Predicting Internet Speed: An Analysis of Demographic Influences

Stephanie Ramos
MSCAPP Candidate
University of Chicago
stephanieramos@uchicago.edu

Macarena Guzmán
MSCAPP Candidate
University of Chicago
macaguzman@uchicago.edu

Mayarak Quintero
MSCAPP Candidate
University of Chicago
mayarakquintero@uchicago.edu

Abstract

The pandemic is affecting the way a myriad of economic and social activities are carried out. Internet communication is solving many of these problems. Therefore, the speed of the Internet is essential to perform these activities successfully. This study implements a machine learning framework to predict Internet speed levels using demographic, economic, technology and housing data. Using regression models (linear, Ridge, and logistic) and a support vector machine, we were able to predict county internet speed levels with 62% accuracy. We found that demographic variables related to education, race, and age are good predictors of internet speed in US counties. Additionally, we do predictions for counties where internet speed data is not available. This kind of analysis is particularly relevant as the nation experiences an unprecedented shift towards work and online learning.

1 Introduction

Since the beginning of the Covid-19 pandemic earlier this year, many activities that are normally carried out in physical presence are taking place online. Consequently, the need for high speed internet connection has increased. Access to a fast internet connection has become crucial in keeping parts of the economy going, allowing large groups of people to work and study from home, enhancing social connectedness, among others.

Internet speed is highly correlated with demographic characteristics such as education level, income and race. Research by the U.S. Census Bureau found that counties with lower median household incomes (<$50,000) had household internet subscription rates on average ten percentage points lower than counties with median household incomes of $50,000 or more.[1] This study implements a machine learning framework based on classification models to predict internet speed levels using demographic, economic, tech-related and housing data from the U.S census bureau and internet speed data from OOkla at the county level.

This analysis is particularly relevant as the nation experiences an unprecedented shift to remote work and online learning. With the COVID outbreak, the labour market and school conditions are particularly weak and broadband connection can help to preserve wealth. Nationwide access to a fast internet connection is crucial to protect and foster economic productivity.

[1] Available at https://www.census.gov/library/stories/2018/12/rural-and-lower-income-counties-lag-nation-internet-subscription.html

Preprint. Under review.

2 Literature review

Literature is scanty assessing demographics and internet speed. Itoku, et al. (2019) use geographic and demographic data on a census block level to predict broadband expansion and find that effective models can be constructed from this data to understand previous broadband expansion and future expansion. They use gradient boosting classifiers to make the predictions.

Gadiraju et al. (2018) use machine-learning techniques to address both the questions of who gets broadband and where. Their study uses the FCC Form 477 data on broadband offerings and the Census's Planning Database data on demographic and economic metrics. They find that median home valuation and percentage of recent home constructions have a strong correlation with the introduction of Internet services with threshold speeds.

Kolko (2010) uses December 2005 FCC data and individual broadband adoption data from Forrester Research to estimate residential broadband availability by ZIP code and finds that population density and average income have strong positive effects to estimate broadband availability.

Gillet et al. (1999) conduct a county-level research project of the MIT Internet Telecoms Convergence Consortium of the patterns of deployment for broadband Internet access in the United States and find that the demographics of certain communities determine the availability of broadband. Where available, broadband is concentrated in areas with higher income and higher density markets, factors that proved to be relevant to know how broadband availability changes over time.

This study is related to the above literature in that it uses demographic data to predict internet speed. Contrary to most existing literature, we use internet data and demographic data at the county level to make predictions. Similar to the existing literature, we found evidence that demographic, economic and housing features are relevant in predicting internet speed. Further, we found that including additional tech-related variables, such as number of smartphones per home and number of laptops per home, adds additional predictive power to the models.

3 Data

This study uses demographic data to predict internet speed in U.S counties. The internet speed data comes from the *Speedtest by OOkla Global Fixed and Mobile Network Performance Map Tiles*. We used data from the second quarter of 2020. Ookla aggregates the hundreds of millions of speed tests that are taken on their platform each month into tiles, which are approximately 610.8 meters by 610.8 meters at the equator. To obtain data at the county level, we calculated the weighted mean, weighted by test count, of the internet speed in all tiles belonging to a each county. This gives us the overall mean download speed of all tests performed in a county if the data had not been aggregated to tiles in the first place. Then we separated internet speed data into five buckets which are the labels that we will predict.

Our second source of data is the U.S Census Bureau. The *American Community Survey (ACS)* contains demographic, economic, and housing data, which we use to predict internet speed. This data includes factors such as percentage of white residents, percentage of black residents, median income, median age, percentage of high-school graduates, number of computers and smartphones, and median home value. We used the 2014-2018 ACS 5-Year Estimates.

4 Machine Learning Approach

Using the speedtest data from OOkla we can determine how fast is the internet connection in each county. We separated the internet speed data into 3 buckets: 0 to 50 Mbps, 50 to 100 Mbps, and more than 100 Mbps. We trained our models to predict which of these buckets a county belongs to. We use demographic, economic and housing data from the Census Bureau as predictors. In particular, we use eight features in our models: percentage of white residents, percentage of black residents, median income, median age, percentage of high school graduates, and median home value, number of households with computers and number of households with smartphones.

We implemented 4 classification models: linear regression, Ridge regression, logistic regression and support vector machines. We assigned a number (1, 2 or 3) to each internet speed bucket and we used those numbers as the labels. Given that the linear regression and Ridge regression

model give continuous predictions, we used a classification rule to calculate the predicted label.[2] For parametric models, we used grid-searching to calculate the best hyperparameters. We used 10 fold Cross Validation to calculate the accuracy score of each model (percent of correct predictions). Finally, we used the best performing model, the one with highest accuracy, to predict the internet speed of counties that don't appear in the OOkla dataset.

5 Evaluation and results

As explained in earlier sections, this study aims to accurately predict the level of internet speed in each county. The models were trained using the following features: percentage of white residents, percentage of black residents, median income, median age, percentage of high school graduates, and median home value, number of households with computers and number of households with smartphones. We ranked the features according to the magnitude of their weights. Table 1 summarizes feature importance.

Table 1: Feature ranking

Feature Name
1. Percentage of high school graduates
2. Percentage of black residents
3. Median age
4. Households with smartphones
5. Median home value
6. Median income
7. Percentage of white residents
8. Households with computers

We used the accuracy score to evaluate the performance of the ML models. Accuracy is the number of correctly predicted data points out of all the data points. Table 2 summarizes the results of all training models after tuning the hyperparameters of each model. The model with highest accuracy is a Multinomial Logistic Regression with a penalization parameter of 10^1, this model correctly predicted 62% of the testing samples. The worst performing models are Linear regression and Ridge Regression, these models correctly predicted only 49% of the testing samples.

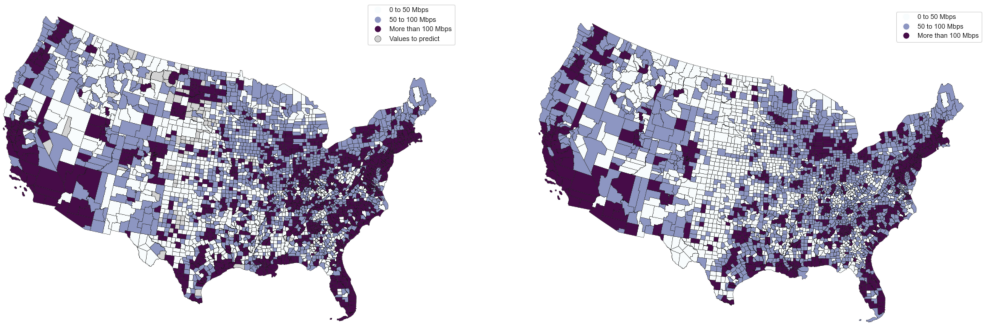
Table 2: Accuracy of predictive algorithms

Model	Accuracy
Linear Regression	49%
Ridge Regression	49%
Logistic Regression	62%
Support Vector Machine	63%

Logistic Regression with multiclass labels works better than other linear models because it performs $n-1$ independent binary logistic regressions against a given label (one-vs-rest scheme), where n represents the number of labels. Then, the model constructs a coefficient from a set of weights that are linearly combined, taking the dot product between those weights and the features. On the other hand SVM tries to find the separating hyperplane that maximizes the distance of the closest point to the margin.

Considering that SVM has higher accuracy than the other models and doesn't penalize examples for which the correct decision is made with confidence like the other models, which is good for generalization, we selected SVM to predict the level of internet speed for counties that do not appear in the OOkla dataset. Figure 1 shows the predicted internet speed levels for the training and testing data and compares them to the true internet speed levels. It is important to note that the accuracy of the model is lower when predicting internet speed levels of counties that are located in the central parts of the country while accuracy is very high when predicting counties located along the coasts.

[2]Predicted values were rounded to obtain the predicted labels.



(a) True internet speed (b) Predicted internet speed

Figure 1: True and predicted internet speed levels across U.S counties

6 Conclusion

Since the COVID-19 pandemic began, access to a fast internet connection became crucial in keeping parts of the economy going, allowing large groups of people to work and study from home and enhancing social connectedness. The speed of the Internet is now essential to perform these activities successfully.

In this study, We analyzed various different demographic, economic and housing features to determine the internet speed level of the U.S counties. We trained four different ML models and we were able to predict the internet speed level with accuracy of 62%. We found that demographic variables related to education, race and age have the most predictive power in our models. In addition, we were able to use our models to predict the internet speed level of new counties that did not appear in the internet speed dataset. Our model predicted more accurately internet speed in counties that are located along the coasts than counties located in the center of the country which rises some concerns related to algorithmic bias.

Broader Impact

This analysis unveils differences on internet access and broadband speed across income segments and races. We believe access a fast internet connection is paramount to develop a more equal and prepared society, thus, this analysis should put pressure not only on internet providers, but also to local governments to improve connectivity for the under-serviced counties.

We expect both a positive and a negative outcome. A positive outcome would be better public policies and private initiatives that will help reduce the gap between high and low income, and among different races. A negative outcome might be that due to COVID-19, companies are leaning towards online and remote work. This new trend might lead companies to hire people that live in or close to counties with better internet access so they can be fully connected anytime.

This analysis could be more accurate if we could go deeper from counties to a block level, as more granular data would help us to better identify where exactly in each county is the disadvantaged population and which counties are being truly under-serviced. One concern is that our model is not fully capturing the relations between demographic characterising and internet speed for counties that are located in the middle part of the country, making less accurate predictions for this sector. The model might not be accounting for counties that might have low-density populations or are geographically unable to have reliable internet access (e.g. mountains, or lakes), or we are not accounting for counties with people less interested in better internet access (e.g. senior citizens).

References

[1] Eisner Gillett, Sharon and Lehr, William, *Availability of Broadband Internet Access: Empirical Evidence* (September 27, 1999). Available at SSRN: https://ssrn.com/abstract=3215923 or http://dx.doi.org/10.2139/ssrn.3215923

[2] Gadiraju, Vamsi and Panat, Anthony and Poddar, Raghav and Sherriff, Zain and Kececi, Sam and Schulzrinne, Henning, *Who Gets Broadband When? A Panel Data Analysis of Demographic, Economic and Technological Factors Explaining U.S. Broadband Deployment* (August 15, 2018). Available at SSRN: https://ssrn.com/abstract=3142479 or http://dx.doi.org/10.2139/ssrn.3142479

[3] Itoku, E. K. and Mantena, Aman Varma and Sadholz, Aaron and Zhou, Anna and Schulzrinne, Henning, *Predicting Broadband Expansion: An Analysis of Geographic and Demographic Influences* (August 21, 2019). TPRC47: The 47th Research Conference on Communication, Information and Internet Policy 2019, Available at SSRN: https://ssrn.com/abstract=3440399 or http://dx.doi.org/10.2139/ssrn.3440399

[4] Kolko, J. *A new measure of US residential broadband availability* (2010). Telecommunications Policy, 34(3), 132–143.