e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:07/July-2022 Impact Factor- 6.752 www.irjmets.com

PREDICTING SALES IN FOOD MANUFACTURING INDUSTRY USING LINEAR REGRESSION

Rahul Modi[*1], Reshma Gulwani[*2]

[*1] Department of Information Technology, RAIT, Nerul, India.

[*2] D.Y. Patil Deemed to be University, Ramrao Adik Institute of Technology, Nerul, Navi Mumbai, India.

ABSTRACT

Food Manufacturing Industry is a Billion-dollar industry in a month more than 100,000 tonnes is produced, these results to more than million-dollar in sales. A manufacturer can produce in range of 1500 – 10,000 tonnes. In India Biscuits alone have a turnover of Rs.3000 crores. A major problem for manufacturer is to predict the sales according to which the production of biscuits will happen, Ordering and usage of raw materials happens on a large amount a single batch of production can use up-to 150 – 1000 Kg of raw materials. These manufacturers have to order these materials on monthly bases, with a lot these materials having low shell life and if not kept in proper conditions like Temperature and Humidity, can cause high hygiene problems. Using Supervised Learning algorithms like Linear regression and Multi Linear regression will try to predict the sales according to which manufacturer can order the raw materials needed. Linear regression and Multi Linear regression algorithms are popular algorithms used for predictive analysis and sentimental analysis.

Keywords: Sale Prediction, Linear regression, Multi Linear regression.

I. INTRODUCTION

Food Manufacturing Industry is a big industry where in Biscuits and Confectionary are big part of this industry. There is hardly a anyone in this world who has not eaten a biscuit or a candy or a chocolate. All of these food items are consumed by people on day-to-day bases. The market in this industry is not very volatile but has a lot of factors under consideration like current trends in market with respect to taste, quantity, ingredients, etc. This market is also divided in different ways some people like sour biscuits some like sweet, some like salt, the trends also keep changing with different flavors that coming in the market. These changes take a heavy toll on manufacturers as the machinery for different shapes and different types of biscuits are different and are at high cost. In India minimum amount to setup a factory that will have a production of 150 kg of biscuits will have a cost of 8 to 10 crores. With this cost changes in trends can lead to heavy losses. A lot of companies have factories that caterer to a single product, so these trends hurt them even more as they have to order raw materials for these plants while the demand in the market is less.

This leads to a lot of wastage, as the more the production happens and the demand is less financial stability will decrease. This also leads to rise in unemployment , to understand this problem lets take an example : A manufacturer sees rise in a product of company A he then convinces the company to let him fulfil a portion of that demand , let's assume he didn't have to change the machinery or he started with a new plant, now after 2 months the demand of the product decreases exponentially this leads to company dropping manufacturers, Now the manufacturer is left with a 2 months old machinery and no work so he will also reduce the labor. Now in this example the company and manufacturer took the decision looking at the trend but both had to deal with heavy losses. opposite to this example a manufacturer new or old should look at all things not just trends, but also at the items that have gradual increase in demand. To solve this problem, we will make model using linear regression with different combination of factors to see sales and profits of different products to decide which will be more beneficial for the manufacturer. Predictive analysis there are various models available like Random Forest, Linear Regression, Logistic Regression, K-means and many other algorithms are prepared according to the data available [7]. In our case we will look to see if there is a linear relationship between our independent variable and dependent variable. If we have a linear relationship, we will use Linear regression. The concept of linear regression was first proposed by Sir Francis Galton in 1894 [8]. Linear regression is test applied on a dataset to define and quantify the relation between considered variable. In today's world linear regression is used by Insurance company or to calculate effect of supplement in weight loss or the impact of advertisement on sales or water bill and amount of water used or connection between work experience and salary, etc. All these applications of Linear regression models are to see how value of a dependent variable i.e., salary, sales, water used, weight loss, Insurance premiums and changes with respect to independent variables like work experience,

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:07/July-2022 Impact Factor- 6.752 www.irjmets.com

water bill, advertisements, Risk assessment [13]. As the predictive target also helps in generating investments required in Research and development of that industry.

Firstly, we will see simple linear regression in which we have one dependent and one independent variable this regression is based on same formula as a line which defines relationship between x and y variable.

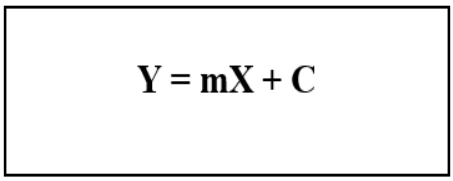

Figure 1 Basic Formula of a Line.

In above fig.1 we can see the formula used by linear regression when we have 1 dependent and 1 independent variable. This formula is used to describe a line with X and Y as points, m as slope of line and C as constant. For linear regression in this formula Y is a dependent variable and X is independent variable with b0 as y-intercept and b1 becoming beta coefficients or parameters and e is the error term also known as the residual errors. This formula is written as following.

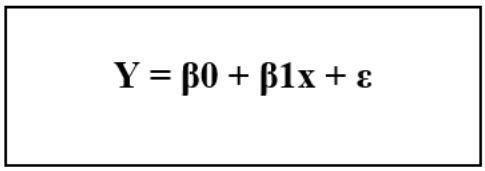

Figure 2 Formula for Simple Linear Regression.

If we have more than one independent variable let's say X1 and X2 then we have use one more beta coefficient b2 in the equation if we have X variables from 1 to p we will use beta parameters from 0 to p. The following formula shows how the equation will be with p independent variables in an equation.

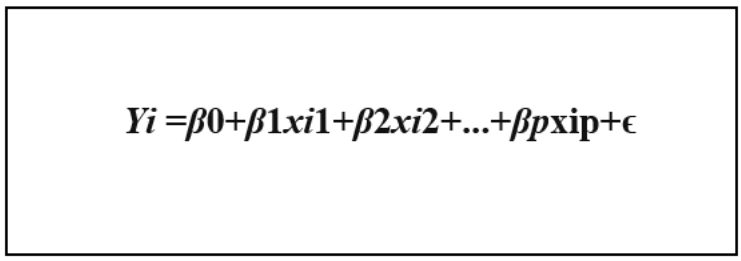

Figure 3 Formula for Multi Linear Regression.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:07/July-2022 Impact Factor- 6.752 www.irjmets.com

To find beta variables we will use built-in python functions from libraries that already exists in python.

```
In [5]: from sklearn.preprocessing import PolynomialFeatures

In [6]: polynomial_converter = PolynomialFeatures(degree=2,include_bias=False)

In [7]: polynomial_converter.fit(X)

Out[7]: PolynomialFeatures(degree=2, include_bias=False, interaction_only=False,
                           order='C')

In [10]: poly_features = polynomial_converter.transform(X)

In [11]: poly_features[0]

Out[11]: array([2.301000e+02, 3.780000e+01, 6.920000e+01, 5.294601e+04,
                8.697780e+03, 1.592292e+04, 1.428840e+03, 2.615760e+03,
                4.788640e+03])
```

Figure 4 Program For getting Polynomial values in python.

Above figure is example of how we use transform function to generate values for the equation to find Yi or the value of dependent variable. This value will be used to train and test the model.

This Linear Regression model will help in getting a better sense of market trends that are good for long term profits and trends that will not be there in few months [12].

When we are making a model for a particular industry, we have to make sure that we are using the model that shows the data in best way possible with respect to accuracy, complexity and different variables.

To decide which model to use we can try if the relation between dependent and independent variable. If the relationship is linear, we can use Linear regression model as using other models like random forest will give a similar result. Random forest is used when we have a complex relationship and do not have any time constraint [14].

Sometimes the data we have is highly volatile, the trends shown by this data can confuse a lot of people on which model will fit the most. Using models like SVR, decision tree or random forest could help, even these models have limitations.

Support Vector Regression works in similar way to SVM where in a hyperplane is created and the points closest to the hyperplane are support vectors. In regression a error or margin of tolerance is set in approximation, but in support vector regression it would have requested from the problem.

In this project we are using a dataset which not very volatile but has large differences in values, so using SVM or any other algorithm can give drastic change in results as compared to real values which means error rate can be high.

Food Industry in India is one of the biggest industries, according to a survey after USA, India produces highest number of biscuits and confectionary. India is also leading exporter of biscuits with countries From Africa and UAE importing biscuits and confectionary [15].

II. METHODOLOGY

In this process of making a model we have following steps:

[1]. Importing Libraries
[2]. Data loading
[3]. Data cleaning
[4]. Data Separation
[5]. Model Training
[6]. Model Testing

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:07/July-2022 Impact Factor- 6.752 www.irjmets.com

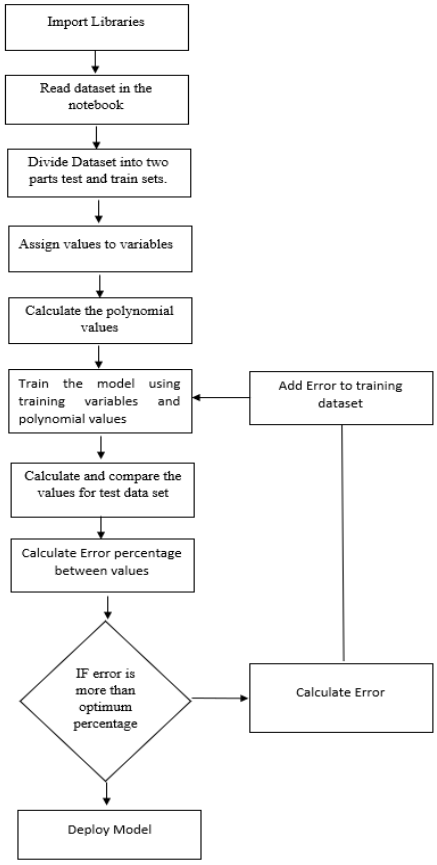

Figure 5 Process of making linear regression model.

1] Importing Library: we will import all the libraries that we are going to use directly into our project these libraries will be used for various functions from importing data to visualizations.

A. NumPy: NumPy is a Python programming language library that is used for making multi-dimensional arrays and matrices, along with this it is also used for mathematical functions.
B. Pandas: Pandas is another Python library that is used for data manipulation and analysis. It also helps in Data Structures and in manipulation of numerical tables and time series.
C. Matplotlib: Matplotlib is a python library used for plotting python programming language and its numerical mathematics extension NumPy. It also has "pylab" interface based on a state machine.
D. Seaborn: Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

```
In [2]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
```

Figure 6 libraries Imported.

2] Data Loading or Data Reading: In this project we have made our own data set using this dataset we will create the model. In this step we will just use the read function of Pandas library and load our dataset to our project file.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:07/July-2022 Impact Factor- 6.752 www.irjmets.com

```
In [26]: df1.drop(['S.No'],axis =1)
```

Out[26]:

	Year	Month	Sugar	Flour	Chemicals	Glucose	Profits	Sales
0	2017	Jan	3150 KG	5400 KG	180 KG	3870 KG	180000	3600000
1	2017	Feb	3000 KG	5720 KG	180 KG	3741 KG	167000	3300000
2	2017	March	3150 KG	5400 KG	180 KG	3870 KG	180000	3600000
3	2017	April	3150 KG	5400 KG	180 KG	3870 KG	180000	3600000
4	2017	May	3150 KG	5400 KG	180 KG	3870 KG	180000	3600000
5	2017	June	3150 KG	5400 KG	180 KG	3870 KG	180000	3600000
6	2017	July	3000 KG	5720 KG	180 KG	3741 KG	167000	3300000
7	2017	August	3000 KG	5720 KG	180 KG	3741 KG	167000	3300000
8	2017	Sep	3000 KG	5720 KG	180 KG	3741 KG	167000	3300000

Figure 7 Importing Dataset.

3] Data Cleaning: Since this dataset is made by us there are no missing or null values, but in some external dataset there are missing and null values which are to be addressed.

4] Data separation: The model we are creating we will have to divide the dataset into 2 parts one part will be used to train the data and the second data will be used to test the data and check the error values. In our project we have divide the data set into 80 – 20 percent which means 80% percent for training and 20% percent for testing purpose.

5] Model Training: We will assign the data into four parts X_Test, Y_Test, x_train and y_train. Then the train data variables will be passed to the model in order to undergo training.

Before we start training, we will need to find the polynomial data for that we will use polynomial_converter function.

```
In [6]: polynomial_converter = PolynomialFeatures(degree=2,include_bias=False)

In [7]: polynomial_converter.fit(X)
Out[7]: PolynomialFeatures(include_bias=False)

In [8]: poly_features = polynomial_converter.transform(X)

In [9]: poly_features[0]
Out[9]: array([2.301000e+02, 3.780000e+01, 6.920000e+01, 5.294601e+04,
               8.697780e+03, 1.592292e+04, 1.428840e+03, 2.615760e+03,
               4.788640e+03])
```

Figure 8 Polynomial Values.

After getting the polynomial values we have pushed these values in to our regression model.

```
In [10]: from sklearn.model_selection import train_test_split

In [11]: X_train, X_test, y_train, y_test = train_test_split(poly_features, y, test_size=0.3, random_state=42)

In [12]: from sklearn.linear_model import LinearRegression

In [13]: model = LinearRegression()

In [14]: model.fit(X_train,y_train)
Out[14]: LinearRegression()
```

Figure 9 Training Model using Train data.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:07/July-2022 Impact Factor- 6.752 www.irjmets.com

6] Model Testing: Using the model we create we will give X test variable as input and then we will get the value of Y this value is then checked with original Y value in Y test variable. The error is also calculated using functions that exists in python errors like Mean Square Error are calculated.

In this step we have tested the data to check whether the model is working properly or not. Also in this step we have calculated the error values.

```
In [15]: test_predict = model.predict(X_test)

In [16]: from sklearn.metrics import mean_absolute_error,mean_squared_error

In [17]: MAE = mean_absolute_error(y_test,test_predict)

In [18]: MSE = mean_squared_error(y_test,test_predict)

In [19]: RMSE = np.sqrt(MSE)

In [20]: MAE
Out[20]: 0.590597483380803

In [21]: MSE
Out[21]: 0.5231944949055424

In [22]: RMSE
Out[22]: 0.7233218473857557
```

Figure 10 Testing and Error calculation.

III. RESULTS AND DISCUSSION

In this experiment we have created a dataset using real world values in order to train the model with respect to the given parameters. In our dataset we had columns for different quantities of ingredients required in a month by a factory producing biscuits or confectionary. These quantities vary from one production plant to another as it depends on the machine used by manufacturer. In our dataset we have kept sales and profit columns in which profit is usually close to 5 percent of total sales.

The model that we have created will help us in making business expansion decisions. We can predict the number of sales or the number of raw materials we need to cater to future demands. This model will help manufacturer in also keeping a check on the sales happening in month with respect to other months and also with respect to sales for same month last year.

Year	Month	Sales	Predicted Sales	Error Percentage
2017	Jan	3600000	3528000	-5
2019	Jan	3800000	36,10,000	-5
2018	Mar	3600000	37,80,000	5
2017	Dec	3800000	38,87,400	2.3
2020	Oct	3300000	32,74,590	-0.77

Figure 11 Final prediction and error percentage.

Above given figure shows how different values for different months are predicted it also shows how much difference is there in predicted sales and sales original value. This difference is kept as error percentage. With increase in the dataset the error percentage should further decrease as more the model learns about this the better the prediction will be there.

IV. CONCLUSION

Food production industry is one of the biggest industries in the world. Manufacturing food requires heavy investment and to get good returns we need to take good statistical decision. These decisions cannot be taken without taking trends, cost and production rate into consideration. Linear regression is widely used in different industries to make a prediction model, but in food industry people still rely on basic statistics like a normal bar graph that will show how the present is going, but the relation and the effect of different trends creates complication in creation of such graphs.

In Our dataset complications are less as compared to some other industries, and if in such cases the error rate is high than tree algorithms like Decision tree or Random Forest etc, should be used. The model we have created has low error and will decrease with increase in datapoints as it learns more from the data the prediction will become better and the error rate will increase. Ensure that abstract and conclusion should not same.

V. REFERENCES

[1] Daniela Şchiopu, Elia Georgiana Petre, Catalina Negoina "Weather Forecast using SPSS Statistical Methods" This paper presents a case study of using SPSS 13.0 in weather prediction. Vol. LXI No. 1/2009 Gas University of Ploiesti, Bd. Bucuresti, No. 39, 100680, Romania.

[2] Wuthrich, Mario V. and Buser, Christoph, Data Analytics for Non-Life Insurance Pricing (October 27, 2021). Swiss Finance Institute Research Paper No. 16-68, Available at SSRN: https://ssrn.com/abstract=2870308 or http://dx.doi.org/10.2139/ssrn.2870308

[3] Abdulraheem Sal, Mohammed Raja, The Impact of Training and Development on Employees Performance and Productivity (july 10, 2016). International Journal of Management Sciences and Business Research, Vol. 5, Issue 7, July 2016, Available at SSRN: https://ssrn.com/abstract=2849769.

[4] Paras and Sanjay Mathur "A Simple Weather Forecasting Model Using Mathematical Regression" Department of Electronics & Communication Engineering, College of Technology, G.B. Pant University of Agriculture & Technology, Pantnagar, (India) 263 145.

[5] Goutami Bandyopadhyay "The Prediction of Indian Monsoon Rainfall: A Regression Approach" 1/19 Dover Place Kolkata-700 019 West Bengal India.

[6] Uddin, Gazi and Alam, Md. Mahmudul, The Impacts of Interest Rate on Stock Market: Empirical Evidence from Dhaka Stock Exchange (2010). South Asian Journal of Management Sciences, Vol. 4(1), pp. 21-30, (2010), Available at SSRN: https://ssrn.com/abstract=2941287

[7] Kaushal, Anirudh and Shankar, Achyut, House Price Prediction Using Multiple Linear Regression (April 25, 2021). Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2021, Available at SSRN: https://ssrn.com/abstract=3833734 or http://dx.doi.org/10.2139/ssrn.3833734.

[8] Fuller, James R., Investment Forecasting with Multivariate Linear Regression (January 12, 2007). Available at SSRN: https://ssrn.com/abstract=532002 or http://dx.doi.org/10.2139/ssrn.532002.

[9] Timothy DelSole and J. Shukla "Linear Prediction of Indian Monsoon Rainfall".

[10] Linear Regression. This article is about Linear regression… | by Renu Khandelwal.

[11] ML | Linear Regression – GeeksforGeeks.

[12] Linear Regression (Definition, Examples) | How to Interpret?

[13] Linear regression review (article) | Khan Academy.

[14] jeas_0617_6115.pdf (arpnjournals.org).

[15] www.indianmirror.com-indian-industries-biscuit.