Predictive Analytics:
Time Series Forecasting

</talentlabs>

Overview

This case study aims to give an idea of applying Machine Learning to a time series forecasting problem. As you know, time series is a set of observations taken at a specified time usually at equal intervals. It is used to predict future values based on previously observed values. Here, in this case study we will tackle one such time series forecasting problem using ML.

Problem Statement

As a junior data analyst intern in a firm. For one of the client XYZ marketings, your manager asked you to use time-series forecasting to forecast store sales on data from Corporación Favorita, a large Ecuadorian-based grocery retailer. Here, you'll build a model that more accurately predicts the unit sales for thousands of items sold at different Favorita stores. Your manager believes that it will help you practice your machine-learning skills with an approachable training dataset of dates, store, and item information, promotions, and unit sales.

Dataset

Download the dataset from here: https://www.kaggle.com/competitions/store-sales-time-series-forecasting/data

train.csv

The training data comprises a time series of features store_nbr, family, and promotion as well as the target sales.

store_nbr identifies the store at which the products are sold.

family identifies the type of product sold.

sales give the total sales for a product family at a particular store at a given date. Fractional values are possible since products can be sold in fractional units (1.5 kg of cheese, for instance, as opposed to 1 bag of chips).

onpromotion gives the total number of items in a product family that were being promoted at a store on a given date.

test.csv

The test data has the same features as the training data. You will predict the target sales for the dates in this file. The dates in the test data are for the 15 days after the last date in the training data.

You don't need to use this file in this assignment. However, if you are planning to submit the project the Kaggle, then you will need to use this testing set.

stores.csv

Store metadata, including city, state, type, and cluster (grouping of similar stores).

oil.csv

Daily oil price. Includes values during both the train and test data timeframes. (Ecuador is an oil-dependent country and its economic health is highly vulnerable to shocks in oil prices.)

holidays_events.csv

Holidays and Events, with metadata

Expected Outcomes

Create a machine learning time series model that predicts the target sales of the store.

- Understand the given problem and explore the data set and do the necessary data pre-processing.
- Do exploratory data analysis to understand and visualize the data.
- Find how the data are related to each other.
- Work on ML algorithms for time series forecasting to select an ML model that performs best for the data at hand.
- Evaluate your results and find the best algorithm that predicts the “target sales”. (You can build 5 models and evaluate the model using a subset of your training data. If you forgot about splitting the data into training and testing set, you can refer to this page: https://www.geeksforgeeks.org/how-to-do-train-test-split-using-sklearn-in-python/)

Note: You need to submit one/two Ipython notebooks which clearly explains the thought process behind your ML model (either in comments or markdown text), code and relevant plots. Also, all the visualizations and plots must be done in Python(should be present in the Ipython notebook).

Thank you.

All The Best :)

</talentlabs>