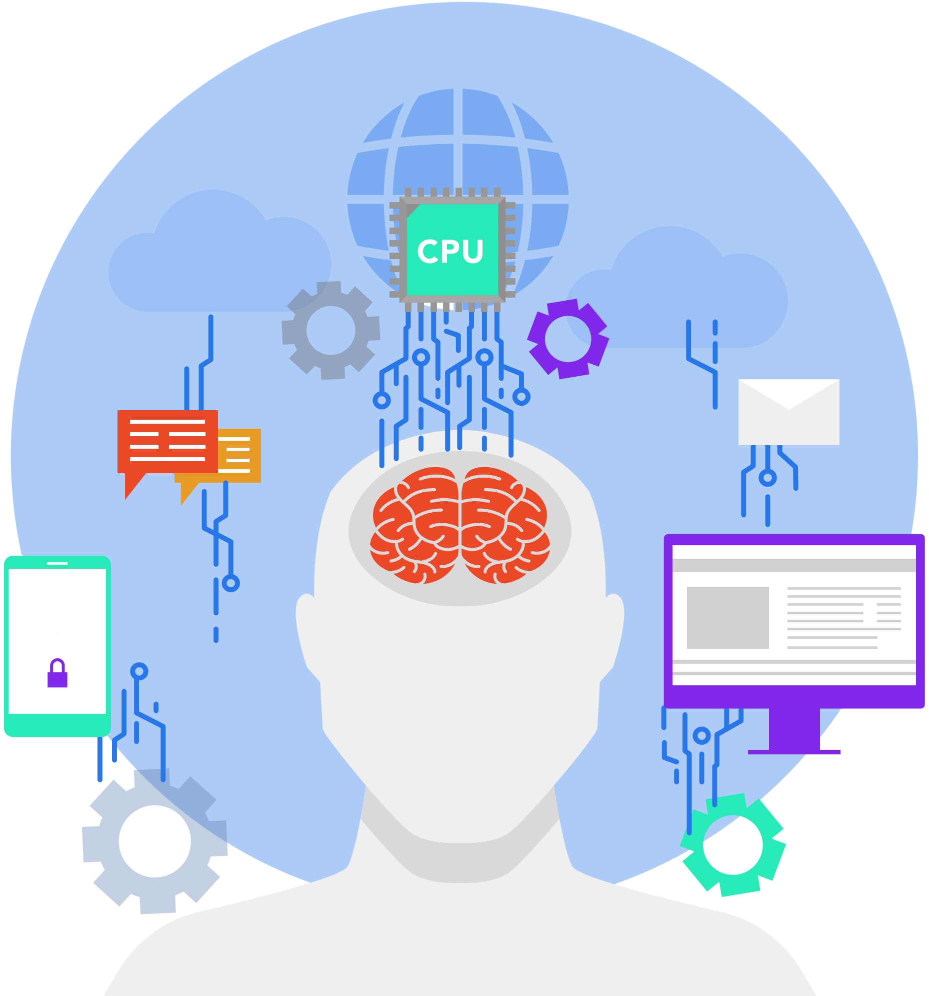


Walmart Sales Prediction

A sales forecasting using machine learning

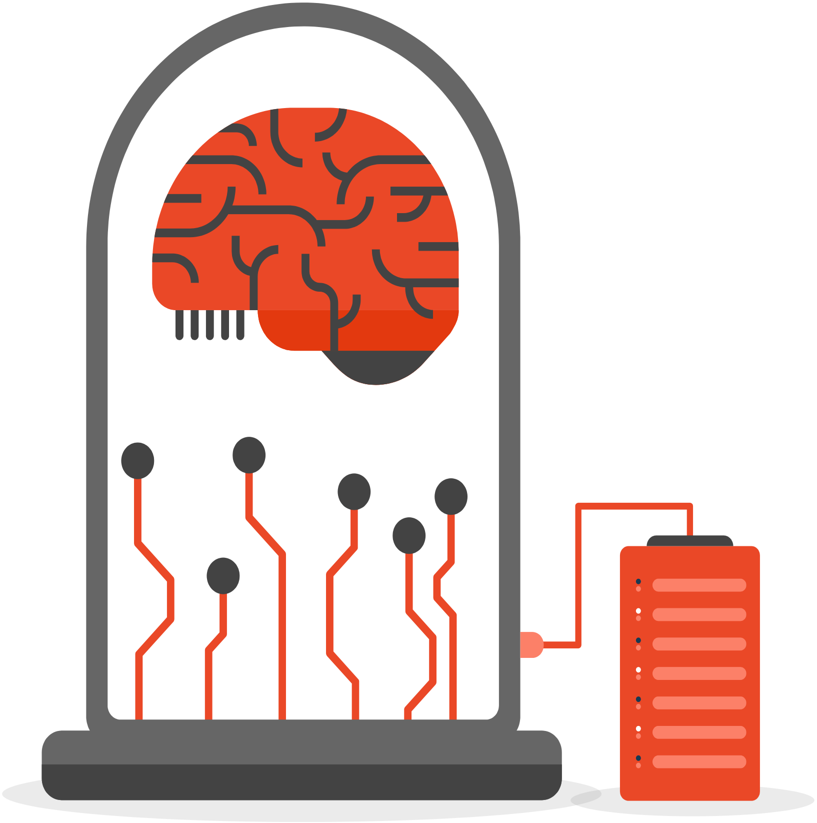
- **Suraj Kumar Prajapati**

Machine Learning

What is Machine Learning ?
Machine Learning is subset of artificial intelligence that mainly concerned with the development of algorithms which allows a computer from a past experiences on their own.

What are the fields where machine learning is used ?
Currently weather prediction, climate change detection, disease detection and recommendation etc. are the various fields where machine learning found very useful.
Most of the businesses are also using Machine Learning as Detection fraud transactions, Customer segmentation and churn prediction etc.

Sales Prediction is also a business task, can be done using machine learning.



Identification of Business Problem

Sales forecasting is process of prediction how much revenue a company will generate within a timeframe.

01

An accurate sales forecasting allows companies to efficiently allocate resources for future growth and manage their cash flows.

02

Sales prediction facilitate strategic planning and tell us how soon we will be able to execute and implementing our plans.

Identification of Data

To create a sales forecasting model, I use dataset **(6435 x 8)** of USA based walmart stores for the time period of 2010 - 2012.

The sales data available for **45 store** of walmart. There are the features of **CPI (Cumulative Price Index), Unemployment** which consider the economic condition of the areas.

For a store sales the **temperature** and **fuel price** also make impact to the purchasing value of the stores.

Predicted **sales** are weekly and there is a holiday column which shows the **holidays** in a week.

Follow the link for dataset - Kaggle Link of Dataset

Roadmap

Data Exploration

01

Explore data to check the present abnormalities

Preprocessing

02

Do preprocessing to make data for model training

Models

03

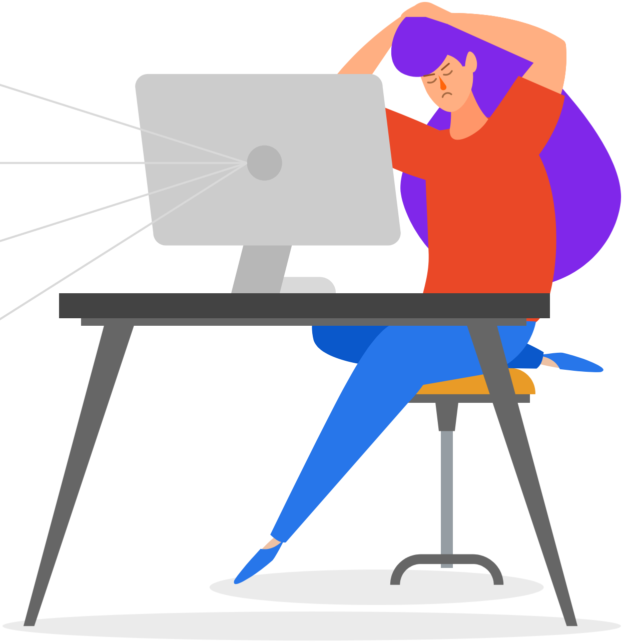
Using various scikit learn models to make predictions

Evaluation

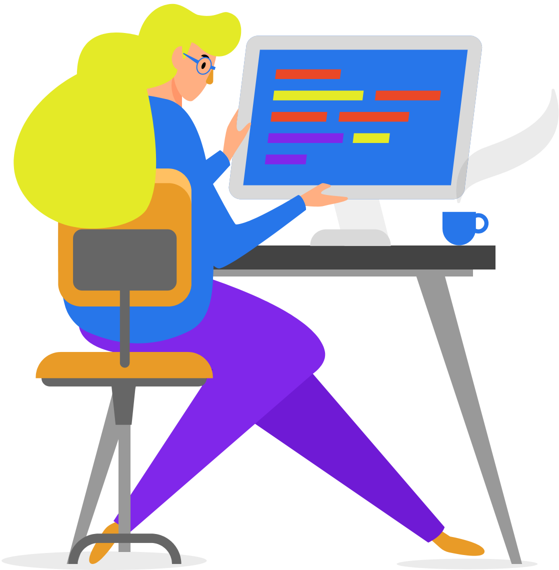
04

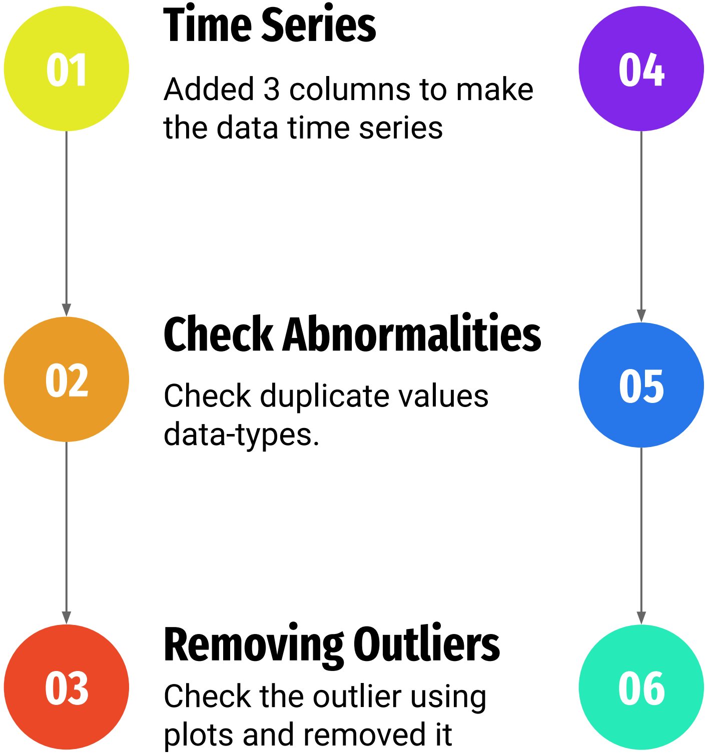
Evaluate the output of the model and make improvements

Model Creation Path



Preprocessing of Data





Time Series

Added 3 columns to make the data time series

Check Abnormalities

Check duplicate values data-types.

Removing Outliers

Check the outlier using plots and removed it

Encoding

Converted the categorical columns into numerical

Standardization

Normalize the values of all features.

Feature Selection

Select the important features from the pool of features

Train - Test Split

Dataset

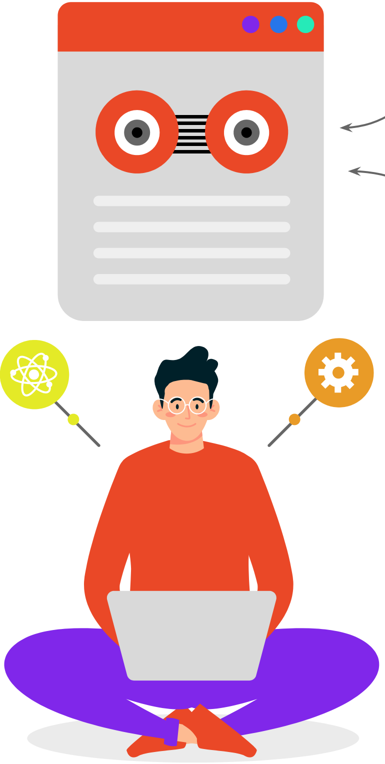
Out of all observations 80% are used to train the model and 20% is used to test the model.

Training Dataset

It has 4760 observation and used for training purpose

Test Dataset

It has 1191 observations and used to test the model performance



Input Features

There is total 68 features used as an input to the model after preprocessing

Target Feature

It is the weekly sales of the corresponding store

Machine Learning Models

Tried various models out of which Linear Regression and XGBoost model gives the best accuracy on train and test both dataset.

Linear Regression

- It is a linear approach for modelling the relationship between scalar response and one or more variables.
- It creates n-1 dimensional plane using n dimensional dataset which fit the given distribution.
- At the end the optimal weights corresponding to the features are used to predict the values.
- Accuracy : Training : 93%
  Testing : 92%

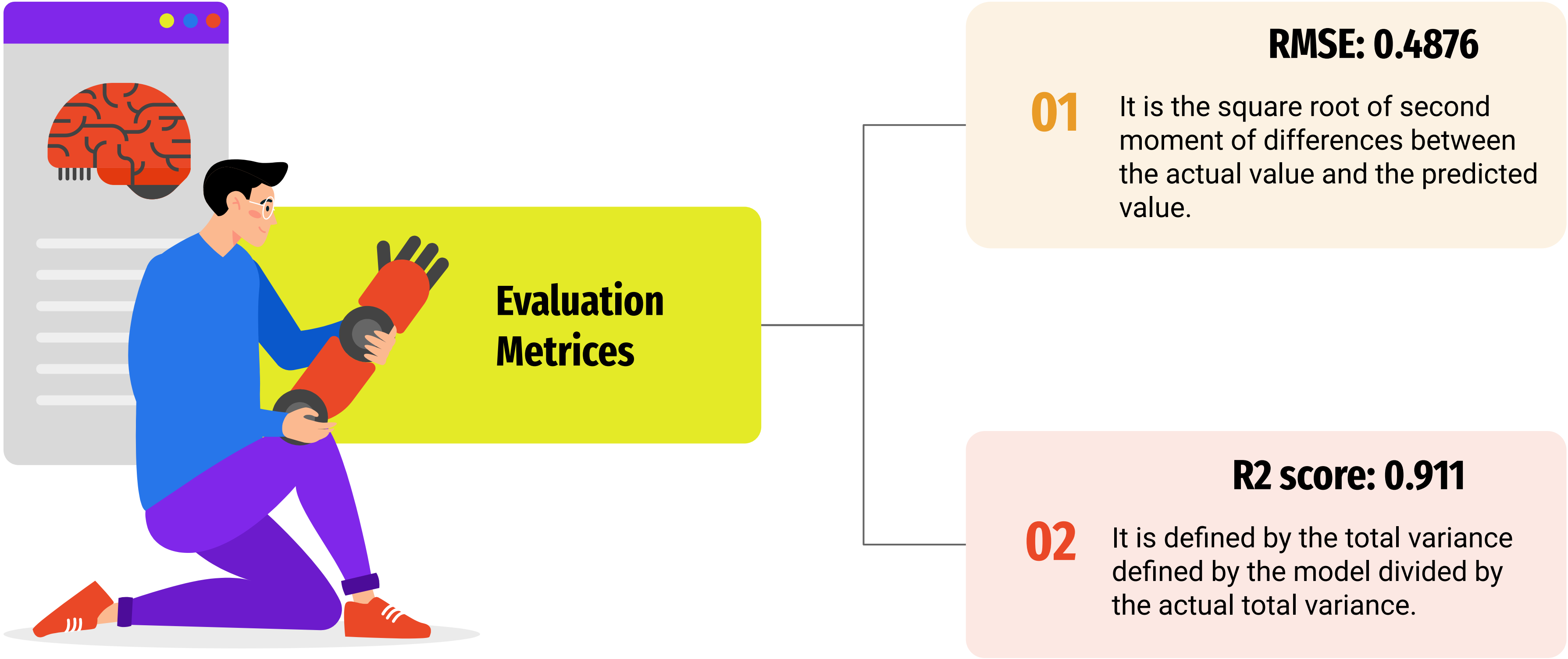
Vs

XG Boost Regressor

- It is a ensemble learning based model which creates a strong regression model using number of weak models.
- Firstly it built a base model on the training data.
- Then second model is build, which tries to correct the error present in the first model.
- Accuracy : Training - 96%
  Testing - 91%

Evaluation

Due to robustness of the XGBoost model, choosed xgboost regressor as optimal model for the prediction



Evaluation Metrices

RMSE: 0.4876

01 It is the square root of second moment of differences between the actual value and the predicted value.

R2 score: 0.911

02 It is defined by the total variance defined by the model divided by the actual total variance.

Experiments and Adjustment

Number of input features

After the preprocessing there are total 78 feature columns, out of which 10 found useless and get removed. This is done by varying the number of features parameter of RFE model

Maximum depth

We tried xgboost model with various depth of ranges and found depth 9 gives the best accurate prediction

Degree of polynomial

In the generation of polynomial model, i used various degrees of polynomial to test the various.

Thank You