Big Mart Sales Prediction using Machine Learning

VIVEK KUMAR GUPTA[1], SHREYA TIWARI[2], PIYUSH ANAND[3], VISHESH RAJ SRIVASTAVA[4]

[1] [2] [3] [4] (Computer Science & Engineering, Raj Kumar Goel Institute of Technology, Ghaziabad)

---***---

Abstract - Supply and demand are two fundamental concepts of sellers and customers. Predicting demand accurately is so difficult. The background behind the Big-mart Data analysis is the purpose of being the determination of the properties of products and stores that help increase the sales. In this paper we have designed a novel predictive model that determine the sales of every product at a specific store. It can analyze and predict outlet production sales for Big Mart. We used the concept of Machine Learning through Python to create predictive models. It will help us to understand the customer needs and provides better estimation of future sales.

Key Words: Machine learning, Sales forecasting, Sales prediction, Python.

1.INTRODUCTION

Day by day competition among different shopping malls as well as big marts is getting more serious and aggressive only due to the rapid growth of the global malls and on-line shopping. Every mall or mart is trying to provide a personalized and short-time offer for attracting more customers depending upon the day, such that the volume of sales for each item can be predicted for inventory management of the organization, logistics, and transport service, etc. With the help of machine learning, we can predict the future sale of a particular product. We are providing a forecast for the sales data of big mart in a number of big mart stores across various location types which are based on the historical data of sales volume. Depth knowledge of the past is required for enhancing and improving the likelihood of marketplace irrespective of any circumstances especially the external circumstance, which allows to prepare the upcoming needs for the business. According to the characteristics of the data, we can use the method of multiple linear regression analysis and random forest to forecast the sales volume.
Machine learning techniques can not only handle non-linear data but also huge data-set efficiently.

2. PROBLEM STATEMENT AND SOLUTION APPROACH

The data scientists at Big Mart have collected 2013 sales data for 1559 products across 10 stores in several cities. Also, certain attributes of every product and store are defined. The aim is to create a predictive model and determine the sales of every product at a specific store.

3. PROPOSED WORK

We will explore the problem in following stages-

1.Hypothesis Generation – understanding the matter better by brainstorming possible factors which will impact the result.

2.Data Exploration – watching categorical and continuous feature summaries and making inferences about the info .

3.Data Cleaning – imputing missing values in the data and checking for outliers.

4.Feature Engineering – modifying existing variables and creating new ones for analysis.

5.Model Building – making predictive models on the data.

3.1. Hypothesis Generation

This is a very pivotal step in the process of analyzing data. This involves understanding the problem and making some hypothesis about what could potentially have a good impact on the outcome. This is done before looking at the data, and we end up creating a laundry list of the different analysis which we can potentially perform if data is available.

THE HYPOTHESIS

We came up with the following hypothesis while thinking about the problem. Since we're talking about stores and products, let's make different sets for each.

❖ Store Level Hypothesis-

1.City type- Stores located in urban or Tier 1 cities should have higher sales because of the higher income levels of people there.

2.Population Density- Stores located in densely populated areas should have higher sales because of more demand.

3.Store Capacity- Stores which are very big in size should have higher sales as they act like one-stop-shops and people would prefer getting everything from one place

4.Competitors- Stores having similar establishments nearby should have less sales because of more competition.

5.Marketing- Stores which have an honest marketing division should have higher sales because it are going to be ready to attract customers through the proper offers and advertising.

6.Location- Stores located within popular marketplaces should have higher sales because of better access to customers.

7.Customer Behaviour- Stores keeping the right set of products to meet the local needs of customers will have higher sales.

8.Ambiance- Stores which are well-maintained and managed by polite and humble people are expected to have higher footfall and thus higher sales.

❖ Product Level Hypothesis-

1.Brand- Branded products should have higher sales because of higher trust in the customer.

2.Packaging- Products with good packaging can attract customers and sell more.

3.Utility- Daily use products should have a higher tendency to sell as compared to the specific use products.

4.Display Area- Products which are given bigger shelves in the store are likely to catch attention first and sell more.

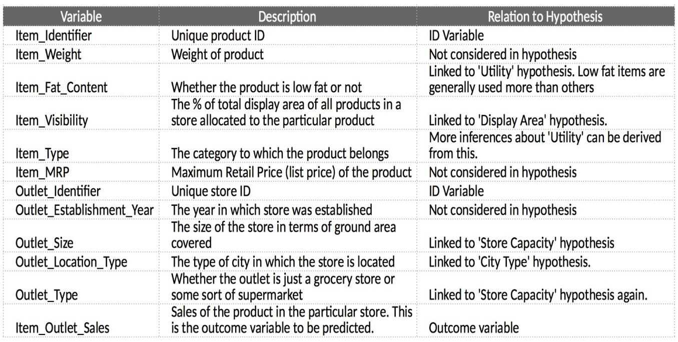
5.Visibility in Store- The location of product in a store will impact sales. Ones which are right at entrance will catch the attention of customer first instead of those in back.

6.Advertising- Better advertising of products in the store will should higher sales in most cases.

7.Promotional Offers- Products accompanied with attractive offers and discounts will sell more.

3.2 Data Exploration

We'll be performing some basic data exploration here and are available up with some inferences about the info . We'll try to figure out some irregularities and address them in the next section. The first step is to look at the data and try to identify the information which we hypothesized vs the available data. A comparison between the data dictionary on the competition page and out hypotheses is shown below-



Variable	Description	Relation to Hypothesis
Item_Identifier	Unique product ID	ID Variable
Item_Weight	Weight of product	Not considered in hypothesis
Item_Fat_Content	Whether the product is low fat or not	Linked to 'Utility' hypothesis. Low fat items are generally used more than others
Item_Visibility	The % of total display area of all products in a store allocated to the particular product	Linked to 'Display Area' hypothesis.
Item_Type	The category to which the product belongs	More inferences about 'Utility' can be derived from this.
Item_MRP	Maximum Retail Price (list price) of the product	Not considered in hypothesis
Outlet_Identifier	Unique store ID	ID Variable
Outlet_Establishment_Year	The year in which store was established	Not considered in hypothesis
Outlet_Size	The size of the store in terms of ground area covered	Linked to 'Store Capacity' hypothesis
Outlet_Location_Type	The type of city in which the store is located	Linked to 'City Type' hypothesis.
Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket	Linked to 'Store Capacity' hypothesis again.
Item_Outlet_Sales	Sales of the product in the particular store. This is the outcome variable to be predicted.	Outcome variable

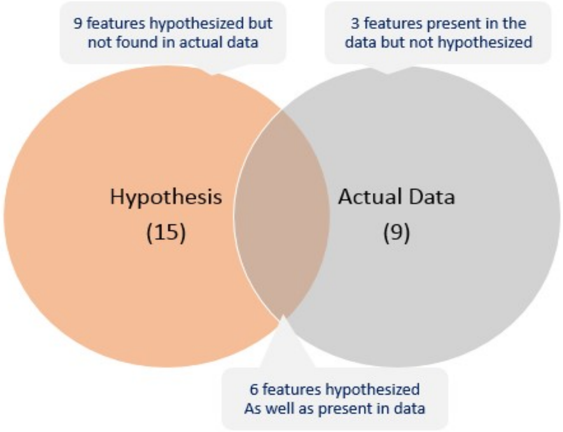

9 features hypothesized but not found in actual data
3 features present in the data but not hypothesized
Hypothesis (15)
Actual Data (9)
6 features hypothesized As well as present in data

Fig.4.2.1 Venn diagram of the comparison

3.3 Data Cleaning

After understanding the character of the data and finding a correlation between different features and target variable i.e. sales. The erroneous values within the info set need to get replaced with values that add up , the missing values need to get replaced with appropriate numerical or categorical value relying on the type of feature . The redundant information within the info set is to be removed. This fills the gaps within the data set and makes it wholesome, which enables better results.

3.4 Feature Engineering

Data cleaning gives us a wholesome error-free data set to figure with, Feature Transformation is that the family of algorithms wont to create new features from existing features, during this we use a linear combination of two or more features to form a replacement feature, this new feature gives better results with reference to target variable i.e. sales. this technique also uses categorical feature transformation to numerical feature transformation. The redundant features are dropped from the data-set for the new ones.

3.5 Model Building

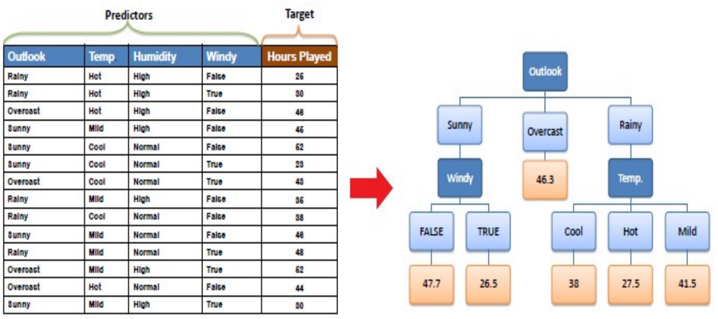
Now that we have the info ready, it's time to start out making predictive models. There are many models but here, we'll mention the rectilinear regression Model, Decision Tree Model, and Random Forest Model.

❖ LINEAR REGRESSION MODEL

Linear Regression is employed for predictive analysis. it's a way that explains the degree of relationship between two or more variables (multiple regression, therein case) employing a best fit line/plane. Simple rectilinear regression is employed once we have, one experimental variable and one variable.

❖ DECISION TREE MODEL

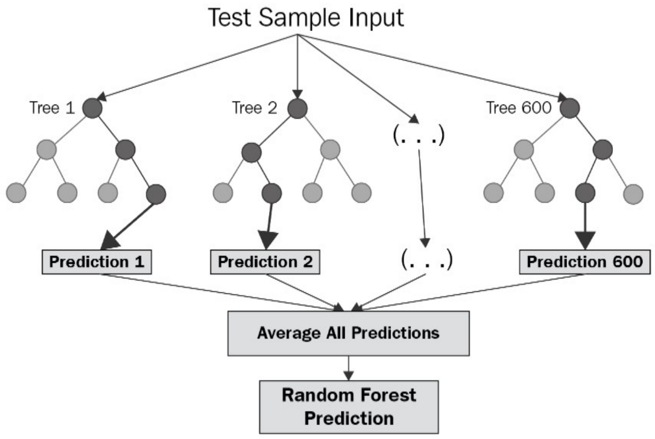
The decision tree builds regression or classification models within the sort of a tree structure. It breaks down a dataset into smaller and smaller subsets while at an equivalent time an associated decision tree is incrementally developed. the ultimate result's a tree with decision nodes and leaf nodes. a choice node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast, and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a choice on the numerical target. The topmost decision node during a tree which corresponds to the simplest predictor called the root node. Decision trees can handle both categorical and numerical data.



Predictors				Target
Outlook	Temp	Humidity	Windy	Hours Played
Rainy	Hot	High	False	25
Rainy	Hot	High	True	30
Overcast	Hot	High	False	46
Sunny	Mild	High	False	45
Sunny	Cool	Normal	False	52
Sunny	Cool	Normal	True	23
Overcast	Cool	Normal	True	43
Rainy	Mild	High	False	35
Rainy	Cool	Normal	False	38
Sunny	Mild	Normal	False	46
Rainy	Mild	Normal	True	48
Overcast	Mild	High	True	52
Overcast	Hot	Normal	False	44
Sunny	Mild	High	True	30

❖ RANDOM FOREST MODEL

- Random forest is a Supervised Learning algorithm which uses ensemble learning method for classification and regression.
- Random forest may be a bagging technique and not a boosting technique. The trees in random forests are run in parallel. there's no interaction between these trees while building the trees.
- It operates by constructing a mess of decision trees at training time and outputting the category that's the mode of the classes (classification) or means prediction (regression) of the individual trees.
- A random forest may be a meta-estimator (i.e. it combines the results of multiple predictions) which aggregates many decision trees.



4. CONCLUSIONS

In this paper, we examine the matter of demand forecasting on an e-commerce internet site. We proposed a stacked generalization method consists of sub-level regress. we've also tested results of single classifiers separately alongside the overall model. Experiments have shown that our approach predicts demand a minimum of nearly as good as single classifiers do, even better using much less training data (only of the data-set). we expect that our approach will predict far better when more data is employed. Because the difference isn't statistically significant between the proposed model and random forest, the proposed method is often wont to forecast demand thanks to its accuracy with fewer data. within the future, we'll use the output of this project as a part of the worth optimization problem which we are getting to work on.

ACKNOWLEDGEMENT

We express esteemed gratitude and sincere to our project guide Ms. Neera Chaudhary, who has been an excellent mentor, helped and encouraged us with extreme sincerity and affection. We are much obliged to our honorable Head of Department Dr. Sachi Gupta whose support and cooperation were always helpful and motivating. And also many thanks to Project Coordinator Dr. Lalit Saraswat who supported us in numerous ways and Our parents to whom we are forever grateful for constant support and encouragement.

REFERENCES

1. Archisha Chandel, Akanksha Dubey, Saurabh Dhawale, Madhuri Ghuge, “Sales Prediction System using Machine Learning”, International Journal of Scientific Research and Engineering Development-– Volume2 Issue 2, Mar –Apr 2019.

2. Gopal Behera and Neeta Nain, “A Comparative Study of Big Mart Sales Prediction”.

3. Punam, K., Pamula, R., Jain, P.K.: A two-level statistical model for big mart sales prediction. In: 2018 International Conference on Computing, Power and Communication Technologies (GUCON). pp. 617–620. IEEE (2018) .