NYC Taxi Fare Predictive Analysis

Benjamin Abrahamsen Hagen

1 INTRODUCTION

In this analysis, the primary objective is first to answer three research questions and then develop a predictive model in order to estimate taxi fares in New York City based on historical trip data. The secondary objective of the study is to identify key factors that influence the total fare amount in order to provide insight into the relationships between these factors.

This analysis is based on a dataset that was obtained from the New York City Taxi and Limousine Commission (TLC)[1]. The dataset contains details such as pickup and dropoff locations, trip distances, fare amounts, total amounts, and other related information. In order to prepare the data for analysis, several steps of preprocessing were performed. These steps include converting datetime columns to pandas[8] datetime format, calculating trip durations in minutes, dropping unnecessary columns, and converting categorical variables to numerical values. Additionally, exploratory data analysis was performed in order to better understand the data and identify any patterns or trends.

In the following sections, we will discuss the data analysis process, perform regression analysis using the preprocessed data, and interpret the results to draw conclusions and make recommendations for improvement.

2 DATA ANALYSIS

The data analysis was conducted using various Python libraries to examine taxi demand patterns and revenue generation.
In order to manipulate data and perform calculations, Pandas was used, while Matplotlib and Seaborn[2] were used to create visualizations that helped in understanding the insights obtained from the data.

2.1 Question 1

What is the average demand for the taxis in the days of the week. Which of the days has the highest and which lowest demand?

As a result of this analysis of taxi demand throughout the week, we discovered distinct patterns that can assist the transportation industry in making informed decisions.

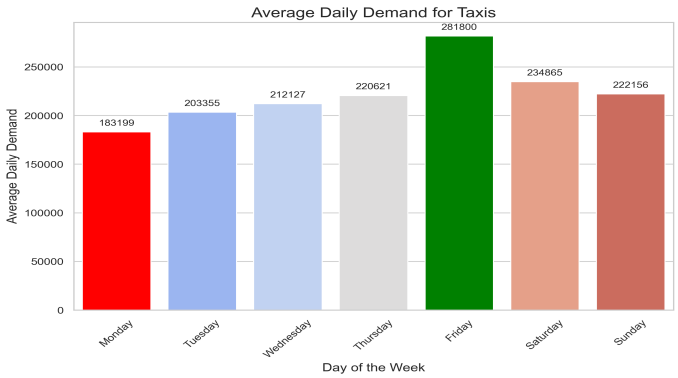

Figure 1: Average Daily Demand for Taxis

In Figure 1 we can see that the highest demand for taxis occurs on Fridays, with an average of 281,800 trips. At the end of the workweek, people may go out for leisure activities or social events, resulting in an increase in demand. On the other hand, Mondays have the lowest demand for taxis, with an average of 183,199 trips. This reduced demand might be due to people staying in or working from home after the weekend.

The demand for taxis gradually increases throughout the rest of the week, as seen in figure 1 with Tuesday having an average of 203,355 trips, Wednesday with 212,127 trips, and Thursday with 220,621 trips. Over the weekend, demand remains relatively high, with Saturday having an average of 234,865 trips and Sunday with 222,156 trips.

As a result of understanding these trends, the company can allocate resources more effectively and optimize its services so that it can meet the needs of its customers. For example, by increasing the number of taxis available on Fridays, the company can better accommodate higher demand and potentially increase revenue. Similarly, resources can be adjusted on Mondays when demand is lower, ensuring efficient use of resources and cost management.

2.2 Question 2

Which time of the day (morning, afternoon, evening, and night) is likely be a peak period for the taxis operation from the data?

Based on this analysis of taxi demand during different times of the day, we can draw valuable conclusions regarding peak periods for taxi services. The days are divided into four time periods: morning (4:00 AM to 11:59 AM), afternoon (12:00 PM to 5:59 PM), evening (6:00 PM to 11:59 PM), and night (12:00 AM to 3:59 AM). The number of trips during each time period can be used to determine when taxis are in high demand.

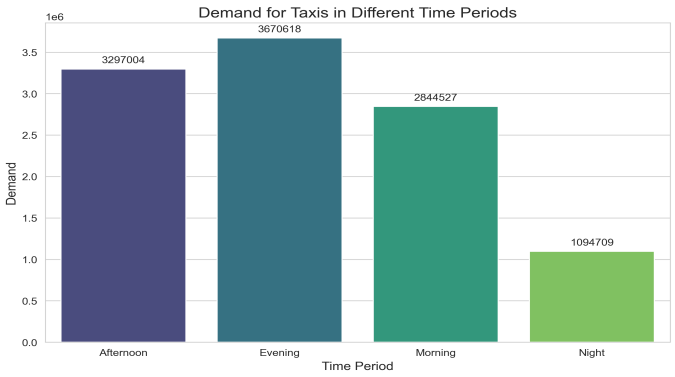

Figure 2: Demand for Taxis in Different Time Periods

Based on the results, in figure 2 we can see that the evening is the peak period for taxi operations, with the highest number of trips. People may be commuting home from work, attending social events or dinners, and participating in evening activities as a result of this increased demand. As the evening is the busiest time for taxis, the transportation business should allocate more resources during this period to accommodate the increased demand and maximize revenue.

On the other hand, taxi demand is lower in the morning, afternoon, and night. By adjusting the number of taxis available during non-peak periods, the transportation business can optimize resource allocation and manage costs more effectively. Overall, knowing the peak period for taxi operations allows the company to make informed decisions regarding resource allocation, cost management, and service optimization to meet customer needs better and improve overall business performance.

2.3 Question 3

On average, how much revenue was generated in the weekdays and weekends for the business for the period covered in the dataset?

The analysis of the dataset reveals that the average daily revenue generated by the taxi business differs significantly between weekdays and weekends.

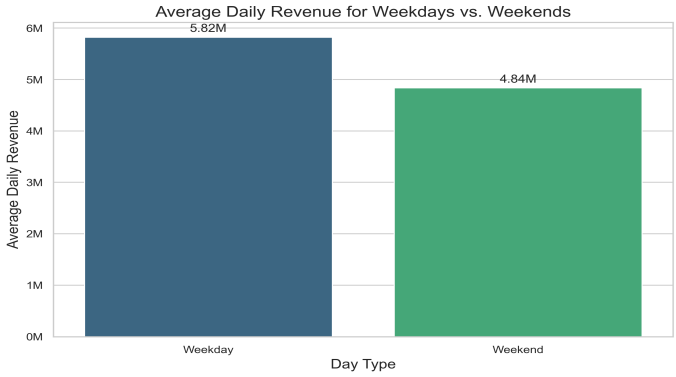

Figure 3: Average Daily Revenue for Weekdays vs. Weekends

As we can see in figure 3 on weekdays, the average daily revenue is approximately 5,821,181, while on weekends, it is approximately 4,835,367. The earnings are higher on weekdays than on weekends, which could be attributed to the increased demand for taxis during the workweek due to work-related travel, business activities, and other events.
In order to optimize operations and maximize profits, the taxi business must take these insights into consideration when making decisions regarding resource allocation, pricing, and scheduling.

3 REGRESSION ANALYSIS

Throughout this section, we will discuss regression analysis, including the modeling process, model output, and the rationale for selecting the particular model.
In addition, we will interpret the results and highlight any interesting findings from the coefficients. Based on its simplicity, ease of interpretation, and suitability for predicting continuous target variables, such as taxi fares, a linear regression model was selected to predict taxi fares.
This analysis begins with linear regression, which assumes a linear relationship between the input features and the target variable.

In this analysis, the dataset was divided into two sets: a training set consisting of 80% of the data and a testing set consisting of 20% of the data. To begin with, all available features were included in the model. Following the training of the model on the training data, predictions were made on the test data, and performance metrics such as root mean squared error (RMSE) and R2 score were calculated[6]. Based on the initial model, an RMSE of 0.0106 and an R2 score of 0.999999 were achieved, indicating a near-perfect fit, raising concerns about possible overfitting.

Table 1: Regression performance metrics (First Model)

Metric	Value
RMSE	0.010586417381644977
R2 Score	0.9999993552960786

Table 2: Predicted total amount values for the New Sample file

Index	Predicted Total Amount
1	5.799968
2	21.299968
3	11.499997
4	7.799993
5	25.300004
6	17.299966
7	9.359987
8	7.799968
9	9.799982
10	17.299969
11	11.759990
12	17.299971
13	8.999984
14	17.999986
15	12.359998
16	6.960000
17	20.159999
18	21.960001
19	36.339847
20	10.790009
21	68.799978
22	53.299640
23	12.739997
24	8.759986
25	28.560009
26	15.359996
27	18.960001
28	10.299982
29	20.160015
30	12.800001
31	8.299993
32	17.159995
33	9.799968
34	6.799969
35	25.560005
36	9.299986
37	4.799973
38	29.750019
39	7.239982
40	8.799968

Consequently, a correlation heatmap as seen in figure 4, was generated in order to identify the features that are most strongly correlated with the 'total_amount' variable.

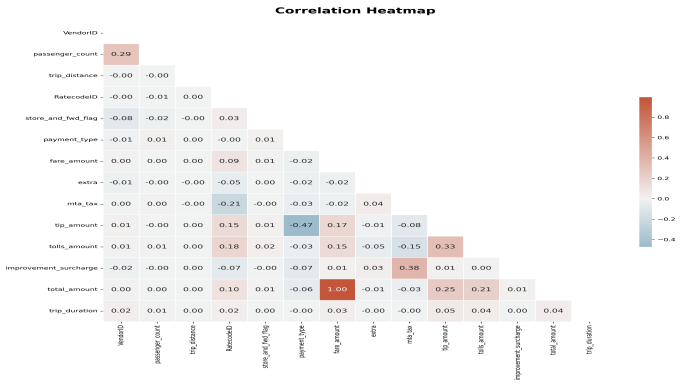

Figure 4: Correlation Heatmap

Based on the heatmap seen in figure 4, the 'fare_amount' variable was found to have the highest correlation. Therefore, a simplified linear regression model was constructed using only 'fare_amount' as the predictor variable.

According to the second model, the RMSE was 3.1786 and the R2 score was 0.9419, which indicates a good fit, although it is not as exceptional as the initial model. As a result of this simplified model, overfitting risks are reduced, and taxi fares can be predicted more accurately.

Table 3: Regression performance metrics (Second Model)

Metric	Value
RMSE	3.1786375916091534
R2 Score	0.9418774958621607

The linear regression model's coefficient for the 'fare_amount' feature indicates how the total amount changes as the fare amount changes. We found that the 'fare_amount' coefficient was positive in our simplified model, suggesting that as the fare amount increases, the total amount increases as well.

Table 4: New Predicted Total Amount Values for the New Sample File (Using Fare Amount)

Index	Predicted Total Amount
1	8.047228
2	23.283826
3	12.618207
4	9.570888
5	23.283826
6	19.220733
7	9.570888
8	9.570888
9	10.586661
10	19.220733
11	11.602434
12	19.220733
13	9.570888
14	18.204960
15	12.110321
16	7.539341
17	18.712847
18	20.236507
19	32.933672
20	10.078774
21	63.914756
22	43.599291
23	11.602434
24	9.063001
25	25.823260
26	14.649754
27	17.697074
28	11.094547
29	18.712847
30	13.126094
31	10.078774
32	16.173414
33	11.602434
34	8.555114
35	23.283826
36	9.570888
37	6.523568
38	25.823260
39	8.047228
40	10.586661

As a result of the regression analysis performed, a simplified linear regression model can be developed that predicts taxi fares using the 'fare_amount' feature.

4 DISCUSSION

The linear regression model developed provided a reasonable prediction of taxi fares using the 'fare_amount' feature. In this simplified model, RMSE was 3.1786 and R2 was 0.9419, indicating a good fit. According to the results, the 'fare_amount' variable is a significant predictor of the 'total_amount' variable.
Although the simplified model performed well, its limitations must also be acknowledged. Linear regression assumes a linear relationship between input features and the target variable, which may not capture complex relationships or interactions between variables[5]. Furthermore, the model relied only on a single variable ('fare_amount'), which might not account for other factors that may have an impact on the fare amount.

The following suggestions can be considered in order to improve the analysis:
Exploring additional features and creating new ones based on existing ones, such as calculating the distance between pickup and dropoff locations or identifying peak hours in order to account for possible surcharges.
The use of more sophisticated models, such as decision trees, random forests [3], or neural networks [4], will enable to capture nonlinear relationships and interactions between variables. Using these models, one may be able to make better predictions and gain a deeper understanding of the data.
Implement cross-validation to assess the model's performance and minimize the risk of overfitting. As a result, the generalization capabilities of a model can be more accurately represented.
Study different feature selection techniques, to identify the most important variables for the model and reduce the risk of overfitting.

As a result, the simplified linear regression model was able to provide reasonable predictions of taxi fares using the 'fare_amount' feature. There are, however, a number of improvements that can be made to the analysis in order to enhance it and potentially achieve better predictions. Future analyses can be made more accurate and insightful by considering the suggestions mentioned. [7].

5 CONCLUSION

The objective of this analysis was to predict taxi fares using the given dataset. In order to understand the data and identify potential relationships among variables, an exploratory data analysis was conducted. The 'fare_amount' feature was then used as the primary predictor for the 'total_amount' variable in a linear regression model.
As a result of the model, the RMSE of the data was 3.1786 and the R2 score was 0.9419, indicating that the 'fare_amount' plays a significant role in determining the total amount. It can be beneficial for both taxi companies and passengers, as this insight can assist in optimizing pricing strategies and providing better estimates of fares.

Although the simplified model provides satisfactory results, there is room for improvement in the analysis. Adding additional features, experimenting with more sophisticated models, and utilizing techniques such as cross-validation and feature selection can enhance the predictive capabilities of the model.
The results of this analysis demonstrate the value of data-driven approaches in predicting taxi fares and lay the groundwork for future research.

REFERENCES

[1] [n. d.]. NYC Taxi and Limousine Commission Trip Record Data. https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

[2] [n. d.]. Seaborn: statistical data visualization. https://seaborn.pydata.org/.

[3] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

[5] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science Business Media.

[6] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning: with Applications in R*. Springer.

[7] John D. Kelleher, Brian Mac Namee, and Aoife D'Arcy. 2015. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. MIT Press.

[8] Wes McKinney. 2017. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython* (2 ed.). O'Reilly Media.