Comparative Analysis of Sales Prediction using different Machine Learning Techniques

Pratheek Gogate
National College of Ireland
School of Computing
Dublin, Ireland
X22159789@student.ncirl.ie

***Abstract*—The objective behind this paper is to find, if it is possible to apply different machine learning techniques on Retail data-sets to predict sales and to choose the best algorithm based on the result comparing to all machine learning techniques applied. So, in our projects we have taken 3 data-sets from the retail sector for sales prediction, the first one is related to the sales from Mega Mart where sales will be predicted using features like weight of the product, MRP, Fat content and type of the store and many more and the second one was related to sales from super store on the basis of factors like country, region, category and many more and third one was weekly sales from Walmart where sales is predicted on the basis of CPI, Unemployment, fuel price and other factors. The common thing about all the three data-sets was sales prediction and it's just that the features on which the result depends are different. In our research we used Linear Regression, Ridge Regression, Decision Tree, Random Forest and XG Boost for sales forecasting in each data-set and based on the results we chose the best performing algorithms for each data-set based on different evaluation metrices like Root Mean Square Error, Average Absolute Error, Mean Absolute Error, R squared value.**

***Index Terms*—Linear Regression, Ridge Regression, Decision Tree, Random Forest and XG Boost**

I. INTRODUCTION

Life is meaningful when we have goals; If we want to achieve something in life first we need to set a goal and learn about the domain and then try to learn from past experience and then plan and execute, similarly when it comes to retail industry it's very important to plan everything in advance so that there is no hiccups in the future, so for doing that first need to analyse the past data and predict the future and based on that take the necessary decisions. In any field of business, the survival depends on how well we manage the resources and especially in marts it is very critical as over stocking and under stocking impacts the business as product's durability varies. For example, the grocery items won't survive for long as they have very a smaller number of days of expiry date and if over stocking happens all those stocks will be spoilt, and it will be a loss for the company and if under stocking is done customer may lose faith towards that store thinking that this store is not dependable. So since childhood I always wondered how this things will be managed and how the customer's expectation is met, and it's not only about how much stocks should be maintained but it's also how much profit that organization is going to earn as this will help them to plan the future and try for the expansion of business and how products can be advertised and many more things like this, and also company should take care about brands of the product, for example in cloth section it should not be like some random branded cloths are kept in stock instead analyse the previous sales and analyse the market trend and take the decision. So, to explore more about these things and help the company to choose the best Machine learning technique based on comparing the result of all the 5 algorithms we used and help the business to survive and grow, I chose this topic, so by this research it will help these 3 organisations to know about the future sales based on our results and also will get to know which algorithm suits best for each data-set.

A. Data-set1: Big-Mart Product Sales Factors [14]

If we go to a store and if we want buy something we consider many things before deciding to actually buy it. For example if we want to buy a laptop we see many features such as RAM,processor,Graphic card, Memory and many more things and we will decide based on the requirements looking at features which one to buy. So this holds good in all the industry. In the above mentioned data-set our AIM is to predict the sales based on different features like fat content, visibility, Maximum Retail Price, Weight of the item, type of the store, location and many more. So the thing is that some people will be very particular about diet and there is high chance that they may not buy products with more fat content and some may choose to buy products based on their budget and that is where MRP is considered and some people don't like to carry heavy goods so in that case weight of the item comes into picture and some products are more likely to be sold more on city than rural area and vice-versa so in those cases the factors related to region and store comes into picture. This data-set was built in 2013 by Big mart using 1559 products from 10 different stores which are located in different cities to analyse and build models in a such a way that it helps their business.This data-set was taken from Kaggle and this has 14204 rows and 12 columns.

B. Data-set2: Walmart Store Sales [15]

Stock management is very important in retail sector because over stocking may impact the business as there is a chance that some products may get spoilt and where under stocking will also impact the business as it may impact the customer which in hindsight may lose customer's faith on a store and he may start buying goods from other stores. So accurate prediction of sales has a very important say in business. In this data-set we have a Walmart data-set which has weekly sales which can be based on many factors like CPI, temperature, unemployment, seasonal offers and fuel price. For example if the temperature is extremely high or low people may not come out of the house at all or people may also considering going out to shop based on the fuel price and also there is a chance that people may buy more goods because of festival discounts but also there is a chance that this may be false also. So in this data-set we all have these things and objective is to predict the weekly sales, this data-set was prepared by Walmart in US for 45 stores located in different cities. This has 421570 rows and 17 columns and we got this data-set from kaggle.

```
#Importing the dataset
path = 'D:/NCI_Masters/Data mining/Project/Project/dataset1/superstore.csv'
data = pd.read_csv(path)
data.shape
data.head()
```

	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City	State	Postal Code	Region	Product ID	Category	S Categ
0	1	CA-2017-152156	08/11/2017	11/11/2017	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420.0	South	FUR-BO-10001798	Furniture	Bookca
1	2	CA-2017-152156	08/11/2017	11/11/2017	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420.0	South	FUR-CH-10000454	Furniture	Ch
2	3	CA-2017-138688	12/06/2017	16/06/2017	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	California	90036.0	West	OFF-LA-10000240	Office Supplies	La
3	4	US-2016-108966	11/10/2016	18/10/2016	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311.0	South	FUR-TA-10000577	Furniture	Ta
4	5	US-2016-108966	11/10/2016	18/10/2016	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311.0	South	OFF-ST-10000760	Office Supplies	Stor

Fig. 1. Sample code for importing and displaying the data-set

C. Data-set3: Superstore Sales Data-set [16]

Internet has made things easy, People used to write letters and now just a message through whats-app will serve the purpose. People used to ask people for routes while travelling and now we have google maps. Similarly online Shopping is also a area where internet has its biggest contribution. So here we have a data-set related to online Shopping of a Super store. This data-set has sales related data which basically contains the the information related to each product which was sold. So things related to geographic factors like country, state, postal code, region ,city etc and also factors related to product type like product id, name, category, sub category. also regarding delivery of the product and also about customers. So with all this data we need to predict sales and see if this given factors really have their say on the final result. We have 9800 rows and 17 columns in this and this data-set was taken from Kaggle.

II. RELATED WORK [17]

Sales prediction is something which is very important in terms of progress of the company. The Paper [1] is about prediction of sales in big-mart and help to take a business decision, so SMO regression, simple linear regression, linear regression, additive regression, multi-layer perceptron, random forest, and M5P were used for model building and M5P gave them the best possible result. The basis to choose it as a best was based onMAE, RMSE, RAE, RRSE and correlation. In any industry brand value of a company depends on what does they have, their future plans, sales and many more other factors, in [2] aim was to demonstrate the available space in warehouse to predict the future sales in sales-mart and for that they used XGBoost, Linear regression, Polynomial regression, as well as Ridge regression approaches and observed that XGBoost has lowest RMSE score and most efficient. To spend on something we need funds and its very important to manage the funds, in retail sector it depends on how much sales we are going to have in the future. The Objective of the paper [3] was to predict the sales based on different factors. Here they came to a conclusion that gradient boosting algorithm is the best method compared to linear regression, adaboosting regression based on least RMSE value. Advertising is something very influential in terms of how people buy things, so first company should find which are all the items where advertisement is needed, so In [4] it was about finding the items where sales pattern based on factors like price tag, outlet type, outlet location and for that they got the best result using Random Forest regressor than linear regression and the result was based on checking the same with real world data-set. Without predicting the sales in retail sector its very difficult to keep up the stocks without knowing what are the items going to be sold fast and what are the items going to be sold in slow fashion, The vision of this paper [5] was to find the patterns in data- set and help the business to succeed by finding the future sales, so here they used linear regression, decision tree, random forest, ridge regression, and XGBoost and later they found that XGBoost had the best accuracy among above algorithms used. Different Machine learning techniques are used in prediction of future sales so similarly in [6] it was specifically built for multi-linear regression for identifying the sales of different products and to eliminate products which are not in demand. The model built had a good accuracy. Irrespective of field its always the accuracy which is looked at and especially when it comes to business the impact will of accuracy of prediction is more as it will be used to take many big decisions which basically will decide the future of the company, The idea of this project [7] was to predict the sale of the goods on studying the last years sales and hence they used Linear Regression, K-Neighbors Regressor, XGBoost Regressor and Random Forest Regressor to predict and found Random Forest

Algorithm as the best method as it had a accuracy of 93 Not all the algorithms support all the data-sets as sometimes it depends on how well that algorithm suits that data-set hence sometimes it will be a good idea to apply 4-5 algorithms and look for the best result, The motive of this publication [8] was to find the balance between sales volume and profit for which sales was predicted using Linear Regression, Decision Tree, Gradient Boosted Tree,Self Organizing Map (SOM) and Kmean clustering and saw that Gradient Boosted Tree had the best accuracy. Expiry date to a item is very important in the aspect of durability and when it comes to medicine it becomes even more critical. In a medical store if more stock is kept it will be a waste and if the stock is less it can be very problematic in emergency situations, The paper [9] was written on the basis of drug store to predict the sales to avoid under stocking or over stocking and ARIMA Model, Facebook's Prophet Model and XGBoost Model were used for comparison purpose and ARIMA model was chosen as the best based on Root mean Square Error. Walmart is a huge organization who are very good at meeting the expectation of customers that is because they plan everything in advance, The goal of this paper [10] was to forecast the sales of Walmart on weekly basis and choose the best algorithm out of the three based on r squared value and in this they had applied Decision Tree Regressor,Random Forest Regressor, and K Neighbors Regressor for prediction and Random Forest Regressor had the best possible result among them. Comparing result is a best way to get the best possible result, In this paper [11] they wanted to check which algorithms performs best in different situations and hence they chose 5 algorithms namely linear regression, K – nearest neighbor, Auto ARIMA, Prophet, and Support Vector Machine and Stock market, earth and sales forecasting data was used and they could see that Auto ARIMA was the best as it supported 2 of the algorithms and K neighbor was the worst case for time series forecasting. Food items have least durability compare to other items hence prediction of sales is very important in this set of business. Accuracy score, mean absolute error and max error were used in [12] to predict the food sales and for which they had used Modest Direct Regression, Incline Increasing Lapse, Provision Course Lapse, Accidental Forest Lapse, Gradient Boosting Regression, and Random Forest Regression. Forest Regression fared well among all those mentioned methods. As we know sales prediction plays an important role in Supply chain management as the sales is the product of the process. So, predicting that in advance will always be an advantage for the organisation. So, in [13] this paper they have taken a Citadel POS dataset which ranges from 2013 to 2018 which is a cloud based application and it helps store to carryout transactions, manage stocks, vendors, customers, viewing the reports, managing the sales. They applied Linear regression, Random Forest Regression, Gradient Boosting Regression, ARIMA LSTM for sales forecasting and found XGboost had the best result with RMSE of 0.63 and MAE of 0.516.

III. METHODOLOGY

CRISP-DM stand for Cross Industry Standard Process for Data Mining. Its very easy,useful and standardised methodology used in data mini projects as it covers all the steps which are required for a successful project in a proper order which makes things easier for the one who is making the project also for the people who are not directly involved in it.

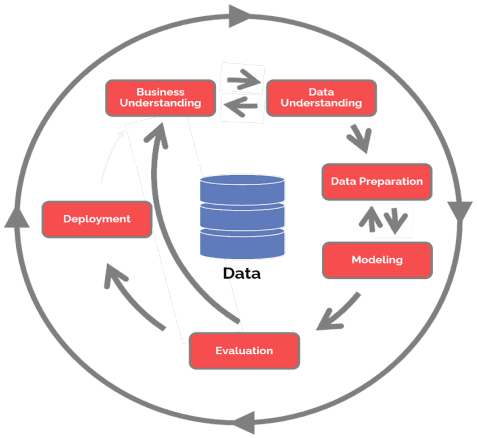

Fig. 2. CRISP-DM Flowchart [18]

This process contains 6 steps:

1.Business Understanding: This steps involves about understanding what the business is about, what is the expectation from this project and how will it help the customer in their business.

2.Data understanding: In this step the data is collected and will try to understand about the data-set by looking at the different features given and to check if the data has enough quality to proceed further.

3.Data preparation : This is very important step before we actually start building the model as first we need to clean the data by removing duplicate values, null values and transform it as per our need so that model can be built.

4.Modeling: it is the actual phase we apply machine learning algorithm based on the the given data and predict the results.

5.Evaluation: In this step the built model is evaluated using different metrics to check if the result is satisfactory or not.

6.Deployment: In this step, the model which was built and evaluated successfully will be deployed in real world and its performance will be monitored.

When it comes to making data mining projects CRISP-DM is the most preferred and used by most of the companies around the world. The reason is simple as it covers all the aspects of Data mining project and addressed properly which ensures smooth completion of the project.

A. Business Understanding:

Big-Mart Product Sales Factors is a data-set which contains the sales related data based on different factors. As we know its one of the biggest name in retail sector and it has its business spread across the globe. So here our goal is to predict the sales

based on different factors using different machine learning technique and select the best performing algorithm for that data-set and thus helping the organisation to predict the sales value so that it would help them to take decisions and also it would help them to understand which factor is having more influence on the sales and it will help the organisations to back their strength factor and work on the non performing factor.

Walmart is a famous retail store in US and it had a difficulty of accurate sales prediction which was affecting their business, so here again our objective is to predict the sales accurately by using best possible machine learning technique to predict the sales on economic and social factors on which company doesn't have control but at least it will help them predict the sales and be prepared with the required amount of stock and plan the business accordingly.

This is a online store where they sell different kind of goods and our objective is to find the future sales based on the historical data which contains mostly the geographic factors. So we need to predict them using ML and help the organisation to answers the questions like, Which area the demand is more for goods and what are the highest selling goods based on area and many more factors like this help them to build their business by taking necessary actions.

B. Data understanding:

In Big-Mart Product Sales we have around 11 factors which may be contributing to the sales. For example, there are fields related to items like visibility, fat content, type, weight which basically means if the item is in a place where its very visible to everyone there is high chance that it gets sold like chocolates are kept near billing section and people to tend buy on the basis of fat content when they are choosy about the diet plans and heavy weight items are less likely to be sold as it will be difficult to carry and also goods will be sold as per requirements, for example: fruits and vegetables are something which are used everyday so they may be sold more in number but where as items like umbrella are seasonal and they will be sold only during rainy seasons. There are some more features which are related to geography which also usually plays its part in sales.We have also done visualisations to understand more about the data. From the figure 3 we can see that there is bit of correlation between MRP and sales, and also there is high negative correlation for outlet age and item_visibility.

In Walmart sales data-set we have around 16 factors which may be contributing to weekly sales. For example there are columns like unemployment, fuel price, CPI, temperature, store, department, seasonal offers etc. For example, if the store is very far from the city then there is less chance of people going there and if the fuel price is high then that chance decreases further and CPI may also affect the business and unemployment to a person also have a impact on his or her buying behaviour as they may choose not to buy goods other than essential things and also Walmart gave seasonal offers during festivals which may also had surge in sales and our aim is to check all this things and see if there is any relation for the same in this data-set using ML techniques and we

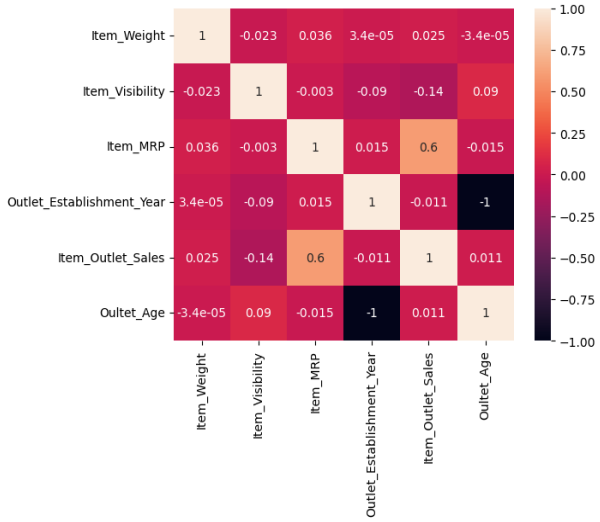

Fig. 3. Correlation Matrix

plotted many graphs visually to understand more about the data. in figure 4 we have plotted a graph which shows top 10 stores based on highest sales.

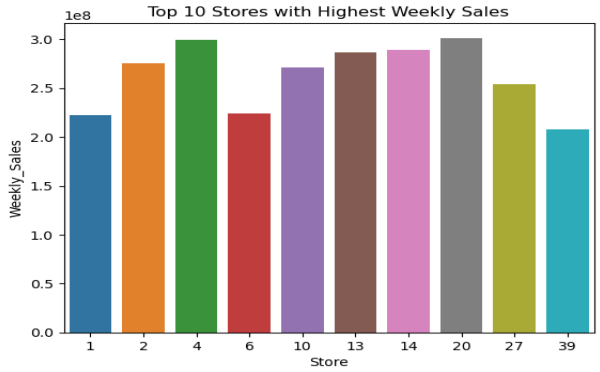

Fig. 4. Top 10 stores based on Sales

In super store sales data, by looking at the data we could tell that its related to online shopping data where sales was the target variable based on different factors related to orders like order date, shipment date, shipment mode and product type. For example if the product is currently available in nearest warehouse and if there is good travel transport is present then order may get delivered quickly and make customers happy and they may tend to shop more and thus sales automatically increase and there are also geographic factors which will have huge impact in online shopping as some places don't have good transport and delivery services and vice versa. In figure 5 we could see a screenshot which gives the output of the code which tells about the data types.

C. Data preparation:

Data preparation is a very important step in machine learning project as it is the step in which data is prepared as per

```
#Trying to understand the dataset
#This is a method that we can use to get general
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9800 entries, 0 to 9799
Data columns (total 18 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Row ID          9800 non-null   int64  
 1   Order ID        9800 non-null   object 
 2   Order Date      9800 non-null   object 
 3   Ship Date       9800 non-null   object 
 4   Ship Mode       9800 non-null   object 
 5   Customer ID     9800 non-null   object 
 6   Customer Name   9800 non-null   object 
 7   Segment         9800 non-null   object 
 8   Country         9800 non-null   object 
 9   City            9800 non-null   object 
 10  State           9800 non-null   object 
 11  Postal Code     9789 non-null   float64
 12  Region          9800 non-null   object 
 13  Product ID      9800 non-null   object 
 14  Category        9800 non-null   object 
 15  Sub-Category    9800 non-null   object 
 16  Product Name    9800 non-null   object 
 17  Sales           9800 non-null   float64
dtypes: float64(2), int64(1), object(15)
memory usage: 1.3+ MB
```

Fig. 5. Information about datatypes

our need to build model and predict the result. This phase includes removing null values and duplicates and transforming the variables as per our need and changing the data-types and many more.

In Big-Mart Product Sales after importing the data-set and understanding it, i checked if there are any null values and found that there were many missing values in item-weight and outlet size and as outlet size was a categorical value i replaced it with mode of that column and for item-weight i plotted a box-plot and found that there are no outliers and mean is exactly placed at the center as you can see the figure 6, stating that data is spread across. So i replaced the missing values with mean. Then in item fat content and item type, there were some duplicate values so we replaced them with proper names having the knowledge about the data-set and then removed outliers from sales column.Then calculated the age of the store based on establishment year. As Item-Fat-Content,Outlet-Type,Outlet-Location-Type,Outlet-Size and item-type are ordinal values, we replaced them with numerical values by transforming them and then later we removed all the duplicates and unnecessary columns and then divided the data-set into train and test and made the data-set ready for model building.

In Wall-mart data-set after importing it and understanding it when we checked for duplicate and null values as we can see in figure 7 and we could observe there were no duplicates nor no null values which indicates that it was a cleaned data-set and then converted Date into date-time format and then fetched Month and Year and then removed outliers from weekly sales and then transformed the type and is-holiday type to numerical value using OneHotEncoder technique and then dropped all the unnecessary columns and split the data-set into training and testing.

In Super sales data-set after going through the data first changed the datatype of shipping date and order date to date-time format from object. Then when checked for missing values we had missing values in Postal Code and as it was categorical variable i replaced it with the mode and then after

```
In [10]: #Lets look into Item_weight
         #Since it is a numerical value, we need to explore a bit
         sns.boxplot(x=data['Item_Weight'], palette='muted')
         plt.title('Item Wieght Distribution')

Out[10]: Text(0.5, 1.0, 'Item Wieght Distribution')
```

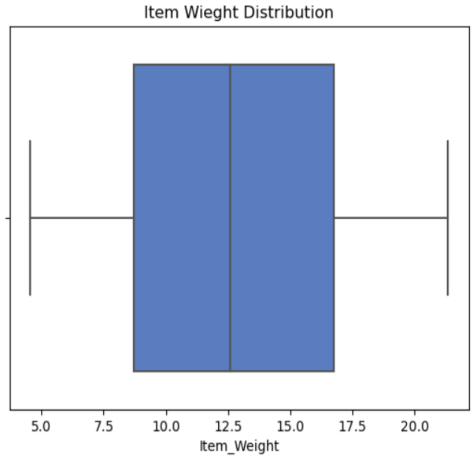

Fig. 6. Item weight Distribution

```
In [32]: #First will check if there are any duplicate values present
         duplicate_rows = wallmart_data.duplicated()
         print(wallmart_data[duplicate_rows])

         Empty DataFrame
         Columns: [Unnamed: 0, Store, Date, IsHoliday, Dept, Weekly_Sales, Temperature, Fuel_Price, MarkD
         kDown4, MarkDown5, CPI, Unemployment, Type, Size]
         Index: []

In [33]: #Finding the missing values
         print('Missing Values by Count: \n\n',
               wallmart_data.isnull().sum().sort_values(ascending=True),'\n\nMissing Values by %:\n\n',
               wallmart_data.isnull().sum().sort_values(ascending=True)/wallmart_data.shape[0] * 100)

         Missing Values by Count:

          Unnamed: 0      0
         Unemployment    0
         CPI             0
         MarkDown5       0
         MarkDown4       0
         MarkDown3       0
         MarkDown2       0
         Type            0
         MarkDown1       0
         Temperature     0
         Weekly_Sales    0
         Dept            0
         IsHoliday       0
         Date            0
         Store           0
         Fuel_Price      0
         Size            0
         dtype: int64
```

Fig. 7. Finding Duplicates and Null values

visualisation removed all the unnecessary columns and also transformed Ship Mode, Segment, Category, and Sub-Category by replacing them with dummy values and then split the data into training and testing for model building purpose, in figure 8 we could see the coding part for transformation which i mentioned above.

D. Modeling:

Model building is very important phase for which all the above phases needs to performed properly. Model building is something where we usually predict or classify the values

```
#We have categorical values such as Ship Mode, Segment, Category, and Sub-Category, we can transform them.
# Select the columns to keep
data = data[['Ship Mode', 'Segment','Region', 'Category', 'Sub-Category', 'Sales']]

# One-hot encode categorical variables
data_encoded = pd.get_dummies(data, columns=['Ship Mode', 'Segment', 'Category', 'Sub-Category','Region'])

# Viewing the encoded dataframe
data_encoded.head()
```

	Sales	Ship Mode_First Class	Ship Mode_Same Day	Ship Mode_Second Class	Ship Mode_Standard Class	Segment_Consumer	Segment_Corporate	Segment_Home Office	Category_Furniture
0	261.9600	0	0	1	0	1	0	0	1
1	731.9400	0	0	1	0	1	0	0	1
2	14.6200	0	0	1	0	0	1	0	0
3	957.5775	0	0	0	1	1	0	0	1
4	22.3680	0	0	0	1	1	0	0	0

5 rows × 32 columns

Fig. 8. Transforming the variables

based given data-set and help the organisations to get some insights and help them to take business decisions.

```
#Model Building 2: Ridge Regression
#Assigning the function
model_ridge_regression = Ridge(normalize=True)

#fitting the model built against training data
model_ridge_regression.fit(X_train, y_train)

#predicting the data based on the model with the help of test data
y_prediction = model_ridge_regression.predict(X_test)

#Evaluation metrices
rmse_rd = np.sqrt(mean_squared_error(y_test, y_prediction)) * 100 / np.mean(y_test)
mae_rd = mean_absolute_error(y_test, y_prediction) * 100 / np.mean(y_test)
aae_rd = mae_rd / (np.mean(y_test)) * 100
r2_rd = r2_score(y_test, y_prediction)

#printing the result
print('Root Mean Squared Error:', rmse_rd,)
print('Mean Absolute Error:', mae_rd,)
print('Average Absolute Error:', aae_rd,)
print('R-squared:', r2_rd,)
```

```
Root Mean Squared Error: 52.95429402091838
Mean Absolute Error: 42.1077549888623
Average Absolute Error: 2.0796764178062
R-squared: 0.40347785696206906
```

Fig. 9. Sample code for decision tree model building

as all the three data-sets were chosen from same domain and almost same functionality its just that factors affecting the result were different. So the common thing about all these three data-sets were sales prediction based different factors like geography, products, order related or delivery related and so on. So for prediction of values based on our domain knowledge and looking at the data-sets and also after referring to many research papers thought Linear Regression, Ridge Regression, XGBoost, Decision Tree and Random Forest were the best 5 possible algorithms for the same and so based on the result we could choose the best performing algorithm for each data-set. Figure 9 shows a sample code for decision tree model building.

Linear Regression: Linear Regression is Supervised learning technique which is used mainly for prediction, In this we predict the value of dependent variable using a independent variable. In this we assume that There should be linear relationship between independent variable and output variable. The equation of linear regression is: y=c+b*x where y is estimated dependent variables c is constant and b is regression coefficient and x is independent variable. Since we have a similar case where we need to predict the future sales based on different factors given in the data-set and felt its wise to use the same in our project and also same has been used in [1].

Decision Tree: Decision Tree is Supervised machine learning technique used predict or categorise values based on how previous questions were answered. The two most important things we consider in this are Entropy and Information gain. 1.Entropy is impurity in the data and the column which has least impurity will be chosen as Root node. 2.Information Gain is something which measures reduction in impure data and its used for splitting the root-node to take decisions. Usually Leaf nodes will be the possible outcomes. As we have a data where our final aim is to predict the sales and we know that this depends on different factors and for that we need to split our factors, hence we are planning to proceed with this method, we had similar case in [5] where its used.

Random Forest: Random Forest is powerful machine learning technique which is both used in classification and regression tasks, its a ensemble technique where multiple decision trees are constructed and combined to make accurate predictions. In this firstly many decision trees are constructed and for classification they use majority voting and for regression they take average of predictions and as our aim is also to predict a value, we chose this for our project and in [7] its used as the situation was similar.

Ridge Regression: Ridge Regression is a standardized techniques used in linear regression, the best thing about this technique is it reduces computation cost and its mainly used when the coefficient value of regression is high. Here Penalty term is added to the cost function of linear regression which shrinks the value of the coefficient to almost zero which will make sure that result is majorly not dependent on any single factor thus makes sure that model is not over-fitted and keeping in mind this we are planning to use the same in our project so that result will be efficient which has been already used in [5].

XGBoost: Extreme Gradient Boosting id a open source library used for optimization purpose in regression and it is basically for structured data and it helps to achieve scalability and efficiency. The special thing about this method is, it can deal the missing values in data-set. Basically it adds new features to ensemble model and help to improve the previous model iteratively which make sure that error is reduced and

help to achieve high accuracy and this can be used in both regression and classification tasks which makes it more reliable thus convinced me to use this model in our project and on referring to [9] it made me more confident about using this.

E. Evaluation:

In any field evaluation is a must needed process as it tells whether we have got the desired result or not. Similarly in this case we have built the model as per our requirements and we need to check if those algorithms are meeting our expectation or not,here in case we have built 5 models for each data-set and choose the best model for each data-set.

RMSE:Root Mean Square Error is the standard measure used for evaluation of models in Machine learning, basically it is a difference between actual and predicted value, its calculated using square root of difference of predicted value and actual value, the less the value of RMSE the better is the result but the problem with this method is that it is vulnerable against the outliers and it doesn't exactly tells on which side the error is, hence its used with other evaluation methods. on reading different research papers we could see the same has been used in many papers and its also mentioned in [11].

AAE: Average Absolute Error is used to measure the accuracy of a model, it is absolute difference between predicted and expected value, the good thing about this measure is that its not affected by outliers but still it doesn't give us much information about the direction of the error but its effective measure when its used with other measures its useful for the evaluation. When I made research on this I felt this metric can be used for this prediction accuracy.

MAE: Mean Absolute Error is commonly used metric to measure used in regression models, it is the average difference between predicted and expected values and compared to RMSE and AAE its less sensitive to Outliers, but it also has a same problem of not informing about the direction of error, but still over all its a good measure when used with other metrics. its superiority over RMSE and AAE convinced me to use this metric in our project and its already been used in [1].

R Squared value: R-squared value is evaluation metric especially used in regression. R Square basically indicates the proportion of variance in the result that can be explained using independent variables in the model. Usually this value lies in the range of 0 to 1, where value near to 1 indicates its a good model and if the value is very near to zero, it can be considered as not a perfect model. But R square value alone can not evaluate the model as there is a chance of over fitting hence when this is used with other metric model can be evaluated more thoroughly.

In Big-mart sales data-set when we applied 5 algorithms we could see that Decision tree, Random forest and XG Boost gave better result compared to Linear and Ridge regression as they had better RMSE,AAE value and better R squared value. SO looking at the result we can say that Random Forest had the best values compared to rest of the algorithms with RMSE value of 37.8 and AAE of 1.23 and r squared value with 0.69. Looking at the output graph in figure 11 for random forest

```python
import pandas as pd

# Create a list of model names
models = ['Linear Regression', 'Ridge Regression', 'Decision Tree', 'Random Forest', 'XGBoost']

# Create a dictionary of model metrics
metrics = {
    'RMSE': [rmse_lr, rmse_rd, rmse_dt, rmse_rf, rmse_xgb],
    'MAE': [mae_lr, mae_rd, mae_dt, mae_rf, mae_xgb],
    'AAE': [aae_lr, aae_rd, aae_dt, aae_rf, aae_xgb],
    'R-Squared': [r2_lr, r2_rd, r2_dt, r2_rf, r2_xgb]
}

# Create a pandas DataFrame from the dictionary of model metrics
df = pd.DataFrame(metrics, index=models)

# Print the DataFrame
print(df)
```

	RMSE	MAE	AAE	R-Squared
Linear Regression	45.340007	33.842267	1.671449	0.562692
Ridge Regression	52.954294	42.107755	2.079676	0.403478
Decision Tree	37.912948	25.296000	1.249354	0.694227
Random Forest	37.804578	25.101960	1.239771	0.695972
XGBoost	39.578981	26.202811	1.294141	0.666763

Fig. 10. Combined result for data-set1

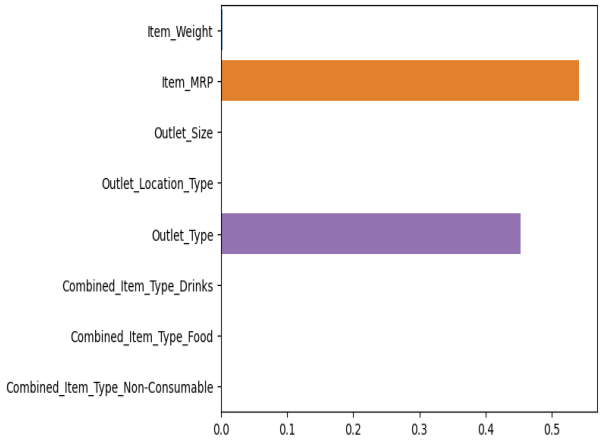

Fig. 11. Output Visualisation for Random Forest

we can say that item-MRP and item-outlet-type had a higher importance in the result.

In Wall-mart sales data-set looking at the results in figure 12, other than linear and ridge regression other 3 algorithms have performed very well with r square value is well over 80 for all of the three and among them XGboost has given the best result with RMSE value of 37, MAE of 19, AAE of 0.12 and R squared value of 0.93. Looking at the figure 13 which is a bar-graph plotted for output of XGboost model tells that sales has highest dependency on the Department and then store and size and other features don't matter much when it comes to contributing to the result.

In Super sales data-set we have applied all the 5 algorithms which we mentioned above and looking at figure 14 which has the results, comparatively result is best in Random forest with RMSE of 274 and AAE of 32 and MAE of 88 and r squared

```python
import pandas as pd

# Create a list of model names
models = ['Linear Regression', 'Ridge Regression', 'Decision Tree', 'Random Forest', 'XGBoost']

# Create a dictionary of model metrics
metrics = {
    'RMSE': [rmse_lr, rmse_rd, rmse_dt, rmse_rf, rmse_xgb],
    'MAE': [mae_lr, mae_rd, mae_dt, mae_rf, mae_xgb],
    'AAE': [aae_lr, aae_rd, aae_dt, aae_rf, aae_xgb],
    'R-Squared': [r2_lr, r2_rd, r2_dt, r2_rf, r2_xgb]
}

# Create a pandas DataFrame from the dictionary of model metrics
df = pd.DataFrame(metrics, index=models)

# Print the DataFrame
print(df)
```

	RMSE	MAE	AAE	R-Squared
Linear Regression	135.841507	91.177358	0.569781	0.088684
Ridge Regression	135.841506	91.177343	0.569781	0.088684
Decision Tree	53.929076	27.691076	0.173046	0.856368
Random Forest	50.109870	25.716562	0.160707	0.875992
XGBoost	37.216707	19.887033	0.124277	0.931596

Fig. 12. Combined result for data-set2

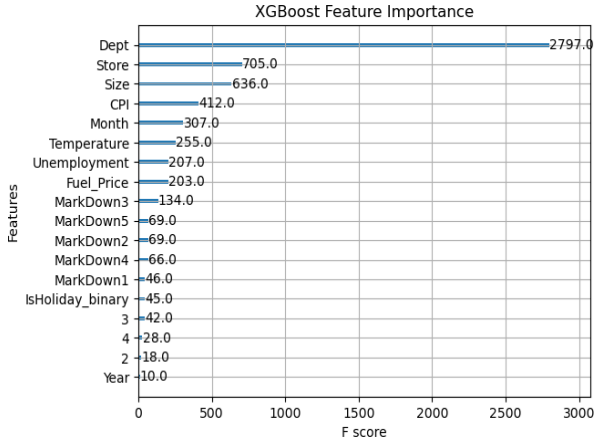

Fig. 13. Output Visualisation for XGBoost

value of 0.17. But with whatever domain knowledge we have these are not the best results by any means so may be some other algorithms like Time series could be applied here to get a better result.

F. Deployment:

This models can be implemented in the real world to help the business to predict the sales in advance so that it will help the organisations to take major decisions and help to grow the business.

IV. Conclusion

When we started the project our goal was to find 3 related suitable data-sets from the same domain and choose 5 algorithms that may be suitable for those data-sets and find the best performing algorithm for each data-set. So we chose Big-mart sales, Wall-mart Sales and Super store sales as our data-set and

```python
import pandas as pd

# Create a list of model names
models = ['Linear Regression', 'Ridge Regression', 'Decision Tree', 'Random Forest', 'XGBoost']

# Create a dictionary of model metrics
metrics = {
    'RMSE': [rmse_lr, rmse_rd, rmse_dt, rmse_rf, rmse_xgb],
    'MAE': [mae_lr, mae_rd, mae_dt, mae_rf, mae_xgb],
    'AAE': [aae_lr, aae_rd, aae_dt, aae_rf, aae_xgb],
    'R-Squared': [r2_lr, r2_rd, r2_dt, r2_rf, r2_xgb]
}

# Create a pandas DataFrame from the dictionary of model metrics
df = pd.DataFrame(metrics, index=models)

# Print the DataFrame
print(df)
```

	RMSE	MAE	AAE	R-Squared
Linear Regression	274.977632	88.704004	32.920059	0.178659
Ridge Regression	287.660069	91.208369	33.849485	0.101148
Decision Tree	289.118118	91.329649	33.894495	0.092013
Random Forest	274.977632	88.704004	32.920059	0.178659
XGBoost	280.060066	89.556577	33.236468	0.148016

Fig. 14. Combined result for data-set3

after understanding the data and referring to different research papers we decided to go ahead with Linear Regression, Ridge Regression, Decision tree, Random Forest and XGBoost and we chose RMSE,AAE,MAE and R squared values as our evaluation metrics by our domain knowledge. So after doing the data engineering and model building when we evaluated, for Big-mart sales data-set, Random forest gave the best result and in Wall-mart data-set, XGBoost gave the best result and in Super store sales dataset,Random forest had the best result compared to other algorithms but over all the result was not that satisfactory as the RMSE, AAE values were on higher side and R square value was near to 0. so may be if we try with some other algorithms related to time series it may give a better result which can be done as a future work. Overall we have achieved our goal what we had planned.

References

[1] Comparative Analysis of Regression Algorithms used to Predict the Sales of Big Marts by M.Saad Bin Ilyas, Atif Ikram, Muhammad Aadil Butt and Iqra Tariq published on 2023-03-02 http://jicet.org/index.php/JICET/article/view/53

[2] Sales prediction and analysis of supermarkets using ridge and polynomial regression techniques by Arun B Prasad, Ram Bhawan Singh, Manish Garg in 2023 https://businessmanagementeconomics.org/pdf/2023/512.pdf

[3] Sales prediction for big mart outlets by J.Jyothi, U.Chandana Madhuri, B.Ome Sai Srija, J.Sravani, P.Mounica in 2021 https://www.journal-dogorangsang.in/no_2_Book_21/23.pdf

[4] Analyzing big market sales using machine learning algorithms by V.Ramyasri, SK.Manisha,V.Padma Priya,B. Mery Prasanna,Rajesh Yamparala in 2021. https://www.journal-dogorangsang.in/no_2_Book_21/14.pdf

[5] Mega Mart Sales Prediction Using Machine Learning Techniques by Gopal Gupta, Kanchan Lata Gupta and Gaurav Kansal on 03 July 2022. https://link.springer.com/chapter/10.1007/978-981-19-1142-2_35

[6] Data visualization and multiple linear regressions for bigmart sales prediction by Lakshmi Karanam in December 2022 http://junikhyatjournal.in/no_2_Online_22/12_dec.pdf

[7] Sales prediction using machine learning algorithms by Purvika Bajaj, Renesa Ray, Shivani Shedge, Shravani Vidhate, Prof. Dr. Nikhilkumar Shardoor on 06 — June 2020 https://docplayer.net/188314966-Sales-prediction-using-machine learning-algorithms.html
[8] Fusing Clustering and Machine Learning Techniques for Big-Mart Sales Predication by S. N. Gunjal,D. B. Kshirsagar,B.J Dange,H.E Khodke in 2022 https://ieeexplore.ieee.org/abstract/document/9935906
[9] Sales Prediction Using ARIMA, Facebook's Prophet and XGBoost Model of Machine Learning by Sushila Ratre & Jyotsna Jayaraj on 01 January 2023 https://link.springer.com/chapter/10.1007/978-981-19-5868-7_9
[10] Walmart Sales Prediction Based on Decision Tree, Random Forest, and K Neighbors Regressor by Bo Yao in 2023 https://drpress.org/ojs/index.php/HBEM/article/view/5100
[11] Experimental Performance Analysis of Machine Learning Algorithms by Ganesh Khekare, Anil V. Turukmane, Chetan Dhule, Pooja Sharma and Lokesh Kumar Bramhane on 13 July 2022 https://link.springer.com/chapter/10.1007/978-981-19-2456-_104
[12] Machine Learning based Food Sales Prediction using Random Forest Regression by Hruthvik Naik,Kakumanu Yashwanth,Suraj P,N. Jayapandian in 2022 https://ieeexplore.ieee.org/abstract/document/10009277
[13] Predictive Analysis of Retail Sales Forecasting using Machine Learning Techniques Muhammad Sajawal1 , Sardar Usman2 , Hamed Sanad Alshaikh3 , Asad Hayat4 and M. Usman Ashraf5 in 2023 Predictive-Analysis-of-Retail-Sales-Forecasting-using-Machine-Learning-Techniques.pdf (researchgate.net)
[14] BigMart Product Sales Factors by devastator: https://www.kaggle.com/datasets/thedevastator/bigmart-product-sales-factors
[15] Walmart Sales https://www.kaggle.com/datasets/ujjwalchowdhury/walmartcleaned
[16] Superstore Sales Dataset by Rohith: https://www.kaggle.com/datasets/rohitsahoo/sales-forecasting
[17] Comparative Analysis of Sales Prediction using different Machine Learning Techniques by me file://C:/Users/gogat/OneDrive/Desktop/Project_Proposal
[18] CRISPDM image https://www.datascience-pm.com/crisp-dm-2/