Data Science – Retail Forecasting

Group Name	RFuMFM (Retail Forecasting Using Combined ML and Deep Learning Multivariate Forecasting Models)
Name	Tanmay Potbhare
Email	tupotbhare@gmail.com
Country	Ireland
College	Dublin City University
Specialization	Data Science

Problem Description

The large company who is into beverages business in Australia. They sell their products through various super-markets and also 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.

Business Understanding

Large beverage company in Australia wants to forecast the demand for their products at the item level every week in weekly buckets. The demand is influenced by various factors such as seasonality, promotions, holiday, and other variables. The company currently uses an inhouse software for forecasting but it often produces inaccurate results. Therefore, the company wants to explore the power of AI/ML based predictions to replace their current solution and improve the accuracy of their forecasts. The objective is to build at least 4-5 multivariate forecasting models using ML or deep learning-based approaches in PySpark, and demonstrate best-in-class forecast accuracy. Additionally, the company wants the model to be efficient and run in the least amount of time possible, while also providing explainability and insight into the contribution of each variable in the forecasting model.

Project Lifecycle with Deadlines

1. **Project Initiation**: In this stage, the project goals and objectives are defined, and stakeholders are identified. The resources required for the project are also determined.
2. **Data Collection and Preparation**: In this stage, historical data on sales, promotions, holidays, and other variables that may impact demand are collected and cleaned. The data is formatted and transformed as necessary to prepare it for modelling. Feature engineering techniques are used to derive more variables to gain accuracy improvement.
3. **Model Selection and Training**: In this stage, 4-5 multivariate forecasting models are built, which includes ML or Deep Learning based Model in PySpark leveraging parallel computing techniques. The models are trained on the historical data to learn the patterns and relationships between the variables and the target variable.
4. **Model Evaluation and Refinement**: In this stage, the models' performance is evaluated using appropriate metrics such as Wt. MAPE, and the models are refined based on the evaluation results. This may include adjusting hyperparameters, adding or removing features, or using different modelling techniques.
5. **Model Deployment and Monitoring**: In this stage, the final model is deployed, and its performance is monitored over time to ensure that it continues to produce accurate

forecasts. This may include updating the model periodically as new data becomes available or as the business environment changes.

Data Intake Report

Name: Retail Forecasting Using Combined ML and Deep Learning Multivariate forecasting models
Report date: 17th March 2023
Internship Batch: LISUM18
Version:<1.0>
Data intake by: Tanmay Umesh Potbhare
Data intake reviewer:<intern who reviewed the report>
Data storage location: https://github.com/TanmayPotbhare/Retail-Forecast-Using-Multivariate-Forecasting-Models
Dataset Details:

Total Number of Observations	1219 Rows
Total Number of Files	1
Total Number of Features	12 Columns
Base format of File	.XLSX
Size of Data	73.8 KB (75,647 bytes)

Proposed Approach

1. Data Preparation - Cleaning, formatting, and processing the data for input to the machine learning or deep learning models.
2. Model Selection - Models can include linear regression, random forest, gradient boosted trees, artificial neural networks, and recurrent neural networks.
3. Model Training - Involves optimizing the model parameters to minimize the prediction error.
4. Model Evaluation - Evaluation using the testing dataset.
5. Hyperparameter tuning - The parameters that control the behaviour of the machine learning or deep learning models. Tuning the hyperparameters can help to improve the accuracy of the model. PySpark provides tools for hyperparameter tuning, such as Grid Search and Random Search.
6. Model Refinement - The refinement process will be iterative, with multiple rounds of model training, evaluation, and refinement.
7. Interpretability and Explainability - Analysing the contribution of each variable to the forecast. The contribution of each variable will be determined using feature importance or Shapley values.
8. Also check for optimization as the code should be run in least amount of time.

GitHub Repo Link - https://github.com/TanmayPotbhare/Retail-Forecast-Using-Multivariate-Forecasting-Models