**Project Report**

**CSE 445.2 – Machine Learning**

**Title:** **BigMart Sales Prediction**



**Submitted By**

1712336642 - Md Sharif Hossain

1711786642 - Meherab Mamun Ratul

1711316642 - Faijul Abedin

**Submitted To**

Syed Athar Bin Amir – SAA3

Lecturer

**ELECTRICAL AND COMPUTER ENGINEERING**

**NORTH SOUTH UNIVERSITY**

Fall 2020

# Abstract

The data scientists at BigMart have collected 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.

Table of Contents

[Abstract 1](#_Toc49694867)

[Introduction 3](#_Toc49694868)

[Background 3](#_Toc49694869)

[Design Methodology 3](#_Toc49694870)

[Results & Discussion 3](#_Toc49694871)

[Conclusion 3](#_Toc49694872)

[References 3](#_Toc49694873)

# Introduction

This is a data training and testing project from a set of secondary data. The dedicated company might use the results to understand the properties of products and stores which play a key role in increasing sales. The project will show how different variables can affect the sales quantities and how to increase sales effectively. We have both train (8523) and test (5681) data sets. Train data set has both input and output variable(s). We need to predict the sales(output) for test data set.

# Background

# **Hypothesis Generation:**

# This is a very important stage in any machine learning process. It basically involves brainstorming and coming up with as many ideas as possible about what could affect the target variable. It helps us in exploring the data at hand more efficiently and effectively. Hypothesis Generation should be done before seeing the data or else we will end up with biased hypotheses. Following are some of the hypotheses based on the problem statement.

# - Sales are higher during weekends.

# - Higher sales during morning and late evening.

# - Higher sales during end of the year.

# - Store size affects the sales.

# - Location of the store affects the sales.

# - Items with more shelf space sell more.

# **Variable & Description:**

# 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 particular product

# Item\_Type | The category to which the product belongs

# Item\_MRP | Maximum Retail Price (list price) of the product

# Outlet\_Identifier | Unique store ID

# Outlet\_Establishment\_Year | The year in which 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.

# Design Methodology

<This section should be used to go through what you implemented. If you are doing a paper review only, you should use this section to clearly describe the design used in your own words.>

# Results & Discussion

<You should discuss the findings here, and cover what future steps are being taken to advance this topic>

# Conclusion

We have applied four regression models to analyze the result. The models are Linear Regression Model, Random Forest Model, XgBoost model and Gradient Boost Decision Tree. Among all the applied models, Gradient Boost Decision Tree gave the most accurate result.

# Authorship: Mehrab Mamun Ratul (MM), Md Sharif Hossain (SH), Faijul Abedin (FA)

|  |  |
| --- | --- |
| **Task** | **Taken Care By** |
| Background Analysis | SH |
| Design Methodology | MM |
| Linear Regression Model | FA |
| Random Forest Model | FA |
| XgBoost model | MM |
| Gradient Boost Decision Tree | SH |
| Result and Discussion | MM |
| Conclusion | SH |
| Prersentation Slide | SH, MM, FA |
| Project Report | SH, MM, FA |

# References

**There are no sources in the current document.**

<You should **cite any and all external references** you used to build this project and drew information in this report. Standard IEEE format will be fine, as generated by MS Word as shown above.>