Predicting Internet Access Among Low-Income Seattle Residents

By Brittney Nitta-Lee

Date: May 2, 2023

Introduction

This project aims to bridge the digital divide in Seattle and offer opportunities to those who may otherwise be left behind. The final model, trained on a dataset from the City of Seattle, solves the problem with 97% accuracy.

Seattle is a technology-driven city with existing research on digital equity and technology access, including data on the digital divide among low-income households. As an employee of a Public Housing Authority that serves low-income residents, I have witnessed firsthand the negative impact of a lack of technology access on individuals and their ability to participate in society. The motivation for this project is to contribute to ongoing efforts to address the digital divide and promote digital equity among low-income residents in Seattle.

Data Understanding

The data used in this project was obtained from the publicly available City of Seattle's Data Portal from 2018. The dataset comprises responses from a random survey of 4,315 Seattle residents. To focus on low-income residents, the analysis will concentrate on survey respondents whose household income is below $90,000, using the 2018 King County Income Limits from the U.S. Department of Housing and Development as a reference.

The dataset has certain drawbacks, such as a poor representation of Seattle's population, as the survey was administered only in English and Spanish, omitting speakers of other languages. This exclusion is unfortunate since data on technology accessibility among low-income residents who speak other languages could be highly informative.

The dataset comprises 479 columns, each containing binary data, and was gathered via a survey that involved responding to 38 questions.

Additional Resources

2018 Technology Access and Adoption Survey

Technology Access and Adoption Survey Codebook

Data Collection

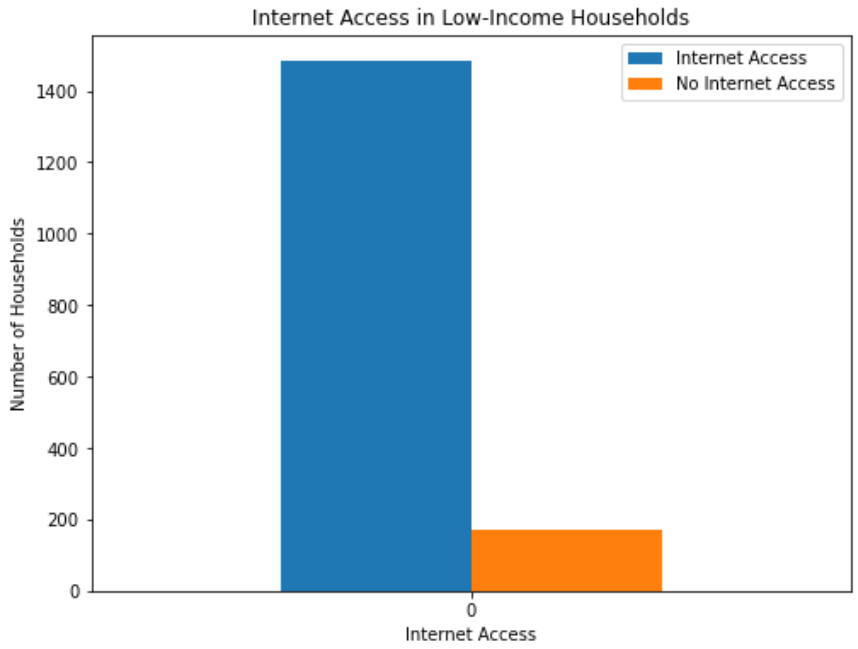
A total of 4,315 surveys were collected from May 23rd through June 25th, 2018, representing 4,315 Seattle households and 10,358 Seattle residents.

- Conducted as a multi-mode survey: across mail, online, telephone, and in-person.
- Completed in both English (4,312 surveys) and Spanish (3 surveys).

- The overall average length of the online surveys was 34.0 minutes.
- The overall survey response was 18% (e.g. 18% of those invited to respond returned a survey).

All eligible respondents are:

- Individuals living within Seattle city limits; Able to conduct the survey in either English or Spanish;
- Has the ability to complete the survey via mail with paper and pencil or pen, online via computer, tablet, or smartphone, or in-person via paper and pencil or pen;
- Able to answer on behalf of the whole household on their use of technology and the internet (though if they needed help completing the survey, they could ask another household member or the survey helpline to assist them).



There are 1483 households with internet access and 171 households without internet access. In comparison, the number of households with internet access is relatively low.

#SMOTE To address the class imbalance, I utilized SMOTE (Synthetic Minority Over-sampling Technique), which is a technique for oversampling imbalanced datasets. This helped improve the performance of the machine learning models and helps with better representation of the minority class, reduced bias and no loss of information.

Modeling

The baseline model that was used is a Logistic Regression model. The evaluation metrics that were used to compare with the advanced models are precision, recall, F1-score and accuracy.

```
Training Report Matrix
              precision    recall  f1-score   support

           1       1.00      0.99      0.99      1190
           2       0.99      1.00      0.99      1190

    accuracy                           0.99      2380
   macro avg       1.00      0.99      0.99      2380
weighted avg       1.00      0.99      0.99      2380

Test Report Matrix
              precision    recall  f1-score   support

           1       0.99      0.98      0.98       293
           2       0.87      0.89      0.88        38

    accuracy                           0.97       331
   macro avg       0.93      0.94      0.93       331
weighted avg       0.97      0.97      0.97       331
```

The baseline model performs well on both the training and test datasets. Class 1 are households who do have internet access and class 2 are households who do not have internet access.

The Training Report Matrix shows Class 1 has perfect precision and recall of 1.00 and 0.99, while Class 2 has a lower precision and perfect recall of 1.00. The macro avg and weighted avg f1-score for both classes are high at 0.99, which shows excellent peformance.

The Test Report Matrix shows the same metrics for both classes on the test dataset. Class 1 has a precision of 0.99 and a recall of 0.98, while class 2 has a precision of 0.87 and a recall of 0.89. The lower performance for class 2 may indicate that the model is struggling to accurately predict this class due to a smaller number of samples or class imbalance in the test dataset. The macro avg and weighted avg f1-score are slightly lower at 0.93, indicating good performance.

XG Boost Model

The final model that was used and evaluated was the XG Boost model. The XG Boost model is a popular and powerful tool for machine learning and can be used for a wide range of problems. After conducting a grid search and hyptertuning the parameters the results were very close to the baseline model.

```
Best Hyperparameters: {'base_XGB__colsample_bytree': 0.7, 'base_XGB__learning_rate': 0.3, 'base_XGB__max_depth': 3, 'base_XGB__min_child_we
ight': 1, 'base_XGB__n_estimators': 200, 'base_XGB__subsample': 0.7}
Best Score: 0.9974789915966387
Training Data Matrix Report:
              precision    recall  f1-score   support

           1       1.00      1.00      1.00      1190
           2       1.00      1.00      1.00       133

    accuracy                           1.00      1323
   macro avg       1.00      1.00      1.00      1323
weighted avg       1.00      1.00      1.00      1323

Training Data Confusion Matrix:
[[1190    0]
 [   0  133]]
Test Data Matrix Report:
              precision    recall  f1-score   support

           1       0.99      1.00      0.99       293
           2       0.97      0.92      0.95        38

    accuracy                           0.99       331
   macro avg       0.98      0.96      0.97       331
weighted avg       0.99      0.99      0.99       331

Test Data Confusion Matrix:
[[292   1]
 [  3  35]]
Training Accuracy: 1.0
Test Accuracy: 0.9879154078549849
```

The model performed very well, with an accuracy of 1.0 on the training data and 0.987 on the test data. The best score obtained through the grid search is 0.997, which is close to 1.0 and indicates that the model is doing a good job of fitting the data.

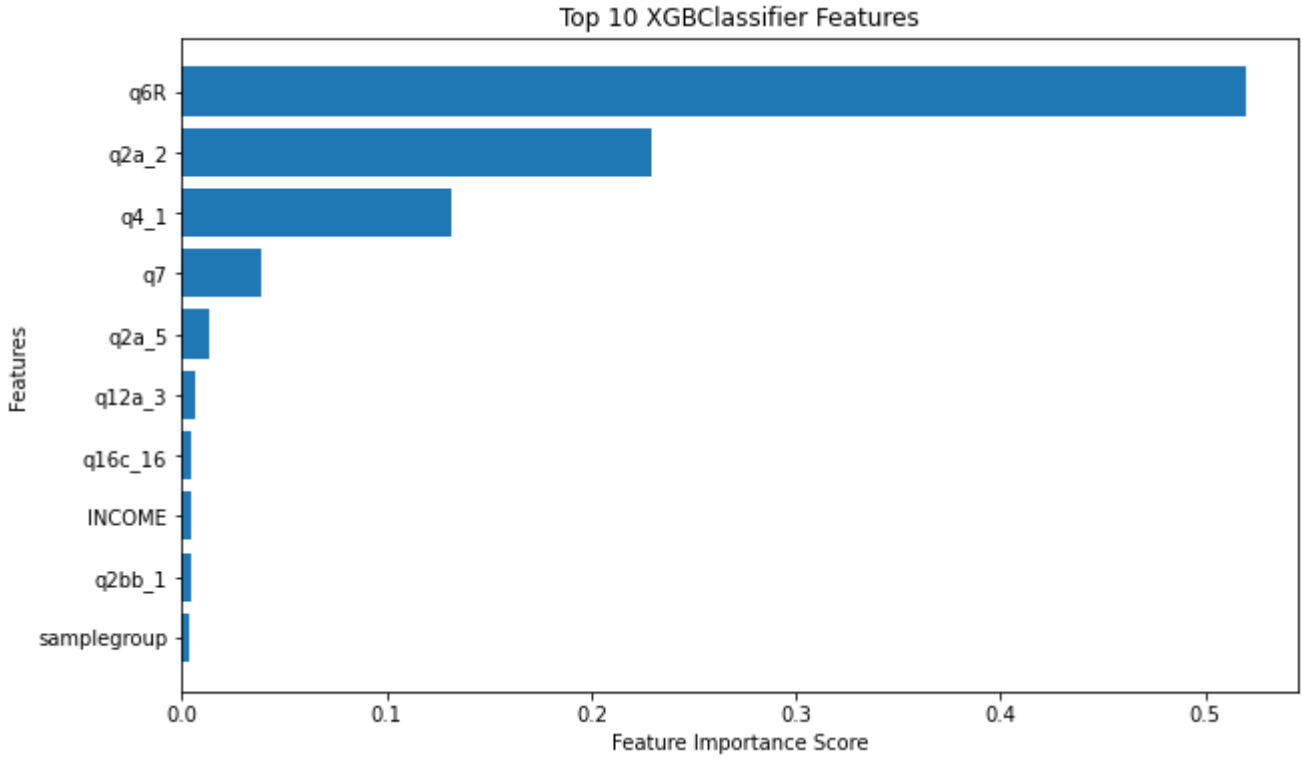
In the training data, both classes have a perfect precision, recall, and F1-score of 1.00, indicating that the model predicted all positive cases accurately.

In the test data, class 1 has a high precision, recall, and F1-score of 0.99, indicating that the model accurately predicted most of the positive cases. However, class 2 has a lower precision, recall, and F1-score of 0.97, 0.92, and 0.95, respectively, indicating that the model misclassified some of the positive cases in this class.

The confusion matrix show that the model is correctly identifying the majority of the observations. In the training data, all 1190 of the class 1 observations are correctly classified, as are all 133 of the class 2 observations. In the test data, there is one false negative (class 2 observation misclassified as class 1) and three false positives (class 1 observations misclassified as class 2), but overall the model is correctly identifying the vast majority of observations.

Feature Importance

From the XGBoost model, the top 10 features were selected based on feature importance. The feature importance provides a score for each feature that indicates how useful or valuable each feature was in the build of the boosted decision trees within the model. The more a feature is used to make decisions, the higher its relative importance.

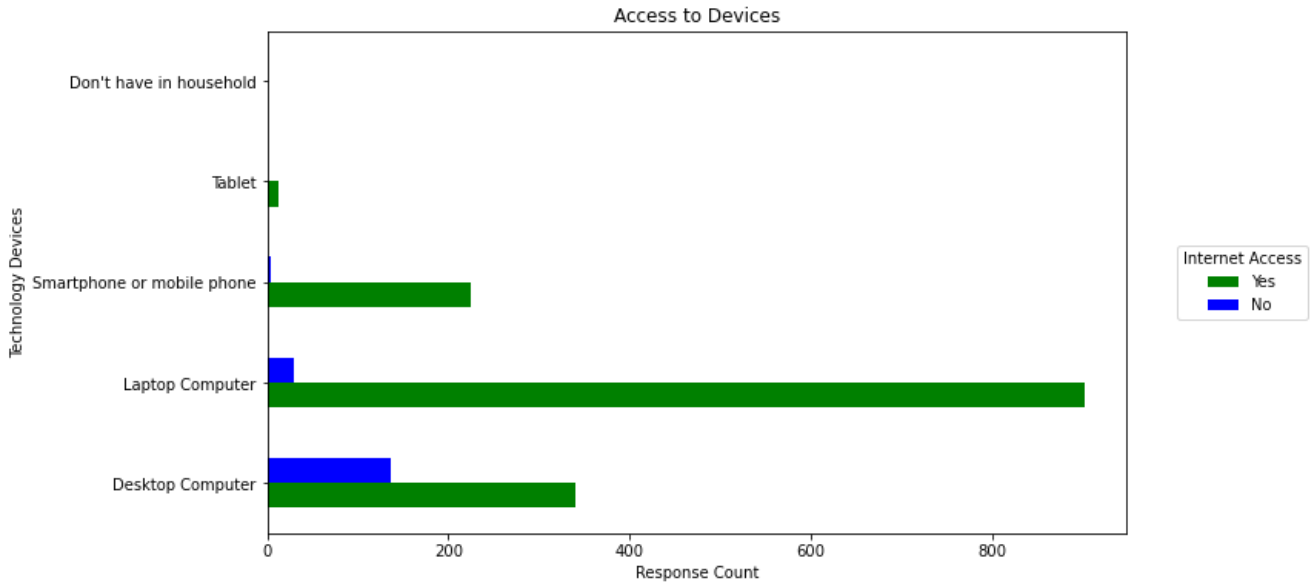


List of Features

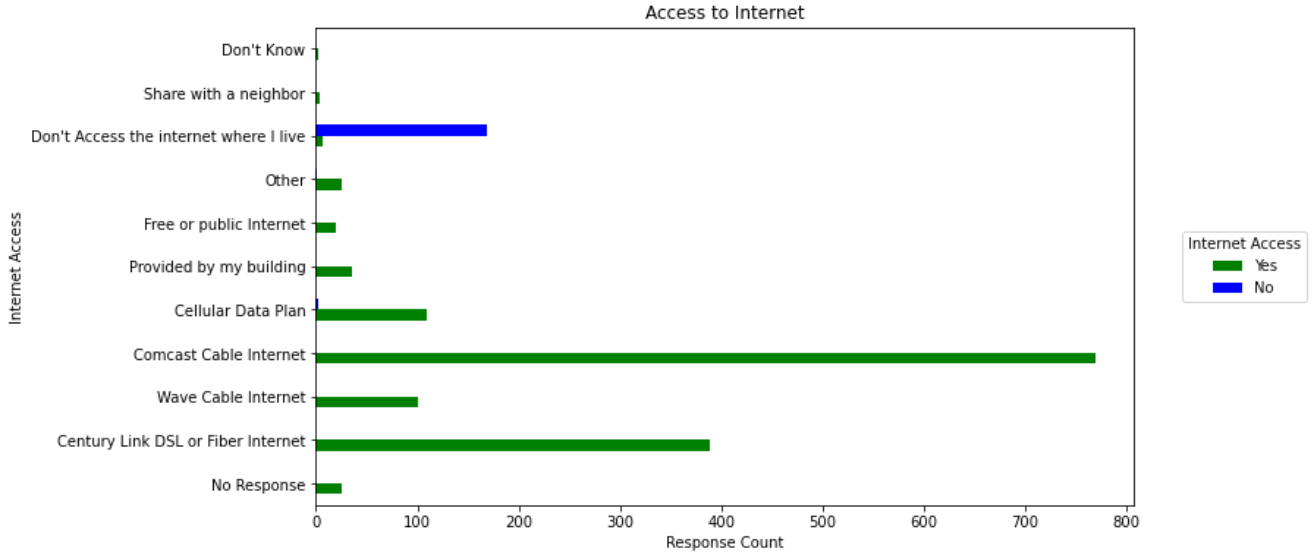
- `q6R` : How would you rate the adequacy of the internet connection and speeds in the place where you live when it comes to your ability to do the tasks you want and need to do on the internet?
- `q2a_2` : Please tell us about the technology devices you have in the place where you live. Does your household have one or more of each of these?
- `q4_1` : What are all the ways you get internet in the place where you live?
- `q7` : Q7. What is the download speed of the internet connection in the place where you live? If you have more than one source, please select the fastest speed you have access to.
- `q2a_5` : Please tell us about the technology devices you have in the place where you live. Does your household have one or more of each of these?
- `q12a_3` : Know of or use low cost internet service - Mobile Citizen / InterConnection: $120 per year internet.
- `q16c_16` : How often does anyone in your household - Look for answers to computer problems online.
- `INCOME` : Income categories
- `q2bb_1` : Thinking about each type of device you have in the place where you live; how did your household get each type of device?
- `samplegroup` : Sample group of survey respondents

Visualizations

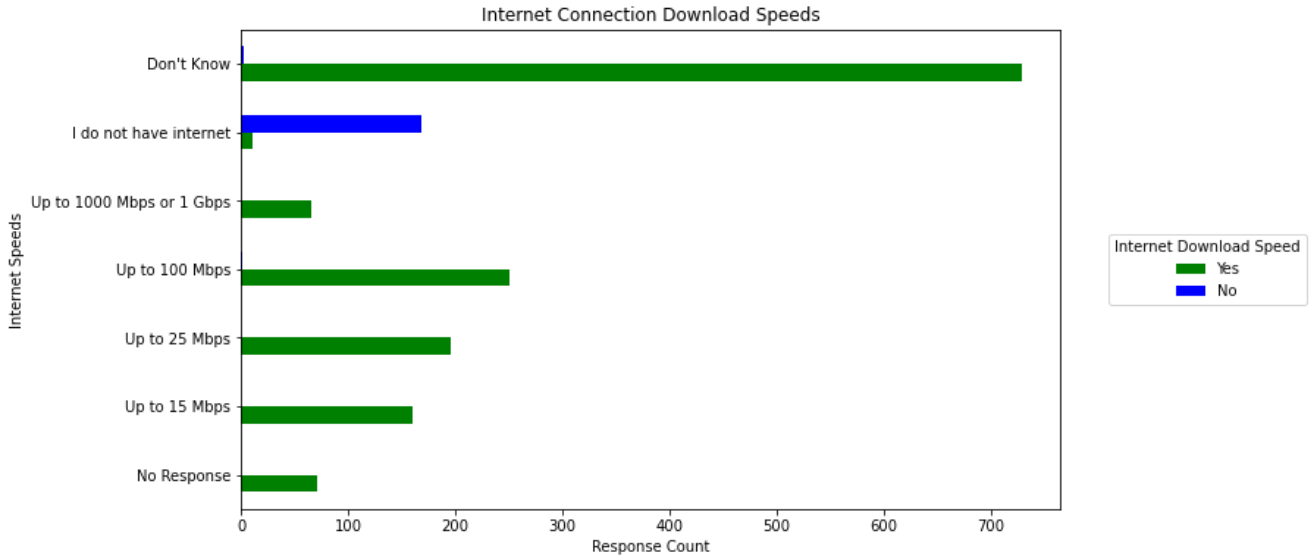
Although individuals without internet cannot respond to this survey question, the high response count for "Mostly Adequate" and "Completely Adequate" among those with internet access suggests that they are generally satisfied with their internet connection speeds.



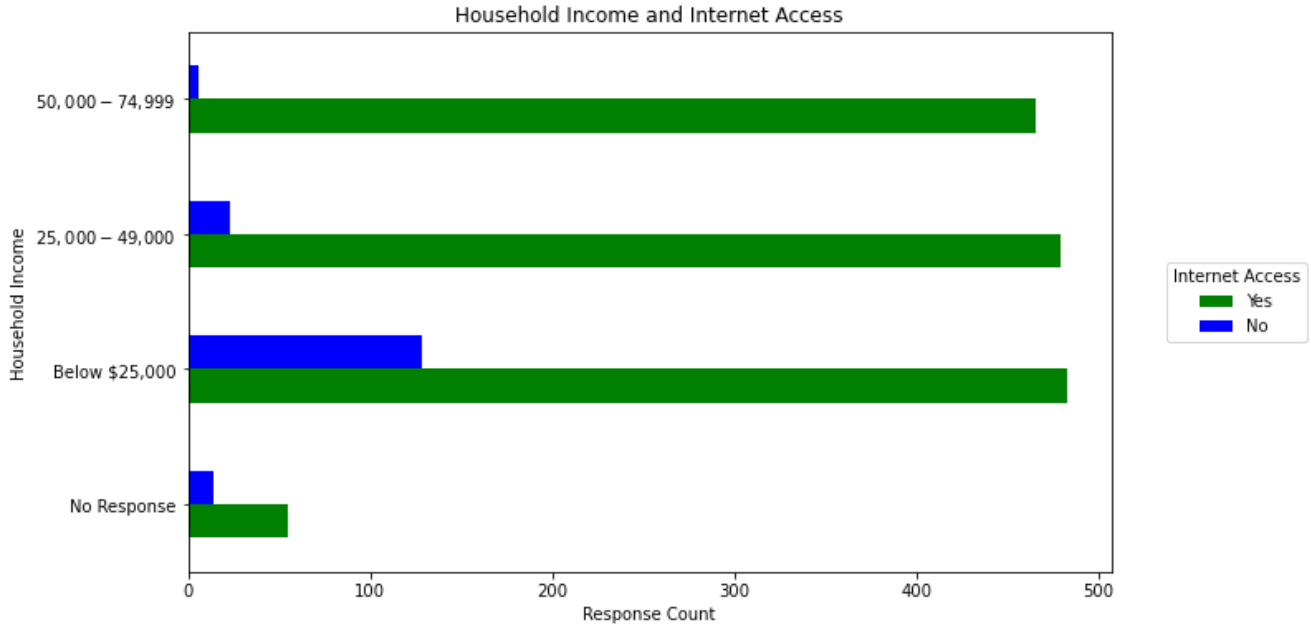
All low-income survey respondents have access to devices and the highest count of devices are laptop computers, following desktop computers. Households who do not have internet access to have some type of device to access the internet.



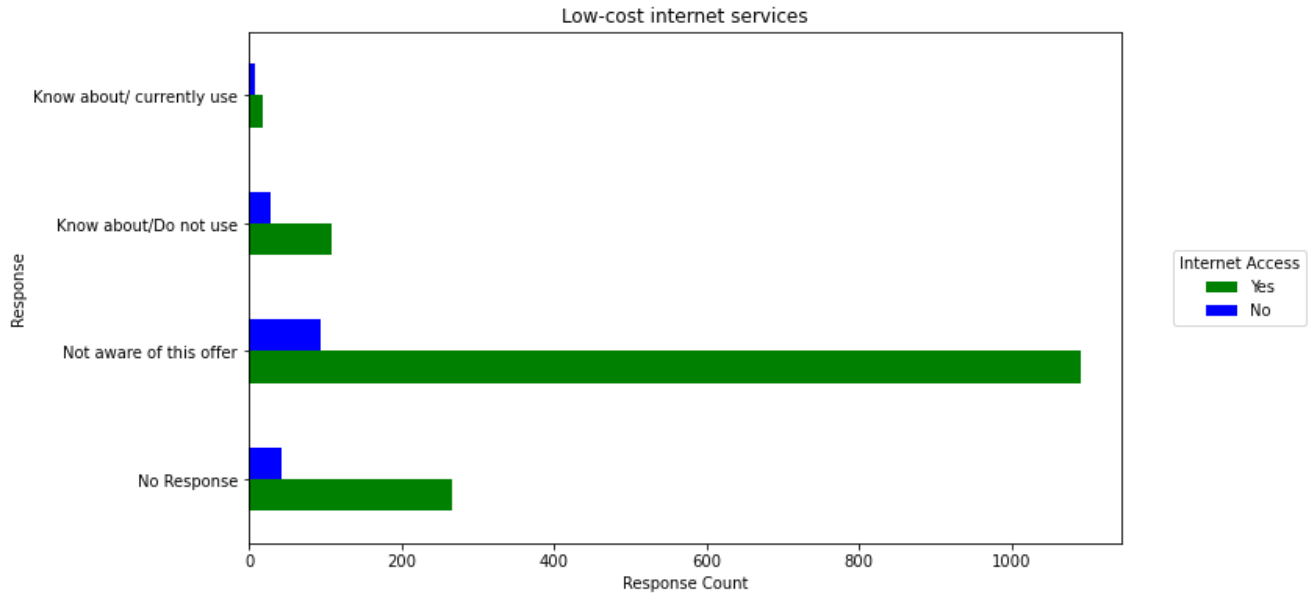
Households who do have internet access in their home use Comcast as their service provider. A very small percentage of those who do access the internet in their home uses a Cellular Data plan.



This was really interesting to see that majority of the survey respondents don't know their internet speed. But, those who do know have at the most 100 Mbps, which is not adequate for large household sizes.



The low-income dataset comprises individuals in Seattle who responded to the survey. Most of these individuals have access to the internet, but the cost of internet services may be burdensome to their monthly income. Additionally, it is noteworthy that the majority of residents reported internet speeds of 100 Mbps provided by Comcast.



The majority of respondents were unaware of the availability of affordable internet services. At that time, Mobile Citizen and InterConnection were providing internet services for $120 per year. The respondents who were unaware of this offer might be using internet services from other providers and possibly paying the full monthly price.

Recommendations

I strongly recommend that Congress make the internet a public utility, especially in light of the COVID-19 pandemic. The data may be from 2018, but the pandemic has underscored the importance of affordable and universal internet access. Without access, individuals are left behind in society and are unable to access essential services such as healthcare, education, work and more.

Here are my other recommendations:

1. Survey should be conducted in multiple languages
2. Look into free in-unit connectivity for low-income households
3. Bring awareness to free or low-cost internet programs
4. Limit the number of survey questions

Next Steps

To bridge the digital gap among low-income households, it is imperative to collect comprehensive data from all such households in the US. This will enable Congress to take proactive steps towards providing free internet access to these households, complete with sufficient download speeds that are tailored to their respective household sizes and number of devices.

For More Information

Contact Brittney Nitta-Lee at [bnittalee@gmail.com]

Repository Structure

```
├── Data
├── IMAGES
├── PDFS
├── .DS_Store
├── .gitattributes
├── EDA.ipynb
├── Evaluation.ipynb
└── Models.ipynb
└── Notebook.ipynb
└── README.md
└── requirements.txt
```

Additional resources

Links to Notebooks and presentation are below:

- Notebook
- Exploratory Data Analysis
- Models
- Evaluation
- Presentation

Reproduction Instructions

Reproduction

Download the dataset and code below:

2018 Technology Access and Adoption Survey

Technology Access and Adoption Survey Codebook

The enivronment requirements is in the `requirements.txt` file

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 100.0%