**Data Mining Methods to Predict University Enrollment**

Our analysis mainly consists of five parts: data exploration, data transformation, modeling, model comparison and conclusion. We based our choice of the best model on the validation misclassification rate. We also identified the most important variables for prediction and obtained a scored dataset.

**Data exploration:**

First, we assessed the original dataset and found missing values in some variables. This is important to note because imputation may be needed for certain models to function correctly. We also analyzed the distributions of the interval variables (see appendix Figure 1). All of the interval variables are highly skewed with the exception of “satscore”. We investigated making transformations to these variables to reduce the skewness and increase normality.

**Data imputation and transformation:**

In the process of imputation, we replaced missing class variables values with the mode and missing interval variables with mean. We used the transformation “max normal” option to assess the appropriate transformations for our skewed variables. Log transformations were used for self initiated contacts, total contacts and average income. The distance variable was square rooted and initial span received a power function. These transformations resulted in both skewness and kurtosis decreasing (See appendix Figure 2). The other variables were not transformed because no transformation made their distributions significantly better. This also aids in interpretation of models if these variables are used. All of the transformed distributions can be found in the appendix in Figure 3.

**Modeling:**

In order to better assess our model development, we partitioned that data into 70% selected to be training data and 30% as validation data. From this point forward, we considered models such as regression, decision trees, and neural networks to predict a prospective student’s likelihood for enrollment. Each model was then compared based on the misclassification rate of the validation dataset.

Decision trees are not that sensitive to missing values and skewness, therefore we added a decision tree node to the partitioned data directly. We experimented with restricting the tree size and the types of variables entered, such as including interactions. The best tree that was obtained is shown in Figure 4 in the appendix. The misclassification rates for the training and validation data are shown in Table 1. These numbers are low meaning this model could be beneficial.

|  |  |  |
| --- | --- | --- |
| *Table 1* | | |
| Data | Train | Validation |
| Misclassification Rate | 0.042936 | 0.043226 |

Due to the target ‘Enroll’ being a binary variable, we explored logistic regression models with both backward and forward selection methods. Since cases with missing values get ignored in regression models we used the imputed dataset to fit regression model. We obtained the misclassification rates shown in Table 2.

|  |  |  |  |
| --- | --- | --- | --- |
| *Table 2* | | | |
| Selection Method | Data | Train | Validation |
| Forward | Misclassification Rate | 0.113019 | 0.104516 |
| Backward | Misclassification Rate | 0.057064 | 0.069032 |

We also explored the effect of the skewness of variables on the performance our model. We fit the model using the transformed dataset described earlier. The misclassification rates as shown in Table 3. Both rates are greatly reduced which means that skewness is affecting the performance of our logistic regression model.

|  |  |  |  |
| --- | --- | --- | --- |
| *Table 3* | | | |
| Selection Method | Data | Train | Validation |
| Forward | Misclassification Rate | 0.101385 | 0.092258 |
| Backward | Misclassification Rate | 0.060942 | 0.072258 |

Neural networks were also considered. We considered using all the variables in the model as well as selecting variables via regression.

Similar to regression, neural networks require a complete record for prediction and scoring. Neural networks also function best with transformed and imputed data so the modifications made for logistic regression were used for our neural network model. There were 3 hidden layers considered. The misclassification rates for training and validation data are shown in Table 4.

|  |  |  |
| --- | --- | --- |
| *Table 4* | | |
| Data | Train | Validation |
| Misclassification Rate | 0.055956 | 0.207097 |

With the previous model, the train misclassification rate is low but the validation rate is significantly high. This could be due to over fitting. We continued by using a neural network based on variables selected by regression. The misclassification rate of the validation dataset is significantly reduced from the previous model, as shown in Table 5.

|  |  |  |
| --- | --- | --- |
| *Table 5* | | |
| Data | Train | Validation |
| Misclassification Rate | 0.095014 | 0.082581 |

**Model Comparison**

After fitting these different models, we used the model comparison node to compare both the training and validation misclassification rates. From the results given in Table 6, we can conclude that decision tree performs best. The results of the model comparison node came to the same conclusion and can be found in Figure 8 in the appendix. Therefore, we used this model to make predictions on our scoring dataset. We obtained the scored dataset which was submitted with this paper.

|  |  |  |
| --- | --- | --- |
| *Table 6* | | |
| Model Misclassification Rate | Train | Validation |
| Decision Tree | 0.042936 | 0.043226 |
| Regression with Non-transformed data –Forward | 0.113019 | 0.104516 |
| Regression with Non-transformed data –Backward | 0.057064 | 0.069032 |
| Regression with Transformed data - Forward | 0.101385 | 0.092258 |
| Regression with Transformed data - Backward | 0.060942 | 0.072258 |
| Neural Network | 0.055956 | 0.207097 |
| Neural Network with variables selected | 0.095014 | 0.082581 |

**Conclusion**

In summary, the decision tree model is the best predictive model to identify prospective students who would most likely enroll as new freshmen. Based on the model, we concluded that the variables: self\_init\_cntcts, stuemail, hscrat, satscore, telecq and init\_span are important predictors for students’ enrollment as shown in Figure 5. In the original scoring dataset, 3.15% of students enrolled. After making predictions with our chosen model on this dataset, we predicted about 11% of the observations as possibly enrolling (see Appendix Figure 6). A copy of our diagram is provided in the appendix, Figure 7.

We can also get some insights on how to achieve the administrations’ goals. Based on the important predictor variables we found in the decision tree model, we can distinguish prospective students from those that may not enroll mainly by their self-initiated contact. Using this segmentation, the university could further gather information from these prospective students to obtain information on pre-existing trends in the other business goals areas, diversity and gender. The administration may also concentrate more advertising to them, in order to increase new freshmen enrollment. With the same level of these predictor variables, the administration can pay less attention to other requirements for students.

In addition, our model results in a majority of students enrolling as ethnicity=C and sex=1. The university will be able to enroll students of diversified background if we continue to gather information on those in our prospective student segment that do not fall in those two categories. A survey could prove useful as means to obtain what these students value most in their university. Once such information is gathered, the marketing department could target these areas to increase OSU’s appeal to more diverse applicants.

Similarly, for students in the prospective category can be targeted by their SAT scores. The administration can set a higher score as the baseline to receive further marketing. These marketing methods can then be modified appeal to what students with a higher SAT find most desirable in a university. The diversity and increased SAT score categories could be combined into one group in order to decrease marketing expenses and increase overall effectiveness.

**Appendix:**

  
  




  
  




