Final Project: Diabetes Prediction in the Pima Tribe

Intro:

The Native American people of the Pima tribe (known in the referenced academic literature as Pima Indians) are Native to the Arizona region of the US. The people of this heritage participated in one of the longest running diabetes studies in existence. During this study, it was found that this population has one of the highest incidents rates of diabetes in the world, at 34.2% of men and 40.8% of women having type 2 diabetes. Diabetes can be a very dangerous condition if not managed correctly, so it is imperative that we can diagnose it accurately and determine risk factors for this disease. To that end, we will train several logistic regression models predict diabetes based some predictor variables. The dataset we will use to train these models originates from the National Institute of Diabetes and Digestive and Kidney Diseases. This data has several constraints placed on it for ease of use, most notably that all observations are of females at least 21 years of age. The data set has 768 observations, 8 features and one binary response variable. The names and data types can be seen in Figure 1 below. In this dataset, 0 values represent NA except for in pregnancies and in the outcome variable. We will impute these data to streamline our model building.

```
data                    768 obs. of 9 variables
  Pregnancies : int 6 1 8 1 0 5 3 10 2 8 ..
  Glucose : int 148 85 183 89 137 116 78 11
  BloodPressure : int 72 66 64 66 40 74 50
  SkinThickness : int 35 29 0 23 35 0 32 0
  Insulin : int 0 0 0 94 168 0 88 0 543 0 .
  BMI : num 33.6 26.6 23.3 28.1 43.1 25.6 3
  DiabetesPedigreeFunction: num 0.627 0.351
  Age : int 50 31 32 21 33 30 26 29 53 54 .
  Outcome : int 1 0 1 0 1 0 1 0 1 1 ...
```

Figure 1 Snippet of Diabetes Dataset

In Figure 2, we can see the histograms of the various variables in the dataset. Glucose, blood pressure, skin thickness, and BMI seem to be normally distributed. Pregnancies, Insulin, Diabetes Pedigree Function, and Age seem to be right skewed. The outcome variable is whether a patient has diabetes (1) or not(0). It seems to be a slightly imbalanced classification problem, though this slight negative skew should not affect the analysis too much. All data values are numerical, so no dummy variables are needed to train these logistic regression models.

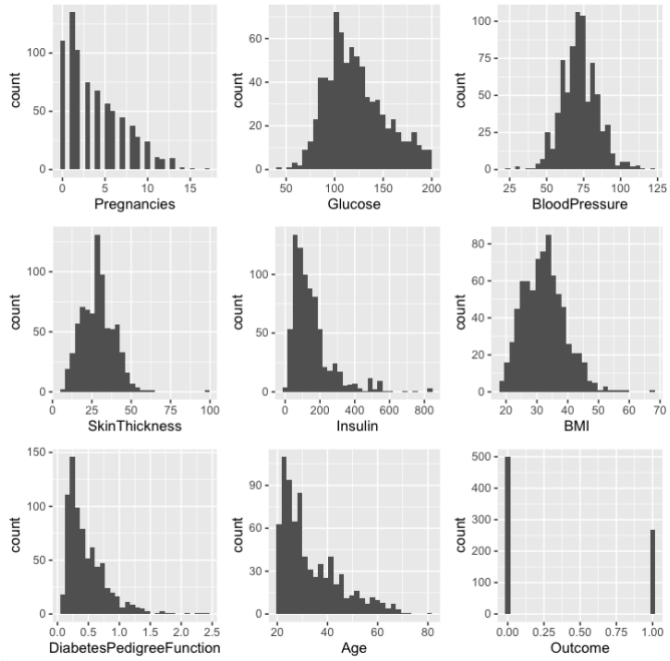

Figure 2 Histograms of Dataset

Model Building:

To accomplish the goal of the project, I built 3 logistic regression classifier models. Model 1 was a standard model using all of the predictor variables. To build model 2, I dropped the insignificant coefficients and only kept the ones with very low p-values (Pregnancies, Glucose, BMI, and DiabetesPedigreeFunction). To build model 3, I did a log transformation on the Insulin and DiabetesPedigreeFunction to correct the skewness. Originally, I attempted to build a lasso regression model, though I was getting the same results as dropping the insignificant coefficients so I abandoned that route. I also tried to log transform the Pregnancies and Age variables in model 3, but this was not correcting the skewness so I opted to leave them as is.

Conclusion:

When analyzing the models, it is clear that the DiabetesPedigreeFunction variable has the strongest predictive power amongst all of the variables. Each model performed similarly well, though Model 3 performed the weakest of the bunch. Model 2 seemed to perform the best with both the highest AUC: 85.6% and lowest Misclassification Rate: 23.04%. This model is both the best performing and most parsimonious off all of the models. If we had to choose, Model 2 would be the ideal candidate for predicting diabetes in this subset of the population of the Pima Tribe. Overall, these models performed reasonably well, though if we wanted to put a model into production, it would be prudent to try other classifiers as well, such as a Random Forest Model.

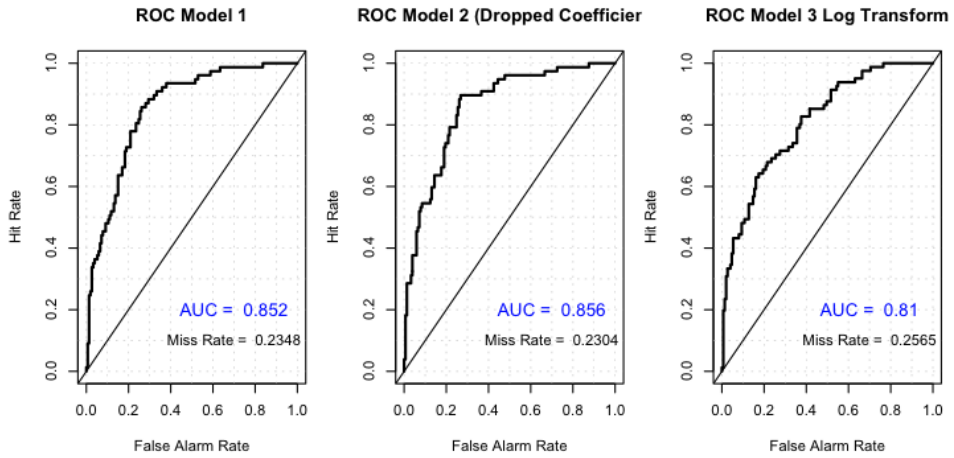

Figure 3 ROC Plots

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -10.07657    1.04482   -9.64  < 2e-16 ***
Pregnancies               0.10835    0.03918    2.77   0.0057 **
Glucose                   0.04314    0.00531    8.13  4.4e-16 ***
BloodPressure            -0.00819    0.01105   -0.74   0.4584
SkinThickness            -0.00148    0.01465   -0.10   0.9197
Insulin                  -0.00192    0.00120   -1.60   0.1093
BMI                       0.10721    0.02416    4.44  9.1e-06 ***
DiabetesPedigreeFunction  1.00753    0.37832    2.66   0.0077 **
Age                       0.01421    0.01139    1.25   0.2124
---
```

Model 1

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -9.64763    0.86229  -11.19  < 2e-16 ***
Pregnancies               0.13154    0.03314    3.97  7.2e-05 ***
Glucose                   0.03974    0.00443    8.97  < 2e-16 ***
BMI                       0.09136    0.01780    5.13  2.9e-07 ***
DiabetesPedigreeFunction  0.96105    0.37681    2.55    0.011 *
```

Model 2

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -7.92427    1.11465   -7.11  1.2e-12 ***
Pregnancies               0.10588    0.03926    2.70   0.0070 **
Glucose                   0.03904    0.00525    7.44  1.0e-13 ***
BloodPressure            -0.01003    0.00998   -1.00   0.3151
SkinThickness            -0.00410    0.01381   -0.30   0.7667
Insulin                  -0.15811    0.20764   -0.76   0.4464
BMI                       0.10073    0.02259    4.46  8.2e-06 ***
DiabetesPedigreeFunction  0.47095    0.18100    2.60   0.0093 **
Age                       0.02031    0.01124    1.81   0.0707 .
```

Model 3

References

Learning, UCI Machine. “Pima Indians Diabetes Database.” *Kaggle*, 6 Oct. 2016, https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/discussion/230264.

Schulz, Leslie O, and Lisa S Chaudhari. “High-Risk Populations: The Pimas of Arizona and Mexico.” *Current Obesity Reports*, U.S. National Library of Medicine, Mar. 2015, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4418458/.