

# Lab 02 - Store Sales - Time Series Forecasting









220144P – Diwakar J.S.P.

**Leaderboard Score = 1.18935**

Store Sales - Time Series Forecasting

Submit Prediction

OverviewDataCodeModelsDiscussionLeaderboardRulesTeamSubmissions

415	Muhammad Uzair abbasi		1.15772	3	2mo
416	Oriol Raventós Toribio		1.16130	1	2mo
417	UOM_220144P		1.18935	5	2h
<div> Your Best Entry! Your most recent submission scored 1.18935, which is an improvement over your previous score of 1.49793. Great job!</div> <div>Tweet this</div>					
418	TonyGu715		1.19648	9	2mo
419	hehcccc		1.24703	10	1mo
420	Antonio Pacheco		1.25164	1	2mo
421	Linyh		1.28048	6	2mo

## Solution Description

The dataset contains over 3 million rows of daily sales information spanning from 2013 to 2017. The challenge is to forecast sales for 33 distinct product families across 54 different stores located throughout Ecuador. To solve this, a regression based approach was implemented instead of traditional time series models. All provided datasets in the competition, which are training data, store information, oil prices, holiday events, and transaction counts were merged into a single dataframe. This unified dataset allows the model to learn complex relationships between sales and various external factors simultaneously.

A comprehensive Exploratory Data Analysis was conducted to understand the underlying patterns in the data. Key insights identified include,

- A clear increasing trend in overall sales over the 5 year period.
- Dominant product families: GROCERY I, BEVERAGES, PRODUCE, CLEANING, and DAIRY consistently represent the majority of sales.
- Strong yearly seasonality, with sales showing a month over month increase.
- A clear payday effect, with sales spiking around the 15th and the end of each month.

- A positive correlation between the number of promotions and sales volume, confirming the value of the onpromotion feature.

Based on these insights, a set of features was engineered to capture these patterns. They are,

- Basic date based features such as year, month and day.
- Cyclic features using Sin and Cos transformations to represent the cyclical nature of weeks and months.
- Holiday and Special Event features, including location specific holiday flags.
- A Payday feature to capture the sales spikes at the middle and end of month.
- An earthquake flag to account for the anomaly in sales data following the 2016 earthquake.
- Lag and rolling window features: This set of features provides recent historical context, including sales from the previous 16, 21, and 28 days, and rolling averages and standard deviations over 7, 14, and 28 day windows. These features proved to be highly correlated with the sales target.

Categorical features like the 54 store numbers and 16 unique city names were preprocessed in two ways. For LightGBM and CatBoost, they were converted to a category data type for efficient internal handling. For XGBoost and RandomForest, they were one hot encoded.

After feature engineering, the unified dataset was separated back into training and testing sets. Several individual models were tested, including LightGBM, RandomForest, and XGBoost. These are the public scores they achieved,

1. LightGBM Score: 3.51051
2. RandomForest Score: 3.27794
3. XGBoost Score: 1.86968

A key finding was that 939,130 records, representing around 30% of the training data, had a sales value of zero, indicating days with no sales for a given product. To address this, a more advanced two stage model was implemented. This approach first uses a classification model to predict whether a sale will occur or not. If a sale is predicted, a second regression model then predicts the actual sales amount. This technique was implemented first with LightGBM and then with XGBoost. The two stage XGBoost model yielding the best results.

1. Two-Stage LightGBM Score: 1.49793
2. Two-Stage XGBoost Score: 1.18935

## Data Cleaning and Preprocessing

These actions were taken to create a complete and robust dataset for better model training.

- **Data Integration:** The five distinct data sources provided which are the main training data, store metadata, daily oil prices, holiday and event data, and daily transaction counts were merged into a single, unified dataframe. Instead of a simple merge, holidays were first separated by their geographical scope.
  1. National holidays were applied universally to all stores based on the date.
  2. Regional holidays were merged using both date and state, ensuring they only affected stores within the correct province.
  3. Local holidays were merged on date and city, precisely targeting only the intended locations.

This multi level merging strategy ensured that the impact of holidays was modeled with geographical accuracy, improving the quality of the features.

- **Missing Value Imputation:** For the `dcoilwtico` column in oil dataset, a combination of forward fill and backward fill was used to populate missing values. For the `transactions` column in transaction dataset, missing values were filled with 0, based on the assumption that a missing record corresponds to a day the store was closed or had no transactions.
- **Data Type Conversion:** The date column was converted to a proper datetime object, which is essential for time series analysis. Furthermore, object columns representing categories were converted to the efficient category data type for models that could handle it.
- **Data Preservation:** No rows were removed from the dataset. Since this is a time series problem, deleting data points would create gaps in the timeline, which would disturb the calculation of crucial lag and rolling window features.
- **Feature Engineering:** New features were engineered based on the insights gained from the EDA. This allowed the model to explicitly learn the patterns related to trends, seasonality, holidays, and promotions that were identified.

## Additional Evaluation Metrics

While the official competition metric is the Root Mean Squared Logarithmic Error (RMSLE), additional metrics were used during validation to provide a better understanding of the model's performance.

- **Root Mean Squared Error (RMSE):** This metric calculates the square root of the average of the squared differences between the predicted and actual sales.

- **Mean Absolute Error (MAE):** This metric provides the most direct interpretation of the model's error. It represents the average absolute difference between the predicted and actual sales, giving a clear sense of the average error margin in sales units.

## **Alternative Solution**

An effective alternative to the two stage model would be to create a stacking ensemble of the best performing individual models such as LightGBM, XGBoost, and CatBoost. In this approach, the individual models which are called as the base models, would make their own predictions. A second model which is called as a meta model, would then be trained to learn how to best combine the predictions from the base models. This technique often leads to a more robust and accurate final prediction by leveraging the diverse strengths of each individual model.

## **Issues and Improvements**

While the two stage model performed well, there can be several issues in this approach, and addressing those can lead to further improvement of the predictions.

### **Current Issues**

- **Missing Data:** The approach of filling missing transaction data with 0 is a conservative assumption. An alternative strategy can be explored.
- **Sub-Optimal Model Parameters:** The parameters used for the models were effective starting points but were not systematically optimized.

### **Improvements**

- **Hyperparameter Tuning:** A more systematic hyperparameter tuning process, using a library like Optuna, can be conducted to find the optimal set of parameters for both the classifier and the regressor, which would likely lead to a significant performance boost.
- **Advanced Feature Engineering:** Further improvements could be gained by engineering a more sophisticated set of features. This would involve creating new variables designed to capture more complex and subtle patterns within the data. By representing the interactions between different factors and how sales trends evolve over time in a better way, the model could be provided with more effective insights.