Time-series Sales Prediction Model

Business context can goals

- **Helps solve business problems and decision making:**
    - Demand planning: Plan inventory & stock, prepare resources to meet sudden spikes
    - Improve marketing strategies: Focus on which aspects to drive sales
- **What decisions will it support?**
    - Adjusting promotions,
    - Managing stock levels,
    - Planning resources,
    - Phasing ad spend, marketing activities,
    - …

Features ['date', 'traffic', 'impressions', 'product_ad_spend', 'shop_ad_spend', 'product_page_bounce_count',
'traffic_from_search', 'run_product_ad', 'run_shop_ad', 'wm_yr_wk', 'wday', 'month', 'doubleday',
'near_dday', 'end_of_month', 'weekend', 'other_commercial_sale', 'day_offs', 'day_of_year',
'week_of_month', 'est_avg_price', 'avg_price', 'promotion_on', 'promotion_price', 'discount_rate',
'comment_received', 'product_rating', 'avg_category_comment', 'avg_category_rate', 'high_rating',
'high_comment', 'high_discount', 'wday_sin', 'wday_cos', 'month_sin', 'month_cos', 'wom_sin', 'wom_cos',
'day_of_year_sin', 'day_of_year_cos', 'payment_lag_3', 'payment_lag_28', 'product_ad_spend_lag_3',
'product_ad_spend_lag_28', 'shop_ad_spend_lag_3', 'shop_ad_spend_lag_28', 'traffic_rolling_7d_mean',
'impressions_rolling_7d_mean', 'payment_rolling_7d_mean', 'product_ad_spend_rolling_7d_mean',
'shop_ad_spend_rolling_7d_mean', 'traffic_rolling_28d_mean', 'impressions_rolling_28d_mean',
'payment_rolling_28d_mean', 'product_ad_spend_rolling_28d_mean', 'shop_ad_spend_rolling_28d_mean',
'cat_daily_avg_product_ad_spend', 'cat_daily_avg_traffic', 'cat_month_avg_product_ad_spend',
'cat_month_avg_traffic', 'cat_week_avg_product_ad_spend', 'cat_week_avg_traffic',
'continuous_zero_sales_days', 'zero_sales_rolling_mean_14d', 'zero_sales_rolling_max_14d',
'cat_rolling_avg_zero_sales_14d', 'cat_zero_sales_x_promotion', 'cat_rolling_avg_zero_sales_trend_14d',
'cat_rolling_avg_zero_sales_pct_change_14d', 'cat_rolling_avg_zero_sales_lag_7d',
'cat_rolling_avg_zero_sales_lag_14d', 'cat_rolling_avg_zero_sales_lag_28d', 'normalized_avg_price',
'normalized_promotion_price', 'ad_spend_during_promotion', 'weekend_ad_spend', 'CTR', 'product_category',
'brand', 'price_bin', 'product_id']

2 main base feature groups: Time

Group	Features
Present	**Categorical features:** `'date'`, `'wday'`, `'month'`, `'day_of_year'`, `'week_of_month'`,etc. **One-hot encoding:** `'doubleday'`, `'near_dday'`, etc.
Periodicity and cyclic patterns of time	`'wday_sin'`, `'wday_cos'`, `'month_sin'`, `'month_cos'`, `'wom_sin'`, etc.
Past	**Lagging features:** `'payment_lag_3'`, `'payment_lag_28'`, `'product_ad_spend_lag_3'`, `'product_ad_spend_lag_28'`, etc. **Moving features:** `'traffic_rolling_7d_mean'`, `'impressions_rolling_7d_mean'`, `'payment_rolling_7d_mean'`, `'product_ad_spend_rolling_7d_mean'`

2 main base feature groups: Product

Group	Features
Product level	**Categorical features**: 'product_id' **Numeric features**: 'traffic', 'impressions','payment', 'avg_price', 'product_rating', 'comment_received', etc.
Product category level	**Categorical features**: 'product_category' **Numeric features**: 'cat_daily_avg_product_ad_spend', 'cat_daily_avg_traffic', 'cat_month_avg_product_ad_spend', 'cat_month_avg_traffic', etc.
Other dimensions	'brand'

Target

`'payment'`

Number of successfully processed payments.

Dataset

Source: Internal eCommerce data

Data range: Since 1st May, 2024

Caveats

- Small data size
- Only use internal data
- Still is a work in progress:
    - Build multiple models to compare between models
    - Optimize feature engineering
    - Gathering more data
    - Tailor to dataset's inherent natures: For example: Sporadic sales patterns of high-ticket products.
   - …

Evaluation metric

RMSE (Root Mean Squared Error):

- MSE calculates the average size of the errors (differences) between the predicted and actual sales, emphasizing larger errors because it squares the differences before averaging.
- Finally, it takes the square root, so the result is in the same unit as the target variable (e.g., number of payments).
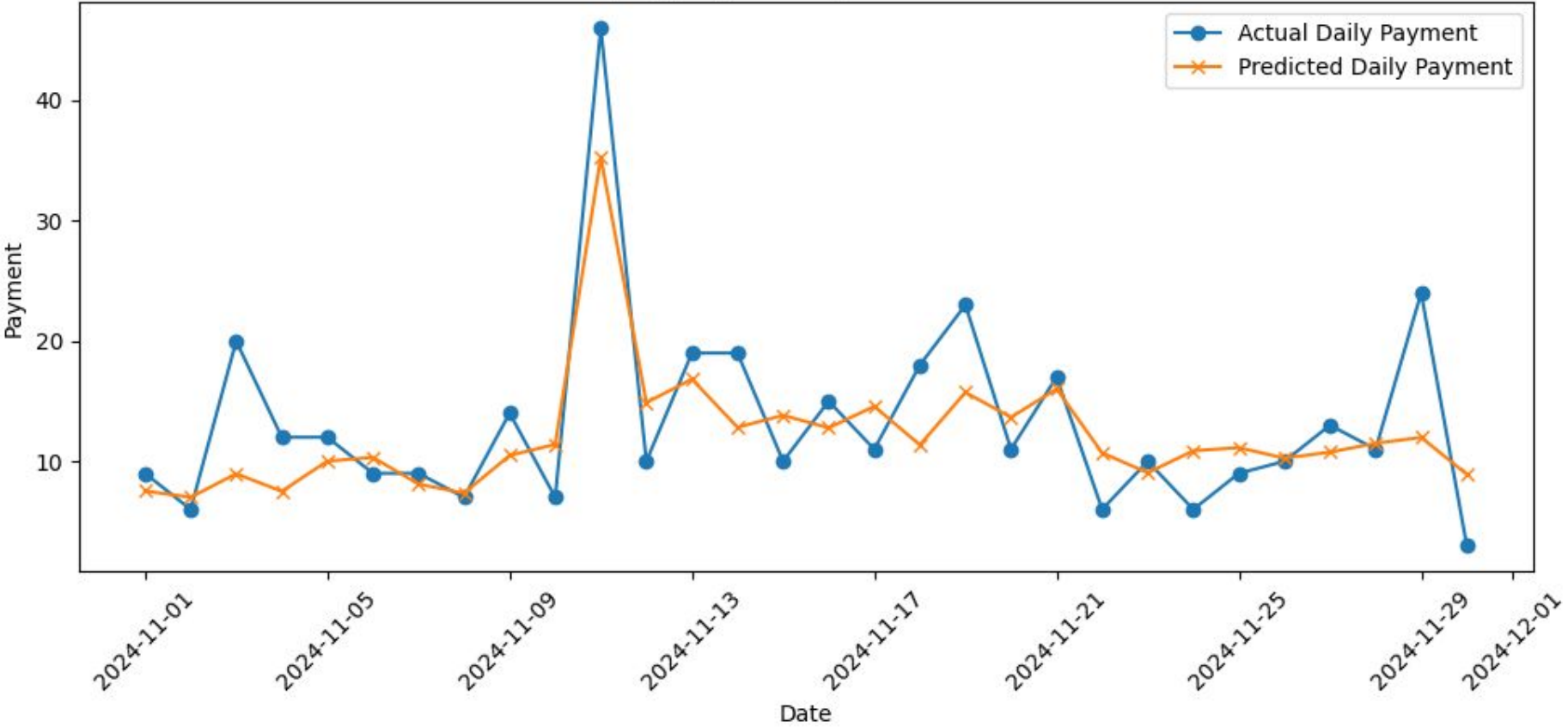
Modeling

XGBoost ("Extreme Gradient Boosting") is a powerful machine learning algorithm used for both regression and classification tasks.

- Imagine you're guessing sales for tomorrow. You might start with a rough estimate.
- XGBoost creates a series of "decision trees" that refine this estimate step by step.
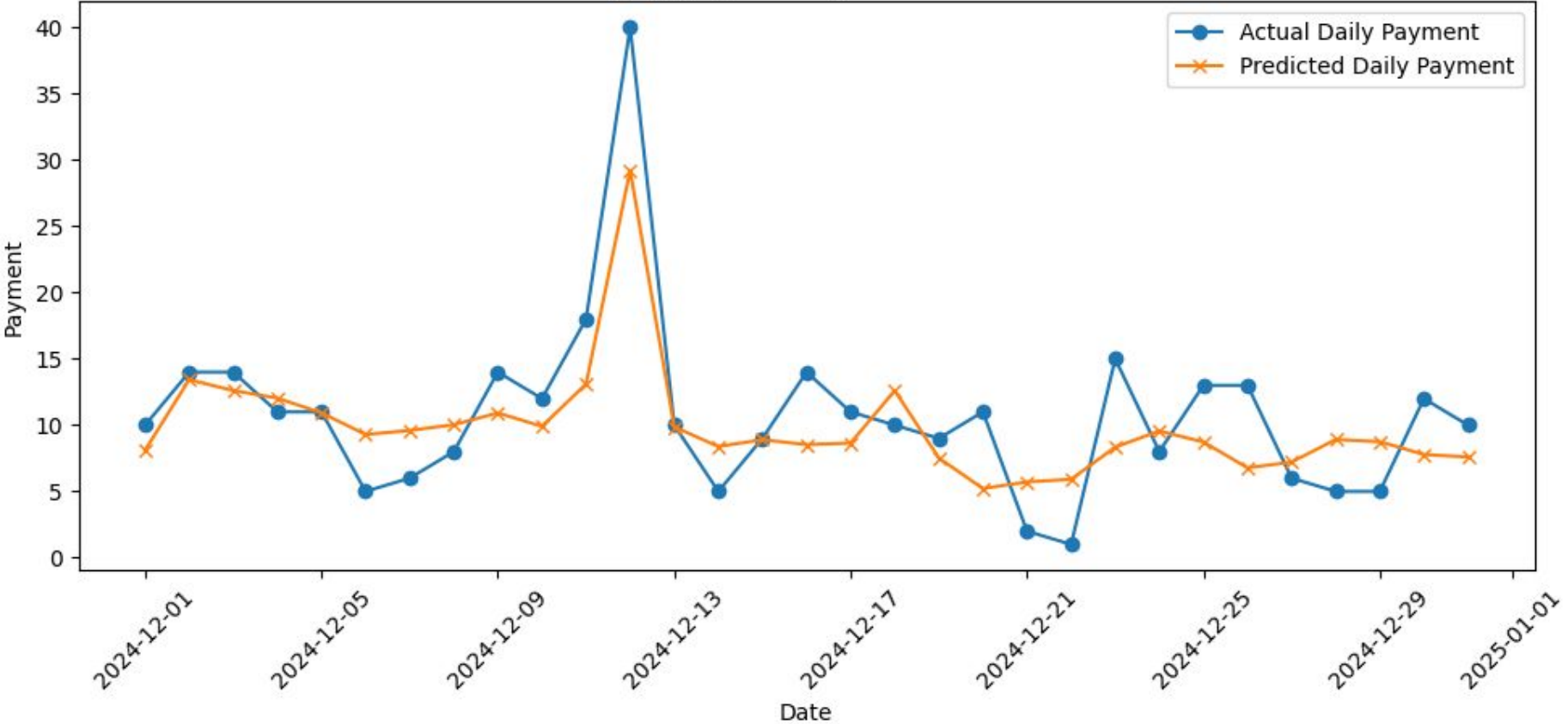- Each tree learns from the mistakes of the previous trees, improving accuracy with every step.
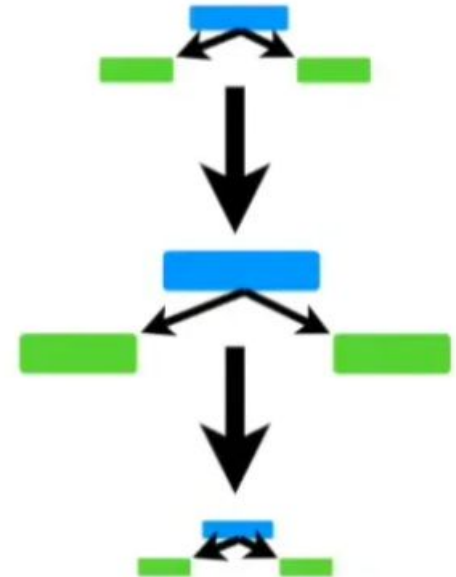
2024-11 RMSE: **0.538**

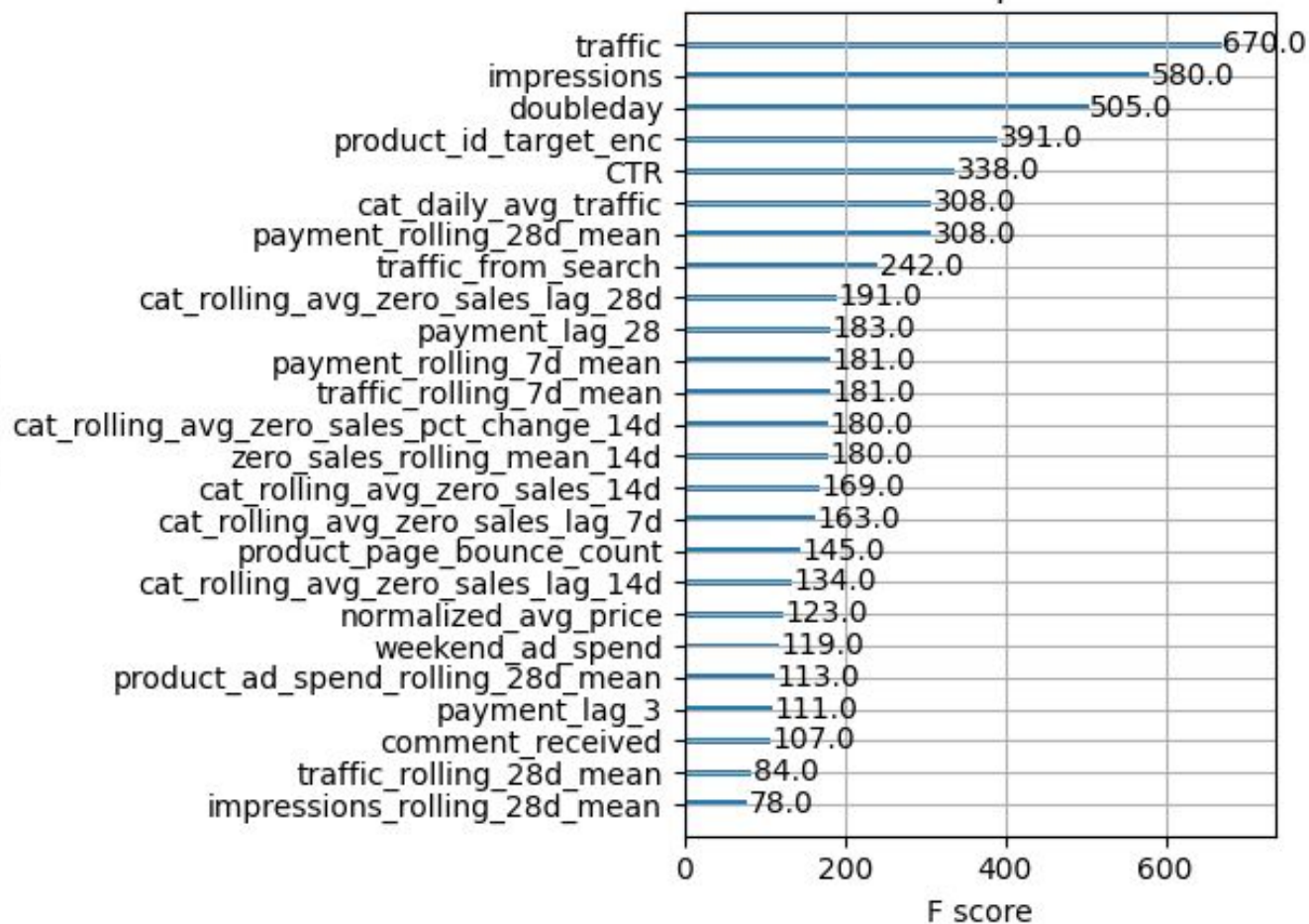November Daily Aggregated Payment: Actual vs. Predicted

2024-12 RMSE: **0.645**

December Daily Aggregated Payment: Actual vs. Predicted

Feature Importance

Features	F score
traffic	670.0
impressions	580.0
doubleday	505.0
product_id_target_enc	391.0
CTR	338.0
cat_daily_avg_traffic	308.0
payment_rolling_28d_mean	308.0
traffic_from_search	242.0
cat_rolling_avg_zero_sales_lag_28d	191.0
payment_lag_28	183.0
payment_rolling_7d_mean	181.0
traffic_rolling_7d_mean	181.0
cat_rolling_avg_zero_sales_pct_change_14d	180.0
zero_sales_rolling_mean_14d	180.0
cat_rolling_avg_zero_sales_14d	169.0
cat_rolling_avg_zero_sales_lag_7d	163.0
product_page_bounce_count	145.0
cat_rolling_avg_zero_sales_lag_14d	134.0
normalized_avg_price	123.0
weekend_ad_spend	119.0
product_ad_spend_rolling_28d_mean	113.0
payment_lag_3	111.0
comment_received	107.0
traffic_rolling_28d_mean	84.0
impressions_rolling_28d_mean	78.0

Average RMSE: **0.569**

Dec 2024 RMSE: **0.645**