IJARIIT
INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 6.078
(Volume 6, Issue 4)
Available online at: https://www.ijariit.com

Accelerating sales for Citrix Education Services using Logistic Regression

A. Kalyan Aravind Kumar
aravindmcs1@gmail.com
Great Lakes Institute of Management, Chennai, Tamil Nadu

ABSTRACT

This paper aims to establish an efficient model for predicting company sales by leveraging logistic regression's strengths. A real dataset of Citrix used to figure out a significant variable affecting sales acceleration and to find an appropriate metric to measure unstructured information. To build an efficient model, we used two statistical methods; logistic regression and linear discriminant analysis. The classification accuracy of the models compared using Fisher Analysis, ROC curves, and confusion matrix. In regression analysis, it is evident that response and predictors some times may suffer from correlation issues. By definition, Multicollinearity is that two or more predictors are correlated; if this happens, the coefficients' standard error will increase. Increased standard errors mean that the coefficients for some or all independent variables may be significantly different from 0. In other words, Multicollinearity makes some variables statistically insignificant by overinflating the standard errors when they should be significant. In this paper, we concentrate on logistic regression analysis, linear discriminant analysis, Multicollinearity, fisher analysis, and consequences and effects on the reliability of the regression model.

Keywords: -*Logistic Regression, Linear Discriminant Analysis, Multicollinearity, Fisher Analysis*

1. INTRODUCTION

1.1 About Company

Citrix strives to influence a world where people, businesses, and things are securely connected and available to make the extraordinary feasible. Citrix helps customers reimagine the prospect of work by providing the most extensive secure digital workspace that unifies the applications, data, and services. People need to be productive and simplify IT's ability to adapt and manage complicated cloud environments. With 2016 yearly revenue of $3.42 billion, Citrix solutions are in use by higher than 400,000 companies, including 99 percent of the Fortune 100 and 98 percent of the Fortune 500.

Table 1: Facts about the company

Data	Numbers
Year Founded	1989
Nasdaq	CTXS
Revenue	FY16 $3.42 billion
Partners	10,000 in 100 countries
Customers	400,000

1.2 About the Education Services Industry

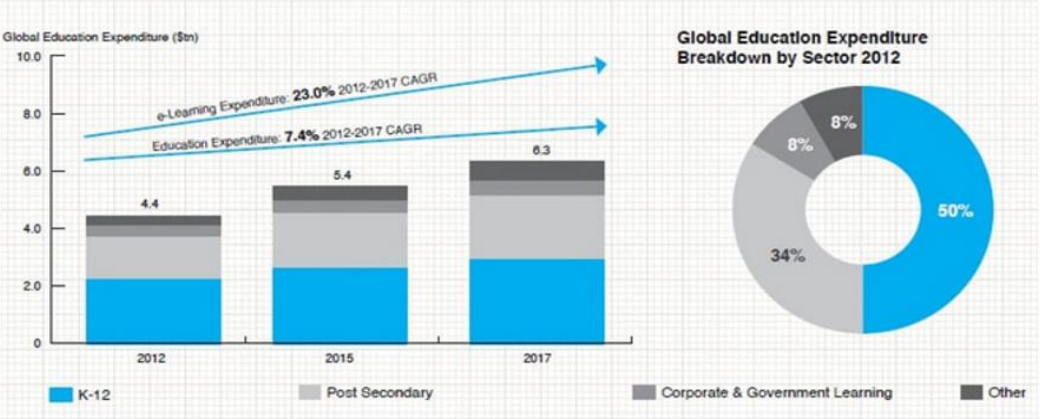


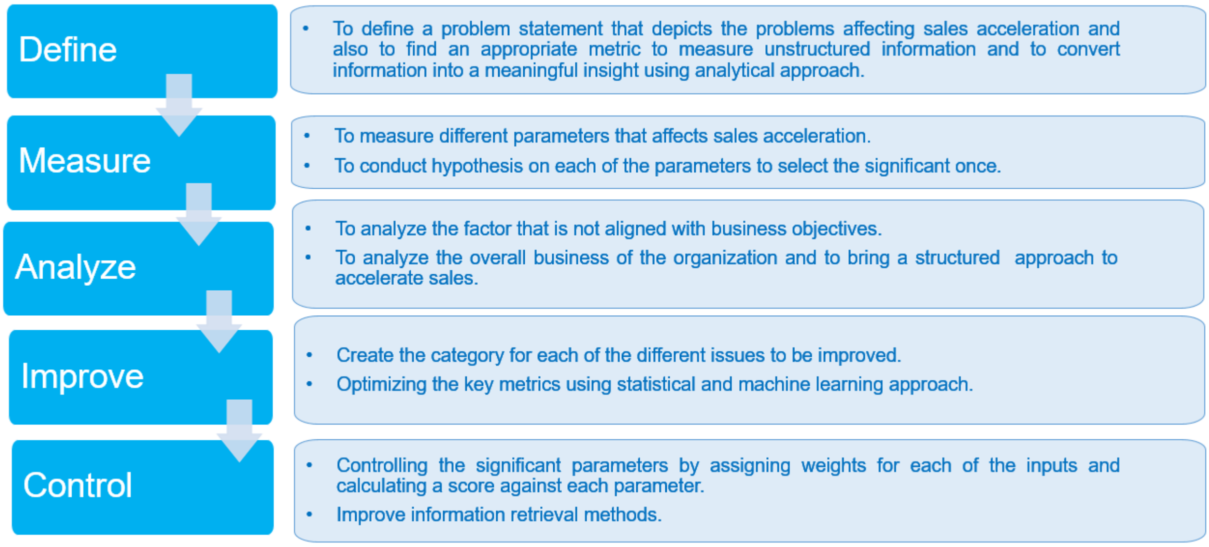
Fig. 1: Statistics about the Education Services Industry

© 2020, www.IJARIIT.com All Rights Reserved

Education Services provides Product Training and Certification to cater to the evolving needs of IT professionals and organizations. As per inputs from the London Based Investment bank IBIS Capital, the education industry is the fastest growing industry in the markets, with its E-learning sector having a growth of 23% in 2017.

1.3 Project Approach

Figure 2 below shows the start to end steps for this project approach. Mainly steps consist of Define, Measure, Analyze, Improve, and Control. In each stage, the right side columns explain output to be achieved.



- To define a problem statement that depicts the problems affecting sales acceleration and also to find an appropriate metric to measure unstructured information and to convert information into a meaningful insight using analytical approach.
- To measure different parameters that affects sales acceleration.
- To conduct hypothesis on each of the parameters to select the significant once.
- To analyze the factor that is not aligned with business objectives.
- To analyze the overall business of the organization and to bring a structured approach to accelerate sales.
- Create the category for each of the different issues to be improved.
- Optimizing the key metrics using statistical and machine learning approach.
- Controlling the significant parameters by assigning weights for each of the inputs and calculating a score against each parameter.
- Improve information retrieval methods.

Fig. 2: Project approach steps

1.4 The Business Case (Why are we doing this project?)

- Citrix Professional Services revenue comprising of Consulting and Education (Product Training and Certification) remained consistent. $132 million as of December 31, 2017, compared to $131 million as of December 2016.
- In particular, the Education Services revenue has declined by 41% for the fiscal year 2017 compared to 2014.
- Overall, Sales bookings for Education Services declined by 32% for the fiscal year 2017 compared to 2014.
- 37% of Opportunities translating to $50 million were not converted into Sales.

1.5 Problem Statement (What is the problem we intend to resolve?)

In the last three years (2015 – 2017), Education Services bookings have not met targets consistently and end up with a difference of $ 5.7 M per year on an average, impacting financial statements. To provide Visibility and Predictability of Sales Motion to Education Services Management by transforming sales data to insights that can drive sales acceleration using the DMAIC approach.

2. LITERATURE REVIEW

The techniques used in earlier research on modeling company growth relied on conventional statistical methods such as multiple regression, logistic regression, and linear discriminant analysis. Delmar et al. [1] used correlations and regression investigation to model company growth. Geroski et al. [2] used static and dynamic optimizing methodologies for company output choice, modeled production functions for corporate learning, constructed R&D competition and diversification, and tested their influence on corporate growth rates. Ma and Tang et al. [3] used a step by step process of the logistic regression model to project prognosticate a default counting the misclassification loss. The method of Logistic regression was also compared with and integrated into specific machine learning methods to predict company crisis or bankruptcy. Hua et al. [4] combined The Logistic Regression and support vector machines (SVM) to forecast corporation financial crisis by using an integrated binary discriminant rule (IBDR) that decreases the experiential risk of support vector machine outputs and interpretations and transforms the outputs according to the result of Logistic Regression (LR) analysis. Other machine learning/deep learning methods, such as ANNs, have not yet been investigated to adequately model company growth.

3. LOGISTIC REGRESSION DEFINITION

Logistic regression will model the odds of a result based on individual characteristics. Because change is a ratio, what will be truly modeled is the logarithm of the change given by:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_m x_m$$

Where $\boldsymbol{\pi}$ indicates, the probability of an event and $\boldsymbol{\beta_i}$ are the regression coefficients linked with the reference group and $\boldsymbol{x_i}$ analytical variables. At this state, an essential concept must be highlighted. The reference group represented by$\boldsymbol{\beta_0}$, is established by those individuals presenting the reference level of every variable$\boldsymbol{x_{1\ldots m}}$.

4. DATA PREPARATION AND EXPLORATION

The data has taken from Citrix CRM Tool (Live Project). Contents that are made from the CRM tool are Opportunity, Bookings, Service Offering, Route-To-Market, Customer, and Partner from 2015 to 2017. The below image (Figure 3) shows the process flow for the opportunity to the status of closed.

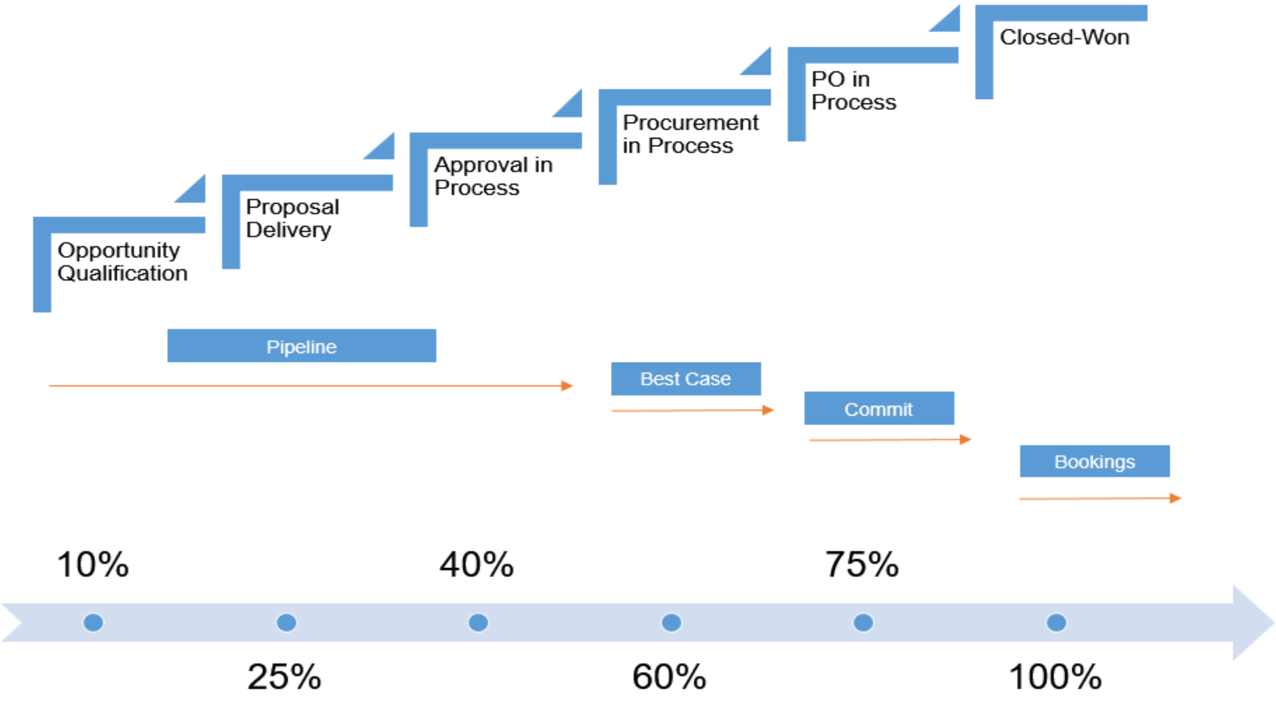
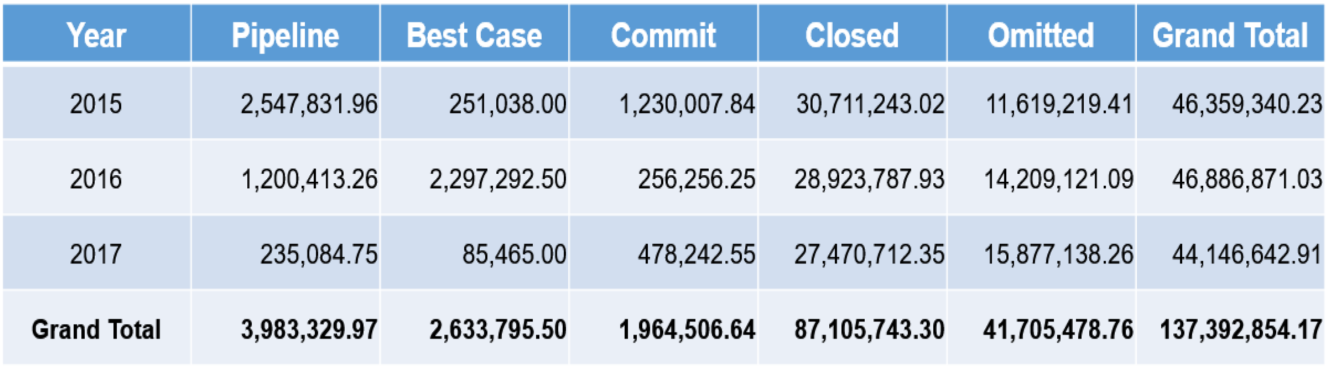
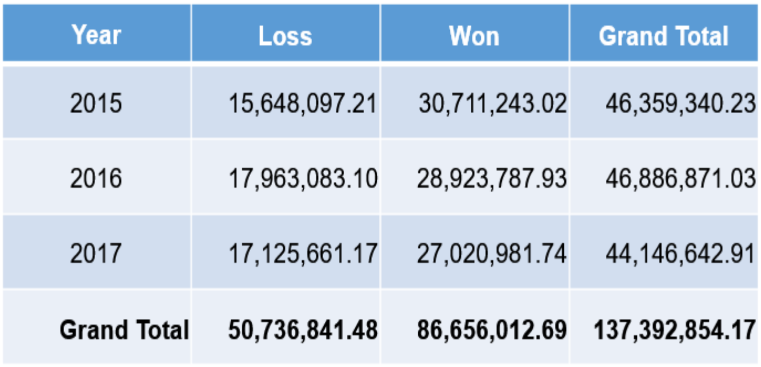


Fig. 3: As-Is Process Map Opportunity

The end status of every opportunity is either Won or Loss. Won means the opportunity has been converted to business successfully, and Loss means Citrix has lost the opportunity. The image below (Figure 4) shows the statistical analysis for every opportunity in terms of dollar values.



Year	Pipeline	Best Case	Commit	Closed	Omitted	Grand Total
2015	2,547,831.96	251,038.00	1,230,007.84	30,711,243.02	11,619,219.41	46,359,340.23
2016	1,200,413.26	2,297,292.50	256,256.25	28,923,787.93	14,209,121.09	46,886,871.03
2017	235,084.75	85,465.00	478,242.55	27,470,712.35	15,877,138.26	44,146,642.91
Grand Total	**3,983,329.97**	**2,633,795.50**	**1,964,506.64**	**87,105,743.30**	**41,705,478.76**	**137,392,854.17**



Year	Loss	Won	Grand Total
2015	15,648,097.21	30,711,243.02	46,359,340.23
2016	17,963,083.10	28,923,787.93	46,886,871.03
2017	17,125,661.17	27,020,981.74	44,146,642.91
Grand Total	**50,736,841.48**	**86,656,012.69**	**137,392,854.17**

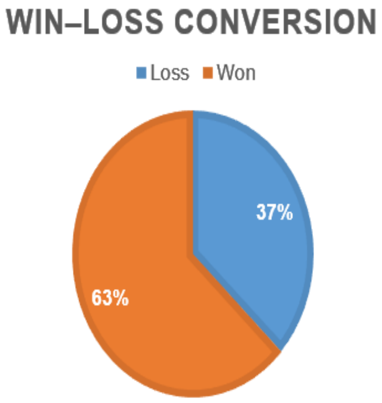


Fig. 4: Opportunity Analysis.

4.1 Basic Data Analysis

The below image (Figure 5) illustrates the distribution of Won and Loss from 2015 to 2017. As per the graph, wining opportunity decreases; in 2015, the won Amount is nearly 30,711,243 USD, whereas, in 2017, it decreased to 27,020,982 USD.

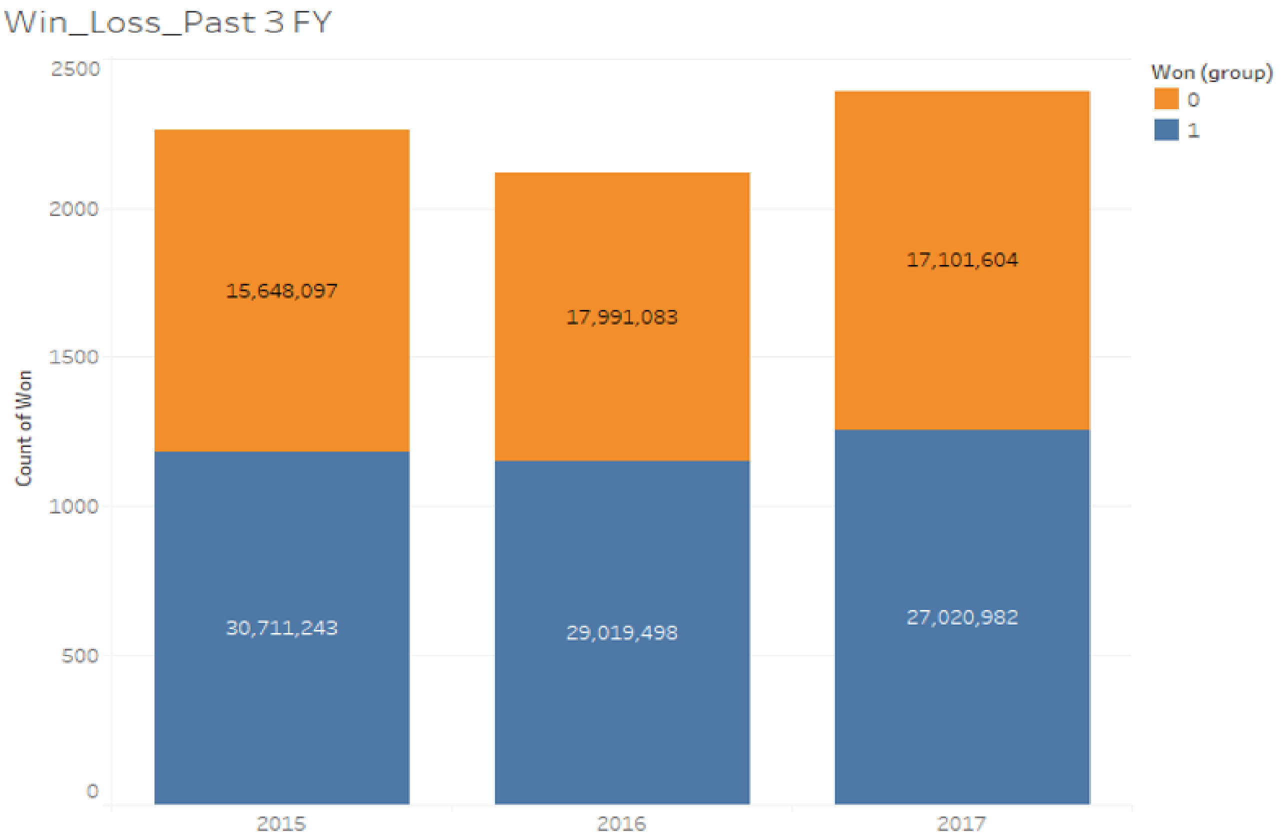


Fig. 5: Basic Data Analysis for Win-Loss from 2015 to 2017.

The below image (Figure 6) shows the distribution of Won and Loss by opportunity source. As per the graph, wining opportunity is high from Citrix Education services followed by Citrix Enterprise Sales.

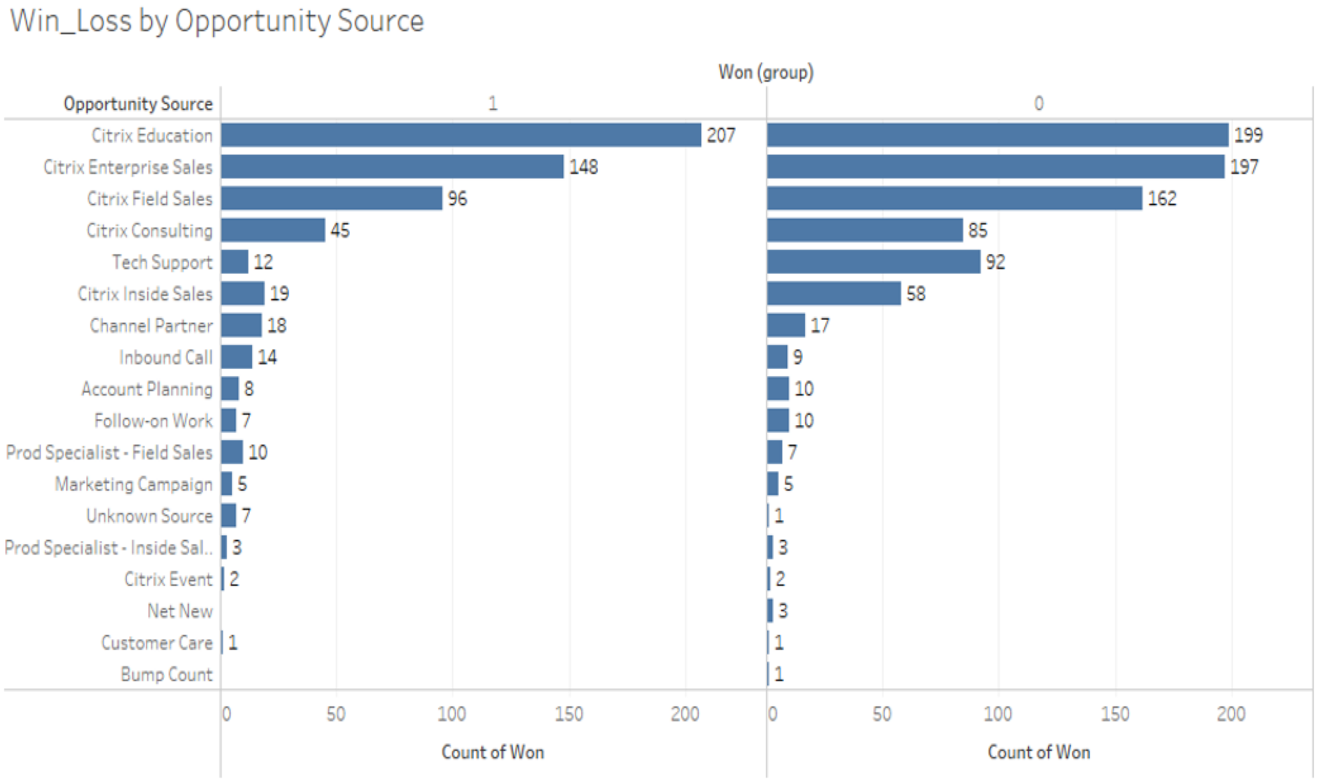


Fig. 6: Basic Data Analysis for Win-Loss by opportunity source.

The image below (Figure 7) is a Pareto chart that describes Citrix's relationship with the customer is the highest significant factor for winning the opportunity.

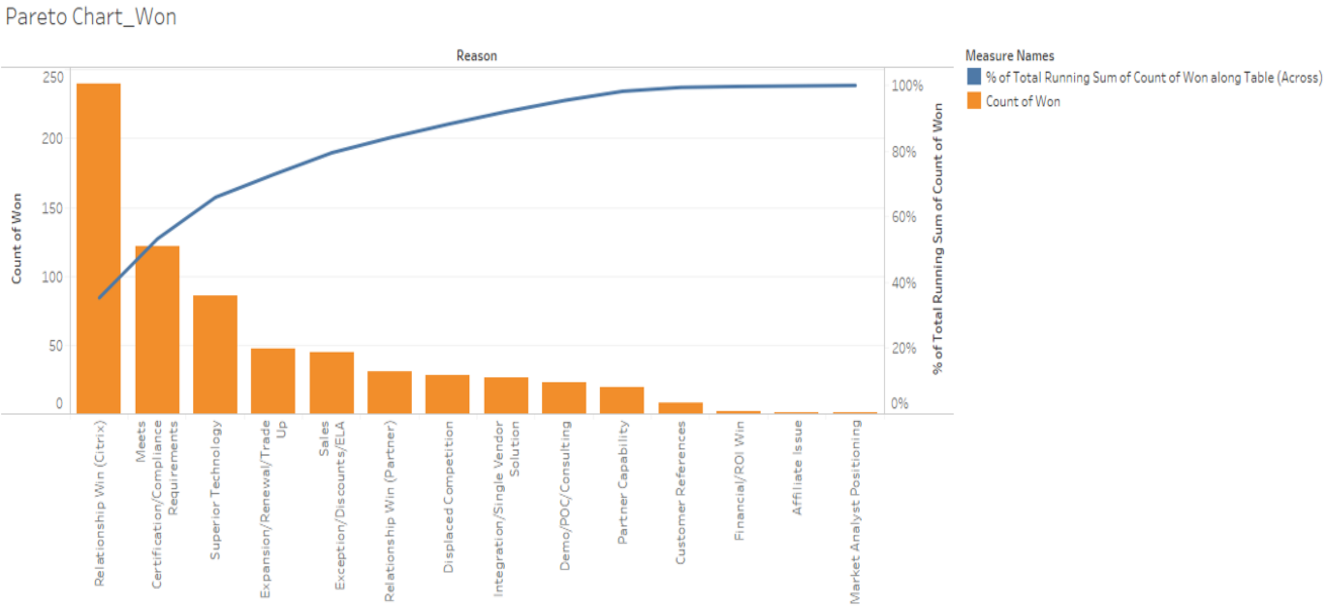


Fig. 7: Pareto Chart by Opportunities Won

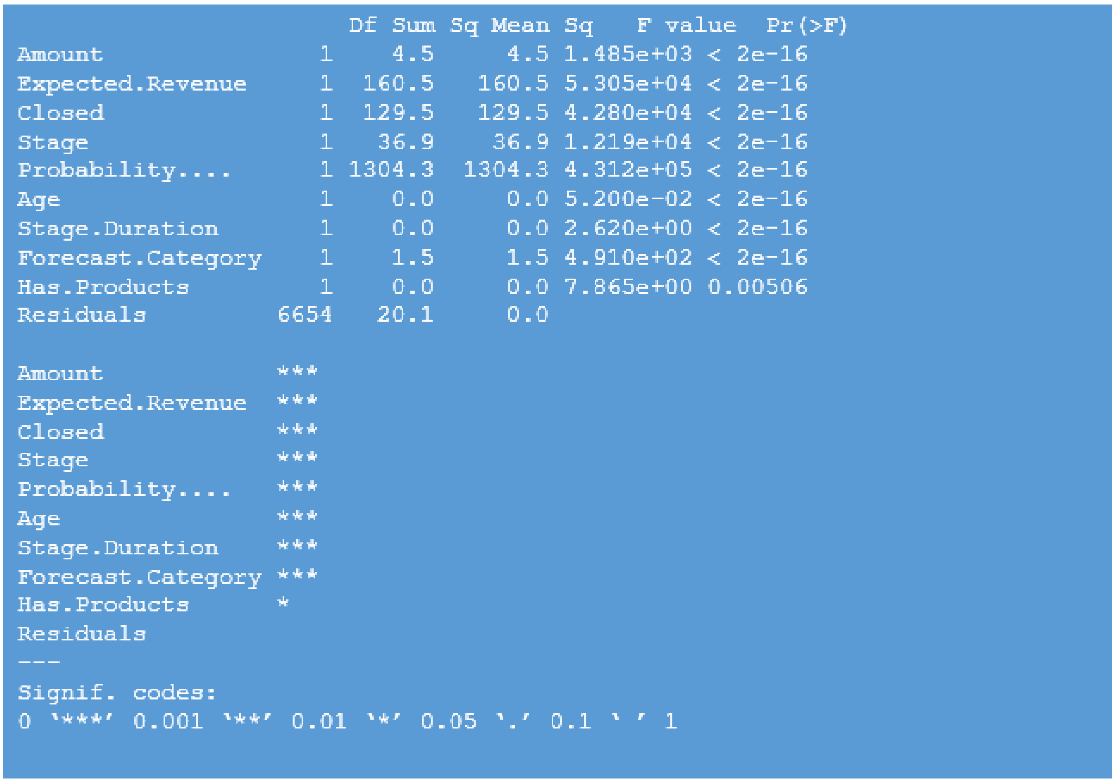
5. MODEL BUILDING

5.1 Hypothesis Testing

Hypothesis:

- Null Hypothesis – All independent variables have no impact on dependent variable “Won”
- Alternative Hypothesis – At least one independent variables affect dependent variable "Won"

The below image (Figure 8) shows the statistical interpretation of the relationship among all variables.



```
                        Df  Sum Sq Mean Sq   F value   Pr(>F)
Amount                   1     4.5     4.5 1.485e+03 < 2e-16
Expected.Revenue         1   160.5   160.5 5.305e+04 < 2e-16
Closed                   1   129.5   129.5 4.280e+04 < 2e-16
Stage                    1    36.9    36.9 1.219e+04 < 2e-16
Probability....          1 1304.3  1304.3 4.312e+05 < 2e-16
Age                      1     0.0     0.0 5.200e-02 < 2e-16
Stage.Duration           1     0.0     0.0 2.620e+00 < 2e-16
Forecast.Category        1     1.5     1.5 4.910e+02 < 2e-16
Has.Products             1     0.0     0.0 7.865e+00 0.00506
Residuals             6654    20.1     0.0

Amount                ***
Expected.Revenue      ***
Closed                ***
Stage                 ***
Probability....       ***
Age                   ***
Stage.Duration        ***
Forecast.Category ***
Has.Products          *
Residuals
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

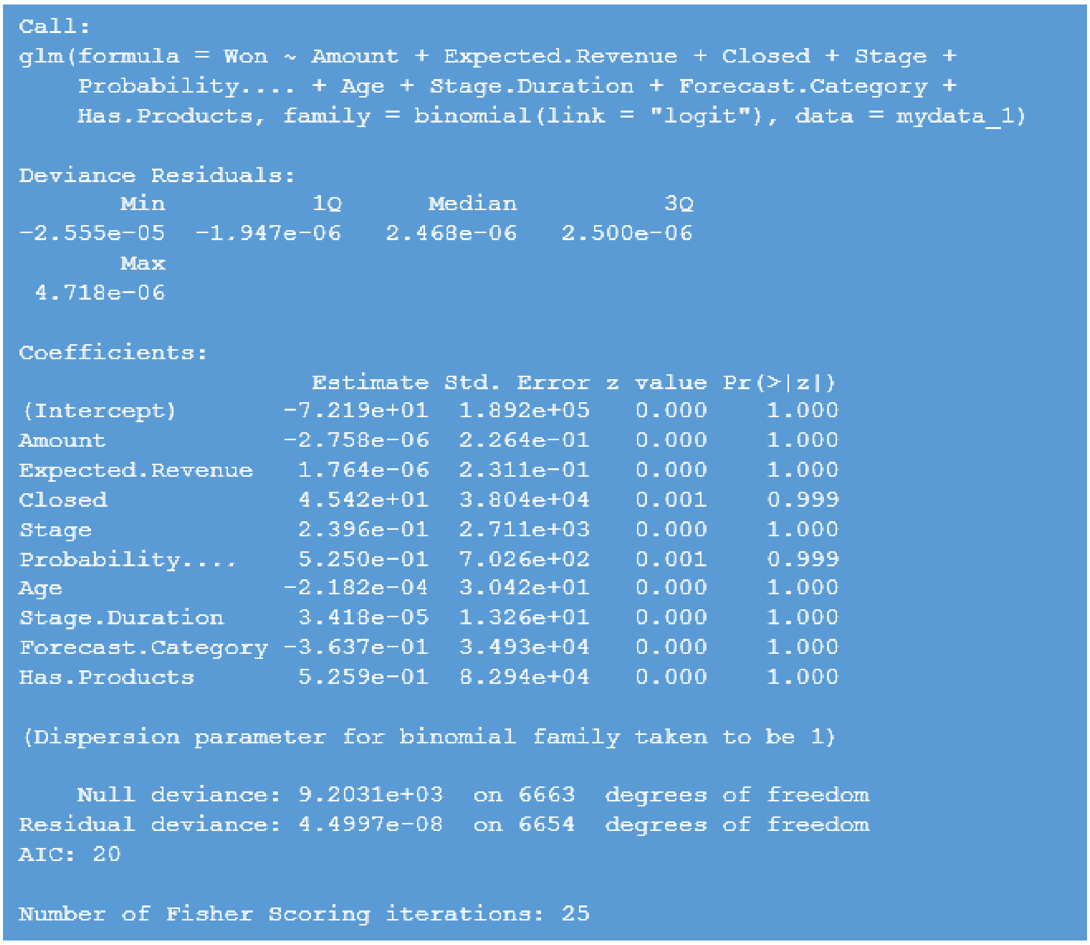
Fig. 8: Statistical values by R software

Interpretation:

Regarding the results, we can conclude that all the variables are significant (with a p-value less than 0.05) other than 'Has products'. Since most of the parameters are significant, we can go with further analysis after checking the problem of Multi-collinearity.

5.2 Modeling – Logistic Regression

Logistic Regression – Given the response variable "Won" is categorical and binary, we are applying Logistic Regression to check the significance of independent variables.



```
Call:
glm(formula = Won ~ Amount + Expected.Revenue + Closed + Stage +
    Probability.... + Age + Stage.Duration + Forecast.Category +
    Has.Products, family = binomial(link = "logit"), data = mydata_1)

Deviance Residuals:
       Min          1Q      Median          3Q
-2.555e-05  -1.947e-06   2.468e-06   2.500e-06
       Max
 4.718e-06

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -7.219e+01  1.892e+05   0.000    1.000
Amount            -2.758e-06  2.264e-01   0.000    1.000
Expected.Revenue   1.764e-06  2.311e-01   0.000    1.000
Closed             4.542e+01  3.804e+04   0.001    0.999
Stage              2.396e-01  2.711e+03   0.000    1.000
Probability....    5.250e-01  7.026e+02   0.001    0.999
Age               -2.182e-04  3.042e+01   0.000    1.000
Stage.Duration     3.418e-05  1.326e+01   0.000    1.000
Forecast.Category -3.637e-01  3.493e+04   0.000    1.000
Has.Products       5.259e-01  8.294e+04   0.000    1.000

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9.2031e+03  on 6663  degrees of freedom
Residual deviance: 4.4997e-08  on 6654  degrees of freedom
AIC: 20

Number of Fisher Scoring iterations: 25
```

Fig. 9: Logistic Regression full model by R software

Interpretation:
The result explains none of the Variables are significant, and the model may be suffering from Multicollinearity.

5.2 Check for Multicollinearity

Multicollinearity is nothing but the correlation between predictors in the model. The existence of Multicollinearity can adversely affect regression results. The VIF evaluates how much the variance of a regression coefficient is increased due to Multicollinearity in the model. VIFs are measured by taking a predictor and regressing it against each other predictor in the model [5] [6].

$$VIF = \frac{1}{1 - R_i^2}$$

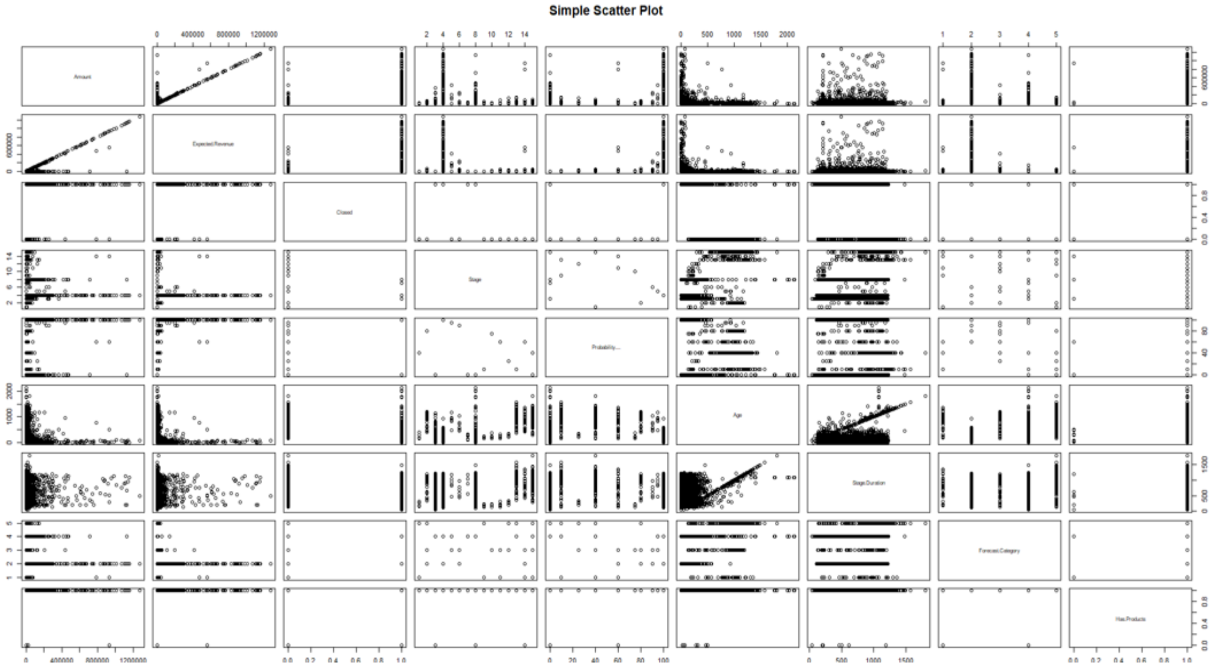
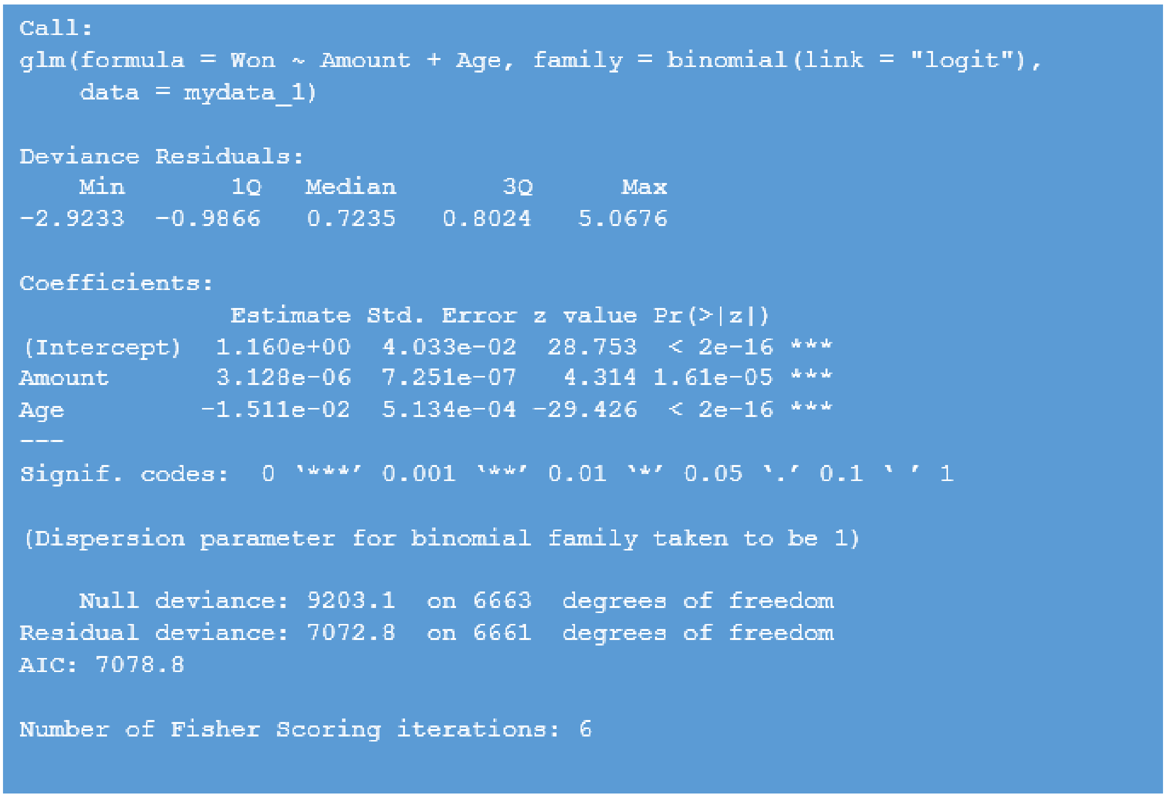


Fig. 10: Correlation table among all predictors

Interpretation:
Based on the VIF values and using the above graph, we concluded that the dependent variables, Amount, and Age are the only significant variables to predict the target variable Won.

5.3 Reduced Model

After removing all correlation predictors, we reran the logistic regression model by considering only the Amount and Age variables.



```
Call:
glm(formula = Won ~ Amount + Age, family = binomial(link = "logit"),
    data = mydata_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9233  -0.9866   0.7235   0.8024   5.0676

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.160e+00  4.033e-02  28.753  < 2e-16 ***
Amount       3.128e-06  7.251e-07   4.314 1.61e-05 ***
Age         -1.511e-02  5.134e-04 -29.426  < 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: 9203.1  on 6663  degrees of freedom
Residual deviance: 7072.8  on 6661  degrees of freedom
AIC: 7078.8

Number of Fisher Scoring iterations: 6
```

Fig. 11: Reduced model after removing correlated predictors

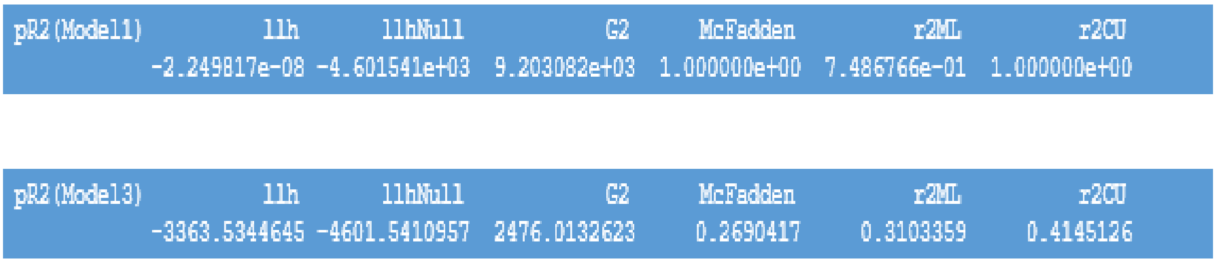
Interpretation:
The dependent variable is highly significant with independent variables. +Ve Sign of coefficient Amount suggests as Amount increases, chances of individual transactions getting converted are high. -Ve Sign of coefficients Age indicates as Age increases chances of individual transaction getting converted is low. The probability of winning an opportunity concerning the 'Amount' variable is 92.3%, which implies that there is a high chance of winning an opportunity for a higher value of Amount. The probability of winning an opportunity concerning the 'Age' variable is 98%, Which implies that there is a high chance of winning an opportunity for less Age.

5.3 Assessing Model Fit - McFadden pseudo-R squared Test

The Logistic regression (LR) models are implemented using the method of maximum likelihood. The parameter estimates values that maximize the likelihood of the data which have been observed. McFadden's R squared Test is defined as

$$R^2_{McFadden} = 1 - \frac{log(L_c)}{log(L_{Null}}$$

As per our R code, McFadden scores shown below.



```
pR2(Model1)           llh        llhNull            G2      McFadden          r2ML          r2CU
           -2.249817e-08 -4.601541e+03  9.203082e+03  1.000000e+00  7.486766e-01  1.000000e+00
```

```
pR2(Model3)           llh        llhNull            G2      McFadden          r2ML          r2CU
           -3363.5344645 -4601.5410957  2476.0132623     0.2690417     0.3103359     0.4145126
```

Fig. 12: McFadden pseudo-R squared Test results

Interpretation:
The McFadden value for the model is 27%, and Model robustness is slightly better.

5.4 ROC Curve

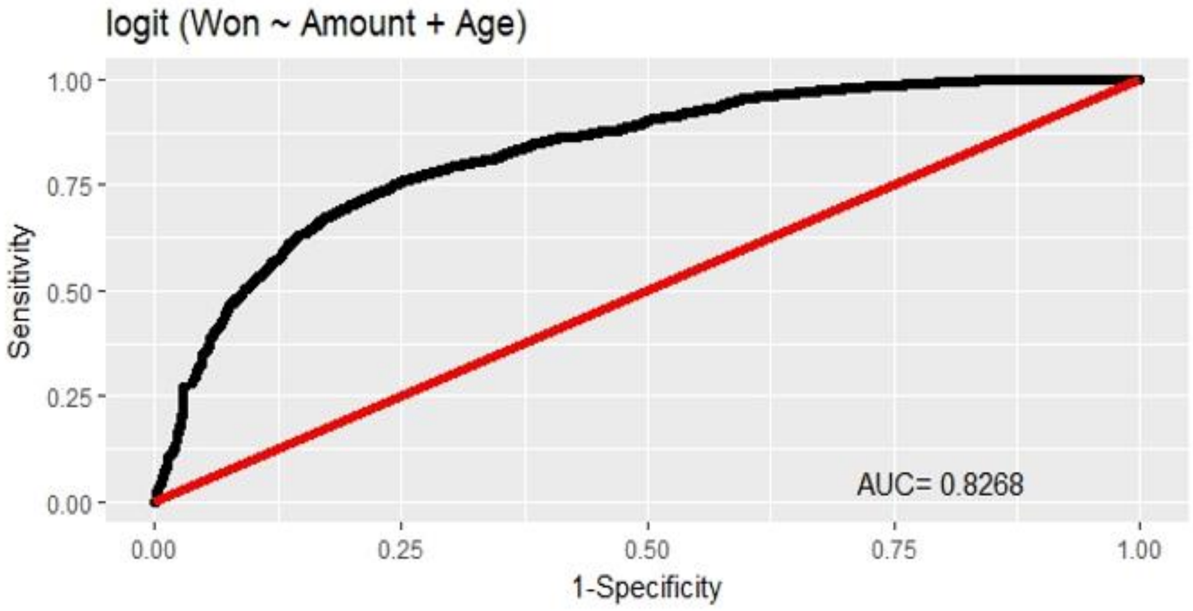
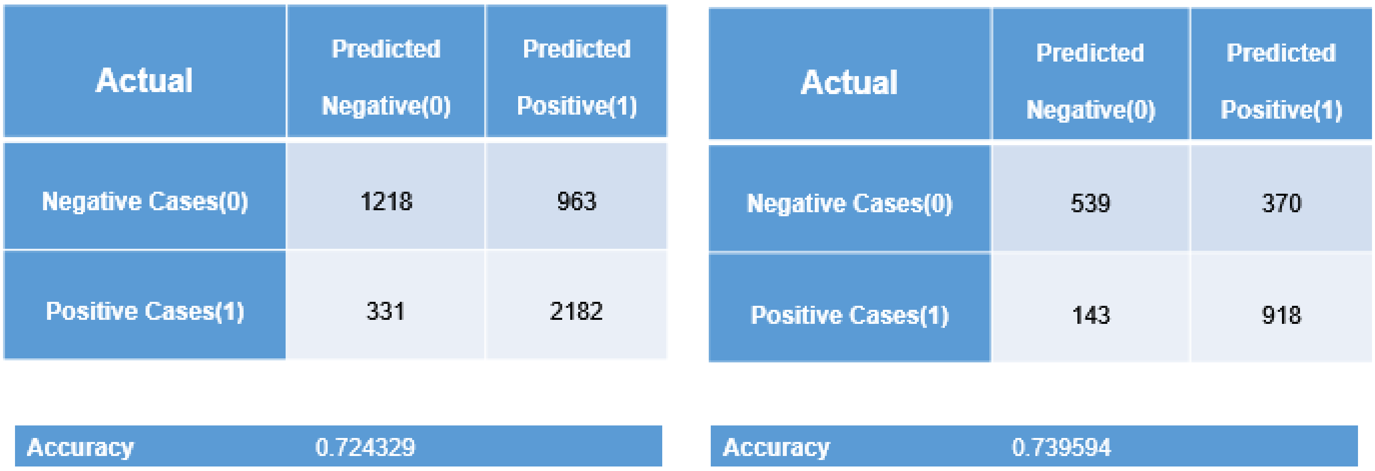


Fig. 13: ROC Curve Test results

Interpretation:
The ROC curve is a vital tool for diagnostic test evaluation. In a ROC curve, the actual positive rate (Sensitivity) is plotted in function of the false positive rate (Specificity) for distinct cut-off points of a parameter. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. The area under the curve or ROC is a measure of how well a parameter can distinguish between two diagnostic groups (Win/Loss).

5.5 Confusion Matrix



Actual	Predicted Negative(0)	Predicted Positive(1)
Negative Cases(0)	1218	963
Positive Cases(1)	331	2182

Accuracy	0.724329

Actual	Predicted Negative(0)	Predicted Positive(1)
Negative Cases(0)	539	370
Positive Cases(1)	143	918

Accuracy	0.739594

Fig. 14: Confusion Matrix for Training and Test data results

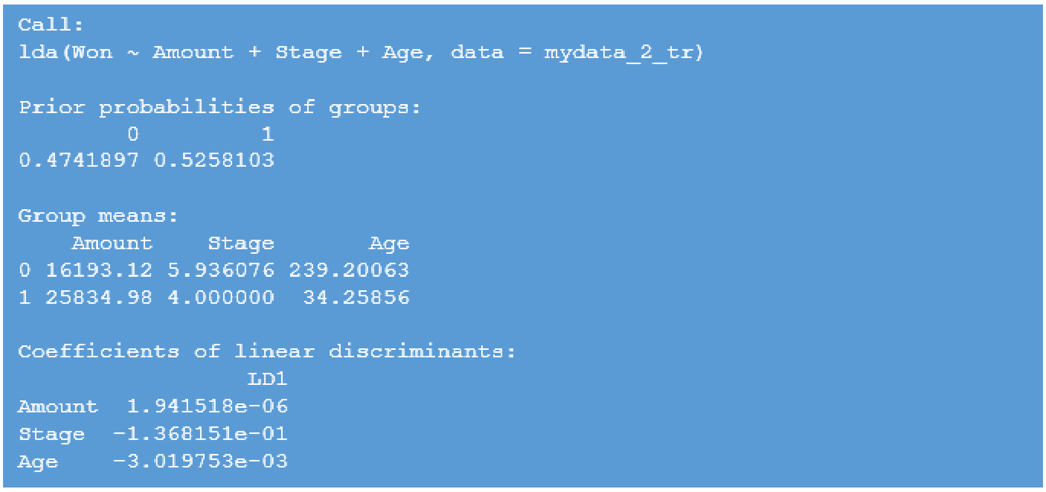
Interpretation:
Confusion Matrix determines the accuracy of the model. Confusion matrix of the model with the predicted accuracy of 72% for the training data (1st Table). Confusion matrix of the model with the expected accuracy of 73% for the test data (2nd Table).

5.6 Modeling – Linear Discriminant Analysis (LDA)

Linear discriminant analysis (LDA) is not just a dimension reduction method but also a powerful classification method. With or without data normalcy assumption, we can arrive at the same Linear discriminant analysis (LDA) features. If we consider data comes from multivariate Gaussian distribution, that is, the population of X can be distinguished by its mean (µ) and covariance (Σ), explicit forms of the above allocation rules can be achieved. Following the Bayesian rule, we divide the data x to class j if it has the highest likelihood amongst all K classes for i = 1,…, K:

$$\boldsymbol{\delta_i(x) = log f_i(x) + log\pi_i}$$

The above function is known as the discriminant function, and the discriminant function informs us how possibly data x is from each class. The decision boundary dividing any two classes, k and l

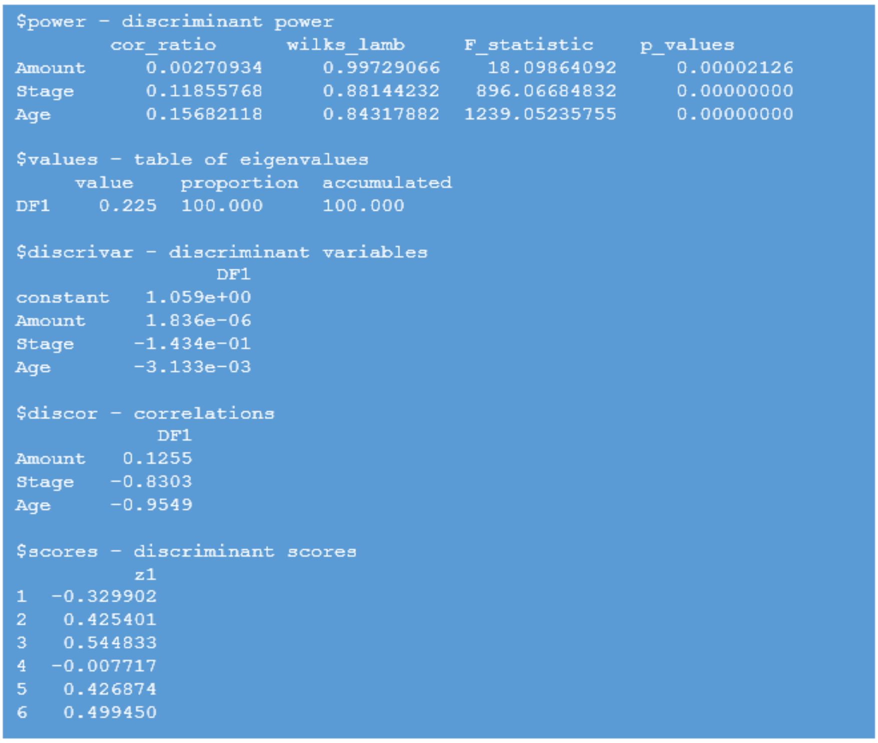


```
Call:
lda(Won ~ Amount + Stage + Age, data = mydata_2_tr)

Prior probabilities of groups:
        0         1
0.4741897 0.5258103

Group means:
    Amount    Stage       Age
0 16193.12 5.936076 239.20063
1 25834.98 4.000000  34.25856

Coefficients of linear discriminants:
                 LD1
Amount  1.941518e-06
Stage  -1.368151e-01
Age    -3.019753e-03
```

Fig. 15: Modeling – Linear Discriminant Analysis (LDA) - R code

Interpretation:
The main approach is separating the Win from Loss in an opportunity by these predictor variables, which will help further classification and differentiation. The Amount is a highly significant independent variable in differentiating the dependent variable (Won/Loss). Age is second in order as an independent variable in distinguishing the dependent variable (Won/Loss). Amount and Age are essential for separating the opportunity "Win" or "Loss".

5.7 Fisher Analysis – Linear Discriminant Analysis (LDA)

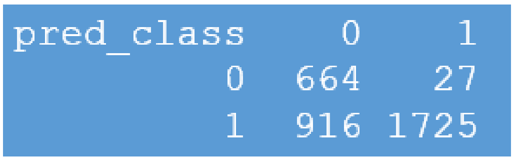


```
$power - discriminant power
        cor_ratio    wilks_lamb    F_statistic    p_values
Amount   0.00270934   0.99729066    18.09864092   0.00002126
Stage    0.11855768   0.88144232   896.06684832   0.00000000
Age      0.15682118   0.84317882  1239.05235755   0.00000000

$values - table of eigenvalues
     value   proportion  accumulated
DF1    0.225  100.000     100.000

$discrivar - discriminant variables
                 DF1
constant   1.059e+00
Amount     1.836e-06
Stage     -1.434e-01
Age       -3.133e-03

$discor - correlations
            DF1
Amount   0.1255
Stage   -0.8303
Age     -0.9549

$scores - discriminant scores
         z1
1  -0.329902
2   0.425401
3   0.544833
4  -0.007717
5   0.426874
6   0.499450
```

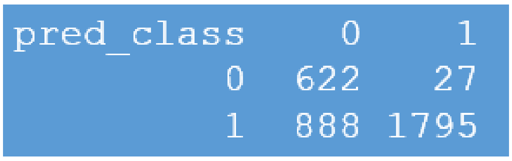
Fig. 16: Fisher Analysis test for Model – Linear Discriminant Analysis (LDA) - R code.

Interpretation:
the p-value is statistically significant for all three independent variables, indicating Amount, Stage, and Age are excellent predictors. Relative importance can be inferred from the correlation ratio for each variable. Age is rank one followed by Stage and Amount as far as their ability to differentiate the group.

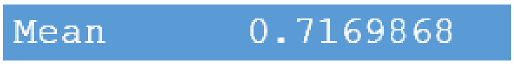
5.8 LDA Model Accuracy – Confusion Matrix



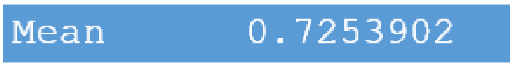
```
pred_class    0    1
         0  664   27
         1  916 1725
```



```
pred_class    0    1
         0  622   27
         1  888 1795
```



```
Mean         0.7169868
```



```
Mean         0.7253902
```

Fig. 17: LDA Model Accuracy – Confusion Matrix

Interpretation:
Confusion Matrix determines the accuracy of the model. Confusion matrix of the model with the predicted accuracy of 72% for the training data (1st Table). Confusion matrix of the model with the expected accuracy of 73% for the testing data (2nd Table).

6. CONCLUSION

This paper showed a comparison of Logistic Regression and Linear Discriminant Analysis (LDA) to predict Citrix education services sales with 73% accuracy. We also discussed the VIF method to avoid Multicollinearity and conducted different tests like McFadden, ROC curve, and confusion matrix. We did this study to identify the significant variables for the sales accretion on the factor opportunity conversion. By this experiment, the key highlighted points are:

- Amount and Age plays a vital role in Opportunity Conversion.
- With every $1000 increase in Opportunity Amount; probability is increased by 0.06%
- With an additional day of increase in Age, the probability is decreased by 0.3%.

7. REFERENCES

[1] Delmar, F., Davidsson, P., Gartner, W. B. (2003). Arriving at the high-growth firm, Journal of Business Venturing, 18(2), pp. 189-216.

[2] Geroski, P.A. (2005). Understanding the implications of experimental work on corporate growth rates, Managerial and Decision Economics, 26(2), 129–138.

[3] Ma, R., Tang, C. (2007). Building up Default Predicting Model based on Logistic Model and Misclassification Loss, Systems Engineering - Theory & Practice, 27(8). 33-38.

[4] Hua, Z., Wang, Y., Xu, X., Zhang, B., Liang, L. (2007). Predicting corporate financial distress based on the integration of support vector machine and logistic regression, Expert Systems with Applications, 33(2), 434-440.

[5] Dodge, Y. (2008). The Concise Encyclopedia of Statistics. Springer.

[6] Everitt, B. S.; Skrondal, A. (2010), The Cambridge Dictionary of Statistics, Cambridge University Press.