https://doi.org/10.51483/IJDSBDA.1.2.2021.31-38 ISSN: 2710-2599

SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Data Science and Big Data Analytics

Publisher's Home Page: https: / / www.svedbergopen.com /

Research Paper **Open Access**

Restaurant tip prediction using linear regression

Alex Mirugwe[1*]

[1]Department of Statistical Sciences, Faculty of Science, University of Cape Town, Cape Town, South Africa. E-mail: Mrgale005@myuct.ac.za

Article Info

Volume 1, Issue 2, May 2021
Received : 12 December 2020
Accepted : 10 April 2021
Published : 05 May 2021
doi: 10.51483/IJDSBDA.1.2.2021.31-38

Abstract

The objective of this paper is to build a linear model for predicting the average amount of tip in dollars a waiter is expected to earn from the restaurant given the predictor variables, i.e., total bill paid, day, the gender of the customer (sex) time of the party, smoker and size of the party. The model was based on the data created by one waiter at a certain restaurant in California who recorded information about each tip he received. This model can be applied at any restaurant with similar predictor variables to determine the amount of tip. The final result from this analysis proved a regression model with a minimum prediction Root Mean Square Error (RMSE) of 1.1815.

Keywords: *Machine learning, Linear regression, Mean Squared Error (MSE), Root Mean Squared Error (RMSE)*

1. Introduction

In this information era, machine learning (Lim, 2019) is one of the big things and it's being widely used to automate and solve computational problems. Machine learning refers to the approach of automating the detection of meaningful patterns in data (Shalev-Shwartz and Ben-David, 2014). The main idea of machine learning is to use previously known data to train a model that can be used to predict the solutions for the problem (Lim, 2019). It's being applied in almost every sector, i.e., education, agriculture, health, transport, and many others (Zeng, 2016).

This study aims to use a linear regression analysis to build a model for predicting the average amount of tip in dollars a waiter can expect to get from the restaurant given a number of predictor variables. Linear regression is one of the most widely used predictive methods (Lim , 2019; James *et al*., 2013) for predicting quantitative responses. This research is based on the data created by one waiter who recorded information about each tip he received over a few months working at one restaurant in California. So, we intend to find out which predictor variables played a vital role in determining the amount of tip received by the waiter.

2. Background

Over years, researchers have used statistical modeling approaches to discover patterns in datasets. And linear regression analysis (Kologlu *et al*., 2018) is one of those approaches that have played a fundamental role in the field of machine learning. Linear regression (Zeng, 2016; James *et al*., 2013) is a modeling approach used in predicting a scalar response y on a basis of X_i independent variables. This model takes a form,

* *Corresponding author: Alex Mirugwe, Department of Statistical Sciences, Faculty of Science, University of Cape Town, Cape Town, South Africa. E-mail: Mrgale005@myuct.ac.za*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \ \beta_p X_p + \varepsilon \quad \ldots(1)$$

In Equation (1), Y represents a response variable, β_0 is model intercept and $\beta_{(1\ldots p)}$ are the slope terms, $X_{(1\ldots p)}$ are the independent variables and finally ε denote a mean-zero random error (residual) term (or the difference between the predicted and actual values). Figure 1, shows how a model with two predictor variables x_1 and x_2 and several observations can be represented on a 3D plane.

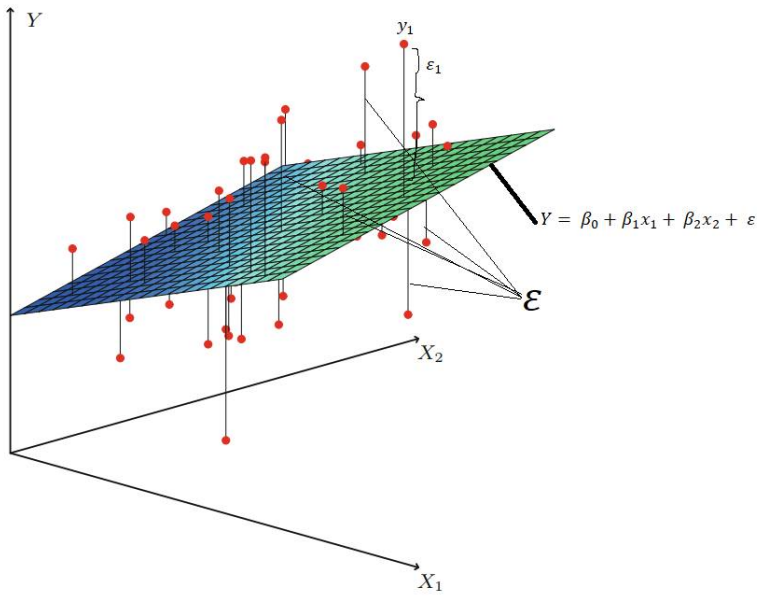

Figure 1: A 3-D representation of two predictors and one response variable, then the least-squares regression line becomes a plane. This plane is chosen to minimize the squared sum of residual values

The accuracy of a linear regression is assessed (Doan and Kalita, 2016) using Mean Absolute Error (MEA), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 score metrics. The smaller the value of these metrics, the better the model. The best model (Kavitha *et al*., 2017) can be identified as one with minimized RMSE, MEA, and MSE values.

MEA is the mean of the absolute value of the errors,

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \tilde{y}_i| \quad \ldots(2)$$

MSE is the mean of the squared errors,

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \quad \ldots(3)$$

RMSE is the square root of the mean of the squared errors,

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2} \quad \ldots(4)$$

In Equations (1,2,3), n represents the number of observations, y is the actual response value and $\tilde{y}$ is the predicted value.

The machine learning model goes through several stages and shown in Figure 2. Keen attention must be paid at every stage for one to have a better predictive model, i.e., a model with minimized MSE and RMSE values.

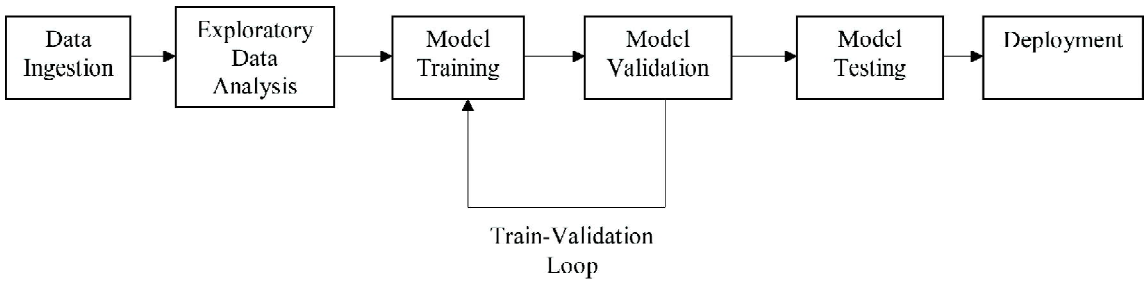

Figure 2: The procedure for a linear regression analysis

3. Data and implementation

3.1. Dataset description

The model was fitted using the dataset of 200 observations (restaurant customers) and seven variables. Of the seven variables, three numerical were variables and these are total bill, tip amount, and size of the party. Four were categorical variables and that is sex (with two levels female and male), smoker (yes or no), day (Thursday, Friday, Saturday, and Sunday), and time (lunch and dinner). Of the 200 customers, 66 were female 134 were male. This means 67% of the total bill payers were male and 33% were females. It was also observed that the majority of the restaurant customers paid a bill below $25 and also most waiters received tip amount less than $4. The figures below show the distribution of the total amount of bill paid and the amount of tip received by the waiter.

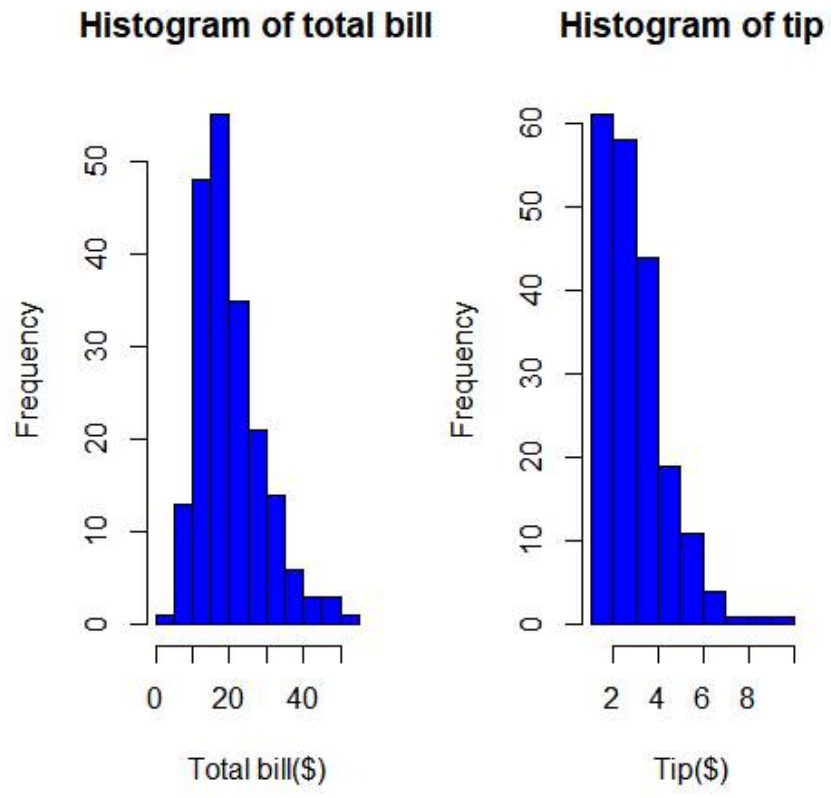

Figure 3: Histograms showing the distribution of the total bill paid by the customer and tip received by the waiter

We went further to analyze how day, gender, smoking, and time variables are interrelated to each other, and this is being shown in Figure 4.

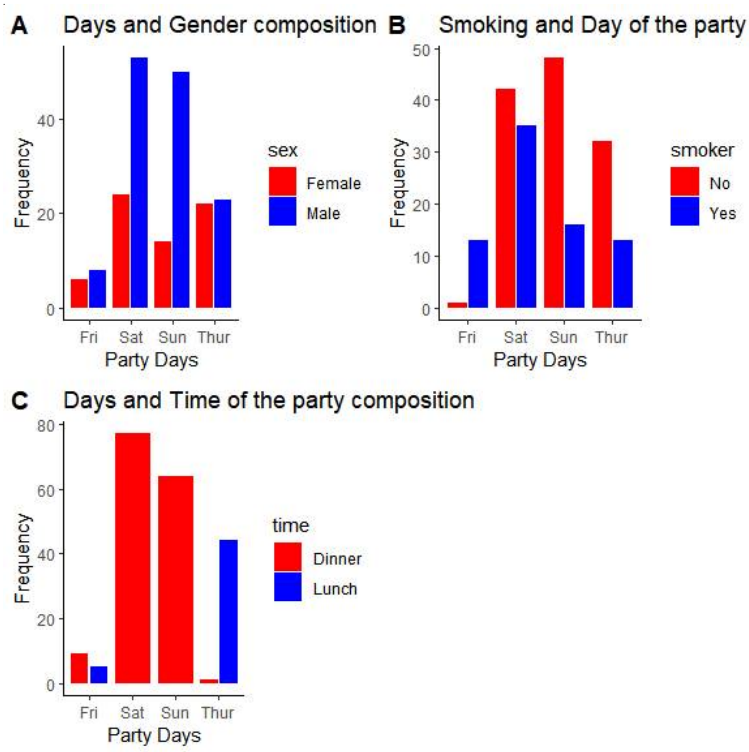

Figure 4: The figures above show the relationship between the day of the party and different variables like sex, time of the party, and the smoking status of the customer

In the graphs above,

Graph A shows that men's attendance was quite high compared to that of females on every day of the week. Over the days of the week, non-smokers were many apart from only Friday which had more of the smokers than the non-smokers and this is demonstrated by Graph B. Graph C shows that they were no parties organized during lunchtimes on Sunday and Saturday. On Thursday parties were organized more during lunchtime than the dinner time and Friday dinner had more parties.

3.2. Model Training

The dataset was split into two sets, i.e., training and testing sets in a ratio of 4:1 respectively (80% of the data for training and 20% for testing).

- Training set. This was used for model training.
- Testing set. Used to assess the performance of a model at predicting the amount of tip received by the waiter.

The linear model was built using the R programming environment with the help of its ***lm() function.***

The model obtained from the regression analysis is;

$$Tip\_amount = 0.77083 + 0.09192total\_bill - 0.05458sex + 0.15709smoker - 0.10900time + 0.15676size - 0.21973daySat + 0.01615daySun + 0.1627dayThur \quad ...(5)$$

In Equation (5), the restaurant waiter receives $0.77083 as a tip whenever a null hypothesis for all the predictors (total_bill, sex, smoker, time, size, and day) cannot be rejected. This means when there is no relationship between the tip and all the predictors. The level of statistical significance of a predictor is often expressed as a *p*-value and it's between 0 and 1 (Berk, 2020). The smaller the *p*-value, the stronger the evidence that you should reject the null hypothesis. If the *p*-value is less than 0.05, then it's statistically significant. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct or at least a 95% confidence interval in the predictor. Therefore, we reject the null hypothesis and accept the alternative hypothesis.

Total bill is the most statistically significant predictor in all the variables with level three significance and 99.99% confidence interval. For other variables is quite hard to reject the Null hypothesis due to high *p*-values. Predictors like smoker male and size also have relatively small *p*-values though still greater than the significance level of 0.05. And for the smoke predictor, the model shows statistical evidence of a difference in the average tip amount received by the waiter based on the smoking status of the customer.

On the other hand, sex, time, and day predictors have extremely large *p*-values and because of that, their null hypothesis cannot be rejected. The impact of these variables on the tip received by the restaurant waiters seems negligible. A change in these three variables don't impact on the tip amount. The model shows no statistical evidence of a difference in average tip received by the waiter between the days of the week but there exists some statistical evidence in the average tip given to the waiter based on the time of the party and the gender of who is paying the bill.

Table 1: Coefficients and *p*-values of the model predictors

Predictors	Coefficient estimate	*p*-Values
Total bill	0.09192	8.87e–11
SexMale	–0.05458	0.768
SmorkerYes	0.15709	0.436
TimeLunch	–0.10900	0.846
Size	0.15676	0.155
daySat	–0.21973	0.609
daySun	0.01615	0.972
dayThur	0.16276	0.752

Looking at Equation (5) and Table 1, we can conclude that the total bill is the most important tip predictor in the model followed by the size of the party.

The estimated predictor coefficients shown in the table indicate the increase in the tip amount, for example, a unit dollar increase on the total bill, increases the tip received by the waiter by $0.09191 keeping other predictors constant. Similarly, the size of the party increases the tip amount by $0.15676 whenever an additional person attends the party.

It's noticed from Table 1 that although the model was fitted on six (6) predictors, the model expression has eight (8) coefficients. This has happened because of the factors variables (sex, day, time, and smoker). The linear model has applied a dummy code to each factor predictor. For example, the smoker variable has two categories, Yes and No. The model has split this into smokerYes and smokerNo. The model is assigned 0 whenever the customer's smoking status is No and 1 for smokerYes. Sex has been split into sexMale and sexFemale and model assigned sexFemale with 0 and sexMale with 1. Similarly, the day variable is also split into dayLunch and dayDinner with assigned numerical values of 1 and 0 respectively. The four-level day predictor has been split into dayThur, dayFri, daySat and daySun.

Therefore we can say that if a bill is paid by a male customer, the waiter's tip decreases by $0.05458 and it neither increases nor decreases when a bill is paid by a female customer. And the amount of $0.15709 for every additional increase on the smoking customers and an additional increase on the non-smoking customers, the tip amount doesn't change. And whenever the bill is paid by a male customer, the waiter's tip decrease by $0.05458 than what he/she gets when the bill is paid by a female customer.

The time of the party also affects the average amount of tip received by the waiter in the restaurant whereby the tip decreases by $0.10900 whenever an additional party is organized during lunchtime and does not affect the tip amount received by the waiter for parties organized during lunchtime.

The overall effect of the day on the tip received by the waiter is ($0.01615 + $0.16276 – $0.21973) = –$0.04082. Though individual days on which the party is organized affect the tip amount independently. A party on Sunday increases the tip amount by $0.01615 and $0.16276 on Thursdays but decreases by $0.21973 for a unit increase in Saturday numbers. The model also that Friday seems not to affect the tip amount.

3.3. Model fitness

Table 2 illustrates how well the model fits the data.

Table 2 Shows more information about the least-squares of the model for the regression of the amount of tip receiver by the waiter

Quantity	Value
Residual standard error	1.042
R^2	0.5248
F-statistic	23.88
Adjusted R^2	0.5028
Overall *p*-value	0.5028

It's observed from the fitted model output which is summarized in Table 2 above that the model Residual Standard Error (RSE) is 1.042. This means that the actual tip received by the wait deviates by $1.042 from the true value, on average. And from this, we can say that the model doesn't fit the data quite well due to a relatively large RSE.

The R^2 value shows that there is 52.5% less variation around the model than the mean. And this indicates that the model did not explain approximately 47.5% of the variability in the tip response. The fitted model only explains 52.5% of the variability in the amount of tip received by the waiter which is still small.

The *F*-statistic of 23.88 is far from 1, which informs us that the null hypothesis (the situation where is there is no relationship between tip amount and its predictors is zero) of the model can be rejected. And based on this, we can conclude that at least one of the predictors is statistically significant in determining the amount of tip received by the waiter.

3.4. Model Testing

The model's performance was measured against the test set to monitor overfitting (Neal *et al.*, 2018). Overfitting (Briscoe and Feldman, 2011) is a scenario when a model tends to memorize patterns in the training set hence performing well when evaluated against the train set but poorly on the data it has not encountered.

We evaluated the model against the test set and accuracy metrics were analyzed. The RMSE of the model is 1.197048 and this indicates that the average deviation of the predicted tip from an actual tip received by the waiter is $1.197048. This error is extremely large making the model less suitable to determine how much tip a waiter receives.

The model's mean square of 1.432924 is quite large indicating also a bigger deviation from the actual amount of tip hence making the model less efficient in predicting the amount of tip that should be given to the waiter.

3.5. Final Model

The final model is,

$$Tip = 0.38277 + 0.09194totalbill + 0.26966size + 0.98489smoker - 0.45433(size * smoker)$$

Table 3: Coefficients and *p*-values of the final model

Predictors	Coefficient estimate	*p*-Values
Total bill	0.09194	1.8e–11
Smoker	0.98489	0.0455
Size	0.26966	0.0228
Size:smokerYes	–0.45443	0.0546

The inclusion of an interaction between size and smoker makes the smoker predictor statistically significant which is not the case when the interaction is excluded. And also the interaction between the two is statistically significant.

And also the size-smoker interaction increases the model R^2 value from 0.4832 to 0.4887 therefore making it a better interaction in the model.

From the table above, the total bill is the most important predictor of the tip amount and a unit increase on the total bill increases the tip amount by \$0.09194 and if smoking is allowed at the party the tip increases by \$0.98489, each additional person added on the party also increases the tip by \$0.26966 and finally if the additional person added on the party smokes, then tip amount decreases by \$0.45443.

The improved model gives a slightly smaller Root Mean Square Root of 1.181519 compared to that of the initial fitted model of 1.197048. This means that the final model's predictions deviate on average from the actual tip received by the waiter by \$1.181519. And therefore, we can conclude that the model *Tip = 0.38277 + 0.09194totalbill + 0.26966size + 0.98489smoker –0.45443(size**smoker) is more suitable and efficient for estimating the amount of tip given to the wait because it reduces the RMSE of the initial model by 0.015529 (1.55%).

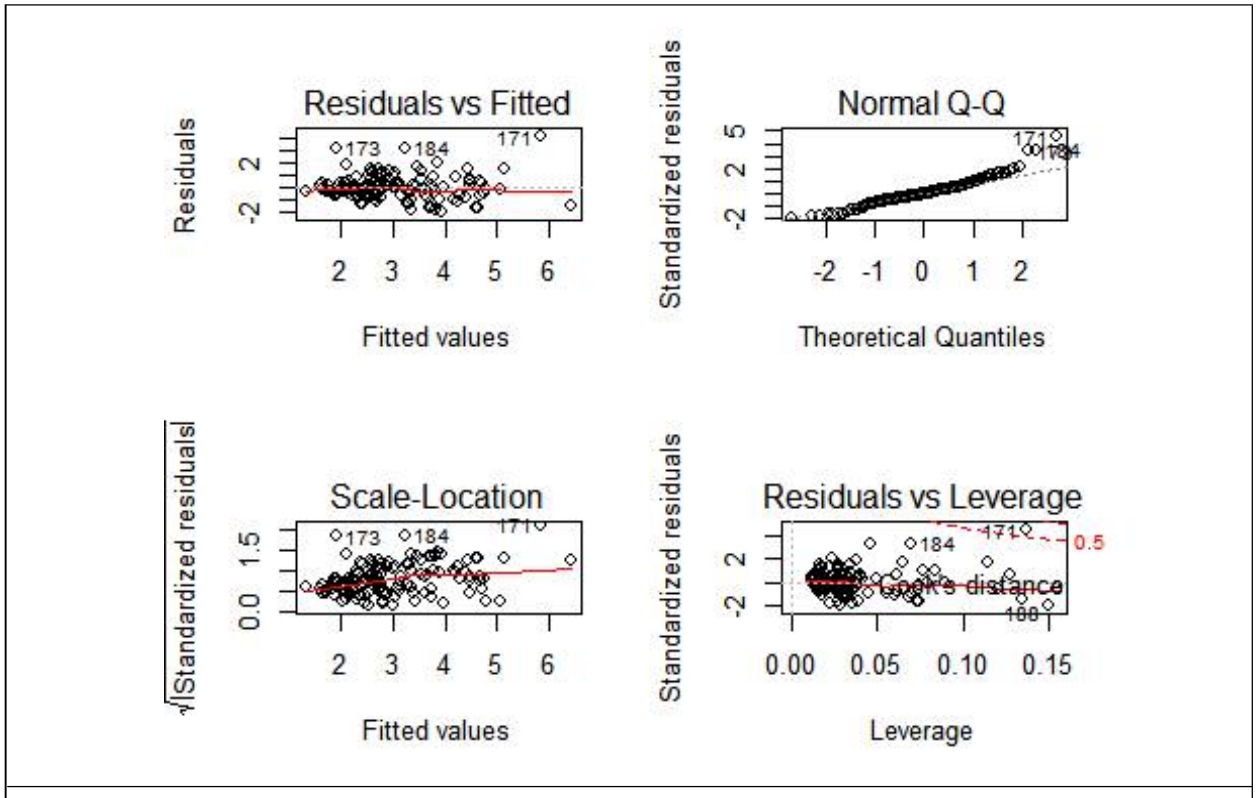

Figure 5: The plot of residuals versus fitted values shows that the variability is not the same across the board where the variance of the residuals increases with fitted values and this violates the assumption of constant variance

Checking Model Assumptions with Residual Plots: The residual plots below were used to evaluate the assumptions of normality of residuals, homoscedasticity, and independence (Jarque and Bera, 1980). These assumptions were made while fitting the model.

We can see from the Normal Quantile-Quantile (Q-Q) plot that most points do lay along the diagonal red line which means that residuals are approximately normally distributed and therefore this validates our use of the parametric method (i.e., Linear Regression model) in designing a model to predict the tip amount. Overall, from the above plots, we can say that there is a moderate linear relationship between the tip variable and the predictors.

In inclusion, the total bill paid by the customer is the most determinant of the tip received by waited followed by the size of the party. And the interaction between size and smoker predictors makes the model more efficient in estimating the tip amount.

4. Conclusion

- The total bill paid by the customer is the most influential variable in predicting the amount of rate received by the waiter. The best model for predicting the tip rate is *Tip = 0.38277 + 0.09194totalbill + 0.26966size + 0.98489smoker -0.45443(size**smoker). This means each additional dollar paid by the customer, the tip amount increases by $0.09194, increases by $0.26966 for each additional person at the dining party, and the rate does increase by $0.98479 for an additional smoking restaurant customer. The interaction between size and smoker decreases the tip rate by $0.45443.
- It was also observed that there is much variation in the amount of tip given to the by the waiter by smoking and the non-smoking customers. The attendance of male customers was higher than that of female counterparts on every day of the week.

References

Berk, R.A. (2020). *Statistical Learning from a Regression Perspective*. Springer International Publishing.

Briscoe, E. and Feldman, J. (2011). Conceptual complexity and the bias/variance tradeoff. *Cognition*, 118 (1), 2-16.

Doan, T. and Kalita, J. (2016). Selecting machine learning algorithms using regression models. Proc. - 15th IEEE Int. Conf. Data Min. Work. ICDMW 2015, pp. 1498–1505. doi: 10.1109/ICDMW.2015.43

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An introduction to statistical learning*, Vol. 112. Springer.

Jarque, C.M. and Bera, A.K. (1980) Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Econ. Lett.*, 6 (3), 255-259.

Kavitha, S., Varuna, S. and Ramya, R. (2017). A comparative analysis on linear regression and support vector regression. *Proc. 2016 Online Int. Conf. Green Eng. Technol. IC-GET 2016*. doi: 10.1109/GET.2016.7916627.

Kologlu, Y., Birinci, H., Kanalmaz, S.I. and Özyýlmaz, B. (2018). A multiple linear regression approach for estimating the market value of football players in forward position. *arXiv*, 1-12.

Lim, H. Il. (2019). A linear regression approach to modeling software characteristics for classifying similar software. *Proc. - Int. Comput. Softw. Appl. Conf.* 1, 942-943. doi: 10.1109/COMPSAC.2019.00152.

Neal, B. *et al.* (2018). A modern take on the bias-variance tradeoff in neural networks. *arXiv Prepr. arXiv1810.08591*.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

Zeng K (2016). Integration of machine learning and human learning for training optimization in robust linear regression Xiaohua Li , Yu Chen State University of New York at Binghamton Department of ECE , Binghamton , NY 13902, *Icassp*. 2613-2617

Cite this article as: Alex Mirugwe (2021). Restaurant tip prediction using linear regression. *International Journal of Data Science and Big Data Analytics*. 1(2), 31-38. doi: 10.51483/IJDSBDA.1.2.2021.31-38.