Roll No: 41453

Data Analytics: Assignment 3

Title:
Bigmart Sales Analysis

Problem Statement:
For data consisting of transaction records of a sales store. The data has 8523 rows of 12 variables. Predict the sales of a store.

Learning Objectives:

- Learn Regression algorithms
- Learn to summarize the properties in the training dataset.
- Learn to split the dataset into training and test datasets.
- Learn to develop a predictive regression model

Learning Outcome:

Be able to :

develop a predictive model for sales of an item at BigMart.

Software and Hardware Requirements :

- 64 bit CPU
- 4 GB RAM
- Ubuntu 18 OS
- Anaconda Environment
- Jupyter Notebook
- Python Libraries : Pandas, Sklearn, Matplotlib, Numpy

Steps to run:

- Start Ubuntu 18
- Install Anaconda on your PC from https://docs.anaconda.com/anaconda/install/linux/
- Create new environment using : conda create --name env_name
- python=3.5
- Install requires libraries, matplotlib, pandas using : conda install PACKAGENAME
- Open jupyter notebook using new enivironment in terminal
- Run the ipynb in jupyter notebooks.

Theory:

Linear Regression

In statistics, linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.

The relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models.

If the goal is prediction, forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.

Given a dataset of n statistical units, a linear regression model assumes that the relationship between the dependent variable y and the p-vector of regressors x is linear. This relationship is modeled through a disturbance term or error variable ε — an unobserved random variable that adds "noise" to the linear relationship between the dependent variable and regressors.

Dataset = $\{y_i,\, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$

Model Equation :

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^\mathsf{T} \boldsymbol\beta + \varepsilon_i, \qquad i = 1, \ldots, n,$$

Matrix Notation :

$$\mathbf{y} = X\boldsymbol\beta + \boldsymbol\varepsilon,$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$X = \begin{pmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Dataset Description

The data scientists at BigMart have collected 2013 sales data for 1559 products across 10 stores in different cities. Also, certain attributes of each product and store have been defined. The aim is to build a predictive model and find out the sales of each product at a particular store.

● Item_Identifier: Unique product ID

● Item_Weight: Weight of product

● Item_Fat_Content: Whether the product is low fat or not

● Item_Visibility: The % of total display area of all products in a store allocated to the particular product

● Item_Type: The category to which the product belongs

● Item_MRP: Maximum Retail Price (list price) of the product

● Outlet_Identifier: Unique store ID

● Outlet_Establishment_Year: The year in which store was established

● Outlet_Size: The size of the store in terms of ground area covered

● Outlet_Location_Type: The type of city in which the store is located

● Outlet_Type: Whether the outlet is just a grocery store or some sort of supermarket

● Item_Outlet_Sales: Sales of the product in the particular store. This is the outcome variable to be predicted.

Handling missing Data :

Predicting The Missing Values

Using the features which do not have missing values, we can predict the nulls with the help of a machine learning algorithm. This method may result in better accuracy, unless a missing value is expected to have a very high variance. We will be using linear regression to replace the nulls in the feature 'age', using other available features. One can experiment with different algorithms and check which gives the best accuracy instead of sticking to a single algorithm.

Pros:

- Imputing the missing variable is an improvement as long as the bias from the same is smaller than the omitted variable bias
- Yields unbiased estimates of the model parameters

Cons:

- Bias also arises when an incomplete conditioning set is used for a categorical variable.
- Considered only as a proxy for the true values.

Autocorrelation:

Autocorrelation is a mathematical representation of the degree of similarity between a given time series and a lagged version of itself over successive time intervals. It is the same as calculating the correlation between two different time series, except autocorrelation uses the same time series twice: once in its original form and once lagged one or more time periods.

Template Matching:

Template matching is a technique in digital image processing for finding small parts of an image which match a template image. It can be used in manufacturing as a part of quality control, a way to navigate a mobile robot, or as a way to detect edges in images.

Data Mart:

A data mart is a structure / access pattern specific to data warehouse environments, used to retrieve client-facing data. The data mart is a subset of the data warehouse and is usually oriented to a specific business line or team.

Conclusion:

Used various regression techniques to develop a predictive model for sales of items in a BigMart.