Big Mart Sales Prediction

Anirban Pal
DSC680
Fall 2019
https://anirbanpaldsc.github.io/

Contents

Business Question

The Big Mart is an international brand established in 2007 with free home delivery services of food and grocery. The purpose of the project is to analyze the store and product-based sales data collected from the Big Mart stores around 2013.

Through the analysis, I am planning to predict the impact of other factors on sales of a product in a store. The train file will be used to explore the data and train the model. The model will be applied on test data in the test file and model performance will be evaluated.

Data Source

The dataset contains the sales data of the year 2013 for 1559 products across 10 stores in different cities. The data has been collected from:

https://code.datasciencedojo.com/tshrivas/dojoHub/tree/a152a17dee24dcfcc10bb75c77c2e88cdcf90212/Big%20Mart%20Sales%20DataSet

Codebook:

Column Position	Atrribute Name	Definition	Data Type	Example	% Null Ratios
1	Item_Identifier	It is a unique product ID assigned to every distinct item. It consists of an alphanumeric string of length 5	Alphanumeric	FDN15	0
2	Item_Weight	This field includes the wieght of the product	Numeric (float)	17.5	17.16531738
3	Item_Fat_Content	This attribute is categorical and describes whether the product is low fat or not. There are 2 categories of this attribute: ['Low Fat', 'Regular']. However, it is important to note that 'Low Fat' has also been written as 'low fat' and 'LF' in dataset, whereas, 'Regular' has been referred as 'reg' as well	Alpha	Low Fat	0
4	Item_Visibility	This field mentions the percentage of total display area of all products in a store allocated to the particular product	Numeric (float)	0.01676	0

Column Position	Atrribute Name	Definition	Data Type	Example	% Null Ratios
5	Item_Type	This is a categorical attribute and describes the food category to which the item belongs. There are 16 different categories listed as follows: ['Dairy', 'Soft Drinks', 'Meat', 'Fruits and Vegetables', 'Household', 'Baking Goods', 'Snack Foods', 'Frozen Foods', 'Breakfast', 'Health and Hygiene', 'Hard Drinks', 'Canned', 'Breads', 'Starchy Foods', 'Others', 'Seafood']	Alpha	Meat	0
6	Item_MRP	This is the Maximum Retail Price (list price) of the product	Numeric (float)	141.618	0
7	Outlet_Identifier	It is a unique store ID assigned. It consists of an alphanumeric string of length 6	Alphanumeric	OUT049	0
8	Outlet_Establishment_Year	This attribute mentions the year in which store was established	Numeric (Integer)	1998	0
9	Outlet_Size	The attribute tells the size of the store in terms of ground area covered. It is a categorical value and described in 3 categories: ['High', 'Medium', 'Small']	Alpha	Medium	28.27642849
10	Outlet_Location_Type	This field has categorical data and tells about the size of the city in which the store is located through 3 categories: ['Tier 1', 'Tier 2', 'Tier 3']	Alpha	Tier 3	0
11	Outlet_Type	This field contains categorical value	Alpha	Supermarket Type2	0

Column Position	Atrribute Name	Definition	Data Type	Example	% Null Ratios
		and tells whether the outlet is just a grocery store or some sort of supermarket. Following are the 4 categories in which the data is divided: ['Supermarket Type1', 'Supermarket Type2', 'Grocery Store','Supermarket Type3']			
12	Item_Outlet_Sales	This is the outcome variable to be predicted. It contains the sales of the product in the particulat store	Numeric (float)	2097.27	0

Store Level Hypotheses:

1. City Type: Here, location mainly refers to cities. Stores located in bigger cities should have higher sales because of the higher income levels.
2. Population Density: Stores located in densely populated areas should have higher sales because of more demand.
3. Store Capacity: Stores which bigger floor area should have higher sales as they act like supermarkets where people would prefer getting everything from one place.
4. Competitors: Outlets with similar business in the vicinity will face competition and potential impact on sales.
5. Effect of promotion: Stores with better strategic promotion and product marketing should have higher sales by generating higher footprints and appealing to the correct target customer segment.
6. Accessibility and location: Location of a store is a key factor since accessibility and visibility will drive higher sales.
7. Customer Behavior: Stores keeping the right set of products to meet the local needs of customers will have higher sales.
8. Store ambience: Stores which are well-maintained, clean and sufficient walking space has aesthetic appeal to people, generating higher traffic.
9. Customer Service: Friendly and polite customer service is very important for generating returning customer and positive feedback.
10. Store Rating: Online rating of stores in company website, Google or related app is an important feedback that customers check before visiting any store.

Product Level Hypotheses:

1. Branded Product: Branded products of each product category usually have higher sales due to better recognition, brand value and advertisement.

2. Product placement: Analyzing co-occurrence of sales for specific products, in short, effective Affinity Analysis or Market Basket Analysis can boost sales performance.
3. Packaging: Attractive packaging can boost product sales.
4. Product Type: Products that are used more frequently, such as toothpaste, milk, etc. are sold more than occasional products like raincoat.
5. Seasonality: Based on the weather, stocking seasonal product is an effective way to promote sales, e.g. storing more beer in summer, school supplies before the start of school season etc.
6. Stocking and Shelf life: While stocking, calculating shelf life is important (as products could be perishable or take revenue generating space for longer period of time). Effective stocking strategy can increase sales.
7. Product Visibility: Products that require more promotion or new products could be given better shelf area so that customers can see them easily and buy them.
8. Advertising and Promotional Offers: Products that require more sales are usually listed with discounts and promotional offers. Advertising boosts sales in general.

Methods used in the project

1. Exploratory Data Analysis
   The first method applied to the data after it is loaded, is Exploratory Data Analysis (EDA). This includes data profiling and generating descriptive statistics at attribute level. This will give a fair idea of how the data is distributed and what type of transformation is required.
   For this, I used `pandas_profiling`. Instead of just getting a single output, pandas-profiling quickly generated a very broadly structured HTML file containing most of what I might need to know before diving into a more specific and individual data exploration. This enabled me in faster EDA, variable specific EDA, analyzing histogram, correlation plot and descriptive statistics.
2. Data Cleansing
   The second method applied to the data is data cleansing (as well as data preparation). This includes tasks like handing blank or null value, eliminating duplicates, shaping and joining data to easily use in the model.
   I made the following assumptions:
   - `Item_Weight` had 17.2% missing values. I decided to populate those values with average item weight to have a cleaner dataset without losing data.
   - `Outlet_Size` has *28.3%* missing values. But unlike `Item_Weight`, `Outlet_Size` is a categorical variable, hence, I used the most frequent value, i.e. mode to fill up the blanks.
   - `Item_Visibility` has 6.2% zero values. I treated this same as `Item_Weight` and used mean weight to fill up.
3. Feature Engineering
   I used feature engineering to create proper input dataset for the machine learning model. I took the following data decisions while performing feature engineering.
   - I created a broad category `Item_Category` based on `Item_Identifier` (Food, Drink, Non-Food).
   - I standardized the value of the variable `Item_Fat_Content` to Low Fat, Regular and Non-Food.
   - I calculated age of an outlet from year established, and converted year to numeric variable, age.

- I performed label encoding, followed by one hot encoding to convert the categorical variables to numeric variable.
- I divide the data into test and training data set.

4. Visualization
   The third method applied to the data is visualization. Data visualization is an effective way of finding relationship between variables. This further aids in model building.
   I did not plot any specific variable distribution since `pandas_profling` did most of that. The final result of the model though is plotted using matplotlib.
5. Model Building
   Model building includes deciding what algorithm works the best for the data as well as for the question we are trying to answer. It also includes one hot encoding, feature engineering, applying machine learning algorithm and storing the model in some fashion. The entire process deals with the training dataset.
6. Model Evaluation and Application
   The last step applied to the data is the application of the trained model and evaluating its performance with measures like RMSE, feature importance etc. The test data is something the model has not seen so far. That is why the model is applied on test data for fair evaluation.
   Multiple models are evaluated in the process. Initially I wanted to start with a regression model to see how the model is performing. While deciding on which regression model to apply, I considered that the data has enough features to cause computational challenge and somewhat mitigate the problem of overfitting. I picked **Ridge Regression** over **Lasso Regression** because Ridge follows **L2** regularization (sum of the square of the weights) whereas Lasso follows **L1** regularization (sum of weights). The intention is to get a better result. I was able to improve the RMSE slightly by reducing the alpha (Y intercept of the regression line) in the model. But the improvement was nominal.
   Finally, I applied Random Forest classifier. I wanted to generalize the model to treat the test file better and eliminate overfitting, if any. Considering there is no situation where Random Forest is not at least somewhat useful, I decided to evaluate Random Forest. The RMSE improved quite a lot. By tuning the estimator, maximum tree depth, leaf nodes etc. parameter, this score could be slightly improved. But for the sale of concluding remark, I take this as the final result.

Potential Issues?

At the beginning of the project, following were identified as potential issues:

- The data might not be sufficient to establish all of the hypothesis.
- The % of missing values in outlet size and item weight will cause some data loss.
- Hypothesis that might not be evaluated due to unavailability of data:
  - Population will require additional data element since the core data does not have any census or population information.
  - Finding competitors will require additional data element since the core data does not have any competitor information.
  - Affinity Analysis is part of a much larger research with more granular data.
  - There is no way to measure effect of promotion, effect of product placement, effect of product packaging, seasonality, stocking, visibility, advertising etc.

Some of the hypothesis evaluations were restricted due to unavailability of data. But all of the partially missing data has been filled up during the data cleansing phase.

Concluding Remarks

From the final model, the following are the most dominant features when it comes to item sales price: `Item_MRP`, `Outlet_Type_0`, `Outlet_Years`, `Outlet_5` and `Outlet_Type_3`. However, apart from `Item_MRP` and `Outlet_Years`, the other features are resulted from encoding during feature engineering.

There were 10 different `Outlet_Identifier` in the data and they were encoded as `Outlet_0` through `Outlet_9`. Among which, `Outlet_5`, i.e. `Outlet_Identifier` = *OUT027* is the most significant one.

The other encoded variables, `Outlet_Type_0` and `Outlet_Type_3`, were derived from the categorical variable `Outlet_Type`. This variable has three distinct values, *Grocery Store, Supermarket Type1, Supermarket Type2, Supermarket Type3,* which were encoded into `Outlet_Type_0` through `Outlet_Type_3`. The model predicted that `Outlet_Type_0` = *Grocery Store* and `Outlet_Type_3` = *Supermarket Type3* has the most impact on sales price among other outlet types.

GitHub Repository

https://github.com/anirbanpalDSC/Bigmart-Product-Sales-Prediction

Appendix

RMSE

Root Mean Square Error – it is the standard deviation of the residuals (prediction errors).

Scikit Learn

Scikit Learn is a Python machine learning library.

EDA

Exploratory Data Analysis – it is an approach to analyzing data sets to summarize their major characteristics, usually with visual methods.

HTML

Hypertext Markup language – it is the standard markup language for documents designed to be displayed in a web browser

Reference

1. Gaggin, Alex, *Applying machine learning to sales prediction*
   The article, published in R Studio, gives a detailed idea of how previous sales data can be used to predict future sales. It uses R programing language and gives an overview of the exploration and model building process.
2. Dancho, Matt (2017), *Predictive Sales Analytics: Use Machine Learning to Predict and Optimize Product Backorders*
   In this article, the author covers the challenge of dataset imbalance, i.e. when the majority class significantly outweighs the minority class. He created backorder prediction model through example.
3. Mitra, Rudradeb (2019), *How-to-Use Machine Learning for Buying Behavior Prediction: A Case Study on Sales Prospecting*
   In this article from Medium, the author used machine learning algorithms (Neural Networks) to identify sales prospects. He described an end to end sales process and pipeline.
4. Jain, Asrshay (2016), *Approach and Solution to break in Top 20 of Big Mart Sales prediction*
   In this article, the author gave an extensive overview of the top 20 Big Mart sales prediction models.
5. Columbus, L (2018), *10 Ways Machine Learning Is Revolutionizing Sales*
   In this article, author gave statistics of to what degree companies are adopting AI to predict sales. He also cited some of the use cases of AI and machine learning in sales and marketing.
6. *A Machine Learning Approach to Inventory Demand Forecasting*
   This article from gormanalysis.com deals with Inventory Demand Forecasting, - effect of overstocking and understocking.
7. Javier (2018), *How Machine Learning is reshaping Price Optimization*
   In this article, the author describes what price optimization is and how machine learning could be used in price optimization. For explanation, he used a typical scenario of a brick-and-mortar retailer. He also described the advantages of price optimization using machine learning.
8. Book, Adrien (2018), *Machine Learning & Physical Retail: A Love Story Waiting to Happen?*
   In this article author described the concept of forecasting, personalization, stock visibility in retail. He also described cases natural language processing can help in sales. Overall, he has a skeptical tone on the practical application of all these technologies.
9. Bobriakov, Igor (2018), *Top 10 Data Science Use Cases in Retail*
   In this article author described machine learning use cases like Recommendation Engine, Market Basket Analysis, Price Optimization, Store location, Customer Sentiment Analysis etc. and how they can be applied to the retail industry.
10. Guinn, Justin (2018), *FutureProof Your Small Business: Machine Learning In Retail*
    In this article the author presented a nice visual about machine learning. He went on to further explain concepts like Understanding Customer Behavior and Operational Efficiency. He also explained data requirement for machine learning.