Methods

In order to properly consult our client, we build score function and apply it to each school, recommending the top institutions to our client. To create the score function we use ordinary least squares regression (OLS) and cross validation.

OLS is an approach to predicting a quantitative response Y based on a multiple predictor variables X_1 through X_p, where Y and X_1 through X_p are vectors and each value in each X_{ij} (x_{ij}) has a corresponding value in Y (y_i). The model assumes that the relationship between every X_i and Y is linear and that each X_j isn't correlated with any other X_j. OLS can be written as $Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$, where β_0 is the intercept and β_1 through β_p are the slopes of their corresponding predictor variable X_j. The beta values are all constants, unknowns, and together are the model coefficients. The interpretation of β_0 is the expected mean value of Y without a predictor variable and the interpretation of β_1 through β_p is the change in Y for a unit increase in the beta's corresponding X_j. Although the betas are unknown, we can estimate them using the OLS model: solving for the intercept and slopes that produce the plane closest to each point (x_{ij},y_i) in each X_j,Y, which is minimizing the residual sum of squares (RSS). Once we have an estimate for the betas, we can compute using OLS to determine the strength of the relationship between each X_j and Y, if the relationship is statistically significant, and asses how accurately the model predicts the relationship.

Cross-validation is a model validation technique used to assess how the results of an analysis will generalize to an independent data set. To cross-validate a model, we fit its required estimated parameters using k parts of a training dataset and then test the k created models on an unseen test dataset. We then chose the model that minimizes the predictive error. If the model performs well using the test dataset (meaning it has a low predictive error), we are more confident that this model is accurate (since it produces good prediction of data not used to fit the model). For our purposes, we will cross validate the resulting model from our OLS regression and choose the β vector that has the minimal cross-validated mean squared error.