BIG MART SALES PREDICTION MODEL

Submitted by:

Abhilekh Verma

Contents

ACKNOWLEDGMENT

I sincerely thanks to the Data Trained Faculty for the guidance. They have covered the topics like Machine Language, Python & SQL. I had also taken help from YouTube & online videos.

INTRODUCTION

ABSTRACT

Big Mart 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.

PROBLEM STATEMENT

The data scientists at Big Mart 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. Create a model by which Big Mart can analyse and predict the outlet production sales.

OBJECTIVE

- Our main objective is to predict sales of each product at a particular store using machine learning algorithms.
- All the parameters will be analysed through Machine Learning algorithms like Linear Regression, Lasso Regression, Ridge Regression, Elastic Net Regression etc which will help to predict the sales.

DATA DESCRIPTION

- The dataset contains the detailed study of Item Identifier, Item Weight, Item Fat Content, Item Visibility, Item Type, Item MRP, Outlet Identifier, Outlet Establishment Year, Outlet Size, Outlet Type and Outlet Location Type.
- The source of data is taken from GitHub.

METHEDOLOGY

- It gives insights of the dependency of target variables on independent variables using machine learnings techniques to determine the sales because it gives the best outcome.
- The dependent variable is Item Outlet Sales, whereas other variables i.e., Item Identifier, Item Weight, Item Fat Content, Item Visibility, Item Type, Item MRP, Outlet Identifier, Outlet Establishment Year, Outlet Size, Outlet Type and Outlet Location Type are the independent variables.

METRIC USAGE

a. Linear Regression.
b. Lasso Regression.
c. Ridge Regression.
d. Elastic Net Regression.

Hardware and Software Requirements and Tools Used

a) Hardware Requirement:
   i. Intel core i5
   ii. 8 GB Ram

b) Software Requirement:
   i. Python 3.x with packages:
      1. Pandas: Data analysis and manipulation tool

2. NumPy: Provide support for mathematical functions, random number etc.
3. Matplotlib: is a low-level graph plotting library in python that serves as a visualization.
4. Seaborn: is a library mostly used for statistical plotting in python.
5. Scikit-Learn: is an open-source Python library that has powerful tools for data analysis and data mining.

IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES

- R2 score : is used to evaluate the performance of a linear regression mode.
- Linear Regression: Logistic regression is fast and relatively uncomplicated, and it is convenient for you to interpret the results.
- Lasso: The Lasso is a linear model that estimates sparse coefficients with l1 regularization.
- Ridge: Ridge regression is an extension of linear regression where the loss function is modified to minimize the complexity of the model.
- Elastic Net: is a linear regression model trained with both l1 and l2 -norm regularization of the coefficients.
- Cross-Validation-Score: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data.
- Grid Search CV: This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set.
- Mean Squared Error: this metric gives an indication of how good a model fits a given dataset.
- Root Mean Squared error : is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed.

TESTING OF IDENTIFIED APPROACH(Algorithms)

a. Train Test Split
b. Linear Regression
c. Lasso Regression
d. Ridge Regression
e. Elastic Net Regression
f. Grid Search CV
g. Cross Validation

KEY METRICS FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION.

1. Analysed data for any unique values and converted categorical columns into numerical columns.
2. Analysed data for distribution.
3. Cross Validation for cross validates the accuracy-score from overfitting.
4. Hyper parameter tuning using Grid Search CV for making the prediction better.

CONCLUSION

- Built the model with ~56% Accuracy.

REFERENCES

- Data trained course videos.
- Google Search.
- YouTube.
- GitHub.
- UCI Machine learning repository.

LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE

This study gives me opportunity for lots of learning starting from various types of plotting like histograms, boxplot, scatterplot, line chart and many more graphs. These graphs helped me to analyse different aspects of data like outlier, skewness, correlation etc.

It also helped me to learn how to apply various model techniques on data and enable predications.