



BigMart Sales Prediction

Rizwan Mohamed Kareem , Abhinaya Kannan, Kshitiz Goyal

Dr. Saleena B

School Of Computer Science and Engineering (SCOPE)

Motivation/Introduction

The main aim of this project is to find key insights from the sales data of the BigMart chain of stores. Analyzing the dataset enables us to provide conclusions on how various factors affect the sales in the supermarket. The conclusions from the data enables us to make improvements in the sales and also helps us to provide a satisfactory experience for the customer.

Sales prediction is the process of estimating future sales. Accurate sales forecasts enables the company to make informed business decisions and predict short-term and long-term performance.

SCOPE of the Project

- Predict achievable sales revenues
- Identify factors which affect the sales negatively
- Provide insights to efficiently allocate resources
- Help plan for future growth of the company
- Improve customer experience

Methodology

Data Pre-processing

Our data has some missing values in the form of NAN values, NULL values. We need to impute those missing values with suitable measurements like arithmetic mean, mode, etc. Calculating the measurements after grouping the dataset based on the item identifier or outlet identifier provides more accurate values for the missing ones.

Feature Engineering

Item Visibility doesn't have missing values but it has zero value which doesn't make sense, hence those zero values has to be treated as missing values. Years of operation of a store is more useful than the establishment year. Therefore the years of operation can be calculated from the establishment year values. The values representing the item fat content has typographical error which needs to be corrected and uniformized.

Feature Transformations

We can calculate the importance given to a product in each store from the item visibility mean ratio. Many built-in regression algorithms only recognize continuous values, therefore there is a need for us to convert the categorical values to continuous values through One-hot encoding.

Model Building

Different kinds of regression, classification and clustering models need to be built and compared to efficiently predict the sales.

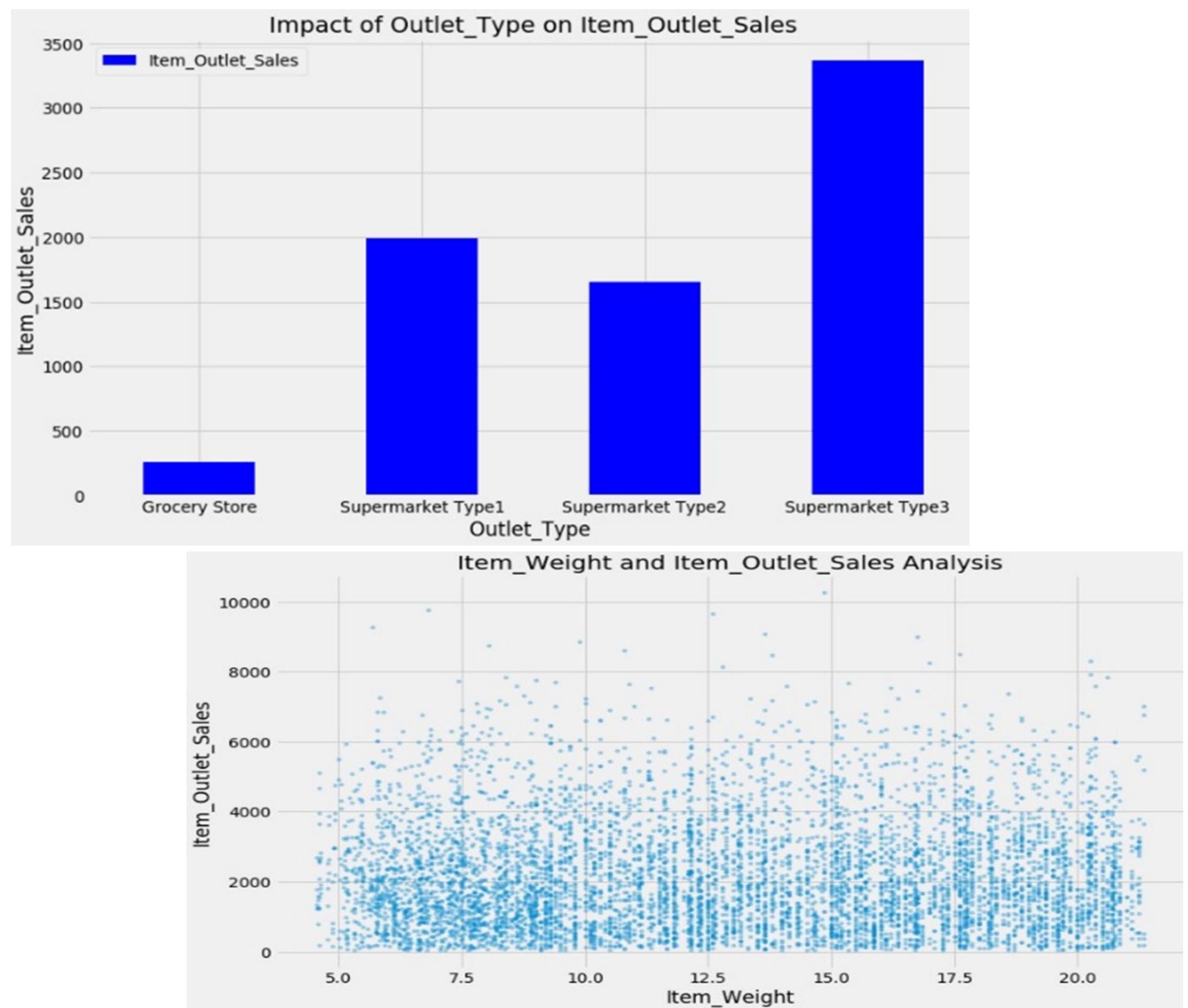
- 1) Decision Trees—They classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance.
- 2) Random Forest—A random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.
- 3) AdaBoost—It helps combine multiple weak learner into a single strong learner.

- 4) Gradient Boosting—It produces a prediction model in the form of an ensemble of weak prediction models.
- 5) Support Vector Machine— It constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression
- 6) K-Nearest Neighbors— It works by finding the distances between a query and all the examples in the data.
- 7) Logistic Regression— It is a statistical model that in its basic form uses a logistic function to model a binary dependent variable.
- 8) DB-Scan— When given a set of points in some space, it groups together points that are closely packed together , marking as outliers points that lie alone in low-density regions .

Results

Decision Tree, Random Forest, Gradient Boosting and KNN performed good prediction and ended up with better RMSE and R2 score. They approximately had an R2 score of around 0.59 and a RMSE value of around 1069. We also have the following inferences—

1. Item weight has low correlation on item outlet sales
2. Shops established on the year 1998 shows very less item outlet sales.
3. Medium size, Supermarket Type 3, Tier 2 outlets shows higher sales.



Conclusion/Summary

We discovered that people search and buy essential products even though they are not in the easily visible shelves. Medium size stores and stores in tier 2 cities show more sales compared to others. The data doesn't include time of purchases which may have a huge impact on our inferences. So, future work can be done with more data points collected over different years.

Contact Details- abhinaya.kannan2019@vitstudent.ac.in

rizwanmohamed.kareem2019@vitstudent.ac.in, kshitiz.goyal2019@vitstudent.ac.in

References - www.analyticsvidya.com, www.medium.com

Acknowledgments - Jupyter Project, VIT University