Machine Learning Model for Predicting Potential Donors Using Logistic Regression

Seow Wei Ling
School of Computing
Asia Pacific University of Technology and Innovation (APU)
Kuala Lumpur, Malaysia
tp057859@mail.apu.edu.my

Nowshath K Batcha
School of Computing
Asia Pacific University of Technology and Innovation (APU)
Kuala Lumpur, Malaysia
nowshath.kb@apu.edu.my

Rajasvaran Logeswaran
School of Computing,
Asia Pacific University of Technology and Innovation (APU)
Kuala Lumpur, Malaysia
loges@ieee.org

***Abstract*—Natural calamities like hurricanes, tsunami and pandemic are tend to happen so often in today's world. Under this scenario, predictive modelling is helpful in terms of resources allocation to achieve the objective effectively. This study intends to construct a prediction model based on logistic regression to predict the possible donors who can help in such tragic situations. Sample dataset is taken from internet source. Initial data exploration being performed to better understand the variables in dataset. To improve the quality of dataset, missing value treatment and feature engineering are performed before the construction of prediction model. During the missing value treatment, various methods being applied with mean imputation has the better performance in terms of variable significance and standard error. Feature engineering including one-hot encoding, categorical grouping, multicollinearity treatment and log transformation being performed. During the modelling phase, normal logistic regression and stepwise logistic regression being performed. The performance of the models was measured by Accuracy, Sensitivity and Specificity of the training and testing dataset. The Stepwise Logistic Regression outperformed the normal Logistic Regression with model accuracy at 58.5% along with sensitivity rate of 54.3% and specificity rate of 62.6%Keywords—Machine Learning, Recommender System & Feature Extraction.**

Keywords—machine learning, potential donors, logistic regression, multiple linear regression

I. INTRODUCTION

Prediction modelling is effective to identify possible donor and to allocate manpower in contacting people who are more likely to donate. Apart from that, predictive model such as logistic regression, decision tree or neural network can help analytical team in understand the relationship between variables in the dataset. As the target variable will be binary, logistic regression will be utilized in this study in predicting the possible donors.

Logistic regression is similar to multiple regression analysis as it utilizes one or more independent variable(s) in predicting a single target variable. Logistic regression is specialized form of regression used to predict binary categorical variable by utilizing logit model in predicting the probability of a particular even existing. Logistic regression is less affected by the heteroscedasticity issue as compared to other multivariate methods such as discriminant analysis. On the other hand, empirical results for logistic regression is easier to interpret as they are parallel with the multiple regression model's results [1].

Before building a model, data pre-processing such as data cleaning, data transformation and dimension reduction is important in improving the quality of the raw dataset. A few models based on different data pre-processing methods will being built, the model performance is then measure and compare by several criteria.

II. LITERATURE REVIEW

Evidences exists to substantiate that many studies are done in understanding the factors influencing donators' behavior in applying regression analysis. The study shows that donation seasons such as Ramadan have significant impact on the willingness to donate. On the other hand, demographic factors such as social class, marital status and education level also play important role in affecting the monetary donations. [2] performed Multiple Linear Regression (MLR), Support Vector Regression (SVR) and Artificial Neural Network (ANN) in predicting the amount of charitable giving in the following years. ANN outperformed MLR and SVR with lowest MSE of 0.01 with three most significant independent variables: population, education level and the charitable amount in previous year. Degasperi & Mainardes [3] had conducted research to understand donor's behavious in Brazil though questionnaire. They utlized exploratory factor analysis and concluded that factors such as environmental influences, personal benefits, future interest, beneficiaries' characteristics are important factors that motivate individual money donation. Snipes et al [4] concluded that charity reputation is important factor in determine the willingness of an individual in donating.

Binary classification task involves classifying the observations into two distinct groups through several different algorithms. Some common classification problems include diagnosis of certain disease, fraud detection and responses on events. There are several algorithms that are widely use when building prediction model for binary classification such as Logistic Regression, Decision Trees, Support Vector Machine, Naïve Bayes and k-Nearest Neighbours. The performance of the algorithm can be varying on different dataset. Studies often applied several machine learning algorithms when building the prediction models and further

evaluate the best model based on several criteria such as accuracy, sensitivity rate and specificity rate.

Dwivedi [5] constructed a prediction model for heart disease dataset by applying Artificial Neural Network (ANN), Support Vector Machine (SVM), Naïve Bayes, Logistic Regression, k-Nearest Neighbour and Classification Tree. Logistic Regression outperform all other models by accuracy of 85% with 89% sensitivity rate and 81% specificity rate. Another study on predictive model for heart disease [6], Decision Tree outperformed Logistic Regression with accuracy of 84%.

On the other hand, Logistic Regression is often used for credit card fraud detection along with other binary classification algorithms. Patil et al [7] built a prediction model for credit card fraud dataset with Random Forest at 76% accuracy rate outperformed Decision Tree and Logistic Regression. Similarly, another study based on European bank data set also resulting in Random Forest performs better with 95.5% accuracy by VS & Deepthi Kavila [8]. However to conclude that Logistic Regression model performed better for fraudulent prediction with accuracy of 91.2%.

There are several common approaches to improve the performance of prediction model. At data pre-processing stage, proper handling of missing value and outlier, feature engineering and feature selection enable the elimination of noise in the dataset. When building the model, hyper parameters tuning can effectively improve the performance of the predictive model. On the other hand, ensemble method is widely used to improve the predictive results by combining multiple algorithms to produce one optimal prediction model.

There are several studies done in predicting diabetes diagnosis based on Pima Indians Diabetes Dataset by logistic regression. Wu et al [9] utilized K-means Algorithm for data pre-processing and able to build logistic prediction model with accuracy of 95.42%. Zhu et al. [10] further enhanced the prediction model by applying Principal Component Analysis (PCA) for dimensionality reduction purpose. With PCA and K-means clustering, and also they build another logistic regression prediction model on same dataset with higher accuracy at 97.4%.

Taslimitehrani et al [11] improve the logistic regression prediction model for Heart Failure survival rate by combining the logistic regression model with loss function. The study successfully improves the prediction accuracy from 89% (Logistic Regression) to 91.4%. On the other hand, Vote technique by combining Logistic Regression with Naïve Bayes in building prediction model for cardiovascular disease. The Vote technique outperform other seven machine learning techniques with the highest accuracy of 87.4% which is 1.54% higher than the original logistic regression built.

III. DATASET & PRE PROCESING

Data preprocessing is done using SAS University edition. For continuous variables, univariate analysis will be performed in SAS Studio by PROC UNIVARIATE. It provides complete report on the variables that useful for data exploration purpose. For categorical variables, PROC FREQUENCY will be utilized to observe the distribution of each category for individual variable. Pie chart and bar chart for the frequency distribution for each categorical variable will be plot for better visualization purpose. The MACRO function is being created in this step to eliminate the repetitive steps required to run the PROC UNIVARIATE, PROC FREQUENCY PROC SGPLOT and PROC GCHART for each variable. The metadata of the dataset is given in Table 1.

TABLE I. METADATA OF DATASET

No.	Variable	Variable Type
1	Donor	Categorical
2	D_ID	Input
3	Donor_D	Continuous
4	DonCntP1	Discrete
5	DONCntAll	Discrete
6	DONCntCardP1	Discrete
7	DONCntCardAll	Discrete
8	DONAvgLast	Continuous
9	DONAvgP1	Continuous
10	DONAvgAll	Continuous
11	DONAvgCardP1	Continuous
12	DONTimeLast	Discrete
13	DONTimeFirst	Discrete
14	CallCntP2	Discrete
15	CallCntP1	Discrete
16	CallCntAll	Discrete
17	CallCntCardP2	Discrete
18	CallCntCardP1	Discrete
19	CallCntCardAll	Discrete
20	Donor_Status	Categorical
21	Donor_Status_Prev_Camp	Categorical
22	DemArea	Categorical
23	Age	Discrete
24	Gender	Categorical
25	DemHomeOwner	Categorical
26	AreaHomeValue	Continuous
27	AreaMedIncome	Continuous

Generally, missing values can be categories into 3 main types: missingness completely at random (MCR), missingness at random (MAR), and missingness that depends on unobserved predictors (MNAR). MCAR indicates that the missing value pattern is completely random and unrelated to all the variables including itself. MAR refers to missing value pattern that is unrelated to the missing variable itself but somehow related to other variables in the model. On the other hand, MNAR implies that the missing pattern is related to the variable itself. It is important to understand the type of missing data before deciding on the missing value treatment to be implement.

There are several more commonly used missing value treatment such as Complete Case (CC), single imputation and multiple imputation. Complete Case method removes observations with missing values where single imputation replace the missing value without defining explicit model. Multiple imputation will be more complex as it is a

simulation-based process which includes imputation phase, analysis phase and pooling phase. However, the efficiency of missing value handling methods might differ for different type of missing value. The percentage of missing value also serves as another important indicator in deciding the missing value treatment options..

Complete case analysis or listwise deletion delete observations with missing value(s). This is easy to apply but it might reduce the statistical power of a dataset as the observations become lesser. Complete Case (CC) should only being applied to missing value with MCAR pattern to avoid bias in the dataset. By applying PROC LOGISTIC, listwise deletion will automatically being performed before the model is build.

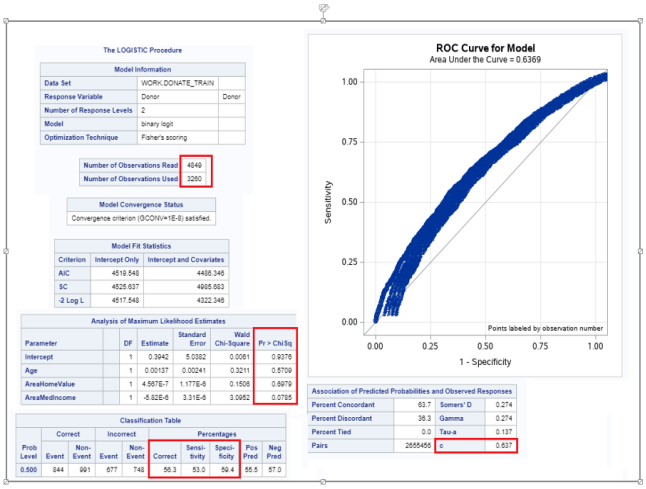


Fig. 1. Logistic Regression Output (Complete case)

From Fig 1., there are only 3,212 observations being used for the logistic regression model after list wise deletion. Overall, the model has 56% of prediction accuracy with 53.1 sensitivity and 58.8 specificity. On the other hand, the model has c value (area under ROC curve) of 0.636.

For the variables AGE AREAHOMEVALUE and AREAMEDINCOME, mean imputation will being use in this section. New dataset being created and named as "Mean_ imputation" to capture the data after the mean imputation.

Mean imputation often reduces the variance of the imputed variables as all the missing value being replaced by the mean. Fig 2. shows the standard deviation of the 3 variables before and after mean imputation.

Fig 2. and 3, shows the result for the logistic regression on the Mean imputation dataset. Overall, the model has 56.2% accuracy with 50.7% sensitivity and 61.4% specificity. Multiple imputation (MI) consists of three main phases: imputation phase, analysis phase and pooling phase. During imputation phase, the missing value are replaced with estimated values and repeated several times depending on the user's configuration. The performance of the imputation will being evaluate using any statistical method of interest. The coefficients obtained from the imputed dataset being utilized for the missing value imputation at pooling phase. Before performing MI, it is important to discover the missing data pattern. By specifying nimpute = 0 for proc mi, the missing data pattern can be obtained. The missing data pattern is non-monotone as shown in Fig 4. Fig 5. show the comparison between the 3 different Multiple Imputation models:

```
344 Title 'Before Mean Imputation';
345 proc means data = work.donate_train;
346   var AGE AREAHOMEVALUE AREAMEDINCOME;
347 run;
348 Title 'After Mean Imputation';
349 proc means data = work.mean_max_imputation;
350   var AGE AREAHOMEVALUE AREAMEDINCOME;
351 run;
```

Before Mean Imputation

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
Age	Age	3699	59.1075968	16.1228985	14.0000000	87.0000000
AreaHomeValue	AreaHomeValue	4821	78372.35	51716.28	7500.00	584600.00
AreaMedIncome	AreaMedIncome	3825	51682.28	17315.98	16564.00	174957.00

After Mean Imputation

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
Age	Age	4849	59.1075968	14.0813816	14.0000000	87.0000000
AreaHomeValue	AreaHomeValue	4849	78372.35	51566.72	7500.00	584600.00
AreaMedIncome	AreaMedIncome	4849	51682.28	15378.88	16564.00	174957.00

Fig. 2. Before and After Mean Imputation

Fig. 3. Logistic Regression Output (Mean Imputation)

```
83 /*Missing Value Pattern*/
84 %let var_miss = Age AREAHOMEVALUE AREAMEDINCOME;
85 proc mi data = work.mode_imputation nimpute=0 seed=1234;
86       FCS;
87     var &var_miss;
88     ods select misspattern;
89 run;
```

The MI Procedure

Missing Data Patterns

Group	Age	AreaHomeValue	AreaMedIncome	Freq	Percent	Group Means		
						Age	AreaHomeValue	AreaMedIncome
1	X	X	X	3260	67.23	59.057975	78997	52025
2	X	X	.	424	8.74	59.601415	82459	.
3	X	.	X	6	0.12	60.666667	.	64176
4	X	.	.	9	0.19	52.777778	.	.
5	.	X	X	559	11.53	.	73399	49547
6	.	X	.	578	11.92	.	76662	.
7	O	O	O	13	0.27	.	.	.

Fig. 4. Missing data pattern

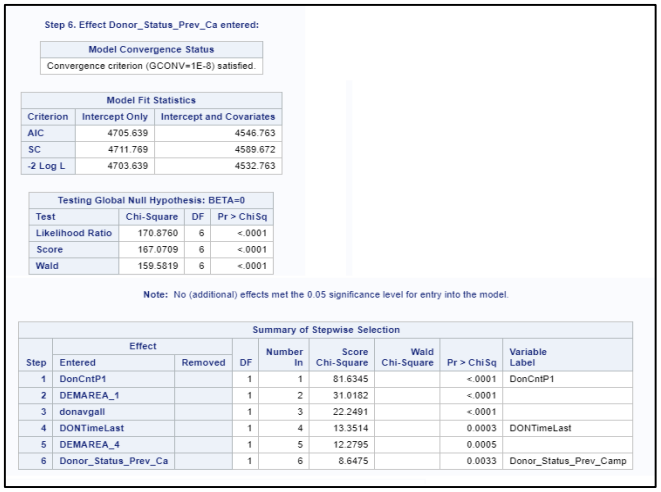
Multiple Imputation									
No of Imputation	5			10			20		
Parameter	Std Error	p-value	FMI	Std Error	p-value	FMI	Std Error	p-value	FMI
intercept	5.032448	0.8948	0.000293	5.032483	0.893	0.00029	5.032291	0.895	0.000223
AGE	0.002191	0.8978	0.208267	0.00214	0.9012	0.159946	0.002107	0.9146	0.133876
AREAHOMEVALUE	0.000000985	0.4041	0.112523	0.000000966	0.3739	0.074075	0.000000971	0.4189	0.083852
AREAMEDINCOME	0.000003111	0.0588	0.26919	0.000003017	0.0389	0.207829	0.000003055	0.0549	0.22463

*FMI: Fraction Missing Information

Fig. 5. Comparison for Multiple Imputation Models

IV. EXPERIMENTATION

After the data pre-processing, the data is being split into training and testing dataset with ratio of 70:30 Training dataset consists of 3,394 (70%) observations whereas testing dataset consists of 1,455 (30%) observations. Fig 6. shows the output for the stepwise logistic regression model. The model stopped at stage 6 as there is no additional effects met the 5% significant level for entry. However, Stepwise Logistic Regression performed better at accuracy rate of 58.5% with only 6 independent variables in the model. The built logistic regression model with 5 independent variables that are significant at critical value of 5%. On the other hand, the model has accuracy of 57.7%, sensitivity of 53.5% and specificity of 61.8% shown in Fig 7.



Step 6. Effect Donor_Status_Prev_Ca entered:

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	4705.639	4546.763
SC	4711.769	4589.672
-2 Log L	4703.639	4532.763

Testing Global Null Hypothesis: BETA=0

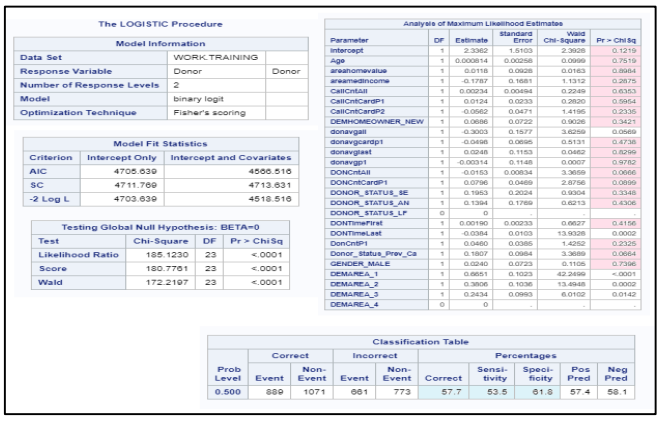
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	170.8760	6	<.0001
Score	167.0709	6	<.0001
Wald	159.5819	6	<.0001

Note: No (additional) effects met the 0.05 significance level for entry into the model.

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	DonCntP1		1	1	81.6345		<.0001	DonCntP1
2	DEMAREA_1		1	2	31.0182		<.0001	
3	donavgall		1	3	22.2491		<.0001	
4	DONTimeLast		1	4	13.3514		0.0003	DONTimeLast
5	DEMAREA_4		1	5	12.2795		0.0005	
6	Donor_Status_Prev_Ca		1	6	8.6475		0.0033	Donor_Status_Prev_Camp

Fig. 6. Output for Stepwise Logistic Regression



The LOGISTIC Procedure

Model Information

Data Set	WORK.TRAINING
Response Variable	Donor
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	4705.639	4566.516
SC	4711.769	4713.631
-2 Log L	4703.639	4518.516

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	185.1230	23	<.0001
Score	180.7761	23	<.0001
Wald	172.2197	23	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.3362	1.5103	2.3928	0.1219
Age	1	0.000814	0.00256	0.0999	0.7519
areahomevalue	1	0.0116	0.0926	0.0163	0.8984
areamedincome	1	-0.1787	0.1681	1.1312	0.2875
CallCntAll	1	0.00234	0.00494	0.2249	0.6353
CallCntCardP1	1	0.0124	0.0233	0.2820	0.5954
CallCntCardP2	1	0.0562	0.0471	1.4195	0.2335
DEMHOMEOWNER_NEW	1	0.0686	0.0722	0.9026	0.3421
donavgall	1	-0.3003	0.1577	3.6259	0.0569
donavgcardp1	1	-0.0498	0.0695	0.5131	0.4738
donavglast	1	0.0248	0.1153	0.0462	0.8299
donavgp1	1	-0.00314	0.1148	0.0007	0.9782
DONCntAll	1	-0.0153	0.00834	3.3659	0.0666
DONCntCardP1	1	0.0796	0.0469	2.8756	0.0899
DONOR_STATUS_SE	1	0.1953	0.2024	0.9304	0.3348
DONOR_STATUS_AN	1	0.1394	0.1769	0.6213	0.4306
DONOR_STATUS_LF	0	0			
DONTimeFirst	1	0.00190	0.00233	0.6627	0.4156
DONTimeLast	1	-0.0384	0.0103	13.9328	0.0002
DonCntP1	1	0.0460	0.0385	1.4252	0.2325
Donor_Status_Prev_Ca	1	0.1807	0.0984	3.3689	0.0664
GENDER_MALE	1	0.0240	0.0723	0.1105	0.7396
DEMAREA_1	1	0.6651	0.1023	42.2499	<.0001
DEMAREA_2	1	0.3806	0.1036	13.4948	0.0002
DEMAREA_3	1	0.2434	0.0993	6.0102	0.0142
DEMAREA_4	0	0			

Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensi-tivity	Speci-ficity	Pos Pred	Neg Pred
0.500	889	1071	661	773	57.7	53.5	61.8	57.4	58.1

Fig. 7. Logistic Model Output

After predicting the results through by the logistic regression on testing dataset, the classification table is being built by comparing the Donor (target variable) with the I_Donor (predicted target variable.

For the testing dataset, the accuracy is slightly lower as compared to the training dataset at 56.4%. This is due to the testing data is data that unseen by the model and it is normal to have testing accuracy which is lower than training accuracy.

Fig 8. shows the overall process of data pre-processing and final model building based on Logistic Regression and Stepwise Logistic Regression. The stepwise logistic regression performed better in terms of accuracy, sensitivity and specificity for both training and testing dataset. Thus, the stepwise logistic regression model is being utilized to predict the DONOR variable in the validation dataset

Data Preprocessing

Missing Value Treatment	Feature Engineering	Accuracy	Sensitivity	Specificity
CC	-	56.3	53	59.4
MI = 5		-	-	-
MI = 10		-	-	-
MI = 20		-	-	-
Single Imputation		56.6	50.9	61.9
Mean Imputation	-	56.2	50.7	61.4
	Categorical Grouping	58.1	52.1	63.7
	Categorical Grouping Multicollinearity Treatment	57.5	51.2	63.5
	Categorical Grouping Multicollinearity Treatment Log Transformation	57.7	51.3	63.7

Final Model Building

Model	Missing Value Treatment	Feature Engineering	Training			Testing		
			Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Logistic Regression	Mean Imputation	Categorical Grouping Multicollinearity Treatment Log Transformation	57.7	53.5	61.8	56.8	53.4	59.9
Stepwise Logistic Regression	Mean Imputation	Categorical Grouping Multicollinearity Treatment Log Transformation	58.5	54.3	62.6	56.9	53.4	60.1

Fig. 8. Summary for Model Building

V. RESULTS

The study focuses on various method of data pre-processing and dimension reduction to improve the logistic regression performance. Initial data exploration is important to understand the nature of each variable in the dataset. For variable AGE, there are few individuals with age below 8 with high values of AreaMedIncome and DONTimeFirst which is considered abnormal. For DONAvgCardP1 (average amount donated with the help of references for last 36 months with the help of references), there are 18.42% missing value. The variable is then being compared to the donation frequency, DONCntCardP1 and concluded that the missing values are all indicating $0 as with donation frequency 0.

The donation dataset with few variables with missing value. Unlike Decision Tree or Neural Network model, Logistic Regression could not take missing value in the regression equation and thus observations with missing value will be removed by default. Thus, several common missing value techniques such as mean imputation, complete case, multiple imputation and single imputation being applied for the missing value treatment. By comparing the standard error and significant of the variable after missing value imputation, mean imputation performs better as compared to other methods in this dataset.

Before building any prediction model, proper feature engineering on the dataset can helps to improve the

performance of the prediction model. In this study, categorical grouping is performed for Donor_status (consists of 6 classes) and DemArea (consists of 54 classes). The grouping for the classes depending on their likelihood in donating. After the grouping, the accuracy of the prediction model increases from 56.2% to 58.1%.

The study further explores the intercorrelation between the independent variables. By stepwise removal, 5 variables that cause the multicollinearity issues being removed. The removal of the variables reduces the model accuracy slightly by 0.6%. The slight decrease in accuracy is acceptable as multicollinearity issue can lead to imprecise estimates of coefficient values for the model. The data pre-processing continued with log transformation to transform skewed data and to reduce the data variability for variable with outliers. The log transformation successfully reduced the skewness and kurtosis for the six variables in the dataset. When compared before and after log transformation, the prediction accuracy slightly increased by 0.2%. This indicates that unlike other multivariate analysis, Logistic Regression is less affected by the normality assumption.

When building prediction model, the dataset being partitioned into training and testing dataset with ratio of 70:30. 70% of the data is being allocated for training as it provides more examples for the algorithms to learn in building the prediction model. The logistic regression model built has accuracy of 57.7% on training dataset and 56.4% on testing dataset. However, the model only consists of 5 variables that are significant at 0.05 critical value in explaining the target variables.

As parsimony is concern, the most representative variables were being chosen in Stepwise Logistic Regression. This is to enable the model to be train fast under acceptable range of model's accuracy. The Stepwise Logistic Regression constructed with 58.5% accuracy on training dataset and 56.9% on testing dataset. In terms of sensitivity and specificity, Stepwise Logistic Regression also outperformed normal Logistic Regression on both training and testing dataset. Under Stepwise Logistic Regression, DonCntP1, DonAvgAll, DONTimeLast, Donor_status_Prev_Camp, DEMAREA_1 and DEMAREA_4 have significant impact on the likelihood of individual in donating. DonAvgAll, DonCntP1 and DemArea_4 have negative relationship on likelihood of individual in donating while DonCntP1, Donor_Status_Prev_Camp and DemArea_1 have positive relationship on the likelihood of individual in donating.

The accuracy for the final model at 58.5% is relatively low as compared to other logistic regression model discussed in Section 2. With all the data pre-processing and stepwise regression, the accuracy of the model merely increased by 2.2% from 56.3% (complete case) to 58.5% (stepwise logistic regression). At this point, other machine learning algorithms such as decision tree, random forest or neural network can be considered as they might provide higher prediction accuracy.

VI. CONCLUSION

In this study, a prediction model to identify possible donor is being built by considering different options for data pre-processing. Prediction model is effective when comes to manpower allocation as the organization able to prioritize possible donor rather than contacting each of them. This also enable the organization to achieve their objectives with minimal time and manpower allocated. Thus, a prediction model with higher accuracy will be preferred for the organization.

Other than model accuracy, sensitivity and specificity rate also another concern for the organization. Test sensitivity measures the ability of the model in correctly identify donors where test specificity measures the ability of the model in correctly identify non-donors. When Sensitivity rate is low, the model tends to have high number of False Negative, which is classifying the donors as non-donors. This will cause the real donor not being prioritize and donation is not obtained from them. On the other hand, low specificity indicates the misclassification of non-donor to donor. This will waste the manpower in contacting them as non-donor are less likely to donate as compared to donor. Both scenarios are equally costly for this case, thus the final prediction model adopted also considering the Specificity and Sensitivity rate.

REFERENCES

[1] Hair Jr, J. F., William, C., Babin, B. J., & Anderson, R. E. (2014). Multivariate Data Analysis Joseph F . Hair Jr. William C. Black Seventh Edition. Pearson Education Limited.

[2] L. Farrokhvar, A. Ansari, B. Kamali, " Predictive models for charitable giving using machine learning techniques," *PLoS ONE*. 13 (10). pp. 1–14, 2018.

[3] N. C. Degasperi and E. W. Mainardes, "What motivates money donation? A study on external motivators," Rev. Adm., vol. 52, no. 4, pp. 363–373, 2017.

[4] R. Snipes and S.Oswald, Charitable giving to not-for-profit organizations: factors affecting donations to non-profit organizations. Innovative Marketing,vol. 6, pp.73-80, 2010.

[5] A. K. Dwivedi, ''Performance evaluation of different machine learning techniques for prediction of heart disease,'' Neural Comput. Appl., vol. 29, no. 10, pp. 685–693, 2018.

[6] J. J. Beunza, E. Puertas, E. García-Ovejero, G. Villalba, E. Condes, G. Koleva, C. Hurtado, and M. F. Landecho, ''Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease),'' J. Biomed. Informat., vol. 97, Sep. 2019

[7] S. Patil, V. Nemade, and P. Soni, "Predictive Modelling for Credit Card Fraud Detection Using Data Analytics," Procedia Computer Science 132, pp. 385-395, 2018

[8] S. V. S. S. Lakshmi and S. D. Kavilla, ''Machine learning for credit card fraud detection system,'' Int. J. Appl. Eng. Res., vol. 13, no. 24, pp. 16819–16824, 2018.

[9] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, ''Type 2 diabetes mellitus prediction model based on data mining,'' Inform. Med. Unlocked, vol. 10, pp. 100–107, Aug. 2018.

[10] C. Zhu, C. U. Idemudia, and W. Feng, ''Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques,'' Informat. Med. Unlocked, vol. 17, 2019, Art. no. 100179.

[11] V.Taslimitehrani, G. Dong, N. L.Pereira, M. Panahiazar, J. Pathak ., " Developing EHR-driven heart failure risk prediction models using CPXR(Log) with the probabilistic loss function," *Journal of Biomedical Informatics*. pp.260–269, 2016.

Available from: http://dx.doi.org/10.1016/j.jbi.2016.01.009