Predicting Preschool Obesity Rates

Milestone 3: Report Draft

DSC 680 Summer 2021

Chris Briggs

Abstract

Childhood obesity is a topic of increasing public health relevance. This is because childhood obesity is associated with negative physical and mental health outcomes, and rates of childhood obesity have been skyrocketing in the first world, and the US in particular. The need for study and remediation of this topic is particularly urgent today, given behavioral pattern shifts during the COVID-19 pandemic. To this end, this study aims to leverage machine learning algorithms to predict the rate of preschool-age obesity within US counties based on other information about the counties such as median income, median age, percent urban, and so on. A successful model will be a valuable public health tool, enabling policymakers to identify counties which may be at higher risk of increased rates of preschool obesity, and to suggest interventions in counties which are already experiencing elevated rates.

Background

Obesity rates in the US have been increasing over time. As a result, the negative health consequences associated with obesity have been increasing as well ("Adult Obesity Facts," 2021). Body Mass Index (BMI) is, historically, the standard tool used to measure obesity ("Body mass index," 2021). The measure of childhood obesity is not straightforward, as the BMI of children varies naturally during the course of their growth ("BMI Calculator," 2021). Nevertheless, childhood obesity has been increasing over time ("CDC Grand Rounds," 2011). For children under five years old, the WHO defines overweight as at least two standard deviations above average, and obese as at least three standard deviations above average ("Obesity and overweight," 2021). Childhood obesity is not uniformly distributed across the globe, but is rather especially a problem in more developed countries. The childhood obesity rate in the US is roughly triple the global rate (Braun, Kalkwarf, Papandonatos, Chen, & Lanphear. 2018).

Childhood obesity is associated with health impacts such as diabetes, high blood pressure, high cholesterol, liver disease, and sleep disturbances. ("What are the Complications of Childhood Obesity?," 2021). Besides the physical impacts, childhood obesity is also associated with poor self-esteem and depression ("Childhood obesity," 2021). As a result, identifying areas at increased risk for experiencing childhood obesity is a matter of public health interest. A UK study found that childhood obesity was more prevalent in less wealthy primary schools, so the question is also one of equity ("Severe obesity four times more likely in poor primary schools," 2018).

Once counties at risk of experiencing higher levels of childhood obesity are identified, there are effective interventions available (Lambrinou, Androutsos, Karaglani, et al., 2020). Some of these may target declining physical activity levels (Strathclyde, 2019) and increasing device usage (Samsom, 2019).

Statement of the Problem

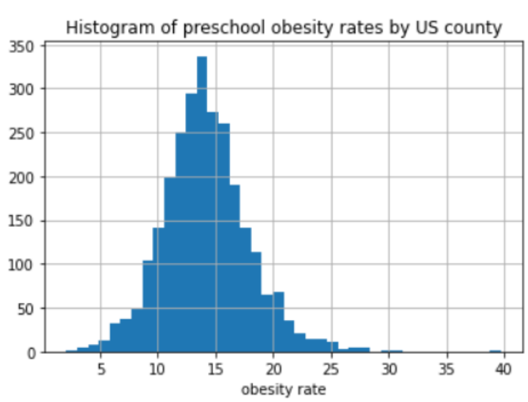
The research question is: what facts about a county will predict an elevated risk of preschool obesity? Conversely, what conditions will predict a lower prevalence of preschool obesity? By answering this question, we may expect to offer a model which will (1) identify counties which may be primed to experience elevated preschool obesity in the near future, and would thus be well-served by intervention efforts, and (2) identify conditions associated with lowered prevalence of preschool obesity, thereby offering suggestions for potential indirect county-wide preventive interventions.

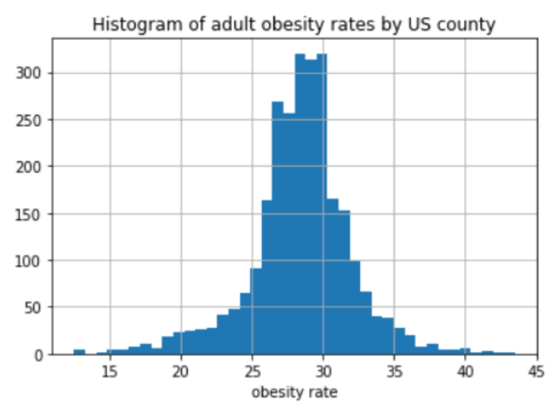
Analysis Approach

I will split the data into training, validation, and testing sets. I will select a number of models to train on the testing set, then optimize hyperparameters on the validation set, and finally test the models on the testing set. This will provide the model most suited to predicting the target variable given the features collected about the counties.

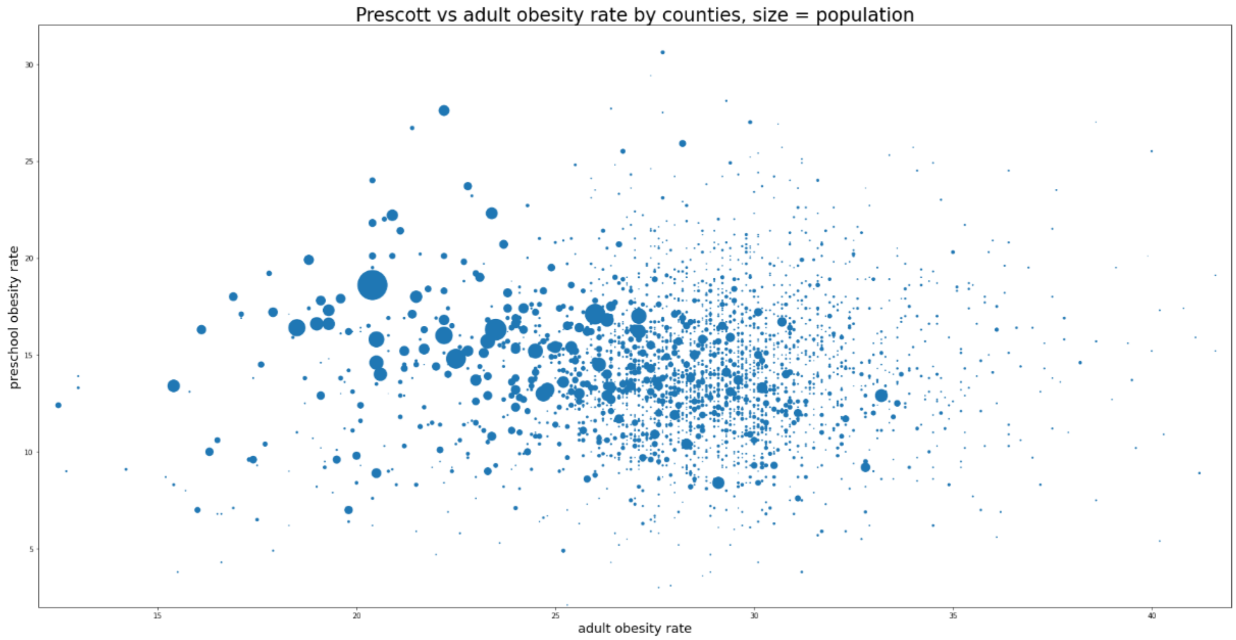
Exploratory Analysis

First, I plotted a histogram for the preschool obesity rate, and for comparison, did the same for adult obesity.





By computation, the mean preschool obesity rate is 14.2% with a standard deviation of 3.7%, and the mean adult obesity rate is 28.4% with a standard deviation of 3.6%. So, while rates of adult obesity are higher, rates of childhood obesity are more variable.
It seems logical that the preschool obesity rate might be related to the adult obesity rate. However, a scatter plot of these two variables shows that they are in fact not strongly correlated.



Exploring correlations further, I am surprised to find that preschool obesity rates correlate weakly with other variables compared to the correlations between adult obesity rates and the other variables. To illustrate, here are the top 20 variables and their correlation with preschool obesity rates:

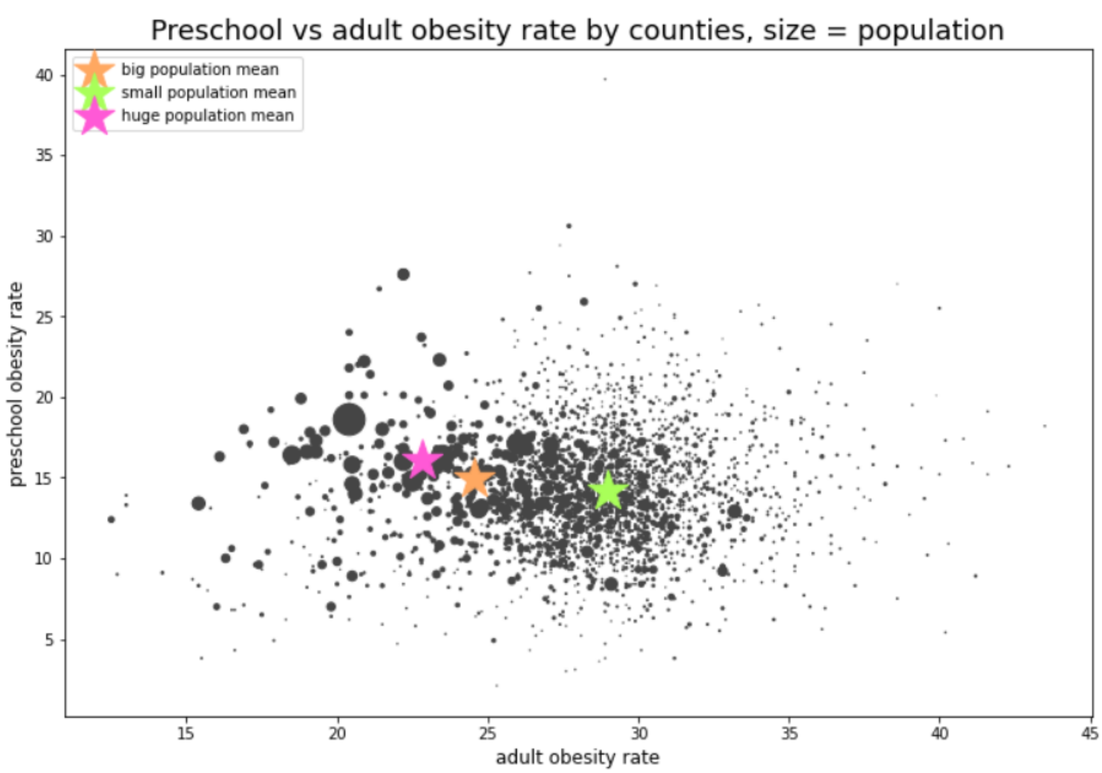
```
 corr      variable
 ----      --------
 0.198    commute_over_sqrt_area
 0.175    foreign_born_pct
 0.12     birth_per_1000_int2
 0.114    longitude
-0.109    land_area_km2
-0.109    land_area_mi2
-0.105    total_area_km2
-0.105    total_area_mi2
 0.105    poverty_pct
-0.095    apartment_rent_ratio
 0.089    birth_per_1000_int1
 0.086    adult_obes_rate
-0.085    latitude
 0.085    rent_1br_usd
 0.081    avg_household_size
 0.081    commute_minutes
 0.081    cost_of_living_usd
 0.079    pop_density
 0.078    urban_pop_density
 0.076    fips
```

The top two variables look as if they may correlate with city size. The 20th largest correlation (FIPS) is of course spurious. None of the variables in the data set correlated very highly with preschool obesity: the coefficients range from -0.1 to 0.2. This is surprising when we compare with the correlates with adult obesity. Here are the top 20:

```
 corr      variable
 ----      --------
-0.602    median_house_value_2017
-0.57     median_house_value_2000
-0.529    rent_3br_usd
-0.52     median_house_income_2017
-0.506    rent_2br_usd
 0.485    poverty_pct
-0.469    median_house_income_ref_val
-0.464    rent_1br_usd
-0.43     cost_of_living_usd
-0.423    income_growth_rate
-0.414    foreign_born_pct
-0.336    pop_growth_rate
-0.313    pop_percent_urban
 0.262    longitude
-0.258    mar_coup_w_children
 0.256    pct_farms_fam_op
-0.254    married_with_kids_ratio
-0.25     pop_in_later_year
-0.25     pop_m
-0.249    pop_f
```

All 20 of these correlates are stronger than the strongest correlation with preschool obesity.

To explore the relationship between county size and child and adult obesity rates, I plotted the mean preschool and adult obesity rates for each of three population size categories: huge (top 10 counties by population), big (at least one standard deviation above average population), and small (smaller than average population).



Although the means in adult obesity vary considerably between the three population size categories, the preschool obesity rate varies comparatively little. It is interesting too to not that there is a light negative correlation between the means - larger cities have higher preschool obesity rates, but lower adult obesity rates.

Analysis Results

I fitted a number of models to the data: linear regression, XGBoost, support vector machine, neural network, random forest, and bagged versions of the last four. The bagged models were optimized for the number of estimators using the validation data set. The resulting accuracies are reported on the validation data set:

```
Accuracy    Model
--------    -----------------
   0.002    linear regression
   0.075    xgb
   0.186    support vector
   0.189    bagging: svr
   0.192    neural network
   0.239    random forest
   0.243    bagging: xgb
   0.247    bagging: neural network
   0.258    bagging: decision tree
```

We see that the most accurate model was the bagged decision tree. The model has been tuned to the validation data, so 0.258 would be a likely upper bound for the performance on yet-unseen data. To get a fair view of the model's accuracy on new data, we run it on the test data set. The result was 0.143, which means that the model explains 14.3% of the variation in preschool obesity rates between counties.

I was surprised by how low this figure was. I trained the same type of model, with the same number of estimators, on the target variable of adult obesity, and this model predicted 68.6% of the variation in adult obesity between counties. So, this demonstrates that preschool obesity rates are difficult to predict from the gathered data, even while adult obesity rates can be predicted rather effectively.

Conclusion

The county-level data has enabled the creation of models which predict preschool obesity rates with very modest success. The best model explains 14.3% percent of the variation of child obesity between counties. The predictive power is not nearly as great as, say, the adult obesity rates, for which the best model can use the same data to explain 68.6% percent of the variation between counties. As a result, an intervention initiative which targets counties at risk of experiencing elevated rates of preschool obesity will not be as effective as those which operate on a smaller level.

Challenges Encountered

Ideally, the product of this study would be an interpretable model. If we can point to a few factors that seem likely to influence preschool obesity rates, then we can begin to explore intervention approaches. The fact that none of the variables is hugely correlated with the target variable means that a model which highlights individual important factors will be more difficult to generate. Indeed, the best-performing model in the analysis was a bagging decision tree, which lacks in interpretability what it offers in accuracy. This is not a total loss: we can still identify at-risk counties by applying the model and viewing the result, then using established effective intervention techniques.

Limitations

I remain surprised at the extent to which preschool obesity is unpredictable at the county level. The positive outcome of this study is: unlike adult obesity rates, preschool obesity rates cannot be effectively predicted from county-level information such as has been gathered in the present data set.

Moving Beyond (how it can be "taken to the next level")

While efficient interventions target adult obesity at the county-level, preschool obesity should be targeted on a more individualized level. A future study should gather household information and attempt to predict preschool obesity rates based on the household data. One potential route is to pair surveyed preschool obesity rates with long-form census data.

References

1. Adult Obesity Facts (2021, June 07). Retrieved from https://www.cdc.gov/obesity/data/adult.html
2. Body mass index. (n.d.). Retrieved Aug 6, 2021 from https://en.wikipedia.org/wiki/Body_mass_index
3. BMI Calculator for Children and Teens. (n.d.). Retrieved Aug 6, 2021 from https://www.stanfordchildrens.org/en/topic/default?id=childrens-bmi-calculator-41-ChildBMICalc
4. CDC Grand Rounds: Childhood Obesity in the United States (2011, Jan 21). Retrieved from https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6002a2.htm#:~:text=during%201963%2D2008.-,In%20the%20United%20States%2C%20childhood%20obesity%20affects%20approximately%2012.5%20million,5%25%20to%20approximately%2015%25.
5. Obesity and overweight (n.d.). Retrieved Aug 6, 2021 from https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight
6. Braun, J., Kalkwarf, H., Papandonatos, G., Chen, A., & Lanphear, B. (2018, May 11). Patterns of early life body mass index and childhood overweight and obesity status at eight years of age. Retrieved from https://bmcpediatr.biomedcentral.com/articles/10.1186/s12887-018-1124-9
7. WHAT ARE THE COMPLICATIONS OF CHILDHOOD OBESITY? (n.d.). Retrieved Aug 6, 2021 from https://childhoodobesityfoundation.ca/what-is-childhood-obesity/complications-childhood-obesity/
8. Childhood obesity (n.d.). Retrieved Aug 6, 2021 from https://www.mayoclinic.org/diseases-conditions/childhood-obesity/symptoms-causes/syc-20354827
9. Severe obesity four times more likely in poor primary schools (2018, Oct 11). Retrieved from https://www.bbc.com/news/health-45822620
10. University of Strathclyde. (2019, Dec 11) All age groups worldwide 'at high risk' of drop in children's physical activity. ScienceDaily. Retrieved from www.sciencedaily.com/releases/2019/12/191211115639.htm
11. Samsom, D. (2019, Feb 20). Screen Time Exposure For Very Young Children Has Doubled Since The Late 90s. Retrieved from https://www.techtimes.com/articles/238729/20190220/screen-time-exposure-for-very-young-children-has-doubled-since-the-late-90s.htm
12. Lambrinou, CP., Androutsos, O., Karaglani, E. et al. (2020) Effective strategies for childhood obesity prevention via school based, family involved interventions: a critical review for the development of the Feel4Diabetes-study school based component. BMC Endocr Disord 20, 52. https://doi.org/10.1186/s12902-020-0526-5

Appendix A

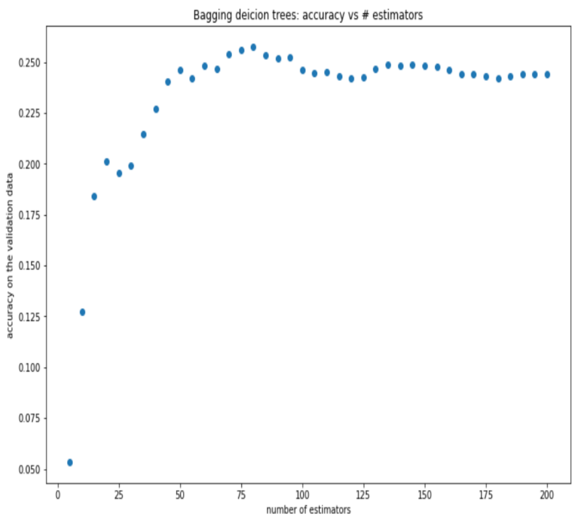
Correlation heatmaps for adult and preschool obesity rates

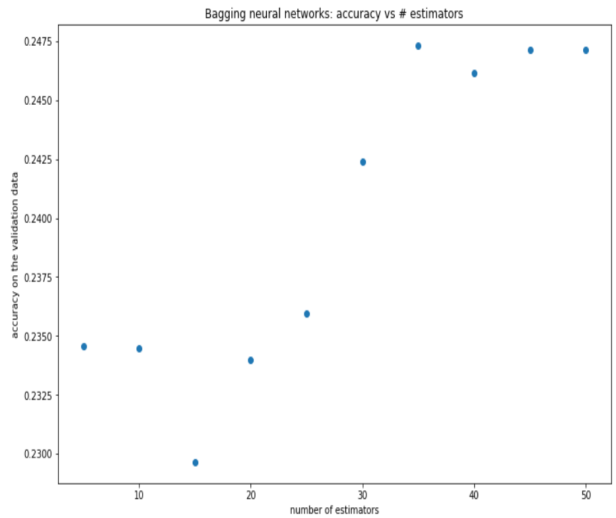
	adult_obes_rate	presch_obes_rate
fips	0.020757	0.076420
pop_in_later_year	-0.249513	0.048470
pop_ref_later_year	-0.048822	-0.025140
pop_f	-0.248906	0.048407
pop_m	-0.250082	0.048524
pop_in_2000	-0.235004	0.044439
median_age	-0.074388	-0.037845
median_age_f	-0.043621	-0.039694
median_age_m	-0.100772	-0.039589
median_house_income_2017	-0.519612	0.020733
median_house_income_ref_val	-0.469243	0.003926
median_house_income_ref_yr	0.019176	0.022352
median_house_value_2017	-0.602284	0.052190
median_house_value_2000	-0.569863	0.022518
avg_household_size	0.021031	0.080824
mar_coup_w_children	-0.257989	0.056783
cost_of_living_usd	-0.430328	0.080954
cost_of_living_yr	nan	nan
poverty_pct	0.485459	0.105175
adult_obes_rate	1.000000	0.085719
presch_obes_rate	0.085719	1.000000
commute_minutes	0.160325	0.080602
pop_per_sq_mi	-0.161027	0.041567
pop_percent_urban	-0.312994	-0.014821
unemploy_rate	0.210449	0.020504
unemploy_date	0.019181	0.022346
rent_1br_usd	-0.464414	0.085128
rent_2br_usd	-0.506202	0.069467
rent_3br_usd	-0.529432	0.054674
avg_farm_size	-0.091784	-0.021160

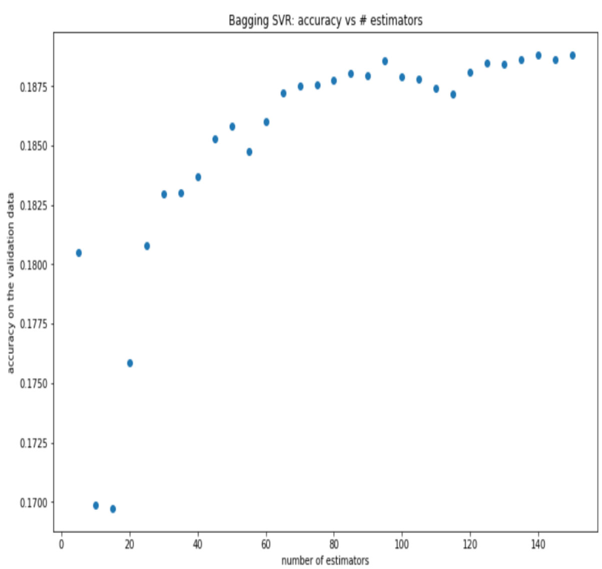
	adult_obes_rate	presch_obes_rate
avg_farm_sales_usd	-0.046212	0.060345
pct_farms_fam_op	0.256303	-0.056612
avg_farm_mach_val_usd	0.050215	-0.020590
birth_per_1000_int1	0.130298	0.088716
births_from_yr_int1	nan	nan
births_from_yr_int2	nan	nan
birth_per_1000_int2	0.126952	0.119521
births_to_yr_int1	nan	nan
births_to_yr_int2	nan	nan
pop_foreign_born	-0.214321	0.075260
land_area_km2	-0.196595	-0.108648
land_area_mi2	-0.196595	-0.108648
water_area_km2	-0.108442	0.007838
water_area_mi2	-0.108442	0.007838
total_area_km2	-0.209419	-0.105306
total_area_mi2	-0.209419	-0.105306
latitude	-0.214833	-0.084585
longitude	0.262488	0.114108
pop_growth_rate	-0.335686	0.008886
pop_density	-0.107603	0.079376
urban_pop_density	-0.108378	0.078388
gender_age_gap	-0.153397	0.001233
income_growth_rate	-0.423371	0.022092
adult_obes_per_pov	-0.140612	-0.050041
married_with_kids_ratio	-0.254343	-0.041360
commute_over_sqrt_area	0.066142	0.197933
apartment_rent_ratio	-0.208755	-0.094744
birth_rate_change	0.000254	0.070625
foreign_born_pct	-0.413849	0.174938

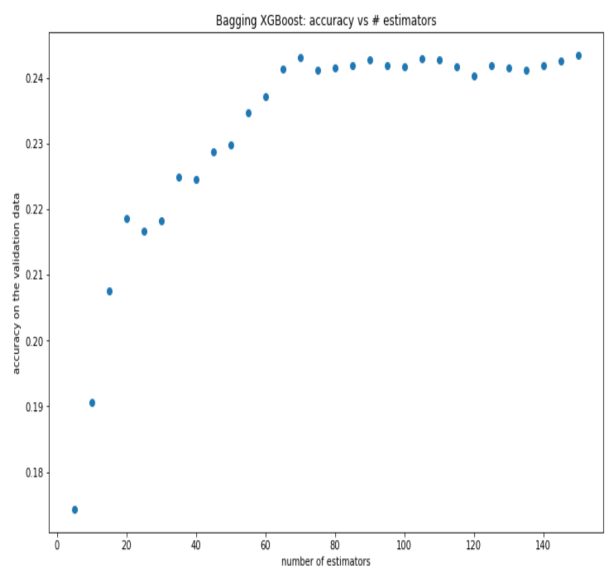
Appendix B

Optimizing the number of estimators in bagged models









Ten Anticipated Questions

1. **Did the data all come from a census?**
   City-data reportedly gathers its data from a mix of public and private sources.

2. **So how reliable is the data?**
   City-data claims that the data is reliable but is not guaranteed to be accurate. As long as there are no systematic biases in the data (as opposed to random noise), our conclusions will be valid, although a bit weaker.

3. **How can the models be made more accurate?**
   More time can be spent trying to optimize the hyperparameters, but I don't think this is the ultimate answer here. Probably the only way to get satisfactory models is to gather more data, and this may mean at the city, district, or household level.

4. **Why is the model so much more accurate for adult obesity rates than preschool obesity rates?**
   I can think of two reasons. First, as noted, preschool obesity is difficult to define. Children do naturally fluctuate a lot in body composition as they grow. This may cause a lot more random noise in the population measurements. Second, where a family lives is probably determined more by the adults than the children, so we might expect the county data to have more to do with any property of the adults there than the children.

5. **Does the study suggest any interventions which might be effective?**
   No, for two reasons. First, none of the features correlated very strongly with preschool obesity. Second, even if they did, this may suggest a route to explore for possible interventions, but it would not prove causation. That is, an intervention acting on a feature correlated with preschool obesity will not necessarily affect the preschool obesity rates.

6. **Why is there a negative correlation between adult and child obesity?**
   Again, this is speculation, but we did note during EDA that adult obesity correlates negatively with county size, while preschool obesity correlates positively with county size. Adults and children live very different lives. It may be that children in smaller counties have more outdoor exercise opportunities, while adults in large counties have better access to gyms and are more likely to commute by means other than a car.

7. **When deriving the extra features, why did you compute commute time over the square root of the county area?**
   I wanted to get a measure of congestion. The linear width of a county goes up roughly as the square root of its area. Commute times increase with linear distance. So, given constant congestion, we'd expect commute times to vary proportionally to the square

root of the county area.

8. **You showed histograms of the preschool and adult obesity rates. Can you comment on the histogram shapes?**
The adult curve looks pretty symmetric, but the preschool curve appears to have some right skew (long right tail) to it. We can't read all that much into this: it's normal for a variable which can't be negative to have a right-skewed distribution. So this could be purely a function of the fact that the preschool obesity rates are closer to 0 than the adult rates are.

9. **Why did you split the data into three sets? I thought usually it's just train and test.**
Given a sufficient quantity of data, it can be useful to split it into three sets. The third set is known as the validation set, and it can be used to tune models - that is, to decide on particular values for hyperparameters. A special case of this which featured prominently in this study was choosing the number of generators in bagging models.

10. **Are there any practical recommendations from this study besides "we should gather better data?"**
Yes, we do see that young children fare just a bit worse in large (by population) counties. This shouldn't surprise us, as large counties may mean high rise apartments, busy parents, and so on. There is a reminder here that children need fresh air and movement outdoors, and school recess may not be giving them enough of it.