Food Delivery Time Prediction

Ananya Garg, Mohmmad Ayaan,

Swara Parekh, Vikranth Udandarao
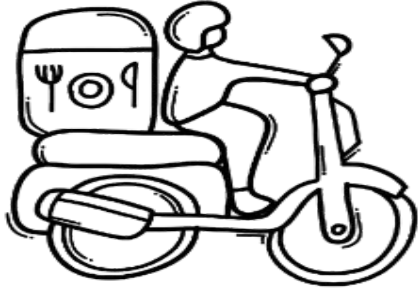
Group 10

ML-end sem project

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Motivation

Running a food delivery service requires ensuring timely and quality deliveries despite challenges like traffic and bad weather.

To tackle this, we are developing a Food Delivery Time Prediction System using machine learning. Our goal is to accurately predict delivery times by analyzing historical delivery data, current traffic conditions, and real-time weather trends.

Literature Review

1. **DergiPark - Comparative Analysis of ML models**
    The research paper titled "<u>A Comparative Analysis of Machine Learning Models for Time Prediction in Food Delivery Operations</u>" highlights that ensemble models like Gradient Boosting outperform others in predicting delivery times, using metrics such as MAE and RMSE. Key factors include traffic volume and order quantity.

2. **DergiPark - Application of Random Forest Algorithm**
    The study uses features like time of day, restaurant info, and traffic conditions to predict food delivery time, with Random Forest (500 trees) achieving ~95% accuracy. Cross-validation, Bootstrap methods, and metrics like confusion matrices confirm the model's high accuracy and reliability.

Dataset Description

- **Initial Rows:** 41,368 (after cleaning and preprocessing)
- **Preprocessing Steps:**
  - Removed **"conditions"** prefix from *Weather Conditions*.
  - Standardized columns into appropriate formats: strings, integers, and floats.
  - Converted *Order Date* to **datetime** format.
  - Extracted time from *Time Ordered* and *Time Order Picked*.
  - Dropped rows with **null values** for consistency.

Key Features Impacting Delivery Times Using Random Forest Analysis

- Traffic Density
- Multiple Deliveries
- Delivery Person's Rating
- Delivery Person's Age
- Festival Season

Dataset Description

The dataset used is publicly available on Kaggle.
The columns of the dataset are:

ID

Delivery_person_ID

Delivery_person_Age

Delivery_person_Ratings

Restaurant_latitude

Restaurant_longitude

Delivery_location_latitude

Delivery_location_longitude

Festival,City

Time_taken(min)

Order_Date

Time_Orderd

Time_Order_picked

Weatherconditions

Road_traffic_density

Vehicle_condition

Type_of_order

Type_of_vehicle

Multiple_deliveries

Dataset Description – EDA

Correlation Heatmap of Numeric Features
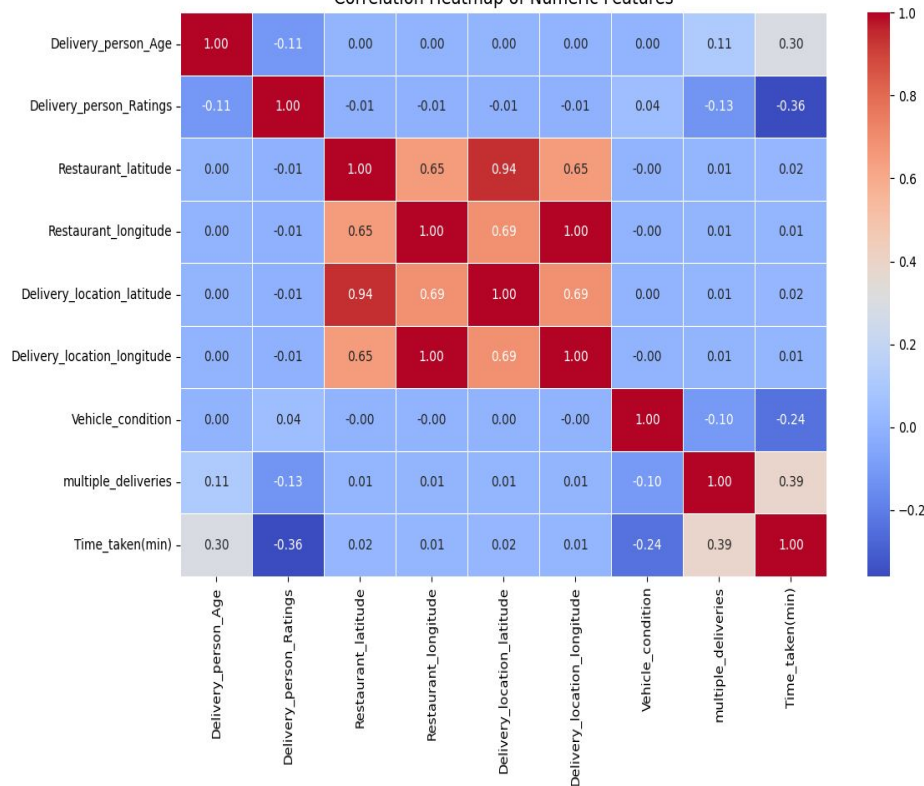
- The **heatmap** visualizes relationships between various **numeric features**.
  - **Red**: Positive Correlations
 - **Blue**: Negative Correlations

Key Observations:

- **Delivery Location Latitude** & **Restaurant Latitude**
  - Strong **positive correlation**: **0.94**
  - Indicates that delivery and restaurant locations tend to align closely in latitude.
- **Delivery Person Ratings** & **Time Taken (minutes)**
  - Moderate **negative correlation**: **-0.36**
  - Suggests that **higher ratings** are associated with **shorter delivery times**.

Dataset Description – EDA


UMAP Projection of Numeric Features
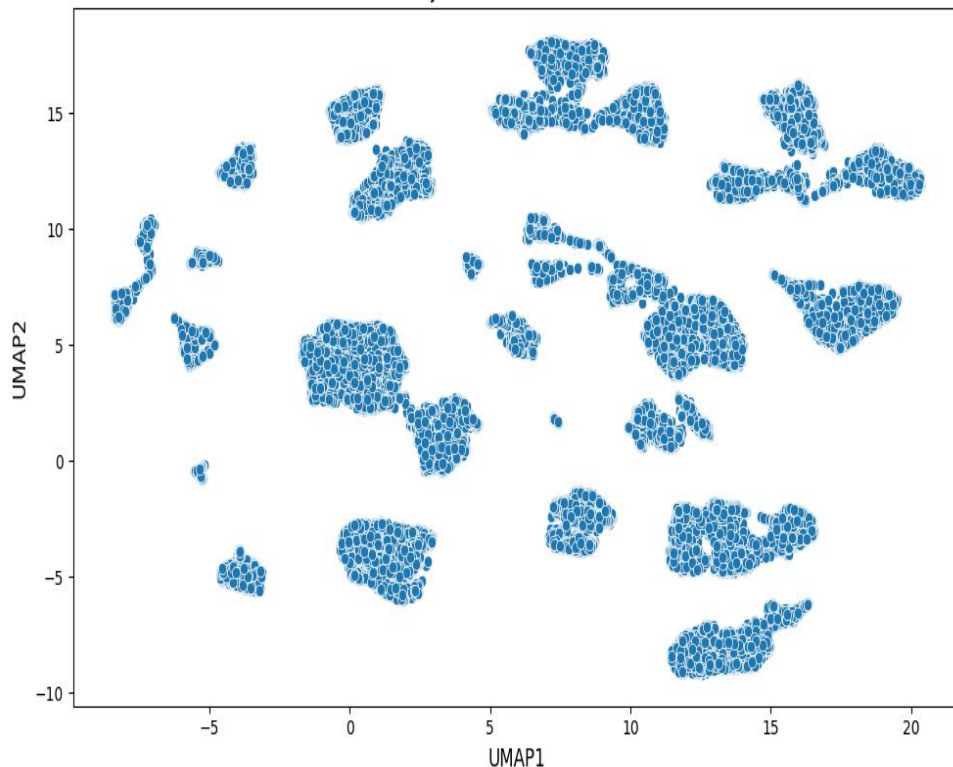
UMAP (Uniform Manifold Approximation and Projection) visualizes the projection of numeric features in the dataset into a **2D space**.

- Each point represents a **data sample**.

Key Observations:

- **Clusters**:
  - Distinct clusters indicate that deliveries with **similar characteristics** (e.g., traffic density, delivery person ratings, time taken) tend to group together.
  - These clusters may represent deliveries with **comparable time predictions** or **delivery conditions**.

Dataset Description – EDA
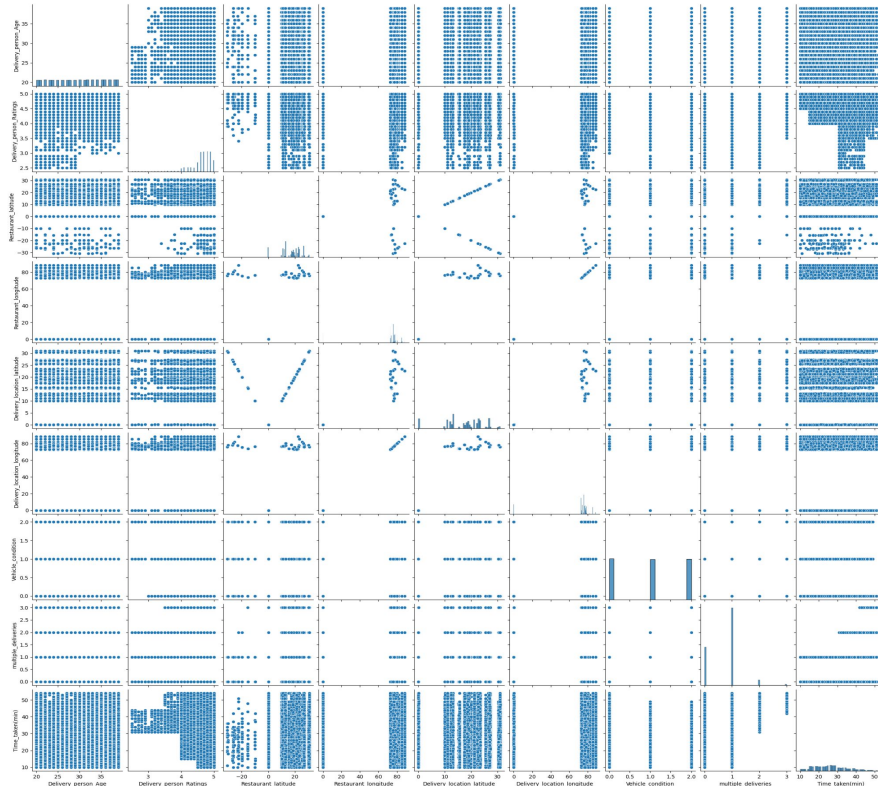
- **Delivery Location Latitude** and **Restaurant Latitude** exhibit a clear **linear trend**.
  - This suggests a **strong geographical alignment** between delivery and restaurant locations, indicating that deliveries typically occur within close proximity to the restaurant's latitude.

Data Preprocessing – Feature Selection

Performed feature selection using **SelectKBest** with **F-regression**, which ranked features by their **linear correlation** with the target. Numerical features were used directly, while categorical features were **one-hot encoded**, and their scores aggregated to reflect group importance.

The k best features achieved are:

Road_traffic_density

Festival

multiple_deliveries

Delivery_person_Ratings

Delivery_person_Age

City

Weatherconditions

Vehicle_condition

Type_of_vehicle

Methodology

Approach:

- **Data Preprocessing**:
  - Handled missing values and standardized the data to ensure consistency and comparability.
  - Performed feature selection using statistical feature selection.
- **Modeling**:
  - Utilized **Ensemble Learning** techniques:
    - **Bagging**, **Random Forest**, **XGBoost**, **LightGBM**, **K-Cross Validation**, and **SVM**.
  - Chosen for their ability to handle **structured data** and capture **nonlinear relationships** effectively.
- **Train-Test Split**:
  - Data split into **80% training** and **20% test** sets to evaluate model performance.
- **Evaluation Metrics**:
  - **Mean Squared Error (RMSE)** and **R2 Score** were used to assess model performance.

Results And Analysis

R² Scores of Models

Mean Squared Error (MSE) of Models
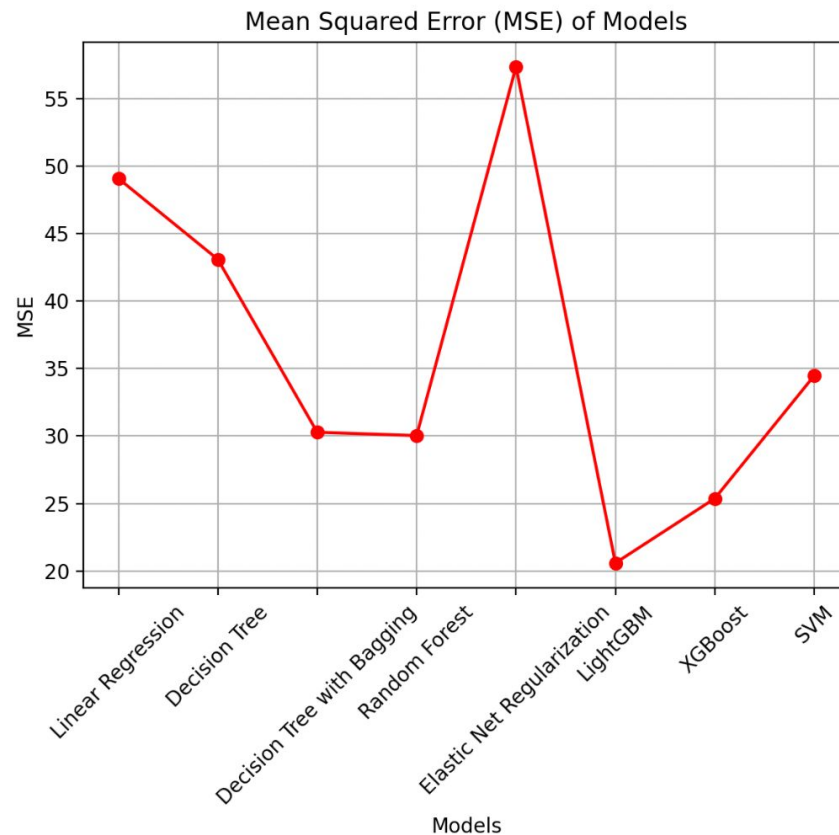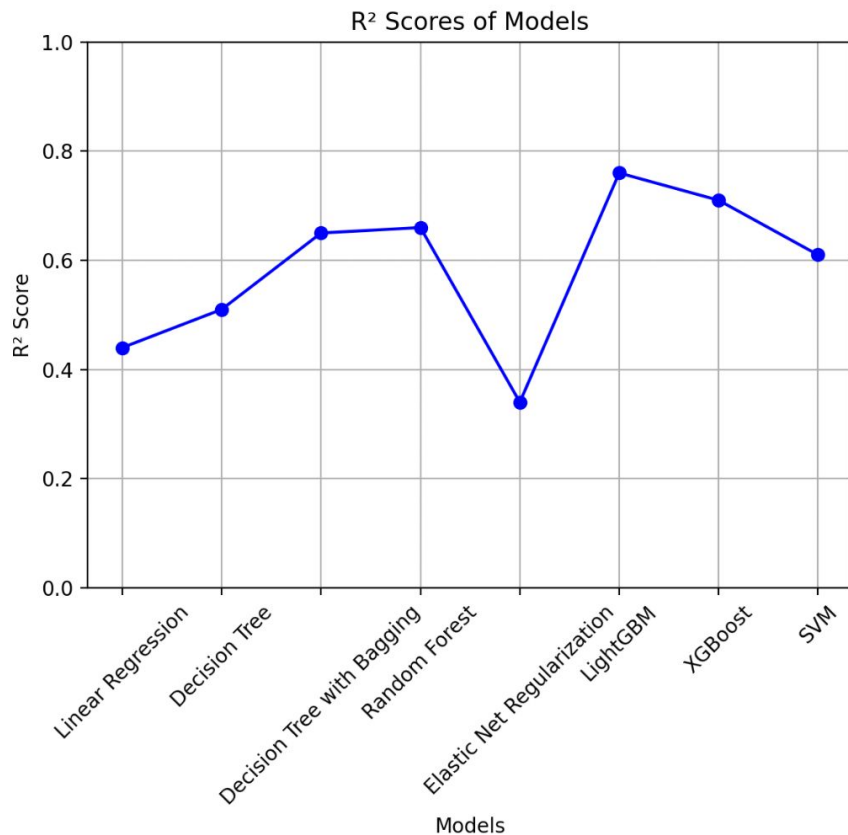
Results, Analysis And Conclusion

R² Scores:

- **LightGBM** and **XGBoost** models have the highest R² scores, indicating better predictive performance with a strong ability to explain the variance in the data.
- **Elastic Net Regularization** shows the lowest R² score, suggesting it struggled to capture the relationships between features and delivery time.

MSE (Mean Squared Error):

- The **LightGBM** and **XGBoost** models have the lowest MSE, meaning they produced the most accurate predictions with minimal error.
- **Elastic Net Regularization** exhibits the highest MSE, indicating it performed poorly compared to the other models.

Conclusion:

- **LightGBM** and **Random Forest** are the top-performing models in terms of both R² and MSE, making them ideal choices for predicting food delivery times in this dataset.
- **Elastic Net Regularization** consistently underperforms across both metrics, showing its limitations for this task.

Timeline

Completed:

- **Data Cleaning & Preprocessing**: Handled missing values, standardized data.
- **Data Analysis**: Performed exploratory data analysis and pattern visualization.
- **Modeling & Evaluation**: Applied ML models and assessed performance using appropriate metrics.
- **Model Refinement:** Enhanced model performance by testing various feature combinations and using k means feature selection.
- **Reporting:** Drafted and finalized the project report, summarizing model performance, key findings, and detailed conclusions.
- **Presentation Preparation:** Created visualizations and completed the final project presentation.

Individual team members' contributions

- **Vikranth Udandarao:** Literature review, Data Collection, EDA, visualization of dataset, model analysis and results.

- **Swara Parekh:** Literature review, Data Collection, EDA, visualization of dataset, model analysis and results.

- **Mohmmad Ayaan:** Literature review, Data Collection, EDA, visualization of dataset, model analysis and results.

- **Ananya Garg:** Literature review, Data Collection, EDA, visualization of dataset, model analysis and results.

References

[1] Comparative Analysis of ML Models for Food Delivery Prediction. Available at: https:4//dergipark.org.tr/en/pub/aita/issue/84471/1459560

[2] Application of Random Forest Algorithm in Food Delivery Time Prediction. Available at: https://dergipark.org.tr/en/pub/forecasting/issue/60291/842180

[3] Gaurav Malik. Food Delivery Dataset. Available at: https://www.kaggle.com/datasets/gauravmalik26/food-delivery-dataset/data

[4] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, pp. 2825-2830, 2011. Available at: https://scikit-learn.org

[5] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T. Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems 30 (NIPS 2017). Available at: https://lightgbm.readthedocs.io

[6] Chen, T. and Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. Available at: https://xgboost.readthedocs.io

Thank You