APPLICATION OF LOGISTIC REGRESSION MODELS IN RISK MANAGEMENT

Arun Velu
Department: Advanced Analytics, Global Consumer Solutions, Equifax
Affiliation: Equifax Inc.
Country: United States

ABSTRACT

Logistic regression is a technique that uses statistics to develop a prediction model on any occurrence that is binary in itself and its nature (Ahmad et al., 2021). When it comes to a binary event, it may either occur or not occur. A binary nature has only two results which are represented in form of 0 (non-occurrence) and 1 (occurrence). Another application for logistic regression is where there are more than two classifications on the dependent variable. Logistic regression can be binary when the classification of the dependent variable is in two groups while it could be multinomial when the dependent variable is two groups or more. Predictive modeling is a technique where the known results are taken and develop a model that will help in predicting the later activities and occurrence (Midi et al., 2010). It uses ancient data to predict events that will occur in the future. Predictive modeling comes in different types which are ANOVA, logistic regression, decision trees, time series, neutral networks, linear regression and ridge regression. It is very critical in selecting the right model for regression to save on time in a project. Selecting incorrect modeling may result in the synthesis of a wrong prediction and non constant mean and varying variances (Hosmer et al., 2013) Variances should be constant and not varying. Consequently, regression analysis predicts continuous variance target from more than one independent variable. Regression analysis utilizes the natural variance and not a variance that have gone through experiments since they are manipulated and cannot produce the correct result. Predictive churn model is used to explain customer churn or a customer stepping down on a product or a service. This model provides quantifiable matrices and alertness to fight the retention effort (Harrell 2015). The probable monthly churn; the number of active users who churned in divided by total number of active user days, this will provide the number of churns in every user day. I hope this study will educate practitioners in the estimation of the independent variable and determining the risk factors. It's helpful because probability produces results for independent variable and result variable in multiple models and/or binary events. When the result of two or more variable is established, it becomes easy to understand and organize a business examination for decision-making.

Keywords: Logistic regression, predictive modeling, regression analysis, ANOVA, predictive churn, dependent variable and independent variable

INTRODUCTION

In the current world, data storage has become quite cheap hence telecommunication companies use different ways of accessing the information which will be beneficial to the company's operation more so pertaining customer analysis. The combination of traditional data stored and the unstructured data in social media which comprises of phone call recordings and social feedback helps in providing reliable information for companies. The use of algorithms such as neutral network, decision tree and logistic regression utilizes the predictive model created by the data (Norton et al 2019). In a telecommunication company, the predictive model will help to locate and identify customers who may be at risk of churn. After identification of the churn score the company may come up with incentives to keep the customers. Despite the fact that the model will provide a reliable data, it is not good to establish who will churn and causes of the churn. This knowledge will help the company to establish a campaign that is aimed at retaining the customer and improving the whole service.

The purpose of this research paper is to predict the reasons for customer churn in American telecommunication companies utilizing logistic regression. The estimates and the prediction figures help in establishing the factors that accelerate customer churn in American telecommunication companies. Another purpose of this paper is to ascertain and develop the quality of the logistic regression models and the analysis of the predictions. It also confirms how logistic regression can be sustained on predictions based on customer churn.

The theoretical background of customer churn, dimensions, its types, its benefits and consequences are explained in the introductory section of this research article. The methodological section contains the definition and explanation of logistic regression, method used to estimate beta coefficient, Hosmer Lemeshow test which is a statistical test, performance statistic, ROC curve and precision. The variables used in the creation of the model are introduced. Some graphical analysis is explored and the estimated result is produced with performance metrices.

LITERATURE REVIEW

There are many existing literatures that explain the predictive models and how the result affects customer churn. From the recent literature, most telecommunication companies lose their customers due to the way they provide their services. Customer loyalty is the most critical part of the growth of the telecommunication industry and other related industries (Kleinbaum et al., 2002). Establishment of this data through the most reliable source helps to prevent customer churn (Wright 1995). Earlier researches have analyzed customer characteristics by the use of regression models. Other research articles utilize clustering to put together and group customers with same characteristics. On this research paper “a review of telecommunication customer churn using a predictive model” this is where we use risk management and predictive modeling to formulate a result. Most telecommunication companies use this model to establish customer behavior and operation of the company (Tranmer and Elliot 2008). Secondly, this review will help companies develop and provide incentives that will help in retaining the loyal customers and the potential customers. Recent researches have ascertained that most customers prefer good service and proper handling from the service providers.

This research article aims at correcting the norm that hinders most customers from being loyal to their service providers. The data will in turn reciprocate from the company to the customers to the potential customers. In this paper, I am going to address in-depth on telecommunication industry and how this literature will help American companies reduce customer churn (Bender and Grouven 1998). Churn predictive technique should be part of a company’s operation as a mirror to improve the customer loyalty and prevent possible exit of the customers.

CONTENT

Logistic regression

Logistic regression is categorized in a model class called generalized linear model. The purpose of this class generalized linear model is to come up with the regression equation which provides for y which is a dependent variable or other variables that are denoted x. This establishment depends on the dependent binary variable. A more aggressive model is to denote y as a linear function of x. Linear regression does not show how x and y are related (Nemes et al., 2010). In other cases, this function can come up with predictions that may be out of the expected range of 0-1. A more reliable form is a nonlinear function which produces a regression model which has a coefficient that is linear and it can change the entire prediction to a range of 0-1. The functions are called generalized linear link functions. Logit and Probit are the main and common link functions in a logistic regression in conducting binary survey (Kain and Verma 2018). The Logit can be put in a model linearly in a regression model.

In [pie(x)/1-pie (x)] =B0 +B1 X1 +….Bk Xk

Where pie (x) represents a conditional probability and y = 1 provided with covariant vector x. B1, B0, Bk are estimated regression coefficient of logit X1, X2, and Xk.

The left side represents the Logit. After the establishment of a coefficient, it then tests the purpose of the explanatory variable. Wald test is used to test the significance of the statistical data.

Z= B1 /SE (B1)

Where SE (B1) is an approximated standard error of the estimated regression co efficient B1. Z is squired to produce a Wald methodology.

The Hosmer-Lemeshow is useful to fit the test and shows how critical the data is and how it blends with the model.

$$X^2_{HL} = \sum_{j=1}^{M} \frac{(O_j-E_j)^2}{E_j(1-E_j/n_j)}$$

Where Nj is the number of observation in the jth pattern, Oj is the number of observation cases, Ej is the expected cases

When classifying the task where the group that we want to predict does not occur regularly, performance matrices can be used. Precision and sensitivity are very critical in any measure and predicts important charecteristics. The four groups of confusion matrices are

True positive, TP: class of interest

True negative, TN: no class of interest

False positive, FP: Incorrectly classified as class of interest

False negative FN: incorrectly classified as no class of interest

From an understanding of a business; sensitivity and predictive measures are the most common predictors.

Precision= TP/TP+FP

Lastly, the critical aspect is to put together all the input variables. After that data analysis should be conducted to know which data needs to be examined and analyzed to give required result. The model then undergoes quality test to prove what is provided is the real reliable information. This can be done through Wald test or Hosmor Lemeshow test (Schober and Vetter 2021). The end result is produced to the result measure that will ascertain if the performance is the one that is required.

Customer churn

Customer churn entails where customers want to stop using a particular subscription or a particular service or a product. For a company to grow there should be customer loyalty and a form of subscription to the company. In the current world there are many competitors and many businesses that strive to survive in the field. Particularly in the United States there are many innovations and technologies that prove to outshine the already existing businesses (Norton 2018). To compete in this environment, it is very critical to come up with a mechanism that will help know the ways that will keep customers in the company. This topic of customer churn is probably the most talked topic in subscription business. Current markets are very saturated and it is very difficult to attract new customers to the company. In the general market it is believed that acquiring a new customer is more difficult than retaining the already existing customers. Most companies prefer to invest in already existing customers than attracting new customers to the business because they come with a high rate of customer churn.

As a result of these mass exodus of customers, there are churn management methodologies which help to find the most important customers and the ones that are likely to churn; this classification will help the firm identify the right way and steps towards preventing exit of the customers. Churn rate is the measure of the number of customers who move out of the business over a particular time frame. Companies should have a clear understanding of the type of churn that comes as a result, voluntary and involuntary churn which may come as a result of the normal switch to another company. If companies can get the predictors and the right reason for the churn they come up with is retention strategies that will help keep the customers intact to the company (Liberman 2005). As mentioned earlier the main reasons for customer churn may be high costs, poor service delivery, in other cases voluntary process, taking a long time to solve a problem, reward for customer loyalty and private concerns. The other type of churn which is involuntary is where the company discontinues the contract due to various reasons. These reasons for discontinuity maybe nonpayment and fraudulent and exploitative customers who do not care about the needs of the company’s service.

In the current world, the cost of storing data has decreased which has prompted companies to look for alternative ways of acquiring data. These ways include the use of predictive measures that bring forth the statistical data that helps to rate the company and its operations. The application of the risk management in customer churn is very important because individual companies are able to prevent the tendency of customer churn and other occurrences that will reduce customer loyalty. With good churn management, companies in America can save a large amount of money. Churn management through identifying the right reasons for customer exit will help in modifying and saving customers and money to the company. Customer retention is very important because they will indeed bring other customers to the business and as a result it will help in the growth of the company’s operations. To distinguish between churners in a company and non churners, churn management team uses different types of classification which are clustering and association rule (Uanhoro et al 2019). Logistic regression and decision trees are a very critical aspect in determining the churn prediction. In many literatures, many writers have used logistic regression to determine churners in a telecommunication company. Customer age and location are some of the factors that are considered to know the main cause of customer churn. As a case study, the following study was carried out in an American telecommunication company, using the Wald test to establish customer churn in the company; the average length of calls and the discount package in the company were primary predictors. The result shows that call quality triggers customer churn in the company. As a result more often users are more likely to churn the company (Monahan et al 2007). In studies that are conducted outside America, for instance Nigeria, the annual churn is 41%. By the use of backward stepwise model in logistic regression, it was concluded that the main cause of the mass churn of the across all the companies are the type of service plan provided, providers service and the mobile connections. These factors cuts across all the companies and drive the churn.

Predictive Modeling Technique

This is a practice where the results of the known occurrence predict the values of the other unknown occurrence. This is applicable when a company or firm wants to predict its original predictive measure. This model uses ancient data to predict the later events. The different types of predictive modeling are linear regression, ANOVA, ridge regression, decision tree, logistic regression, time series and neutral networks.

Linear Regression

This is used when the variables are continuous, categorical and the interdependence between variables are linearly distributed - these are independent variable and the dependent variable. Predictor variable should be presented linearly and limited margin of multi-collinearity should be seen. This model is applicable when one wants to know the unknown predictor in a business or a company.

ANOVA

It is also called an analysis of the variances; the target variance should be continuous and the dependent variable is categorical. The null assumption of this model is that there is no difference in the groups. The population should be spread normally, in other cases, simple samples should be independent and variables should be closely distributed in the group. This is applicable in testing variables which are continuous and categorical.

Ridge Regression

This model helps to analyze multiple variables in regression which undergo multi collinearity. Ridge regression utilizes the square technique and adds degree of bias to reduce standard error during prediction. They assume that they follow the multiple regression model and the plots and points are spread linearly. The variables should be constant and should depict independence. This is applied to predict the multiple linear variables in a firm.

Logistic Regression

The logistic regression does not need linear interdependence between the dependent variable and the target variable. The target variable in most cases in this model is binary. Let’s take an example of (0 or 1); the error that may emerge should be distributed where the variance is not necessarily constant. Just as the target variance, the dependent variable is binary and the result should be independent and stands on its own (Martinez et al., 2017). The size should be big and a little multi-collinearity in the information obtained. In this model the independent variable should be related linearly to the log odds. This is applicable in the determination of non-related linear variances in risk management.

Decision Tree

This is a supervision tree in learning algorithm and asking simple question about simple things. This helps in classifying problems that may emerge as a result of simple decisions. It is general that most people find it hard to make decisions based on simple problems (Chen et al., 2019). Decision tree represents multiple kinds of decisions that are bound to happen. This model helps in establishing the most useful variable and their interdependence on each other. Simple decision making will lead to significant results.

Neutral Networks

This model helps in clustering and grouping data. They are algorithms that are modeled to be loose and to identify different patterns. This model is very complex compared to other models because they have many of algorithms (Moamer et al., 2019). This model is very important in determining complex algorithms and variances; however, it may not be ideal in determining why something happened in the variances.

Time Series Regression

Time series regression is a technique that helps to predict later responses based on the previous responses. Data for this are collected on observation over a period of time and more specifically over different time series. The series should be stationary; this means they are always distributed. Variance and mean do not change over a period of time. Residuals should be spread normally without changing the mean and the variances (Chang and Hoaglin 2017). Outliers should not be contained in the series; therefore, with a random distribution of variances, the mean should be 0 and no changing the variable as mentioned earlier. This model is very critical in a firm or company since it helps in the correlations and comparison of responses for decision making.

Obtaining Results and Discussion on Logistic Regression

This part of the research is aimed at application of two real datasets on logistic regression. This data consists of approximately 50,000 customers from an American telecommunication company. There were two datasets that were independent, data set for training along with the model classification and a test set of data that is used to test the predictive performance. Service usage pattern and demographic groups are input variables that explain this part vividly. Below are the steps that are followed to explain the logistic regression using a set of training data to arrive at a predictive data. The model is analyzed, examined and tested. Performance statistics are set to the testing data to develop the characteristics of the model on the unseen data.

Description of the data

Table 1: Demographic variables

Variable name	Description
Birth year	Customer’s year of birth
Delinquent	Did the customer fall on the delinquency spectrum
Duration	Customer lifetime
Acc type	Type of account
Contract duration	Days till the end of the contract
City	Location of the customer

Source: company database

The data that is used above was downloaded from the company’s database in a warehouse. The two datasets are for about than 50,000 customers. These samples were selected randomly on over a population of over 900 thousand customers. The first dataset was used to examine the logistic regression while the second dataset was used test the fits of the model. The two categories of data are the duration and the birth year which affects the dependent variable.

Nearly all the variables are self-explanatory. From the data in Table 1 the type of the account that a customer uses will greatly affect the predictability of the other variables.

Table 2: Service usage variables

Variable name	Description
AvgInvoice	Average amount of invoice for last year
avgExtra3	Average overpayment for last 3 months
avgExtra6	Average overpayment for last 6 months
avgExtra12	Average overpayment for last 12 months
AvgData	Average monthly data consumed (GBs)
AvgVoice	Average monthly voice consumed (mins)
avgSMS	Average monthly number of sent SMS
AvgMMS	Average monthly number of sent MMS
VAS	Whether the customer has currently value added services
Portout	Whether the customer left the company

Source: company database

The information was obtained from a database of one of the companies in America. All the variables are on average. The *AvgInvoice* for the company is described annually. *AvgExtra3* is described as overpayment in 3 months. This helps to establish the average payments of the company over a short period of time. It continues to 6 months and 12 months respectively. The average data consumed in a month is calculated in GBs. This will help to formulate the average of the people who access the services of the company. This in turn creates an average voice consumed in minutes and is measured on a monthly basis. The other variable is the SMS sent to the company in a particular month. The more the SMS the more engaging the company is to customers. The MMS also is calculated in monthly average. The relationship between these variables is very significant because it helps to create a predictor variable that will provide an insight to the company. VAS variable is independent because individual customers have different judgment of the service to the company. When customer service is improved, it will reduce customer churn.

Portout is a binary variable that is dependent and shows the number of people leaving the company in a 45 day period. As mentioned earlier, customer service delivery and extreme cost are the main causes of customer churn.

Table 3: Descriptive statistics of numerical variables

Variable	Min.	1st quartile	Medium	Mean	3rd quartile	Max
birthYear	1920	1950	1960	1965	1978	2002
contract Duration	-6800	-2580	-1271	-1690	-360	4860
avgInvoice	0.12	330.70	430.6	560.90	377.00	740.60
avgExtra12	0.01	9.35	45.30	90.19	670.85	6130.45
avgExtra6	0.00	6.00	46.1	100.00	23.16	6530.77
avgExtra3	0.00	2.00	40.11	100.87	130.16	9660.13
avgData	0.00	366.40	2500.5	10646.99	8961	60 600.34
avgVoice	0.00	0.00	0.00	48.15	62.00	2525.70
avgSms	0.00	0.00	0.00	2.88	0.00	590.55
avgMms	0.00	0.00	0.00	0.09	0.00	670
duration	15	1330	2112	2630	3835	7060

Source: company database

The above descriptive data was run on the analysis from the company's data and shows the (variable, minimum, 1st quartile, medium, mean, 3rd quartile, max) it explains how the binary variable affects the customer relation and in other occasions it triggers customer churn. The variable describes the descriptive elements. The minimum cannot relate to each other as the binary prediction starts from the minimum statistic. The first quartile and the third quartile differ slightly since it its binary variance relationship with the variable is almost equal.

Description of Estimation model

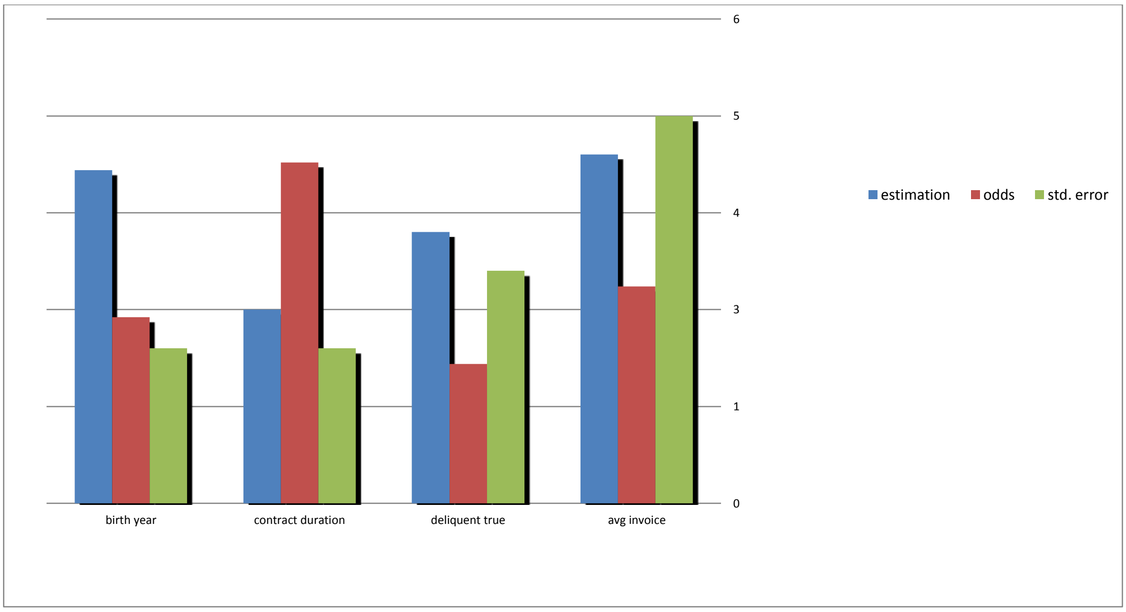
The results of the estimation model - regression coefficient dependent on odd ratio, the value of the Wald test result, standard error and p value are described in the table below. In light of the regression coefficient, high numeric value increases the possibility of a customer churn. Average birth year, average SMS and average data are the most critical parts of the estimation model. Delinquent customers influence a big chance to churn as compared to non-delinquent customers.

We can pose an assumption that slightly younger people who use SMS and mobile data to access the company's website and services are more likely to churn from the company. Again, we can infer that students who use account (B) have 2.56 odds which mean they have a higher probability of leaving the company.

Higher value of the numeric binary variable reduces the probability of customer churn. Companies should always focus on retention of customers to ensure continuity of the business. Customers that hold personal account are likely to churn the company as compared to those with family account.

Table 5: Estimation results

Variable	Estimation	Odds	Std. error	Z value	P value
Intercept	-33.760	0.00	8.567	–3.941	0.000
Birth year	0.017	1.017	0.004	3.839	0.000
Contract duration	-0.004	0.996	0.000	–29.371	< 2e-16
Delinquent true	0.961	2.614	0.262	3.665	0.000
Acc type B	0.962	2.525	0.488	1.899	0.058
Acc type D	-15.345	0.00	1628	–0.009	0.992
Avg Invoice	–2.64e-05	1.00	1.32e-04	–0.200	0.841



Hosmer-Lemeshow Test in Estimation

This test was the first to be used to examine the effectiveness of the predictive model. Hosmer-lemeshow approved the use of the measuring parameter g (subgroups members) as k+1 as the variants number. Lemeshow then dedicated the measure to 18. Test statistical data use x square distribution in which g+2 =16 level of freedom. P value is equivalent to 2.2e-16 which has a low level alpha of 0.05. From the above formula all the observations of the churn and predictive measure are equal across all the groups predicted. Large data groups may cause a negative result or it may be caused by the nonlinearity of the model.

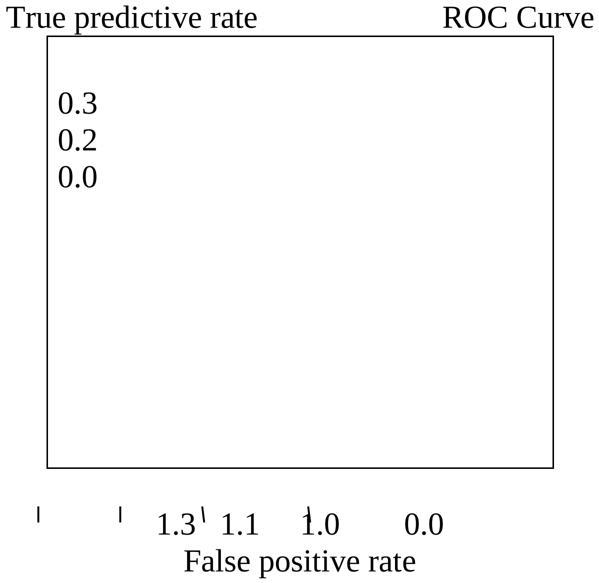
R Square Method is another alternative to the formulae to examine the fitness of the model. It should be noted the R square is only used to assess the legibility and the fitness of the model. It can be very fit to the same group and set of data. Low values of pseudo R squire are compared to the R squire values and the difference should not be understood as not reliable model.

Table 6: Pseudo R squared Matrix

Pseudo R squared	Value
McFadden	0.471
NangelKerke	0.492

Source own: construction

In another test of quality of the model collected data of 46,500 customers were gathered. This data was not used in the event modeling training. To differentiate between classes, a large analysis of the classification model is carried out. This predictor test is made possible by the use of ROC curve and look alike measures of predicting AUC



From the figure, it is clear that the curve a reliable source of classification model. The curve passes at 100 percent true and 0 percent false positive level. ACU provides a value of 0.977 which shows the logic regression classifier as the best model test.

How the Research Will Help American Business

The research will help American companies to cope with the changing trend in customer retention. America being the leading industrialized country, the research will help in developing techniques that will help in maintaining the customer. Learning about the variants helps to cope with the changing trends in the acquisition marketing world. Secondly, the predictive measure is helpful in establishing the position of the firm and how it will improve its constituents.

CONCLUSION

The purpose of this study was to develop a predictive model that will help create awareness on how to handle customers and reduce customer churn. When developing a predictive model charts a lot of care should be taken to produce the right result.

Precaution is taken during data retrieval to avoid compromising the data security policies of companies that the data is taken from. That aside, the three models help to present the data and result that is useful in decision making. The logistic algorithms are the basis of this research and helps in developing predictive measures.

Lastly the predictive model types are very critical in the prediction of performance of a company and how it can retain its clients. Again in this paper I have looked into several important aspects of variances. The dependent variable and its influences by the independent variables are analyzed at length.

REFERENCES

1) Wright, R. E. (1995). Logistic regression.
2) Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). Logistic regression. New York: Springer-Verlag.
3) Menard, S. (2002). Applied logistic regression analysis (Vol. 106). Sage.
4) Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.
5) Tranmer, M., & Elliot, M. (2008). Binary logistic regression. Cathie Marsh for census and survey research, paper, 20.
6) King, J. E. (2008). Binary logistic regression. Best practices in quantitative methods, 358-384.
7) Midi, H., Sarkar, S. K., & Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. Journal of Interdisciplinary Mathematics, 13(3), 253-267.
8) Harrell, F. E. (2015). Binary logistic regression. In Regression modeling strategies (pp. 219-274). Springer, Cham.
9) Nemes, S., Jonasson, J. M., Genell, A., & Steineck, G. (2009). Bias in odds ratios by logistic regression modelling and sample size. BMC medical research methodology, 9(1), 1-5.
10) Norton, E. C., Dowd, B. E., & Maciejewski, M. L. (2018). Odds ratios—current best practice and use. Jama, 320(1), 84-85.
11) Bender, R., & Grouven, U. (1998). Using binary logistic regression models for ordinal data with non-proportional odds. Journal of clinical epidemiology, 51(10), 809-816.
12) Uanhoro, J. O., Wang, Y., & O'Connell, A. A. (2019). Problems with using odds ratios as effect sizes in binary logistic regression and alternative approaches. The Journal of Experimental Education, 1-20.
13) Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression.
14) Liberman, A. M. (2005). How much more likely? The implications of odds ratios for probabilities. American Journal of Evaluation, 26(2), 253-266.
15) Chang, B. H., & Hoaglin, D. C. (2017). Meta-analysis of odds ratios: current good practices. Medical care, 55(4), 328.
16) Norton, E. C., Dowd, B. E., & Maciejewski, M. L. (2019). Marginal effects—quantifying the effect of changes in risk factors in logistic regression models. Jama, 321(13), 1304-1305.
17) Martinez, B. A. F., Leotti, V. B., Nunes, L. N., Machado, G., & Corbellini, L. G. (2017). Odds ratio or prevalence ratio? An overview of reported statistical methods and appropriateness of interpretations in cross-sectional studies with dichotomous outcomes in veterinary medicine. Frontiers in veterinary science, 4, 193.
18) Chen, W., Qian, L., Shi, J., & Franklin, M. (2018). Comparing performance between log-binomial and robust Poisson regression models for estimating risk ratios under model misspecification. BMC medical research methodology, 18(1), 1-12.
19) Schober, P., & Vetter, T. R. (2021). Kaplan-Meier Curves, Log-Rank Tests, and Cox Regression for Time-to-Event Data. Anesthesia & Analgesia, 132(4), 969-970.
20) Moamer, S., Baghestani, A. R., Pourhoseingholi, M. A., Khadem Maboudi, A. A., Shahsavari, S., Zali, M. R., & Mohammadi Majd, T. (2017). Application of the parametric regression model with the four-

parameter log-logistic distribution for determining of the effecting factors on the survival rate of colorectal cancer patients in the presence of competing risks. Iranian Red Crescent Medical Journal, 19(6).

21) Ahmad, R. W., Hasan, H., Jayaraman, R., Salah, K., & Omar, M. (2021). Blockchain applications and architectures for port operations and logistics management. Research in Transportation Business & Management, 100620.

22) Kain, R., & Verma, A. (2018). Logistics management in supply chain–an overview. Materials today: proceedings, 5(2), 381.