Data Scientist Capstone Project Report

Project Overview

This project analyzes the **Supply-to-Line process** in automotive production, focusing on material req

Data Sources

- **Transactional Data**: Daily Excel files containing material request logs.

- **Master Data**: Excel file with PLP characteristics.

- Data merged and stored in SQLite database for analysis.

Problem Statement

The challenge is to predict the time between material requests for PLPs to optimize logistics and reduc

Metrics

- **RMSE (Root Mean Squared Error)**: Measures prediction error in hours.

- **R^2 Score**: Indicates variance explained by the model.

Data Exploration

Two key variables were analyzed:

- **Supply Time**: Duration between request and delivery.

- **Time Between Requests**: Interval between consecutive requests for the same PLP.

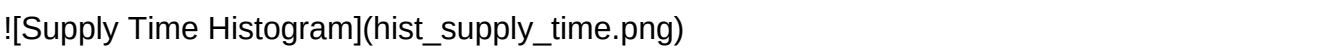
Observations

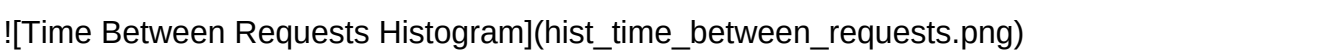
- Supply Time shows high variability across PLPs.

- Time Between Requests varies significantly by PLP and material.

Data Visualization

Below are example histograms illustrating the distribution of Supply Time and Time Between Requests.





Methodology

Data Preprocessing

- Merged transactional and master data.

- Filtered relevant columns and rows.

- Calculated derived features: Supply Time (hours) and Time Between Requests.

- Stored cleaned data in SQLite database.

Implementation

- Built regression pipelines using **RandomForestRegressor** and **GradientBoostingRegressor**.

- Applied **GridSearchCV** for hyperparameter tuning.

Results

Model Performance Comparison

Model	RMSE (hours)	R^2
RandomForest (baseline)	11.94	0.38
RandomForest + GridSearch	10.87	0.44
GradientBoosting	12.10	0.36

Observations

- Tuned RandomForest improved RMSE and R^2 compared to baseline.

- GradientBoosting performed slightly worse than RandomForest.

Conclusion

Predicting time between requests is challenging due to variability across PLPs. Future improvements in

- Collecting more data (beyond 4 weeks).

- Handling outliers (e.g., weekend gaps).

- Grouping PLPs with similar characteristics.

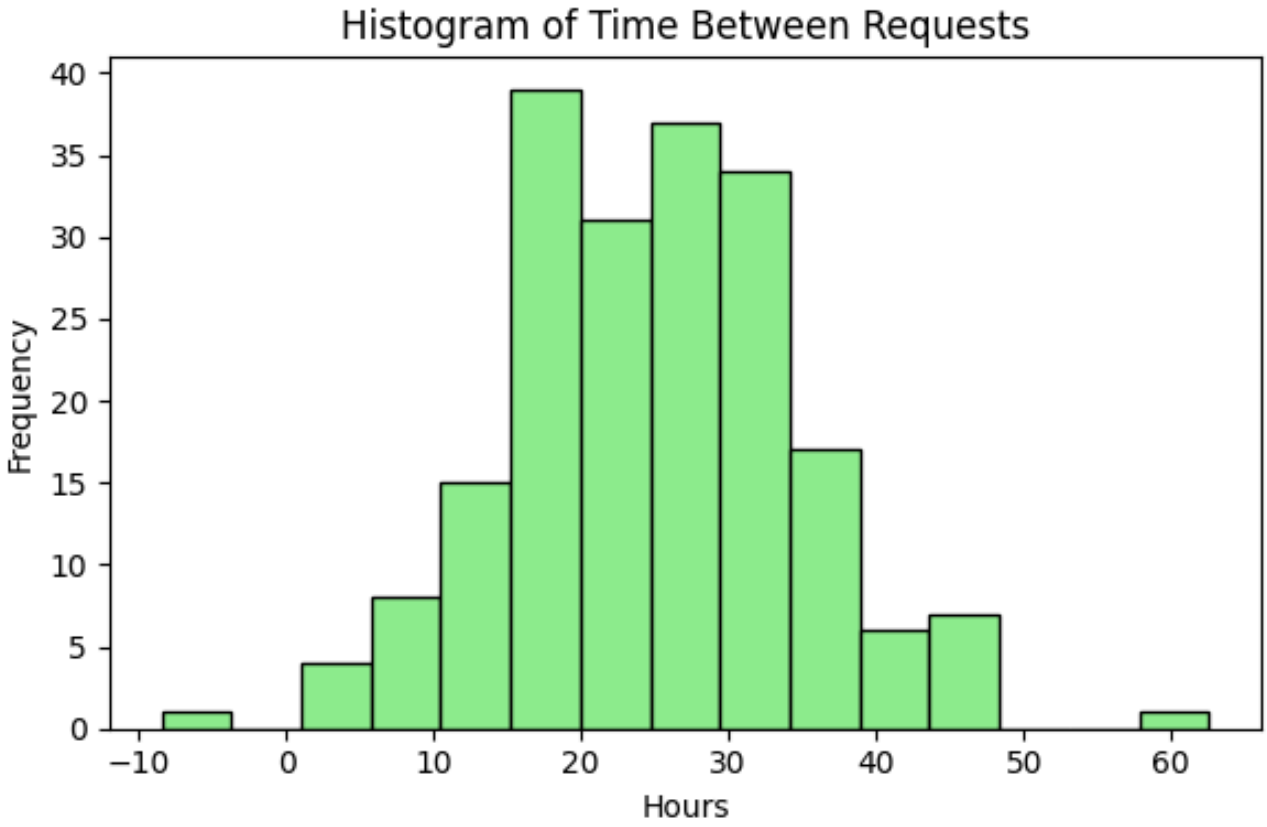
Reflection

The most interesting aspect was the variability in PLP behavior and its impact on prediction accuracy.

Improvement

Improving feature engineering and increasing dataset size could significantly enhance model performar


Histogram of Supply Time
Frequency
Hours
0
5
10
15
20
25
30
0
1
2
3
4
5


Histogram of Time Between Requests
Frequency
Hours
40
35
30
25
20
15
10
5
0
−10
0
10
20
30
40
50
60