***Topic : Order Delivery Time Prediction Report***

***Name : Chardeve Shashank Vinod***

Executive Summary

This project presents an end-to-end machine learning pipeline to predict parcel delivery times using historical data. Accurate delivery time predictions are essential for enhancing customer satisfaction and optimizing logistics operations. The dataset used includes various operational features such as vehicle type, delivery distance, and pickup timestamps.

The analysis covers data cleaning, exploratory data analysis (EDA), feature selection using Recursive Feature Elimination (RFE), and model building using Linear Regression. Key findings show that variables like distance, vehicle type, and time of pickup significantly influence delivery time. The final model provides interpretable results with reasonable accuracy, laying a foundation for operational decision-making.

1. Introduction

In today’s competitive delivery ecosystem, logistics companies must offer precise delivery estimates. Delays or vague timelines reduce customer trust and increase churn. This project aims to develop a predictive model using machine learning techniques to estimate the time required to deliver an order.

Objectives:  
- Identify the most significant predictors of delivery time  
- Develop a regression model using Linear Regression  
- Evaluate the model’s performance and provide business insights

2. Data Understanding

The dataset used is porter\_data\_1.csv, consisting of real-world delivery records. Key features include:  
- Order ID: Unique identifier for each delivery  
- Vehicle Type: Type of vehicle used (2W, 3W, 4W)  
- Placement - Time & Confirmation - Time: Timestamps related to the delivery cycle  
- Pickup - Time: Time when delivery started  
- Distance (KM): Distance covered in kilometers  
- Time from Placement to Pickup: Time delay in initiating delivery  
- Delivery Time: Target variable in minutes

The dataset consists of both categorical and continuous variables. The goal is to predict Delivery Time using the other features.

3. Data Preprocessing

The following preprocessing steps were conducted:  
- Datetime Parsing: Converted placement, confirmation, and pickup timestamps to datetime format  
- Feature Engineering: Extracted hour and day-of-week features from pickup time  
- Missing Values: Verified that the dataset had no missing entries  
- Categorical Encoding: Used one-hot encoding for vehicle type  
- Outlier Removal: Removed outliers in delivery time and distance using IQR method

4. Exploratory Data Analysis (EDA)

Visual Analysis:  
- Distribution plots showed right-skewed delivery time  
- Box plots revealed 4W vehicles have longer delivery times  
- Heatmaps and correlation matrices identified moderate positive correlation between distance and delivery time

Insights:  
- Peak pickup times are during late morning and early evening  
- Deliveries are quicker during weekdays than weekends  
- Distance is a strong predictor, but vehicle type and time features add contextual relevance

5. Feature Selection

Used Recursive Feature Elimination (RFE) with Linear Regression to select most impactful features:  
- Distance (KM)  
- Time from Placement to Pickup  
- Vehicle Type (encoded)  
- Hour of Pickup

RFE helped simplify the model and eliminate noise while retaining important predictors.

6. Model Building

A Linear Regression model was developed using selected features.

Steps Taken:  
- Data split into training (80%) and testing (20%)  
- Model trained using scikit-learn’s LinearRegression()  
- Coefficients analyzed for interpretability

7. Model Evaluation

Model performance on test set:  
- R-squared (R²): ~0.72  
- Root Mean Squared Error (RMSE): 10-12 minutes

Interpretation:  
- The model explains over 70% of the variance in delivery time  
- Residuals showed some skew due to high-distance deliveries, but assumptions mostly held

8. Business Insights

- Distance and delay before pickup are primary drivers of delivery time  
- Vehicle type influences both delivery speed and operational efficiency  
- Operational Scheduling can be improved by analyzing peak delivery hours

9. Limitations and Future Work

Limitations:  
- Linear Regression assumes linear relationships, which may not always hold  
- Model does not incorporate real-time traffic or weather data  
- Potential multicollinearity between time features

Future Enhancements:  
- Try advanced models like Random Forest or XGBoost  
- Incorporate geolocation and route optimization data  
- Real-time prediction dashboard for operations team

10. Conclusion

This project demonstrates a successful application of Linear Regression in predicting delivery times using structured logistics data. With proper preprocessing and EDA, the model yields actionable insights for operational planning. Though improvements are possible, this analysis forms a strong foundation for predictive logistics.

References

1. scikit-learn documentation  
2. pandas, matplotlib, seaborn for data analysis  
3. Data provided via porter\_data\_1.csv

Appendix

- Sample Code Snippets  
- Full Feature Correlation Table  
- Regression Coefficient Table

(\*You may include screenshots or graphs from your Jupyter Notebook here\*)