INNOMATICS RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Smart Physico Risk Monitoring System

Name : jatoth shoban babu
Batch : 436

INNOMATICS RESEARCH LABS

About me

B. Tech in Computer Science . Skilled in Python, SQL, and data analysis with hands-on experience in machine learning projects, EDA, dashboards, and real-world datasets.

Why Data Science:

Passionate about using data to extract insights, make predictions, and solve real-world problems. Data Science blends statistics, programming, and analytics, aligning with my interest in building intelligent, data-driven solutions and pursuing a career as a Data Scientist / ML Engineer.

Introduction

- Physical injuries due to incorrect posture and improper exercises are common.
- Continuous monitoring by physiotherapists is difficult, especially during home rehabilitation.
- Technology can assist physiotherapists by providing real-time posture analysis.

Business & Data Understanding

Swiggy processes millions of food-delivery journeys across hundreds of cities every month. Accurate Estimated Time of Arrival (ETA) at order placement and during delivery is core to the customer experience — it affects conversion, cancellations, ratings, and retention. Today's ETA accuracy suffers from many sources of uncertainty: highly variable traffic patterns, weather ,events, restaurant preparation variability, rider behaviour and skill, order batching/multiple

deliveries, and city-specific operational constraints.

Inaccurate ETAs cause:

- Reduced customer trust and lower repeat orders.
- Increased cancellations and refund costs.
- Higher support volume and operational interventions.
- Sub-optimal rider allocation and increased rider idle or overtime costs

Objective of the Project :

Swiggy's objective is to predict the delivery time (minutes) per order at the moment of order placement (and update it dynamically), with business-grade accuracy and uncertainty estimates so the platform can

(a) show reliable ETAs to customers

(b) optimize rider allocation.

Build a production-ready, scalable Machine Learning system that predicts per-order delivery time (in minutes) using real-time and historical features (order, rider, restaurant, geospatial, traffic, weather, and temporal signals). The system must produce

Delivery Dataset Features Overview

Rider Information

- **Rider_id** - Unique ideir
- **Age** - Rider's age
- **Ratings** - Average customer rating
- **vehicle_condition** - Vehicle condition
- **Type_of_vehicle** - Motorcycle, scooter, etc.

Order Details

- **Type_of_order** - Snack, meal, drinks, buffet
- **multiple_deliveries** - Deliveries per trip
- **pickup_time_minutes** - Restaurant prep time

Order Details

- **Type_of_order** - Snack, meal, drinks, buffet
- **multiple_deliveries** - Deliveries per trip
- **pickup_time_minutes** - Restaurant prep time

Time & Date

- **Order_date, Order_day, Order_month, order_day_of_week**
- **Order_time_hour** - Hour of order (0–23)
- **order_time_of_day** - Morning, afternoon, evening,night
- **Is_weekend** - Indicates weekend order
- **Festival** - Indicates festival day order

Location Data

- **restaurant_longitude**
- **delivery_longitude**
- **Distance** - Distance in km

Time & Date

- **Order_date**, Order_day, Order_month, order_day_of_week
- **Order_time_hour** - Hour of order
- **order_time_of_day** -
- **Festival** - *Indicates festival doy*

City Info

City_type - Urban, HYD, CID, CHEN, etc.

Target Variable

- **Time_taken** - Actual delivrtime

Data Understanding

Category	Details
Total Records	45,502
Total Features	26
Target Variable	time_taken
Numerical Features	13
Categorical Features	10
Identifier	rider_id
Missing Values	Present in age, ratings, traffic, distance, etc.

Missing % Range	Features
~8%	restaurant_latitude, restaurant_longitude, delivery_latitude, delivery_longitude, distance
3–4%	age, ratings, pickup_time_minutes, order_time_hour
2–3%	city_type, multiple_deliveries
<2%	weather, traffic, festival

Descriptive stats

Feature	Key Insight	Interpretation
time_taken	Mean = 26.3 mins	Moderate variation (10–54 mins)
age	Mean = 29.5 yrs	Mostly young riders (20–39 yrs)
ratings	Mean = 4.63	Riders are highly rated
distance	Mean = 9.7 km	Most deliveries under 14 km
pickup_time_minutes	Mean ≈ 10 mins	Stable preparation time
order_time_hour	Mean ≈ 17 (5 PM)	Peak orders in evening
multiple_deliveries	Mean = 0.74	Mostly single deliveries
is_weekend	27% orders	Majority on weekdays
vehicle_condition	Avg ≈ 1	Average vehicle quality
location features	Wide lat-long range	Covers multiple cities

Columns/features data types

Data Type	Features
Discrete (Numerical – Countable)	vehicle_condition, time_taken, order_day, order_month, is_weekend
Continuous (Numerical – Measurable)	age, ratings, restaurant_latitude, restaurant_longitude, delivery_latitude, delivery_longitude, multiple_deliveries, pickup_time_minutes, order_time_hour, distance
Categorical	order_date, weather, traffic, type_of_order, type_of_vehicle, festival, city_type, city_name, order_day_of_week, order_time_of_day

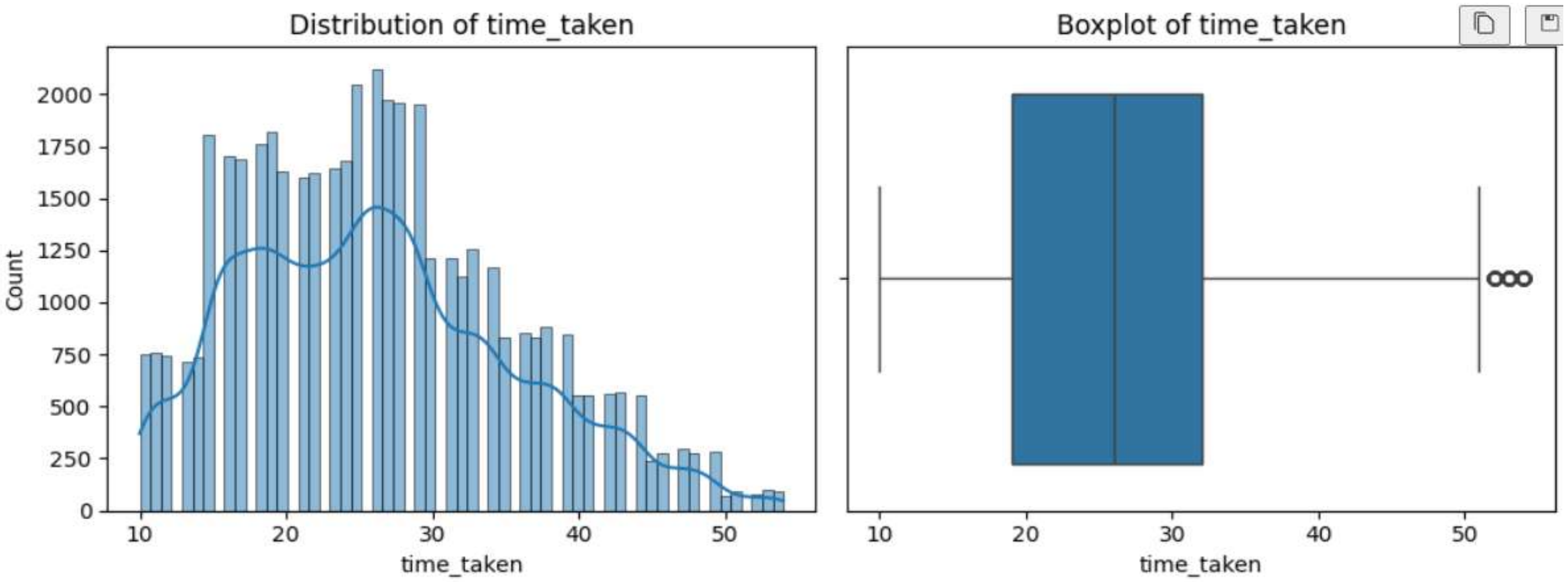
INNOMATICS RESEARCH LABS

Data Cleaning (Missing value handling)

Feature Type	Strategy Used	Reason
Integer Numerical	Median	Robust to outliers
Symmetric Float	Mean	Preserves distribution
Skewed Float	Median	Handles skew & outliers
city_type	Mapping + Mode	Logical inference
weather, traffic, festival	Mode	Categorical feature

INNOMATICS RESEARCH LABS

Exploratory Data Analysis



Key Insights

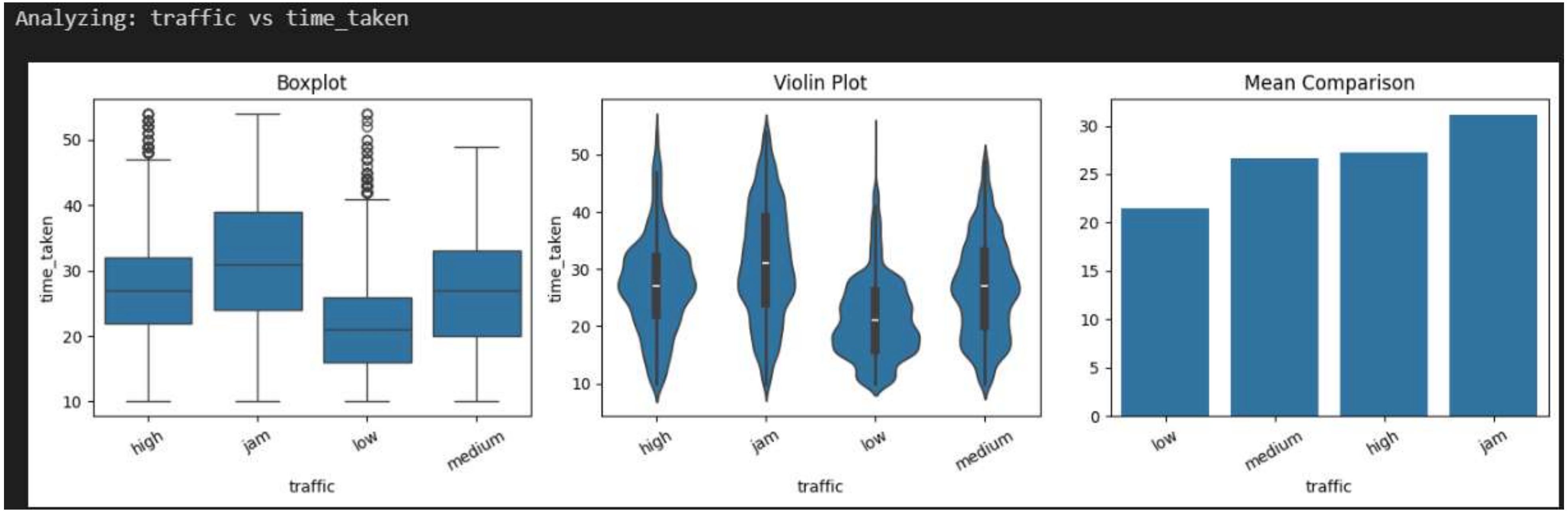
Most deliveries occur between **19–32 minutes**

Mean ≈ Median → distribution fairly balanced

Some extreme delays (up to 54 mins)

Moderate spread (std ≈ 9 mins)

Bivariate Analysis



Strong positive relationship between traffic level and delivery time

Jam traffic significantly increases delay

Traffic is one of the most important predictors (confirmed by feature importance)

Higher traffic → Higher variability + higher average time

Outliers

```
ratings                      Lower IQR cap + Reflect+Log1p
order_time_hour              Lower IQR cap + Reflect+Log1p
age                              no capping | no transform
order_month                      no capping | no transform
pickup_time_minutes              no capping | no transform
vehicle_condition                no capping | no transform
restaurant_latitude              no capping | no transform
delivery_latitude                no capping | no transform
multiple_deliveries              no capping | no transform
order_day                        no capping | no transform
distance                         no capping | no transform
is_weekend                           Upper IQR cap + Log1p
restaurant_longitude                 Upper IQR cap + Log1p
delivery_longitude                   Upper IQR cap + Log1p
```

INNOMATICS RESEARCH LABS

Feature Encoding

Feature	Type	Encoding Method	Category Order
traffic	Ordinal	Ordinal Encoding	low → medium → high → jam
order_day_of_week	Ordinal	Ordinal Encoding	monday → tuesday → wednesday → thursday → friday → saturday → sunday
order_time_of_day	Ordinal	Ordinal Encoding	after_midnight → morning → afternoon → evening → night
city_type	Ordinal	Ordinal Encoding	semi-urban → urban → metropolitan
weather	Nominal	One-Hot Encoding	No order
type_of_order	Nominal	One-Hot Encoding	No order
type_of_vehicle	Nominal	One-Hot Encoding	No order
festival	Nominal	One-Hot Encoding	No order

Feature Selection

Method

Variance Threshold (0.01)

Mutual Information

VIF (Multicollinearity)

Model Feature Importance

Permutation Importance

Final Selected Features

No.	Feature
1	traffic
2	distance
3	ratings
4	age
5	vehicle_condition
6	multiple_deliveries
7	order_time_hour
8	order_day
9	city_type
10	festival
11	pickup_time_minutes
12	weather_sunny
13	weather_stormy
14	weather_windy
15	weather_sandstorms
16	weather_fog

INNOMATICS RESEARCH LABS

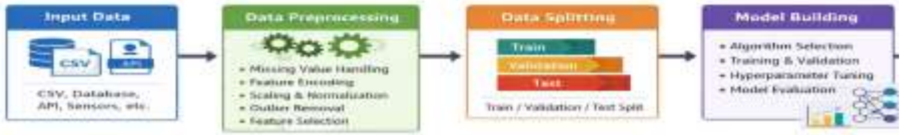
Model Building

```python
from sklearn.ensemble import RandomForestRegressor

rf = RandomForestRegressor(
    min_samples_split=10,
    min_samples_leaf=1,
    max_features='log2',
    n_estimators=300,
    random_state=42,
    n_jobs=1
)

rf.fit(X_train, y_train)

y_pred_rf = rf.predict(X_test)
```



Model Evaluation

	MAE	RMSE	R2
Linear Regression	5.012256	6.314297	0.546727
KNN	3.927935	5.084989	0.706039
Random Forest	3.278211	4.159829	0.803274
tuned Random Forest	3.234162	4.072536	0.811444

Cross validation

Before Tuning

Metric	Value	Interpretation
Cross-Validation R^2 Scores	0.8001 – 0.8040	Very consistent across folds
Average R^2	**0.8003**	Explains ~80% of variance
Stability	High	Very small variation between folds

After Tuning

Metric	Value	Interpretation
Cross-Validation R^2 Scores	0.8084 – 0.8105	Very consistent across folds
Average R^2	**0.8096**	Explains ~81% of variance
Variation Between Folds	~0.002	Extremely low variance → Stable model

Performance improved compared to before tuning (~0.80 → ~0.81)

Very small difference between fold scores → No overfitting

Model generalizes well to unseen data

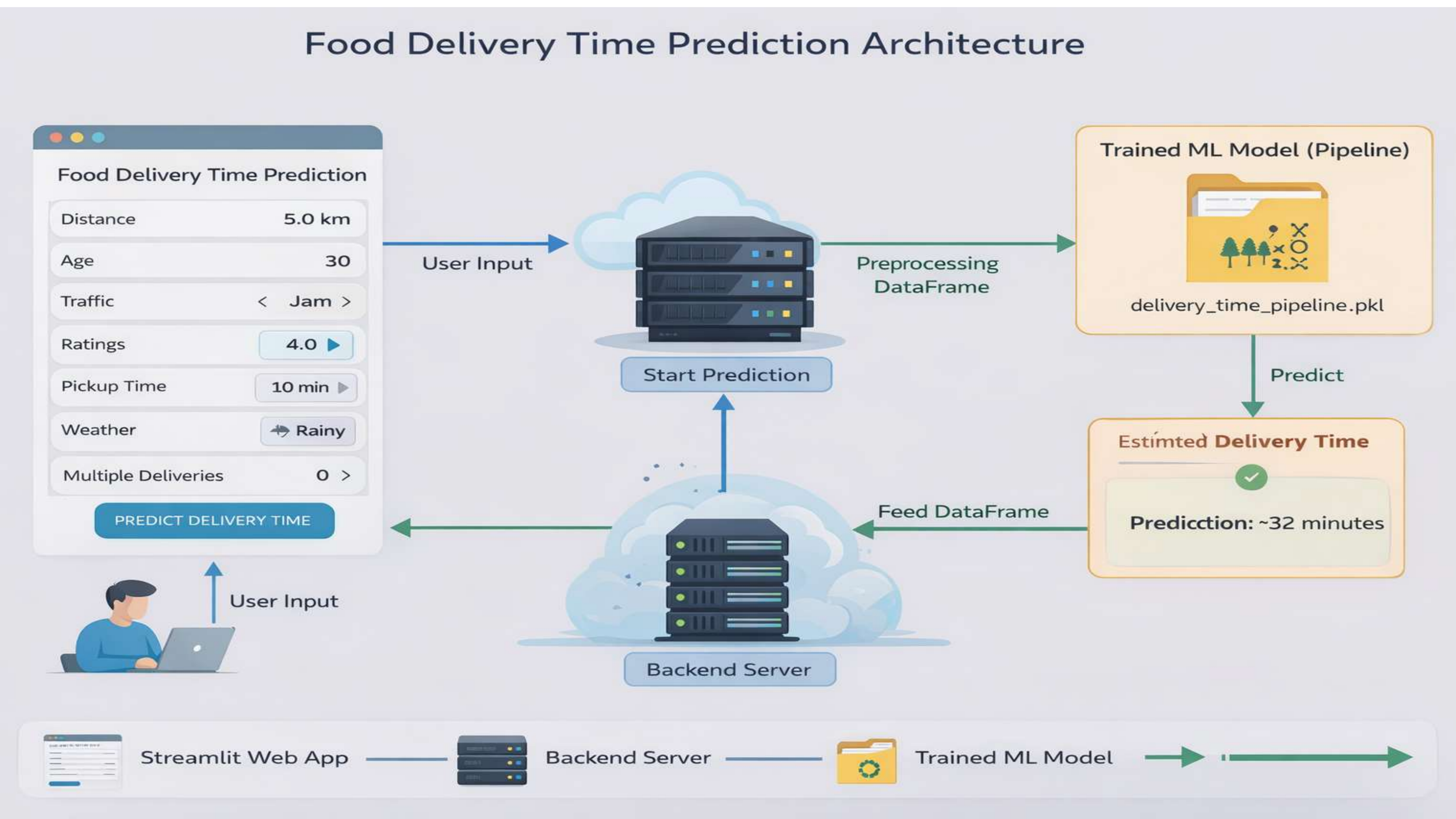
Strong and reliable predictive performance

INNOMATICS RESEARCH LABS

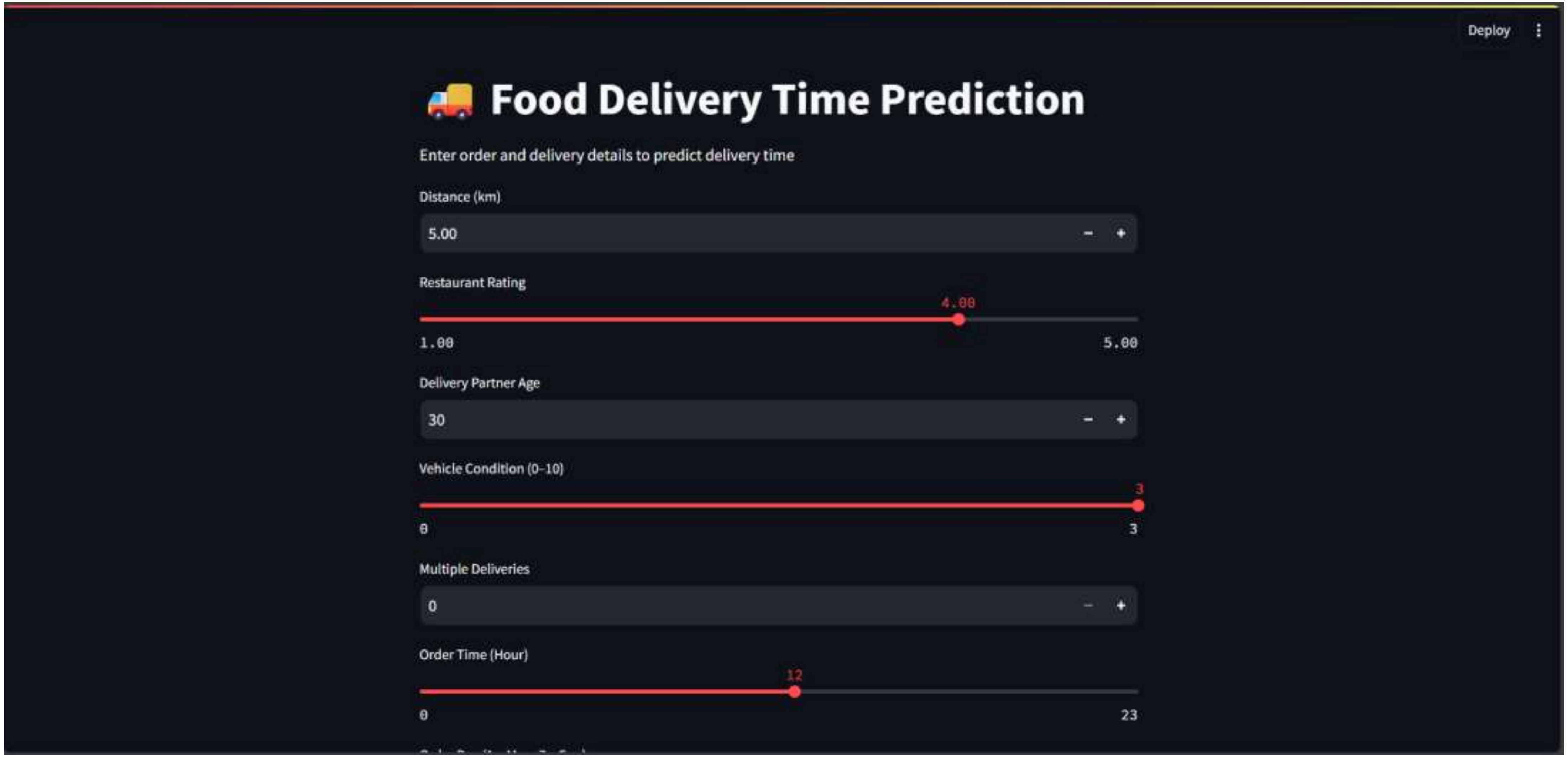
Hyperparameter tuning

Random Search – Best Parameters

Parameter	**Best Value**	**Meaning**
n_estimators	300	Uses 300 trees → More stable predictions
max_depth	None	Trees grow fully (no depth restriction)
min_samples_split	10	Minimum 10 samples required to split a node
min_samples_leaf	1	Leaf node can have minimum 1 sample
max_features	log2	Uses log2(number of features) per split


Food Delivery Time Prediction Architecture
Food Delivery Time Prediction
Distance
5.0 km
Age
30
Traffic
< Jam >
Ratings
4.0
Pickup Time
10 min
Weather
Rainy
Multiple Deliveries
0 >
PREDICT DELIVERY TIME
User Input
User Input
Start Prediction
Preprocessing DataFrame
Trained ML Model (Pipeline)
delivery_time_pipeline.pkl
Predict
Estimted Delivery Time
Predicction: ~32 minutes
Feed DataFrame
Backend Server
Streamlit Web App
Backend Server
Trained ML Model

Web app



INNOMATICS RESEARCH LABS

Applications & Improvements

- Accurate **delivery time estimation** for customers
- Improved **logistics planning & route optimization**
- Better **resource allocation** (riders & vehicles)
- Traffic-aware delivery scheduling
- Enhanced **customer satisfaction & trust**

Improvements

- Use **real-time traffic API data**
- Add weather intensity instead of simple categories
- Try advanced models (XGBoost, LightGBM)
- Deploy as a web/app-based prediction system
- Continuously retrain model with new data

INNOMATICS RESEARCH LABS

Conclusion

This project successfully developed a machine learning model to predict food delivery time with strong performance ($R^2 \approx 0.81$). The model is stable, well-generalized, and shows low prediction error (~3–4 minutes). Traffic and distance were identified as the most important factors influencing delivery time. Overall, the solution is accurate, reliable, and suitable for real-world implementation.

Experience/Challenges

Challenges:

- Handling missing values with different data types
- Managing multicollinearity (latitude, longitude, distance)
- Choosing the right encoding for categorical features
- Feature selection from multiple importance methods
- Hyperparameter tuning without overfitting

Experience Gained:

- Strong understanding of end-to-end ML workflow
- Practical knowledge of preprocessing pipelines
- Experience with feature engineering & selection techniques
- Model evaluation using cross-validation
- Hyperparameter tuning using GridSearch & Random Search

THANK YOU

INNOMATICS RESEARCH LABS