

**Project: Prediction of Taxi Fares and the Behavior Analysis of the Driver**

**Pavan Kumar Duvvuru**

**CS 777**

**Dimitar Trajanov, PHD**

**10/17/2023**

**Contents of the project:**

1. **Overview**
2. **About the dataset**
3. **Exploratory Data Analysis (EDA)**
4. **Fare Prediction**
5. **Behavior Analysis of Taxi Drivers**
6. **Scope**
7. **Conclusion**
8. **Overview:**

In this project, we delved into the intricate dynamics of the taxi industry, using a rich dataset to shed light on drivers' earnings and behaviors. We scrutinized the variations in fares, how often taxis were hailed, and the prevalent payment methods. Our exploration included visual interpretations like heat maps and pie charts, offering a comprehensive understanding of hourly fare trends and their correlations. This endeavor not only highlighted key patterns but also pinpointed anomalies, serving as a guide for future improvements and strategies in the taxi business.

1. **About Dataset:**

The data set itself is a simple text file. Each taxi trip report is a different line in the file. Each trip report includes the starting point, the drop-off point, corresponding timestamps, and information related to the payment. The data are reported by the time that the trip ended, i.e., upon arriving in the order of the drop-off timestamps. The attributes present on each line of the file are, in order:

|  |  |  |
| --- | --- | --- |
|  | **Attributes** | **Description** |
| 0 | medallion | an md5sum of the identifier of the taxi - vehicle bound (Taxi ID) |
| 1 | hack license | an md5sum of the identifier for the taxi license (Driver ID) |
| 2 | pickup datetime | time when the passenger(s) were picked up |
| 3 | dropoff datetime | time when the passenger(s) were dropped off |
| 4 | trip time in secs | duration of the trip |
| 5 | trip distance trip | distance in miles |
| 6 | pickup longitude | longitude coordinate of the pickup location |
| 7 | pickup latitude | latitude coordinate of the pickup location |
| 8 | dropoff longitude | longitude coordinate of the drop-off location |
| 9 | dropoff latitude | latitude coordinate of the drop-off location |
| 10 | payment type | the payment method -credit card or cash |
| 11 | fare amount | fare amount in dollars |
| 12 | surcharge | surcharge in dollars |
| 13 | mta tax | tax in dollars |
| 14 | tip amount | tip in dollars |
| 15 | tolls amount | bridge and tunnel tolls in dollars |
| 16 | total amount | total paid amount in dollars |

1. **Exploratory Data Analysis (EDA):**

Exploratory Data Analysis (EDA) is a crucial preliminary step in data processing and analysis, as it allows us to understand the nature of our data, identify potential anomalies or outliers, ascertain the quality of the data, and prepare it for further analysis or modeling. In this taxi ride project, EDA was extensively performed to comprehend the dataset and extract insights from it.

**Null values analysis:**

|  |  |
| --- | --- |
| Attributes | Null values |
| medallion | 0 |
| hack license | 0 |
| pickup datetime | 0 |
| dropoff datetime | 0 |
| trip time in secs | 0 |
| trip distance trip | 0 |
| pickup longitude | 0 |
| pickup latitude | 0 |
| dropoff longitude | 0 |
| dropoff latitude | 0 |
| payment type | 0 |
| fare amount | 0 |
| surcharge | 0 |
| mta tax | 0 |
| tip amount | 0 |
| tolls amount | 0 |
| total amount | 0 |

a thorough check for missing values was conducted across all attributes, confirming the absence of null values. This ensures data integrity and avoids potential disruptions during subsequent analyses.

**Distribution of Trips:**

**A graph of a trip distance

Description automatically generated**

The histogram illustrates the distribution of trip distances for taxi rides. Most rides have a trip distance between 0.5 to 2.5 miles, peaking around 1 to 1.5 miles. As the trip distance increases beyond 2.5 miles, the frequency of such trips diminishes. There are very few rides with distances above 4.5 miles, indicating that longer trips are less common in this dataset. The bell-like shape suggests a near-normal distribution with a slight right-skew, meaning there are more shorter trips than longer ones.

**Average fare amount per-hour:**

**A graph with a line

Description automatically generated**

the graph depicts the average fare amount for taxi rides throughout different hours of the day. Here's a breakdown:

**Early Morning (0-5 hours):** The average fare begins modestly and then peaks around the 3rd or 4th hour (likely the early morning rush) and sharply drops after that.

**Morning (5-10 hours):** There's a notable dip in the average fare during the early to mid-morning hours.

**Daytime (10-15 hours):** The fare stabilizes and exhibits minor fluctuations throughout the day, indicating consistent pricing.

**Evening (15-20 hours):** Towards the evening, there is a gradual rise in the average fare amount, likely representing the evening rush or increased demand.

**Late Evening (20 onwards):** The graph shows a steep incline in the fare as we approach midnight, suggesting that fares tend to be higher late at night, possibly due to reduced availability of taxis or higher demand.

Overall, the graph provides insights into the taxi fare dynamics throughout the day, highlighting periods of increased fare, likely tied to demand and availability.

**Correlation Heatmap:**

**A screenshot of a graph

Description automatically generated**

This is a correlation heatmap displaying relationships between variables:

* trip\_distance and trip\_time\_in\_secs have a strong positive correlation of 0.74, indicating they tend to increase together.
* amount is strongly positively correlated with both trip\_distance and trip\_time\_in\_secs at 0.92.
* Other variables have weak correlations, close to 0, suggesting minimal relationships between them.

**Pie Chart of Payment Types:**

**A blue and orange circle with numbers

Description automatically generated**

This pie chart shows payment methods: 50.7% use cash (CSH), 49.2% use card (CRD), and 0.0% use another method (N/A)

1. **Machine learning Model:**

In the analysis, we applied machine learning techniques to predict taxi fare amounts based on a set of predictors. The predictors encompassed trip\_distance, the hour of pickup denoted as pickup\_hour, and geospatial data points that include both the latitude and longitude of pickup (pickup\_latitude and pickup\_longitude) and drop-off (dropoff\_latitude and dropoff\_longitude) locations. To prepare the data for model training, we utilized the VectorAssembler function, amalgamating all the selected features into a unified feature vector. This facilitated the subsequent training process and ensured computational efficiency. Once our data was preprocessed, we partitioned it into training and testing subsets with an 80-20 split, ensuring that our model would be validated on unseen data. We employed linear regression as our predictive model due to its simplicity and interpretability, especially beneficial for initial exploratory modeling. After training the model on the designated training dataset, it was time to assess its performance. For this, the model was used to make predictions on the test data. Model performance is a critical aspect of any machine learning task; thus, we utilized the Root Mean Squared Error (RMSE) metric, which provides insight into the model's accuracy by quantifying the difference between the predicted and actual values. An RMSE value essentially communicates the standard deviation of the residuals or prediction errors. In this analysis, our model achieved an RMSE of 1.3756471056022814 on the test data. This value indicates the average magnitude of error between our predicted fare amounts and the actual fare values in the test dataset. While this RMSE value provides a baseline measure of our model's performance, further analyses, model tuning, and potentially incorporating additional features or trying different algorithms might enhance predictive accuracy.

To know the RMSE effect we did the distribution of the fare amount: A graph of blue rectangular bars

Description automatically generated with medium confidence

The histogram displays the distribution of fare amounts in bins of size 2, with the 4-10 range being most frequent. Fare amounts of 2-4 are rare, while fares above 10 declines in frequency. An RMSE of 1.3756471056022814 indicates the model's average prediction deviation. Given the distribution, this error suggests the model performs well.

1. **Behavior Analysis of Taxi Drivers:**

Anomalies in total\_fare\_earned: +--------------------+------------------+---------------+------------------+

| hack\_license| total\_fare\_earned|number\_of\_trips| avg\_tip\_received|

+--------------------+------------------+---------------+------------------+

|F49FD0D84449AE7F7...| 4306.5| 434|0.7935483870967742|

|97F7B431B057B98EA...|3757.8899999999994| 356|0.9266573033707863|

|9E035DBF346FDE01F...|3856.4900000000002| 377|0.8156233421750662|

|51C1BE97280A80EBF...| 4232.239999999999| 411|0.9110462287104623|

|DB1B4490DA4A46A7A...| 4400.29| 411| 0.827712895377129|

|C9674190984BA193F...| 3810.64| 363|0.7948760330578513|

+--------------------+------------------+---------------+------------------+

Anomalies in number\_of\_trips: +--------------------+------------------+---------------+-------------------+

| hack\_license| total\_fare\_earned|number\_of\_trips| avg\_tip\_received|

+--------------------+------------------+---------------+-------------------+

|F49FD0D84449AE7F7...| 4306.5| 434| 0.7935483870967742|

|97F7B431B057B98EA...|3757.8899999999994| 356| 0.9266573033707863|

|9E035DBF346FDE01F...|3856.4900000000002| 377| 0.8156233421750662|

|9112D33A328C37CF6...|3558.5899999999997| 351| 0.8364387464387464|

|51C1BE97280A80EBF...| 4232.239999999999| 411| 0.9110462287104623|

|DB1B4490DA4A46A7A...| 4400.29| 411| 0.827712895377129|

|C9674190984BA193F...| 3810.64| 363| 0.7948760330578513|

|00B7691D86D96AEBD...| 2075.11| 489|0.10118609406952966|

+--------------------+------------------+---------------+-------------------+

Anomalies in avg\_tip\_received:

+--------------------+------------------+---------------+------------------+

| hack\_license| total\_fare\_earned|number\_of\_trips| avg\_tip\_received|

+--------------------+------------------+---------------+------------------+

|B508465FAC4F54A40...| 113.5| 10| 0.0|

|A2AAF2102F3D8AF40...| 18.6| 1| 3.1|

|E9B0F9B61761F9F37...| 12.5| 1| 2.0|

|560DE283367A74CE6...| 15.62| 1| 3.12|

|EEDE57108D76F5B35...| 23.4| 2| 1.95|

|685E69EDB24986567...| 16.0| 1| 1.5|

|D8856767BB390CD43...| 37.5| 3|1.8333333333333333|

|4215A39B9D316E2D7...|1612.8899999999999| 147|1.5434693877551022|

|C2A4BD9771B4A78B4...| 19.0| 2| 1.75|

|CAD29CEA4D2A2B5EA...| 23.0| 2| 1.75|

|66B67569099D08421...| 29.6| 2| 2.05|

|BB6566AE14108C9B8...| 11.4| 1| 1.9|

|993592AE0ADFDD11C...|30.950000000000003| 2| 2.725|

|9794EE180CD9B2FD4...| 12.0| 1| 2.0|

|E8847883EC19A33ED...| 12.6| 1| 2.1|

|729642394A25B2518...| 34.3| 2| 3.15|

|7DEB8806DE87A4007...| 30.7| 2| 2.35|

|F007C1A5716E48DEA...| 41.35| 3|2.4499999999999997|

|AB92B799A4DFDECBF...| 9.0| 1| 1.5|

|B10F570869AF86DF5...| 29.8| 3|1.5999999999999999|

+--------------------+------------------+---------------+------------------+

only showing top 20 rows

In the realm of the taxi industry, understanding driver behavior is paramount for both service providers and policymakers. This analysis delves into the intricate web of factors that influence how taxi drivers perform, earn, and interact with customers. By scrutinizing a dataset encompassing a multitude of attributes, we unravel key insights into the behavior of taxi drivers and their role within the larger transportation ecosystem.

**Total Fare Earned:** One of the pivotal aspects in driver behavior analysis is the examination of the total fare earned by each driver. Our dataset revealed a significant disparity, with top-performing drivers earning as much as $4,400.29, while others earned as little as $9.0. This variance is indicative of diverse driver performance levels and underscores the potential impact of factors such as location, working hours, and the ability to provide quality service. Insights into these discrepancies can aid service providers in optimizing driver incentives and support.

**Number of Trips:** The number of trips completed by drivers serves as another vital metric in assessing their behavior. The dataset portrays a wide spectrum, with some drivers executing 434 trips and others just one. This variance underscores the differing levels of driver activity and availability. Factors like work schedules and geographical coverage zones play a pivotal role in determining the number of trips a driver can undertake.

**Average Tip Received:** Understanding the average tip received by drivers provides deeper insights into customer interactions and satisfaction. Our analysis uncovered a diverse range of tip amounts, from a mere $0.0 to a generous $3.12 on average per trip. This variation indicates that some drivers consistently receive substantial tips, while others receive minimal or no tips. The reasons for these differences could range from the quality of customer service to trip durations.

**Anomalies:** One intriguing aspect of this analysis was the identification of anomalies within the data. These anomalies encompassed drivers who earned exceptionally high fares or received unusually large tips. These outliers might be attributed to special events, lengthy trips, or outstanding customer service. By dissecting these anomalies further, service providers can gain valuable insights into exceptional driver performance and tailor strategies accordingly.

**Clustering:** To gain a deeper understanding of driver behavior, we applied K-Means clustering to group drivers based on their total fare earned, the number of trips, and average tip received. This clustering technique unveiled distinct driver segments, including high earners with moderate tip averages, low earners with minimal tips, and a segment with moderately high earnings and tips. This categorization empowers service providers to customize incentives, training, and support for different driver groups, ultimately enhancing overall driver satisfaction.

1. **Future scope:**

The future scope for this behavior analysis of taxi drivers includes implementing real-time monitoring systems and leveraging machine learning to predict driver behavior, enhancing driver training programs, and refining customer satisfaction strategies to optimize the taxi service industry's efficiency and customer experience.

1. **Conclusion:**

In conclusion, this project provided a comprehensive exploration of taxi driver behavior, highlighting variations in total fare earnings, trip frequency, and average tips. These findings underscore the importance of tailored incentives and support to optimize driver satisfaction and productivity. Exceptional driver performance was revealed through anomaly identification, offering valuable insights for improvement.

K-Means clustering categorized drivers into distinct segments, facilitating targeted strategies. The project lays the groundwork for implementing real-time monitoring and machine learning for enhanced driver behavior analysis. The future of the taxi industry hinges on optimizing driver training, refining customer satisfaction, and improving the overall experience. Leveraging these insights, we can steer the industry toward greater efficiency and customer satisfaction, benefitting all stakeholders.