PREDICTING U.S. BROADBAND AVAILABILITY FROM ECONOMIC CENSUS DATA

ABSTRACT

Does the percentage poverty rate of a region influence broadband internet supplier locations? How about rate of health insurance, use of supplemental food programs or unemployment?

This report summarizes the findings of a Jupyter notebook analysis regarding the accuracy of seven popular modeling algorithms in their ability to identify counties likely to have broadband availability given key economic measures.

CS 677 Data Science with Python

Edward T Myers

Table of Contents

Table of Figures

Introduction

Background

I have a background in local government in central Pennsylvania, and finding gaps in general services within our constituency is especially useful as we have districts with varying degrees of financial distress. Understanding the correlation between several 'basic needs' census categories and how they relate to broadband services has become prevalent with the recent *Get Internet : Affordable Connectivity Program*, championed by President Biden's Administration (The White House, n.d.). As an extension of the American Rescue Plan funding (The White House, 2022), local governments would benefit from gaining knowledge on where broadband services exist and where to supplement efforts.

Project Inspiration Research Articles:

- *Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms* (Lee, Resnick, & Barton, 2019).
- *Machine learning and algorithmic fairness in public and population health* (Mhasawade, Zhao, & Chunara, 2021).
- *A clarification of the nuances in the fairness metrics landscape* (Castelnovo, et al., 2022)

Project Goals

1. Visually analyze data to observe correlations and select appropriate model algorithms.
2. Compare data overall and using several predictive numeric models (K Nearest Neighbor, Logistic Regression, Gaussian Naive Bayes, etc.)
3. Execute training and testing of model classifiers and analyze confusion matrices to determine the most accurate algorithms.
4. Use the models to predict broadband availability in a specific example, my home county (Cumberland County, Pennsylvania).

Dataset

- "US Broadband Availability" (https://www.kaggle.com/datasets/mmattson/us-broadband-availability) (Mattson, 2021)

Preparing the Data

Data Overview

The four leading columns are categorical information (full_name, county, state, state_abr). The rest is quantitative data representing actual amounts or percentages of population. A list of the column header explanations is in the corresponding document “Column Explanations.txt”. This data set comes from a combination of data sources compiled by the author:

1. The Institute of Museum and Library Services (IMLS): that blended indicators about economic status from the U.S Census Bureau American Community Survey (ACS), FCC data from BroadbandNow.com (BBN), and local unemployment from the Bureau of Labor and Statistics (BLS). This data blends instantaneous data from 2019 and aggregate census data from 2014-2018 (Mattson, 2021).
2. An updated dataset from BroadbandNow.com: which contained parallel data regarding population and broadband availability at the time of sample (2020), adding numerous classifications for broadband data such as providers and speeds (Mattson, 2021).

As a result of this data blending, several columns have similar or duplicate data from different data sets or time periods.

Data Cleaning

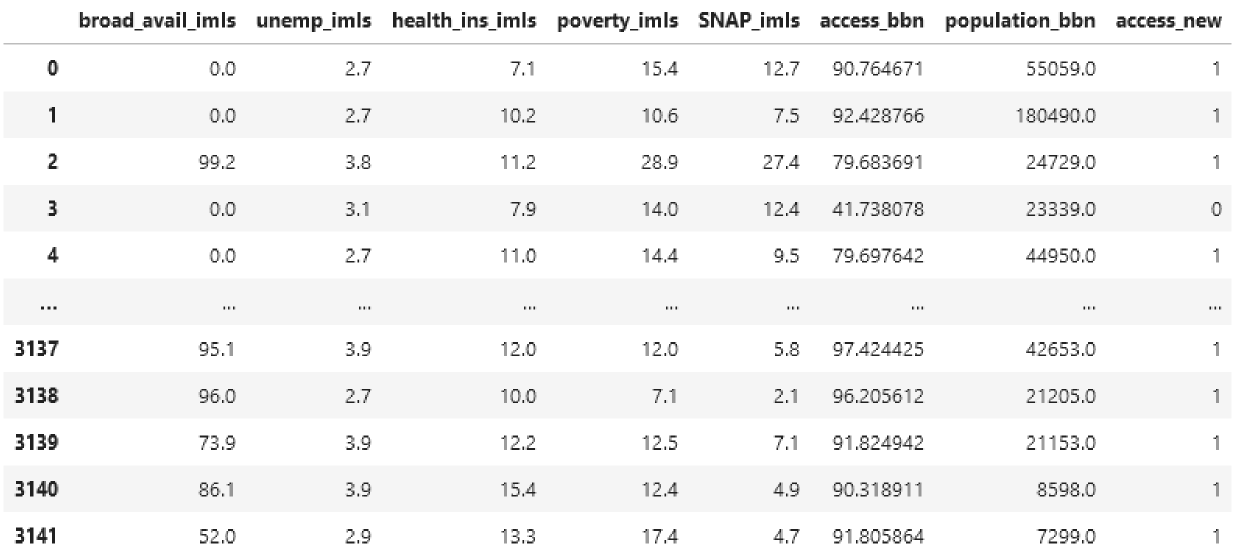
Several of the categorical data columns are redundant, such as full_name and state_abr. Those columns will be removed from the data frame upon import. Several rows do not have data available (NULL or NaN). As these will function as the variables which trains most of the models, rows with NaN entries will be removed. The class determiner derives from the "access_bbn" data column for broadband availability. Other extraneous columns from the BBN and ACS portions of my data will be removed. For clarity, the column names will be appended with their source data abbreviation.

Creating a Subset with Population Metrics

The key quantitative factors that will be compared against broadband availability are:

1. Population count
2. Percentage without health insurance
3. Percentage of poverty rate
4. Percentage rate using the supplemental nutrition assistance program (SNAP) (Benefits.gov, n.d.)

Broadband availability in terms of the data set is measured in a percentage. For the purposes of this project, the percentages will be transformed into a binary class label. Looking at the median behavior of the two columns related to broadband availability, the 2019 data shows a median of 81% while the 2020 data shows a median of 92%. From the statistics, the binary class label created from either set should skew well towards broadband being available than not available. The 'access_bbn' column will be used as it contains more current data. From this column, a new column named 'access_new' will be populated with ‘1’ if the percentage is above or equal to 50%, and ‘0’ if the percentage is below 50%. Figure 1 shows the prepared data set.



	broad_avail_imls	unemp_imls	health_ins_imls	poverty_imls	SNAP_imls	access_bbn	population_bbn	access_new
0	0.0	2.7	7.1	15.4	12.7	90.764671	55059.0	1
1	0.0	2.7	10.2	10.6	7.5	92.428766	180490.0	1
2	99.2	3.8	11.2	28.9	27.4	79.683691	24729.0	1
3	0.0	3.1	7.9	14.0	12.4	41.738078	23339.0	0
4	0.0	2.7	11.0	14.4	9.5	79.697642	44950.0	1
...	...	...	...	...	...	...	...	...
3137	95.1	3.9	12.0	12.0	5.8	97.424425	42653.0	1
3138	96.0	2.7	10.0	7.1	2.1	96.205612	21205.0	1
3139	73.9	3.9	12.2	12.5	7.1	91.824942	21153.0	1
3140	86.1	3.9	15.4	12.4	4.9	90.318911	8598.0	1
3141	52.0	2.9	13.3	17.4	4.7	91.805864	7299.0	1

Figure 1 - Analysis Dataframe w/ 'access_new' Binary Class Label

Training Data & Observations

The analysis will split the training and testing data in a 30%/70% ratio. The results will be stratified to ensure a fair and distributed sample. A scaler will be used for all model classifiers regardless of whether it is needed or not for a particular algorithm to ensure an honest comparison of accuracies. A random state seed will be set to make the results repeatable. Next, a pair plot and correlation matrix will be used to identify strongly correlated variables. If near-perfect correlations exist within a set of data columns, one or more of the duplicates can be omitted as they would not have a discernable impact on the model. Efficiency would also be improved.

Observations (see Figure 2):

- No two dependent variables share a near-perfect positive or negative correlation as to remove a duplicate from consideration. All of the variables will be retained for the model analysis.
- Population demonstrates the weakest correlation to any other variable. This may be due to the nature of the data being a count in the measure of thousands versus percentages that don’t go above 100.
- The strongest correlation is between poverty rate and SNAP at 0.81, which makes sense as use of such a program is often attributed to a lower income.

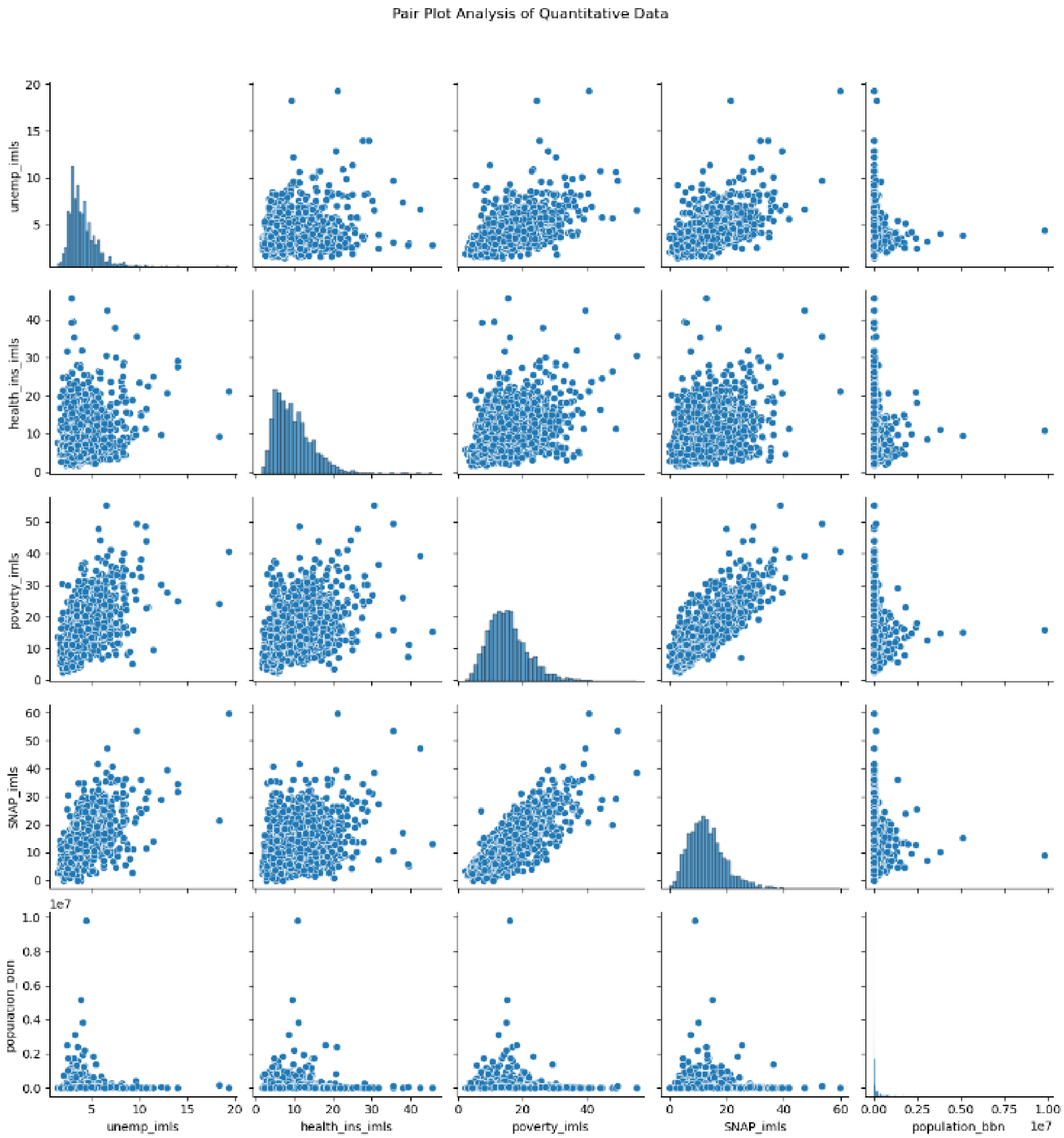


Figure 2 - Pair Plot Analysis of Independent Variable Data

Modeling the Data

Model Selection

As the data dependent variables are quantitative and the independent variable is a binary classifier, the models that will be employed to classify the data are:

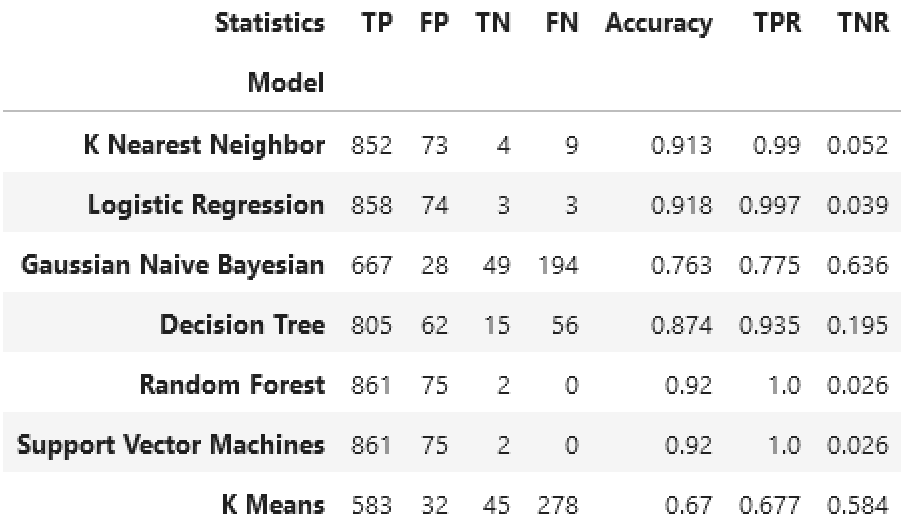
1. K Nearest Neighbor
2. Logistic Regression
3. Gaussian Naive Bayes
4. Decision Tree
5. Random Forest
6. Support Vector Machines
7. K-Means

Model Considerations

Several strategies were used where appropriate to evaluate the ‘fit’ of the model to the accuracy of the test data prediction. When using K Nearest Neighbor, a looping analysis of accuracies using specific numbers of ‘k’ were graphed to determine the best ‘k’ fit without seeing diminishing returns in accuracy. With the Decision Tree classifier, multiple loops were run to compare the data splitting method and the algorithm criterion (scikit learn, n.d.). In Random Forest, values of ‘N’ and ‘d’ were selected from the lowest produced error across multiple algorithm criteria (scikit learn, n.d.). For Support Vector Machines, high accuracy was chosen among the various kernels that define the SVM classifiers (scikit learn, n.d.). Finally, K-Means used a knee plot to show most efficient reduction of inertia with respect to ‘k’ number of clusters using all applicable algorithms (scikit learn, n.d.).

Each model’s section includes an accuracy report on how well it predicted the class label of the test data, as well as a confusion matrix to demonstrate its level of precision with the counts of correct positive outputs.

Model Comparison

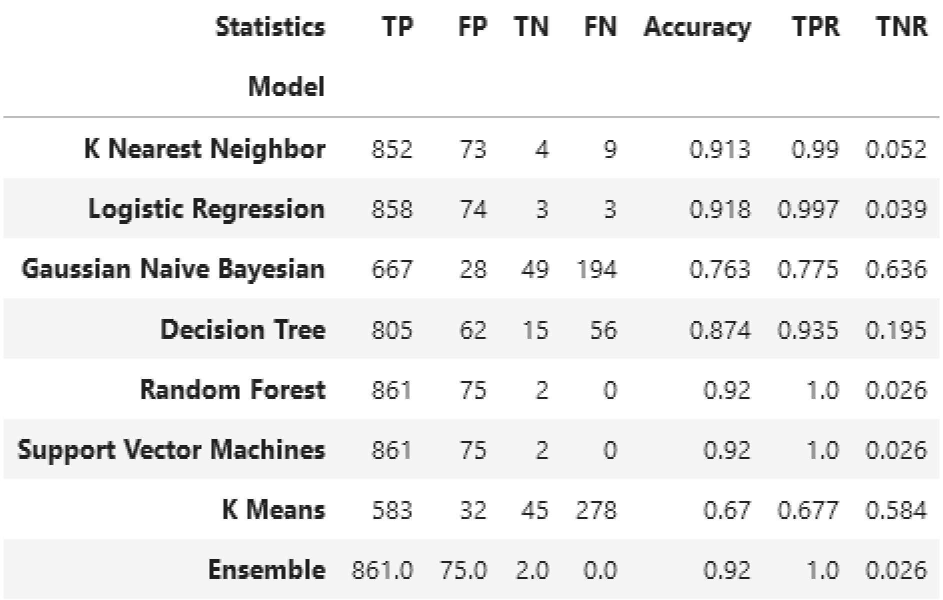


Statistics	TP	FP	TN	FN	Accuracy	TPR	TNR
Model							
K Nearest Neighbor	852	73	4	9	0.913	0.99	0.052
Logistic Regression	858	74	3	3	0.918	0.997	0.039
Gaussian Naive Bayesian	667	28	49	194	0.763	0.775	0.636
Decision Tree	805	62	15	56	0.874	0.935	0.195
Random Forest	861	75	2	0	0.92	1.0	0.026
Support Vector Machines	861	75	2	0	0.92	1.0	0.026
K Means	583	32	45	278	0.67	0.677	0.584

Figure 3 - Precision and Accuracy Comparisons of All Models

The most accurate models were the Random Forest and Support Vector Machines, which tied at a 0.920 accuracy rating. Also with high accuracy were Logistic Regression at 0.918 and K Nearest Neighbor at 0.913. Unsurprisingly, Random Forest and Support Vector Machines showed excellent precision by guessing all true positives. K Means performed the worst of the models at 0.670 accuracy, which was influenced by its k = 2 cluster assignment and the fact that the class label data was not evenly weighted. A point to note is that K Means was used here as a classifier on scaled test data rather than clustering on the original data. This was done on purpose to perform an equal comparison with all algorithms processing on the same test data. Gaussian Naïve Bayes performed only slightly better at 0.763 accuracy. No model had a particularly good TNR, possibly because the class label data is more heavily skewed to the positive ‘1’ rather than the negative ‘0’.

In a final effort to achieve better accuracy, a manual ensemble was created that used a majority vote of the five best-performing algorithms from above (Random Forest, Support Vector Machines, Logistic Regression, K Nearest Neighbor and Decision Tree). Once calculated, the manual ensemble classifier performed just as accurately as Random Forest and Support Vector Machines at 0.920 accuracy.



Statistics	TP	FP	TN	FN	Accuracy	TPR	TNR
Model							
K Nearest Neighbor	852	73	4	9	0.913	0.99	0.052
Logistic Regression	858	74	3	3	0.918	0.997	0.039
Gaussian Naive Bayesian	667	28	49	194	0.763	0.775	0.636
Decision Tree	805	62	15	56	0.874	0.935	0.195
Random Forest	861	75	2	0	0.92	1.0	0.026
Support Vector Machines	861	75	2	0	0.92	1.0	0.026
K Means	583	32	45	278	0.67	0.677	0.584
Ensemble	861.0	75.0	2.0	0.0	0.92	1.0	0.026

Figure 4 - Precision and Accuracy Comparisons of All Models and Ensemble

Predicting Broadband Availability

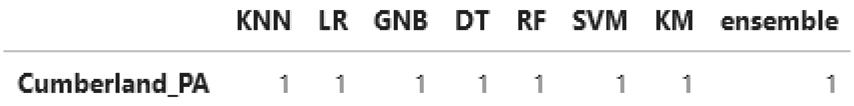
Sample County – Cumberland, PA

Pulled from the original dataset values, my home county of Cumberland, PA was selected to demonstrate how the various models predicted broadband availability given the economic variables.

Cumberland County, Pennsylvania

From dataset (referenced 2019 to 2020):

- percentage of people unemployed = 3.4
- percentage without health insurance = 5.8
- percentage of people in poverty = 7.4
- percentage of people on SNAP = 7.2
- population = 254951



	KNN	LR	GNB	DT	RF	SVM	KM	ensemble
Cumberland_PA	1	1	1	1	1	1	1	1

Figure 5 - Model Predictions for Broadband Availability in Cumberland, PA

All models plus the ensemble unanimously predicted that broadband availability exists within Cumberland County, PA. From the source data, the 'access_bbn' value which related to the 2020 datum from BroadbandNow for Cumberland County, PA is 95.68%, therefore for the purposes of this report the models predicted the availability accurately.

Conclusion

This report summarizes the analysis involved with predicting broadband availability among counties within the United States using several economic census variables. Data cleaning was performed and extraneous variables to the study removed. A binary classifier was established to employ the greatest number of models to the quantitative dependent data that supplied supervised label predictions. All the models were analyzed and tweaked to provide their greatest accuracy and precision, then compared for review. Lastly, a sample county was chosen to validate the models.

References

Benefits.gov. (n.d.). *Pennsylvania Supplemental Nutrition Assistance Program (SNAP)*. Retrieved 04 24, 2023, from Benefits.gov: https://www.benefits.gov/benefit/1169

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., & Cosentini, A. C. (2022, 3 10). *A clarification of the nuances in the fairness metrics landscape*. Retrieved 4 24, 2023, from Nature: https://www.nature.com/articles/s41598-022-07939-1

Lee, N. T., Resnick, P., & Barton, G. (2019, 5 22). *Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms*. Retrieved 4 24, 2023, from Brookings.edu: https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Mattson, M. (2021, 01 04). *US Broadband Availability.* Retrieved 04 24, 2023, from https://www.kaggle.com: https://www.kaggle.com/datasets/mmattson/us-broadband-availability

Mhasawade, V., Zhao, Y., & Chunara, R. (2021, 07 29). *Machine learning and algorithmic fairness in public and population health*. Retrieved 4 24, 2023, from Nature: https://www.nature.com/articles/s42256-021-00373-4

scikit learn. (n.d.). *sklearn.cluster.KMeans*. Retrieved 4 25, 2023, from scikit learn: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

scikit learn. (n.d.). *sklearn.ensemble.RandomForestClassifier*. Retrieved 4 25, 2023, from scikit learn: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

scikit learn. (n.d.). *sklearn.svm.SVC*. Retrieved 4 25, 2023, from scikit learn: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

scikit learn. (n.d.). *sklearn.tree.DecisionTreeClassifier*. Retrieved 4 25, 2023, from scikit learn: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

The White House. (2022, 06 07). *FACT SHEET: Biden-Harris Administration Announces Over $25 Billion in American Rescue Plan Funding to Help Ensure Every American Has Access to High Speed, Affordable Internet*. Retrieved 04 24, 2023, from The White House: https://www.whitehouse.gov/briefing-room/statements-releases/2022/06/07/fact-sheet-biden-harris-administration-announces-over-25-billion-in-american-rescue-plan-funding-to-help-ensure-every-american-has-access-to-high-speed-affordable-internet/

The White House. (n.d.). *Get Internet - Claim Your Affordable Connectivity Program Benefit*. Retrieved 4 24, 2023, from The White House: https://www.whitehouse.gov/getinternet/