**Prediction of Early Readmission of Inpatients: Analysis of 69, 970 Clinical Patient Records**

Abimbola Ogungbire – University of North Carolina at Greensboro

**Introduction**

Having handled the 101,766 records of data collected from Health Facts database, we are left with 69,970 records of data to build a classification model for early readmission of patients with diabetes. In previous assignments, we have made effort to reiterate the model in [1]. This report will attempt to evaluate the model in [1], build a logistic regression using a different approach from the one used in [1] and also consider a different algorithm. In this case, I will employ the use of random forest and compare these two algorithms.

**Methodology**

VALIDATION OF LOGISTIC MODEL IN [1]

In order to ascertain how well the logistic regression of the author predicts, our data of 69,970 observations have been split into 70% train data set and 30% test data set. Having built the model, one question still remain unanswered – is the model good enough?

The final model as in the paper was trained using the training data set. It was used to predict the test data set and the model accuracy was assessed. The ROC curve and AUC value of the both models with and without interaction was checked. The result of the model validation is as presented in the result and discussion section below.

A confusion matrix containing the classification error rate of both classes was also checked. The model accuracy was compared with the out of bag error rate and the discrepancy was noted.

LOGISTIC REGRESSION MODEL USING BEST GLM SELECTION TECHNIQUE

Having done the preliminary analysis of the data as recommended by [1] in previous assignments, I have decided to build a logistic regression using best glm selection approach.

In order to address the primary question which involves the association between A1C result and early readmission of patients, I have built a selected a model while controlling for HbA1c.

The model was developed in stages and each stage was succeeded by an analysis of deviance test. The final model was finally compared with the initial model using an analysis of variance test to check if the two models differ from one another.

First, we employed the use of best glm selection technique to select an optimum model using the Akaike Information Criterion (AIC). Second, we added HbA1c to the selected model. Third, we added possible pairwise interactions to the model in the second step one after the other to take note of only the significant ones. Finally, we added the pairwise interactions of HbA1c and each covariates to the model in the third stage and selected only the significant HbA1c interaction(s).

The above stages of model selection were carried out on the train data. The validation of the model was done using the test data set. The model accuracy, sensitivity, specificity and auc were used to assess how well the model predicts.

RANDOM FOREST

A random forest model unlike the logistic model is more robust and might do a better job in classifying the target variable. However, measures have been put in place to make sure the model does not over fit and does not give misleading result.

The random forest was built in three stages. The first model was built to include all variables except HbA1c. Second, I checked the variable importance plot to take out few variables that are less important in our model. Finally, I added HbA1c to the model obtained in the second stage.

Similar to the logistic regression, the above steps were carried out using the train data set. The test data set was used in model evaluation and similar criteria used in the previous algorithm.

**Result and Discussion**

VALIDATION OF LOGISTIC MODEL IN [1]

A summary of the model fit on the training data is as shown below:

The resulting model was used for prediction on the test data set and the confusion matrix below was generated using a cut off value equivalent to the mean of probabilities of prediction.

From the confusion matrix below, the model is doing a good job in classifying late readmission with a low class error rate of 6.49% on the other hand it is misclassifying an ample amount of early readmitted patients with a very high error rate of 88.1%.

Table 1: Confusion Matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Referenced | | |
| Late Readmission | Early Readmission | Class Error |
| Predicted | Late Readmission | 11645 | 809 | 0.0649590 |
| Early Readmission | 8033 | 1085 | 0.8810046 |

The ROC curve which shows the tradeoff between correctly predicted labels and that of the misclassified labels. We are mostly concerned about the area under the ROC curve. Below is the ROC of the model after prediction. The resulting area under the beneath ROC curve is given as 0.6129. However, it should be noted that the metrics used in quantifying AUC is between the ranges of 0.5 to 1.0. Therefore, it is not enough to say that our AUC of 0.6129 is a good value.

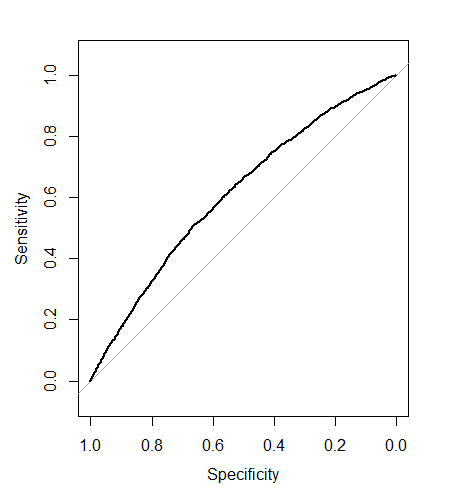


Fig. 1: ROC Curve

LOGISTIC REGRESSION MODEL

The core model fit from the best glm selection method is as presented in table 2 below. Since the optimum model having 7 parameters contain 6 variables, we dropped out race and gender from the core model i.e. the model without the A1C result. We carried out an analysis of deviance test and all 6 variables seem to be relevant to the model.

Table 2: Result from best glm selection method

|  |  |  |  |
| --- | --- | --- | --- |
| Model Size | AIC | Number or parameters | Variables in Model |
| 0 | 29159.82 | 1 | - |
| 1 | 28610.59 | 2 | Pri-diag |
| 2 | 28226.72 | 3 | Pri-diag, discharge id |
| 3 | 28189.61 | 4 | Pri-diag, discharge id, time in hospital |
| 4 | 28176.53 | 5 | Pri-diag, discharge id, time in hospital, age |
| 5 | 28168.24 | 6 | Pri-diag, discharge id, time in hospital, age, medical specialty |
| 6\* | 28163.90 | 7 | Pri-diag, discharge id, time in hospital, age, medical specialty, admission source id |
| 7 | 28164.14 | 8 | Pri-diag, discharge id, time in hospital, age, medical specialty, admission source id, race |
| 8 | 28167.64 | 9 | Pri-diag, discharge id, time in hospital, age, medical specialty, admission source id, race, gender |

The resulting model from best glm selection technique was made to include HbA1c. Analysis of deviance test shows that HbA1c has a slightly large p-value of 0.074. The p-value of HbA1c is relatively large compared with other variables in the model with a small deviance value. This suggest that HbA1c, if removed from the model will not make much impact. However, the data set should tend to shrink p-value but a large p-value of HbA1c indicate that it is significant enough.

Table 3: Test of Deviance Table

|  |  |  |
| --- | --- | --- |
| Attribute Name | Deviance | p-value |
| Discharge Disposition | 224.903 | < 2.2e-16 |
| Admission Source | 9.237 | 0.0098658 |
| Medical Specialty | 18.064 | 0.0028667 |
| Time in Hospital | 36.717 | 1.366e-09 |
| Age | 11.584 | 0.0006651 |
| Diagnosis | 48.435 | 2.118e-07 |
| HbA1c | 6.946 | 0.0736469 |

The covariates in the analysis were interacted in pairs to obtain significant interactions that will otherwise be useful in our analysis. Individual pairwise interactions was done for all possible interactions. The significant interactions were kept in the model and all significant interactions were fitted to check the deviations. HbA1c was also interacted with the covariate and the resulting final model is as presented in table 4 and 5 below.

Table 4: Coefficients of non-interaction terms estimated from the final logistic model

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Estimate | P-value |
|  | Intercept | -2.931e+00 | < 2e-16 |
| Discharge Disposition | Home | Reference |  |
| Others\* | 5.433e-01 | < 2e-16 |
| Age\* | | 3.854e-03 | 0.000695 |
| Medical Specialty | Cardiology | Reference |  |
| Family/General Practice\* | 0.266695 | 0.004507 |
| Internal Medicine\* | 0.198496 | 0.019670 |
| Missing | 0.101087 | 0.197194 |
| Others | 0.061980 | 0.469044 |
| Surgery | 0.003482 | 0.973559 |
| Time in Hospital\* | | 0.033216 | 1.67e-09 |
| Admission Source | Emergency | Reference |  |
| Otherwise\* | -0.131788 | 0.007633 |
| Physician/Clinical Referral | -0.068298 | 0.080021 |
| Primary Diagnosis | Circulatory | Reference |  |
| Diabetes | -1.019e-03 | 0.987435 |
| Digestive\* | -1.302e-01 | 0.037663 |
| Genitourinary | -7.956e-02 | 0.307308 |
| Injury | -2.720e-03 | 0.966483 |
| Musculoskeletal\* | -2.044e-01 | 0.009175 |
| Neoplasm | -1.203e-02 | 0.892112 |
| Other\* | -1.192e-01 | 0.018381 |
| Respiratory\* | -3.293e-01 | 5.06e-09 |
| Unknown | -1.057e+01 | 0.905369 |
| HbA1c | High & Change in Med | Reference |  |
| High & No change in Med | -1.645e-01 | 0.173475 |
| None | 1.029e-02 | 0.884525 |
| Normal | -1.077e-01 | 0.214111 |

Interactions in the final model is as presented in table 5 below.

Table 5: Coefficients of Interaction terms from the final logistic regression model.

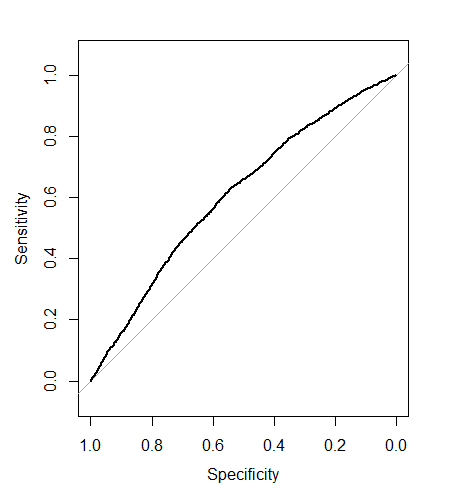
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Feature Name | Value | Feature Name | Value | Estimate | P-value |
| Discharge Disposition | Others | Medical Specialty | Family/General Practice | 2.953e-01 | 0.146237 |
| Internal Medicine | 8.845e-02 | 0.635588 |
| Missing | 1.047e-01 | 0.546350 |
| Others | 2.792e-01 | 0.136597 |
| Surgery\* | 5.317e-01 | 0.017942 |
|  |  |  |
| Time in hospital | | Primary Diagnosis | Diabetes\* | 4.161e-02 | 0.040820 |
| Digestive | 8.715e-03 | 0.683680 |
| Genitourinary\* | 5.684e-02 | 0.031108 |
| Injury | -1.875e-03 | 0.933098 |
| Musculoskeletal | 5.090e-02 | 0.090546 |
| Neoplasm | -3.175e-02 | 0.262238 |
| Other | -1.722e-02 | 0.304166 |
| Respiratory | 2.437e-02 | 0.192961 |
| Unknown | -1.200e-02 | 0.999760 |
| Discharge Disposition | Others | age | | -8.174e-03 | 0.000603 |
| Age | | Primary Diagnosis | Diabetes | -1.620e-03 | 0.664826 |
| Digestive | -8.603e-04 | 0.844249 |
| Genitourinary | 4.368e-03 | 0.431024 |
| Injury\* | 9.286e-03 | 0.047982 |
| Musculoskeletal\* | 1.949e-02 | 0.002237 |
| Neoplasm | -1.682e-03 | 0.814542 |
| Other\* | 9.846e-03 | 0.003788 |
| Respiratory\* | 1.046e-02 | 0.012678 |
| Unknown | 4.547e-03 | 0.999603 |
| Discharge Disposition | Others | Primary Diagnosis | Diabetes | 1.583e-02 | 0.909387 |
| Digestive | 1.580e-01 | 0.248656 |
| Genitourinary | -9.114e-02 | 0.592874 |
| Injury\* | 3.172e-01 | 0.036596 |
| Musculoskeletal | 2.096e-01 | 0.242153 |
| Neoplasm | -3.501e-02 | 0.851323 |
| Other | 1.689e-01 | 0.115789 |
| Respiratory | 1.324e-01 | 0.276132 |
| Unknown | -5.912e-01 | 0.998355 |
| Time in hospital | | Age | | -4.848e-04 | 0.202953 |
| Discharge Disposition | Others | Time in hospital\* | | -2.950e-02 | 0.009983 |
| HbA1c | High & No change in Med | Primary Diagnosis | Diabetes | -4.153e-03 | 0.990286 |
| Digestive | -3.330e-01 | 0.565563 |
| Genitourinary | -6.884e-01 | 0.399326 |
| Injury | 6.173e-01 | 0.371650 |
| Musculoskeletal | 1.061e-01 | 0.877893 |
| Neoplasm | 1.016e+00 | 0.251321 |
| Other | 5.199e-01 | 0.138109 |
| Respiratory | -3.959e-01 | 0.413689 |
| None | Diabetes\* | 5.049e-01 | 0.011016 |
| Digestive | 1.319e-01 | 0.682444 |
| Genitourinary | 5.860e-02 | 0.867404 |
| Injury | 6.839e-01 | 0.100553 |
| Musculoskeletal | -2.850e-01 | 0.488190 |
| Neoplasm | 5.482e-01 | 0.378679 |
| Other | 2.463e-01 | 0.255676 |
| Respiratory | 1.853e-01 | 0.463061 |
| Unknown | -1.214e-01 | 0.999729 |
| Normal | Diabetes\* | 6.381e-01 | 0.016194 |
| Digestive | 4.129e-02 | 0.916192 |
| Genitourinary | 2.726e-01 | 0.514430 |
| Injury | 3.040e-02 | 0.951354 |
| Musculoskeletal | -3.837e-01 | 0.451442 |
| Neoplasm | 1.214e+00 | 0.074763 |
| Other | 3.661e-01 | 0.154599 |
| Respiratory | -1.383e-01 | 0.654783 |

I have conducted a test of deviance to compare the core model and the final model. This is to check whether the interaction added is statistically significant in the model. The deviance test shows a very tiny p-value (4.426e-08) which suggest that we should include the interaction in our model.

Table 6: Test of Deviance

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Residual DF | Deviance DF | DF | Deviance | p-value |
| 47446 | 28119 |  |  |  |
| 47386 | 27981 | 60 | 137.98 | 4.426e-08 |

The area under the ROC curve of the final model was estimated to be 0.6076. Comparing this result with the final model in [1]. These two models are not too different from one another.



The confusion matrix is also shown to check the classification error rate of each classes.

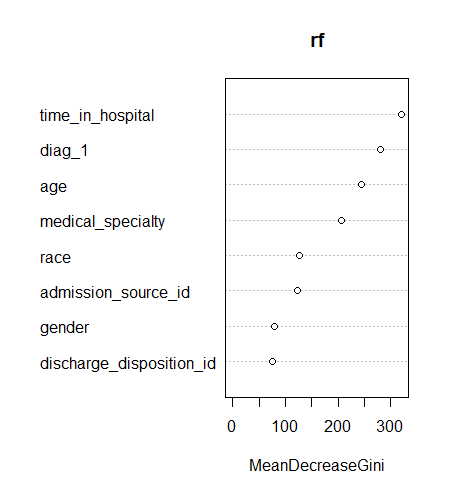
Table 7: Confusion Matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Referenced | | |
| Late Readmission | Early Readmission | Class Error |
| Predicted | Late Readmission | 11618 | 804 | 0.0647239 |
| Early Readmission | 8060 | 1090 | 0.8808743 |

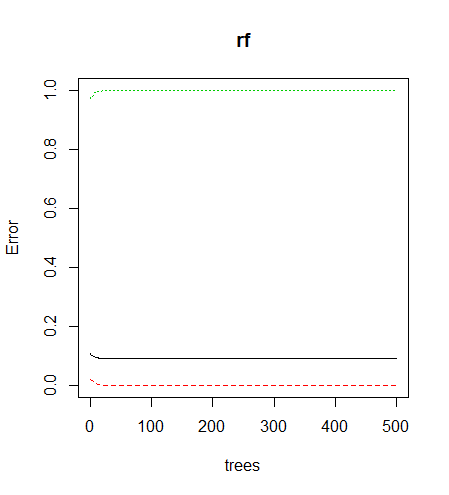
The result in the above table is very similar to the matrix of the final model generated in [1]. Classifying early readmission seem to come with a lot of prediction error.

RANDOM FOREST MODEL

Having fit a random forest model of readmitted with covariates (without HbA1c). The variable importance plot in suggest that gender and discharge disposition id are less important in our model. But the question remain – Is the gini index high enough for us to keep them in the model? I think the gini index is close to 100 and if you would ask me, I would say we keep the two variables in the model.

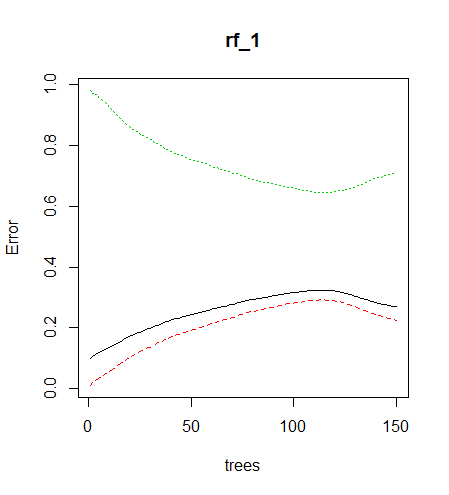


The error rate of the final model is as shown below to include the number of trees



Looking at the trees plot above, the green curve indicates the error rate for early readmission and the red curve represent the error rate for late readmission. While the black curve repres-ent the out of bag error rate.

I saw the need to change the cut off (I made use of the mean probability of both prediction classes). Looking at the curve above, the error rate remains constant while the number of trees increases. We have reduced the number of trees to 200 in the next step and the resulting model is as shown below.



The error rate in the early readmission class has significantly dropped while there is an increase in the error rate of both late readmission and out of bad error rate.

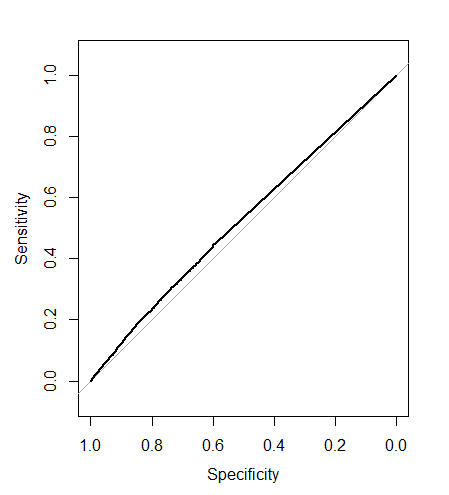
In validating the above model, I have shown the confusion matrix and ROC curve below.

Table 8: Confusion Matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Referenced | | |
| Late Readmission | Early Readmission | Class Error |
| Predicted | Late Readmission | 14784 | 1334 | 0.0827646 |
| Early Readmission | 4894 | 560 | 0.8973231 |

The have high sensitivity of 0.7513 and a low specificity of 0.2957. The accuracy of the model which shows how much right prediction is made in our model is 0.7113. This is quite reasonable because late readmission which happens to have a higher number of votes is well predicted.

The AUC of the ROC curve is given by 0.5238. This model when compared to the model without HbA1c is not too different.



**Conclusion**

From the two model built above, the Logistic regression model does a better job than the random forest model. Also, the significance of HbA1c from both model is not quite pronounced. Comparing the AUC values in the random forest model, the models look alike in al respect.

Although the logistic regression suggest a slightly large p-value from the deviance test which suggest that HbA1c might not have a significant impact in the study of early readmission patients in the hospitals.