# DRAFT: A model to predict chances of matching into Obstetrics and Gynecology Residency

*Tyler M. Muffly, MD*

*07 June, 2019*

## Contents

Objective: We sought to construct and validate a model that predict a medical student's chances of matching into an obstetrics and gynecology residency.

# Introduction

The data set of years 2015, 2016, 2017, and 2018 at University of Colorado. Applicants who were in the SOAP and applied to the preliminary spot were determined to be unmatched. The data is contained in a data frame called all_data.

## Install and Load packages. See Session information at the end.

## Create a dataframe of independent and dependent variables. Download cleaned data from Dropbox.

#Potential code with Drake. I'm not sure if using drake is necessary or if it is overkill.

### Data Quality Check of `all_data`

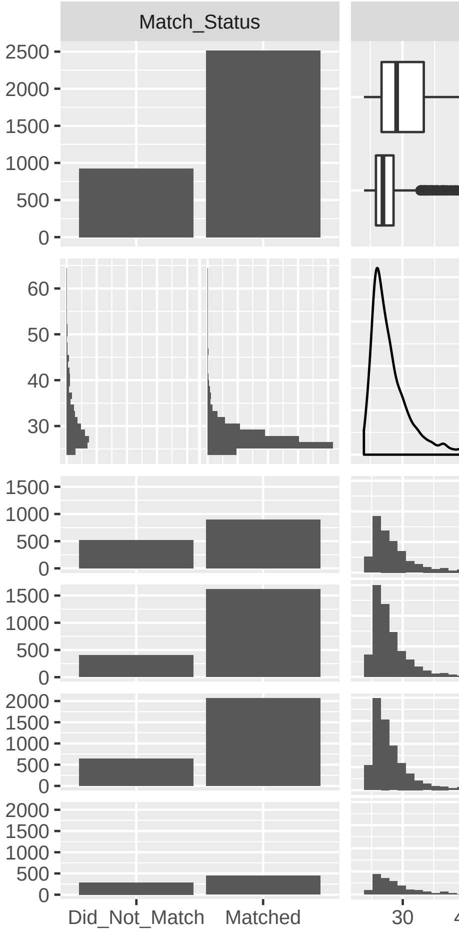
A summary of the 19 variables are listed below:

1. Eleven of the variables were a factor. All factors had two levels except for Alpha_Omega_Alpha had three levels. The target variable is `all_data$Match_Status`.
2. Seven of the variables were integers.
3. One of the variables was a number. Age was calculated as a number.

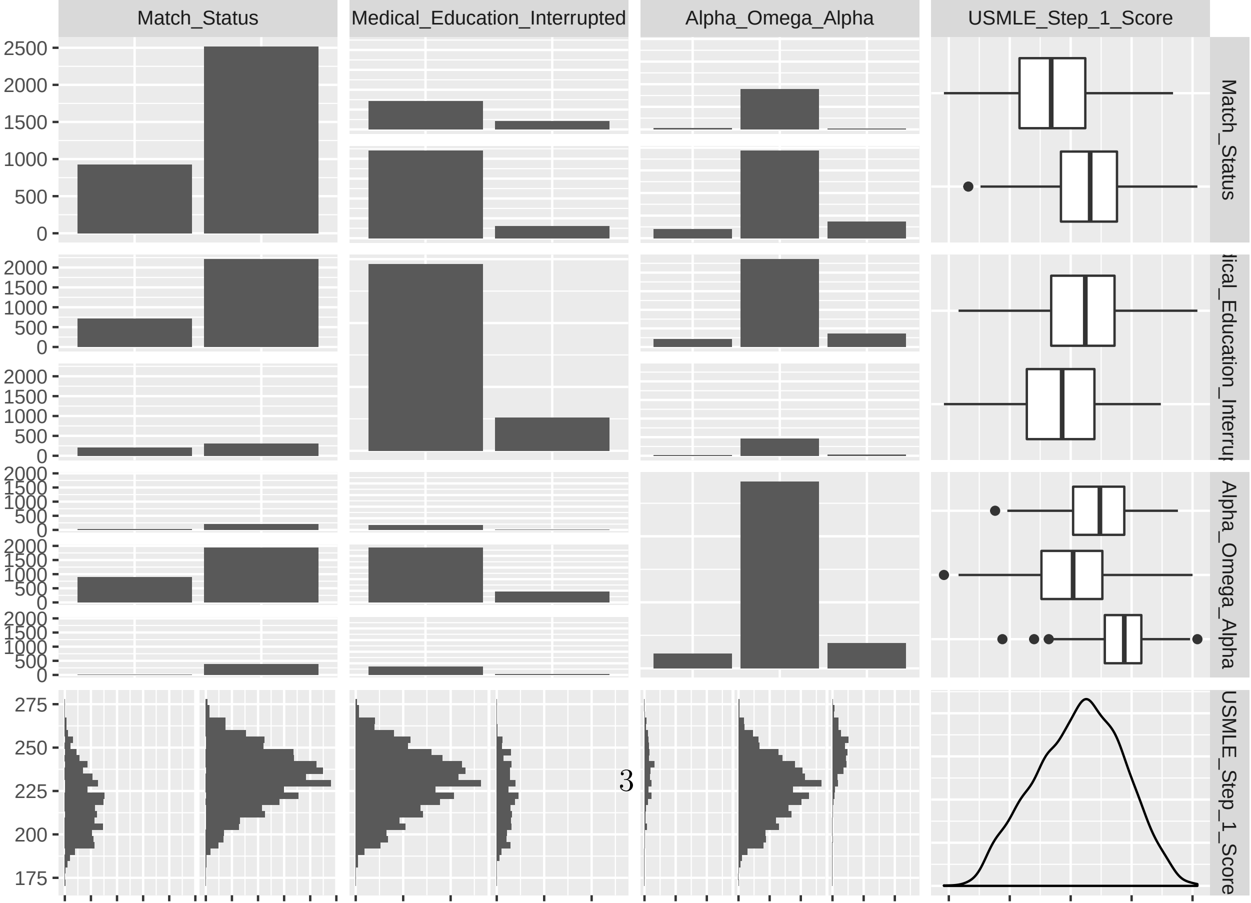
#cleanup.import is not working.

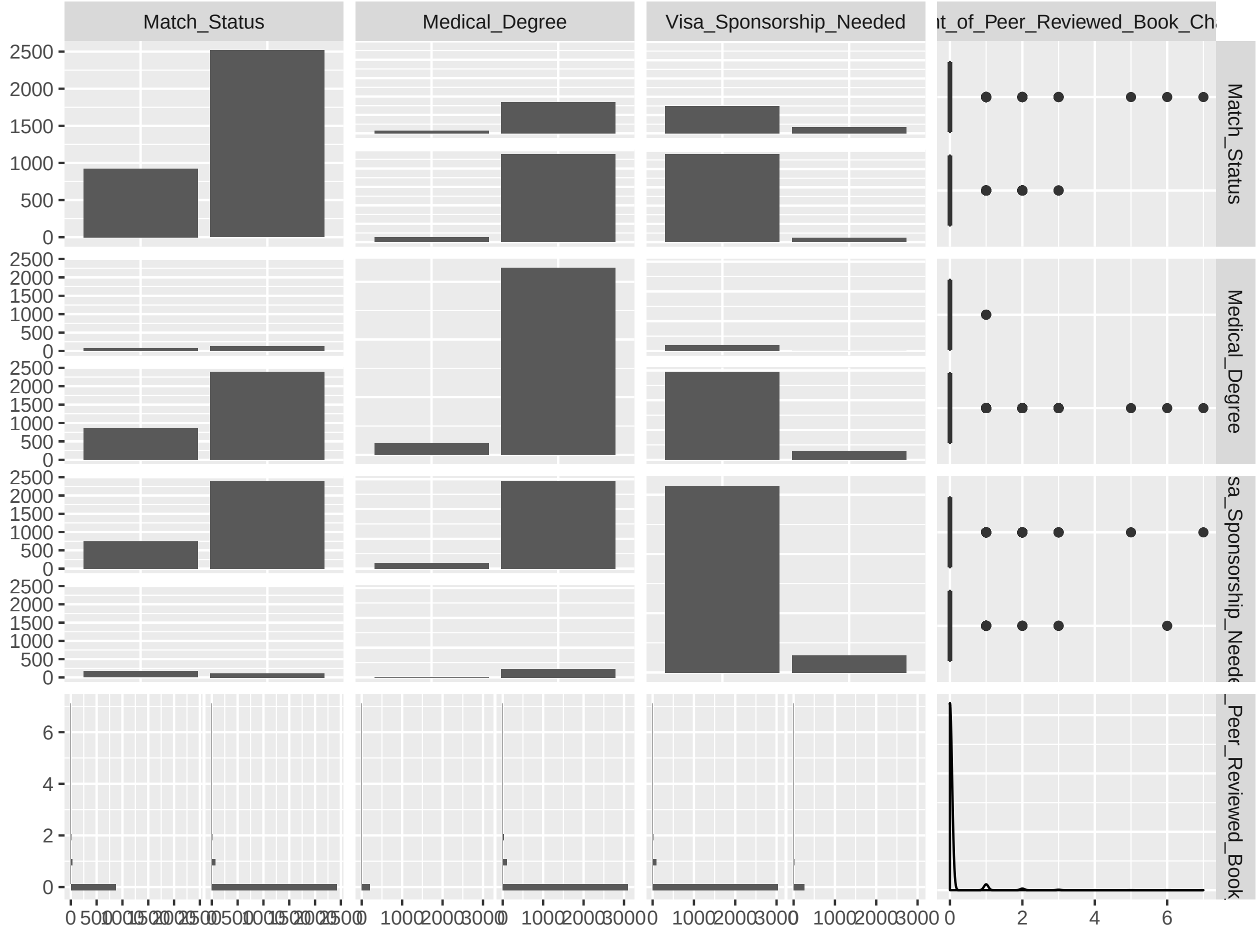
Data size and structure

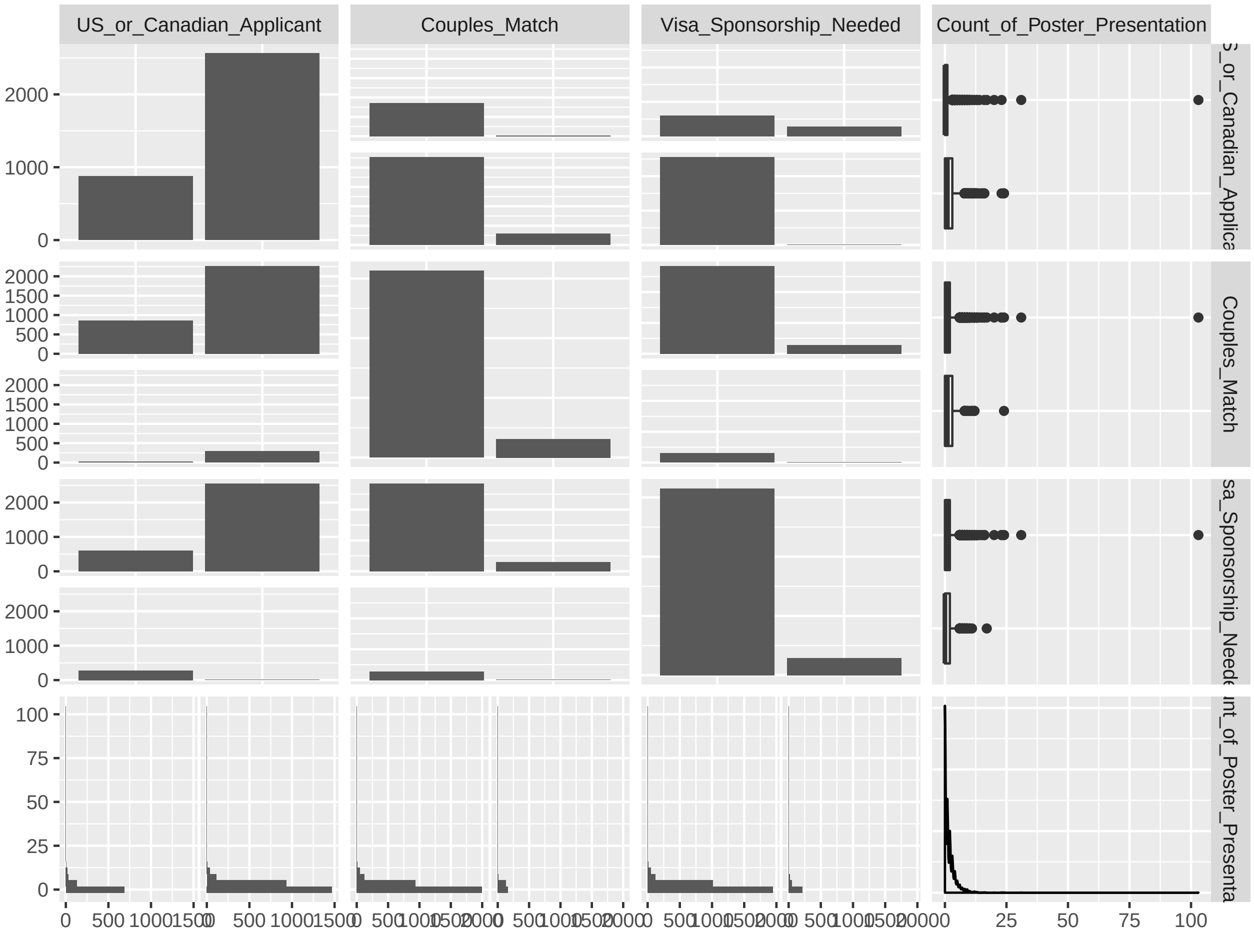
## Univariate Analysis



IS THERE SOMETHING BETTER THAN THESE GGPAIRS PLOTS?




Match_Status
Medical_Degree
Visa_Sponsorship_Needed
nt_of_Peer_Reviewed_Book_Ch
Match_Status
Medical_Degree
sa_Sponsorship_Neede
_Peer_Reviewed_Book
2500
2000
1500
1000
500
0
6
4
2
0
0 500 1000 1500 2000 2500
1000 2000 3000
0
2
4
6



Univariate analysis of the data.

IS THERE A WAY TO SET THE CUTPOINTS BASED ON A STATISTICAL REASONING?

I set the cutpoints based on nothing really. I like this better than the base summary command. See the Appendix at the end of the document for univariate distributions.

## Relaxed Cubic Splines For Continuous Variables

HOW DO YOU DECIDE HOW MANY KNOTS TO USE ON CONTINUOUS VARIABLES WITH RELAXED SPLINES?

Set the Match_Status variable to be a number and a factor.

## Table 1

Table 1: Applicant Descriptive Variables by Matching Success (1) or Failure (0)

ARSENAL VS. STARGAZER TO CREATE A TABLE 1 OF DESCRIPTIVE STATISTICS?

Descriptive summaries of all variables in the dataset are provided in the table.

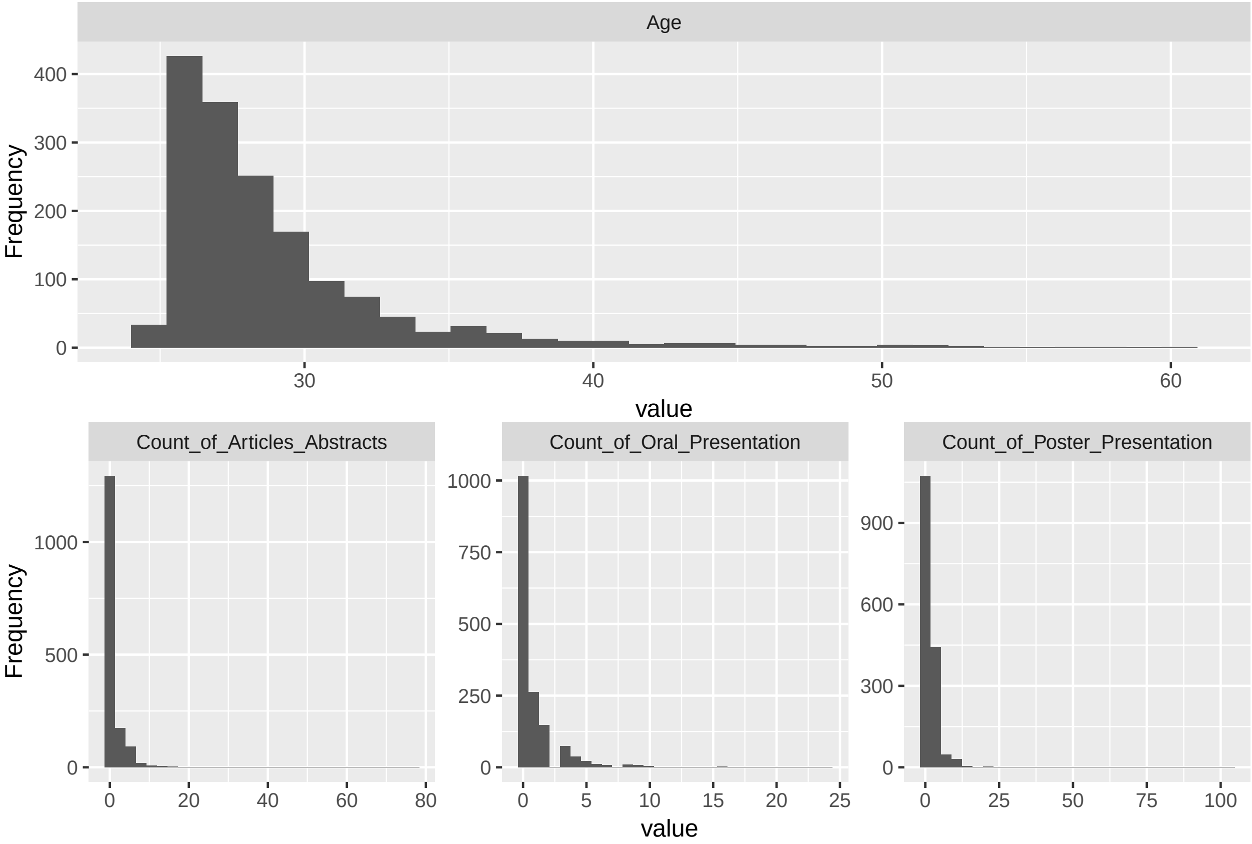
\begin{table}[!htbp] \caption{Descriptive Statistics of Match_Status Data}

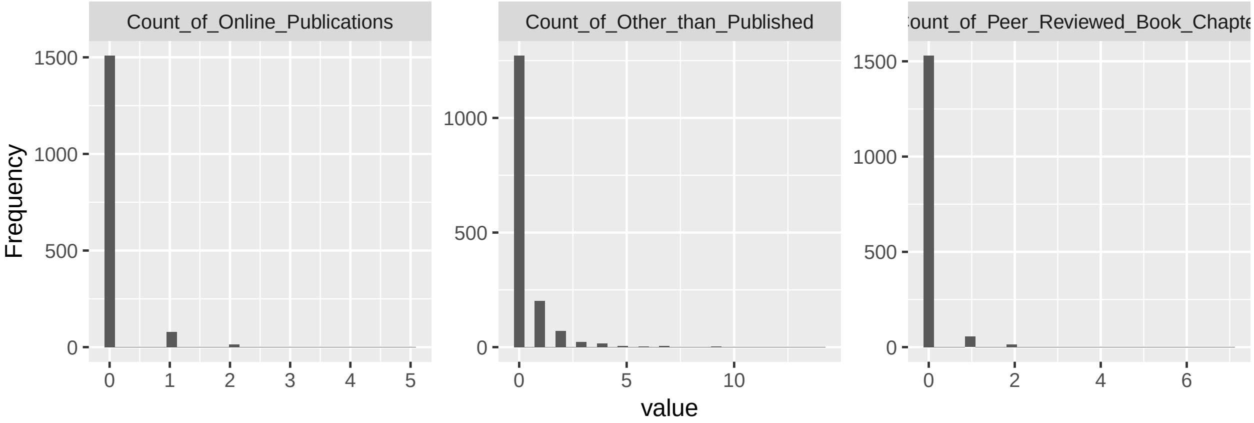
| Statistic | N | Mean | St. Dev. | Min | Pctl(25) | Median | Pctl(75) | Max |
|---|---|---|---|---|---|---|---|---|

\end{table}

## 2) Exploratory data analysis

After the data check was completed, an exploratory data analysis (EDA) was conducted to look for interesting relationships among the variables. Histograms were used to visualize distributions among predictors. Since the assignment was a classification problem, relationships between predictors and the dichotomous outcome were also performed. Distributions of all variables were skewed right. Examples of histograms of seven variables: age, Count of articles and abstracts, count of oral presentations, count of poster presentations, Count of online pubications, count of non-published publications, count of peer-reviewed book chapters, are demonstrated below.





# Normality testing in R

The D'Agostino tests for skewness and the Anscombe tests for kurtosis with numeric variables. There is kurtosis for the Step 1 score data. Therefore only use medians in table 1 and I will stick with non-parametric tests throughout. There is skew in age.

Quantile-Quantile plot is a way to visualize the deviation from a specific probability distribution. After analyzing these plots, it is often beneficial to apply mathematical transformation (such as log) for models like linear regression. DO WE NEED TO TRANSFORM THESE VARIABLES IN ANY WAY? SHOULD WE USE Kolmogorov-Smirnov (K-S) normality test OR Shapiro-Wilk's test???

The null hypothesis of these tests is that "sample distribution is normal". If the test is significant, the distribution is non-normal. From the output, the p-value less than 0.05 implying that the distribution of the data are significantly different from normal distribution. In other words, we can not assume the normality.

This allowed for measuring the associations between continuous predictors using a matrix with correlation coefficients. The scatterplot matrix of a sample of predictors below demonstrated some associations. CAN WE INCLUDE ONLY CONTINUOUS VARIABLES? SHOULD WE INCLUDE THE TARGET VARIABLE: MATCH_STATUS?

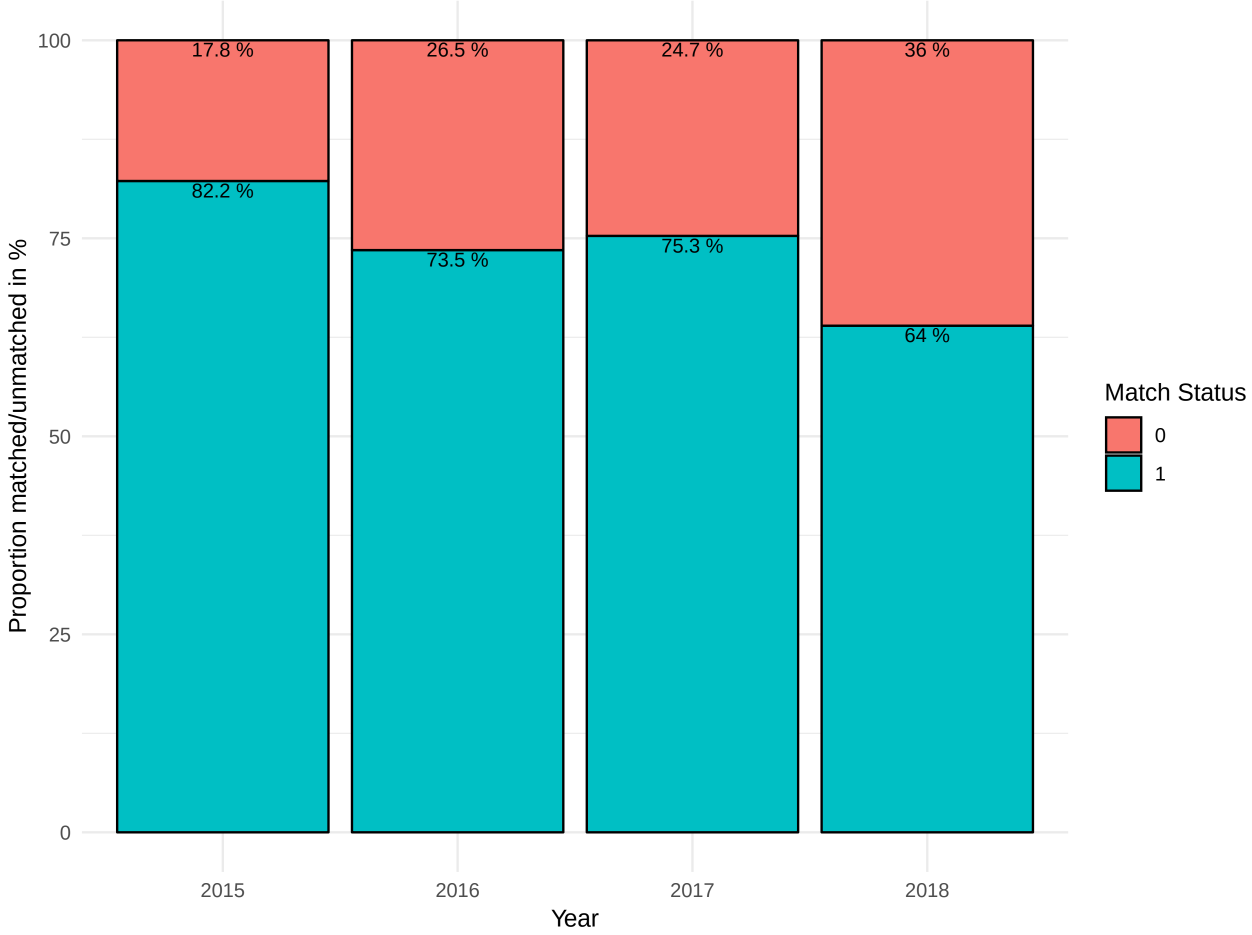
## Correlation overview

In this kind of plot we want to look for the bright, large circles which immediately show the strong correlations (size and shading depends on the absolute values of the coefficients; color depends on direction). This shows whether two features are connected so that one changes with a predictable trend if you change the other. The closer this coefficient is to zero the weaker is the correlation. Anything that you would have to squint to see is usually not worth seeing! CAN ONLY CORRELATE CONTINOUS VARIABLES? SHOULD I RUN CORRELATION ON THE TRAIN DATA OR ON THE ALL_DATA SET?

p-value associated with the null hypothesis of 0 correlation, small values indicate evidence that the true correlation is not equal to 0.

I'M NOT SURE WHY THE PROPORTION OF MATCHED VS UNMATCHED DATA IS DIFFERENT BETWEEN THE TRAIN VS TEST DATA SETS? I THOUGHT ABOUT

REMAKING THE TRAINING SET TO BE 2015 AND 2016 WITH THE TEST SET OF 2017 AND 2018 IF WE CAN'T FIGURE IT OUT. proportion vs year-1.bb



Check Proportions of Matched

##Compare the datasets of train and test

IS IT NECESSARY TO CREATE A MODEL WITH ALL THE VARIABLES FIRST? SPLINES QUESTION FROM ABOVE.

Create a Kitchen Sink model with all factors first. This is essentially a screening model with all variables. I relaxed the cubic splines with the guidance from above.

Are there predictor interactions????? HOW WOULD I RECOGNIZED THE INTERACTIONS AND DISPLAY THE RESULTS IN A READABLE WAY?

# HOW DO I CHECK FOR COLLINEARITY IN THE VARIABLES? THE VARIABLES WITH SPLINES HAVE A VERY HIGH VIF. WHY IS THIS AND SHOULD I KEEP THEM IN?

## Evaluating the signficance of kitchen.sink variables

According to the ANOVA: USMLE_Step_1_Score, Age, and US_or_Canadian_Applicants are predictors. The anova() function for the model object allows to see the null and residuals deviances. The difference between these two deviances shows how well the model is performing against the null deviance. The residuals deviance column allows to see the drop of deviance value by additional respective predictor term added.

#Factor Selection

I LIKED LASSO BECAUSE CHOOSING SOME PEOPLE MAY FIND THE PREDICTORS OR AGE, RACE, GENDER DISCRIMINATORY. I NEED SOME STATISTICAL TEST TO STAND ON ABOUT WHY I CHOSE THESE VARIABLES.

#Factor Selection using a LASSO model (Penalized Logistic Regression)

Here, we use Lasso for simplicity and interpretability. The aim is to avoid over-parametrization and unnecessary model bias by carrying feature selection on-the-go. Key to this task will be cross-validation. Start by creating a custom train control providing the number of cross-validations and setting the classProbs to TRUE for logistic regression.

WOULD YOU DO MORE REPEATS IN MYCONTROL?

Create the LASSO using glmnet within the caret package. Here we are solely using the train dataset to determine what varaiables predict the outcome.

IS THERE A COEFFICIENT CUTOFF THAT YOU WOULD USE TO CHOSE THE FINAL PREDICTORS OR NO?

Plot the results of the lasso.mod so we can see if this is more ridge or more lasso. 0 = ridge regression and 1 = LASSO regression, here ridge is better

DO WE NEED TO SAVE THE LASSO MODEL FOR ANY REASON IN THE FUTURE?

## Plot LASSO factors

Plot the individual variables by lambda. Saves the lasso.mod to an RDS file for later use.

Makes predictions of matching based on the lasso.mod using the training data.

GLMNet to do factor selection with the previously made LASSO model And we use the glmnet library to determine the optimal penalization parameter. Note that this must be assigned through cross validation; here, we use 50-fold cross validation (only suitable in small datasets). GLMnet accepts data in a matrix format so the data format was changed before giving it to glmnet.cv.

```
`%ni%`<-Negate(`%in%`)
# save the outcome for the glmnet model, could use dummyVars with fullRan=FALSE can remove col
```

```r
x <- model.matrix(train$Match_Status~., data=train)

class(x)
```

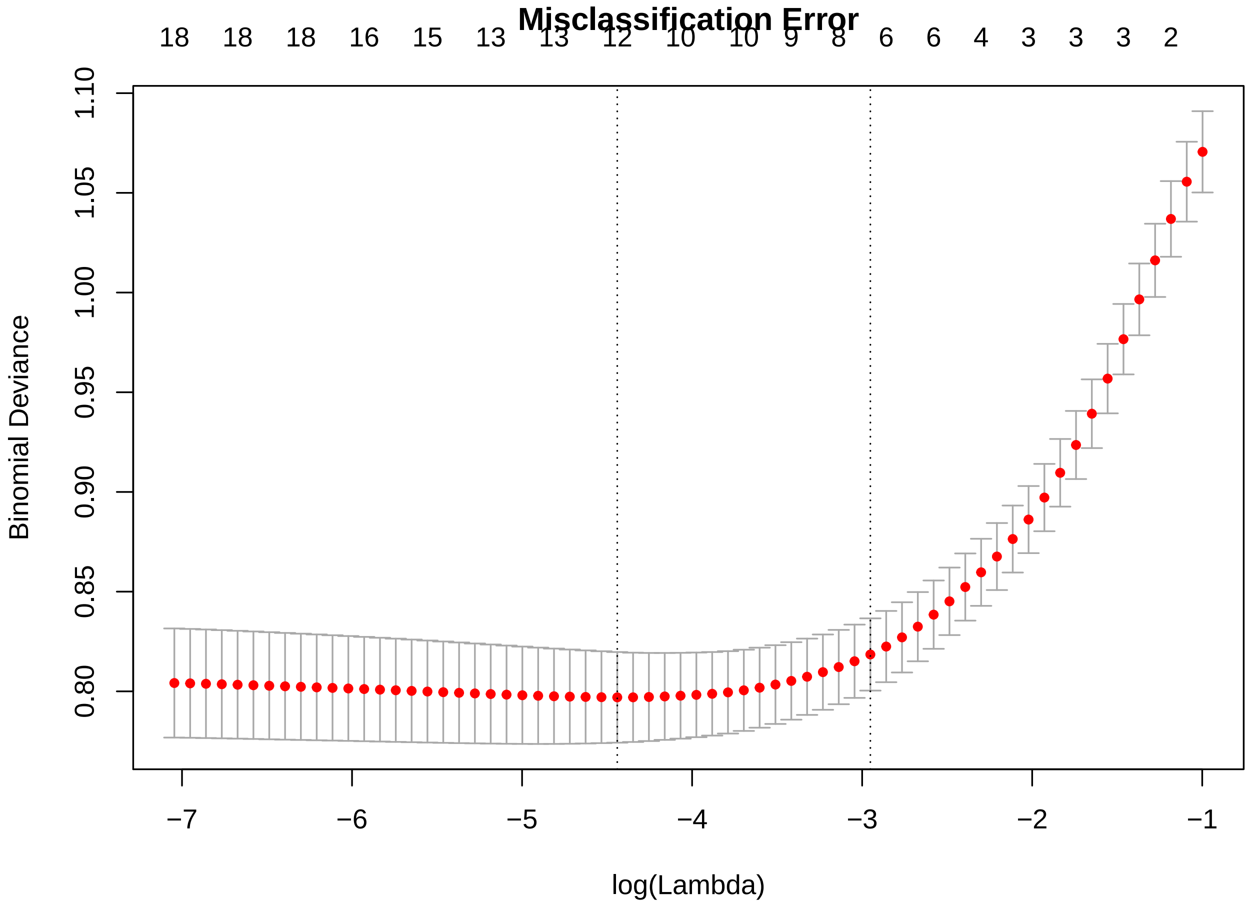
```
[1] "matrix"
```

```r
x <- x[,-1]  #Removes intercept

set.seed(356)

glmnet1 <-
  cv.glmnet(x=x,y=train$Match_Status,nfolds=10,alpha=.5, family="binomial")

plot(glmnet1,main = "Misclassification Error")
```



The left vertical line represents the minimum error, and the right vertical line represents the cross-validated error within 1 standard error of the minimum. LASSO, least absolute shrinkage and selection operator

If you look at this graph we ran the model with a range of values for lambda and saw which returned the lowest cross-validated error. You'll see that our cross-validated error remains consistent until we hit the dotted lines, where we start to see our model perform very poorly due to underfitting with misclassification error. Cross validation is an essential step in studies to help up us not only calibrate the parameters of our model but estimate the prediction accuracy with unseen data.

Variable selection using LASSO in the train dataset

I DON'T REALLY UNDERSTAND WHAT THIS GRAPH MEANS?

## Measuring Strength and Direction of Predictors

I plotted everything on a bar graph so we can easily compare the strongest predictors and the direction they affect the model:

Revise Model with selected factors

Creating a more parsiomonious model using the variables selected by LASSO in the train dataset. I made the model with lrm so I could fit it to a rms::nomogram function.

I THINK IT WAS WRONG TO DROP SOME OF THESE VARIABLES THAT HAD RELAXED CUBIC SPLINES:

## Remove variables with multicollinearity and rebuild model

AUC of this pared down model is 0.8261587.

## Odds ratios of the `train` dataset

#Table 2 of odds ratios in graph form in the train dataset.

Odds ratios for train data

#https://rstudio-pubs-static.s3.amazonaws.com/283447_fd922429e1f0415c89b93b6da6dc1ccc.html

Annotation for Manuscript Table 2: A: Nonlinear component A of the function describing the variable and the probability of matching into OBGYN. B: Nonlinear component B of the function describing the variable and the probability of matching into OBGYN. C: Nonlinear component C of the function describing the variable and the probability of matching into OBGYN.

# Use Model to predict match for Test Data

Shift Gears: Test Accuracy of Model on Training Data, Use glmnet model on 2018 TEST data Here the code is creating a vector called predictorsNames so that we can reuse the model by changing the variables in predictorsNames in the future prn. Run the 2018 data through the train model.

First, we need to fit lrm.with.lasso.variables in GLM, rather than rms to get the AUC. There is probably a better way to do this. Using the test data set. Also built the same model in lrm.

The Receiver Operating Characteristic (ROC) curve is plotted below for false positive rate (FPR) in the x-axis vs. the true positive rate (TPR) in the y-axis. It shows the detection of true positive while avoiding the false positive. This is the same as measuring the unspecificity (1 - specificity) in x-axis, against the sensitivity in y-axis. This ROC curve in particular shows that its very closed to the perfect classifier meaning that its better at identifying the positive values.

## Use Model to predict match Status for Test Data

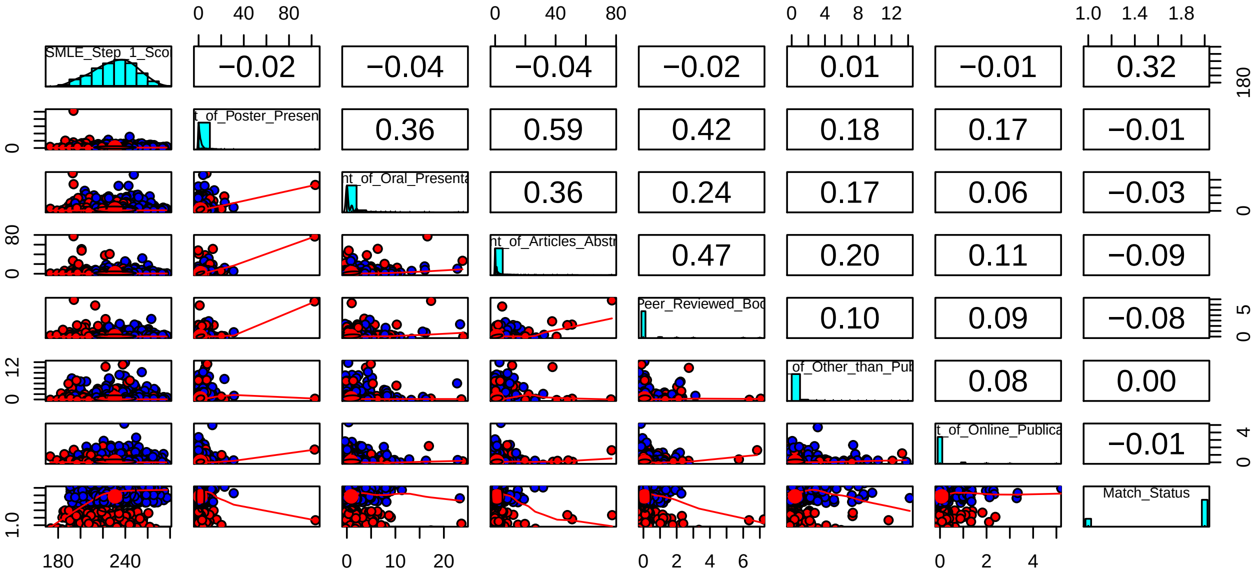
#ROC: Type 1 using ggplot with nice controls

ROC Curve type 2 with nice labels on the x and y

ROC Curve Type 3 with nice diagnal line but half of the formula printed

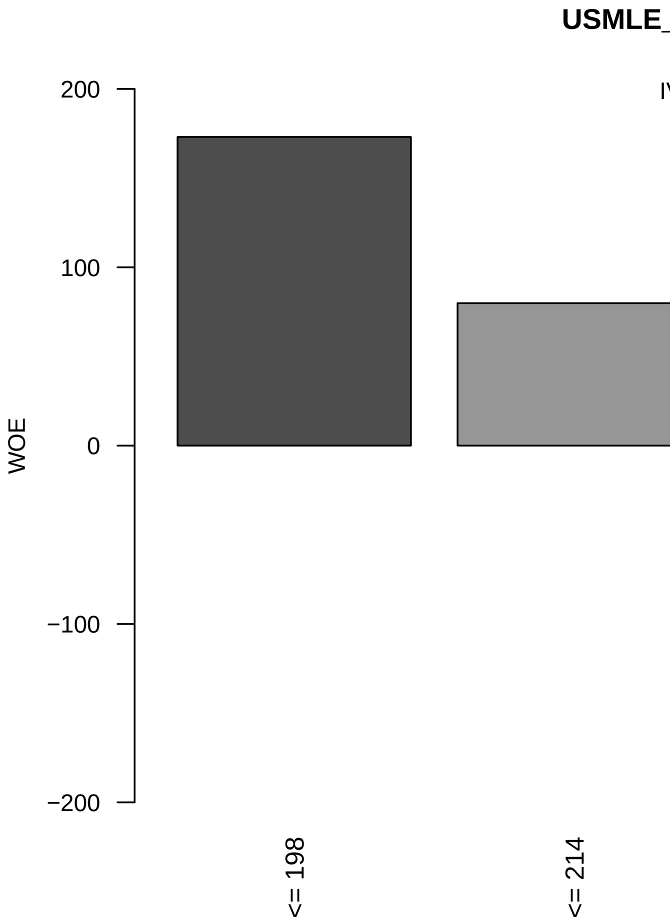
ROC Curve Type 4, ROC in color

I DO NOT UNDERSTAND WOE BINNING AT ALL.



Since the predictors were highly skewed, binning was also explored. This facilitated visualizing associations between binned variables and the outcome using contingency plots. Supervised Weight of Evidence (WOE) binning of numeric variables were explored using the woeBinning package. Fine and coarse classing that merged granular classes and levels step by step was performed. Bins were merged and respectively split based on similar weight of evidence (WOE) values and stop via an information value (IV) based criteria. The figure below demonstrated the top five predictors ranked by information value during binning.

```
   0    1
 364 1240
```



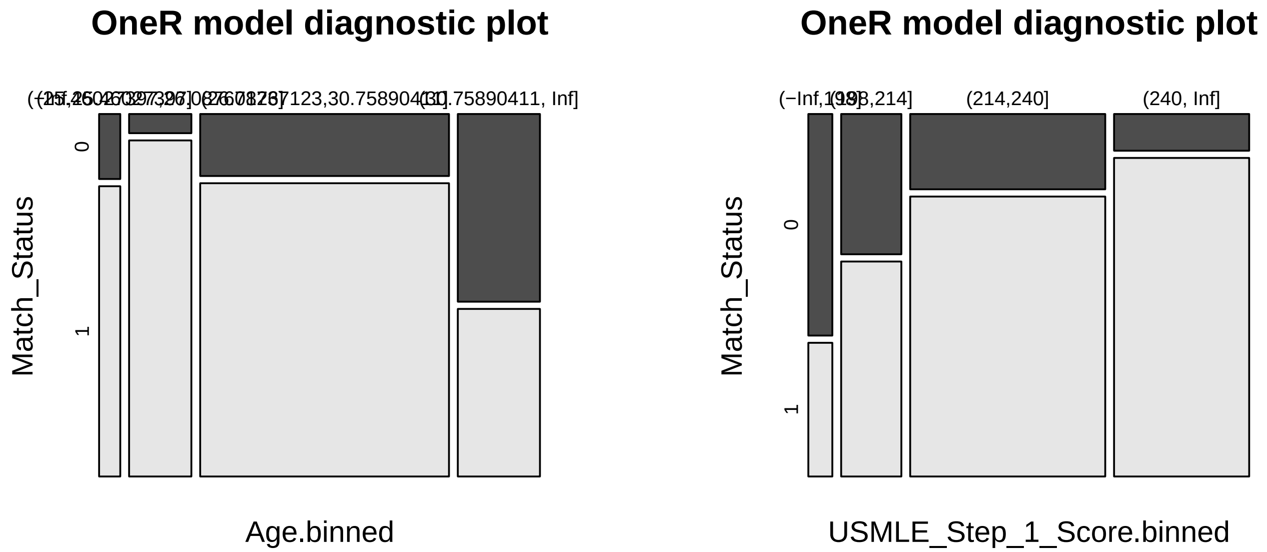
| [](Tyler-2_files/figure-latex/ | unnamed-chunk-49-1.pdf) |
|---|---|
| Age | 0.7613197 |
| USMLE_Step_1_Score | 0.5566617 |
| Count_of_Poster_Presentation | 0.0608892 |
| Count_of_Articles_Abstracts | 0.01508961 |
| Count_of_Oral_Presentation | 0.01043028 |

These top five binned variables were used for the training and test set.

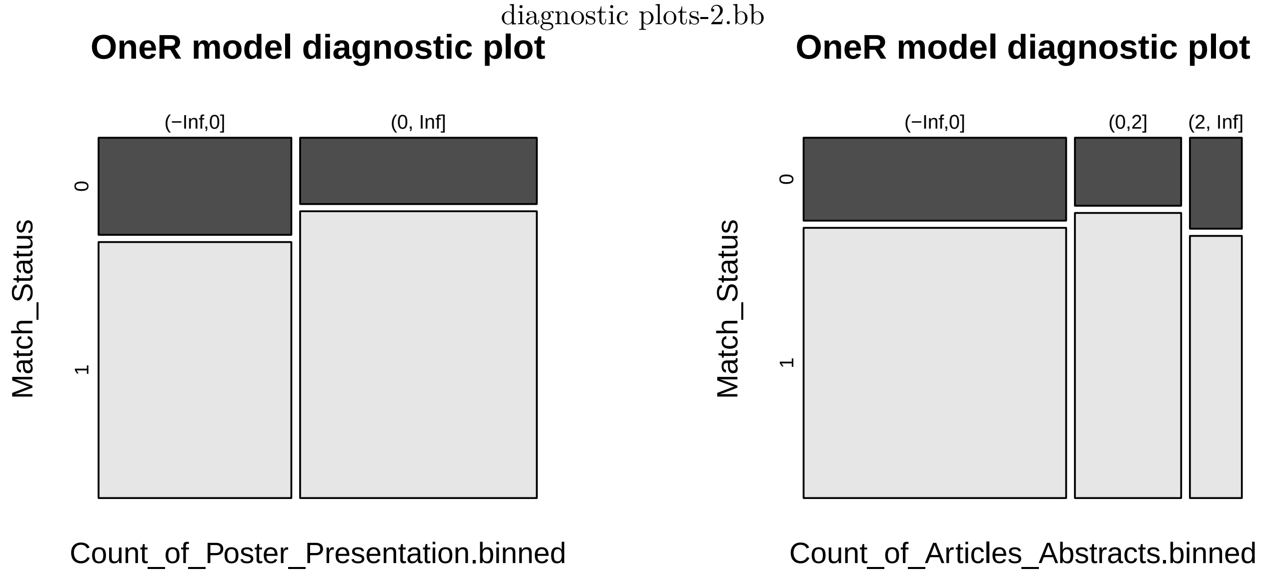
```
  0    1
560 1277
```

Relationships between binned variables and `Match_Status` were explored using mosaic plots to look for interesting bins that aided in discrimination. An example of several binned variables are shown in the plots below.

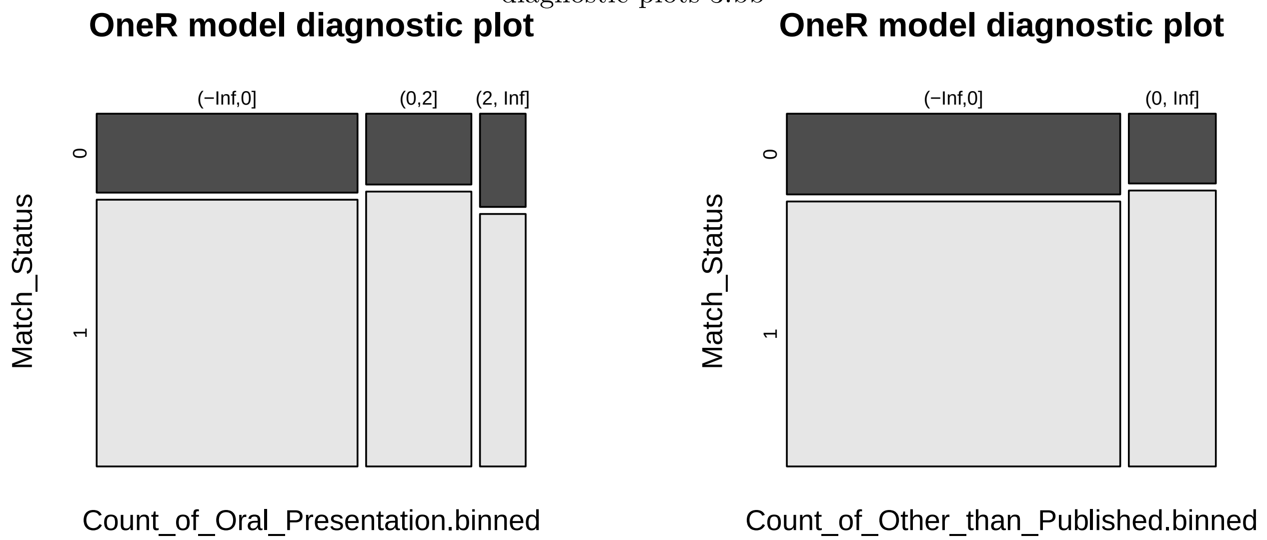
diagnostic plots-1.bb



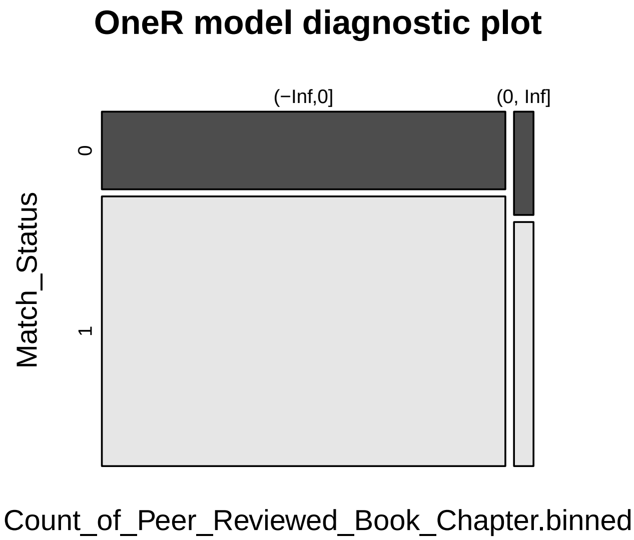
diagnostic plots-2.bb

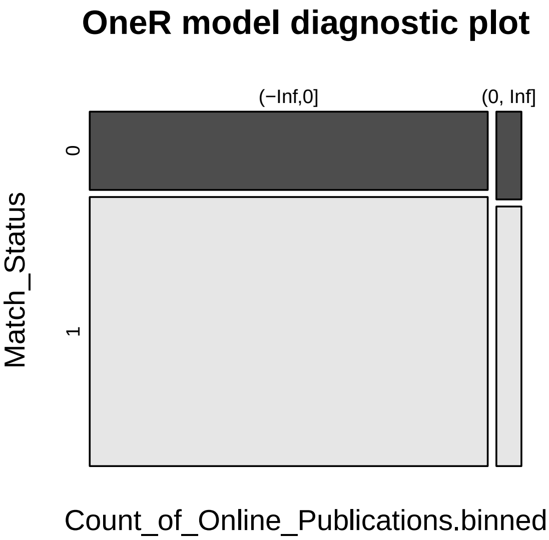


diagnostic plots-3.bb



diagnostic plots-4.bb



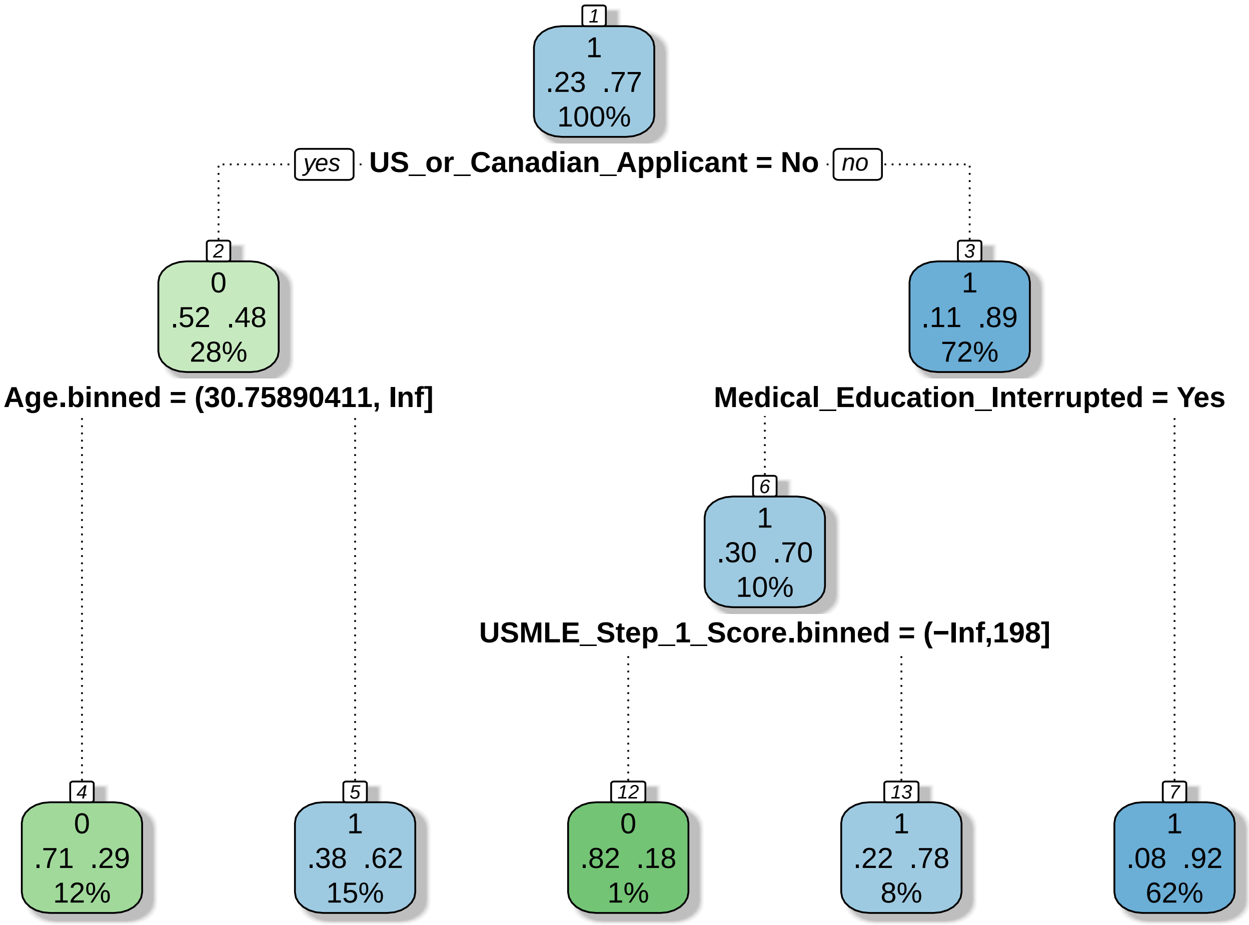


A simple decision tree model was used for exploration. The variable importance summary from the simple tree was used to explore important relationships. The variables a, b, c, and d were the top four variables in importance.

CAN WE USE ONLY FACTORS IN THIS TREE MODEL?

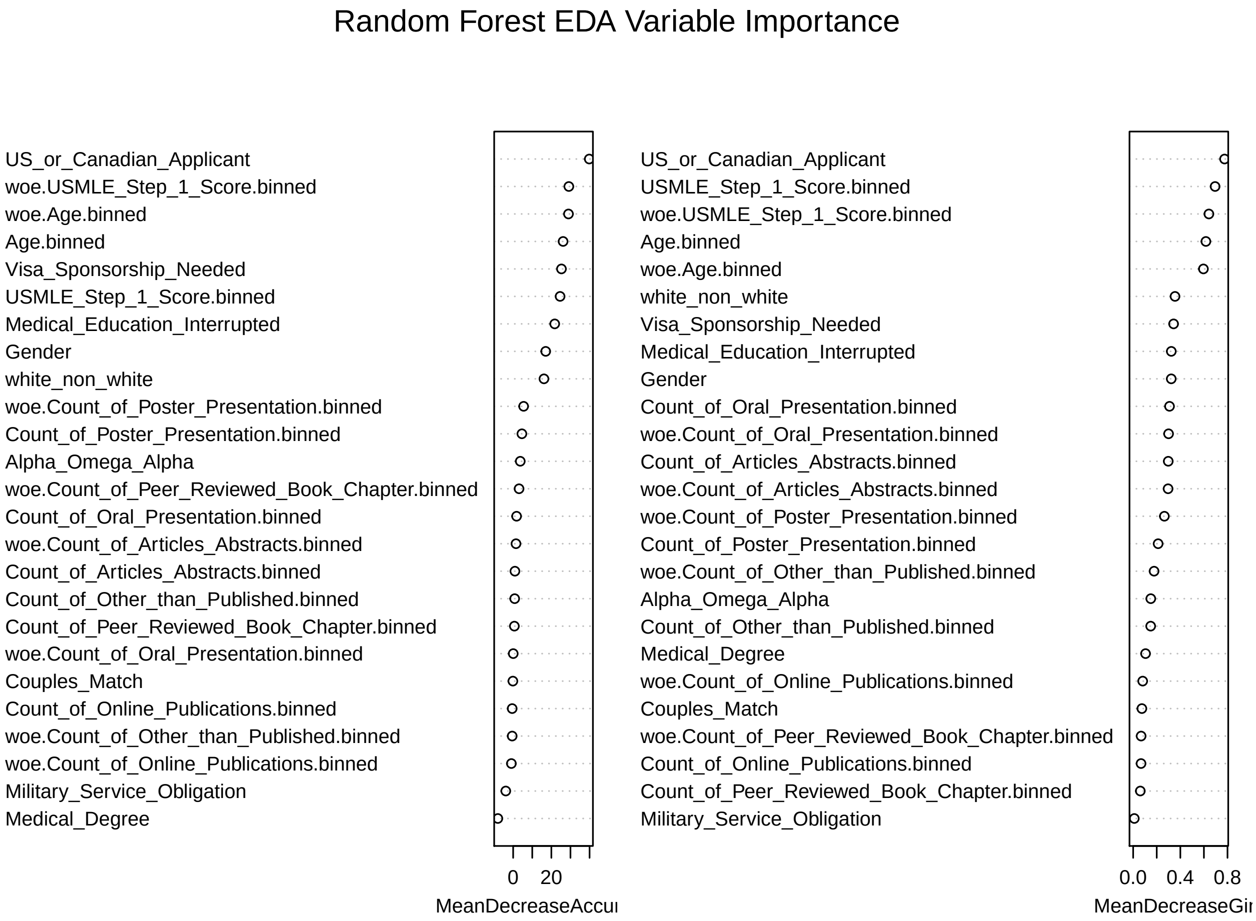
The simple tree was plotted below. The a, b and c variables were near the roots of the tree demonstrating importance. DO ALL THE VARIABLES NEED TO BE BINNED AS FACTORS TO RUN A TREE PLOT?

plot rpart EDA-1.bb


1
1
.23 .77
100%
yes
US_or_Canadian_Applicant = No
no
2
0
.52 .48
28%
3
1
.11 .89
72%
Age.binned = (30.75890411, Inf]
Medical_Education_Interrupted = Yes
6
1
.30 .70
10%
USMLE_Step_1_Score.binned = (−Inf,198]
4
0
.71 .29
12%
5
1
.38 .62
15%
12
0
.82 .18
1%
13
1
.22 .78
8%
7
1
.08 .92
62%

Exploratory random forest was also performed. The variable importance for the random forest model was summarized in the figure below. The variables capital_run_length_longest, capital_run_length_total, char_freq_dollar.binned, word_freq_free and word_freq_your were the top five using accuracy and the Gini index.

EDA-1.bb



## 3) The Model Build

All models were fit using the data labeled train and validated using the data labeled test. 10-fold cross-validation was performed for variable selection and parameter estimation was performed using cross-validation where appropriate.

### (1) Logistic regression using backwards variable selection model

A logistic regression model using backwards variable selection was fit. The summary of the model coefficients for the final model is presented in Table 3. Table 4 demonstrates the confusion matrix for the in-sample performance of the model and Table 5 demonstrates the confusion matrix? or AUC? for the out-of-sample performance.

#I'M NOT SURE IF I INCLUDED ALL THE RIGHT VARIABLES

The in-sample accuracy was 0.8304239 and the out-of-sample accuracy was 0.7713664.

Table 2: Backwards Logistic Regression Model Results

| | *Dependent variable:* |
|---|---|
| | as.factor(Match_Status) |
| white_non_whiteWhite | 0.413$^{***}$ (0.128, 0.697) |
| woe.Age.binned | −0.005$^{***}$ (−0.007, −0.004) |
| GenderMale | −0.249 (−0.564, 0.066) |
| Couples_MatchYes | 0.824$^{**}$ (0.036, 1.611) |
| US_or_Canadian_ApplicantYes | 1.536$^{***}$ (1.203, 1.868) |
| Medical_Education_InterruptedYes | −0.549$^{***}$ (−0.900, −0.198) |
| Visa_Sponsorship_NeededYes | −0.391$^{*}$ (−0.814, 0.032) |
| woe.USMLE_Step_1_Score.binned | −0.007$^{***}$ (−0.009, −0.005) |
| Constant | 0.247 (−0.068, 0.562) |
| Observations | 1,604 |
| Log Likelihood | −624.123 |
| Akaike Inf. Crit. | 1,266.246 |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table 3: Confusion matrix of Backwards LR on train set

| | 0 | 1 |
|---|---|---|
| 0 | 0.4587912 | 0.5412088 |
| 1 | 0.0604839 | 0.9395161 |

### (2) Tree model

A CART model was fit using the rpart package. The final model is presented in the figure below. The majority of final predictors were derived from the binning process. The variables char_free_exclamation.binned, word_freq_removed.binned, and woe.char_freq_dollar.binned had significant influence in the model. Table 6 demonstrates the confusion matrix for the in-sample performance of the model and Table 7 demonstrates the confusion matrix for the out-of-sample performance.

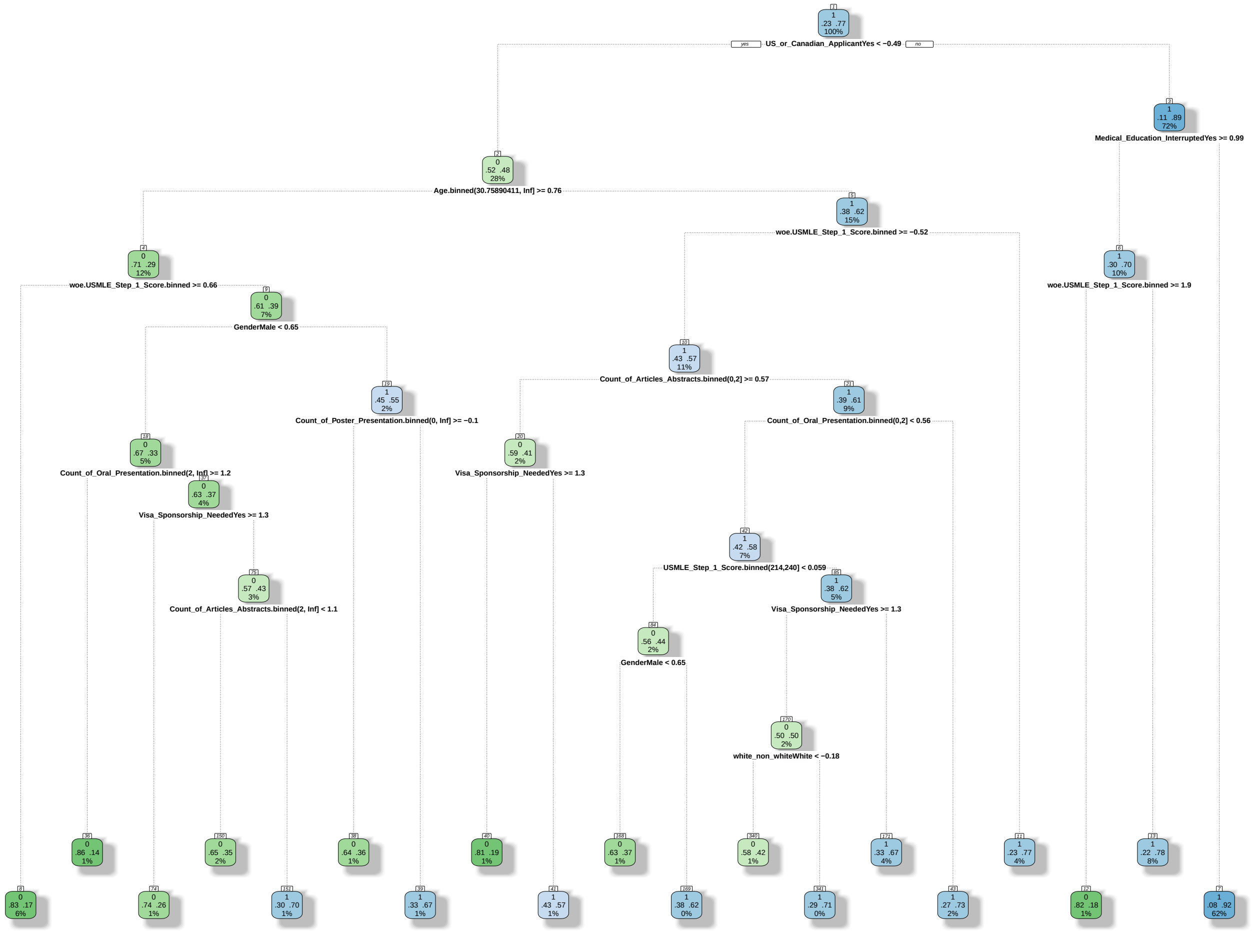
plot -1.bb

Table 4: Confusion matrix of Backwards LR on test set

|  | 0 | 1 |
|---|---|---|
| 0 | 0.3571429 | 0.6428571 |
| 1 | 0.0469851 | 0.9530149 |

Table 5: Confusion matrix of CART on train set

|  | 0 | 1 |
|---|---|---|
| 0 | 0.5109890 | 0.4890110 |
| 1 | 0.0491935 | 0.9508065 |



1
1 .23 .77 100%
yes
US_or_Canadian_ApplicantYes < −0.49
no
2
0 .52 .48 28%
Age.binned(30.75890411, Inf] >= 0.76
3
1 .11 .89 72%
Medical_Education_InterruptedYes >= 0.99
4
0 .71 .29 12%
woe.USMLE_Step_1_Score.binned >= 0.66
5
1 .38 .62 15%
woe.USMLE_Step_1_Score.binned >= −0.52
6
1 .30 .70 10%
woe.USMLE_Step_1_Score.binned >= 1.9
9
0 .61 .39 7%
GenderMale < 0.65
10
1 .43 .57 11%
Count_of_Articles_Abstracts.binned(0,2] >= 0.57
18
0 .67 .33 5%
Count_of_Oral_Presentation.binned(2, Inf] >= 1.2
19
1 .45 .55 2%
Count_of_Poster_Presentation.binned(0, Inf] >= −0.1
20
0 .59 .41 2%
Visa_Sponsorship_NeededYes >= 1.3
21
1 .39 .61 9%
Count_of_Oral_Presentation.binned(0,2] < 0.56
37
0 .63 .37 4%
Visa_Sponsorship_NeededYes >= 1.3
42
1 .42 .58 7%
USMLE_Step_1_Score.binned(214,240] < 0.059
75
0 .57 .43 3%
Count_of_Articles_Abstracts.binned(2, Inf] < 1.1
85
1 .38 .62 5%
Visa_Sponsorship_NeededYes >= 1.3
84
0 .56 .44 2%
GenderMale < 0.65
170
0 .50 .50 2%
white_non_whiteWhite < −0.18
36
0 .86 .14 1%
150
0 .65 .35 2%
38
0 .64 .36 1%
40
0 .81 .19 1%
168
0 .63 .37 1%
340
0 .58 .42 1%
171
1 .33 .67 4%
11
1 .23 .77 4%
13
1 .22 .78 8%
8
0 .83 .17 6%
74
0 .74 .26 1%
151
1 .30 .70 1%
39
1 .33 .67 1%
41
1 .43 .57 1%
169
1 .38 .62 0%
341
1 .29 .71 0%
43
1 .27 .73 2%
12
0 .82 .18 1%
7
1 .08 .92 62%

NOT WORKING, BECAUSE TEST HAS NOT BEEN WOE BINNED???

It is not working because the training model had a different amount of columns then the test data you're feeding into it. Just ensure that they have the correct columns that match. I am commenting this out as it's not something I should fix.

Georg: Should work now

For the CART model, the in-sample accuracy was 0.8509975 and out-of-sample accuracy was

Table 6: Confusion matrix of CART on test set

|   | 0 | 1 |
|---|---|---|
| 0 | 0.5109890 | 0.4890110 |
| 1 | 0.0491935 | 0.9508065 |

0.7675558.

Table 7: Confusion matrix of SVM model on train set

| | 0 | 1 |
|---|---|---|
| 0 | 0.4697802 | 0.5302198 |
| 1 | 0.0620968 | 0.9379032 |

Table 8: Confusion matrix of SVM model on test set

| | 0 | 1 |
|---|---|---|
| 0 | 0.4697802 | 0.5302198 |
| 1 | 0.0620968 | 0.9379032 |

### (3) a Support Vector Machine model

The support vector machine model was fit. Cross validation identified a cost C = 1 using a linear kernel and a sigma = ??? using ??? support vectors.

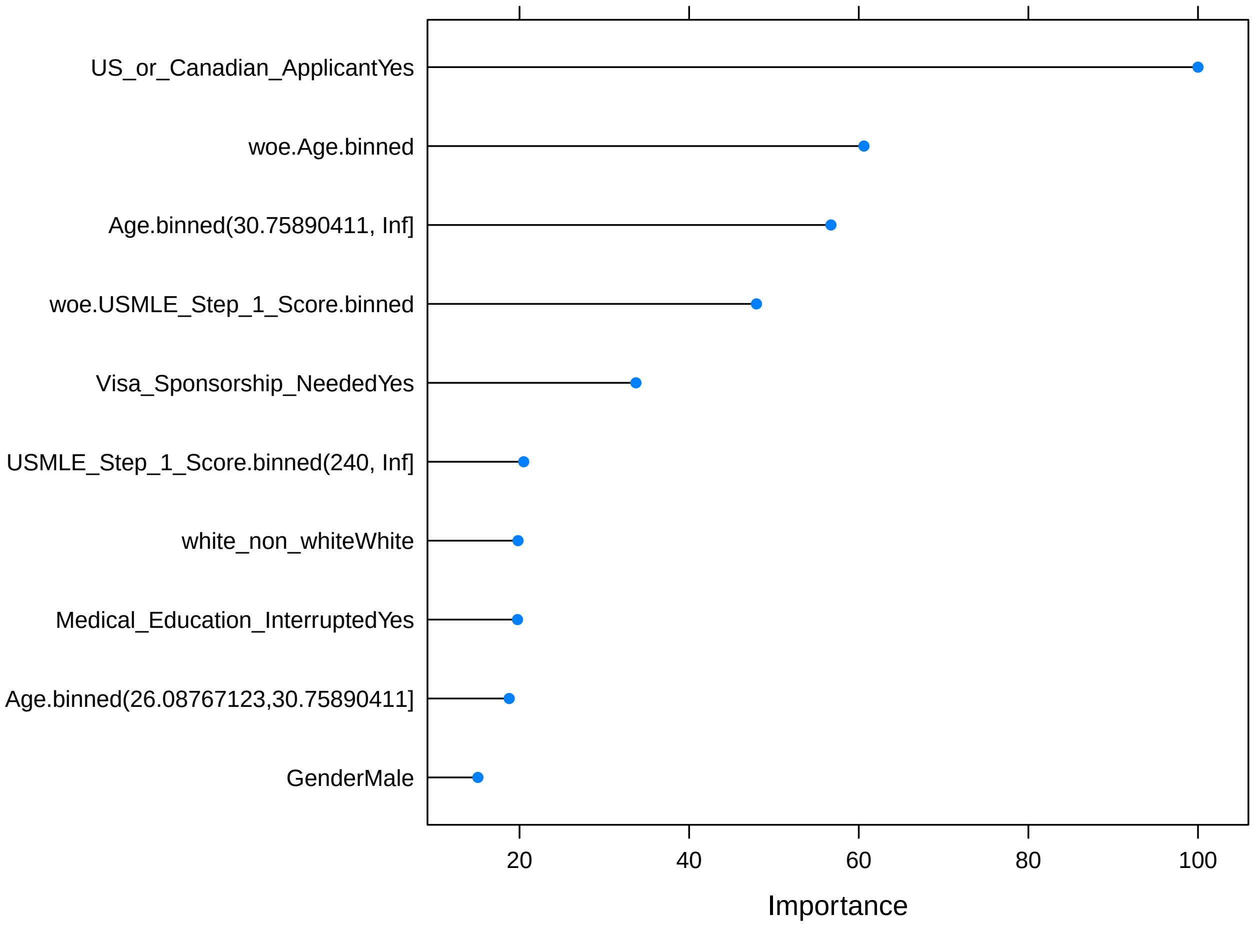
I DO NOT KNOW HOW TO MAKE THIS WORK. Georg:Should be working now

For the SVM model, the in-sample accuracy was 0.8316708 and out-of-sample accuracy was 0.7637452. The in-sample confusion matrix for the SVM model is shown in Table 8 and the out-of-sample confusion matrix for the SVM model is shown in Table 9.

### (4) Random Forest model

A random forest model was fit to the training data. Cross-validation selected the a final value used for mtry = 12 based on optimizing accuracy. The variable importance plot for the random forest model is demonstrated below. Important predictors were similar between the CART model and the RF model. The predictors char_freq_exclamation.binned, woe.char_freq_exclamation.binned, aand woe.char_freq_dollar.binned were top three for variable importance.

imp plot rf model-1.bb



For the random forest model, the in-sample accuracy was 0.8260599 and out-of-sample accuracy was 0.7332608. The in-sample confusion matrix for the RF model is shown in Table 10 and the out-of-sample confusion matrix for the RF model is shown in Table 11.

Table 9: Confusion matrix of Naive Bayes model on train set

| | 0 | 1 |
|---|---|---|
| 0 | 0.0137363 | 0.9862637 |
| 1 | 0.0000000 | 1.0000000 |

## 4) Naïve Bayes with WOE Binning model

Finally, a Naïve Bayes model was fit. Similar to the previous models, the top 5 WOE binned variables were also included in this model. Cross-validation demonstrated that the tuning parameter ‘laplace’ was held constant at a value of 0 and tuning parameter ‘adjust’ was held constant at a value of 1.

For the naive Bayes model, the in-sample accuracy was 0.7761845 and out-of-sample accuracy was 0.6984213. The in-sample confusion matrix for the naive Bayes model is shown in Table 12 and the out-of-sample confusion matrix for the naive Bayes model is shown in Table 13.

# 5) Model Comparison

Table 14 summarizes the overall in-sample and out-of-sample accuracy of each model. The best performing models (highest accuracy) was the random forest model with a test set accuracy of 0.7332608. The Logistic regression model using backwards elimination was second with a test set accuracy of 0.7713664. The Naive Bayes model did not perform as well as the other models. In summary, if accuracy is the most important aspect of the model and interpretion is not a priority then the best model was the random forest model. If interpretability of the model is paramount, then the logistic regression model is recommended.

#Annotation: Manuscript Figure 1: The first row called points assigned to each variable's measurement from rows 2-12, which are variables included in predictive model. Assigned points for all variables are then summed and total can be located on line 13 (total points). Once total points are located, draw a vertical line down to the bottom line to obtain the predicted probability of matching. For non-linear variables (count of oral presentations, etc.) values should be erad from left to right.

# Step 10: Calibration of the model based on the test data.

The ticks across the x-axis represent the frequency distribution (may be called a rug plot) of the predicted probabilities. This is a way to see where there is sparsity in your predictions and where there is a relative abundance of predictions in a given area of predicted probabilities.

The "Apparent" line is essentially the in-sample calibration.

The "Ideal" line represents perfect prediction as the predicted probabilities equal the observed probabilities.

The "Bias Corrected" line is derived via a resampling procedure to help add "uncertainty" to the calibration plot to get an idea of how this might perform "out-of-sample" and adjusts for "optimistic" (better than actual) calibration that is really an artifact of fitting a model to the data at hand. This is the line we want to look at to get an idea about generalization (until we have new data to try the model on).

When either of the two lines is above the "Ideal" line, this tells us the model underpredicts in that range of predicted probabilities. When either line is below the "Ideal" line, the model overpredicts in that range of predicted probabilities.

Applying to your specific plot, it appears most of the predicted probabilities are in the higher end (per rug plot). The model overall appears to be reasonably well calibrated based on the Bias-Corrected line closely following the Ideal line; there is some underprediction at lower predicted probabilities because the Bias-Corrected line is above the Ideal line around $< 0.3$ predicted probability.

The mean absolute error is the "average" absolute difference (disregard a positive or negative error) between predicted probability and actual probability. Ideally, we want this to be small (0 would be perfect indicating no error). This seems small in this plot, but may be situation dependent on how small is small.

# References

Lorrie Faith Cranor and Brian A. LaMacchia. Match_Status! Communications of the ACM. Vol. 41, No. 8 (Aug. 1998), Pages 74-83. Definitive version: http://www.acm.org/pubs/citations/journals/cacm/1998-41-8/p74-cranor/

#Appendix

# Appendix, Exploratory Data Analysis

The funModeling package will first give distributions for numerical data and finally creates cross-plots. This also saves the output of the distributions to the results folder.

# Appendix, Supplemental Table: Descriptive analysis of all variables considered in the training set along with their association to matching.

#Appendix, Data in a private repository to be shared with the journal http://dx.doi.org/10.17632/3rtg46skbd.1

Medical student #1 is a 27.4year old White Male who is a US Senior medical graduate

https://denverhealth.az1.qualtrics.com/WRQualtricsControlPanel/?Section=SV_3QmslHJJmin4xBX&SubSection=&SubSubSection=&PageActionOptions=&TransactionID=1&Repeatable=0&restrictToBrand=1&criteria=&ContextSection=EditSection

# Abstract DRAFT

Background: A model that predicts a medical student's chances of matching into an obstetrics and gynecology residency may facilitate improved counseling and fewer unmatched medical students.

Objective: We sought to construct and validate a model that predicts a medical student's chance of matching into obstetrics and gynecology residency.

Study Design: In all, 3441 medical students applied to a residency in Obstetrics and Gynecology at the University of Colorado from 2015 to 2018 were analyzed. The data set was splint into a model training cohort of 1604 who applied in 2015, 2016, and 2017 and a separate validation cohort of 1837 in 2018. In all, 19 candidate predictors for matching were collected. Multiple logistic models were fit onto the training choort to predict matching. Variables were removed using least absolute shrinkage and selection operator reduction to find the best parsimonious model. Model discrimination was measured using the concordance index. The model was internally valideated using 1,000 bootstrapped samples and temporarly validated by testing the model's performance in the validation cohort. Calibration curves were plotted to inform educators about the accuracy of predicted probabilities.

Results: The match rate in the training cohort was 77.3% (I need help getting 95% CI). The model had excellent discrimination and calibration during internal validation (bias-corrected concordance index,0.83) and maintained accuracy during temportal validation using the separate validation cohort (concordance index,0.84).

# Prose of the paper DRAFT

Materials and Methods: This was an institutional review board exempt retrospective cohort analysis of medical students who applied to Obstetrics and Gynecology (OBGYN) residency from 2015 to 2018. Guidelines for transparent reporting of a multivariable prediction model for individual outcomes were used in this study.(https://www.equator-network.org/reporting-guidelines/tripod-statement/). Eligible students were identified if they applied to OBGYN residency during the study period. The outcome of the model was defined as matching or not matching into residency for the specific application year. Individual predictors of successfully* matching were compiled from a literature review, expert opinion, and judgment then collected from the Electronic Residency Application Service materials.

Once the data set was complete it was divided into a model training and test set. *When an external validation data set is unavailable to test a new model but an existing modeling data set is sufficiently large, as in this case, it is recommended to split by time and develop the model using data from one period and evaluate its performance from data from a future period. We arbitrarily chose to divide the cohort into a training set of 2015 to 2017 data and a training set of 2018 data. In all, ?? candidate risk factors were considered for fitting on the training data set (supplmental table). Variable selection was done using a peenalized logistic regression called least absolute shrinkage and selection operator (LASSO). The LASSO model is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. We elected to use LASSO to choose which covariates to include over stepwise selection because the latter only improves prediction accuracy in certain cases, such as when only a few covariates have a strong relationship with the outcome. The logistic model's discriminative ability was measured by the area under the curve (AUC) for the receiver operating characteristic curve based on the sensitivity and specificity of the model. An AUC value closer to 1 indicates a better prediction of the outcome and an AUC value of 0.5 indicates that the model predicts no better than chance. The AUC is also a representation of the concordance index and measures the model's ability to generate a higher predicted probability of a successful match* occurring in a medical student who has a ????. For example, if we have a pair of medical students, in which one medical student matches and the other does not, the concordance index measures the model's ability to assign a higher risk of not matching to the medical student who successfully matches. All concordance indices and receiver operating characteristic curves were internally validated using a 1,000 bootstrap resample to correct for bias and overfitting within the model. The bootstrapping method of validation has been shown to be superior to other approaches to estimate internal validity. Calibration curves were also plotted to depict the relationship between the model's predicted outcomes against the cohort's observed outcome, where a perfectly calibrated model follows a 45° line. After the best model was selected and internally validated, the model was compared with the best currently available method of estimating risk, that is, an expert medical educator's predictions. To perform these comparisons, a subset of 50 participants was randomly selected for comparing the probability of matching between the model and the panel of experts. These ?? participants were used to compare predictions of the models with experts' predictions and not as a true independent validation subset. The model was rebuilt using the remaining participants in the data set excluding the 50 randomly selected participants. The candidate risk factors of these 50 participants were given to 20 "expert" medical educators with representation from each of the *** for review resulting in 1,000 expert predictions and 50 model predictions for each outcome. All medical educators were considered to be experienced in counseling medical students regarding OBGYN matching. Each of the 20 experts were asked to consider each medical student's data from all ??? variables among the 50 randomly selected students and provide their best estimated outcome by answering the following question: "Out of 100 medical students with these exact characteristics, estimate the number of medical students who would not matching into OBGYN during the 2019 application year." Individual medical educators' predictions were not averaged to yield a single value because incorporating each medical educator's predictions substantially increased statistical power. The model's predictions were compared with the experts' predictions, which included all risk factors, to determine which was most accurate. The difference in accuracy was determined by using a bootstrap method from their respective receiver operating characteristic curves. All analyses were performed using R 3.5. Results: A total of 3441 applied to obstetrics and gynecology residency at the University of Colorado from 2015 to 2018. The overall mean rate of matching in the training cohort was 1240 of 1604 was (77.3%). The unadjusted comparison of the 19 candidate predictors in the training cohort are presented in Supplemental Table 1. To identify predictors from the candidates we employed least absolute shrinkage and selection operator (LASSO). Regularisation techniques change how the model is fit by adding a penalty for every additional parameter you have in the model. 12 variables were included within the final model. Applicants from the United States or Canada, high USMLE Step 1 scores, female gender, White race, no visa sponsorship needed, membership in Alpha Omega Alpha, no interruption of medical training, couples matching, and allopathic medical training increased the chances of matching into OBGYN. In contrast, more oral presentations, increasing age, a higher number of peer-reviewed online publications, an increased number of authored book chapters, and a higher count of poster presentations all decreased the probability of matching into OBGYN (table 2). The nomogram illustrates the strength of association of the predictors to the outcome as well as the nonlinear associations between age, count of Oral Presentations, count of peer−reviewed book chapters and the chances of matching (Figure 1). # Appendix, DynNom Model for Shiny Upload I WOULD LIKE HELP UPLOADING THIS MODEL TO SHINY SERVER AS WELL PLEASE.

# Publish to shiny

# getwd()