Jeanne Reppert
4/27/2020

IAF 606, Solving Problems with Data Analytics, Final Project Report

Introduction: The diabetes dataset used for this study contains 50 variables and 101,766 observations. The data was collected over 10 years from 130 hospitals and contains information regarding patient stays in which a diagnosis of diabetes was indicated. Variables include patient information, admission and discharge information, medications and diagnostic measures and a variety of patient information. The objective of this paper is to analyze the information chosen by the authors of a research paper from the Virginia Commonwealth University (*Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical database Patient Records*), replicate the models used and then attempt to find models that may make predictions in keeping with the goals of the research paper. The objective of the paper was to predict early readmissions (those that occurred within 30 days of discharge) and its association with HbA1c measures using the variables remaining after cleaning the data.

Step One - Data Preparation: The first step in the process was to prepare the data in the same way that the data was configured and cleaned by the authors of the study. Most of the variables were omitted, included medications, patient weight, payer information and a few other metrics. Diagnosis and medical specialty codes were used to consolidate categories and several other categories were reconfigured. Additionally, missing values were noted and listed as "other" in most categories. The HbA1c variable was reconfigured using the change variable to represent patients with high HbA1c measures and changed medications and those with high HbA1c measures with no change to medications. Once this variable was used to alter the HbA1c category, the change variable was also dropped.

The target variable chosen was readmissions. This category was reconfigured to be comprised of two subcategories. The first category consisted of observations with no readmissions and readmissions that occurred after a period of 30 days. The readmission category was comprised of observations where the patient was readmitted within 30 days of discharged. Additionally, to remove bias only one observation per patient was used significantly cutting the overall number of fnal observations. Likewise, observations where patients were discharged to a hospice or where patients died were also omitted. Once the data was cleaned 69970 observations were left and the following variables remained:

Continuous variables:

- Time spent in the hospital
- Age of patient (computed from age variable taking the median of the age in each category and assigning that value to a new variable)

Categorical variables:

- HbA1c (4 categories: no test, high with meds changed, high without meds changed, and normal results)
- Gender (2 categories: male and female)
- Discharge disposition (2 categories: discharged to home, other)
- Admission source (3 categories: admitted from emergency room, admitted from doctor or clinic referral, Other)
- Specialty of Admitting Doctor (6 categories: internal medicine, cardiology, surgery, family/general practice, missing, other)
- Primary diagnosis (9 categories: internal medicine, circulatory, diabetes, digestive, injury, musculoskeletal, genitourinary, neoplasms, other)
- Race (4 categories: African American, Caucasian, other, missing)
- Age (3 categories: <30, 30-60, >30
- Readmission (2 categories: readmission, other)

Final data set: 69### observations - the process (programmed in python) that was used to cleaned the data is found in the "Reppert-IAF-606-Final-Project-Part 1 Cleaning Data.rmd" file.

Building and analyzing the study model: (code found in Reppert-IAF-606-Final-Project-Part-2 Log Regression)

1. *Study model (logistic regression):*

a. The first step the authors used was to create a model with all variables except for HbA1c. Tests of significance of variables were obtained by looking at coefficients as variables were removed one at a time and analyzing the results. Then the HbAc1 variable was added to the core model obtained in the first step. One variable, gender, was found to not be significant and was thus removed from the model.

b. Next interactions were added to the model. Once again, this was done without HbA1c. All pairwise interactions were included and removed one at a time investigating each interaction for significance and eliminating those that were not significant. Once the best pairwise interactions were determined, HbA1c was added back into the model. The final model that was built was as follows:

Readmitted ~ Discharge + Race + Admit + Med_spec + Time_hosp + Age + Diag + HbA1c + Age:Med_spec + Diag:Discharge + Race:Discharge + Discharge:Time_hosp + Med_spec:Discharge + Time_hosp:Med_spec + Time_hosp:Diag + HbA1c:Diag

Jeanne Reppert
4/27/2020

Table 4 values:

Dependent variable: Readmitted	
	Coef.
DischargeTo home	-0.262, p = 0.116
RaceCauc	0.032, p = 0.549
RaceMissing	-0.031, p = 0.815
RaceOther	0.199, p = 0.067*
AdmitFrom ER	0.017, p = 0.587
AdmitOther	-0.115, p = 0.010***
Med_specFam/GenPract	-1.399, p = 0.102
Med_specIntMed	-1.059, p = 0.156
Med_specOther	-1.375, p = 0.055*
Med_specSurgery	-1.738, p = 0.157
Med_specUnknown	-0.438, p = 0.527
Time_hosp	0.067, p = 0.002***
Age> 60	-1.224, p = 0.064*
Age30-60	-1.477, p = 0.027**
DiagDiabetes	-0.623, p = 0.002***
DiagDig	-0.154, p = 0.609
DiagGenit	-0.650, p = 0.044**
DiagInjury	-0.500, p = 0.167
DiagMusc/skel	-0.136, p = 0.720
DiagNeoplasms	-0.743, p = 0.191
DiagOther	-0.125, p = 0.525
DiagResp	-0.512, p = 0.027**
HbA1cHigh no change	-0.159, p = 0.320
HbA1cNo test	-0.150, p = 0.118
HbA1cNormal	-0.196, p = 0.093*

Table 5 values:

Dependent variable: Readmitted	
	Coef.
Med_specFam/GenPract:Age> 60	2.196, p = 0.009***
Med_specIntMed:Age> 60	1.641, p = 0.024**
Med_specOther:Age> 60	1.947, p = 0.006***
Med_specSurgery:Age> 60	2.662, p = 0.028**
Med_specUnknown:Age> 60	1.010, p = 0.133
Med_specFam/GenPract:Age30-60	2.121, p = 0.012**
Med_specIntMed:Age30-60	1.644, p = 0.025**
Med_specOther:Age30-60	2.050, p = 0.004***
Med_specSurgery:Age30-60	2.817, p = 0.021**
Med_specUnknown:Age30-60	-1.092, p = 0.108
DischargeTo home:DiagDiabetes	-0.026, p = 0.814
DischargeTo home:DiagDig	-0.050, p = 0.661
DischargeTo home:DiagGenit	0.162, p = 0.230
DischargeTo home:DiagInjury	-0.306, p = 0.013**
DischargeTo home:DiagMusc/skel	-0.444, p = 0.004***
DischargeTo home:DiagNeoplasms	0.139, p = 0.376
DischargeTo home:DiagOther	-0.227, p = 0.007***
DischargeTo home:DiagResp	-0.131, p = 0.180
DischargeTo home:RaceCauc	-0.018, p = 0.799
DischargeTo home:RaceMissing	-0.285, p = 0.130
DischargeTo home:RaceOther	-0.496, p = 0.001***
DischargeTo home:Time_hosp	0.028, p = 0.003***
DischargeTo home:Med_specFam/GenPract	-0.324, p = 0.073*
DischargeTo home:Med_specIntMed	-0.196, p = 0.235
DischargeTo home:Med_specOther	-0.393, p = 0.018**
DischargeTo home:Med_specSurgery	-0.681, p = 0.001***
DischargeTo home:Med_specUnknown	-0.244, p = 0.113
Med_specFam/GenPract:Time_hosp	-0.062, p = 0.018**
Med_specIntMed:Time_hosp	-0.037, p = 0.108

```
Med_specOther:Time_hosp                      -0.052, p = 0.027**
Med_specSurgery:Time_hos                     -0.110, p = 0.0002***
Med_specUnknown:Time_hosp                    -0.058, p = 0.007***
Time_hosp:DiagDiabetes                        0.034, p = 0.043**
Time_hosp:DiagDig                             0.0003, p = 0.987
Time_hosp:DiagGenit                           0.076, p = 0.0005***
Time_hosp:DiagInjury                         -0.009, p = 0.639
Time_hosp:DiagMusc/skel                       0.057, p = 0.023**
Time_hosp:DiagNeoplasms                      -0.013, p = 0.590
Time_hosp:DiagOther                          -0.023, p = 0.080*
Time_hosp:DiagResp                            0.026, p = 0.093*
DiagDiabetes:HbA1cHigh no change              0.026, p = 0.929
DiagDig:HbA1cHigh no change                  -0.293, p = 0.533
DiagGenit:HbA1cHigh no change                -0.765, p = 0.252
DiagInjury:HbA1cHigh no change                0.457, p = 0.420
DiagMusc/skel:HbA1cHigh no change             0.031, p = 0.959
DiagNeoplasms:HbA1cHigh no change             0.852, p = 0.298
DiagOther:HbA1cHigh no change                 0.399, p = 0.173
DiagResp:HbA1cHigh no change                  0.061, p = 0.862
DiagDiabetes:HbA1cNo test                     0.542, p = 0.002***
DiagDig:HbA1cNo test                          0.033, p = 0.902
DiagGenit:HbA1cNo test                        0.102, p = 0.725
DiagInjury:HbA1cNo test                       0.697, p = 0.045**
DiagMusc/skel:HbA1cNo test                   -0.251, p = 0.471
DiagNeoplasms:HbA1cNo test                    0.652, p = 0.222
DiagOther:HbA1cNo test                        0.278, p = 0.118
DiagResp:HbA1cNo test                         0.219, p = 0.291
DiagDiabetes:HbA1cNormal                      0.595, p = 0.008***
DiagDig:HbA1cNormal                           0.013, p = 0.968
DiagGenit:HbA1cNormal                         0.375, p = 0.274
DiagInjury:HbA1cNormal                        0.152, p = 0.709
DiagMusc/skel:HbA1cNormal                    -0.292, p = 0.483
DiagNeoplasms:HbA1cNormal                     0.948, p = 0.106
DiagOther:HbA1cNormal                         0.258, p = 0.224
DiagResp:HbA1cNormal                         -0.216, p = 0.399
=================================================================
  Note:                                *p<0.1; **p<0.05; ***p<0.01
```

2. Model analysis:

Once the model was created, the dataset was split into training and test set at a 70/30 ratio.

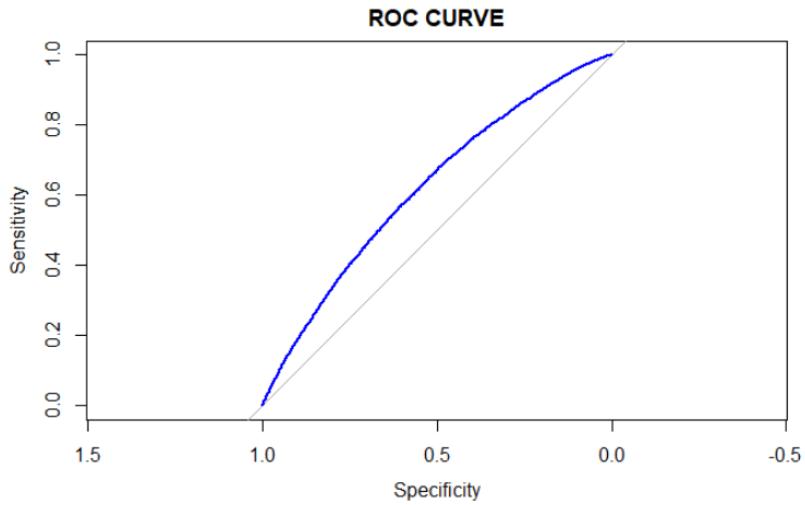
A sampling of model probabilities was taken and are as follows:

```
        1         2         3         4         5         6         7         8         9        10
0.06259691 0.09288183 0.09057588 0.06425587 0.07185845 0.07332110 0.08881165 0.09334290 0.10836918 0.15509099
```

Based on these results a cutoff value of .1 was established for model prediction. The confusion matrix of the model is as follows:

glm.pred	Other	Readmitted
Other	10844	888
Readmitted	7943	1201

The model generated a ROC plot as follows:



The accuracy/efficiency calculated was 0.5769783 and the AUC score was 0.6189. The classification rate was calculated as 57.7. Other measures for this model are as follows:

Precision: 0.9243096
Specificity: 0.5749162
Sensitivity: 0.5772076
F1_Score: 0.7106393
J: 0.1521239

3. *Conclusions for paper model:*

Overall the outcome for predictions gave mixed results. An accuracy of 0.6189 would not appear to be ideal in terms of usefulness. There were many false negatives and false positives within the confusion matrix. While the precision score was relatively high, the specificity scores and sensitivity scores were not quite as impressive leading to a relatively low Youden Index score (J).

Creating my own logistic regression model:

1. *Building the model:*

In my attempts to create a better model, I tried a few approaches. The first was to experiment with a variety of interactions much in the way the paper added interactions. It became clear to me that there was potential with utilizing the admit and diagnosis variable more than the paper model used it. Additionally, I

experimented with using the numeric value for age instead of the age ranges. Several comparisons were made between various models with results reported below. All the models were constructed without the HbA1c variable. This variable was then added to the model that was determined to be best. Variables that were considered to be significant had a Pr(>Chisq) value below 0.05.

The Admit:Age variable was added and appeared to be significant with a Pr(>Chisq) of 0.008508. The ANOVA output for the model with the interaction of Admit:Age added was as follows:

```
Analysis of Deviance Table (Type II tests)

Response: Readmitted
                  LR Chisq Df Pr(>Chisq)
Discharge            340.93  1  < 2.2e-16 ***
Race                   7.10  3   0.068780 .
Admit                  9.63  2   0.008123 **
Med_spec              29.81  5  1.608e-05 ***
Time_hosp             46.81  1  7.803e-12 ***
Age                   41.28  2  1.088e-09 ***
Diag                  60.38  8  3.932e-10 ***
Med_spec:Age          33.70 10   0.000208 ***
Discharge:Diag        23.88  8   0.002399 **
Discharge:Race        14.19  3   0.002662 **
Discharge:Time_hosp    9.10  1   0.002555 **
Admit:Age             13.65  4   0.008508 **
Med_spec:Time_hosp    17.03  5   0.004445 **
Time_hosp:Diag        33.74  8  4.527e-05 ***
Discharge:Med_spec    18.01  5   0.002938 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The results for age_numeric substituted for the age variable did not produce a significant enough improvement over the age category (the age category's values were higher) to merit a substitution although the age_numeric variable and interactions did have significant values.

```
Analysis of Deviance Table (Type II tests)

Response: Readmitted
                    LR Chisq Df Pr(>Chisq)
Discharge             328.85  1  < 2.2e-16 ***
Race                    7.26  3  0.0641513 .
Admit                   9.32  2  0.0094898 **
Med_spec               24.34  5  0.0001870 ***
Time_hosp              42.94  1  5.639e-11 ***
Age_numeric            22.34  1  2.280e-06 ***
Diag                   58.35  8  9.814e-10 ***
Med_spec:Age_numeric   16.16  5  0.0064118 **
Discharge:Diag         29.82  8  0.0002273 ***
Discharge:Race         14.08  3  0.0027976 **
Discharge:Time_hosp     9.31  1  0.0022840 **
Admit:Diag             34.84 16  0.0041770 **
Med_spec:Time_hosp     10.76  5  0.0562980 .
Time_hosp:Diag         36.09  8  1.690e-05 ***
Admit:Age_numeric       7.86  2  0.0196387 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The Admit:Diag variable was added and appeared to be significant with a Pr(>Chisq) of 0.0060984. The ANOVA output for the model with the interaction of Admit:Diag added was as follows:

```
Analysis of Deviance Table (Type II tests)

Response: Readmitted
                    LR Chisq Df Pr(>Chisq)
Discharge             340.31  1  < 2.2e-16 ***
Race                    7.08  3  0.0694450 .
Admit                   9.63  2  0.0081230 **
Med_spec               25.20  5  0.0001277 ***
Time_hosp              41.89  1  9.674e-11 ***
Age                    42.19  2  6.900e-10 ***
Diag                   60.38  8  3.932e-10 ***
Med_spec:Age           33.75 10  0.0002032 ***
Discharge:Diag         24.95  8  0.0015849 **
Discharge:Race         14.01  3  0.0028879 **
Discharge:Time_hosp     8.85  1  0.0029373 **
Admit:Diag             33.52 16  0.0063059 **
Admit:Age              11.41  4  0.0222926 *
Med_spec:Time_hosp     16.32  5  0.0059910 **
Time_hosp:Diag         35.57  8  2.106e-05 ***
Discharge:Med_spec     17.90  5  0.0030765 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Med_spec:Diag variable was added and appeared to be significant with a . Pr(>Chisq) of 0.0023047. The ANOVA output for the model with the interaction of Med_spec:Diag added was as follows:

```
Analysis of Deviance Table (Type II tests)

Response: Readmitted
                    LR Chisq Df Pr(>Chisq)
Discharge             339.36  1  < 2.2e-16 ***
Race                    7.17  3  0.0666202 .
Admit                   9.16  2  0.0102667 *
Med_spec               25.20  5  0.0001277 ***
Time_hosp              40.90  1  1.601e-10 ***
Age                    40.56  2  1.557e-09 ***
Diag                   60.38  8  3.932e-10 ***
Med_spec:Age           25.85 10  0.0039511 **
Discharge:Diag         28.40  8  0.0004040 ***
Discharge:Race         14.38  3  0.0024342 **
Discharge:Time_hosp     8.85  1  0.0029351 **
Admit:Diag             30.06 16  0.0176874 *
Admit:Age              11.69  4  0.0197946 *
Med_spec:Time_hosp     14.50  5  0.0127182 *
Time_hosp:Diag         38.84  8  5.260e-06 ***
Discharge:Med_spec     15.61  5  0.0080596 **
Med_spec:Diag          70.04 40  0.0023047 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interactions that did not improve the model: Admit:Med_Spec (Pr(>Chisq) was 0.0915773), Race:Diag (Pr(>Chisq) was 0.1112454), and Race:Admit (0.2179067). These interactions were all explored with the other interactions above due to the potential with Admit and Diag interactions being potentially significant. Nonetheless, these combinations did not produce significant values in the model results. Additionally, the interaction between time in hospital and medical specialty was removed from the model to explore if this interaction improved it since at times this value was a bit above the cutoff for significance of .05. Despite this concern the anova results demonstrated that the addition of this interaction was indeed an improve with a Pr(>Chisq) of 0.004445 with the interaction added.

The final model that was chosen was model_tb5.4:

Readmitted ~ Discharge + Race + Admit + Med_spec + Time_hosp + Age + Diag + Age:Med_spec + Diag:Discharge + Race:Discharge +Discharge:Time_hosp + Admit:Diag + Admit:Age + Time_hosp:Med_spec + Time_hosp:Diag + Med_spec:Discharge + Med_spec:Diag

Once this model was chosen, HbA1c and the interaction HbA1c and Diagnosis were added to the model. The interaction between HbA1c and Admit (Pr(>Chisq) of 0.7501402) and the interaction between HbA1c and Medical Specialty (Pr(>Chisq) of 0.1053272 were also explored but neither interaction was significant so they were omitted. The ANOVA output for model_tb5.4F is as follows:

```
Analysis of Deviance Table (Type II tests)

Response: Readmitted
                  LR Chisq Df Pr(>Chisq)
Discharge            338.31  1  < 2.2e-16 ***
Race                   7.10  3  0.0689012 .
Admit                 10.09  2  0.0064255 **
Med_spec              26.35  5  7.636e-05 ***
Time_hosp             41.34  1  1.281e-10 ***
Age                   37.14  2  8.621e-09 ***
Diag                  61.97  8  1.911e-10 ***
HbA1c                  8.04  3  0.0451569 *
Med_spec:Age          23.82 10  0.0080790 **
Discharge:Diag        29.17  8  0.0002955 ***
Discharge:Race        14.37  3  0.0024477 **
Discharge:Time_hosp    8.95  1  0.0027680 **
Admit:Diag            30.02 16  0.0179053 *
Admit:Age             11.73  4  0.0194410 *
Med_spec:Time_hosp    14.62  5  0.0121066 *
Time_hosp:Diag        36.74  8  1.285e-05 ***
Discharge:Med_spec    15.58  5  0.0081422 **
Med_spec:Diag         69.04 40  0.0029276 **
Diag:HbA1c            42.25 24  0.0121116 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

2. *Final Model:*

Readmitted ~ Discharge + Race + Admit + Med_spec + Time_hosp + Age + Diag + HbA1c + Age:Med_spec + Diag:Discharge + Race:Discharge +Discharge:Time_hosp + Admit:Diag + Admit:Age + Time_hosp:Med_spec + Time_hosp:Diag + Med_spec:Discharge + Med_spec:Diag + HbA1c:Diag

Figure 3 Results for model_tb5.4F:

```
=================================================================

                                        Dependent variable:

                                    ---------------------------

                                             Readmitted

 -----------------------------------------------------------------

DischargeTo home                          -0.212, p = 0.210
RaceCauc                                   0.026, p = 0.624
RaceMissing                               -0.040, p = 0.764
RaceOther                                  0.197, p = 0.070*
AdmitFrom ER                               0.156, p = 0.530
AdmitOther                                 0.111, p = 0.772
Med_specFam/GenPract                      -0.925, p = 0.303
Med_specIntMed                            -0.800, p = 0.313
Med_specOther                             -0.987, p = 0.195
Med_specSurgery                           -1.599, p = 0.206
Med_specUnknown                           -0.116, p = 0.875
Time_hosp                                  0.069, p = 0.002***
Age> 60                                   -0.873, p = 0.213
Age30-60                                  -1.196, p = 0.090*
DiagDiabetes                              -0.002, p = 0.998
DiagDig                                   -0.280, p = 0.623
DiagGenit                                 -0.439, p = 0.546
```

```
DiagInjury                                    0.270, p = 0.572
DiagMusc/skel                               -10.936, p = 0.934
DiagNeoplasms                               -12.097, p = 0.905
DiagOther                                     0.412, p = 0.200
DiagResp                                     -0.378, p = 0.265
HbA1cHigh no change                          -0.166, p = 0.300
HbA1cNo test                                 -0.142, p = 0.140
HbA1cNormal                                  -0.202, p = 0.084*



================================================================
```

Figure 5 results for model_tb5.4F:

```
================================================================

                                        Dependent variable:

                                    ---------------------------

                                           Readmitted

 ----------------------------------------------------------------

Med_specFam/GenPract:Age> 60                  1.947, p = 0.026**
Med_specIntMed:Age> 60                        1.369, p = 0.074*
Med_specOther:Age> 60                         1.599, p = 0.029**
Med_specSurgery:Age> 60                       2.428, p = 0.050**
Med_specUnknown:Age> 60                       0.743, p = 0.295
Med_specFam/GenPract:Age30-60                 1.891, p = 0.031**
Med_specIntMed:Age30-60                       1.411, p = 0.067*
Med_specOther:Age30-60                        1.746, p = 0.018**
Med_specSurgery:Age30-60                      2.600, p = 0.036**
Med_specUnknown:Age30-60                      0.878, p = 0.219
DischargeTo home:DiagDiabetes                -0.011, p = 0.919
DischargeTo home:DiagDig                     -0.057, p = 0.618
DischargeTo home:DiagGenit                    0.128, p = 0.355
DischargeTo home:DiagInjury                  -0.336, p = 0.007***
DischargeTo home:DiagMusc/skel               -0.498, p = 0.002***
DischargeTo home:DiagNeoplasms                0.169, p = 0.288
DischargeTo home:DiagOther                   -0.263, p = 0.002***
DischargeTo home:DiagResp                    -0.145, p = 0.143
DischargeTo home:RaceCauc                    -0.018, p = 0.802
DischargeTo home:RaceMissing                 -0.287, p = 0.129
DischargeTo home:RaceOther                   -0.503, p = 0.001***
DischargeTo home:Time_hosp                    0.028, p = 0.003***
AdmitFrom ER:DiagDiabetes                    -0.174, p = 0.166
AdmitOther:DiagDiabetes                       0.029, p = 0.871
AdmitFrom ER:DiagDig                         -0.086, p = 0.494
AdmitOther:DiagDig                            0.196, p = 0.263
AdmitFrom ER:DiagGenit                       -0.290, p = 0.055*
AdmitOther:DiagGenit                          0.027, p = 0.905
AdmitFrom ER:DiagInjury                       0.123, p = 0.341
AdmitOther:DiagInjury                        -0.022, p = 0.908
AdmitFrom ER:DiagMusc/skel                    0.171, p = 0.340
AdmitOther:DiagMusc/skel                      0.027, p = 0.900
AdmitFrom ER:DiagNeoplasms                    0.320, p = 0.060*
AdmitOther:DiagNeoplasms                      0.212, p = 0.436
AdmitFrom ER:DiagOther                       -0.281, p = 0.003***
AdmitOther:DiagOther                         -0.055, p = 0.665
AdmitFrom ER:DiagResp                        -0.187, p = 0.113
AdmitOther:DiagResp                          -0.047, p = 0.770
AdmitFrom ER:Age> 60                         -0.057, p = 0.815
AdmitOther:Age> 60                           -0.320, p = 0.395
AdmitFrom ER:Age30-60                        -0.080, p = 0.745
AdmitOther:Age30-60                          -0.029, p = 0.940
Med_specFam/GenPract:Time_hosp               -0.062, p = 0.018**
Med_specIntMed:Time_hosp                     -0.039, p = 0.098*
```

```
Med_specOther:Time_hosp                   -0.058, p = 0.015**
Med_specSurgery:Time_hosp                 -0.101, p = 0.001***
Med_specUnknown:Time_hosp                 -0.060, p = 0.007***
Time_hosp:DiagDiabetes                     0.032, p = 0.061*
Time_hosp:DiagDig                         -0.002, p = 0.922
Time_hosp:DiagGenit                        0.081, p = 0.0003***
Time_hosp:DiagInjury                      -0.012, p = 0.515
Time_hosp:DiagMusc/skel                    0.056, p = 0.031**
Time_hosp:DiagNeoplasms                   -0.023, p = 0.331
Time_hosp:DiagOther                       -0.029, p = 0.032**
Time_hosp:DiagResp                         0.024, p = 0.124
DischargeTo home:Med_specFam/GenPract     -0.367, p = 0.044**
DischargeTo home:Med_specIntMed           -0.233, p = 0.164
DischargeTo home:Med_specOther            -0.461, p = 0.007***
DischargeTo home:Med_specSurgery          -0.634, p = 0.002***
DischargeTo home:Med_specUnknown          -0.288, p = 0.065*
Med_specFam/GenPract:DiagDiabetes         -0.657, p = 0.140
Med_specIntMed:DiagDiabetes               -0.465, p = 0.279
Med_specOther:DiagDiabetes                -0.762, p = 0.086*
Med_specSurgery:DiagDiabetes              -0.228, p = 0.631
Med_specUnknown:DiagDiabetes              -0.444, p = 0.288
Med_specFam/GenPract:DiagDig              -0.053, p = 0.916
Med_specIntMed:DiagDig                     0.268, p = 0.580
Med_specOther:DiagDig                      0.312, p = 0.527
Med_specSurgery:DiagDig                    0.172, p = 0.738
Med_specUnknown:DiagDig                    0.155, p = 0.745
Med_specFam/GenPract:DiagGenit             0.039, p = 0.954
Med_specIntMed:DiagGenit                  -0.202, p = 0.758
Med_specOther:DiagGenit                    0.108, p = 0.869
Med_specSurgery:DiagGenit                  0.777, p = 0.301
Med_specUnknown:DiagGenit                  0.010, p = 0.987
Med_specFam/GenPract:DiagInjury           -0.918, p = 0.019**
Med_specIntMed:DiagInjury                 -0.769, p = 0.026**
Med_specOther:DiagInjury                  -0.912, p = 0.006***
Med_specSurgery:DiagInjury                -0.745, p = 0.046**
Med_specUnknown:DiagInjury                -0.852, p = 0.008***
Med_specFam/GenPract:DiagMusc/skel        10.603, p = 0.936
Med_specIntMed:DiagMusc/skel              10.671, p = 0.936
Med_specOther:DiagMusc/skel               10.801, p = 0.935
Med_specSurgery:DiagMusc/skel             11.261, p = 0.932
Med_specUnknown:DiagMusc/skel             10.672, p = 0.936
Med_specFam/GenPract:DiagNeoplasms        10.614, p = 0.916
Med_specIntMed:DiagNeoplasms              11.511, p = 0.909
Med_specOther:DiagNeoplasms               11.221, p = 0.912
Med_specSurgery:DiagNeoplasms             10.818, p = 0.915
Med_specUnknown:DiagNeoplasms             11.285, p = 0.911
Med_specFam/GenPract:DiagOther            -0.638, p = 0.024**
Med_specIntMed:DiagOther                  -0.108, p = 0.681
Med_specOther:DiagOther                   -0.184, p = 0.489
Med_specSurgery:DiagOther                 -0.411, p = 0.223
Med_specUnknown:DiagOther                 -0.373, p = 0.138
Med_specFam/GenPract:DiagResp             -0.457, p = 0.108
Med_specIntMed:DiagResp                    0.029, p = 0.911
Med_specOther:DiagResp                     0.216, p = 0.420
Med_specSurgery:DiagResp                   0.277, p = 0.526
Med_specUnknown:DiagResp                   0.012, p = 0.963
DiagDiabetes:HbA1cHigh no change           0.041, p = 0.889
DiagDig:HbA1cHigh no change               -0.306, p = 0.515
DiagGenit:HbA1cHigh no change             -0.758, p = 0.257
DiagInjury:HbA1cHigh no change             0.501, p = 0.377
DiagMusc/skel:HbA1cHigh no change          0.090, p = 0.882
DiagNeoplasms:HbA1cHigh no change          0.873, p = 0.289
DiagOther:HbA1cHigh no change              0.367, p = 0.211
DiagResp:HbA1cHigh no change               0.103, p = 0.769
DiagDiabetes:HbA1cNo test                  0.503, p = 0.004***
DiagDig:HbA1cNo test                       0.017, p = 0.951
DiagGenit:HbA1cNo test                     0.021, p = 0.942
DiagInjury:HbA1cNo test                    0.717, p = 0.040**
DiagMusc/skel:HbA1cNo test                -0.210, p = 0.549
DiagNeoplasms:HbA1cNo test                 0.769, p = 0.153
```

```
DiagOther:HbA1cNo test                        0.241, p = 0.176
DiagResp:HbA1cNo test                         0.222, p = 0.287
DiagDiabetes:HbA1cNormal                      0.582, p = 0.010***
DiagDig:HbA1cNormal                           0.022, p = 0.945
DiagGenit:HbA1cNormal                         0.358, p = 0.297
DiagInjury:HbA1cNormal                        0.164, p = 0.686
DiagMusc/skel:HbA1cNormal                    -0.267, p = 0.523
DiagNeoplasms:HbA1cNormal                     1.001, p = 0.090*
DiagOther:HbA1cNormal                         0.256, p = 0.229
DiagResp:HbA1cNormal                         -0.203, p = 0.428
 ---------------------------------------------------------------
 Observations                                        69,987
 ===============================================================
 Note:                                *p<0.1; **p<0.05; ***p<0.01
```

3. Final Model Analysis

Once the model was created, the dataset was split into training and test set at a 70/30 ratio.

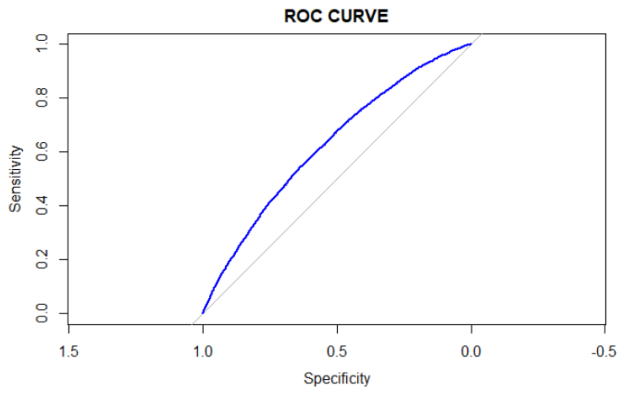
A sampling of model probabilities was taken and are as follows:

```
        4         5         8        11        15        16        19        22        26        27
0.0542601 0.1042721 0.0986191 0.1069317 0.0608526 0.1566370 0.0978859 0.0501543 0.0341524 0.1039197
```

Based on these results a cutoff value of .1 was established for model prediction. The confusion matrix of the model is as follows:

glm.pred2	Other	Readmitted
Other	10859	873
Readmitted	7897	1225

The model generated a ROC plot as follows:



The accuracy/efficiency calculated was 0.5794572 and the AUC score was 0.6251. The classification rate for the final model (model_tb5.4F) was 57.95. Other measures for this model are as follows:

```
Precision: 0.9255881
Specificity: 0.5838894
Sensitivity: 0.5789614
F1_Score: 0.7123458
J: 0.1628508
```

3. *Conclusions for paper mode*

Overall, the outcome for predictions still was rather lackluster. The quantitative measures for the "improved model" showed some small improvements but there were still deficits in the model performance. There were still many false negatives and false positives within the confusion matrix (although there were less). The precision score was still relatively high, but the specificity scores and sensitivity scores were only a little higher than the paper model with a slightly higher Youden Index score of 0.1628508. Overall, more readmissions were predicted, which is good, however total readmissions predicted correctly is still, in my opinion, not a very strong number.

Creating a random forest model: (code found in file - Reppert-IAF-606-Final-Part-3-Random-Forest.rmd)

1. *Building the model:*

The model chosen for further exploring the diabetes dataset was the Random Forest model. The first step in the model building process was to fit the model with all of the variables. The same cutoff parameters were used (.10) so that the analysis of all models would be consistent. Some time was spent using the caret package model for random forest in order to do parameter tuning to find best mtry values and the best number of trees. In the end, after running several tests, it was found that the default values showed the best results. Most of these tests took several hours to run even with additional cores allocated to them using a parallel package to assign 6 additional cores to the model creation. For that reason this code, since it was not used in the final model, is not included. The random forest package in R was used instead since this package had a much more efficient means of creating models that only took a few minutes per model.

The dataset was split into training and test sets at a 70/30 ratio and the model was run using the training set. Predictions were then made using the test set. These predictions were used to create a confusion matrix to analyze results. The model was also used to find variable importance values in order to ascertain

which variables were rated as most important in the model. Based on this information, subsequent models were created by removing the variable with the smallest variable importance value, one at a time.

Results of the full model (Model 1) with training set:

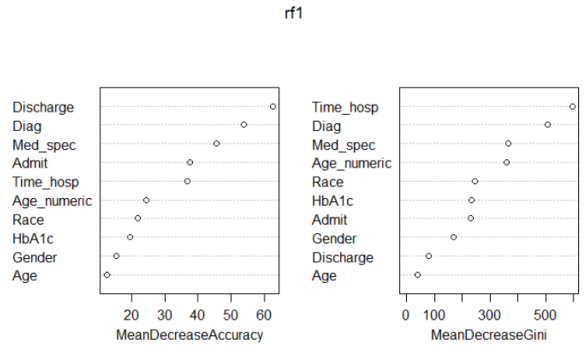
```
Call:
 randomForest(formula = Readmitted ~ ., data = train, importance = TRUE,      cutoff = c(0.9, 0.1))
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 17.97%
Confusion matrix:
            Other Readmitted class.error
Other       39523       5068   0.1136552
Readmitted   3734        666   0.8486364
```

The full model (Model 1) was used to produce a confusion matrix after using the model to make predictions. The model is quite overfitted with all the variables but still has a pretty high error rate in the training set. The quantitative measures are as follows:

```
Confusion Matrix and Statistics

            Reference
Prediction   Other Readmitted
  Other      16864       1612
  Readmitted  2247        273

               Accuracy : 0.8162
                 95% CI : (0.8109, 0.8214)
    No Information Rate : 0.9102
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0237

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8824
            Specificity : 0.1448
         Pos Pred Value : 0.9128
         Neg Pred Value : 0.1083
              Precision : 0.9128
                 Recall : 0.8824
                     F1 : 0.8973
             Prevalence : 0.9102
         Detection Rate : 0.8032
   Detection Prevalence : 0.8800
      Balanced Accuracy : 0.5136

       'Positive' Class : Other
```

	Other <dbl>	Readmitted <dbl>
Race	7.523115	7.523115
Gender	6.558730	6.558730
Age	3.652420	3.652420
Discharge	29.947925	29.947925
Admit	15.849180	15.849180
Time_hosp	19.059204	19.059204
Med_spec	20.543162	20.543162
Diag	22.757935	22.757935
HbA1c	9.137579	9.137579
Age_numeric	5.457839	5.457839

Variable Importance Plot:



Youden Index:

```
Sensitivity
  0.8824237
Specificity
  0.1448276
J: 0.0608244
```

The least important variable in both cases is the age variable (three categories) followed by discharge and gender. A variety of combinations were tried in addition to removing one variable at a time; however, none of the various combinations had measures that were preferable to that of the full model and most had significantly poorer quantitative measures. In particular, after removing anything other than the age category (either numeric or categorical) the specificity measure fell significantly. Ultimately, these observations led to utilizing all of the selected variables but removing the categorical variable for age.

Examples of some other model attempts:

Model with discharge omitted:

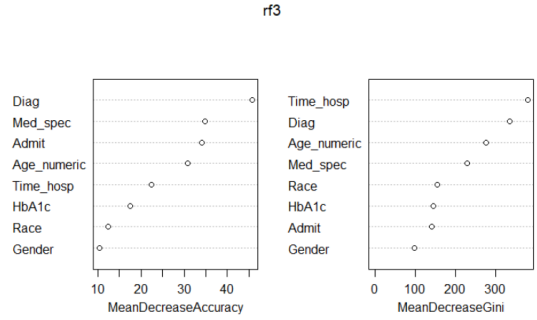
Training output:

```
Call:
 randomForest(formula = Readmitted ~ Time_hosp + Diag + Med_spec +      Age_numeric + Race + HbA1c + Admit +
Gender, data = train,      importance = TRUE, cutoff = c(0.9, 0.1))
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 13.05%
Confusion matrix:
           Other Readmitted class.error
Other      42325       2266  0.05081743
Readmitted  4125        275  0.93750000
```

Test output:

```
Confusion Matrix and Statistics

          Reference
Prediction   Other Readmitted
  Other      18176       1778
  Readmitted   935        107

               Accuracy : 0.8708
                 95% CI : (0.8662, 0.8753)
    No Information Rate : 0.9102
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0098

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.95108
            Specificity : 0.05676
         Pos Pred Value : 0.91090
         Neg Pred Value : 0.10269
              Precision : 0.91090
                 Recall : 0.95108
                     F1 : 0.93055
             Prevalence : 0.91022
         Detection Rate : 0.86569
   Detection Prevalence : 0.95037
      Balanced Accuracy : 0.50392

       'Positive' Class : Other
```

	Other <dbl>	Readmitted <dbl>
Time_hosp	14.764940	14.764940
Diag	20.375188	20.375188
Med_spec	14.934433	14.934433
Age_numeric	11.827089	11.827089
Race	3.711477	3.711477
HbA1c	5.871925	5.871925
Admit	14.312745	14.312745
Gender	4.406117	4.406117



Youden Index:

```
Sensitivity
  0.9510753
Specificity
 0.05676393
J3: 0.00783923
```

Jeanne Reppert
4/27/2020

Model 4 (removing gender):

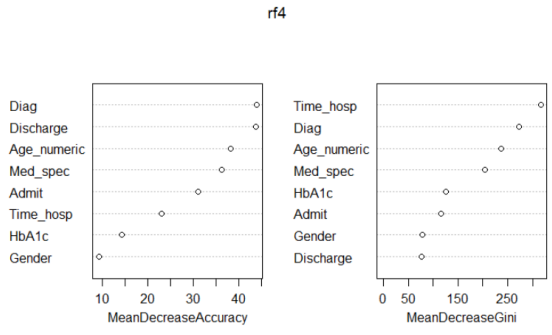
Training output:

```
Call:
 randomForest(formula = Readmitted ~ Time_hosp + Diag + Med_spec +      Age_numeric + Gender + HbA1c + Admit +
Discharge, data = train,      importance = TRUE, cutoff = c(0.9, 0.1))
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 11.56%
Confusion matrix:
           Other Readmitted class.error
Other      43141       1450  0.03251777
Readmitted  4211        189  0.95704545
```

Test output:

```
Confusion Matrix and Statistics

           Reference
Prediction   Other Readmitted
  Other      18214       1782
  Readmitted   897        103

               Accuracy : 0.8724
                 95% CI : (0.8678, 0.8769)
    No Information Rate : 0.9102
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0098

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.95306
            Specificity : 0.05464
         Pos Pred Value : 0.91088
         Neg Pred Value : 0.10300
              Precision : 0.91088
                 Recall : 0.95306
                     F1 : 0.93150
             Prevalence : 0.91022
         Detection Rate : 0.86750
   Detection Prevalence : 0.95237
      Balanced Accuracy : 0.50385

       'Positive' Class : Other
```

	Other <dbl>	Readmitted <dbl>
Time_hosp	11.735735	11.735735
Diag	17.551729	17.551729
Med_spec	14.706231	14.706231
Age_numeric	13.058630	13.058630
Gender	4.589535	4.589535
HbA1c	6.098576	6.098576
Admit	12.974454	12.974454
Discharge	19.201083	19.201083



Youden Index:

```
Sensitivity
  0.9530637
Specificity
 0.05464191
J4: 0.00770561
```

2. Final random forest model analysis:

The only model that appeared to be preferable to the full model was the model where categorical age was dropped:

Readmitted ~ Time_hosp + Diag + Med_spec + Age_numeric + Race + HbA1c + Admit + Gender + Discharge

This model had a higher balanced accuracy and did succeed to correctly classify more readmissions leading to a specificity rate that was higher than the other models. The higher specificity rate appeared to come at a cost of a lower sensitivity rate. Although the Youden Index was lower for this model (0.0427076) the other quantitative measures led me to choose this model over the full model.

3. Final random forest model quantitative measures:

Training output for final model:

```
Call:
 randomForest(formula = Readmitted ~ Time_hosp + Diag + Med_spec +      Age_numeric + Race + HbA1c + Admit +
Gender + Discharge,      data = train, importance = TRUE, cutoff = c(0.9, 0.1))
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 19.81%
Confusion matrix:
             Other Readmitted class.error
Other      38481       6110   0.1370232
Readmitted  3596        804   0.8172727
```

Test output and confusion matrix:

```
Confusion Matrix and Statistics

           Reference
Prediction   Other Readmitted
  Other      16399       1537
  Readmitted  2712        348

               Accuracy : 0.7976
                 95% CI : (0.7921, 0.803)
    No Information Rate : 0.9102
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0333

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8581
            Specificity : 0.1846
         Pos Pred Value : 0.9143
         Neg Pred Value : 0.1137
              Precision : 0.9143
                 Recall : 0.8581
                     F1 : 0.8853
             Prevalence : 0.9102
         Detection Rate : 0.7811
   Detection Prevalence : 0.8543
      Balanced Accuracy : 0.5214

       'Positive' Class : Other
```

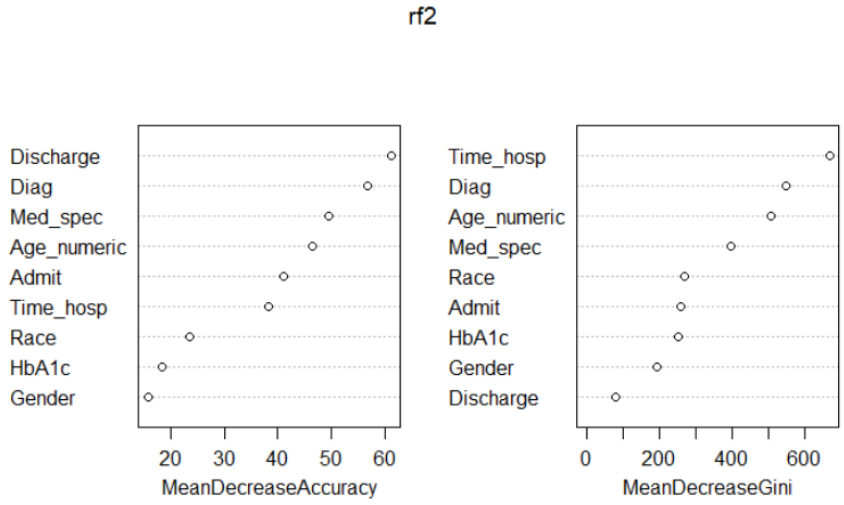
Youden Index:

```
Sensitivity
  0.8580922
Specificity
  0.1846154
J2: 0.0427076
```

Jeanne Reppert
4/27/2020

Final Model Feature Importances:

	Other <dbl>	Readmitted <dbl>
Time_hosp	19.777130	19.777130
Diag	25.318087	25.318087
Med_spec	23.569592	23.569592
Age_numeric	16.030257	16.030257
Race	8.237336	8.237336
HbA1c	8.725223	8.725223
Admit	17.377268	17.377268
Gender	8.098224	8.098224
Discharge	29.634945	29.634945



Final Conclusions:

Overall, the improved logistic regression model appeared to be able to accurately predict more readmissions than either the paper model (although not by a large margin). The logistic regression model proved to be far more effective than the random forest model.

For future study, it could be possible to improve the random forest model by adding in some of the dataset variables that were omitted by the paper. More variable choices could lead to a better performance using the random forest model. Another possible problem with the random forest model might have been the imbalance of the target value.

Once possible solution to this imbalance that could be explored would involve subsetting the data in some fashion and exploring models with a smaller more

balanced set of data. The HbA1c variable could also be used to subset the data by selecting out only the observations where an HbA1c test was performed. It could be possible that some results could be identified by the random forest model given some of these adjustments to the dataset.