CSCI 499: Education Deserts Paper

Lucas Hu, Nikhil Sinha, Jeff Chen, Ashwath Raj

April 2019

1 Abstract

Geographical accessibility of colleges is one of many problems facing students in education deserts. This is a problem because lacking a bachelor's degree severely limits the career options and resulting income of education desert populations. Existing analyses and proposals often provide policy decision on how to increase accessibility of existing colleges. However, there is a gap in research around specific socioeconomic analyses of a fair, optimized framework of where to create new, local colleges.

2 Introduction

Within the United States, areas with limited accessibility to universities are called education deserts. For our purposes, we define them as areas that do not contain any colleges within 25 miles. Although other research has established more rigorous methods of measuring critical distances, for our introductory analysis we simply selected 25 miles as a heuristic for commute time. The reason this is important is because over 30% of undergraduate students live at home and commute to university. Clearly having a college nearby students is very important, both for physical accessibility and for helping them believe college is a feasible, realistic option. Therefore, on an economic and moral basis, it becomes increasingly important to address the limited accessibility of higher education for students who would otherwise attend university.

We seek to help alleviate the issue of geographically limited college accessibility by mapping education deserts, calculating the potential socioeconomic benefit of placing colleges in those areas, and creating a framework for optimizing these benefits.

3 Related Work

Existing work, like from Klasik et al., often works from the perspective of race, income, and social inequality and how they are magnified by the presence of hierarchical college rankings and various types of deserts [1]. Klasik et al is successful in identifying population trends of students in education deserts, in saying, "students in access deserts appear to overcome the limitations of their local college options,while the students in match deserts do not." This is critical information which might inform the algorithmic approaches of further research. Further work, like that of Hillman et al., suggests a hyperawareness of the specific problem of education deserts, and advocates for policy changes to increase accessibility of colleges (as broad-access institutions) and awareness of pathways to higher education [2]. Prior work, however, has not yet approached the issue as an optimization problem, in which one can quantitatively determine the economic benefit of creating new universities to combat education deserts.

4 Data

Our primary method of describing education desert populations was through census tract features, provided from the American Community Survey of 2017's 5 year estimates. A census tract is defined as an area of between 2,500 to 8,000 people; roughly equivalent to a neighborhood established by the Bureau of Census [3]. We opted to use census tract information because of the feature-rich, consistent, and widespread information, in addition to the ability to see differences down to local, neighborhood levels. We obtained the shapes of the census tracts through the Census Bureau [4]. We obtained the aggregated locations of current universities from the Integrated Postsecondary Education Data System [5].

In order to find out which census tracts we wanted to place our universities in, we first had to define an "education desert" quantitatively. Hence, we formally defined an education desert as area with no accessible universities within a 25 miles radius. Figure 1 demonstrates the ubiquity of education deserts across the united states, especially in Midwestern and rural communities.

4.1 Exploratory Data Analysis

According to our definition of education desert, we find that about 40.5M people live in census tracts that are education deserts, whereas 283.9M peo-

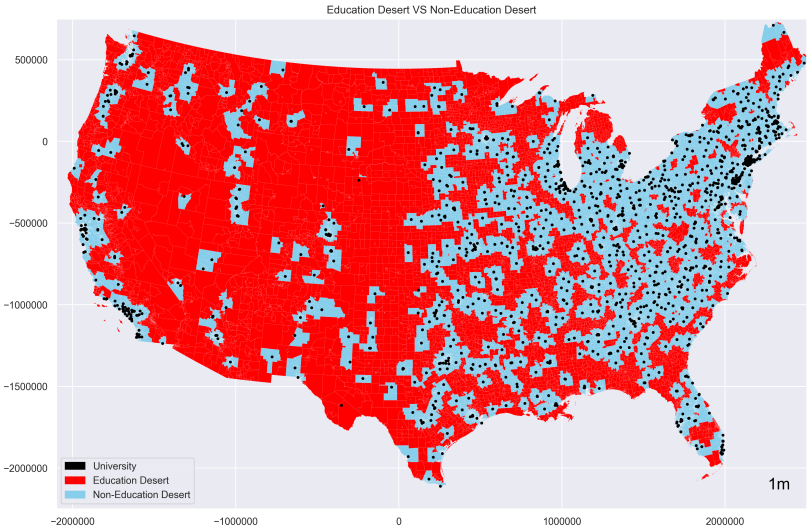

Figure 1: A map showing education deserts across the United States

ple live in census tracts that are not education deserts. We also find that education deserts have an average population density of about 1029 people per sq. mile, whereas non-education deserts have an average population density of about 6122 people per sq. mile, indicating that education deserts tend to appear in more sparsely populated areas.

In general, we see that poverty rates are lower in education deserts than than they are in non-education deserts. Furthermore, both income levels and rent prices are, on average, lower in education deserts.

Surprisingly, racial demographics generally do not seem to differ significantly between education deserts and non education deserts. A few exceptions are that education deserts seem to have a lower population of Asian residents, and are more likely to have a disproportionately high number of white residents (likely due to education deserts being in more rural areas).

5 Tasks and Methodology

5.1 Data Mining: Education Desert Identification

In order to gain a better insight of which census tract features are the most indicative of an education desert, we built 3 ensemble classifiers - Random

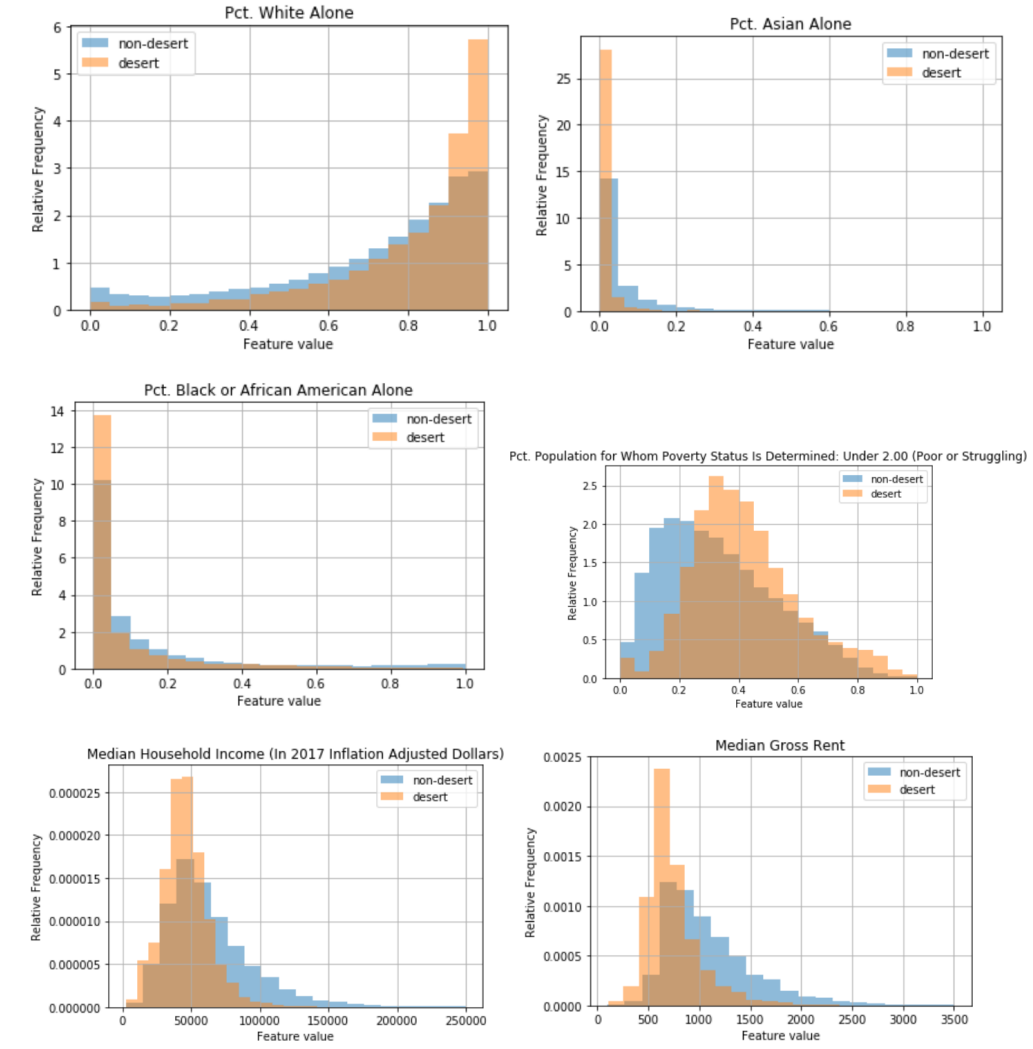

Figure 2: Here, we display various histograms of features that differed significantly between education deserts and non-education deserts

Forest, AdaBoost, and Gradient Boosting - to predict whether a census tract is an education desert or not given its census tract features and the education desert labels from prior the data-preprocessing step.

An imbalance of 10,230 education desert and 63,515 non-education desert census tracts forced us to oversample our data in order to have an equal representation of both education deserts and non-education deserts in our classification.

Out of the 2160 features available from the ACS 2017 (5 year estimates), we chose 113 of them that are likely to help us in our prediction. In addition to the 113 features above, we elected to add on their log-transformed counterparts to the list of features as well. We also applied standard scaling to all features because income related features were on a much larger scale than others

Classification results in Figure 2 indicate there was a good separation of data, given our 5-fold cross-validated Ensemble Classifiers performed quite well, with accuracy >80% and ROC AUC scores >0.8 for the test sets.

	acc-test	auc-test	f1-test
RandomForestClassifier	.805	.805	.800
AdaBoostClassifier	.808	.808	.811
GradientBoostingClassifier	.830	830	.834

To find out which features were more indicative of an education desert, we first looked at the 5 most important features in our random forest classifier, shown in Figure 3. After performing Recursive Feature Elimination with Cross-validation on our Random Forest Classifier, we discovered that the 19 features described here are the most predictive of whether a census tract is an education desert or not.

1. Total Population: White Alone
2. Total Population: Black or African American Alone
3. Total Population: Asian Alone
4. Total Population: 65 to 74 Years
5. Median Household Income (In 2017 Inflation Adjusted Dollars)
6. Median Gross Rent
7. Pct. White Alone

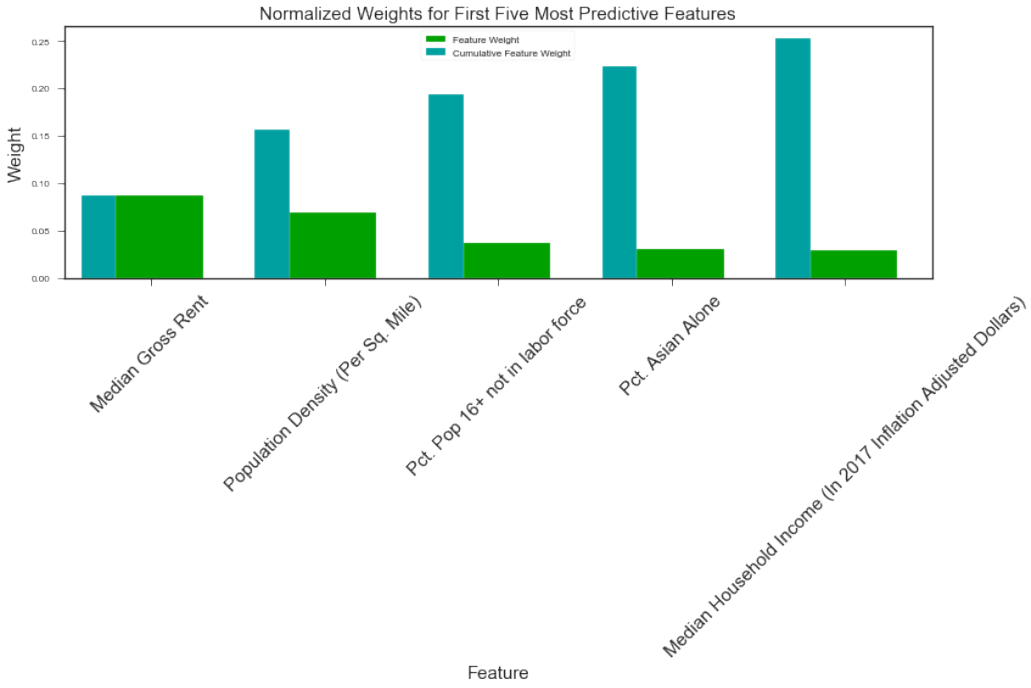

Figure 3: 5 Features Identified From our Random Forest Classifier (Population Density Feature removed)

8. Pct. Black or African American Alone
9. Pct. Asian Alone
10. Pct. 25 to 34 Years
11. Pct. 35 to 44 Years
12. Pct. 65 to 74 Years
13. Pct. 75 to 84 Years
14. Pct. Students enrolled in private school
15. Pct. Population 25 Years and Over: High School Graduate (Includes Equivalency)
16. Pct. Population 25 Years and Over: Some College
17. Pct. Population 25 Years and Over: Bachelor's Degree
18. Pct. Population 25 Years and Over: Master's Degree
19. Pct. Pop 16+ not in labor force
20. Pct. Population for Whom Poverty Status Is Determined: Under 2.00 (Poor or Struggling)

5.2 Predicting Impact of College Locality on College Attainment Rates

Next, we aim to predict whether a local college in an education desert would impact higher-education attainment rates. For every census tract, we compute the percentage of people age 25+ with at least a bachelor's degree. We then use a two-sided t-test to compare these rates between education deserts and non-education deserts, and find that there is a statistically significant difference between the higher education attainment rates between the two groups (Figure 4).

If a new college were built near a current education desert, we would expect graduation rates in those regions to eventually rise to match those of a comparable non-education desert. To forecast this effect, we train a random forest regressor on non-education deserts: the input features include demographic breakdowns, employment levels, income and poverty statistics, and the output is the percentage of the population age 25+ that has a bachelor's

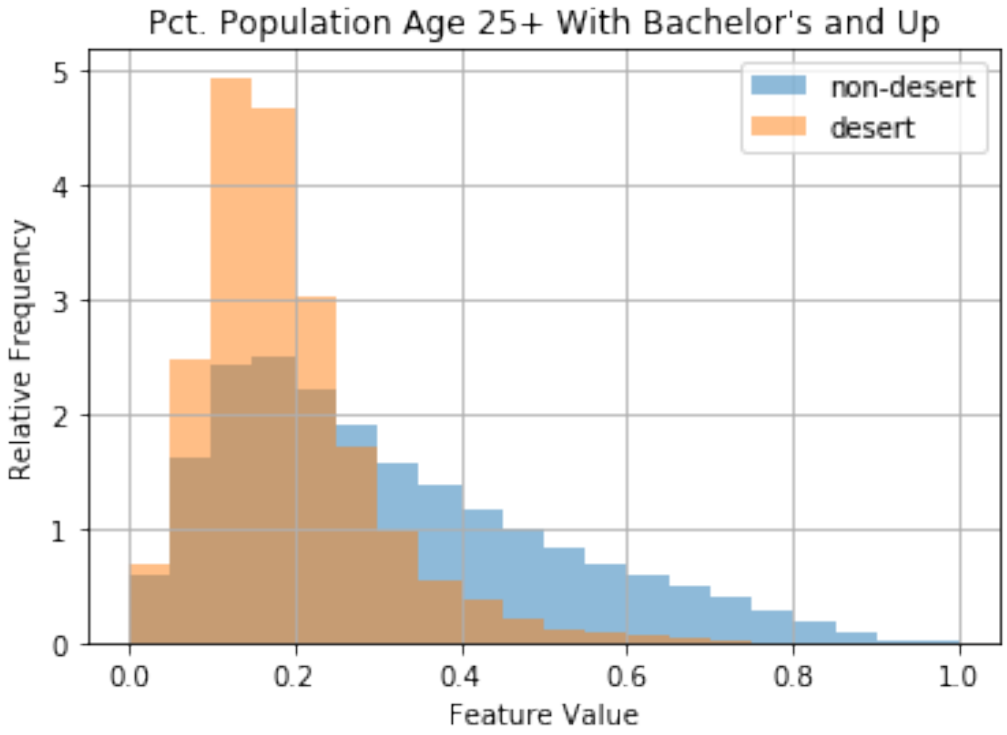

Figure 4: Predicting Bachelor's Degrees in Education Deserts vs. Non-Education Deserts

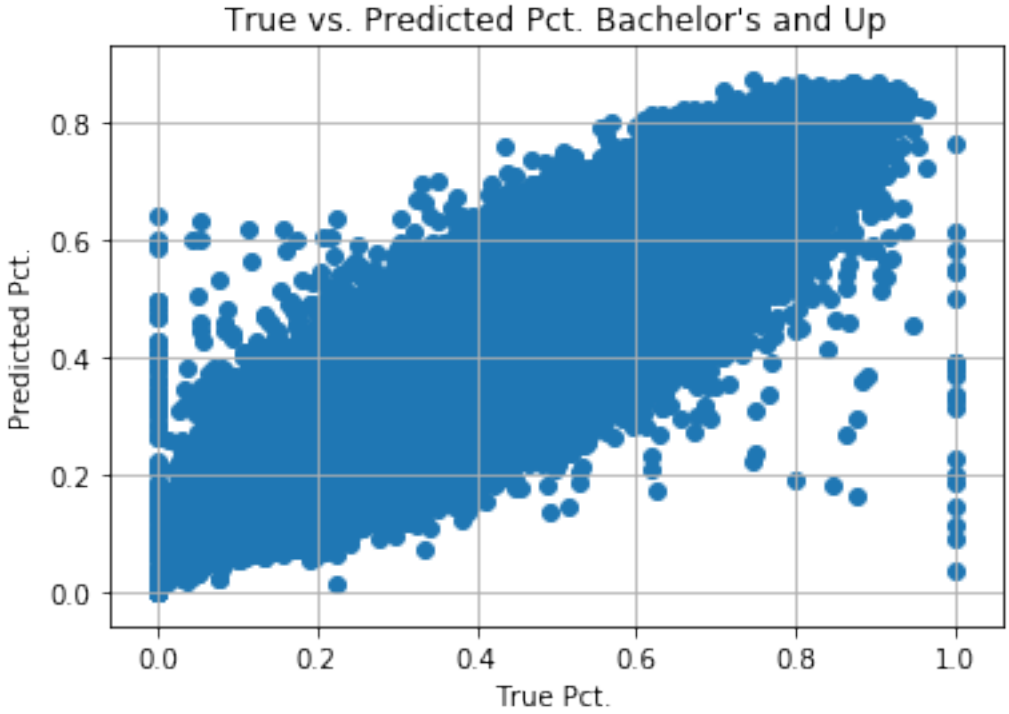

Figure 5: Forecasting Higher Education Attainment Rates Using Existing Colleges For Extrapolating to Education Deserts

degree or higher. We evaluate this regressor using 5-fold cross validation, and achieved an out-of-sample $R\hat{2}$ of 0.83, and a mean absolute error of about 6.0%.

Since this model, in Figure 5, predicts a census tract's higher education attainment rate given that the census tract is not an education desert (i.e. has a college nearby), we can use this model to predict higher future education attainment rates if one were to build a college near a census tract that is currently an education desert. After using this model to forecast higher education attainment rates in current education deserts, we find that it predicts a greater higher education attainment rate than the current rate in 58% of current education deserts (median increase of about 1.0%).

After analyzing the trained model, we discover, in Figure 6, that the most important feature is by far the percentage of households that earn \$200K+/year, followed by the percentage of the census tract's population that is between ages 25-34 and 18-24 respectively, and then the percentage of the population suffering from various levels of poverty. Population density seems to play a role in the predictions as well. Beyond that, other features

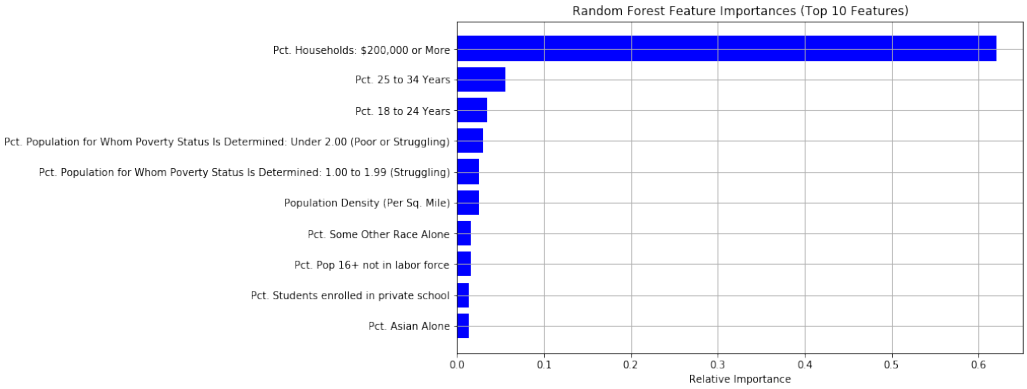

Figure 6: Finding the most important features for educational attainment

do not seem to contribute as much information as the features mentioned above.

These features seem to make intuitive sense, but they are primarily correlative, and not necessarily causal. (E.g. a higher percent of the population aged 25-34 may hint at more college graduates in the area, but this doesn't cause a greater higher education attainment rate.) Since this purpose of the model is meant to forecast the effect of a real-world intervention, we would ideally use a causal model rather than a purely correlative regression. However, for the sake of this study, it does appear that this model does make reasonable enough predictions.

We also attempted to use a different method to forecast the impact of building a college near a school: namely, by using Education Desert as a binary feature (1, if and only if a census tract is an education desert), and then training a similar regression model across both education-desert and non-education-desert census tracts. Since there are more non-education deserts than education deserts, we randomly subsample from the non-education desert group, so that both groups of census tracts are equally represented. Using this method, our RandomForestRegressor achieved a cross-validated $R\hat{2}$ metric of 0.68: significantly lower than before, likely due to the increased variation in the combined dataset.

To generate our final predictions, we then manually set the Education Desert feature to 0 for all education desert census tracts (to emulate the effect of building a college nearby), and then use the same model to predict on this artificially modified data. Using this method, the model predicts that approximately 56% of education deserts will see a rise in their higher education attainment rates: a result quite comparable to the previous method of training a regressor on only non-education deserts. However, upon further

investigation, we see that the feature importance of the Education Desert feature in the RandomForestRegressor is quite low; as a result, the impact of changing the Education Desert feature is very small: the median predicted increase in percent bachelor's and up was only around 0.3%. While this method is theoretically promising, these current limitations led us to stick with the previously described method of training the regressor only on non-education deserts, and then applying that same model to current education deserts.

5.3 Optimizing College Placement

From here we went into choosing the top k locations for our specified optimization objectives, with k specified at runtime. We heavily based these results on the predictions that we made using the outputs of the regressors described in the previous section. After gathering the results of the predicted percentage of the population with a bachelor's degree or higher, we used these values in our different optimization objectives. The different optimization objectives we tried were:

1. The total population with access to a university (as a baseline)
2. The number of added graduates, calculated by multiplying the difference of the predicted percentage of college graduates by the actual percentage of college graduates for a region by the population of the region
3. The added average salary, calculated by multiplying the difference of the predicted percentage of college graduates by the actual percentage of college graduates by the salary of an average college graduate in 2017 according to the National Association of Colleges and Employers, and then adding that to the product of the remaining percentage of the population and the salary of the region, then subtracting the salary of the region from the final value
4. The total added salary, calculated by multiplying the added average salary by the population of the region

If we define the variables that we are inputting into the optimization functions as pop, the total population in a region, pct_pred, the percentage of people in a region who will have a college degree if a college is within 25 miles, pct_real, the current percentage of people in a region with a college degree, grad_sal, the average salary of college graduates from the 2017 NACE

survey mentioned above, and avg_sal, the average salary for a region, the equations for the formulas described above are:

1. total_pop = pop

2. num_added_grads = (pct_pred - pct_real)*pop

3. added_avg_salary = (pct_pred - pct_real)*grad_sal - (1-pct_pred - pct_real)*avg_sal

4. total_added_salary = added_avg_sal * pop

We built a graph where each node was a census tract and each edge represented the census tracts within 25 miles of a specific census tract. We then calculated a weight for each node by taking the sum of the optimization objective for each node and all of its neighbors. After initially sorting the list of nodes by weight, we then picked the top node as our first college. Once a node is chosen, we set the value of the optimization objective for that node and all its neighboring nodes to 0. After this, we re-evaluated each node in the list with the new updates and resorted the list in a lazy fashion, evaluating and resorting until the top node remained unchanged, meaning that no node in the list below it could have a higher marginal benefit. We iteratively used this process to choose the next k-1 nodes.

6 Conclusion and Discussion

In the future, we should experiment with different fairness constraints. A way for us to do this is to add new optimization objectives that explicitly factor in the number of graduates from disadvantaged subgroups that would have access to colleges. Objectives could look both at the number of people with access to a college from a disadvantaged group as well as the expected rise in the number of graduates from disadvantaged groups. Currently, our system may disproportionately place colleges in locations that benefit certain subgroups more so than other subgroups, as we can see in the distribution differences below between the racial demographic breakdown of the entire United States and the breakdown of the census tracts chosen by our optimization model.

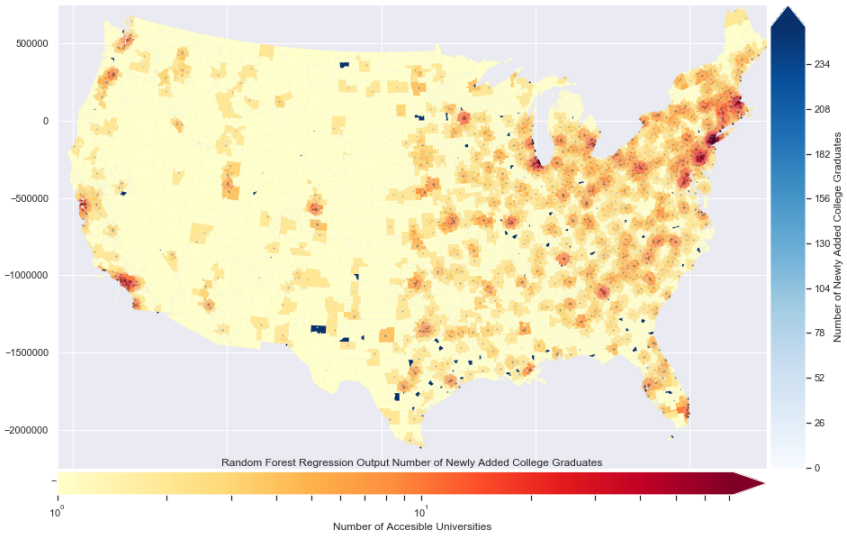

Figure 7: Random Forest Regression Output Number of Newly Added College Graduates

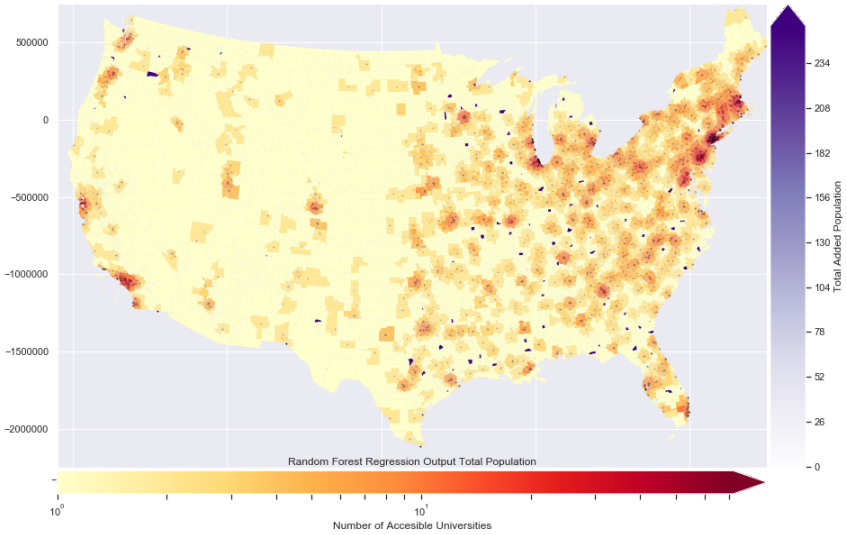

Figure 8: Random Forest Regression Output Total Newly Added Salary

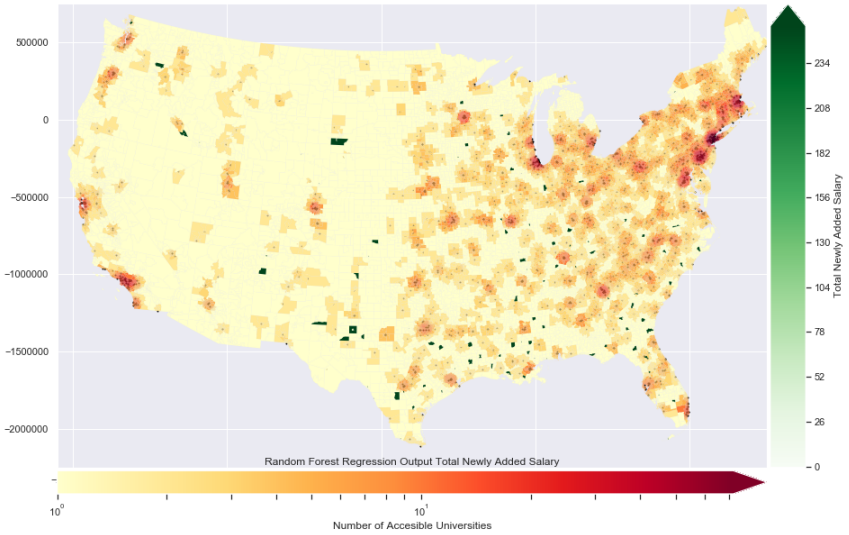

Figure 9: Random Forest Regression Output Total Population

7 References

[1] Klasik, Daniel |& Blagg, Kristin |& Pekor, Zachary. (2018). Out of the Education Desert: How Limited Local College Options are Associated with Inequity in Postsecondary Opportunities. Social Sciences. 7. 165. 10.3390/socsci7090165.
[2] Hillman, Nicholas, and Taylor Weichman. 2016. Education Deserts: The Continued Significance of "Place" in the Twenty-First Century. Viewpoints: Voices from the Field. Washington, DC: American Council on Education
[3] https://libguides.lib.msu.edu/tracts
[4] https://www2.census.gov/geo/tiger/TIGER2018/TRACT/
[5] https://public.tableau.com/s/sites/default/files/media/Resources/IPEDS_data.xlsx

8 Appendix

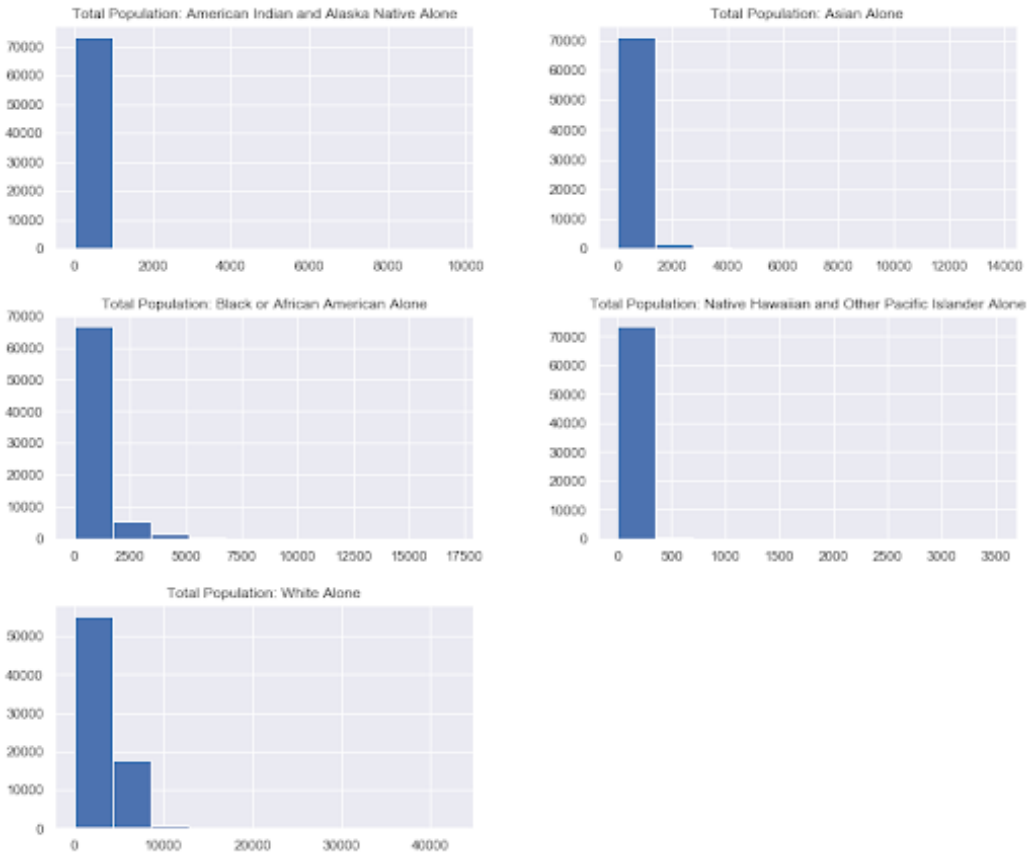

Figure 10: US Population

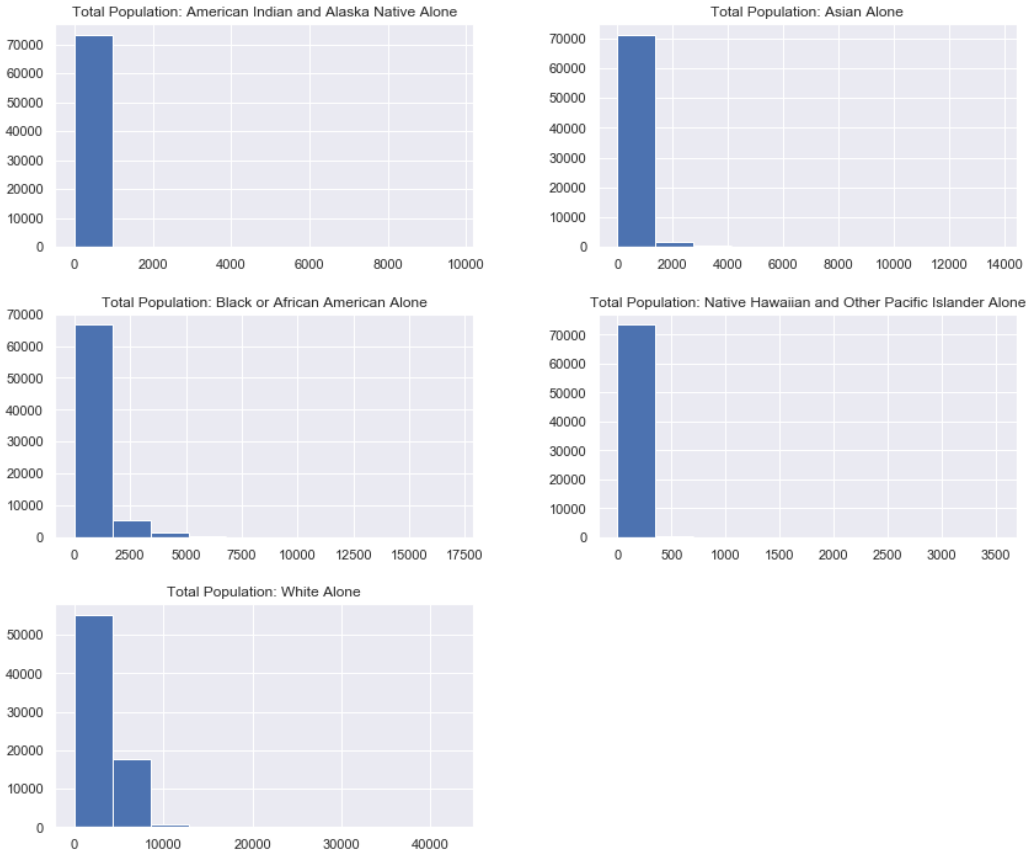

Figure 11: Distribution of University Population

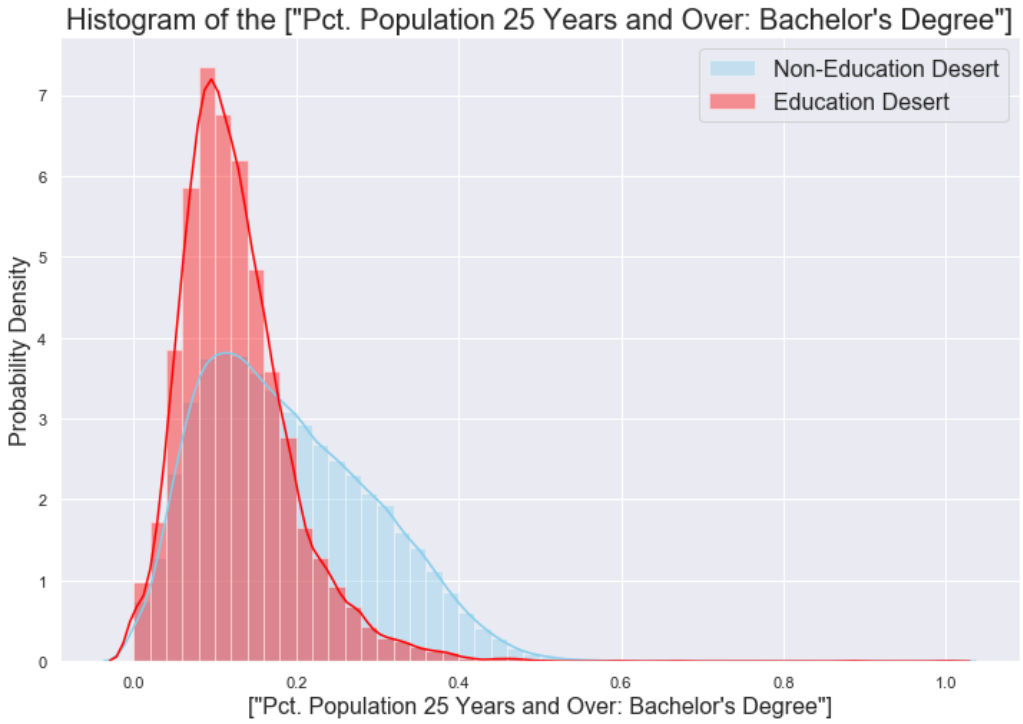

Figure 12: Distribution of Percentage of Population 25 Years and Older with Bachelor's Degree

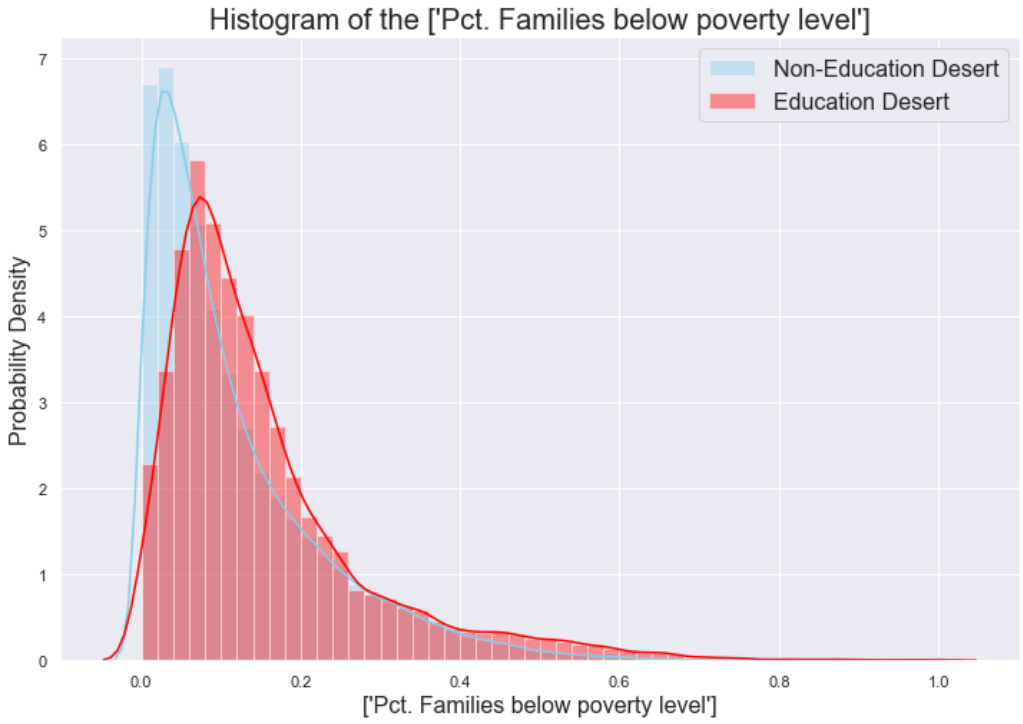

Figure 13: Distribution of Percentage of Families Below Poverty Level

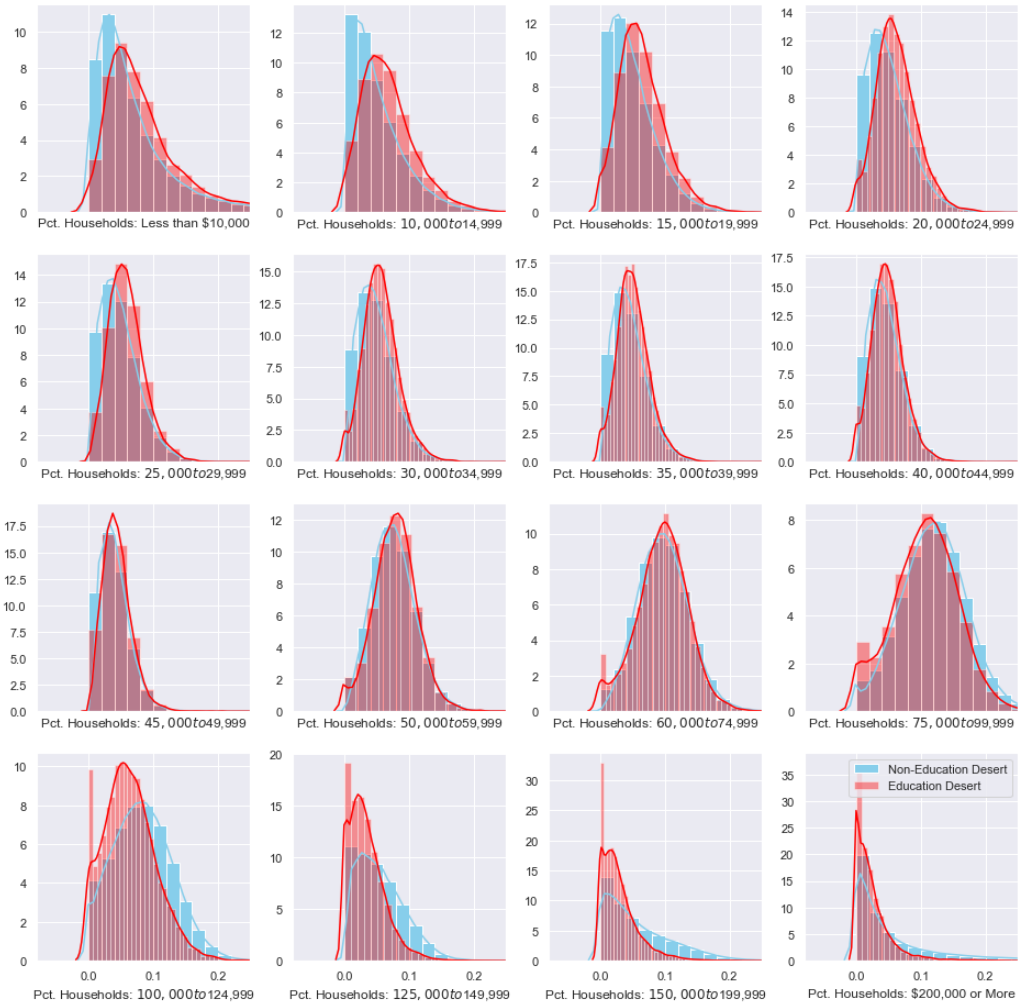

Figure 14: Distribution of Income of Census Tract Households