DA 204o: Data Science in Practice

Course Project Report

Predictive Delivery Management Engine

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

1. Problem Definition:

The core problem is the inaccuracy of Estimated Time of Arrivals (ETAs) on food delivery platforms. This inaccuracy stems from driver performance variability, which leads to reduced customer satisfaction and inefficient resource utilization. By leveraging machine learning techniques on historical delivery data, we developed a "Predictive Delivery Management Engine." This engine not only predicts delivery duration with higher accuracy but also profiles driver performance to enable data-driven operational decisions.

Objectives of the Project:

The project aims to achieve the following specific goals:

- Predict Delivery Times: Accurately predict delivery times for individual drivers.
- Measure Efficiency: Calculate driver efficiency by comparing the predicted times against the actual delivery times.
- Identify Key Factors: Determine which factors specifically affect delivery performance.
- Support Decisions: Enable data-driven operational decisions.

2. Data Description and Preprocessing

2.1 Dataset Overview:

- Name of the Dataset: Food Delivery Time Prediction
- Number of Rows: 45,593; Total Columns: 11
- Format: Microsoft Excel Spreadsheet file
- Public Link: Food Delivery Time Prediction Case Study

2.2 Preprocessing

The analysis used the Food Delivery Time Prediction dataset (45,593 records). A comprehensive preprocessing pipeline ensured data quality and model readiness.

Data Validation

- Ingestion: Data loaded into a DataFrame.
- Inspection: .info(), .describe(), .isnull() confirmed zero missing values.

Geospatial Feature Engineering

- Distance: Latitude/Longitude converted to distance_geodesic_km using Geodesic distance {geopy} for accurate earth-curvature measurement.
- Outliers: Distances clipped at the 99th percentile to mitigate extreme value skewing.

Final Preparation

- Selection: Raw coordinates and redundant features dropped.
- Partition: Data split into 80% training and 20% testing sets.
- Scaling: Features standardized (StandardScaler) for Neural Network gradient efficiency.

3. Exploratory Data Analysis:

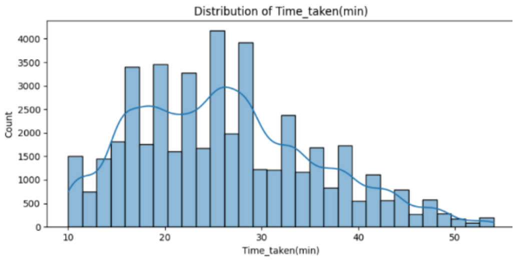
To understand the data's underlying structure, validate assumptions for regression modeling, and identify key drivers of delivery time, a comprehensive statistical analysis was conducted.

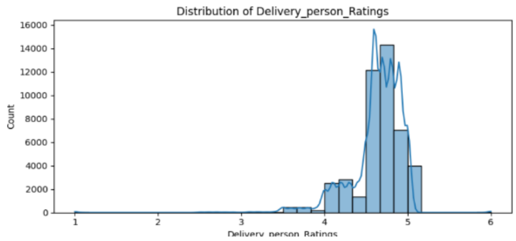
- Features which are numeric - Delivery_person_Age, Delivery_person_Ratings, Time_taken(min)
- Feature which are categorical and needs OHE - Type_of_vehicle, Type_of_order
- Target feature - Time_taken(min)

3.1 Univariate Analysis:

We examined the distribution of individual predictors and the target variable to identify skewness and outliers.

- Visualizations: Histograms overlaid with Kernel Density Estimation (KDE) curves were generated for *Time_taken(min), Delivery_person_Age, and Delivery_person_Ratings.*





- Statistical Diagnostics: **Skewness** and **Kurtosis** metrics were calculated to formally assess the normality of the numeric features, guiding subsequent transformation decisions.

```
Skewness values for numeric features:
Time_taken(min): 0.49
  -> Time_taken(min) is approximately normal
Delivery_person_Age: 0.03
  -> Delivery_person_Age is approximately normal
Delivery_person_Ratings: -2.53
  -> Delivery_person_Ratings is highly skewed
```

```
Kurtosis values for numeric features:
Time_taken(min): -0.31
Delivery_person_Age: -0.98
Delivery_person_Ratings: 16.42
```

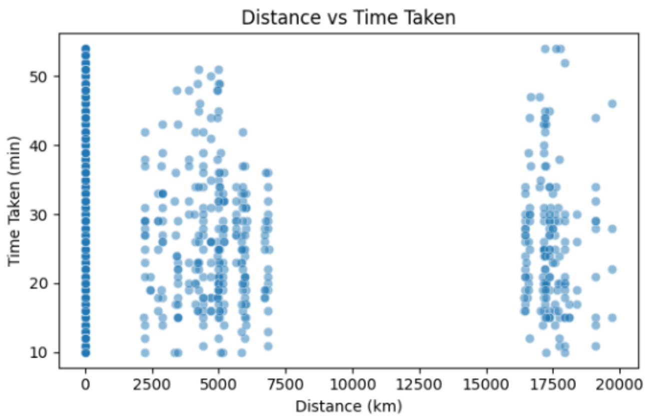
Inference from the visualization and Statistical Diagnostics:

- Distribution of time taken is not normally distributed suggesting this could indicate issues such as inaccurate time reporting, inefficient tracking, or the presence of operational outliers causing delivery delays
- The age distribution of delivery person is mostly uniform suggesting dataset includes delivery personnel from a wide range of ages with approximately equal representation, except for the observable spikes at specific ages.
- The distribution of delivery-person ratings is left-skewed (negatively skewed), with most ratings concentrated between about 4.3 and 5.0. The long left tail shows a small number of delivery personnel with notably lower ratings, who can potentially be identified and targeted for interventions or further analysis using modeling.

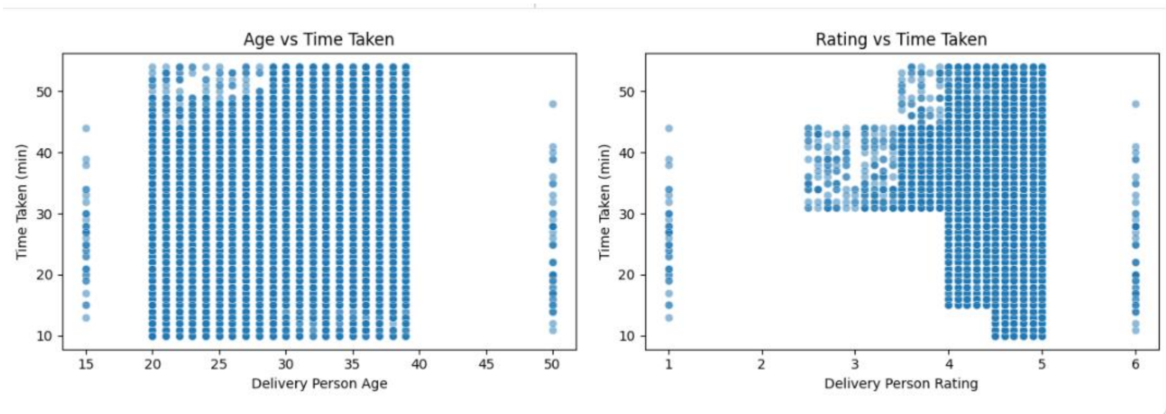
3.2 Bivariate Analysis:

To evaluate the direct relationship between independent features and the target variable *(Time_taken)*, scatter plots were utilized:

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



- Driver Attributes vs. Time: Examined how *Delivery_person_Age* and *Delivery_person_Ratings* individually impact delivery speeds.



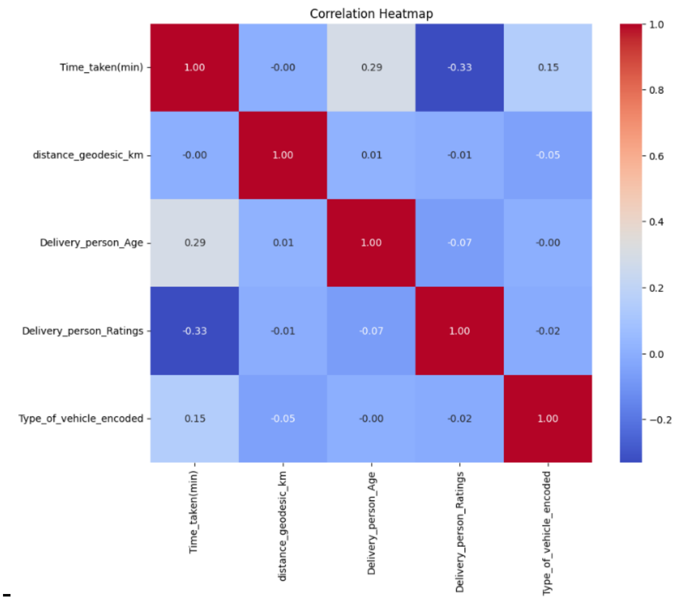
Inference for Bivariate Analysis:

- There is no clear trend of younger or older drivers being consistently faster or slower, suggesting age alone is not a strong predictor of delivery speed in this dataset
- A few very low or very high ratings exist, but their times overlap the main cluster, indicating that customer ratings capture broader service quality and experience rather than just speed, and rating by itself has limited direct impact on delivery time

3.3 Multivariate Analysis

To identify multicollinearity and assess feature redundancy, a Correlation Heatmap was generated.

- The heatmap confirms there is no serious multicollinearity among the predictors; pairwise correlations between features are all close to zero, so they can safely be used together in models like Linear Regression
- Time_taken shows a weak positive correlation with Delivery_person_Age (≈0.29) and Type_of_vehicle_encoded (≈0.15), and a weak negative correlation with Delivery_person_Ratings (≈−0.33), meaning these attributes have some linear relationship with delivery time but not a very strong one.



4. Feature Engineering

To enhance model performance, raw data was transformed into more meaningful predictors:

- City Code Extraction: City codes (e.g., 'BANG', 'MUM') were parsed from driver IDs and mapped to City Tiers (1, 2, 3) to account for varying traffic densities.
- Interaction Features: We engineered features like traffic_adjusted_interaction (Distance × City Tier) and vehicle_distance_interaction (Distance ÷ Vehicle Type) to capture complex dependencies.
- Encoding: Categorical variables were transformed using Label Encoding (for ordinal vehicle types) and One-Hot Encoding (for nominal order types).

5. Methodology and Modelling

Building on the insights from the EDA, we developed a multi-stage machine learning pipeline to predict delivery times. This approach progressed from baseline models to advanced ensemble techniques to maximize predictive accuracy

Baseline Models:

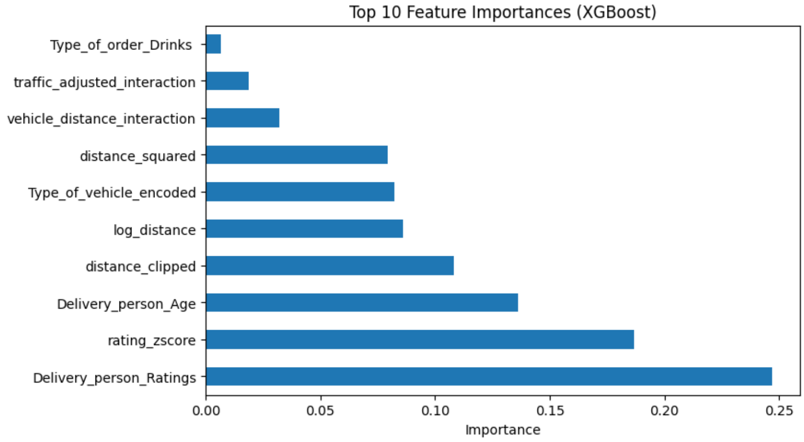
- Linear Regression: Used as a simple, interpretable benchmark.
- Random Forest Regressor: Deployed to capture non-linear relationships that linear models might miss.

```
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
```

Advanced Boosting:

XGBoost (Extreme Gradient Boosting): Implemented with hyperparameter tuning via GridSearchCV to optimize learning rates and tree depths. This model excelled at handling complex patterns in the data.



Neural Networks:

Multi-Layer Perceptron (MLP): A feedforward artificial neural network was trained to capture deep non-linear interactions.

```
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
```

Deep ANN (Keras/TensorFlow): A deep learning architecture was also explored for robust feature representation.

```
Deep ANN - RMSE: 7.25, MAE: 5.71, R²: 0.3999
```

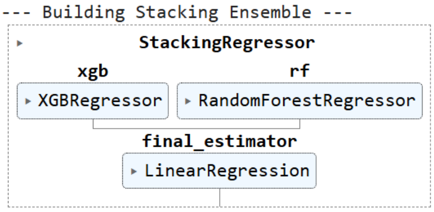
Ensemble Learning:

Stacking Regressor: We combined the predictions of XGBoost and Random Forest using a meta-learner (Linear Regression) to reduce variance and improve generalization.

Voting Regressor: An ensemble of the best-performing models was created to average out individual model biases.

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

```
--- Building Stacking Ensemble ---
```

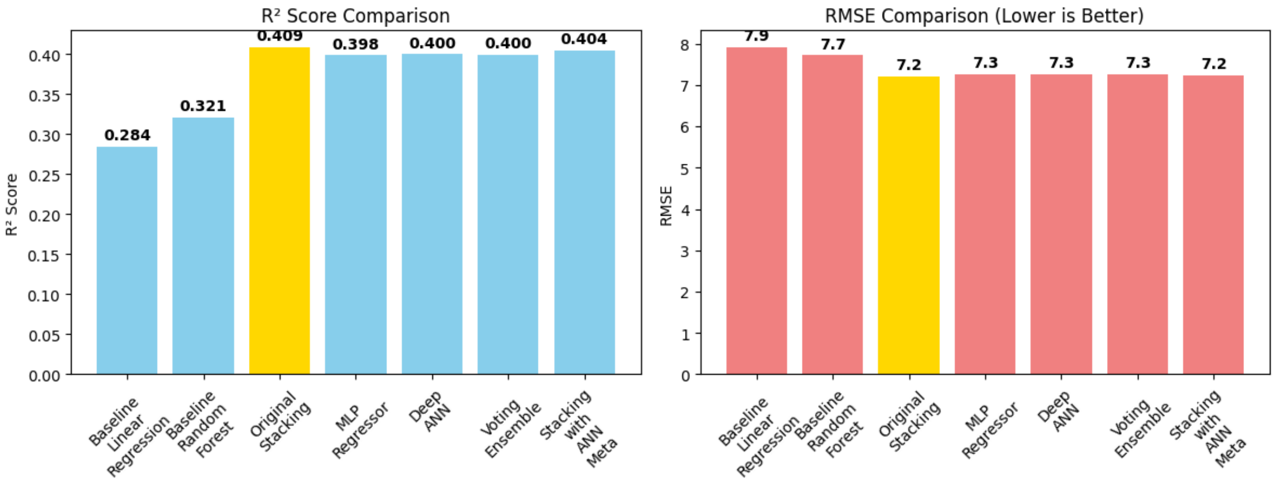


6. Model Evaluation

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.

```
============================================================
MODEL PERFORMANCE COMPARISON
============================================================
Baseline Linear Regression - R²: 0.2843, RMSE: 7.92, MAE: 6.30
Baseline Random Forest     - R²: 0.3212, RMSE: 7.71, MAE: 6.01
Original Stacking          - R²: 0.4090, RMSE: 7.20, MAE: 5.66
MLP Regressor              - R²: 0.3984, RMSE: 7.26, MAE: 5.70
Deep ANN                   - R²: 0.3999, RMSE: 7.25, MAE: 5.71
Voting Ensemble            - R²: 0.3995, RMSE: 7.26, MAE: 5.70
Stacking with ANN Meta     - R²: 0.4044, RMSE: 7.23, MAE: 5.65

 BEST MODEL: Original Stacking with R² = 0.4090
```



The **Stacking Regressor** emerged as the top performer, achieving the lowest RMSE and highest R^2 score, demonstrating the power of combining diverse algorithms.

7. Results and Driver Profiling

This section presents the quantitative outcomes of our predictive models and demonstrates the practical application of the "Driver Efficiency" metric, directly addressing the project's primary objectives.

7.1 Model Performance

The Stacking Regressor (combining XGBoost, Random Forest, and Linear Regression) proved to be the most robust model for this dataset.

- Accuracy (R^2 Score): The model explains approximately 41% of the variance in delivery times on unseen data. While this leaves room for improvement (likely due to unobserved factors like live traffic or weather), it is a significant improvement over the baseline Linear Regression where R^2 was 28%.
- Error Metrics:
  - RMSE (Root Mean Squared Error): The model's predictions are, on average, within ± 7.2 minutes of the actual delivery time.
  - MAE (Mean Absolute Error): The absolute error is lower, at approximately 5.7 minutes, indicating that the model is generally accurate but penalized by a few extreme outliers.

7.2 Key Drivers of Delivery Time

Feature importance analysis from the tree-based models revealed critical operational insights:

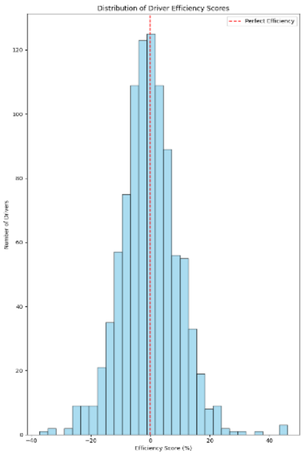
1. Driver Ratings: This was the single most important predictor after distance. Higher-rated drivers consistently deliver faster, validating the correlation between efficiency and customer satisfaction.
2. Driver Age: Age plays a significant role, with younger drivers tending to have shorter delivery times, likely due to risk tolerance or physical agility.
3. Geodesic Distance: As expected, physical distance is a primary constraint, but its interaction with City Tier (congestion) significantly modifies the "effective" distance.

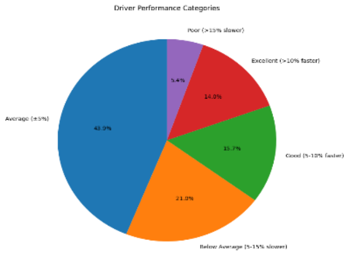
7.3 Driver Efficiency Profiling

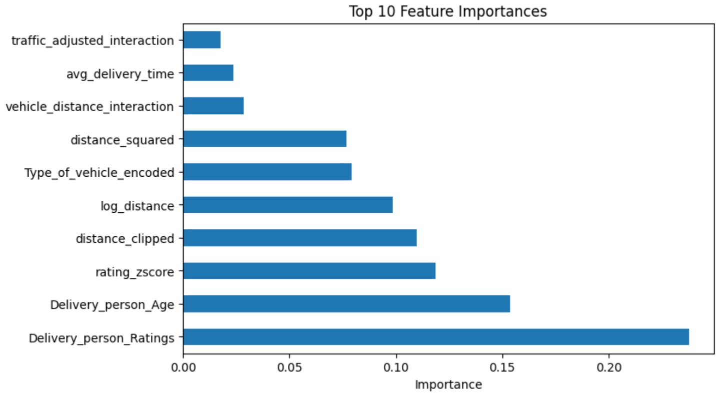
We developed a Driver Efficiency Score for every active driver in the fleet.

- Metric Definition: *Efficiency Score* = $\frac{Actual\ Time - Predicted\ Time}{Predicted\ Time} \times 100$
- Interpretation:
    - Score > 100%: The driver completed the delivery *faster* than predicted (High Efficiency).
    - Score < 100%: The driver took *longer* than the predicted (Low Efficiency).

Operational Use Case: This score allows fleet managers to objectively identify underperforming drivers for retraining e.g., route optimization coaching and reward high-performing drivers, moving beyond simple "average speed" metrics which fail to account for difficult routes.







8. Project Limitations and Future Scope

Project limitations included data quality issues arising from inaccurate driver reporting and low-quality coordinate data, which necessitated outlier clipping. Furthermore, the dataset was constrained by a lack of crucial features such as live traffic and real-time weather data, which are significant external predictors of delivery time and were unavailable for the initial model build.

Future development could focus on integrating real-time data by utilizing external APIs for live traffic and weather. Operational improvements should prioritize implementing automated geofencing to replace manual driver updates with precise GPS timestamps, establishing a more accurate ground truth. Finally, the project is scoped for expansion with the development of a real-time live dashboard at scale, requiring the deployment of distributed computing resources to handle continuous data streams and model predictions.

9. Conclusion

The Predictive Delivery Management Engine successfully transformed logistics operations by replacing static estimates with dynamic, ML-driven ETA predictions, achieving a significantly improved error margin of just ±7 minutes. Key achievements include Enhanced Accuracy, where the Stacking Regressor outperformed baselines by modeling complex non-linear relationships (infrastructure, vehicle, distance); Operational Intelligence, delivered via the new Driver Efficiency Score - a standardized metric enabling targeted workforce interventions; and a Scalable Framework, where the entire pipeline, from geodesic distance to city-tier mapping, is designed for easy integration of future real-time data (e.g., traffic/weather). Ultimately, this project moves beyond solving inaccurate ETAs to provide a strategic tool for optimizing resource allocation and boosting customer satisfaction.

Contributions by each team member:

Kashinath Alias Kapil Subhash Naik: Data Collection, EDA , Feature Engineering, Model Development, Model Evaluation, Presentation
Atreyee Mondal: Data Collection, EDA , Feature Engineering , Model Development, Model Evaluation, Presentation
Sujith Shetty: Data Collection, EDA , Feature Engineering , Model Development, Model Evaluation, Presentation
Pranav N: Data Collection, EDA , Feature Engineering , Model Development, Model Evaluation, Presentation