Time Series Regression Analysis (Corporation Favorita)

Scenario: You are a data scientist in Corporation Favorita, a large Ecuadorian-based grocery retailer. Corporation Favorita wants to ensure that they always have the right quantity of products in stock. To do this you have decided to build a series of machine learning models to forecast the demand of products in various locations. The marketing and sales team have provided you with some data to aid this endeavor. Your team uses CRISP-DM Framework for Data Science projects

This is a **time series regression analysis** problem. In this project, you'll predict store sales on data from Corporation Favorita, a large Ecuadorian-based grocery retailer.

Specifically, you are to **build a model** that more accurately predicts the unit sales for thousands of items sold at different Favorita stores.

The training data includes dates, store, and product information, whether that item was being promoted, as well as the sales numbers. Additional files include supplementary information that may be useful in building your models

File Descriptions and Data Field Information

train.csv

- The training data, comprising time series of features store_nbr, family, and onpromotion 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** gives 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 at a given date.

test.csv

- The test data, having 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.

transaction.csv

- Contains date, store_nbr and transaction made on that specific date.

sample_submission.csv

- A sample submission file in the correct format.

stores.csv

- Store metadata, including city, state, type, and cluster.
- cluster is a grouping of similar stores.

oil.csv

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

holidays_events.csv

- Holidays and Events, with metadata

NOTE: Pay special attention to the transferred column. A holiday that is transferred officially falls on that calendar day but was moved to another date by the government. A transferred day is more like a normal day than a holiday. To find the day that it was celebrated, look for the corresponding row where type is **Transfer**.

For example, the holiday Independencia de Guayaquil was transferred from 2012-10-09 to 2012-10-12, which means it was celebrated on 2012-10-12. Days that are type **Bridge** are extra days that are added to a holiday (e.g., to extend the break across a long weekend). These are frequently made up by the type **Work Day** which is a day not normally scheduled for work (e.g., Saturday) that is meant to payback the Bridge.

- Additional holidays are days added a regular calendar holiday, for example, as typically happens around Christmas (making Christmas Eve a holiday).

Additional Notes

- Wages in the public sector are paid every two weeks on the 15th and on the last day of the month. Supermarket sales could be affected by this.
- A magnitude 7.8 earthquake struck Ecuador on April 16, 2016. People rallied in relief efforts donating water and other first need products which greatly affected supermarket sales for several weeks after the earthquake.

Hypothesis

- Formulate 1 null and alternate hypothesis each.

Analytical Questions: The questions below are to be answered. Do note that, you are free to draw more questions from the data after answering these questions.

1. Is the train dataset complete (has all the required dates)?
2. Which dates have the lowest and highest sales for each year (excluding days the store was closed)?
3. Compare the sales for each month across the years and determine which month of which year had the highest sales.
4. Did the earthquake impact sales?
5. Are certain stores or groups of stores selling more products? (Cluster, city, state, type)
6. Are sales affected by promotions, oil prices and holidays?
7. What analysis can we get from the date and its extractable features?
8. Which product family and stores did the promotions affect.

9. What is the difference between RMSLE, RMSE, MSE (or why is the MAE greater than all of them?)
10. Does the payment of wages in the public sector on the 15th and last days of the month influence the store sales.

Your task is to build a model that more accurately predicts the unit sales for thousands of items.

Important

- Document process from data cleaning, analysis, assumptions, model building etc. Marks will be awarded for documentation.

Rubrics

Hypothesis, EDA & Analytical Questions

- Excellent: Validated the hypothesis and answered all questions listed earlier with appropriate charts. Used relevant diagrams and charts to show analysis/metrics.
- Good: Validated at least 4 hypothesis and answered some of the questions listed with appropriate charts. Used relevant diagrams but might need some improvement and.
- Fair: Lack of clarity on whether the hypothesis was true.
- Poor: Not answered any of the hypothesis

Model Building & Improvement

- Excellent: Model has an RMSLE of 0.2
- Good: Model has RMSLE of 0.3
- Fair: Model has RMSLE of 0.4
- Poor: Model has RMSLE of 0.4 +

Documentation & Key Insight:

- Excellent: Having documentation on the project ie data cleaning, analysis, hypothesis and model.
- Good: Gave a summary on some of the processes
- Fair: Gave a bullet list of the processes with short sentences
- Poor: No documentation