Big Market Sales Prediction

Problem Statement: Details and characteristics of 1559 products and 10 outlets were given and we need to predict sales for given product in the given outlet.

Data: The data contains information about product and outlets. (7 categorical columns)

Product- Weight, fat content, visibility, Type and MRP

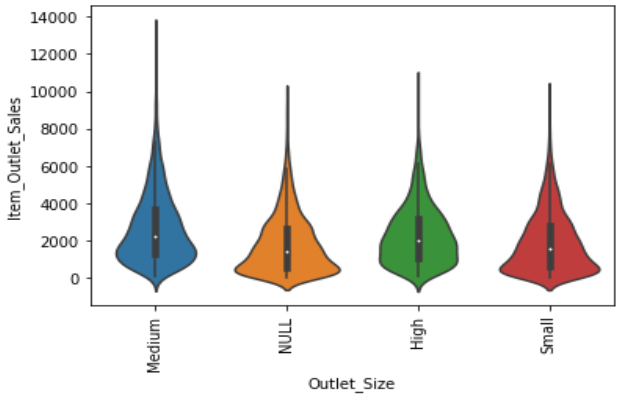
Outlet- Size, location, type and establishment year

Target Variables: Item sales

What can be added: Seasonality (Vaseline is more sold in winters, Powder is more sold in summers), There are some occasional Products ed. Celebration pack during Raksha Bandhan

Data Pre-processing:

1. Handling missing values: There were missing values in item weight and outlet size. As weight was continuous variables, we replaced the null values with mean as it is a standard practice.
   Outlet size is categorical values hence we can replace it by mode. But when we see the stripe plot for categories in Outlet size and Item sales, we see same distribution in small size outlets and null values. Hence, it would be feasible to replace all the null values by ”small”.



2. Data Transformation: As it’s a problem of regression and if we are applying linear regression, with respect to X, Y must follow normal distribution. Hence we used log transformation and cube root transformation to convert the distribution to somewhat normal
3. Checking Correlation between features: Correlation was checked between numerical as well as categorical variables. We need to ensure that there no two features which are highly correlated and it would violate the assumption of Linear regression and may generate misleading results.
   For categorical variables we used chi square test to reject the null hypothesis.
4. Feature Engineering: used Label Encodings (Outlet size) and One-hot encodings (fat content)

5. Normalization: Feature scaling is important step as it saves computation time. Hence normalized the data.

Model Training: I have trained Linear regression, XG Boost and random forests.

Optimization for Random Forests: n_estimators: Number of trees

Optimization for XG Boost: n_estimators and Learning rate.

Evaluation Matrix: As it's a regression problem we can use Root mean squared error or Mean Absolute error. We can also use R^2 Error.

Important Key Concepts: Imputation, Box Plot, quantiles, Distribution plot, log transformation, cube root transformation, Correlation, Chi-square test, p-values, Significance level, Encodings, Normalization, Linear regression, Bagging and Boosting techniques, hyperparameter optimization for Random Forests and XG Boost, R^2, RMSE.
