**DSC630 PA - Project Proposal**

Bellevue University

Winter Term- 2021

Walmart Sales Forecasting

Nitin Mahajan | Ganesh Kale

**Abstract**

**Sales Forecasting** is the process of using a company’s sales records over the past few years to predict the short-term or long-term sales performance of the company in the future. This is one of the pillars of proper financial planning. Sales forecasting is a globally conducted corporate practice where number of objectives are identified, action-plans are chalked out as well as budgets and resources are allotted to them. Here in this project, we are going to build the **Sales Forecast Model** that would learn from the past sales records, events and predict the accurate sales so company will be ready to source appropriate resources before the actual event happens.

The Sales Forecast Model will be machine learning model built using python, trained, and tested on Walmart sales data, the detailed use cases and technical information is provided in the paper.

# CONTENTS

[TABLE OF CONTENTS 2](#_Toc90063920)

[1. INTRODUCTION 4](#_Toc90063921)

[1.1 Background 4](#_Toc90063922)

[1.2 Problem Statement 4](#_Toc90063923)

[1.3 Scope 4](#_Toc90063924)

[1.4 Document Overview 4](#_Toc90063925)

[2. Preliminary REquirement 5](#_Toc90063926)

[2.1 Technical Approach 5](#_Toc90063927)

[2.2 Data Sources 5](#_Toc90063928)

[2.3 Analysis 6](#_Toc90063929)

[2.4 Requirement Development 6](#_Toc90063930)

[2.5 Model Deployment 6](#_Toc90063931)

[2.6 Testing and Evaluation 6](#_Toc90063932)

[3. Expected results 6](#_Toc90063933)

[4. Execution and Management of project 7](#_Toc90063934)

[4.1 Project Plan 7](#_Toc90063935)

[4.2 Project Risk 7](#_Toc90063936)

# INTRODUCTION

## **Background**

Walmart, Inc. is part of the retail and wholesale business and is based in Bentonville, Arkansas. The President, Chief Executive Officer, and Director is C. Douglas McMillon. Walmart operates Walmart, Walmart Neighborhood Market, Wal-Mart, Walmart.com, and Sam’s Club. Retail companies commonly have issues with predicting sales accurately throughout the days, months, and years ahead. There are many varying factors that can cause issues with predicting sales such as holidays, economic factors, temperature, fuel prices, Consumer Price Index (CPI), and unemployment. Sales are the lifeblood of business. With an accurate sales forecast in hand, one can plan wisely. If the varying factors are not predicted correctly, then there could be staffing issues at stores, financial implications, and the business could become obsolete if customer satisfaction goes down.

Business Sales Executives often find themselves scrambling for answers when it comes to sales forecasting during business reviews with their leaderships team. The Sales Forecast Model will help sales executives to find such answers upfront and be ready with numbers and predictions to share with leaderships team. This model would help individual stores to upscale their customer satisfaction by stocking the right products at right time and decrease overstocking and wastage of food products.

## **Problem Statement**

The goal of this analysis is to predict future sales for the Walmart stores based on the varying features and events mentioned in the introduction.

In addition, Walmart runs several promotional markdown events throughout the year. These markdowns precede prominent holidays, the four largest of which are the Super Bowl, Labor Day, Thanksgiving, and Christmas. The weeks including these holidays are weighted five times higher in the evaluation than non-holiday weeks.

* Build the Machine Learning model that would learn from past records and predict the accurate outcomes.
* Modeling the effects of markdowns on the holiday weeks in the absence of complete/ideal historical data and predict the sales.

## **Scope**

The project will follow the CRISP-DM methodology and complete all the different phases of it in defined project milestones and each phase of CRISP-DM process is discussed in detail in this paper. The project scope is limited to Walmart Markdown data for different stores and model will be trained and tested on this data.

The plan with this data is to analyze the trends of weekly sales during the time mentioned in the data set (February 5, 2010, through November 1, 2012.). This will help with understanding how the sales will increase or decrease in the future. It will be necessary to analyze and determine which features are the significant predictors for weekly sales, why they are important, and their relationship with one another.

**1.3.1 Assumptions**:

* The Walmart historical data for 40+ different stores will be used for building the Sales Forecast Model.
* The ML algorithms and evaluation criterions will be followed as mentioned in the paper.
* The model will be built using Python programming language and ML algorithms may vary based on outcomes of the models.
* Since data used for training and testing model is historical, it may product slightly different results if used on current data.

**1.3.2 Limitations:**

* The **Sales Forecast Model** can be used for another departmental stores or company data, but result may vary, and data needs to be in same format as Walmart data.

## **Document Overview**

The present document provides information on building **Sales Forecast Model** using python and machine learning algorithms. The document shades light on how this model will be built, the background, the reason of building this model, who would be benefited from this model, what methodology is used to build this model, technical approach, data sources, techniques used to explore the data, model evaluation and testing process. This document has detailed information about each process followed to complete this project.

# Preliminary REquirement

## **Technical Approach**

To build Sales Forecast Model, python programming language and its ecosystems will be used for data collection, EDA, visualizations, model evaluation, deployment etc.

* **Programming**: Python
* **EDA Packages**: pandas, numpy
* **Visualizations**: seaborn, matplotlib, plotly & Tableau
* **ML Algorithms**: Regression Algorithms, random forest, XGBoost, SVM, KNN, ANN.
* **Methodology**: CRISP-DM.
* **Time Series Models**: Auto ARIMA Model and Holt-Winters Model.

Machine learning algorithms mentioned above such as logistic regression, random forest, XGBoost, neural networks (ANN) etc. will be applied to develop and evaluate models for the prediction of sales. Also, as the data have the **date** as a column, we would like to build a model considering the dataset as a Time Series.

## **Data Sources**

The Walmart dataset is retrieved from Kaggle (<https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting> and the data is historical sales data for 45 Walmart stores located in different regions. Each store contains several departments, and we are tasked to predict the department wide sales for each store.

The data ranges from February 5, 2010, through November 1, 2012. This file contains anonymized information about the 45 stores, indicating the type and size of store.

**stores.csv:** This file contains anonymized information about the 45 stores, indicating the type and size of store.

**train.csv:** This is the historical training data, which covers to 2010-02-05 to 2012-11-01. Within this file we will find the following fields:

* Store - the store number
* Dept - the department number
* Date - the week
* Weekly\_Sales - sales for the given department in the given store
* IsHoliday - whether the week is a special holiday week

**test.csv:** This file is identical to train.csv, except we have withheld the weekly sales. Our goal is to predict the sales for each triplet of store, department, and date in this file.

**features.csv:** This file contains additional data related to the store, department, and regional activity for the given dates. It contains the following fields:

* Store - the store number
* Date - the week
* Temperature - average temperature in the region
* Fuel\_Price - cost of fuel in the region
* MarkDown1-5 - anonymized data related to promotional markdowns that Walmart is running. Markdown data is only available after Nov 2011 and is not available for all stores all the time. Any missing value is marked with an NA.
* CPI - the consumer price index
* Unemployment - the unemployment rate
* IsHoliday - whether the week is a special holiday week

**Note** - The four holidays fall within the following weeks in the dataset (not all holidays are in the data):

* Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13
* Labor Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13
* Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13
* Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

## **Analysis**

In this phase, data will be cleaned by following different data preparation methods such as Identifying and handling missing data, handling duplicate data, formatting data to form model can accept it and after this EDA will be performed on the data to understand data using different statistical methods, different chart will be plotted to see distribution, trend, correlation etc. After exploratory data analysis and data preprocessing, we would split data into train, and test and different machine learning algorithms will be trained and evaluated using training data and test data. At the end we would compare all the different algorithms result to finalize the model.

## **Requirement Development**

To train and test different ML algorithms to build model, data will be acquired from appropriate sources and stored in table or flat file, python and supporting libraries will be installed to build Sales Forecast Model. Below is the list of items that would be needed to complete the project.

* The data source should be cited and credible
* Python & ecosystem
* R
* We might use cloud-based notebook e.g., Google colab
* Communication –Teams Meeting minimum two times in a to discuss the plan and progress. Exchange of files, literature and code will be done using the Teams chat or email

## Model Deployment

The finalized model demo will be given to stakeholders and detailed plan will be made on deployment. The model will be deployed to production environment to make real time sales forecasting. The monitoring system will be placed to monitor the model and test report will be generated to determine model performance.

## **Testing and Evaluation**

In model testing phase, we would evaluate the performance of the model using following metrics and model with better result will be finalized.

* Accuracy
* Precision
* F1 Score
* Confusion Matrix

After deployment of model in production environment, the model performance reports will be generated to evaluate the model and based on it, changes will be performed to the model.

# Expected results

Expecting to extract features which have significant impact on predicting the sales. Using this data an accurate model will be developed that can predict future sales.

# Execution and Management of project

## **Project Plan**

This subject is for 12 weeks, so we plan to finish the project in 4 milestones having 3 week per milestones

* Milestone 1 - Identify dataset, describe the objective, scope of the project, and decide what interesting things can be predicted from data set.
* Milestone 2 - Data Wrangling, clean up data, handle outliers, rename any fields, create new fields if required.
* Milestone 3 - Feature extraction and create various Machine Learning models for predictions.
* Milestone 4 - Fine tuning of models till satisfaction and summarize the findings and writing the paper.

## **Project Risk**

There are no specific risks that would impact the project based on the analysis done so far but there could be risk of data based on the model accuracy, additional data may need to pull to train and test model to improve the performance of model.