

Inquiry to Applicant Predictive Models

Final documentation behind the process, creation, and optimization.

**Analytical Plan**

November 2021, Merrimack College Admissions data analysts tasked their data scientist intern, Noah Foilb, to create a successful predictive model by manipulating data mining and machine learning techniques. Through supervision and guidance, this intern created multiple successful predictive models. On November 15th, 2021, Christopher Cormio and Noah Foilb queried a Big Dataset from the inquiry pool from Fall 2021 and the current inquiry pool from Fall 2022 to predict the Fall 2022 Applicant Pool look like. By creating this model, Merrimack College could label all inquiries by the probability they will apply, let Merrimack take the initiative, and give those who are more likely to use extra initiative, hoping to make them an applicant. This model will also create a baseline applicant pool for Merrimack College, giving Merrimack College Admissions a jump over other institutes. They can now act based on applicant goals before they even apply. While this may sound skeptical, the research and validation below will support such decisions.

*Data Cleaning*

With the Fall 2021 dataset of 36,515 rows (inquires) and 50 unique features and the Fall 2022 dataset of 38,752 rows and 50 unique features, cleaning and sorting the data had the highest priority. The first step was to look at how much data we had per inquiry, which led to a drastic decline in features (now only 22). The reasoning behind this was either confounding factors, meaningless features, or too many NA’s. We looked into which features left still have NA’s and how to deal with them by removing the most insignificant parts.

* Sex not only had NA’s but is-inputted/outlying values such as O, N, and NA. By not only translating the values into numeric (M =1, F=2) but by combing the unique values with the NA’s and labeling them as 3 (other/Na), we can create a third value for Sex that has meaning without deleting meaningful rows.
* Applicant had a few errors. This feature was only supposed to be binary, yet a few students with applicant values greater than 1 were fewer than a dozen of these cases; simply filtering these out would suffice.
* Intended Major had many NA’s as well. As I believed this to be a significant feature, I translated the NA’s into “Undecided/Unknown.”
* School Address US 5 Digit ZIP Code’s NA’s were addressed by substituting NA’s with “No Recorded ZIPCode.”
* Distance from 01845 was the first feature we kept that removed significant values. There were not a lot of NA’s in this column, yet addressing these are still important. Being a numerical feature, creating a substitute value would not suffice. Noah deleted these rows to fix the data at the cost of a few thousand rows. Outliers also existed, so deleting the distances greater than 500 solved this issue.
* Age had some values that were unreasonable such as 10 – 16. Age was filtered for those values greater than 15 to be safe.

After fixing the data, more valuable data was created and transformed, such as:

* Income data per ZIP code. This was taken from the government's online census and was pulled and merged into the dataset to solve and optimize variability in future models.
* Created four equally distanced bins from Distance from 01845 and one hot encoded them.
* Merged outlying IPEDS Classifications and one hot encoded them.
* Inquiry date was very significant, yet I could not see how. I created more significant models by one-hot encoding this by its Year, year to the day value, quarter, month, day, any day of the week.

*Machine Learning Applied*

To accurately predict which inquiries will apply, the manipulation of Null, Logistic Regression, Decision Tree, Random Forest, XGBoost, Neural Network, Average Ensemble, Weighted Ensemble, and Majority Voting models were used. To validate such models, these models were validated and analyzed from k-cross-validations, confusion matrices, A/B testing, train test splits, AUC/ROC curve analysis, and null model comparisons. After this analysis is completed, the most efficient model will be optimized and micro-managed to create a long-term functional model. The biggest measurements from the confusion matrix used to compare models will be accuracy, balanced accuracy, recall, and precision.

**Initial Results**

The first model created was the Null model. As one would guess, this model is what it would predict and how accurate it would be if it were purely assuming. This model was 65.02% accurate if it guessed every applicant would not apply. We will base all models on this.

The second model created was the logistic regression model. This binomial model had an accuracy of 73.69 and a balanced accuracy of 65.48%. This makes this model better than the Null.

The third model created was the decision tree model. This model had an accuracy of 75.96% and a balanced accuracy of 70%, making this the best model so far.

The fourth model created was the Random Forest model. This model was the best so far, with a baseline accuracy of 77.22% and a balanced accuracy of 83.44%. This superior balanced accuracy may be due to its low positive predictions, yet it is viable as it will be consistent and better than the Null model.

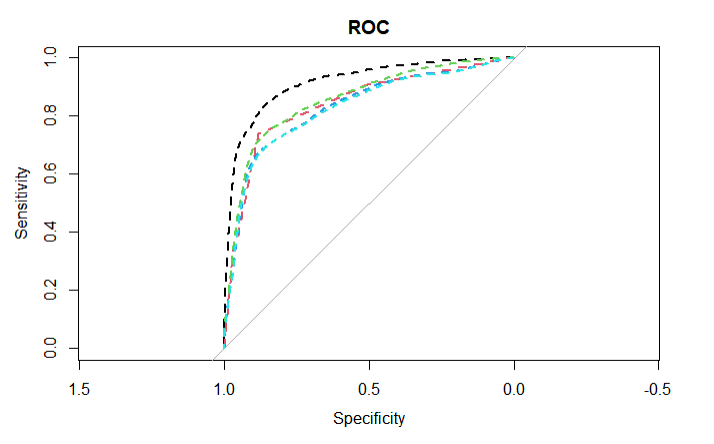
The fifth model is the XGBoost model. This model results in an accuracy of 77.2% and a balanced accuracy of 70.8%. This is not better than the Random Forest.

The sixth model was the Neural Network. This had a precision of 77.37% and a balanced accuracy of 71.77%. This was also not better than the Random Forest.

The ensemble method models were created, yet no new insight was found after analysis beyond what the Random Forest brought. Even after manipulating probabilities and a mixture of ensemble methods were used, nothing could beat the Random Forest model.

*Random Forest Analysis/Validation*

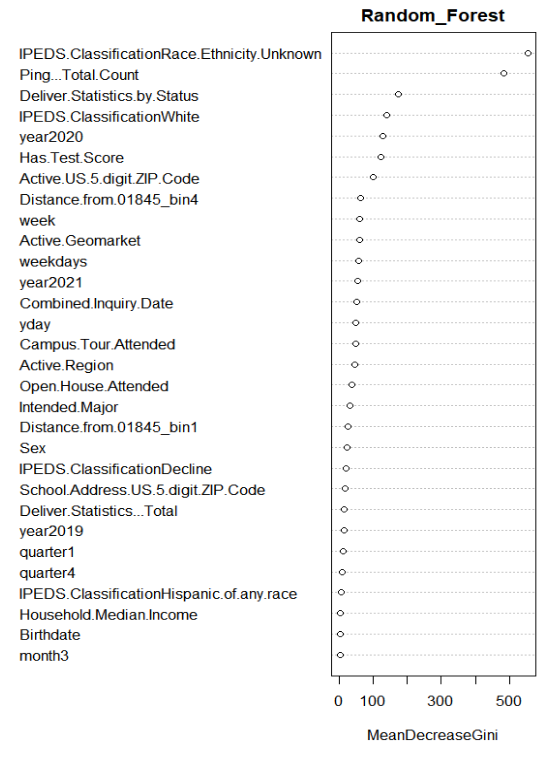
Only the Random model looked valid after training these models and testing them. Even looking at the ROC/AUC curve, this differentiation is evident.

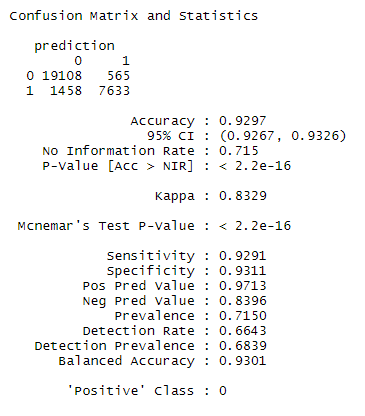


The closer the curve gets to the coordinate (1,1) in this graph, the more valid and further away from randomly guessing. These models have a lot of validity based on this graph, but the Random Forest (Blackline) is undeniably unique compared to the other models.

The final measurement used to show the model’s validity is the k-cross-validation, which coincidentally enough, according to stat.berkeley.edu, it is estimated internally. Hence, there is no need for cross-validations when using Random Forests.

**Final Results and Optimization**

 The model shown to work the best was the Random Forest model. This model had an accuracy of 77.22% and balanced accuracy of 83.44%. Yet when the trained and tested model from the Fall 2021 dataset was tested on the Fall 2022 dataset, the results were significantly more excellent. The accuracy was 92.97%, and the balanced accuracy was 93.01%! While the jump is huge, this information cannot be analyzed until the start of the Fall 2022 semester, as this data is constantly changing. Below is the confusion matrix and the importance of this model.



As we can see from the importance graph (right), the most significant predictor of applying is if their ethnicity is unknown, the Ping…Total.Count (pings received of them going on different links from MyMack), Deliver.Statistics.by.Status, race is white; they were inquired in 2020, has a test score on file. The other predictors are valid but not as significant as these.

<https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#ooberr>