# Predicting readmission probability for diabetes inpatients

*Juan Manubens*

*Due: April 7, 2017 at 11:59PM*

## Executive Summary

Diabetes, while treatable, is a very invasive medical condition affecting millions of people in the United States. Diabetes can be treated and managed by healthful eating, regular physical activity, and medications to lower blood glucose levels. The 2014 National Diabetes Statistics Report estimated that 29.1 million people or 9.3% of the U.S. population have diabetes, but only 21 million are diagnosed, leaving 27.8% of all cases concealed. The rates of new cases of diagnosed is highest for people aged 45-64, and poor management of the disease can lead to hospital readmission, which is a burden for both clinics and patients. The cost of readmissions is quite high [1], and is a key risk factor for hospitals, particularly since the Centers for Medicare and Medicaid Services announced in 2012 that they would no longer reimburse hospitals for services provided if a patient was readmitted with complications within 30 days of being released from care. Thus, building predictive models in this domain has clear value.

This project uses a data set from the Center for Clinical and Transnational Research at Virginia Commonwealth University. It covers data on diabetes patients across 130 U.S. hospitals from 1999 to 2008. There are over 100,000 unique hospital admissions in this data set, from ~70,000 unique patients.

The goal is to test the performance of different models on predicting patient readmission. For these models, the prediction output is the readmission variable, which has 3 levels indicating whether or not a readmission occurred, and the number of days between hospital release and readmission. We are specifically interested The patients of concern are those for which this predictor has a value of "<30".

Predictive modeling in this domain has been conducted in the past, and there is evidence supporting the efficacy of these [2]. From the paper by Futoma et al, I picked two models I thought would be effective and fit the scope of the class: Elastic Net and Random Forest. These models, considered up to 28 predictors, including demographic variables, admission and discharge details, medical history, clinical results, and medication details. Both models performed well, replicating earlier research.

Model selection was done after comparing the model's performance through the area under ROC (Receiver Operating Charactistics) curve plots, which graphically illustrates the performance of a binary classifier system as its discrimination threshold is varied, and each model's MCE (Mean Classification Error). Since false positives have a higher cost, sensitivity was more valued in the decision. While the Elastic Net model had a higher MCE, it had a higher AUC and overall showed higher sensitivity, thus outperforming the Random Forest model. Our final model then is:

$$\hat{Y}_{\text{Elastic Net}} = \hat{p}_{\text{Readmission}} = \hat{p}(x_1, x_2, x_3, x_4) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3}} \tag{1}$$

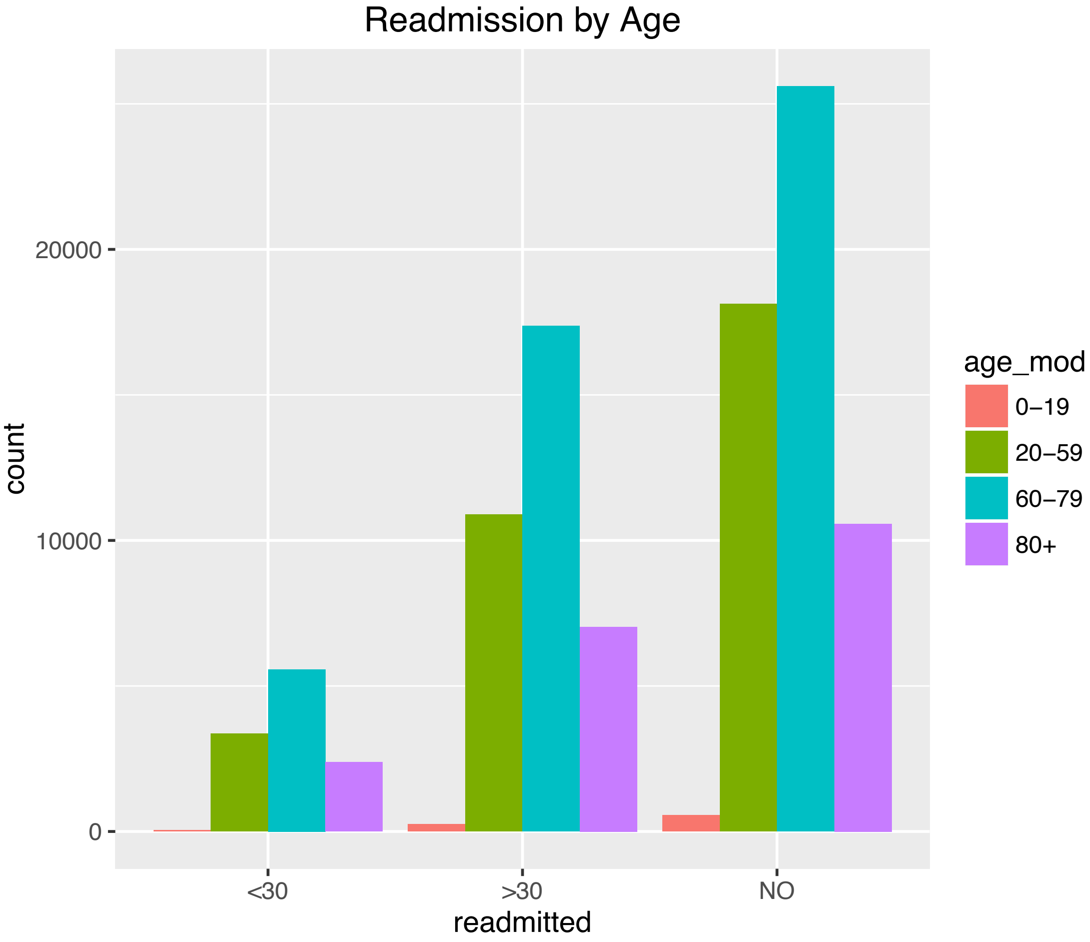
# The Data

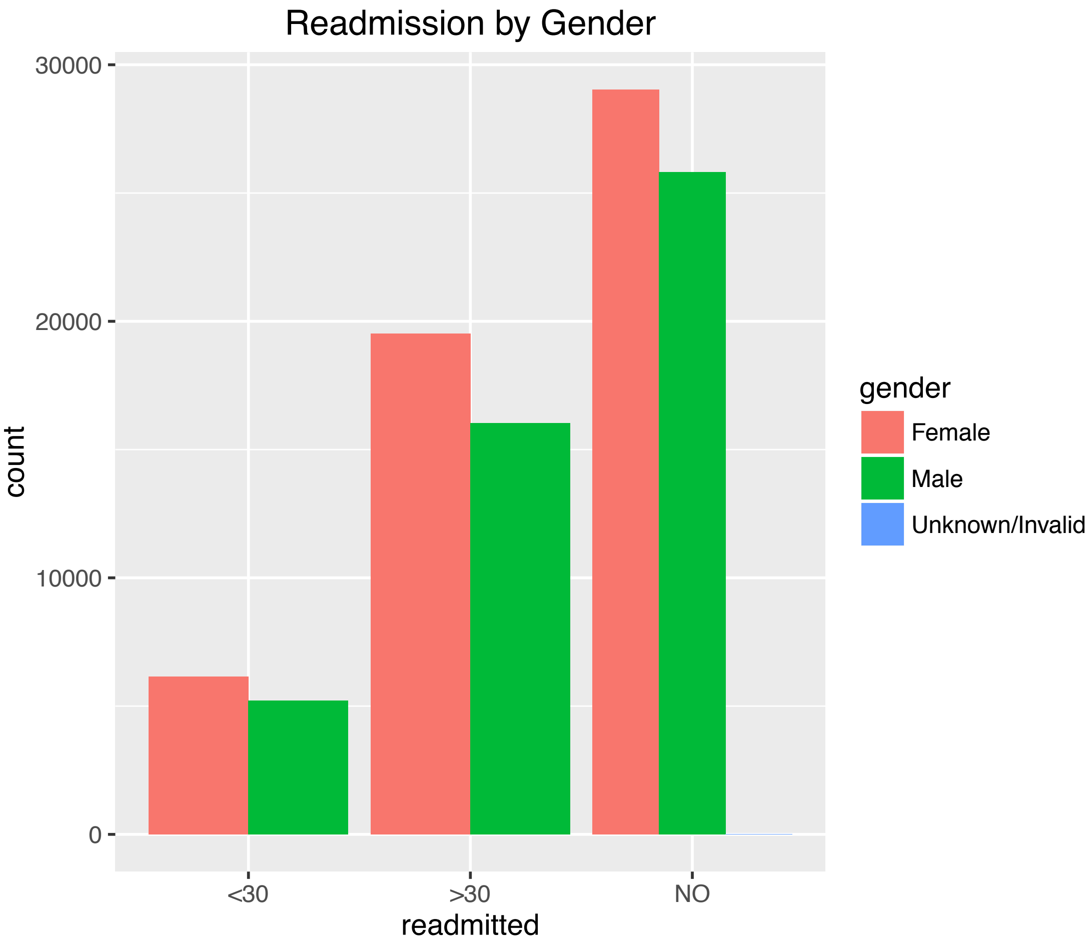
## Overview

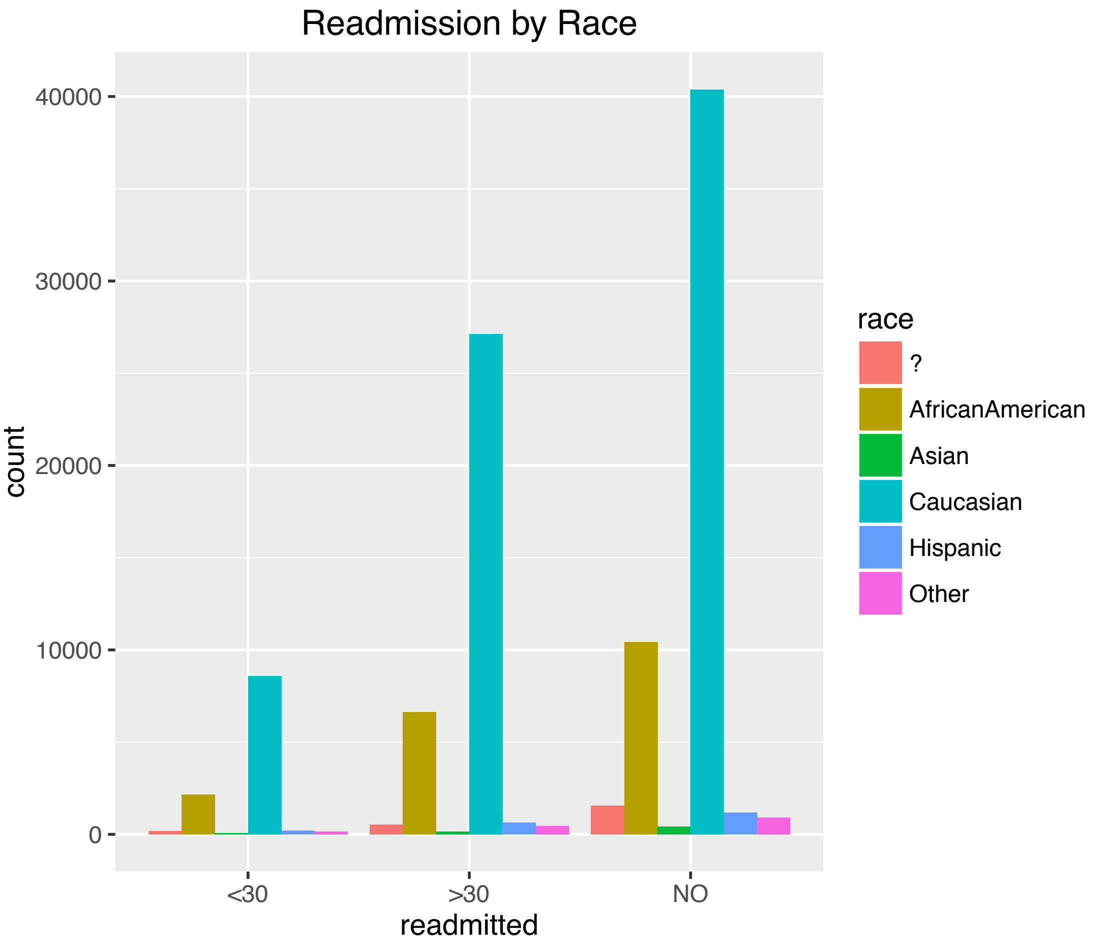
Our data, *readmission.csv*, is a cleaned, modified version of the original dataset. It contains 101,766 observations accross and 31 predictors. Among the predictors, 10 are numerical, and the remaining 21 are categorical. The majority of the patients (~88.8%) have either not been readmitted, or have been readmitted >30 days after their release. In the interest of time, I will analyze demographic variables and their relationship with our outcome of interest.

### Demographic Variables: age, gender and race

While there are no immediately discernable differences in the within-group proportions of the readmission levels by age or gender, we notice that the majority of the patients in this dataset (1) female (~53.8%) (2) in the senior age range, 60-79 (~47.7%), followed by adult age range (~31.8%). This is expected: incidence of diabetes increases with age. Nevertheless, within-age-group differences have been concealed since the data is binned - since we are dealing with censored data in the case of age, we will only be able to use age as a categorical variable in our analysis. The majority of the patients in the dataset are Caucasian, followed by African Americans, which is not surprising as this is sampled from the United States, and these are the two largest ethnicities representing 72.4% and 12.6% of the population respectively, according to the 2010 census. It is surprising to see that only ~2% of the dataset rows relate to hispanic patients, considering that Hispanic and Latino Americans represented 16.3% of the population in the same census.




Readmission by Gender
count
30000
20000
10000
0
<30
>30
NO
readmitted
gender
Female
Male
Unknown/Invalid



## Data Cleaning & Featurization

(1) Two variables, `encounter_id` and `patient_nbr`, have unique values per patient and will be excluded from the analysis, as they will not contribute to our efforts.

(2) I will also omit all observations without gender or race information

(3) I will perform take a stratified random sample from the data to minimize the runtime of the algorithms used while minimizing information loss. As I was performing the exercises, I found the runtimes to be a considerable obstacle, and I consider this to be a reasonable approach.

(4) The prediction output, readmission, comes from the `readmitted` variable. In order to be able to use it for predictive modeling, it needs to be re-featurized into two binary values:

- Readmission within <30 days will be assigned as "1"
- No readmission or readmission after 30 days will be assigned as "0"

```
### (3)

# Stratified Random Sampling
set.seed(20)

## Sort
readmission.clean.strat <- data.table(readmission.clean)
# readmission.clean.strat$age_mod %>% table readmission.clean.strat$race %>%
```

```r
# table
setkey(readmission.clean.strat, age_mod, gender)
readmission.clean.strat[, .N, , keyby = list(age_mod, gender)]
```

```
##    age_mod gender     N
## 1:    0-19 Female   481
## 2:    0-19   Male   361
## 3:   20-59 Female 16325
## 4:   20-59   Male 15345
## 5:   60-79 Female 24522
## 6:   60-79   Male 22934
## 7:     80+ Female 12247
## 8:     80+   Male  7277
```

```r
proportions <- as.vector(readmission.clean.strat[, .N, , keyby = list(age_mod,
    gender)][, 3])
proportions <- as.vector(unlist(floor(proportions/5)))

## Select sample
set.seed(22)
readmission.clean.stratdf <- data.table(strata(readmission.clean.strat, c("age_mod",
    "gender"), proportions, "srswor"))
N.srs <- readmission.clean.stratdf$ID_unit %>% unlist %>% as.vector
N.sample <- length(N.srs)
readmission.clean <- readmission.clean.strat[N.srs, ]
```

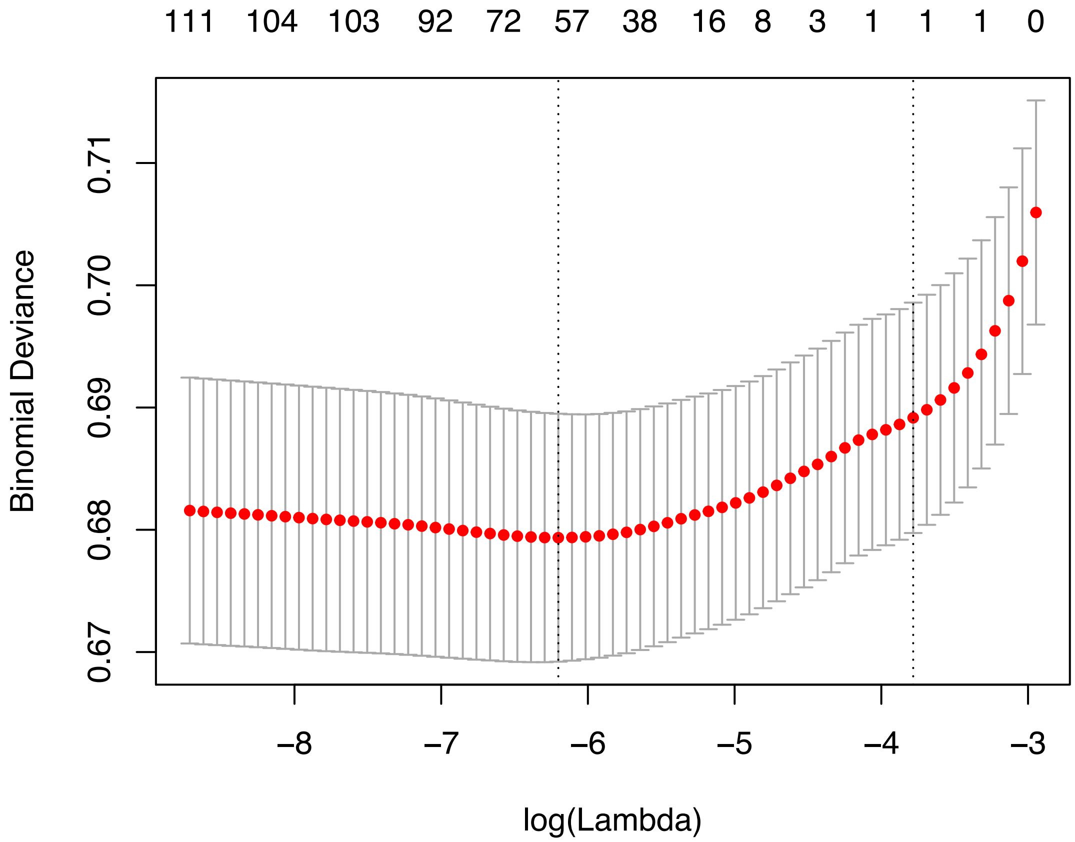
This finalizes the pre-processing stage, and we can proceed to the analysis.

### Model 1 - Elastic Net / Lasso

This model is fit using elasticnet. We set Alpha at 0.99 to obtain a simpler model, closer to Lasso, and set cross-validation folds at $ k = 10 $.

```r
X <- as.matrix(model.matrix(readmitted ~ ., readmission.clean)[, -1])
Y <- as.matrix(readmission.clean[, 29])
```

Elastic Net plot:



```
fit.elastic.1se <- glmnet(X, Y, alpha = 0.99, family = "binomial", lambda = fit.elastic.cv$lambda.1se)
fit.elastic.1se.beta <- coef(fit.elastic.1se)
beta.elastic <- fit.elastic.1se.beta[which(fit.elastic.1se.beta != 0), ]
beta.elastic <- as.matrix(beta.elastic)
rownames(beta.elastic)
```

```
## [1] "(Intercept)"      "number_inpatient"
```

### Logistic regression fit with beta.elastic output

```
fit.elastic <- glm(readmitted ~ time_in_hospital + number_inpatient + disch_disp_modified,
    readmission.clean, family = binomial)
```

### Anova analysis:

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

Response: readmitted
                    LR Chisq Df Pr(>Chisq)
time_in_hospital       13.47  1   0.000242 ***
number_inpatient      379.08  1  < 2.2e-16 ***
disch_disp_modified    76.72  3  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All the selected variables are significant at even below Alpha = 0.001, which is a good indicator.

### Model 2 - Random Forest

The stratified sample is split into training and test sets at a 4:1 proportion. The Random Forest model is then built.

```
set.seed(10)
fit.rf.train <- randomForest(readmitted ~ ., readmission.train, mtry = 4, ntree = 500)
predict.rf.y <- predict(fit.rf.train, newreadmission = readmission.test)
predict.rf <- predict(fit.rf.train, newdata = readmission.test, type = "prob")
```

## Model Evaluation

We estimate that false positives cost twice as much false negatives. Thus, we use

$$\frac{\frac{1}{2}}{1+\frac{1}{2}} \tag{2}$$

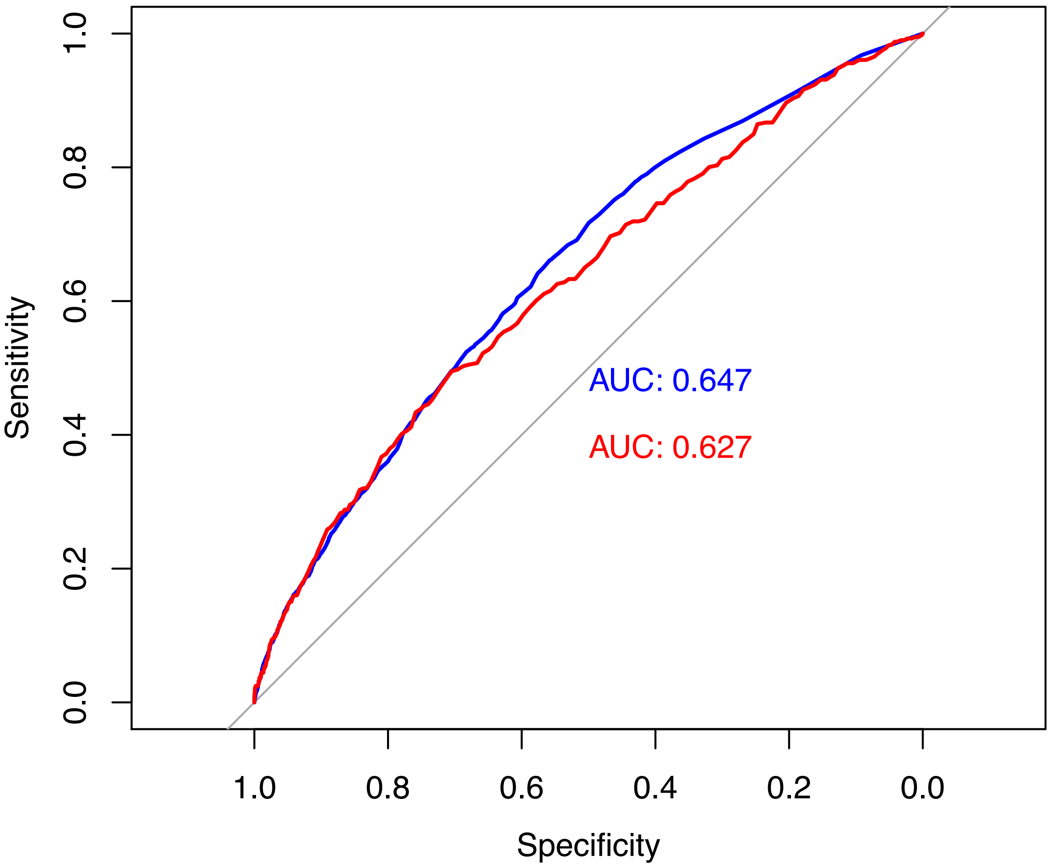
as our decision boundary, with the following classification rule for our fitted values:

$$\hat{Y} = 0 \quad \text{if} \quad \hat{p} < \frac{1}{3}; \hat{Y} = 1 \quad \text{if} \quad \hat{p} > \frac{1}{3} \tag{3}$$

# Selecting our final model and classifier

```
                             Elastic Net Random Forest
   Area Under Curve            0.6472340     0.6267742
   Mean Classification Error   0.1167571     0.1026640
```

```
finalROC <- plot(fit.elastic.roc, print.auc = TRUE, col = "blue")
finalROC <- plot(fit.randomF.roc, print.auc = TRUE, col = "red", print.auc.y = 0.4,
    add = TRUE)
```



Model selection was done after comparing the model's performance through the area under ROC (Receiver Operating Charactistics) curve plots, which graphically illustrates the performance of a binary classifier system as its discrimination threshold is varied, and each model's MCE (Mean Classification Error). Since false positives have a higher cost, sensitivity was more valued in the decision. While the Elastic Net model had a higher MCE, it had a higher AUC and overall showed higher sensitivity, thus outperforming the Random Forest model. Our final model then is:

$$\hat{Y}_{\text{Elastic Net}} = \hat{p}_{\text{Readmission}} = \hat{p}(x_1, x_2, x_3, x_4) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3}} \tag{4}$$

## Limitations & suggestions

*(1)* Computational constraints: results could change if they were run with the entire sample, thought stratified random sampling provided a good avenue for faster computation in this case.

*(2)* The binning of age data mutes the true effect of age, and having a discrete or numeric age variable would have been much more valuable.

*(3)* The low number of hispanic people might indicate that the original sample is not fully representative.

*(4)* Additional variables, such as income and employement data, or data relating to patient lifestyle habits and other medical conditions could provide valuable insights.

*(5)* Other models could be tested and replicated: Neural Nets and Support Vector Machines could be tested in future projects.

# Bibliography

[1] Sushmita, S.; Khulbe, G.; Hasan, A.; Newman, S.; Ravindra, P.; Basu Roy, S.; De Cock, M.; Teredesai, A.. Predicting 30-Day Risk and Cost of “All-Cause” Hospital Readmissions. AAAI Workshops, North America, mar. 2016.

[2] Futoma, J., Morris, J., Lucas, J.: A comparison of models for predicting early hospital readmissions. J. Biomed. Inform. 56, 229–238 (2015)