Can Predictions with R Help A Small Start-Up Company Increase Its Potential Sales?

Mircea Georgescu[1] and Ionuț-Daniel Anastasiei[2]

Abstract

ERP solutions that include prediction modules are very expensive. According to a survey (Ly, 2019), small to medium-sized businesses can expect to pay between $75,000 and $750,000 for the implementation. The R statistical language is easy to use when it comes to implementing regressions on the actual data of companies, and the costs are nearly zero. Logistic regression was the statistical formula used for predictions and it intends to create a model to predict the probability of buying a product based on the annual income of a costumer. To make these concepts easier to explain in this article, we considered a "*toy problem*" where you only have one customer characteristic (the customer's annual income) and a data scientist from a small company that wants to predict if the customer will make the purchase. This matter can be expanded in future studies that can make predictions for multiple independent variables, either binomial or multinomial. Our article also accepts the vast importance of using digital marketing to reach potential customers; however, it is more important to predict a potential client's behavior in terms of whether or not they will buy our solution, so that the company is able to set its own expectations.

Keywords: R Predictions, Logistic Regression, Sales, Digital Marketing

JEL Codes : C87, D47, D83

[1]**Accounting, Business Informatics and Statistics** Department, *Alexandru Ioan Cuza University of Iași*, Iași, Romania. mirceag@uaic.ro
[2]**Accounting, Business Informatics and Statistics** Department, *Alexandru Ioan Cuza University of Iași*, Iași, Romania. ionut_daniel_anastasiei@yahoo.com

1. Introduction

There are many things that cause a small business to fail, but one of the main reasons is related to insufficient capital which can also be associated with improper planning (Chaney, 2016). These two issues can be associated as a sort of recipe for failure. When it comes to ERP (Enterprise Resource Planning) costs, which can be quite substantial, every businessman knows that the importance of ERP systems far outweighs the initial cost, time and effort involved in implementation if you choose the right solution (O'Shaughnessy, 2019). *But what can you do when you have much less money than you need in order to buy an ERP system? Should an entrepreneur completely ignore CRM (Customer Relationship Management)?*

We may never find the answer for that, but entrepreneurs should take in consideration other methods, such as Digital Marketing. *Why Digital Marketing?* Well, the answer can be found at Clutch (a survey company from the USA), which surveyed 501 digital marketers in businesses across the U.S. to discover how they use digital marketing (Herhold, 2018). Actually, the top three digital marketing channels that businesses are currently using are social media marketing (81%), a website (78%) and email marketing (69%). And this is not all there is! There are multiple things to consider, especially for a new firm opened in 2019, as you cannot possibly sustain a business without taking those three digital marketing channels into consideration. We must admit that Digital Marketing is not an option, it is mandatory for any business (Bhuiyah, 2017).

Opening a business page on Facebook or Twitter is free of charge, but it does entail some costs when you want to sponsor certain campaigns. It is very cheap to create a website, because there are many third-party companies that allow you to build your own website with minimal costs, especially when you are not familiar with HTML, CSS or JavaScript web design. Also, there are free email marketing platforms that can help you reach out to old or current customers or to potential leads.

The first thought would be that if we have a small company we really do not need to pay thousands of dollars for ERP solutions, but it is not all that simple. Anyway, this subject can be included in a future analysis for another study, as this study assumes that the company for which the study was conducted already has a customer portfolio in its database and that email channels are currently in use.

2. Literature review

ERP solutions allow companies of all sizes to support key business processes by leveraging virtualization. The implementation of cloud ERPs is not straightforward and there are many issues that need to be taken into consideration when launching am ERP solution, one primary issue being the costs (Uri Sørhell, V., *et al.*, 2018). The most popular companies in the ERP systems' market are SAP (a German company with customers in more than 190 countries and a 20.8 billion Euro annual turnover in 2015) (SAP Company, 2016) and Oracle (an US-based company, also known for their database managements systems, which has more than 420,000 customers and a current annual turnover of 37 billion Euros) (Oracle, 2016). Basically, we can see that the market is led mainly by big companies, which can be quite intimidating.

ERP implementation may differ from any traditional systems' implementations in terms of project costs and the need for business process re-engineering (Somers, T., *et al.*, 2001). The percentage of ERP implementation failures exceeds 60%, and half of the top 10 failures stem from market leading ERP vendors, such as the ones mentioned above (Morris, M.G., Venkatesh, V., 2010). This means that implementation success can be quite intriguing for any entrepreneur and that success is not a guarantee. This begs the question: *Should a small business try to create a mini-project from a small ERP process with nearly no costs?*

This article may not provide a complete answer, due to its theoretical limitations, but the implemented model can help any businessman make predictions in R.

The model selected for this study is based on the logistic regression formula. If the actual data does not meet the assumed conditions of the model or if it has a significant error, then it is not feasible. This is mainly because a default distribution, such as the normal distribution for the response variable or the linearity of the proposed relationship for error variance, is considered to be among the limitations of certain classical methods (Sedehi, M., *et al.*, 2010).

This is why the advantages of using the logistic regression model, in addition to observations' modeling and the predicted probability of each person belonging to each level of the dependent variable, can help us discover the possibility to directly calculate the probability ratio of using the coefficients of the model.

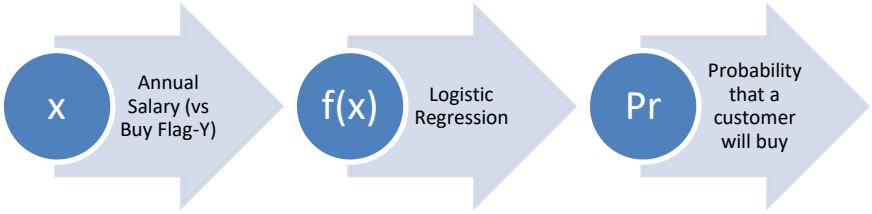
The model is mainly comprised of two different variables. The dependent variable is the one being tested and it is called "dependent" because it depends on the independent variable. The other one is the independent variable, which is the one you change or control in an experiment (Helmenstine, 2019). Those two are very easy to interpret and almost any person with analytics skills can apply this in the context of statistical formulas.

3. Research methodology

Before getting into statistical details, this study was conducted using the RStudio software, which supports any statistical analysis and prediction formulas. Our study also uses the *ggplot*, which is an absolute representation of the normal distribution of observations taken into consideration. Using that plot, we can determine whether the customers that purchased the product tended to earn higher incomes or not and, similarly, if the customers that did not purchase our product tended to have lower incomes.

The analysis contains a simple model to predict if a customer is going to buy a product after receiving an email, due to a marketing campaign. First, we must explain the context and the basis of the model that we are going to use in our example. There is a variety of formulas that can be used for a prediction, the simplest one being linear regression. Unfortunately, as simple as linear regression is, it is very difficult to meet all the statistical assumptions of that formula (Solutions, 2019), which is why we are choosing the logistic regression.

Figure 1. Model of regression used to predict if a customer will buy our product



Source: https://www.statisticssolutions.com/assumptions-of-linear-regression/

As shown in the figure above, our model (or function) will take the characteristics of a customer (in our case that is the annual income) and the marketing campaign to predict if the customer will buy via the logistic regression.

The main hypothesis is that *a small start-up company can predict using logistic regressions if a customer is going to buy a product or not, based on the email marketing campaign targeting the customer portfolio.*

3.1. Data collection

The company for which the study has been conducted is currently part of a *toy problem* obtained from Kaggle. This decision was made as it is mandatory to make these concepts easier to explain in order to accept the hypothesis. Also, the data consists of 859 observations and 15 variables.

The most important variables are *BD*, which is a binomial variable, and *Income*, which is the independent variable. These two will be used further on in our predictions.

In many real cases, this kind of categories are often used to show if certain clients do or do not buy a product, based on how much they earn, on their sex, wealth, etc. Of course, you can add other columns to this formula in order to build a multiple regression, but in our study we considered the logistic regression after the ETL (Extract-Transform-Load) part was concluded. Our data comprises 859 customers, of which only 309 have bought a product, while the rest were contacted via email, but did not buy anything.

3.2. Instrument Design

The instrument design considers that our example has two variables:

- **Y** or the so-called responding variable, that is the binomial variable (buy flag);
- **X** or the so-called manipulated variable, in our case that is the annual income.

$$\Pr(Y = 1 \mid X = x) = \frac{e^{a+bx}}{1+e^{a+bx}} \quad (1)$$

The logistic regression equation can be seen above (1) and it is used to calculate the predicted probabilities in our study. The variables of the formula consist of:

- Pr is the probability;
- Y = 1, if the customer will buy (or not = 0);
- X = x, yearly income (=x);
- a is the y intercept of the line;
- b is the slope of the line.

$$b = r\,\frac{S_y}{S_x} \quad (2)$$

The equation for slope (2) also takes in consideration the following variables:

- r = correlation;
- S_y = standard deviation of Y;
- S_x = standard deviation of X.

$$a = M_y - bM_x \quad (3)$$

The equation for the intercept of Y (3) also takes in consideration the following variables:

- M_y= mean of y;
- M_x= mean of x.

Once the values of coefficients "*a"* and *"b"* are obtained (R can do this automatically), the model can then predict a customer's probability to buy a product by substituting their corresponding annual salary.

In our case, the model takes a 0.5 cut-off value. For customers that bought the product, the predicted probability of buying has to be above the cut-off value (0.5), therefore the prediction is that they will buy.

In R, the equations from 1, *2 and 3* are translated as follows:

Table 1. The translation of equations in R language

Equation	Statistic equations translated into R
Logistic regression	model = glm(formula = BD ~ Salary, data = sales, family = "binomial") Prob = predict(model, newdata = sales, type = "response")
Slope	a = coef(model)["(Intercept)"]
Intercept	b = coef(model)["Salary"]
Manual prediction	pred_logit(a, b, x) new_vals = data.frame(Salary = x) predict(model, newdata = new_vals, type = "response")

Implementation of the cut-off	cutoff = 0.5 #Cutoff for probability
	sales = sales %>%
	cbind(Prob) %>%
	mutate(Prediction = ifelse(Prob > cutoff, 1, 0))

Source: Author Translation of Equation in R - Output

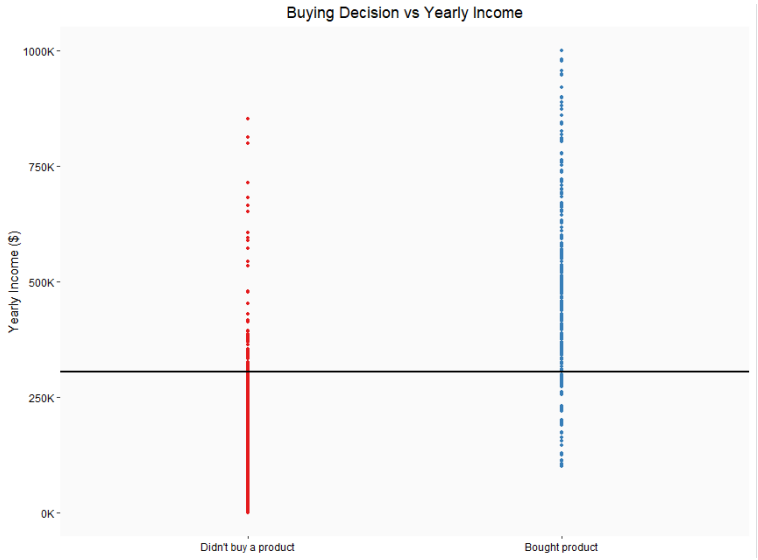
Next to the formulas translated into R, our study has some charts that were also created in Rstudio. The syntax for those charts is very long, but they are based on the *ggplot* command.

4. Results and Discussions

4.1. Predicting if a customer will or will not buy

The first plot obtained in RStudio can be seen in *figure 2*. The conclusion is that customers who bought the product tend to have a higher income. The black line is the mean of the income (salary) and we can see it as the turning point in our data set. Similarly, customers who did not buy the product tend to have a lower income, i.e. less than $300K per year.

Figure 2. Buying Decision vs. Yearly Income

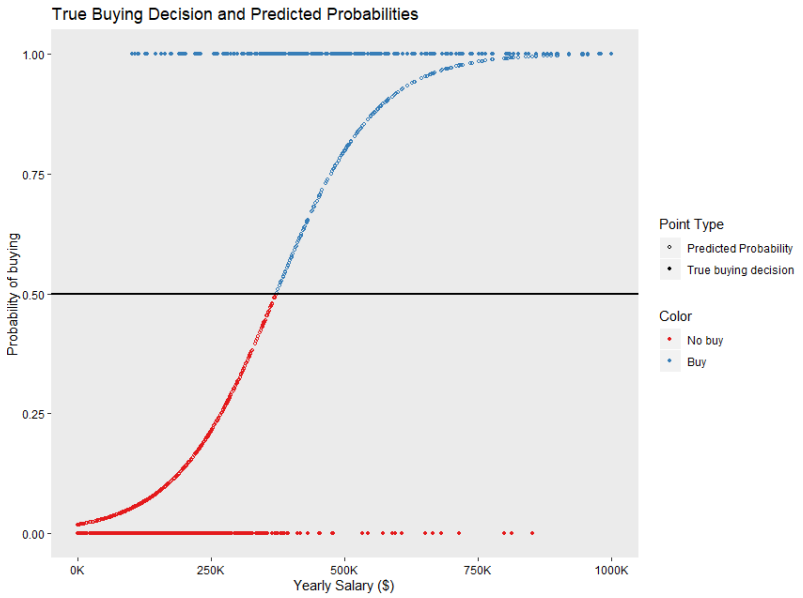


Source: RStudio Output for the used database

Thus our data set has a normal distribution and the prediction is the next step in the analysis of this study.

As established, the cut-off for the analysis is 0.5, which means that all scores higher than this number can convert our potential customers. The red dots in *figure 3* are nothing but the scores lower than 0.5.

Figure 3. True buying decision and Predicted Probabilities



Source: RStudio Output for the used database

To see how the points in the plot were created, we have to look at an example of how the formula works in RStudio. The random example was the amount of 200,000 applied in the regression formula, next to the slope and the intercept (4).

$$\Pr(Y = 1 \mid X = 150{,}000) = \frac{e^{-3.96+1.07(200{,}000)}}{e^{-3.96+1.07(200{,}000)}} = 0.14 \qquad (4)$$

Strictly in this example we can predict that a customer who earns $ 200,000 has a lower probability than the cut-off, therefore the company should not focus on customers with this amount of income.

4.2. Hypothesis Testing

The decision on whether there is any significant relationship between the independent variable *Y* and the dependent variable *X* can be made based on the logistic regression equation. The chi-square test tells if the null hypothesis is valid, then *X* is statistically insignificant in our regression model.

In order to measure the dependency relation between the variables, the significance level should be not higher than *0.05*.

The *glm* function applied to a formula describes whether the customer did or did not buy a product based on the annual income. This creates a generalized linear model, a so-called *glm*, in the binomial family. The summary can be printed out in RStudio and the check-up of p-values can be sorted out without SPSS. As the p-values of the annual income are less than 0.05, our model is significant in the logistic regression model as seen in *figure 4* below. There is a 95% confidence interval that our model is statistically significant.

Figure 4. Summary result of the GLM model in RStudio

```
Call:
glm(formula = BD ~ Salary, family = "binomial", data = sales)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2028  -0.5628  -0.3360   0.4807   2.4238

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.963e+00  2.464e-01  -16.08   <2e-16 ***
Salary       1.067e-05  7.200e-07   14.82   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1122.30  on 858  degrees of freedom
Residual deviance:  655.86  on 857  degrees of freedom
AIC: 659.86

Number of Fisher Scoring iterations: 5
```

Source: RStudio Output for the used database

The research results showed how statistical equations can help any business predict customer behavior, as well as how the combination with R language can reduce costs and also increase sales. The primary findings can be summarized as follows:

- The logistic regression formula is easy to implement in R, because any calculation is performed in the background. The only task of the person in charge is to make sure they select the correct variable;
- In fact, predictions in R can help small businesses increase their potential sales if the next emailing campaign is carried out according to the results of the first analysis. For example, in our study the entrepreneur should focus on customers with higher incomes, which can be viewed as the target market;
- Rstudio can also help test out the predictions' confidence interval, which can be easily obtained;
- Finally, an expensive ERP system is not always necessary for small businesses. Of course, this can be viewed as a matter for further study, such as a comparison between Rstudio and SAP, Oracle or Tableau. PowerBI also works very well with R and it is fairly affordable alongside the Office Suite.

5. Conclusions

In fact, the direction of this research is entirely based on the rest of the process that has to be taken in consideration. The main issue is the class imbalance problem. This happens when the relative frequency of a particular class (which in our case is comprised of the customers who bought the product) is low compared to the other class (the customers who did not buy the product).

There are many scenarios where this could happen, but in the context of digital marketing, the only problem is that the click rate is comprised only of a small proportion of the actual customers. This may affect predictions if the ETL processes are not taken into consideration.

As a general conclusion of this study, the results obtained herein reveal that small businesses can adapt to the latest market assumptions without expending substantial costs. This is indeed only a small part of a larger process within a company, and there are still a lot of things to take in consideration.

References

Bhuiyah, P. (2017, November 27). *Digital marketing is not an option in 2018 it's mandatory for any business!* Retrieved from Microsoft Partner Community: https://www.microsoftpartnercommunity.com/t5/Scale-Your-Business/Digital-marketing-is-not-an-option-in-2018-it-s-mandatory-for/td-p/2998

Chaney, P. (2016, July 28). *10 Reasons Small Companies Fail*. Retrieved from Small Business Trends: https://smallbiztrends.com/2016/07/small-companies-fail.html

Helmenstine, A. (2019, May 15). *DRY MIX Experiment Variables Acronym*. Retrieved from ThoughtCo: www.thoughtco.com/dry-mix-experimental-variables-acronym-609095

Herhold, K. (2018 , July 18). *How Businesses Use Digital Marketing in 2018*. Retrieved from Clutch. Firms that deliver: https://clutch.co/agencies/digital-marketing/resources/how-businesses-use-digital-marketing-2018

Ly, A. (2019, June 6). *How Much Does an ERP System Cost? 2019 Pricing Guide*. Retrieved from Better Buys: https://www.betterbuys.com/erp/erp-pricing-guide/

Morris, M.G., Venkatesh, V. (2010). Job characteristics and job satisfaction: understanding the role of enterprise resource planning system implementation. *MIS Quarterly, Vol. 34, No. 1 (March 2010)*, 143-161.

Oracle. (2016). *Oracle, Oracle Fact Sheet: Empowering and Accelerating the Modern Business*. Retrieved from Audientia Gestion: http://www.audentia-gestion.fr/oracle/oracle-fact-sheet-079219.pdf

O'Shaughnessy, K. (2019). *8 Reasons Why ERP Systems are Important in 2020*. Retrieved from Select Hub: https://selecthub.com/enterprise-resource-planning/why-erp-systems-are-important/

SAP Company, S. (2016). *About SAP*. Retrieved from SAP-Global Company Information: http://go.sap.com/corporate/en/company.fast-facts.html

Sedehi, M., Mehrabi, Y., Kazemnejad, A., Hadaegh, F. (2010). Comparison of artificial neural network, logistic regression and discriminant analysis methods in prediction of metabolic syndrome. *Iranian Journal of Endocrinology and Metabolism (6)*, 11.

Solutions, S. (n.d.). *Assumptions of Linear Regression*. Retrieved from Statistics Solutions. Advancement Through Clarity: https://www.statisticssolutions.com/assumptions-of-linear-regression/

Somers, T., M., Nelson, K. (2001). The impact of critical success factors across the stages of enterprise resource planning implementations. *Proceedings of the 34th Annual Hawaii International Conference, -6 Jan. 2001,*. System Sciences.

Uri Sørhell, V., Jørgensen Høvik, E., Hustad, E., Vassilakopouloua, P. (2018). Implementing cloud ERP solutions: a review of sociotechnical concerns Procedia Computer Science. *Procedia Computer Science, vol 138*, 470-477.