Delivery Time Predictions

SZCZEPAN POLAK

TABLE OF CONTENTS

INTRODUCTION

Accurate delivery time predictions are crucial for efficient route planning, better customer experience, and optimal use of drivers' time.

Currently, our system predicts delivery durations by simply averaging the times of all past deliveries. While straightforward, this method ignores important contextual factors and often results in inaccurate estimates.

The objective of this analysis is to explore historical delivery data to identify patterns or correlations that could lead to more precise predictions. Although we lack direct information about building types, the goal is to uncover other useful indicators—such as sector, weight, or order size—that can serve as proxies or additional features in improving our prediction model.

Hypothesis: Sector-based Predictions

Hypothesis Statement

Deliveries in certain geographic sectors may consistently take more or less time than others. Therefore, predicting delivery times based on sector-specific averages could improve accuracy compared to using a global average.

Methodology to Validate the Hypothesis

To test this idea, the following steps can be taken:

01

Group data

Group historical deliveries by sector_id

02

Calculate average

Calculate the **average actual delivery time** for each sector.

03

Compare predictions

Compare sector-based predictions with the global average prediction by evaluating:

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)

Results and Interpretation

	Global mean	Sectors mean
MSE	12480.67	10794.52
RMSE	111.72	103.90
MAE	87.24	82.10

As we can see from the table, the **sector-based prediction model performed slightly better** across all evaluated metrics. Although the improvement is modest, this simple experiment supports the hypothesis that incorporating geographic sector information into delivery time estimation can enhance prediction accuracy.

Proposed Alternative Prediction Algorithm

There are numerous approaches to improving delivery time predictions. From simple rule-based or statistical models to more advanced machine learning techniques.

The problem at hand is a **classic regression task**, where we aim to predict a continuous target variable (delivery time) based on multiple explanatory features such as:

- Total package weight
- Number of items in the order
- Delivery sector

A variety of regression algorithms could be applied to this task, including:

Algorithm	Description
Linear regression	A simple and fast model that finds a straight-line relationship between features and the target. Best for baseline comparison and data with linear patterns.
Polynomial regression	An extension of linear regression that models curved trends by adding polynomial terms. Suitable when relationships between variables are nonlinear.
Support Vector Regression (SVR)	A robust algorithm that tries to fit the data within a margin of tolerance. It performs well on noisy datasets and can handle outliers better than basic regression models.
Random Forest Regression	A flexible, tree-based ensemble model that captures complex feature interactions. Delivers strong performance on tabular data with minimal preprocessing.

Among these, **Random Forest** is particularly well-suited as a strong baseline model.

It offers solid predictive power out of the box, handles nonlinearities and interactions between variables, and - importantly - is interpretable. With tools such as feature importance scores and SHAP values, we can not only predict delivery times more accurately but also understand the underlying drivers behind those predictions.

Model Building Methodology

To develop the prediction model, historical delivery data was extracted from the database. Only orders with a valid order ID were included to ensure data quality. The dataset was then cleaned by removing duplicate records and correcting clear errors, such as negative delivery durations or unusually long delivery times that extended for several hours.

With access to detailed information from the database, it was possible to calculate and create additional useful features to enrich the model. Ultimately, the model used the following key factors: delivery sector, day of the week, hour of the order, total quantity and weight of products, and the assigned driver. The goal was to predict the total delivery time in seconds.

Because building a predictive model requires learning from past examples, the dataset was split into two parts: one for training the model and one for testing it. Approximately 30% of the data was set aside as a test set to evaluate how well the model performs on new, unseen orders. The model itself was based on a Random Forest approach, using 100 decision trees to capture patterns in the data and make accurate predictions.

Model Validation Strategy

To evaluate how well this model performs, the dataset should be **split into training and test sets**. The training data will be used to train the model, while the test set will serve as unseen data to evaluate its generalization performance.

For each order in the test set, the model will generate a predicted delivery time. These predictions will be compared to the actual times using standard regression evaluation metrics like MSE, RMSE, MAE.

	RF	**RF with driver id**
MSE	8255.93	471.22
RMSE	90.86	21.71
MAE	72.42	14.95

The comparison table clearly shows that using a **Random Forest** model **significantly improved** prediction **accuracy** compared to the simpler, sector-based baseline. Error metrics such as **MSE** and **MAE** are **noticeably lower**, confirming the model's ability to capture complex patterns in the data. Additionally, a second version of the model was tested, which included the driver ID as an input feature. For this specific dataset, incorporating this information led to even better results, suggesting that individual driver performance may play a role in delivery duration.

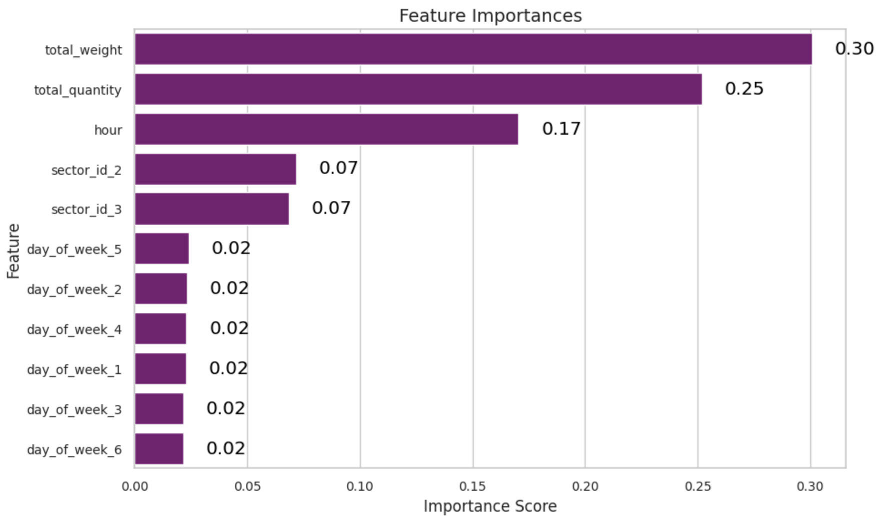
This not only validates the value of richer feature engineering but also highlights how even simple models can bring **meaningful improvements** when supported by thoughtful data representation.

Feature Importance Analysis

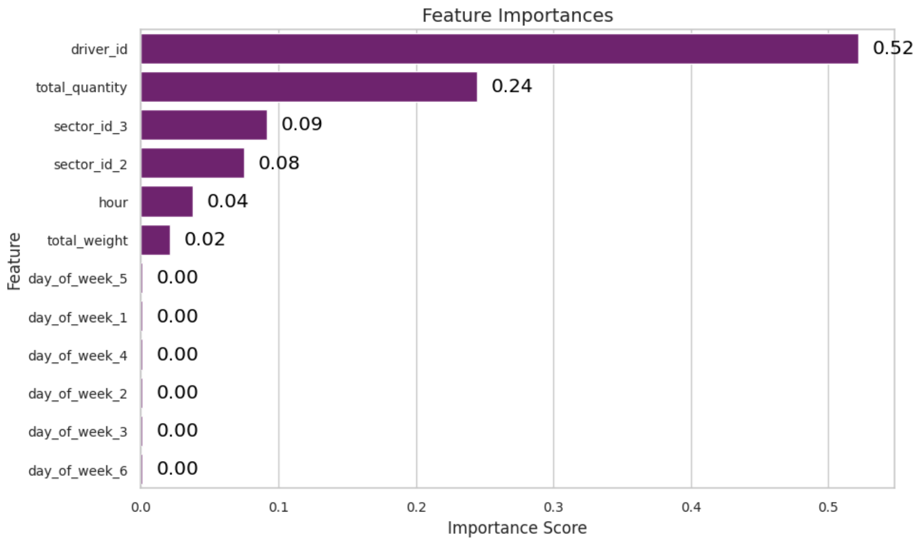
Understanding which features have the strongest influence on delivery time predictions is crucial both for improving model accuracy and for gaining business insights.

After training the Random Forest Regression model, we extracted feature importances to identify the most predictive variables. The model evaluates how much each feature contributes to reducing prediction error, which helps us prioritize relevant factors.

Below, two charts compare feature importance: one excluding the driver ID and one including it. They illustrate how the most relevant variables shift depending on the input features.



The chart highlights the top features impacting delivery time predictions. **Order weight**, **number of products**, and **order hour** stand out as key drivers, showing that both order size and time of day significantly affect delivery duration.



After including the driver ID, the feature importance distribution has shifted significantly.
Driver ID became the dominant factor, reducing the relative impact of other features like order weight and time. This suggests that individual driver behavior strongly influences delivery duration.

By using the Random Forest model, we were able to not only improve prediction accuracy but also identify which features play the most important role in estimating delivery time. This insight significantly supports our understanding of the problem and guides further improvements and data collection efforts.

Hypothesized Reasons for Longer Deliveries

There are several potential factors that could explain why some deliveries take longer than others. Here are a few hypotheses that we believe may influence delivery times:

1. **Worker Fatigue**: As the day progresses, the efficiency of drivers may decrease due to fatigue. This could lead to slower deliveries as drivers become more tired with each order they complete.
2. **Order Size and Weight**: The number of items in an order, combined with their weight, is another likely factor. It is much easier and faster to deliver smaller, lighter packages than larger and heavier ones. Orders with many items or heavy products may require more time for handling, loading, and delivery.
3. **Building Type**: The type of building being delivered to may also play a role. Deliveries to apartment buildings or multi-story complexes can take longer due to factors such as difficulty locating the correct apartment, longer walking distances from parking areas, or the need to use stairs if an elevator is unavailable. In contrast, deliveries to single-family homes are typically quicker due to easier access.
4. **Weather Conditions**: Adverse weather, such as rain or snow, can significantly affect delivery time. Drivers may need to navigate through difficult conditions, slowing down their progress and increasing delivery duration.
5. **Day of the Week**: Although not as obvious, the day of the week could influence delivery times. For instance, weekends may see more traffic or higher demand, leading to potential delays.

Valuable Additional Data to Collect

To enhance the accuracy of delivery time predictions and better understand the factors influencing delivery duration, it would be beneficial to collect additional data points.

By incorporating more detailed and context-specific information, we could develop a more robust model that accounts for real-world variables that are currently not captured. The following data types could be valuable:

- **Building Type**: Understanding the type of building (e.g., apartment complex, single-family home, office building) can offer insights into delivery difficulties, such as accessibility issues or the need to navigate through complex buildings.
- **Presence of an Elevator**: Knowing whether a building has an elevator is crucial. Deliveries to higher floors in buildings without elevators take significantly more time and effort.
- **Weather Conditions**: Weather can have a substantial impact on delivery times. Rain, snow, extreme heat, or cold can slow down the process, especially when delivering to areas with a lot of outdoor or exposed walking routes.
- **Size and Number of Packages**: The complexity of handling the shipment is often determined by the number and size of the items. Large or multiple packages may require more time and effort to transport, especially if they are difficult to carry.
- **Driver Data**: Collecting information about the delivery driver could help in understanding how personal factors, such as experience or age, affect delivery times. Experienced drivers may be faster, and age-related factors may influence delivery speed, especially on multi-floor deliveries.
- **Availability of Parking**: The availability of parking near the delivery address can have a significant impact on the time it takes to complete a delivery. If a driver has to walk a considerable distance to reach the address due to lack of parking, the overall delivery time will increase.

By collecting and incorporating these factors into the model, we can better predict delivery times and optimize the process for more accurate and efficient operations.

Risks of Poor Time Estimation

Inaccurate delivery time predictions can have varying impacts on both customers and the supply chain, depending on whether the prediction overestimates or underestimates the actual delivery time.

1. **Overestimated Delivery Time**
   - Overestimating delivery times typically results in a less severe impact on customer satisfaction. If the actual delivery occurs earlier than expected, customers are likely to be pleased with the faster service. However, this situation can still pose certain challenges. For example, overestimation might lead to suboptimal resource utilization. Drivers could be underused or assigned deliveries that do not match their full capacity, ultimately reducing overall efficiency in the delivery system. Additionally, if the predicted delivery time is visibly displayed to customers during the order process, an overestimation could discourage potential buyers who may be concerned about the longer wait.
2. **Underestimated Delivery Time**
   - Underestimating delivery times tends to result in more significant negative consequences. Customers may experience dissatisfaction due to delays, leading to frustration and a potential loss of trust in the service. From a supply chain perspective, this miscalculation can also cause logistical problems. The system might assign too many deliveries to a single driver, which they may struggle to complete on time. This can lead to increased pressure on delivery personnel, potentially causing burnout, the need for overtime, or, in extreme cases, missed deliveries. In addition, underestimating delivery times can have a cascading effect on the entire logistics operation, making it harder to meet customer expectations and damaging the overall service quality.

In both cases, the accuracy of delivery time predictions is crucial to ensure customer satisfaction and maintain a well-functioning supply chain. Effective forecasting helps optimize resources, improve operational efficiency, and foster a better customer experience.

Summary

This analysis explored various approaches to improving delivery time predictions by examining the current method and proposing alternative models. The Random Forest Regression model was used to identify key features such as order weight, number of products, and order time as significant factors influencing delivery duration. Potential reasons for longer delivery times were discussed, including factors like worker fatigue, order size, building type, and weather conditions. The risks associated with overestimating and underestimating delivery times were also highlighted, emphasizing the importance of accuracy in forecasting to maintain both customer satisfaction and operational efficiency. The findings provide valuable insights for refining delivery time predictions and optimizing the delivery process.

SZCZEPAN POLAK

2025