Project Proposal

Vaishali Maheshwari

[vmaheshwari@regis.edu](mailto:vmaheshwari@regis.edu)

+1 3034725791

**Title Of the Project**: - Walmart Sales Prediction

**High-Level Description: -**

I have taken the Walmart Sales data from “https://data.world/tommywilczek/walmart”. The idea is to explore the two data sets for 45 Walmart stores and weekly sales data. The store dataset has fields store ID, Size and Type of the store. The second dataset “Walmart Sales” has fields Store ID, Temperature, Fuel Price, Weekly sales, Date, Holiday, Department, 5 Markdowns, CPI, and unemployment.

I will find the impact of holidays on the sales of the store. The dataset includes four holiday weeks which are Christmas, Thanksgiving, Super Bowl and Labour Day, I will take these four also into consideration for my analysis. Other fields like temperature, fuel prices etc also impact on the sales of each store or department so will analyse these to predict the results.

The aim is to find out the features impacting sales across Walmart, the stores with maximum sales and analysis based on departments in each store and predictions using machine learning algorithms. I will use different data visualization techniques like Python libraries (Matplotlib, Seaborn, etc.) and Tableau dashboards to present the data.

**Type Of Data Science Task**: -

The plan is to explore all the phases of a Data Science lifecycle as explained below: -

1. **Understanding the business of Walmart** 🡪 This was quite easy as Walmart is a very well-known retail corporation that operates a chain of hypermarkets and grocery stores across the United States.
2. **Understanding the available data sets** 🡪 I explored quite a few data sets and then selected the two datasets for Walmart sales from “Data World”, the first dataset has store details of 45 Walmart stores across different regions of the US.

The other dataset has sales details that are various factors impacting the sales at Walmart. The dataset contains multiple departments, and, in this project, I will depict the sales for each department in the stores. Walmart also runs a few markdown events for promotion throughout the year, the details are available in the dataset and the weeks including these holidays are expected to weigh higher than other working day weeks in terms of evaluation of sales.

1. **Prepare the data** 🡪 This will be the most important phase for my project as I have picked up complete raw datasets and we need to join both the data sets as well. The store number field can be used to join the two datasets.

After joining the datasets, the next step is to clean up the data. The first column has no label so that needs to be corrected or removed if not required for analysis. The date fields need to be converted into multiple columns to check on the month to predict the monthly sales., days with maximum sales etc. Markdown columns have multiple null values that should be filled in appropriately. The negative values available for weekly sales columns should be corrected and temperatures in a few records are extremely high, need to check on those records as well.

I will spend some time on this part and convert the data into a more meaningful dataset before starting the actual analysis.

1. **Data Modelling** 🡪 The idea is to split the data into test and train data sets, figure out the features that will impact the sales in Walmart and perform the EDA. EDA can be performed on data based on markdown fields, store-wise sales, monthly sales etc. can be projected.

The next step is to use machine learning on this data to predict the sales trend. Will train and test the below models: -

1. First will start with a regression model to fit the trained data set and check on the score.
2. Linear regression
3. Random Forest Classifier
4. Gradient Booster algorithm
5. Explore the decision tree for supervised learning.
6. Feature importance, Categorical data interpretation.
7. **Evaluation and Data Visualizations** 🡪 Once the data modelling is completed, will evaluate the results from different models and draw the final predictions. Data Visualization will be another focus area for me, and I will use different data visualization techniques like Python libraries (Matplotlib, Seaborn, Plotly etc.) and Tableau dashboards to present the data in different formats to show the sales trends attractively. Will also include these in my presentation.

**Data: -**

I was able to download the dataset for 45 Walmart stores and sales data with 423326 records. The primary key for applying join on these two datasets is store ID. Using this column will merge both datasets for further analysis.

**Data Analysis: -** I will start with data cleaning and joining both the datasets to link store details with the sales data. Further will train and test machine learning models and evaluate these to predict the most accurate results.

I will use different data visualization techniques like Python libraries (Matplotlib, Seaborn, etc.) and Tableau dashboards to present the data.

**Anticipated difficulties and problems**: - Data clean up and EDA as the data needs to be categorized into high-level categories which will take some time.

**Timelines: -** This is high level and might change if any step takes more time than expected

Week 2🡪 I will work on joining the datasets and start with the data cleaning and EDA.

Week 3 🡪 Once the Data is cleaned, will start working on EDA based on markdown fields and check on the correlation of all the fields in the dataset.

Week 4 🡪 Split the data into test and train datasets. Check on the monthly sales and other analyses.

Week 5 🡪 Complete evaluation of regression model and Linear regression models.

Week 6 🡪 Work on other machine learning models like Random Forest Classifier, Gradient Booster etc.

Week 7 🡪 Work on Data visualizations using Python libraries and Tableau dashboards.

Week 8 🡪 Prepare the presentation.

**GitHub Repository: -** I have uploaded the dataset and proposal to the below repository in GitHub.

https://github.com/vaishaliM01/AmazonProductData\_Practicum1