Modeling Take Rate for Internet Services

By: Jason Rutberg

Introduction:

XXX is a leading internet service provider operating across New England, with a strategic objective to build new networks and expand its services to underserved areas. To achieve this goal, XXX aims to develop a robust model to predict the internet take rate—defined as the ratio of the number of subscribers to the total number of subscriber passings in a given region. The primary objective is to accurately and reliably model take rate for specific population areas by leveraging available data on various relevant parameters.

The process will involve collecting and analyzing data from multiple sources, including demographic, economic, and infrastructure-related variables, to identify the key factors influencing take rate. By applying advanced statistical and machine learning techniques, XXX plans to create a mathematical model that can effectively predict take rate based on these parameters.

Once the model is developed, it will be rigorously validated using existing data from similar regions or historical data within New England. This validation step is crucial to ensure the model's accuracy and reliability in predicting take rates in real-world scenarios. The final outcome will be a powerful predictive tool that will guide XXX in making informed decisions about network expansion and service offerings, ultimately enabling the company to better serve its customer base across New England.

Methodology:

In order to begin constructing a model, several potential variables were identified that could impact take rate, such as population, median income, median age, number of cable providers, percent of the population with high school diplomas/Bachelor's degrees, percent of the population under twenty-five, percent of the population over sixty-five years old, percent who own their own homes, and percent who work at home for a given region. With these variables in mind, the search for publicly available data to build the model began. Through American FactFinder, the variables for which data could be collected were identified. Consequently, the search was narrowed down to ten variables with easily accessible data: (1) population, (2) median age, (3) percent of the population under 21, (4) percent of the population over 65, (5) percent of the population who own their own homes, (6) percent of the population with a high school degree (or higher), (7) percent of the population with a Bachelor's degree (or higher), (8) percent of the population who work at home (where workers are at least 16 years old), (9) median household income, and (10) percent of the population with 3+ internet providers (with speed > 25/3 Mbps as reported by the FCC as of December 2017). The data for the first nine variables is available from the 2010 U.S. census, while the data for the tenth variable is available on the FCC's website. The most challenging part of building this model was finding reliable take rate data for a given region; the first large-scale data source found regarding take rate was the Brian Webster Consulting and The Gadberry Group's November 2009 report concerning nationwide data. This report included a table containing the observed take rate for all fifty states in the country. Hence, all the data necessary to establish the first model had been collected.

A table was constructed with data from the ten variables (discussed above) and the observed take rate for each state, and this data was imported into MATLAB. Using the least squares regression method, a multilinear model with an adjusted R-squared value of 0.746 was formulated. Additionally, predicted values, residuals, confidence intervals, and prediction intervals for each state were determined. By examining the p-values for each variable, it was possible to conclude which ones were significant (and which ones were not) for this particular model. It was found that three variables were statistically significant at a level of 0.05, with one additional variable close to significance.

With this information, a second model was constructed, including only the four variables that appeared to significantly impact take rate from the first model. These four variables were: (1) percent of the population over 65, (2) percent of the population who work at home, (3) median household income, and (4) percent of the population with 3+ internet providers (speed > 25/3 Mbps). Additionally, a "discount factor" was introduced to counteract the overly inflated values for the percent of the population with 3+ internet providers with speed > 25/3 Mbps and the observed take rate. Specifically, the FCC reports that only 26 million Americans lack access to internet speeds of at least 25/3 Mbps, while Microsoft reports that 126 million Americans do not have access to internet speeds at this benchmark. With this in mind, it was found that the FCC overinflates its internet speed data by around fifty percent, so all the data for the percent of the population with 3+ internet providers (speed > 25/3 Mbps) was multiplied by 0.66 to accurately reflect internet speeds. Regarding the take rate, the mean state take rate from the Brian Webster Consulting and The Gadberry Group data was 73%, which is much higher than the generally accepted average take rate of 44% for the United States. Therefore, all the observed

state take rate data was multiplied by 0.6 to better represent the actual take rate in these states (as opposed to the overly inflated values reported by internet companies to the FCC).

With these adjustments, the second take rate model was ready to be assembled, using only the four variables from the first model that were found to be significant and the adjusted speed and observed take rate data. Once again, the least squares regression method was utilized to construct this new model, resulting in an adjusted R-squared value of 0.649. With this solid model, it was time to use "test data" to evaluate the model's accuracy in predicting take rate. To find this "test data," it was necessary to find take rate data for communities across the country, which proved to be quite difficult to obtain. However, municipal broadband deployments are funded by local governments, which are legally required to publicly report take rates. By examining the municipal broadband networks in Tennessee, Virginia, Vermont, and Utah, take rate data for several small towns on these networks was collected. Hence, the "test data" necessary to evaluate the validity of the model was now available.

Results and Discussion:

Using the first state take rate model, an adjusted R-squared value of 0.746 was found, and the p-values for each variable were examined to determine statistical significance. It was found that the percent of the population over 65, median household income, and the percent of the population with 3+ internet providers with speed > 25/3 Mbps were significant at the 0.05 level (p-value < 0.05). Additionally, the percent of the population who work at home had a p-value of 0.065. Given that these variables were significant or nearly significant, these four variables were chosen to build a second state take rate model.

In the second model, the percent of the population with 3+ internet providers with speed > 25/3 Mbps was adjusted using a “discount factor” due to the overinflated values reported by the FCC. The adjusted speed values in the second model were determined by multiplying the original speed values by a scaling factor of 0.66. Additionally, the observed take rates were adjusted by a scaling factor of 0.6 in the second model. After performing a linear regression with these four variables, median household income and the adjusted percent of the population with 3+ internet providers with speed > 25/3 Mbps were found to be significant. This second model had an adjusted R-squared value of 0.649, indicating a moderate correlation between take rate and the four variables under consideration.

Subsequently, "test data" was collected on various small towns across the country with take rate data reported by municipal broadband networks. Mixed results were observed in the reliability of the state model, as several communities had large take rate residuals, indicating substantial differences between observed take rates and those predicted by the model. This suggests that the state take rate model might not be reliable for small towns, indicating the need for further refinement. Moving forward, more data should be collected from small towns on various municipal broadband networks, and a new model should be developed using town-level take rate data rather than state-level data. Additionally, the variables excluded from the second state model should be reintroduced in this new model to verify whether they are significant at the town level.

Next Steps:

Moving forward, more data needs to be collected regarding observed take rates in individual towns. This data can be obtained by researching the various municipal broadband networks throughout the country. Since these services are required to report take rates, this information can be found through public disclosures. As previously discussed, the state take rate model might not be the most accurate, so it would be beneficial to create a new model using data exclusively from individual communities rather than entire states. Additionally, including all ten variables used in the first model would be advisable to determine which variables are significant at the town level, as these may differ from those at the state level. All relevant census data is easily accessible through American FactFinder, even at the town level.

Moreover, the “discount factor” related to the percent of the population with 3+ internet providers with speed > 25/3 Mbps should be refined. This data is reported to the FCC by various internet service providers and is often inflated, so adjustments are necessary to improve the model's accuracy. Refining this “discount factor” will require further research into the actual availability of internet speeds versus the overinflated reports.

References/Sources:

1. https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml (American FactFinder)
2. "Evolving Metrics: New Levels of Accuracy Reveal Increased Take Rates," Brian Webster Consulting and The Gadberry Group, November 2009 ("Take-Rate-Brief" in XXX dropbox)
3. "Broadband Adoption Rates and Gaps in U.S. Metropolitan Areas," Metropolitan Policy Program at Brookings, Adie Tomer and Joseph Kane, December 2015 ("Broadband-Tomer -Kane-12315" in XXX dropbox)
4. https://broadbandmap.fcc.gov/#/ (FCC-Federal Communications Commission-Fixed Broadband Deployment)
5. http://ftpcontent2.worldnow.com/wrcb/pdf/091515EPBFiberStudy.pdf (Tennessee Municipal Broadband--Hamilton County, Tennessee)
6. https://ilsr.org/wp-content/uploads/2012/04/muni-bb-speed-light.pdf (Virginia Municipal Broadband--Bristol City, Virginia)
7. Penetration and Take Rate in VT (in XXX dropbox)
8. https://media.rainpos.com/442/utopia_info_history_as_of_oct2015.pdf (Utah Municipal Broadband (UTOPIA)--Lincoln, Utah)
9. https://www.deseret.com/2013/2/23/20514916/do-two-no-votes-to-fund-utopia-signal-trouble-for-agency (Utah Municipal Broadband (UTOPIA)--Tremonton, Utah)