CAN PREDICTIONS WITH R HELP A SMALL START-UP COMPANY INCREASE ITS POTENTIAL SALES?

Mircea Radu GEORGESCU
Alexandru Ioan Cuza University of Iași
Iași, Romania
mirceag@uaic.ro

Ionuț-Daniel ANASTASIEI
Alexandru Ioan Cuza University of Iași
Iași, Romania
ionut.anastasiei@student.uaic.ro

Abstract: *The ERP solutions which include the predictions modules are very expensive and very hard to comprehend whenever a new company starts its activity. According to a survey (Ly, 2019), small to medium-sized businesses can expect to pay somewhere between $75,000 and $750,000 for implementation and this expenditure grows even larger for large businesses. Fortunately, there are some free tools available which can be used to implement a small part of an ETL professional process. Of course, we may have to admit that you also need some technical skills in order to learn how R, but also some statistical ones. The R language is very easy to use when it comes to implement regressions on the actual data of companies, and it comes at zero costs. Also, there is almost none ETL (extract-transform-load) technics needed because the client portfolio of small businesses is not large enough to be worth investing into. The statistical formula used for predictions was logistic regression and it intends to create a model to predict the probability of buying a product based on the yearly income of a costumer. To make these concepts easier to explain in this article, we have considered a toy problem where you only have one customer characteristic (the customer's yearly income) and a data scientist from a small company wants to predict if the customer will buy. This matter can be extended in future studies which can conduct the predictions of multiple independent variables, binomial or multinomial. Mainly, this article also admits that the use of digital marketing to reach the potential customers is very important, but more important is to predict the behaviour of a potential client whether it will buy or not our solution so that the company may set its own expectations.*

Keywords: *R, predictions, logistic regression, sales, digital marketing, ERP*

INTRODUCTION

There are many things that makes a small business to fail, but one of the main reasons is related to *insufficient capital* which can also be associated to *improper planning* (Chaney, 2016). These two issues can be associated like some sort of the ingredients of a fail recipe. When it comes to ERP (Enterprise Resource Planning) costs, which can be quite considerable, 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 you can do when you have much less money than you need in order to but an ERP system? Should an entrepreneur completely ignore the CRM (Customer Relationship Management)?

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

It is free to open a business page on Facebook or Twitter, but it comes with costs when you want to sponsor some campaigns. It is very cheap to create a website because there are lots of third-party companies that allows you to build a website at small costs, especially when you don't know HTML, CSS or JavaScript. Also, there are free email marketing platforms that can help you contact old customers, actual customers or potential leads.

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

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 and one main issue is the cost (Sorheller, 2017). The most popular companies in the market of ERP systems are SAP, which is a German company with customers in more than 190 countries and an annual turnover of 20,8 Billion Euro in 2015 (SAP, 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). Therefore, we can see that big firms mainly conduct the market, which can be quite intimidating.

ERP implementation may differ from any traditional systems implementations in project costs and the need for business process reengineering (Somers & Nelson, 2001). The percentage of ERP implementation failures is over 60%, and half of top-10 failures are from market leading ERP vendors, like the ones mentioned above (Morris & Venkatesh, 2010). This means that the success of the implementation can be quite intriguing for any entrepreneur and the success is not guaranteed. So, the question is: *should a small business try to create a mini project from a small ERP process, with almost zero costs?*

The answer may be not entirely be found in this article, due to theoretical limitations, but the model implemented can help any businessman build predictions in R.
The model chose for this study is based on the logistic regression formula. If the actual data does not have the assumed conditions of the model, then it is not feasible or alongside with a significant error. Mainly because a default distribution, like the normal distribution for response variable or the linearity of the proposed relationship of the variance of errors,

are among the limitations of some of the classical methods (Sedehi et al., 2010) (Amiri et al., 2018).

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 of the levels of the dependent variable, can help us find out the possibility of directly calculating the probabilities ratio to use the coefficients of the model.

Mainly, the model has 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 (Hermenstine, 2019). Those two are very easy to interpret and almost any person with analytics skills can apply this in the context of statistical formulas.

RESEARCH METHODOLOGY

Before getting into statistical details, the present study was conducted using the RStudio program which supports any statistical analysis and prediction formulas. Of course, in our study is also shown the *ggplot* which is an absolute representation of the normal distribution of observations taken in consideration. With that plot, we can see if the customers who bought the product tended to have a higher income or not and, similarly, if the customers who did not buy our product tend to have a lower income.

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 is linear regression. Unfortunately, as simple a linear regression is, as harder it is to fulfill all statistical assumption of that formula (Statistic Solutions, 2020) and this is the reason for choosing the logistic regression.

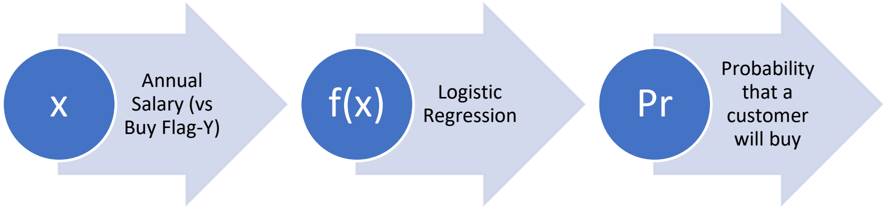

x
Annual Salary (vs Buy Flag-Y)
f(x)
Logistic Regression
Pr
Probability that a customer will buy

Figure 1. Model of regression followed to predict if a customer will buy our product (Mezquita, 2017)

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

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

DATA COLLECTION

The company for which the study has been conducted is currently a part of a *toy problem* obtained from Kaggle. This decision was made since it is mandatory to make these concepts easier to explain in order to accept the hypothesis. Also, the data consists has 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 further be used in our predictions.

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

Instrument design

The instrument design takes in consideration that our example has two variables:

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

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

Figure 2. **The logistic regression equation**

The equation can be seen in the figure above and it is used to calculate the predicted probabilities in our study. The formula's variables 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.

$$b = r\,\frac{S_y}{S_x}$$

Figure 3. The equation for slope

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

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

$$a = M_y - bM_x$$

Figure 4. The equation for the intercept

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

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

Once the values of the coefficients a and b are obtained (R can do this automatically), then the model can predict the probability of buying a product for a customer by substituting its corresponding yearly salary.

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

In R, the equations from *figure 2,3 and 4* are translated as follows:

Table 1. The translation of equations in R language

Equation	Statistic equations translated in 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 cutoff	cutoff = 0.5 #Cutoff for probability
	sales = sales %>%
	cbind(Prob) %>%
	mutate(Prediction = ifelse(Prob > cutoff, 1, 0))

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

RESULTS AND DISCUSSIONS

Predicting if a customer will buy or not

The first plot obtained in RStudio can be observed in *figure 5*. The conclusion is that the customers who bought the product tend to have a higher income. The black line represents the mean of the income salary and we can perceive it as the turning point in our data set. Similarly, the customers who did not buy the product tend to have a lower income, less than $300K per year.

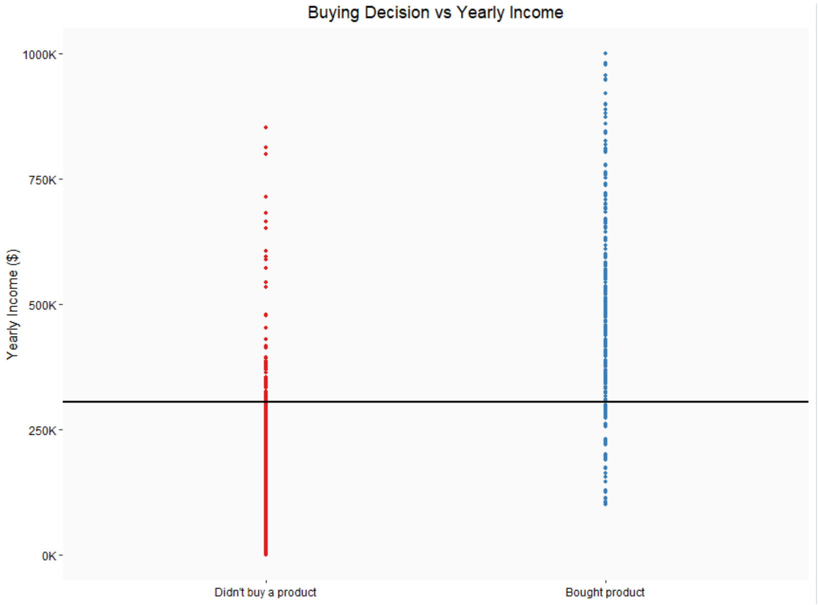


Figure 5. Buying Decision vs Yearly Income

Therefore, our data set has a normal distribution and the prediction is the next step in the study's analysis. As established, the cutoff of the analysis is 0.5; this means that all scores higher then this number can convert our potential customers. The red dots from the *figure 6* are nothing but the scores smaller than 0.5.

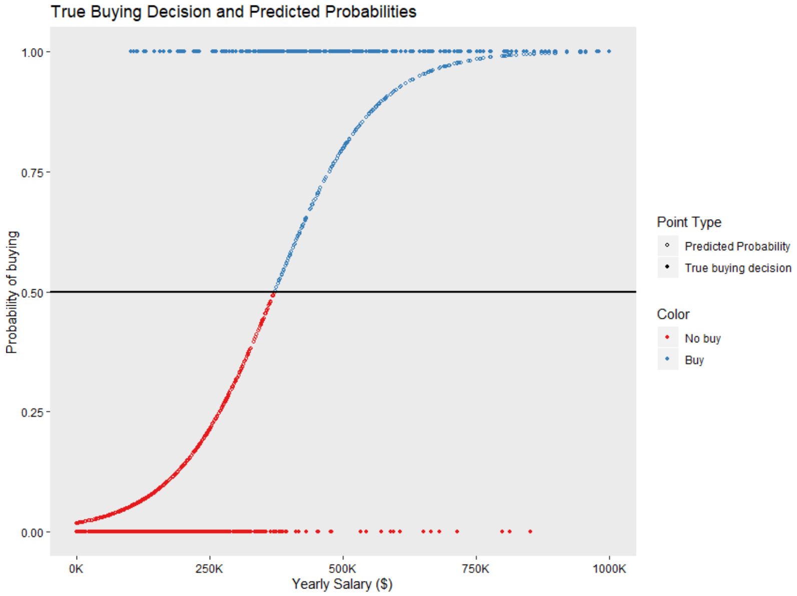


Figure 6. True buying decision and Predicted Probabilities

In order to see how the points from the plot were created, we must see an example of how the formula works in RStudio. The random example was the sum of 200,000 applied in the regression formula, next to the slope and the intercept.

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

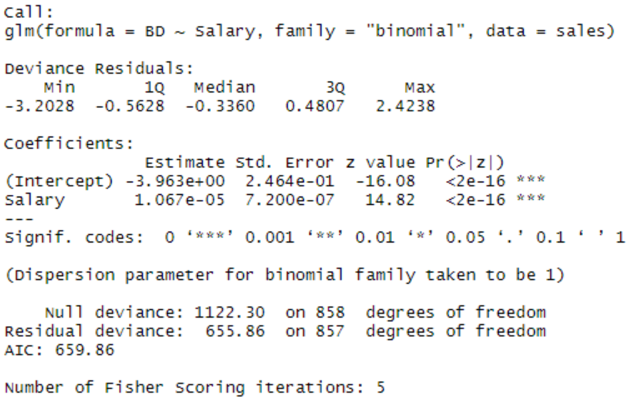
Figure 7. The formula applied for an income of $200,000

In this example only, we can predict that a customer who earns $200,000 has a probability lower than the cutoff; therefore, the company should not focus on customers with this amount of income.

Hypothesis Testing

The decision whether there is any significant relationship between the independent variable *Y* and the dependent variable *X* can be taken 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 that describes if the customer bought a product or not, by the annual income. This creates a generalized linear model, so called *glm*, in the binomial family. The summary can be printed out in RStudio and the check of the p-values can be sorted out without SPSS. As the p-values of the annual income is less than 0.05, then our model is significant in the logistic regression model as seen in the *figure 8* below. There is a 95% confidence level that our model is statistically significant.



```
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
```

Figure 8. Summary result of the GLM model in RStudio

RESULTS AND DISCUSSIONS

The results of the research showed how statistical equations can help any business predict the behaviour of the customer and also how the combination with R language can reduce the costs and also can increase its sales. The main findings can be summarized as follows:

- The logistic regression formula is very easy to be implemented in R because any calculation is being done in the background. The only task that the person in cause has is to make sure it selects the correct variable.
- In fact, the predictions in R can help a small business increase its potential sales if the next emailing campaign is taking place according to the results of the first analysis. For example, in our study, the entrepreneour should focus on the customers that have a higher income, which can be viewed as the targeted market.

- Rstudio can also help to test the predictions confidence level interval, which can be easily be obtained.
- Finally, the need of an expensive ERP system is not always needed for a small business. Of course, this can be seen as a future study, like a comparison between Rstudio and SAP, Oracle or Tableau. Also, PowerBI works very well with R and it is also accessible at a fair price, along with the Office package.

As a general conclusion of the study, the results obtained in this study reveal that small businesses can adapt to the latest assumptions of the market even without big costs. Certanly, this is only a small piece of an entire process of a company, there are still a lot of things to take in consideration by expanding the number of independent variables.

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 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 when this could happen, but in the context of digital marketing, the only problem is that the click rate is constituted from only a small proportion of the customers. This may affect the predictions if the ETL processes are not taken in consideration.

References

1. Amiri, M. J., Karami, P., Chichaklu, A.H., Jangan, E.H., Amiri, M.J., Owrang, M., Khaledi, A., (2018). *Identification and isolation of Mycobacterium tuberculosis from Iranian patients with recurrent TB using different staining methods.* J Res Med Dent Sci, 6 (2) (2018), pp. 409-414
2. Anonymous, (2016). *AG S. SAP Company Information | About SAP*, 2016. Retrieved from http://go.sap.com/corporate/en/company.fast-facts.html, 2017.
3. Anonymous, (2016). *Oracle. Oracle Fact Sheet: Empowering and Accelerating the Modern Business 2016.* Retrieved from http://www.audentia-gestion.fr/oracle/oracle-fact-sheet-079219.pdf
4. Anonymous, (2020). *Assumptions of Linear Regression,* Retrieved from https://www.statisticssolutions.com/assumptions-of-linear-regression/
5. Bhuiyah, P. (2017). *Digital marketing is not an option in 2018 it's mandatory for any business!* Retrieved from https://www.microsoftpartnercommunity.com/t5/Scale-Your-Business/Digital-marketing-is-not-an-option-in-2018-it-s-mandatory-for/td-p/2998
6. Chaney, P., (2016). *10 Reasons Small Companies Fail.* Retrieved from https://smallbiztrends.com/2016/07/small-companies-fail.html
7. Helmenstine, A. M., (2019). *DRY MIX Experiment Variables Acronym.* Retrieved from https://thoughtco.com/dry-mix-experimental-variables-acronym-609095.
8. Herhold, K., (2018). *How Businesses Use Digital Marketing in 2018.* Retrieved from https://clutch.co/agencies/digital-marketing/resources/how-businesses-use-digital-marketing-2018
9. Ly, A., (2008). *How Much Does an ERP System Cost? 2019 Pricing Guide.* Retrieved from https://www.betterbuys.com/erp/erp-pricing-guide/
10. Morris M. G., Venkatesh V., (2010). *Job characteristics and job satisfaction: understanding the role of enterprise resource planing system implementation.* MIS Quarterly. 2010:143-161.
11. O'Shaughnessy, K., (2019). *8 Reasons Why ERP Systems are Important in 2020.* Retrieved from https://selecthub.com/enterprise-resource-planning/why-erp-systems-are-important/
12. 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), p. 11

13. Sorheller, V., Hovik, E., Vassilakopoulou, P., (2017). *Implementing cloud ERP solutions: a review of sociotechnical concerns* Procedia Computer Science, Volume 138, 2018, Pages 470-477
14. Somers T., M., Nelson K., (2001). *The impact of critical success factors across the stages of enterprise resource planning implementations.* Paper presented at: System Sciences, 2001. Proceedings of the 34th Annual Hawaii International Conference on; 3-6 Jan. 2001.
15. Mezquita, O., (2017). *Using R to predict if a customer will buy.* Retrieved from https://www.masterdataanalysis.com/r/using-r-predict-customer-will-buy/