MarketPlace -
Exploratory Data Analysis & Benchmark Models

Initial Alerts Detected

carrier_to_customer is highly overall correlated with purchase_to_customer	High correlation
customer_state is highly overall correlated with geolocation_lat and 1 other fields	High correlation
freight_component is highly overall correlated with price	High correlation
geolocation_lat is highly overall correlated with customer_state	High correlation
geolocation_lng is highly overall correlated with customer_state	High correlation
price is highly overall correlated with freight_component and 1 other fields	High correlation
product_height_cm is highly overall correlated with product_weight_g and 1 other fields	High correlation
product_length_cm is highly overall correlated with product_weight_g and 2 other fields	High correlation
product_weight_g is highly overall correlated with price and 4 other fields	High correlation
product_width_cm is highly overall correlated with product_length_cm and 2 other fields	High correlation
purchase_to_customer is highly overall correlated with carrier_to_customer	High correlation
total_size is highly overall correlated with product_height_cm and 3 other fields	High correlation

These high correlation alerts make sense.

order_status is highly imbalanced (91.1%)	Imbalance
order_delivered_carrier_date has 1713 (1.9%) missing values	Missing
order_delivered_customer_date has 2807 (3.1%) missing values	Missing
approved_to_carrier has 1728 (1.9%) missing values	Missing
carrier_to_customer has 2808 (3.1%) missing values	Missing
purchase_to_customer has 2807 (3.1%) missing values	Missing
purchase_to_approved is highly skewed (γ1 = 55.11416993)	Skewed
purchase_to_approved has 1473 (1.6%) zeros	Zeros ✅

Missing values in order lifecycle were discussed and it should be investigated further

The skewness corresponds exactly to what the product manager has raised.

Current Estimation Method

- Significant **overestimation** for fast deliveries
- **Underestimation** for slow deliveries
- Pearson Correlation: 0.37
- **Unusually long** estimations and actual deliveries can be observed

Default Estimation - Evaluation

Model: Default Estimation Mechanism

Imputation: Unavailable

Feature Space: Unavailable

Status: Integrated to product (?)

Mean Absolute Error: 12.7 Days
Median Absolute Error: 12.2 Days

Benchmark Model - Evaluation

Model: Linear Regression

Imputation: mean

Feature Space (11 features):

- `approval_time`
- `approved_to_carrier`
- `price`
- `freight_value`
- `freight_component`
- `product_weight_g`
- `product_length_cm`
- `product_height_cm`
- `product_width_cm`
- `total_size`
- `inter_state`

Status: Research

Mean Absolute Error: 5.8 Days
Median Absolute Error: 4.4 Days

Benchmark VS Default Estimation Mechanism

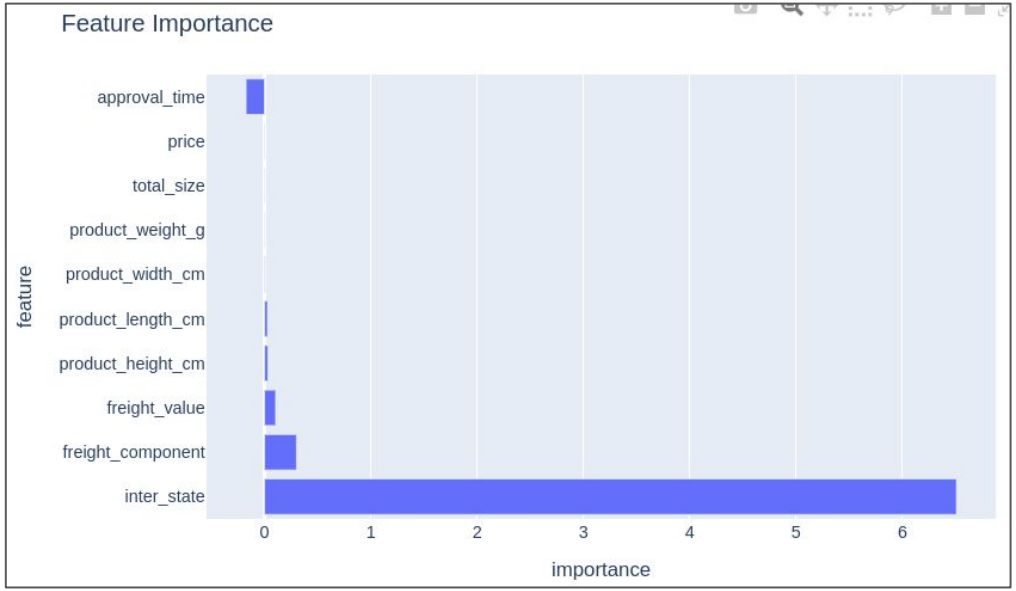
Mean Absolute Error: 12.7 Days
Median Absolute Error: 12.2 Days

Mean Absolute Error: 5.8 Days
Median Absolute Error: 4.4 Days

Highlights

- Major accuracy improvement - **over X2 reduction** in errors, ~6 days more accurate
- Model is very slim - **fast** training and inference
- Whitebox model - **easy to interpret**

Benchmark Model - Feature Importance



`Freight` value and `Inter-State` transports are the strongest predictors for delivery time (linear predictive power, we did not check for non-linearities yet).

Hence, let's try to **enhance the physical distance** element to improve the model.

Improvement #1

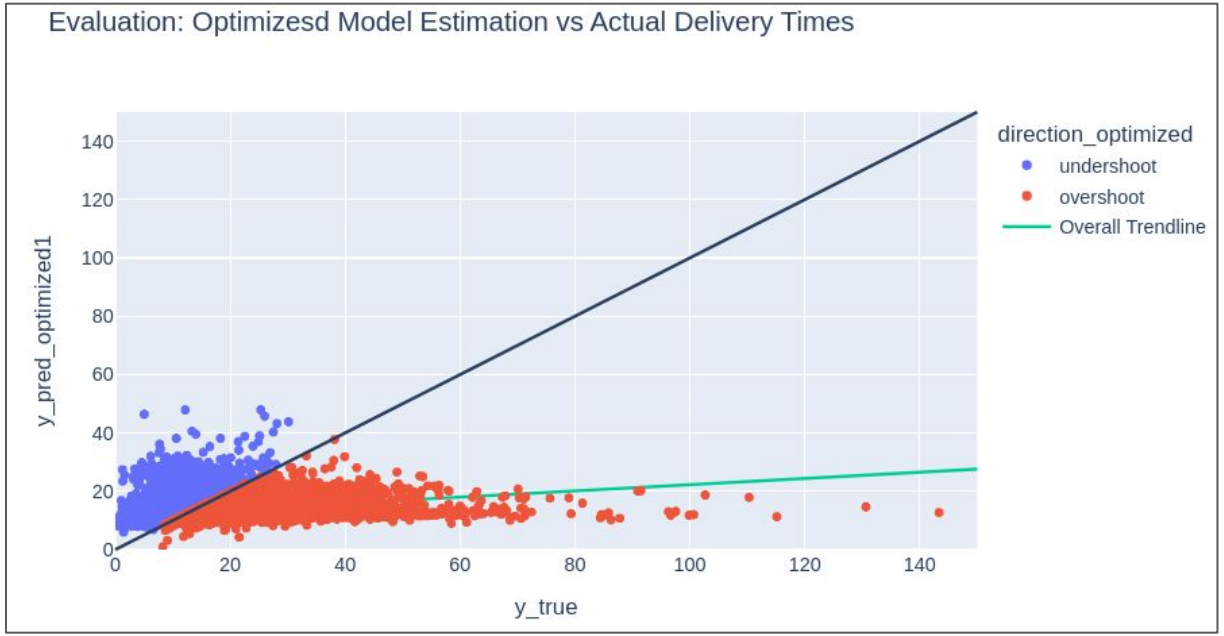
Model: Linear Regression

Imputation: mean

Feature Space (12 features):

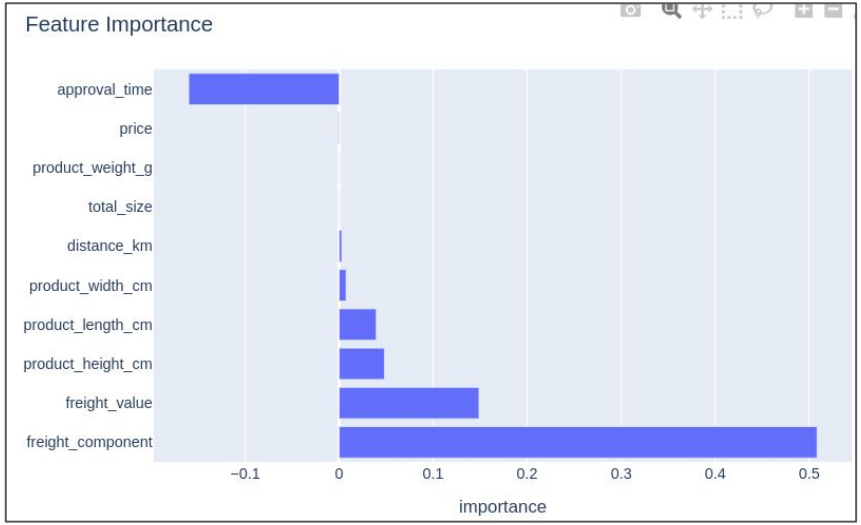
- `approval_time`
- `approved_to_carrier`
- `price`
- `freight_value`
- `freight_component`
- `product_weight_g`
- `product_length_cm`
- `product_height_cm`
- `product_width_cm`
- `total_size`
- `distance`

Status: Research



Mean Absolute Error: 6 Days
Median Absolute Error: 4.6 Days

Optimized Model - Feature Importance



Interestingly, the model performance **degraded** a little bit, and the `distance_km` feature hardly adds any linear predictive power to the model.
It looks like `inter_state`, even though it was a boolean feature, had a significant impact of the model but trying to decipher it using distance in KM might cause **overfitting**.

Future Directions

- Target encoding of sellers
- Better understanding of the freight process and which additional data can be collected
- Optimizing the feature-space with feature selection.
- Non-linear model like CatBoost
- Integrate temporal events like bad weather and holidays data.
- Evaluate benchmark model on test set and consider silent mode