**Machine Learning – Portfolio Project**

**Description:**

One of the leading retail stores in the US, Walmart, would like to predict the sales and demand accurately. There are certain events and holidays which impact sales on each day. There are sales data available for 45 stores of Walmart. The business is facing a challenge due to unforeseen demands and runs out of stock sometimes, due to the inappropriate machine learning algorithm. An ideal ML algorithm will predict demand accurately and ingest factors like economic conditions including CPI, Unemployment Index, etc.

Walmart runs several promotional markdown events throughout the year. These markdowns precede prominent holidays, the four largest of all, which are the Super Bowl, Labour Day, Thanksgiving, and Christmas. The weeks including these holidays are weighted five times higher in the evaluation than non-holiday weeks. Historical sales data for 45 Walmart stores located in different regions are available.

**Objective:**

* Understand the Dataset, perform EDA and engineer new features if required.
* Build Regression models to predict the sales w.r.t single & multiple features.
* Also evaluate the models & compare their respective scores like R2, RMSE, etc.

**About the dataset:**

The dataset contains historical data that covers sales from 2010-02-05 to 2012-11-01, consisting of 8 features & 6435 samples.

Columns are explained below:

* Store - the store number
* Date - the week of sales
* Weekly\_Sales - sales for the given store
* Holiday\_Flag - whether the week is a special holiday week
* 1 – Holiday week
* 0 – Non-holiday week
* Temperature - Temperature on the day of sale
* Fuel\_Price - Cost of fuel in the region
* CPI – Prevailing consumer price index
* Unemployment - Prevailing unemployment rate
* Holiday Events – on given dates
* 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

**Steps required:**

**We expect you to solve the problem statement by working on these necessary steps:**

1. Data Exploration

* Import the basic libraries
* Importing the dataset
* Check the data types of all the columns
* Check the null values and number of unique values in each categorical feature
* Check the stats of all the numerical columns
* Write a few lines summary of your findings so far.

1. Exploratory Data Analysis (EDA)

* First analyze the distribution of the target variable.
* Visualize the distribution of data by using counter plots.
* Visualize the numeric features using box plot.
* Understand the relationship between all the features by creating pair plots.
* Mention any multicollinearity in the features and outliers detected during EDA process.

1. Data Pre-processing

* Remove any Duplicate rows.
* Impute null values if any.
* Convert categorical Columns to Numeric by using one-hot encoding
* Removal of outlier
* Analyze and write a short summary of the preprocessing done so far, how many rows were dropped and what is the size of the dataset now.

1. Data Manipulation

* Split the data intro training & testing sets
* Feature Scaling (Standardization)

1. Feature Selection/Extraction

* Check the correlation of all features
* Choose which features to keep for model training

1. Predictive Modelling

* Multiple Linear Regression
* Decision Tree Regressor
* Random Forest
* Ensemble Technique
* Mention the R2-Scores Comparison for various Regression Models

1. Project Outcomes & Conclusion

* Write a short summary of your findings through different predictive models and why you’ll chose any specific for future predictions.