MACHINE LEARNING ALGORITHIMS

Topic

Sales Prediction using Regression Analysis

Dataset

E – Commerce Sales Data

Sandeep Kannaujiya

Reg. No. – 11712528

Roll. No. - 25

Saurabh Shukla

Reg. No. – 11710440

Roll No. – 5

Submitted to:

Puja Rana

Work distribution

Data Pre-processing, Imputation, Label Encoder - Saurabh Shukla

Linear Regression, Predicting output, Backword Elimination – Sandeep Kannaujiya

ABSTRACT

In this paper, we study the usage of machine-learning models for sales predictive analytics. The main goal of this paper is to consider main approaches and case studies of using machine learning for sales forecasting. The effect of machine-learning generalization has been considered. This effect can be used to make sales predictions when there is a small amount of historical data for specific sales time series in the case when a new product or store is launched. A stacking approach for building regression ensemble of single models has been studied. The results show that using stacking techniques, we can improve the performance of predictive models for sales time series forecasting.

Intelligent Decision Analytical System requires integration of decision analysis and predictions. Most of the business organizations heavily depend on a knowledge base and demand prediction of sales trends. The accuracy in sales forecast provides a big impact in business. Data mining techniques are very effective tools in extracting hidden knowledge from an enormous dataset to enhance accuracy and efficiency of forecasting. The detailed study and analysis of comprehensible predictive models to improve future sales predictions are carried out in this research. Traditional forecast systems are difficult to deal with the big data and accuracy of sales forecasting. These issues could be overcome by using various data mining techniques. In this paper, we briefly analysed the concept of sales data and sales forecast. The various techniques and measures for sales predictions are described in the later part of the research work. On the basis of a performance evaluation, a best suited predictive model is suggested for the sales trend forecast. The results are summarized in terms of reliability and accuracy of efficient techniques taken for prediction and forecasting. The studies found that the best fit model is Gradient Boost Algorithm, which shows maximum accuracy in forecasting and future sales prediction.

Sales prediction is rather a regression problem than a time series problem.

Practice shows that the use of regression approaches can often give us better results compared to time series methods. Machine-learning algorithms make it possible to find patterns in the time series. We can find complicated patterns in the sales dynamics, using supervised machine-learning methods.

INTRODUCTION

One of the major objectives of this research work is to find out the reliable sales trend prediction mechanism which is implemented by using data mining techniques to achieve the best possible revenue. Today's business handles huge repository of data. The volume of data is expected to grow further in an exponential manner. The measures are mandatory in order to accommodate process speed of transaction and to enhance the expected growth in data volume and customer behaviour.

The E-commerce industry is badly in need of new data mining techniques and intelligent prediction model of sales trends with highest possible level of accuracy and reliability. Sales forecasting gives insight into how a company should manage its workforce, cash flow and resources. It is an important prerequisite for enterprise planning and decision making. It allows companies to plan their business strategies effectively.

Accurate predictions allow the organization to improve market growth with higher level of revenue generation. Data mining techniques are very effective in tuning huge volume of data into useful information for cost prediction and sales forecast, it is the basic of sound budgeting [1]. At the organizational level, forecasts of sales are essential inputs to many decision-making activities in various functional areas such as operations, marketing, sales, production and finance. In order to serve an organization's internal resources effectively, predictive sales data is important for businesses when looking for acquiring investment capital. The studies proceed with a new perspective that focuses on how to choose an appropriate approach to forecast sales with high degree of precision. Initial dataset considered in this research had a large number of entries, but the final dataset used for analysis having much smaller size compared to the original due to the riddance of non-usable data, redundant entries and irrelevant sales data.

The data mining techniques and predictions methods are discussed in Section I. The review of various literatures about sales forecasts are stated in Section II. In Section III, data tuning process and predictions are highlighted with visual representation of generated results. The predictive analytics and methodology on sales price also discussed. The performance evaluations of various prediction algorithms using machine learning approaches are stated. Finally, the result is analysed and concluded by summarizing the research findings and future scope.

We need to have historical data for a long time period to capture seasonality. However, often we do not have historical data for a target variable, for example in case when a new product is launched. At the same time, we have sales time series for a similar product and we can expect that our new product will have a similar sales pattern.

Sales data can have a lot of outliers and missing data. We must clean outliers and interpolate data before using a time series approach.

We need to take into account a lot of exogenous factors which have impact on sales.

PROPOSED METHODOLOGY:

Data Set:

This dataset consists of real-world Sales data of an E – Commerce store

ORDERNU	QUANTI	PRICEEACI	ORDERLIN	SALES	ORDERDA	STATUS	QTR_I	MONTH	YEAR_I	PRODU	MSRP	PRODU	CUSTOME	PHONE	ADDRES	ADDRE	CITY	STATE	POSTALCO	COUNTRY	TERRITOR	CONTACTI	CONTACTF	DEALSIZE
10107	30	95.7	2	2871	2/24/2003	Shipped	1	2	2003	Motorc	95	S10_16	Land of To	2.13E+09	897 Long Airpo		NYC	NY	10022	USA	NA	Yu	Kwai	Small
10121	34	81.35	5	2765.9	########	Shipped	2	5	2003	Motorc	95	S10_16	Reims Coll	26.47.155	59 rue de l'Abb		Reims		51100	France	EMEA	Henriot	Paul	Small
10134	41	94.74	2	3884.34	########	Shipped	3	7	2003	Motorc	95	S10_16	Lyon Souv	+33 1 46 6	27 rue du Colo		Paris		75508	France	EMEA	Da Cunha	Daniel	Medium
10145	45	83.26	6	3746.7	8/25/2003	Shipped	3	8	2003	Motorc	95	S10_16	Toys4Grov	6.27E+09	78934 Hillside I		Pasadena	CA	90003	USA	NA	Young	Julie	Medium
10159	49	100	14	5205.27	########	Shipped	4	10	2003	Motorc	95	S10_16	Corporate	6.51E+09	7734 Strong St.		San Franci	CA		USA	NA	Brown	Julie	Medium
10168	36	96.66	1	3479.76	10/28/200	Shipped	4	10	2003	Motorc	95	S10_16	Technics S	6.51E+09	9408 Furth Circ		Burlingam	CA	94217	USA	NA	Hirano	Juri	Medium
10180	29	86.13	9	2497.77	########	Shipped	4	11	2003	Motorc	95	S10_16	Daedalus I	20.16.155	184, chausse de		Lille		59000	France	EMEA	Rance	Martine	Small
10188	48	100	1	5512.32	11/18/200	Shipped	4	11	2003	Motorc	95	S10_16	Herkku Gif	+47 2267 3	Drammen 121,		Bergen		N 5804	Norway	EMEA	Oeztan	Veysel	Medium
10201	22	98.57	2	2168.54	########	Shipped	4	12	2003	Motorc	95	S10_16	Mini Whee	6.51E+09	5557 North Per		San Franci	CA		USA	NA	Murphy	Julie	Small
10211	41	100	14	4708.44	1/15/2004	Shipped	1	1	2004	Motorc	95	S10_16	Auto Cana	(1) 47.55.6	25, rue Lauristo		Paris		75016	France	EMEA	Perrier	Dominique	Medium
10223	37	100	1	3965.66	2/20/2004	Shipped	1	2	2004	Motorc	95	S10_16	Australian	03 9520 45	636 St K	Level 3	Melbourne	Victoria	3004	Australia	APAC	Ferguson	Peter	Medium
10237	23	100	7	2333.12	########	Shipped	2	4	2004	Motorc	95	S10_16	Vitachrom	2.13E+09	2678 Kir	Suite 1	NYC	NY	10022	USA	NA	Frick	Michael	Small
10251	28	100	2	3188.64	5/18/2004	Shipped	2	5	2004	Motorc	95	S10_16	Tekni Colle	2.02E+09	7476 Moss Rd.		Newark	NJ	94019	USA	NA	Brown	William	Medium
10263	34	100	2	3676.76	6/28/2004	Shipped	2	6	2004	Motorc	95	S10_16	Gift Depot	2.04E+09	25593 South Ba		Bridgewat	CT	97562	USA	NA	King	Julie	Medium
10275	45	92.83	1	4177.35	7/23/2004	Shipped	3	7	2004	Motorc	95	S10_16	La Rochell	40.67.855	67, rue des Cin		Nantes		44000	France	EMEA	Labrune	Janine	Medium
10285	36	100	6	4099.68	8/27/2004	Shipped	3	8	2004	Motorc	95	S10_16	Marta's Re	6.18E+09	39323 Spinnake		Cambridge	MA	51247	USA	NA	Hernandez	Marta	Medium
10299	23	100	9	2597.39	9/30/2004	Shipped	3	9	2004	Motorc	95	S10_16	Toys of Fir	90-224 85	Keskuskatu 45		Helsinki		21240	Finland	EMEA	Karttunen	Matti	Small
10309	41	100	5	4394.38	10/15/200	Shipped	4	10	2004	Motorc	95	S10_16	Baane Min	07-98 955	Erling Skakkes g		Stavern		4110	Norway	EMEA	Bergulfsen	Jonas	Medium
10318	46	94.74	1	4358.04	########	Shipped	4	11	2004	Motorc	95	S10_16	Diecast Cla	2.16E+09	7586 Pompton		Allentown	PA	70267	USA	NA	Yu	Kyung	Medium
10329	42	100	1	4396.14	11/15/200	Shipped	4	11	2004	Motorc	95	S10_16	Land of To	2.13E+09	897 Long Airpo		NYC	NY	10022	USA	NA	Yu	Kwai	Medium
10341	41	100	9	7737.93	11/24/200	Shipped	4	11	2004	Motorc	95	S10_16	Salzburg C	6562-9555	Geislweg 14		Salzburg		5020	Austria	EMEA	Pipps	Georg	Large
10361	20	72.55	13	1451	12/17/200	Shipped	4	12	2004	Motorc	95	S10_16	Souveniers	+61 2 9495	Monitor	Level 6	Chatswoo	NSW	2067	Australia	APAC	Huxley	Adrian	Small
10375	21	34.91	12	733.11	########	Shipped	1	2	2005	Motorc	95	S10_16	La Rochell	40.67.855	67, rue des Cin		Nantes		44000	France	EMEA	Labrune	Janine	Small
10388	42	76.36	4	3207.12	########	Shipped	1	3	2005	Motorc	95	S10_16	FunGiftIde	5.09E+09	1785 First Stree		New Bedf	MA	50553	USA	NA	Benitez	Violeta	Medium
10403	24	100	7	2434.56	########	Shipped	2	4	2005	Motorc	95	S10_16	UK Collect	(171) 555-	Berkeley Garde		Liverpool		WX1 6LT	UK	EMEA	Devon	Elizabeth	Small
10417	66	100	2	7516.08	5/13/2005	Disputed	2	5	2005	Motorc	95	S10_16	Euro Shop	(91) 555 94	C/ Moralzarzal,		Madrid		28034	Spain	EMEA	Freyre	Diego	Large
10103	26	100	11	5404.62	1/29/2003	Shipped	1	1	2003	Classic	214	S10_19	Baane Min	07-98 955	Erling Skakkes g		Stavern		4110	Norway	EMEA	Bergulfsen	Jonas	Medium
10112	29	100	1	7209.11	3/24/2003	Shipped	1	3	2003	Classic	214	S10_19	Volvo Moc	0921-12 35	Berguvsv„gen		Lule		S-958 22	Sweden	EMEA	Berglund	Christina	Large
10126	38	100	11	7329.06	5/28/2003	Shipped	2	5	2003	Classic	214	S10_19	Corrida Au	(91) 555 22	C/ Araquil, 67		Madrid		28023	Spain	EMEA	Sommer	Martín	Large

Regression:

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed.

In this project we have used multiple linear Regression.

Multiple Linear Regression is one of the important regression algorithms which models the linear relationship between a single dependent continuous variable and more than one independent variable.

```
#fitting
from sklearn.linear_model import LinearRegression
regressor=LinearRegression()
regressor.fit(x_train,y_train)
```

Label Encoder:

Categorical data is data which has some categories such as, in our dataset; there are many categorical variables like State, Address etc.

Since machine learning model completely works on mathematics and numbers, but if our dataset would have a categorical variable, then it may create trouble while building the model. So it is necessary to encode these categorical variables into numbers. We have used **LabelEncoder()** class from **preprocessing** library.

```
#encodeing
from sklearn.preprocessing import LabelEncoder,OneHotEncoder
labelencoder_x=LabelEncoder()
x[:,5]=labelencoder_x.fit_transform(x[:,5])
x[:,6]=labelencoder_x.fit_transform(x[:,6])
x[:,10]=labelencoder_x.fit_transform(x[:,10])
x[:,12]=labelencoder_x.fit_transform(x[:,12])
x[:,13]=labelencoder_x.fit_transform(x[:,13])
x[:,14]=labelencoder_x.fit_transform(x[:,14])
x[:,15]=labelencoder_x.fit_transform(x[:,15])
x[:,16]=labelencoder_x.fit_transform(x[:,16].astype(str))
x[:,17]=labelencoder_x.fit_transform(x[:,17])
x[:,18]=labelencoder_x.fit_transform(x[:,18].astype(str))
x[:,19]=labelencoder_x.fit_transform(x[:,19].astype(str))
x[:,20]=labelencoder_x.fit_transform(x[:,20])
x[:,21]=labelencoder_x.fit_transform(x[:,21].astype(str))
x[:,22]=labelencoder_x.fit_transform(x[:,22])
x[:,23]=labelencoder_x.fit_transform(x[:,23])
labelencoder_y=LabelEncoder()
y=labelencoder_y.fit_transform(y)
```

Imputer:

If our dataset contains some missing data, then it may create a huge problem for our machine learning model. Hence it is necessary to handle missing values present in the dataset.

By calculating the mean: In this way, we will calculate the mean of that column or row which contains any missing value and will put it on the place of missing value. Here, we will use this approach.

To handle missing values, we will use **Scikit-learn** library in our code, which contains various libraries for building machine learning models. Here we will use **Imputer** class of **sklearn.preprocessing** library.

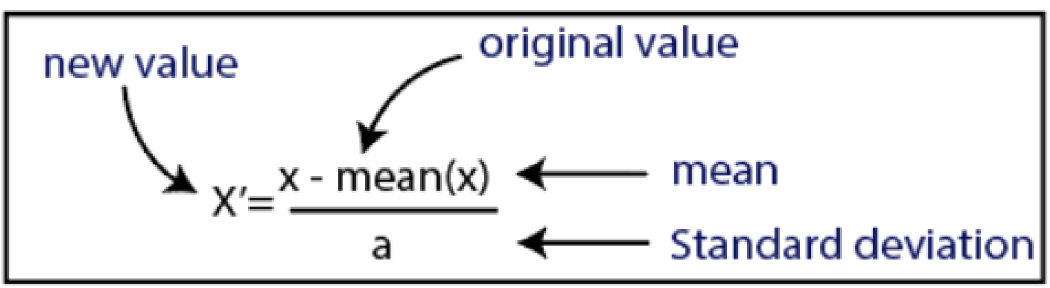
Below is the code for it:

```
#Imputation
from sklearn.preprocessing import Imputer
imputer=Imputer(missing_values="NaN",strategy='mean',axis=1)
x=imputer.fit_transform(x)
```

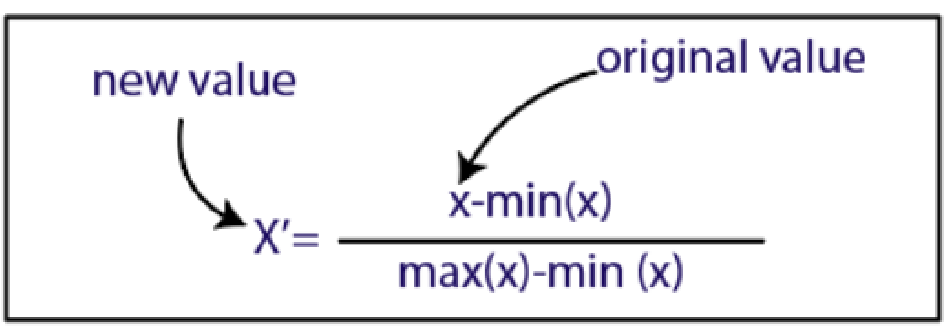
Feature Scaling:

It is a technique to standardize the independent variables of the dataset in a specific range. In feature scaling, we put our variables in the same range and in the same scale so that no any variable dominates the other variable. There are two ways to perform feature scaling in machine learning

Standardization



Normalization



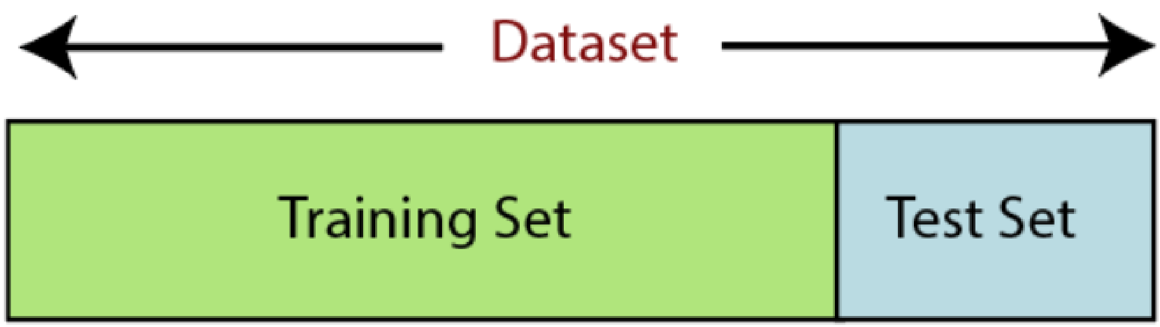
Here, we will use the standardization method for our dataset.

For feature scaling, we will import ***StandardScaler*** class of ***sklearn.preprocessing*** library as:

```
#standardlization
from sklearn.preprocessing import StandardScaler
st_x=StandardScaler()
x_train=st_x.fit_transform(x_train)
x_test=st_x.transform(x_test)
```

Splitting the Dataset into the Training set and Test set:

If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So, we always try to make a machine learning model which performs well with the training set and also with the test dataset. Here, we can define these datasets as:



Training Set: A subset of dataset to train the machine learning model, and we already know the output.

Test Set: A subset of dataset to test the machine learning model, and by using the test set, model predicts the output.

```
#split
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=0)
```

Backword Elimination:

Backward elimination is a feature selection technique while building a machine learning model. It is used to remove those features that do not have a significant effect on the dependent variable or prediction of output.

```
#backword elemination
import statsmodels.api as sm
x = np.append(arr = np.ones((2823, 1)).astype(int),  values = x, axis = 1)
#x_opt=x[:,[0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24]]
#x_opt=x[:,[2,3,5]]
x_opt=x[:,[0]]
ols=sm.OLS(endog=y,exog=x_opt).fit()
print(ols.summary())
```

Backword Elimination Summary:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                  0.000
Method:                 Least Squares   F-statistic:                       nan
Date:                Thu, 02 Apr 2020   Prob (F-statistic):                nan
Time:                        17:36:13   Log-Likelihood:                -2527.6
No. Observations:                2823   AIC:                             5057.
Df Residuals:                    2822   BIC:                             5063.
Df Model:                           0
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.3985      0.011    125.411      0.000       1.377       1.420
==============================================================================
Omnibus:                      211.904   Durbin-Watson:                   1.299
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              133.325
Skew:                          -0.406   Prob(JB):                     1.12e-29
Kurtosis:                       2.311   Cond. No.                         1.00
==============================================================================
```

Output:

```
Train score ::  0.756399911161953
Test score ::  0.7479824562750076
Mean Squared error ::  0.0896171148672l353
```

RESULT AND DISCUSSION

Train Score: 0.75639911161953

Test Score: 0.747982456275007

Mean Squared Error: 0.08961711486721353

Accuracy: 98%

Some of the data was not in proper way, so there was a need of data preprocessing. During data processing we divided the dataset into training and testing sets so that our model can be tested properly. After applying different techniques, we found that our data was neither overfitted nor underfitted, it was giving the result as mentioned above. Accuracy was more than 98%, if we further apply polynomial regression with 3-degree, accuracy can be achieved till 100% also.

CONCLUSION

As we can see that an intelligent sales prediction system is required for business organizations to handle enormous volume of data. Business decisions are based on speed and accuracy of data processing techniques. Machine learning approaches highlighted in this research paper will be able to provide an effective mechanism in data tuning and decision making. In order to be competent in business, organizations are required to work and adopt with modern approaches to accommodate different types of customer behaviour by forecasting attractive sales turn over. In my project, I used almost 15,000 records for the comparison of algorithms. Since the time of execution was huge and to manage such a large set of records are complex, some of the records were discarded, during the analysis phase. At the same time, fields and attributes, used in this analysis were insufficient for the further analysis. It was the major challenge we faced during the research. However, we had thoroughly weighed our works by implementing efficient ML techniques as like Imputer, Label Encoder, Backward Elimination for prediction and forecasting. The current studies can be expedited by using Big Data as a tool for the predictive analytics in sales forecasting. The big data analysis and forecasting are measured as the vital fields in the modern business scenario.

Reference

I. *Machine Learning Tutorial*

https://www.javatpoint.com/machine-learning

II. *Dataset*

https://www.kaggle.com/datasets

https://towardsdatascience.com/