Prediction on Readmission to Hospital using a Popular Diabetes Dataset

Yijia Liu

December 22, 2020

Abstract

It is important to know if a diabetes patient will be readmitted to some hospital. The reason is that you can change the treatment, in order to avoid a readmission. We will apply Generalized linear model(GLM) (Nelder and Wedderburn 1972), Generalized Linear Mixed Models (GLMMs) (McCulloch and Neuhaus 2014) and Backward Selection (Hocking 1976) to select variables and decide the finalized model. Finally, 10 variables are found to help predict readmission rate and the prediction accuracy is around 0.88, which is an accepted result.

Keywords: Diabetes, Readmission, GLM, AIC, BIC

Data: https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

Github: https://github.com/LesleyyijiaLiu/Prediction-on-Readmission-to-Hospital-using-a-Popular-Diabetes-Dataset.git

1. Introduction

Diabetes is a lifelong condition where your body does not produce enough insulin or your body cannot use the insulin it has effectively. It can increase the risk of high blood pressure, narrowing of the arteries (atherosclerosis), coronary artery disease and stroke (Government 2019). Thus, it is important to identify patients with worse outcomes.

In general, it is hard to measure directly. However, readmission can provide a low cost, stable estimate. If the readmission is none, which means the treatment is great; if the readmission is less than 30 days, which means the treatment may not be appropriate; if the readmission is more than 30 days, which means the treatment is not very good, but the reason could be the patients.

In this project, we would like to predict 'readmission' using the variables from the dataset (University 2012). Firstly, we recombine 'readmission' to a binary variable and apply the 'GLM' model to find a linear relationship. Secondly, variable selection is applied to find more efficient variables. Thirdly, we construct a better model with conditioning on 'patient_nbr'. Finally, model validation and prediction testing are made to make the project more convincing.

The structure of this project is: we simply describe the data and explain the variables firstly section, and then give an introduction to our model. After that, the procedures for getting our final model will be provided. In the end, we will talk about the prediction results and discussion parts.

Our report is built using `R` (R Core Team 2020), with packages `MASS` (Venables and Ripley 2002), `car` (Fox and Weisberg 2019), `lmtest` (Zeileis and Hothorn 2002), `lme4` (Bates et al. 2015), `tidyverse` (Wickham et al. 2019), `magrittr` (Bache and Wickham 2020), `qqtest` (Oldford 2020), `dplyr` (Wickham et al. 2020), `GGally` (Schloerke et al. 2020), and `pROC` (Robin et al. 2011).

2. Data

2.1. Source of the Data

The data set represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. We find this data set from Kaggle but the raw data are submitted on behalf of the Center for Clinical and Translational Research, Virginia Commonwealth University, a recipient of NIH CTSA grant UL1 TR00058 and a recipient of the CERNER data. John Clore, Krzysztof J. Cios, Jon DeShazo, and Beata Strack. This data is a de-identified abstract of the Health Facts database (University 2012).

It includes over 50 features representing patient and hospital outcomes. Information was extracted from the database for encounters that satisfied the following criteria.

- It is an inpatient encounter (a hospital admission).
- It is a diabetic encounter, that is, one during which any kind of diabetes was entered into the system as a diagnosis.
- The length of stay was at least 1 day and at most 14 days.
- Laboratory tests were performed during the encounter.
- Medications were administered during the encounter.

The data contains such attributes as the patient number, race, gender, age, admission type, time in the hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, number of outpatients, inpatient, and emergency visits in the year before the hospitalization, etc.

2.2. Description of the Data

The data contains 101766 observations from 71518 patients. Some of the 50 features representing patient and hospital outcomes are numerical and some are categorical.

- Encounter ID: Unique identifier of an encounter
- Patient number: Unique identifier of a patient
- Race Values: Caucasian, Asian, African American, Hispanic, and other
- Gender Values: male, female, and unknown/invalid
- Age Grouped in 10 -year intervals: $0,10), 10, 20), \ldots, 90, 100)$
- Weight: Weight in pounds
- Admission type: Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available
- Discharge disposition: Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available
- Admission source: Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital
- Time in hospital: Integer number of days between admission and discharge
- Payer code: Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay Medical
- Medical specialty: Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon
- Number of medications: Number of distinct generic names administered during the encounter
- Number of outpatient visits: Number of outpatient visits of the patient in the year preceding the encounter
- Number of emergency visits: Number of emergency visits of the patient in the year preceding the encounter
- Number of inpatient visits: Number of inpatient visits of the patient in the year preceding the encounter
- Diagnosis 1: The primary diagnosis (coded as first three digits of ICD9); 848 distinct values
- Diagnosis 2: Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values
- Diagnosis 3: Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values
- Number of diagnoses: Number of diagnoses entered to the system 0%
- Glucose serum test result: Indicates the range of the result or if the test was not taken. Values: $''>200,$"$''>300,$ " "normal," and "none" if not measured
- A1c test result: Indicates the range of the result or if the test was not taken. Values: $">8$" if the result was greater than 8%, $''>7''$ if the result was greater than 7% but less than 8%, "normal" if the result was less than 7%, and "none" if not measured.
- Change of medications: Indicates if there was a change in diabetic medications (either dosage or generic name). Values: "change" and "no change"

- Diabetes medications: Indicates if there was any diabetic medication prescribed. Values: "yes" and "no"
- 24 features for medications For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride- pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: "up" if the dosage was increased during the encounter, "down" if the dosage was decreased, "steady" if the dosage did not change, and "no" if the drug was not prescribed
- Readmitted: Days to inpatient readmission. Values: $''<30''$ if the patient was readmitted in less than 30 days, $''>30''$ if the patient was readmitted in more than 30 days, and "No" for no record of readmission

2.3. Manipulation of the Data

To clean the data and keep the most important variables, we will do the following steps:

(1) For the response variable (readmission), since the distribution is imbalanced (11357 less than the 30s, 35545 over 30s and 54864 no), additionally, the distinction of the less than 30 readmission and over 30 readmission is not great, we decide to use two levels that are 'no readmission' and 'readmission'. Thus, we transform the response variable 'readmission' to be binary: combine 'less than 30 (total 11357)' and 'over 30 (total 35545)' to '1', which means the patient does have readmission; change 'no' to '0 (54864)', which means there is no readmission. This transformation also helps to keep the data balanced.

(2) Remove some seriously imbalanced or NA variables: Since 'weights', 'payer_code' and 'medical_specialty' contain almost all NAs, we will remove them. Also, 'examide', 'metformin_rosiglitazone' and 'citoglipton' are removed due to the seriously imbalanced.

(3) Remove other rows that contain NAs.

(4) Correlation Matrix: For all rest numerical variables, we can construct the correlation matrix as following (Figure 1):

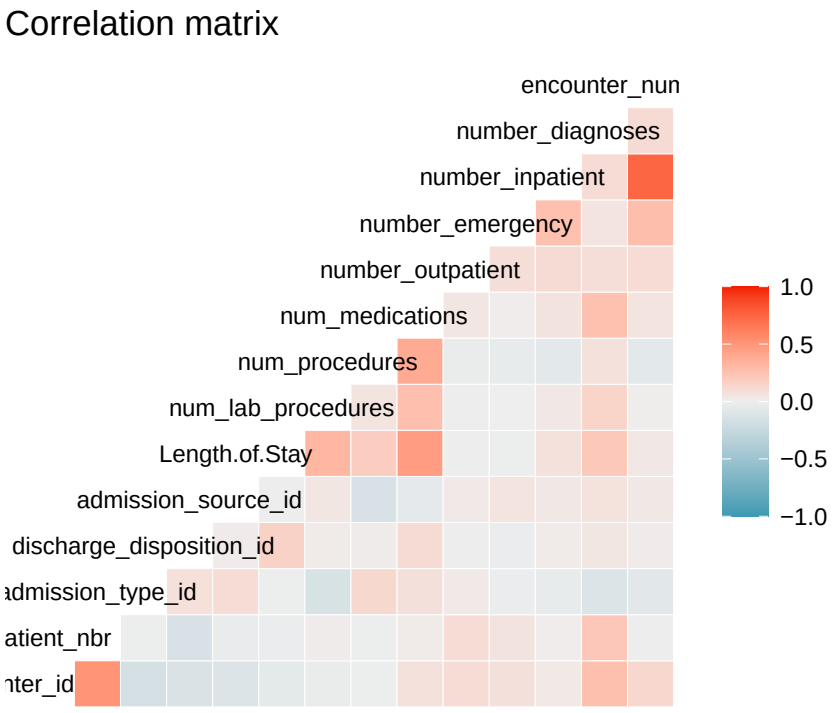

Figure 1: Correlation Matrix

Figure 1: Regession and Ratio Estimation Comparison

From the figure above, we can find 'num_medications' has a relatively strong relationship with several variables and 'number_inpatient' has a strong relationship with 'encounter_num'.

3. Models

3.1. Introduction to Models

Model1: GLM with Binomial Response

When the response variable has only two outcomes, which follows a binomial distribution $\text{Bin}(m_i, \pi_i)$, we can use

$$\text{P}(Y_i = y_i) = \begin{pmatrix} m_i \\ y_i \end{pmatrix} \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i}$$

to express. We further assume that the Y_i are independent. The individual trials that compose Y_i are subject to the same q predictors $(x_{i1}, \ldots, x_{iq})$

As in the binary case, we construct a linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq}$$

We can use a logistic link function $\eta_i = \log\left(\pi_i / \left(1 - \pi_i\right)\right)$ and the log-likelihood is given by

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \eta_i - m_i \log\left(1 + e^{\eta_i}\right) + \log \begin{pmatrix} m_i \\ y_i \end{pmatrix} \right]$$

The deviance is given by:

$$D = 2 \sum_{i=1}^{n} \left[\frac{y_i \log\left(y_i\right)}{\hat{y}_i} + \frac{\left(m_i - y_i\right) \log\left(m_i - y_i\right)}{m_i - \hat{y}_i} \right]$$

Model2: Generalized linear Mixed Models (GLMMs)

After considering random effects, we can use generalized linear mixed models. The interpretation of GLMMs is similar to GLMs; however, there is an added complexity because of the random effects. On the linear metric (after taking the link function), interpretation continues as usual. However, it is often easier to back transform the results to the original metric. For example, in a random-effects logistic model, one might want to talk about the probability of an event given some specific values of the predictors. Likewise in a Poisson (count) model, one might want to talk about the expected count rather than the expected log count. These transformations complicate matters because they are nonlinear and so even random intercepts no longer play a strictly additive role and instead can have a multiplicative effect. When the response follows an exponential family distribution,

$$f\left(y_i \mid \theta_i, \phi\right) = \exp\left\{ \frac{y_i \theta_i - b\left(\theta_i\right)}{a(\phi)} + c(y, \phi) \right\}$$

Let $\mathbb{E}\left(Y_i\right) = \mu_i$ and we connect it to the linear predictor η_i using the link function g by $\eta_i = g\left(\mu_i\right)$. If the random effect γ has distribution $h(\gamma \mid V)$ for parameters V. The fixed effects are β. Condition on the random effects γ, we have $\theta_i = x_i^\top + z_i^\top \gamma$. Then the likelihood function is

$$L(\beta, \phi, V \mid y) = \prod_{i=1}^{n} \int f\left(y_i \mid \beta, \phi, \gamma\right) h(\gamma \mid V) d\gamma$$

Where $\gamma \sim N(0, D)$

3.2. Variable Selection

Some variables in the original model are redundant or there exists Multicollinearity. So we need to develop some variable selection method:

(a) AIC and BIC Criterion:

AIC (Sakamoto, Ishiguro, and Kitagawa 1986) and BIC (Schwarz and others 1978) are Information criteria methods used to assess model fit while penalizing the number of estimated parameters. Let k be the number of estimated parameters in the model. Let L be the maximum value of the likelihood function for the model. Then the AIC value of the model is the following.

$$\text{AIC} = 2k - 2\ln(L)$$

The formula for BIC is similar to the formula for AIC, but with a different penalty for the number of parameters.

$$\text{BIC} = \ln(n)k - 2\ln(\text{L})$$

(b) Stepwise selection

Stepwise regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure.

- Forward selection: Forward selection is a type of stepwise regression that begins with an empty model and adds in variables one by one.
- Backward selection: The backward selection model starts with all candidate variables in the model. At each step, the variable that is the least significant is removed. We applied the backward selection model in this project.

3.3. Process of Obtaining Final Model

(1) Split the data to train set and test set: I create a test dataset that contains a random selection of 20000 patients by using the 'patient_nbr' variable. The rest are set as a train set.

(2) Remove other variables that cannot be used as covariates: I continue to remove "encounter_id", "patient_nbr", "admission_source_ _id","encounter_num" before fitting models.

(3) Fitting GLM with Binomial Response: since the response variable "readmitted" is a binary variable, we can let $Y_i \sim \text{Bernoulli}(\pi_i)$, where π_i is the probability of readmission. And then we fit a GLM as $\text{logit}(\pi_i) = X_i\beta + e_i$, where $e_i \sim N(0, \sigma^2)$

- The fitted formula for model1 is:

formula = readmitted ~ race + gender + age + num_lab_procedures + num_procedures + num_medications + number_outpatient + number_emergency + number_inpatient + number_diagnoses + max_glu_serum + A1Cresult + metformin + glipizide + pioglitazone + rosiglitazone + acarbose + insulin + change + diabetesMed + Length.of.Stay (totally 21 variables)

(4) Backward Stepwise Selection: From the regression results above, I find half of the covariates are not significant. So I decide to use the backward stepwise selection method to find a more proper and simpler model but with a reasonable explanation.

- The final result is:

readmitted ~ race + gender + age + num_lab_procedures + num_procedures + number_outpatient + number_emergency + number_inpatient + number_diagnoses + max_glu_serum + A1Cresult + metformin + glipizide + rosiglitazone + acarbose + insulin + diabetesMed + Length.of.Stay (totally 18 variables)

(5) Generalized Linear Mixed Model: After selecting variables using Stepwise, we still find some of the variables that are not significant. Since one patient may have more than one encounter, the generalized linear mixed model condition on "patient_nar" maybe a better choice. We can let $Y_{ij} \mid U_i \sim \text{Bernoulli}(\pi_{it})$ where π_{it} is the probability of readmission. And then we fit a GLMM as $\text{logit}(\pi_{it}) = X_{it}\beta + U_i$

- The fitted formula for model 1 is:

Formula: readmitted ~ (1 | patient nbr)+ race + gender + age + num_lab_procedures + num_procedures + number_outpatient + number_emergency + number_inpatient + number_diagnoses + acarbose + insulin + diabetesMed + Length.of.Stay

- From the regression results, we continue to remove the variables that are not significant, the final model is:

final formula = readmitted ~ (1 | patient nbr)+ race + gender + age + num_procedures + number_outpatient + number_emergency + number_diagnoses + insulin + diabetesMed + Length.of.Stay

3.4. Model Validation/ Dignostics

After selecting the variables, we need to check the following assumptions:

(1) Normality of residuals (Figure 2); (QQ-plot)

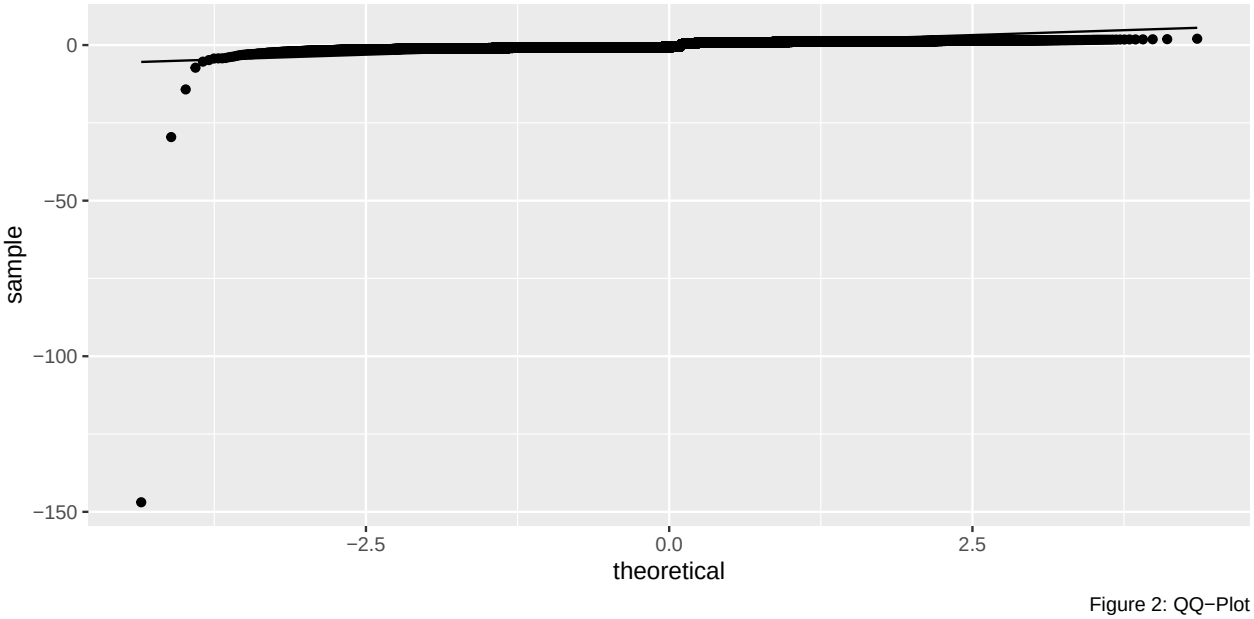

Figure 2: Regession and Ratio Estimation Comparison

From the QQ plot above, we can find most of the points are on the straight line, which means the residuals follow normal distribution approximately.

(2) Homoscedasticity (Figure 3); (Fitted Values v.s. Residuals)

From the figure below, we can find the points have some trends even if they are on the two sides around zero. Thus, it is not very confident to state the variance is constant, which also means there may not exist homoscedasticity.

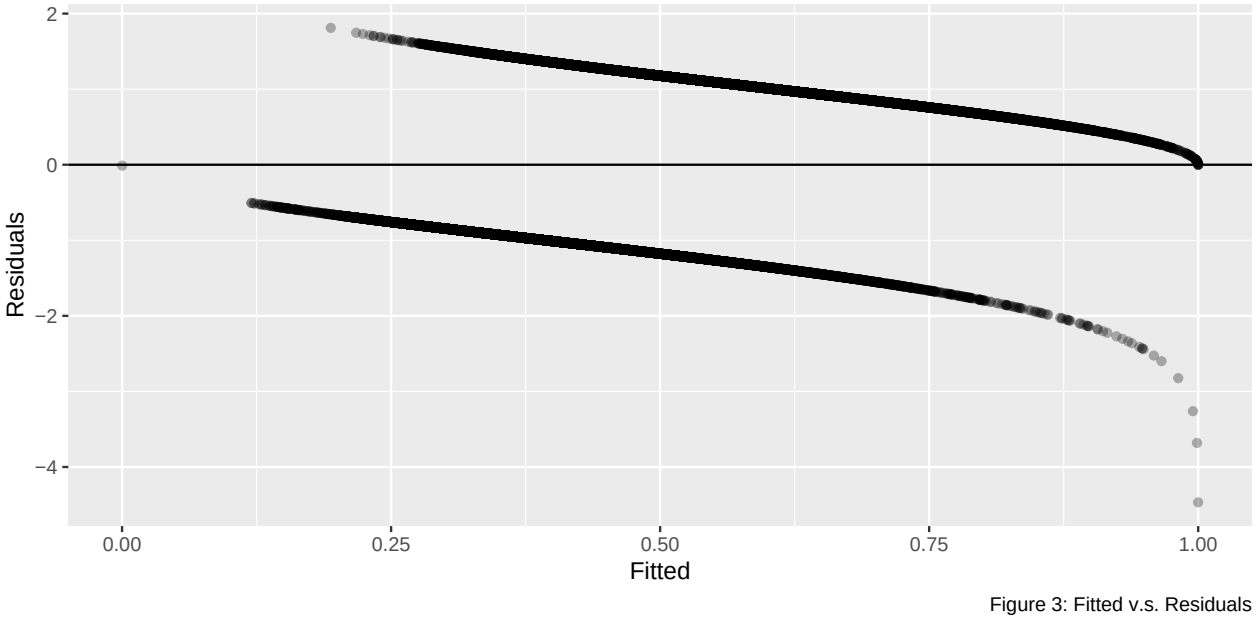

Figure 3: Regession and Ratio Estimation Comparison

4. Results

- Finally, we build the GLMM model condition on "patient_nbr", totally of 10 covariates, which is:

Final Formula = readmitted $\sim$ (1 | patient nbr)+ race + gender + age + num_procedures + number_outpatient + number_emergency + number_diagnoses + insulin + diabetesMed + Length.of.Stay

- After fitting this model on the train set, we use it on the test set. The AUC is 0.8854 and the ROC curve is as below (Figure 4).

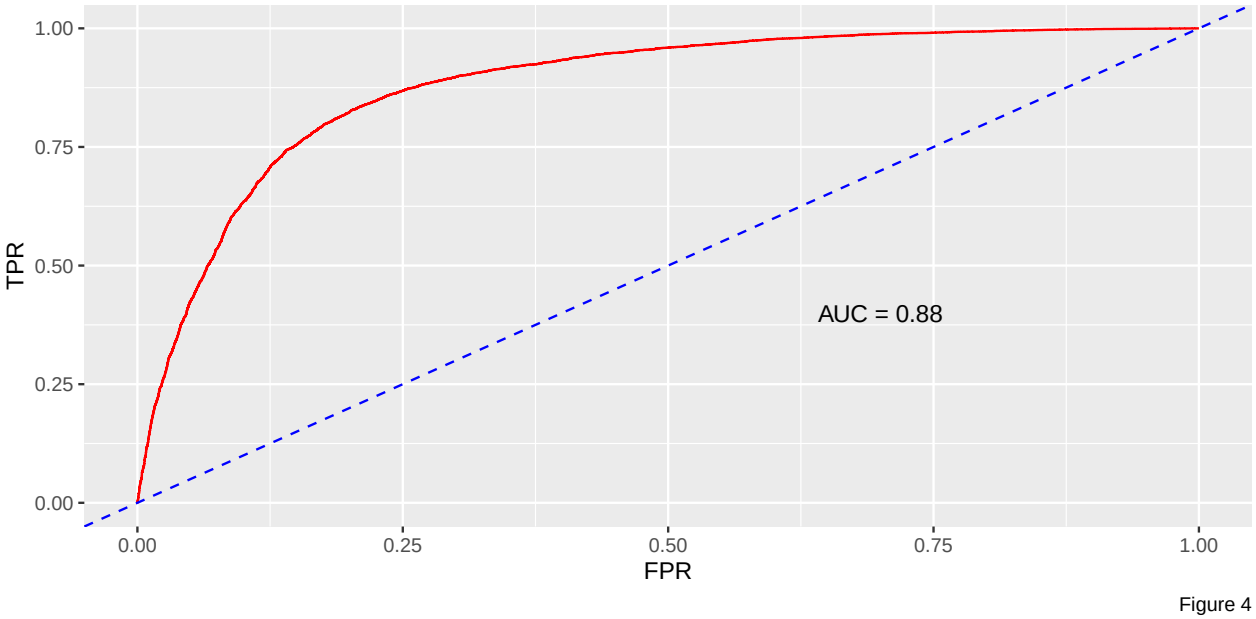

Figure 4: Regession and Ratio Estimation Comparison

From the accuracy, we can find the predicted results are acceptable and even good. Thus, the model may be a suitable model for the data.

5. Discussion

5.1. Results Interpretation

The result shows that there are many causes for readmission. The first one is the race. If a person is an Asian, he or she is more likely readmitted. The possibility of man a readmitted is higher than a woman. Patients who are more than 60 years old but less than 90 years old can be easily readmitted. The higher number of procedures, outpatient, emergency and diagnoses mean the higher possibility to get readmitted. The use of insulin is also a crucial reason for readmitted. Moreover, diabetes has a positive influence on readmission. Besides, if a patient has already stayed a long time in the hospital, he or she has a greater probability of readmission.

5.2. Limitations

Our model gives some guidance to predict the readmission. However, we have several limitations here in our model.

- BIC is a good criterion to select variables. However, it may give too much penalty that leads to an under-fit issue.
- No interaction terms are considered in our model (due to the large data set and computing time). We are not sure how this could affect our model. But we believe all these variables cannot be independent of each other. There should be some correlations underneath.
- The model is under the linear assumption. We may consider involving the smooth pattern (model GAM). However, it may make the model too complex.
- There are some problems with the stepwise selection method, so it is a meaningful future work on more efficient variable selection methods.

(a) It yields R-squared values that are badly biased to be high.

(b) It yields p-values that do not have the proper meaning, and the proper correction for them is a difficult problem.

(c) It yields p-values that do not have the proper meaning, and the proper correction for them is a difficult problem.

(d) The stepwise selection allows us to think too much about the statistical model but not the original problem.

Reference

Bache, Stefan Milton, and Hadley Wickham. 2020. *Magrittr: A Forward-Pipe Operator for R*.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. https://doi.org/10.18637/jss.v067.i01.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Government, Ontario. 2019. "Heart and Strok." https://www.heartandstroke.ca/heart-disease/risk-and-prevention/condition-risk-factors/diabetes.

Hocking, Ronald R. 1976. "A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression." *Biometrics* 32 (1): 1–49.

McCulloch, Charles E, and John M Neuhaus. 2014. "Generalized Linear Mixed Models." *Wiley StatsRef: Statistics Reference Online*.

Nelder, John Ashworth, and Robert WM Wedderburn. 1972. "Generalized Linear Models." *Journal of the Royal Statistical Society: Series A (General)* 135 (3): 370–84.

Oldford, Wayne. 2020. *Qqtest: Self Calibrating Quantile-Quantile Plots for Visual Testing*.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. "PROC: An Open-Source Package for R and S+ to Analyze and Compare Roc Curves." *BMC Bioinformatics* 12: 77.

Sakamoto, Yosiyuki, Makio Ishiguro, and Genshiro Kitagawa. 1986. "Akaike Information Criterion Statistics." *Dordrecht, the Netherlands: D. Reidel* 81.

Schloerke, Barret, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Jason Crowley. 2020. *GGally: Extension to 'Ggplot2'*. https://CRAN.R-project.org/package=GGally.

Schwarz, Gideon, and others. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics* 6 (2): 461–64.

University, Virginia Commonwealth. 2012. "Datasource." https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Fourth. New York: Springer. http://www.stats.ox.ac.uk/pub/MASS4/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain Fran?ois, Lionel Henry, and Kirill Müller. 2020. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Zeileis, Achim, and Torsten Hothorn. 2002. "Diagnostic Checking in Regression Relationships." *R News* 2 (3): 7–10. https://CRAN.R-project.org/doc/Rnews/.