iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.43869

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Analysis of Prudential Life Insurance Customer Evaluator

Soham Mitra[1], Soumya Banerjee[2]
[1]*Department of Statistics, Amity University Kolkata, West Bengal, India*
[2]*Assistant Professor, Department of Statistics, Amity University Kolkata, West Bengal, India.*

Abstract: ***Life Insurance is one the most important part in today's world. And with many varieties of insurance policies, the objective is to provide the most suitable policy which is beneficial to both the customer and the insurance company. This is generally done by the insurance agents based on their domain knowledge. But our goal here is to do it using a statistical solution. One method is to use Customer Scoring based on information provided by them when filling their insurance application form. Based on this score they are provided a set of insurance policies to select from. For this we tried building regression models (Linear and Multinomial Logistic) based on the customer information and scores we already have. These scores were provided based on the policy they already have. Different models based on different variable combinations were compared using Stepwise AIC Method for both regression models. The final model has an accuracy of 44% which can of course further be improved. This kind of statistical modelling will be useful in filtering the large number of policies to select from. After which the customer or agent may select from the smaller number of choices suitable for them. This will make the job of the agents as well as the customer much easier.***

Keywords: Life Insurance, Customer Scoring, Statistical Solution, Logistic Regression, Multinomial Logistic Regression, Stepwise AIC Method

I. INTRODUCTION

The insurance is a contract or an agreement between the insurer and insured, whereby the insurer promises to pay benefits (sum of money) to the insured person or on their behalf to a third party in exchange for a premium, if certain defined events occur. Life Insurance covers such certain defined event like death of the policyholder under given conditions. However, if policy holder remains alive till the maturity period of the insurance policy, then the policy holder gets only maturity benefit and bonus sum. Suneja and Sharma (2009) in their work *"Factors Influencing Choice of a Life Insurance Company"* have discussed the factors influencing the choice of a life insurance company. The insurance applicant must fill an application form that includes certain personal information. This includes applicants' demographic information, medical information, income and family information. The insurance provider or insurance company recommends the best suitable insurance policy to the beneficiary and then processes the app. Chaudhury and Kaur (2016) in their work "*Consumer Perception Regarding Life Insurance Policies: A Factor Analytical Approach"* have studied the attributes consumers consider while purchasing insurance policies.

This study aims at to reduce the application processing time and effort, by providing a statistical model to assign a risk score to the applicants (Customer Evaluation) and thus provide the appropriate insurance policy based on the scores.

II. DATA DESCRIPTION

Data has been taken from sample provided by an American Insurance and financial products company: ***Prudential Financial.*** The data is collected from test data uploaded at www.kaggle.com. It was provided with a reward offer for providing a statistical solution to reduce application process time and labour.

The provided dataset consists of 128 variables. The response variable (risk score) is a categorical variable (ordinal data) with eight levels of risk.

The data provided may contain sensitive customer data and the customer may have privacy issues with data such as medical history or income being shared too public. Thus, to overcome this issue the company provided data such as medical records, family history etc. in the form of masked and normalized variables. For example, Medical_History1 to Medical_History41. This variable set of Medical_History in actuality contains 41 different medical parameters whose names have been masked to maintain privacy of customers.

©IJRASET: All Rights are Reserved

Some variables were removed while cleansing the dataset. In addition, few variables provided in the data were already normalized, e.g., normalized age.

Below is the description for each data element.

Table1: Data Description

Variable	Description
Id	A unique identifier associated with an application.
Product_Info_1-7	A set of normalized variables relating to the product applied for
Ins_Age	Normalized age of applicant
Ht	Normalized height of applicant
Wt	Normalized weight of applicant
BMI	Normalized BMI of applicant
Employment_Info_1-6	A set of normalized variables relating to the employment history of the applicant.
InsuredInfo_1-6	A set of normalized variables providing information about the applicant.
Insurance_History_1-9	A set of normalized variables relating to the insurance history of the applicant.
Family_Hist_1-5	A set of normalized variables relating to the family history of the applicant.
Medical_History_1-41	A set of normalized variables relating to the medical history of the applicant.
Medical_Keyword_1-48	A set of dummy variables relating to the presence of/absence of a medical keyword being associated with the application.
Response	This is the target variable, an ordinal variable relating to the final decision associated with an application

The following variables are all categorical (nominal):
Product_Info_1, Product_Info_2, Product_Info_3, Product_Info_5, Product_Info_6, Product_Info_7, Employment_Info_2, Employment_Info_3, Employment_Info_5, InsuredInfo_1, InsuredInfo_2, InsuredInfo_3, InsuredInfo_4, InsuredInfo_5, InsuredInfo_6, InsuredInfo_7, Insurance_History_1, Insurance_History_2, Insurance_History_3, Insurance_History_4, Insurance_History_7, Insurance_History_8, Insurance_History_9, Family_Hist_1, Medical_History_2, Medical_History_3, Medical_History_4, Medical_History_5, Medical_History_6, Medical_History_7, Medical_History_8, Medical_History_9, Medical_History_11, Medical_History_12, Medical_History_13, Medical_History_14, Medical_History_16, Medical_History_17, Medical_History_18, Medical_History_19, Medical_History_20, Medical_History_21, Medical_History_22, Medical_History_23, Medical_History_25, Medical_History_26, Medical_History_27, Medical_History_28, Medical_History_29, Medical_History_30, Medical_History_31, Medical_History_33, Medical_History_34, Medical_History_35, Medical_History_36, Medical_History_37, Medical_History_38, Medical_History_39, Medical_History_40, Medical_History_41

The below mentioned variables are continuous:

Product_Info_4, Ins_Age, Ht, Wt, BMI, Employment_Info_1, Employment_Info_4, Employment_Info_6, Insurance_History_5, Family_Hist_2, Family_Hist_3, Family_Hist_4, Family_Hist_5

The following variables are discrete:

Medical_History_1, Medical_History_10, Medical_History_15, Medical_History_24, Medical_History_32, Medical_Keyword_1-48 are dummy variables.

III. METHODOLOGY

The Data Pre-processing was handled in the following six stages:

1) Handling Missing Data:

a) Deleted columns: I deleted columns/variables containing more than 40% missing values.

b) Replaced missing values: I replaced the missing values with the column median in the remaining missing data columns.

2) Dummy variable treatment: The categorical variables were broken down into dichotomous ones and dummy variables were formed to make data handling easier.

3) Normalization: The variables that required scaling were already provided as normalized variables.

4) Outlier Treatment: The outliers for each column were calculated using boxplots and then replaced by the column median. (Note: We use the median because it is not greatly affected by the value of the outlier. Thus, if we replace the outlier data it helps in avoiding loss of other valuable parameters.)

5) Dimensionality Reduction: It is the process of reducing the number of random variables by considering the principal variables, which are of utmost importance. This way we reduce the number of columns. This was performed using three filtering methods:

a) Missing value ratio (1st Stage)

b) Low variance filter: Variables with extremely low variance are completely removed since it does not affect the model in a significant way.

c) P-value check: Checking and removal of the variables that were least affecting the response variable.

6) Split the 59831 rows of clean data into Training set and testing set in a 70:30 ratio.

In the raw form, it had 127 independent variables that were brought down to 36. The following is the list of independent variables used in creating the model:

A. *Continuous*

Product_Info_4, Ins_Age, Ht, Wt, Employment_Info_1, Employment_Info_6, Family_Hist_4, Medical_History_1,
Discrete:
Id, Medical_Keyword_25, Medical_Keyword_37

B. *Categorical*

Product_Info_2.A8, Product_Info_2.D1, Product_Info_2.D3, Product_Info_2.D4, Product_Info_3.10, Product_Info_6.1, Employment_Info_2.1, Employment_Info_2.9, Employment_Info_2.14, InsuredInfo_3.2, InsuredInfo_3.3, InsuredInfo_3.6, InsuredInfo_3.8, InsuredInfo_3.11, InsuredInfo_4.2, InsuredInfo_6.1, Insurance_History_1.1, Insurance_History_3.1, Insurance_History_7.3, Medical_History_2.112, Medical_History_2.491, Medical_History_4.1, Medical_History_26.2, Medical_History_26.3, Medical_History_39.1

IV. MODELLING

A. *Linear Regression*

The first step to fit a model was to apply linear regression in R on the training dataset.

B. *Standard Error*

The standard error is the standard error of our estimate, which allows us to construct marginal confidence intervals for the estimate of that particular feature. If s.e.(B_i) is the standard error and B_i is the estimated coefficient for variable i, then a 95% confidence interval is given by $B_i + 1.96*$s.e.(B_i). Note that this requires two things for this confidence interval to be valid:

- The model assumptions hold.
- The have enough data/samples to invoke the central limit theorem, as we need B_i to be approximately normal.

That is, assuming all model assumptions are satisfied, we can say that with 95% confidence the true parameter B_i.
From the C.I., we calculated we can see Medical_History_26.2 and Medical_History_26.3 have a much larger effect size than compared to the other variables.

C. Multiple and Adjusted R^2

These tell us about how good a fit the model is and whether any of the coefficients are significant. R^2 tells us what proportion of the variance is explained by our model, and is given as follows:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$
$$= 1 - \frac{\sum_i \hat{\epsilon}_i^2}{\sum_i (y_i - \bar{y})^2}$$

Both R^2 and the residual standard deviation tells us about how well our model fits the data.
The adjusted R^2 deals with an increase in R^2 spuriously due to adding features, essentially fitting noise in the data. It is given by

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

thus, as the number of variables p increases, the required R^2 needed will increase as well to maintain the same adjusted R^2.

Next, we went for Stepwise AIC method to select a model by eliminating each variable and checking the Akaike Information Criterion (AIC) value. The best model will be with the lowest AIC score.
AIC= 2*k - 2*ln(L)
Where 'k' is the no. of independent variables estimated and 'L' is the maximum value of the Likelihood function of the model.

Finally, the model was tested with the testing dataset by predicting the response based on the model.
Multinomial models are linear statistical models for which the response variable is a factor with more than two levels. These models (also termed as generalized logit models) are extensions to the more familiar binomial regression models (logistic regression or logit models). Multinomial models are linear statistical models for which the response variable is a factor with more than two levels. These models (also termed as generalized logit models) are extensions to the more familiar binomial regression models (logistic regression or logit models).
Multinomial models are linear statistical models for which the response variable is a factor with more than two levels. These models (also termed as generalized logit models) are extensions to the more familiar binomial regression models (logistic regression or logit models).
Multinomial models are linear statistical models for which the response variable is a factor with more than two levels. These models (also termed as generalized logit models) are extensions to the more familiar binomial regression models (logistic regression or logit models).
Multinomial models are linear statistical models for which the response variable is a factor with more than two levels. These models (also termed as generalized logit models) are extensions to the more familiar binomial regression models (logistic regression or logit models).
Multinomial Logistic Regression: We used R to fit a Multinomial Logistic Regression model to the training dataset.
Multinomial models are linear statistical models for which the response variable is a factor with more than two levels. These models (also called generalized logit models) are extensions to the more familiar binomial regression models (logistic regression or logit models).
In a traditional logistic regression model, the response variable is a discrete variable that comes either as a binary response (zeroes and ones) or as a binomial response (number of successes and failures given number of trials). For example, a binary response variable could be coded as 0 for healthy and 1 for ill.

Now, in a multinomial model, these principles are just extended to the multi-category case; that is, the response variable has more than two response categories. Importantly, these models usually work with a baseline category, that is, one of the categories is selected to be the baseline to which all other categories are then compared. A multinomial model can essentially be expressed as a series of individual logistic regression models. Mathematically, each response category enters with the baseline category in a logit transformation, such that: Where η_j is the linear predictor for response category j, $ß_0$ and $ß_1$ are the intercept and the slope respectively, and x is a numeric explanatory variable. The linear predictor η_j is related to the explanatory variables using the logit link function:

logit = log[p/(1 - p)]

The multinomial regression can be thought of as multiple independent logistic regression models to compare the same baseline. Like good v/s. bad and good v/s. very good, where the baseline is good and bad, good and very good are the three qualities have been compared.

D. Iteration Values

First to be mentioned is the iteration count in the regression process. Multinomial logistic regression uses maximum likelihood estimation which is an iterative process. The first iteration (called iteration 0) is the log likelihood of the "null" or "empty" model; that is, a model with no predictors. At the next iteration, the predictor(s) are included in the model. At each iteration, the log likelihood increases because the goal is to maximize the log likelihood. When the difference between successive iterations is very small, the model is said to have "converged", the iterations are stopped.

E. Coefficients

These are the multinomial logistic regression coefficients. An important feature of the multinomial logit model is that it estimates k-1 models, where k is the number of levels of the outcome variables. Since the parameter estimates (coefficients) are relative to the referent group, the standard interpretation of the multinomial logit is that for a unit change in the predictor variable, the logit of outcome m relative to the referent group is expected to change by its respective parameter estimate (which is in log-odds units) given the variables in the model are held constant.

F. Assessing Overall fit and Significance

For a multinomial logistic model there is no specific or very appropriate measure for understanding the overall fit of the model unlike linear regression which has multiple and adjusted R^2 values that directly explain how much of the variance can be explained by the estimators through the model.

Next, we find out and list the most important variables in the model from most important to least important based on just the logistic model we built. For classification, ROC curve analysis is conducted on each predictor. For two class problems, a series of cutoffs is applied to the predictor data to predict the class. The sensitivity and specificity are computed for each cutoff and the ROC curve is computed. The trapezoidal rule is used to compute the area under the ROC curve. This area is used as the measure of variable importance. For multi-class outcomes, the problem is decomposed into all pair-wise problems and the area under the curve is calculated for each class pair (i.e. class 1 vs. class 2, class 2 vs. class 3 etc.). For a specific class, the maximum area under the curve across the relevant pair-wise AUC's is used as the variable importance measure.

Based on this we eliminate the least important factor and retrain the model. Hence compare the AIC scores. Now, The Akaike Information Criterion (AIC) has the same role here as before, just a no. to compare two or more related models.

The testing of this model was done with the same test dataset as before. we used the same method to calculate accuracy rate as the linear model.

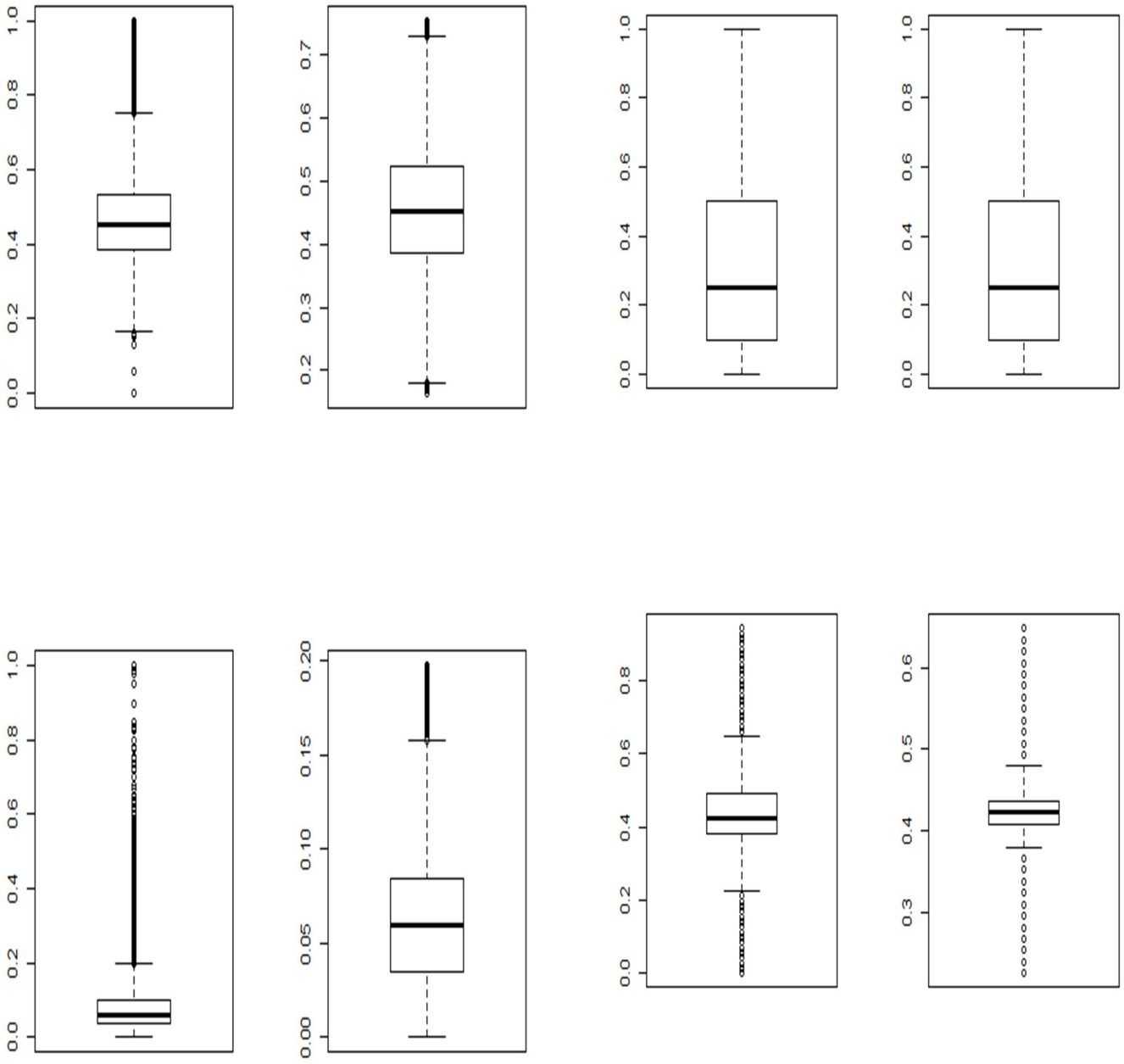
V. RESULTS AND DISCUSSION:

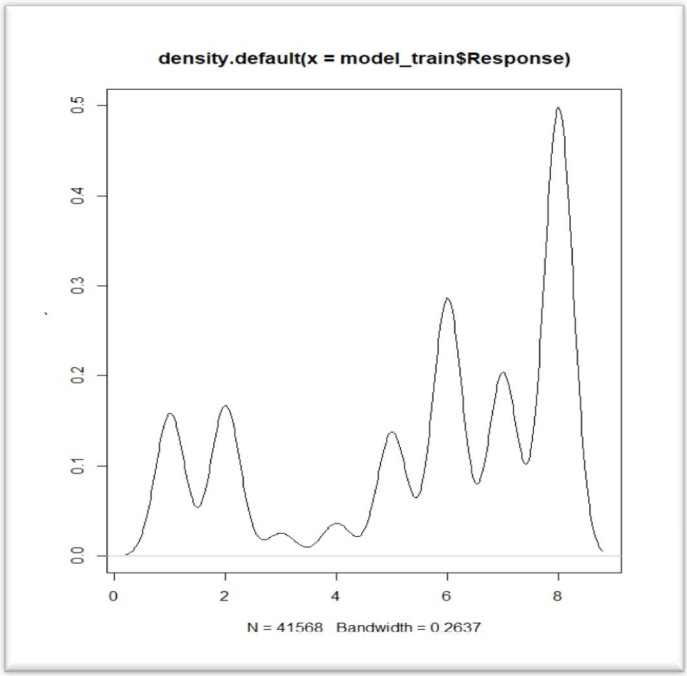
A. Data Pre-Processing

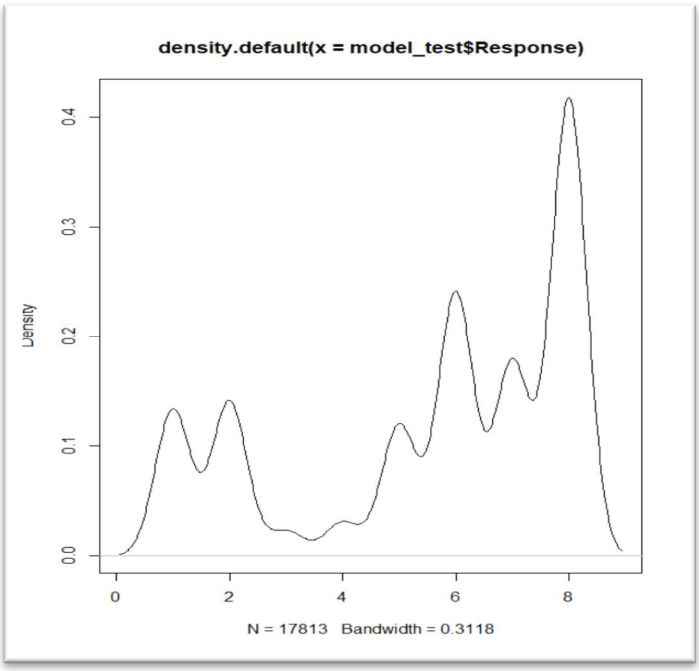
Although cleansing of the data and reducing the dimension may produce mathematically satisfactory results but in all practical sense, we can find many gaps in these processes. For example, the variable BMI has been removed in the reduction process although it makes more sense to keep BMI in the place of height and weight separately. BMI is ultimately the ratio of height and weight so using it would have been enough but due to the mathematical nature of the cleansing process these small practical aspects go unnoticed.

The categorical variables like InsuredInfo_4.2 or Product_Info_2.D4 etc. were previously not there. These were created after a disjunctive table was created to break down the categorical variables into dichotomous ones. For example, InsuredInfo_4.2 indicates the presence of the second category of InsuredInfo_4 or not through zero or one.
Some comparative boxplots are given below to show the effects after and before imputing the outliers of the continuous variables with the median value. In the raw form, it had 127 independent variables that were brought down to 36.

The following figure shows the boxplots of some factors before and after removal of outliers:
Fig:1 Boxplots







B. Modelling

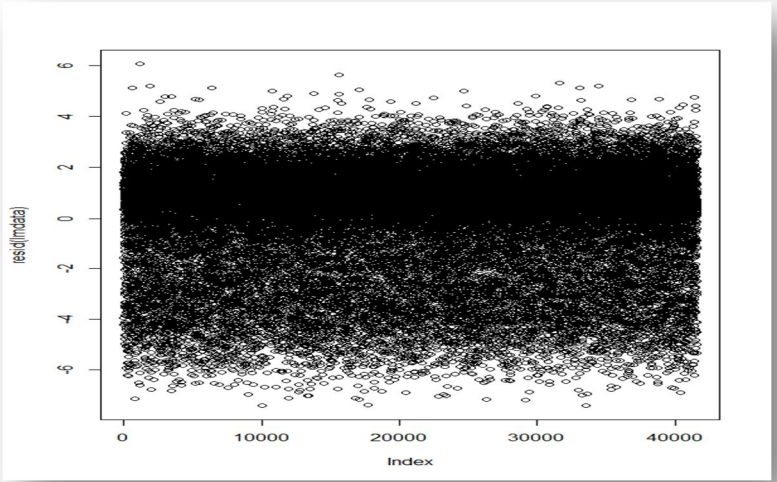
In our case the R^2 value for the linear regression is very low (only 24%). Also, the model has a low Variance inflation factor (VIF<10) so the chance of multicollinearity is very low.

After applying the stepwise AIC method, there is no improvement in the model i.e., the initial or maximal model seems to have the lowest AIC score (63153.57).

Testing the model yielded results with very low accuracy rates (17%) calculated by taking the average of the total no. of predicted values that match the actual response of the test data.

Plots: The following plots depicts similarities between train and test data sets.

Residual Plot after linear regression



In case of the Multinomial Logistic Regression, it converges after the 260th iteration with the log-likelihood of 62373.264338.

Factors affecting Customer Score sorted by their importance level are given below:

Table 2: Factors affecting Customer score and their importance

Variables	Importance
Medical_History_26.3	65.0630301
Medical_History_26.2	64.6992571
Wt	51.6472344
Ht	27.2481230
Employment_Info_1	24.3381443
Insurance_History_7.3	17.1188694
Ins_Age	15.8734389
Insurance_History_3.1	15.7607742
Medical_History_4.1	8.4903306
Medical_History_39.1	7.8913497
Family_Hist_4	6.4524806
Product_Info_4	6.0613013
Medical_Keyword_37	3.8315263
Employment_Info_2.1	2.9903341
Product_Info_2.D1	2.8572687
Product_Info_2.D4	2.5492367
Employment_Info_2.14	2.5010811
InsuredInfo_6.1	2.1409256
Product_Info_2.A8	2.0644041
Employment_Info_6	1.9294022
InsuredInfo_3.11	1.8039393
Medical_History_2.491	1.6390076
Product_Info_3.10	1.4674741
InsuredInfo_3.2	1.4479890
InsuredInfo_4.2	1.3081583
Employment_Info_2.9	1.2361772
Medical_History_2.112	1.2354892
InsuredInfo_3.6	1.2081598
Insurance_History_1.1	1.1785946
Product_Info_2.D3	0.9182471
InsuredInfo_3.3	0.9102812
InsuredInfo_3.8	0.9061783
Product_Info_6.1	0.7542763
Medical_Keyword_25	0.7214543
Medical_History_1	0.1717611

©IJRASET: All Rights are Reserved

As we have seen Medical_History_1 comes out as least important or least affecting the model. So, we tried re-training the data and building another multinomial model without the variable Medical_History_1. This yielded results with 10 iterations less but higher AIC score (125441.3) than the previous model (125250.5)
Hence, we did not proceed with the model retraining any further as the present one seems to be the best multinomial logistic model that can be obtained.
Testing the model with the same test dataset yielded predictive accuracy rate of 44.04%

VI. CONCLUSION

The objective of the study is to reduce the application processing time and effort, by providing a statistical model to assign a risk score to the applicants. Data has been taken from sample provided by an American Insurance and financial products company – Prudential Financial. Different regression models (Linear and Multinomial Logistic) are developed based on the customer information and scores**.** It is observed that the Linear Regression model has a predictive accuracy of 17% while the Multinomial Logistic Regression model has a predictive accuracy of 44.04%. Hence, the latter is a pretty good improvement over the first model and is acceptable based on industry standards for model accuracy. This study has allowed us brought into light the various important factors on which choice of life insurance depends upon. Also, it shows us that automated and computer - based choices and suggestions for the customer is feasible. It shows how much important customer evaluation is to provide customers with good choices. This study paves the way for some future research. This study can be compared to far more efficient methods that can be used to model this problem. For cleaning the data, we can use Principal Component Analysis and Random Forests, while for fitting models SVM, XG-boost and decision trees may be performed to yield more accurate prediction.

A. Conflict of Interest

The author declares that there is no conflict of interest for this publication.

B. Acknowledgement

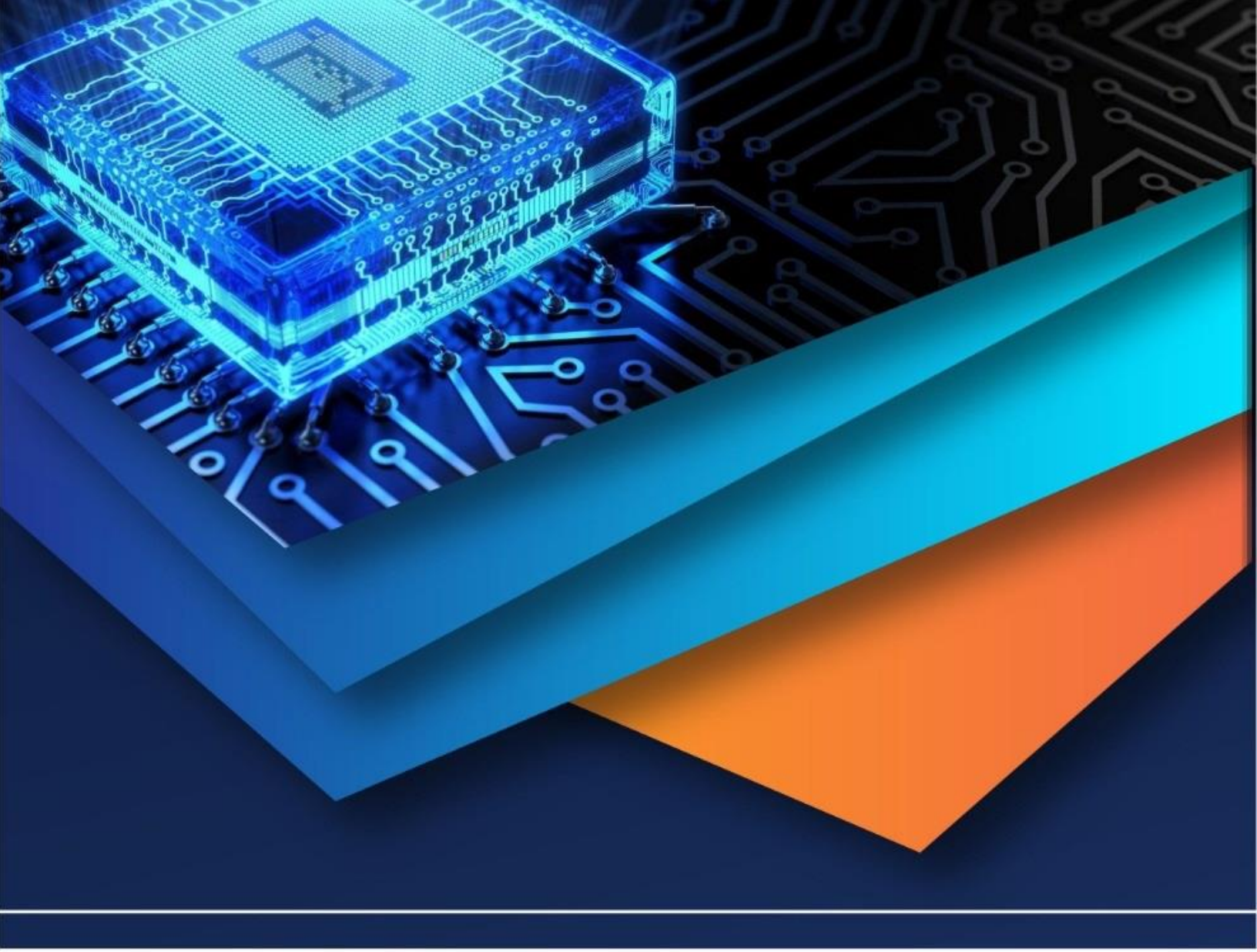
The author extends his appreciation to the anonymous reviewers for their valuable suggestions.

C. Funding

The authors did not receive support from any organization for the submitted work

REFERENCES

[1] Chaudhary S and Kaur J (2016), “Consumer Perception Regarding Life Insurance Policies: A Factor Analytical Approach”, International Journal of Information Movement, Volume I, Issue III (ISSN: 2456-0553 (online))
[2] Das S.C. and Gope A.K. (2014), “Impact of Demographic Features of Employees on HRD in Life Insurance Corporation of India: The Multinomial Logistic Regression Modeling”, Review of HRM, Vol. 3, New Delhi (ISSN: 2249-4650)
[3] Jain R, Alzubi Jafar A., Jain N and Joshi P (2019), “Assessing risk in life insurance using ensemble learning”, Journal of Intelligent & Fuzzy Systems, vol. 37, no. 2, pp. 2969-2980.
[4] Nena, S. (2013). Performance evaluation of Life Insurance Corporation (LIC) of india. International Journal, 1(7), 113-118.
[5] Nguyen, H. T., Nguyen, H., Nguyen, N. D., & Phan, A. C. (2018). Determinants of customer satisfaction and loyalty in Vietnamese life-insurance setting. Sustainability, 10(4), 1151.
[6] Suneja A and Sharma K (2009), “Factors Influencing Choice of a Life Insurance Company”, LBS Journal of Management & Research.
[7] Wu, Z., Lin, W., Zhang, Z., Wen, A., & Lin, L. (2017, July). An ensemble random forest algorithm for insurance big data analysis. In 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC) (Vol. 1, pp. 531-536). IEEE.
[8] https://www.kaggle.com/c/prudential-life-insurance-assessment/data
[9] https://www.wikipedia.org/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)