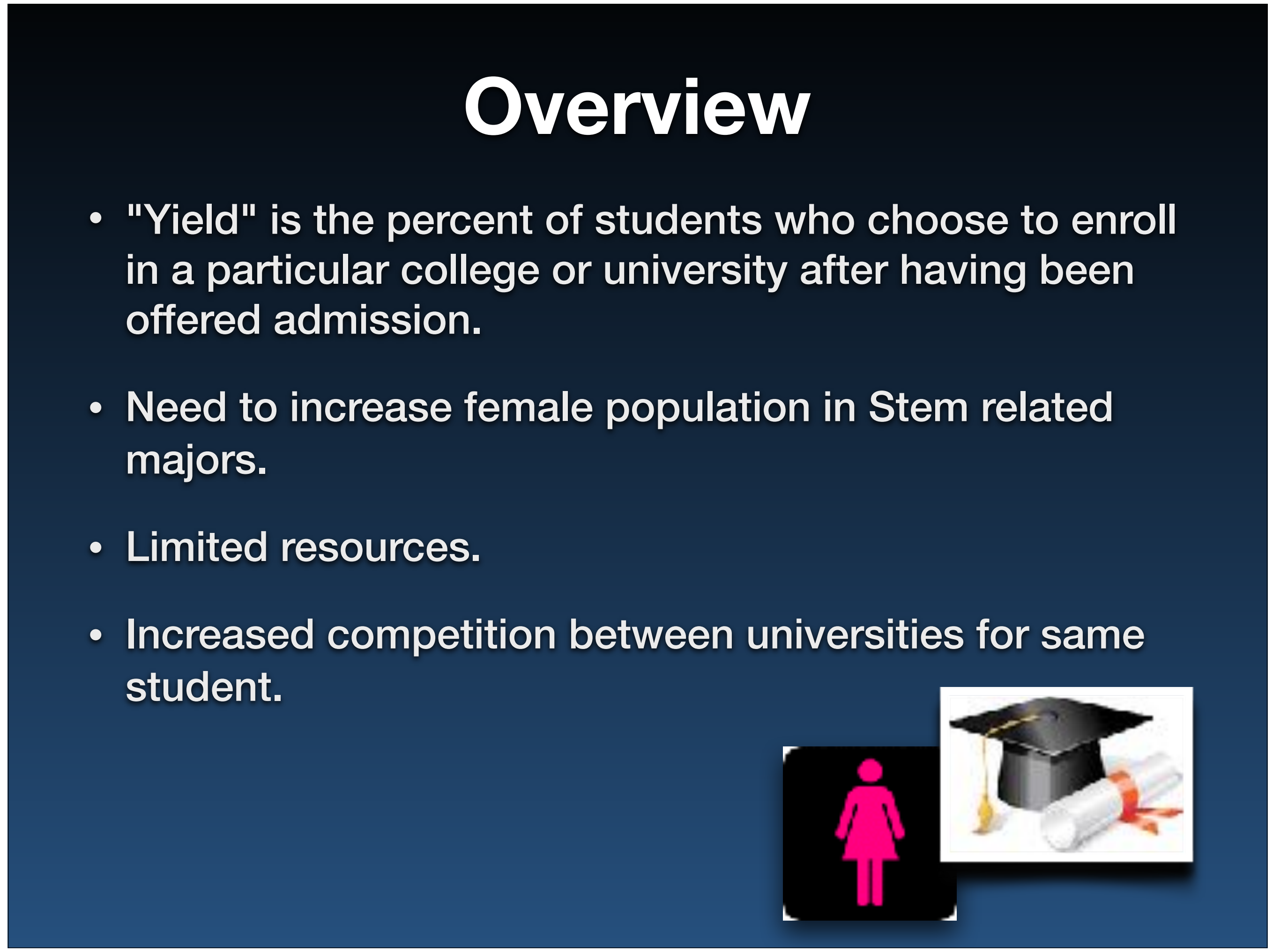
Admissions' Yield Prediction by Gender in a Colorado Higher Education Institution.

by Vanessa Gonzalez



The term "Yield" in college admissions is the percent of students who choose to enroll in a particular college or university after having been offered admission. Higher Education Institutions and Institutions with focus in STEM in particular need to have a better handle of the yield between admitted and enroll students. With the need of increasing female population and limited resources the admissions office needs to know what students have a better chance to enroll so they can invest these resources and attention to increased yield rates.

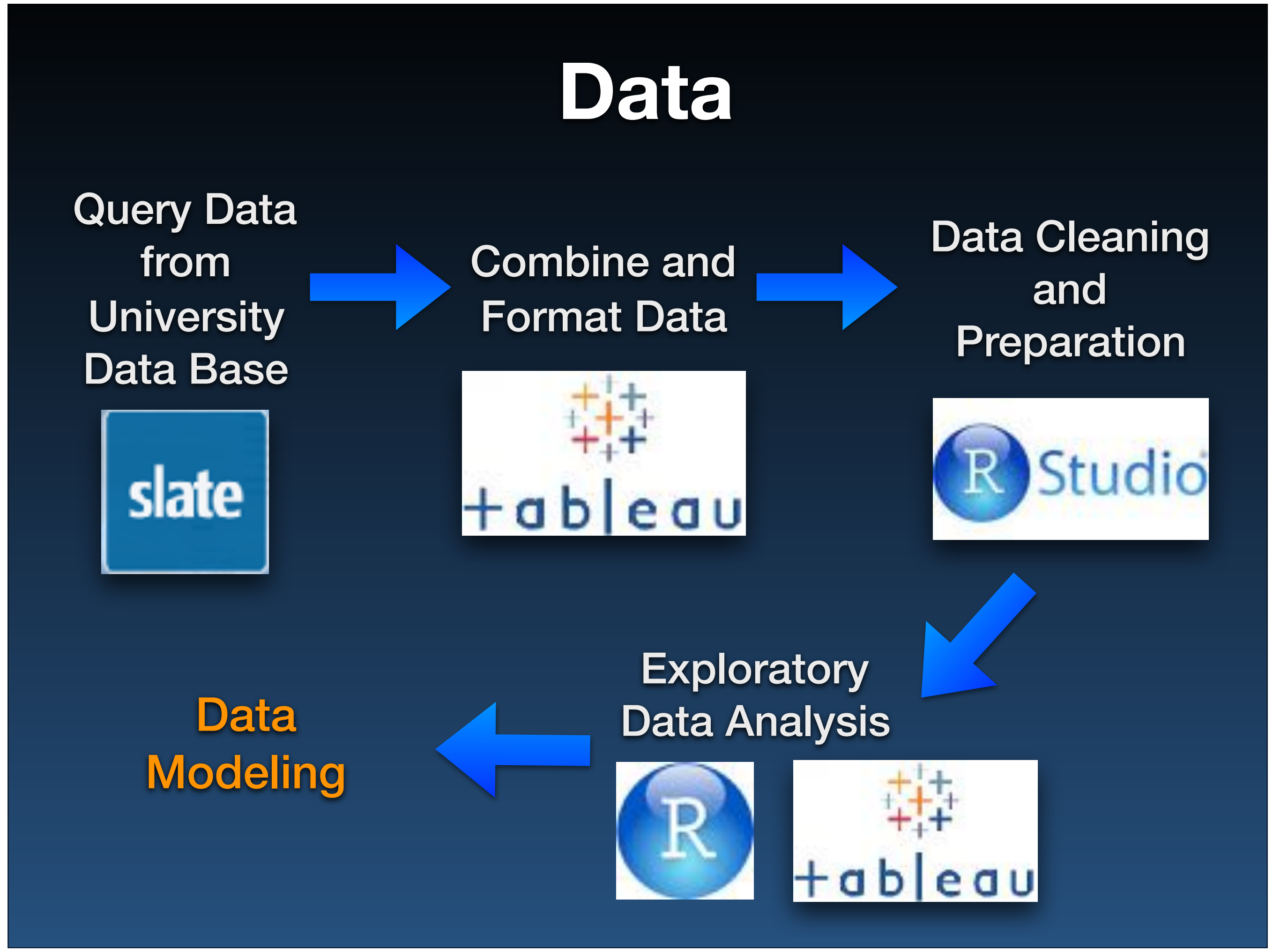
In this case we will utilize data from last year (complete cycle) and will be able to increase the data set to two years after Census of Fall 2018.

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?

With this analysis we will try to answer the following questions:
Which are the accepted students with a higher possibility to enroll?
Which are the accepted female students with a higher possibility to enroll?

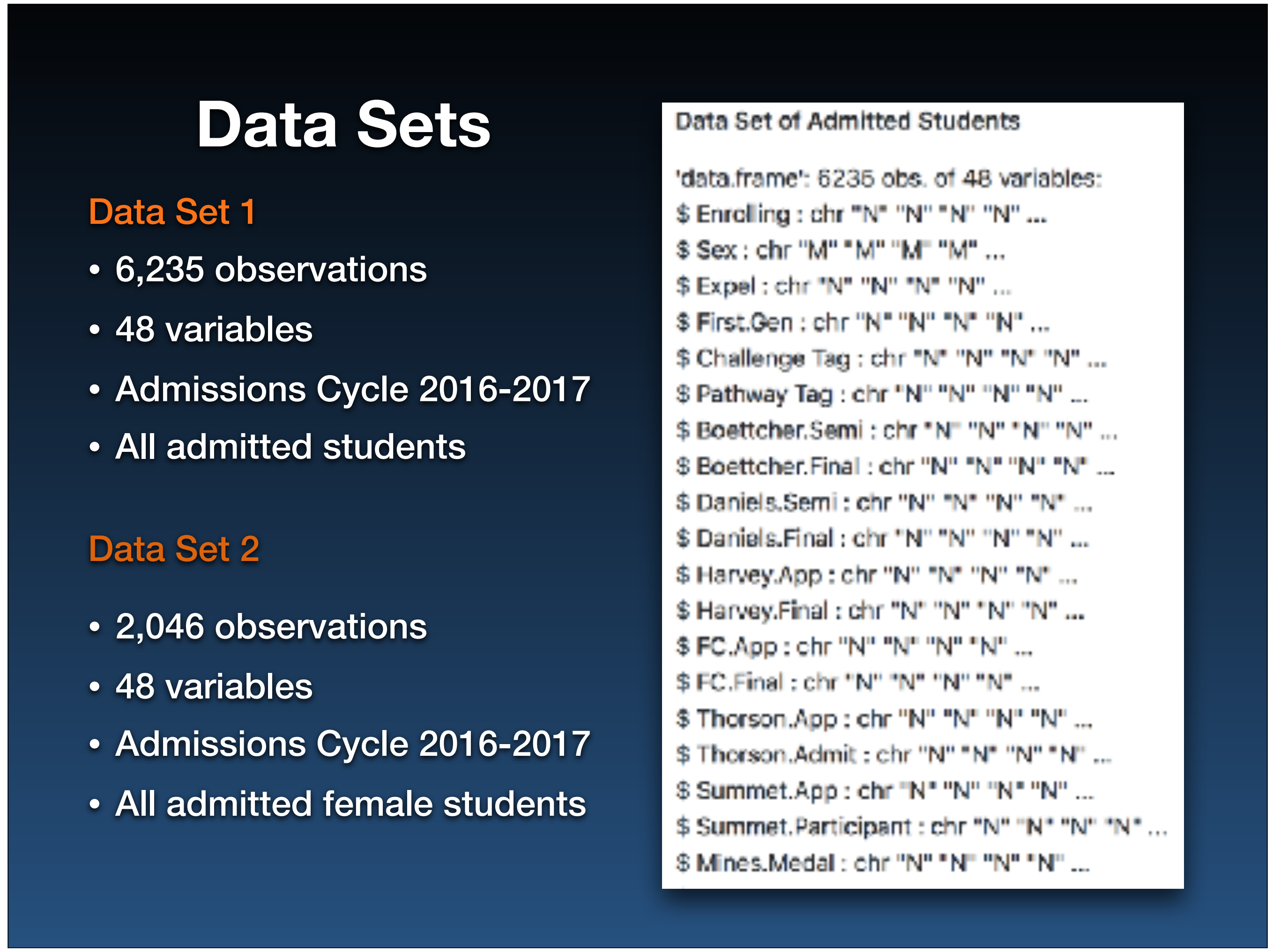


In this project several tools were used:
Slate, where a report was built to query the admissions data base and extract the needed information.
Tableau, to combine this reports, manipulate the data and export as a result two data sets to be used in our analysis. Part of EDA was done in Tableau as well.
R Sudio, where further cleaning and preparation of data sets and data subsets happened, where further exploratory data analysis was done and
Machine Learning Models were built.

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.

Several changes were made to the original data including substitution of missing values for the words “Missing”or “None” as appropriate. Other NA values were substituted utilizing kNN method. Data types were transformed to the appropriate type. Factor variables with more than 21 levels were omitted.
Dummy variables were created.
Variables other than “Enrolling” were transformed into number variables.
Numerical values were normalized.



We will be talking as Data Set 1 of the data set used to answered the first question and Data Set 2 as the Data Set used to answered the second question.

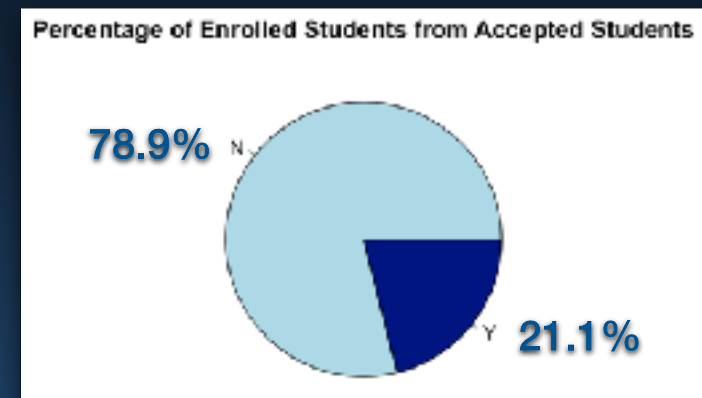
The first data set consisted of all admitted students in the admissions cycle from Fall 16 to Summer 2017. In the second data set are included all female admitted students in this same period of time.

Exploratory Analysis (EDA)

6,235 Students in the Data Set 1

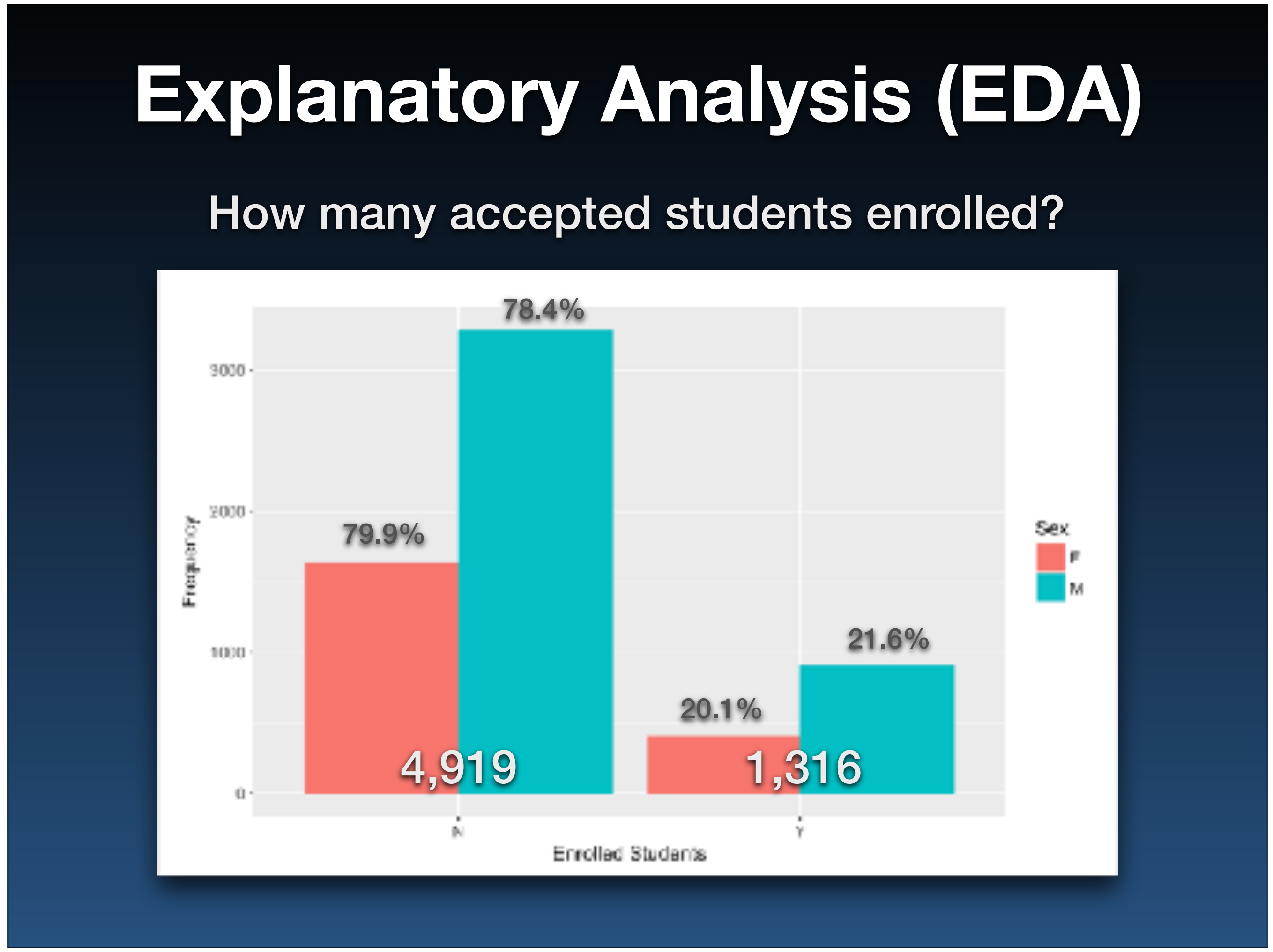
Admitted Students by Gender

	Admitted		
Sex	No	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%

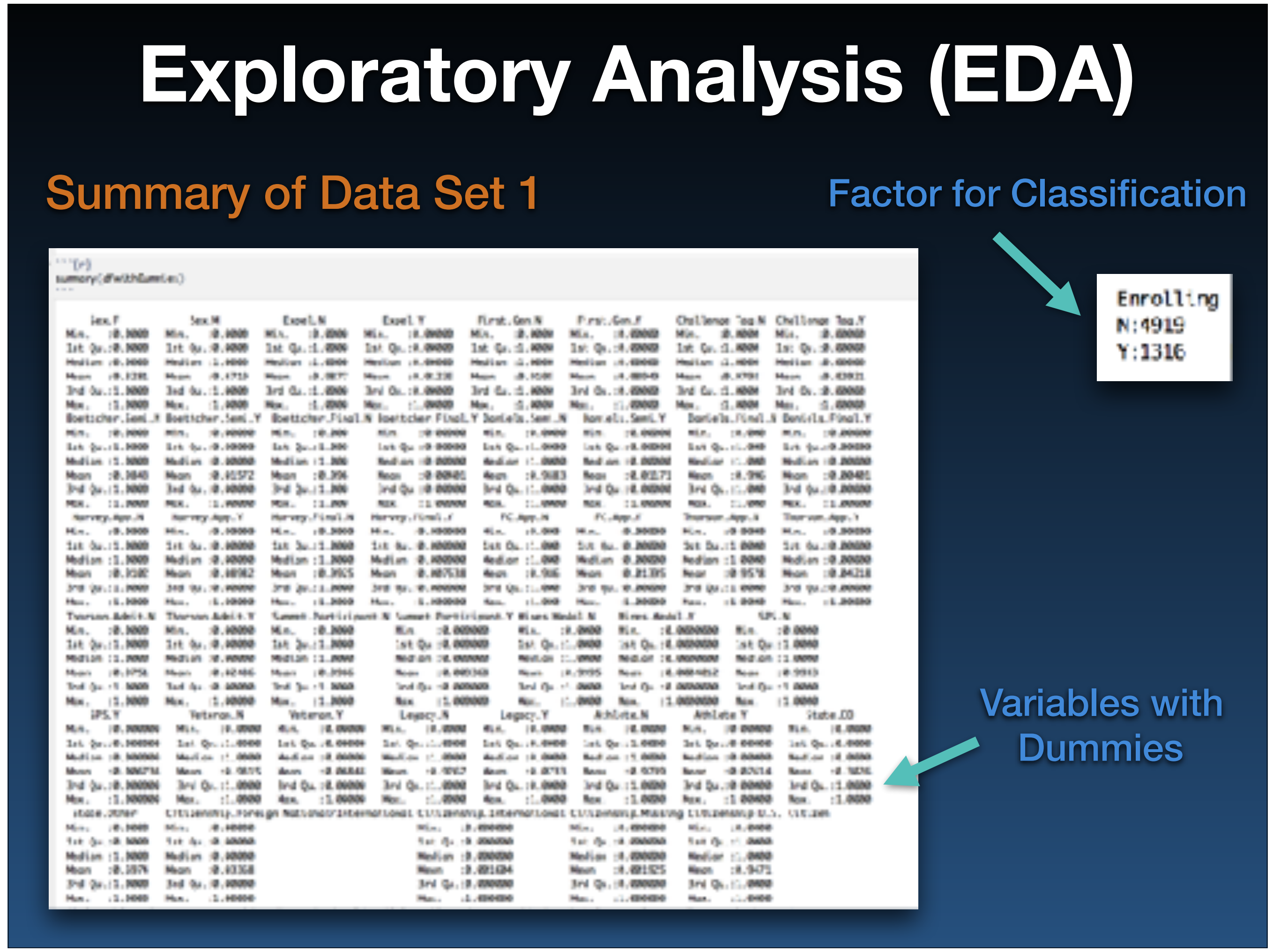


2016-2017 Cycle

The exploratory analysis was done mainly in R studio and Tableau. Of the 6,235 students considered 21.1% enrolled the institution in the 2017 2018 academic year.



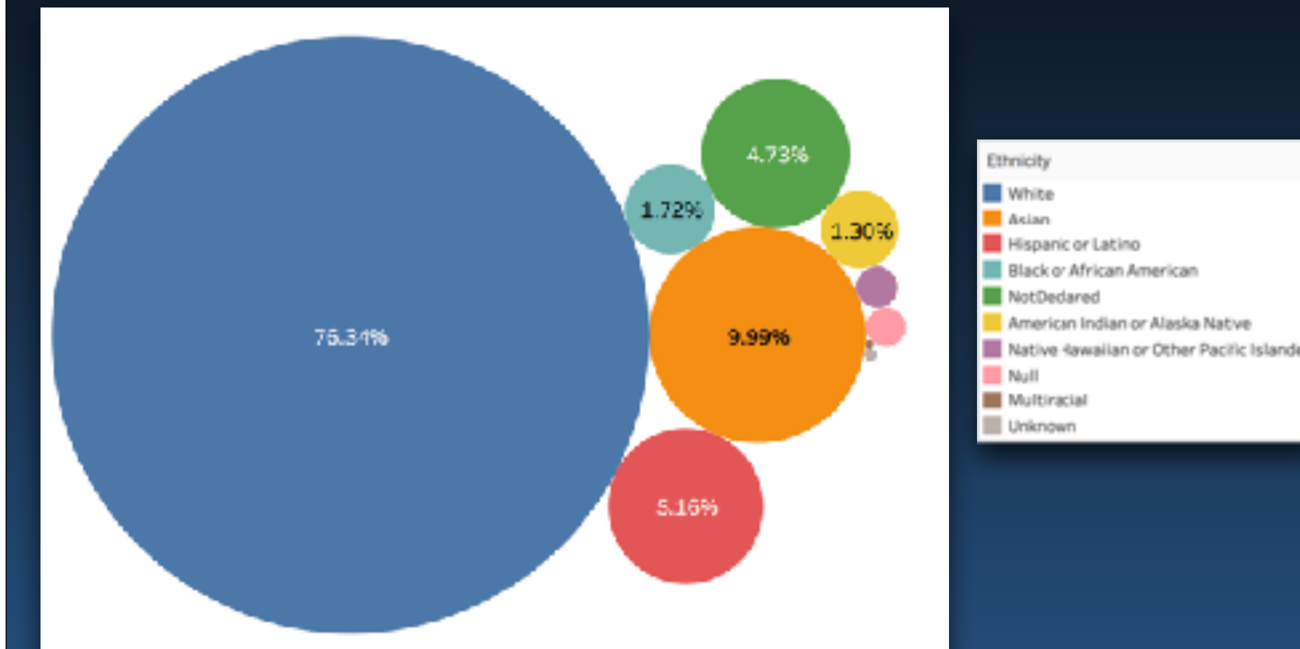
20.1% of the female students enrolled and 21.6% of the male students enrolled.



Summaries were made for both data sets and the Enrolling variable was used as a factor for classification.

Exploratory Analysis (EDA)

Ethnicity of Enrolled Students

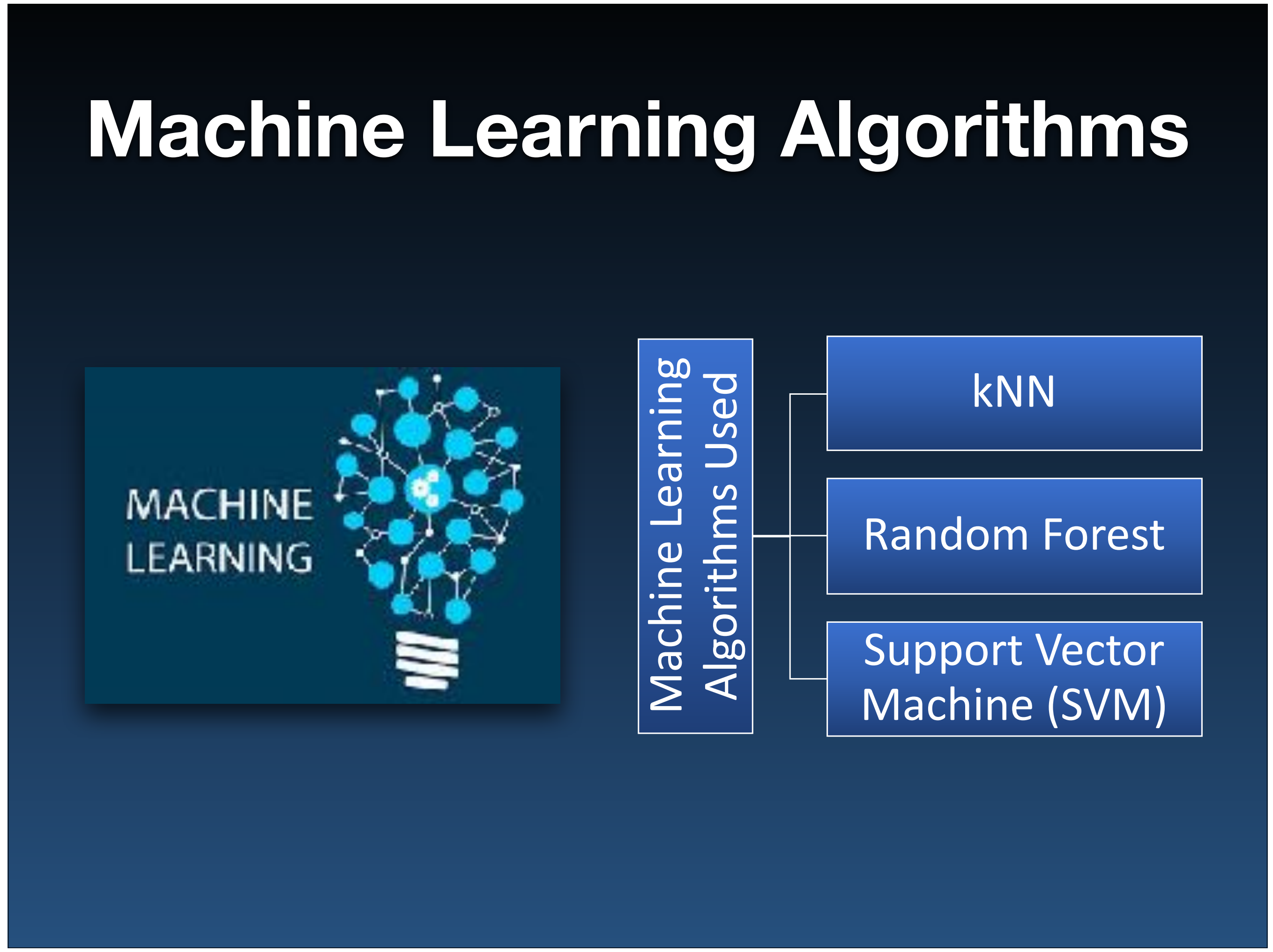


76.34% of the enrolled students where white, 9.99% Asian and 5.15% Hispanic.

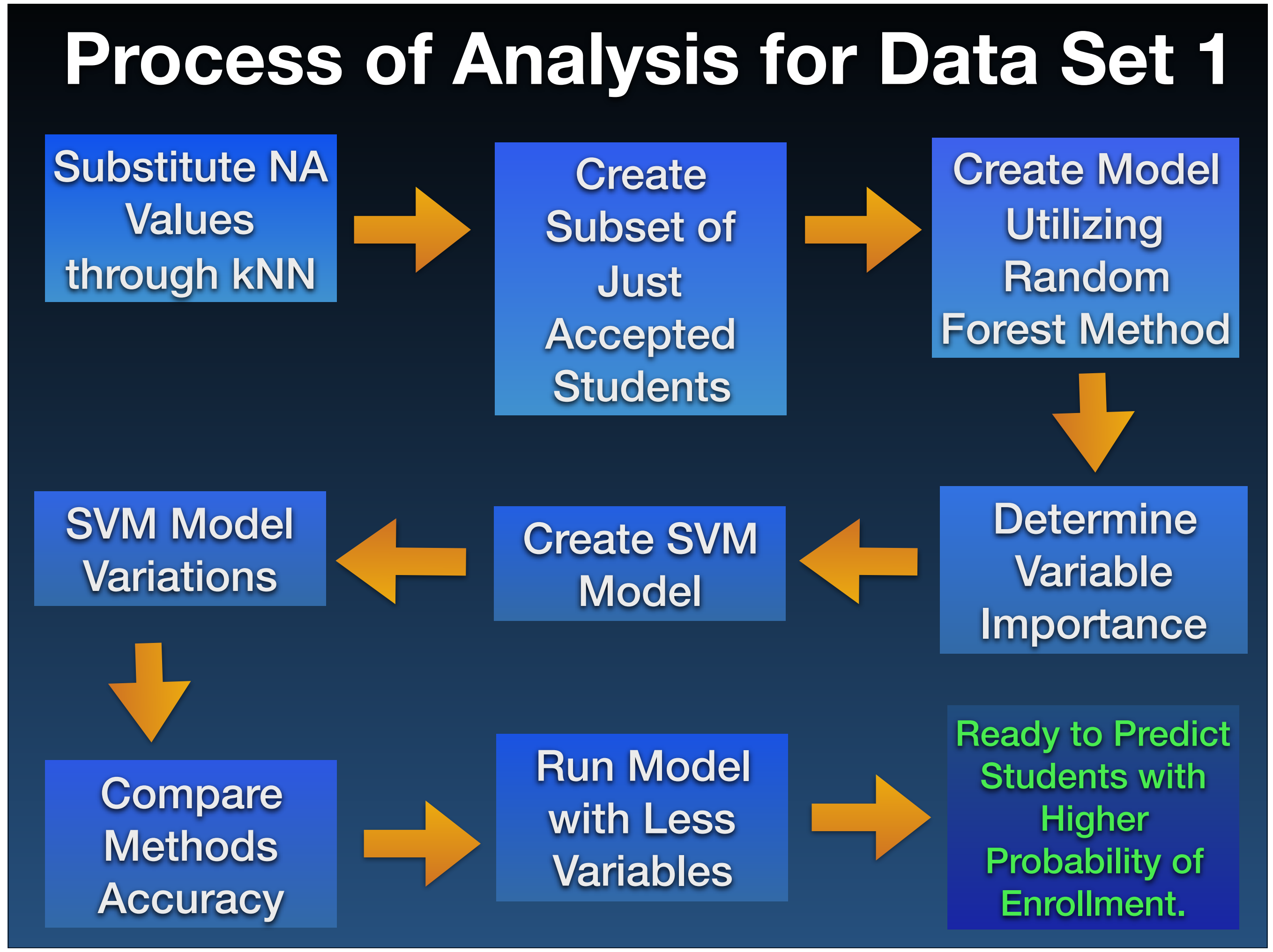
Analysis for Data Set 1

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

The first part of the analysis tried to answer our first question. Which are the accepted students with a higher possibility to enroll?



Different methods of machine learning were used.
kNN to substitute missing values.
Random Forest and Support Vector Machine for prediction,
Random Forest for Variable Importance.
Of all this methods Random Forest produced the highest accuracy and we will see the results in a few slides.

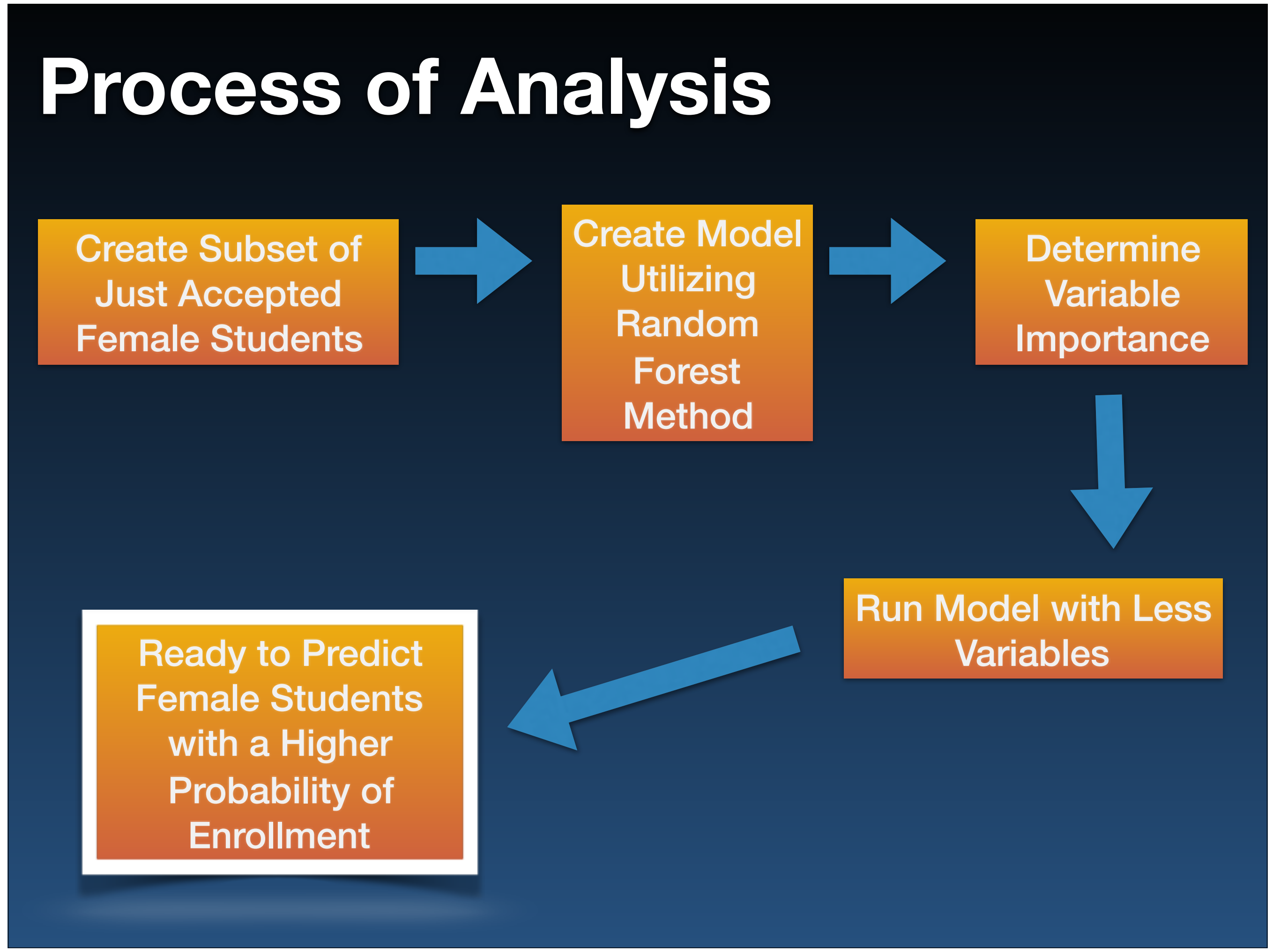


The main steps fallowed for Data Set 1 were: Substitute NA values through kNN method, Create subset for just accepted students.
A predictive model, using Random Forest algorithm was used and Importance of variables was determined. The different methods' accuracy was compared using SVM Models and a Random Forest model was created for less variables.

Analysis for Data Set 2

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

The second part of the analysis tried to answer our second question. Which are the accepted female students with a higher possibility to enroll?



For the second set, a Random Forest Model was built, variable importance determined and then a Random Forest Model was built using less variables.

Results Data Set 1

Random Forest All Variables Analysis

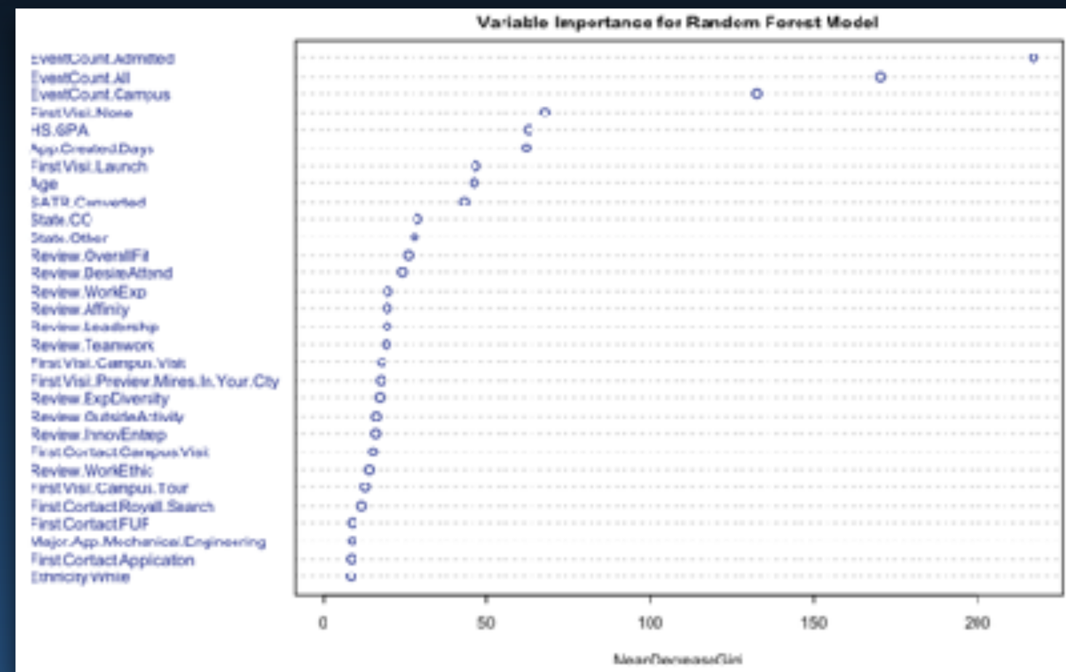
```
Confusion Matrix and Statistics


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
```

90.76% Accuracy

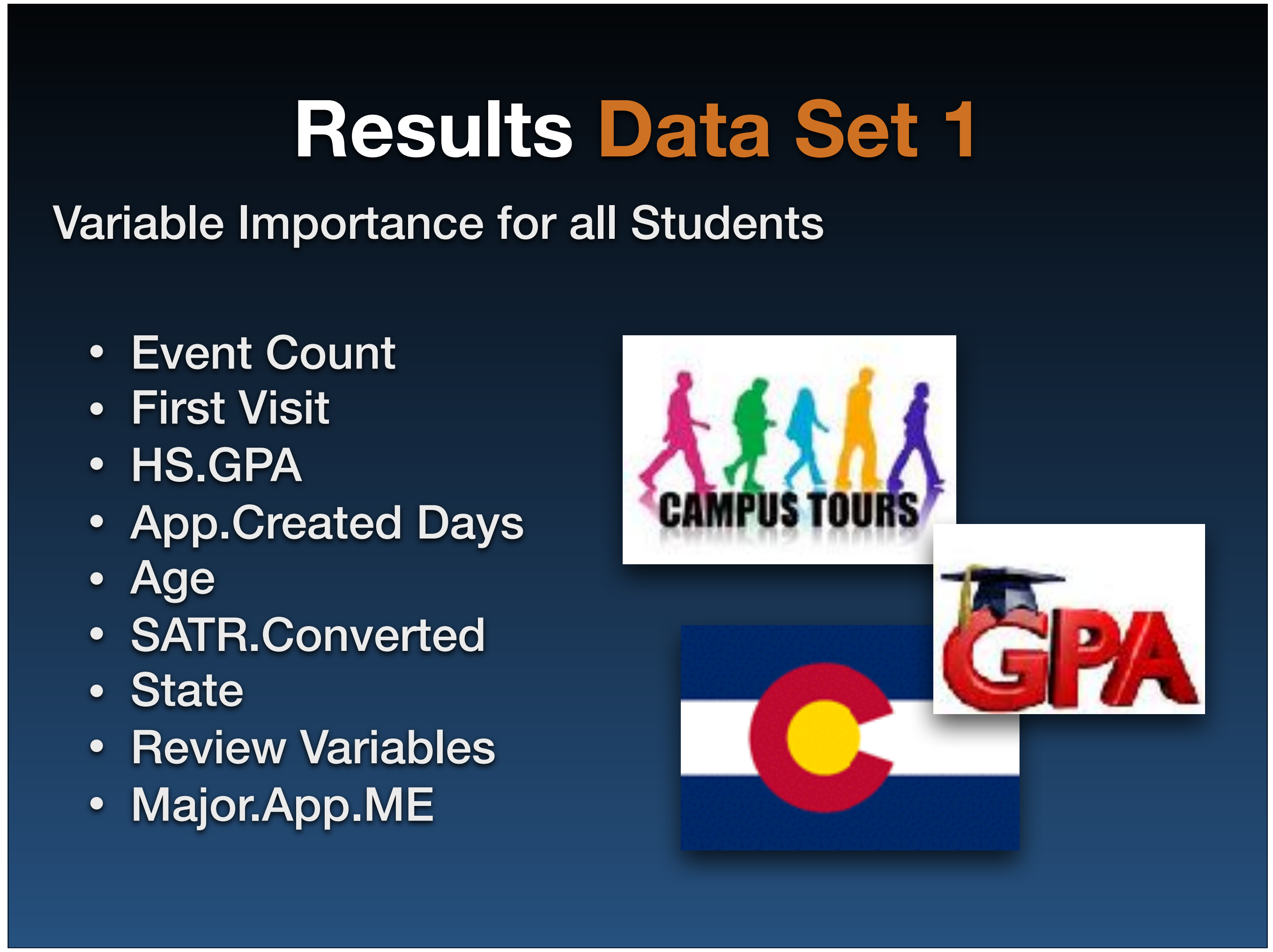
For the Random Forest Model with all variables for all admitted students the accuracy to predict of the model was of 90.76% with a kappa of .0.7069.

Results Data Set 1

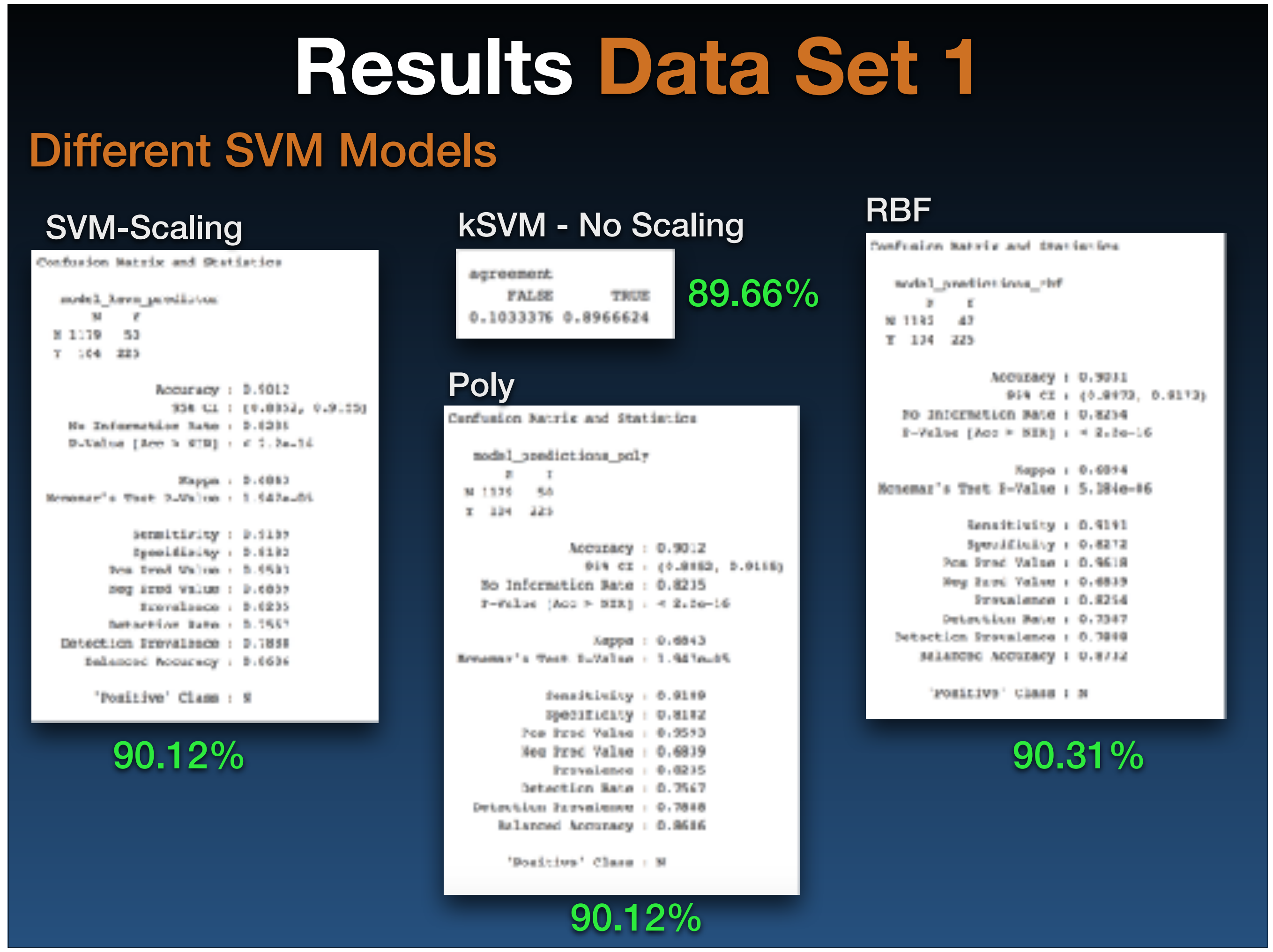
Variable Importance for all Students



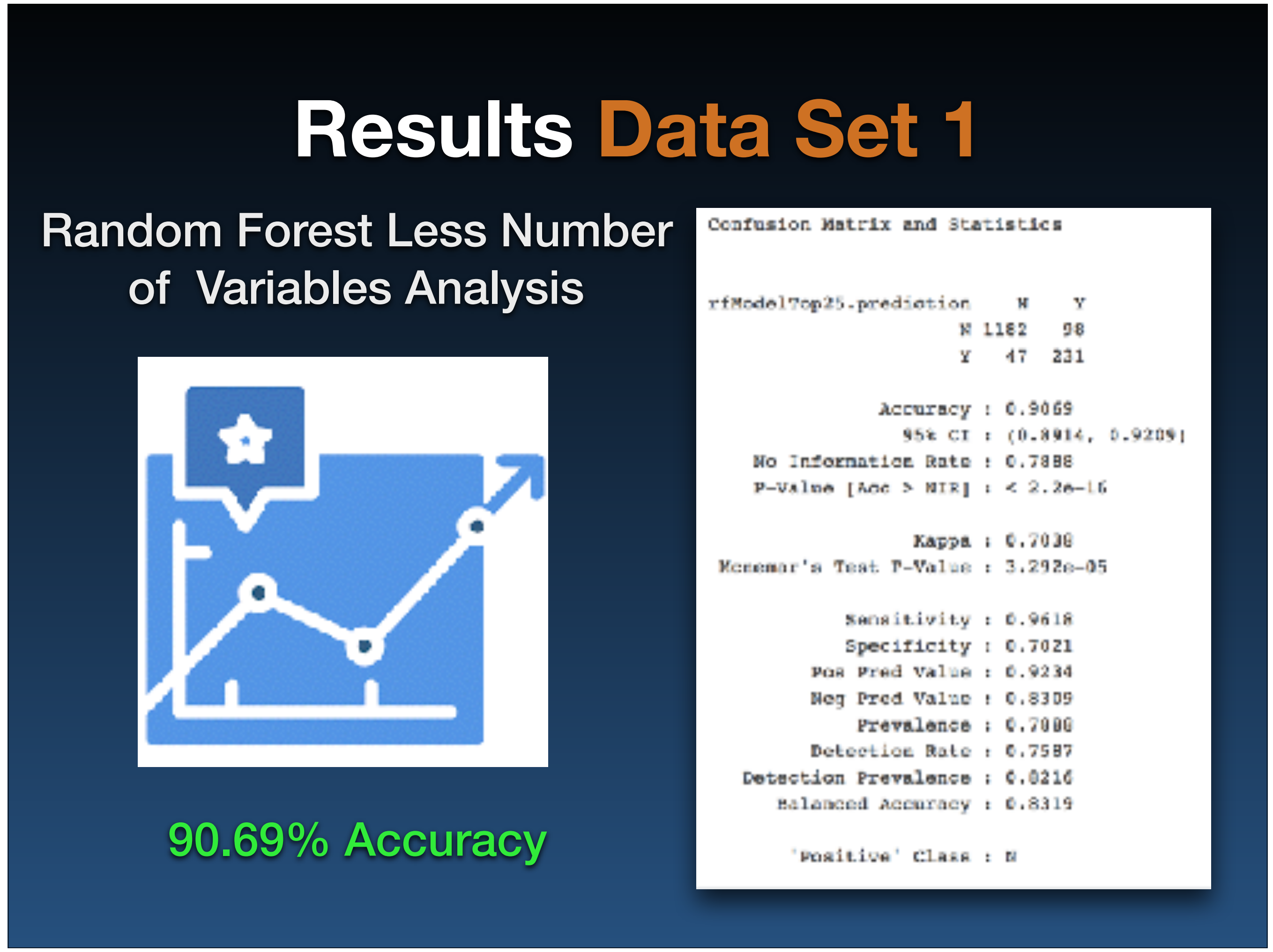
Variable importance was determined.



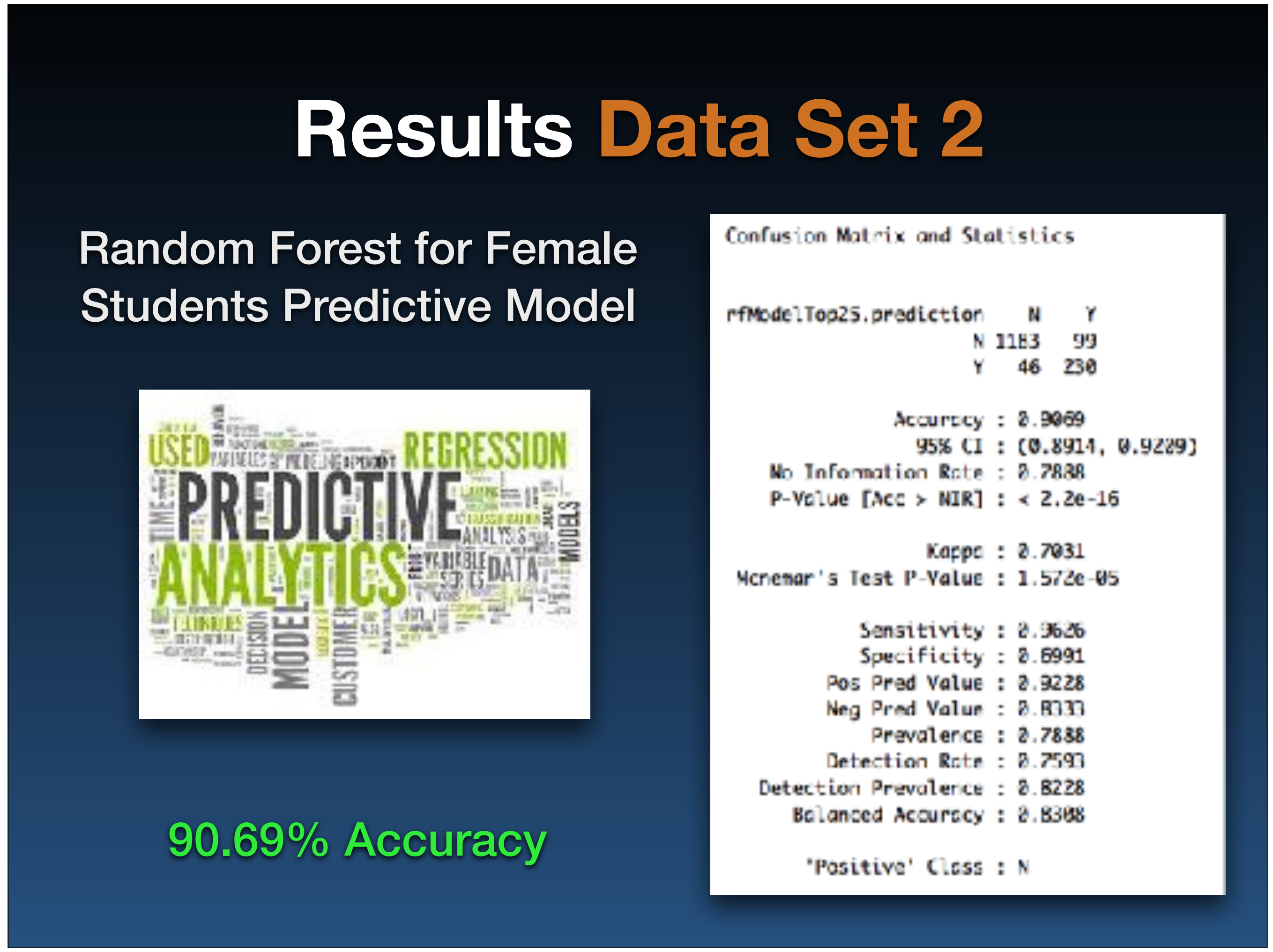
The most important variables used by the model were: Event count, first visit, high school GPA, application since created in days, age, SAT converted, state, review variables and Major.



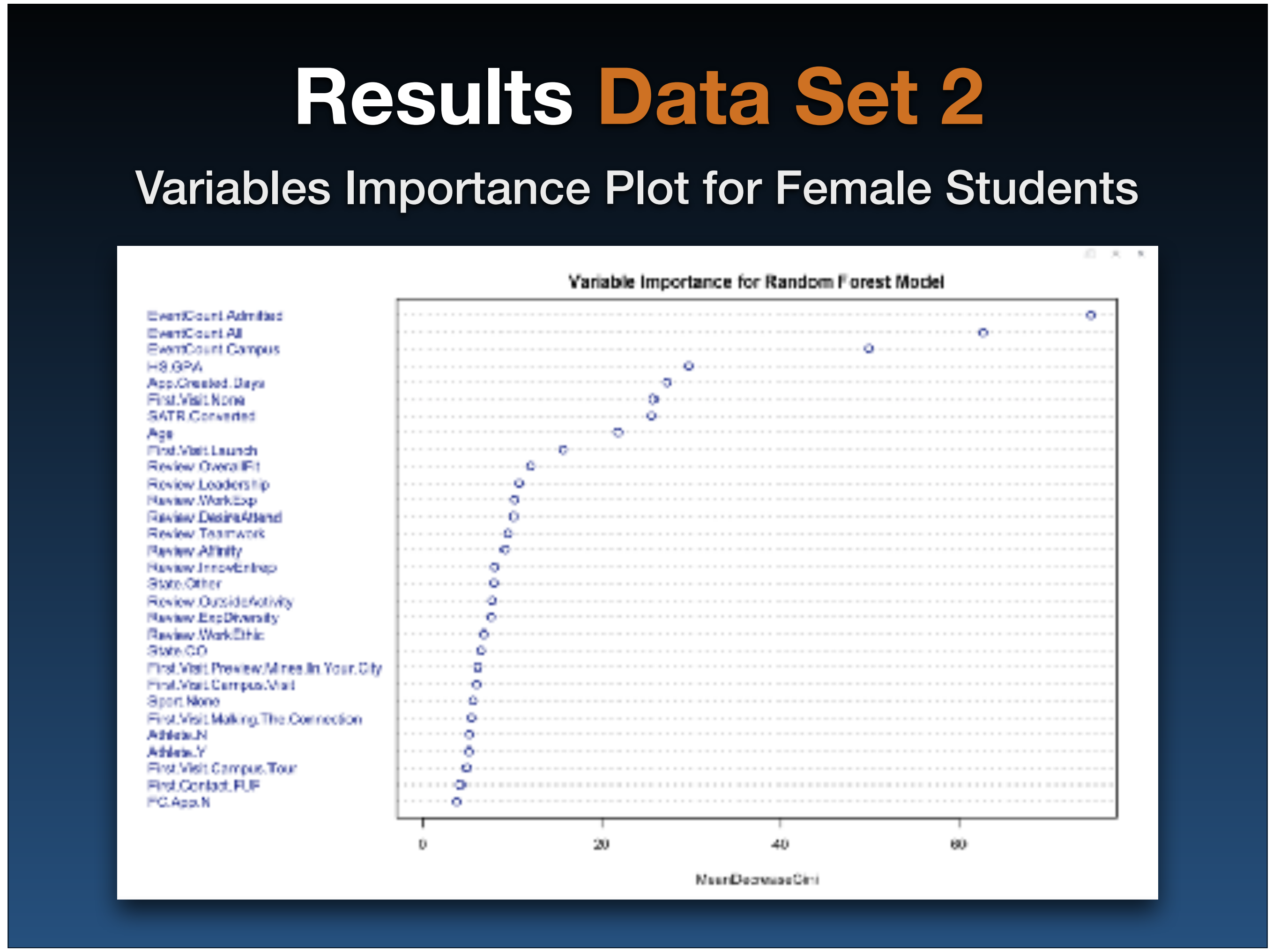
Different variations of the SVM models were done and they all had similar results.



Utilizing less variables the accuracy of the prediction for the Random Forest method was 90.69% with a Kappa value of 0.7038.



Utilizing all variables for the Female Student subgroup the accuracy of the prediction for the Random Forest method was 90.69% with a Kappa value of 0.7031.



Variable importance was determined.

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

The most important variables were similar for this group compared to all admitted students.

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

Accuracy and Kappa results were very similar for all variables, less variables and just female students subgroup.

Conclusions

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

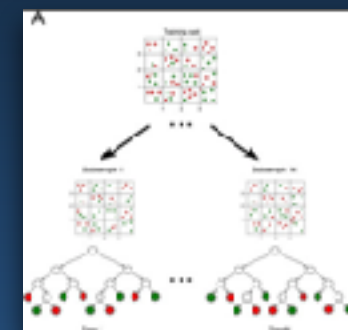
- By using the Random Forest Algorithm with all the variables we were able to predict with a 90.8% accuracy which of the accepted students had a higher probability of enrolling the fallowing year.

- 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!