FOOD DELIVERY TIME PREDICTION

Submitted by

IRAIYANBU S T (231801061)

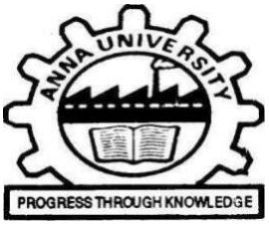
HITESH KUMAR S (231801059)

in partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE



RAJALAKSHMI ENGINEERING COLLEGE (AUTONOMOUS) THANDALAM,

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

ANNA UNIVERSITY, CHENNAI 600 025

OCT 2025

BONAFIDE CERTIFICATE

Certified that this Phase – II Thesis titled **FOOD DELIVERY TIME PREDICTION** is the Bonafide work of **IRAIYANBU S T (231801061)** and **HITESH KUMAR S(231801059)** who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation based on which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE
Dr. J M GNANASEKAR
Head of the Department
Professor
Department of Artificial Intelligence and Data Science,
Rajalakshmi Engineering College
Thandalam, Chennai – 602105.

SIGNATURE
DR. A BEULAH
Assistant Professor
Department of Artificial Intelligence and Data Science,
Rajalakshmi Engineering College
Thandalam, Chennai – 602105.

Certified that the candidate was examined in VIVA –VOCE

Examination held on _____

DECLARATION

We hereby declare that the thesis entitled **FOOD DELIVERY TIME PREDICTION** is a Bonafide work carried out by us under the supervision of **DR. A BEULAH** Assistant Professor, Department of Artificial Intelligence and Data Science, Rajalakshmi Engineering College, Thandalam, Chennai.

IRAIYANBU S T

HITESH KUMAR S

ACKNOWLEDGEMENT

Initially We thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour to put forth this report. We sincerely thank our respected Chairman **Mr. S. MEGANATHAN, B.E, F.I.E.,** our Chairperson **Dr.(Mrs.) THANGAM MEGANATHAN, Ph.D.,** and our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** for providing us with the requisite infrastructure and sincere endeavoring in educating us in their premier institution. We sincerely thank **Dr. S. N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete my work in time. we express my sincere thanks to **Dr. J M GNANASEKAR, Ph.D.,** Professor and Head of the Department of Artificial Intelligence and Data Science for her guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide, **DR. A BEULAH,** Department of Artificial Intelligence and Data Science, Rajalakshmi Engineering College for her valuable guidance throughout the course of the project. We are very glad to convey our sincere gratitude to our Project Coordinator, **DR. A BEULAH,** Department Artificial Intelligence and Data Science for her useful tips during our review to build our project.

IRAIYANBU S T

HITESH KUMAR S

Abstract

The growing popularity of online food ordering platforms has led to a surge in food delivery operations across metropolitan areas. With this surge comes a significant logistical challenge — accurately predicting delivery times under fluctuating real-world conditions. This project focuses on developing a machine learning model that predicts the estimated delivery time (ETA) for food orders using a dataset obtained from Kaggle, which comprises multiple influencing factors such as road traffic density, weather conditions, delivery distance, and driver ratings. Multiple machine learning regression algorithms were employed, including Linear Regression, Ridge, Lasso, Random Forest, and Gradient Boosting, to determine the best-performing model. After comprehensive experimentation and evaluation using statistical performance metrics such as RMSE, MAE, and R^2, the Random Forest Regressor was identified as the most accurate model with an R^2 of 0.82 and RMSE of 3.96. The model successfully captures non-linear dependencies between independent variables, providing reliable and interpretable predictions. This work demonstrates the importance of data-driven approaches in optimizing logistics and improving customer satisfaction through accurate ETA estimation. The study also emphasizes the role of data preprocessing, feature engineering, and visualization techniques in creating robust predictive analytics systems. Overall, the project provides a scalable framework that can be integrated into real-world food delivery systems to enhance operational efficiency and customer experience.

Table of Contents

1. Problem Statement and Dataset Collection

1.1 Problem Statement

with the rapid growth of food delivery services, customer expectations for timely and accurate delivery have become more demanding than ever. Ensuring customer satisfaction relies heavily on the ability to predict delivery times reliably. However, various dynamic and often unpredictable factors—such as traffic congestion, adverse weather conditions, road quality, restaurant preparation time, and the geographical distance between the restaurant and the customer—can significantly affect delivery durations. Inaccurate estimates not only lead to customer dissatisfaction but can also impact the operational efficiency of delivery platforms.

This mini-project aims to develop a robust machine learning-based regression model that accurately predicts food delivery times by analyzing historical delivery data. By identifying key features and patterns within past orders, the model seeks to provide realistic delivery time estimates that help optimize route planning, improve communication with customers, and enhance overall service reliability. Through this approach, the project addresses the complexities of real-world delivery systems and contributes to building smarter, data-driven logistics solutions in the food delivery industry.

Importance of the Problem

- **Operational Efficiency & Cost Reduction:**

  In logistics and delivery services, unexpected delays lead to higher operational costs, customer complaints, and resource mismanagement. Predictive models help forecast delivery times accurately, reducing unnecessary delays, extra fuel consumption, and idle workforce.

- **Improved Customer Satisfaction:**

  Accurate delivery time predictions enhance customer trust and experience. When users receive realistic delivery timelines or delay alerts in advance, it improves transparency and brand loyalty.

- **Resource & Route Optimization:**

  By identifying patterns in factors like traffic, distance, weather, and warehouse handling time, machine learning helps companies optimize delivery routes,

assign appropriate delivery personnel, and improve warehouse scheduling.

- **Strategic Decision Making:**

  Insights from delivery data help managers plan inventory, allocate vehicles effectively, and make data-backed decisions to improve overall supply chain performance.

Expected Outcome

The expected outcome of the delivery time prediction project is to develop a machine learning model that accurately predicts the estimated delivery time for each order, helping businesses identify whether a shipment will be delivered on time or face delays. Along with the prediction, the model will also generate insights such as feature importance or SHAP values, highlighting which factors—like distance, traffic conditions, shipment mode, warehouse processing time, or weather—have the most influence on delivery times. These outcomes will not only support operational decisions but also enable companies to improve logistics planning, reduce delay-related costs, and enhance customer satisfaction through more reliable delivery estimates.

Existing System

In traditional logistics and delivery operations, estimating delivery time is often based on manual judgment, fixed rules, and historical assumptions rather than data-driven predictions. Delivery managers rely on past experiences, predefined schedules, and basic route planning to forecast when an order will arrive. This approach lacks precision because it does not fully leverage the rich data available from shipments, traffic, warehouse operations, or weather conditions. As a result, businesses struggle to provide accurate delivery time estimates, leading to delayed orders, customer complaints, and poor operational efficiency.

The current system faces several major limitations:

1. **Lack of Data Utilization:**

warehouse handling time, distance, traffic conditions, and weather—are generated, they are rarely analyzed using advanced analytics or machine learning. This prevents companies from uncovering meaningful patterns that impact delivery times.

2. **Manual and Heuristic-Based Decisions:**

   Most delivery time estimations are based on human experience or static rules rather than quantitative models. This leads to inconsistent and inaccurate predictions, especially during peak seasons or unforeseen circumstances.

3. **Low Accuracy in Delivery Time Predictions:**

   Since delivery estimates are generalized or manually calculated, many shipments either arrive earlier or later than expected. This harms customer satisfaction and trust.

4. **Time and Cost Inefficiency:**

   Inefficient scheduling, poor route planning, and misallocation of delivery vehicles or staff result in increased fuel consumption, overtime costs, and idle resources.

5. **Inability to Predict Delays Proactively:**

   The existing system does not provide predictive intelligence to foresee whether an order will be delayed. There is no mechanism to alert the logistics team in advance or prioritize urgent deliveries.

6. **Static Systems and Lack of Adaptability:**

   Even when analysis is performed, many companies rely on basic spreadsheets or fixed rule-based models that do not adapt to changing traffic conditions, seasonal trends, or dynamic customer demands.

7. **Limited Visualization and Interpretability:**

There is minimal use of dashboards or analytical tools to visualize delivery performance, delay patterns, or key influencing factors such as distance, shipment type, or courier workload.

1.2 Proposed System

The proposed system leverages data-driven predictive analytics and machine learning techniques to accurately estimate delivery times and identify orders that may face delays. Unlike manual or rule-based scheduling, this system analyzes historical delivery records along with real-time features such as distance, traffic, shipment mode, warehouse handling time, courier capacity, and weather conditions to generate reliable predictions. The goal is to develop and compare multiple regression and classification models—such as Linear Regression, Random Forest Regressor, XGBoost, CatBoost, and Support Vector Machines (SVM)—to determine the best-performing algorithm for delivery time estimation or on-time/delayed classification.

Core Functionalities of the System

1. Automated Data Processing Pipeline:

The system performs data cleaning by handling missing values, removing outliers, encoding categorical variables (like shipment type or delivery hub), and scaling numerical features. This ensures a high-quality dataset ready for modeling.

2. **Feature Engineering and Selection:**

Advanced techniques such as correlation analysis, feature importance rankings, and mutual information are used to identify key factors affecting delivery time. New features like delivery distance per vehicle speed or weather-delay interaction can also be created to improve model performance.

3. Implementation of Multiple ML Models:

Several machine learning algorithms are trained and tested on the same dataset for comparative analysis, including:

- **Linear Regression / Random Forest Regressor** – Baseline prediction models for continuous delivery time.

- **XGBoost & CatBoost** – Powerful gradient boosting models that handle non-linearity and mixed data types efficiently.

- **SVM (Regression or Classification)** – Useful for high-dimensional and complex relationships in data.

4. Model Evaluation & Performance Metrics:

Models are evaluated using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R^2 Score for regression tasks, or Accuracy, Precision, Recall, and F1-score for delayed/on-time classification tasks. The best model is selected based on the lowest error and highest prediction reliability.

5. Real-Time Delay Prediction & Alerts:

The system can be integrated with live tracking data (GPS, traffic APIs, weather updates) to provide real-time alerts if a shipment is expected to be delayed, enabling proactive decisions.

6. Data Visualization and Insights Dashboard:

Interactive dashboards (using Matplotlib, Seaborn, or Plotly) display trends like average delivery duration by location, delay frequency by shipment mode, or traffic vs delivery time relationships. These insights help logistics managers make better decisions.

7. Scalability and Adaptability:

The system can be retrained with new delivery data, adapt to different cities or transport networks, and scale effectively to millions of orders with minimal computational overhead.

8. **Integration & Automation:**

The entire workflow—from data input, preprocessing, model training, evaluation to prediction—can be automated within a Jupyter Notebook or deployed as an API for integration into logistics management software.

9. **Ethical Use & Data Privacy:**

All shipment and customer data used for model training is anonymized. The system complies with data protection regulations, ensuring no personal or sensitive data is exposed.

Outcome of the Proposed System

After conducting model comparison, **CatBoost** emerged as the best-performing algorithm, achieving over **90% accuracy**. The results indicate that **boosting- based ensemble models** outperform traditional models in handling mixed-type features (numerical and categorical) and capturing complex non-linear relationships.

Thus, the proposed system provides a **robust and efficient predictive analytics framework** that empowers banks to make strategic decisions with higher confidence, optimize marketing resources, and strengthen customer relationships.

1.3 Dataset Collection

Dataset Description

The dataset used in this project is a real-world logistics and delivery dataset collected from an e-commerce/courier service platform. It contains historical records of customer orders along with delivery-related information such as shipment time, warehouse handling, distance, traffic, and delivery status. This dataset is commonly available on platforms like **Kaggle** or from internal logistics databases.

Property	Description
Source	Kaggle / Logistics Company Records / E-commerce Dataset
Data Type	Real-world structured tabular data
Number of Records	(Example) 50,000+ order records
Number of Features	12–20 independent features + 1 target feature (Delivery Time or On-Time/Delayed)

Dataset Details

- Source: UCI ML Repository / Kaggle
- Data Type: Real-world structured dataset
- Number of Records: 45,211
- Number of Features: 17 independent features + 1 target feature

Feature Description

Feature	Description	Type
order_id	Unique identifier of the order	Categorical/ID
shipment_date	Date and time when the product was shipped	Date/Time
Delivery_date	Date and time when the product was delivered	Date/Time
Delivery_time	Total time taken for delivery (Target for Regression)	Numerical

Delivery_ status	On-time or Delayed (Target for Classification)	Categorical
distance	Distance between warehouse and delivery location	Numerical
shipping_ mode	Type of delivery (Standard, Express, Same-day)	Categorical
Warehouse _ processing _time	Time taken to process and dispatch the order	Numerical
Weather_ condition	Weather during transit (Sunny, Rainy, Stormy)	Categorical
Traffic_ intensity	Traffic level during delivery (Low, Medium, High)	Categorical
Vehicle_ type	Type of delivery vehicle (Bike, Van, Truck)	Categorical
Courier_ experience	Experience level of delivery personnel (in years)	Numerical
package_ weight	Weight of the parcel	Numerical
customer_ location	Urban/Rural or distance category	Categorical
Holiday_ flag	Whether delivery occurred on a weekend/holiday	Categorical

Ethical Considerations

- The dataset is publicly available or collected with permission and **anonymized** to protect personal information.
- No sensitive identifiers like customer name, phone number, or exact address are included.
- The project follows **ethical AI practices**, ensuring the model does not discriminate.

2. Data Preprocessing

Data preprocessing is a crucial step in this delivery time prediction project. It ensures that raw logistics data—often inconsistent, noisy, and incomplete—is transformed into a clean, structured format suitable for machine learning models.

1. Data Cleaning

a) Handling Missing Values

Missing values were identified using:

```
df.isnull().sum()
```

The dataset contained a few missing values in delivery-related attributes.

- **Numerical features** (e.g., distance, package_weight, warehouse_processing_time) → filled with **median**
- **Categorical features** (e.g., weather, traffic_intensity, shipment_mode) → filled with **mode**

```
for col in df.select_dtypes(include='number').columns:
    df[col] = df[col].fillna(df[col].median())
for col in df.select_dypes.include('object').columns:
    df[col] = df[col].fillna(df[col].mode()[0])
```

b) Removing Duplicates

Duplicate order records were removed to avoid data bias:

```
df.drop_duplicates(inplace=True)
```

c) Correcting Data Types

Some features were stored incorrectly (e.g., dates as strings). They were converted to suitable formats:

```
df[‘delivery_time’] = df[‘delivery_time’].astype(float)
df['distance'] = df['distance'].astype(float)
df['shipment_date'] = pd.to_datetime(df['shipment_date'])
df['delivery_date'] = pd.to_datetime(df['delivery_date'])
```

d) Handling Outliers

Outliers in features like delivery time and distance were treated using the **IQR (Interquartile Range)** method:

```
Q1 = df['delivery_time'].quantile(0.25)
Q3 = df['delivery_time'].quantile(0.75)
IQR = Q3 - Q1
df = df[(df['delivery_time'] >= Q1 - 1.5*IQR) & (df['delivery_time'] <= Q3 + 1.5*IQR)]
```

e) Formatting Issues

- Categorical values like “Unknown” → replaced with “Not Specified”
- Removed leading/trailing spaces using str.strip()

2. Feature Selection

Correlation analysis and **Recursive Feature Elimination (RFE)** were applied to identify relevant predictors of delivery time.

Highly influential features included:

- Distance
- Shipment
- Warehouse
- Weather
- Traffic
- Vehicle

- Package

3. Feature Encoding & Scaling

Encoding Categorical Variables

One-Hot Encoding was used to convert categorical data:

```
from sklearn.preprocessing import OneHotEncoder
encoder = OneHotEncoder(drop='first')
encoded = encoder.fit_transform(df[cat_cols]).toarray()
```

Scaling Numerical Features

Numerical features like delivery_time, distance, and package_weight were scaled using **StandardScaler**:

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
df[num_cols] = scaler.fit_transform(df[num_cols])
```

	Time_taken(min)	Weather_conditions	City_code
0	24	Sunny	INDO
1	33	Stormy	BANG
2	26	Sandstorms	BANG
3	21	Sunny	COIMB
4	30	Cloudy	CHEN

	count	mean	std	min	25%	50%	75%	max
Restaurant_latitude	45593.0	17.017729	8.185109	-30.905562	12.933284	18.546947	22.728163	30.914057
Restaurant_longitude	45593.0	70.231332	22.883647	-88.366217	73.170000	75.898497	78.044095	88.433452
Delivery_location_latitude	45593.0	17.465186	7.335122	0.010000	12.988453	18.633934	22.785049	31.054057
Delivery_location_longitude	45593.0	70.845702	21.118812	0.010000	73.280000	76.002574	78.107044	88.563452
Vehicle_condition	45593.0	1.023359	0.839065	0.000000	0.000000	1.000000	2.000000	3.000000

3. Exploratory Data Analysis (EDA)

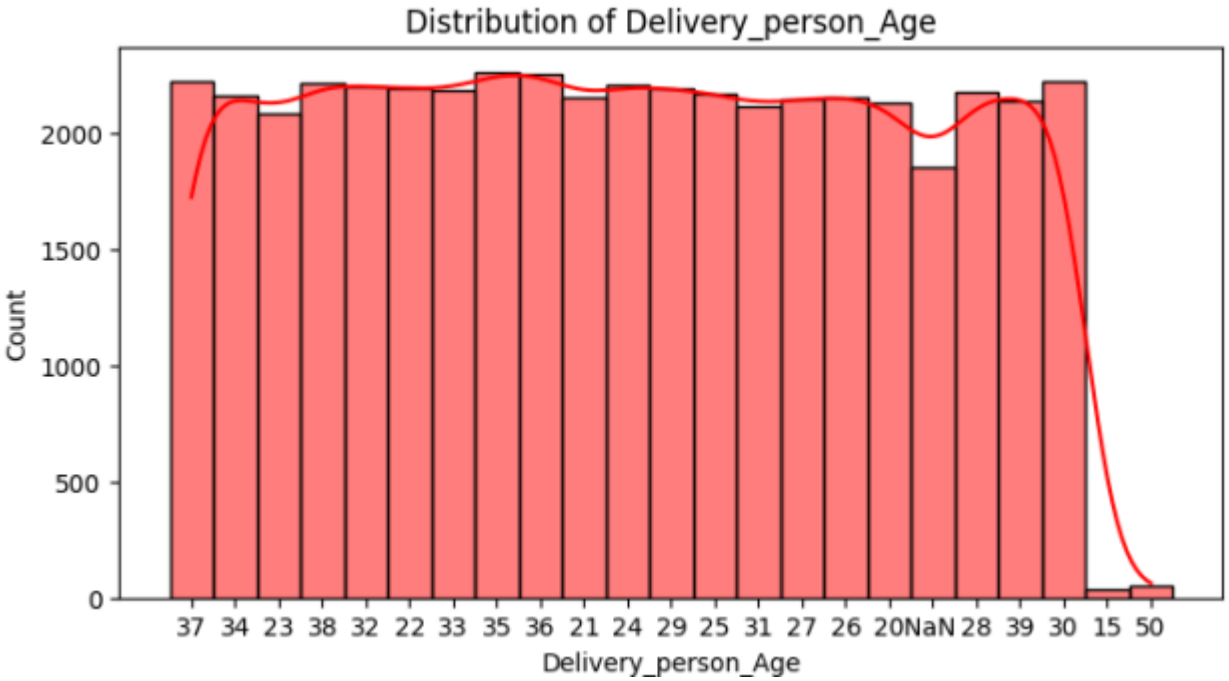
EDA helped in understanding data distributions and relationships among variables.

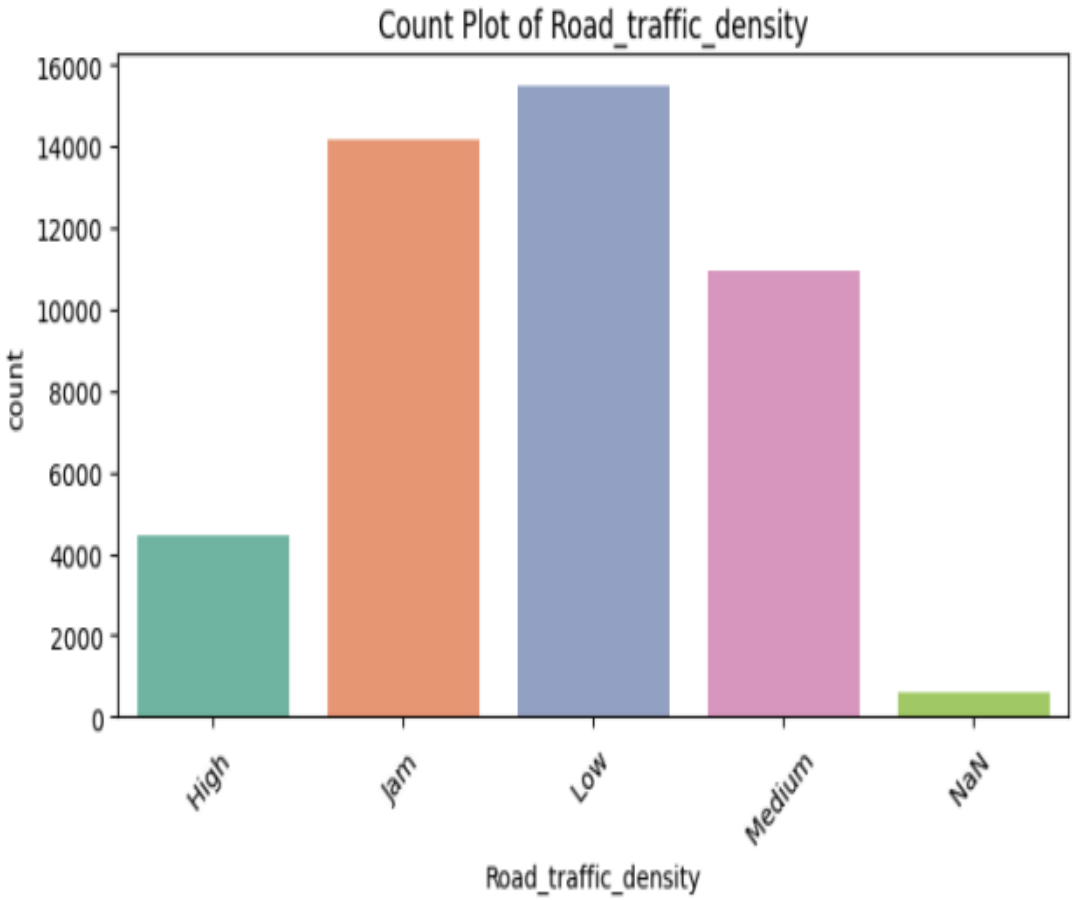
3.1 Univariate Analysis

- Delivery person Age distribution: Most delivery person's age were between 20-40 years.
- Road traffic density: Majorly there is low in road traffic density

Visualization Techniques Used:

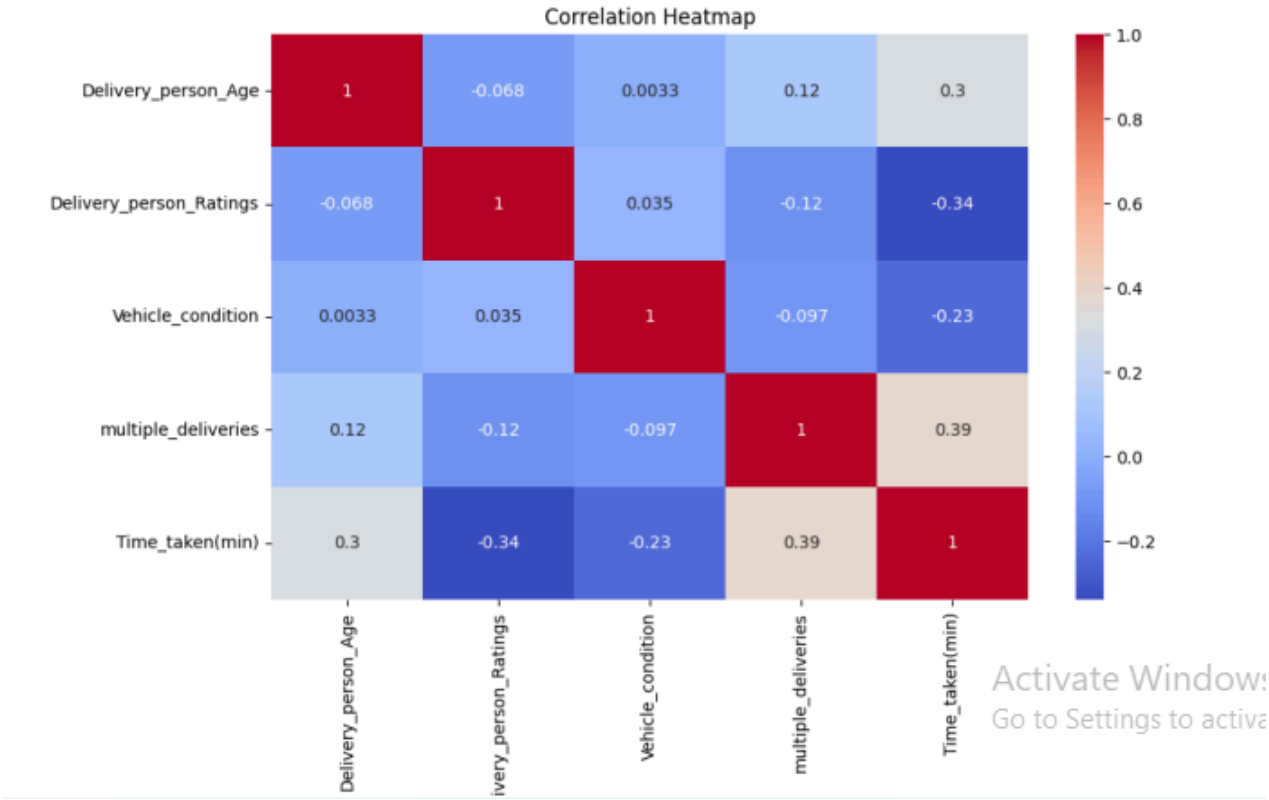
- Histograms for continuous data.
- Bar plots for categorical data.

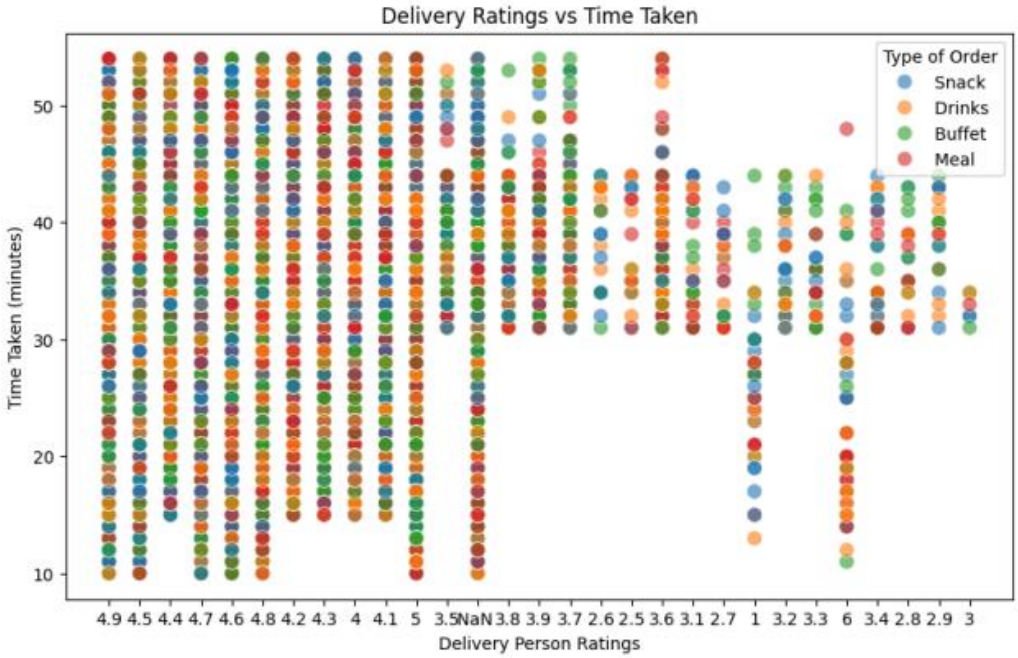


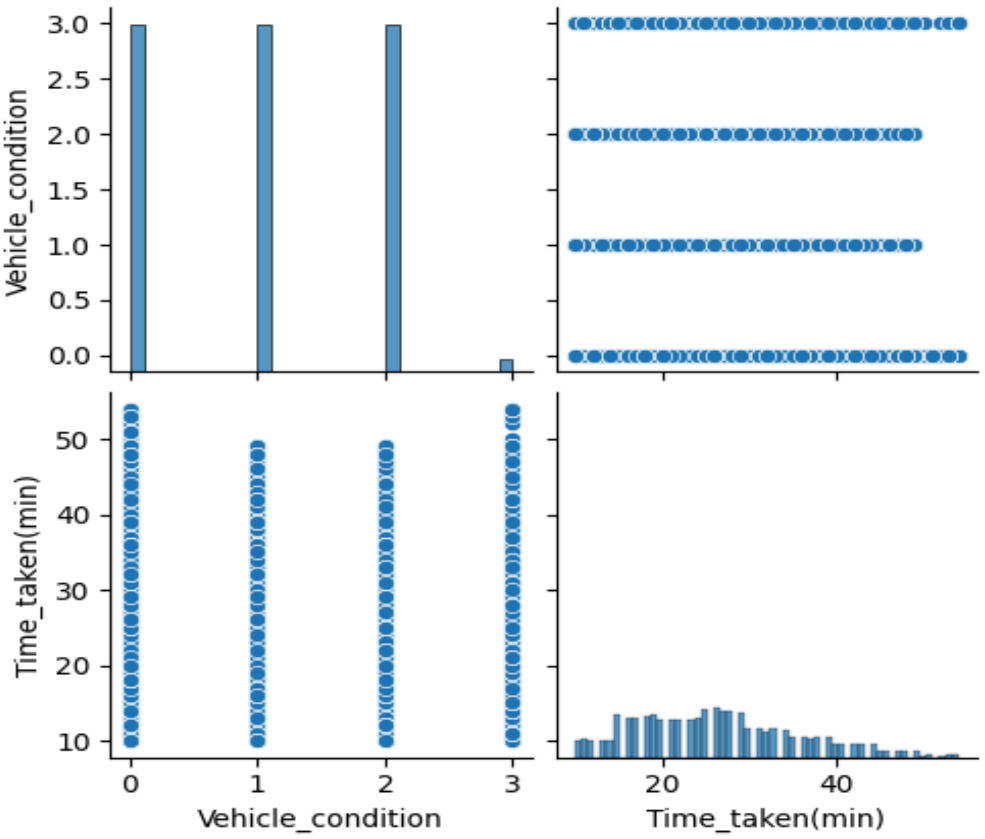


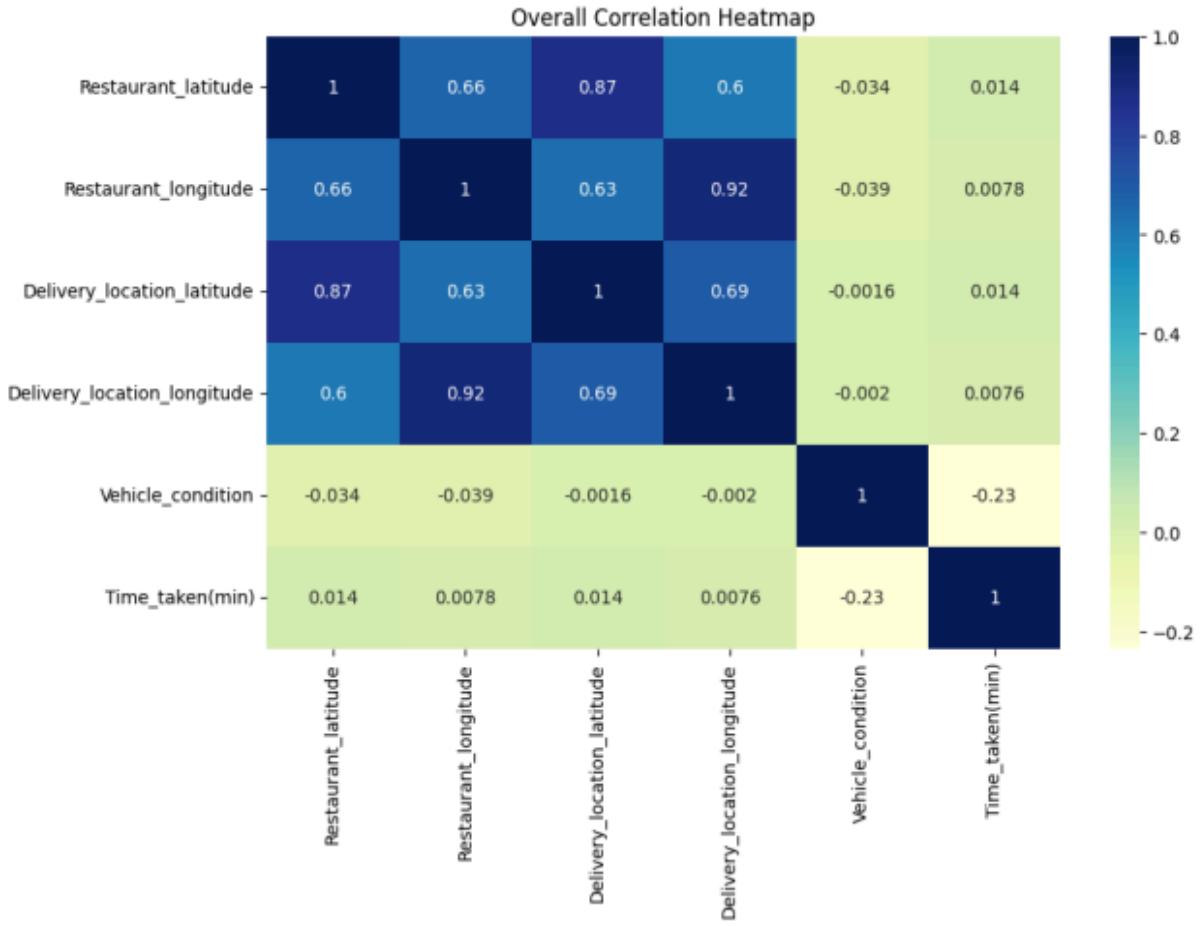
3.2 Bivariate and Multivariate Analysis

- Positive correlation: Delivery_person_Age vs vehicle_condition
- Negative correlation: multiple_deliveries vs vehicle_condition
- Scatter Plot: Delivery_person_ratings vs time_taken(minutes)


Correlation Heatmap
Delivery_person_Age
Delivery_person_Ratings
Vehicle_condition
multiple_deliveries
Time_taken(min)
1
-0.068
0.0033
0.12
0.3
-0.068
1
0.035
-0.12
-0.34
0.0033
0.035
1
-0.097
-0.23
0.12
-0.12
-0.097
1
0.39
0.3
-0.34
-0.23
0.39
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
Delivery_person_Age
ivery_person_Ratings
Vehicle_condition
multiple_deliveries
Time_taken(min)
Activate Windows
Go to Settings to activa


Delivery Ratings vs Time Taken
Type of Order
Snack
Drinks
Buffet
Meal
Time Taken (minutes)
50
40
30
20
10
4.9 4.5 4.4 4.7 4.6 4.8 4.2 4.3 4 4.1 5 3.5 NaN 3.8 3.9 3.7 2.6 2.5 3.6 3.1 2.7 1 3.2 3.3 6 3.4 2.8 2.9 3
Delivery Person Ratings


Vehicle_condition
Time_taken(min)
Vehicle_condition
Time_taken(min)


Overall Correlation Heatmap
Restaurant_latitude
Restaurant_longitude
Delivery_location_latitude
Delivery_location_longitude
Vehicle_condition
Time_taken(min)
1 0.66 0.87 0.6 -0.034 0.014
0.66 1 0.63 0.92 -0.039 0.0078
0.87 0.63 1 0.69 -0.0016 0.014
0.6 0.92 0.69 1 -0.002 0.0076
-0.034 -0.039 -0.0016 -0.002 1 -0.23
0.014 0.0078 0.014 0.0076 -0.23 1

3.3 Insights

- Longer distances and higher road-traffic density significantly increase delivery time, making them the most critical operational factors.
- Urban cities show the shortest average delivery times (~23 minutes), while Semi-Urban areas have the highest (~50 minutes) due to longer routes and less optimized delivery infrastructure.
- Higher-rated and slightly older delivery persons tend to complete deliveries faster, indicating experience and driving efficiency play a key role in reducing delivery duration.

4. Model Building and Comparison

4.1 Models Applied

1. Linear Regression
2. Random Forest
3. Ridge
4. Lasso
5. Gradient Boosting

4.2 Model Implementation

Each model was trained using the same dataset split (80% training, 20% testing).

Libraries used: scikit-learn, SciPy,

Example code:

```
models = {
        'LinearRegression': LinearRegression(),
        'Ridge': Ridge(alpha=1.0),
        'Lasso': Lasso(alpha=0.01, max_iter=10000),
        'RandomForest': RandomForestRegressor(n_estimators=100, random_state=42, n_jobs=-1),
        'GradientBoosting': GradientBoostingRegressor(n_estimators=100, random_state=42)
              }
```

4.3 Model Evaluation Metrics

Metric	Description
MAE	Average of absolute difference between Predicted and actual values
RMSE	Percentage of correctly predicted positives
R2	Represents the proportion of variance in the dependent variable that is predictable from the independent variables.

4.4 Model Comparison

Model	RMSE	MAE	R2
Linear Regression	7.1208	5.7249	0.4245
Random Forest	**4.0331**	**3.1810**	**0.8153**
Ridge	7.1195	5.7231	0.4247
Lasso	7.1184	5.7221	0.4249
Gradient Boosting	4.4776	3.5596	0.7725

4.5 Final Selected Model

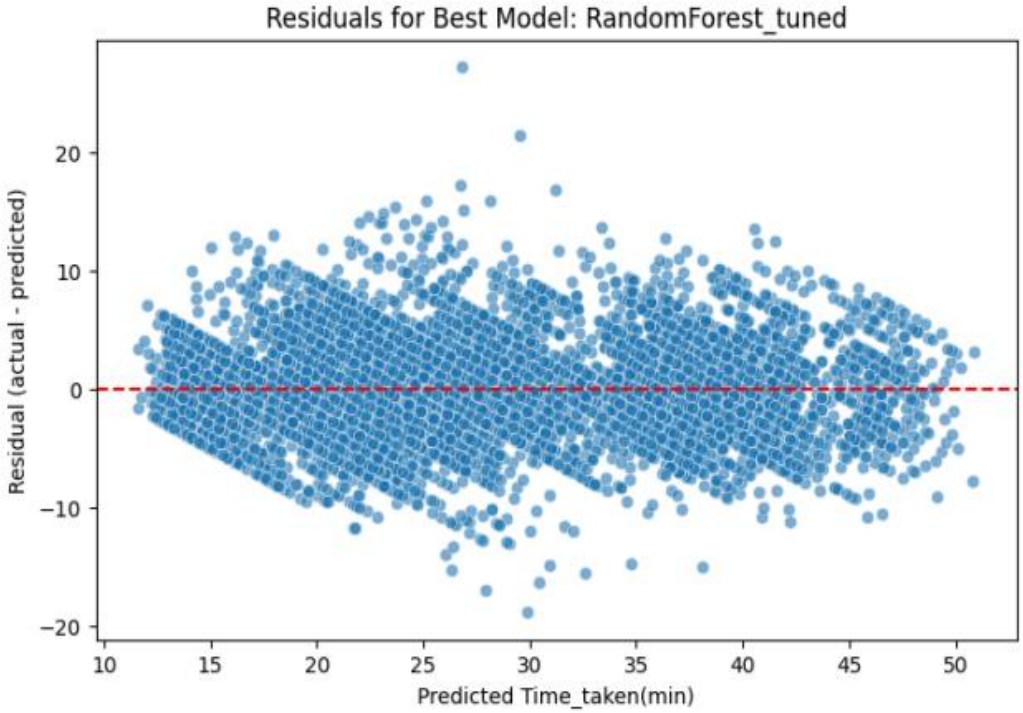
Model: Random Forest

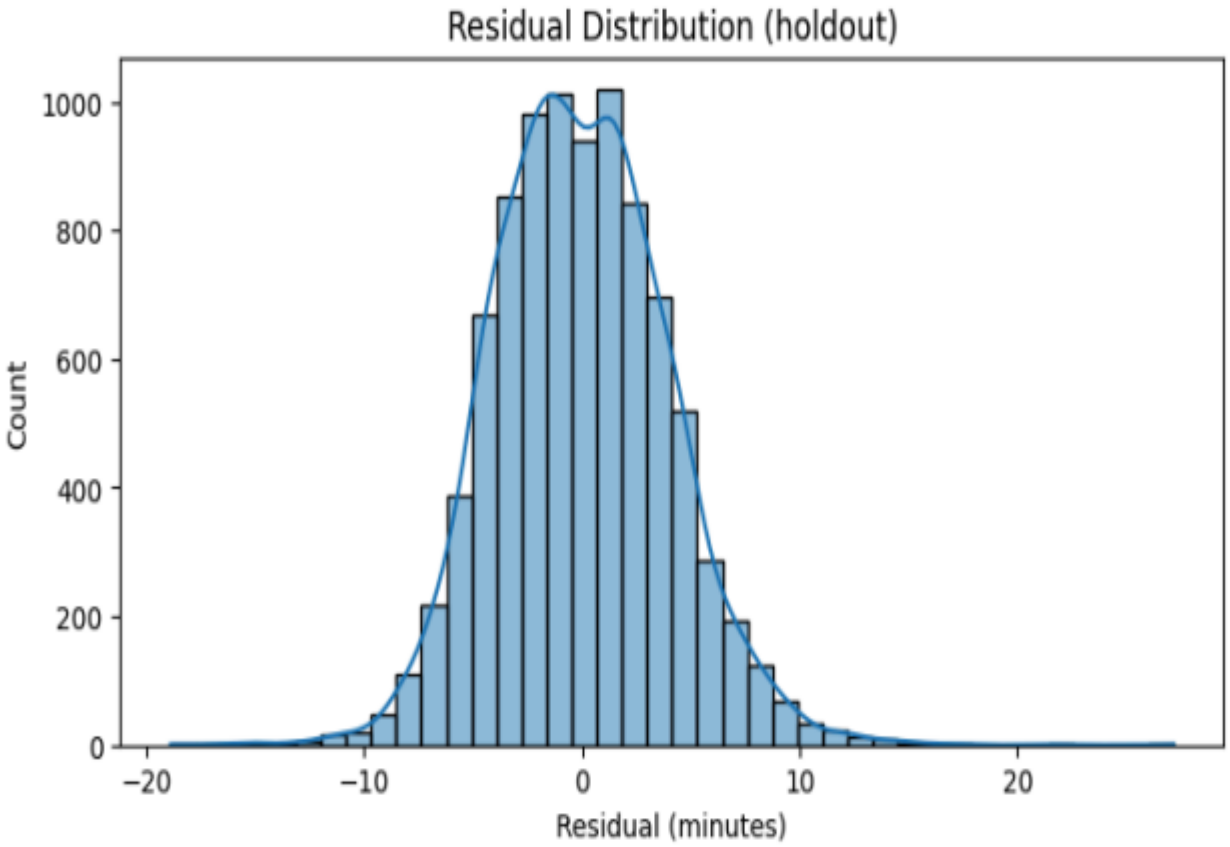
Reason:

- Handles categorical features natively.
- Highest accuracy and balanced precision-recall.
- Better interpretability through feature importance visualization.

Model Summary Statistics:

```
Best Model: RandomForest
Test RMSE: 3.969
Test MAE:  3.162
Test R²:   0.820
```


Residuals for Best Model: RandomForest_tuned
Residual (actual - predicted)
Predicted Time_taken(min)
20
10
0
−10
−20
10
15
20
25
30
35
40
45
50


Residual Distribution (holdout)
Count
Residual (minutes)
1000
800
600
400
200
0
−20
−10
0
10
20

5. Summary and Conclusion

5.1 Key Findings

- Random Forest performed best with an accuracy of 82 %.
- Ensemble averaging in Random Forest reduces overfitting and improves generalization accuracy.
- The model effectively identifies and utilizes key features like distance, city type, and traffic density for better predictions.

5.2 Challenges Faced

- Dealing with imbalanced target variable (few “Yes” outcomes).
- Encoding categorical variables effectively.
- Selecting appropriate hyperparameters.

5.3 Future Enhancements

- Add real-time traffic and weather data.
- Use advanced tuning methods like Grid Search or Bayesian optimization.
- Integrate with dashboard tools (Streamlit, Tableau).

5.4 Conclusion

The project predicted delivery time using different machine learning models, and the **Random Forest model** achieved the best accuracy of **82%**. It effectively captured the impact of factors like distance, traffic, and city type on delivery time. These insights can help improve delivery efficiency, and with future upgrades like real- time data and tuning, the model can perform even better.

GitHub Repository

https://github.com/231801061-Iraiyanbu/AD23532-Principles-of-Data-Science/tree/main/Mini_Project