

Store Sales - Time Series Forecasting

1. Domain Background

The focus of my upcoming project will be on Time Series Forecasting, a statistical method that leverages past data to predict future values of a variable. This technique finds widespread use across various fields, such as finance, economics, engineering, and marketing, and holds great significance for my personal work.

Essentially, time series forecasting involves analyzing historical data to identify patterns and trends that can be used to forecast future values of the variable.

2. Problem Statement

The aim of the project is to employ time series forecasting techniques to predict the sales of a food retailer named Corporación Favorita, which is a prominent retailer based in Ecuador. The project's specific objective is to develop a model that can provide precise forecasts for the unit sales of a vast array of products sold across various Favorita stores.

3. Datasets and Inputs

The primary objective of this project is to predict the sales of numerous product families sold in Favorita stores situated in Ecuador. To accomplish this, I will be using the training data, which comprises details such as dates, store and product information, sales figures, and whether or not a particular item was being promoted. Additionally, there are other files containing supplemental information that could aid in the development of the forecasting models.

These files are available on the Kaggle platform, and the link to access them is provided here:

<https://www.kaggle.com/competitions/store-sales-time-series-forecasting/overview>.

4. Solution Statement

The proposed solution for this problem would be to develop an accurate time series forecasting model that can predict the sales of various product families sold at different Favorita stores in Ecuador.

To achieve this, the model would need to take into account the historical sales data, store and product information, promotional activities, and any other relevant data provided in the training dataset. The model would need to be able to capture any patterns, trends, seasonality, and irregularities in the data to make accurate predictions for future sales.

Some possible techniques that could be used for this project include simple moving averages, exponential smoothing, ARIMA models, and machine learning algorithms such as random forests and neural networks. The choice of the appropriate technique would depend on the characteristics of the data.

The final solution would be a well-tuned forecasting model that can provide precise predictions for the sales.

5. Benchmark Model

A simple and commonly used benchmark model for time series forecasting is the naive model, which assumes that the next value in the time series will be equal to the most recent observed value. This means that the model does not take into account any trends, seasonality, or other patterns in the data. While this model is extremely basic, it provides a useful benchmark against which other, more sophisticated models can be compared.

6. Evaluation Metrics

While common metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) can be utilized to assess the performance of various models, the primary metric that will be used in this project is the Root Mean Squared Logarithmic Error (RMSLE). This metric is specifically employed in evaluating the Kaggle competition and provides a more accurate measure of the prediction error when the target variable has a wide range of values.

7. Project Design

Following the data preparation and analysis, my plan is to train a model using classic statistical methods such as ARIMA, as well as more complex machine learning models like Random Forest or XGBoost. I am considering using AutoGluon for this purpose, although I am not certain whether it will be feasible. Additionally, I am interested in implementing the Amazon Forecast

AWS solution and comparing the different metrics to conduct an evaluation. Once the best model has been identified, I plan to implement an endpoint for inference.