Predictive Analytics for NYC Taxi Trip Duration

DS-670: Capstone Bigdata & Data science

Veera Reddy

Sridivya Gorantla

Bhargavi

Karim

Saleem

Nikhil

Introduction

Objective:

- Overview of the NYC taxi industry and its significance
- Importance of accurate trip duration predictions for various stakeholders
- Dataset source: Kaggle competition "NYC Taxi Trip Duration"
- Project goal: Develop a machine learning model to predict taxi trip durations
- Potential impact: Improved service efficiency and customer satisfaction
- Challenges: Complex urban environment, traffic patterns, and external factors

Problem Statement

- **Primary objective**:
  - Predict the duration of taxi trips in New York City
  - Build predictive models to estimate trip durations accurately.
  - Analyze influential factors like weather, trip distance, and traffic patterns
- Key questions to address:

  1. What factors most significantly influence trip duration?

  2. How can we accurately model the relationship between these factors and trip time?

  3. Can we create a robust model that generalizes well to unseen data?

- Evaluation metric:

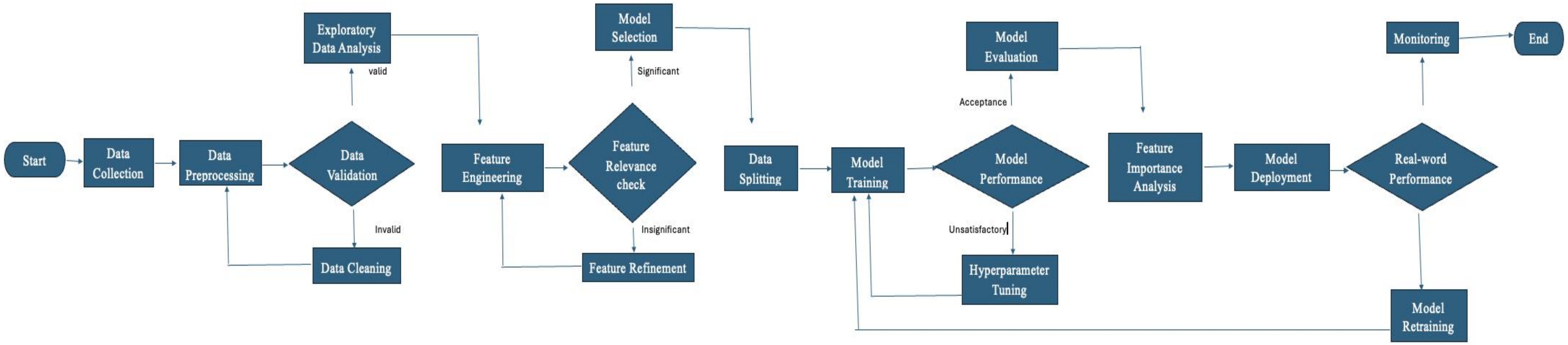
  RMSE, R-squared, MAE
- Additional goals: Gain insights into NYC taxi operations and travel patterns

Dataset Overview

- Source: 2016 NYC Yellow Cab trip record data
- Features: id, vendor_id, pickup_datetime, dropoff_datetime, passenger_count, pickup_longitude, pickup_latitude, dropoff_longitude, dropoff_latitude, store_and_fwd_flag
- Target variable: trip_duration (in seconds)
- Dataset size: 1,458,644 trip records
- Time period covered: January 1 to June 30, 2016
- Geographical scope: New York City's five boroughs

_id	pickup_datetime	dropoff_datetime	passenger_count	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	store_and_fwd_flag	trip_duration
2	2016-03-14 17:24:55	2016-03-14 17:32:30	1	-73.982155	40.767937	-73.964630	40.765602	N	455
1	2016-06-12 00:43:35	2016-06-12 00:54:38	1	-73.980415	40.738564	-73.999481	40.731152	N	663
2	2016-01-19 11:35:24	2016-01-19 12:10:48	1	-73.979027	40.763939	-74.005333	40.710087	N	2124
2	2016-04-06 19:32:31	2016-04-06 19:39:40	1	-74.010040	40.719971	-74.012268	40.706718	N	429
2	2016-03-26 13:30:55	2016-03-26 13:38:10	1	-73.973053	40.793209	-73.972923	40.782520	N	435

Flow Chat



Start
Data Collection
Data Preprocessing
Data Validation
valid
Exploratory Data Analysis
Invalid
Data Cleaning
Feature Engineering
Feature Relevance check
Significant
Model Selection
Insignificant
Feature Refinement
Data Splitting
Model Training
Model Performance
Acceptance
Model Evaluation
Unsatisfactory
Hyperparameter Tuning
Feature Importance Analysis
Model Deployment
Real-word Performance
Monitoring
End
Model Retraining

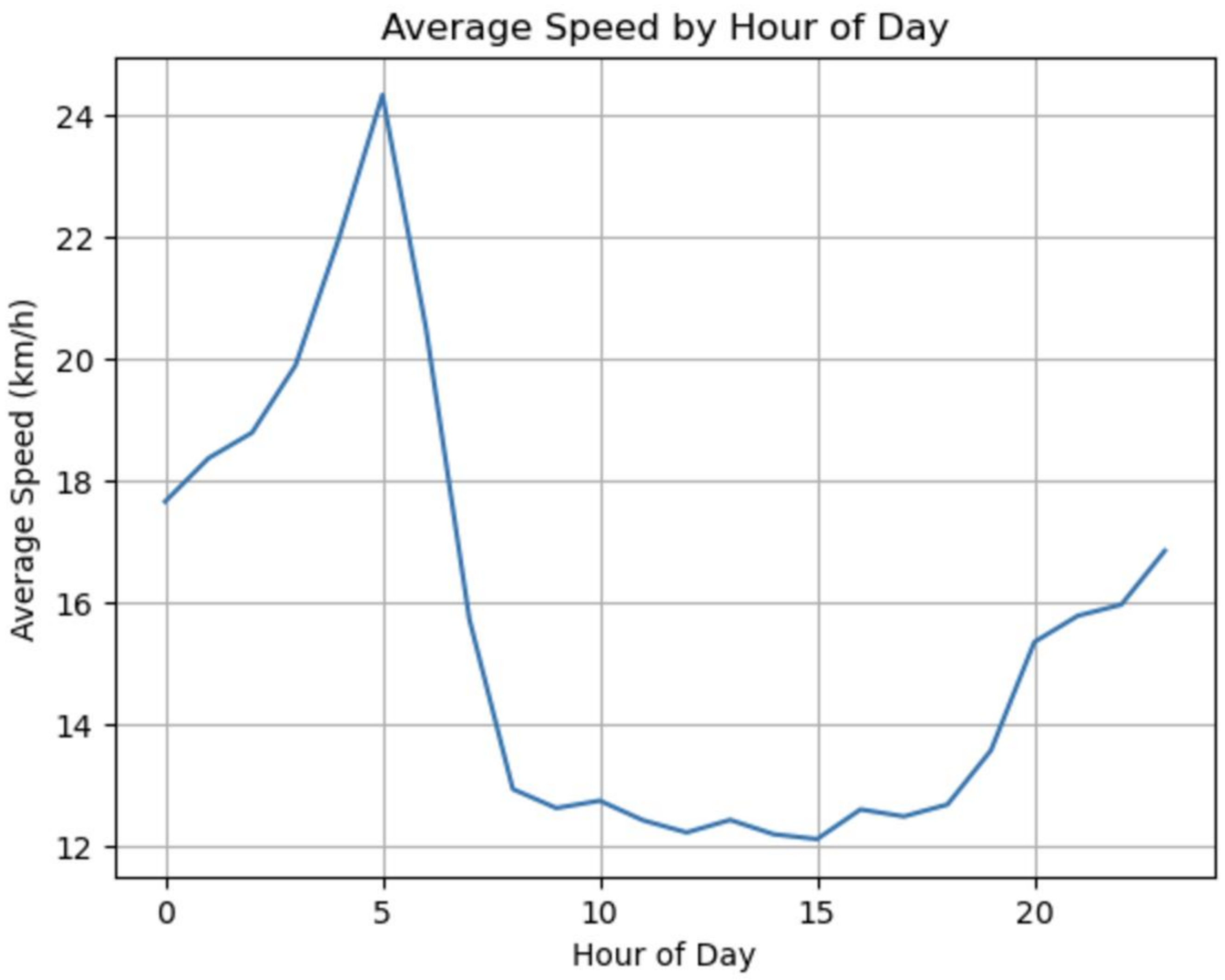
Data Preprocessing

- Handling missing values and outliers
- Feature engineering: Extracting time-based features (hour, day, month, weekday)
- Calculating trip distance using Haversine formula
- Creating trip categories (short, medium, long) based on duration
- Coordinate transformation and normalization
- Encoding categorical variables (e.g., vendor_id, store_and_fwd_flag)
- Splitting data into training and validation sets (80/20 ratio)

dropoff_datetime	passenger_count	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	store_and_fwd_flag	trip_duration	trip_distance
2016-03-14 17:32:30	1	-73.982155	40.767937	-73.964630	40.765602	N	455	1.498523
2016-06-12 00:54:38	1	-73.980415	40.738564	-73.999481	40.731152	N	663	1.805510
2016-01-19 12:10:48	1	-73.979027	40.763939	-74.005333	40.710087	N	2124	6.385107
2016-04-06 19:39:40	1	-74.010040	40.719971	-74.012268	40.706718	N	429	1.485500
2016-03-26 13:38:10	1	-73.973053	40.793209	-73.972923	40.782520	N	435	1.188590

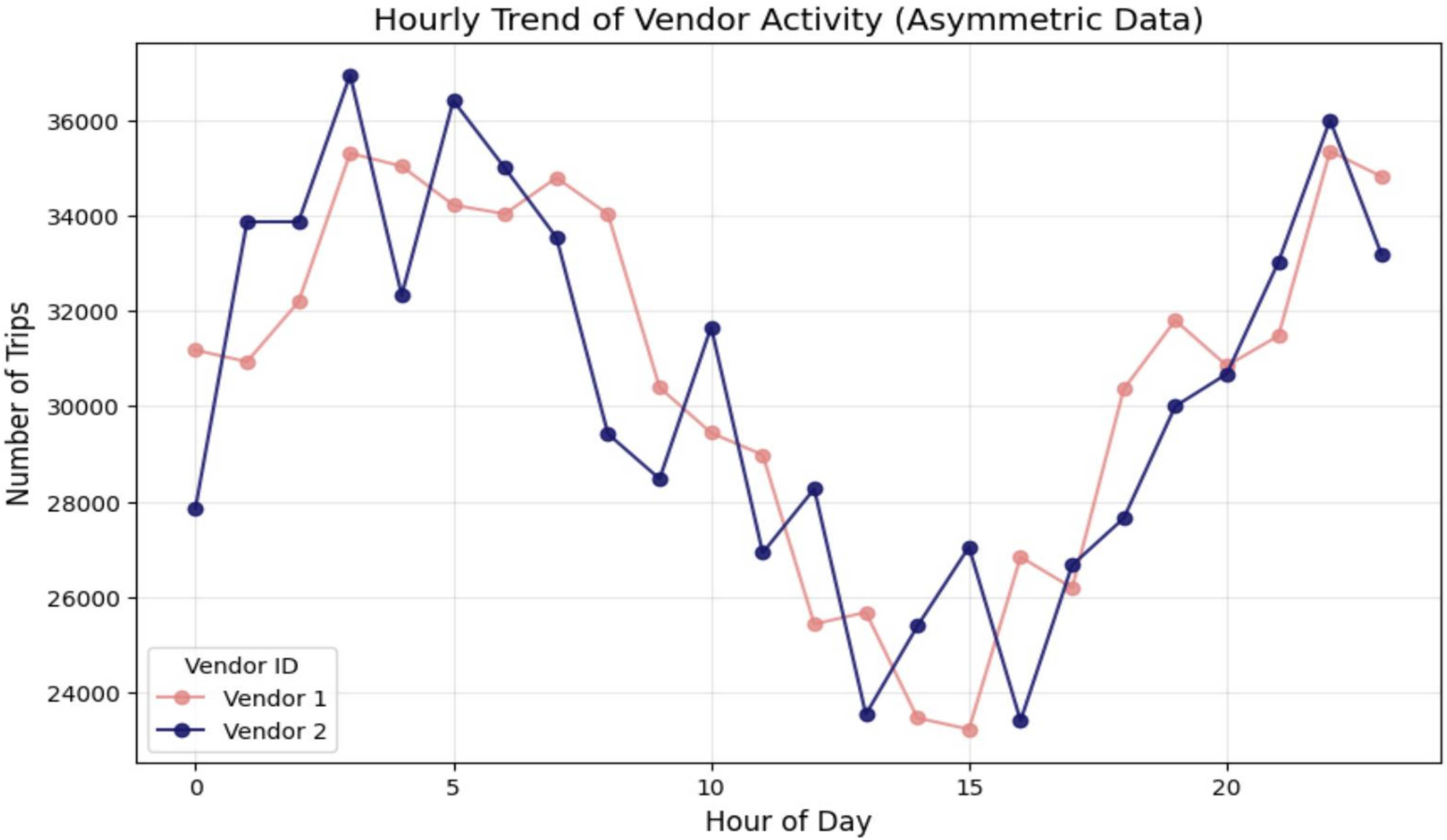
Exploratory Data Analysis (EDA)

- Distribution of trip durations and distances
- Temporal patterns: hourly, daily, and monthly trends
- Geographical analysis: popular pickup/dropoff locations
- Correlation analysis between features and trip duration
- Passenger count impact on trip duration
- Vendor comparison and performance analysis

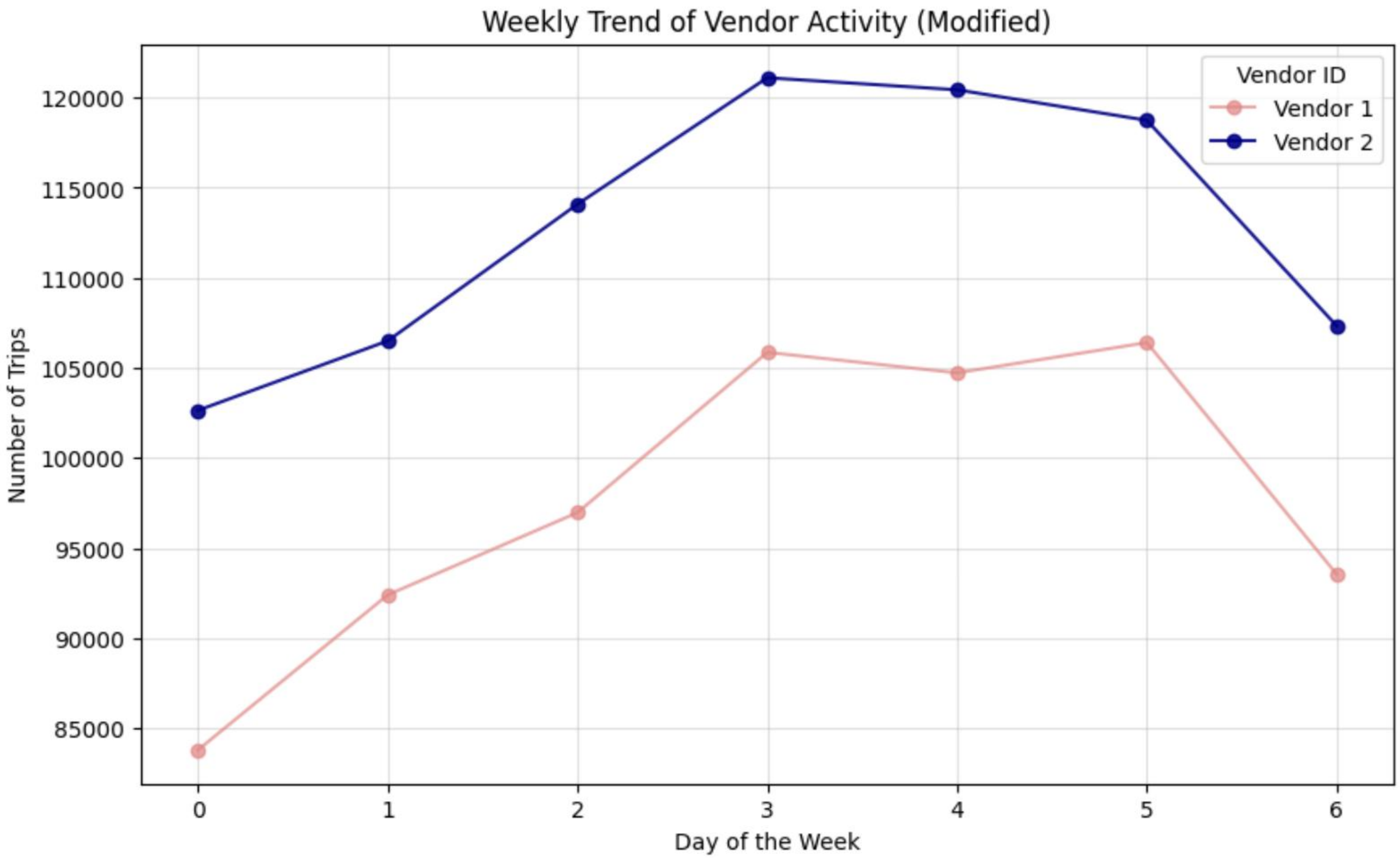


Average Speed by Hour of Day
24
22
20
18
16
14
12
Average Speed (km/h)
0
5
10
15
20
Hour of Day

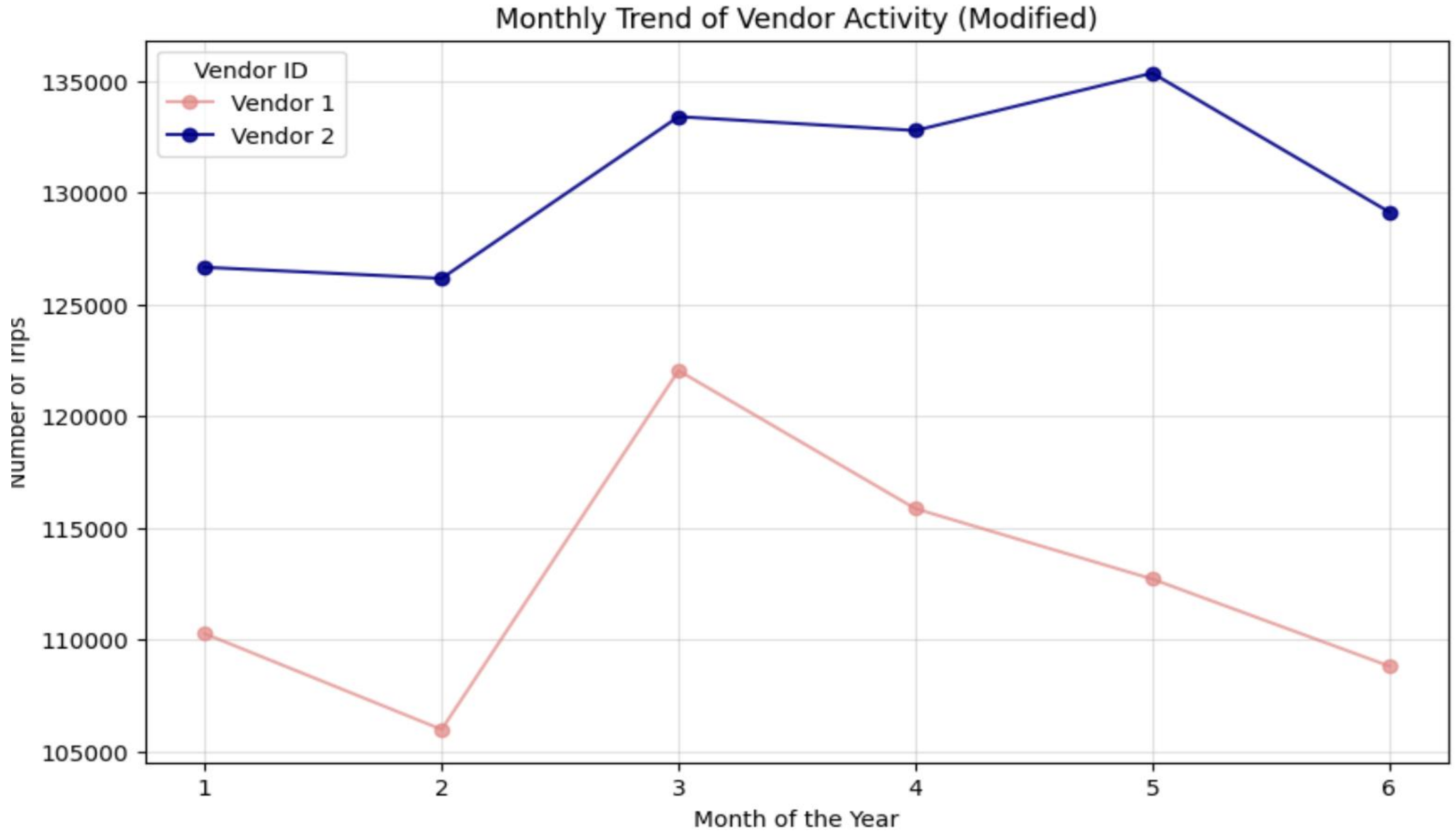
Hourly Trend of Vendor Activity



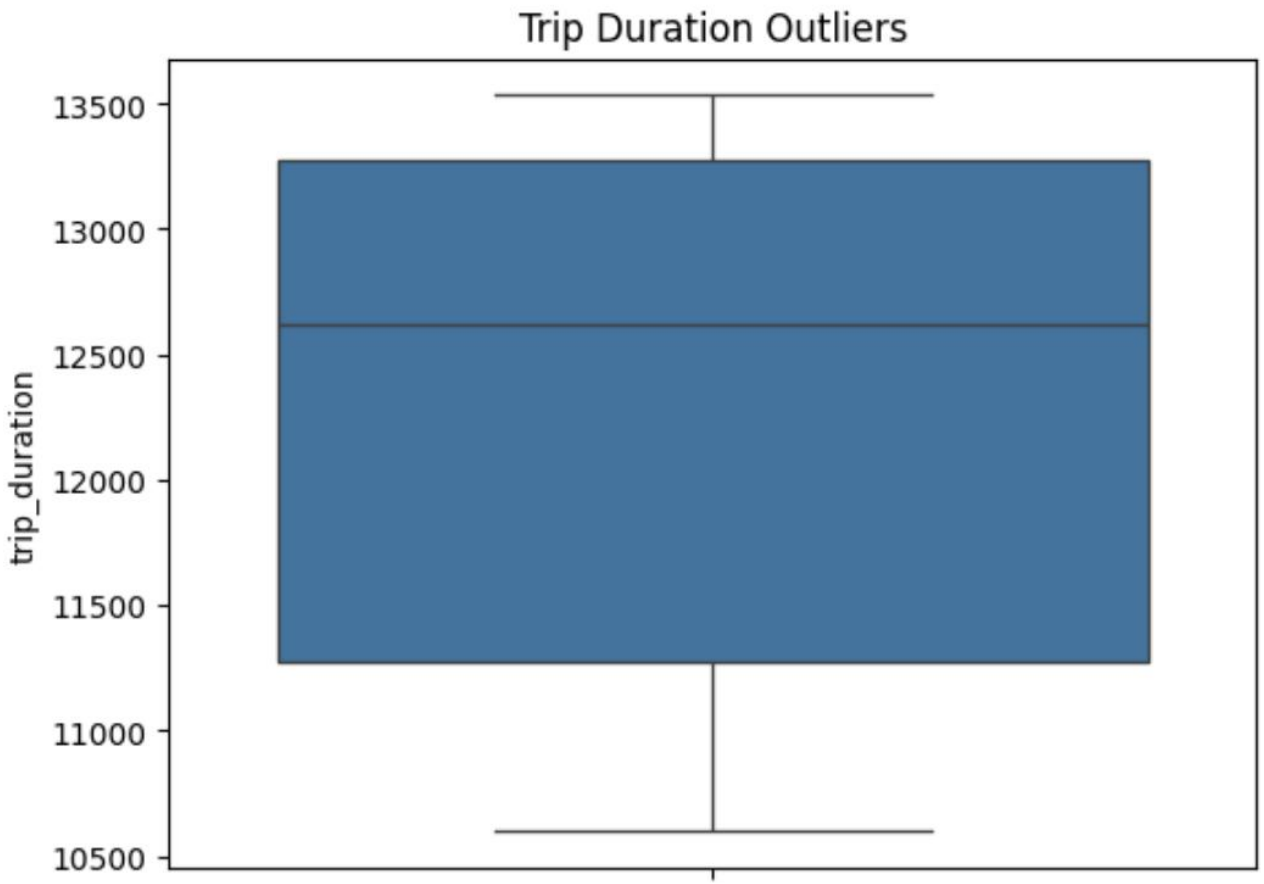
Weekly Trend of Vendor Activity

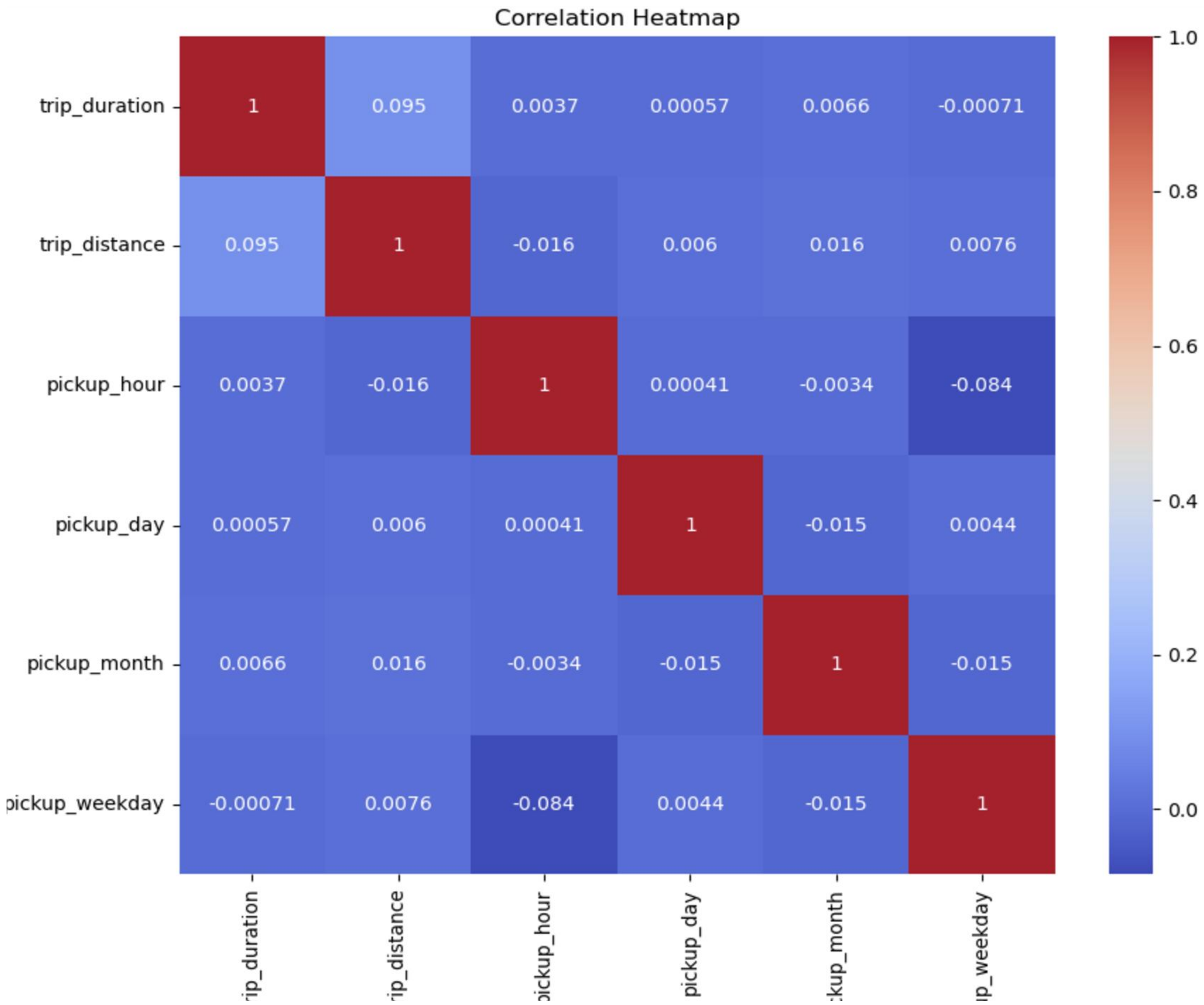


Monthly Trend of Vendor Activity



Trip Duration Outliers





Correlation Heatmap
trip_duration
trip_distance
pickup_hour
pickup_day
pickup_month
pickup_weekday
1
0.095
0.0037
0.00057
0.0066
-0.00071
0.095
1
-0.016
0.006
0.016
0.0076
0.0037
-0.016
1
0.00041
-0.0034
-0.084
0.00057
0.006
0.00041
1
-0.015
0.0044
0.0066
0.016
-0.0034
-0.015
1
-0.015
-0.00071
0.0076
-0.084
0.0044
-0.015
1
1.0
0.8
0.6
0.4
0.2
0.0

Feature Selection and Engineering

- Importance of relevant features for prediction
- Techniques used:
  - Correlation analysis with target variable
  - Feature importance from tree-based models (Random Forest)
  - Domain knowledge and intuition
- Selected features and rationale behind each
- Additional engineered features:
  - Time-based features (rush hour, weekend/weekday)
  - Geographical features (borough, neighborhood)

Model Selection

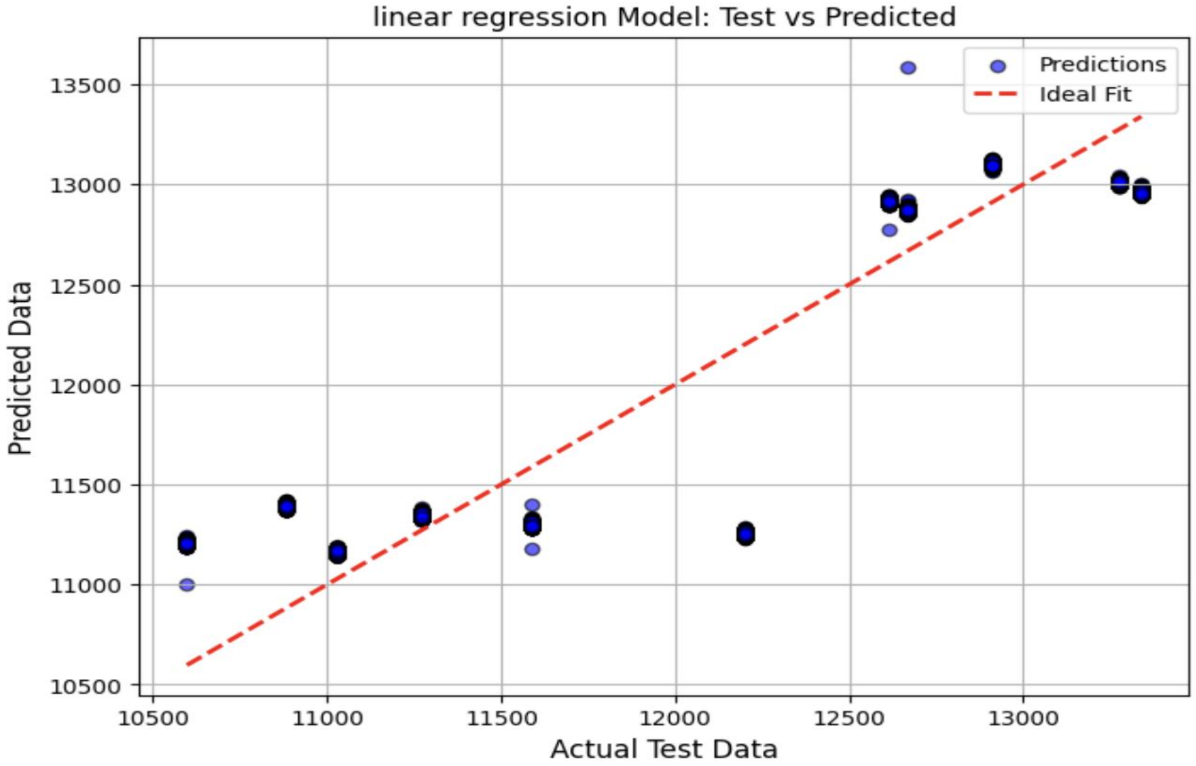
- Comparison of various machine learning algorithms:
  - Linear Regression (baseline model)
  - Random Forest Regressor
  - Gradient Boosting (XGBoost, LightGBM)
  - Neural Networks (if applicable)
- Evaluation metrics:
  - RMSE, R-squared, MAE
- Cross-validation strategy: Time-based splitting to prevent data leakage
- Hyperparameter tuning using GridSearchCV or RandomizedSearchCV

Model Training and Optimization

- Training process for each model type
- Hyperparameter optimization results
- Learning curves analysis
- Feature importance analysis from tree-based models
- Ensemble methods: Stacking top-performing models

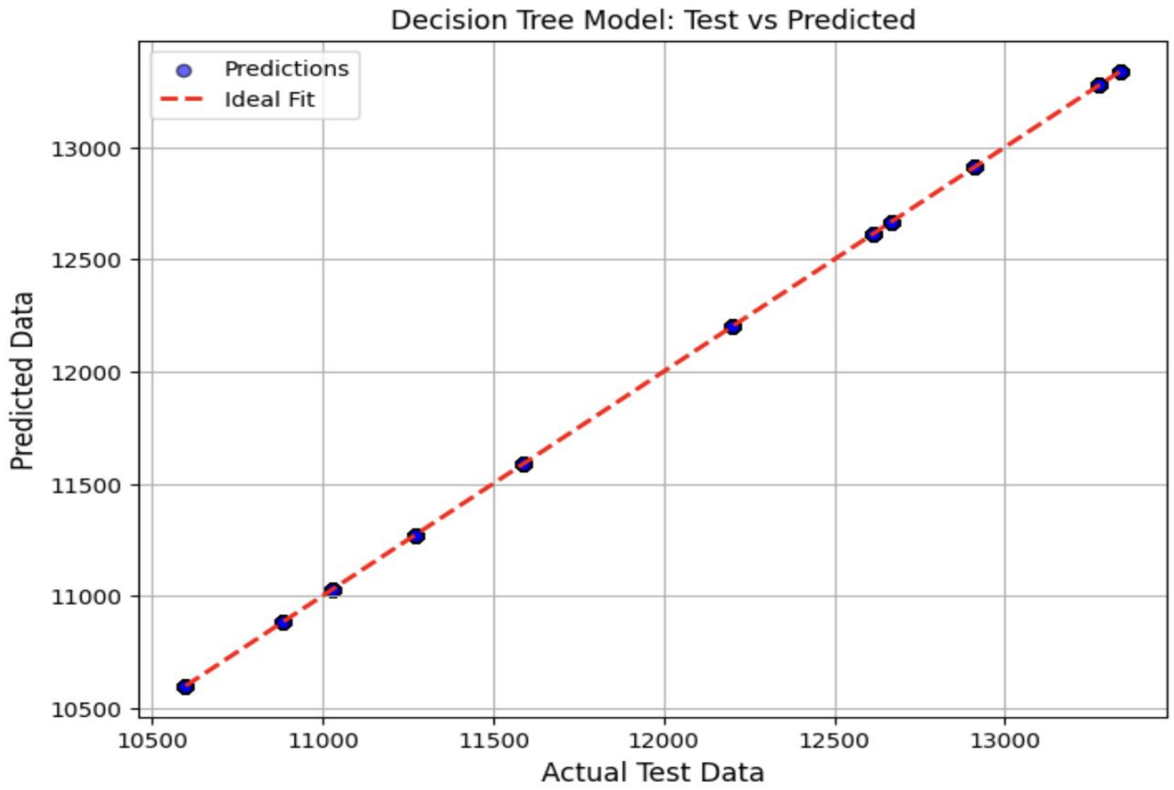
Linear Regression

```
Mean Absolute Error (MAE): 352.4231
Mean Absolute Percentage Error (MAPE): 2.95%
Root Mean Squared Error (RMSE): 424.5169958502619
R-squared (R²): 0.8004161005121733
```



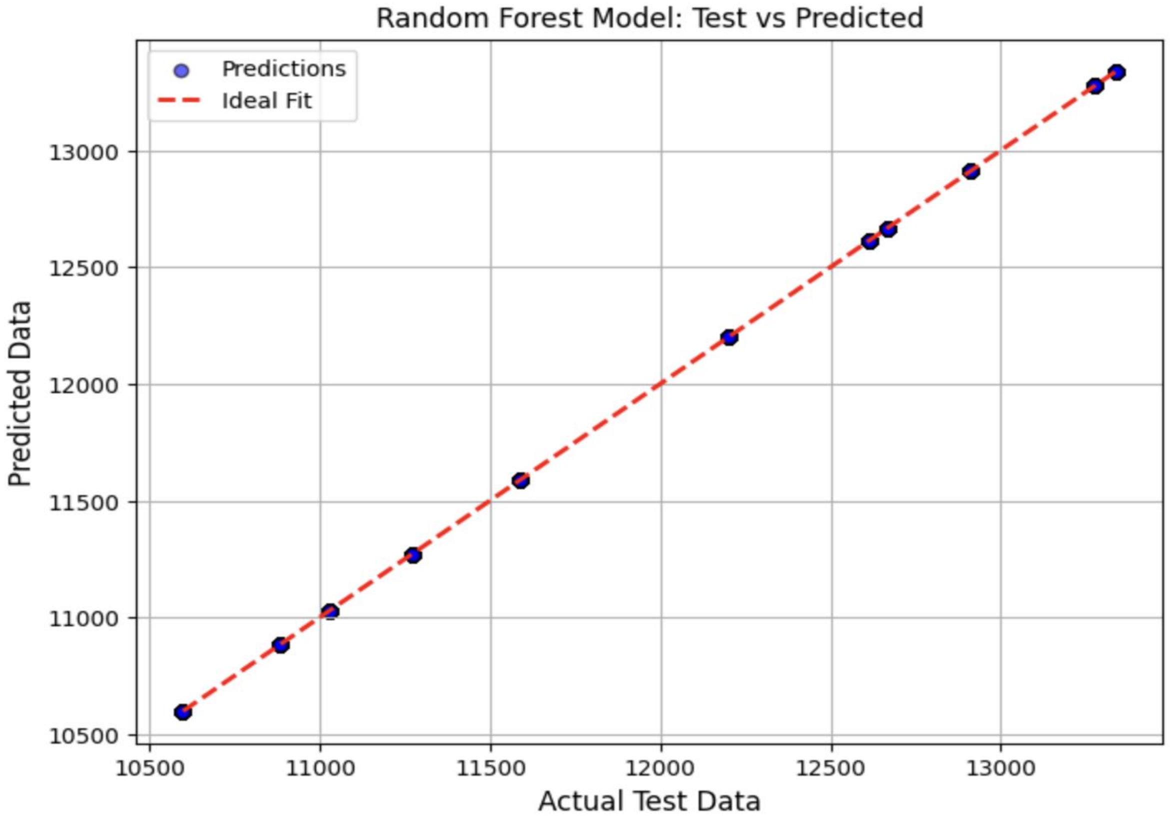
Decision Tree

Mean Absolute Error (MAE): 0.0000
Mean Absolute Percentage Error (MAPE): 0.00%
Root Mean Squared Error (RMSE): 0.0
R-squared (R^2): 1.0



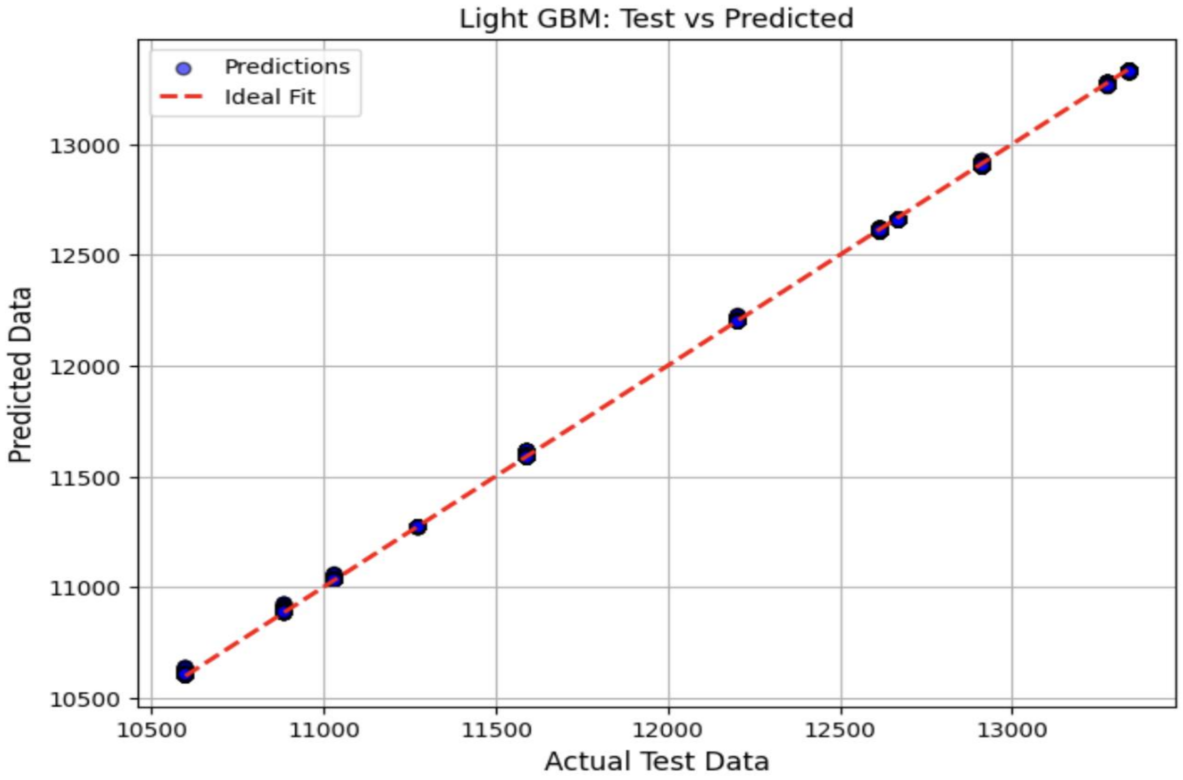
Random Forest

```
Mean Absolute Error (MAE): 0.0000
Mean Absolute Percentage Error (MAPE): 0.00%
Root Mean Squared Error (RMSE): 0.0
R-squared (R²): 1.0
```



LightGBM

Mean Absolute Error (MAE): 7.1243
Mean Absolute Percentage Error (MAPE): 0.06%
Root Mean Squared Error (RMSE): 8.02819167353052
R-squared (R^2): 0.9999286209432355



Results and Model Evaluation

- Performance comparison of different models
- Best model selection based on RMSE and other metrics
- Analysis of prediction errors and residuals
- Model interpretability: SHAP values or partial dependence plots

Future Work and Improvements

- Incorporating additional data sources (weather, events, traffic)
- Exploring advanced techniques (e.g., LSTM networks for time series)
- Developing a real-time prediction system
- Extending the model to other cities or transportation modes
- Potential for a production-ready application or API

Thank You