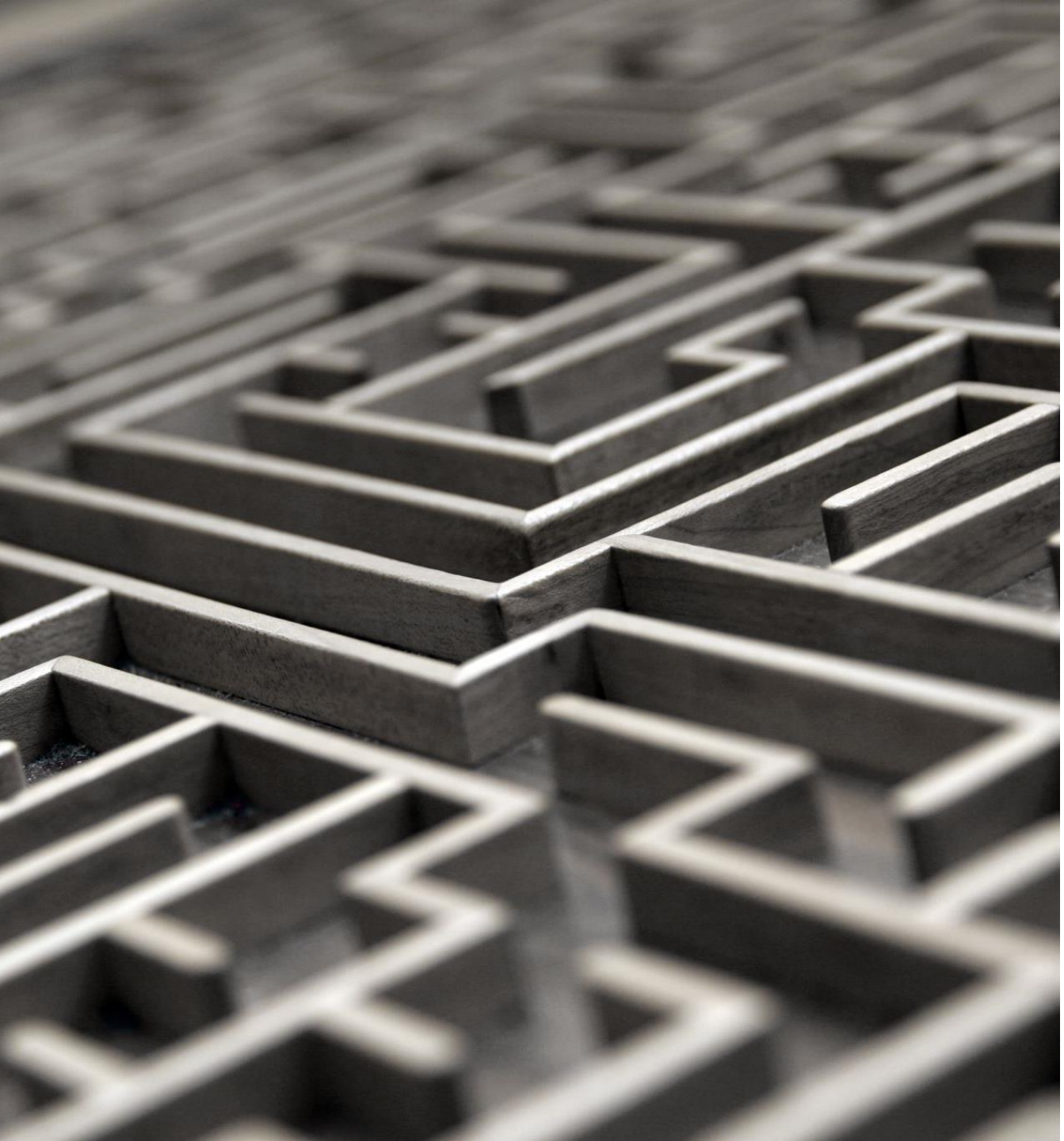
Uber Data Analytics

By Dev Singhmar

21IT3017

Mentored by Dr. Susham Biswas



Problem Statement

Predicting Fare and Tip Amounts for Uber Trips

Background: Uber, a popular ride-hailing service, aims to provide reliable fare estimates to both passengers and drivers. However, accurately predicting fare and tip amounts for each trip remains a challenge due to various factors such as trip distance, number of passengers, tolls, and additional charges.

Objective: The objective of this project is to develop robust predictive models to estimate fare amounts and tip amounts for Uber trips based on trip-related features.

Approach

Data Preprocessing: Initial data cleaning involves removing redundant columns and handling missing or duplicate entries. Feature engineering may be employed to extract relevant information or create new features that could enhance model performance.

Model Selection: Support Vector Regression (SVR) is chosen as the predictive modeling technique due to its ability to handle non-linear relationships and high-dimensional data. SVR models will be trained separately for predicting fare and tip amounts.

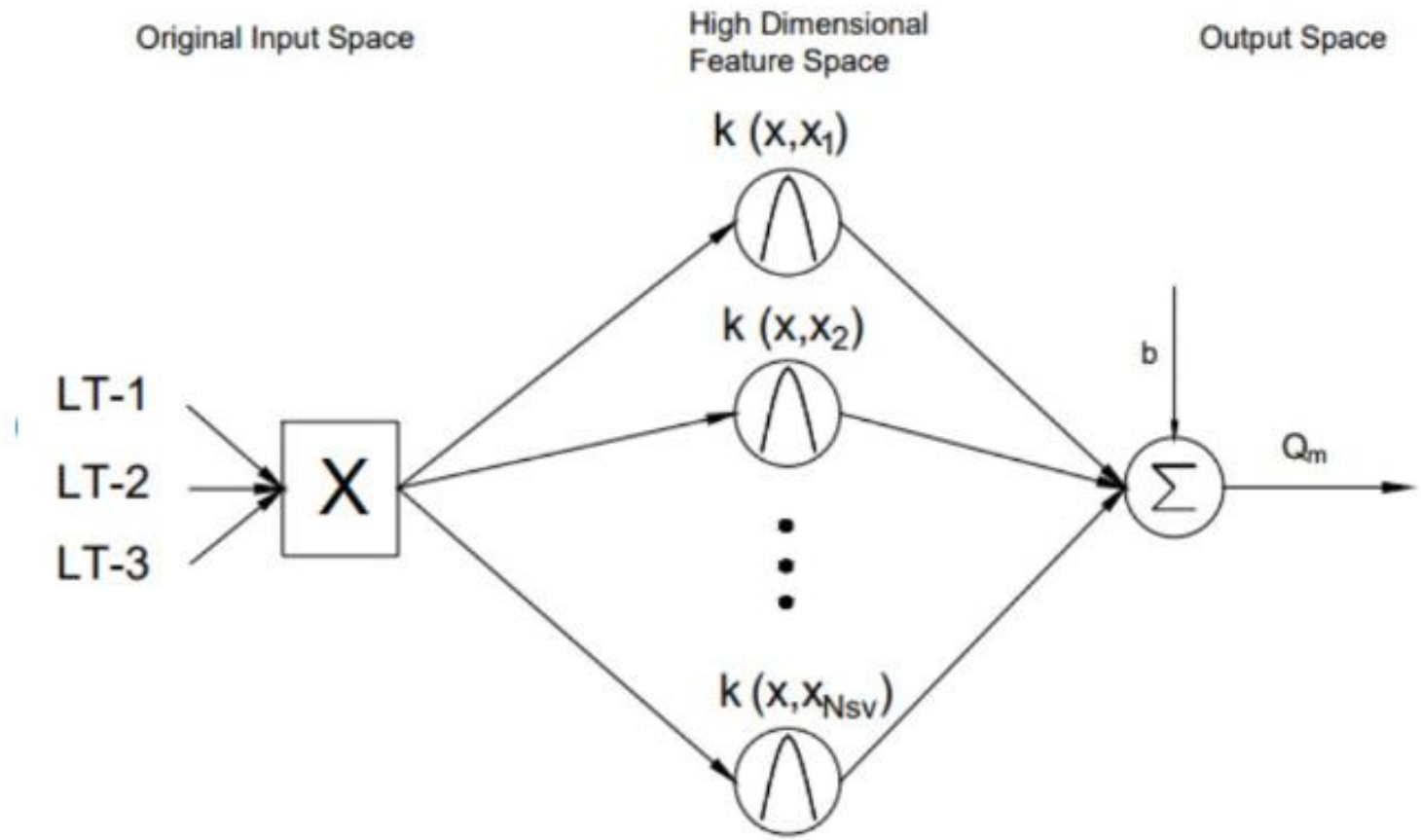
Feature Scaling: Features are standardized using StandardScaler to ensure that all features contribute equally to the model and to improve convergence during training.

Model Training: The dataset is split into training and testing sets. SVR models are trained on the training data, optimizing hyperparameters such as the kernel function and regularization parameters to minimize prediction errors.

Model Evaluation: The performance of the trained models is evaluated using Root Mean Squared Error (RMSE) as the primary evaluation metric. Additionally, other metrics such as Mean Absolute Error (MAE) and R-squared may be considered to assess model performance comprehensively.

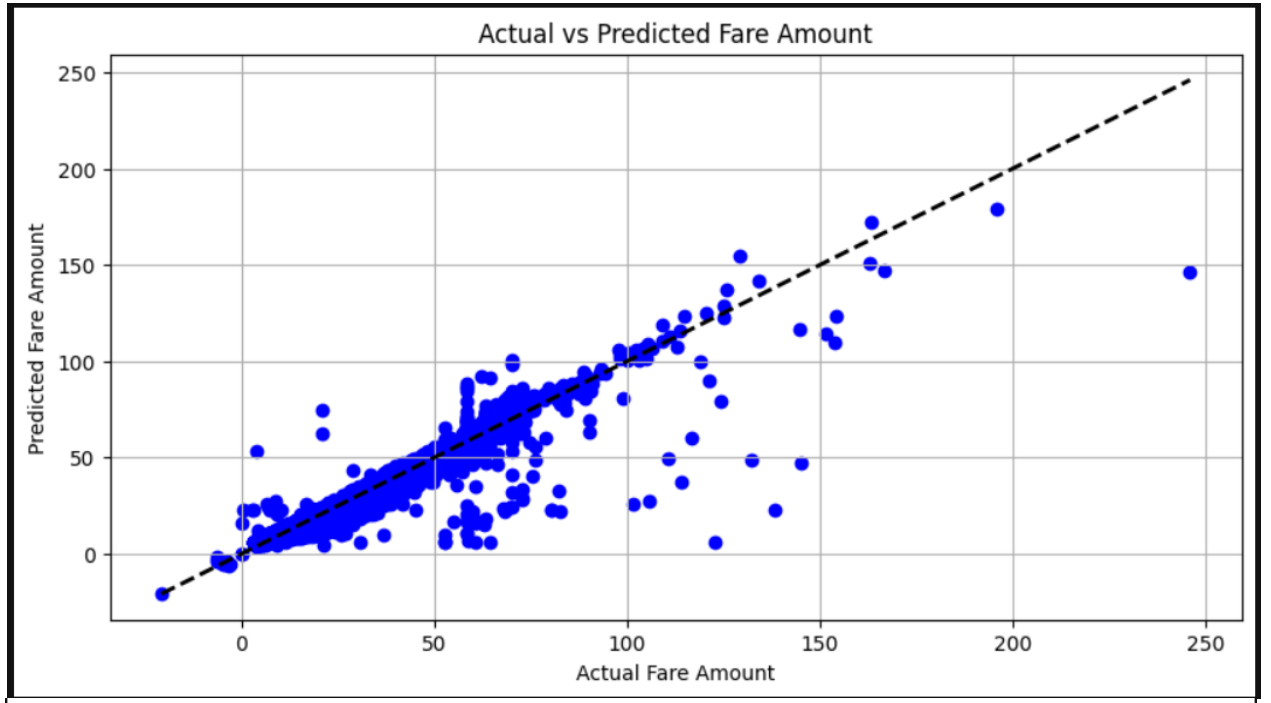
Visualization: Scatter plots are generated to visualize the relationship between actual and predicted fare amounts, as well as actual and predicted tip amounts, providing insights into model accuracy and potential areas for improvement.

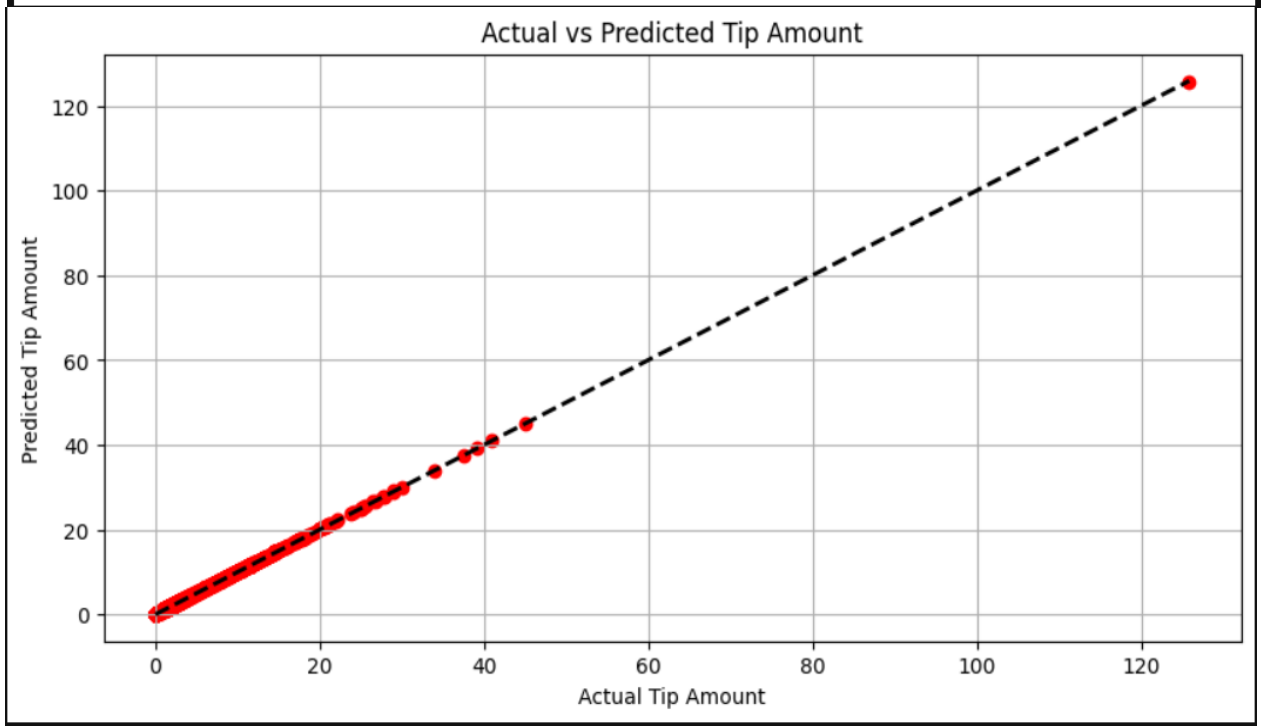
Architecture



Outcome

```
Fare amount RMSE: 4.6371134008545285
Tip amount RMSE: 0.09436176044830155
```





Scope of Improvement for Higher Accuracy

Feature Engineering:

Explore additional features that may impact fare and tip amounts, such as time-related features (e.g., time of day, day of the week), weather conditions, or special events in the area.

Consider interactions between existing features or transformations to capture non-linear relationships.

Model Selection and Hyperparameter Tuning:

Experiment with different regression algorithms beyond SVR, such as Random Forest Regression or Gradient Boosting Regression, which may capture complex patterns better.

Perform thorough hyperparameter tuning for SVR and other algorithms using techniques like grid search or randomized search to find the optimal parameters.

Ensemble Learning:

Implement ensemble methods like Random Forest Regression or Gradient Boosting Regression to combine the predictions of multiple models, potentially improving accuracy through model averaging or boosting.

Data Preprocessing:

Explore alternative methods for feature scaling or normalization to improve the convergence of SVR models.

Consider handling outliers or skewed distributions in the data using robust scaling techniques or transformations.

Cross-Validation and Model Evaluation:

Utilize more advanced cross-validation techniques such as nested cross-validation to obtain more reliable estimates of model performance.

Evaluate models using additional metrics beyond RMSE, such as Mean Absolute Error (MAE) or R-squared, to gain a comprehensive understanding of prediction accuracy.

Domain-specific Insights:

Collaborate with domain experts or stakeholders in the ride-sharing industry to gain insights into factors influencing fare amounts and tipping behavior, which can inform feature selection and model interpretation.

References

- TLC Trip Record Data (Uber Dataset used officially provided by NYC government)

  TLC Trip Record Data - TLC (nyc.gov)

- Dictionary - This data dictionary describes yellow taxi trip data. For a dictionary describing green taxi data.

  data_dictionary_trip_records_yellow.pdf (nyc.gov)