NYC TLC Trip Dataset

Ninad Shah | Sai Varun Kanduri | Sumit Mamtani
1229719714 | 1230541249 | 1231155720
MS DSAE | MS DSAE | MS DSAE

Introduction:

The New York Taxi dataset serves as a captivating gateway into the intricate dynamics of urban transportation within one of the world's most bustling and complex environments – the iconic cityscape of New York. This rich dataset, meticulously sourced from the Taxi and Limousine Commission (TLC) Trip Record Data, provides an exhaustive and granular view of taxi journeys, encapsulating vital information such as precise timestamps, geographical coordinates, trip distances, fare breakdowns, rate classifications, payment methods, and passenger counts. Through the lens of this extensive dataset, our project embarks on a profound exploration, seeking to unravel the nuanced behavior of tipping within the distinctive yellow cabs that navigate the vibrant streets of New York City.

The study of tipping dynamics takes center stage as a fascinating inquiry, offering a unique perspective into the cultural and behavioral facets of service gratuity within the taxi sector. Beyond its intrinsic intrigue, this examination carries significant implications for both taxi drivers and passengers. Leveraging advanced statistical and machine learning techniques, our overarching objective is to develop predictive models that not only estimate the likelihood of receiving a tip but also discern the probable amount. This analytical endeavor contributes to a deeper understanding of the economic and sociocultural underpinnings embedded in the heart of New York City's dynamic urban fabric.

Data Collection:

The dataset curated by the TLC stands as a testament to meticulous record-keeping, encompassing a diverse array of attributes for each taxi trip. From the precision of timestamps to the spatial coordinates marking pick-up and drop-off locations, and from trip distances to fare breakdowns and payment methods – the dataset unfolds as a comprehensive tapestry of the myriad factors influencing the taxi experience. It is within this wealth of information that our exploration finds its foundation, promising a nuanced understanding of tipping dynamics in this bustling urban environment.

Summary of the Dataset:

The TLC (Taxi and Limousine Commission) Trip Record Data stands as an extensive and multifaceted dataset, providing a comprehensive overview of taxi and limousine journeys across the dynamic landscape of New York City. This rich repository encompasses a diverse array of intricate data points, encapsulating the essence of each trip with meticulous detail. The dataset includes:

- Timestamps: Precise timestamps capturing the nuanced moments of both pick-up and drop-off, offering a temporal dimension to the dataset.

- Geographic Coordinates: Latitude and longitude coordinates intricately mapping the spatial trajectory of each journey, providing a geospatial context to the dataset.

- Trip Distances: Information detailing the distances covered during each trip, shedding light on the spatial dynamics of travel within the city.

- Fare Breakdown: A detailed breakdown of fare components, offering transparency into the pricing structure, unveiling the financial intricacies of each journey.

- Rate Types: The dataset classifies the rate applied to each trip, recognizing variations based on factors such as the time of day, contributing to a nuanced understanding of fare dynamics.

- Payment Methods: Insightful data revealing how passengers choose to settle their fares, whether through credit card transactions, cash payments, or alternative means, reflecting diverse payment preferences.

- Passenger Counts: The number of passengers on each trip, as reported by the driver, serves as a crucial indicator of the occupancy dynamics within the vehicles.

This expansive and detailed dataset serves as a cornerstone for analyses, offering researchers and practitioners a wealth of information to unravel the intricate tapestry of transportation behaviors within the vibrant and bustling cityscape of New York.

Hypothesis:

Our research is centered on a well-crafted hypothesis that seeks to disentangle the complex interactions among variables that affect tipping behavior. We propose that tipping is a multidimensional outcome that is influenced by several factors such as trip duration, temporal considerations, and the socio-economic context entrenched in the pick-up and drop-off sites. Tipping is a traditional behavior strongly rooted in the service sector. Our hypothesis functions as a beacon, directing our analytical process in the direction of identifying the minute details that add up to the observed variety in tipping patterns in the dataset.

Importance of Solution:

The significance of understanding tipping behavior extends far beyond the realm of academic curiosity. For taxi drivers navigating the labyrinthine streets of New York, these insights offer tangible benefits. They provide drivers with a panoramic view of potential income streams, enabling them to strategize and optimize their routes and service quality. On a broader scale, our insights contribute to the ongoing discourse surrounding urban mobility, customer satisfaction, and the economic dynamics that define the broader transportation sector within the city.

In the subsequent sections, we embark on a detailed journey of exploration, employing analytical techniques to unravel the complexities of tipping patterns. As our models evolve, so too does our understanding of the intricate dance between passengers and taxi drivers within the unique urban tapestry of New York City.

Data Analysis and Visualization:

Summary Statistics for Key Variables in the Taxi Trip Dataset:

	trip_distance	RatecodeID	fare_amount	payment_type	Airport_fee
count	5.670178e+06	5.670178e+06	5.670178e+06	5.670178e+06	5.670178e+06
mean	3.552810e+00	4.000710e-01	2.010128e+01	1.187945e+00	1.307321e+00
std	4.704921e+00	1.490626e+00	1.928603e+01	5.749371e-01	8.036970e-01
min	0.000000e+00	0.000000e+00	-2.823000e+02	0.000000e+00	0.000000e+00
25%	1.030000e+00	0.000000e+00	9.300000e+00	1.000000e+00	1.000000e+00
50%	1.800000e+00	0.000000e+00	1.420000e+01	1.000000e+00	1.000000e+00
75%	3.520000e+00	0.000000e+00	2.330000e+01	1.000000e+00	1.000000e+00
max	5.722200e+02	7.000000e+00	3.243000e+02	4.000000e+00	4.000000e+00

	tip_amount	pickup_weekday	pickup_hour	pickup_week_hour
count	5.670178e+06	5.670178e+06	5.670178e+06	5.670178e+06
mean	3.517125e+00	2.888152e+00	1.421218e+01	8.352784e+01
std	4.176793e+00	1.886374e+00	5.808972e+00	4.525191e+01
min	-3.308800e+02	0.000000e+00	0.000000e+00	0.000000e+00
25%	0.000000e+00	1.000000e+00	1.000000e+01	4.400000e+01
50%	2.800000e+00	3.000000e+00	1.500000e+01	8.400000e+01
75%	4.450000e+00	4.000000e+00	1.900000e+01	1.190000e+02
max	4.111000e+02	6.000000e+00	2.300000e+01	1.670000e+02

	pickup_minute	dropoff_weekday	dropoff_hour	dropoff_week_hour
count	5.670178e+06	5.670178e+06	5.670178e+06	5.670178e+06
mean	2.957056e+01	2.891712e+00	1.425701e+01	8.365809e+01
std	1.734417e+01	1.888900e+00	5.927075e+00	4.525322e+01
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	1.400000e+01	1.000000e+00	1.100000e+01	4.400000e+01
50%	3.000000e+01	3.000000e+00	1.500000e+01	8.400000e+01
75%	4.500000e+01	4.000000e+00	1.900000e+01	1.190000e+02
max	5.900000e+01	6.000000e+00	2.300000e+01	1.670000e+02

	dropoff_minute
count	5.670178e+06
mean	2.958335e+01
std	1.740166e+01
min	0.000000e+00
25%	1.500000e+01
50%	3.000000e+01
75%	4.500000e+01
max	5.900000e+01

Table1. Summary Statistics of the features used for training of the model

Inferences on the summary:

- Trip Distance: The average trip distance is approximately 3.55 miles, with a relatively high standard deviation of 4.70, indicating a wide variation in trip distances.The minimum trip distance of 0 raises concerns.

- RatecodeID: Most trips (around 40%) have a RatecodeID of 0, suggesting that a significant portion of the trips is likely standard fare trips.

- Fare Amount: The average fare amount is $20.10, with a standard deviation of $19.29, indicating considerable variability in fare prices.The presence of negative minimum fare amounts and a maximum fare of $324.30 suggests potential data anomalies that require investigation.

- Payment Type: The majority of payments (about 75%) are made using payment type 1, indicating a prevalent payment method.

- Airport Fee: The average airport fee is approximately $1.31, and most records have a non-zero airport fee, suggesting that a significant portion of trips may involve airport-related charges.

- Tip Amount: The average tip amount is $3.52, with a considerable standard deviation of $4.18. The presence of negative tip amounts and extremely high positive tip amounts may indicate potential issues or anomalies in the data.

- Pickup and Dropoff Time: Trips are distributed across different weekdays and hours, with a peak around 14:00 (2:00 PM) for both pickup and dropoff times. The similar patterns in pickup and dropoff times suggest symmetry in the distribution of trips throughout the day.

- Weekday and Hour Combos: The combination of pickup weekday and hour shows a relatively even distribution, with some peak hours, indicating consistent demand during specific time periods.

- Pickup and Dropoff Minutes: The minutes for pickup and dropoff show a fairly uniform distribution, suggesting no strong patterns in specific minutes of the hour.

- General Observations: Further investigation is needed to understand the anomalies in trip distance, fare amount, and tip amount, which may impact the overall quality of the dataset.The dataset reveals significant variability in trip-related metrics, such as trip distance, fare amount, and tip amount. This suggests a diverse range of taxi trips, including both short and long-distance journeys, with varying fare structures and tipping behavior.

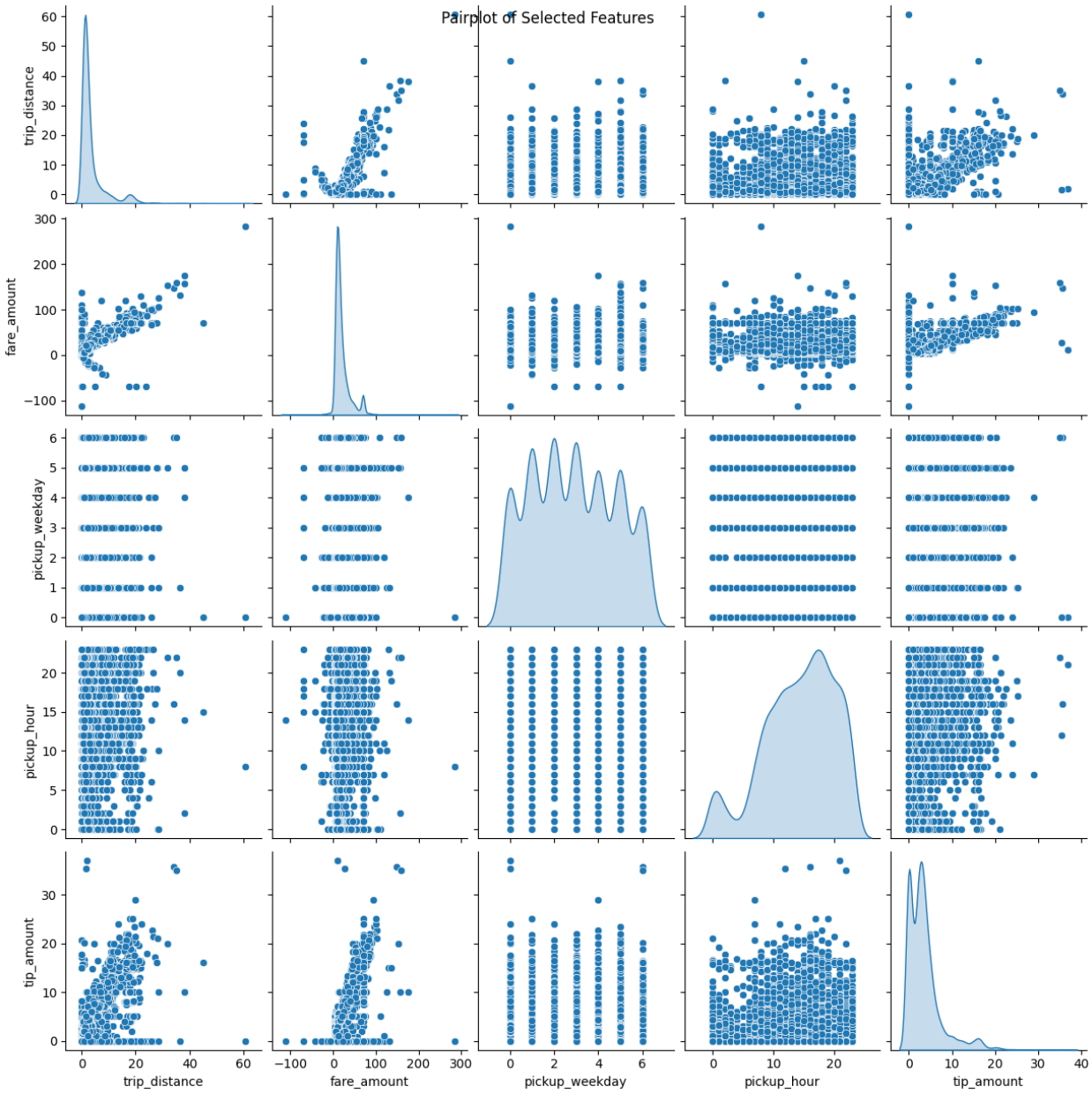

Fig1. Pairplot of Selected Features of the dataset

The graph above shows the correlation between the different features, and really helps to understand the dataset. We can see that there are many outliers within the dataset with negative fare amounts and tip amounts. We have removed almost all excessive outliers that would be bad for the training of the model but not entirely to help prevent the overfitting of the model on the dataset.

Problems found with Dataset and how we tackled them:

Data Imbalance: The target variable (tip_amount) is highly skewed and contains outliers. We kept some outliers to prevent overfitting of data. Sometimes outliers might seem like extreme values that could potentially skew the analysis, but here, we removed the outliers that would heavily influence the training of the model.

Feature Importance: Analyzed correlations and distributions of features to understand their relationship with the target variable. Features with low correlation didn't contribute much to the model's prediction, and we removed them.

Overfitting: After monitoring the model's performance on both training and validation sets, large differences between training and validation losses caused overfitting. We used hyperparameter tuning methods and avoided overfitting.

GPU Usage: Since the dataset is too large, GPU availability was a must for training the neural network. We used the T4 GPU on google colab. We couldn't use the dataset to its full potential as even more powerful computation was required to perform statistical analysis on it.

Model Building Process:

- Data Exploration: Descriptive statistics, information about the dataset, and a correlation heatmap provide an overview of the dataset.

- Data Preprocessing: Data from the NYC TLC yellow taxi for August and September 2023 is loaded using Pandas, with date columns parsed. Time-related features are extracted from the pickup and dropoff timestamps. Categorical variables are encoded using LabelEncoder. The remove_outliers_IQR_df function is defined to remove outliers based on the Interquartile Range (IQR). Outliers are removed separately for features and the target variable (tip_amount).

- Data Sampling: The data_sampling function is defined to take a random sample of 9999 records from the dataset. This sample is used for visualization purposes later in the code.

- Outlier Handling: Outliers are removed during data preprocessing to improve model robustness.

- Neural Network Model: A Sequential neural network model is created using Keras, consisting of input, hidden, and output layers. The model is compiled with the mean squared error loss function and the Adam optimizer. GPU acceleration is utilized for training the model.

➢ Model Training and Evaluation: The model is trained using the training set (X_train_scaled and y_train) for 50 epochs with a batch size of 8196. The training history is stored in the history variable. The model is evaluated on the test set (X_test_scaled and y_test), and the mean squared error is printed as the test loss.

➢ Evaluation: Model performance is assessed using the mean squared error, and visualizations help in understanding the model's behavior.

➢ Visualization: A KDE (Kernel Density Estimation) plot is created to visualize the distribution of true and predicted tip amounts.

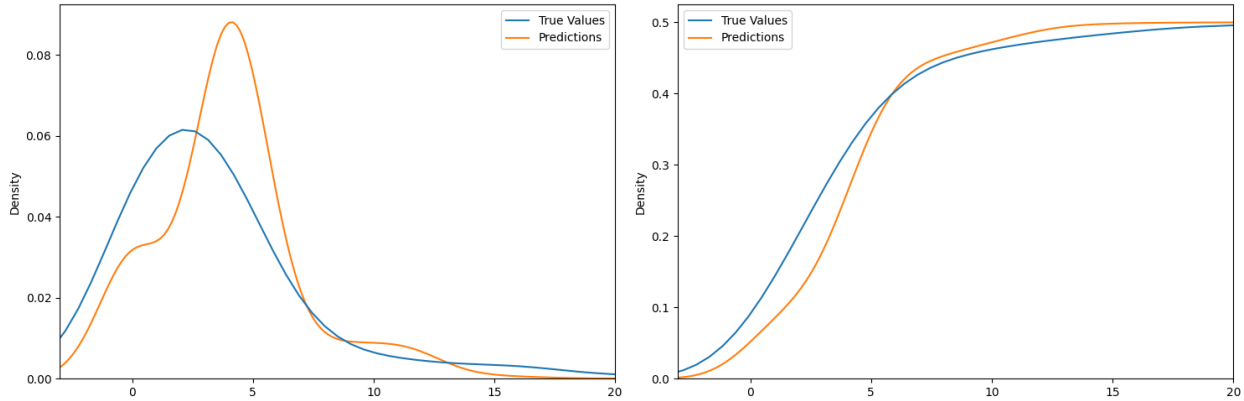

Fig2. KDE Plot of True values of the tip amount vs the Predicted amount

➢ Model Comparison: True and predicted tip amounts are compared using scatter plots, and a heatmap is created to visualize the correlation between features and tip amounts.

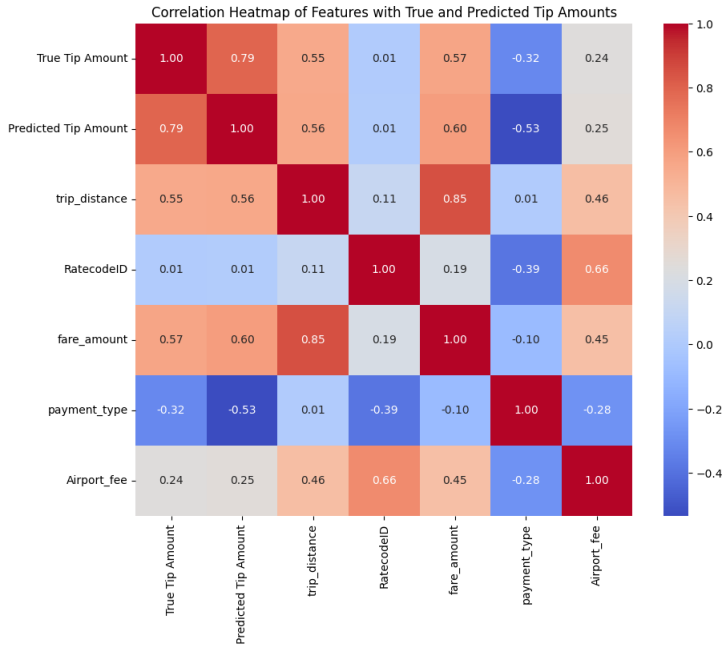

Fig3. Correlation Heatmap of the features with true and predicted tip amounts

➢ Further Visualization: Scatter plots are generated for trip distance vs tip amount and fare amount versus tip amount, with outliers removed.

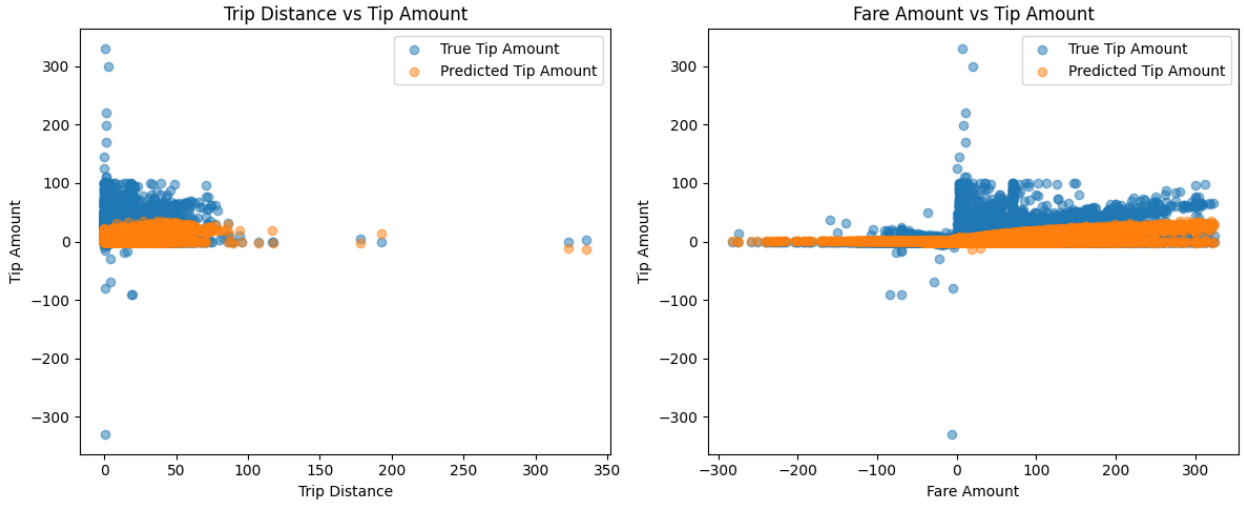

Fig4. Scatter Plots of True tip amounts vs Predicted tip amounts vs Trip distance and Fare amount

Reason of choosing Neural Network model over other Potential Models:

➢ Linear Regression: Assuming a linear relationship between features and tip amounts, linear regression could be suitable if the underlying patterns are relatively straightforward. For instance, linear regression might perform well if tip amounts primarily depend on factors like trip distance and fare amounts in a directly proportional manner. However, the NYC taxi dataset is likely to exhibit more complex relationships, considering the dynamic and multifaceted nature of taxi rides in a bustling metropolis.

➢ Support Vector Machine(SVM): If Support Vector Machine (SVM) models were employed, interpretability might be enhanced, and computational efficiency could be advantageous for datasets with moderate sizes. SVM's robustness to overfitting and suitability for capturing non-linear relationships, especially with the proper kernel, might offer competitive performance. Ultimately, the choice between the neural network and SVM models would hinge on the dataset's specific characteristics, like size of the dataset. Since the dataset is too large, we discounted the SVM model due to lack of computational power.

Conclusion:

In examining the intricate New York City Taxi dataset, our focus centered on unraveling tipping behavior within this bustling metropolis. The dataset, sourced from TLC Trip Record Data, presented comprehensive details of taxi journeys. Challenges like data imbalances and outliers were addressed through robust preprocessing.

Employing a Sequential neural network model in Keras, we aimed to predict tip amounts based on diverse trip variables. Despite computational limitations, the model was chosen for its capacity to handle complex relationships within the dataset. Evaluation via mean squared error and visualizations like KDE plots and scatter plots unveiled insights into tipping patterns and feature importance.

Our findings shed light on the nuanced interplay of factors influencing tipping behavior in NYC taxi rides, vital for drivers optimizing earnings and broader discussions on urban transportation economics. This exploration, despite limitations, offers a stepping stone for further research, providing a richer understanding of tipping dynamics within New York City's vibrant urban landscape.

Future Work:

Expanding the scope of our analysis beyond the limitations of computational resources presents several promising avenues for further exploration and enhancement of predictive models for tip amounts in New York City's taxi rides.

- Seasonal Variations: Extending the dataset to encompass a full year's worth of taxi trip data for all twelve months can provide a more comprehensive understanding of tipping behavior across different seasons. Analyzing tipping patterns concerning seasonal changes, such as summer tourism influxes or winter holiday periods, would yield valuable insights into how tipping behaviors fluctuate throughout the year.

- Geospatial Integration: Incorporating the geographical coordinates of pickup and drop-off locations as additional features in predictive models could significantly enhance predictive accuracy. This geospatial integration can illuminate how tipping behavior correlates with specific neighborhoods, landmarks, or even socio-economic factors in different areas of the city. Understanding how location influences tipping practices can be pivotal for drivers to strategize their routes and service quality.

- Model Refinement and Optimization: Exploring more advanced machine learning algorithms or ensemble techniques beyond neural networks, such as gradient boosting or random forests, can potentially improve prediction accuracy and generalizability. Fine-tuning hyperparameters and conducting more extensive model validations can lead to more robust and reliable predictive models.

➢ Interactive Visualization and Interpretability: Developing interactive visualizations or dashboards can facilitate better interpretation of model predictions and feature importance. This would enable stakeholders, including taxi drivers and policymakers, to grasp tipping behavior trends intuitively and make informed decisions based on these insights.

By delving deeper into these aspects, future analyses can provide a richer understanding of tipping behaviors in NYC taxi rides, contributing not only to the drivers' strategies but also to broader discussions on urban transportation dynamics and passenger preferences.

References:

➢ Visualizing and analyzing New York City yellow ... Accessed November 30, 2023. https://www.stat.cmu.edu/capstoneresearch/spring2022/315files_s22/team23.html.

➢ Zhong, Haonan. "Unraveling NYC Yellow Taxi Patterns: An in-Depth Exploratory Data Analysis." Medium, November 28, 2023. https://medium.com/@haonanzhong/new-york-city-taxi-data-analysis-286e08b174a1.

➢ "Bodo: Machine Learning Series: NYC Yellow Taxi Tips Prediction." bodo.ai. Accessed December 2, 2023. https://www.bodo.ai/blog/machine-learning-series-nyc-yellow-taxi-tips-prediction.

➢ ChatGPT: OpenAI. (2023). ChatGPT (Mar 14 version) [Large language model]. https://chat.openai.com/chat

➢ "TLC Trip Record Data." TLC Trip Record Data - TLC. Accessed December 4, 2023. https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

➢ Anand, Sejal. "Exploratory Data Analysis on NYC Taxi Trip Duration Dataset." Analytics Vidhya, Accessed December 1, 2023. https://www.analyticsvidhya.com/blog/2021/01/exploratory-data-analysis-on-nyc-taxi-trip-duration-dataset/.

Appendix

Code Used to get the Inferences, Graphs, and Predictions

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, LabelEncoder
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Dropout
from tensorflow.keras.optimizers import Adam
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import mean_squared_error, r2_score
import tensorflow as tf
from scipy import stats
august_data =
pd.read_csv('/content/drive/MyDrive/TLC_Yellow/yellow_tripdata_2023-08.csv',
parse_dates=['tpep_pickup_datetime', 'tpep_dropoff_datetime'])
september_data =
pd.read_csv('/content/drive/MyDrive/TLC_Yellow/yellow_tripdata_2023-09.csv',
parse_dates=['tpep_pickup_datetime', 'tpep_dropoff_datetime'])

combined_data = pd.concat([august_data, september_data], ignore_index=True)
selected_features = ['trip_distance', 'RatecodeID', 'fare_amount','payment_type',
'Airport_fee', 'tip_amount']

taxi_filtered = combined_data[selected_features]

taxi_filtered['pickup_weekday'] = combined_data['tpep_pickup_datetime'].dt.weekday
taxi_filtered['pickup_hour'] = combined_data['tpep_pickup_datetime'].dt.hour
taxi_filtered['pickup_week_hour'] = (taxi_filtered['pickup_weekday'] * 24) +
taxi_filtered['pickup_hour']
taxi_filtered['pickup_minute'] = combined_data['tpep_pickup_datetime'].dt.minute

taxi_filtered['dropoff_weekday'] = combined_data['tpep_dropoff_datetime'].dt.weekday
taxi_filtered['dropoff_hour'] = combined_data['tpep_dropoff_datetime'].dt.hour
taxi_filtered['dropoff_week_hour'] = (taxi_filtered['dropoff_weekday'] * 24) +
taxi_filtered['dropoff_hour']
taxi_filtered['dropoff_minute'] = combined_data['tpep_dropoff_datetime'].dt.minute

def remove_outliers_zscore(df, columns):
    z_scores = stats.zscore(df[columns])
    abs_z_scores = abs(z_scores)
    filtered_entries = (abs_z_scores < 3).all(axis=1)
    return df[filtered_entries]
columns_to_check = ['trip_distance', 'fare_amount']
taxi_filtered_no_outliers = remove_outliers_zscore(taxi_filtered, columns_to_check)

label_encoders = {}
categorical_columns = ['RatecodeID', 'payment_type','Airport_fee']

for col in categorical_columns:
    label_encoders[col] = LabelEncoder()
    taxi_filtered_no_outliers[col] =
label_encoders[col].fit_transform(taxi_filtered_no_outliers[col])

X = taxi_filtered_no_outliers.drop('tip_amount', axis=1)
```

```
y = taxi_filtered_no_outliers['tip_amount']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=42)

scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
def data_sampling(taxi_filtered):
    sample = taxi_filtered.sample(9999)
    return sample
model = Sequential()
model.add(Dense(128, input_dim=X_train.shape[1], activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(64, activation='relu'))
model.add(Dense(1))
optimizer = tf.keras.optimizers.Adam(learning_rate=0.001)
model.compile(loss='mean_squared_error', optimizer=optimizer)

with tf.device('/GPU:0'):
    history = model.fit(X_train_scaled, y_train, epochs=50, batch_size=8196,
validation_split=0.2, verbose=1)

loss = model.evaluate(X_test_scaled, y_test)
print(f"Test Loss: {loss}")

predictions = model.predict(X_test_scaled)
predictions_df = pd.DataFrame({'True Values': y_test.values, 'Predictions':
predictions.flatten()})
sample = data_sampling(predictions_df)
fig, axs = plt.subplots(1, 2, figsize=(15, 5))

axs[0].set(xlim=[-3, 20])
axs[1].set(xlim=[-3, 20])

sns.kdeplot(data=sample, ax=axs[0], bw_adjust=3)
sns.kdeplot(data=sample, ax=axs[1], bw_adjust=3, cumulative=True)

fig.tight_layout()
print(taxi_filtered_no_outliers.info())
print(taxi_filtered_no_outliers.describe())
print(taxi_filtered_no_outliers.isnull().sum())
plt.figure(figsize=(8, 6))
sns.histplot(taxi_filtered_no_outliers['tip_amount'], bins=30, kde=True)
plt.xlabel('Tip Amount')
plt.ylabel('Frequency')
plt.title('Distribution of Tip Amount')
plt.show()
plt.figure(figsize=(10, 8))
sns.heatmap(taxi_filtered_no_outliers.corr(), annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Heatmap')
plt.show()
sns.pairplot(taxi_filtered_no_outliers.sample(5000), vars=['trip_distance',
'fare_amount', 'pickup_weekday', 'pickup_hour', 'tip_amount'], diag_kind='kde')
plt.suptitle('Pairplot of Selected Features')
plt.show()
```

```
selected_features_without_tip = [feature for feature in selected_features if feature
!= 'tip_amount']

comparison_df = pd.DataFrame({
    'True Tip Amount': y_test.values,
    'Predicted Tip Amount': predictions.flatten()
})
comparison_df[selected_features_without_tip] =
X_test.reset_index(drop=True)[selected_features_without_tip]

correlation_matrix = comparison_df.corr()

plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Heatmap of Features with True and Predicted Tip Amounts')
plt.show()
visualization_df = pd.DataFrame({
    'Trip Distance': X_test['trip_distance'],
    'Fare Amount': X_test['fare_amount'],
    'True Tip Amount': y_test.values,
    'Predicted Tip Amount': predictions.flatten()
})

plt.figure(figsize=(12, 5))

plt.subplot(1, 2, 1)
plt.scatter(visualization_df['Trip Distance'], visualization_df['True Tip Amount'],
label='True Tip Amount', alpha=0.5)
plt.scatter(visualization_df['Trip Distance'], visualization_df['Predicted Tip
Amount'], label='Predicted Tip Amount', alpha=0.5)
plt.xlabel('Trip Distance')
plt.ylabel('Tip Amount')
plt.title('Trip Distance vs Tip Amount')
plt.legend()

plt.subplot(1, 2, 2)
plt.scatter(visualization_df['Fare Amount'], visualization_df['True Tip Amount'],
label='True Tip Amount', alpha=0.5)
plt.scatter(visualization_df['Fare Amount'], visualization_df['Predicted Tip
Amount'], label='Predicted Tip Amount', alpha=0.5)
plt.xlabel('Fare Amount')
plt.ylabel('Tip Amount')
plt.title('Fare Amount vs Tip Amount')
plt.legend()

plt.tight_layout()
plt.show()
```