Predicting future sales from marketing campaign

Bang-Trinh Tran-To

Nhat Anh Truong

Faculty of Information Technology

University of Science, HCMVNU

April 6, 2022

Abstract: Sales forecasting is an important aspect of marketing and business. Predicting the sales based on the budget allocated to each advertising channel helps businesses manage their operations and change their marketing strategy accordingly for each period of time in the future to optimize profits. This paper proposes an approach to predicting future sales of marketing campaigns using a multivariate regression model. We describe multivariate linear regression in the more approachable way, which is the combination of forming a mathematical regression analysis model, explaining its mechanics, and applying theory into practice by addressing the real-life example proposed in this paper. Thus, conquering regression models is essential for the work of anticipation and estimation that can be an advantage in various fields.

Keywords: Linear regression, multiple linear regression, forecast, sales prediction, marketing.

1 Introduction

Sales prediction is a crucial part of marketing and business that can provide enterprises with foreseeing information to boost their revenue. Due to the rapid flow of messages, people are attempting to anticipate and comprehend a company's earnings by employing a variety of patterns and methodologies. Various methods have been introduced and researched to handle forecasting tasks, such as linear [1] and nonlinear [2] regression, LSTM [3], etc. Inheriting from previous work, this paper proposes an approach to predict future sales of marketing campaigns using multi-variate regression models.

Examining the connection between the independent and dependent variables is known as regression. We can use multivariate linear regression to forecast sales of the company's product based on the budget allocated to each marketing channel, which is beneficial to save money on excess inventory, appropriately plan for productions in the future, and enhance profits. Subsequently, the regression model has reference value in marketing areas and many other industries.

The study offers several main contributions. To begin with, we present a multiple linear regression model in theory with the mathematical basis and explain how it functions. Furthermore, we perform the process of applying multiple linear regression in solving realistic problems. Also, the paper evaluate the model and give implications.

We believe that linear methods can be as precise as the more complex nonlinear methods in some uncomplicated problems or even faster due to the simplicity of the linear model. However, when much more in-depth analysis is required, nonlinear methods may prove superior. More research is needed in sales forecast modeling, including the more complicated model that can feed more structural elements, up-to-date with the latest changes in current market and marketing trends.

2 Related Work

Kausthub, K. (2021) [1] showed that the study could complete the error prediction approach and examine if its variables fit into the supplied model appropriately or not by using multiple linear regression techniques. We seem to agree that the paper is well-defined and intelligible. However, there was solid convergent evidence that they only focused on data processing. The study started with analyzing the data then paying attention to applying an algorithmic approach known as multiple linear

regression. The process was beneficial to those who already knew about the model. But, of course, this was not the case; a lack of consensus exists on the point of missing the process of building the model.

Rong, S., Bao-wen, Z. (2018) [4] introduced the algorithm and model of the field of machine learning and demonstrated how to use Python 3.6 to investigate how temperature affects the selling of iced items. We supported the claim that the paper made data analysis in data mining more accessible. In comparison, earlier work focused on how to cleanse data and its applications. This study spent more than three-quarters of its research describing the theoretical. The investigation introduced Python 3.6 then instructed data mining, linear regression, and multiple linear regression. The process may give a limitation for us to have a better view of its application.

Inheriting the strengths of the previous studies, our paper proposes a new way to describe linear regression models as a combination of them. Instead of focusing on either side, we will present the mathematics theory, then visualize it with the graph to clarify the impact of data on the results, and ultimately implement it in Python to explain how it works from the computer science point of view. In addition, we have discovered the already cleansed data set, which means we will put great effort into building models rather than manipulating data. Furthermore, we apply the theory into practice by building the model with real data to tackle the work of predicting sales based on the marketing campaign. This paper is not the discovery about linear regression, but we believe it could be the more approachable way to grasp linear regression profoundly and employ it effectively.

3 Background

In this section, we present some fundamentals of multiple linear regression in theory, which are the basis for applying it with the actual data to tackle the work of predicting.

Multiple regression is a method where the dependent variable shows a linear relationship with two or more independent variables. When you want to know how strong the association is between two or more independent factors and one dependent variable and the value of the dependent variable at a given value of the independent variables, you may use multiple linear regression.

Using knowledge about another variable, statisticians may predict the value of one variable using simple linear regression. Linear regression seeks to construct a straight-line relationship between the two variables. Multiple regression is a type of regression in which the dependent variable has a linear relationship with two or more independent variables.

Multiple linear regression equations have the same structure as simple linear regression equations, but with extra terms [5]:

$$Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + ... + \beta_p X_p^{(i)} + \epsilon^{(i)}$$

where, for $i = n$ observations:

$Y^{(i)} = i^{th}$ dependent variable
$X^{(i)} = i^{th}$ explanatory variables
p = number of explanatory variables (features)
X_p = explanatory variable p^{th}
β_0 = y-intercept (constant term)
β_p = slope coefficients for each explanatory variable
$\epsilon^{(i)}$ = the model's error term (also known as the residuals)

and the formula in terms of matrix is as below:

$$Y^{(i)} = \beta^T x^{(i)} = [\beta_0 \beta_1 ... \beta_p] \begin{bmatrix} x_0^{(i)} \\ x_1^{(i)} \\ ... \\ x_p^{(i)} \end{bmatrix}$$

Multiple linear regression is extensively used in econometrics and financial research. The dependent variable y has a value for each value of the independent variable x. The population regression line is defined as $Y^(i) = \beta_0 + \beta_1 X_1^{(i)} + \beta_1 X_2^{(1)} + ... + \beta_p X_p^{(i)} + \epsilon^{(i)}$ for p explanatory variables x_1, x_2,..., x_p. This graph depicts how the explanatory variables influence the mean answer y. The standard deviation is expected to be the same as the observed values for y vary around their means y. The dependent variable is the one we

want to predict, whereas the independent or explanatory factors are used to anticipate the dependent variable's value.

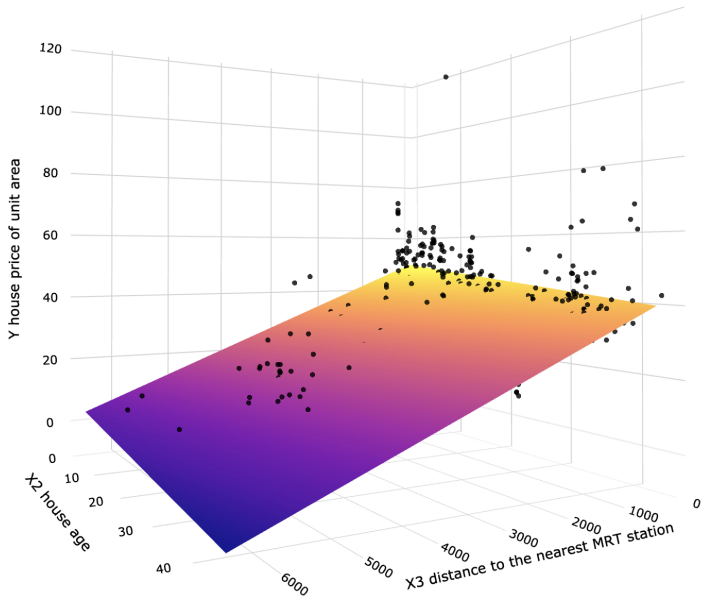

Figure 1: Multiple linear regression — observation 3D scatter plot with a prediction plane. Graph by Saul Dobilas

For example, in this figure, we will see the relationship between distance to the nearest MRT station, house age, and house price of unit area. X-direction show X_3 distance to the nearest MRT station, Y direction displays Y house price of unit area, and Z direction tells us the X_2 house age. We can use the values of X_3 and X_2 to predict the value of Y.

4 Methodology

The data set used in this study, which is analyzed clearly in below section, contains four marketing channels, which are: `TV`, `Radio`, `Social Media`, and `Influencer`. Therefore, the equation for sales will be:

$$Sales = \beta_0 + (\beta_1 \times TV) + (\beta_2 \times Radio) + (\beta_3 \times SocialMedia) + (\beta_4 \times Influencer) + \epsilon$$

$$Sales = \beta^T x^{(i)} = [\beta_0 \beta_1 \beta_2 \beta_3 \beta_4] \begin{bmatrix} 1 \\ TV \\ Radio \\ SocialMedia \\ Influencer \end{bmatrix}$$

We first generate the random number for β with small values, from β_0 to β_4. After having the results of predicted sales, we calculate the difference between the predicted and the absolute value (cost function). Our model will be trained by updating the weight (β). The weight updated by the formula:

$$\beta = \beta - \alpha \times \frac{\partial}{\partial \beta} J(\beta)$$

With $J(\beta)$ is the cost function and $\frac{\partial}{\partial \beta} J(\beta)$ is the derivative of its, which indicates the slope of the $J(\beta)$. This process is called **Gradient Descent** that helps in updating the weight to minimize the difference between the predicted value and the actual data. The results gained from the training process are the intercept (β_0) and coefficients ($\beta_{1..p}$), which form the line predicting future sales.

The process of experiment is carried out by several steps:

- *Step* 1: Data pre-processing and visualization.
- *Step* 2: Implementing the model with train data set.
- *Step* 3: Testing the model with test data set.
- *Step* 4: Calculating the accuracy and error of the model to evaluate the effectiveness of the model.

5 Experimental design

We build the multiple linear regression model in Python programming language on Google Colaboratory environment. The data used in the experiment is from Kaggle [6]. The data has four marketing channels and is visualized later with the help of `matplotlib` and `seaborn` library. We import `scikit-learn` library to use the built-in linear regression algorithm and metrics to build and evaluate models. Other popular libraries that support the study are `numpy` and `pandas`.

5.1 Data analysis

Before building model, we need to explore and understand the data to know what we have and what we can do to accustom our model to data.

	TV	Radio	Social Media	Influencer Mega	Influencer Micro	Influencer Macro	Influencer Nano	Sales
0	16.0	6.566	2.907	1	0	0	0	54.732
1	13.0	9.237	2.409	1	0	0	0	46.677
2	41.0	15.886	2.913	1	0	0	0	150.177
3	83.0	30.020	6.922	1	0	0	0	298.24
4	15.0	8.437	1.405	0	1	0	0	56.594
5	29.0	9.614	1.027	1	0	0	0	105.88
6	55.0	24.893	4.273	0	1	0	0	198.67
7	31.0	17.355	2.289	0	0	0	1	108.73
8	76.0	24.648	7.130	0	0	1	0	270.18
9	13.0	0.431	2.229	1	0	0	0	48.280

Table 1: First 10 rows of data

Visually, the data has four marketing channels: `TV`, `Radio`, `Social Media`, and `Influencer`. The considered marketing channels are popular and have a significant impact on sales in daily lives. Because the value of `Influencer` is not the number but the word, so we divide this column into four zones, each of which has the value of 1 or 0 to mark which `Influencer` is chosen for the corresponding campaign. This process is called one-hot encoding.

Viewing data with tables cannot give us a general picture of data. Therefore, we need to visualize data to get insights into it. Visualization gives us a deeper understanding of how data vary and the dependency of each feature on the results.

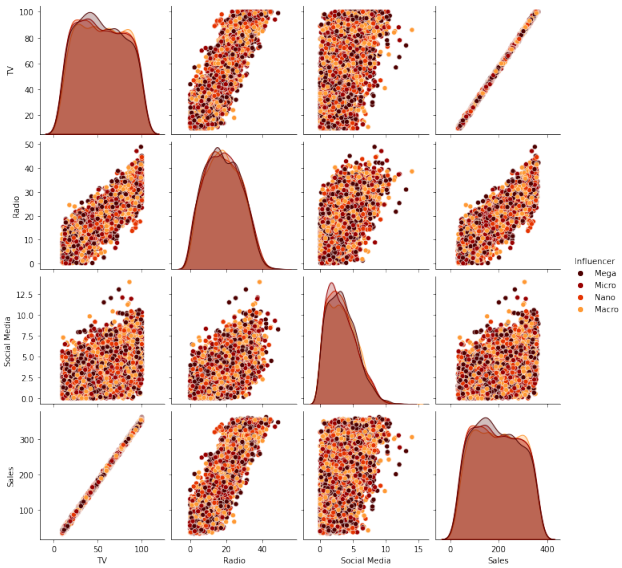

Figure 2: General data visualization

From the graph above, we can see that the data of TV feature almost formed a line. Hence, we can imply that the impact of TV is directly proportional to sales.

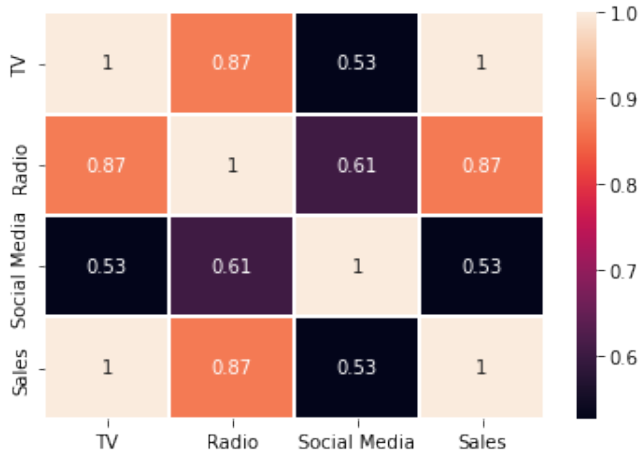

Figure 3: Dependency of each features on output using heatmap

A correlation matrix is visualized using a heat map to tell us the sales dependency on each feature. From the heat-map above we can see that the sales are affected mainly by the TV marketing channels.

5.2 Implementation

Before building the model, we randomly select 70% of the data set for training data and the rest, 30%, is reserved for the prediction step.

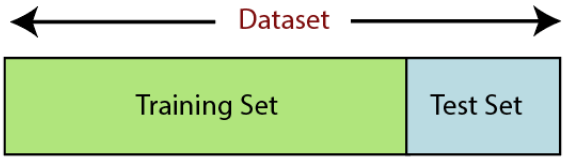

Figure 4: Data set is split for training and testing

We create an instance of `LinearRegression()` class and use `x_train`, `y_train`, which are features and labels, respectively, to train the model using the `fit()` method of that class to fit the features to the label, or is known as the process of training.

```
1 from sklearn import linear_model
2 model = linear_model.LinearRegression()
3 model.fit (x_train, y_train)
```

The result of the model is the Intercept and Coefficient of the linear regression, which forms the line. Then, we will extract the intercept and coefficient of

the model to see the result after training.

Intercept	-0.1782040902
Coefficient	
`TV`	3.5636
`Radio`	-0.0025
`Social Media`	-0.0062
`Influencer_Mega`	-0.0083
`Influencer_Micro`	-0.0032
`Influencer_Macro`	0.0866
`Influencer_Nano`	-0.0750

Table 2: The intercept and coefficient of the linear regression model after training

$$\begin{aligned} Sales = &-0.178 + (3.5636 \times TV) \\ &- (0.0025 \times Radio) \\ &- (0.0062 \times SocialMedia) \\ &- (0.0083 \times Influencer\_Mega) \\ &- (0.0032 \times Influencer\_Micro) \\ &+ (0.0866 \times Influencer\_Macro) \\ &- (0.0750 \times Influencer\_Nano) \end{aligned}$$

From the above-obtained equation for the Multiple Linear Regression Model, we can see that the value of intercept is -0.178, which shows that if we keep the money spent on `TV`, `Radio`, `Social Media`, and those four `Influencers` for advertisement as 0, the estimated average sales will be -0.178 and a single unit of sales increase in the money spent on:

- `TV`: increases sales by 3.5636.
- `Radio`: decreases sales by 0.0025.
- `Social Media`: decreases sales by 0.0062.
- `Influencer_Mega`: decreases sales by 0.0083.
- `Influencer_Micro`: decreases sales by 0.0032.
- `Influencer_Macro`: increases sales by 0.0866.
- `Influencer_Nano`: decreases sales by 0.0750.

6 Results

After training is the process of testing to evaluate the accuracy and the error of our model with respect to actual data.

index	**Actual value**	**Predicted value**
4070	296.3267871	295.5892337974368
2374	320.8831438	320.46259468403866
6	198.6798248	195.73092679429774
9	48.28058223	46.12600814902145
10	224.9610192	220.60065165315496
19	329.3505396	327.47275014979857
25	192.4619083	192.0894778955264
27	96.74937343	96.00560570315241
28	76.99292807	74.60794273649677
32	40.79219533	42.47946009444598
38	281.5789422	281.36538030826125
39	231.6700477	231.46368024837983
40	148.7069665	153.09761258185125
44	91.92229739	92.53292584616433
48	101.8731014	103.23231393672152
51	326.3817124	323.96101073521936
57	50.98848158	49.68911032608148
58	70.47172273	71.13948221950021
59	58.59009126	60.281799234342884
62	181.0796122	185.06457605206487
63	145.9192	145.83137128415194
65	288.9822393	291.9299556962976
66	49.81567405	49.62951929310184
69	55.00971238	53.25604649850727
71	247.1394014	249.29805640227175

Table 3: The comparison between the predicted and actual values

The table shows the values of the sale and the predicted values. The first column in the table shows the actual or absolute value of the data used for comparison. The second column shows the predicted values that we forecast. From the table, it can be observed that this method denotes values closer to the actual values in most of the period. The prediction performance is evaluated in terms of 'hit ratio,' which is the percentage of times our methods have been applied in the correct direction.

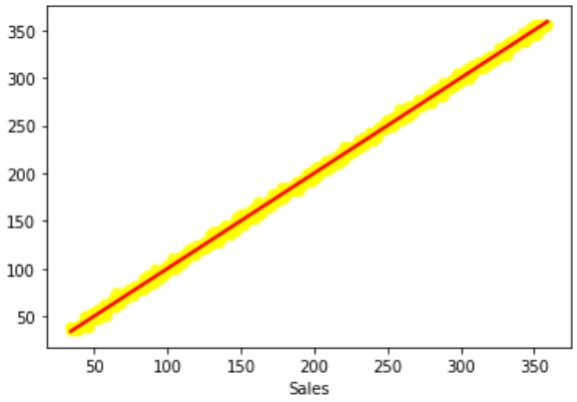

Figure 5: The linear regression sketched by predicted values with respect to the actual value

The graph displays a regression model that assesses the actual and forecast value relationship. Based on the figure, the yellow path is our absolute value of the data, while the red line is the result that our model can predict. The absolute value results were found to match very closely with the results reported by the expected model.

To demonstrate how well our model predict data, we use metrics `R-squared`, `MSE`, `MAE`, and `RMSE` to measure the accuracy as well as the errors, which are all provided by `sklearn` library.

R squared	99.90
Mean Absolute Error	2.3446477605060436
Mean Square Error	8.426457713486725
Root Mean Square Error	2.9028361499551996

Table 4: Evaluation results

This study uses `R-squared` to indicate how many points fall on the regression line. The value of `R-squared` is 99.90, which suggests that 99.90% of the data fits the regression model. The prediction performances are evaluated using the standard evaluation metrics: Mean Squared Error (`MSE`), Mean Absolute Error (`MAE`), and Root Mean Square Error. Mean Absolute Error is the absolute difference between the actual and predicted values, and `MAE` calculated on our model is 2.34. Mean Squared Error is calculated by taking the average of the square of the difference between the original and predicted values of the data. In this research, we estimate `MSE` and have a result of about 9.03. Root Mean Squared Error is the same as Mean Squared Error, but the root of the value is considered while determining the model's accuracy. The `RMSE` measured on our model is 2.90.

7 Conclusion

In this paper, we have described multivariate linear regression through many aspects, including theory, method's structure, and building the model to experiment with its application in forecasting the sales formulated on marketing campaigns. We have presented the mathematical theory of multiple linear regression, and explained how it functions to predict sales. After that, we have shown the process to build a model from exploratory data analysis (plotting data to understand the impact of the data on the results) to model training with the `scikit-learn` library in Python.

The evaluation results show that the goodness-of-fit is up to 99.90%, which implies a minimal difference between the observed sales and the one predicted by our model. We also measure the error of our model with some metrics, including `MSE`, `MAE`, and `RMSE`. The resulting error measured by those three metrics is 8.43, 2.34, and 2.90, respectively. Our findings indicate that the magnitudes of our model give relatively large errors, so we plan to improve our model and deal with it accordingly.

The results imply that the model works fine in prediction tasks in marketing and other fields with unsophisticated data. Nevertheless, the data will be getting more complicated due to the market fluctuations and the expansion of different marketing channels. Thus, the multiple linear regression is no more effective in predicting recent sales; there is essential to research other non-linear regression models and time series to have a more grounded result and better performance on sales projection.

References

[1] Kausthub, K. (2021). "COMMERCIALS SALES PREDICTION USING MULTIPLE LINEAR REGRESSION." *International Research Journal of Engineering and Technology (IRJET)*. *08*(03).

https://www.researchgate.net/publication/350398584_COMMERCIALS_SALES_PREDICTION_USING_MULTIPLE_LINEAR_REGRESSION

[2] Hossain, I., Esha, R., Alam Imteaz, M. (2018). "An Attempt to Use Non-Linear Regression Modelling Technique in Long-Term Seasonal Rainfall Forecasting for Australian Capital Territory." *Geosciences (Switzerland)*.

https://www.researchgate.net/publication/326708581_An_Attempt_to_Use_Non-Linear_Regression_Modelling_Technique_in_Long-Term_Seasonal_Rainfall_Forecasting_for_Australian_Capital_Territory

[3] K. Chen, Y. Zhou and F. Dai, "A LSTM-based method for stock returns prediction: A case study of China stock market," 2015 *IEEE International Conference on Big Data (Big Data)*, 2015, pp. 2823-2824.

https://ieeexplore.ieee.org/abstract/document/7364089

[4] Rong, S., Bao-wen, Z. (2018). "The research of regression model in machine learning field." *MATEC Web of Conferences*.

https://www.researchgate.net/publication/326121964_The_research_of_regression_model_in_machine_learning_field

[5] Adam Hayes. "Multiple Linear Rergression (MLR)."

https://www.investopedia.com/terms/m/mlr.asp#:~:text=Key%20Takeaways-,Multiple%20linear%20regression%20(MLR)%2C%20also%20known%20simply%20as%20multiple,uses%20just%20one%20explanatory%20variable.

[6] HARRIMAN SAMUEL SARAGIH. (2021). "Dummy Marketing and Sales Data." *Kaggle*.

https://www.kaggle.com/harrimansaragih/dummy-advertising-and-sales-data