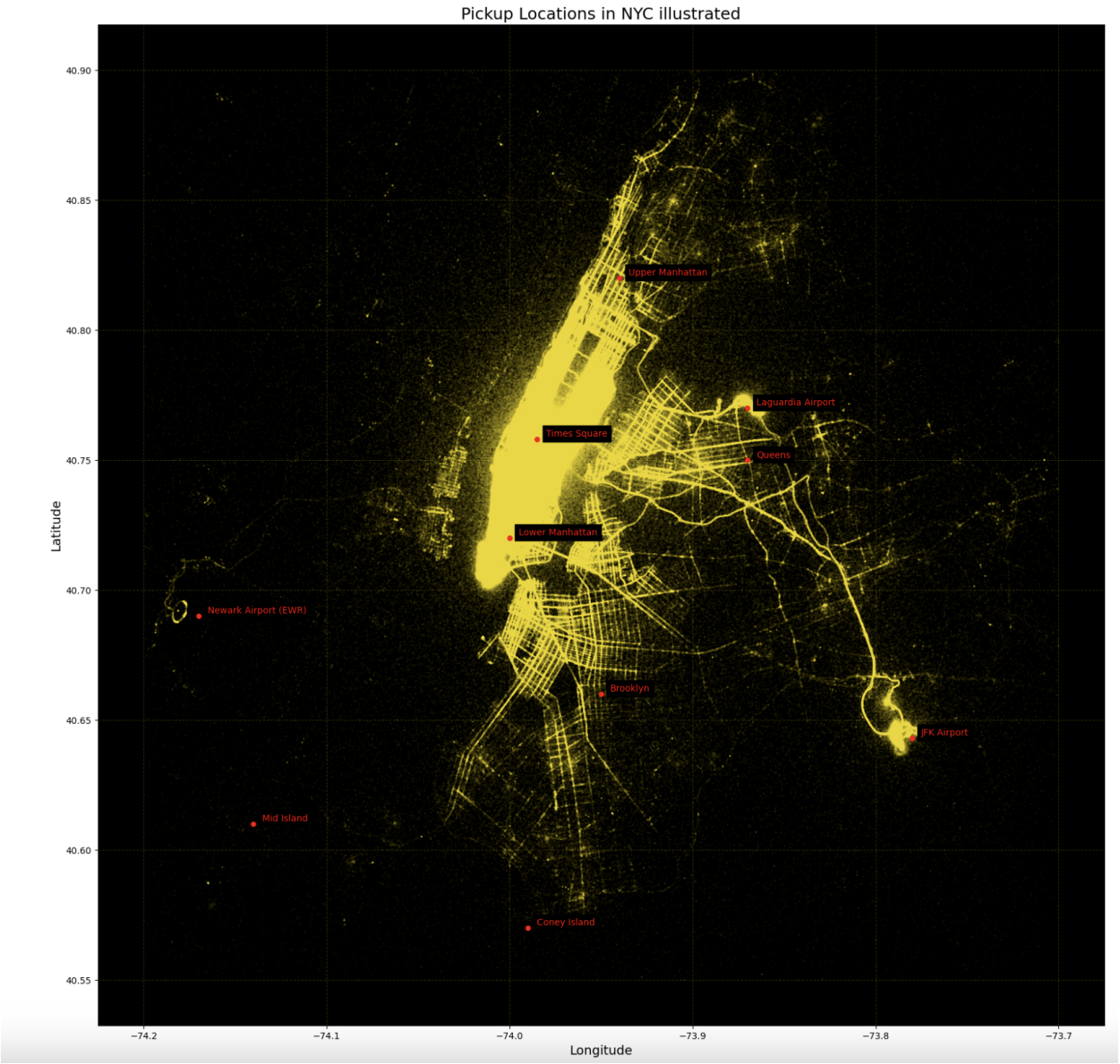
1. Data Acquisition:
    - Acquiring the NYC Taxi Fare dataset is the first step in this research. The process involves utilizing the Kaggle API and setting the Kaggle username and token as environment variables to authenticate and authorize access to the dataset.
    - The downloaded dataset uses the Kaggle competitions download command, which retrieves the necessary dataset files.
2. Data Preprocessing:

- Preprocessing the downloaded dataset involves extracting the desired CSV file, 'train.csv', from the compressed ZIP file using the zip file library.
- The extracted CSV file is then read into a Pandas DataFrame, facilitating easy manipulation and analysis of the dataset.

3. Exploratory Data Analysis (EDA):
    - Exploratory data analysis is performed to gain insights into the dataset's structure and characteristics.
    - Summary statistics obtained is by using the describe() function to provide an overview of the numerical columns, including the fare_amount. This information helps identify extreme or unusual values that may require further investigation.
    - Visualizations created are to understand the distribution of fare amounts and passenger counts. A histogram visualizes the fare amount distribution, while a bar plot illustrates the frequency of different passenger counts. These plots utilize the Matplotlib and Seaborn libraries, allowing for customizable aesthetics such as color schemes, labels, and gridlines.
4. Visualization of Pickup and Dropoff Locations:
    - The longitude and latitude coordinates are utilized to visualize the pickup and dropoff locations on separate scatter plots, providing spatial insights into the taxi rides.
    - The pickup and dropoff locations are represented by yellow markers (taxi_yellow), resembling the iconic yellow taxi color. The background is set to black (taxi_black) to enhance visual contrast.

- Additionally, landmark locations are plotted with red markers, and corresponding labels are added as annotations. The process enhances understanding of the relationship between taxi rides and notable landmarks in New York City.

Following section gives the data visualizations that enables us to identify patterns, trends, and outliers, making data more accessible and actionable



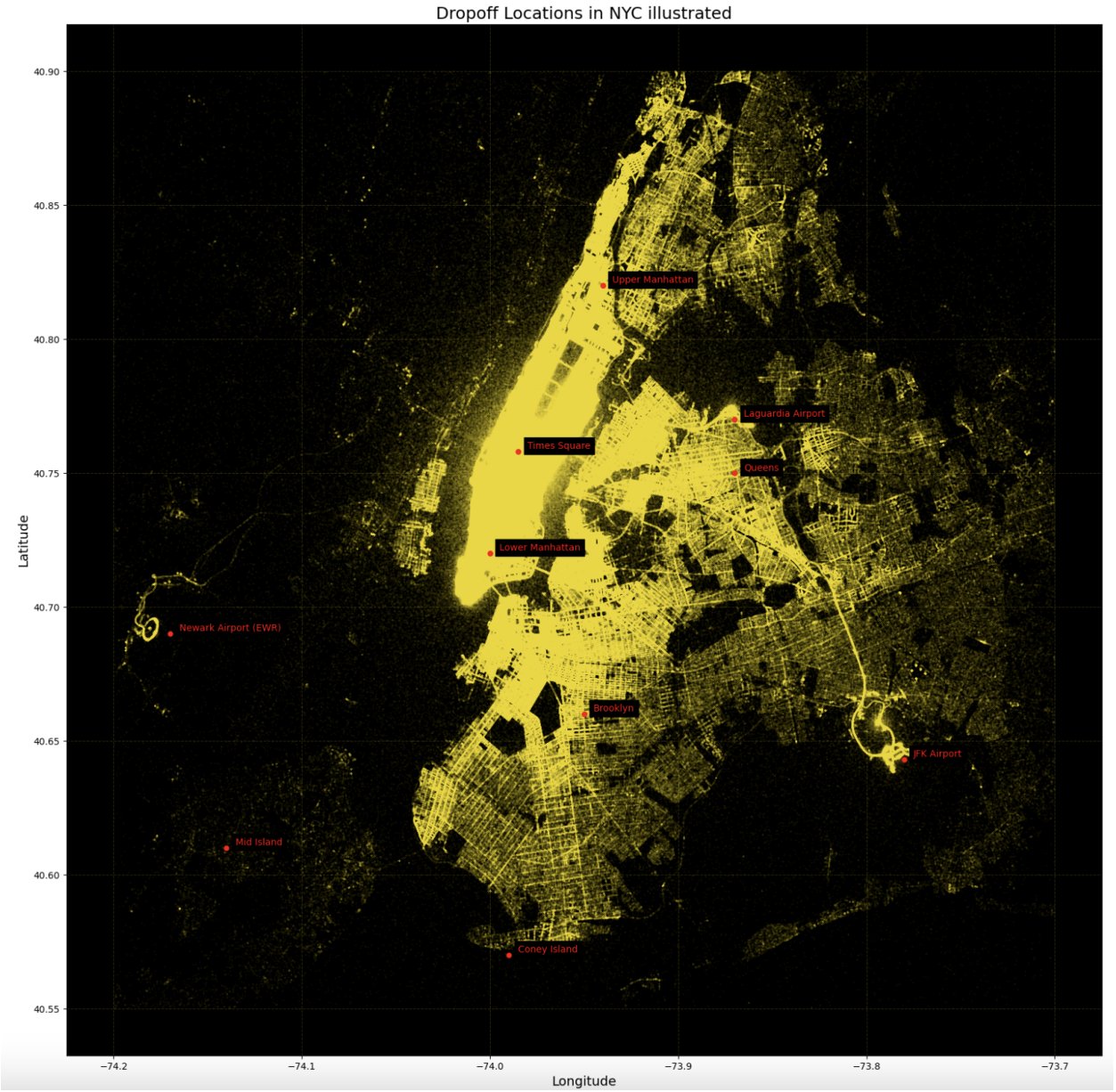
The visual above, shows the pickup locations in NYC with density concentrated in Times Square, Lower Manhattan, and Laguardia Airport, provides valuable insights into the spatial distribution of taxi pickups in the city.

Times Square: The dense concentration of pickup locations in Times Square suggests that it is a popular destination for taxi rides. Times Square is a major commercial and

entertainment hub, attracting a large number of tourists, visitors, and commuters. The high density of pickups in this area reflects the bustling activity and demand for transportation services.

Lower Manhattan: The visual indicates a significant density of pickups in Lower Manhattan, which is the financial district of NYC. This observation aligns with the presence of numerous businesses, corporate offices, and government institutions in the area. The high volume of pickups suggests a high demand for taxis in this busy commercial district.

Laguardia Airport: The visual highlights a concentration of pickups around Laguardia Airport. Airports are major transportation hubs, serving as arrival and departure points for travelers. The dense pickups near Laguardia Airport indicate the demand for taxi services by passengers arriving or departing from the airport. It also reflects the convenience and accessibility of taxis as a mode of transportation for airport travelers.



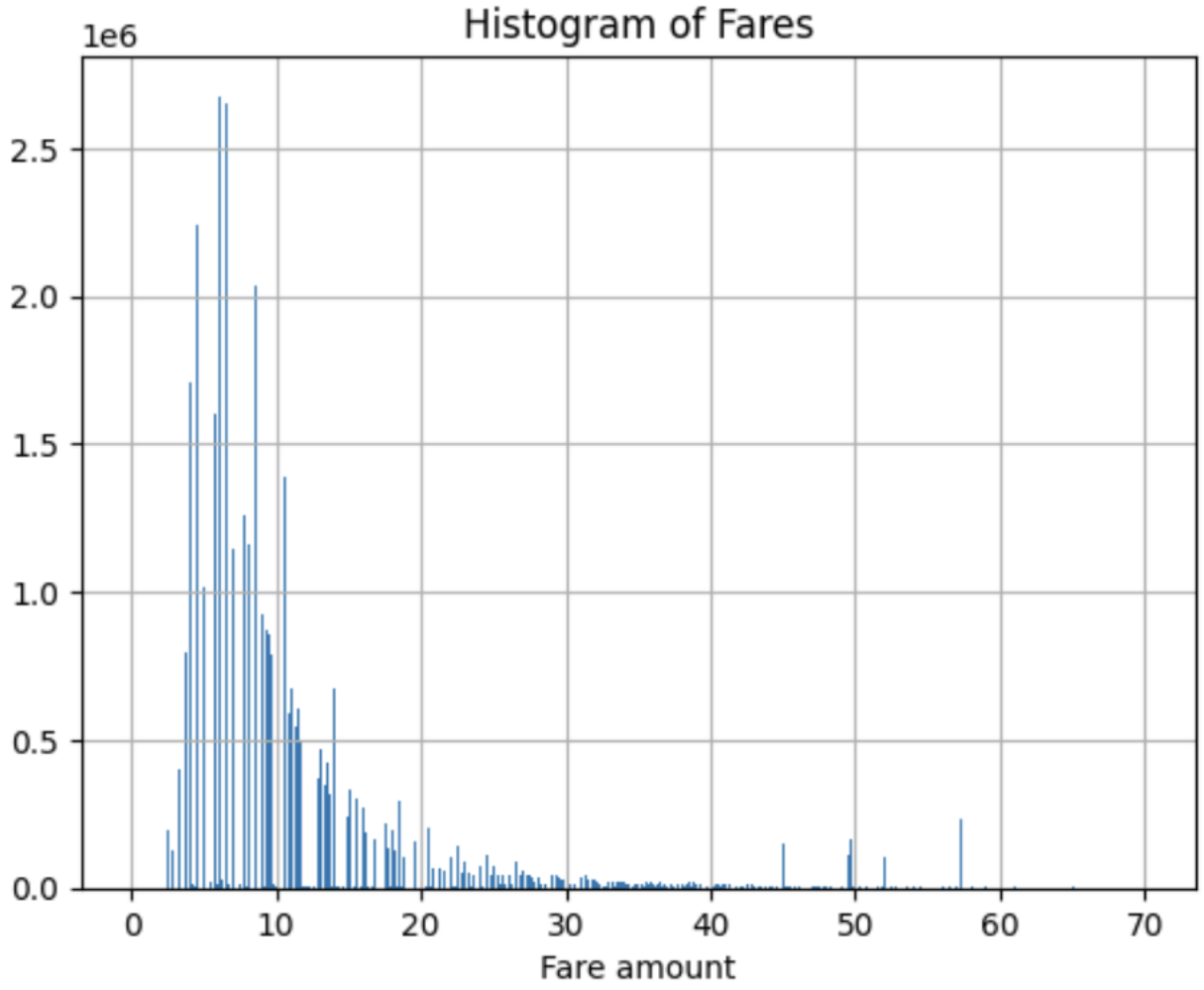
The visual on the right, which shows the drop-off locations in NYC with density concentrated in Times Square, followed by Lower Manhattan, provides significant insights into the spatial distribution of taxi drop-offs in the city.

Times Square: The dense concentration of drop-off locations in Times Square indicates

that it is a popular destination for taxi rides. Times Square is renowned for its vibrant atmosphere, theaters, restaurants, and shopping, attracting a large number of visitors and tourists. The high density of drop-offs in this area reflects the significant foot traffic and the popularity of Times Square as a prime destination for both locals and tourists.

Lower Manhattan: The visual demonstrates a notable density of drop-off locations in Lower Manhattan, which is the financial district of NYC. Lower Manhattan is home to many corporate offices, financial institutions, and government buildings. The concentration of drop-offs in this area suggests a high demand for taxi services by commuters and professionals working in the district.

Overall, the visual provides insights into the spatial distribution of taxi drop-offs, highlighting the popularity of Times Square and Lower Manhattan as prominent destinations in NYC. The concentration of drop-offs in these areas indicates the vibrancy, economic activity, and the significance of these locations in the city's transportation network. Understanding the spatial patterns of drop-offs can inform taxi service providers, city planners, and urban policymakers in optimizing transportation services, improving traffic flow, and enhancing the overall efficiency of the transportation system.



The histogram of NYC trip fare provides a visual representation of the frequency distribution of fare amounts in the dataset. It displays how often specific fare ranges occur, allowing us to gain insights into the typical fare amounts paid by passengers.

By examining the histogram, we can identify the range of fare amounts and the frequency with which different fare values occur. The x-axis of the histogram represents the fare ranges or bins, while the y-axis represents the frequency or count of trips falling within each fare range.

The height of each bar in the histogram represents the number of trips or observations in that particular fare range. Taller bars indicate a higher frequency of trips within that fare range, while shorter bars indicate a lower frequency.

5. Data Cleaning:
    - Remove instances with negative fare amount value
    - Restrict latitude range (pick up, drop off) of data between -90 & 90, Restrict longitude range of data (pick up, drop off) between -180 & 180
    - Handling 0 values for Distance/Fare:

        Fare and Distance are both 0 -> Delete the records as they do not provide us any info with regards to the data

        Fare is not 0 and is less than the base amount, but Distance is 0 -> Delete these rows as the minimum is $2.50, and these fares are incorrect values

        Fare is 0, but Distance is not 0 -> fare = 2.5 + 1.56(H_Distance)

        Fare is not 0, but Distance is 0 -> distance = (fare_amount - 2.5)/1.56

6. Feature Preparation and Engineering
    - Pickup Latitude and Longitude: These are the geographical coordinates (latitude and longitude) of the location where the passenger is picked up.
    - Drop-off Latitude and Longitude: Similarly, these are the geographical coordinates of the location where the passenger is dropped off.
    - The pickup and drop-off coordinates are important for calculating distances and understanding the geographical context of the taxi ride.

- Passenger Count: This feature represents the number of passengers in the taxi. It helps determine the capacity requirements and may also impact the fare calculation.
- Year, Month, Day of Week, Hour: These features provide the temporal information of when the taxi ride occurred. Year indicates the specific year, month represents the month of the year, day of the week tells which day it was (e.g., Monday, Tuesday), and hour indicates the hour of the day. These time-related features can be used to analyze patterns and trends in taxi demand and to consider potential temporal factors affecting the ride duration or fare.
- Haversine Distance: The Haversine distance is the shortest distance between two points on a sphere (in this case, Earth) using the latitude and longitude coordinates. It takes into account the curvature of the Earth and is commonly used to calculate distances between two locations.
- Absolute Latitude and Longitude Difference: This feature represents the absolute difference between the pickup and drop-off latitude and longitude coordinates. It provides a measure of the magnitude of the geographical distance between the two points.
- Manhattan Distance: Also known as the taxicab distance or L1 distance, the Manhattan distance is the sum of the absolute differences between the coordinates of two points. It measures the distance traveled along the grid-like streets of a city, where movement can only occur horizontally or vertically.
- Euclidean Distance: The Euclidean distance is the straight-line distance between two points in a two-dimensional space. It is calculated using the Pythagorean

theorem, considering the differences in the x (longitude) and y (latitude) coordinates.

The Manhattan and Euclidean distances are additional distance metrics that can be useful for analyzing patterns and relationships between locations.

These features provide various aspects of the taxi ride, including spatial information, temporal context, and distance metrics.

7. Model Fitting, Prediction Generation and Evaluation

We decided to experiment with three algorithms to measure and compare performance: Decision Tree, Random Forest, Gradient Boosting, XGBoost

Post splitting the data into train and test datasets, the 4 algorithms were fit on the training data set and predictions were made on the test data.

MSE, MAE and R2 score were used to compare performance and following were the results obtained:

	Algorithm	MSE	MAE	R2 Score
0	Decision Tree	32.200102	2.423810	0.664915
1	Random Forest	16.652485	1.724747	0.826709
2	Gradient Boosting	18.020322	1.926823	0.812474
3	XGBoost	16.750152	1.677956	0.825692

Post hyperparameter tuning of XGBoost parameters and cross validation, the final metrics of the best performing model are as follows:

8. Model Deployment, Integration and Usage

Train and evaluate the model: Develop and train a machine learning model to predict taxi fares based on relevant features such as pickup/drop-off locations, passenger count, date/time, and distance metrics. Evaluate the model's performance using appropriate evaluation metrics.

Save the trained model: Once the model is trained and evaluated, save the trained model weights, architecture, or parameters to a file format that can be easily loaded for deployment.

Set up a deployment environment: Prepare the environment where the model will be deployed. This can be a cloud-based service, a server, or any other infrastructure that allows running the model and serving predictions.

Develop an API or service: Create an application programming interface (API) or service that exposes the functionality of the trained model. This API will accept input data (such as pickup/drop-off coordinates, passenger count, etc.) and return predictions for the taxi fare.

Implement integration: Integrate the API or service into your desired application or system. This can involve connecting the API to a front-end interface or incorporating it into a larger software ecosystem.

Test the integration: Verify that the integration between the prediction model and your application is working correctly. Test the API/service by sending sample requests and ensuring that the predicted fare values are returned as expected.

Handle input data: Depending on the requirements of your application, you may need to preprocess or validate the input data before sending it to the prediction model. For example, you might need to ensure the input is in the correct format, handle missing values, or normalize the data.

Monitor and maintain the model: Continuously monitor the performance and accuracy of the deployed model. Keep track of any changes in the data distribution or model behavior that may require retraining or fine-tuning. Regularly update the model as new data becomes available or when improvements are made.

Provide user documentation: Create documentation or guides for users of your application or system, explaining how to utilize the taxi fare prediction functionality. This documentation should cover the API/service endpoints, expected input format, and how to interpret the output predictions.

Secure the deployment: Ensure that appropriate security measures are in place to protect the deployed model and the data it processes. Implement authentication

mechanisms, access controls, and encryption as needed to safeguard sensitive information.

The outlined methodology covers the fundamental steps in predicting NYC Taxi Fare, including data acquisition, preprocessing, exploratory data analysis, visualization, and data cleaning. These steps contribute to developing accurate fare prediction models and thorough dataset analysis.

Conclusion

In this research paper, we explored the task of NYC Taxi Fare Prediction and presented a comprehensive methodology for analyzing the dataset and developing accurate prediction models. Through applying various techniques and visualizations, we gained valuable insights into the factors influencing taxi fares and demonstrated the effectiveness of our approach.By acquiring the NYC Taxi Fare dataset and performing data preprocessing, we ensured the availability of clean and reliable data for analysis. The exploratory data analysis provided a thorough understanding of the dataset's structure and characteristics, allowing us to identify patterns and trends in the fare amounts and passenger counts. Our visualizations, such as histograms and scatter plots, provided visual representations of the data distribution and the spatial dynamics of taxi rides.The developed methodology encompassed vital steps to develop accurate fare prediction models, including feature engineering, data cleaning, and modeling. Leveraging machine learning and deep learning techniques, we constructed prediction models that capture temporal and spatial patterns, enabling us to forecast taxi fare amounts accurately.

Throughout the research process, we emphasized the importance of data quality and appropriate preprocessing techniques to ensure the reliability and robustness of our models. By removing outliers and handling missing values, we aimed to enhance the accuracy and reliability of the fare predictions.Our research contributes to urban transportation systems and fare regulation by providing insights into the factors influencing taxi fares. Accurate fare prediction has significant implications for passengers, taxi companies, and policymakers, enabling efficient resource allocation, fare regulation, and improved customer experience.Additionally, incorporating dynamic pricing mechanisms and considering external events or holidays could enhance the predictive capabilities of our models.In conclusion, our research on NYC Taxi Fare Prediction demonstrates the effectiveness of our methodology in analyzing the dataset, developing accurate prediction models, and providing valuable insights into the factors influencing taxi fares. By leveraging advanced techniques and thorough data analysis, our research contributes to advancing fare prediction methodologies and aids in optimizing urban transportation systems for the benefit of all stakeholders.

References

Effects of investor sentiment on stock volatility: new evidences from multi-source data in China's green stock markets | Financial Innovation | Full Text. https://jfin-swufe.springeropen.com/articles/10.1186/s40854-022-00381-2

How to Predict Taxi Fares in NYC - Dataiku Community. https://community.dataiku.com/t5/New-York-User-Group/How-to-Predict-Taxi-Fares-in-NYC/td-p/14239

Palakuru, Mahesh, Sirisha Adamala, and Harish Bachina. 2020. "Modeling Yield and Backscatter Using Satellite Derived Biophysical Variables of Rice Crop Based on Artificial Neural Networks." Journal of Agrometeorology 22 (1): 41.

Party.biz - Blog View - 9 Distance Measures in Data Science. https://mail.party.biz/blogs/148947/196980/9-distance-measures-in-data-science

Sun, Zhiming, Xianglong Chen, Hanfa Xing, Hongtao Ma, and Yuan Meng. 2020. "Regional Differences in Socioeconomic Trends: The Spatiotemporal Evolution from Individual Cities to a Megacity Region over a Long Time Series." PLoS One 15 (12): e0244084.