UE20CS312: Data Analytics

BIG-MART SALES PREDICTION ANALYSIS

Team Details:

Pratheek Babu	PES1UG20CS674
Pruthvi S Kodmurgi	PES1UG20CS676
Vijay Ram Enaganti	PES1UG20CS700
Richa Ramesh	PES1UG20CS712

Introduction:

Sales prediction is an important aspect of any business, as it allows companies to make informed decisions about their operations. By accurately predicting future sales, businesses can ensure that they have sufficient inventory on hand to meet customer demand, and can also make informed decisions about hiring and training new staff. Additionally, sales prediction can help businesses identify trends and patterns in consumer behavior, allowing them to tailor their marketing and sales strategies to better target their audience. Furthermore, accurate sales prediction can help businesses avoid financial pitfalls by allowing them to anticipate slow periods and adjust their spending accordingly. Overall, sales prediction is a vital tool for any business looking to stay competitive and successful in today's market.

Sales prediction can be a difficult and challenging task for businesses. One of the main problems is the lack of accurate and reliable data. Often, sales data can be affected by external factors such as weather, competition, and market trends, making it difficult to accurately predict future sales. Additionally, sales can vary greatly depending on the product or service being sold, making it difficult to create a one-size-fits-all approach to prediction. Another problem is the potential for human error in the prediction process, as sales forecasting is often based on subjective estimates and assumptions. Overall, predicting sales can be a complex and unpredictable process, requiring careful analysis and consideration of a wide range of factors.

The goal of this project was to identify patterns and trends in sales at a retail outlet, and to develop a model for predicting future sales. To achieve this, we collected data on various factors that could impact sales, such as the location of the outlet, the size of the store, the target customer demographic, and the types of products sold. We used a combination of statistical techniques and machine learning algorithms to analyze the data, and to develop a predictive model. The results of our analysis and the recommendations for improving sales are presented in this report.

Methodology:

1. Data collection and exploration

We used a dataset of sales data from a broad market that comprises 12 attributes for our investigation. The fundamental characteristics of the predicted data are defined by these 12 criteria. Predictors and answer variables are two different kinds of attributes. We will make use of a dataset that has 8523 objects from different locations and cities. Our dataset focuses on hypotheses at the shop and product levels as its main elements.

Attributes including area, population density, retail capacities, location, and others have been included at the store level. The dataset is then divided into training and testing halves.

2. Statistical data computation

- Univariate analysis
- Bivariate analysis
- Multivariate analysis

3. Regression

3.1 Linear Regression

A parametric method known as regression is used to forecast a continuous or dependent variable from a set of independent variables. This method is known as parametric since different assumptions are made based on the data set.

For simple linear regression, apply the

$\mathbf{Y = \beta o + \beta 1X} + \in$ (1)

equation presented in eq. (1).

These parameters are as follows:

Y -Predictable variable.

X -Variable(s) that will be used to make a forecast

o -When X = 0, it is referred to as the prediction value or intercept

3.2 Random forest regression

The random forest algorithm is incredibly effective at predicting sales. The process of anticipating the results of machine learning activities is straightforward and easy to understand. Because they have hyper parameters that are similar to decision trees, random forest classifiers are used in sales prediction. The tree model resembles a tool for making decisions.

Dataset Link: BigMart Sales Data | Kaggle

3.3 Lasso regression

Penalized regression method is another name for lasso regression. Typically, this approach is used in machine learning to choose the subset of variables. Compared to other regression models, it offers better prediction accuracy. Lasso Regularization improves the readability of models.

3.4 XGBoost regressor

A decision-tree-based ensemble machine learning method called XGBoost makes use of gradient boosting. Artificial neural networks beat all current algorithms or frameworks in prediction challenges involving unstructured data (pictures, text, etc.). However, decision tree-based algorithms do better with small to medium-sized structured/tabular data.

3.5 Ridge regressor

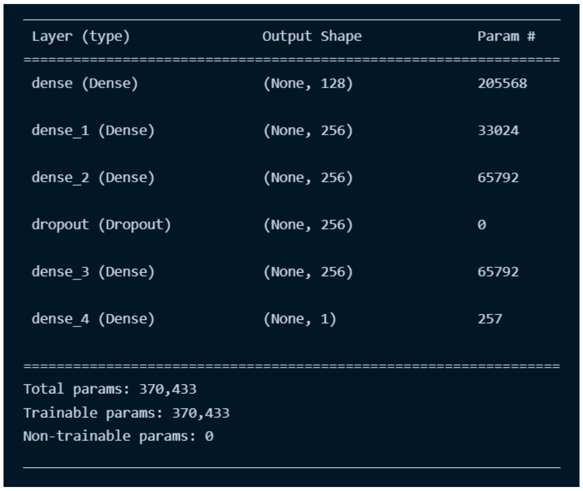
When a data set exhibits multicollinearity or when there are more predictor variables than observations, ridge regression can be used to build a parsimonious model (correlations between predictor variables).

4. Artificial Neural Networks

In order to establish the mapping of inputs to outputs by identifying the underlying correlations between them, artificial neural networks connect a number of neurons (input/output units) in many layers. A data-driven learning process is required because of the enormous number of layers and neurons in such networks, as well as the wide range of their kinds (linear or nonlinear). Each unit (neuron) will be given a weight, which will be adjusted as the stage of training progresses. The least number of neurons in a weighted network capable of turning inputs into outputs with the smallest fitting error deviation is the last step.

An input layer, a hidden layer (maybe one), and an output layer make up the network architecture. It also goes by the moniker MLP due to the many layers (Multi Layer Perceptron). As a "distillation layer," the hidden layer extracts some of the most important patterns from the inputs and passes them on to the next layer so they may be seen. By identifying and rejecting all except the most important data from the inputs, it increases the network's speed and efficiency. The activation function serves two crucial purposes. Non-linear input-output interactions are captured. It helps to turn the input into an output that is more useful. The output of the sigmoid activation function is a number between 0 and 1.

There might be more activation mechanisms like Tanh, softmax, and RELU. As a novelty, we have also used the tanh and softmac functions to check the efficiency and compare with other activation functions. The aforementioned network design is described as a "feed-forward network" since input signals only move from inputs to outputs in one way. Additionally, we are able to create "feedback networks" with bidirectional signal flow. A high-accuracy model generates forecasts that closely match the measured values. As a result, in the table above, the values in Column X should be very close to the values in Column W. The prediction error is the difference between column W and column X. The goal is to create a model utilizing sales information from 1559 items sold across 10 Big Mart locations in various cities that can predict sales for each product at various merchants.The sequential arrangement of the Keras layers is the fundamental idea behind Sequential API, hence the name. Data flows from one layer to the next in the designated order until it reaches the output layer in the majority of ANNs, which contain layers that are structured in a sequential sequence.



```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 128)               205568

 dense_1 (Dense)             (None, 256)               33024

 dense_2 (Dense)             (None, 256)               65792

 dropout (Dropout)           (None, 256)               0

 dense_3 (Dense)             (None, 256)               65792

 dense_4 (Dense)             (None, 1)                 257

=================================================================
Total params: 370,433
Trainable params: 370,433
Non-trainable params: 0
_________________________________________________________________
```

Image-1: Model Summary

5. Metrics and Parameters:

We use a few metrics to measure the correctness of the predictions made by the Regression and ANN Models.

Optimal predictions are made by altering or tuning the parameters of the models used. Metrics help in identifying the optimal set of predictions and hence the optimal set of values for the parameters.

For Regression Models:

1. Linear Regressor
   a. Loss = Mean Absolute Error
2. XGBoost Regressor
   a. N estimators = 1000
   b. Learning Rate = 0.05
   c. Loss = Mean Absolute Error
3. Lasso Regressor
   a. Alpha = 0.01
   b. Loss = Mean Absolute Error
4. Ridge Regressor
   a. Alpha = 1
   b. Loss = Mean Absolute Error

For ANN Model:

- Optimizer = Adam
- Loss = Mean Absolute Error
- Kernel Initializer = Normal
- Epochs = 100
- Batch Size = 64
- Validation Split = 0.2
- Activation Function
  - Relu and Linear
  - Tanh and Linear
  - Sigmoid and Linear

Result and Conclusion:

As a result, exploratory data analytics techniques are employed in this work to forecast bigmart store sales. The suggested system will function more precisely and efficiently. The company's revenue followed the prediction in precise proportion. As a result, superior predictions are obtained as compared to single model predictive approaches. More instance settings and other components can be used to make this sales prediction more creative and effective. Raising the amount of parameters used can significantly increase accuracy, which is crucial in prediction-based systems. The efficiency of the system can also be increased by investigating how the submodels function.

Sales extrapolations are cheap and frequently sufficient for the necessary judgments. Causal approaches are advised when significant changes are anticipated or when it would be beneficial to investigate other tactics.

Also note that the introduction section of this paper is written by AI developed by OpenAI.

Github Link:
[VjayRam/BigMart_Sales_Analysis (github.com)](github.com)