Data Glacier

Your Deep Learning Partner

Week 7 Presentation: Retail Forecasting

Group Name: Light of Summer

Member's Details

Member 1: Alexander Quesada Quesada
alexander.quesadaquesada@ucr.ac.cr
Costa Rica
Universidad de Costa Rica
Data Science

Problem Statement

The large company who is into beverages business in Australia. They sell their products through various super-markets and engage into heavy promotions throughout the year. Their demand is also influenced by various factors like holiday, seasonality**. They needed forecast of each of products at item level every week in weekly buckets**.

The time series data showed a range of patterns, some with trends, some seasonal, and some with neither. At the time, they were using their own software, written in-house, **but it often produced forecasts that did not seem sensible**. Company wanted to explore power of AI/ML based forecasting to replace their in house local solution

1. **Build at least 4-5 multivariate forecasting model, which included ML or Deep Learning**, based Model in PySpark leveraging parallel computing techniques (You can develop models without Pyspark if you are not comfortable with pyspark and parallel computing).

2. **Demonstrate best in class forecast accuracy** (Forecast Accuracy = 1 - Wt. MAPE where Wt. MAPE = sum(Error)/sum(Actual)

3. Write a code in such a way **you run the model in least time**

4. Demonstrate explainability in the form of **contribution of each variables**

5. Leveage Feature Engineering concepts **to derive more variables to gain accuracy improvement** (You can build model and demonstrate accuracy for Q3-Q4 of 2020)

Business understanding

The first insight approach for this task consists on:

1. **Data cleaning**, as in handling missing values, outliners, and format issues.

2. **Sorting each variable by its relevance for sales**, not only this makes easier the process of feature contribution demonstration on point number 4, but also it returns the data columns which the model would perform its gradient descent most of its training time, this can be performed with multiple feature impact analysis techniques such as **SelectKBest**, **Elastic Net Regularization, PCA, Gradient Boosting Regression, Recursive Feature Elimination, etc…**

3. **Performing EDA with Time Series emphasis**, this is both an essential part of Feature Engineering (See point number 5) but also would give insights on the patterns, seasonality, trends, and potential outliers within the time series data; it would also graphically support relationships between variables and its impact for sales, and other statistical facts about the data behavior.

4. **Model selection, training, error estimation and tuning**, this is the most extensive part of the project pipeline as it covers points 1, 2 and 3. It involves choosing the best model to perform the forecast predictions

based on model options detailed in future deliveries, training with the lowest variable cost possible to optimize time and resources, estimating the error as class forecast accuracy and tuning the model to save the best of all executions weights based on valid predictions.

Project lifecycle along with deadline

13 – 19 September 2023	Project and Dataset Details
20 – 26 September 2023	Data understating and cleaning
27 – 2 October 2023	Data transformation
3 – 9 October 2023	EDA and Feature Engineering
10 – 16 October 2023	EDA Presentation
17 – 23 October 2023	Model selection and tuning
24 – 30 October 2023	Final Presentation of Best Solution