Int. J. Nonlinear Anal. Appl.
Volume 12, Special Issue, Winter and Spring 2021, 1611-1630
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2021.5848

Predicting periodical sales of products using a machine learning algorithm

A. Bhuvaneswari[a], T.A. Venetia[a,]

[a]Department of Computer Applications PSG College of Technology Coimbatore, India

(Communicated by Madjid Eshaghi Gordji)

Abstract

Today, online shopping has evolved as a prominent business and there are very few opportunities for vendors to improve their sales. A machine learning algorithm can be used to predict what should be sold in a particular month so that sales can be increased. Once the Prediction is done a dashboard will be created to display which products should have been offered to have high sales. Billing the sales and analyzing with help of an expert is done. But in this case, not all people have the resources to get help from the experts. Vendors rely on their experiences. People who have started businesses for a few years lack experience and need support. To Help the vendors in improving their business a prediction of sales is done for each month and a dashboard will display the items to be sold in a particular month for an offer. To do Prediction Machine Learning Algorithms Random Forest Algorithm is used. This Algorithm is the best algorithm to do prediction and it is based on decision trees. The Scope of this project is developing the random forest model for predicting the sales of the products in each month from the year January 2013 to October 2015.

Keywords: E-commerce, Machine learning, Artificial intelligence, Online advertising, Random forest algorithm

1. Introduction

In this era, e-commerce is becoming prominent in all business sectors as well as part of life. Though the products are intangible, customers tend to shop online because shopping becomes easy and there are many products available on discounts and require less time for purchase. There are many vendors who are facing difficult times in selling their products because of e-commerce sales; due to this many vendors had to stop their business. Business vendors who have experience in how to

Email addresses: abh.mca@psgtech.ac.in (A. Bhuvaneswari), venetiaadbe@gmail.com (T.A. Venetia)

Received: August 2021 *Accepted:* Novamber 2021

market the products, increase sales, when to give discounts and on which products to invest, survive during these hard times. Some vendors approach business experts, who know all about business, for advice on how to improve sales, what products to purchase and when to give discounts. Experts specialized in these areas are in great demand; however, not all vendors can afford to hire such an expert. Vendors who are new to the business will also face this problem. This project is focused on solving this problem, using machine learning algorithms for predicting the market trends and increasing sales.

A. Challenges in sales of products

E-commerce means doing business with the help of the internet and buying products from anywhere in the world. People have started to often buy daily essentials online, especially during this COVID-19 period, while staying at home. Many businesses try to sell their products through e-commerce websites with help from the business management experts and information technology (IT) solutions. There are many IT solution providers working on e-commerce technologies to support their clients on how to sell their products and develop e-commerce websites. Emerging technologies enable online customers to customize the required product like selecting clothes by virtual trials, designing interiors of a room, and booking tickets through virtual tours. These features make customers more happy and excited in online purchases. While designing a website for e-commerce, the designer must consider what features must be included in the website, see to that the website is user-friendly, how to display the products and so on. The major challenge in e-commerce is that all products are fashionable, updated and sold at less cost when compared with local sales. There are big teams who work just to sell and buy the products. They find new strategies to be implemented and all the strategies will be developed,tested and will be launched in the market.

B. Machine learning algorithms

Machine Learning(ML) is the subset of Artificial Intelligence (AI) and has the ability to learn from datas and test the knowledge. ML aims to make the computer learn and with its knowledge complex problems will be solved without human intervention or actions. There are four main category which is used in Machine Learning

- Supervised learning
- Unsupervised learning
- Semi-Supervised learning
- Reinforcement Learning

Each of these categories in machine learning can be used to build many applications which will solve complex problems. ML Algorithms will help the client to predict what should be sold in the upcoming month so that the client will be able to stock the products in advance and plan how to sell. There are different types of machine learning algorithms which is based on their category and they are

- Regression algorithms
- Decision tree algorithms
- Bayesian algorithms

- Clustering algorithms
- Artificial neural network algorithms
- Deep learning algorithms
- Predictive algorithms

There are so many where machine learning algorithms can be used to build models and some of the applications are

- Recommender systems
- Natural language understanding
- Online advertising
- Handwriting recognition
- Economics
- DNA sequence classification
- Fraud detection
- Bioinformatics
- Time series forecasting

Machine learning has helped many industry complex processes and we can see at least one Machine Learning application in our daily life. Algorithms learn from data and can predict what the future will be like. These kinds of features have helped business people to know their sales values. And with the help of machine learning algorithms there is great understanding between the customer and the business owners. Machine Learning started evolving way back in the 70's and 60's but people were not sure or they neither cared about it. But now without machine learning complex works cannot be reduced.In this project machine learning algorithms will be used to do prediction of sales.

C. AI for sales

Artificial Intelligence(AI) techniques will help the business to operate and communicate with the customer or client smartly. Because AI, companies were able to produce more profit and reduce the expenses where it is needed by replacing them with AI technology. By extracting data, analysts can analyze the data, come up with marketing strategies and do decision making.

AI techniques are used in forecasting the future sales. Business owners were able to see and analyze their products sales in each month and view the future sales. Because of this feature , marketing teams can identify where the problem is when there is loss and plan different marketing strategies to improve the sales. AI will help in making prediction, product recommendations while analyzing the dataset. There are many advanced AI techniques which can increase the sales profit and scale the business growth. Some of the AI techniques for sales are

- Predictive forecasting
- Price optimization

- Upselling and cross-selling
- Performance management

AI can be used for Business to Consumer and Business to Business. There is so much research which has been going on related to AI. There is no need to rely on any software where human work to input the data instead AI can help and improve the process much better.

D. Need for prediction

Prediction allows us to make decisions which is a very important part in managing a business. The accuracy rate of the prediction must be high so that correct decisions could be taken and if the accuracy is low then there are chances to affect the business when it is implemented. To do prediction there are so many algorithms in Machine Learning and they are

- Random forest
- Generalized linear model(GLM) for two values
- Gradient linear model(GBM)
- K-Means
- Prophet

In E-commerce with the help of prediction each and every product in shopping websites are able to predict which product is going to sell high. There are many factors that need to be considered while implementing prediction, the factors can be climate, temperature, holidays, festivals, discounts and so on. So when a prediction of sales is to be implemented then all the factors must be taken into account. At present all the e-commerce websites make predictions to know about the sales and the future.

the e-commerce websites make predictions to know about the sales and the future.

2. Literature survey

In this session, system study will be elaborated. Predicting algorithms are used in this project. Before learning about the system the basic fundamental id the decision tree which will be used in this project. Decision trees can be used as classification and regression. One of the decision trees is Classification and Regression Tree(CART). The representation of CART is Binary Tree and it can be stored to file as a graph or a set of rules. Each root node represents a single input variable (x) and a split point on that variable. The leaf nodes of the tree contain an output variable (y) which is used to make a prediction. An example (fig 1)of CART is classification of Male and Female based on height and weight.

Creating a binary decision tree is actually a process of dividing up the input space. A greedy approach is used to divide the space called recursive binary splitting. The most common stopping procedure is to use a minimum count on the number of training instances assigned to each leaf node. If the count is less than some minimum then the split is not accepted and the node is taken as a final leaf node. The complexity of a decision tree is defined as the number of splits in the tree. Simpler trees are preferred. Variable importance in a CART model helps in judging the influence of a variable in a model. It is calculated by summing the split improvement score for each variable across all splits in a tree. CART is the best model for small dataset and there are chances for overfitting of data when large dataset are used for training.

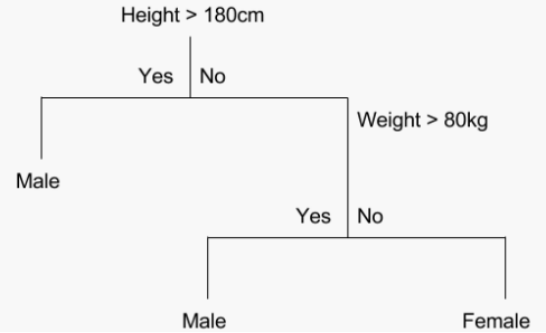

Figure 1: Single decision tree

A. Existing system

Predicting the future sales of the products are available in e-commerce websites and it is also an important feature among others. Because e-commerce business vendors who run locally are affected. The sales of their products are in loss and at present they are trying to sell the products by using different strategies. They started to sell their products by offering "Free Home Delivery" and this will be displayed in the advertisement. "Free Home Delivery" is being used because customers are expecting to buy products from home and this feature is available on e-commerce websites. But not all shop owners are able to sell the products in e-commerce instead the shops are closed forever.

While building an e-commerce website managers will try to analyze how the business is going on the website. They try to analyze what kind of improvements need to be made. Data Analyst and data scientist work on predicting the sales because of this developers will be able to develop or modify their portal to make users shop more. Since many business vendors are affected in this e-commerce evolution , this project will help the vendors in predicting the sales of the products.

B. Proposed system

To help the business vendors prediction of sale of products will be done.In this one of the prediction algorithms will be used to predict the sales and that product will be recommended to the business owner. For Example in the month January if the dataset has high sales of TV in the month January then the application must recommend to the owner that he/she must have more stock of TV products in January because high sales that have been calculated in historical data. This feature can also be used by owners to plan which product needs to be offered so that there is high sales of the products.

This can be done with the help of Random Forest Algorithm. In the introduction part this session has explained about how classification and regression trees works. Random Forest algorithm is the building block of decision trees and a single decision tree is equal to one CART decision tree and a random forest algorithm will be explained in further sessions. So for prediction RandomForest algorithm will be used.

3. Proposed system design

Systems design is the process of defining the architecture, modules, interfaces, and data for a system to satisfy specified requirements. Systems design could be seen as the application of systems theory to product development. System design is the phase that bridges the gap between problem domain and the existing system in a manageable way. Logical design pertains to an abstract representation of the data flow, inputs, and outputs of the system. It describes the inputs , outputs, databases, procedures all in a format that meets the user requirements. While preparing the logical

design of a system, the system analyst specifies the user needs at a level of detail that virtually determines the information flow into and out of the system and the required data sources. Data flow diagram, E-R diagram modeling are used.

A. System architecture

A system architecture is the conceptual model that defines the structure, behavior, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system. A system architecture can consist of system components and the sub-systems developed, that will work together to implement the overall system.

A representation of a system, including a mapping of functionality onto hardware and software components, a mapping of the software architecture onto the hardware architecture, and human interaction with these components.An architecture consists of the most important, pervasive, top-level, strategic inventions, decisions, and their associated rationales about the overall structure and associated characteristics and behavior. System architecture conveys the informational content of the elements consisting of a system, the relationships among those elements, and the rules governing those relationships. The architectural components and set of relationships between these components that an architecture description may consist of hardware, software, documentation, facilities, manual procedures, or roles played by organizations or people.A system architecture primarily concentrates on the internal interfaces among the system's components or subsystems, and on the interfaces between the system and its external environment, especially the user.

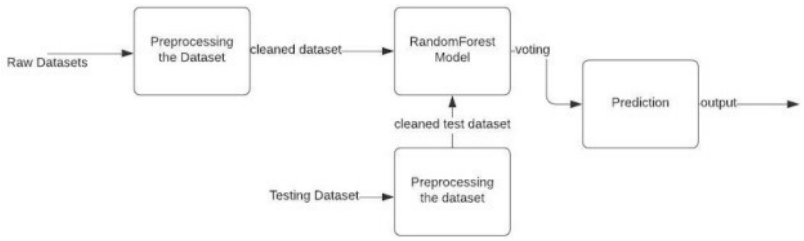

Figure 2: System architecture for predicting the sales of the Products

In this project fig 2 explains the architecture of the whole system. There are totally three stages in this architecture and they are

- Preprocessing the dataset
- Random forest model
- Prediction

The data will be preprocessed by removing unwanted data from the dataset . There can be merging two datasets for predicting which is explained in Development session. The next stage is the Random forest model where the output of the preproceed stage will be taken as input in the random forest model. Then the test dataset will be sent to the preprocessing stage where the data are preprocessed and at the final stage the prediction takes place.

The training dataset is the sales of the products for each day which will be predicted. The prediction is for finding which item has been sold at the highest according to the historical data. From this prediction model , to find the month which the item needs to be sold is found out.

Fig 3 is an architecture diagram for random forest models. In the random forest model there are totally two stages and there are Bootstrapping, decision trees. A detailed study for Random Forest is given in Development session. The dataset will be passed onto the bootstrapping process where n number of decision trees will be built. Decision trees are useful for classification and regression problems.

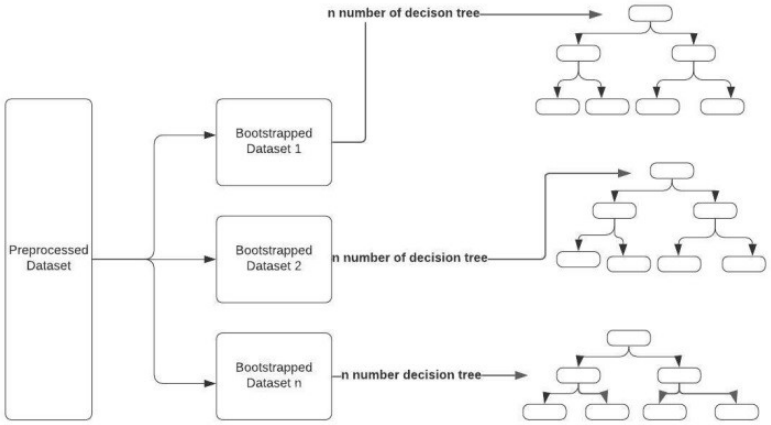

Figure 3: Architecture of random forest model

These are the architecture design of the system and the flow of the system is explained below.

B. System flow

System flowcharts are a way of displaying how data flows in a system and how decisions are made to control events. Flow diagram is a collective term for a diagram representing a flow or set of dynamic relationships in a system.A flowchart is a graphical depiction of a sequence of activities in a process.A variety of standard symbols are used system flowcharts, with the shape of the symbol indicating its function. The symbols are connected by lines that show the direction of flow.System flowcharts use the diamond symbol to represent yes/no decisions, with a separate line leaving the diamond for each response.Programmers need to thoroughly understand a task before beginning to code. System flowcharts were heavily used in the early days of programming to help system designers visualize all the decisions that needed to be addressed. Other tools have since been introduced that may be more appropriate for describing complex systems. Systems flowcharts have several symbols used to represent different media, such as disks, tapes, documents, terminal inputs and terminal outputs. Lines with arrowheads indicate the flow of data. The flow of data generally goes from top to bottom and left to right and depicts the sequence of processing steps along these data flow lines.

In fig 4 the flow diagram for this project is given. Initially the preprocessing stage will be implemented. Then the cleaned data will be sent for generating random samples. After that the random forest will generate a bootstrapped dataset which is formed from random samples. From a single bootstrapped dataset a multiple decision tree can be formed. The testing dataset will be passed to the model where the prediction takes place.

C. Data flow

A data-flow diagram is a way of representing a flow of data through a process or a system (usually an information system).The DFD also provides information about the outputs and inputs of each entity and the process itself. A data-flow diagram has no control flow, there are no decision rules and

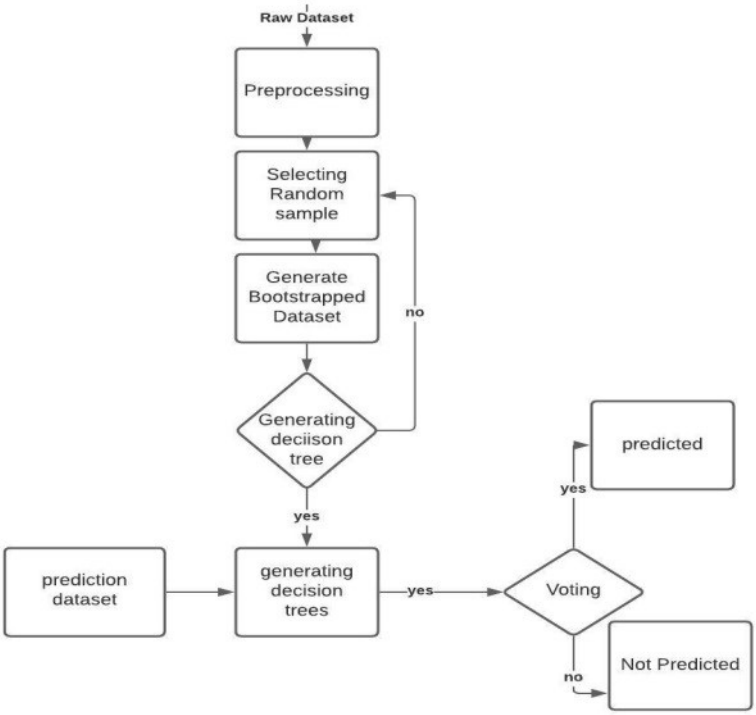

Figure 4: System flow chart

no loops. Specific operations based on the data can be represented by a flowchart. The data-flow diagram is part of the structured-analysis modeling tools. When using UML, the activity diagram typically takes over the role of the data-flow diagram. A special form of data-flow plan is a site-oriented data-flow plan. DFD consists of processes, flows, warehouses, and terminators. There are several ways to view these DFD components.

In this session there are two levels DFD level 0 in the fig 5 and DFD level 1 in the fig 6. In fig 5 the input is user input given to the random forest model . Once the Random Forest model does prediction the output will be displayed. In this project the input will be the item name which is passed to the Random Forest Model and the output is the month number which means that the item name which is passed as an input must be sold in the month that is output because there was high sales in the products according historical data.

In DFD level 1 the detailed process of how the dataset is processed into a Random Forest Model is given in the fig 5 .Initially the raw dataset is sent as input to the preprocessing function. In preprocessing all the unwanted datas will be removed which is explained in detail in Development session. The whole set of Dataset is separated into 70% Training dataset and 30% Testing Dataset. In the training dataset, random samples will be selected and bootstrapped dataset will be generated. There can be n number of bootstrapped dataset because of this feature only random forest models can work really well. The bootstrapped dataset will be converted into multiple decision trees which is combined together to form a Random Forest Model. The testing dataset will be passed as input to the model to check whether classification or regression is taking place correctly. Once the testing dataset has been predicted this model can be used to predict the real time input.

This is the flow of the data from preprocessing to prediction process. The type of dataset used is explained below.

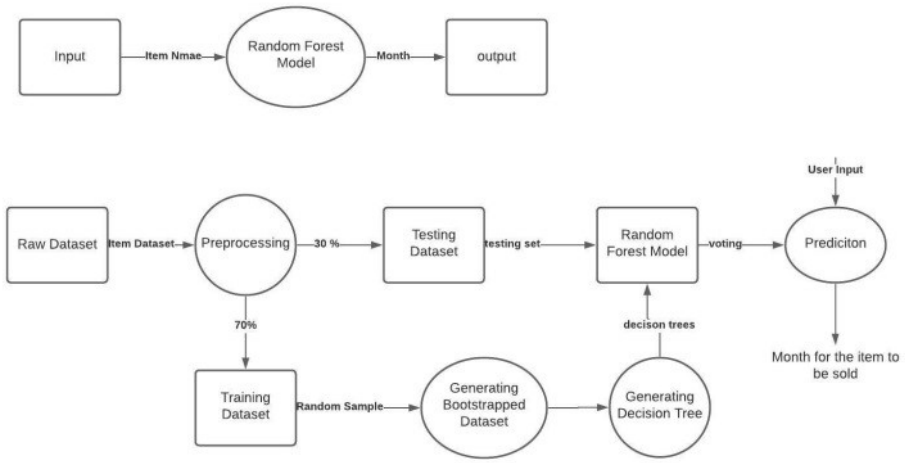

Figure 5: DFD level 0 for predicting sales of the products

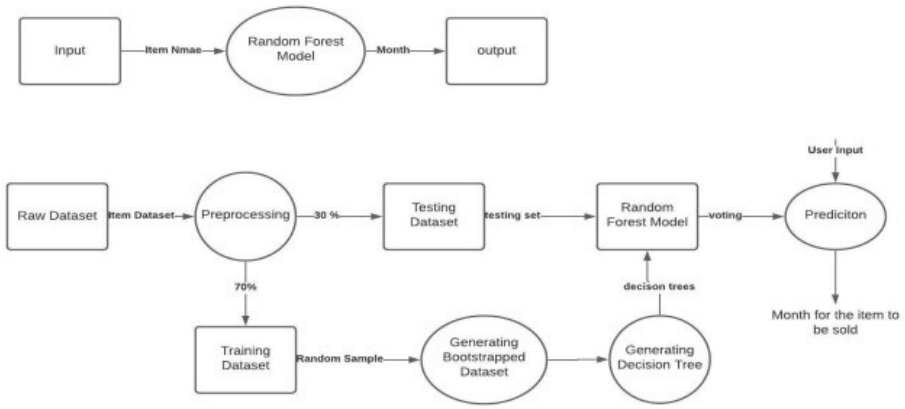

Figure 6: DFD level 1 predicting sales of the products

4. System development

In this project Machine Learning Algorithms have been used to do prediction and the algorithm is Random Forest Algorithm. The system design and the flow of the project is explained in system study session. In this session the implementation of the system design is done. Each and every process in the system design is explained and implemented here.

A. Data preprocessing module

In any Machine Learning or AI the first step will be preprocessing the dataset. Each and every algorithm accepts input in different formats, it is the responsibility of the developer to convert the raw data into a standardized data. When the dataset is not preprocessed or when they are not preprocessed correctly then there are high chances for the model to work correctly.

When there are multiple attributes in a dataset then data analysts must check which attributes are needed for the model. This can be done by visualizing the data in graphical model or bar plot. This will help in optimizing the machine learning algorithm. To do data preprocessing there are techniques which can be implemented and some of the techniques are

- Data quality assessment

- Feature aggregation
- Feature sampling
- Dimensionality reduction
- Feature encoding

In this project while analysing all the datasets it found that there are no missing or abnormal data. It means that half of the work in data preprocessing need not to be done. Removal of unwanted data must removed in the dataset so that algorithm will be able to learn efficiently.

B. Random forest algorithm

Random Forest inherits properties of CART-like variable selection, missing values and outlier handling, nonlinear relationships, and variable interaction detection. It creates multiple CART trees based on "bootstrapped" samples of data and then combines the predictions. The combination is an average of all the predictions from all CART models. A bootstrap sample is a random sample conducted with replacement. Random Forest has better predictive power and accuracy than a single CART model.

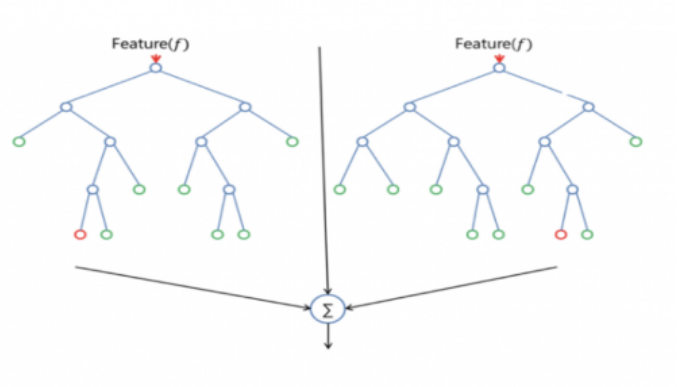

Figure 7: Random forest diagram

C. Advantages of random forest

- It can come out with very high dimensional (features) data, and no need to reduce dimension, no need to make feature selection
- It can judge the importance of the feature
- Can judge the interaction between different features
- Not easy to overfit
- Training speed is faster, easy to make parallel method
- It is relatively simple to implement
- For unbalanced data sets, it balances the error.
- If a large part of the features are lost, accuracy can still be maintained.

Random forests are used in Classification of discrete values, Regression of continuous values, Unsupervised learning clustering, Abnormal point detection. So by implementing Random Forest will be helpful in predicting the sales. It is best when compared with CART because CART will be predicting and taking decisions based on one decision tree. But in random forest the prediction can be done very accurately because the prediction result will be based on comparing multiple decision tree results. This is a very important reason why random forest will be used in this project to do prediction. From the original dataset a bootstrapped dataset will be created.Bootstrapped dataset are the datasets which are formed by selecting random datas from the original dataset. The size of the bootstrapped dataset can be defined by the developer or can be the same size of the original dataset.This means that there will be a repeated number of datas because the same sample/data can be selected multiple times. Below the example(fig 8) of heart disease(not part of the project) is used to explain the Random Forest Algorithm.

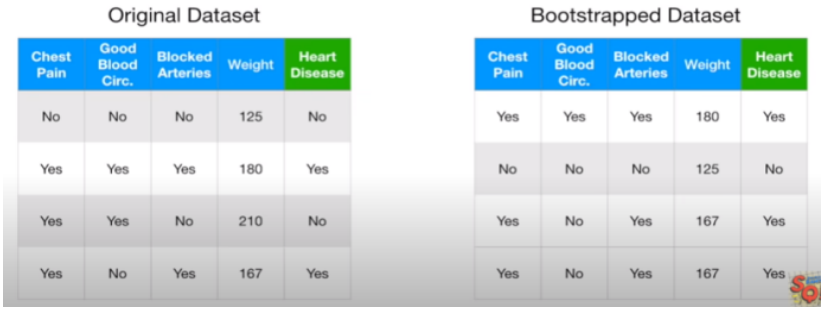

Figure 8: Example of random forest

Now from the created bootstrapped dataset multiple decision trees will be created. The decision trees will be created based on the number of variables a dataset has. The root node of a decision tree will vary many times. The accuracy of the random forest model is based on the root node. Therefore many numbers of bootstrapped dataset will be created and from that n number of decision trees will be created. Because of this feature it is known as random forest and the classification and regression accuracy is much accurate. A Multiple decision tree is built and the structure will look like fig 9.

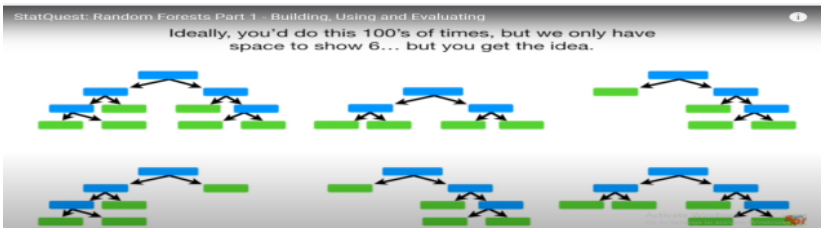

Figure 9: Multiple decision tree

After creating the decision trees it's time to do prediction. In the Random Forest algorithm, a single sample from the testing dataset will be passed to all the decision trees. The Label which has the highest vote will be the class for that sample. So it means that a single sample will be passed onto all the decision trees. This bootstrapping of a dataset can also be called "BAGGING" . After the prediction happens,the testing set will be passed to the Random Forest model and the input will be passed to all trees in fig 10. In this example the heart disease class will be predicted.

One thing to remember is that in a bootstrapped dataset a lot of sample data is being reused multiple times. This can lead in false prediction and there are chances that 1/3rd part of data are

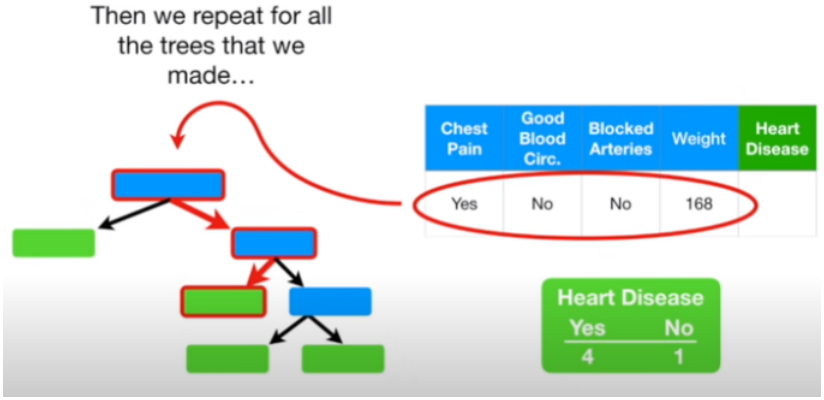

Figure 10: Test set passed to the model

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	No	168	YES

Figure 11: Test set passed to the model

not trained. To avoid this "OUT-OF-BAG" is done.That means that the data which have not been trained will be taken into account and a seperate bootstrapped dataset will be formed. Then decision trees will be created,the same sample will be passed to this decision tree for prediction. The example for out-of-bag is displayed in fig 12 and the output is displayed in fig 13.

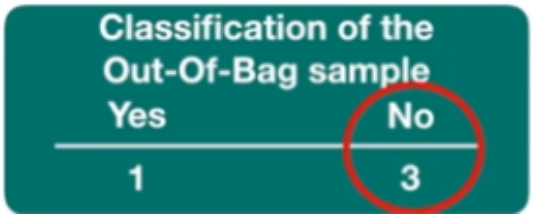

Figure 12: Out-of-bag high frequency class label

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	Yes	No	210	No

Figure 13: Predicting the class label with out-of-bag

The implementation of the Random Forest algorithm is explained in further session.

D. MapReduce

While Random Forest is implemented for large dataset the process time is increased even though the algorithm works well with large dataset. And for this project since the aim is to find the sales

of the product in a month then we can separate the data in month wise and reduce the size of the dataset. This will help in improving the performance of the algorithm and the process time for the random forest algorithm is decreased because the dataset has been sorted.

MapReduce is used in big data and is a method to map the data and also helps in reducing the size of large data.Reduce the big data and extract only meaningful data. There are two phases in MapReduce

- Map
- Reduce

Both the phases will be explained with a real time example; Amazon wants to calculate its total sales city wise for the year 2015 in India.

MAP: The data will be passed as input to map function. According to the month in the data will be distributed to 12 different map functions. And each month the mapping function will be processed accordingly. The output of the map function will be the intermediate key and value.

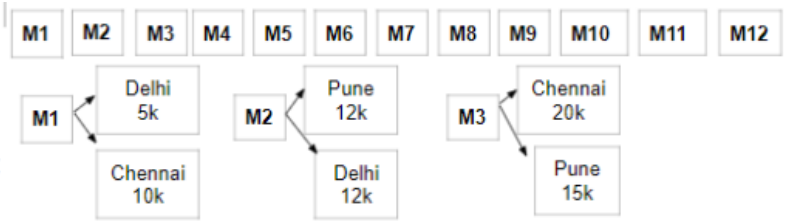

Figure 14: Amazon mapReduce-MAP

REDUCE: The reduce function will be implemented where the similar ones will be merged so that size will be reduced and reduce function uses some techniques to merge them.In amazon problem, the reduce function will be processed based on the north , south,east and west direction.

Figure 15: Amazon MapReduce-REDUCE

The output will be cities with a total number of sales.

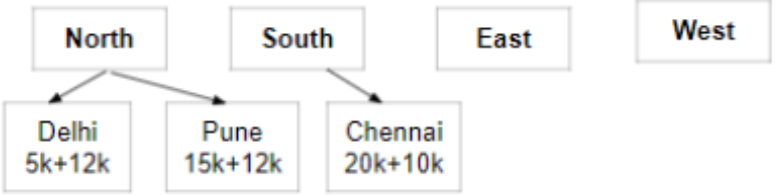

Figure 16: Amazon MapReduce-OUTPUT

The design when MapReduce Algorithm is implemented will be as fig 17.

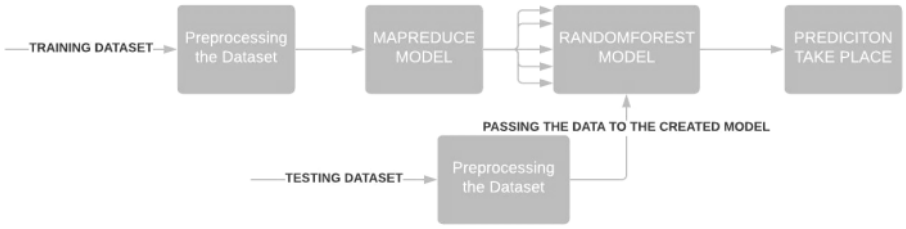

Figure 17: System design with MapReduce

When Mapreduce was implemented with Random Forest there was not much difference between the model which has only Random Forest and the model which has both MapReduce and Random Forest. So in this project implementation of MapReduce will not be implemented. While analyzing the dataset for the model only with the Random Forest algorithm and the dataset with both MapReduce and Random Forest algorithm is the same. Even though MapReduce helps in reducing the size of the dataset, analyzing whether the dataset is correct for the model or not is very important. MapReduce is one of the best algorithms used in many applications to reduce the task and size of the data.

5. Implementation and testing

In Random Forest there are two ways to implement one is classification model and other is regression model. In this project the available dataset can be applied for both the models and that is the reason for implementing both the models but there is much difference in the accuracy level even when both are implemented.

A. System implementation

The first model will be Random Forest Classification. Its accuracy rate is displayed in fig 20. In fig 18 from sklearn model train_test_split class is imported for splitting the dataset. The X array will hold the features from the cleaned sales_train dataset and y will hold class labels of sales_train dataset. Then the dataset will be splitted into 70% training dataset and 30% testing dataset.

```
[ ]  from sklearn.model_selection import train_test_split
     X=sales_train_dataset[['item_id','item_cnt_day']]
     y=sales_train_dataset['date_block_num']
     X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3)
```

Figure 18: Splitting the Sales_train dataset into training and testing dataset

In fig 19 from sklearn RandomForest classifier is imported. Initially the Random Forest model is fixed with parameters as n_estimators as 100 and n_jobs as 1.The splitted training dataset dataset is fitted into the random forest model with the help of fit function. The prediction of the testing dataset is done.

In fig 20 the accuracy of the model is checked. There are 5 types of accuracy checker and they are accuracy_score,log_loss,meansquared_error,root mean square.

```
[ ]  from sklearn.ensemble import RandomForestClassifier
     clf=RandomForestClassifier(n_estimators=200,n_jobs=1)
     clf.fit(X_train,y_train)
     y_pred=clf.predict(X_test)
```

Figure 19: RandomForest model

```
[ ]  from sklearn import metrics
     import numpy as np
     print("Accuracy:",metrics.accuracy_score(y_test,y_pred))
     clf_probs = clf.predict_proba(X_test)
     print("Log Loss Accuracy:",metrics.log_loss(y_test,clf_probs))
     print("Mean Absoulte Error:",metrics.mean_absolute_error(y_test,y_pred))
     print("Mean Square Error:",metrics.mean_squared_error(y_test,y_pred))
     print("Root Mean Squared Error",np.sqrt(metrics.mean_squared_error(y_test,y_pred)))
```

Figure 20: Accuracy checking for training dataset

The output for the accuracy and the prediction of the input is displayed in fig 21.

```
Accuracy: 0.22208900318476762
Log Loss Accuracy: 2.8471641813423303
Mean Absoulte Error: 3.320369455751032
Mean Square Error: 21.457769754358477
Root Mean Squared Error 4.63225320490563

[ ]  clf.predict([[5037,1]])

     array([8])
```

Figure 21: Accuracy and the prediction result.

After implementing the model it's important to check the weightage of the features so that the least feature will be removed from the dataset. By doing this the accuracy rate is increased drastically. This concept is shown in 22 and fig 23.

```
[ ]  import pandas as pd
     features_imp=pd.Series(clf.feature_importances_,index=['item_id','item_cnt_day']).sort_values(ascending=False)
     features_imp

     item_id        0.986021
     item_cnt_day   0.013979
     dtype: float64
```

Figure 22: Feature importance for sales_train dataset

The unimportant feature is removed from the model and the accuracy level is checked from this there is little improvement in the model.This is shown in fig 24.

The Second Model is RandomForest Regression Model. The initial step will be loading the random forest model with fixed estimators and fighting the dataset into the model and prediction model.This is shown in the fig 25.

In fig 26 the accuracy rate for the regression model is being checked and it varies a lot from classification accuracy rate.

```
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
sns.barplot(x=features_imp,y=features_imp.index)
plt.xlabel("Feature Importance Score")
plt.ylabel("Features")
plt.title("Visualizing Important Features")
plt.legend()
plt.show()
```

```
No handles with labels found to put in legend.
```

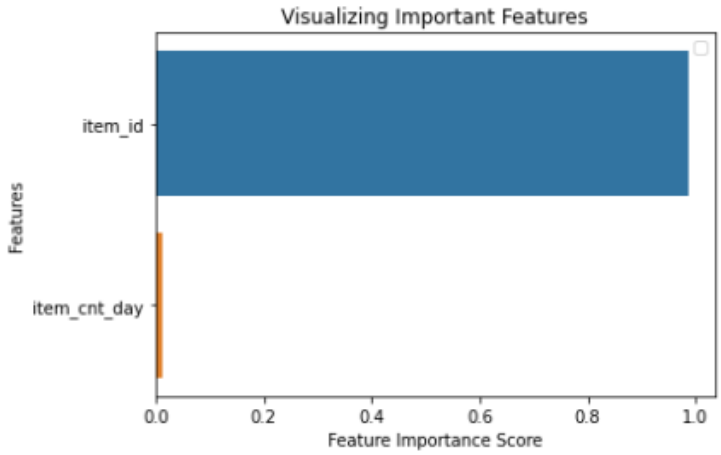

Figure 23: Visual representation of feature importance

```
from sklearn.model_selection import train_test_split
X=sales_train_dataset[['item_id']]
y=sales_train_dataset['date_block_num']
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3)
```

```
from sklearn.ensemble import RandomForestClassifier
clf=RandomForestClassifier(n_estimators=200,n_jobs=-1,bootstrap= True,max_features='sqrt',oob_score= True)
clf.fit(X_train,y_train)
y_pred=clf.predict(X_test)
clf_probs = clf.predict_proba(X_test)
```

```
print(y_pred)
print(clf_probs)
```

```
[5 4 9 ... 7 2 7]
[[0.07200408 0.15288735 0.13944466 ... 0.02103047 0.01357107 0.01148327]
 [0.07411069 0.0813782  0.04899425 ... 0.03227682 0.01562443 0.09678815]
 [0.1105944  0.08491032 0.08339358 ... 0.11346539 0.0874842  0.10334064]
 ...
 [0.08371797 0.08533128 0.06210677 ... 0.1519326  0.02014314 0.1482988 ]
 [0.05315765 0.03649476 0.24305821 ... 0.05076686 0.04238523 0.05111377]
 [0.03987031 0.04295338 0.03858344 ... 0.08495944 0.08309676 0.07182662]]
```

Figure 24: Removing unwanted features in sales_train dataset

B. Test cases

In this Random Forest regression model different item id's are passed to find out which month has the highest sale of that item and the model was able to predict the month. After prediction from the item id , the item category can be found from item_combine_item_cateogry dataset and this dataset was created for finding the item category.

The item id are randomly given based on the item dataset and the predicted result is given in fig 27.

```
[ ] from sklearn.model_selection import train_test_split
    X=sales_train_dataset[['item_id']]
    y=sales_train_dataset['date_block_num']
    X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3)

[ ] from sklearn.ensemble import RandomForestRegressor
    regressor=RandomForestRegressor(n_estimators=100,random_state=0)
    regressor.fit(X_train,y_train)
    y_pred=regressor.predict(X_test)

    print(y_pred)

    [3.51720948 6.97496594 5.30818942 ... 5.55131116 5.04237742 5.01900802]
```

Figure 25: Random forest regression model

```
[ ] from sklearn import metrics
    print("Mean Absoulte Error:",metrics.mean_absolute_error(y_test,y_pred))
    print("Mean Square Error:",metrics.mean_squared_error(y_test,y_pred))
    print("Root Mean Squared Error",np.sqrt(metrics.mean_squared_error(y_test,y_pred)))

    Mean Absoulte Error: 2.709937246851121
    Mean Square Error: 10.799108225974239
    Root Mean Squared Error 3.286199663132817

[ ]
```

Figure 26: Accuracy calculation in random forest regression model

```
[7] clf.predict([[5037]])

    array([8])

    clf.predict([[15757]])

    array([0])

    clf.predict([[7440]])

    array([6])
```

Figure 27: Prediction for month January, August,Map

The Month Numbering starts from 0-11 i.e 0 for January, 1- February and 12 for December. The items are sold on three different months and the items are sold based on the seasonal sales. In December items which are used indoors are sold more because of winter.

6. Result

Result will tell how much accurate the model has worked. By analyzing the result one can take decision and there are types to display the result and they are

- Quantitative Outcome
- Qualitative Outcome

A. Quantitative outcome

The Accuracy rate for bothe Random Forest Classification and Random Forest Regression was calculated and the comparison is displayed below in the table TABLE 1.

Table 1: Comparison of random forest classification and random forest regression accuracy

Accuracy type	Ramdom forest classification	Random forest regression
Mean Absolute Error	3.320369455751032	2.709937246851121
Mean Square Error	21.457769754358477	10.799108225974239
Root Mean Squared Error	4.63225320490563	3.286199663132817

B. Qualitative outcome

In Qualitative outcome the prediction of the random forest classifiers and the random forest regression is being displayed below. The random forest classifiers are displayed in fig 29 and the random forest regression is displayed in fig 31.

Random forest classifier

```
import numpy as np
import matplotlib.pyplot as plt

X_grid = np.arange(min(X_train['item_id']), max(X_train['item_id']),1000)
X_grid = X_grid.reshape((len(X_grid), 1))


plt.scatter(X, y, color = 'blue')

# plot predicted data
plt.plot(X_grid, clf.predict(X_grid),
         color = 'green')
plt.title('Random Forest Classifier')
plt.xlabel('item id')
plt.ylabel('month')
plt.show()
```

Figure 28: Plot coding for random forest classifier

7. Conclusion

Random Forest classifiers and regression have been implemented and the result shows that when the dataset is passed to the both models their accuracy and other metrics vary a lot. So it is important to know which model needs to be used and in a data analytics project the initial step must be identifying the correct dataset for the project and correct model to solve the problem. Random Forest Regression works best in this project even though there are only 12 class labels. With the help of regression ,in which month the item needs to be sold out is found out.

Now with the help of these features business vendors will have knowledge which products need to be sold without the help of experts and these features can be integrated with any kind of billing

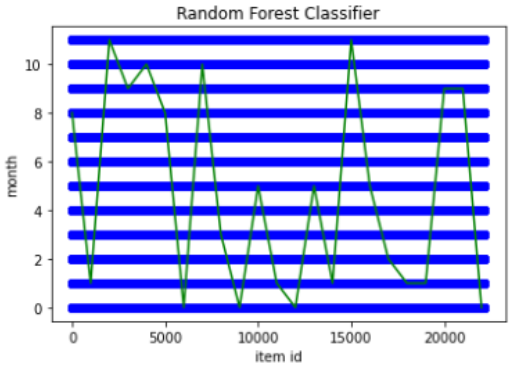

Figure 29: Plot result for random forest classifier

```
# Visualising the Random Forest Regression results
import numpy as np
import matplotlib.pyplot as plt

X_grid = np.arange(min(X_train['item_id']), max(X_train['item_id']),1000)
X_grid = X_grid.reshape((len(X_grid), 1))


plt.scatter(X, y, color = 'blue')

# plot predicted data
plt.plot(X_grid, regressor.predict(X_grid),
         color = 'green')
plt.title('Random Forest Regression')
plt.xlabel('item id')
plt.ylabel('month')
plt.show()
```

Figure 30: Plot coding for random forest regression

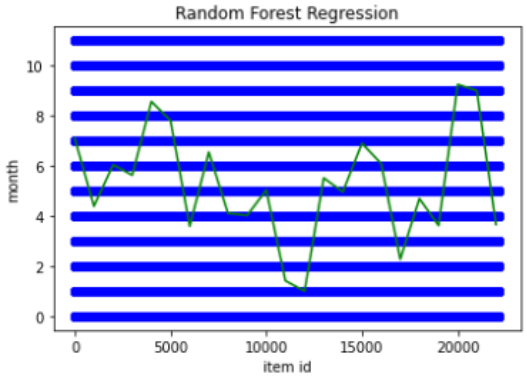

Figure 31: Plot result for random forest regression

application so that process will work much better. There are additional features which can be implemented like predicting the sales of products in each shop. Shop dataset was ignored in this project because the only aim was to find the sales of the products but with the help of shop dataset we will be able to solve different kinds of problems. The problems can be business owners will be able to identify which product has the lowest sale in a particular shop, or finding the products which have highest sales compared with all other branches. This feature can be extended to this project.

References

[1] T.F. Cootes, M.C. Ionita, C. Lindner and P. Sauer, *Robust and Accurate Shape Model Fitting Using Random Forest Regression Voting*, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2012.

[2] U. Groömping, *Variable importance assessment in regression: Linear regression versus random forest*, Amer. Statistic. 63 (2009) 308–319.

[3] J. Han, Y. Liu and X. Sun, *A scalable random forest algorithm based on MapReduce*, 2013 IEEE 4th Int. Conf. Software Engin. Serv. Sci. Beijing (2013) 849–852.

[4] Q. He, T. Shang, F. Zhuang and Zh. Shi, *Parallel extreme learning machine for regression based on MapReduce*, Neurocomput. 102 (2013) 52–58.

[5] V. Svetnik, A. Liaw, Ch. Tong, J. Christopher Culberson, R.P. Sheridan and B.P. Feuston, *Random forest: A classification and regression tools for compound classification and QSAR modeling*, J. Chem. Inf. Comput. Sci. 43(6) (2003) 1947–1958.

[6] W. Zhao, H. Ma and Q. He, *Parallel K-Means Clustering Based on MapReduce*, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2009.