TITLE HERE PLEASE

Hannah Gautreau

Methods

Selection of Class Variable

There were a number of candidate variables to use as a proxy for "success" in the first work term. The main candidates were when the student got a job, and their work term evaluation.

Figure **??** shows the distribution of evaluations. The vast majority of the students have received evaluations of excellent or outstanding, and there was no correlation that emerged between the work term evaluation and any feature in the feature list. Due to the fact that the evaluations are close to random, and heavily biased toward excellent and outstanding, they were removed from consideration in the model.

Figure 2 shows the distribution of when first work term students were matched with a job. This is a much more interesting class variable because there are significantly fewer students who were matched in the first round of the job search. In addition to this, there more correlation with getting a job in the first round and other features in the model.

Feature Selection

The features with the best predictive value were selected using the Recursive Feature Elimination (RFE) Algorithm. This algorithm recursively removes features, builds a model using the remaining features, and calculates the accuracy of the model. The result of this is the combination of attributes that are the most useful to predicting the target variable (Bakharia). All features were used in the algorithm.

To determine which features should be used in the model, numerous iterations of the RFE algorithm were run to determine the optimal number of features. To determine the accuracy of the resulting model, a simple Logistic Regression was run with 5-fold cross validation. Figure shows that the optimal number of features used is 9. Here is the ranked list of features chosen by the RFE Algorithm:

1. HS_JOB

Copyright © 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

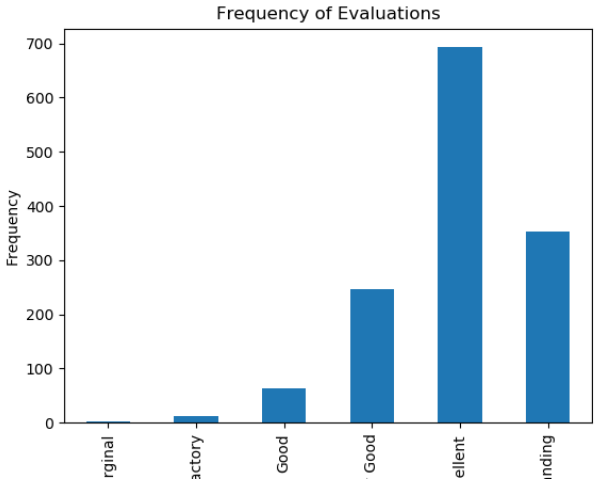

Figure 1: Frequency of First Work Term Evaluations

2. PROGRAMMING
3. EA_COUNT
4. COM_EA
5. DRA_EA
6. MUS_EA
7. OTH_EA
8. PRJ_EA
9. SOC_EA

Algorithm Selection and Evaluation

The following 6 classification algorithms were evaluated as candidates for the final model.

1. Logistic Regression (LR)
2. Linear Discriminant Analysis (LDA)
3. K Nearest Neighbours Classifier (KNN)
4. Decision Tree Classifier (DT)
5. Gaussian Naive Bayes (NB)
6. Support Vector Machine (SVM)

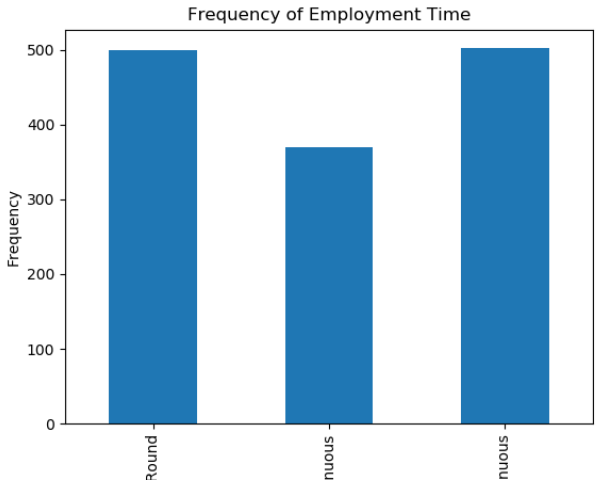

Figure 2: Frequency of First Work Term Employment Timing

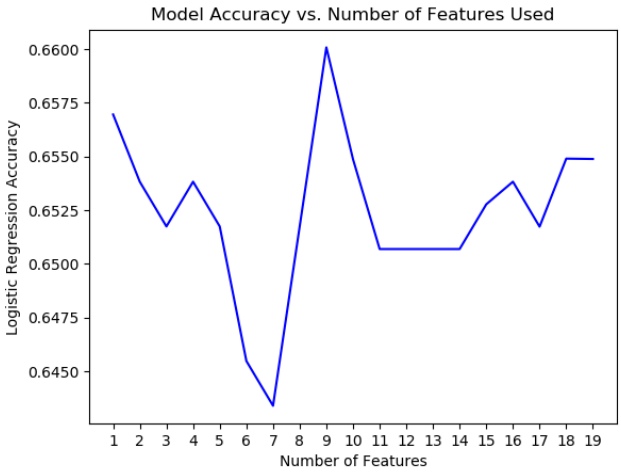

Figure 3: Model Accuracy vs. Number of Features Used

All algorithms were evaluated on their overall accuracy using 10-fold cross validation on the features listed in the previous section.

Results

Table 1 shows the results of this analysis.

Algorithm	Accuracy	Standard Deviation
LR	0.660534	0.048550
LDA	0.666897	0.054624
KNN	0.604904	0.049695
DTC	0.579349	0.042431
NB	0.633086	0.056725
SVM	0.663244	0.045802

Table 1: Algorithm Evaluation Results

Figure 4 shows a box and whisker plot of the evaluation to give more insight into the results. This plot shows tat there is too much overlap in the LR, LDA, NB, and SVM algorithms to make a definitive decision about which algorithm produced the best model.

The next four classifiers were evaluated using 20% of the dataset

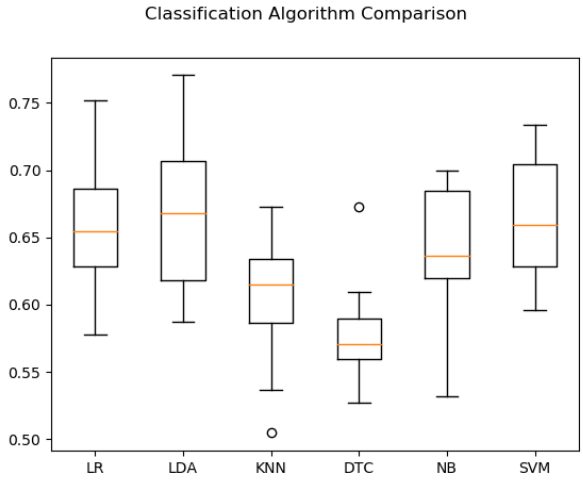

Figure 4: Algorithm Comparison Results

Logistic Regression

Class	Precision	Recall	F1-Score	Support
!FR	0.65	0.87	0.75	171
FR	0.53	0.24	0.33	104
total	0.61	0.63	0.59	275

Table 2: Logistic Regression Evaluation Metrics

Linear Discriminant Analysis

Class	Precision	Recall	F1-Score	Support
!FR	0.66	0.86	0.75	171
FR	0.55	0.28	0.37	104
total	0.62	0.64	0.60	275

Table 3: Linear Discriminant Analysis Evaluation Metrics

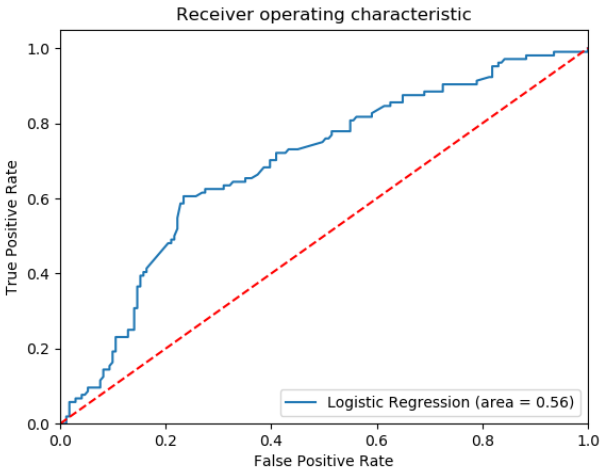

Figure 5: ROC - Logistic Regression

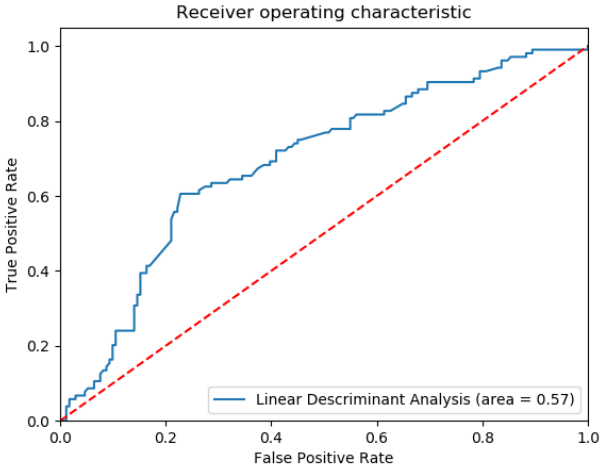

Figure 6: ROC - Linear Discriminant Analysis

Class	Precision	Recall	F1-Score	Support
!FR	0.69	0.80	0.74	171
FR	0.55	0.39	0.46	104
total	0.63	0.65	0.63	275

Table 4: Naive Bayes Evaluation Metrics

Class	Precision	Recall	F1-Score	Support
!FR	0.65	0.90	0.76	171
FR	0.56	0.21	0.31	104
total	0.62	0.64	0.59	275

Table 5: Support Vector Machine Evaluation Metrics

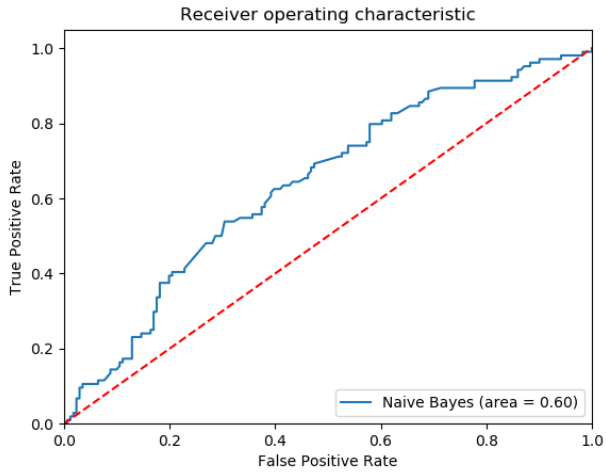

Figure 7: ROC - Naive Bayes

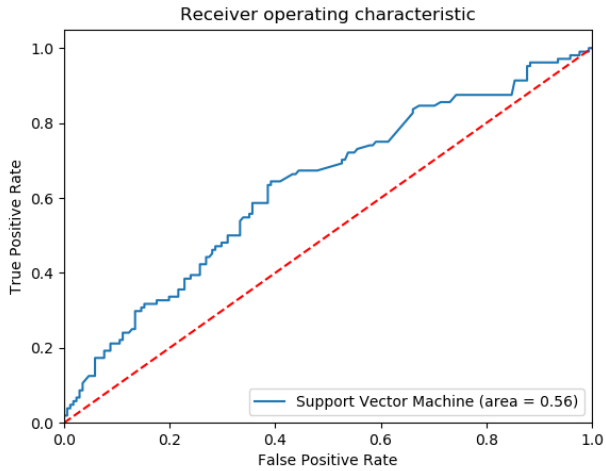

Figure 8: ROC - Support Vector Machine

Naive Bayes

Support Vector Machine

References

Bakharia, A. Recursive feature elimination with scikit learn. https://medium.com/@aneesha/recursive-feature-elimination-with-scikit-learn-3a2cbdf2 Accessed: 2017-11-20.