An Analysis of Stop Store Sales Predictions based on Machine Learning Algorithms

Kaukab Fatima[1] Dr. Shaik Shavali Tailor Kanekal[2]

[1]Research Scholar, Dept. of Computer Science and Engineering, Lords Institute of Engineering & Technology, Hyderabad, Telangana

[2]Professor, Dept. of Computer Science and Engineering, Lords Institute of Engineering & Technology, Hyderabad, Telangana

Abstract— **Large supermarket run-centers, also known as Big Marts, now keep track of the sales volume and revenue figures for each individual product in order to estimate potential domestic demand and update inventory management. Inconsistencies and wide trends are frequently discovered by examining the data warehouse's database server. Businesses like Big Mart may use analytics to anticipate possible product sales using several machine learning techniques. To predict the sales of the products in the Big Mart, we used a variety of machine learning algorithms in this project, including Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regressor, Adaboost Regressor, and XGBoost Regression. We find that, of the methods described, XGBoost Regression performs the best in forecasting sales volume. In order to further increase the accuracy, we built a model using XGBoost Regression and fine-tuned it. This model is available on a flask application, where users may log in, enter the details of a product, and receive accurate forecasts of its sales.**

Keywords—Linear Regression, Polynomial Regression, Ridge Regression, Xgboost Regression.

I. INTRODUCTION

The competition between various retail stores and huge supermarkets is growing more ferocious and fierce on a daily basis as a result of the quick expansion of international retailers and online purchases[1]. In order to manage inventory, shipping, and operational tasks for the business, it must be able to pull in a sizable number of clients in a brief period of time and predict the amount of income for each product. In order to outperform low-cost methods used for prediction, the current machine learning technology is highly advanced and offers approaches for estimating or forecasting sales patterns for just about any type of firm[3]. Always more precise forecasting is helpful for developing and improving the business plan, which is also very advantageous.

There has been a lot of work done up to this point that was specifically intended for the field of transaction forecasting[2]. The substantial research on big-mart agreements that has been done is briefly summarised in this section. Several other Measurable approaches, including regression, auto-regressive integrated moving average (ARIMA), and auto-regressive moving average (ARMA), have been used to build a few deals prediction standards. A. S. Weigend et al. [6]suggested combining the occasional quantum relapse strategy with the (ARIMA) Auto-Regressive Integrated Moving Average method to manage daily food deals anticipating. Predicting transactions is a difficult problem that is affected by both internal and external factors

.

II. Research Background

A. Problem Statement

Individuals' purchasing power has been rapidly increasing in both offline and online settings in recent years[4]. Large supermarkets occasionally offer a variety of deals during holidays like Christmas and the New Year. Since it is clear that sales will be quite strong during this time, it is crucial for management to accurately forecast product sales and maintain proper inventory management[5]. The superstore should purchase sufficient product inventory and sell it off within the time. If it fails to do so and its projections prove to be inaccurate, the mart will incur significant losses. There are currently no methods available that can correctly forecast product sales based on historical data. It is only carried out by managers, who carefully examine historical data and attempt to gauge sales volume in light of several events and other pertinent facts[7]. This activity cannot be completed by an automated system; instead, human assistance is required.

B. Aim Of The Project

This project's main goal is to forecast how many things will be sold at a superstore. For this, we used the Big Mart sales data set from Kaggle. The R2 score of this data set is compared after preprocessing, analysis, and feeding it to many regression algorithms. Stock booking and inventory maintenance are outside the purview of this project.

Technical Approach

The technological strategy to solve the issue is listed below:

1. Dataset identification
2. Exploratory Analysis of Data
3. Dataset preparation
4. Running the dataset through many algorithms to see which one best fits the situation.
5. Developing a final classifier model and training the final classifier
6. Validating the ultimate classifier and recording the outcomes.

III. System Analysis

A. Research Gap

It goes without saying that sales would be quite strong during special days of the year like new year's, Christmas, etc. Therefore, it is crucial for management to accurately anticipate product sales and manage their inventory without any problems. The superstore should purchase sufficient product inventory and sell it off within the time. If it fails to do so and its projections prove to be inaccurate, the mart will incur significant losses. There are currently no methods available that can correctly forecast product sales based on historical data. It is solely carried out by managers who carefully examine historical data and attempt to anticipate the sales volume based on several events and other pertinent information. Predictions require human effort.

B. Proposed System

In this project, we suggest using a variety of machine learning algorithms, such as Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regressor, Adaboost Regressor, and XGBoost Regression, to predict the sales of the products in the Big Mart. We then use the algorithm that performs best to build a model to predict sales volume. We want to host this model on a flask application where users can log in, enter the details of the product, and receive the pertinent forecasts regarding the product's sales.

This suggested approach will evaluate consumer behavior based on their past conduct. These client records are collected to form a data set. We make predictions about whether or not the customer's loan will be approved using these data sets and a machine learning model that has been trained. These computer algorithms forecast the likelihood that a consumer will be able to pay back the lending credit or not.

After testing, the model determines whether the new application is a good candidate for loan approval or not based on the inference it draws from the training sets of data to determine if a client would indeed be capable of repaying his loan or not

Advantages of Proposed System:

- High precision
- Extendable to real-time settings.

IV. Algorithmic Process

A. Creating Model

The algorithm used: XGBoost Regressor

Below is the technical approach used for the Loan recommendation system using the Loan review dataset

1. Data cleaning and visualization
2. Feature Extraction
3. Data Preparation
4. Splitting the data set into test data and training data
5. Data modeling using Sequential and Decision Tree algorithm classifier
6. Data Evaluation and prediction
7. Prediction of a sales in big marts systems

1) Feature extraction:

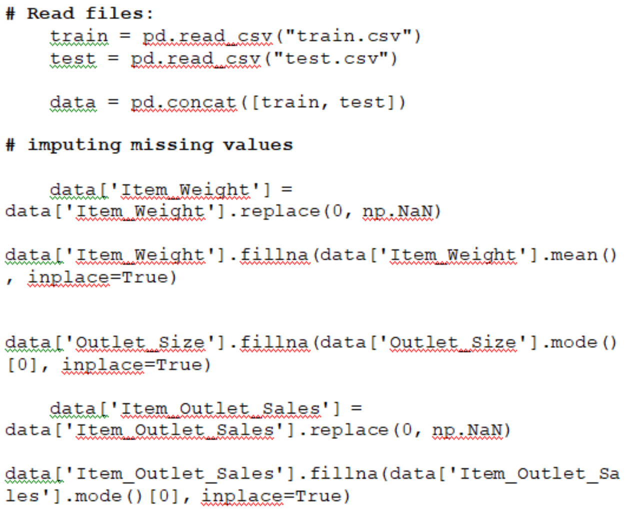
Below are the features present in the dataset:

Feature Name	**Type**
Item_Identifier	String
Item_Weight	String
Item_Fat_Content	String
Item_Visibility	String
Item_Type	String
Item_MRP	String
Outlet_Identifier	String
Outlet_Establishment_Year	String
Outlet_Size	String
Outlet_Location_Type	String
Outlet_Type	String
Item_Outlet_Sales	String

Feature Extraction Details:

A. Imputing Missing Values

When no information is given for one or more elements, a whole unit, or both, this is known as missing data. Missing data is a major issue in real-world situations. In pandas, missing data can also refer to NA (Not Available) values. Many datasets in DataFrame occasionally arrive with missing data, either because the data was never collected or because it was present but was not captured. We can use the fillna(), replace(), and interpolate() functions to fill in any null values in a dataset by replacing NaN values with one of their own. Each of these functions aids in filling in null values in a data frame's datasets. Interpolate() function is basically used to fill NA values in the data frame but it uses various interpolation techniques to fill the missing values rather than hard-coding the value. Although it uses a variety of interpolation algorithms rather than hard-coding the value, the Interpolate() function is mostly used to fill NA values in data frames.

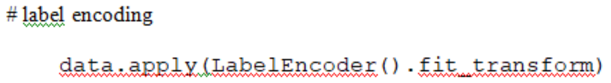


```
# Read files:
    train = pd.read_csv("train.csv")
    test = pd.read_csv("test.csv")

    data = pd.concat([train, test])

# imputing missing values

    data['Item_Weight'] =
data['Item_Weight'].replace(0, np.NaN)

data['Item_Weight'].fillna(data['Item_Weight'].mean()
, inplace=True)


data['Outlet_Size'].fillna(data['Outlet_Size'].mode()
[0], inplace=True)

    data['Item_Outlet_Sales'] =
data['Item_Outlet_Sales'].replace(0, np.NaN)

data['Item_Outlet_Sales'].fillna(data['Item_Outlet_Sa
les'].mode()[0], inplace=True)
```

B. Label Encoding

In machine learning, we usually deal with datasets that contain multiple labels in one or more than one column. These labels can be in the form of words or numbers. To make the data understandable or in human-readable form, the training data is often labeled in words.

Label Encoding refers to converting the labels into a numeric form so as to convert them into a machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.



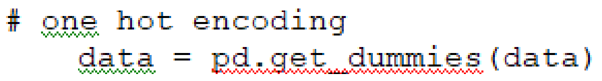
```
# label encoding

    data.apply(LabelEncoder().fit_transform)
```

C. One Hot Encoding

Most Machine Learning algorithms cannot work with categorical data and needs to be converted into numerical data. Sometimes in datasets, we encounter columns that contain categorical features (string values) for example parameter Gender will have categorical parameters like Male, Female. These labels have no specific order of preference and also since the data is string labels, machine learning models misinterpreted that there is some sort of hierarchy in them.

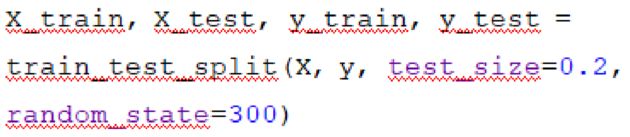
One approach to solve this problem can be label encoding where we will assign a numerical value to these labels for example Male and Female mapped to 0 and 1. But this can add bias in our model as it will start giving higher preference to the Female parameter as 1>0 and ideally both labels are equally important in the dataset. To deal with this issue we will use the One Hot Encoding technique.

One Hot Encoding:
In this technique, the categorical parameters will prepare separate columns for both Male and Female labels. So, wherever there is a Male, the value will be 1 in the Male column and 0 in the Female column, and vice-versa.



```
# one hot encoding
    data = pd.get_dummies(data)
```

D. Defining features and Labels



```
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.2,
random_state=300)
```

E. Generating Synthetic Samples

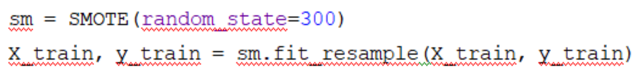
Imbalanced Data Distribution is a phrase used frequently in machine learning and data science and refers to situations where observations in one class are significantly greater or lower than those in the other classes. Machine learning algorithms do not take the class distribution into account since they prefer to improve accuracy by decreasing the error.

SMOTE (Synthetic Minority Oversampling Technique) – Oversampling

One of the most popular oversampling techniques to address the imbalance issue is SMOTE (synthetic minority oversampling technique).

By increasing minority class samples at random and duplicating them, it seeks to balance the distribution of classes.

SMOTE combines already existing minority instances to create new minority instances. For the minority class, it creates virtual training records using linear interpolation. For each example in the minority class, one or more of the k-nearest neighbours are randomly chosen to serve as these synthetic training records. Following the oversampling procedure, the data is rebuilt and can be subjected to several categorization models.



```
sm = SMOTE(random_state=300)
X_train, y_train = sm.fit_resample(X_train, y_train)
```

F. Scaling the data

A data preparation technique for numerical features is called data scaling. Data scaling is necessary for the success of many machine learning techniques, including the KNN algorithm, linear and logistic regression, and gradient descent approaches.

The MinMax Scaler reduces the data inside the specified range, often between 0 and 1. By scaling features to a predetermined range, it changes data. It scales the values to a particular value range while preserving the original distribution's shape..

V. PROJECT IMPLIMENTATION

A. *Proposed Modular Implementation*

Below is the proposed modular implementation of the project. It consists of modules:

1. Admin
2. User

1. Admin Module:

The admin of the system is responsible for the activities like:

1. Uploading the dataset
2. The dataset's data analysis
3. Data Preparation
4. Divvying up the dataset into training and test portions
5. Conditioning the model for various regression techniques
6. Examine how well the algorithms performed on the provided dataset.
7. Use the XGBoost regressor algorithm to build the model.

2. User Module:

The system's user may take advantage of the following available machine learning services:
logging in and entering item/product information to forecast future sales

.

B. *SYSTEM DESIGN*

1. Data Flow Diagram: Admin

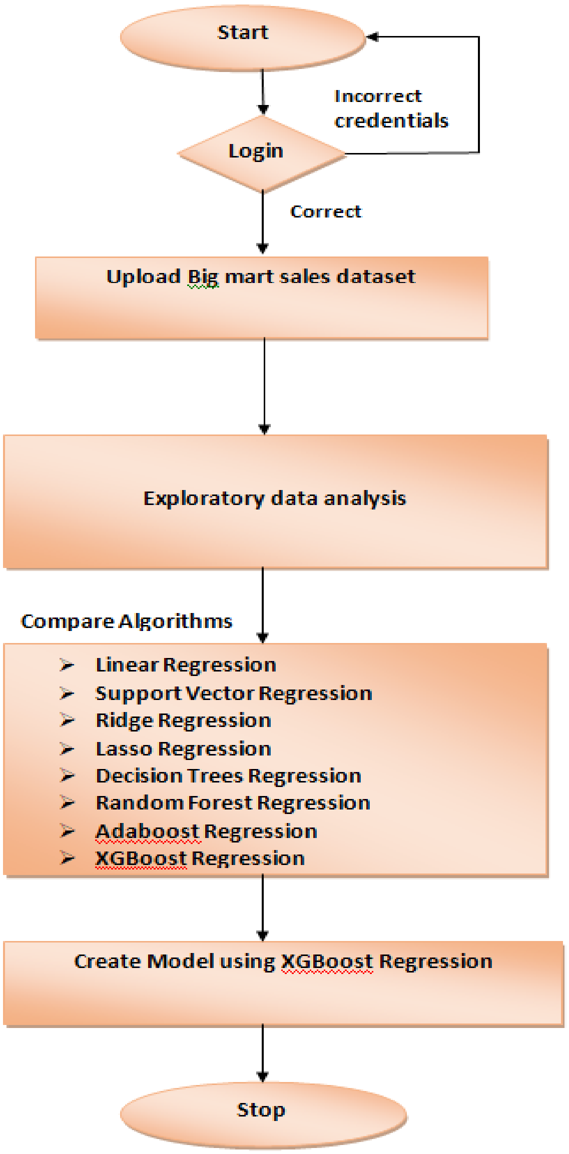


Figure 1: A Data Flow Diagram for Admin

2. Data Flow Diagram: User

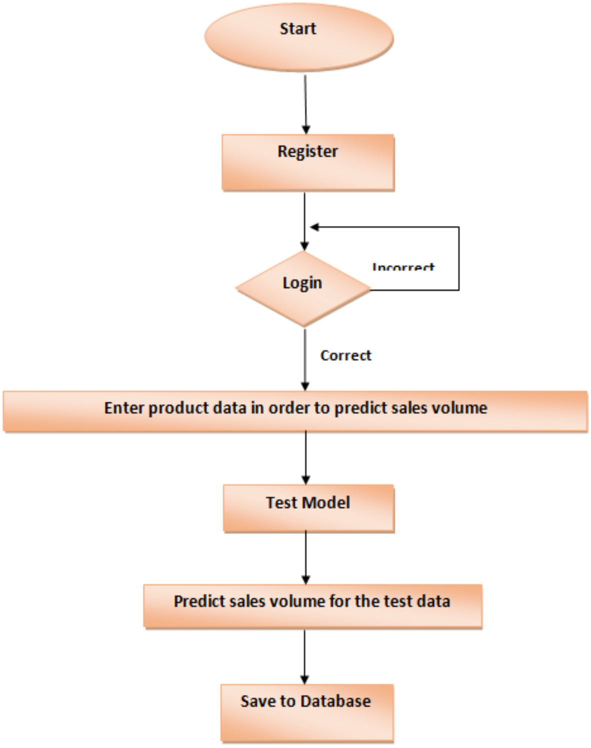

Start
Register
Login
Incorrect
Correct
Enter product data in order to predict sales volume
Test Model
Predict sales volume for the test data
Save to Database

Figure 2: A Data Flow Diagram for Admin

VI. Implementation And Result Analysis

A. Project execution process:

1. Upload Dataset

The system administrator can upload datasets that are used to train machine learning models on this page. To upload a file to a server, an administrator must first choose the file by clicking the Choose file button, then click the Upload button. A success message indicating that the file was successfully uploaded would be shown once the upload was finished. We are utilising the dataset train.csv for this project.

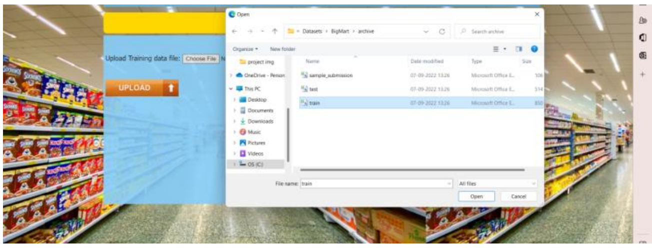
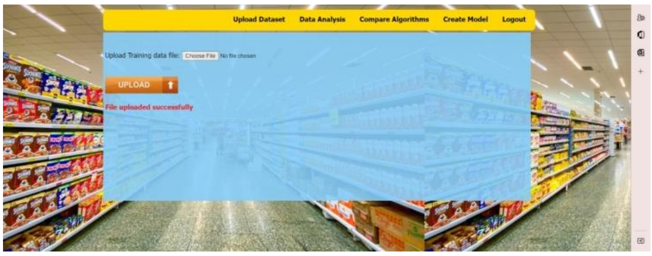


Figure 3: Upload Dataset



File uploaded successfully

Figure 4: File uploaded

2. Data Analysis

Exploratory Data Analysis is performed on the dataset in order to clean the dataset for any missing data, identify patterns, and identify the relationships of various parameters of the outputs with the help of graphs, statistics, etc. so that Data Analysis can be performed.

a) Item Type Analysis:

The below graph shows the Education Analysis of an individual from the Training dataset Loan_Train.csv File.

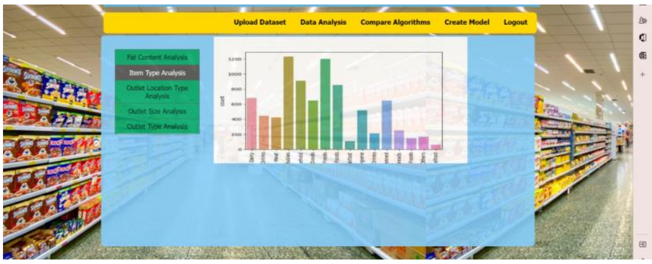


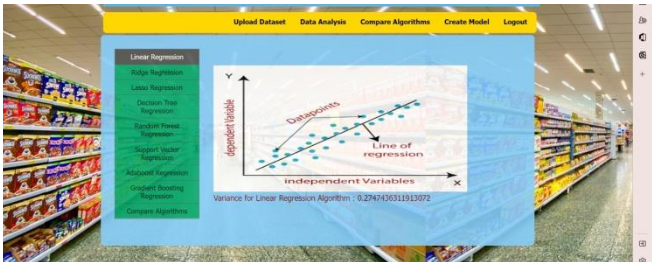
Figure 5: Item Type Analysis

3. Compare Algorithms

On this page, the admin can feed the dataset to various Algorithms to train them and get the test accuracy for each algorithm.

a) Linear Regression:

When the dataset is fed to Linear Regression algorithm we observe that the variance is 0.27

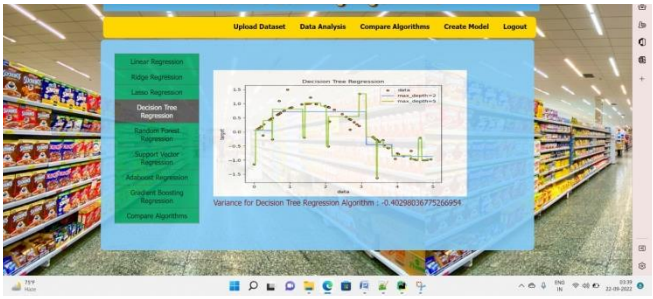


Variance for Linear Regression Algorithm : 0.2747436311913072

Figure 7: Linear Regression

b) Decision Trees Regression

We see that the variance is -0.4 when the dataset is fed into the Decision Trees Regression algorithm.

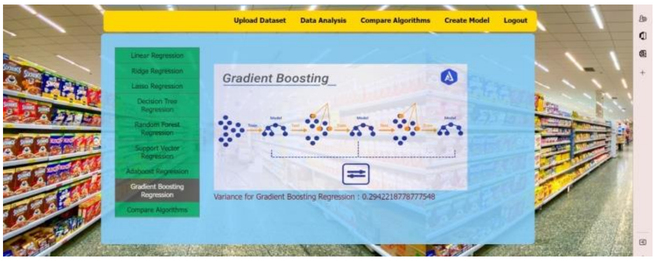


Variance for Decision Tree Regression Algorithm : -0.40298036775266954

Figure 8: DT Regressor

c) XGBoost Regressor

We see that the dataset's XGBoost Regression algorithm's variance is 0.29.

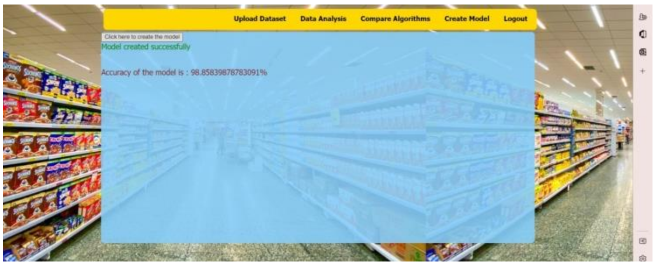


Variance for Gradient Boosting Regression : 0.2942218778777548

Figure 9: XGBoost Regression Algorithm

4. Create Model

The Generate Model button can be used to create the Model. After pressing the button, a success message is presented and the model is built. Our model's precision is 98.85%.



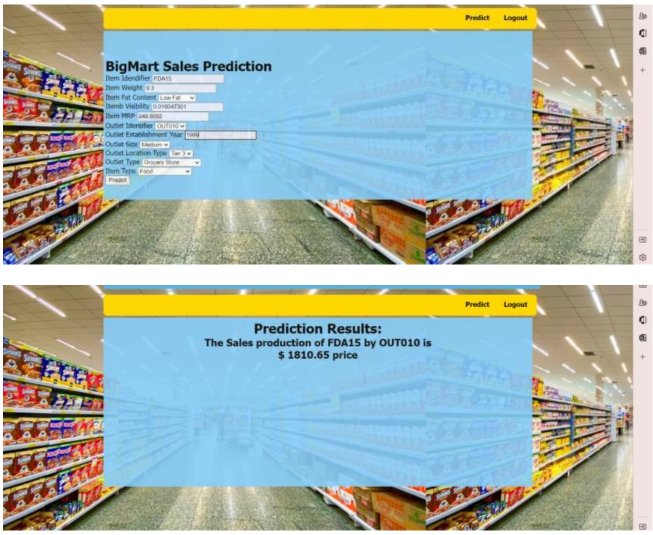
Model created successfully

Accuracy of the model is : 98.85839878783091%

Figure 10: Create Model

5. Test Model:

This is in User Home Page for the user module. The user need to login into the system with his credentials in order to facilitate prediction over a of new applicant Loan data over dataset.



BigMart Sales Prediction

Item Idendifier FDA15
Item Weight 9.3
Item Fat Content Low Fat
Itemb Visibility 0.016047301
Item MRP 249.8092
Outlet Identifier OUT010
Outlet Establishment Year 1999
Outlet Size Medium
Outlet Location Type Tier 3
Outlet Type Grocery Store
Item Type Food

Prediction Results:

The Sales production of FDA15 by OUT010 is $ 1810.65 price

Figure 11: Test model

B. *Metrics Evaluation :*

Mean Squared Error (MSE):

The term Mean Squared Error (MSE) refers to the square of the difference between actual and estimated values in statistics. You may determine how closely a regression line resembles a set of points using the mean squared error (MSE). This is accomplished by squaring the distances between the points and the regression line (also known as the "errors"). The squaring is required to eliminate any unfavourable indications. Additionally, it emphasises bigger discrepancies. Since you're averaging a collection of errors, this error type is known as the mean squared error.

The forecast is more accurate the lower the MSE. A crucial component of the estimation of the statistics is the mean-squared error. Through the use of a special method to determine the square of the mistakes and their average, it aids in calculating the differences between the estimated value and the actual value and provides insight. The value of the MSE is determined by the square and the average. The difference between the value being estimated and the estimated value is the basis for calculating mistakes. It is a risk function because it predicts the amount of the loss. This quantification aids in determining the loss's cause, which might also be the estimator. MSE is largely favourable. There are several methods for determining MSE. For instance, the variance and mean-squared error will be equal if the estimator is not biassed. The unit changes depending on how the quantity is primarily measured.

What is variance?

In terms of linear regression, variance is a measure of how much the mean of the predicted values and the observed values vary from one another. The objective is to have a low value. The r2 score measures what low implies (explained below).

```
****************************************

Linear Regression variance score: 0.2898929014039604

Linear Regression accuracy score: 0.2898929014039604

Model Accuracy: 28.989

The mean squared error (MSE) on test set: 1527402.9605

****************************************

Ridge variance score: 0.2897475632976064

Model Accuracy: 28.975

The mean squared error (MSE) on test set: 1527715.5751

****************************************

Lasso variance score: 0.29025278613992467

Model Accuracy: 29.025

The mean squared error (MSE) on test set: 1526628.8674

****************************************

Decision Tree variance score: -0.4477602327585175

Model Accuracy: -44.776

The mean squared error (MSE) on test set: 3114055.9924

****************************************

Random Forest variance score: 0.2201021623422761

Model Accuracy: 22.010

The mean squared error (MSE) on test set: 1677519.1636

****************************************

Support Vector variance score: -0.25841716778740387

Model Accuracy: -25.842

The mean squared error (MSE) on test set: 2706789.0342

****************************************

Adaboost variance score: 0.19297485037034323

Model Accuracy: 19.297

The mean squared error (MSE) on test set: 1735868.5826

****************************************

XGBoost variance score: 0.30810725637279024

Model Accuracy: 30.811
```

The mean squared error (MSE) on test set: 1488224.8425

****

VII. Conclusion

We were successful in creating and implementing a machine learning model for this project that forecasts sales of various products in a superstore. We have used the Big Mart sales data set from kaggle.com, an open source data set, in order to accomplish this. The data has been cleaned and preprocessed, divided into training and test data, and fed to a variety of regression algorithms, including Linear regression, Support vector regression, Ridge regression, Lasso regression, Decision Trees regression, Random Forest regression, Adaboost regression, and Xgboost regression. That transaction, which we have seen, provides the highest level of accuracy. As a result, we improved the xG boost regressor, which now provides an accuracy of roughly 98%. For this data set, our model performs better than any other machine learning models we have used. Future plans call for expanding this research to include time series analysis using ARIMA.

References

[1] Ching Wu Chu and Guoqiang Peter Zhang, “A comparative study of linear and nonlinear models for aggregate retails sales forecasting”, Int. Journal Production Economics, vol. 86, pp. 217-231, 2003.

[2] Wang, Haoxiang. "Sustainable development and management in consumer electronics using soft computation." Journal of Soft Computing Paradigm (JSCP) 1, no. 01 (2019): 56.- 2. Suma, V., and Shavige Malleshwara Hills. "Data Mining based Prediction of D

[3] Suma, V., and Shavige Malleshwara Hills. "Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics." Journal of Soft Computing Paradigm (JSCP) 2, no. 02 (2020): 101- 110

[4] Giuseppe Nunnari, Valeria Nunnari, “Forecasting Monthly Sales Retail Time Series: A Case Study”, Proc. of IEEE Conf. on Business Informatics (CBI), July 2017.

[5]https://halobi.com/blog/sales-forecasting-five-uses/. [Accessed: Oct. 3, 2018]

[6] Zone-Ching Lin, Wen-Jang Wu, “Multiple LinearRegression Analysis of the Overlay Accuracy Model Zone”, IEEE Trans. on Semiconductor Manufacturing, vol. 12, no. 2, pp. 229 – 237, May 1999.

[7] O. Ajao Isaac, A. Abdullahi Adedeji, I. Raji Ismail, “Polynomial Regression Model of Making Cost Prediction In Mixed Cost Analysis”, Int. Journal on Mathematical Theory and Modeling, vol. 2, no. 2, pp. 14 – 23, 2012.

[8] C. Saunders, A. Gammerman and V. Vovk, “Ridge Regression Learning Algorithm in Dual Variables”, Proc. of Int. Conf. on Machine Learning, pp. 515 – 521, July 1998.IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 56, NO.

7, JULY 2010 3561.

[9] ”Robust Regression and Lasso”. Huan Xu, Constantine

Caramanis, Member, IEEE, and Shie Mannor, Senior Member, IEEE. 2015 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration.”An improved Adaboost algorithm based on uncertain functions”.Shu Xinqing School of Automation Wuhan University of Technology.Wuhan, China Wang Pan School of the Automation Wuhan University of Technology Wuhan, China.

[10] Xinqing Shu, Pan Wang, “An Improved Adaboost Algorithm based on Uncertain Functions”, Proc. of Int. Conf. on Industrial Informatics – Computing Technology, Intelligent Technology, Industrial Information Integration, Dec. 2015.