To Predict sales using historical markdown data of 45 Walmart stores located in different regions.

Abstract:

The purpose of this research is to construct a sales prediction model for retail stores using the machine learning approach, which has gained significant attention in recent years. Using such a model for analysis, an approach to store management could be formulated. The present study uses two years' worth of point-of-sale (POS) data from different retail stores to construct a sales prediction model. In addition, Walmart runs several promotional markdown events throughout the year. These markdowns precede prominent holidays, a challenge in this study is to take the holidays into consideration by modeling the effects of markdowns on these holiday weeks in the absence of complete/ideal historical data.

1.Business problem

1.1 Introduction

Every Business needs to be able to see their progress and the factors affecting their sales. Predicting future sales for a company is one of the most important aspects of strategic planning and a way to scale progress. Walmart uses its data for time series analysis and prediction methods for forecasting its sales across its different stores across the United States. As per the company's distinguished data scientist and director of data science, John Bowman, Walmart estimates the predictions for a full 52-week horizon in weekly increments, and produces a new set of forecasts every week, over the weekend. Additionally, the company has a 12-hour window to complete all of the forecasting tasks, and about three days to evaluate all of the training tasks. The methods include Deep Learning models as well and Walmart also assesses the forecasts which it has received from its vendors for comparison and evaluation. We shall also consider similar methodology of forecasting such as Time-series forecasting models like ARIMA, Machine learning models like Random forests and XGboost, Deep Learning models like Neural Prophet and SilverKite.

1.2 Challenges

This dataset contains anonymized information about the 45 stores. It has additional data related to the store, department, and regional activity such as store number, date, avg temp. of the region, fuel price per region, Markdown – unidentified promotional data for each store. 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- consumer price index Unemployment - the unemployment rate, IsHoliday - whether the week is a special holiday week.

2. **Methods**

We shall discuss possible methodology under this section for forecasting taking the data into consideration.

- **ARIMA**, short for 'Auto Regressive Integrated Moving Average' is actually a class of models that 'explains' a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values. Any 'non-seasonal' time series that exhibits patterns and is not a random white noise can be modeled with ARIMA models. Because, term 'Auto Regressive' in ARIMA means it is a linear regression model that uses its own lags as predictors. Linear regression models, as you know, work best when the predictors are not correlated and are independent of each other. Hence , we need stationary data. The most common approach to do so is to difference it. That is, subtract the previous value from the current value. Sometimes, depending on the complexity of the series, more than one differencing may be needed.
- **Random Forest** is a popular and effective ensemble machine learning algorithm. It is widely used for classification and regression predictive modeling problems with structured (tabular) data sets, e.g. data as it looks in a spreadsheet or database table. It is an extension of bootstrap aggregation (bagging) of decision trees and can be used for classification and regression problems. In bagging, a number of decision trees are made where each tree is created from a different bootstrap sample of the training dataset. A bootstrap sample is a sample of the training dataset where an example may appear more than once in the sample. This is referred to as "sampling with replacement". Random forest involves constructing a large number of decision trees from bootstrap samples from the training dataset, like bagging. Unlike bagging, random forest also involves selecting a subset of input features (columns or variables) at each split point in the construction of the trees. By reducing the features to a random subset that may be considered at each split point, it forces each decision tree in the ensemble to be more different. The effect is that the predictions, and in turn, prediction errors, made by each tree in the ensemble are more different or less correlated. When the predictions from these less correlated trees are averaged to make a prediction, it often results in better performance than bagged decision trees.
- **XGBoost** is an efficient implementation of gradient boosting for classification and regression problems. XGBoost is short for Extreme Gradient Boosting and is an efficient implementation of the stochastic gradient boosting machine learning algorithm. The stochastic gradient boosting algorithm, also called gradient boosting machines or tree boosting, is a powerful machine learning technique that performs well or even best on a wide range of challenging machine learning problems. It is an ensemble of decision trees algorithm where new trees fix errors of those trees that are already part of the model. Trees are added until no further improvements can be made to the model.

- **NeuralProphet** bridges the gap between traditional time-series models and deep learning methods. With a traditional time series model like ARIMA, all this seasonality would require us to specify orders (look back indices) that corresponds to whatever level of seasonality we observe. The prophet models, on the other hand, automatically encapsulate this sinusoidal motion with Fourier Series', so both Prophet models should effectively leverage the seasonality.
- The **Silverkite algorithm** ships together with the Greykite library released by LinkedIn. It was explicitly designed with the goals in mind of being fast, accurate, and intuitive. The algorithm is described in a 2021 publication ("A flexible forecasting model for production systems", by Reza Hosseini and others). According to LinkedIn, it can handle different kinds of trends and seasonalities such as hourly, daily, weekly, repeated events, and holidays, and short-range effects. Within LinkedIn, it is used for both short-term, for example, a 1-day head, and long-term forecast horizons, such as 1 year ahead. Use cases within LinkedIn include optimizing budget decisions, setting business metric targets, and providing sufficient infrastructure to handle peak traffic. Furthermore, a use case has been to model recoveries from the COVID-19 pandemic. The time-series is modeled as an additive composite of trends, change points, and seasonality.

3. **Approach expected to be followed in this project:**

With the availability of so many algorithms for a forecasting problem it becomes even more crucial to opt for a fast, accurate and apt model for our problem. In this project we shall perform all the necessary data cleaning and then do a mandatory EDA and visualize the problem better. Looking at the results of the EDA we shall decide for the best suited model for our problem.

3.1 **Plausible Metrics**

Since Our project is a problem of forecasting and can be dealt with different models. There are a few Metrics we should compare for the best results such as:

- **Mean Absolute Error (MAE)**

  The MAE is defined as the average of the absolute difference between forecasted and true values. Where yi is the expected value and xi is the actual value (shown below formula). The letter n represents the total number of values in the test set.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

  The MAE shows us how much inaccuracy we should expect from the forecast on average. MAE = 0 means that the anticipated values are correct, and the error statistics are in the original units of the forecasted values.The lower the MAE value, the better the model; a value of zero indicates that the forecast is error-free. In other words, the model with the lowest MAE is deemed superior when comparing many models.However, because MAE does not reveal the proportional scale of the error, it can be difficult to distinguish between large and little errors. It can be

combined with other measures to see if the errors are higher (see Root Mean Square Error below). Furthermore, MAE might obscure issues related to low data volume.

- **Root Mean Squared Error(RMSE)**
  This measure is defined as the square root of mean square error and is an extension of MSE. Where y' denotes the predicted value and y denotes the actual value. The number n refers to the total number of values in the test set. This statistic, like MSE, penalizes greater errors more.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

  This statistic is likewise always positive, with lower values indicating higher performance. The RMSE number is in the same unit as the projected value, which is an advantage of this technique. In comparison to MSE, this makes it easier to comprehend. The RMSE can also be compared to the MAE to see whether there are any substantial but uncommon inaccuracies in the forecast. The wider the gap between RMSE and MAE, the more erratic the error size. This statistic can mask issues with low data volume.
- **Model F1 score**
  It represents the model score as a function of precision and recall score. F-score is a machine learning model performance metric that gives equal weight to both the Precision and Recall for measuring its performance in terms of accuracy, making it an alternative to Accuracy metrics (it doesn't require us to know the total number of observations)
  F1 Score = 2* Precision Score * Recall Score/ (Precision Score + Recall Score/)

  F1-score is harmonic mean of precision and recall score and is used as a metrics in the scenarios where choosing either of precision or recall score can result in compromise in terms of model giving high false positives and false negatives respectively.

4. **Dataset**

The data comes in 4 .csv files : stores.csv, train.csv, test.csv, features.csv

- 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.
- test.csv
  This file is identical to train.csv, except we have withheld the weekly sales. You must 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

4.1 Source of Dataset

For this project we are going to use the data set provided on the Kaggle website where this problem has been posted. The dataset consists of actual store data provided Walmart and hence has all the scenarios of noise as well anonymized data.

Link for the data set:

https://www.kaggle.com/competitions/walmart-recruiting-store-sales-forecasting/data

5. **Expectations**

Through this project we aim to accurately predict the sales for several Walmart stores located in different regions. In doing so we can provide important insights for the strategic management/ planning of the company. A sales forecast helps every business make better business decisions. It helps in overall business planning, budgeting, and risk management. We can figure what features of the store really affect the sales, i.e the store size, store type, the department or any other factor. Decision makers rely on these forecasts to plan for business expansion and to determine how to fuel the company's growth. So, in many ways, sales forecasting affects everyone in the organization.

6.**References**

https://www.kaggle.com/code/yasirhussain1987/eda-and-store-sales-predictions-using-xgb/notebook

https://www.researchgate.net/publication/228255537_Sales_Forecasting

https://gallery.azure.ai/Experiment/Walmart-Sales-Forecasting-Using-Regression-Analysis-1

https://www.researchgate.net/publication/328246040_Walmart's_Sales_Data_Analysis_-_A_Big_Data_Analytics_Perspective

https://forecasters.org/wp-content/uploads/gravity_forms/7-c6dd08fee7f0065037affb5b74fec20a/2017/08/Seaman_ISF-2017.pdf

https://www.projectpro.io/project-use-case/walmart