Colleges and Coronavirus

Aziz, Katharine, Manny, Max

Background

“There is... an emerging confidence among at least some college administrators that they have learned much about managing the pandemic on their campuses.” (NYTimes)

Problem Statement

What attributes of colleges contribute to an increased probability that the campus will see greater than 5% of the population infected with the Coronavirus?

Agenda

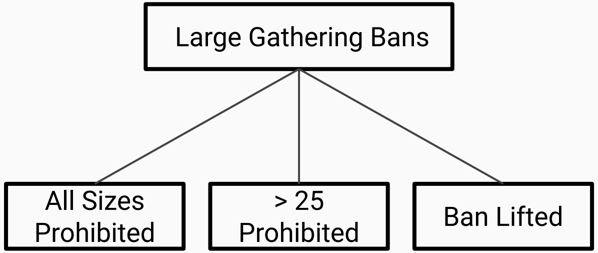
1. Datasets
2. Data Clean-up and EDA
3. Key Feature Engineering
4. Modeling
5. Conclusions and Next Steps

Datasets

- New York Times College COVID Tracker
- College In-person Classes Plan
- State Social Distancing Mandates & Policies
- College Admission Statistics

Examples of Variables Analyzed

State Policies Dataset



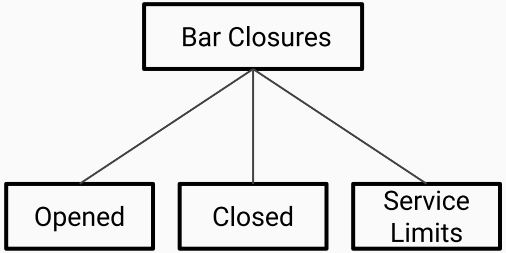
Large Gathering Bans

All Sizes Prohibited

> 25 Prohibited

Ban Lifted

State Policies Dataset



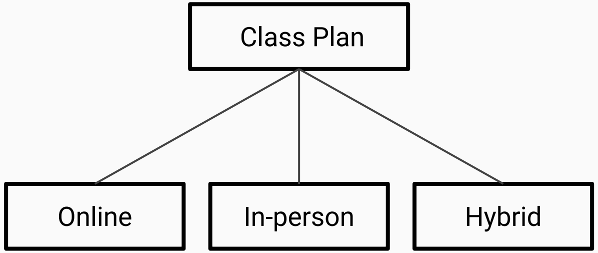
Bar Closures

Opened

Closed

Service Limits

Class Plan Dataset



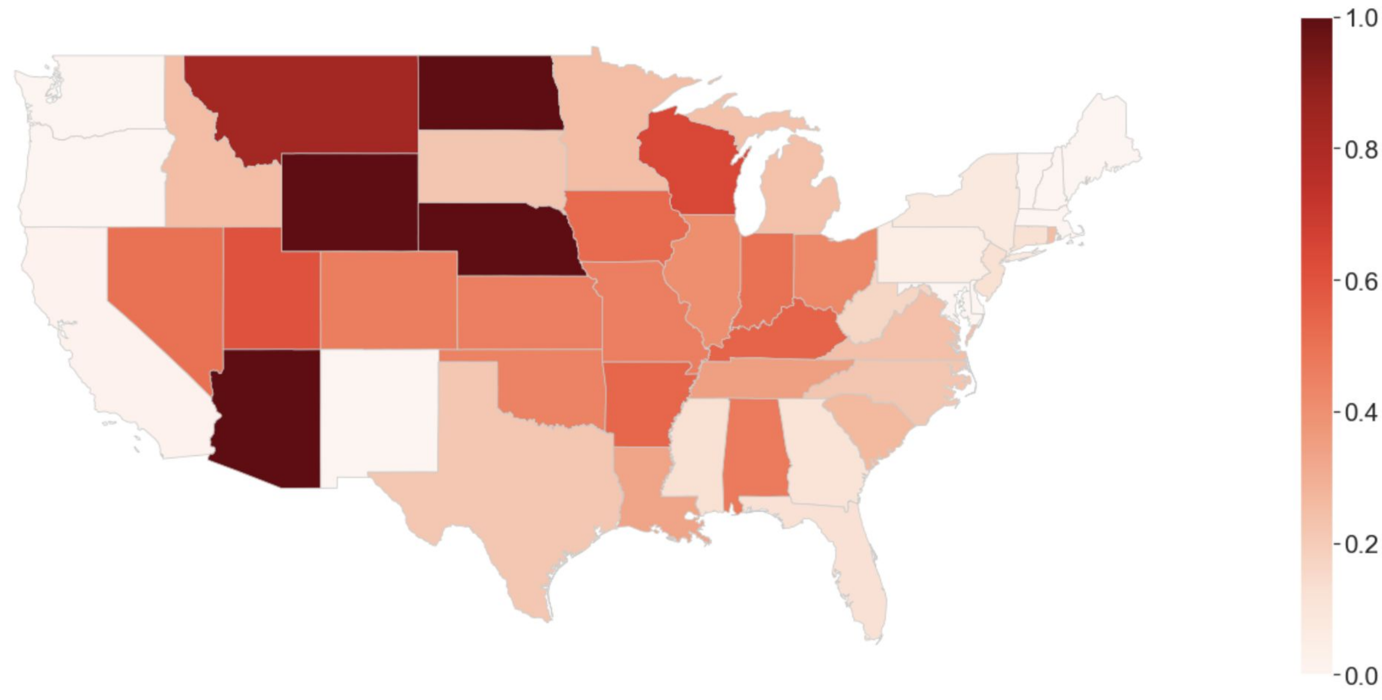
Class Plan

Online

In-person

Hybrid

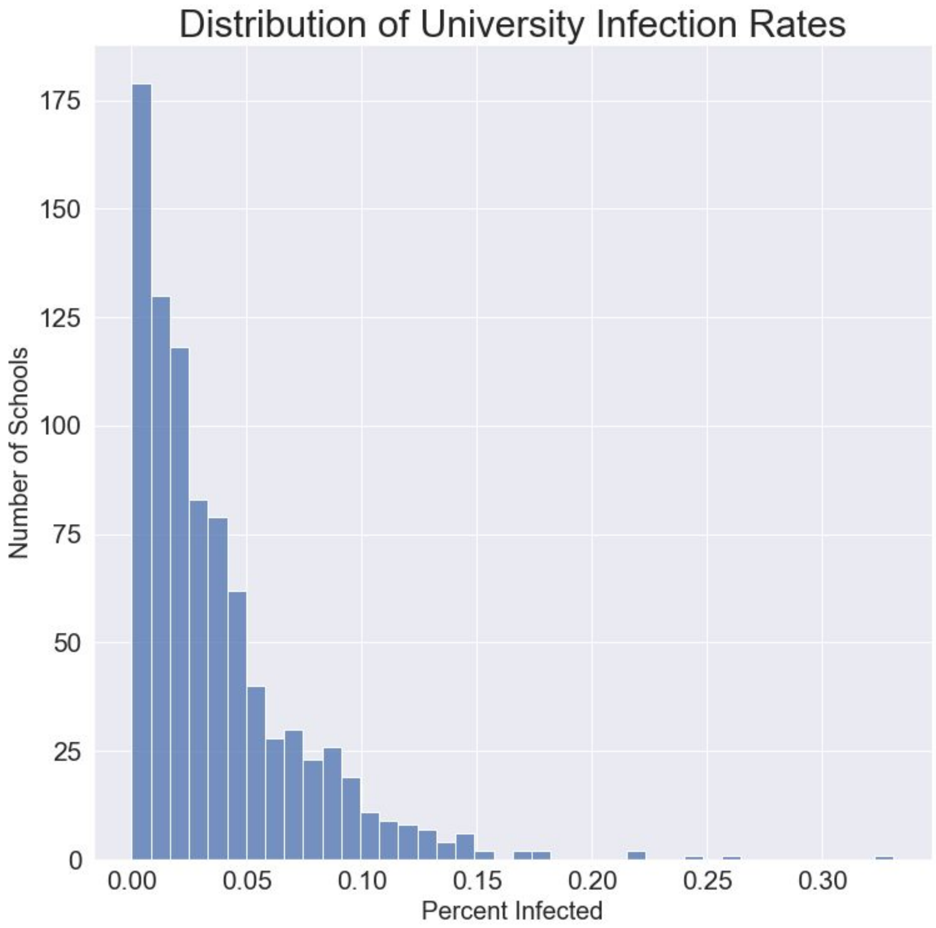
College Outbreaks Mapped



EDA

653 schools had infection rate of 5% or less.

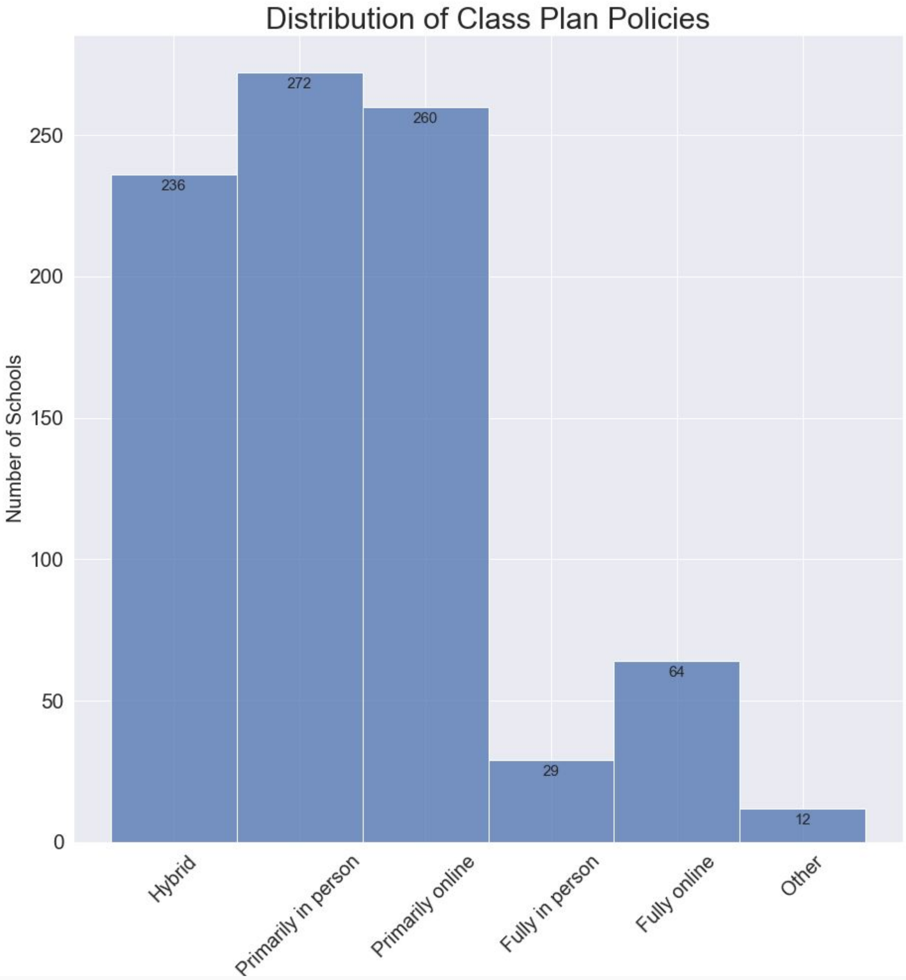
220 schools had infection rate > 5%.



EDA

Most schools are operating in a hybrid fashion.

Minority of schools are operating in pure online/in person format.



Target Column

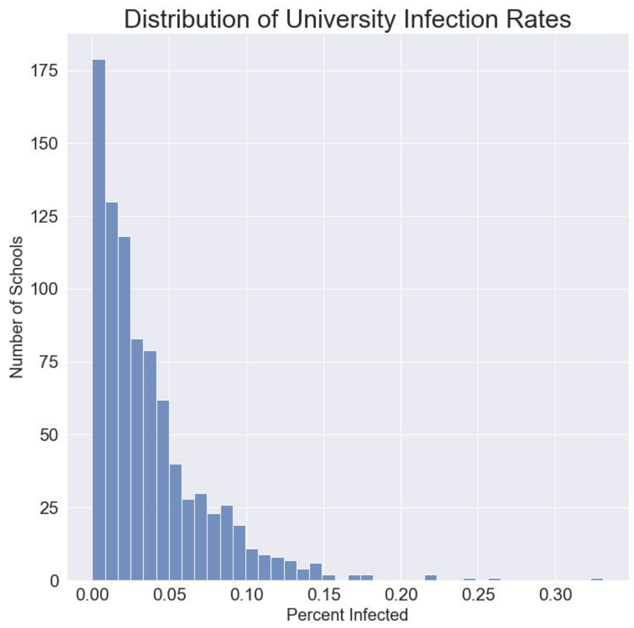
Classification Problem

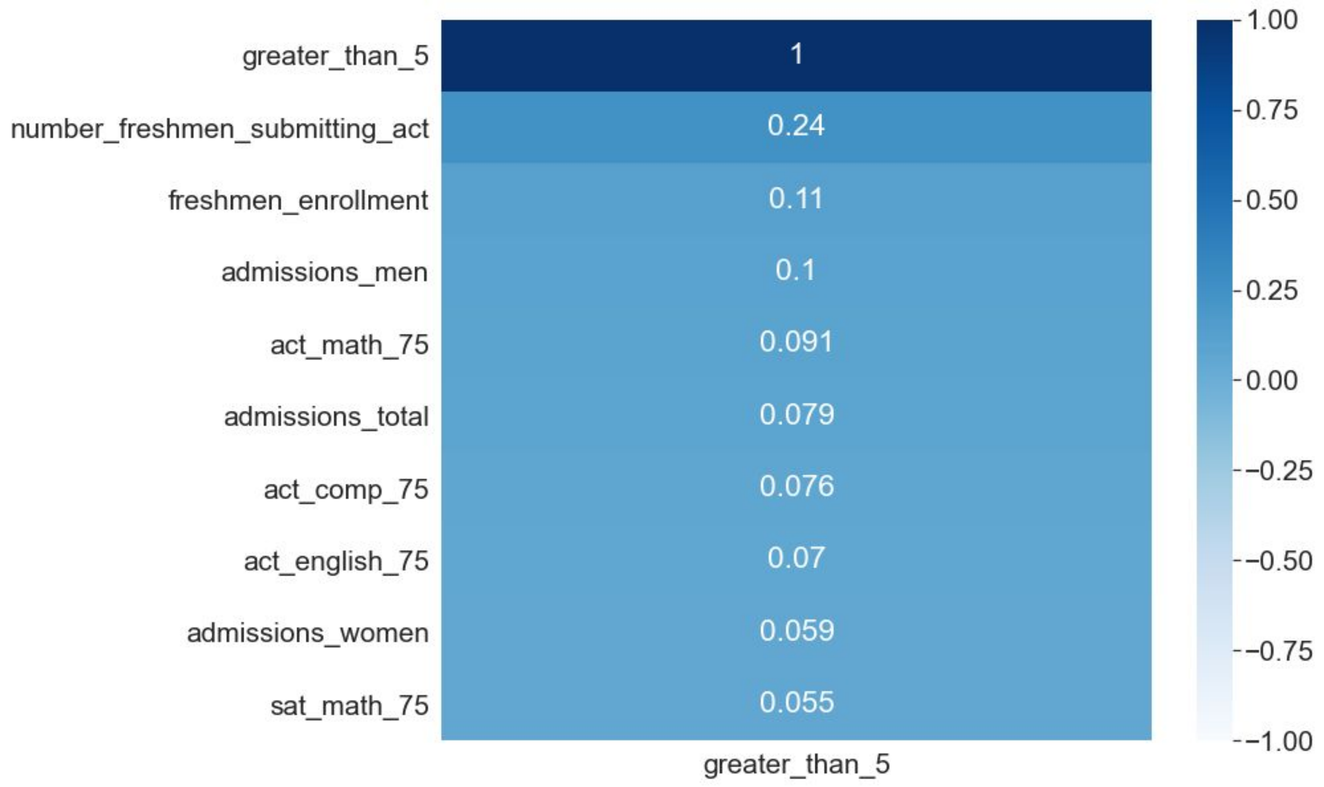
High infection schools:

- Infection rate > 5%

Low infection schools:

- Infection rate <= 5%





Challenge: Low correlation between our numerical features and target variable. Will need to dummify categorical features + feature engineer.

Key Feature Engineering

Packed Bars/Empty Bars

- **Packed Bars -** Class Plan - Primarily in Person * Bar Closures - Reopened
- **Empty Bars -** Class Plan - Primarily online * Bar Closures - New Service Limits

Interaction Feature for Admissions Statistics

- **Test Scores 75** - 75th Quartile for all SAT & ACT scores
- **Test Scores 25** - 25th Quartile for all SAT & ACT scores

Correlation to Target (with Dummied Variables)

Greater_than_5 (Target)	1
Number freshmen submitting act	0.24
Packed bars	0.21
Class plan primarily in person	0.20
Restaurants - Reopened to Dine-in Service	0.19

Modeling

Baseline Accuracy - 0.75

- Models that performed best - Random Forest, Adaboost, Neural Net, Logistic Regression
- Others tested - KNN, BaggingClassifier, SVC

Best Model: Logistic Regression

- With Football Conference Dummies - 0.80 training, 0.80 testing score (135 Features)
- Without Football Conference Dummies - 0.82 training, 0.79 testing score (43 Features)

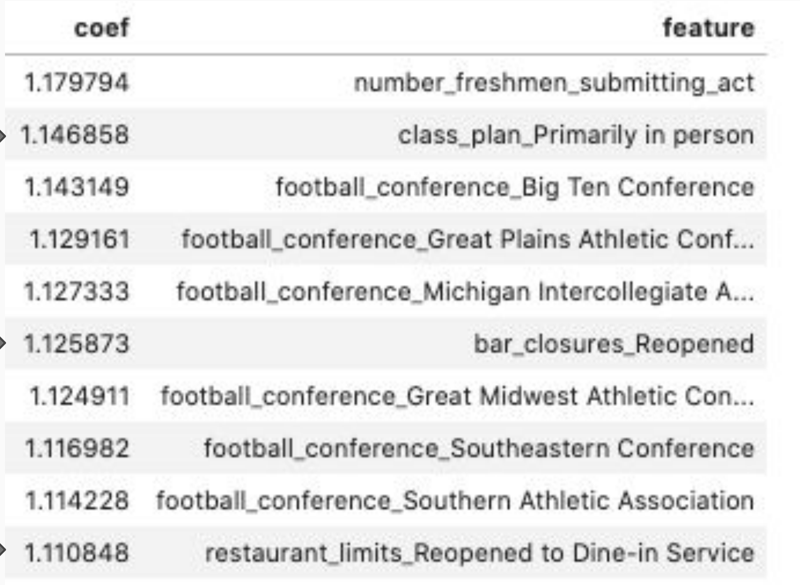
Key coefficients and interpretations

Quantitative Features

- For every 1 unit increase in number of freshmen submitting ACT, institution ~1.179 times as likely to have a significant amount of covid cases, all else held constant.

Categorical Features

- If an institution's class plan is primarily in person, the institution is ~1.147 times as likely to have a significant amount of covid cases, all else held constant.



coef	feature
1.179794	number_freshmen_submitting_act
1.146858	class_plan_Primarily in person
1.143149	football_conference_Big Ten Conference
1.129161	football_conference_Great Plains Athletic Conf...
1.127333	football_conference_Michigan Intercollegiate A...
1.125873	bar_closures_Reopened
1.124911	football_conference_Great Midwest Athletic Con...
1.116982	football_conference_Southeastern Conference
1.114228	football_conference_Southern Athletic Association
1.110848	restaurant_limits_Reopened to Dine-in Service

Conclusions/ Next Steps

- Public health data has many confounding variables and can be difficult to model
- Risk largely seems to correspond to:
  - Regional shifts in policies and your own state’s policies for social distancing guidelines
  - Having classes primarily in person increases risk of infections (and vice versa for online classes)
- Next Steps
  - Additional feature engineering
  - Time Series analysis

Sources

NYT Repository

IPED Database