INFO 370 BigMart Sales

Yasmine Hejazi, Kari Nasu, Zico Deng,
Frederick Wijaya, Sarah Park
University of Washington, Informatics

Abstract

BigMart has amassed a large set of data from their existing outlet stores and has publicized them for the purpose of machine learning. Identifying key features of stores can enable for effective outlet management of items and assist in solidifying sales tactics. We first conducted exploratory data analysis to identify areas of interest within the data as well as linear regression with a selected set of features. Thereafter receiving our results of the linear regression, we turned to various machine learning algorithms before finalizing on with a simple linear regression model. Overall our methods and analysis have found two features that companies should look to prioritize when they want to sell an item effectively; we found that to have a successful item, in terms of sales, it needs to be within a medium-sized Supermarket Type 3 that lies in a Tier 2 / 3 location.

Keywords: BigMart, machine learning, linear regression

Introduction

Provided by Analytics Vidhya, an online data science competition host, we have been participating in the BigMart Sales Practice Problem which began on May 25th, 2016, and ends on December 31st, 2018. Data scientists at BigMart have "collected 2013 sales data for 1559 products across 10 stores in different cities", which participants use to build a model to predict product sales by store. With this, BigMart will try to gain understanding of product and store properties that lead to increased sales.

Using data to increase profitability is now an intuitive and common practice in business; the type, number, and even physical placement of products is no longer arbitrary, rather, it is determined by data. Therefore, this study echoes the common business approach to increasing profitability: data-driven decision making.

Problem Statement: "The data scientists at BigMart have collected 2013 sales data for 1559 products across 10 stores in different cities. Also, certain attributes of each product and store have been defined. The aim is to build a predictive model and find out the sales of each product at a particular store. Using this model, BigMart will try to understand the properties of products and stores which play a key role in increasing sales."

The problem space we are interested in is the relationship between the profitability of a store and the factors that influence it such as the items a store offers, their placement on shelves, the size of stores, where the stores are located, etc. Through this analysis of the Big Mart dataset, we aim to identify key factors that should be considered for when companies like Big Mart advances with opening new outlet stores or advancing their sales in existing ones.

Some of the questions we will try to answer specifically are:

1. What properties of a store are key in determining store profitability?
2. What properties of store items are key in determining product profitability?
3. What items are most profitable in each location?
4. What category of products are most profitable in each location?
5. What types of location are most profitable?

Specific hypotheses we will be testing are:

1. Item visibility affects the sale of the product
2. Outlet size, type and outlet location type affects the profitability of a store
3. Stores that are located in tier 1 cities or urban areas should have higher sales because people who live in these areas tend to have higher levels of income

Related Work

An attempt to the Big Mart challenge is described in the paper titled *An Ensemble Based Predictive Modeling in Forecasting Sales of Big Mart*, written by T. Leo Alexander and D. Delwin Christopher, both of whom are in the Department of Statistics at Loyola College Chennai. These analysts participated in this challenge with the premise that it "can be beneficially adopted for the wholesale and retail vendor joints in India," since it will help with "understanding the factors that influence the sales of similar products in a better manner" (T. Leo Alexander, D. Delwin Christopher, *AN ENSEMBLE BASED PREDICTIVE MODELING IN FORECASTING SALES OF BIG MART*).

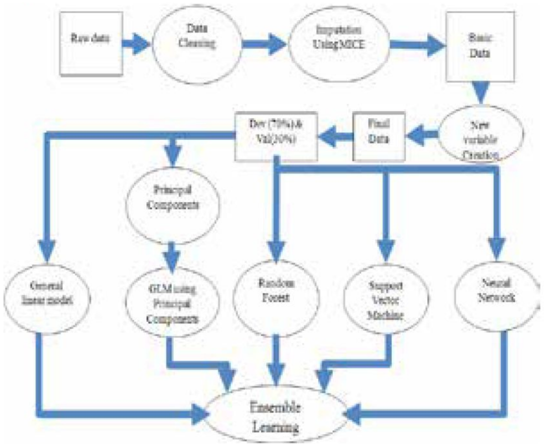


Figure 1.
Diagram depicts the procedural breakdown, which ends with ensemble learning.

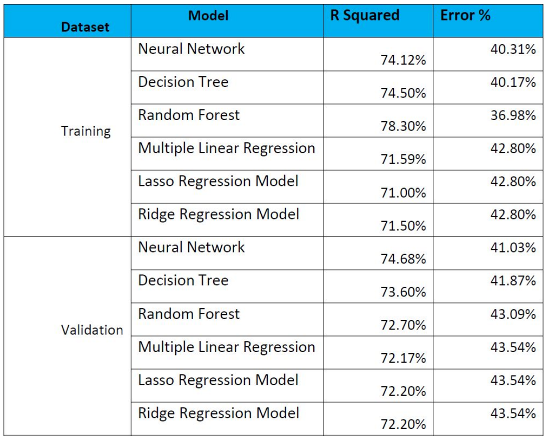
In their analysis, they used "basic statistical predictive models like general linear model, principal component analysis based model and other machine learning techniques like random forest, support vector machine and neural network" (T. Leo Alexander, D. Delwin Christopher). They utilized the "linear model, random forest, support vector machine and neural network [as] ensemble methods for regression, where a *committee* of trees each cast a vote for the predicted values" (T. Leo Alexander, D. Delwin Christopher). This allowed the strengths of each model to be used. Figure 1 depicts the procedural breakdown of this teams' analysis.

Multivariate imputation by chained equations (MICE), also known as joint modeling (JM) and fully conditional specification (FCS), was utilized to handle missing data (this process occurred in the *Imputation Using MICE* phase as seen in Figure 1). This method creates "multiple imputations" to "[fill] in the missing values multiple times, creating multiple 'complete' datasets" (Azur, Melissa J., et al., *Multiple Imputation by Chained Equations: What Is It and How Does It Work?*). This accounts for statistical uncertainty in each of the imputations. In addition to this, "the chained equations approach is very flexible and can handle variables of varying types (e.g., continuous or binary)" (Azur, Melissa J., et al., *Multiple Imputation by Chained Equations: What Is It and How Does It Work?*). Utilizing MICE to handle missing data sounds like a reasonable approach, as the data is continuous and statistical uncertainty is dealt with from the multiple imputations.

The data was also segmented, which is necessary to build the regression model. There were categorical variables with many categories within, which made this process complex. The analysts did not detail on how segmentation was specifically done, but declared that it was performed.

The team found that the general linear model that used the principal component analysis and the random forest produced the best score of 1171.800. It is not surprising that the general linear model based on a tree-like model performed well, given that we found in our analysis that regression performed the best, followed by the decision tree.

Another approach to the Big Mart challenge is described in the University of Connecticut Masters in Business Administration and Project Management (MSBAPM) May 2016 Newsletter. The team determined that the top influencing independent variables were: Item MRP (Maximum Retail Price), Outlet Age, Outlet Size, Outlet Location Type, and Outlet Type. Explanation for the selection was not provided.



Dataset	Model	R Squared	Error %
Training	Neural Network	74.12%	40.31%
	Decision Tree	74.50%	40.17%
	Random Forest	78.30%	36.98%
	Multiple Linear Regression	71.59%	42.80%
	Lasso Regression Model	71.00%	42.80%
	Ridge Regression Model	71.50%	42.80%
Validation	Neural Network	74.68%	41.03%
	Decision Tree	73.60%	41.87%
	Random Forest	72.70%	43.09%
	Multiple Linear Regression	72.17%	43.54%
	Lasso Regression Model	72.20%	43.54%
	Ridge Regression Model	72.20%	43.54%

Figure 2.
Table shows summary for each model in training and validation datasets.

Figure 2 depicts the behavior of each model for both training and validation data. It's seen that the highest performing models are the Neural Network, Decision Tree, and Random Forest.

The team came up with five conclusions. Three insights overlapped with our analysis, being:

1. Medium-sized stores Tier-2 stores perform the best. Therefore, Tier-2 cities should be first choice when opening new stores.
2. Supermarkets always perform better than grocery stores. Therefore, when opening new stores, supermarkets should take priority.
3. Medium-sized stores are the highest performing.

Two findings that this team discovered that our team did not account for are as follows:

1. Increasing the MRP of a product does not translate to increase in sales.
2. As a store ages, the sales increase. This could put priority on renovating older stores compared to opening new ones, since customers loyalty develops after years of purchasing from a particular store.

Methods

Exploratory Data Analysis

We first began our analysis with a preliminary exploratory data analysis, looking into the data to identify areas of interest for our further analysis. This portion had no predictions and we simply calculated the mean values for a given category and compared the results with each other. Looking at the first couple rows of the data and doing a few scatter plots, we decided to look more closely into Item_Visibility, Outlet_Size, Outlet_Location_Type, and Outlet_Type to investigate a further relationship between these features and sales.

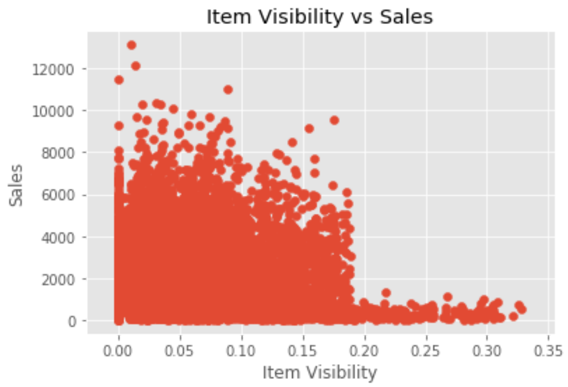


Figure 3.
Diagram of all item sales by item visibility.

Linear Regression

After we selected four features, we utilized a linear regression model on the dependent variables (Item_Visibility, Outlet_Size, Outlet_Location_Type, and Outlet_Type) and the independent variable (Item_Outlet_Sales).

Hypothesis 1: Item visibility affects the sale of the product

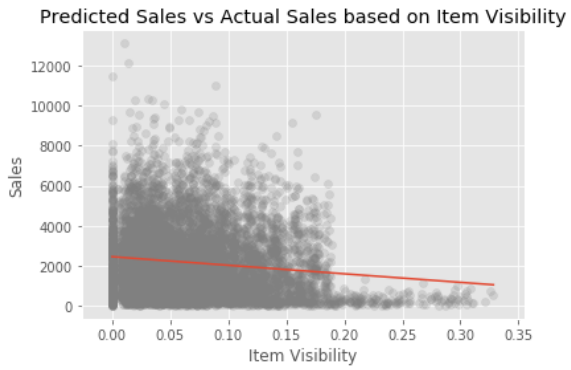


Figure 4.
A predicted negative relationship is shown between sales and item visibility.

For Item_Visibility, we observed that each additional unit of visibility is associated with an -$355.310 change in outlet sales in general. As a result, we have observe a negative relationship between Sales and Item Visibility. Item Visibility was shown to be statistically significant.

Hypothesis 2: Outlet size and outlet location type affects the profitability of a store

For Outlet_Size, we observed that each additional unit of visibility is associated with an +$916.8250 change in outlet sales for medium-sized outlets. The standard deviation of the sampling distribution is (+/-)$39.403. There is a 95% chance the true value lies between $839.582 and $994.068

For Outlet_Location_Type, we observed each additional unit is associated with an +$867.2392 change with similar standard error in outlet sales for outlets located in Tier 3 Locations. Confidence Intervals are between $790.009 and $944.469.

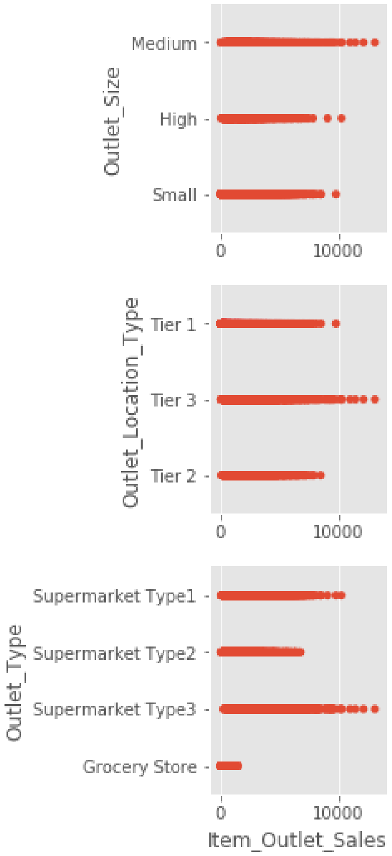


Figure 5.
The relationship of sales to outlet type, location, and size

For Outlet Type, we observed that each additional unit of visibility is associated with a +$2402.5746 change in outlet sales for Type 3 Supermarkets. The standard deviation of the sampling distribution is (+/-)$96.386 and the confidence intervals are between $2213.624 and $542.587

For Item Visibility, each additional unit of visibility is associated with an -$266.7973 change in outlet sales in general. Standard Error: The standard deviation of the sampling distribution is (+/-)$409.997. And the confidence intervals are between -$1070.536 and $591.525

Overall, we can observe a negative relationship between Sales and Item Visibility. However, the inclusion of the covariates Outlet Type, Outlet Location Type and Outlet Size make it so that the relationship is not statistically significant as observed by the p-value of 0.515. We instead observe statistically significant relationships between Outlet Sales and these covariates. We also observe Outlet_Type to have a larger influence on Item Outlet Sales compared to other covariates based on the coefficient value.

Hypothesis 3: Stores that are located in tier 1 cities or urban areas should have higher sales because people who live in these areas tend to have higher levels of income.

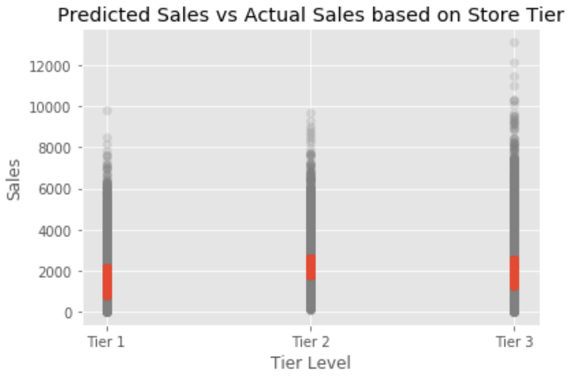


Figure 6.
Diagram depicts the relationship of sales with location by store tier.

The multilinear regression summary shows that each additional unit of visibility is associated with an +$406.0533 change in outlet sales for outlets located in Tier 2 Locations. The standard deviation of the sampling distribution is (+/-)$47.086 with a a confidence interval between $313.754 and $473.17.

In general, each additional unit of visibility is associated with an -$4035.2495 change in outlet sales. The standard deviation of the sampling distribution is (+/-)$354.458 and there is a 95% chance the true value lies between -$1070.536 and $473.171

We can observe a negative relationship between Sales and Item Visibility. The relationship is statistically significant. We also observe statistically significant, positive relationship between Outlet Sales and Outlet Location Type per unit of item visibility. The trend is higher for outlets in Tier 2 Locations compared to Tier 3 or Tier 1 Locations.

Machine Learning Models

Finally, we used a linear regression model to predict sales. The data is split into testing and training. This allows us to train our program with training data and then test our program with testing data. We will keep adjusting predictors to find the point when the program has the highest accuracy predicting sales.

- Scikit-learn (SKLearn) - machine learning library for Python
- NumPy - package for scientific computing with Python
- Machine learning algorithms

Results

Multilinear Regression

Based on the nature of our preliminary regressions, we have concluded that the first hypothesis (testing the relationship between item visibility and product sale) returns somewhat meaningful results whereas when other covariates are introduced the findings become negligible. Based on the results of this study, we conclude that an ideally successful item in terms of sales would be placed in a lower visibility setting within a medium-sized Supermarket Type 3 that lies in a Tier 2 or 3 location.

Machine Learning

After trying a few different algorithms, we ended up with choosing linear regression as our main model. Not surprisingly, regression model gave us the highest score compared to all other classification algorithms we have tried because this machine learning problem is trying to predict a continuous quantity rather than classification.

Discussion

Multilinear Regression

The findings of the regression were as we predicted as there are many other factors that we were not able to assess due to the lack of time, resources, and skills within the team for this analysis. The results of the data was lacking due to the type of data, where we may have desired continuous variables but we had left them as categorical. Our regression analysis could have been better in terms of accuracy had the columns been transformed into data types that matched what our regression analysis was looking for. Furthermore the relationships with Item_Visibility could have been better explained given better context of the variable.

Machine Learning

Interestingly, RandomForest, although similar to DecisionTree, performed significantly worse. Considering that RandomForest utilizes the DecisionTree process (by creating a number of decision trees to build the random forest), overfitting may be one of the reasons for the poor performance.

Also, typically well-performing classifiers like GradientBoost and XGBoost did not perform well with this smaller dataset. The biggest reason is that gradient boosting algorithms, while often being very efficient and accurate, are usually not for predicting quantities or continuous values. They are usually for classifying observations. We failed to identify what exactly we were trying to predict at early stages. Therefore, we wasted a lot of time trying algorithms that did not fit this problem space. Recently, the trial and errors showed us that we pursued a wrong direction by picking wrong algorithms. The hardest part of solving a machine learning problem can be picking the right algorithm that does right job. Not all algorithms are applicable for a wide range of contexts, which is a main takeaway from this project. Each of them have distinct pros and cons depending on different situations. It is important for researchers to deeply understand the machine learning problem and identify the exact predictions needed for solving the problem before training a model.

Future Work

One of our next steps for our multilinear regression analysis would be to refine our hypotheses as well as the features we inspect for our regression analysis; we picked those areas of interest prior to assessing the dataset fully and we should have opted to review the dataset and identify key areas of interest. Also we may want to investigate other distributions and transform our existing data to match those distributions.

When tuning parameters for different machine learning models, we could spend more time understanding parameters and their potential impact on the model before tuning them. This will actually save us time because tuning parameters that are not particularly useful for improving model accuracy are time consuming, considering we have such a large dataset.

Strategies for evaluating model accuracy can be also improved in the future. We simply used log loss for evaluating trained model. However, this is not the best strategy for model evaluation in this competition because it does not accurately reflect our submitted scores. We could have totally simulated how the competition evaluated our models because the competition has shared their model evaluation equation to the public.

Reviewing other reports on this same data science challenge brought other approaches to light.

For example, if we were to re-prep the data, utilizing MICE could be a plausible option. Imputation of data, if done wisely, is a productive and meaningful approach to handling missing data. This is opposed to setting the value to mean without account for the circumstances.

Utilizing MICE could set up the next new approach we could try in the future: splitting up the training data into training and validation. This would allow us to early on gain understanding of the model’s behavior.

Lastly, both papers discussed in the *Related Work* section of this report include the neural network model. In the project completed by the University of Connecticut Masters student team, the neural network was the highest performing model in both the training and validation datasets. This could be, that training encompasses running individual cases through one at a time, while updating the weights based on error. Over time, this typically causes the networks to be attuned to the data. This minimizes the ultimate error in predictive model (Kellett, Dan, *Making Data Science Accessible - Neural Networks*).

References

Azur, Melissa J., et al. “Multiple Imputation by Chained Equations: What Is It and How Does It Work?” International Journal of Methods in Psychiatric Research, U.S. National Library of Medicine, 1 Mar. 2011.

Kellett, Dan. “Making Data Science Accessible – Neural Networks.” KDnuggets Analytics Big Data Data Mining and Data Science, KDnuggets,

T. Leo Alexander, D. Delwin Christopher AN ENSEMBLE BASED PREDICTIVE MODELING IN FORECASTING SALES OF BIG MART. International Journal of Scientific Research, Vol : 5, Issue : 5 MAY 2016

UConn MSBAPM Newsletter. University of Connecticut. May 2016.