Admissions' Yield Prediction by Gender in a Colorado Higher Education Institution.

by Vanessa Gonzalez

Overview

- "Yield" is the percent of students who choose to enroll in a particular college or university after having been offered admission.
- Need to increase female population in Stem related majors.
- Limited resources.
- Increased competition between universities for same student.

Questions to be Answered

- Which are the accepted students with a higher possibility to enroll?

- Which are the accepted female students with a higher possibility to enroll?

Data

Query Data from University Data Base

Combine and Format Data

Data Cleaning and Preparation

Exploratory Data Analysis

Data Modeling

Data Prep and Cleaning

- Some NA values: substituted by “Missing” or “None” as appropriate.
- Other NA values: substituted with KNN apmputation method.
- “state”variable: all different than “CO” substituted by “Other”.
- Data types: transformed to appropriate type.
- Factor variables with more than 21 levels were omitted.
- Dummy variables created.
- Variables other than “Enrolling” were transformed into number variables.
- Numerical values: normalized.

Data Sets

Data Set 1

- 6,235 observations
- 48 variables
- Admissions Cycle 2016-2017
- All admitted students

Data Set 2

- 2,046 observations
- 48 variables
- Admissions Cycle 2016-2017
- All admitted female students

Data Set of Admitted Students

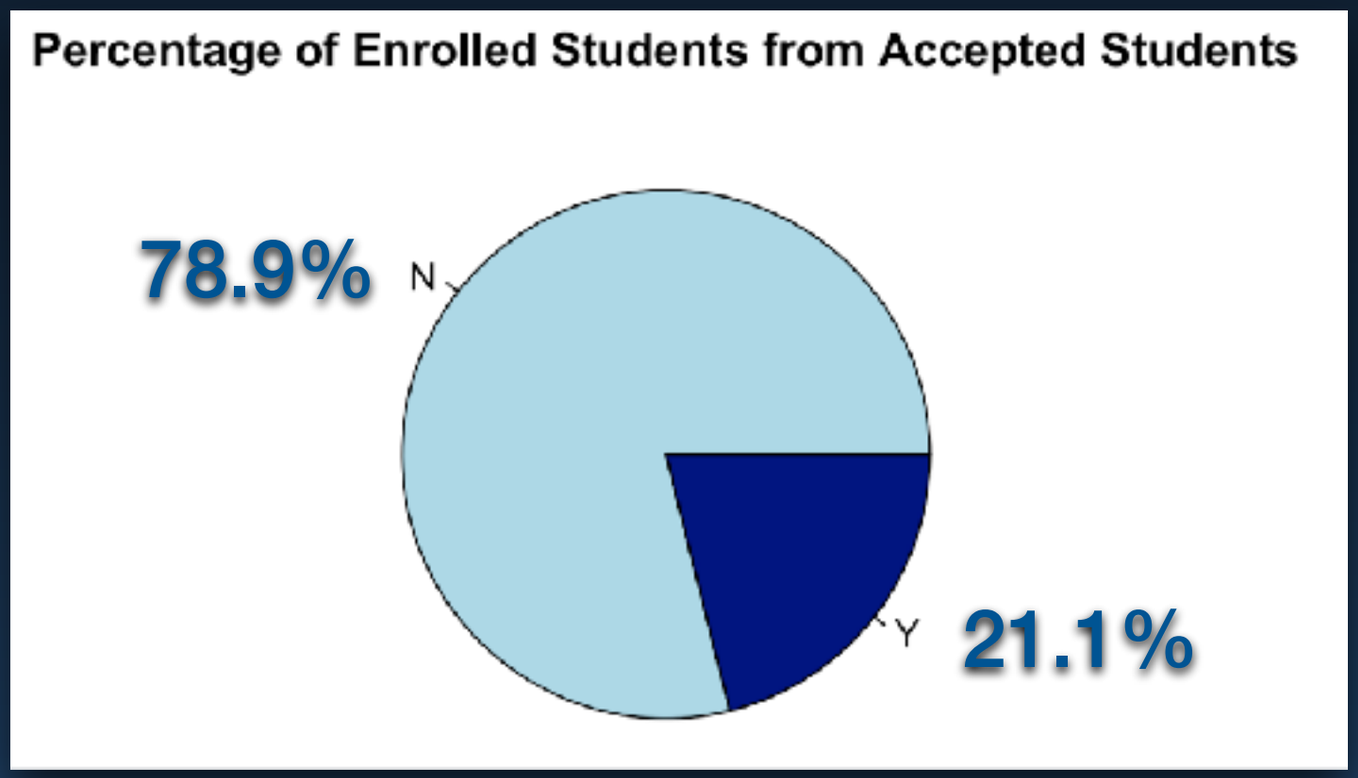
'data.frame': 6235 obs. of 48 variables:
$ Enrolling : chr "N" "N" "N" "N" ...
$ Sex : chr "M" "M" "M" "M" ...
$ Expel : chr "N" "N" "N" "N" ...
$ First.Gen : chr "N" "N" "N" "N" ...
$ Challenge Tag : chr "N" "N" "N" "N" ...
$ Pathway Tag : chr "N" "N" "N" "N" ...
$ Boettcher.Semi : chr "N" "N" "N" "N" ...
$ Boettcher.Final : chr "N" "N" "N" "N" ...
$ Daniels.Semi : chr "N" "N" "N" "N" ...
$ Daniels.Final : chr "N" "N" "N" "N" ...
$ Harvey.App : chr "N" "N" "N" "N" ...
$ Harvey.Final : chr "N" "N" "N" "N" ...
$ FC.App : chr "N" "N" "N" "N" ...
$ FC.Final : chr "N" "N" "N" "N" ...
$ Thorson.App : chr "N" "N" "N" "N" ...
$ Thorson.Admit : chr "N" "N" "N" "N" ...
$ Summet.App : chr "N" "N" "N" "N" ...
$ Summet.Participant : chr "N" "N" "N" "N" ...
$ Mines.Medal : chr "N" "N" "N" "N" ...

Exploratory Analysis (EDA)

6,235 Students in the Data Set 1

Admitted Students by Gender

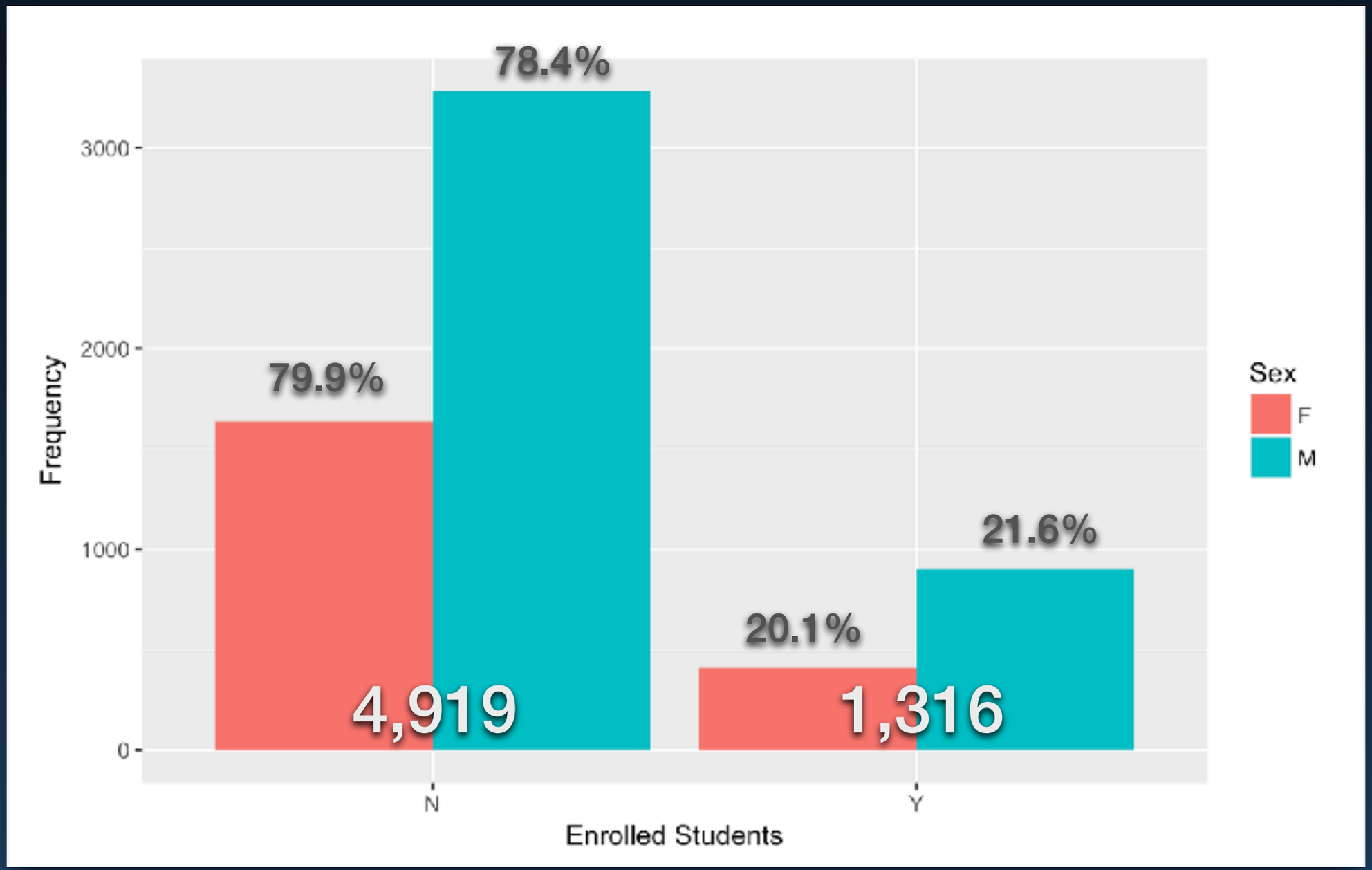
Sex	Admitted: No	Admitted: Yes	Grand Total
Female	1,267 25.72%	2,046 32.81%	**3,313** **29.68%**
Male	3,660 74.28%	4,189 67.19%	**7,849** **70.32%**
Grand Total	**4,927** **100.00%**	**6,235** **100.00%**	**11,162** **100.00%**



2016-2017 Cycle

Explanatory Analysis (EDA)

How many accepted students enrolled?



Exploratory Analysis (EDA)

Summary of Data Set 1

Factor for Classification

Enrolling
N:4919
Y:1316

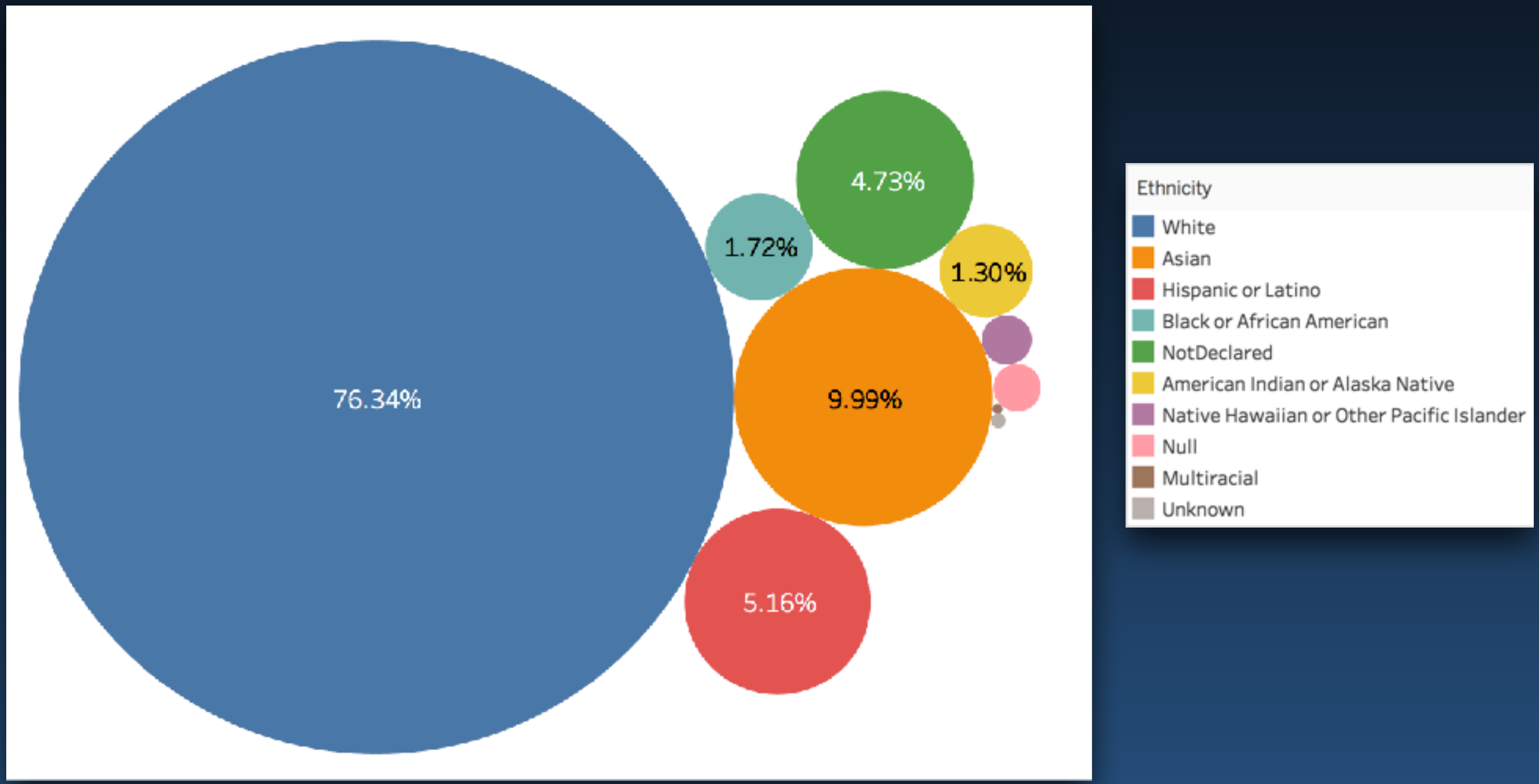
Variables with Dummies

```{r}
summary(dfwithDummies)
```

```
     Sex.F            Sex.M            Expel.N          Expel.Y          First.Gen.N      First.Gen.Y      Challenge Tag.N  Challenge Tag.Y
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:0.00000
 Median :0.0000   Median :1.0000   Median :1.0000   Median :0.00000   Median :1.0000   Median :0.00000   Median :1.0000   Median :0.00000
 Mean   :0.3281   Mean   :0.6719   Mean   :0.9877   Mean   :0.01235   Mean   :0.9185   Mean   :0.08949   Mean   :0.9798   Mean   :0.02021
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.00000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.00000
 Boettcher.Semi.N Boettcher.Semi.Y Boettcher.Final.N Boettcher.Final.Y Daniels.Semi.N   Daniels.Semi.Y   Daniels.Final.N Daniels.Final.Y
 Min.   :0.0000   Min.   :0.00000  Min.   :0.000     Min.   :0.00000   Min.   :0.0000   Min.   :0.03000  Min.   :0.000   Min.   :0.00000
 1st Qu.:1.0000   1st Qu.:0.00000  1st Qu.:1.000     1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:0.03000  1st Qu.:1.000   1st Qu.:0.00000
 Median :1.0000   Median :0.00000  Median :1.000     Median :0.00000   Median :1.0000   Median :0.03000  Median :1.000   Median :0.00000
 Mean   :0.9843   Mean   :0.01572  Mean   :0.996     Mean   :0.00401   Mean   :0.9883   Mean   :0.01171  Mean   :0.996   Mean   :0.00401
 3rd Qu.:1.0000   3rd Qu.:0.00000  3rd Qu.:1.000     3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.03000  3rd Qu.:1.000   3rd Qu.:0.00000
 Max.   :1.0000   Max.   :1.00000  Max.   :1.000     Max.   :1.00000   Max.   :1.0000   Max.   :1.03000  Max.   :1.000   Max.   :1.00000
  Harvey.App.N     Harvey.App.Y     Harvey.Final.N   Harvey.Final.Y      FC.App.N         FC.App.Y       Thorson.App.N    Thorson.App.Y
 Min.   :0.0000   Min.   :0.00000  Min.   :0.0000   Min.   :0.000000  Min.   :0.000   Min.   :0.00300  Min.   :0.0000   Min.   :0.00000
 1st Qu.:1.0000   1st Qu.:0.00000  1st Qu.:1.0000   1st Qu.:0.000000  1st Qu.:1.000   1st Qu.:0.00300  1st Qu.:1.0000   1st Qu.:0.00000
 Median :1.0000   Median :0.00000  Median :1.0000   Median :0.000000  Median :1.000   Median :0.00300  Median :1.0000   Median :0.00000
 Mean   :0.9102   Mean   :0.08982  Mean   :0.9925   Mean   :0.007538  Mean   :0.986   Mean   :0.01395  Mean   :0.9578   Mean   :0.04213
 3rd Qu.:1.0000   3rd Qu.:0.00000  3rd Qu.:1.0000   3rd Qu.:0.000000  3rd Qu.:1.000   3rd Qu.:0.00300  3rd Qu.:1.0000   3rd Qu.:0.00000
 Max.   :1.0000   Max.   :1.00000  Max.   :1.0000   Max.   :1.000000  Max.   :1.000   Max.   :1.00300  Max.   :1.0000   Max.   :1.00000
 Thorson.Admit.N  Thorson.Admit.Y  Summet.Participant.N Summet.Participant.Y Mines.Medal.N    Mines.Medal.Y        SPS.N
 Min.   :0.0000   Min.   :0.00000  Min.   :0.0000       Min.   :0.000000     Min.   :0.0000   Min.   :0.0003000  Min.   :0.0000
 1st Qu.:1.0000   1st Qu.:0.00000  1st Qu.:1.0000       1st Qu.:0.000000     1st Qu.:1.0000   1st Qu.:0.0003000  1st Qu.:1.0000
 Median :1.0000   Median :0.00000  Median :1.0000       Median :0.000000     Median :1.0000   Median :0.0003000  Median :1.0000
 Mean   :0.9751   Mean   :0.02486  Mean   :0.9966       Mean   :0.003368     Mean   :0.9995   Mean   :0.0004812  Mean   :0.9933
 3rd Qu.:1.0000   3rd Qu.:0.00000  3rd Qu.:1.0000       3rd Qu.:0.000000     3rd Qu.:1.0000   3rd Qu.:0.0003000  3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.00000  Max.   :1.0000       Max.   :1.000000     Max.   :1.0000   Max.   :1.0003000  Max.   :1.0000
     SPS.Y          Veteran.N        Veteran.Y         Legacy.N         Legacy.Y         Athlete.N        Athlete.Y          State.CO
 Min.   :0.000000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0300   Min.   :0.00000   Min.   :0.0000
 1st Qu.:0.000000   1st Qu.:1.0000   1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:1.0300   1st Qu.:0.00000   1st Qu.:0.0000
 Median :0.000000   Median :1.0000   Median :0.00000   Median :1.0000   Median :0.0000   Median :1.0300   Median :0.00000   Median :0.0000
 Mean   :0.006736   Mean   :0.9315   Mean   :0.06848   Mean   :0.9267   Mean   :0.0733   Mean   :0.9739   Mean   :0.02614   Mean   :0.3025
 3rd Qu.:0.000000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0300   3rd Qu.:0.00000   3rd Qu.:1.0000
 Max.   :1.000000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0300   Max.   :1.00000   Max.   :1.0000
  State.Other      Citizenship.Foreign National/International Citizenship.International Citizenship.Missing Citizenship.U.S. Citizen
 Min.   :0.0000   Min.   :0.00000                             Min.   :0.000000         Min.   :0.000300   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.00000                             1st Qu.:0.000000         1st Qu.:0.000300   1st Qu.:1.0000
 Median :1.0000   Median :0.00000                             Median :0.000000         Median :0.000300   Median :1.0000
 Mean   :0.6974   Mean   :0.03368                             Mean   :0.001604         Mean   :0.001925   Mean   :0.9471
 3rd Qu.:1.0000   3rd Qu.:0.00000                             3rd Qu.:0.000000         3rd Qu.:0.000300   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.00000                             Max.   :1.000000         Max.   :1.000300   Max.   :1.0000
```

Exploratory Analysis (EDA)

Ethnicity of Enrolled Students



Analysis for Data Set 1

Which are the accepted students with a higher possibility to enroll?

Machine Learning Algorithms

Machine Learning Algorithms Used

- kNN
- Random Forest
- Support Vector Machine (SVM)

Process of Analysis for Data Set 1

Substitute NA Values through kNN

Create Subset of Just Accepted Students

Create Model Utilizing Random Forest Method

Determine Variable Importance

Create SVM Model

SVM Model Variations

Compare Methods Accuracy

Run Model with Less Variables

Ready to Predict Students with Higher Probability of Enrollment.

Analysis for Data Set 2

Which are the accepted female students with a higher possibility to enroll?

Process of Analysis

Create Subset of Just Accepted Female Students

Create Model Utilizing Random Forest Method

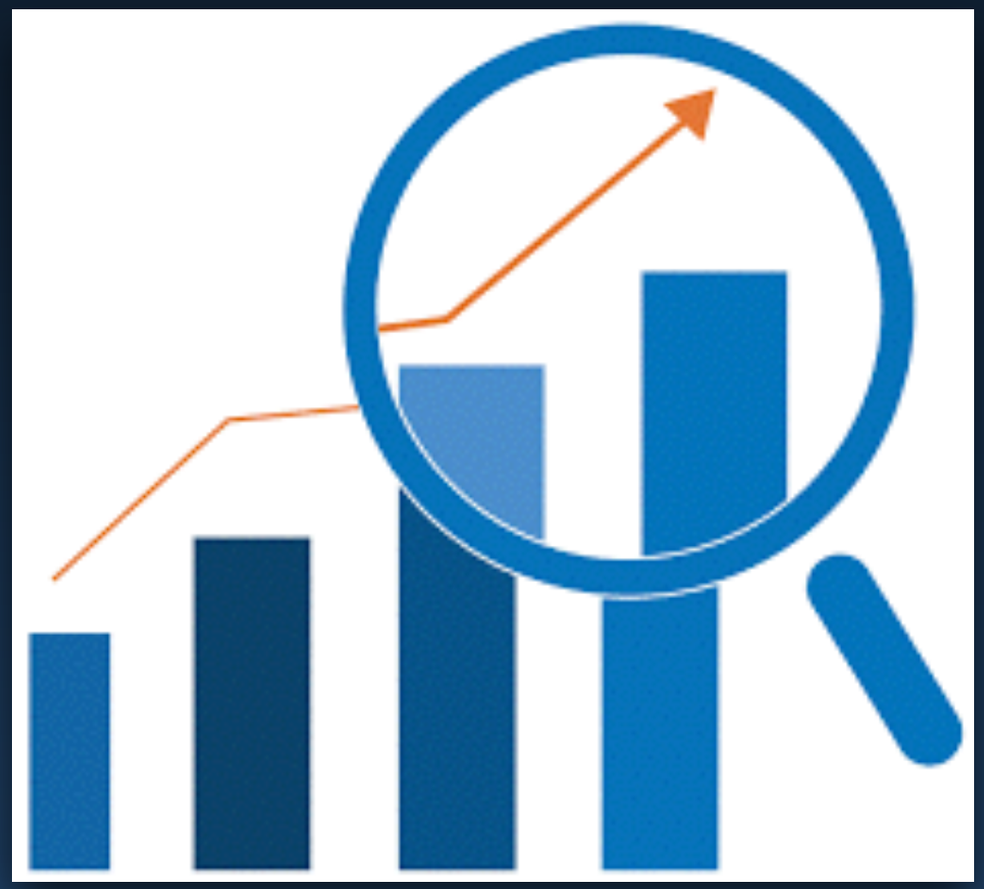
Determine Variable Importance

Run Model with Less Variables

Ready to Predict Female Students with a Higher Probability of Enrollment

Results Data Set 1

Random Forest All Variables Analysis

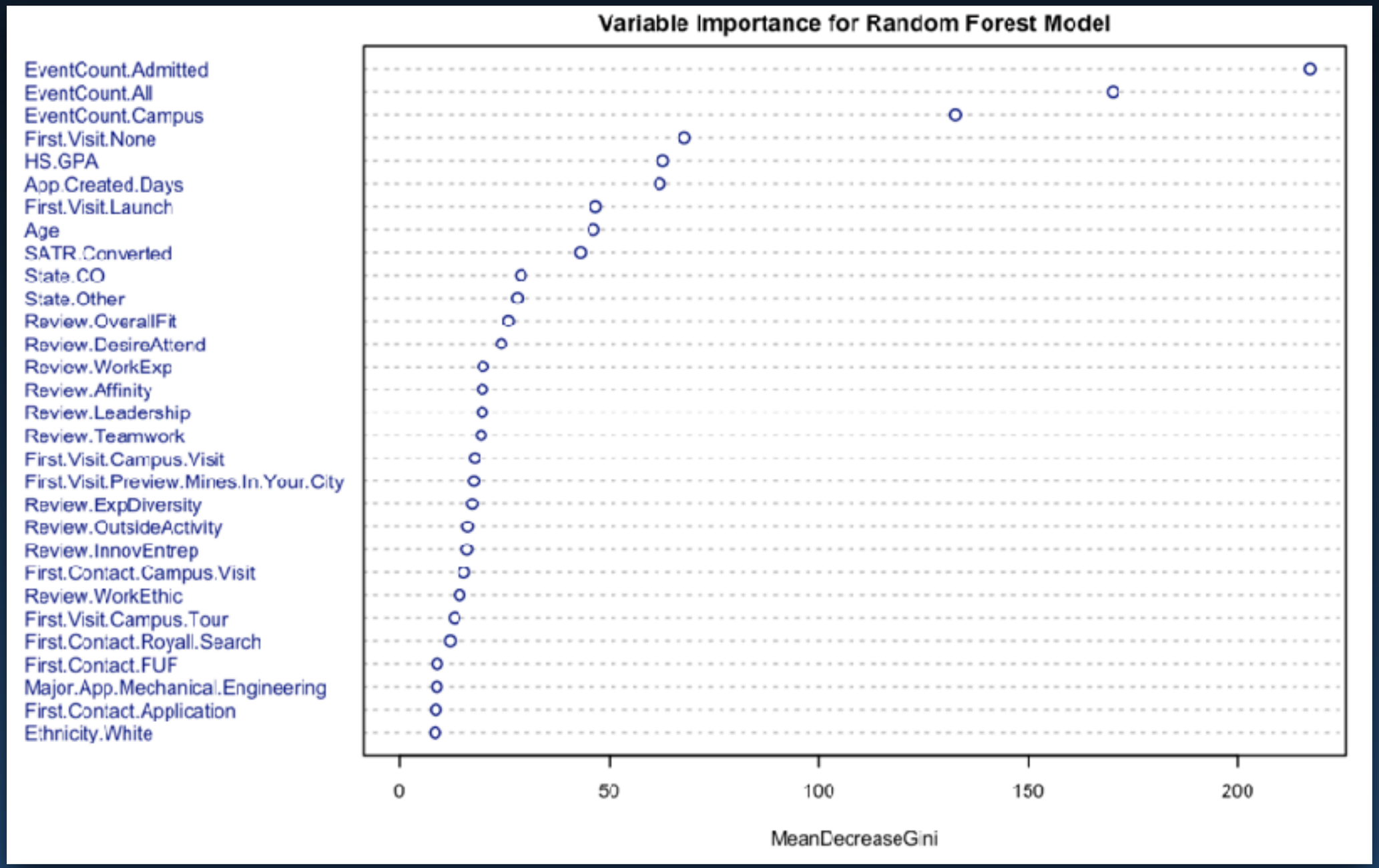


90.76% Accuracy

```
Confusion Matrix and Statistics


rfModel.prediction    N    Y
                 N 1181   96
                 Y   48  233

               Accuracy : 0.9076
                 95% CI : (0.8921, 0.9215)
    No Information Rate : 0.7888
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7069
 Mcnemar's Test P-Value : 8.978e-05

            Sensitivity : 0.9609
            Specificity : 0.7082
         Pos Pred Value : 0.9248
         Neg Pred Value : 0.8292
             Prevalence : 0.7888
         Detection Rate : 0.7580
   Detection Prevalence : 0.8196
      Balanced Accuracy : 0.8346

       'Positive' Class : N
```

Results Data Set 1

Variable Importance for all Students



Results Data Set 1

Variable Importance for all Students

- Event Count
- First Visit
- HS.GPA
- App.Created Days
- Age
- SATR.Converted
- State
- Review Variables
- Major.App.ME

Results Data Set 1

Different SVM Models

SVM-Scaling

```
Confusion Matrix and Statistics

          model_ksvm_predictor
             N    Y
        N 1179   50
        Y  104  225

                    Accuracy : 0.9012
                      95% CI : (0.8852, 0.9155)
       No Information Rate : 0.8235
       P-Value [Acc > NIR] : < 2.2e-16

                     Kappa : 0.6843
 Mcnemar's Test P-Value : 1.947e-05

                Sensitivity : 0.9189
                Specificity : 0.8182
             Pos Pred Value : 0.9593
             Neg Pred Value : 0.6839
                 Prevalence : 0.8235
             Detection Rate : 0.7567
       Detection Prevalence : 0.7888
          Balanced Accuracy : 0.8686

           'Positive' Class : N
```

90.12%

kSVM - No Scaling

```
agreement
    FALSE      TRUE
0.1033376 0.8966624
```

89.66%

Poly

```
Confusion Matrix and Statistics

          model_predictions_poly
             N    Y
        N 1179   50
        Y  104  225

                    Accuracy : 0.9012
                      95% CI : (0.8852, 0.9155)
       No Information Rate : 0.8235
       P-Value [Acc > NIR] : < 2.2e-16

                     Kappa : 0.6843
 Mcnemar's Test P-Value : 1.947e-05

                Sensitivity : 0.9189
                Specificity : 0.8182
             Pos Pred Value : 0.9593
             Neg Pred Value : 0.6839
                 Prevalence : 0.8235
             Detection Rate : 0.7567
       Detection Prevalence : 0.7888
          Balanced Accuracy : 0.8686

           'Positive' Class : N
```

90.12%

RBF

```
Confusion Matrix and Statistics

          model_predictions_rbf
             N    Y
        N 1182   47
        Y  101  225

                    Accuracy : 0.9031
                      95% CI : (0.8873, 0.9173)
       No Information Rate : 0.8254
       P-Value [Acc > NIR] : < 2.2e-16

                     Kappa : 0.6894
 Mcnemar's Test P-Value : 5.184e-06

                Sensitivity : 0.9191
                Specificity : 0.8272
             Pos Pred Value : 0.9618
             Neg Pred Value : 0.6839
                 Prevalence : 0.8254
             Detection Rate : 0.7587
       Detection Prevalence : 0.7888
          Balanced Accuracy : 0.8732

           'Positive' Class : N
```

90.31%

Results Data Set 1

Random Forest Less Number of Variables Analysis

90.69% Accuracy

```
Confusion Matrix and Statistics

rfModelTop25.prediction    N    Y
                       N 1182   98
                       Y   47  231

               Accuracy : 0.9069
                 95% CI : (0.8914, 0.9209)
    No Information Rate : 0.7888
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7038
 Mcnemar's Test P-Value : 3.292e-05

            Sensitivity : 0.9618
            Specificity : 0.7021
         Pos Pred Value : 0.9234
         Neg Pred Value : 0.8309
             Prevalence : 0.7888
         Detection Rate : 0.7587
   Detection Prevalence : 0.8216
      Balanced Accuracy : 0.8319

       'Positive' Class : N
```

Results Data Set 2

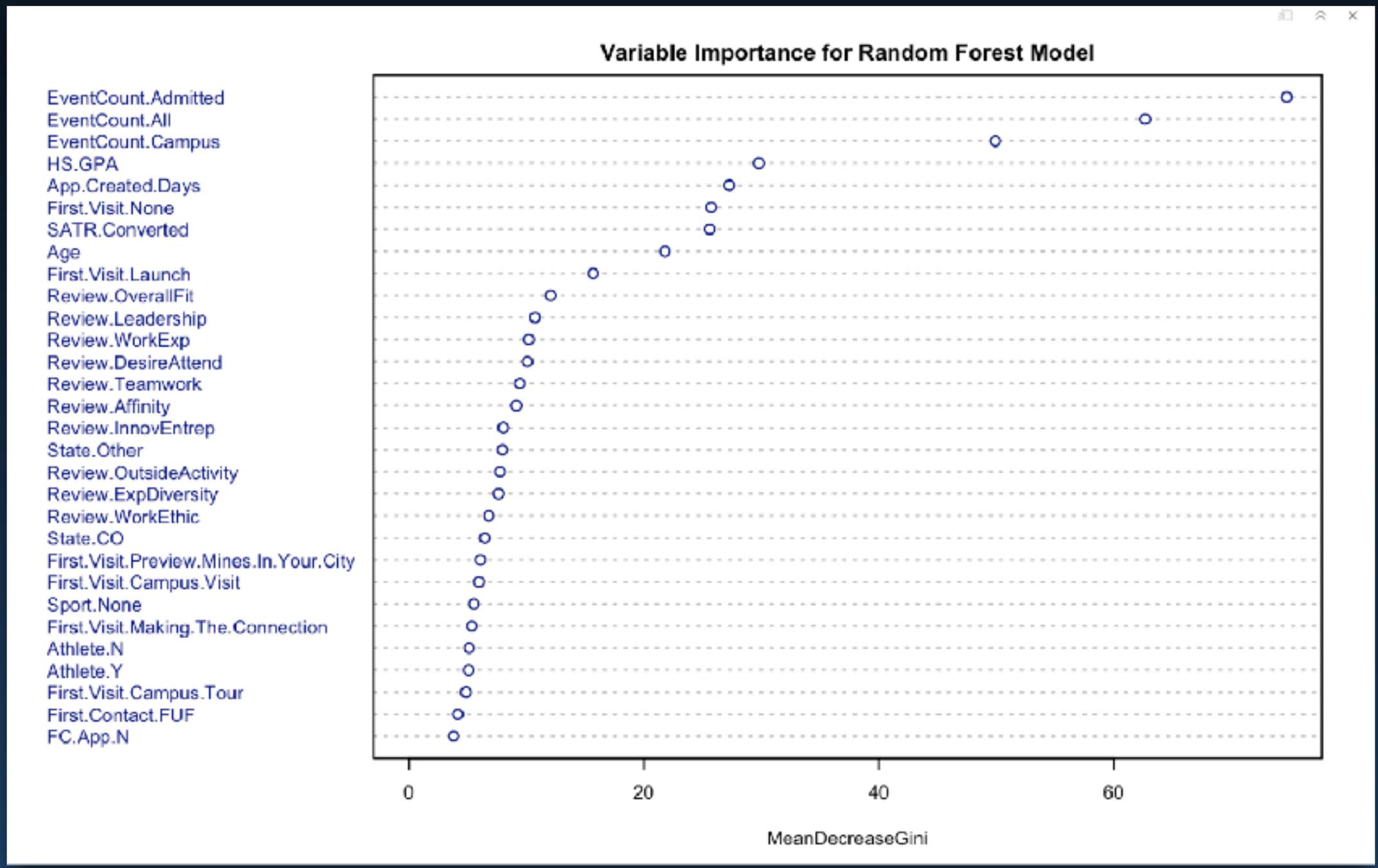
Random Forest for Female Students Predictive Model

90.69% Accuracy

```
Confusion Matrix and Statistics


rfModelTop25.prediction    N    Y
                      N 1183   99
                      Y   46  230

               Accuracy : 0.9069
                 95% CI : (0.8914, 0.9209)
    No Information Rate : 0.7888
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7031
 Mcnemar's Test P-Value : 1.572e-05

            Sensitivity : 0.9626
            Specificity : 0.6991
         Pos Pred Value : 0.9228
         Neg Pred Value : 0.8333
             Prevalence : 0.7888
         Detection Rate : 0.7593
   Detection Prevalence : 0.8228
      Balanced Accuracy : 0.8308

       'Positive' Class : N
```

Results Data Set 2

Variables Importance Plot for Female Students



Results Data Set 2

Variable Importance for Female Students

- Event Count
- HS.GPA
- App.Created Days
- First Visit
- Age
- SATR.Converted
- State
- Review Variables
- Major.App.ME

Results Summary

Different Machine Learning Methods Results

Machine Learning Method	Accuracy	Kappa	P-Value
Random Forest - All Variables	90.8%	0.7069	0.00008978
Random Forest - Less Variables	90.7%	0.7038	0.00003292
kSVM-Scaling	90.1%	0.6843	0.00001947
kSVM-No Scaling	89.0%		
RBF	90.3%	0.6894	0.00005184
Poly-SVM	90.1%	0.6843	0.00001947
Random Forest - Female Students	90.7%	0.7031	0.00001572

Conclusions

- 6,235 Students were admitted on the 2016-2017 admissions cycle. 1,316 of these students enrolled for the 2017-2018 academic year.

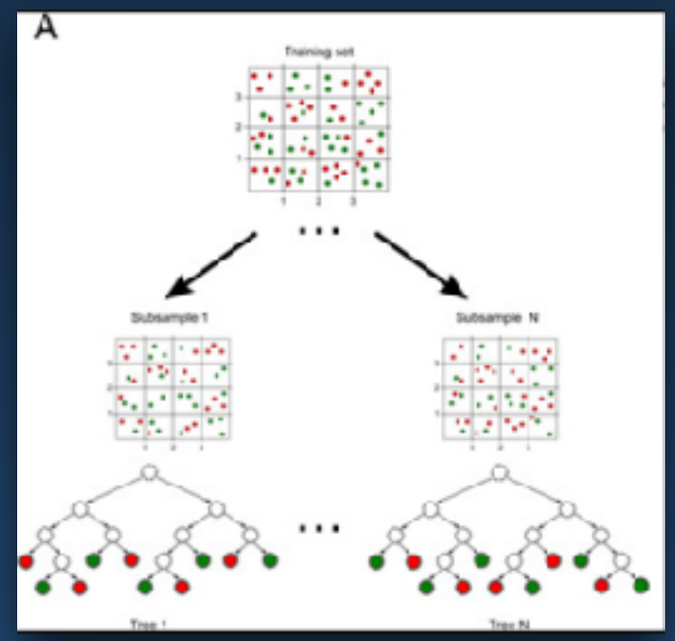
- By using the Random Forest Algorithm with all the variables we were able to predict with a 90.8% accuracy which of the accepted students had a higher probability of enrolling the fallowing year.

- Additional resources may be directed to these students with a higher probability of enrollment increasing yield.

Conclusions

- By using a smaller number of variables (25) the accuracy of the model did not decreased significantly (Accuracy of 90.7% and Kappa of 0.7038) suggesting the use of the Model with less variables in the future.

- SVM Models were similar but not more accurate than the Random Forest model.



Conclusions

- 2,046 Female Students were admitted on the 2016-2017 admissions cycle. 412 of these students enrolled for the 2017-2018 academic year.

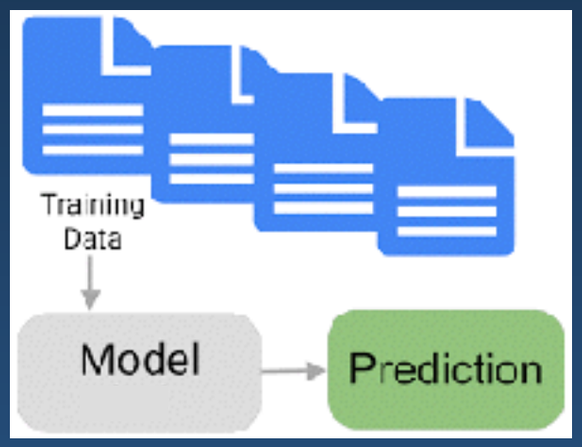
- By using the Random Forest Algorithm with all the variables we were able to predict with a 90.7% accuracy and a Kappa of .7030 which of the female accepted students had a higher probability of enrolling the fallowing year.

Steps Forward

- There is a lot more to be done. More questions to to be answered and other angles to be explored. It would be interesting to experiment with the use of less variables or a combination of some of them to increase accuracy.

- There is more work to be done in the development of the report that queries the database to avoid some of the data clean-up. Adjustments will be made on this area.

- It will be interesting to direct resources to students with higher probability to enroll and measure results.



Student Support

Thank You!