Walmart Sales Prediction

Rupali Shah | Springboard | May 31, 2019

Overview

Sale Forecasts:

Reliable sales forecasting is essential for a business to enable it to produce the required quantity at the right time. Further, it allows to make the arrangement in advance for raw materials, equipment, labor, etc. Some firms manufacture on the order basis, but in general, firms produce the material in advance to meet the future demand.

Methodologies for deriving a forecast

Companies can base their forecasts on historical data, industry-wide comparisons, custom surveys, competition, economic trends, or a combination of the above. In general, statistics models can be fed with such data to forecast sales.

This project uses historical data to gain an insight into the sales of products at Walmart stores, and builds a model to predict future sales.

The variables analyzed within this set of data include:

- Price
- Fat Content
- Store visibility/placement
- Store Location
- Period/Year
- Item Type
- Outlet Type and Size

Data Source

This dataset was downloaded from Kaggle, and it provides information on the historical sales data for Walmart stores in different cities for the year 2013. It has two files in CSV format:

- train.csv - The train dataset contains 11 independent variables and 1 target variable.
- test.csv - The test dataset contains the same set of independent variables but no target variable because that variable will be predicted.

Data Wrangling

- The train and the test datasets were first imported into a Pandas Dataframe, and then merged into one Dataframe. The combined dataset has 14204 records and 13 columns.
- The Fat content of the Items has different names for the same categories. They were combined into Low Fat and Regular.
- The establishment year of the outlets was subtracted from the current year, and added as a new variable, Outlet_Year.
- There are some items with zero visibility. Perhaps these items are placed on top shelves or far behind other items, and not visible. So, the value was left unchanged.
- The item type was categorized into three broad categories : Food, Drink, Non-consumable.
- The missing values of the Item Weight were imputed based on the weight of other similar items, and those of the Outlet Size were replaced based on the Outlet Type.
- One-hot encoding was performed on the Categorical variables.
- Columns that were grouped into new features were deleted.
- The cleaned data was split into train and test datasets and saved as a csv file. The Outlet Sales column was deleted from the test dataset.
- Size of the new data:

Dataset	Number of Records	Number of Columns
Train	8523	14
Test	5681	13

Description of variables in the dataset

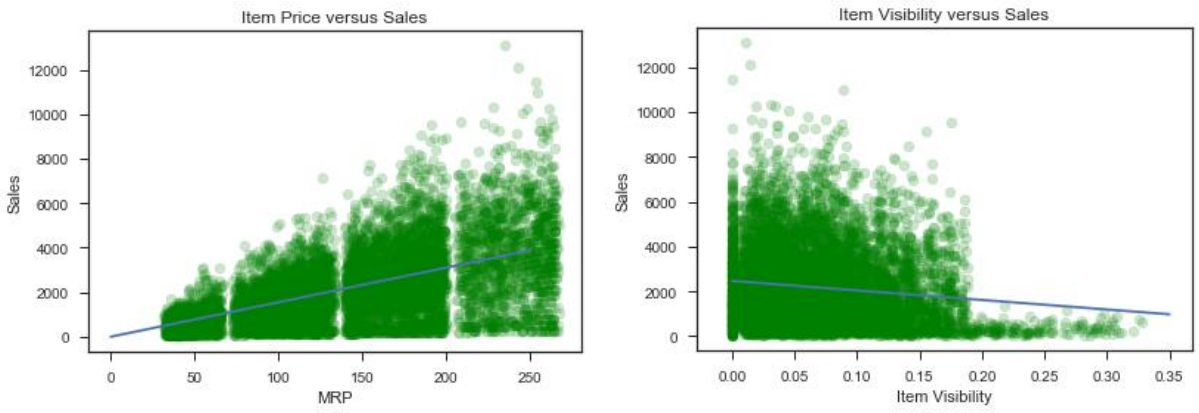
Item_Identifier	Unique product ID
Item_Weight	Weight of product
Item_Fat_Content	Whether the product is low fat or not
Item_Visibility	The % of total display area of all products in a store allocated to the product
Item_Type	The type of item (dairy, fruits, household items,..)
Item_MRP	Maximum Retail Price (list price) of the product
Item_Category	The category to which the product belongs (Food, Drink, Non-Consumable)
Outlet_Identifier	Unique store ID
Outlet_Establishment_Year	The year in which the 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 particulat store. This is the outcome variable to be predicted.
Outlet_Age	Age of the outlet

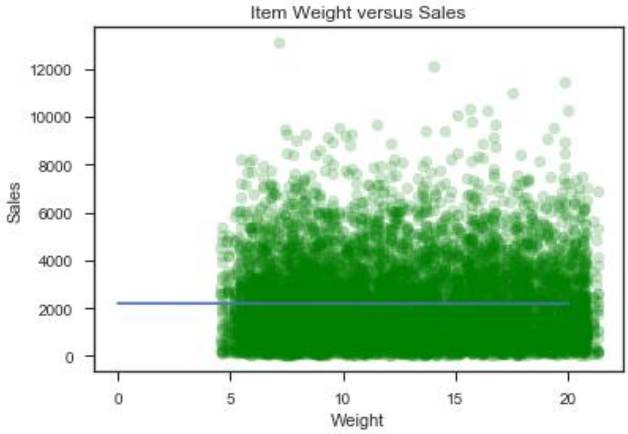
Exploratory Data Analysis

Our data questions

1. Are the higher income levels of people in urban cities driving the sales of Tier 1 stores?
2. Do supermarkets have higher sales because customers prefer to shop most of the items from one place?
3. Do the older stores have better sales because of customer loyalty?
4. Does product visibility impact sales?
5. Do necessities sell more than other items?

Relationship between item price, visibility, weight AND sales

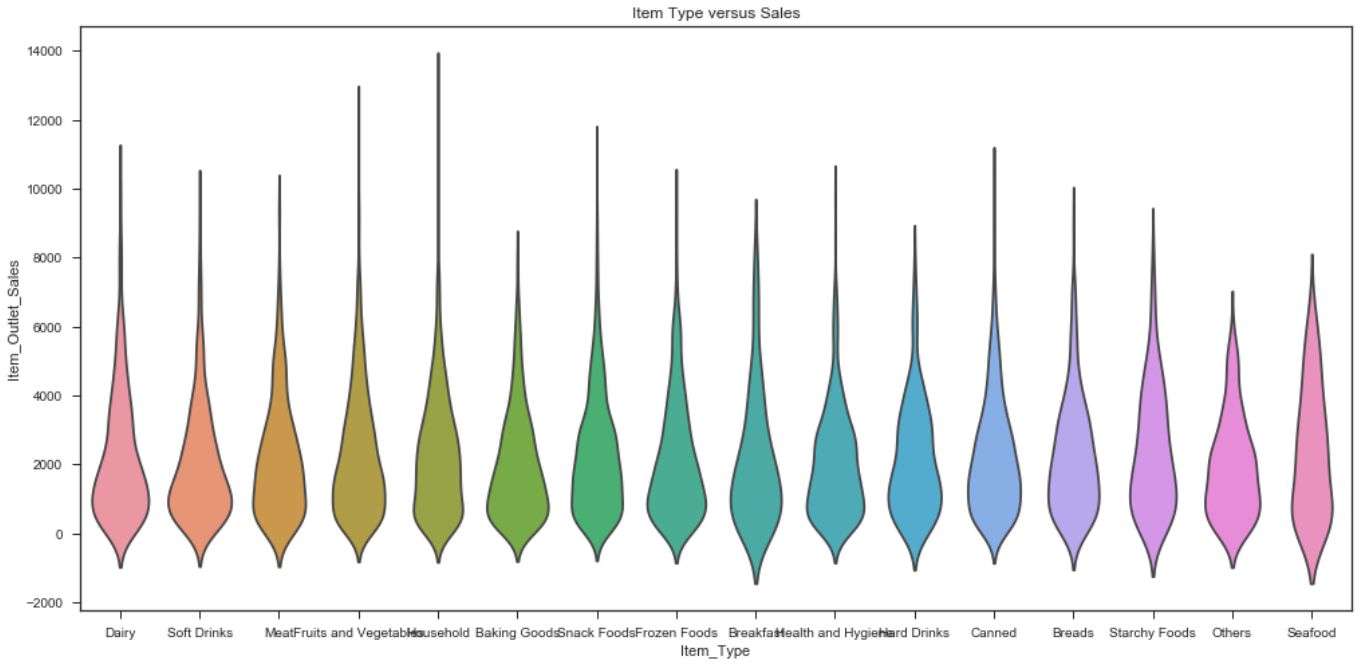


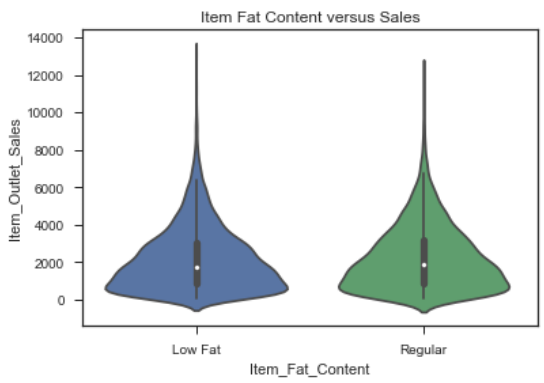


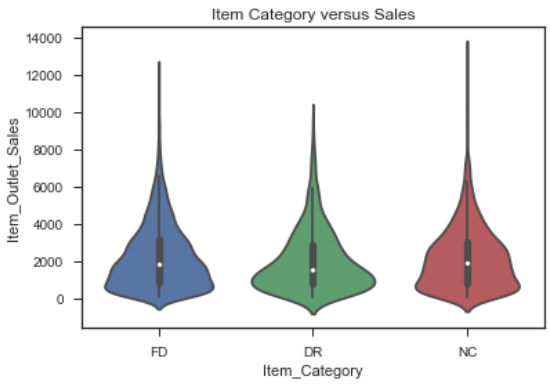
- The item price shows four different groups of prices .
- Items with very high visibility are seen to have lower sales.
- There is no clear trend between the weight of the item and its price. The sales are spread across the entire range of weight.

Distribution of sales across all categories

The distribution of sales for the different categories was visualized with a violin plot. The width of the violin plot indicates the concentration of data at that level. The height shows the range of sales.

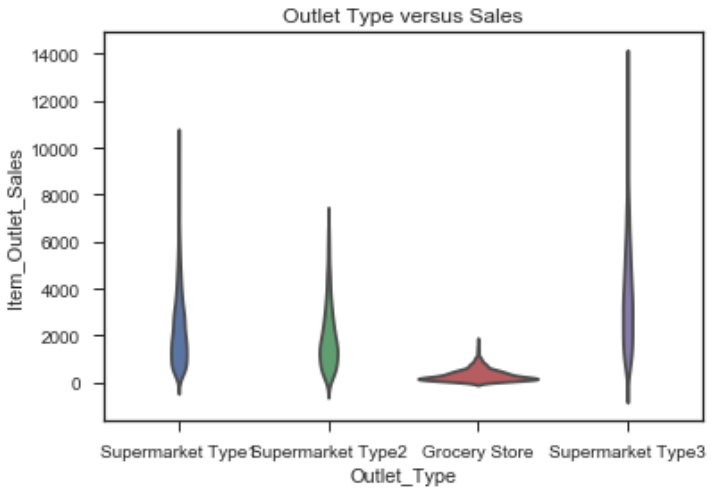




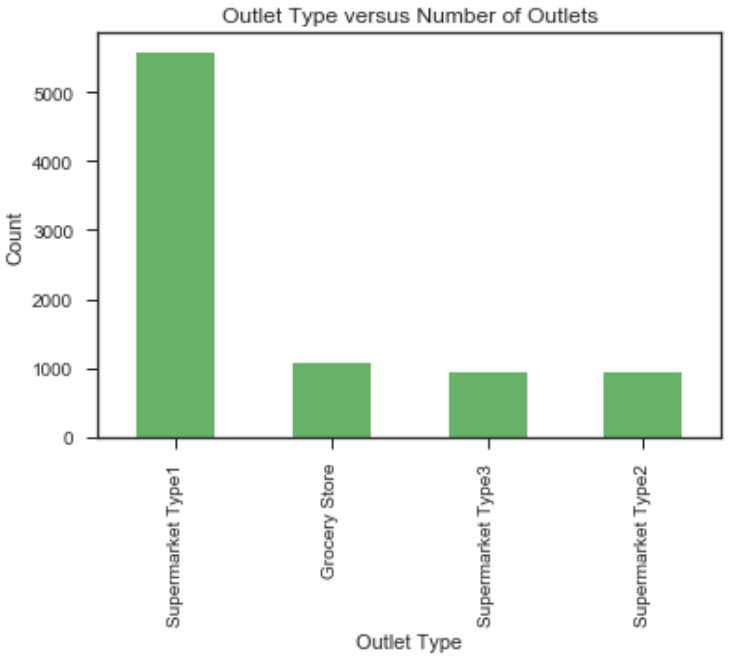


- The distribution of sales across the item types is not very distinct.
- The distribution of sales of Regular and Low Fat items is very similar.
- Drinks have a wider distribution and lower sales as compared to food and non-consumable items. Non-consumables have the largest range of sales.

Outlet Type

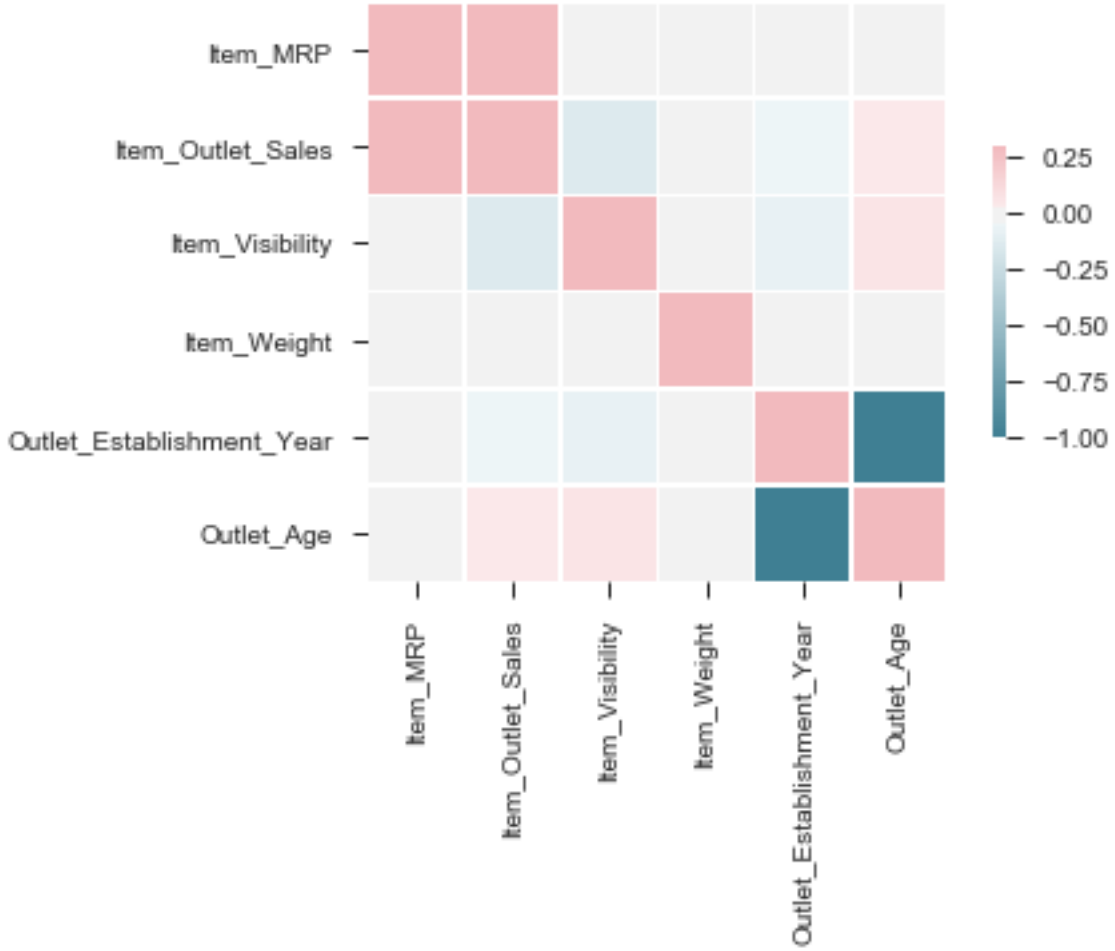


Outlet Type	Outlet Size
Grocery Store	Small
Supermarket Type 1	Small
Supermarket Type 2	Medium
Supermarket Type 3	Medium



- Grocery stores have a higher concentration of data points around the lower sales while the supermarkets have a wide range of item sales.
- Supermarket Type3 has the highest sales and Grocery Store has the lowest. Since Grocery Stores are small, they carry fewer types of items.
- Even though supermarkets Type1 are small, their sales are high because they are larger in number as compared to the other types of outlets.

Correlation between the variables



There is a strong correlation between the item price and sales, and a negative correlation between age and establishment year; age was derived from establishment year.

Machine Learning

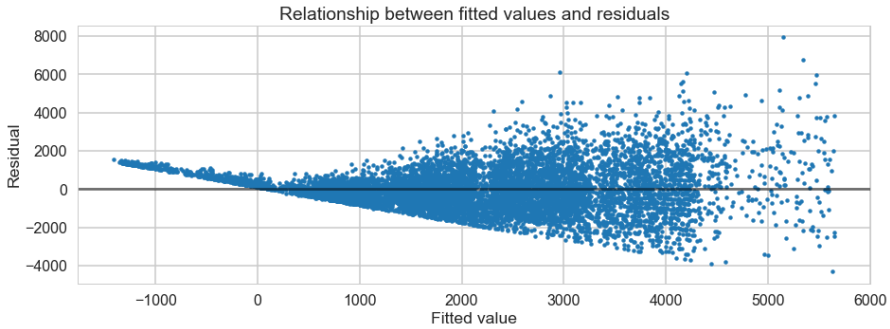
Models

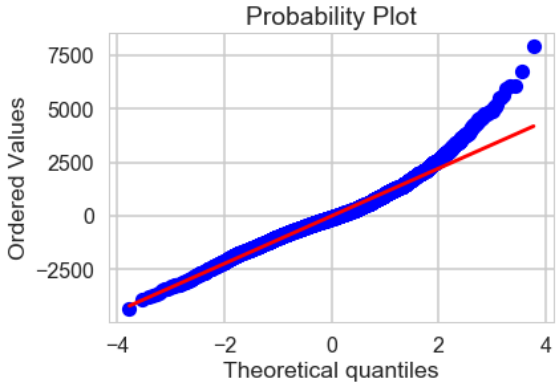
The data was fitted to three different models :

- Linear Regression
- Lasso Regression
- Random Forest Regression

A linear model was fitted and evaluated using the OLS (ordinary least squares) method from Python library's statsmodel. The statsmodel summary report gave an R2 score of 0.56. The p-value and the coefficients indicated that the Outlet Age, Outlet Size, Outlet Type, and Item MRP are statistically significant predictors of sales while the Item_Visibility and Item_Weight were not significant in predicting sales.

Fitted values plotted against residuals produced an outward funnel shaped pattern indicating an increasing variance of errors. The standard Linear Regression assumption is that the variance is constant across the entire range. The probability plot shows that standard errors were not distributed normally in the third quartile. Data was positively skewed.





After evaluating the model, the Item_Visibility and Item_Weight columns were dropped.

The test dataset downloaded from Kaggle, does not have the Outlet_Sales column. Since the actual Outlet_Sales data is needed to calculate accuracy scores against the predicted sales, this dataset was not used. Instead the training dataset was split into train and test after creating Features and Label arrays.

Model Building

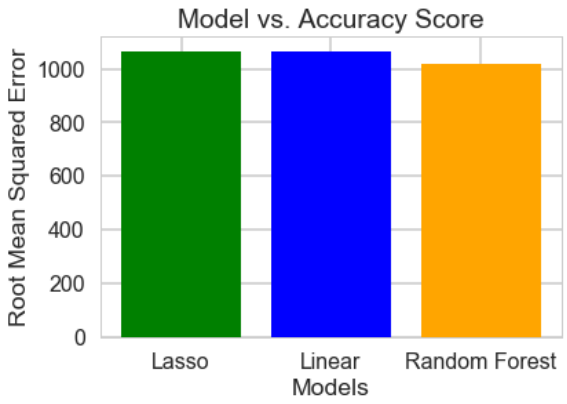
The models were imported from scikit-learn, instantiated, and fitted. The models were set to random state to make the results reproducible. After training the models to learn the relationships between the features and the labels, these models were used to make predictions on the test data, and the accuracy scores for each model were computed.

Hyperparameter Tuning

Since Random Forest Regressor gave the best accuracy score, Hyperparameter Tuning was performed on this model using RandomizedGridSearchCV. The model was built with 500 trees with a maximum depth of 5.

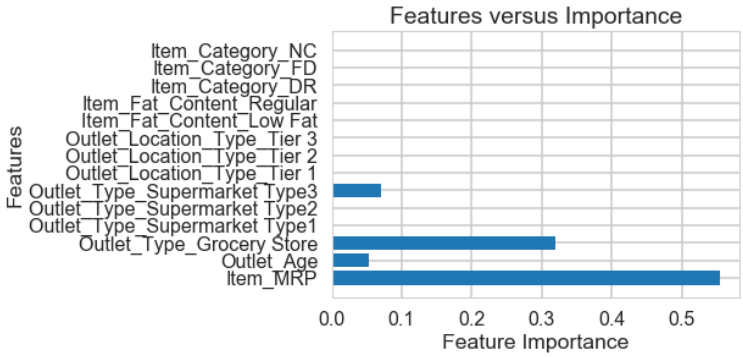
Model Evaluation

The R2 score and RMSE were almost the same for the Linear and Lasso Regression models. R2 score was higher for the Random Forest Regressor. The Random Forest Regressor gave the best Root Mean Squared Error of 1020.



	model	r2 score	root mean sqr err
0	Linear	0.580125	1068.273195
1	Lasso	0.580377	1067.952716
2	Random Forest	0.616902	1020.416567

Feature importance



	features	importance
0	Item_MRP	0.555105
1	Outlet_Age	0.053701
2	Outlet_Type_Grocery Store	0.318748
3	Outlet_Type_Supermarket Type1	0.000878
4	Outlet_Type_Supermarket Type2	0.000595
5	Outlet_Type_Supermarket Type3	0.070620
6	Outlet_Location_Type_Tier 1	0.000011
7	Outlet_Location_Type_Tier 2	0.000035
8	Outlet_Location_Type_Tier 3	0.000025
9	Item_Fat_Content_Low Fat	0.000115
10	Item_Fat_Content_Regular	0.000108
11	Item_Category_DR	0.000005
12	Item_Category_FD	0.000042
13	Item_Category_NC	0.000012

Results

- Random Forest Regressor predicted MRP as the most important predictor of Sales followed by Supermarket Type and Outlet Age. An increase in price per unit will result in an increase in sales.
- Outlet Age is an important feature proves our hypothesis that older stores have customer loyalty.
- Even though both Supermarket Type2 and Type3 are medium size supermarkets, Type 3 is a strong predictor as it carries a wider breadth of products as compared to Type2.
- Grocery stores are also a strong predictor
- The location of the outlet does not impact sales.
- Overall, adding Grocery stores and medium size Supermarkets with a larger variety of products will help increase future sales.

Future Work

- Other models such as XGBoost can be tested.
- More models can be tested with a different subset of features.
- Outliers can be removed to reduce skewness, and improve model performance.