Delivery Time Forecasting Through Data Analytics and Machine Learning

Tuan Vu

January 25, 2025

1 Introduction

Wolt is renowned for fast deliveries and seamless customer experiences. This project presents a data-driven approach to improving delivery time estimates through comprehensive **data analytics**, **feature engineering**, and **predictive modeling**. The analysis leverages a real-world dataset containing delivery times, geographic coordinates, weather information, and order metadata to uncover patterns and inform model design.

The objective is not only to build accurate prediction models but also to derive actionable insights through exploratory data analysis and performance evaluation across different modeling approaches. This report outlines:

- Data exploration and key insights
- Feature engineering and preparation
- Predictive modeling using:
    1. Random Forest
    2. A basic neural network
    3. An advanced neural network
- Performance evaluation and comparison
- Discussion of results and future improvements

2 Data Exploration and Visualization

This section presents key insights from the dataset through exploratory data analysis (EDA) and visualization. The goal was to uncover patterns in delivery times, explore relationships with weather conditions and geographical distances, and identify potential factors influencing delivery performance.

2.1 Data Description

The dataset (`orders_autumn_2020.csv`) includes the following columns:

- **TIMESTAMP**: Date and time of the order.
- **ITEM_COUNT**: Number of items in the order.
- **USER_LAT**, **USER_LONG**: GPS coordinates of the customer.
- **VENUE_LAT**, **VENUE_LONG**: GPS coordinates of the vendor.
- **ESTIMATED_DELIVERY_MINUTES**: Delivery time estimate (from the dataset).
- **ACTUAL_DELIVERY_MINUTES**: Actual delivery time.
- **CLOUD_COVERAGE**, **TEMPERATURE**, **WIND_SPEED**, **PRECIPITATION**: Weather conditions.

2.2 Initial Cleaning and Outlier Removal

Several steps were taken to clean the data:

- **Dropping nulls:** Rows missing critical information (e.g., lat/long) were dropped.
- **Filtering unrealistic values:** Excluding extreme or negative times (e.g., if `ACTUAL_DELIVERY_MINUTES` ¡ 0).
- **Parsing** `TIMESTAMP`: Converted timestamps to a standard `DateTime` format.

2.3 Overview of Exploratory Analysis

The exploratory phase used SQL for efficient data aggregation and querying, followed by modern visualization techniques to analyze the results. Key areas of focus included:

- Delivery time deviations: Patterns in the difference between actual and estimated delivery times.
- Time-based patterns: Trends based on day of the week and hour of the day.
- Weather impacts: Correlations between weather variables and delivery times.
- Geographical distance: Relationship between user-venue distance and delivery time.

2.4 Delivery Time Deviations

To understand delivery accuracy, we analyzed the deviation:

$$(\texttt{ACTUAL\_DELIVERY\_MINUTES} - \texttt{ESTIMATED\_DELIVERY\_MINUTES})$$

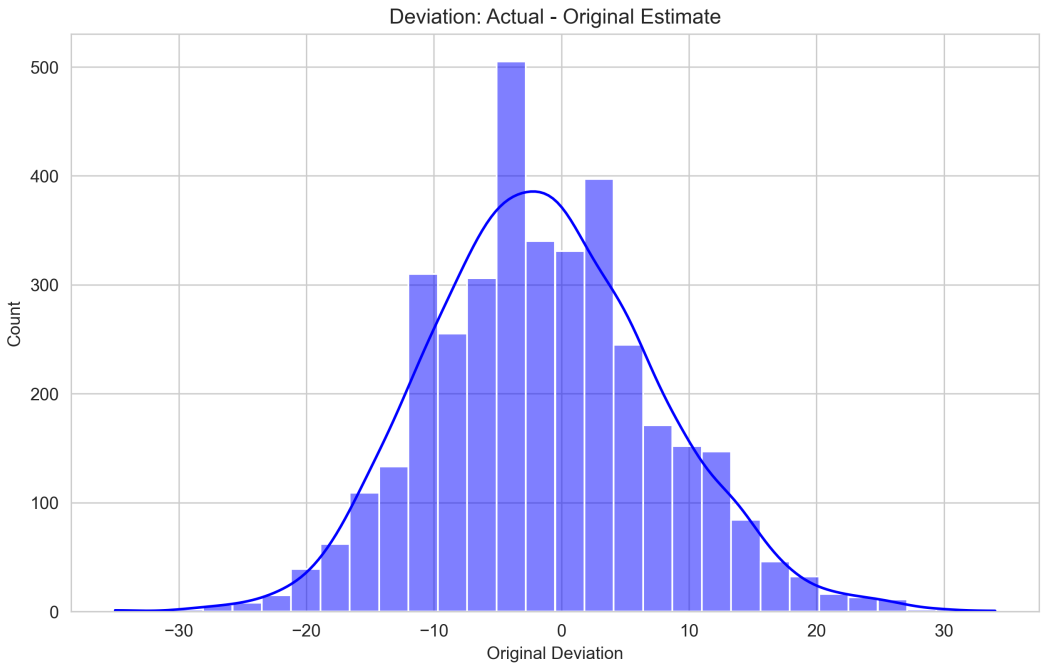


Figure 1: Histogram of delivery time deviations

Key Observations:

- **Left Skew:** The distribution is slightly skewed to the left, indicating more delays than early arrivals.
- **Positive Deviations:** Early deliveries are still notable, reflecting occasional overestimation in the original model.
- **Centered Around Zero:** Errors near 0 dominate, showing the original estimates are often accurate.

2.5 Time-Based Patterns

Aggregating delivery times by day of the week and hour of the day revealed significant trends.

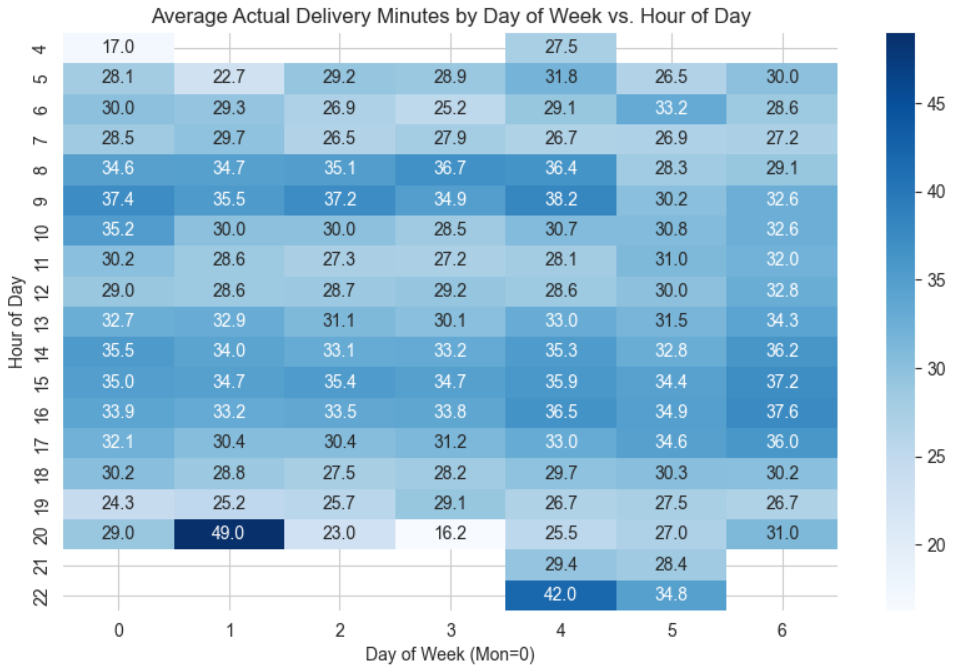


Figure 2: Heatmap of average actual delivery times by day of the week and hour of the day.

Key Observations:

- **Peak Hours:** Deliveries during 8:00–10:00 and 13:00–17:00 take longer on average, reflecting increased activity during these timeframes.
- **Outliers:** Specific days and times occasionally show very high delivery times (e.g., 49 or 42 minutes), but these are rare occurrences.
- **Weekend Patterns:** There is no significant increase in delivery times during weekends compared to weekdays.

2.6 Weather Effects on Delivery Times

Weather variables, including precipitation, temperature, wind speed, and cloud coverage, were analyzed to understand their impact on delivery times.

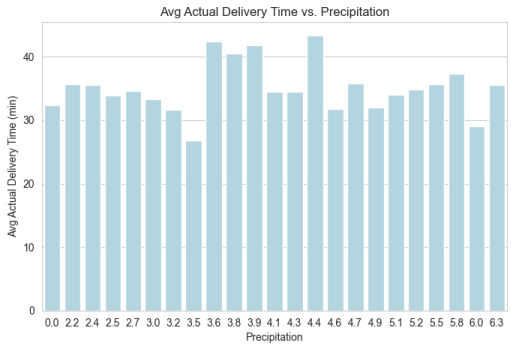


Figure 3: Avg. delivery times vs. precipitation levels.

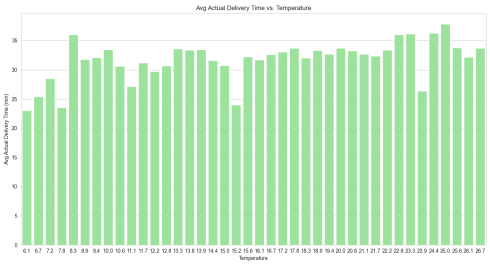


Figure 4: Avg. delivery times vs. temperature.

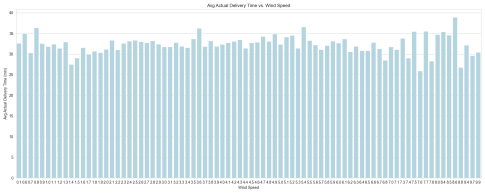


Figure 5: Avg. delivery times vs. wind speed.

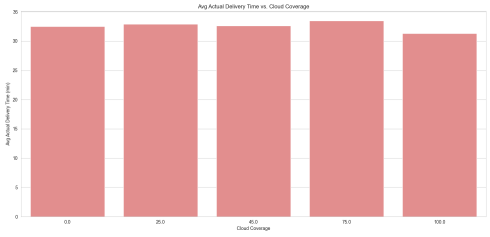


Figure 6: Avg. delivery times vs. cloud coverage.

Key Observations:

- **Precipitation (Figure 3):** Precipitation levels between 3.6 and 4.4 result in higher delivery times, while other ranges show similar times with minimal deviation.
- **Temperature (Figure 4):** Low temperatures are associated with shorter delivery times, while high temperatures do not significantly affect delivery times and remain similar to other ranges.
- **Wind Speed (Figure 5):** While wind speed has some effect on delivery times, the high times are likely influenced by other factors in combination with wind speed.
- **Cloud Coverage (Figure 6):** Cloud coverage shows limited impact on delivery times, with no strong correlation observed.

Conclusion: While weather variables have an impact on delivery times, the effects are most likely due to a combination of factors rather than individual elements.

2.7 3D Weather Analysis

A 3D scatter plot was generated to visualize normalized weather variables—temperature, precipitation, and wind speed—with color representing actual delivery times (Figure 7).

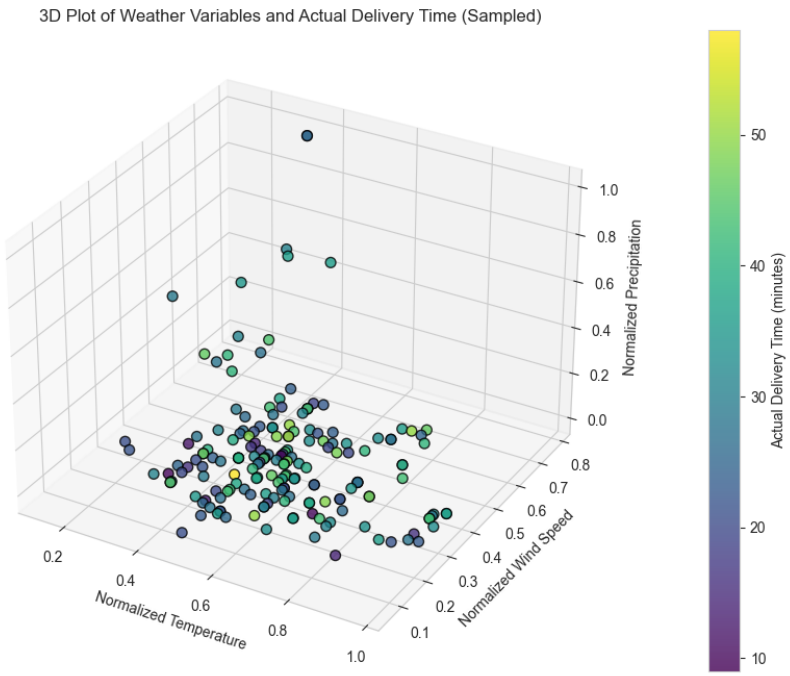


Figure 7: 3D plot of normalized weather variables and actual delivery times (randomly sampled for clarity).

Key Observations:

- **Simplified Visualization:** Cloud coverage was excluded due to its minimal impact, highlighting relationships between key weather variables.
- **Weather's Role:** While weather conditions affect delivery times, they are not the sole determinants.
- **Other Factors:** Significant differences in delivery times under similar weather conditions suggest additional influences like traffic, courier availability, or order complexity.

This analysis underscores the complexity of delivery time prediction and the importance of incorporating non-weather variables for better accuracy.

2.8 Geographical Distance

The distance between users and venues was calculated using the Haversine formula and compared against actual delivery times to assess the relationship.

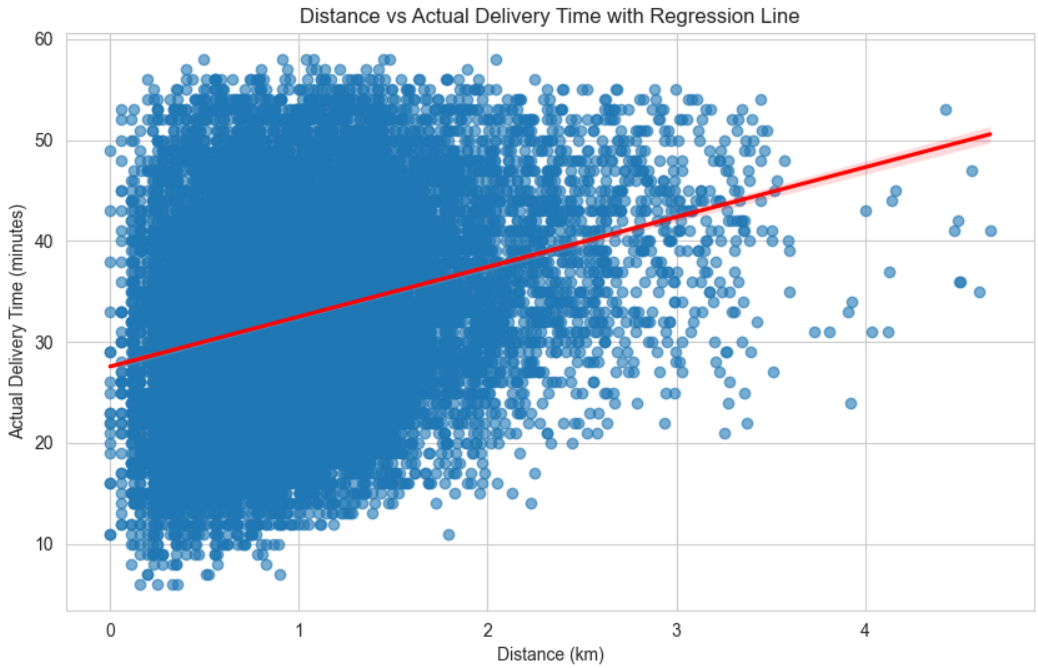


Figure 8: Regression plot: Distance (km) vs. actual delivery time.

Key Observations:

- **Overall Trend:** As indicated by the regression line, longer distances generally lead to longer delivery times. This aligns with the expected physical constraints of delivery logistics.
- **Variability Within Distances:** The data shows significant variability in delivery times for the same distance, suggesting that other factors (e.g., traffic, courier availability, order complexity) influence delivery times as well.
- **Operational Insight:** The correlation between distance and delivery time can inform route optimization strategies, such as assigning closer couriers to reduce variability and improve overall efficiency.

This analysis demonstrates that while distance is a key factor influencing delivery times, it is not the sole determinant, and additional contextual factors must be considered in predictive modeling.

2.9 Correlation Analysis

A correlation matrix was used to identify relationships between key numeric features, such as order size, weather variables, and delivery times.

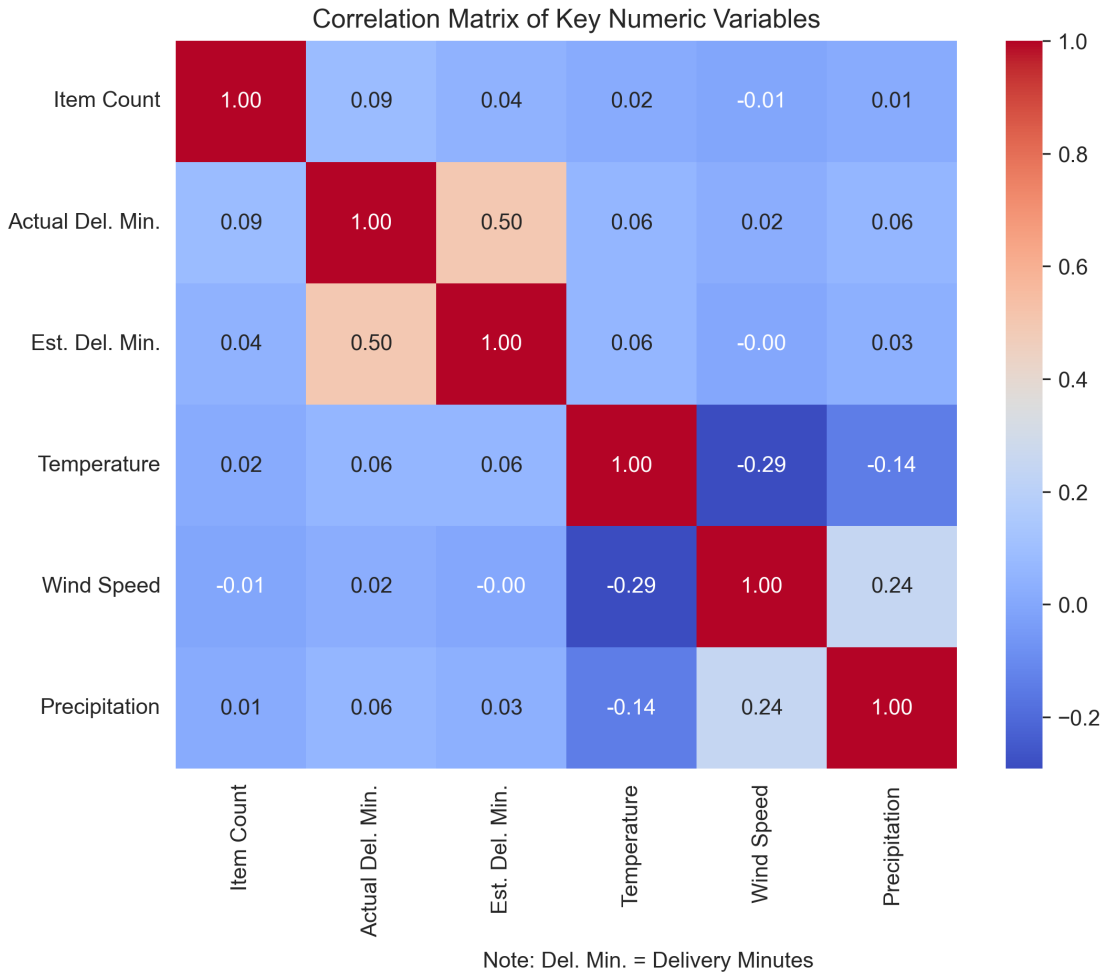


Figure 9: Correlation matrix of key numeric variables.

Key Observations:

- **Actual Delivery Minutes** and **Estimated Delivery Minutes** show a moderate positive correlation (**0.50**), reflecting the relationship between estimated and actual times.
- **Item Count** has a weak positive correlation (**0.09**) with actual delivery times, indicating that larger orders slightly increase delivery durations.
- Weather variables (**Temperature**, **Precipitation**, and **Wind Speed**) individually show weak correlations with delivery times (**0.06** or lower), though their combined effects may still influence variability.
- **Wind Speed** correlates moderately with **Precipitation** (**0.24**), suggesting interdependence among weather factors that could collectively impact deliveries.

3 Feature Engineering

- **Distance Computation**: We used the Haversine formula to compute `DISTANCE_KM` from `USER_LAT`, `USER_LONG`, `VENUE_LAT`, `VENUE_LONG`.
- **Time Features**: Extracted `HOUR_OF_DAY`, `DAY_OF_WEEK`, `MONTH` from the timestamp.
- **Weather Indicators**: Scaled continuous weather variables (e.g., `TEMPERATURE`, `WIND_SPEED`) and introduced a binary `RAIN_INDICATOR` for precipitation.
- **Normalization/Scaling**: Min-max scaling to keep features in comparable ranges.

These engineered features were used as inputs for all models.

4 Modeling Approaches

4.1 Modeling Task

The **primary task** is predicting **Estimated Delivery Minutes** for a new order, with the goal of improving the accuracy of the baseline estimates provided in the dataset. This was approached using machine learning and deep learning techniques.

4.2 Model Choices

Three models were selected to balance interpretability, simplicity, and capacity for complex relationships:

- **Random Forest:** A classical machine learning model that combines predictions from multiple decision trees. Its robustness to noise and ease of interpretation make it a strong baseline for structured data.
- **Basic MLP:** A simple feed-forward neural network with two hidden layers, designed to capture non-linear relationships while remaining computationally efficient.
- **Advanced MLP:** A deeper neural network with three hidden layers (128, 128, 64), providing greater capacity to learn complex patterns in the data, albeit with a higher risk of overfitting.

4.3 Training Procedure

The dataset was split into `80%` training and `20%` testing, and models were trained using the following configurations:

- **Random Forest:** Implemented with Scikit-learn's `RandomForestRegressor`, using `n_estimators=50` for a balance between speed and performance.
- **Basic MLP:** A feed-forward neural network with two hidden layers of 64 neurons each, optimized using Adam (`lr=1e-3`) and trained for 200 epochs. Dropout was applied to prevent overfitting.
- **Advanced MLP:** A deeper architecture with three hidden layers (128, 128, 64), using the same optimizer and training setup as the Basic MLP. This model was designed to explore more complex data relationships.

5 Evaluation and Results

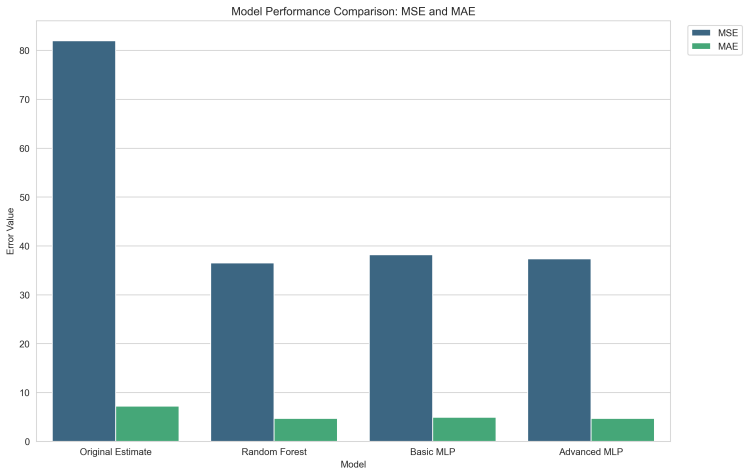
We evaluated the models by comparing their predictions against:

1. The **actual** delivery times (`Actual Delivery Minutes`).
2. The **original** estimates provided in the dataset (`Estimated Delivery Minutes`).

The performance metrics used were:

- **Mean Squared Error (MSE)**: Captures the average squared difference between predicted and actual values.
- **Mean Absolute Error (MAE)**: Measures the average absolute difference between predicted and actual values.

5.1 Global Metrics and Visualization



(a) Visualization of model performance metrics.

Model	MSE	MAE
Original Estimate	81.97	7.23
Random Forest	36.50	4.68
Basic MLP	38.16	4.91
Advanced MLP	37.35	4.67

(b) Comparison of average MSE and MAE across the test set.

Figure 10: Side-by-side comparison of visualization and metrics table.

Key Observations:

- All three models significantly improve performance compared to the original estimates.
- The **Random Forest** achieves the lowest MSE and MAE, performing slightly better than the deep learning models. This may be due to the dataset size being insufficient for the deep learning models to fully leverage their capacity.
- The **Advanced MLP** performs comparably to the Random Forest but is marginally less effective in reducing MSE.

5.2 Comparison by Day of Week

To further analyze performance, we compared the models' predictions to the actual delivery times by grouping data by `Day of Week`.

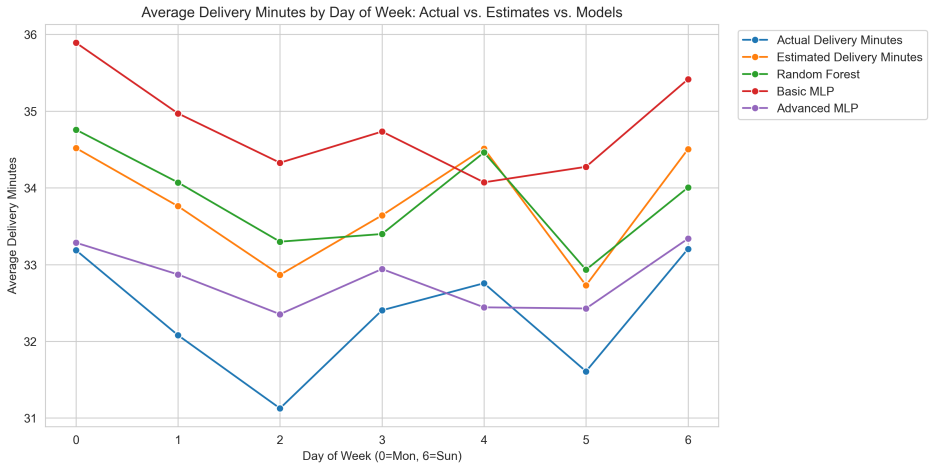


Figure 11: Average predicted vs. actual delivery times by day of week.

Key Observations:

- The **Random Forest** model performs close to the original estimates but shows better accuracy, resulting in improved performance across all days of the week.
- The **Advanced MLP** closely resembles the actual delivery times, particularly on weekends, suggesting that it holds promise with additional data or better feature engineering.

5.3 Deviation Distribution Analysis

The deviation distribution (Actual - Prediction) was analyzed to visualize the prediction errors for each model.

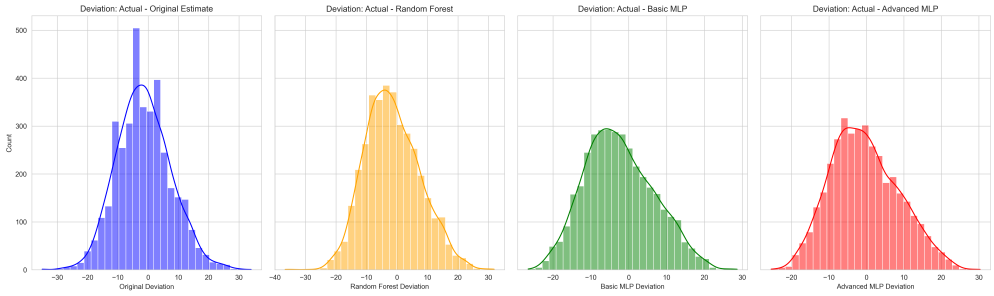


Figure 12: Deviation distributions for all models.

Key Observations:

- The **Random Forest** model produces the most tightly centered deviations around zero, indicating consistent and accurate predictions.
- The **Advanced MLP** demonstrates a slightly broader distribution, reflecting potential for improvement with more training data or refined hyperparameters.

6 Discussion and Conclusion

The results demonstrate that all proposed models outperform the baseline, with the Random Forest achieving the best overall performance on the current dataset. The Advanced MLP model shows strong potential, closely resembling actual delivery times, particularly on weekends, indicating promise with additional data and refined hyperparameters. Despite these improvements, several opportunities remain for further enhancements:

- **Hyperparameter Optimization**: Advanced tuning (e.g., grid search, Bayesian optimization) for learning rate, batch size, and model architecture can refine model performance.
- **Sequential Modeling**: Exploring time-series models such as LSTMs or Transformers could capture temporal dependencies and improve predictions.
- **Feature Enrichment**: Adding contextual features like traffic patterns, holiday indicators, or courier-specific factors can provide more insight.
- **Deployment Considerations**: For production use, considerations such as real-time data streaming, system latency, and scalability must be addressed.

In conclusion, this project demonstrates the potential of machine learning and deep learning to enhance delivery time predictions. By addressing the outlined next steps, Wolt could deploy more accurate, data-driven models at scale, further improving operational efficiency and customer satisfaction.

Please reach out if you'd like to discuss any aspect of the project.