The Higher-Ed Coronavirus Response in the Public Sector: The Usage of Dimensionality Reduction Techniques and Feature Importance Algorithms to Analyze Fall Re-Opening Plans

Adam Hearn

December 2020

Abstract

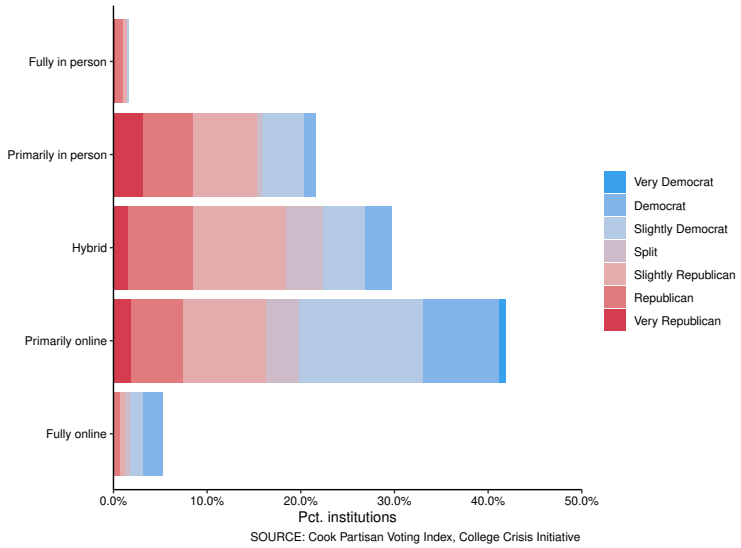
Using a combination of data from the College Crisis Initiative (C2i), the Integrated Postsecondary Data System (IPEDS), and CDC coronavirus statistics, this paper investigates the factors which play significant roles in predicting Fall 2020 mode of instruction in the public, 4-year sector of higher education. Using principal components analysis (PCA), I am able to scale down nearly 30 institutional features to a handful of orthogonal vectors while maintaining relatively high predictive accuracy. Further, I find that the primary factors in predicting in-person instruction were state political leaning and intercollegiate athletics.

1 Introduction

Throughout June and July this past summer, there were many uncertainties surrounding the fall plans of colleges and universities across the United States. The decisions of how to hold classes across the country were highly varied across institutions: some indicated they would instruct their classes entirely online, some decided to hold "hybrid" format classes alternating between in-person and virtual instruction, while others conducted classes fully in-person. This project digs into these decisions of various institutions in the public, four-year or above sector of higher education and uncovers some of the underlying factors in this decision.

As this is a novel topic, very little research has been conducted on these decisions, especially in the public sector. A research team out of Davidson College, the College Crisis Initiative (C2i), has compiled a robust database on these plans (The College Crisis Initiative, 2020). Chris Marsicano, head of the College Crisis Initiative, found strong evidence of political control of the state impacting an institution's decision on these plans: colleges that reside in states carried by Trump in 2016 were more likely to conduct in-person classes

Figure 1: State Political Leaning and Mode of Instruction:
Public, 4-year or above



during Fall 2020, while blue states were more likely to be online (Amour, 2020).

Ad Astra, a higher-ed consulting firm, also finds heavy evidence of state politics impacting these reopening plans (Amour, 2020). The firm conducted a survey on 57 institutions and found the state health department to be the largest influence in this decision and the governor as the third largest influence. However, with this small sample and qualitative nature, these results may not be generalizable to the public higher-education sector as a whole.

The approach for this project differs from that of Marsicano and Ad Astra in that I employ a latent factor approach alongside regression analysis and machine-learning classification to analyze these decisions. Further, my analysis focuses on the public, 4-year sector of higher education whereas others have pooled across all sectors and levels. By filtering my data to include only public, 4-year institutions I can investigate the partisan trends on specifically state and locally controlled institutions.

By conducting this research, I hope to answer the following three questions:

1. Can highly correlated institutional metrics be scaled down to a handful of eigenvectors?

2. What were the key factors in determining Fall 2020 mode of instruction?

3. Can Fall 2020 mode of instruction be accurately predicted using a combination of administrative and CDC data?

2 Data

Cross-sectional, institutional-level data are ideal for this research. Luckily, these data are widely available for researchers and stakeholders. In addition, including state- and county-level measures of partisanship and coronavirus infection rates can aid in classification and causal inference of institutions' instruction mode. That said, the data for this research come from six key sources, outlined below:

1. **College Crisis Initiative (C2i)**: The College Crisis Initiative, known colloquially as "C2i", has generously produced a dataset on each institution's fall reopening plan. This variable can take on several categories, including "fully online," "primarily online," "hybrid," "primarily in-person," and "fully in-person" (The College Crisis Initiative, 2020). For the sake of this project, I condense these plans into three categories: online, hybrid, and in-person.

2. **The Integrated Postsecondary Data System:** Also known as IPEDS, the Integrated Postsecondary Data System is the primary administrative data survey used in the higher education sphere (US Department of Education, National Center for Education Statistics, 2020a). From this data source, I am able to import information on several institutional factors, including enrollment numbers, percentage of students who are full-time, percentage of students who are from in-state or international, admission rate, and other institutional characteristics. Unfortunately, the IPEDS data contains missing values distributed across several variables. As a result, dropping all institutions with missing values leads to a small samples size. As such, I impute missing values using a K-nearest neighbors imputation method to fill in the missing data where needed.[1]

3. **Equity in Athletics Data Analysis:** The Equity in Athletics Disclosure Act requires all institutions who receive Title IV funding and participate in intercollegiate athletics to submit data on participation, revenues, and expenses for each team at their institution, available in the Equity in Athletics database (US Department of Education, Office of Postsecondary Education, 2020). The NCAA offers championships in eight sports during the Fall season: Field Hockey (W), Cross Country (M/W), Football (M), Soccer (M/W), Volleyball (W), and Water Polo (M). To analyze the effect of athletics on instruction plans, I incorporated

[1] The methodology behind this imputation can be found in the Appendix.

data on revenues and participation at the institutional level for these athletic offerings.

4. **National Conference of State Legislatures:** Leveraging hypotheses generated from Marsicano, Ad Astra, and others, I hypothesize state party control to impact institutions' decisions on how to conduct classes during the pandemic. The virus became quickly politicized, especially in the summer-months when these decisions were taking place. President Trump downplayed the threat of the virus and Republican governors and state legislatures followed his lead. As such, institutions in Republican controlled states found a much higher proportion of institutions conducting classes in-person as opposed to Democratic states (Figure 1). This data are available at the state-level and merged with each institution, based on 2020 representation in state legislatures by party (National Conference of State Legislatures, 2020).

5. **Cook Partisan Voting Index:** The same theoretical framework holds that republican-leaning states are more likely to be in-person than democratic leaning states, so the Partisan Voting Index can give a measure how strongly a state leans towards one party or another (Cook Political Report, 2017).

6. **The New York Times**: Based on the research from Ad Astra, it appears CDC guidance was the driving factor in the decisions of their small subset of institutions. Counties where the novel coronavirus was hitting particularly hard during the summer months may have had an effect on institutional plans for fall 2020. County-level data was downloaded from the New York Times GitHub repository (New York Times, 2020). A per-capita measure of daily positive cases in July 2020, the month where institutions were making these decisions, was calculated.

Together, each of these sources combine into a robust dataset with metrics available from multiple domains that can explain these instructional decisions each institution had to face. Summary statistics of variables collected from each of these sources are displayed in Table 1, which I draw from to develop my analytical plan.

3 Methodology

Principal Components analysis is an increasingly popular dimension reduction technique seen in data science circles, yet is rarely seen nowadays in the higher education literature. PCA, an unsupervised learning technique, is a method which derives a small, low-dimensional set of features from a large set of variables (James, Witten, Hastie, & Tibshirani, 2017). The goal of PCA is to

explain a relatively large portion of the variance in the data, while uncovering latent attributes formed by a combination of metrics.

In order to run the PCA algorithm, the data must be scaled such that each variable has a mean of zero and a standard deviation of one. This must be performed so that each variable contributes equally to the analysis. Next, a covariance matrix is produced and principal components are generated.

In essence, principal components are linear combinations or mixtures of the initial variables. These eigenvectors are constructed such that each principal component is orthogonal (uncorrelated) with the others (Abdi & Williams, 2010). Each principal component represents an attribute, at times latent, that is comprised of the original variables and weights that correspond with each variable. PCA loads metrics that are highly correlated with one another into the same principal component, so there is little need to worry about using correlated features in our data.

Aside from dimensionality reduction, principal components can be used alongside regression analysis to help fight multicollinarity (Alibuhtto & Peiris, 2015). Multicollinarity becomes an issue when two or more covariates are highly correlated with one another. When this occurs, standard errors are inflated and causal inference becomes harder to achieve. PCA in regression analysis has been applied in a variety of empirical research, from predicting household food waste (Qi & Roe, 2016) to chemometrics (Næs & Martens, 1988).

However, when using PCA alongside a logistic regression, a method of principal components regression (PCR), I will forgo a portion of the variance in the data. Further, each singular feature can no longer be attributed to changes in the dependent variable. Since some of this interpretability is sacrificed, each eigenvector will be labeled such that interpretability will be easier to come by.

Aside from Kosar and Scott (2018), who used PCA to examine Carnegie Classifications in research universities, the technique has yet to be adopted heavily in the higher education causal inference sphere. However, PCA is highly utilized in the public health sector for dimensionality reduction with highly correlated variables (Batis et al., 2016; Schaik, Peng, Ojelabi, & Ling, 2019). Similarly to Schaik et al. (2019), I will be using the principal components to run a variety of algorithms to generate feature importance for each eigenvector. The techniques involved will be a logistic regression and a random forest decision tree, similar to the analytical methods of Mendez, Buskirk, Lohr, and Haag (2008) and Wager and Athey (2018).[2]

[2]While a multi-class algorithm could be conducted on this data, I choose to split it up into three separate models so that feature importances can vary across modes of instruction.

3.1 Analytical Plan

3.1.1 Principal Components Analysis

To answer question (1), I will run a principal components analysis on the data to scale down several correlated features into a handful of orthogonal vectors. This will involve standardizing the data, running the principal components algorithm, and choosing the ideal number of eigenvectors to keep. The variables mapped to each principal component will be inspected to determine what attribute each eigenvector represents.

3.1.2 Feature Importances

To evaluate feature importances and answer question (2), I will be investigating the coefficient estimates of each principal component in the logistic regression. This can be represented with the following equation, where Y can represent either in-person instruction, hybrid instruction, or online instruction.

$$Y_i = \beta_0 + \beta_1 Eig_{1i} + ... + \beta_8 Eig_{8i} + \epsilon_i$$

By standardizing the coefficients in this regression, I can easily compare each feature's importance by looking at the magnitude of each coefficient generated in the logistic classifier. By using PCA alongside a logistic regression, the coefficients cannot be interpreted directly (e.g., analyzing a hypothetical increase in enrollment with an increased probability of conducting classes online). However, I can interpret each principal component as singular feature that impacts mode of instruction.

Also to aid in the determination of feature importance, I will run a random forest classifier to inspect the key components in this decision. This random forest model will create an ensemble of decision tree classifiers and generate the Gini importance, a derivation of the Gini index. This value represents the impurity of samples that are split at the parent node of the decision tree (Zhang & Ma, 2012). Random forests cut down on some of the innate noise found in regular decision tree classifiers by creating an ensemble of smaller trees. However, it can be prone to overfitting and typically gives a lower prediction accuracy compared to other models.

Both of these methods for determining feature importance can be useful for answering question (2). However, the feature importances from each of these methods may differ. In the logistic regression, the influence of all eigenvectors are studied as they work simultaneously, whereas the random forest decision tree looks at the influence of the variables one at a time. To evaluate the primary predictors of reopening plans, these feature importances should be inspected holistically.

3.1.3 Classification and Evaluation

Lastly, to evaluate these results and determine if Fall 2020 mode of instruction can be accurately predicted, I will be inspecting the classification power of both the logistic regression and the random forest calculator by generating a receiver operating characteristic (ROC) graph and confusion matrices. The ROC curve will provide insight into predictive capabilities of a certain estimator. The area under this curve (AUC) represents the probability that the classifier will rank the "wrong" classification above the "true" classification. As such, this can be used to evaluate my classification algorithms.

4 Findings

4.1 Principal Components Analysis

To answer question (1), I first conducted principal components analysis on the data. This was achieved by transforming the dependent variables as scaled and then running the principal components algorithm. Eight principal components were chosen to run my analysis.[3] The corresponding variables and their respective weights that are mapped to these eigenvectors are displayed in Table 5. I identify the latent attributes associated with each of these eigenvectors as:

1. **Institutional Prestige:** 150% graduation rate, FY 2018 EOY endowment (logged),[4] Open access admissions, percentage of students full-time, percent of faculty who are full-time, NCAA Division I, in-state tuition (logged).

2. **State Partisanship:** Percent Democratic/Republican in State House of Representatives, Percent Democratic/Republican in State Senate, Cook Partisan Voting Index.

3. **Institutional Testing Capacity:** Undergraduate enrollment (logged), student-faculty ratio, total faculty, 2019 county-level population estimate (logged), average county-level cases per 100,000 in July.

4. **Institutional Diversity:** Percentage of undergraduates who were awarded Pell Grants, percentage of undergraduate who are minority, admission rate.

5. **Institutional Geographic Diversity:** Percentage of students in-state, geographic diversity index.[5]

[3] This value was chosen as eight principal components was the lowest number in which athletic data was successfully pulled into its own eigenvector.

[4] Variables that were seen to have a high level of skewness were log-transformed before standardization.

[5] See Appendix for how this was generated.

6. **Athletics:** Number of fall athletes, fall sports revenue (logged), football offered at institution.
7. **State Governorship:** Republican/Democratic Governor.
8. **Proportion of nontraditional students:** Percent of undergraduates degree-seeking, percent of undergraduates international, percent of undergraduates age 25 or older.

Together, these eight eigenvectors account for 78.73% of the variation in my data and group correlated features together. As such, I can scale down the data from 29 features to 8 uncorrelated eigenvectors.

Running principal components analysis on this dataset significantly cut down on the number of features and correlation within the data. For example, Figure 3 displays the correlation matrix of the original, standardized data alongside the new, transformed data using principal components. The correlation between each new eigenvector is zero, meaning that there will be no threat of multicollinarity in the logistic regression model.

As mentioned, PCA groups highly related features together. These variable groupings are intuitive at times (e.g., Fall sports revenue grouped with Fall athletic participation), yet can require greater inspection at other times. For example, admissions rate being loaded onto the same principal component as percent of undergraduates Pell-eligible and percent of students from a minority background. However, an unfortunate side-effect of stratification in the higher education sphere is a positive correlation between admissions rates and proportion of underrepresented students (Hearn, 1984), thus providing insight unto this grouping.

4.2 Feature Importances

4.2.1 Logistic Regression

To answer question (2), A logistic regression model was run on the complete dataset to generate coefficients and uncertainty estimates for each eigenvector. These results are displayed in Table 1. When predicting in-person instruction, I find that state-partisanship, athletics, and testing capacity yield the greatest effect on this classification. Each of these coefficients are significant at $p < 0.01$. State governorship is also significant at $p < 0.05$.

On the opposite side of the spectrum, predicting online instruction, state partisanship is by far the most significant predictor. Testing capacity is also a highly significant indicator.

4.2.2 Random Forest Classifier

Next, the data was split into a training and test set to generate feature importances using the random forest model. To improve uncertainty and fight

Table 1: Logistic Regression Results

	Dependent variable:		
	In-person	Hybrid	Online
	(1)	(2)	(3)
Institutional prestige	0.162	0.0004	−0.039
	(0.108)	(0.086)	(0.083)
State partisanship	−0.757***	−0.347***	0.823***
	(0.106)	(0.088)	(0.090)
Testing capacity	−0.280***	−0.198**	0.360***
	(0.101)	(0.087)	(0.084)
Diversity	0.161	0.091	−0.194**
	(0.098)	(0.088)	(0.085)
Geographic Diversity	0.008	−0.337***	0.155*
	(0.095)	(0.091)	(0.084)
Athletics	−0.397***	0.206**	0.035
	(0.108)	(0.086)	(0.083)
Governorship	−0.259**	0.043	0.149*
	(0.101)	(0.085)	(0.080)
Nontraditional students	0.109	0.141*	−0.192**
	(0.106)	(0.086)	(0.085)
Constant	−1.455***	−0.956***	−0.166**
	(0.108)	(0.088)	(0.083)
Observations	713	713	713
Log Likelihood	−337.527	−408.944	−427.467
Akaike Inf. Crit.	693.053	835.888	872.935
Note:			*p<0.1; **p<0.05; ***p<0.01

random sampling errors, this model was run using a Monte Carlo simulation of 1,000 random samples with each tree representing 5,000 estimators. Next, the average feature importance was calculated for each random sample.

Using the random forest classifier, I find state legislative makeup, athletics, and state governorship to be the three most significant indicators for in-person instruction. For online instruction, the three most significant indicators are state legislative makeup, institutional testing capacity, and institutional prestige. These results are displayed in Table 2, which displays feature importances for each eigenvector.

4.3 Evaluation

Lastly, a new logistic model was trained on a subset of units and tested on the remaining institutions to evaluate predictive accuracy of the algorithm. The

Table 2: Random Forest Feature Importance

	Classification model:		
	In-person	Hybrid	Online
	(1)	(2)	(3)
Institutional prestige	0.120	0.120	0.114*
State partisanship	0.154***	0.130*	0.210***
Testing capacity	0.114	0.132***	0.123**
Diversity	0.127	0.128	0.113
Geographic Diversity	0.110	0.125	0.108
Athletics	0.131**	0.123	0.108
Governorship	0.127*	0.128	0.113
Nontraditional students	0.109	0.130**	0.192
# Trees	5000	5000	5000
# Random Samples	1000	1000	1000
Note:	*Rank #3; **Rank #2; ***Rank #1		

same was repeated for the random forest classifier. This evaluation step answers question (3).

Table 3 displays the evaluation measures for my classification algorithms, with columns (1) and (2) representing the algorithm run with the principal components and columns (3) and (4) include the algorithms run with the full set of features.

In the models predicting in-person and online instruction, the logistic model outperforms that of the random forest. In these models, though the random forest model yields a higher predictive accuracy, the the area under the ROC curve of the logistic regression is greater than that of the random forest. Surprisingly, predictive accuracy of hybrid instruction was relatively high.

Lastly, conducting the classification algorithms with only the principal components yields similar, if not better, predictive accuracy than the full set of features. This is likely due to overfitting when using all variables. It appears that highly correlated institutional metrics can be successfully scaled down to a handful of eigenvectors while retaining relatively high predictive accuracy, helping to fight multicollinarity and prevent overfitting.

5 Conclusion

These results highlight the politicization of the novel coronavirus and its impact in the public, higher education sector. Politics are often overlooked in higher

Table 3: Evaluation of Classification Algorithms

		Principal Components		All Features	
		Logistic	Random Forest	Logistic	Random Forest
Model	Measure	(1)	(2)	(3)	(4)
In-person	Accuracy	0.58	**0.79**	0.54	**0.77**
	Precision	0.31	**0.5**	0.27	**0.38**
	Recall	**0.83**	0.23	**0.7**	0.17
	F-Measure	**0.46**	0.31	**0.39**	0.23
	ROC AUC	**0.67**	0.59	**0.6**	0.55
Hybrid	Accuracy	0.51	**0.76**	0.51	**0.81**
	Precision	0.28	**0.52**	0.28	**0.71**
	Recall	**0.61**	0.31	**0.58**	0.42
	F-Measure	0.38	**0.39**	0.38	**0.52**
	ROC AUC	0.54	**0.61**	0.53	**0.68**
Online	Accuracy	**0.68**	0.66	0.65	**0.69**
	Precision	0.71	**0.72**	0.69	**0.79**
	Recall	0.65	**0.67**	**0.61**	0.55
	F-Measure	**0.68**	0.63	**0.65**	0.65
	ROC AUC	**0.68**	0.66	0.65	**0.69**

education policy analysis, and this research shows its quantifiable effects on institutions and students.

While both significant, the magnitude of state partisanship on online instruction were greater than that on in-person instruction. This could signal that democratic-controlled legislatures had a more significant impact in pressuring institutions to go online than their republican counterparts had in pressuring institutions to remain open. Further, my results suggest that governorship plays a less significant role than general state legislative makeup.

In addition, my results contradict the qualitative survey put out by Ad Astra where institutions ranked athletics low on their influences to remain open (Amour, 2020). Perhaps this is a result of their smaller sample size, or maybe there was a "halo" effect present where institutions were reluctant to signal that this is a priority. However, this is not surprising, considering the revenue institutions bring in from Fall athletics. The average institution in my sample received just over $7 million in Fall athletic revenue, while this number reaches a max of a whopping $159 million.

I also find evidence that testing capacity goes both ways; institutions with inadequate testing capacities were more likely to be closed, just as well-equipped institutions were more likely to remain open. This cannot be said for other significant variables, such as athletics or proportion of nontraditional students.

For example, having low athletic participation did not pressure institutions to be closed, and having few nontraditional students did not pressure institutions to remain open.

As for classification and evaluation, it appears that institutional characteristics can be scaled down to fewer features while retaining predictive accuracy. That said, PCA may become increasingly popular within the higher education literature as correlated features such as graduation rate, tuition, and selectivity can be scaled down to a singular eigenvectors. Researchers and higher education policy analysts should look to PCA as an alternative for dimensionality reduction and as a side-step to multicollinarity.

Bibliography

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), 433–459.

Alibuhtto, M., & Peiris, T. (2015). Principal component regression for solving multicollinearity problem.

Amour, M. S. (2020). *State politics influenced college reopening plans, data show.*

Batis, C., Mendez, M., Gordon-Larsen, P., Sotres-Alvarez, D., Adair, L., & Popkin, B. (2016). Using both principal component analysis and reduced rank regression to study dietary patterns and diabetes in chinese adults. *Public health nutrition*, *19 2*, 195-203.

Cook Political Report. (2017). *2017 Cook Political Report Partisan Voter Index.*

Hearn, J. C. (1984). The relative roles of academic, ascribed, and socioeconomic characteristics in college destinations. *Sociology of Education*, 22–30.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning: with applications in R.* Springer.

Kosar, R., & Scott, D. W. (2018). Examining the carnegie classification methodology for research universities. *Statistics and Public Policy*, *5*(1), 1–12. doi: 10.1080/2330443x.2018.1442271

McLaughlin, G., Howard, R., & McLaughlin, J. (2011). Forming and using peer groups based on nearest neighbors with ipeds data. *Association for Institutional Research (NJ1).*

Mendez, G., Buskirk, T. D., Lohr, S., & Haag, S. (2008). Factors associated with persistence in science and engineering majors: An exploratory study using classification trees and random forests. *Journal of Engineering Education*, *97*(1), 57–70.

Morris, E. K., Caruso, T., Buscot, F., Fischer, M., Hancock, C., Maier, T. S., ... others (2014). Choosing and using diversity indices: insights for ecological applications from the german biodiversity exploratories. *Ecology and evolution*, *4*(18), 3514–3524.

Næs, T., & Martens, H. (1988). Principal component regression in nir analysis: viewpoints, background details and selection of components. *Journal of chemometrics*, *2*(2), 155–167.

National Conference of State Legislatures. (2020). *Legislator data.* National Conference of State Legislatures.

New York Times. (2020). *Coronavirus (Covid-19) Data in the United States.* The New York Times. Retrieved from `https://github.com/nytimes/covid-19-data`

Qi, D., & Roe, B. (2016). Household food waste: Multivariate regression and principal components analyses of awareness and attitudes among u.s. consumers. *PLoS ONE*, *11*.

Schaik, P. V., Peng, Y., Ojelabi, A., & Ling, J. (2019). Explainable statistical learning in public health for policy development: the case of real-world suicide data. *BMC Medical Research Methodology*, *19*(1). doi:

10.1186/s12874-019-0796-7

The Chronicle of Higher Education. (2020). *Here's Our List of Colleges' Reopening Models.*

The College Crisis Initiative. (2020). *Covid-19 dashboard.* Davidson Unviersity. Retrieved from `https://collegecrisis.org/`

US Department of Education, National Center for Education Statistics. (2020a). *The Integrated Postsecondary Education Data System.* Retrieved from `https://nces.ed.gov/ipeds/`

US Department of Education, National Center for Education Statistics. (2020b). *IPEDS Survey Methodolology.*

US Department of Education, Office of Postsecondary Education. (2020). *Equity in Athletics Disclosure Act (EADA) Survey.* Retrieved from `https://surveys.ope.ed.gov/athletics/`

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Zhang, C., & Ma, Y. (2012). *Ensemble machine learning: methods and applications*. Springer.

6 Implementation Appendix

6.1 *Chronicle of Higher Education* webscraper

While data on institutional plans had been generously compiled by the College Crisis Initiative, this data was not available for distribution on their RShiny dashboard. However, C2i and the *Chronicle of Higher Education* have a contract with one another, and this data was more easily accessble through the *Chronicle* website (The Chronicle of Higher Education, 2020). Within the HTML source code, the data was stored in a JSON file that could easily be converted to a dataframe.

6.2 IPEDS UnitID Matching

The data collected from the *Chronicle* was perfect, except for the fact that IPEDS UnitIDs were not included and not all institutions names matched up with those in the IPEDS universe. This makes merging several data sources challenging. To combat this problem, I created a function using cosine text-similarity metrics to match an incomplete institution name with the official institution name in the IPEDS universe and obtain the official UnitID numbers for merging.

6.3 K-nearest neighbors missing data imputation

When using the PCA algorithm, all units of observation need to have complete data. Otherwise, these units will be dropped. Without the missing-data imputation, I severely lost institutions cutting down on my sample size. As a result, I was not able to reach the 500-unit threshold required. To adjust for this, I conducted a KNN imputation method using Euclidean distance finding the 15 nearest institutions and weighting these values by distance. This method was chosen as it is the preferred imputation method for the IPEDS survey (McLaughlin, Howard, & McLaughlin, 2011; US Department of Education, National Center for Education Statistics, 2020b).

6.4 Geographic Diversity Index

The geographic diversity index (GDI) is derived from the IPEDS Fall Enrollment (EF) survey which includes the state information on the state of residence of the 2018 first-year class. This variable is calculated using Simpson's Diversity Index (Morris et al., 2014):

$$SDI = \frac{N(N-1)}{\sum n(n-1)}$$

N represents the number of students in the 2018 first-year class, while n_i represents the number of students from a particular state (US Department of Education, National Center for Education Statistics, 2020a). This value ranges from

0 to 1, with 1 being the most geographically diverse. A GDI of zero means that the institution only serves students from one state, where a GDI of 1 means that an institution serves an equal number of students across multiple states.

6.5 Variable Transformation

Variables that were seen to have a high skewness were transformed before conducting PCA. These variables include undergraduate enrollment, tuition, number of fall athletes, total faculty, fall sports revenue, endowment, and 2019 population estimate.

7 Tables and Figures

Table 4: Summary Statistics

Statistic	Mean	St. Dev.	Min	Max
Open access (Y/N)	0.31	0.46	0	1
Adm. rate	0.80	0.19	0.12	1.00
Undergrad. enrollment	10,283	9,960	271	59,483
% Full-time	0.76	0.19	0.01	1.00
% Degree-seeking	0.65	0.14	0.00	0.98
% International	0.02	0.03	0.00	0.35
% minority	0.47	0.21	0.05	1.00
% Pell	0.36	0.13	0.07	0.80
% age 25+	0.21	0.16	0.00	0.94
% students in-state	0.84	0.17	0.00	1.00
Geo. Div. Index	0.25	0.22	0.00	1.00
150% Graduation rate	0.48	0.19	0.04	1.00
In-state tuition	8,367	3,602.25	480	23,400
FY2018 Endowment	204,030,698	832,648,299	40,226	12,688,650,784
Student-Faculty Ratio	17.31	3.98	6	34
Total faculty	730.20	709.43	35	5,911
% Faculty full-time	0.57	0.18	0.00	1.00
2019 Population est.	647,175	1,379,660	6,972	10,039,107
Avg. cases per 100k (Jul.)	16.08	15.45	0.38	98.10
% Democrat: Senate	0.46	0.17	0.10	0.96
% Republican: Senate	0.53	0.18	0.00	0.90
% Democrat: House	0.49	0.16	0.15	0.90
% Republican: House	0.50	0.17	0.10	0.84
Democrat Gov. (Y/N)	0.53	0.50	0	1
Republican Gov. (Y/N)	0.47	0.50	0	1
Partisan Voting Index	−2.09	9.37	−25	25
Fall sports revenue	7,031,945	18,355,374	0	159,274,447
# Fall athletes	107.58	74.62	1	286
Offer football (Y/N)	0.49	0.49	0	1
NCAA Div-1 (Y/N)	0.32	0.47	0	1
In-person (Y/N)	0.23	0.42	0	1
Online (Y/N)	0.46	0.50	0	1
Hybrid (Y/N)	0.29	0.45	0	1

Figure 2: Correlation Matrix: All Features

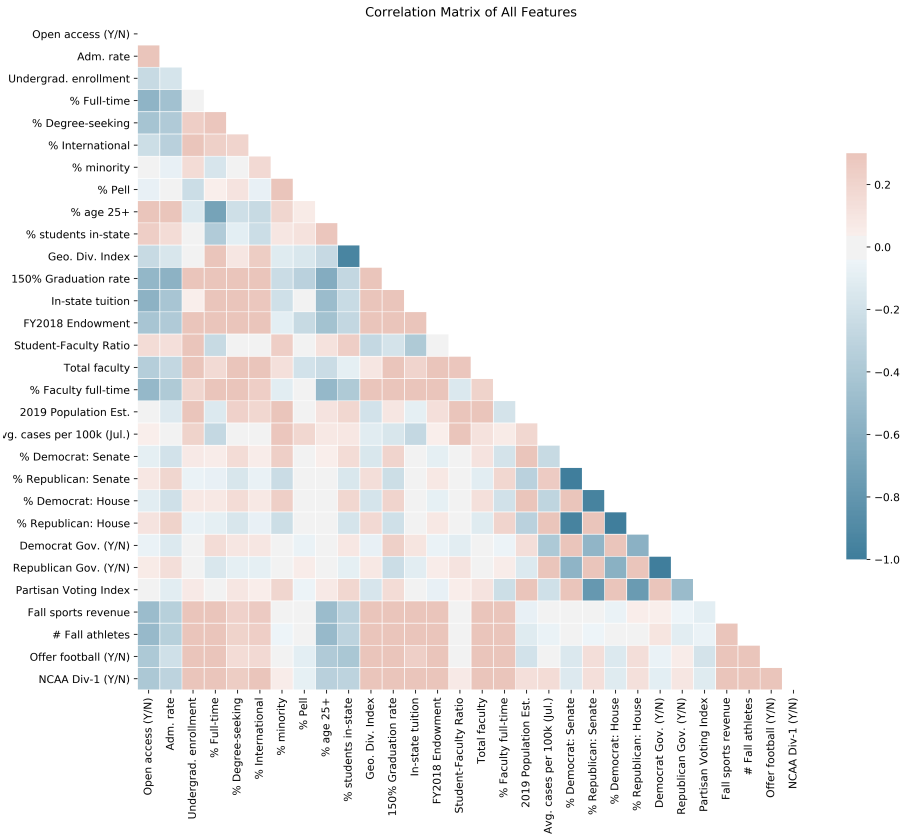


Figure 3: Correlation Matrix: Principal Components

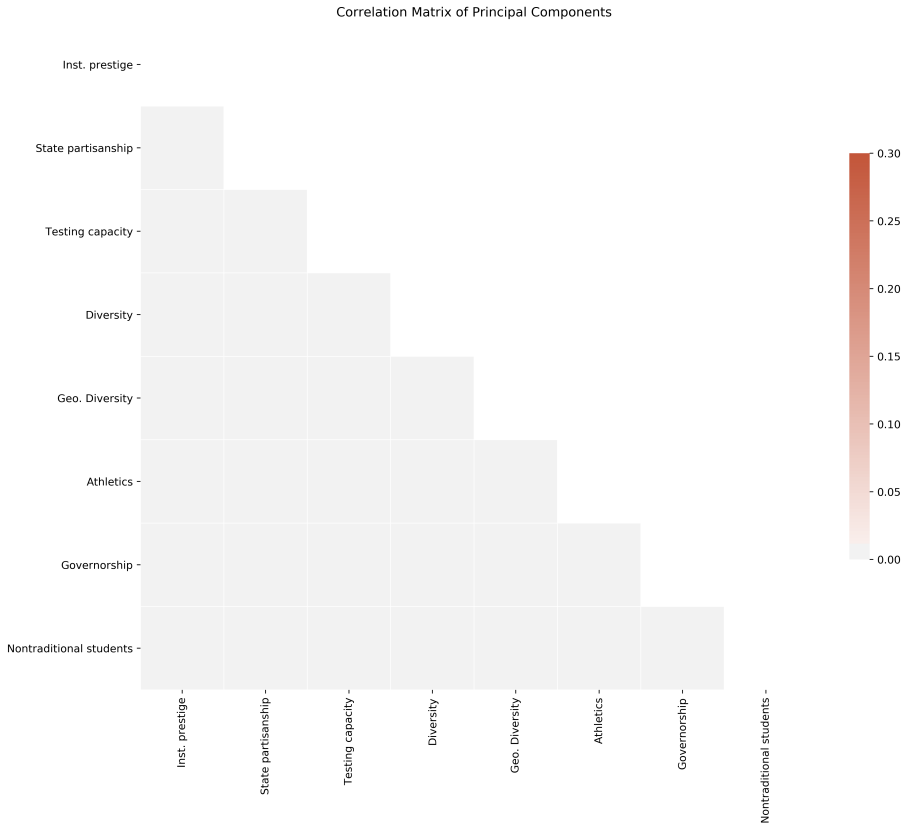
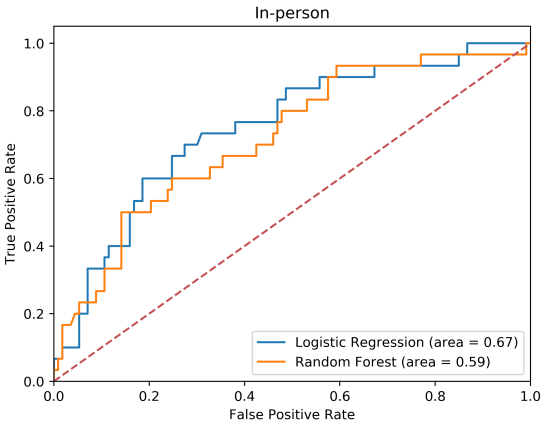
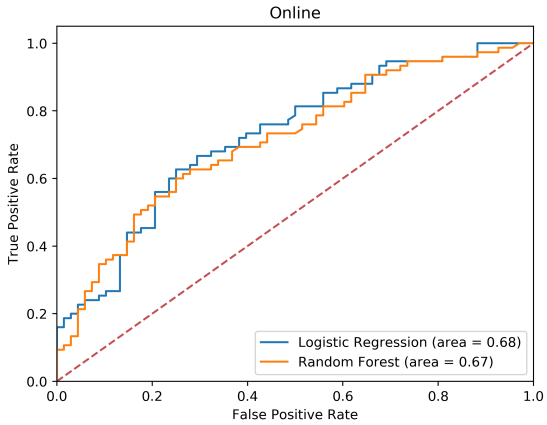


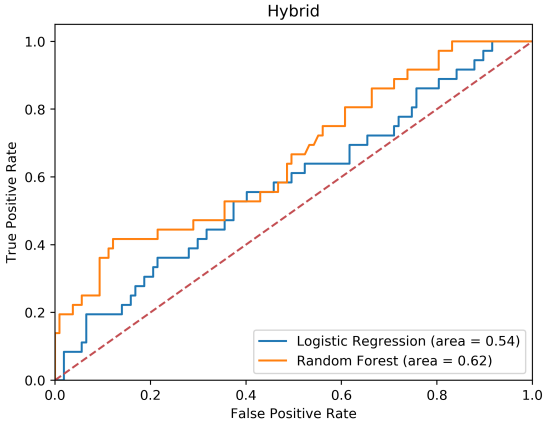
Figure 4: Receiver Operating Characteristics



(a) In-person



(b) Online



(c) Hybrid

Table 5: Principal Component Feature Loadings

Principal Component	*Feature*	*Eigenvector* (1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Institutional Prestige	150% graduation rate	**0.29**	0.08	-0.07	-0.11	-0.21	0.15	-0.05	0.12
	FY2018 Endowment	**0.28**	-0.05	0.15	-0.16	-0.10	0.06	0.01	0.05
	Open access (Y/N)	**-0.26**	-0.02	0.04	-0.25	0.13	0.00	0.09	0.18
	Pct. full-time	**0.26**	0.01	-0.21	0.14	-0.07	-0.06	-0.03	0.13
	% Faculty full-time	**0.25**	-0.11	-0.06	0.13	-0.01	-0.04	0.03	-0.05
	NCAA Div. I	**0.24**	-0.07	0.19	-0.05	0.05	0.09	0.07	-0.11
	In-state tuition	**0.23**	0.01	-0.19	0.11	-0.22	0.17	0.08	0.02
State Partisanship	% Democrat: House	0.02	**0.39**	-0.01	0.02	0.07	0.00	-0.18	0.02
	% Democrat: Senate	0.02	**0.38**	-0.01	0.02	0.10	0.02	-0.23	0.00
	% Republican: House	-0.02	**-0.38**	0.02	-0.03	-0.07	0.00	0.19	-0.02
	% Republican: Senate	-0.02	**-0.38**	0.01	-0.02	-0.11	-0.01	0.24	-0.02
	Cook PVI	-0.01	**0.34**	0.02	-0.04	0.05	0.06	-0.15	0.04
Institutional Testing Capacity	Undergrad. enrollment	0.18	0.03	**0.42**	-0.21	-0.04	-0.06	0.02	-0.07
	Student-faculty ratio	-0.04	0.01	**0.39**	-0.04	0.03	-0.20	-0.04	0.06
	Total faculty	0.22	0.05	**0.35**	-0.17	-0.05	0.02	0.02	-0.11
	2019 population est.	0.01	0.17	**0.33**	0.02	-0.05	0.28	0.14	-0.03
	Avg. cases per 100k (Jul.)	-0.03	-0.13	**0.32**	0.25	0.04	0.04	-0.09	0.01
Institutional Diversity	% Pell	-0.05	0.00	0.02	**0.62**	0.15	-0.14	0.18	-0.08
	% minority	-0.04	0.10	0.25	**0.37**	0.36	0.04	0.18	0.21
	Adm. rate	-0.22	-0.08	0.02	**-0.26**	0.15	-0.17	-0.01	-0.09
Geographic Diversity	% in-state	-0.18	0.11	0.16	0.11	**-0.51**	-0.27	0.03	0.11
	Geo. Div. Index	0.18	-0.11	-0.17	-0.13	**0.49**	0.29	-0.02	-0.12
Athletics	# Fall athletes	0.28	-0.01	0.00	0.02	0.13	**-0.44**	-0.07	0.05
	Fall sports revenue	0.28	-0.04	0.06	0.00	0.11	**-0.39**	-0.03	0.05
	Football offered (Y/N)	0.22	-0.09	0.00	-0.06	0.25	**-0.26**	-0.10	-0.14
State Governorship	Democrat Gov. (Y/N)	0.04	0.29	-0.14	-0.13	0.01	-0.19	**0.51**	-0.12
	Republican Gov. (Y/N)	-0.04	-0.29	0.14	0.13	-0.01	0.19	**-0.51**	0.12
Nontraditional Students	% Degree-seeking	0.16	0.07	0.07	0.21	-0.17	0.20	0.05	**-0.61**
	% Int'l	0.17	0.04	0.10	-0.01	0.09	0.28	0.36	**0.52**
	% Age 25+	-0.25	0.03	0.11	-0.09	0.16	0.08	0.14	**-0.34**