Uber: Predict the fare amount of future rides using regression analysis

Analysis by Shubham Soni

Introduction & Business Problem

Uber Technologies, Inc. is an American mobility-as-a-service provider.

The case study aims to analyze the factors influencing Uber's market performance and predict fare amounts using regression analysis.

Uber aims to improve its pricing strategy by accurately predicting ride fares. The fare amount for an Uber ride is determined by multiple factors such as distance traveled, ride duration, traffic conditions, time of day, and demand. To enhance fare predictions and provide better pricing for customers, we need to develop a robust regression model.

The goal is to create a predictive model that uses historical ride data to forecast the fare amount based on the aforementioned factors. This model will leverage a dataset containing historical ride data, including fare amounts and relevant features such as distance, pickup and dropoff locations, timestamps, and passenger counts. By accurately predicting fare amounts, Uber can optimize its pricing strategy, improve customer satisfaction, and increase operational efficiency.

Data Overview

Description of the dataset:

- 200,000 entries with 9 columns
- Data types: float64, int64, object
- Minimal missing values in dropoff_longitude and dropoff_latitude

```
df.head()
```

	Unnamed: 0	key	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude
0	24238194	2015-05-07 19:52:06.0000003	7.5	2015-05-07 19:52:06+00:00	-73.999817	40.738354	-73.999512	40.723217
1	27835199	2009-07-17 20:04:56.0000002	7.7	2009-07-17 20:04:56+00:00	-73.994355	40.728225	-73.994710	40.750325
2	44984355	2009-08-24 21:45:00.00000061	12.9	2009-08-24 21:45:00+00:00	-74.005043	40.740770	-73.962565	40.772647
3	25894730	2009-06-26 08:22:21.0000001	5.3	2009-06-26 08:22:21+00:00	-73.976124	40.790844	-73.965316	40.803349
4	17610152	2014-08-28 17:47:00.000000188	16.0	2014-08-28 17:47:00+00:00	-73.925023	40.744085	-73.973082	40.761247

```
#checking missing %tage
100*df.isnull().mean()
```

```
Unnamed: 0           0.0000
key                  0.0000
fare_amount          0.0000
pickup_datetime      0.0000
pickup_longitude     0.0000
pickup_latitude      0.0000
dropoff_longitude    0.0005
dropoff_latitude     0.0005
passenger_count      0.0000
dtype: float64
```

Data Preprocessing

```python
# Impute missing values with median
df['dropoff_longitude'].fillna(df['dropoff_longitude'].median(), inplace=True)
df['dropoff_latitude'].fillna(df['dropoff_latitude'].median(), inplace=True)

# Confirm no missing values remain
print(df.isnull().sum())
```

```
Unnamed: 0           0
key                  0
fare_amount          0
pickup_datetime      0
pickup_longitude     0
pickup_latitude      0
dropoff_longitude    0
dropoff_latitude     0
passenger_count      0
dtype: int64
```

Steps taken to preprocess the data:

Imputed ‘dropoff_longitude’ & 'dropoff_latitude’ with median Value

Converted pickup_datetime to datetime object

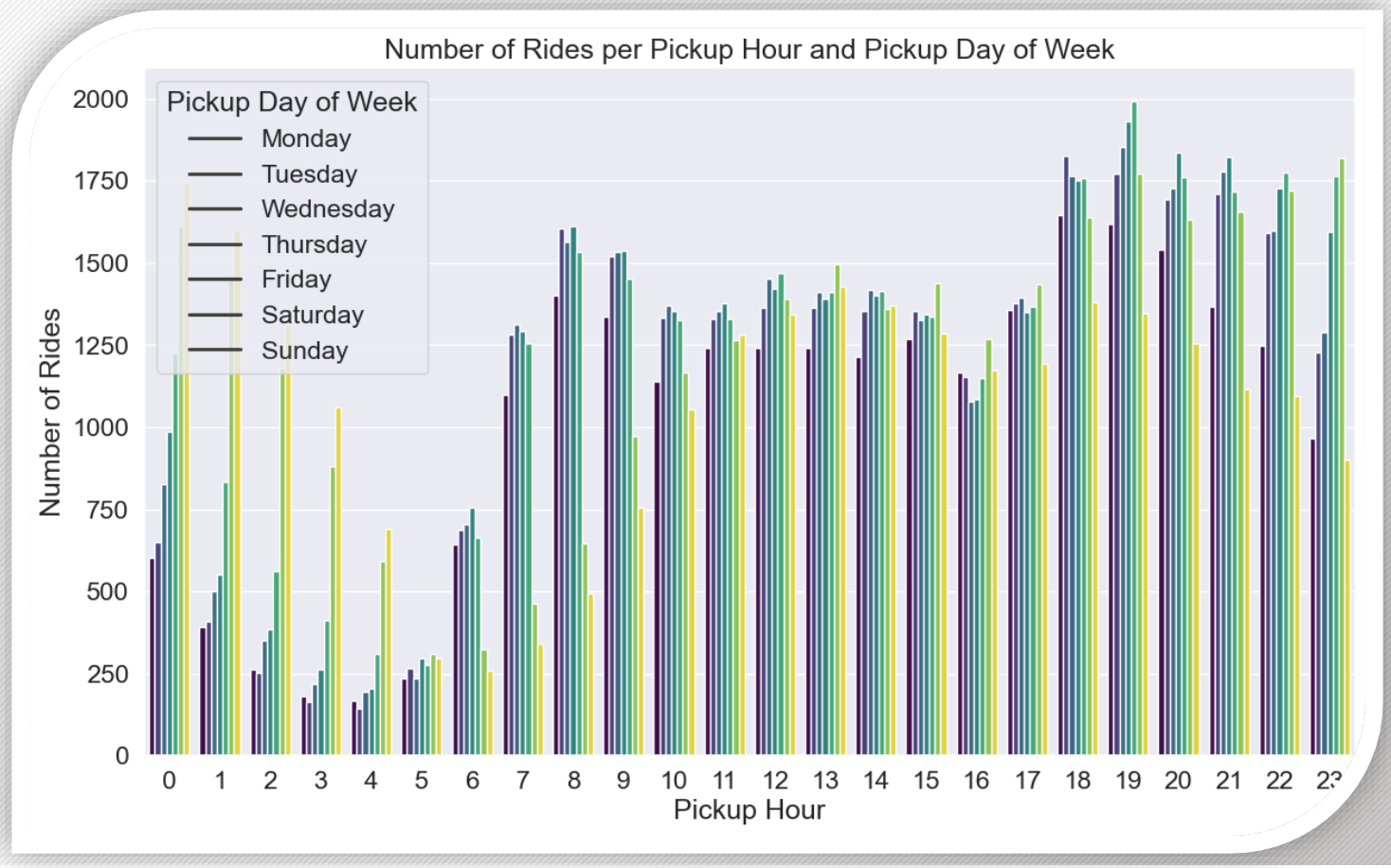
Extracted features: year, month, day, hour, minute, day of the week

	fare_amount	passenger_count	pickup_hour	pickup_day_of_week	pickup_month	trip_distance
0	7.5	1	19	3	5	1.683323
1	7.7	1	20	4	7	2.457590
2	12.9	1	21	0	8	5.036377
3	5.3	3	8	4	6	1.661683
4	16.0	5	17	3	8	4.475450
...	...	...	...	...	...	...
199995	3.0	1	10	6	10	0.112210
199996	7.5	1	1	4	3	1.875050
199997	30.9	2	0	0	6	12.850319
199998	14.5	1	14	2	5	3.539715
199999	14.1	1	4	5	5	5.417783

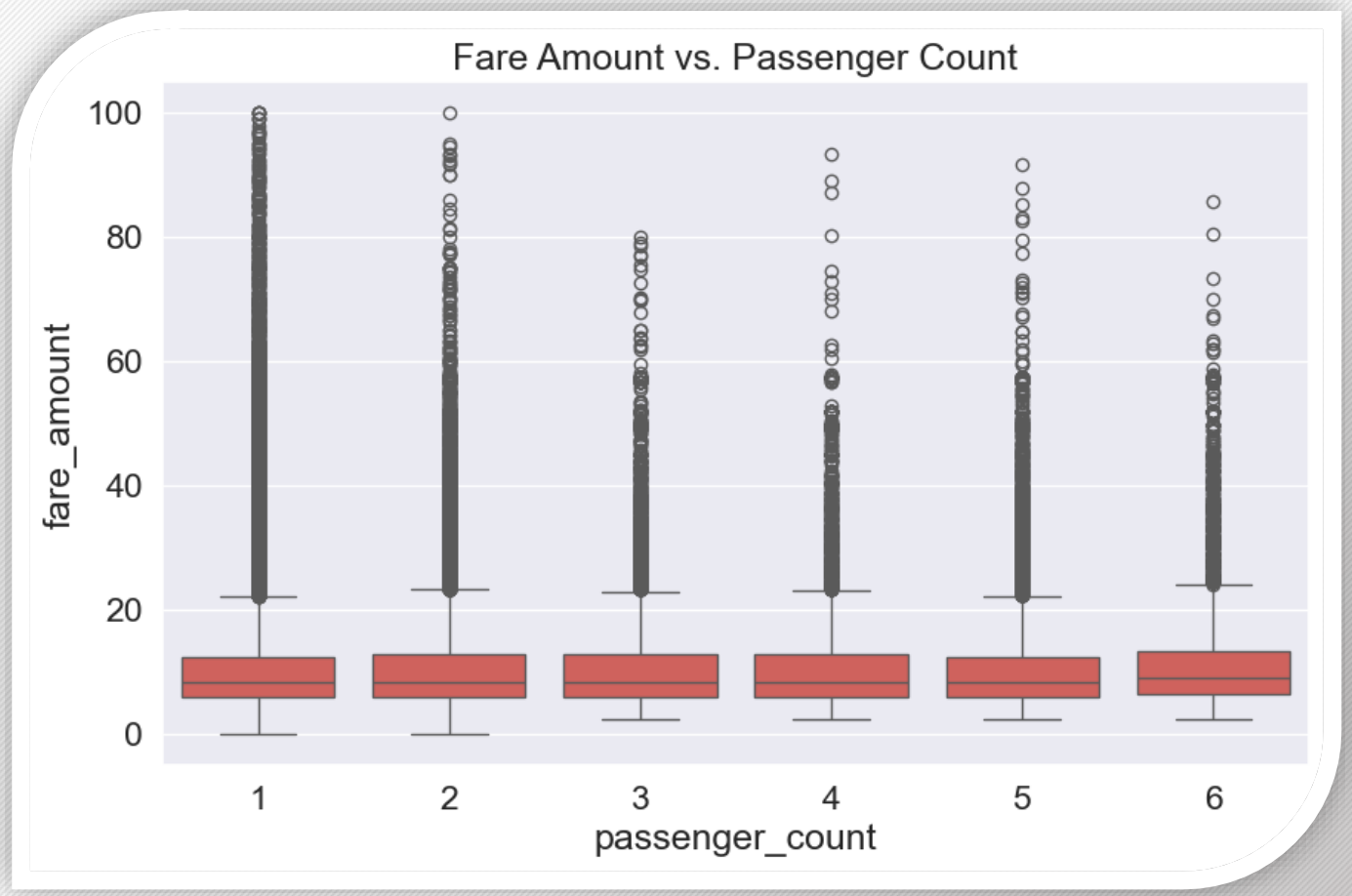
Exploratory Data Analysis

Key findings from EDA:

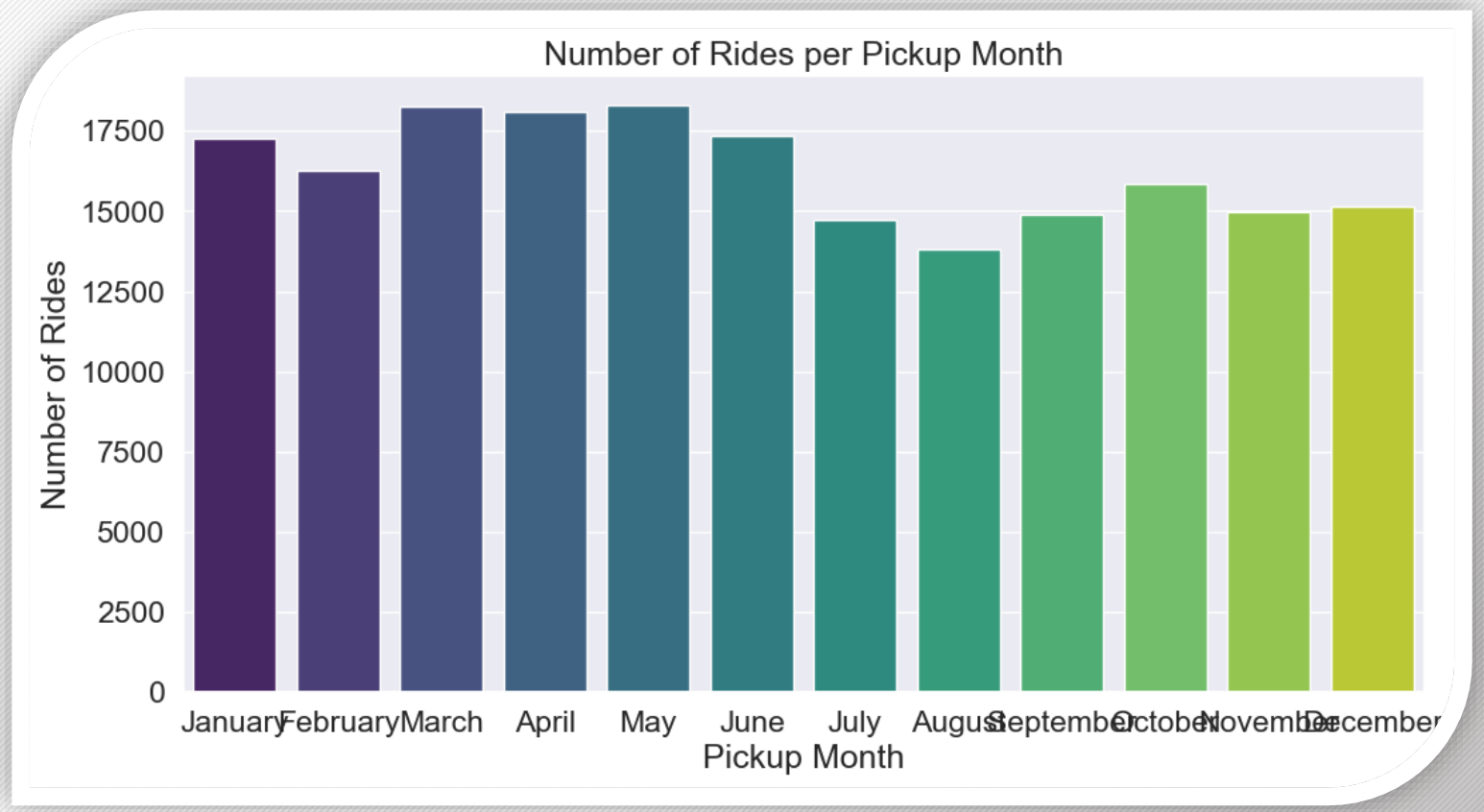
- Distribution of fare amount and passenger count
- Analysis of Number of Rides per Pickup Hour and Pickup Day of Week
- Number of Rides per Pickup Month
- Fare Amount vs. Trip Distance
- Distribution of Rides by Distance Category
- Correlation Matrix



- **Weekend Late-Night Surge Significant Ride Demand**
- **Daytime Booking Spike**
- **Evening Peak (6 PM to 11 PM) High Demand, Except Sundays**
- **Consistent High Demand at 7 PM Daily Peak**

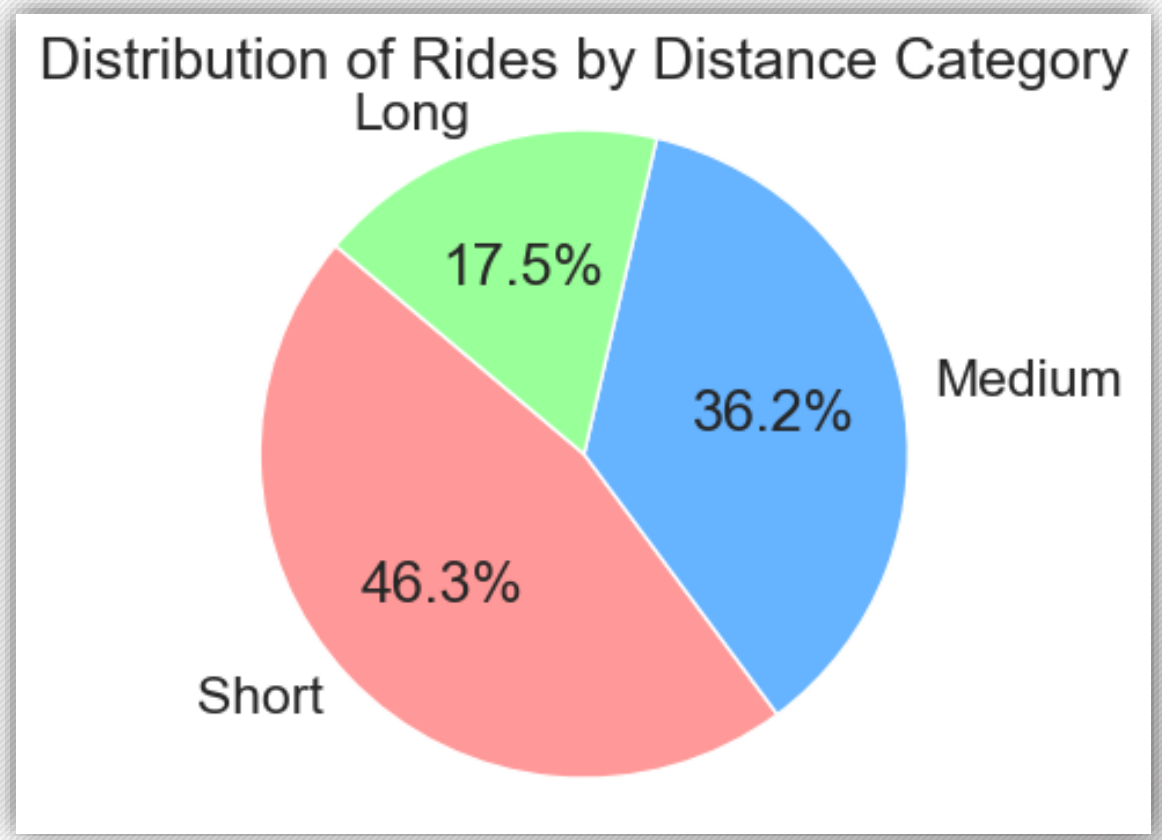


- **Median fare remains consistent**
- **Increased spread with more passengers**
- **Presence of outliers**

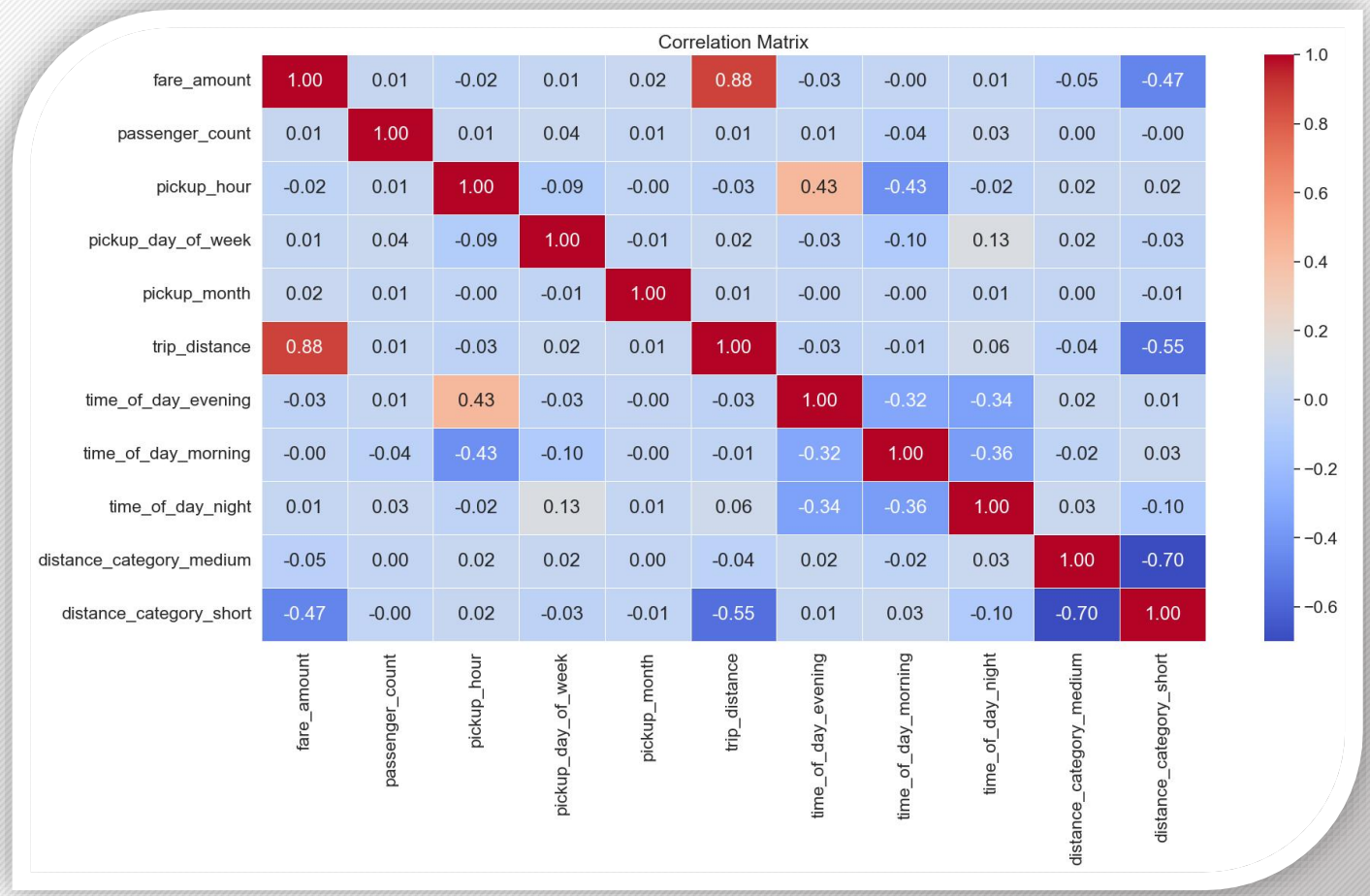


- **Peak Ridership in Early Months**
- **Consistent Ridership Throughout the Year**

- **Positive Correlation Between Fare Amount and Trip Distance**
- **Non-Linear Relationship**
- **Outliers: There are a few data points that deviate significantly from the general trend**

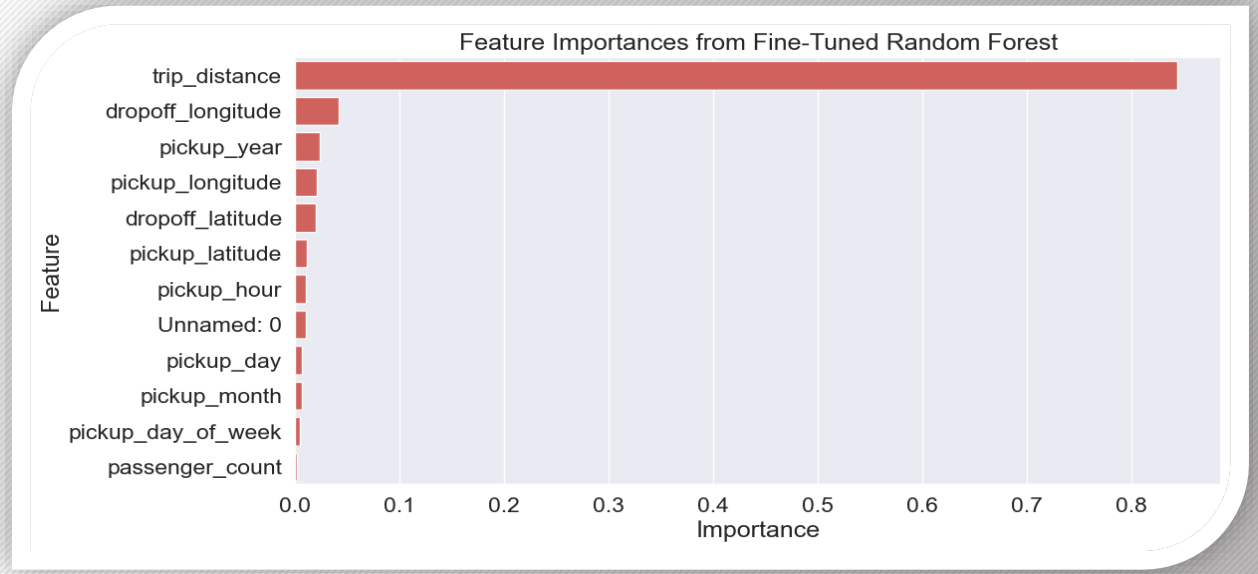


- **Short Rides Dominate**
- **Medium Rides are Less Frequent**
- **Long Rides are Least Common**



- **Fare amount: There is a weak positive correlation between fare amount and trip distance (0.88) which means that longer trips tend to cost more.**
- **Passenger count: There is a weak positive correlation between passenger count and distance category (medium and long). This means that longer trips tend to have more passengers.**

Regression Analysis



Models used and evaluation metrics:

- Linear Regression: MSE: 21.4579, R2: 0.7626
- **Random Forest: MSE: 21.2795, R2: 0.7646**
- Decision tree: MSE: 39.3784, R2: 0.5644

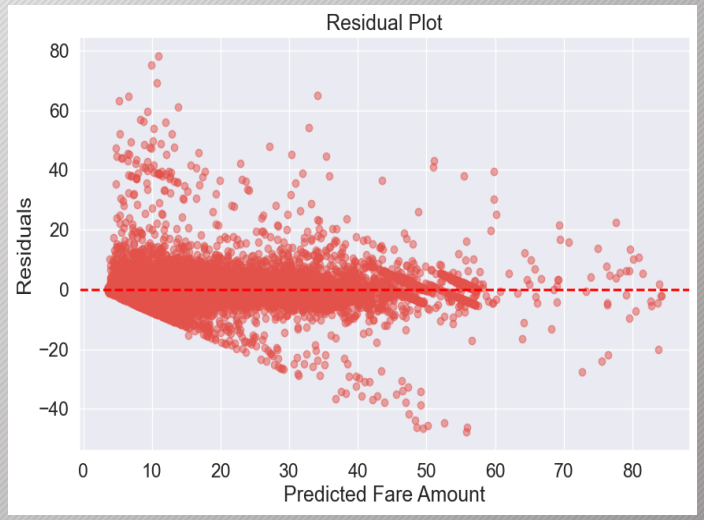
Random Forest tends to be best model as it have low MSE Value and Hight R^2.

- **Evaluated using R^2, MAE, MSE:** Fitting 2 folds for each of 10 candidates, totalling 20 fits Best Hyperparameters: {'n_estimators': 200, 'min_samples_split': 10, 'min_samples_leaf': 2, 'max_depth': None} Fine-Tuned Random Forest - MSE: 13.4969, R2: 0.8507
- **Comparison of model performances**
- **Feature importance from the fine-tuned Random Forest model:** Trip distance is the most significant factor influencing fare amount, with a very high importance score. This aligns with the conventional pricing model where the cost of a ride is directly proportional to the distance traveled. Therefore, accurately measuring and incorporating trip distance is crucial for fare predictions with an importance score of 0.886642. The other features have much lower importance scores, suggesting that they contribute less to the model's predictions.

Residual Analysis

Analysis of residuals:

- Detected heteroscedasticity
- Identified and analyzed outliers
- **Detected Heteroscedasticity**: Variability in fare predictions increases with fare amount.
- **Identified Outliers**: Outliers can skew revenue predictions and financial planning.



Predicted Business Outcomes:

- More accurate fare predictions for higher amounts.
- Improved fare pricing and revenue optimization.
- Enhanced forecasting for ride demand and operational costs.

Factors Influencing the Fare Price

- **Trip Distance:** Highest Influence (Importance Score: 0.843665)Longer trips directly lead to higher fares.Recommendation: Accurately measure and incorporate trip distance in fare calculations.
- **Dropoff Longitude:** Moderate Influence (Importance Score: 0.041446)Geographic location impacts fare due to route complexity and traffic conditions.
- **Pickup Year:** Moderate Influence (Importance Score: 0.023971)Fares may vary slightly year-over-year due to inflation, policy changes, or demand fluctuations.
- **Pickup/Dropoff Latitude & Longitude:** Moderate Influence (Combined Importance Scores: 0.031045)Specific pickup and dropoff locations can affect the fare due to distance and route efficiency.
- **Pickup Hour:** Moderate Influence (Importance Score: 0.010539)Time of day affects fare due to peak and off-peak pricing strategies.
- **Other Factors:** Lower Influence (Combined Importance Scores: 0.018628)Includes passenger count, pickup day, and pickup month. Smaller contributions to fare variations but still relevant.

Recommendation:

- Focus on Trip Distance for fare prediction accuracy.
- Incorporate real-time traffic and geographic data to refine fare calculations.
- Monitor and adjust for temporal variations like year, month, and hour.

Recommendations for Predicting Future Ride Fare Amounts

- **Key Insights from Data Analysis:**

1. **Weekend Late-Night Surge:**
    1. **Observation:** Significant increase in ride requests from 12 AM to 2 AM on weekends.
    2. **Recommendation:** Implement surge pricing during these hours to optimize revenue.
2. **Daytime Booking Spike:**
    1. **Observation:** Moderate rise in ride requests from 8 AM to 5 PM reflecting commuting patterns.
    2. **Recommendation:** Ensure adequate driver availability to meet the demand and reduce wait times.
3. **Evening Peak (6 PM to 11 PM):**
    1. **Observation:** High demand during evening hours on most days, especially at 7 PM.
    2. **Recommendation:** Schedule promotions or discounts to attract more riders and optimize vehicle distribution.
4. **Consistent High Demand at 7 PM:**
    1. **Observation:** Daily peak in ride requests at 7 PM.
    2. **Recommendation:** Prepare for this peak by encouraging drivers to be active during this time to minimize surge pricing and ensure customer satisfaction.
5. **Peak Ridership in Early Months:**
    1. **Observation:** Increased ride requests during initial months of the year, peaking in February and May.
    2. **Recommendation:** Plan marketing campaigns and driver incentives around these peaks to capitalize on increased demand.
6. **Consistent Year-Round Demand:**
    1. **Observation:** Steady high demand throughout the year, despite early months' peak.
    2. **Recommendation:** Maintain a robust driver base and continually monitor demand trends to adjust pricing and availability dynamically.

CONCLUSION

- **Fare Predictions:** The regression model accurately predicts fare amounts with high R^2 value (0.8507) on the testing set.
- **Key Predictor:** Trip distance is the most significant factor influencing fare amount.
- **Potential Overfitting:** Slight performance drop from training to testing indicates potential overfitting.
- To enhance Uber’s ability to predict fare amounts and optimize ride pricing, it is crucial to leverage the insights from data analysis. By addressing peak demand times, refining the prediction model, and continuously updating our strategies based on real-time data, we can ensure a more efficient and profitable operation. These recommendations will help in maintaining customer satisfaction, improving driver experience, and maximizing revenue.

Thank you