Page
Date : / /

Assignment : C3

- TITLE : Bigmart Sales Analysis
- Problem Statement : Bigmart Sales Analysis: For data comprasion of transactions records of a sales store. The data has 8523 rows of 12 variables. Predict the sales of a store.
- Objective : i] To understand the sales analysis and prediction.
  ii] To understand the concept of supervised learning in Data Analysis.
- Outcome : i] learn the supervised learning Algorithm
  ii] Understand sales prediction.
- Software and Hardware Requirement : OS: windows / ubuntu distribution 4GB RAM, 500GB HDD, Python libraries, Python framework. Bigmart sales DataSet.
- Theory :

Bigmart Sales Analysis :

According to the information provided, Bigmart is a big supermarket chain, which stores all around the country and its current board set out a challenge to all data scientist out there, to help they create a model that can predict the sales per product, for each store. Bigmart has

Page :
Date : / /

Collected sales data from the year 2013 from 1559 products across 10 stores in different cities. With this information the corporation hopes we can identify the products and stores which play a key role in their sales and uses that information to take the correct measures to ensure success of thier business.

we will explore the problem in following steps:

1.] Hypothesis generation:

This is a very pivotal step in the process of analysing the data. This involves understanding the problem and making some hypothesis about what could potentially have a good impact on the outcome. This is done before looking at the data and we end up creating a laundary list of the different analysis which we can potentially perform if data is available.

2] Data Exploration: we will be performing some basis data exploration here and come up with some inferences about the data. we will figure out some irregularities and address them in the next session. The first step is to look at the data and try to identify the information which we hypothesized vs the data available. A comparision between the data dictionaries on the competition page and out hypothesis. we will invariable find features which are hypothesized but data doesn't carry and vise versa.

train = pd.read_csv("train.csv")
test = pd.read_csv("test.csv")
train['source'] = train
test['source'] = test
data = pd.concat([train, test], ignore_index = True)

Some observations:

1. Item visibility: has a min value of zero. This makes no practical sense because when a product is being sold in a store, the visibility cannot be 0.
2. Outlet_Establishment_Years: vary from 1985 to 2009. The values might not be apt in this form.

3. Data cleaning: This step typically involves imputing missing values and treating outliers. Though outlier removal is very important in regression techniques, advanced tree based algorithm are impervious to outliers. Imputing Missing Values: we have found two values with missing values: Item_weight, Outlet_Size.

4. Feature Engineering: we explored some nuances in the data in the data exploration section. we will create some new variables using existing ones in this section:

Step 1: Consider Combining Outlet_Type:

During exploration we decided to consider combining the supermarket Type 2 and Type 3 variables.

Step 2: Modify Item_Visibility: we noticed that the minimum value here is 0 which makes

no Pratical Sence. we can use the visibility_avg_item variable to achive this.

Step 3: Create a broad category of Type of item.
Earlier we saw that the Item_Type variable has 16 categories which might prove to be very usefull in analysis. So its good idea to combine them. If you look at the Item_identifier i.e. the unique ID of each item it starts either FD, NC, or DR. If you see the categories these look like beign Food, Drinks and Non-consumables.

Step 4: Determine the years of operation of a store:
we wanted to make a new column depicting the year of operation of store

Step 5: Modify Categories of Item_Fat_Content
we found types and difference in represen-tation in categoreis of Item_Fat_Content variables.

Step 6: Numerical and One hot Coding of Categorical values: I have converted all categories of nominal variables into numeric types. Also I wanted Outlet_Identifier as a variable, as well. One-hot coding refers to creating dummy variables, one for each category of a categorical variable. For example the Item_fat_content.

Page :
Date : / /

- Test cases :

	Description	Result Values	Pass/fail
i)	Imputing Missing Values : Item-weight outlet-size	Item_weight : original missing : 2439 Final missing : 0	Pass
		outlet-size : original missing : 4016 Final missing : 0	Pass
ii)	Item-visibility	original zeros : 879 Final zeros : 0	Pass
iii)	Model Training	linear regression : accuracy score : 56 Decision tree : accuracy score : 62 XGB Regressor : accuracy score : 67	Pass

- Conclusion : We are successfully done bigmart Sales analysis and predicted the Sales of a Store.