Research Paper / Proposal

TITLE

Assessing the Digital Divide in Massachusetts through Data Modeling

ABSTRACT

As a student who has been working on an initiative to provide low-income neighborhoods with proper tech education for the past three semesters, I would like to utilize machine learning techniques to investigate the association between internet access and socioeconomic status Doing so may further reveal ways to remedy the already growing gap in tech education between low- and high-income neighborhoods of Boston. Through this project I would like to:

1) Be able to build a model that can characterize the relationship between socioeconomic status and internet access.
2) Pinpoint which school districts in Massachusetts, and Boston specifically, need more attention based on **internet** access, education and income status.
3) Discover whether or not the prospect of social mobility status is based solely on initial socioeconomic status or access to technology or both.
4) Overall, I want this research project to contribute to advocating for lessening the digital divide and hopefully find ways to determine which neighborhoods need help so that we may find ways to remedy that.

So far we have collected zip code level information from ACS (American Community Survey) [found on census.gov] on income and internet access. In addition to this, we are collecting zip code level data from GreatSchools.org regarding grade-level, rating, and number of awards.

INTRO/LITERATURE REVIEW

MBI or Massachusetts Broadband Institute is a Massachusetts-based organization that aligns very closely with our mission. Their mission is to make high-speed internet access affordable for all. "MBI works closely with the Administration, the state legislature, municipalities, broadband service providers, and other key stakeholders to bridge the digital divide in Massachusetts" (https://broadband.masstech.org/about-mbi). We connected with Jody Jones, a senior program director at MBI, to work closely with their data science team on this issue, and have had a few very productive meetings with her already.

We acknowledge that previous research projects have also attacked this problem of internet access in schools. The most relevant to our project is "Internet Deserts Prevent

Remote Learning During Covid (https://data.aclum.org/2020/05/13/internet-deserts-prevent-remote-learning-during-covid-19/?fbclid=IwAR0Ij2gfLPrFJu110UgR1BmvJJzeUpWjNV91gnp9DWy5wTvYG-gSw7KYbbY)." This article discusses information regarding broadband internet access in low income communities. They essentially connect this lack of high-speed internet to be the cause of lessened opportunity for remote learning during covid.

Other articles that have dealt with the digital divide's impact on health as well: (https://www.samhsa.gov/blog/digital-access-super-determinant-health, The Relationship Between Internet Usage, Socioeconomic Status, Subjective Health and Social Status | Business & Economic Review (bereview.pk)). The first article: "Digital Access: A Super Determinant of Health" entails how important internet access is in the health industry as it addresses how the digital divide affects 19 million Americans. The pandemic increased telehealth use but it was difficult to implement this equitably and in turn impacted the doctor-patient relationship. In the end it connected back to the shared national "Internet for All" initiative. The second study then examines the digital divide in Pakistan by analyzing internet usage patterns based on a variety of factors. The research revealed that there is a significant relationship between internet usage and socioeconomic status, health, and social status. This overall just revealed how all factors of the digital divide are intricately woven together and it becomes very complex in determining which factors actually have a causational relationship with internet access and usage— all over the world.

In the 21st century, especially after the effects of the covid-19 epidemic on schools, the use of the internet and devices has skyrocketed in an academic setting. Unlike other studies, my research aims to discuss the relationship between different factors that could possibly affect this lack of internet access and subsequent lower test scores. We want to investigate if being in a low income community and lack of high-speed internet have a cyclic relationship and how that relates to school performance in the greater Boston area.

PROPOSED METHODS

Initially, I will make scatter plots and box plots to visually assess the relationship between each feature and then use numeric summaries to characterize them, and make basic inferences using these summary statistics.

My machine learning analysis will begin with implementing unsupervised machine learning techniques, such as Principal Component Analysis and/or Factor Analysis to more rigorously investigate the relationships between the features. Our hope is that this analysis can both provide proper context for how our features impact each other, while also potentially uncovering hidden, but important, relationships between them.

The next step would be to utilize the results of the unsupervised learning to guide in building data models that predict specific features of interest; for example, predicting income status based on internet access and education features. More work is needed to properly define the questions we intend to answer at this stage, but we have ample data to approach any number of problems of interest.

DISCUSSION

We strongly believe that internet access, education, and economic status are associated in some way, and expect to learn a great deal about the details of that association through our analysis. Discovering the correlation and which factors cause a higher stress on this issue of the Digital Divide can help us and MBI identify which areas are the most pressing in terms of closing this issue. This will then further Massachusette's presence in the national "Internet for All '' initiative, and in turn, my own ethnographic analysis of this issue in the Boston area.

As for pitfalls that we have already faced, we have had issues in collecting the education data about the nearby school districts. More specifically, finding issues in using methods of gaining this information through online sources where the data has not been made available for public use or issues with web scraping datasets into readable format. As well as this, we could expect difficulty in interpreting the results if there is minimum correlation. This could raise questions regarding the validity of the features we are choosing in terms of the entire feature itself or various aspects of the features. For example, should there be minimal correlation we have to question whether or not we chose a wide enough data set in terms of number of area codes— Should we factor the entire state of Massachusetts? or just a few neighborhoods around Boston?

If we should choose to evaluate or analyze other features to replace these as well, there could be additional technical issues in collecting these data.

What excites you about being a member of the PEAK Experiences Award cohort? What do you look forward to doing or discussing together? (We are open to your ideas as we shape our group activities!) In thinking about these questions, you might draw on previous experiences you've had in groups that generated meaningful connections, as well as your hopes and aspirations for the future. (max. 100 words)

Recalling my time spent as a member of WISE and IDEA, I'm thrilled by PEAK's potential for meaningful connections and shared growth. I am especially looking forward to learning from each other's diverse perspectives— further sharing challenges we've faced and giving any advice we may have. Overall I'm excited to join a community in which we are able to inspire one another's aspirations and help each other leave lasting community impact. Since IDEA and WISE are both communities of structure, I've found

that I love brainstorming with like-minded people to help refine the projects that they are so passionate about!

Recalling my time spent as a member of WISE and IDEA, I'm thrilled by PEAK's potential for meaningful connections and shared growth. I am especially looking forward to learning from each other's diverse perspectives— further sharing challenges we've faced and giving any advice we may have. Overall I'm excited to join a community in which we are able to inspire one another's aspirations and help each other leave lasting community impact. Since IDEA and WISE are both communities of that structure, I've found that I love brainstorming with like-minded people to help refine the projects that they are so passionate about!

Notes from Research Meetings

- Three sources of data
    - Zip Code + income level
    - Presence of computer and type of internet subscription
        - How many people in this zip code have a broadband subscription or internet access

For GreatSchools:

- Will contact for their data

Analysis:

- Information on a zip code level
- Some schools exist on the same zip code
- Figure out how to adapt indicators and sum them up
- How to weight each school by importance
    - Whether they are private/public
- How should we combine schools in the same zip code?
- We start with:
    - Looking at basic correlations between every part of the data
    - Display with plots
    - Explore their linear relationships or no relationship
        - bivariate level
    - What do we want to finally measure?
        - Predict internet access ?
        - Predict school performance with internet access and income level?
        - Or do we not want to predict anything?
            - Unsupervised machine learning
            - Are there clusters of schools in zip codes that have / do not have internet access
                - Random?

Our Research Question:

- Can we predict school performance with internet access indicators while accounting for income
    - Multivariate regression
    - Try to see if the decision tree/random force?
- What is the one indicator you think is most important for assessing school performance?
    - THINK ABOUT JUST ONE

Meeting with scholarships

- MAKE SURE TO TALK ABOUT MBI
- Sai will help you with applying for the grants and the scholarships

Take JSON and clean it into a dataframand output it into a csv using pandas

- Have to go from the json to the data frame

https://medium.com/nerd-for-tech/web-scraping-beautiful-soup-and-selenium-468ed6e0dbef
https://broadbandmap.fcc.gov/home
DR. GERBER

- Looking at great schools data

ME

- Look at the two files with internet access and income data
    - Cleanup the data, work on it

SAI

- Looking into the FFC data and finding which data points to keep/figuring things out with that one
- Look at what can be helpful
-

JODY

- Finished online Survey-not really satisfied with it, not sure if the data is useful or not
- Will send this online survey to us as well for another data source
- Might be a little skewed by bias because not everyone was supposed to finish it (not mandatory)

Jan 22, 2024

- Meeting with MBI
    - Role is mainly analysts
        - Answer questions + statistical analysis on that data
        - https://data.aclum.org/2020/05/13/internet-deserts-prevent-remote-learning-during-covid-19/?fbclid=IwAR0Ij2gfLPrFJu110UgR1BmvJJzeUpWjNV91gnp9DWy5wTvYG-gSw7KYbbY
        -

See what has been done based on the article

- New path that does a better job

- Write a couple questions about what they did + how they did
- Before working on real time data
    - Find real questions with older data (existing data to practice)
        - Can simulate data based on the real data (fake data that we controlled)
- IF they can't give data:
    - We use census data
    - Create data
    - Find other data
- Otherwise use theirs or what they point to

As a student who has been working on an initiative to provide low-income neighborhoods with proper tech education for the past two semesters, I wanted to delve deeper into the relationship between internet access and socioeconomic status in Boston. In specific I would like to investigate the association between the two so I may further predict ways to remedy the already growing gap in tech education between low- and high-income neighborhoods. In specific, through this project I would like to:

1. Be able to build a model that can characterize the relationship between socioeconomic status and internet access.
2. Pinpoint which school districts in Massachusetts, and Boston specifically, need more attention based on internet access, education and income status.
3. Discover whether or not the prospect of social mobility status is based solely on initial socioeconomic status or access to technology or both.

Overall, I want this research project to contribute to advocating for lessening the digital divide and hopefully find ways to determine which neighborhoods need help so that we may find ways to remedy that.

- Mainly we want to investigate a few different factors
    - The relationship between socioeconomic status and internet access
    - If the prospect of social mobility is affected by either of those two things
    - Based on this we can look into specific neighborhoods and specific school districts
        - For example, if we find that a specific neighborhood or town is lacking in technology access, we can take a look at what school district they are a part of and see if that correlates with the quality of the student's scores on standardized exams
        - So that's another thing we can investigate, whether or not students are being affected (academically) based on their access to technology.
        - GOOD SCHOOLS

Data sources

- Broadband now
- Roberto Gallardo Purdue center for Research Development
-
- Come up with a couple pages of report for some correlations between the data
- Gerber will provide the books/resources for Principle Data Analysis
-

Think about questions you wanna answer

- With what measurements do we have match up to the questions hat we have
    - Mental feature reduction

    - Do we really need to look at all the data?
- Columns that are more helpful
    - Median income > total number of households

Keep options open when working with missing data

- Keep a master data set that is separate from any data manipulations that you do

Things to do in the next couple weeks

1. Take the data that Sai provided
    1. Exploratory data analysis + thinking about the questions
        1. Which features in the data sets seem to align with those questions
            1. Don't get too many features and not too little
        2. Pick 3 or 4 plots that you create that seem to demonstrate what we might find
            1. Create a 2 page report, create a plot, what you think the plot means, plans for further investigating
            2. After get back in tough with jody and then share with jody your entry level investigations
2. Scatterplot or two, box plots, distribution of features on the same scale
    1. Fleshed out understanding
    2. Across all the zip codes we are seeing these relationships
3. Can look at everything but education because everything else is merged
    1. Look at income and internet data
    2. Can also sit tight and wait for the cleaning and merging of the data

We want to combine the schools via zip code (all schools combined within the same zip code)

- Make sure that we have the ratings and the number of students
- Rating, award, zip code + extract from first column

Feb 5, 2024

- Get the data in the form where we can start making basic summaries
- Create scatterplots
- (Weeks 1-2) The education data should be cleaned so that:
    1. each row of the data is a single zipcode (i.e. it goes from school-level data to zipcode-level data)
    2. as part of that process, we should:
        1. average the ratings of the schools in each zipcode
        2. sum the number of awards of the schools in each zipcode
        3. count how many public and private schools are in each zipcode
        4. use the total students and student teacher ratio columns to calculate the number of teachers at each school, then sum the number of teachers and total students in each zipcode, then recalculate a student teacher ratio for each zipcode

    3. perhaps consider getting more data from all of Massachusetts in this data set, since our other two data sets have income and internet access for all zipcodes in the state, not just around Boston
- (Week 3) Update Jodi and MBI about our progress and suggest that we'll have some results in a week or two to go over with her. Make sure that all three data sets (income, internet access, education) are cleaned so that they are on the zipcode-level and that the zipcode column is formatted the same so that we can join them using that column
- (Weeks 4-5) Select a few (about 5-10) columns from across the three data sets to focus on building some basic data exploration summaries (both numeric and graphical) including investigating correlations and scatter plots, means, distributions, boxplots, etc.
- (Week 6) Once that is completed, writing up a short ~2 page report with the key data exploration takeaways to share with Jodi.
- (Week 7-9) Moving on to the unsupervised learning aspect of the project.
- (Week 10) Write up a final report/poster/abstract and discuss where to present/publish results (potentially in collaboration with Jodi and MBI)