Contents

[Chapter 1: Project Plan 2](#_Toc169278695)

[Project Title 2](#_Toc169278696)

[Research Question 2](#_Toc169278697)

[Objectives 2](#_Toc169278698)

[Background and Summary 2](#_Toc169278699)

[List of References 2](#_Toc169278700)

[Chapter 2: Task List and Project Time Line 4](#_Toc169278701)

[Chapter 3: Data Management Plan 5](#_Toc169278702)

[Summary of Dataset 5](#_Toc169278703)

[Data collection 5](#_Toc169278704)

[Document control 5](#_Toc169278705)

[Ethical requirements 5](#_Toc169278706)

[References 5](#_Toc169278707)

# Chapter 1: 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, 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

Retail stands as one of the industries heavily reliant on sales forecasting skill— impacting staff, marketing, and inventory decisions. Being able to predict sales trends can allow major corporations such as Walmart to cost-effectively make their business while also customer-oriented. The Walmart dataset is all-encompassing: it includes weekly sales figures coupled with details on when holidays occurred, abnormal temperatures recorded among other local weather-related anomalies all over America's retail outlets and their gasoline prices. On top of that, it features various macroeconomic indicators like CPI and unemployment rates for good measure.

The adoption of machine learning (ML) and deep learning approaches for sales prediction has always led to a significant advancement in the field of trade; more so retail. While traditional statistical methods such as ARIMA (Pathak, 2020) have been instrumental in addressing trends and seasonality within time series forecasting, they prove incompetent at capturing complex nonlinear relationships— a task achieved much more effectively by ML models. Prophet is a contemporary forecasting tool engineered by Facebook (Choudhary, 2018), which is well suited for business time series data that exhibit strong seasonal patterns along with holidays and outliers— common among retail data. On another front, deep learning techniques including LSTM and GRU (Phi, 2018) show prowess in detecting intricate patterns within sequence-based data, managing large datasets with diverse data types all at once — capabilities essential for retail environments.

## List of References

Latha, S. B. *et al.* (2023) “An Adaptive Machine Learning model for Walmart sales prediction,” in *2023 International Conference on Circuit Power and Computing Technologies (ICCPCT)*. IEEE, pp. 988–992.

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

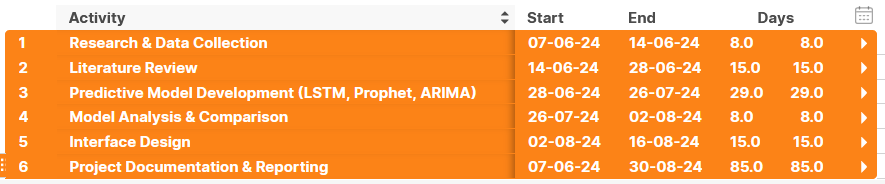
Niu, Y. (2020) “Walmart Sales Forecasting using XGBoost algorithm and Feature engineering,” in *2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*. IEEE, pp. 458–461.

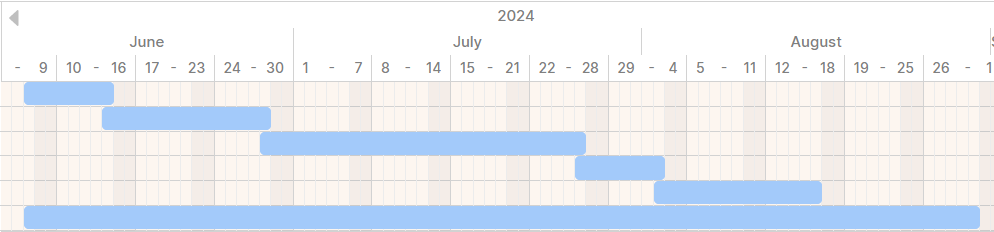
https://ieeexplore.ieee.org/document/9403812

Qiao, Z. (2020) “Walmart sale forecasting model based on LightGBM,” in *2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*. IEEE, pp. 76–79.

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

# Chapter 2: Task List and Project Time Line





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: Conduct 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

# Chapter 3: 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. Weekly sales fluctuate greatly between approximately $209986 and $3818686 with an average value around $1046965. About 7% records are marked as holiday weeks which indicate special sale conditions. The database also contains environmental determinants such as temperatures that range between -2.06 and 100.14-degrees Fahrenheit in addition to fuel prices ranging from $2.472 to $4.468 per gallon 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:   
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).