Using Twitter and Other Recruit Data to Predict the Probability of Going to Michigan

Wolverine Sports Analytics

Overview

Using Twitter data and recruit data, models were created to project the probability of a recruit going to Michigan. Models predicted with over 90% accuracy whether a recruit who has or has not taken an official visit to Michigan whether they will attend Michigan or not for the 2017 Recruiting Season after being offered; however, these models had a bias of classifying recruits to not go to Michigan. Using only recruits who have taken an official visit, models predicted with over 50% accuracy whether a recruit will go to Michigan or not and were unbiased in their classification.

Methodology

Data was gathered for all 2016 and 2017 football recruits that were offered by Michigan. Using manual data entry and 247 sports, data was gathered on the distance from their high school to Ann Arbor, and whether Michigan was their first offer, last offer, whether they took an official visit, if the official visit they took was their last official visit, and whether they were an in-state recruit. Furthermore, using Twitter, tweets were processed for analysis until their commitment date. Metrics were gathered on the ratio of Michigan tweets to overall tweets, Michigan retweets to overall tweets, the average number of favorites recruits received on Michigan tweets compared to the average number of favorites recruits received on all tweets, and the average number of retweets recruits received on Michigan tweets compared to the average number of retweets recruits received on all tweets. For recruits whose twitter data was not available (account was set private, do not have twitter, did not tweet before commitment date, or deleted tweets), those recruits were not evaluated.

With these metrics as inputs, machine learning models ran to predict the probability that a recruit would attend Michigan. Furthermore, a second run of the model was done to predict recruits who had visited Michigan on an official visit. The types of machine learning models used were a Logistic Regression Model with an L1 and L2 Normalization penalty, a Random Forest Model, and an Extreme Gradient Boosted Model. 2016 Recruit data was used as a training set and 2017 Recruit data was used to evaluate the model. Cross validation was done to tune the model parameters and to evaluate whether the model held up across all data so that it can be applied to future and current recruiting cycles.

Results

The model ran for two types of datasets: recruits who have and have not taken an official visit to Michigan and recruits who have only taken an official visit to Michigan.

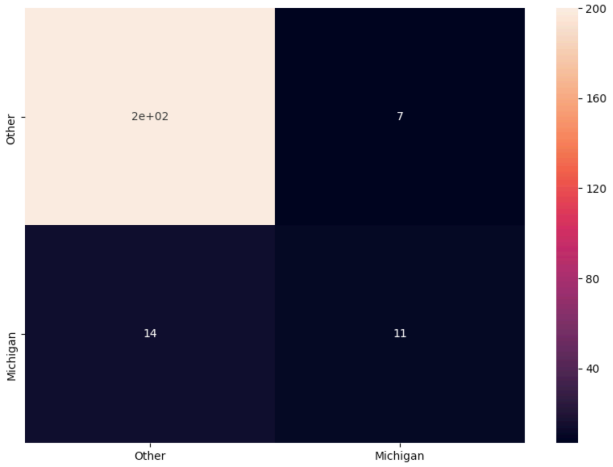
Recruits Who Have and Have Not Taken an Official Visit to Michigan

For recruits who have and have not taken an Official Visit to Michigan, our model predicted with over 90% accuracy whether a recruit will go to Michigan or not go to Michigan; however, the model had a strong bias towards classifying recruits not to go to Michigan who ended up going.

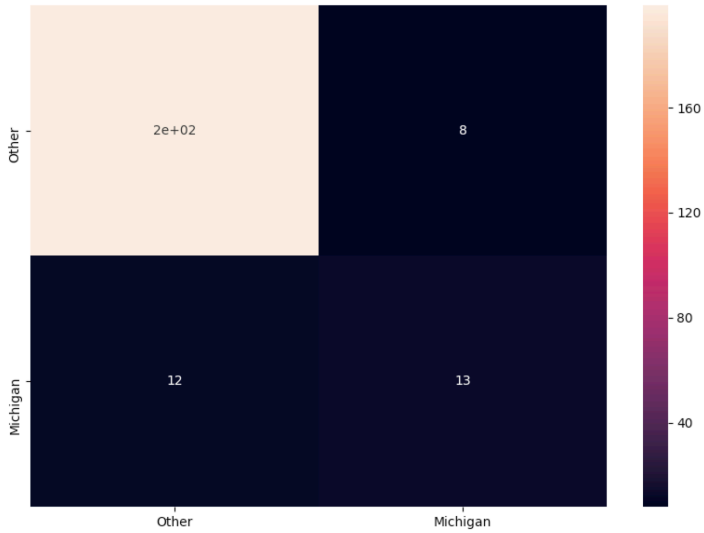
Model	Accuracy
L1 Log Regression	90.95%
L2 Log Regression	91.38%
Random Forest	90.05%

Below are confusion matrixes for the respective models (see Appendix I for the definition of a confusion matrix)

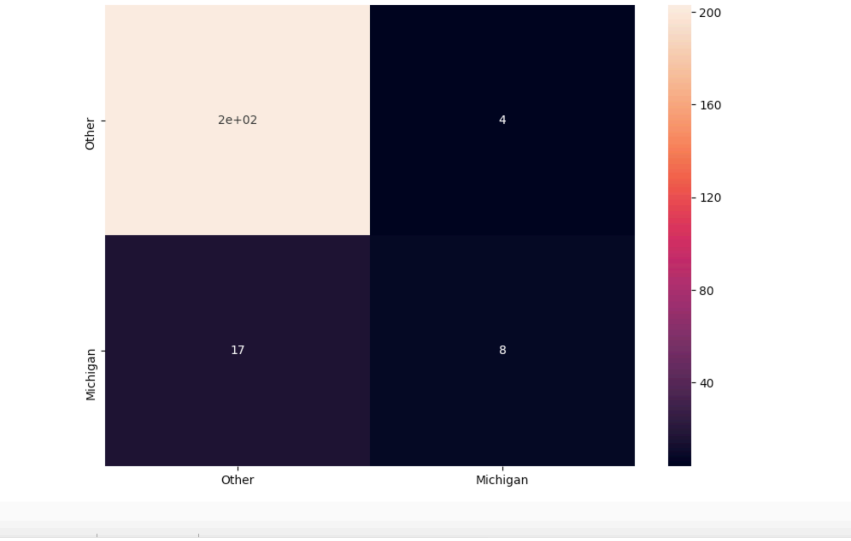
L1 Logistic Regression Confusion Matrix



L2 Logistic Regression Confusion Matrix

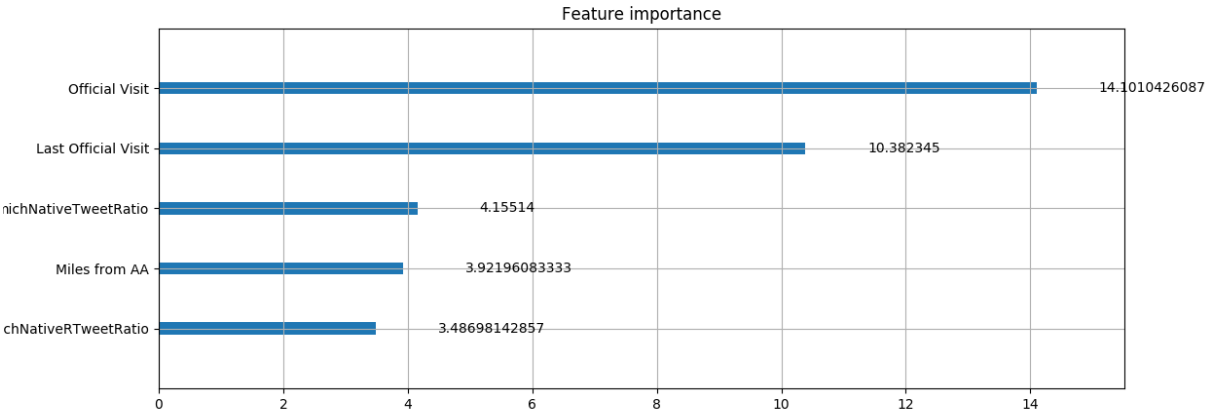


Random Forest Confusion Matrix



By viewing the confusion matrixes for the above models, the L2 Logistic Regression Model performed the best when classifying recruits who actually attended the University of Michigan. This model had an accuracy of 52% for classifying recruits who attended Michigan. Furthermore, these models are skewed to classify recruits to not attend Michigan because of the vast amount of recruits Michigan offered who did not attend Michigan.

Also, below is a graph depicting the information gain each feature provides the model:

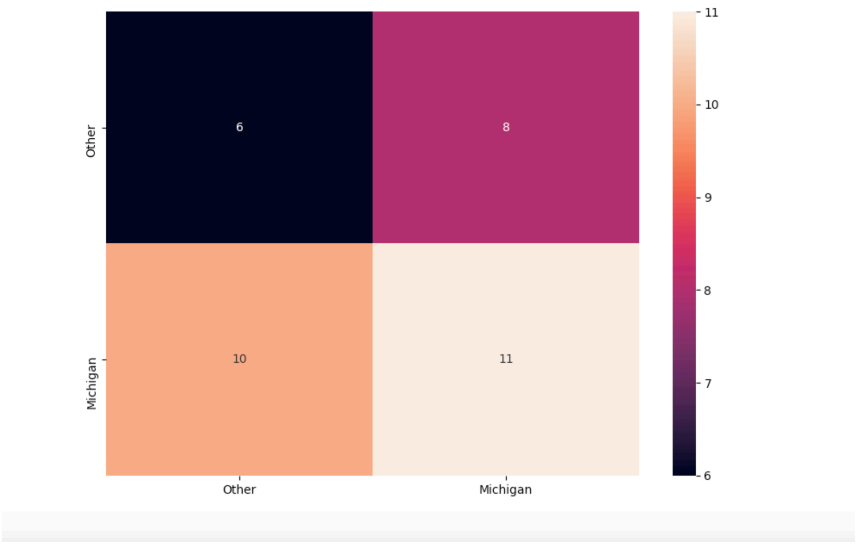


Recruits Who Have Taken an Official Visit to Michigan

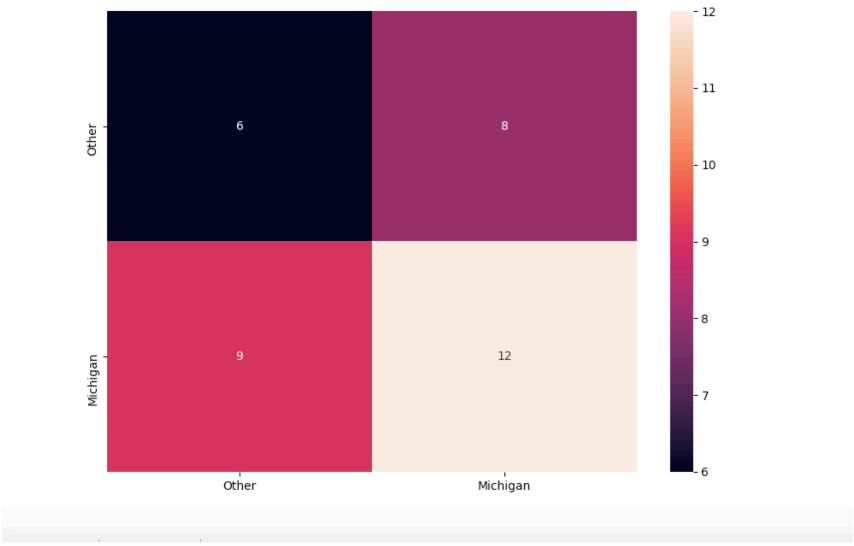
For recruits who have taken an official visit, models using the same features except for whether they have taken an official visit or not predicted with 54% accuracy that a recruit will go to Michigan or not.

Model	Accuracy
L1 Logistic Regression	48.57%
L2 Logistic Regression	51.43%
Random Forest	54.28%
Extreme Gradient Boosting	45.71%

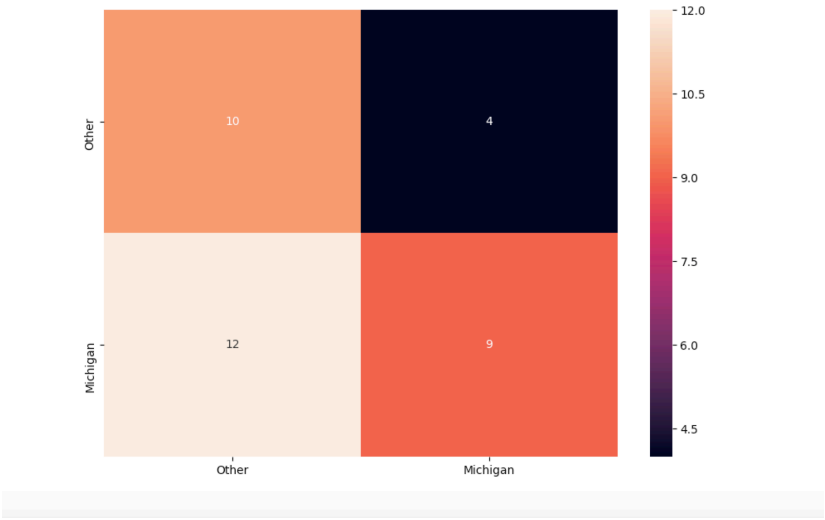
L1 Logistic Regression Official Visit Confusion Matrix



L2 Logistic Regression Official Visit Confusion Matrix

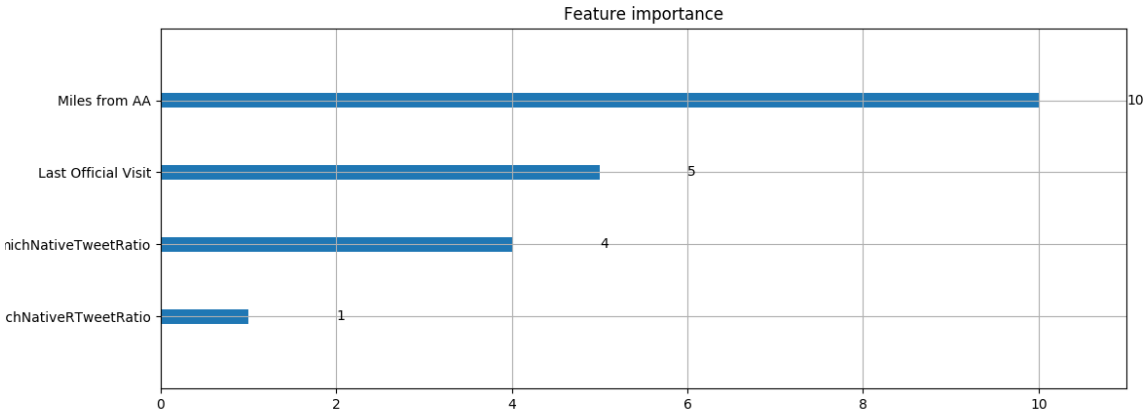


Random Forest Official Visit Confusion Matrix



For predicting recruits who actually went to Michigan, the L2 Logistic Regression Model performed the best, however, overall, the Random Forest model classified recruits the best. By eliminating the official visit feature, the model put more emphasis on twitter data and other recruit factors. Furthermore, there is little bias in classifying recruits to not commit to Michigan or to commit to Michigan unlike using the previous dataset.

Below is a graph depicting the weights assigned to each feature for recruits who only took an official visit provides the model:



Conclusion

Using Twitter data can be insightful to predicting whether recruits go to Michigan or not, especially comparing the amount of times they tweet or retweet about Michigan compared to all their other tweets. By using twitter data along with other recruit data, a general probability can be formed whether a recruit will go to Michigan or not. In order to future expand upon this model and to improve model accuracy, especially for recruits who end up going to Michigan, more features should be examined. These features include the amount of people at their position currently at Michigan, the amount of people that go to Michigan from their surrounding area, and their recruit ranking. Also, Natural Language Processing can be implemented to get a recruits' sentiment about Michigan compared to other schools. As social media becomes more pervasive in teenagers lives, using Twitter data and other social media data will become key in predicting whether a recruit will go to Michigan. This will help football programs allocate more money and resources to target recruits who have a higher probability of going to Michigan and allow them to get better recruiting classes.

Appendix

I: Confusion Matrix

Below is an example of a confusion matrix

n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	