Walmart Weekly Sales Forecasting

Sasmit Khokale | Capstone Project | 2/23/2019

PROJECT OVERVIEW

Retail giant Walmart is an American multinational retail corporation that operates a chain of hypermarkets, discount department stores, and grocery stores. For retail of this scale it is crucial to have best possible demand numbers or sales forecast available at each given time. For better forecasting it is important to invest in data science efforts.

Knowing better forecast numbers will help Walmart in following ways –

1) Balancing demand and supply, this consists of replenishing products at right times, right places to ensure availability when customer needs it.
2) Plan in advance by partnering with manufacturers to have the products available at given time

Forecasting can be achieved based on historical sales, and the data for this is available on Kaggle - https://www.kaggle.com/iamprateek/wallmart-sales-forecast-datasets

Also, here are academic papers where machine learning concepts are applied for similar problem

1) http://cs229.stanford.edu/proj2015/216_report.pdf
2) https://www.mdpi.com/397796

PROBLEM STATEMENT

For this project, goal is to use Walmart's historical sales data to develop models that predicts future sales, steps include –

1) Download data from available source
2) Perform Exploratory analysis by visualizing and analyzing important dimensions
3) Preprocess data by adding more dimensions as needed
4) Train different models using preprocessed data and choose best model
5) Predict sales for future using test data
6) Use test data to analyze final model performance

We can apply supervised learning to this use case; specifically this is a regression problem where we will be predicting future sales based on other known factors

DATASETS AND INPUTS

Our target variable for this model will be weekly sales (future) and Input is going to be several features from given data and derived based on given data set –

Continuous Variables -

1) Historical sales
2) Temperature
3) Fuel Price
4) Mark down price (if any)
5) CPI – Consumer Price Index
6) Unemployment rate

Categorical Variables -

7) Whether there was holiday or not
8) Store Type
9) Markdown exist (Yes/No)

We have total about 421k rows in available data which is historical info for 36 stores and depts.

SOLUTION STATEMENT

Preprocessing and Model Preparation – In this process I will consider if we can derive additional dimensions based on given data and if we need any more information to come up with best predictions

Model Creation – During this phase I will explore effects of several models such as Linear Regression, KNeighbors Regression and Random Forest Regression the on given data.

Using model tuning techniques I will identify model and parameter values which gives best values for evaluation metrics

BENCHMARK MODEL

Here, as a benchmark model we will run simple linear regression on given data, results of which will be used as base results and later we will use more advanced models such as K Neighbors Regressor and Random Forest for which we are hoping to have better results as compared to results from our base model.

EVALUATION METRICS

Root Mean Squared Error is a common evaluation metric in this context as we are trying to predict the value which should be equal to or very close to the actual value.

RMSE = $1/n \sum$ (Predicted value – Actual value)^2

PROJECT DESIGN

EXPLORATORY DATA ANALYSIS

Perform exploratory analysis on available data to find:

- Data distribution
- Important features
- Missing data

PREPROCESSING AND DATA PREPARATION

Here I will merge given data files if needed.

Handling missing data – Based on data quality I will explore if I need to infer any missing data.

Dummy Variables – I'll create dummy variables for all the categorical data

Feature Creation – I plan to add the previous week's sales data as an additional feature to the model

MODEL CREATION

As benchmark model I will first run linear regression on preprocessed data and results of which will be used as base results, all the other results later in the process will be compared to these base results and aim will be to improve the results as compared to base numbers.

Later I will apply various models on preprocessed data such as Linear regression, KNeigbors regression, random forest etc. and will tune the model parameters.

MODEL EVALUATION

For model evaluation I will use RMSE as evaluation metric.

PREDICT TEST DATA VALUES

Final model will then be used to predict sales for future timeframe given in test data set.

COMPARE WITH BENCHMARK MODEL

These results will then be evaluated with respect to those from base model by comparison