A Supervised Learning Approach to Predicting Academic Performance Based on Self-Esteem & Social Acceptance

Naitik Poddar & Tony Sabella

Mar 24, 2024

Introduction:

Can we predict the type of grades 13-17 year old students will earn based on self-esteem & social acceptance? We are working as a teenage psychologist and want to understand the relationship between a student's perception of their own worth and their academic performance. We predict grades based on self-esteem & social acceptance and other control covariates like age, race and gender. The reason we are predicting grades based on self esteem & social acceptance is to identify potential correlations or patterns that could help in developing interventions or support systems to improve academic outcomes, enhance student well-being, and promote overall success for future students.

Data:

Our primary dataset for this analysis is "Social Media, Social Life: How American Teens View Their Digital Lives, United States, 2012 (ICPSR 37960)" dataset. This dataset is survey data that provides a comprehensive view of teenagers' social media habits alongside demographic and academic performance indicators. Our population of interest are US teens between 13-17 years of age. U.S. teens between 13-17 years old is the only population that these estimates will be accurate for.

Our original purpose for this dataset was to focus on social media use in teens and how that relates to the types of grades earned, however once we removed all variables with more than 10 missing values the variables that related to social media usage were also removed. Instead of

focusing on survey questions related to social media usage this survey also asked the teenagers questions related to self-esteem & social acceptance which had many observations. We changed gears and used questions related to self esteem & social acceptance to help predict what type of grades students earn.

The variables in the data that will be relevant for predicting teenagers' grades based on self esteem are Q28, Q27_A-Q27_M, and other demographic control covariates. The outcome variable of interest is Q28, representing the kind of grades students usually get in school. This variable categorizes academic performance from "1" for mostly A's to "7" for mostly D's or lower, with an additional categories "8" for students in schools that do not use traditional grades and “-1” for students who refused to answer the question. We omitted these values from the predictions. Figure 1 in the appendix shows the distribution of responses for Q28. Notice that a majority of the respondents chose “1” and “2” indicating that the majority of the students taking the survey do well in school and only few teens who are taking the survey do poorly. This would mean that this sample is not very representative of students who do poorly in school and more representative of the successful students. Another possible reason for the skewed distribution is Social desirability bias where students overestimate their grades because they feel pressure to present themselves in a positive light or avoid appearing unsuccessful. This bias can lead to inflated estimates or inaccurate perceptions of certain variables, such as academic performance. These problems presented above show some limitations of survey data.

To capture self-esteem, we utilized variables that represent how each student feels about themselves. In the survey of 13-17 year olds under self-esteem questions it said, “How well do each of the following statements describe you”, and students' choices were categorized from “1”

for “A lot like me” to “4” for “Not at all like me”. A small fraction of students refused to answer the question and we omitted these responses from the estimates.

This would include variables such as:

- Q27_A: I have a lot of friends
- Q27_B: I am lonely
- Q27_C: Compared to other people my age I feel normal
- Q27_D: I often feel rejected by people my age
- Q27_E: I get along well with my parents
- Q27_F: I get in trouble a lot
- Q27_G: There are lots of things I can do well
- Q27_H: I like myself
- Q27_I: I am happy with my life
- Q27_J: I often feel sad or depressed
- Q27_K: I’m outgoing
- Q27_L: I’m shy
- Q27_M: I find it easy to make friends

Methodology:

We used three different methods to make predictions on this dataset lasso, ridge and random forest. We started the analysis with Lasso regression by splitting the data with 80% representing the training set and 20% representing the test set. The glmnetUtils library is helpful in generating this regression. LASSO stands for Least Absolute Shrinkage and Selection Operator and it works by adding a penalty term to the standard OLS regression that is proportional to the sum of the absolute values of the coefficients. This addition of the penalty

term helps minimize the deviance of the regression. Lasso regulation selects betas from variables that are least associated with the outcome and shrinks them to 0 in order to minimize the residual sum of squares. The lasso helps find an ideal number of betas to have a good balance variance tradeoff. The lasso regularization technique is more useful in high dimensional data where the number of variables far exceeds the number of observations. In this project we are working with low-dimensional data but are still trying to demonstrate the lasso method.

Using the same split of 80% of the data representing the training data and 20% of the data representing the test data we ran a regression using ridge regularization which has a similar style to the lasso. Ridge regularization works by adding a penalty term that is proportional to the sum of the squares of the coefficients rather than absolute value like in lasso. Since the ridge model is adding a penalty that is squared it can only shrink betas close to zero but not all the way to zero.

Random forests is the third technique used in this analysis and it consists of decision trees, hence the name forest. Decision trees include two types of nodes: decision and leaf nodes. The decision nodes contain conditions that split the data. Leaf nodes are the endpoints of the decision tree branches and they represent the final outcome or prediction of the tree.The CART algorithm summarizes the decision tree procedure where first choose a leaf in the tree, Then determine the optimal split X_{IJ} across all observations i and covariates j that that minimizes the IS deviance computed using only observations in the leaf. Then this procedure is repeated for a new leaf in the tree, and continues recursively. The final step of creating a decision tree is to determine where to stop. Finding a good number of bins to stop at helps control the complexity of the estimator and the bias variance tradeoff. Given a decision tree with the right number of nodes and stopping criteria this can predict new data and run it through the tree to compute the

fraction of observations in each bin that belong to each category to get a predicted probability for each category.

The CART algorithm uses a greedy approach, meaning it chooses the lowest-deviance split from the beginning of the tree rather than choosing a high-deviance split now which could lead to even better lower deviance splits later. As a result, the tree may not always find the best possible tree structure because CARTs tend to have high variance. A popular solution that seems to work very well in practice is to modify CARTs using bagging. This is where random forest comes into play.

Random forests use bootstrap aggregation of the same decision tree with n observations and averages them together to help improve the estimate of the decision tree. CARTs tend to have high variance. It is important to pick m = square root d at random which helps decorrelate the trees and helps with overfitting. Figure 3 shows OOS deviance for the random forest model used for this data set at different levels of m and when m = 3 it seems to generate the lowest OOS deviance. Each decision tree in the forest provides a numerical prediction. In regressions, the average of the predictions from all trees is taken as the final prediction. We used cross-validation to compare our OOS deviance between lasso and random forest.

Main Results:

Capturing the relationship between self esteem/ social acceptance and academic ability was the focus of this project. We used three different methods to evaluate this relationship. We used the Lasso, Ridge and Random Forest methods for generating a model that can predict out of sample. The Lasso model generated an in-sample R^2 of 0.2284 and an out-of-sample R^2 of 0.0476. This means that our model tells us about 23% of the variance in academic performance is explained by our self-esteem/social acceptance regressors in sample. Using the training data to

train the model and then test it on the test data 4.76% of the variance in academic performance is explained by our self-esteem/social accpetance out-of-sample. This is not a very good out-of-sample estimate because the OOS R^2 is so low. The reason for this low R-squared value is because we are only regressing the outcome on 16 out of the 86 regressors in this dataset.

The Ridge model generated similar results to the Lasso method which makes sense because of how similar they are to each other. The Ridge model generated an in-sample R^2 of 0.24 and an out-of-sample R^2 of 0.065. While this model does slightly better at predicting out of sample than the Lasso model it is not a model that we would want to use to develop support systems or enhance learning overall. This model only considers self-esteem/feeling of social acceptance, age, race and gender while there are many other variables that affect educational ability besides these factors.

The limited success of Lasso and Ridge regression approaches in our study can also be attributed to the specific characteristics of our dataset. With a relatively low number of regressors and observations, the dataset may have lacked the necessary characteristics for these regularization techniques to effectively capture the relationship between self-perception/social acceptance and academic ability. The reduced dimensionality of the predictor space and the limited variability in the data may have constrained the ability of Lasso and Ridge regression models to identify and leverage meaningful patterns and relationships. Furthermore, the small sample size may have increased the risk of overfitting. Consequently, despite their potential utility in high-dimensional datasets, the limited scope and size of our dataset made Lasso and Ridge regression less suitable for accurately modeling the prediction in our study.

The Mean Squared Error (MSE) against the logarithm of the regularization parameter (Log(λ)) graph seen in Figure 5 in the appendix also suggests that Lasso regression was not a

good approach for this dataset because increasing the regularization parameter λ did not improve the model's performance but instead led to a higher error rate. For lower values of Log(λ) (from -7 to around -4), the MSE remains relatively constant, suggesting that for these values of λ, the Lasso regression is not significantly improving the model's performance. As Log(λ) increases beyond -4, the MSE starts to increase, indicating that as the penalty on the coefficients increases (i.e., as λ increases), the model's performance worsens, and leads to a higher error rate. This is contrary to the expected behavior of Lasso regression, where increasing λ should help in reducing overfitting and improving the model.

Finally the results of the random forest model generated the lowest OOS deviance and the largest OOS R^2 of the three models. Figure 2 below compares the OOS deviance between the lasso and random forests models and the random forest model has about half the amount of OOS deviance as the lasso model on average. This means that the random forest predicts about two times better on new data than the lasso model. Although the random forest predicts much better than the lasso model we are still skeptical about using this model to help us make decisions on implementing new protective measures to help teen development.

Conclusion:

In this machine learning final project we demonstrated many aspects of the class using a dataset that we found using the resources provided. The population of interest is US teens who are 13-17 years old who earn A's and B's in school on average. The methods used to predict educational ability based on self-esteem and social acceptance in this project are the lasso and ridge regularization methods along with random forest. Of the three methods, the random forest model gave us the lowest OOS deviance meaning it was the best at predicting OOS. Of the 13

different self esteem/ social acceptance indicators the one most correlated with higher grades is “I like myself” and the indicator correlated with lower grades is “I get into trouble a lot”.

We would not use even the random forest model to help us change our strategy of elevating student success because the OOS deviance is too low. Figure 4 highlights the regressors most correlated with doing poorly in school and we might be able to use this for insight. Question 27 part F asked students how much they feel like this represents them “I get in trouble a lot”. This question is by far the most correlated with doing poorly in an academic context. As teen psychologists we should try to understand why teens that tend to get in trouble a lot also tend to have lower grades and what can we do to support the teens who get in trouble the most to help them also succeed in school.

Appendix:

Figure 1: Frequency of Response for Q28

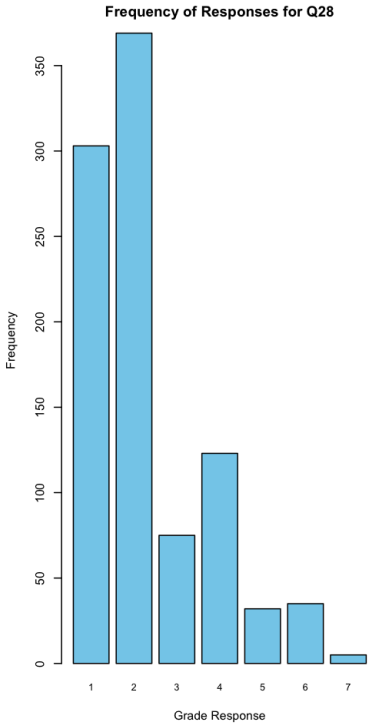

Figure 2: Lasso, Random Forest OOS Deviance Comparison

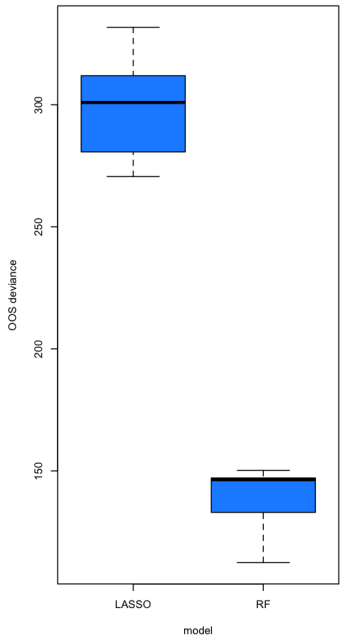

Figure 3: Comparing different m values in the random forest

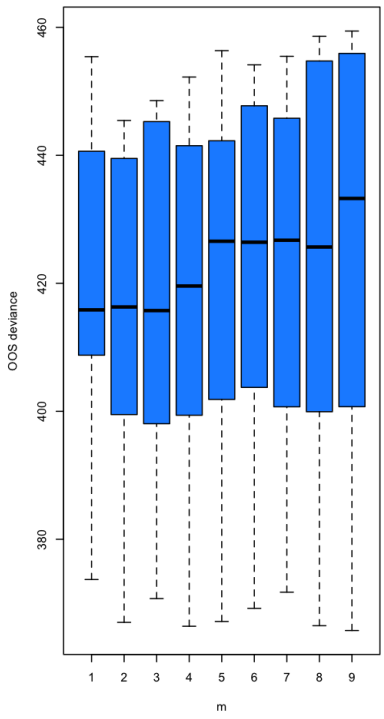

Figure 4: Regressor Relationship on Grades Comparison

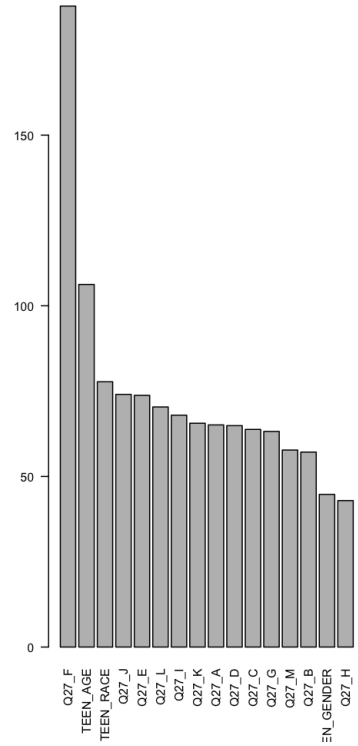

Figure 5: Mean Squared Error against Coefficients (lasso)