Porter Delivery Time Prediction: Machine Learning Analysis

Utkarsh Karambhe
Final Year, B.Tech in Computer Science and Engineering (Data Science)

August 2025

Contents

Abstract

This report presents a machine learning analysis for predicting delivery times at Porter, based on a dataset of 197,428 delivery records from 2015. Following extensive data preprocessing and outlier removal, four algorithms were evaluated, with XGBoost achieving the best performance (Mean Absolute Error: 10.14 minutes, Root Mean Squared Error: 12.78 minutes). The analysis highlights significant improvements in prediction accuracy, offering actionable recommendations for deployment and future enhancements to optimize Porter's delivery operations.

1 Executive Summary

This study develops a robust machine learning framework to predict delivery times for Porter, utilizing a comprehensive dataset of 197,428 delivery records from 2015. After addressing data quality issues, including missing values and outliers, four models—Linear Regression, Random Forest, XGBoost, and TensorFlow Neural Network—were evaluated. XGBoost emerged as the top performer, achieving a Mean Absolute Error (MAE) of 10.14 minutes and a Root Mean Squared Error (RMSE) of 12.78 minutes. Outlier removal significantly enhanced model performance, with RMSE improvements up to 74.95 minutes for Random Forest. The findings provide a foundation for improving delivery time estimates, optimizing resource allocation, and enhancing customer satisfaction.

2 Dataset Overview

2.1 Dataset Characteristics

- Total Records: 197,428 delivery transactions
- Time Period: 2015
- Features: 14 original columns
- Target Variable: Delivery time (difference between actual delivery time and order creation time)

2.2 Column Descriptions

3 Data Preprocessing

3.1 Missing Value Treatment

1. Critical Missing Values: Rows with missing values in market_id, actual_delivery_time, and order_protocol were removed due to their critical role in analysis.
2. Categorical Imputation: Missing values in store_primary_category (2.41%) were imputed with 'Unknown'.
3. Numerical Imputation: Missing values in total_onshift_partners, total_busy_partners, and total_outstanding_orders (8.24%) were imputed using K-Nearest Neighbors (KNN) with $k = 5$.

Table 1: Dataset Columns and Missing Values

Column Name	Description	Missing Values
market_id	Market identifier	987 (0.50%)
created_at	Order creation timestamp	0 (0.00%)
actual_delivery_time	Delivery completion timestamp	7 (0.00%)
store_id	Unique store identifier	0 (0.00%)
store_primary_category	Store cuisine type (70+ categories)	4,760 (2.41%)
order_protocol	Order processing method	995 (0.50%)
total_items	Number of items in order	0 (0.00%)
subtotal	Order subtotal amount	0 (0.00%)
num_distinct_items	Count of unique items	0 (0.00%)
min_item_price	Lowest priced item	0 (0.00%)
max_item_price	Highest priced item	0 (0.00%)
total_onshift_partners	Available delivery partners	16,262 (8.24%)
total_busy_partners	Busy delivery partners	16,262 (8.24%)
total_outstanding_orders	Pending orders count	16,262 (8.24%)

3.2 Data Quality Results

- Final Dataset Size: 195,926 records (99.2% retention)
- Missing Values: 0% across all columns post-preprocessing

4 Exploratory Data Analysis

4.1 Outlier Detection and Treatment

Outliers were identified using the Interquartile Range (IQR) method:

- Q1 (25th percentile): Calculated for delivery times
- Q3 (75th percentile): Calculated for delivery times
- IQR: $Q3 - Q1$
- Upper Bound: $Q3 + 1.5 \times \text{IQR} = 88.32$ minutes

Extreme outliers, including delivery times exceeding 140,000 minutes (97 days), were removed, resulting in 189,708 records for model training.

5 Feature Engineering

5.1 Temporal Features

Derived from created_at timestamp:

- hour: Hour of order placement (0–23)

- days_of_week: Day of week (Monday=0, Sunday=6)
- is_weekend: Binary indicator (1 for Saturday–Sunday, 0 otherwise)

5.2 Categorical Encoding

- Store Categories: 70+ unique cuisine categories
- Encoding Method: One-hot encoding applied to store_primary_category, creating 75 binary columns

5.3 Final Feature Set

- Numerical Features: 9 (temporal and order metrics)
- Categorical Features: 75 one-hot encoded store categories
- Total Features: 84

6 **Model Development**

6.1 Algorithm Selection

Four algorithms were evaluated:

1. Linear Regression (baseline)
2. Random Forest Regressor (100 estimators, max_depth=10)
3. XGBoost Regressor (100 estimators, learning_rate=0.1, max_depth=5)
4. TensorFlow Neural Network (with early stopping)

6.2 Neural Network Architecture

- Input Layer: 84 features
- Hidden Layer 1: 128 neurons (ReLU activation) + 20% Dropout
- Hidden Layer 2: 64 neurons (ReLU activation) + 20% Dropout
- Hidden Layer 3: 32 neurons (ReLU activation)
- Output Layer: 1 neuron (regression)
- Optimizer: Adam
- Loss Function: Mean Squared Error
- Training: 50 epochs with early stopping (patience=10)

Table 2: Model Performance Before Outlier Removal

Model	MAE (minutes)	RMSE (minutes)
Linear Regression	13.00	17.87
Random Forest	12.60	87.99
XGBoost	11.81	16.82
TensorFlow NN	11.90	16.81

7 Performance Analysis

7.1 Model Performance Before Outlier Removal

7.2 Model Performance After Outlier Removal

Table 3: Model Performance After Outlier Removal

Model	MAE (minutes)	RMSE (minutes)	MAE Improvement	RMSE Improvement
Linear Regression	10.64	13.36	2.36	4.51
Random Forest	10.36	13.04	2.24	74.95
XGBoost	10.14	12.78	1.67	4.04
TensorFlow NN	10.27	12.94	1.63	3.87

7.3 Cross-Validation Results (XGBoost)

- 5-Fold CV MAE: 10.11 minutes
- 5-Fold CV RMSE: 12.75 minutes

7.4 Overfitting Analysis

- XGBoost:
  - Train MAE: 10.00 minutes | Test MAE: 10.14 minutes
  - Train RMSE: 12.61 minutes | Test RMSE: 12.78 minutes
  - Generalization Gap: Minimal (0.14 minutes MAE, 0.17 minutes RMSE)
- TensorFlow NN:
  - Train MAE: 10.07 minutes | Test MAE: 10.21 minutes
  - Train RMSE: 12.78 minutes | Test RMSE: 12.94 minutes
  - Generalization Gap: Minimal (0.14 minutes MAE, 0.16 minutes RMSE)

8 Results and Recommendations

8.1 Key Findings

1. Outlier Impact: Removing outliers (>88.32 minutes) significantly improved model performance, with Random Forest showing the largest RMSE reduction (74.95 minutes).
2. Model Ranking:
   - Best: XGBoost (MAE: 10.14, RMSE: 12.78)
   - Second: TensorFlow NN (MAE: 10.27, RMSE: 12.94)
   - Third: Random Forest (MAE: 10.36, RMSE: 13.04)
   - Baseline: Linear Regression (MAE: 10.64, RMSE: 13.36)
3. Generalization: XGBoost and TensorFlow models exhibit minimal overfitting, ensuring robust performance.

8.2 Business Impact

- Prediction Accuracy: 10-minute average error enhances delivery time estimates.
- Operational Benefit: Improves customer satisfaction and resource planning.

8.3 Recommendations

1. Deploy XGBoost Model: Use as the primary prediction engine due to superior performance and efficiency.
2. Real-time Monitoring: Implement data quality checks to detect outliers.
3. Feature Enhancement: Incorporate weather, traffic, and seasonal trends.
4. Model Updating: Schedule regular retraining to maintain accuracy.
5. A/B Testing: Gradually deploy the model with performance monitoring.

9 Conclusion

This machine learning analysis successfully developed a high-accuracy delivery time prediction model for Porter, with XGBoost achieving an MAE of 10.14 minutes. Rigorous data preprocessing, including outlier detection and missing value imputation, was critical to performance. The model's minimal overfitting and strong generalization make it suitable for production deployment, enhancing delivery estimation accuracy and customer experience. Future improvements through additional features and regular retraining will further optimize Porter's logistics operations.