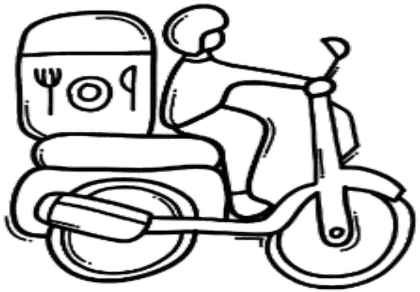
Food Delivery Time Prediction

Ananya Garg, Mohmmad Ayaan,
Swara Parekh, Vikranth Udandarao
Group 10
ML-mid 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 "A Comparative Analysis of Machine Learning Models for Time Prediction in Food Delivery Operations" 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

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

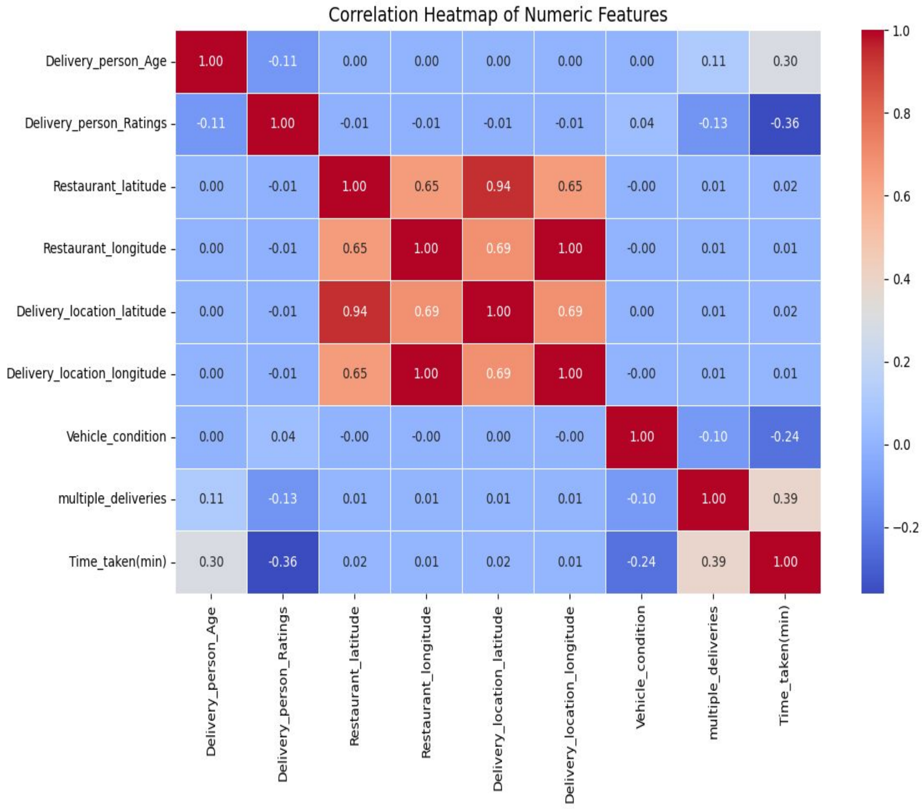
Dataset Description

- **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 – EDA

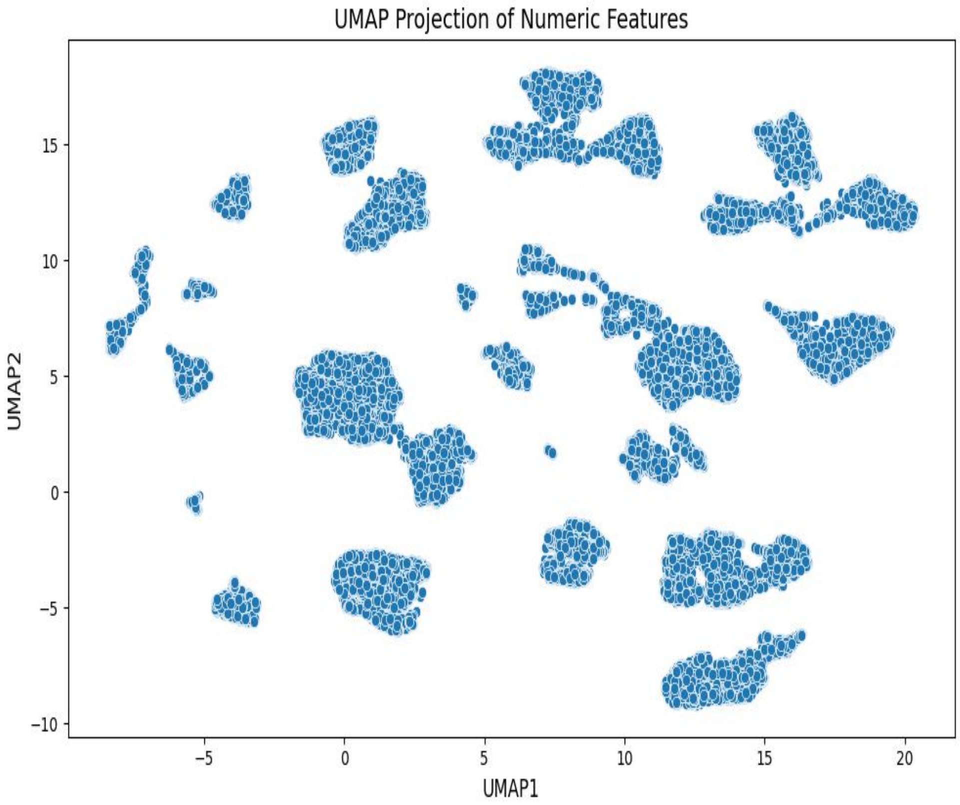


- The **heatmap** visualizes relationships between various **numeric features**.
    - **Red**: Positive Correlations
    - **Blue**: Negative Correlations

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



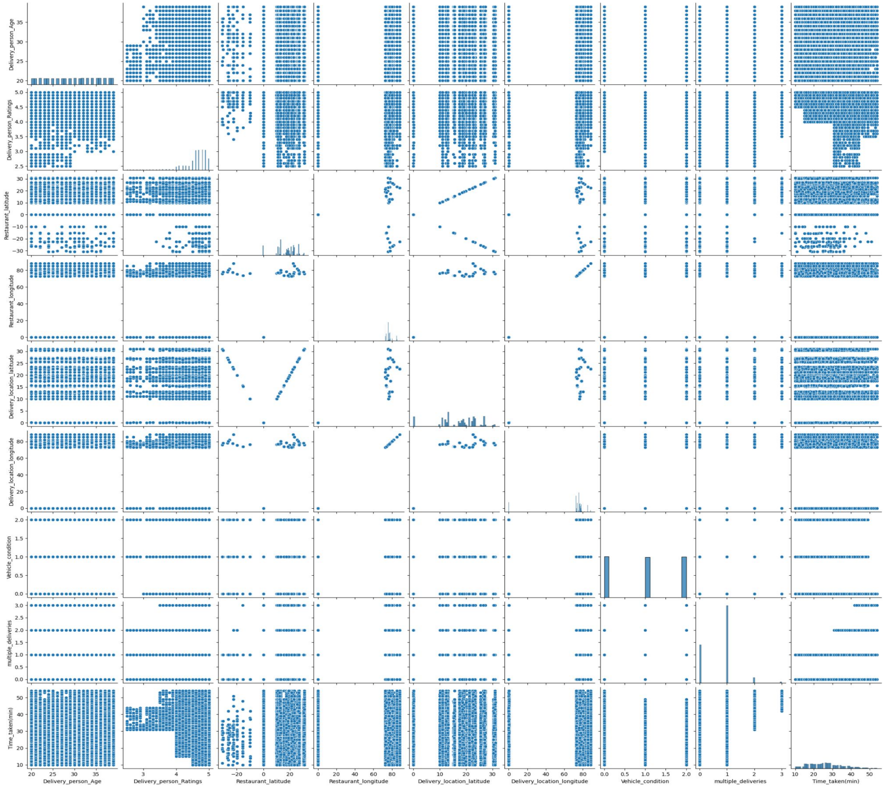
UMAP (Uniform Manifold Approximation and Projection) visualizes the projection of numeric features in the dataset into a **2D space**.

- 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



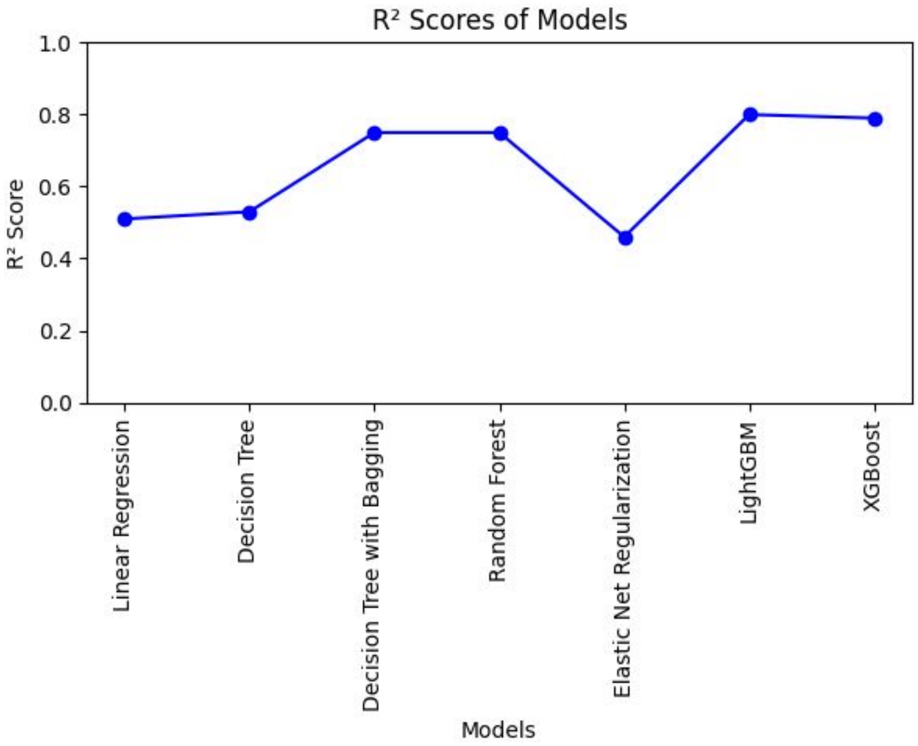
- **Delivery Location Latitude** and **Restaurant Latitude** exhibit a clear **linear trend**.
    - This suggests a **strong geographical alignment** between delivery and restaurant locations, indicating that deliveries typically occur within close proximity to the restaurant's latitude.

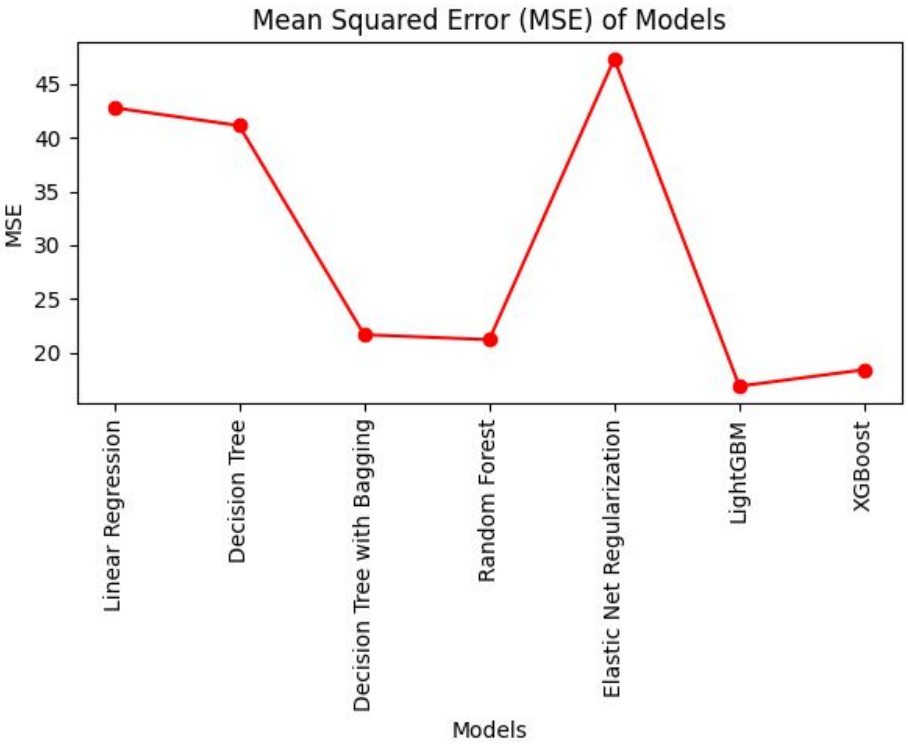
Methodology

Approach:

- **Data Preprocessing**:
  - Handled missing values and standardized the data to ensure consistency and comparability.
- **Feature Engineering**:
  - Created new features, such as the **distance** between the delivery location and the restaurant.
- **Modeling**:
  - Utilized **Ensemble Learning** techniques:
    - **Bagging**, **Random Forest**, **XGBoost**, **LightGBM**, and **K-Cross Validation**
  - 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 Absolute Error (MAE)** and **R^2 Score** were used to assess model performance.

Results And Analysis





Results, Analysis And Conclusion

R^2 Scores:

- **LightGBM** and **XGBoost** models have the highest R^2 scores, indicating better predictive performance with a strong ability to explain the variance in the data.
- **Elastic Net Regularization** shows the lowest R^2 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 R^2 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.

Future Work

Future Work:

- **Feature Selection**:
  - Utilize techniques like **K-Means Best Features** to refine and narrow down the most impactful features used in model training.
  - Integrating **SVM** can be used to evaluate the performance of the refined feature set by identifying optimal decision boundaries.
- **Further Enhancements**:
  - Perform **additional data preprocessing** and explore more advanced **ensemble methods** to further improve **prediction accuracy** and overall model performance.

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.

Upcoming Timeline:

- **Week 5-6**:
  - Refine models based on feedback.
  - Integrate new features for performance enhancement.
- **Week 7-8**:
  - Test and implement different feature combinations to further improve results.
- **Week 8-9**:
  - Write initial drafts for report sections.
  - Summarize model performance and key findings in the report.
- **Week 10-11**:
  - Complete the final report with detailed conclusions and visualizations.
  - Prepare for 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.

Thank You