Predictive Delivery Management System

DA 204o: Data Science in Practice - Course Project

Team Members:

Kashinath Alias Kapil Subhash Naik kashinathn@iisc.ac.in

Atreyee Mondal atreyeem@iisc.ac.in

Sujith Shetty sujith1@iisc.ac.in

Pranav N pranavn@iisc.ac.in

Problem Statement & Motivation

Business Challenge:

- Inaccurate ETAs on food delivery platforms
- High driver performance variability
- Customer dissatisfaction due to late deliveries
- Reduced platform trust and loyalty
- Inefficient resource and driver allocation
- Challenges in order batching and routing
- Negative impact on restaurant partnerships
- Higher operational costs for the platform

Why this Matters?

- Accurate ETAs enhance customer satisfaction and trust
- Reliable timing improves service quality and brand reputation
- Enables fair and data-driven driver performance evaluation
- Optimizes resource allocation
- Reduces operational costs and increases profitability
- Strengthens competitive advantage in the market

Objective

Predict delivery times accurately for individual orders.

Measure driver efficiency by comparing predicted vs. actual times.

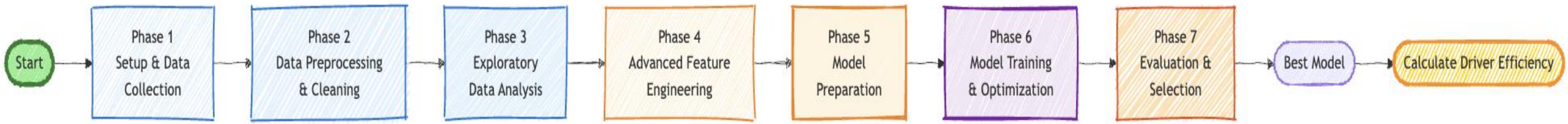
Identify key factors influencing delivery speed.

Enable data-driven Business decisions.

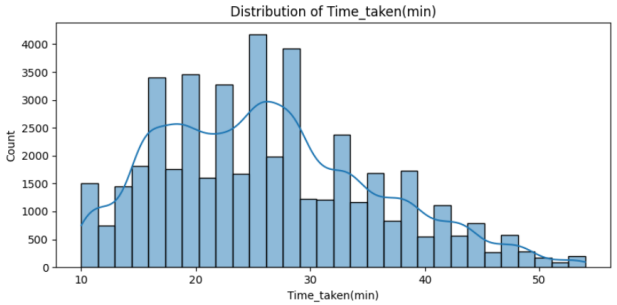
Data Acquisition & Preprocessing

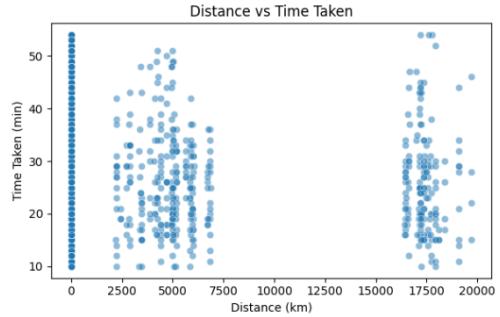
- **Dataset** : Food Delivery Dataset; comprising 45,593 records and 11 features.
- **Data Integrity**: No missing values detected, ensuring robust analysis.
- **Feature Engineering**: Geospatial features transformed to geodesic distances (km) for delivery route length.
- **Noise Reduction**: Redundant attributes (e.g., raw coordinates) removed to minimize noise and multicollinearity.
- **Categorical Encoding**: Categorical variables encoded for ML integration.
- **Outlier Detection**: Outliers detected and mitigated.
- **Data Partitioning**: Dataset split into 80% training, 20% testing

System Overview Architecture



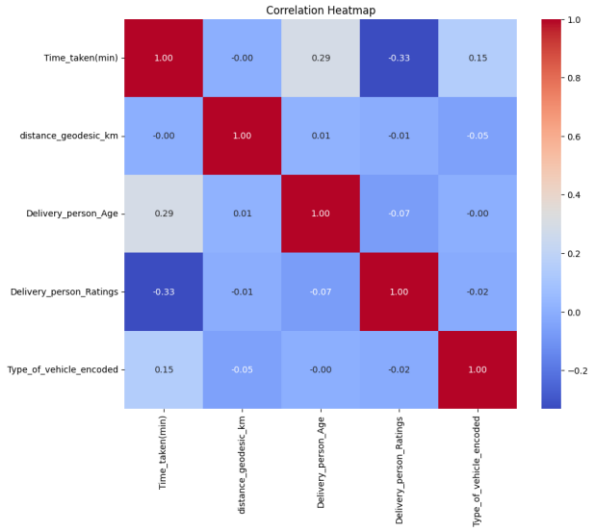
Exploratory Data Analysis





Univariate: Analysis of Time_taken, Age, and Ratings distributions showed Time_taken is nearly normal.

Bivariate: Geodesic Distance vs. Time - Analyzed to confirm the expected positive correlation between travel distance and duration.



Multivariate: To identify multicollinearity and assess feature redundancy, a Correlation Heatmap was generated. This matrix visualized the Pearson correlation coefficients between *Time_taken(min) distance_geodesic_km Delivery_person_Age Delivery_person_Ratings Type_of_vehicle_encoded*

Feature Engineering

- To enhance model performance, raw data was transformed into more meaningful predictors:
- Outliers: Distance clipped at the 99th percentile to remove anomalies.
- City Tier Segmentation: We parsed the *Delivery_person_ID* to extract city codes (e.g., 'BANG', 'MUM'). A custom mapping logic was applied to categorize these into City Tiers (1, 2, and 3), creating a proxy variable for traffic density and infrastructure quality.
- Interaction Features: We engineered features like traffic_adjusted_interaction (Distance × City Tier) and vehicle_distance_interaction (Distance ÷ Vehicle Type) to capture complex dependencies.
- Ordinal Encoding: The *Type_of_vehicle* feature was mapped to ordinal integers (e.g., Bicycle=0 to Motorcycle=3) to reflect the inherent order of vehicle speed and capacity.
- One-Hot Encoding: The *Type_of_order* feature (e.g., Snack, Meal, Buffet) was converted into binary dummy variables to allow models to treat each order type independently.

Models

- **Baseline Models: Linear Regression and Random Forest Regressor**

```
Training Linear Regression model...

 Linear Regression Performance:
RMSE: 7.92 min
MAE:  6.30 min
R²:   0.2843
```

```
Training Random Forest model...

 Random Forest Performance:
RMSE: 7.71 min
MAE:  6.01 min
R²:   0.3212
```

- **Ensemble Learning: Stacking Regressor**

```
RMSE: 7.21 min
MAE:  5.67 min
R²:   0.41
```

```
--- Building Stacking Ensemble ---
▸ StackingRegressor
  xgb                    rf
  ▸ XGBRegressor         ▸ RandomForestRegressor
  final_estimator
  ▸ LinearRegression
```

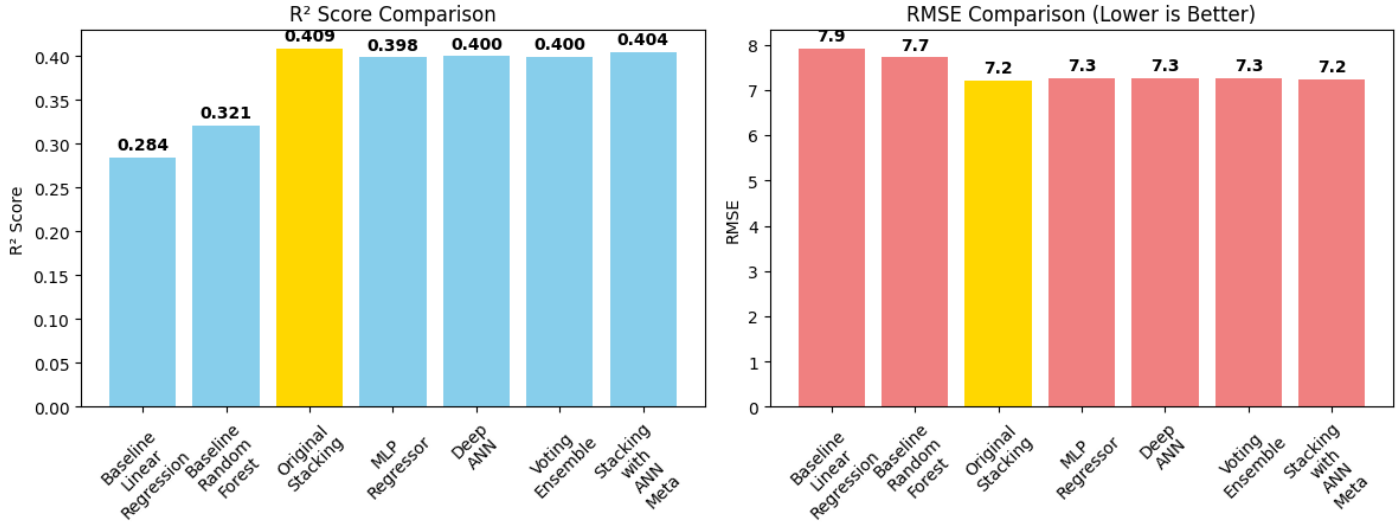
- **Neural Networks**
  - MLP

```
Best MLP Parameters: {'max_iter': 1000, 'learning_rate': 'adaptive', 'hidden_layer_sizes': (150, 100, 50), 'alpha': 0.0001, 'activation': 'tanh'}
MLP - RMSE: 7.26, MAE: 5.70, R²: 0.3984
```

  - ANN: Deep ANN - RMSE: 7.25, MAE: 5.71, R^2: 0.3999

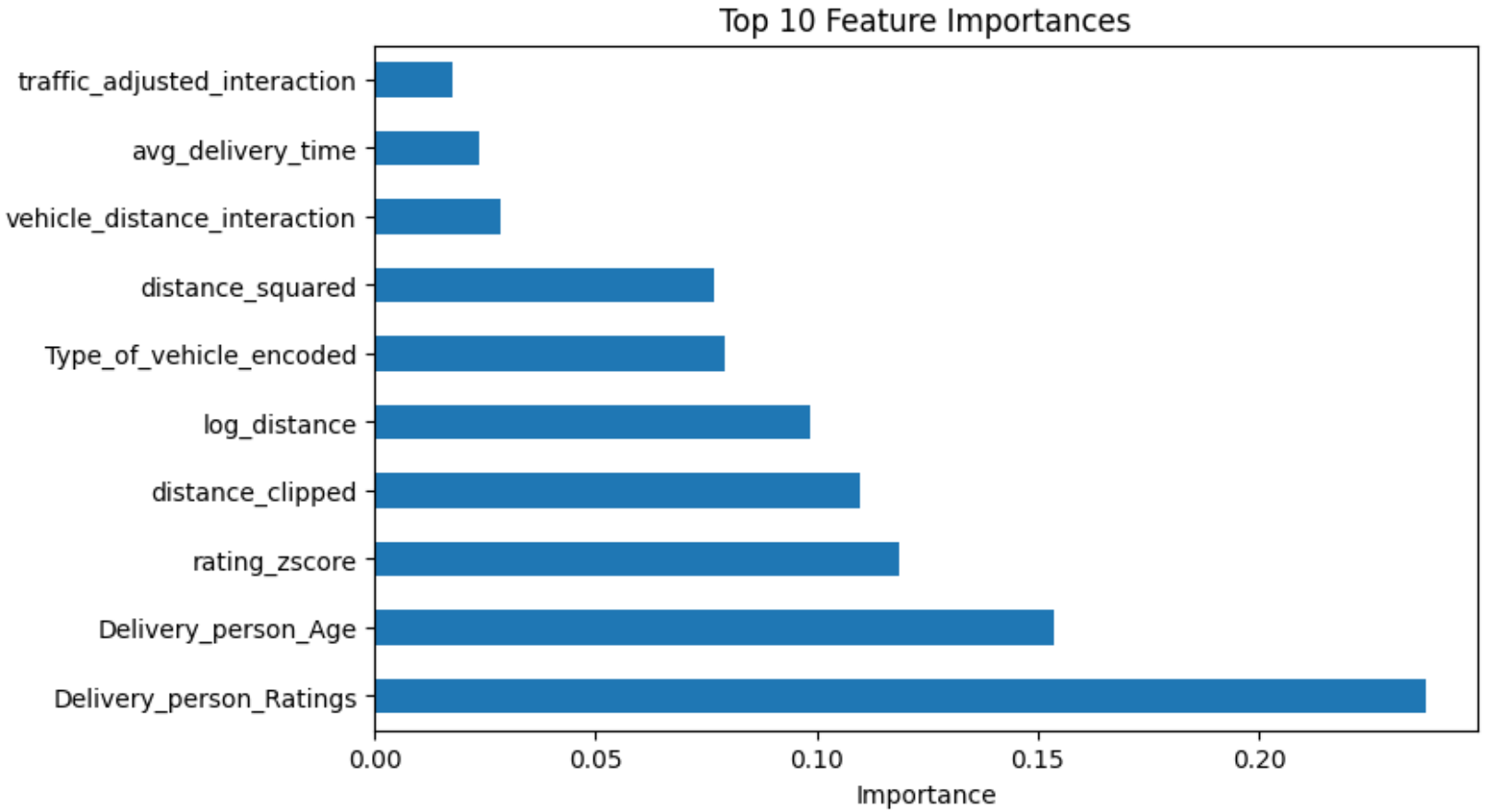
Model Performance

- Models were rigorously evaluated on a 20% hold-out test set using standard regression metrics:
  - RMSE (Root Mean Squared Error): To penalize larger errors, providing a clear measure of prediction accuracy in minutes.
  - MAE (Mean Absolute Error): To understand the average magnitude of errors.
  - R^2Score: To determine the proportion of variance in delivery time explained by the model.



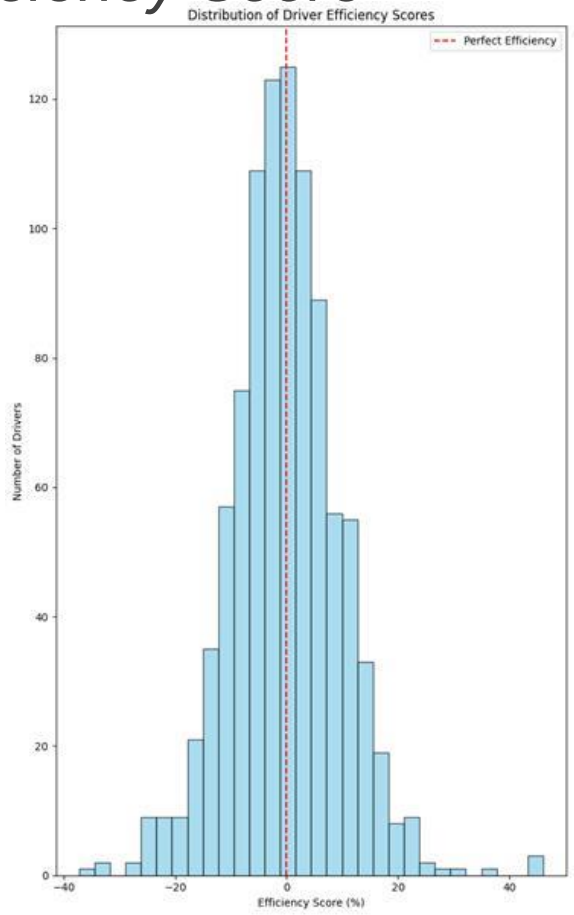
Analysis

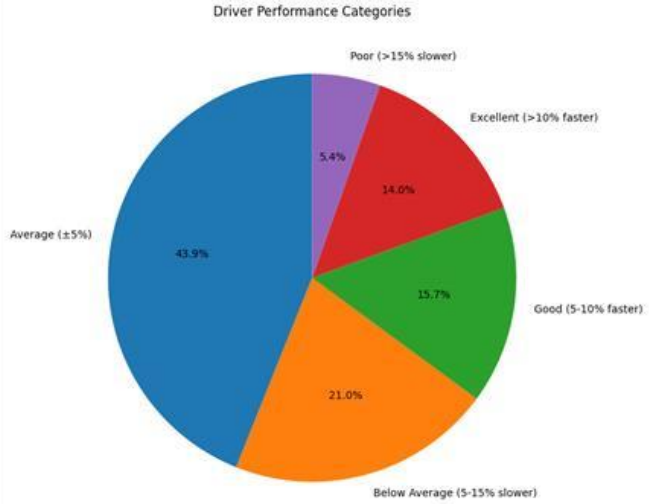
- Factors Affecting Delivery Performance



Inference

- We developed a Driver Efficiency Score for every active driver in the fleet.
- Metric Definition:
- *Efficiency Score* = $\frac{Actual\ Time - Predicted\ Time}{Predicted\ Time} \times 100$





Limitations and Future Scope

- Limitations
  - Coordinates are not accurate and some data needed to be clipped out as they were outliers.
  - Data quality issues due to driver reporting
  - Limited features in dataset. Live traffic and weather data were missing.

- Future Scope
  - Real-time Data: Integrate live traffic and weather APIs.
  - Automated Geofencing: Replace manual driver updates with GPS timestamps for accurate ground truth.
  - Realtime live dashboard at scale using distributed compute