International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 03 | Mar 2021 www.irjet.net p-ISSN: 2395-0072

COMMERCIALS SALES PREDICTION USING MULTIPLE LINEAR REGRESSION

Kodamagulla Kausthub[1]

1 Student, Department of computer science and engineering,Vidya Jyothi Institute of Technology,Telangana,India

---

Abstract - Commercials have always been one of the most important medium for a company to extensively promote their brand or to increase the sales of their product. In case of commercial sales managers, data analysis and data visualization play an important role because it helps the organizations to derive insights from given data and accordingly make decisions which can profoundly influence their product market and also sales. Using this analysis, many brands also improve the commercials content and relevance based on the customer's interest. This paper mainly aims to address the usage of multiple linear regression technique in case of predicting sales related to commercials which are displayed in mainly three forms of media namely TV, Radio & Newspaper. We also use Yellow- Brick library here so that it extends the Sci-kit learn API to make model selection and hyper-parameter tuning of the model easily(for visualization also)

***Key Words*: Multiple Linear Regression, Pandas, Sci-kit, Yellow-Brick, Sales Revenue**

1.INTRODUCTION

Commercials have always been the most important source of marketing for many brands. Brands mainly use commercials to market their product and also improve their sales. Commercials Sales Manager are primarily concerned with making proper decisions in order to increase their sales revenue[1]. This paper presents a technique of devising an algorithm for predicting the sales of a commercial. We use Multiple Linear Regression technique in order to predict the sales revenue for various forms of medium and subsequently use Yellow-brick library to make visualizations on our output.

Yellow-Brick library helps us to identify the error and also present various parameters like accuracy, bias and loss-error functions in a precise manner. This also helps us to understand our model function from basic and eventually get the steps required to increase the efficiency of our model.

2. BACKGROUND

2.1 COMMERCIALS DATA-SET CLASSIFICATION

In order to be in with the accordance of prediction methodology, we first download the commercials sales data-set from the kaggle repository. The data-set contains 202 rows and 4 columns in it. It also contains the distribution and %revenue of sales of each type. The 4 columns are namely TV,Radio, Newspaper & sales. The classification of data-set is mainly done for the purpose of understanding the forms of medium present in the data-set. The data-set contains 3 categorical variables and 1 continuous variable. The classification of the data-set helps us to understand the particular revenue for a particular medium.

2.2 LIBRARIES AND DATA CLEANING

Machine learning methodology has to be implemented in a step-by-step manner. This includes data-cleaning, data transformation and data pre-processing. Then a algorithm can be devised based on the input of model and finally testing phase comes into picture for handling errors and also accuracy of the model.

2.2.1 LIBRARY LOADING AND DATA CLEANING

The preliminary step involved in devising an model for commercials sales is loading the required libraries. In this case, we mainly load four libraries namely pandas,numpy , sci-kit and Yellow-brick.

Pandas: Pandas is a python package which is quite quick, easy to use and structured in nature. Pandas data-frame is mainly used for data analysis purposes. Pandas also helps us to handle missing data, data to be reshaped and data transformation methodologies.

Numpy: Numpy is essentially used for creating very powerful and intuitive n-dimensional arrays. It offers various mathematical functions and also supports various kinds of computing hardware and software requirements. It is an open-source project and also contains various array-objects which are generally quick for usage.

Sci-kit: Sci-Kit is one of the most efficient and useful machine learning libraries in python. It provides various statistical and mathematical tools. It also have various techniques like regression, clustering etc. in it.

Yellow-Brick: It is an open-source project which extends the functionality of sci-kit library for data error reporting and visualization purposes. It's main advantage is data exploration. It also gives us vital information about parameters like mode stability, model accuracy etc.

3. PROPOSED ALGORITHM METHODOLOGY

This paper mainly addresses the usage of an algorithmic technique name Multiple Linear Regression[2]. We achieve a greater accuracy on sales revenue using multiple linear regression. Along with this, a library named yellow-brick is also used for noting down the error prediction value.

3.1 DATA-SET DESCRIPTION

Data-set description is very essential for understanding our data. In order to get insights from our data, we mainly use two commands. The first one is .info() command which gives us information about the number of rows and columns in the data-set and the other one is .describe() which explains various parameters like count(),min(),standard deviation(), max() etc.

	TV	radio	newspaper	sales
count	200.000000	200.000000	200.000000	200.000000
mean	147.042500	23.264000	30.554000	14.022500
std	85.854236	14.846809	21.778621	5.217457
min	0.700000	0.000000	0.300000	1.600000
25%	74.375000	9.975000	12.750000	10.375000
50%	149.750000	22.900000	25.750000	12.900000
75%	218.825000	36.525000	45.100000	17.400000
max	296.400000	49.600000	114.000000	27.000000

Fig A: Data-set Description

3.2 DATA-SET VISUALIZATION

The given data-set is then visualized using a library named seaborn. Seaborn helps us to create plots and live-interactive statistical plots and images. It also contains matplotlib which helps us to view our data in a much more intuitive manner. Data-set visualization is very useful for understanding the data contents and also trying to express the given data in form of various charts, figures, plot etc.

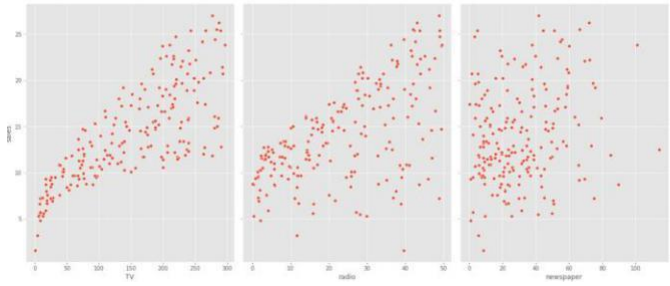

Fig B. Data-set Visualization

3.3 MULTIPLE LINEAR REGRESSION

3.3.1 ALGORITHM EXPLANATION

Linear regression is basically defined as a prediction based analysis. It is mainly done to model a relationship between a dependent variable and set of independent variables. Multiple Linear Regression comes under it. It is defined as a technique which tries to model relationship between two or more features. The main point of multiple linear regression lies in evaluation of algorithm. Some important points in are:

A) Multiple linear regression tries to fit points into multi-dimensional space region.

B) For it to work, the dependent variable has to be continuous and independent variable can either be continuous or categorical.

3.3.2 MULTIPLE LINEAR EQUATION:[2]

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p ,$$

where is the predicted or expected value of the dependent variable, X1 through Xp are p distinct independent

or predictor variables, b0 is the value of Y when all of the independent variables (X1 through Xp) are equal to zero, and b1 through bp are the estimated regression coefficients. Each regression coefficient represents the change in Y relative to a one unit change in the respective independent variable. In the multiple regression situation, b1, for example, is the change in Y relative to a one unit change in X1, holding all other independent variables constant (i.e., when the remaining independent variables are held at the same value or are fixed). Again, statistical tests can be performed to assess whether each regression coefficient is significantly different from zero.

3.3.3 MULTIPLE LINEAR REGRESSION STEPS

A) Firstly select your variables: Make sure that you select proper predictor variables.

B) Refine your model: Try to refine your or fine-tune your model with methods like rmse(root mean square error method). It helps you to get the estimation of standard deviation for random error prediction.

C) Test your model assumptions: Make sure that your data has no major outliers, better relationship between the variables and also your data should be independent of auto-correlation.

D) Validate the model: Cross validate the results by splitting the data into two forms. Use first form for model parameter checking and other for prediction based modelling. Also make sure that the results predicted are according your initial estimation.[3]

E) Yellow-Brick for error reporting: Use Yellow-Brick library for error report generation and also for getting access to various diagnostic tools.

4. OBSERVATIONS AND RESULTS

Accuracy of the multiple linear regression algorithm for this model is calculated and was found to be 91.7%. A correlation matrix in form of a heat-map is built in order to understand the dependency and also data relation.

Fig C: Correlation Heat-map

4.1 ERROR PREDICTION

Error prediction is done by using a library named yellow-brick. It contains a function named PredictionError which helps us to get prediction based error value while fitting the model.

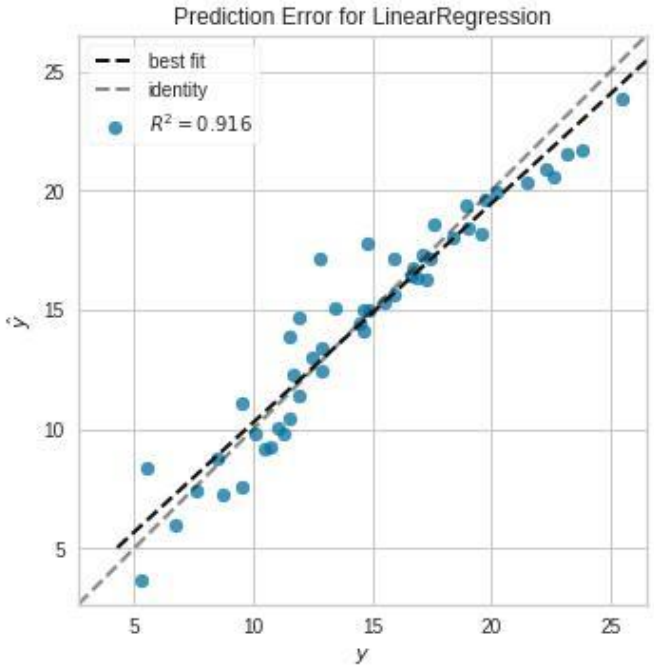

Fig D: Prediction error for the algorithm

5. CONCLUSION

As we finally complete the error prediction method, we come to know that the rmse value stands at 0.916. This paper helped us to analyzed whether our variables fit accordingly or not in the given model and we also achieved an overall accuracy of 92% . In future, we can extend this project by taking various other prediction based algorithms into account and eventually rate them based on their accuracy levels and error values. We can also take a much larger data-set with more categorical variables being added into our data-set. This also helps us to understand and implement an algorithm with much more accuracy and less error values respectively. A survey paper regarding all the algorithms can also be done using this particular data-set.[4]

REFERENCES

[1] Dawes, John & Kennedy, Rachel & Green, Kesten & Sharp, Byron. (2018). Pre-publication version to: Forecasting advertising and media effects on sales: Econometrics and alternatives. International Journal of Market Research. 60.10.1177/1470785318782871.

[2] Kologlu, Yunus & Birinci, Hasan & Ilgaz, Sevde & Ozyilmaz, Burhan. (2018). A Multiple Linear Regression Approach For Estimating the Market Value of Football Players in Forward Position.

[3] Schinazi, Rinaldo. (2012). Multiple Linear Regression. 10.1007/978-0-8176-8250-7_10.

[4] Singh, Manpreet & Ghutla, Bhawick & Jnr, Reuben & Mohammed, Aesaan & Rashid, Mahmood. (2017). Walmart's Sales Data Analysis - A Big Data Analytics Perspective. 114-119. 10.1109/APWConCSE.2017.00028.