Data Science: Retail Forecasting

Group Name: Ranalytics

Name: Rana Fakeeh

Email: ranafakeeh1987@gmail.com

Country: Saudi Arabia

College/Company: Data Glacier

Specialization: Data Science

Batch code: LISUM28

Submission date: 1/19/2024

I. Problem description

Project Title: Retail Forecasting of Beverages Company in Australia

Problem Statement: a large company who is into beverages business in Australia, selling its products through various super-markets and engaging into heavy promotions throughout the year. This company faces the challenge of accurately forecasting the demand (i.e., quantity sold) for each product on a weekly basis, considering various influencing factors such as holidays, promotions, and multiple patterns of seasonality and trend. To enhance the accuracy of their forecasts and replace their in-house solution, the company is seeking a data-driven approach, leveraging the power of machine learning.

Data Characteristics: the data for analysis is already collected and uploaded on Google Docs platform[1] as a spreadsheet. The data consists of 12 features (columns), and 1,218 observations (rows) distributed over 6 products each has 204 rows (except for one product having 198 rows). The features include information about the products, sales, discount, promotions, and holidays.

Challenges: challenges include few features and insufficient data size for the regression problem; therefore, feature engineering and augmentation could be applied. Different trend and seasonality patterns as well as minimal execution time are also obvious challenges for multivariate time series modeling.

[1] https://docs.google.com/spreadsheets/d/1sOTsmkY4ZeNzww_yDGePGYt1iXtZjNHb/edit?usp=sharing&ouid=110600711982317630177&rtpof=true&sd=true

II. Business understanding

The main objective of this project is to develop advanced and efficient forecasting models that can provide accurate predictions of product demand on a weekly basis. This will enable the company to optimize inventory management, production planning, and promotional strategies. Accurate forecasts will also help in minimizing stockouts, reducing overstock situations, and ultimately improving customer satisfaction and profitability.

III. Solution description

The solution is to build at least 4-5 multivariate forecasting models which include Machine Learning, and Deep Learning techniques and also leverage parallel computing to ensure efficient and minimal execution time (e.g. PySpark could be implemented). The contribution of each variable into the forecast will be explained via visuals and correlation metrics. The evaluation will demonstrate best in-class forecast accuracy, measured by the weighted Mean Absolute Percentage Error (Wt. MAPE) as following:

$$Forecast\ Accuracy = 1 - Wt.\ MAPE$$

$$Wt.\ MAPE = sum(Error)/sum(Actual)$$

IV. Project lifecycle along with deadline

Week 1-2: Project Setup and Data Preparation

- Understand data quality and perform data cleaning as necessary.
- Exploratory Data Analysis (EDA): explore and analyze data to identify patterns, seasonality, and potential feature engineering opportunities.

Week 3-4: Feature Engineering and Model Development

- Leverage feature engineering techniques to derive additional variables that may improve forecasting accuracy.
- Build 4-5 forecasting models, including ML and deep learning models, using PySpark or other suitable tools.
- Train and validate models on the available historical data.

Week 5-6: Model Optimization and Execution

- Optimize code for parallel computing and efficient execution.
- Fine-tune hyperparameters of the models to achieve the best possible accuracy.
- Evaluate model performance on a holdout dataset to ensure generalization.

Week 7: Explainability and Documentation

- Develop methods to explain the contributions of each variable in the models.
- Document the entire project, including data preprocessing steps, model development, and optimization techniques.
- Prepare a presentation summarizing the project and highlighting key findings and recommendations.

Deadline: The project is expected to be completed within 7 weeks from the project initiation date, with regular progress updates and milestone reviews. The final deliverables, including the forecasting models, codebase, documentation, and presentation, should be ready for presentation to the client by the end of the 7th week.