PROJECT PLAN

Project Title

Analysing the role of store characteristics and economic factors in Walmart Sales using Machine Learning

Research Question

Is there a difference in performance between Deep Learning, Statistical Methods, and Prophet in forecasting sales data in Walmart weekly data?

Objectives

- Create a ML pipeline to implement three types of forecasting automatically on passing the dataset.
- Give a comparison graph of the three methods using a dashboard.
- Train and test the Walmart weekly data.
- Create an automated system that takes input data as csv, pre-process it, train the model, and give the results.
- Create an interactive UI for this system.

Background and Summary

One of the sectors most dependent on accurate sales forecasting is retail, which affects decisions about personnel, marketing, and inventory. Forecasting sales patterns enables large companies like Walmart to operate in a way that is both customer-focused and economical. The Walmart dataset is comprehensive; it contains weekly sales data together with information on holidays, unusual temperatures, and other local weather-related anomalies across all American retail locations, as well as information on petrol costs. Additionally, it includes a number of macroeconomic data for good measure, such as the CPI and unemployment rates.

Trade has always advanced significantly as a result of the use of machine learning (ML) and deep learning techniques for sales prediction, particularly in the retail sector. Traditional statistical techniques, such ARIMA, have proved helpful in time series forecasting by addressing trends and seasonality, but they are ineffective at capturing complicated nonlinear interactions; ML models are far better at this task. Facebook developed Prophet, a modern forecasting tool that works well with business time series data that show significant seasonal trends, holidays, and outliers—features that are typical of retail data. On another hand, deep learning methods like as LSTM and GRU have exceptional ability to recognize complex patterns in sequence-based data and to handle big datasets with many data types simultaneously—skills that are critical in retail settings.

List of References

Latha, S.B., Dastagiraiah, C., Kiran, A., Asif, S., Elangovan, D. and Reddy, P.C.S., 2023, August. An Adaptive Machine Learning model for Walmart sales prediction. In 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT) (pp. 988-992). IEEE.

https://ieeexplore.ieee.org/document/10245029

Vyas, R. and As, R., 2022, March. Seasonal Sales Prediction and Visualization for Walmart Retail Chain Using Time Series and Regression Analysis: A Comparative Study. In 2022 International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN) (pp. 1-6). IEEE.

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9761294

Qiao, Z., 2020, October. Walmart Sale Forecasting Model Based On LightGBM. In 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI) (pp. 76-79). IEEE.

https://ieeexplore.ieee.org/document/9360930

Task List and Project Time Line

Task	7-Jun	25-Jun	10-Jul	25-Jul	09-Aug	24-Aug	29-Aug
Research and Data Collection							
Literature Review							
Predictive Model Development							
Model Analysis and comparison							
Interface Design							
Project Documentation & Reporting							

Here's a brief explanation for each step depicted in the Gantt chart:

- Research & Data Collection: Gather necessary data and research materials to lay the foundational knowledge and resources for the project.
- Literature Review: Perform a thorough review of existing literature to understand previous work and identify gaps in the current knowledge.
- Predictive Model Development (LSTM, Prophet, ARIMA): Develop predictive models using LSTM, Prophet, and ARIMA techniques to forecast sales data.
- Model Analysis & Comparison: Analyse and compare the performance of the developed models to determine their accuracy and efficacy.
- Interface Design: Design a user interface that allows users to interact with the predictive models and view forecasting results.
- Project Documentation & Reporting: Document all aspects of the project and prepare a comprehensive report detailing methodologies, findings, and conclusions. This step overlaps with all other activities, highlighting its ongoing nature throughout the project duration

Data Management Plan

Summary of Dataset

The Walmart dataset includes eight columns, and a total of 6,435 entries. This weekly sales data is distributed among 45 different stores during a period of 143 days starting from February 5, 2010. These consist of several metrics like store ID, date, weekly sales, holiday flag, temperature, fuel price (in dollars), Consumer Price Index (CPI), and unemployment rates. About 7% records are marked as holiday weeks which indicate special sale conditions. The database also contains environmental determinants such as temperatures in addition to fuel prices sold in all Walmart's nationwide plus the entire Consumer Price Index (CPI) measuring inflation rates and unemployment percentages for each year's first quarter within the United States.

Data collection

Kaggle is the source of the dataset under discussion, a popular open-source data platform. The website gives access to several free datasets that are used for analytics and modeling purposes. Kaggle is well known among data science enthusiasts because it has databases across multiple disciplines, which means researchers and practitioners can interact with them in various forms such as hands-on projects or collaborations. Here is the source to the dataset: https://www.kaggle.com/datasets/yasserh/walmart-dataset/data

Document control

GitHub: https://github.com/mp22abw/Final-Project

The above repository is used to maintain the records of data and code changes made in the duration of this research.

Ethical requirements

1. Does the data meet GDPR requirements? - Yes

2. Does the project conform to UH ethical policies? – Yes

3. Do you have permission to use the data for your proposed research project? - Yes

4. Are you assured that the data was collected ethical (i.e. by the original people who gathered/collected/ collated/made the data)? - Yes

References

Choudhary, A. (2018) *Generate Quick and Accurate Time Series Forecasts using Facebook's Prophet (with Python & R codes)*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2018/05/generate-accurate-forecasts-facebook-prophet-python-r/ (Accessed: June 14, 2024).

Pathak, P. (2020) *Building an ARIMA model for time series forecasting in python*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2020/10/how-to-create-an-arima-model-for-time-series-forecasting-in-python/ (Accessed: June 14, 2024).

Phi, M. (2018) *Illustrated Guide to LSTM's and GRU's: A step by step explanation*, *Towards Data Science*. Available at: https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21 (Accessed: June 14, 2024).