BIG MART SALE PREDICTION

PRESENTED BY:

JOSEPH DAVIS
&
ROHIT MATHEW

DATA OVERVIEW

According to the information provided, Bigmart is a big supermarket chain, with stores all around the country and its current board set out a challenge to all Data Scientist out there to help them create a model that can predict the sales, per product, for each store. BigMart has collected sales data from the year 2013, for 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 use that information to take the correct measures to ensure success of their business.

SPECIFICATIONS

Variable	**Description**
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.

GOAL

The aim is to build a predictive model and find out the sales of each product at a particular store

APPROACH

Hence, this analysis will be divided into five stages:

1)Exploratory data analysis (EDA)
2)Data Preprocessing
3)Modeling
4)Hyperparameter tuning

Exploratory Data Analysis (EDA)

We've made our first assumptions on the data and now we are ready to perform some basic data exploration and come up with some inference. Hence, the goal for this section is to take a glimpse on the data as well as any irregularities so that we can correct on the next section,ie Data Pre-Processing

Data Preprocessing

We've seen previously on the EDA section that the itemweight and the outletsize had the missing values.Hence we impute missing values with mean and mode for each corresponding variable.Since the data contains outliers.So we found upper limit and lower limit using IQR method.We replaced values higher than Q1 with Q1 and lower than Q2 with Q2 where Q1 and Q2 are upper limit ,lower limit respectively. Since scikit learn only accepts numerical variables, we did convert all categories variable into numeric types

Modelling

First we split our data into training and testing.We assign test size as 0.33, so 67% of data used for training and 33% of data used for testing. Then we import our regression model from scikitlearn tool kit.After that we fit our model using trained data

Hyperparameter tuning

We import Gridsearch cv for parameter tuning and found out the best parameters that fits our model

MODELS USED

1. Logistic Regression
2. Ridge Classifier
3. Random Forest Classifier
4. XGBoost

RESULT

The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed

$$RMSE=\sqrt{\frac{\sum_{i=1}^{N}(y-x)^2}{N}}$$

RMSE-Root mean square error i- variable N-number of non missing datapoints

y-predicted variable x-actual variable

Random forest has the least Root Mean Square Error.RMSE= 964

So we select that Random Forest as our final model