Journal of Industrial and Systems Engineering
Vol. 14, Special issue: 18th International Industrial Engineering Conference
Winter (January) 2022, pp. 19-30

JISE

17 November 2021
Iran Institute of Industrial Engineering

Diagnosis the dependence of revenue sources of communication service companies on specific services using machine learning

Mahmoud Tajik Jangali[1*], Ahmad Makui[1], Narges Taheri[2], Ehsan Dehghani[1], Somaye Kazemi[3]

[1]*School of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran*
[2]*strategic & business development, Telecommunication Infrastructure Company, Tehran, Iran*
[3]*Master of Engineering, Tehran University*

mahmoud_tajik@ind.iust.ac.ir,amakui@iust.ac.ir,ntaheri@tic.ir,ehsan_dehghani@mail.iust.ac.ir, s.kazemy@tic.ir

Abstract

Nowadays, Telecommunication has a vital role in both developed and emerging economies countries. Especially after coronavirus epidemic, the importance of telecommunication service like internet in education, research, economy and other areas is evident. Due to the alluring market of providing internet services to the main customers of IT industry and its significant profit, the demand of the other services has decreased sharply. Hence, a large part of the revenues of the IT industry be related to internet services. In this study, balancing of revenue sources has investigated as one of the important diagnosis facing the IT industry. In order to overcome this problem, introducing low-demand services along with internet service in the form of a package to the main customers is analyzed with a best-known machine learning algorithm, Generalized Linear Model. In order to validate the applicability of our study, a case study of a company providing telecommunication infrastructure and internet network bandwidth in Iran, is presented.

Keywords: IT industry, internet services, telecommunication, machine learning

1-Introduction

In the new global economy, competitiveness has become a crucial issue for organizations. Due to high growing consumer demands and several tools and methods of management, it is vital that companies achieve a real competitive advantage. To stand out in this environment, companies have tendency to optimize their processes. Therefore, they improve their performance through a proper diagnosis, implementing or adapting to change management. However, companies often fail in their improvement efforts because of conducting inappropriate companies' diagnosis. The diagnostic activity can be implemented in various area such as healthcare (De Ramon Fernandez, Ruiz Fernandez and Sabuco Garcia, 2020), manufacturing industry (Hossain, Abu-Siada and Muyeen, 2018; Soualhi and Razik, 2020), new industry such as 3D printing (Kim *et al.*, 2018), supply chain activities (Wu and Hsiao, 2021), tourism (Félix, Garcia and Vera, 2020), education (Wang *et al.*, 2020), and many other services such as Telecommunication (Sorrentino *et al.*, 2019). The business process is a set of activities that receives inputs and creates a valuable output to the customer.

[*]Corresponding author
ISSN: 1735-8272,

In the past decade, business processes have received much attention, but there is a lack of established diagnosis standards. Ko, Lee and Lee (2009) provided an extensive literature review and propose a classification. This classification includes execution, interchange, graphical, and diagnostic standards. Kohlbacher (2010) provide a complete review that investigated the impact of process orientation. They declare that focusing on business processes rather than emphasizing functional structure have positive effects on customer satisfaction, quality improvement, cost reduction, financial performance. In another work studied by Varela-Vaca et al., (2019) a risk assessment method is presented to evaluate a set of activities in a business process model. First, an algorithm is defined to verify the level of risk. Then, an algorithm is designed to diagnose the risk of the activities. Finally, the tool is used to support the described proposal. Furthermore, they validate the automation of the algorithms by a real case study. There are a lot of effective methods which has been employed to identify the relationship between the monitoring data and the health of the system to help root cause detection and analysis. Typically, the diagnosis of the faults is detected by experience and expert knowledge of engineers and this is a time-consuming procedure. This important limit has led to the presentation of several intelligent methods to shorten the cycle and improve the diagnosis. Machine learning algorithms are a part of these methods (Lei et al., 2020). The application of these algorithms is common in various areas, especially in analyzing and predicting problems and faults. According to the growth of complexity, the application of machine learning algorithms in this area, has become more popular than before. For instance, Generalized Linear Model (GLM) (Emami *et al.*, 2020), Random Forest (Liu *et al.*, 2020; Sun et al., 2020), KNN (Shahabi *et al.*, 2020; Wardani, Sihombing and others, 2020), Naive Bayes (Yao and Ye, 2020), and Logistic Regression (Chen et al., 2020) are some proposed algorithms which has been applied to diagnose faults. Moreover, GLM is used as the solution approach in this study. Nowadays, Telecommunication has a vital role in both developed and emerging economies countries. Especially after Coronavirus epidemic, the importance of telecommunication service like internet in education, research, economy and other areas is evident. In the work conducted by Tetteh (2012), paid attention to diagnosis as a way to rectify and improve performance of the telecommunication industry. Both qualitative and quantitative methods were applied to collect data in the organization's diagnosis, the areas of change, and the telecom environment assessment. The result of a real case shows that all three telecommunication companies, paid a lot of attention to the diagnosis and accepted proposing new products and services as an important diagnose that require to be discovered. Moreover, this approach helped propose practicable change and advance. Like Tetteh, Carrera et al., (2014) studied a telecommunications network. In their work, a hybrid diagnostic technique is proposed to provide competitiveness for the telecommunications networks. They achieved cost savings and customer retention by introducing the Fault Diagnosis Multi-Agent System. The collected data of system operation shows a reduction in the average incident solution time and the mean diagnosis time. In a study investigating diagnosis system in LTE network, Khatib et al., (2015) reported that Self-Organizing Networks (SON) can be a solution to reduce operational cost. SON can be used in the radio access network to detect faults, analyze root cause, balance fault, and recover them automatically. Consequently, the downtime reduces the impact on the user experience. They proposed a supervised, data driven learning decision making algorithm to analyze root cause. Finally, they validated their method by a real case study. A systematic procedure based on data driven method can be employed as a fault monitoring and diagnosis. Echraibi et al., (2020) paid attention to diagnosis of root causes in a telecommunication network. An infinite Multivariate Categorical Mixture Model is proposed for the clustering patterns of faults from data which is gathered from telecommunication networks. In addition, Quiñones-Grueiro, Llanes-Santiago and Neto (2021) have investigated diagnosis application in industrial systems. They classified faults with Data-Driven Methods. This work examines the various method of introducing services to customers and their performances in sales and profits in a telecommunications company as a diagnostic approach. Moreover, to the best our knowledge, there has been a lack of good and applicable references in this field. This work has been organized in the following way. Section 2, explains the problem definition in detail. The research methodology is presented in section 3. Furthermore, section 4 present the case study and validation of our proposed method. The results are analyzed in section 5. Finally, the conclusion future lines of the research are provided in section 6.

2-Problem definition

The development of the digital economy and the change in the approach and lifestyle of human has changed the way governments look at the Information Technology (IT) Industry. It is obvious that the development of digital economy in the country will increase productivity, increase transparency, promote creativity and improve the quality of government services (Covas, Silva and Dias, 2013; Daim, Bhatla and Mansour, 2013). IT industry in countries offers various services such as «provide Internet bandwidth», «provide transmission cloud bandwidth», «and provide internal Internet exchange point (IXP) bandwidth and International capacity transit». As mentioned before, due to the change in the lifestyle of people and the attitude of governments, IT industries encounter the significant increase in demand of Internet service. In addition, meeting these demands need not only providing the necessary infrastructure to meet these demands but also, establishing secure communication (Büyüközkan and şakir Ersoy, 2009). It is important to note that establishing the necessary infrastructure is time consuming and costly, which will be implemented in the form of a multi-year plan. On the other hand, due to the alluring market of providing Internet services to the main customers of IT industry and its significant profit, the demand of the other services has decreased sharply and has caused a large part of the revenues of the IT industry to be related to Internet services. Therefore, one of the important diagnosis facing the IT industry is the balancing of revenue sources. One of the methods that can be used to overcome this problem is to introduce low-demand services along with Internet service in the form of a package to the main customers. Obviously, in this situation, more discounts should be considered for Internet service customers to encourage them to buy other services. In order to complete the explanation, the following questions are asked:

- ✓ Which of the low demand services along with the Internet service is offered to the customer in the form of a package?
- ✓ How to offer the package to customers?
- ✓ How to determine if the proposed package has a positive effect on sales of low-demand services?

In this study, we developed Machine Learning methods to predict customer behavior and examine the effectiveness or ineffectiveness of the proposed packages. Also in this study, effective methods to provide the proposed packages to customers are examined and their impact on the effectiveness of the offers is also examined. Finally, using the real data of the Telecommunication Infrastructure Company (TIC), the proposed model is evaluated and the results are compared with each other.

3-Methodology

Generalized linear models (GLMs) are employed to develop linear regression models which study linear relationship between the response variable and a set of explanatory variables. GLM utilizes random component, linear predictor, and link function to evaluate non-normal response distributions and possibly nonlinear functions of the mean (Nelder and Wedderburn, 1972). This method applied iterative weighted linear regression to provide an adequate description of the data. In addition, this method can be used to establish an appropriate estimate for parameters with exponential distribution, such as the Normal, Binomial, Poisson, gamma, etc. A suitable transformation provides a linear model for this kind of parameters (Agresti, 2015). GLM is used to fit generalized linear models, specified by giving a symbolic description of the linear predictor and a description of the error distribution. GLM consists of different inputs and outputs. These outputs are exploited to analyze the proposed GLM. Table 1 summarizes main inputs and outputs.

Table 1. The explanation of GLM inputs and outputs

Main Inputs	
Formula	A typical predictor has the form response ~ terms where response is the (numeric) response vector and terms are a series of terms which specifies a linear predictor for response.
Family	A description of the error distribution and link function to be used in the model. For GLM, this can be a character string naming a family function, a family function or the result of a call to a family function. Binomial and Quasibinomial families the response can also be specified as a factor.
Data	An optional data frame, list or environment that we assign to apply in the proposed model.
Outputs of GLM	
Coefficients	A named vector of coefficients.
AIC	A version of Akaike's an Information Criterion (AIC), minus twice the maximized log-likelihood plus twice the number of parameters, computed via the aid component of the family.
Null.deviance	The deviance for the null model, compared with deviance. The null model will include the offset, and an intercept if there is one in the model.
Deviance	Up to a constant, minus twice the maximized log-likelihood. Where sensible, the constant is chosen so that a saturated model has deviance zero.
Df.null	The residual degrees of freedom for the null model.
Df.residual	The residual degrees of freedom.

4-Case study

Among the countries of the Middle East, the highest growth rate in the use of mobile phones and fixed telephones is related to Iran, and this country can be introduced as one of the countries that have huge communication networks (Badri Ahmadi, Hashemi Petrudi and Wang, 2017). One of the most important companies that has a key role in creating communication and information infrastructure in Iran is TIC. The company in the field of communication infrastructure such as core network bandwidth development, the development of Internet connection ports, development of traffic, transit network and in the field of information infrastructure, development of traffic exchange centers, and creating cloud infrastructure. According to the issues raised, it can be clear that this company is active in all areas of communication and provide services such as provide Internet bandwidth, provide transmission cloud bandwidth, Virtual Private Network (VPN), Internal Internet Exchange Point (IXP) bandwidth, and International IP transit. It is noteworthy that the share of Internet bandwidth revenue is higher than other services and this comparison is demonstrated in figure1.

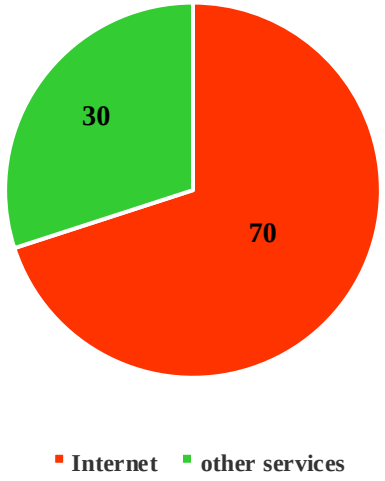

Fig1. Income shares of services

According to figure1, if the forecast of the realization of the income from the Internet service is not fulfilled, the company's income will face a significant decrease, which can even lead to the bankruptcy of the company. Therefore, the senior managers in the TIC decided to investigate this diagnostic more carefully and offer a solution to overcome this problem. One of the decisions that all managers agreed with it, providing more discounts for customers who buy the Internet service along with other services. Obviously, various combinations of services along the Internet service can be created. In this study, the effect of providing Internet service and VPN service packages compared to providing VPN service is analyzed. It is necessary to explain the details of the discounts and conditions of each of the two packages introduced, due to the confidentiality of the information, are unpublishable.

5-Results

As mentioned in the previous section, the real information and data about TIC is considered as a case study in this study. Therefore, in this section, information about the packages introduced by the company is collected and analyzed. In this section, in order to perform calculations and analysis, R software and R-studio is used in system with (Intel® Core™ i7-6500U CPU 2.50GHz), which is expressed in the following extracted results. The data of this case study consist of four main terms. The first part of data introduces the code of each of the packages "PackageID" provided to the main customers, all the created codes are unique. The second part of data is Connection, which introduces how to present the introduced packages to customers. Namely, introduced packages can be provided to customers through attendance at conferences and in-person meetings, via email or through social networks. Therefore, "Connection" in this study is divided into three categories "Direct", "Email" and "Social Network". The third part of data is information about the division of packages provided by TIC. One of the packages offered represents the VPN service and the other type of package is related to providing Internet service along with VPN service. Therefore, two categories "No Package" and "Package" are considered for "IFPackage", which "No Package" represents the VPN service package and "Package" denotes the hybrid package. The fourth part of data is about the use or non-use of the offered packages by the customers as "IFActivePackage", where the answers are categorized as "Yes" and "No". Data are categorized in Figure 2.

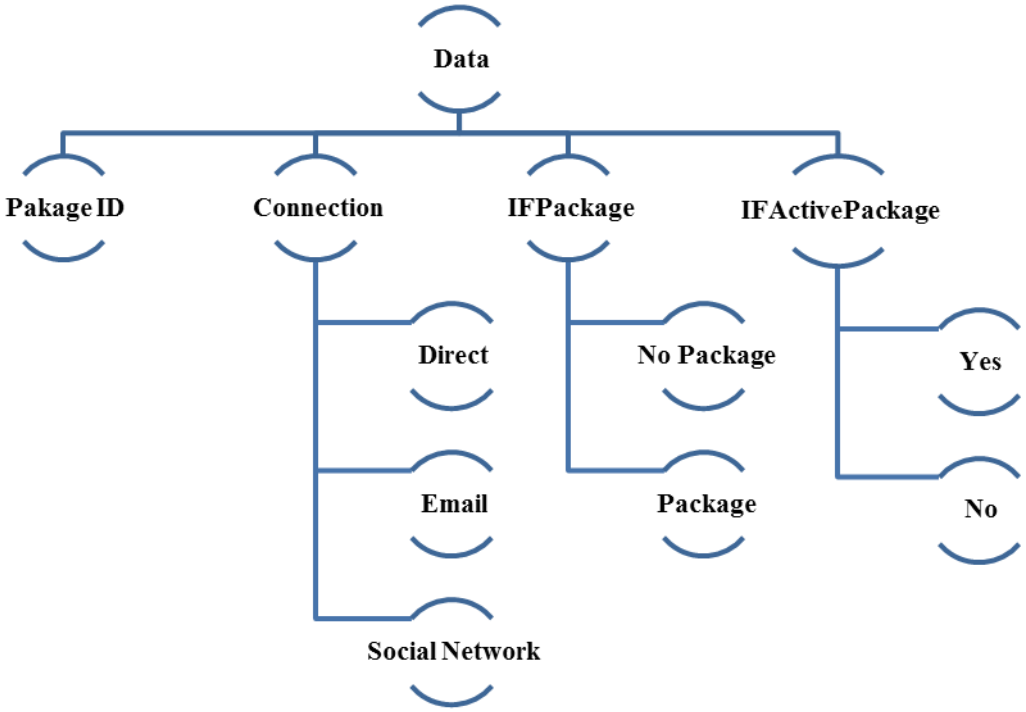

Fig 2. Category of data

In this section, an attempt has been made to examine the interaction of variables with respect to each other. First, cross tabulation of two variables "IFActivePackage" and "Connection" is investigated, which is shown in table 2.

Table2. Cross Tabulation Analysis

		Connection		
		Direct	Email	Social Network
IFActivePackage	No	290	550	727
	Yes	903	85	601

According to table 2, customers who did not use the offered packages often received this offer from social networks. On the other hand, those who have activated the offered packages have often received the offer directly. This analysis for "IFActivePackage" and "IFPackage", "Connection" and "IFPackage" is also presented in tables 3 and 4.

Table 3. Cross Tabulation Analysis

		IF Package	
		No Package	Package
IFActivePackage	No	812	755
	Yes	670	919

Table 4. Cross Tabulation Analysis

		IF Package	
		No Package	Package
Connection	Direct	333	862
	Email	512	121
	Social Network	637	691

Now, according to the information obtained, an attempt has been made to check in the form of a hypothesis test whether there is relation between "IFPackage" and "IFActivePackage" or not?
Therefore, hypotheses test is considered as follow equation1:

$$\begin{Bmatrix} H0: "IFActivePackage" \text{ is independent of } "IFPackage". \\ H1: "IFActivePackage" \text{ is not independent of } "IFPackage". \end{Bmatrix} \quad (1)$$

Using the chi-square test, we analyze apply hypotheses test. The P-value of chi-square test is 4.552e-08, which, since it is less than 0.05, rejects H0 and accept H1. Therefore, it can be concluded that there is a logical relationship between "IFPackage" and "IFActivePackage", which is also confirmed by statistical tests.
We can also define the hypothesis test to examine the effect of "Connection" on "IFActivePackage" as follows:

$$\begin{Bmatrix} H0: "IFActivePackage" \text{ is independent of } "Connection". \\ H1: "IFActivePackage" \text{ is not independent of } "Connection". \end{Bmatrix} \quad (2)$$

The P-value for this test is less than 3.451e-13 and due to being smaller than 0.05, we reject *H0* and accept *H1*. Therefore, it can be said that "Connection" and "IFPackage" have an effect on" IFActivePackage".
Now, owning to that there is relation between "Connection" and "IFActivePackage", "IFPackage" and " IFActivePackage", we create GLM for this case study as follows:

$$IFActivePackage = \beta * "IFPackage" + \varepsilon \quad (3)$$

According to equation 3, "IFPackage" is defined as predictor variable and "IFActivePackage" as Predictable variable. Also, β shows the regression coefficient "IFPackage" and ε shows deviance of GLM. According to the results of the GLM, β value of "IFPackagePackage" is 0.34479, which indicates that "IFPackagePackage" has a positive effect on "IFActivePackage". Other details of GLM are demonstrated as follow:

Table 5. GLM outputs

	Deviance Residuals	degrees of freedom
Null deviance	4375.0	3155
Residual deviance	4325.4	3154

According to table 5, it is clear that the deviance of GLM is less than the chance model. Also, the value of Akaike's an Information Criterion (AIC) for GLM is 4131.4. It is necessary to mention about AIC. When comparing models fitted by maximum likelihood to the same data, the smaller the AIC, is the better the fit. In order to confirm that the proposed GLM works better than the chance model, we perform the hypothesis test as follows:

$$\begin{cases} H0: \text{ the model is not better than chance at predicting the " } IFActivePackage \text{ ".} \\ H1: \text{ the model is better than chance at predicting the " } IFActivePackage \text{ ".} \end{cases} \quad (4)$$

According to the results obtained from the chi-square test, the P-value is 5.302866e-08, which indicates that the *H0* is rejected. Therefore, it can be concluded that the superiority of GLM over the chance model is also confirmed by the chi-square test.

As previously discussed, one of the predicator variables whose cross over effect was confirmed to be is "Connection.″

In the new GLM, we also add "Connection" to the model and analyze the results. The new GLM is also expressed as follow:

$$IFActivePackage = \beta_1 * "IFPackage" + \beta_2 * " Connection" + \varepsilon \quad (5)$$

According to the results of the GLM, β_1 value of "IFPackagePackage" is (-0.56022), which indicates that "IFPackagePackage" has negative effect on "IFActivePackage". Also, β_{21} of "ConnectionDirect" and β_{22} of "ConnectionEmail" are (1.41234) and (-2.23651). Other details of GLM are demonstrated as follow:

Table 6. GLM outputs

	Deviance Residuals	degrees of freedom
Null deviance	4375.0	3155
Residual deviance	3490.2	3152

According to table 6, the deviation of proposed GLM has been drastically reduced, which indicates the improvement of the proposed model, and the AIC has been 3475.2, which has decreased and improved compared to the initial GLM.

According to the obtained results, in the initial GLM, the coefficient of "IFPackagePackage" is positive, and in the second GLM, this coefficient is negative, which indicates the existence of conflict in the results. Namely, when the GLM was created with "IFPackage", the coefficient of "IFPackagePackage" was positive, and when "ConnectionEmail" was added to the GLM, this coefficient was negative. Here, in view of the conflict, the Simpson paradox arises.

The Simpson's Paradox occurs when at least three variables are interacted in the problem. The variable that is explained ("IFActivePackage"), the variable that is explainer ("IFPackage"), and the third variable ("Connection") that, although effective, the effect is neglected. The Simpson's Paradox is obtained when the effect of the explainer variable on the explained variable is reversed by considering the third variable (Simpson, 1949; Bandyoapdhyay *et al.*, 2011). Therefore, due to the conflict raised in the GLM, the relation between predictor variables must be considered in the GLM so that we can remove the Simpson paradox from the model. Finally, the final GLM is introduced as follows:

$$IFActivePackage = \beta_1 * "IFPackage" + \beta_2 * " Connection" + \beta_3 * "IFPackage" : " Connection" + \varepsilon \quad (6)$$

After solving the final GLM, the regression coefficients become logical and are presented in table 7:

Table 7. regression coefficients of final GLM

No	regression coefficients	Value	P-value	significant
1	IFPackagePackage	-0.87379	9.88e-15	*
2	ConnectionDirect	1.50145	< 2e-16	*
3	ConnectionEmail	-3.14401	< 2e-16	*
4	IFPackagePackage:ConnectionDirect	0.16937	0.41255	
5	IFPackagePackage:ConnectionEmail	2.98084	< 2e-16	*

In the GLM, the default mode of ("IFPackage") is considered Package and default mode of "Connection" is considered SocialNetwork. According to Table 7, the regression coefficient of "IFPackagePackage" is negative, which is interpreted as having a negative effect on the model output when the package is presented to customers through SocialNetwork. Regression coefficient of "ConnectionDirect" is positive. Namely, when packages are offered directly to customers, it has a positive effect on the use of the offer. The reason can be stated that usually when there is a face-to-face meeting with each other, it will have a greater impact on customers. But on the other hand, the regression coefficient of "ConnectionEmail" is negative, which indicates the negative effect of on the output variable. Due to the Simpson's Paradox in the proposed model, the relationship between predictor variables is also considered in the model. Therefore, when the regression coefficient of "IFPackagePackage: ConnectionEmail" is negative, it means that when the hybrid package is sent to customers via official email, it will have a significant positive effect on the output variable. The details of final GLM are also collected in table 8.

Table 8. GLM outputs

	Deviance Residuals	degrees of freedom
Null deviance	4375.0	3155
Residual deviance	3393.5	3150

As can be seen in table 6, the deviation of final GLM has also been significantly reduced, which indicates an improvement in proposed GLM model. In order to confirm that the proposed GLM works better than the chance model, we perform the hypothesis test as follows:

$$\begin{cases} H0: \text{ the model is not better than chance at predicting the " } IFActivePackage\text{".} \\ H1: \text{ the model is better than chance at predicting the " } IFActivePackage\text{".} \end{cases} \quad (7)$$

As it was predictable, the final GLM has a high capability, and after performing the chi-square test, the P-value is considered close to zero, which is a rejection of H0 and acceptance of the H1. Also, the value of AIC for the final GLM is 3211.5, which has been improved compared to other models in this output.

Therefore, it can be concluded that the best proposed model is related to final GLM, which is better and more acceptable than other models in every respect.

A comparison between the three GLM presented in the Table 9:

Table 9. Comparison between the three GLM

GLM	Deviance	Degrees of Freedom	AIC
1	4345.4	3154	4131.4
2	3490.2	3152	3475.2
3	3393.5	3150	3211.5

6-Conclusion

Nowadays, telecommunication services have a crucial role in many social and economic areas. As mentioned earlier, a large part of the revenues of the IT industry be related to Internet services. Therefore, a change in the trend of internet service demand can lead to a dramatic decrease in profits. Hence, offered service by IT industry was analyzed to balance revenue sources. In this work, introducing low-demand services along with Internet service in the form of a package to the main customers was proposed as a potential solution. The best-known machine learning algorithm, Generalized Linear Model (GLM), was employed to analyze the impact of the three ways of package introduction to customers and the type of the offered packages on usage of suggested packages. Validation of applicability of our study is confirmed by a case study of a Telecommunication

company in Iran**.** The data of this case study consists of four main terms "PackageID", "Connection", "IFPackage", and "IFActivePackage". By Chi-square test, we show that "Connection" and "IFPackage" have an effect on" IFActivePackage". According to this result, GLM was created for case study. First, "IFPackage" is considered as a predictor variable of "IFActivePackage". The results present positive effect of "IFPackagePackage" on "IFActivePackage". Then, "Connection" was added to the model. Surprisingly, Analysis indicates the negative effect of "IFPackagePackage" in "IFActivePackage". Next, conflict outcome leads us to consider Simpson paradox. According to the Simpson's Paradox, the relationship between predictor variables ("Connection". "IFPackage") is added to previous GLM as the third variable. Moreover, in each examines, the superiority of GLM over the chance model is also confirmed by the chi-square test. Our finding shows hybrid package is sent to customers via the official email have a significant positive effect on the output variable. Finally, in future investigations, it might be possible to:

1. Use other machine learning methods and compare their performance.
2. Determine other predictor variables that can be effective in evaluating the model.
3. Evaluate model using time series techniques.

Acknowledgments

This research is the result of the efforts of many people, which is presented in the form of an 11 pages article. Special thanks to all the senior managers of the Infrastructure Communications Company for working with our team.

References

Agresti, A. (2015) *Foundations of linear and generalized linear models*. John Wiley \& Sons.

Badri Ahmadi, H., Hashemi Petrudi, S. and Wang, X. (2017) 'Integrating sustainability into supplier selection with analytical hierarchy process and improved grey relational analysis: A case of telecom industry.', *International Journal of Advanced Manufacturing Technology*, 90.

Bandyoapdhyay, P. S. *et al.* (2011) 'The logic of Simpson's paradox', *Synthese*, 181(2), pp. 185–208.

Büyüközkan, G. and şakir Ersoy, M. (2009) 'Applying fuzzy decision making approach to IT outsourcing supplier selection', *system*, 2, p. 2.

Carrera, Á. *et al.* (2014) 'A real-life application of multi-agent systems for fault diagnosis in the provision of an Internet business service', *Journal of Network and Computer Applications*, 37, pp. 146–154.

Chen, X. *et al.* (2020) 'A Novel Fault Diagnosis Method for High-Speed Railway Turnout Based On DCAE-Logistic Regression', in *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pp. 318–323.

Covas, M. T., Silva, C. A. and Dias, L. C. (2013) 'Multicriteria decision analysis for sustainable data centers location', *International Transactions in Operational Research*, 20(3), pp. 269–299. doi: 10.1111/j.1475-3995.2012.00874.x.

Daim, T. U., Bhatla, A. and Mansour, M. (2013) 'Site selection for a data centre--a multi-criteria decision-making model', *International Journal of Sustainable Engineering*, 6(1), pp. 10–22.

Echraibi, A. *et al.* (2020) 'An Infinite Multivariate Categorical Mixture Model for Self-Diagnosis of Telecommunication Networks', in *2020 23rd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN)*, pp. 258–265.

Emami, M. *et al.* (2020) 'Generalization error of generalized linear models in high dimensions', in *International Conference on Machine Learning*, pp. 2892–2901.

Félix, A., Garc\'\ia, N. and Vera, R. (2020) 'Participatory diagnosis of the tourism sector in managing the crisis caused by the pandemic (COVID-19)', *Revista Interamericana de Ambiente y Turismo*, 16(1), pp. 66–78.

Hossain, M. L., Abu-Siada, A. and Muyeen, S. M. (2018) 'Methods for advanced wind turbine condition monitoring and early diagnosis: A literature review', *Energies*, 11(5), p. 1309.

Khatib, E. J. *et al.* (2015) 'Data mining for fuzzy diagnosis systems in LTE networks', *Expert Systems with Applications*, 42(21), pp. 7549–7559.

Kim, J. S. *et al.* (2018) 'Development of data-driven in-situ monitoring and diagnosis system of fused deposition modeling (FDM) process based on support vector machine algorithm', *International Journal of Precision Engineering and Manufacturing-Green Technology*, 5(4), pp. 479–486.

Ko, R. K. L., Lee, S. S. G. and Lee, E. W. (2009) 'Business process management (BPM) standards: a survey', *Business Process Management Journal*.

Kohlbacher, M. (2010) 'The effects of process orientation: a literature review', *Business process management journal*.

Lei, Y. *et al.* (2020) 'Applications of machine learning to machine fault diagnosis: A review and roadmap', *Mechanical Systems and Signal Processing*, 138, p. 106587.

Liu, P. *et al.* (2020) 'Optimization of Edge-PLC-Based Fault Diagnosis With Random Forest in Industrial Internet of Things', *IEEE Internet of Things Journal*, 7(10), pp. 9664–9674.

Nelder, J. A. and Wedderburn, R. W. M. (1972) 'Generalized linear models', *Journal of the Royal Statistical Society: Series A (General)*, 135(3), pp. 370–384.

Quiñones-Grueiro, M., Llanes-Santiago, O. and Neto, A. J. S. (2021) 'Fault Diagnosis in Industrial Systems', in *Monitoring Multimode Continuous Processes*. Springer, pp. 1–14.

De Ramon Fernandez, A., Ruiz Fernandez, D. and Sabuco Garcia, Y. (2020) 'Business Process Management for optimizing clinical processes: A systematic literature review', *Health informatics journal*, 26(2), pp. 1305–1320.

Shahabi, H. *et al.* (2020) 'Flood detection and susceptibility mapping using sentinel-1 remote sensing data and a machine learning approach: Hybrid intelligence of bagging ensemble based on k-nearest neighbor classifier', *Remote Sensing*, 12(2), p. 266.

Simpson, E. H. (1949) 'Measurement of diversity', *nature*, 163(4148), p. 688.

Sorrentino, M. *et al.* (2019) 'A Novel Energy Efficiency Metric for Model-Based Fault Diagnosis of Telecommunication Central Offices', *Energy Procedia*, 158, pp. 3901–3907.

Soualhi, A. and Razik, H. (2020) 'Diagnostic Methods for the Health Monitoring of Gearboxes', *Electrical Systems 1: From Diagnosis to Prognosis*, pp. 1–43.

Sun, Y. *et al.* (2020) 'A new convolutional neural network with random forest method for hydrogen sensor fault diagnosis', *IEEE Access*, 8, pp. 85421–85430.

Tetteh, V. K. (2012) *Organisational Diagnosis--A Management Tool for Change in the Telecommunication Industry*.

Varela-Vaca, Á. J. *et al.* (2019) 'Automatic verification and diagnosis of security risk assessments in business process models', *IEEE Access*, 7, pp. 26448–26465.

Wang, F. *et al.* (2020) 'Neural cognitive diagnosis for intelligent education systems', in *Proceedings*

of the AAAI Conference on Artificial Intelligence, pp. 6153–6161.

Wardani, S., Sihombing, P. and others (2020) 'Hybrid of Support Vector Machine Algorithm and K-Nearest Neighbor Algorithm to Optimize the Diagnosis of Eye Disease', in *2020 3rd International Conference on Mechanical, Electronics, Computer, and Industrial Technology (MECnIT)*, pp. 321–326.

Wu, J.-Y. and Hsiao, H.-I. (2021) 'Food quality and safety risk diagnosis in the food cold chain through failure mode and effect analysis', *Food Control*, 120, p. 107501.

Yao, J. and Ye, Y. (2020) 'The effect of image recognition traffic prediction method under deep learning and naive Bayes algorithm on freeway traffic safety', *Image and Vision Computing*, 103, p. 103971.