Estimating NYC Internet Access at the Block Group Level Using Machine Learning

Philipp Kats, Nikhil Kishore, Erwan LeCun, Alan Polson

5/1/2016

Contents

Abstract

The US Census is an authoritative source of granular data for decision-makers at local levels of government. However due to various reasons, certain characteristics such as Internet access figures are not currently available at lower levels of granularity. In this paper, we modelled the relationships between Internet access and certain other demographical characteristics available at the Block Group level. We trained, validated and tested various machine learning algorithms on PUMA level data, selected the Neural Network model, and used it to impute Internet access characteristics from the PUMA level to the Block Group Level. We also cross validated our results and plotted the results on an interactive map that is available for decision-makers to utilize.

1. Introduction

1.1. Background

The U.S. Census conducts a decennial survey of various demographic features for the entire country and releases this data at various levels of aggregation, with the smallest levels being at the Census Block, Census Block Group, and Census Tract levels. From the 1940 census through the 2000 census, a subset of all Americans received a "long-form questionnaire" containing additional questions. Many Americans found filling out the long-form questionnaire to be burdensome and intrusive, and its unpopularity was a factor in the declining response rate of the decennial census.[1]

Because of this declining response rate, in 1994 the Census Bureau began the process of changing the means of obtaining demographic, housing, and socioeconomic information from the census long-form questionnaire to the ACS. In 2010, the ACS produced its first set of estimates for areas of all population sizes, using information collected from January 2005 through December 2009.[2]

The ACS now releases 5 year estimates at a block group level, and one year estimates at the more aggregated, Public Use Microdata Area (PUMA) level.[3]

In 2013, for the first time, the ACS included questions about Internet access and computer ownership in its questionnaire. This information is critical to understanding the availability of internet in a city and for mapping 'broadband-access deserts'.[4] However, since this data is currently only available in one-year estimates, (i.e. for 2013, 2014 and 2015) it is only available at a PUMA level of aggregation, and will only be published at the Block Group level as 5 year estimates in 2018. [5]

1.2. Motivation

Understanding Internet Access (or lack thereof) at a Block-Group level of granularity will be especially useful for decision-makers who have to allocate budgets for Internet-Equality programs such as the Red Hook initiative[6] and the Downtown Brooklyn Alliance[7]. It will also be crucial to decision regarding the siting of future Links in the LinkNYC program.[8]

This will also enable private companies to act immediately to install Wi-Fi hotspots in areas of New York that may need it, enabling seamless corridors of internet connectivity as envisioned by programs such as Hotspot 2.0.[9]

There is thus ample reason to correctly estimate internet access in New York City now, at the Block Group level, both as an aid to decision-makers and as an input for other research and citizen scientists who want to address the inequalities in New York.

1.3. Goal

In order to estimate the availability of Internet access at the Block Group level today, our team planned to use features from the ACS that are available at both the Block Group level and the PUMA level, and then use machine learning techniques to map their relationship to a target variable describing the percentage of people with an Internet subscription, which is only available at the PUMA level of aggregation across the US. We would then use these relationships to impute percentage of Internet subscription values to Block Groups in New York State and New York City.

1.4. Literature Review

A 2014 paper by the Open Technology institute on the costs of connectivity showed that the US lags behind the rest of the world in terms of affordable internet access, with New Yorkers with the fastest available broadband paying four times as much for half the speed as other cities.[10] A policy brief by NYC Comptroller Scott Stringer highlighting the issue quoted the 2013 ACS as its only source of information, but could only quote Boroughs, Congressional Districts and PUMAs as its spatial units of measurement.[11]

Technical Review: The concept and technique of using relationships between features at higher levels of aggregation to impute values to those features when they are unknown at lower levels of aggregation has been done before by at least two other parties.

In September 2015, Enigma Labs published a technique for mapping where houses with a higher risk of having a dysfunctional smoke alarm were located at the Census Tract level, using features that were drawn from the American Housing Survey (AHS) and the American Community Survey (ACS). They used a random forest model to determine the relationship between the features and documented and published their methodology on GitHub.[12,13]

In March 2016, analysts at CartoDB were able to use a similar technique to map the proportion of people who suffered from both sight and hearing impairments, (which they were only able to extract at the PUMA level) down to the Census Block Group level. They used a neural network to achieve their results and also published their methodology on GitHub.[14]

2. Data

The data used in this project came primarily from three sources:

1. American Community Survey and American Housing Survey Data at the Public Use Microdata Area (PUMA) level for the United States. (2008 - 2012)

2. American Community Survey and American Housing Survey Data at the Block-Group level for the State of New York. (2013)
3. Lion shapefiles for the Block Groups of NY State from the 2010 Census.

Meta-data about the above datasets, including spreadsheets detailing what features were available at what level of aggregation were also used to select our features we could use and how to wrangle the features.

3. Methodology

3.1. Data Cleaning

The data was collected from the ACS FTP server. The original data was provided in the form of numerous .txt files. Each .txt file represents answers to specific questions for all geographies in the chosen state. In order to aggregate the data into a useable tabular format, we had to concatenate it both vertically (for different states) and horizontally (for different questions and answers). The data was then filtered by geography using a geographical lookup table. In order to interpret and filter the features, answer-series dictionary was created. As the series ID was different for 1 year and 5 year summary files, unique cross-data ID's were defined.

After that, we proceeded to examine both of the resulting tables. We computed the percentage of internet subscription for PUMAS, and extracted this as a separate CSV file. Then, both PUMAs and Block Group features were filtered so that each of two tables only contained features present in both tables with less than 50% of cells empty. Finally, technical features (such as imputation flags) were dropped.

The data was then split into training (80%) and test (20%) sets. The training set was then split into train (66% of overall data) and validation (16% of overall data) and distributed among our team members in order to develop models using the same datasets.

The entire data processing pipeline was automated and well-documented in order to comply with reproducibility requirements and to be easily replicated to predict other variables. A list of all such variables that are presented in 1 year estimates, but not presented in 5 year estimates (e.g. at the Block Group level) is presented in Table 3.1.

3.2. Machine Learning Implementation

Our model uses census characteristics available at both the Block Group and PUMA level as our independent input features and the percentage of internet subscriptions as the dependant target we are trying to predict. In order to build our predictive model, we tested and compared several machine learning algorithms to see which one produced the best results. These algorithms include LASSO Regression, Support Vector Machine Regression, Random Forest Regression, and Neural Networks.

3.2.1 LASSO Regression

The cleaned data was split into training, testing and validation datasets, according to the criteria mentioned in section 3.1. Since some values were missing, these were imputed as the men of the field, using Scikit-learn's imputer function. The training data was then fitted to a LASSO model and, using the parameters calculated on the training data, the validation data was tested. Based on the results from the validation set, the alpha value was adjusted to 0.9, for which the scores for the training, validation and test data were 0.9335, 0.9058 and 0.8959 respectively.

3.2.2 Support Vector Machine Regression

The cleaned data was split into training, testing and validation datasets, according to the criteria mentioned in section 3.1. Since some values were missing, and SVR requires all data to be present, these values were imputed with average feature values. Then, each feature was normalized to have mean 0, and standard deviation 1.

Three main modes, *rbf, linear* and *poly,* were applied, using the validation set to optimize the hyperparameters. The best model, *rbf*, resulted in R-squared values of 0.845 for the training set, 0.823 for the validation set, and 0.831 for the test set.

3.2.3 Random Forest Regression

The cleaned data was split into training, testing and validation datasets, according to the criteria mentioned in section 3.1 . All the NaN values present in the data were replaced by Numpy NaN values for purposes of analysis. Then, a Random forest regression was used to train the data on the training dataset and was then tested on the validation and test sets to determine the R-squared score. It was found that test data set had an R-squared score of 92.30 and 90.45 for validation set. 500 decision trees were built in our Random Forest regression analysis for purposes of training.

Similarly, Random forest classification was also performed on the model. However, one challenge in doing classification is that the target needs to be a binary value. Since the counts of the internet users in each PUMA was in the form of continuous data, we had to define certain thresholds to say whether or not a certain PUMA had Internet or not. Therefore, the median of the percentage of internet subscribers in the PUMA data is taken. Any value above the median is considered to have Internet and any value below the median is considered to be without internet. Our Random Forest classifier was trained with 500 trees on the training data and tested on the validations and testing datasets. The validation dataset had an accuracy of 93.51 and AUC score of 98.55. Similarly, the test dataset had an AUC score of 98.70 and accuracy of 94.37.

3.2.4 Neural Network

The cleaned data was split into training, testing and validation datasets, according to the criteria mentioned in section 3.1. Because there are so many different features present in our dataset, a fair bit of preprocessing was involved to get the features into useable condition. First, the logarithm of each feature was taken in order to have a somewhat smaller range of values for each feature. Then, each feature was normalized to have mean 0, and standard deviation 1. Finally, each NaN value was replaced with 0. Since the data had already been normalized, it is not too huge of an assumption to set NaN values equal to 0, since that is theoretically the mean value of that feature.

Next, the neural network was built in Python using the Keras neural network library. We defined our neural network as taking in an input vector of 1965 dimensions (for each Block Group/PUMA feature), and producing an output vector of 1 dimension (percentage of people with an Internet subscription). The neural network consisted of 80 hidden units, a dropout regularizer of 0.2, and a *tanh* activation function. The loss is defined by mean squared error and the method of training is stochastic gradient descent. The results are detailed in *Table 3.2.*

Overall, the errors for the training, validation, and test datasets are quite low, indicating that our model works very well for the PUMA dataset. Shown in *Figure 3.2.4* is a plot of the loss over time for the train and validation data. This plot illustrates the the loss for both sets eventually converge to 0.

Shown in *Figure 3.2.4 (b)* are plots of the actual vs predicted values for both the validation and test data. These plots visually illustrate that the actual and predicted values are very similar. Finally, shown *Figure 3.2.4 (c)* is a series of histogram plots illustrating the distribution of Internet subscription percentages for PUMA and Block Group level data. From looking at the distribution of our Block Level predictions versus the distribution for our known PUMA level data, we can see that they are very similar. Overall, the neural network model proved to be the most successful of those tried in our analysis.

4. Results

Since the Neural Network model gave us the best results on the test data-set, we decided to use it to predict Internet Access at the Census Block-Group level. We the Cross-Validated our results with the ACS estimates at the PUMA level, and estimate our results to be 74% accurate. (Details below) Using our model we were able to generate a map of internet access at the census block-group level for the entire state of New York. The results are presented as maps (plots 4.a, 4.b.)

4.1. Cross-Validation

In order to evaluate the true predictive accuracy of our models, we aggregated a weighted average of the neural network predictions for each Block Group to a PUMAs level in New York State, and compared it with the actual value of known internet subscription percentages for those PUMAs. Computed in this way, we acquired an R-squared value equal to 74.6%. 95% of PUMAs had an error below 10%, and 68 has an error below 5%.

Interestingly, the PUMAs with the highest error were those of Brighton Beach and Midwood in Brooklyn. We speculate, this error may indicate recent socioeconomic development in those areas.

4.2 Mapping

The resulting Block Group predictions were merged with Block Group boundary shapefiles and saved as a GeoJSON. For the purposes of visualisation, Block Groups with land area smaller than water area (both features were provided in boundaries shapefile) were removed. The resulting shapefiles were uploaded to CartoDB in order to generate an interactive map[15]. Colors were mapped using quantiles for existing range (from 65 to 100%) and with a cubehelix colormap using python *plottable* module
(plots 4 a, 4 b).

4.3 Limitations

Assumptions we were forced to make include:

1. The relationship between the characteristics are the same at the PUMA level as at the Block Group level
2. The relationships between the characteristics is more or less equivalent across PUMAs in the U.S., and do not differ much, regardless of whether the PUMAs are in urban or rural areas
3. The errors associated with the measurements are not significant enough to throw off our model. We have not taken the Margin of error estimates into consideration

4. That our cross validation data is accurate. The ACS survey goes out to a sample of the population and the respondents are given weights based on how many characteristics they share with the general populous, based on previous responses. Since this is an estimate, made in a similar manner to our own, it is difficult to say – without actually knocking on doors – which estimate is the accurate one, when the results diverge.

5. Conclusion

Our final conclusion is that it is possible to map internet access and many other features of the ACS and AHS down to the Block Group level. The granular understanding of these demographic and household features enable us to get a clearer idea of the conditions of streets. This is especially helpful for informing policy and making decisions on new developments, such as LinkNYC kiosks. We found that although Neural Networks provide better accuracy, it is difficult to draw any inference from these models, since it is more or less a black box. The Random Forest, however, provide us with the relative importance of each feature, along with a score for its importance. This showed us that "Number of households receiving Food Stamps in last 12 month and having at least one person with disabilities" was the most important feature, rated almost twice as important than the next feature. It would be interesting to do some further investigation into this finding.

Reproducibility and Reusability

Our code and results are available on our GitHub repository, and might be reused to predict (granulate) other features. URL to Repository: https://github.com/AlanPolson/ML_Project

6. Personal Contributions

While the team was constantly collaborating, each member had his own responsibility, listed below:

Philipp Kats
Data acquisition and processing, SVR model, cross-validation, mapping, presentation.

Nikhil Kishore
Random forest model, feature importance rating.

Erwan LeCun
Neural network model, distribution cross-validation.

Alan Polson
LASSO regression, literature review, report writing.

7. Tables and Figures

Table 3.1 : Topics covered in 1-year summary estimates and not presented in 5-year estimates (thus, not presented on block group level)

N	Topic
1	**Sex-Age distribution**
2	**Detailed race, minorities distribution**
3	**Ancestry**
4	**Place of birth, year of entry for foreign-born population**
5	**Geographical mobility and means of transportation**
6	**School enrolment and types of school**
7	**Poverty**
8	**Occupation, Industry**
9	**Household vacancy status, tenure, population per unit, household income, price**
10	**Health care**
11	**Internet Connection and presence of electronic devices**

Table 3.2 : Summary of Machine Learning Models and their R^2 scores on different sections of Data

Model	**Train R^2**	**Validation R^2**	**Test R^2**	**OutBound R^2**
Lasso	**0.9335**	**0.9058**	**0.8959**	**-**
SVR	**0.845**	**0.822**	**0.831**	**-**
Random Forest	**0.9806**	**0.9045**	**0.9230**	**-**
Neural Networks	**0.998717**	**0.99809**	**0.997946**	**.743**

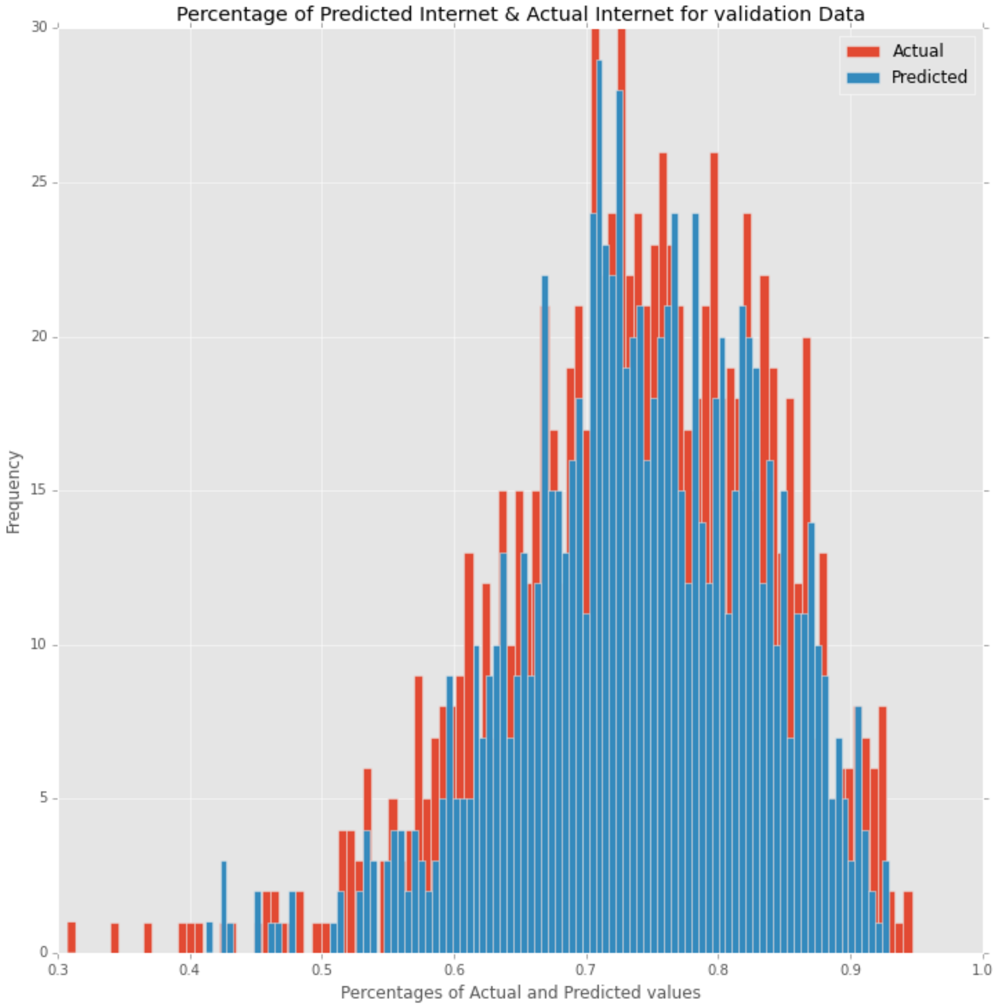

Figure 3.2.3 (a) : Random Forest Model
Predicted vs Actual Values for Validation Data

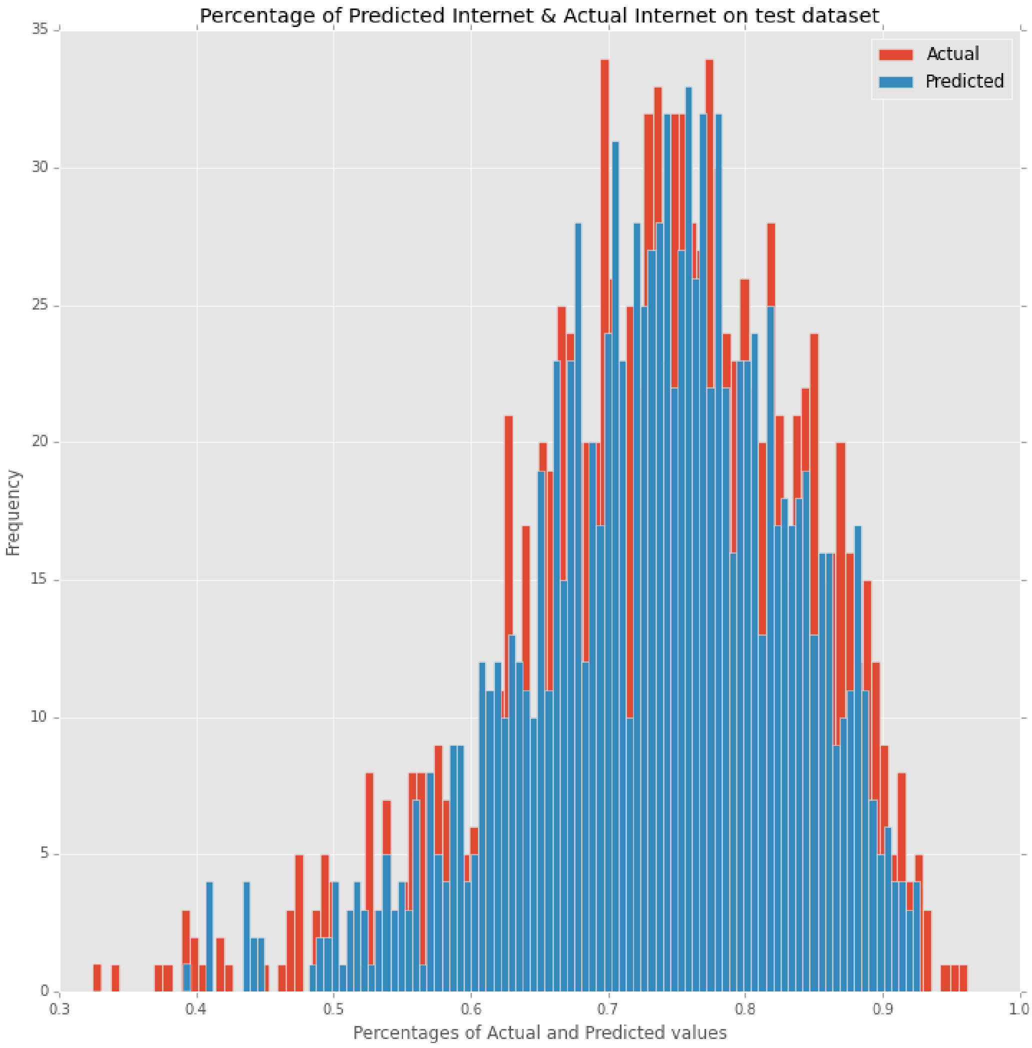

Figure 3.2.3 (b) : Random Forest Model Predicted vs Actual Values for Test Data

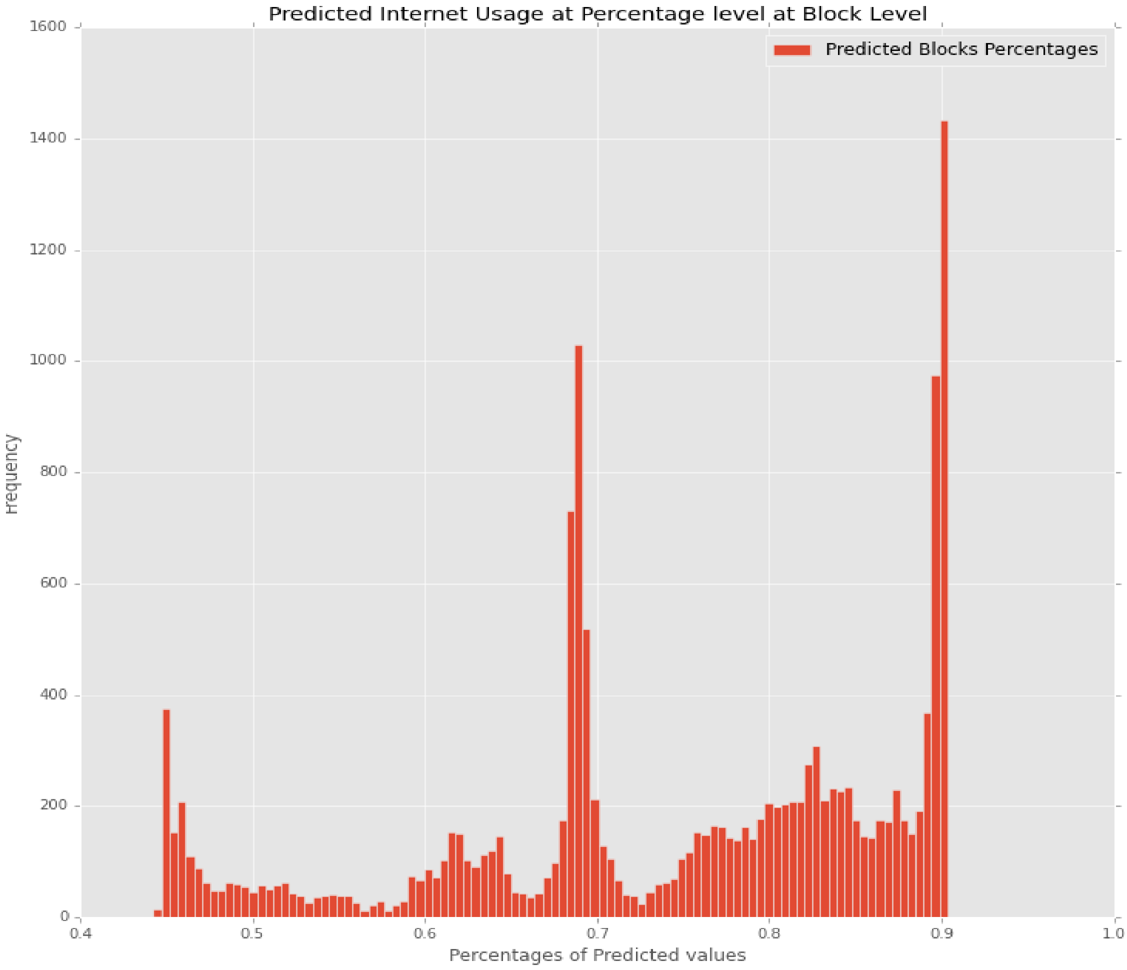

Figure 3.2.3 (c) : Random Forest Model
Histogram of predicted percentages (Expressed as Values between 0 and 1)
of Internet Access for Block groups of NYS

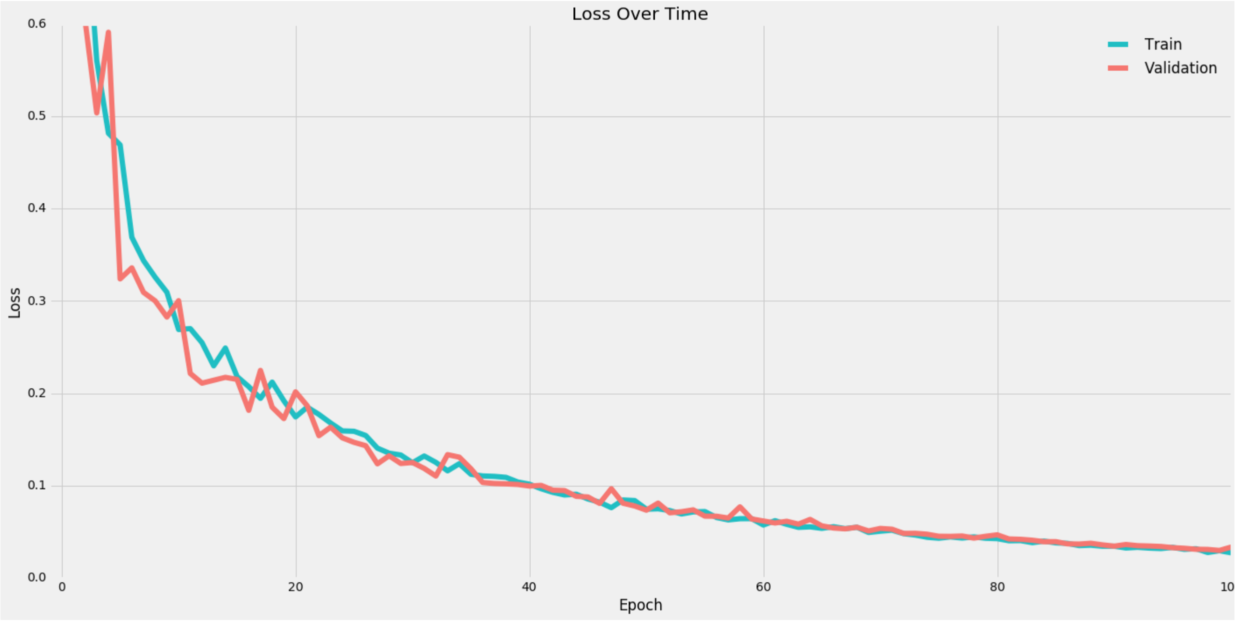

Figure 3.2.4 : Neural Network Model

Loss of Mean Squared Error over subsequent Epochs for Validation and Train

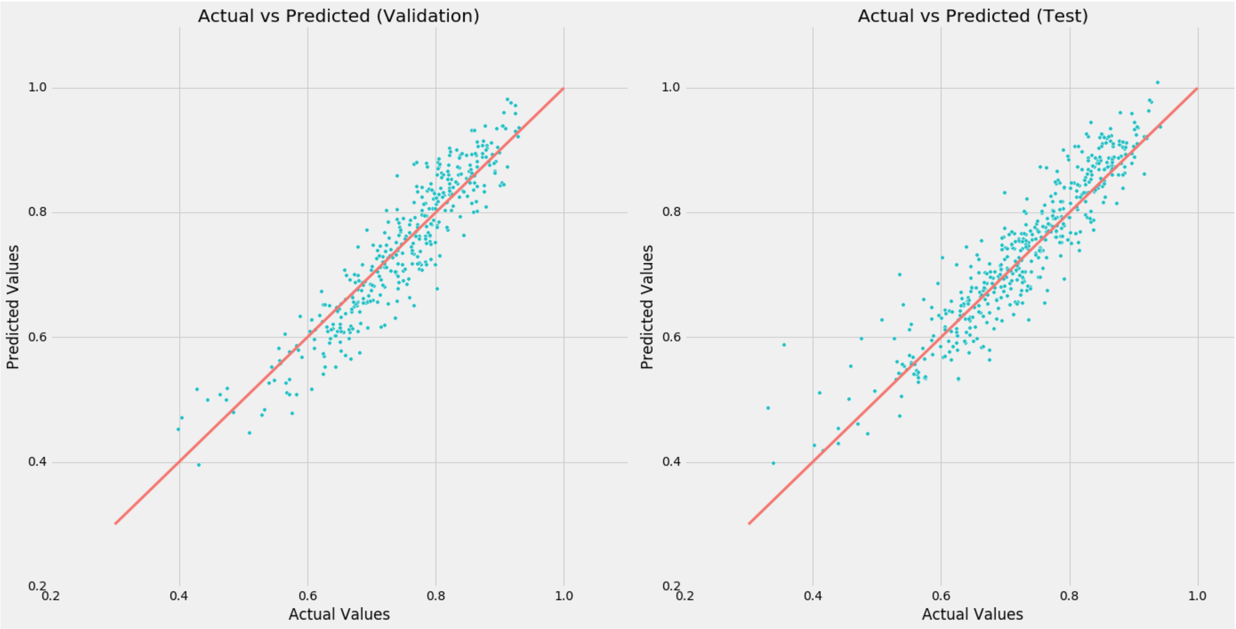

Figure 3.2.4 (b) : Neural Network Model

Scatter Plot of Actual vs Predicted values for Validation and Test Sets

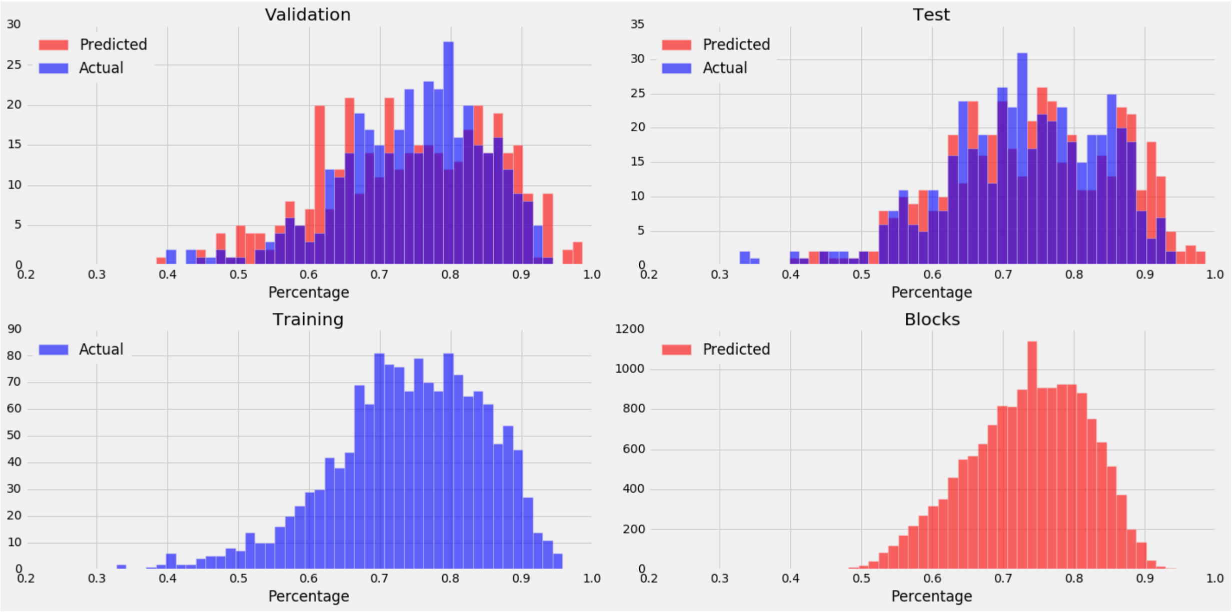

Figure 3.2.4 (c) : Neural Network Model

Histogram of Prediction results superimposed over Actual results, when available.

Clockwise from bottom-left: Training, Validation, Test and prediction for Block Groups of NYS

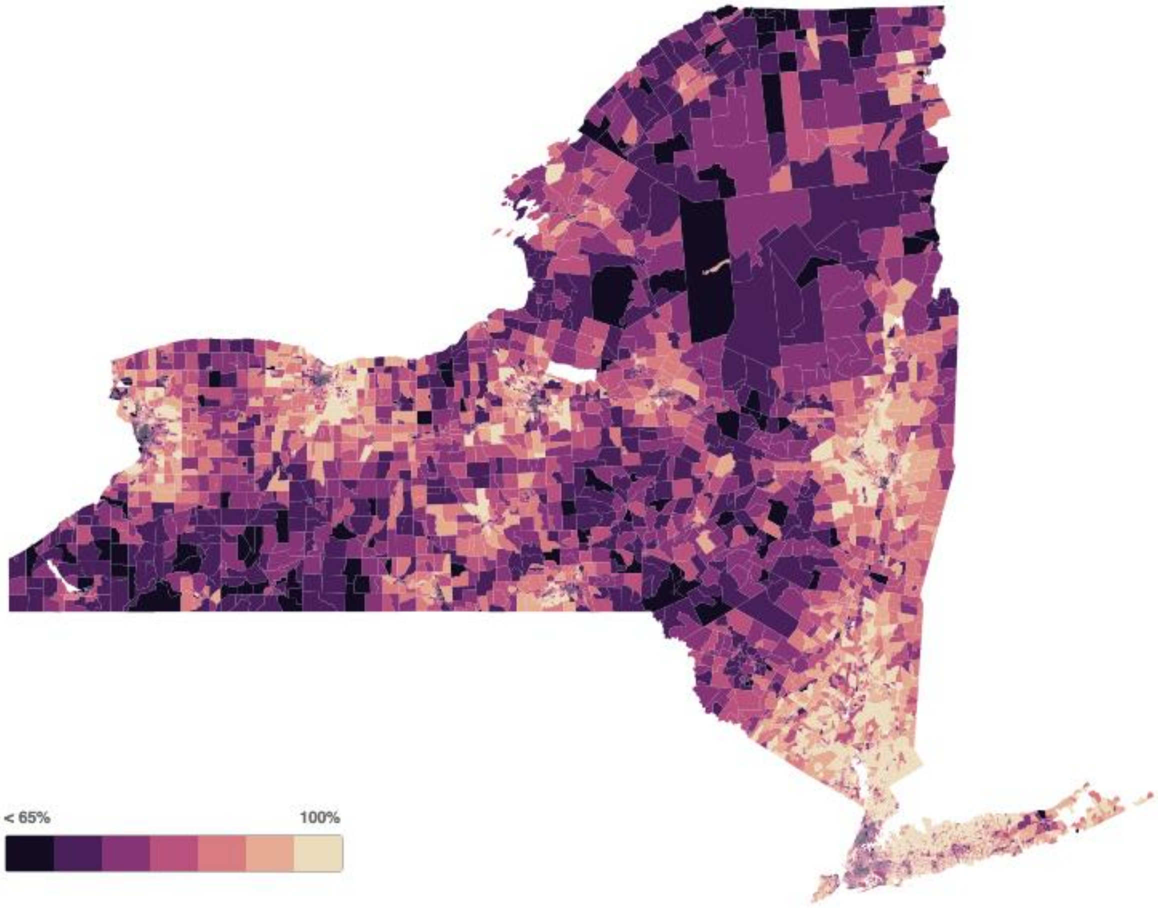

Figure 4 (a): *New York State Map, % of internet subscription per block group, 2013, predicted*

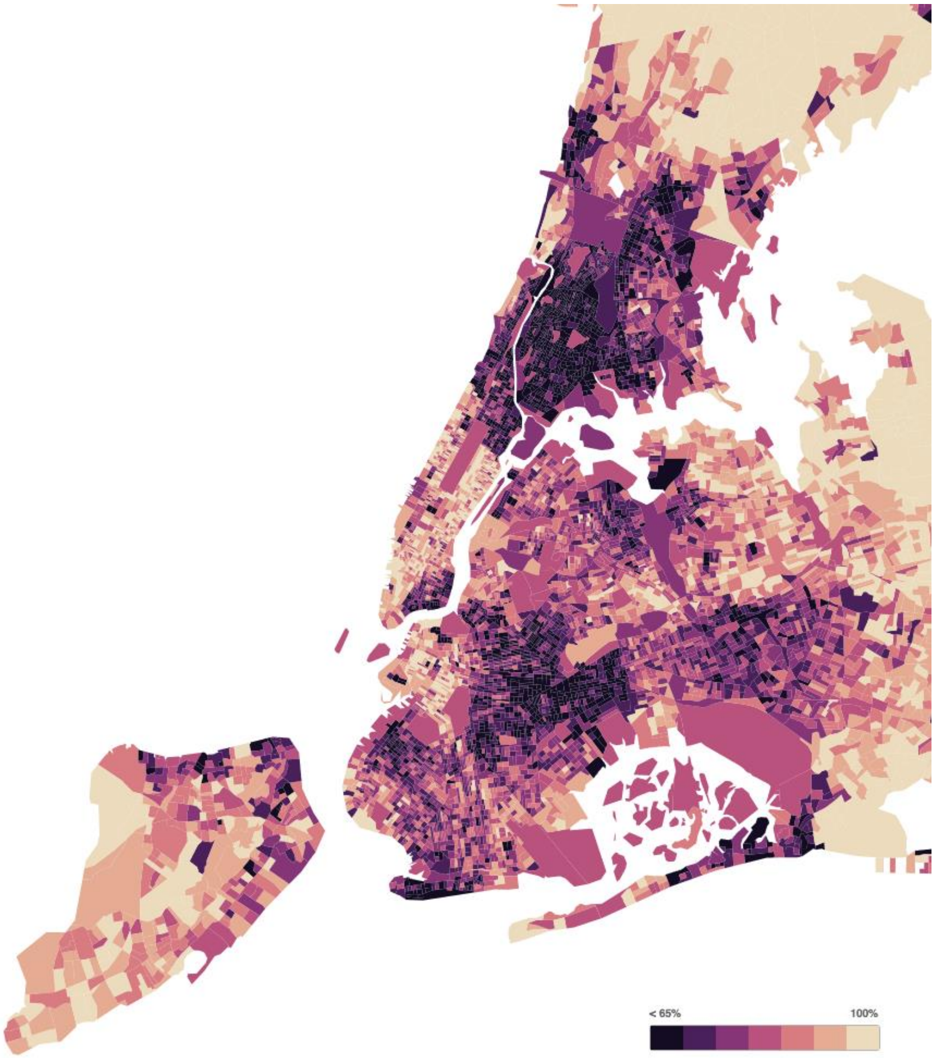

Figure 4 (b): *New York City map, % of internet subscription per block group, 2013, predicted*

8. References

[1] "1940 (Population) - History - U.S. Census Bureau." 2009. 30 Apr. 2016 <https://www.census.gov/history/www/through_the_decades/index_of_questions/1940_population.html>

[2] "ACS Information Guide - Census.gov." 2015. 30 Apr. 2016 <https://www.census.gov/programs-surveys/acs/about/information-guide.html>

[3] "ACS-TP-67 cover.indd - Census.gov." 2009. 30 Apr. 2016 <https://www.census.gov/history/pdf/ACSHistory.pdf>

[4] "INTERNET INEQUALITY: - New York City Comptroller - NYC ..." 2015. 30 Apr. 2016 <https://comptroller.nyc.gov/wp-content/uploads/documents/Internet_Inequality.pdf>

[5] Source: Phone Call to Anthony G. Jr Tersine, Mathematical Statistician at the US Census

[6] "ABOUT | RED HOOK WIFI." 2015. 1 May. 2016 <http://redhookwifi.org/about/>

[7] "Wi-Fi on the Streets of Brooklyn? A Progress Report | Local ..." 2014. 1 May. 2016 <http://thebrooklynink.com/2014/11/30/54213-wi-fi-on-the-streets-of-brooklyn-a-progress-report/>

[8] "Inequality In The Digital World | Equality Indicators." 2016. 1 May. 2016 <http://equalityindicators.org/blog/2016/03/inequality-in-the-digital-world/>

[9] "Wi-Fi CERTIFIED Passpoint | Wi-Fi Alliance." 2013. 1 May. 2016 <http://www.wi-fi.org/discover-wi-fi/wi-fi-certified-passpoint>

[10] "The Cost of Connectivity 2014 - New America." 2014. 1 May. 2016 <https://www.newamerica.org/oti/the-cost-of-connectivity-2014/>

[11] "INTERNET INEQUALITY: - New York City Comptroller - NYC ..." 2015. 1 May. 2016 <https://comptroller.nyc.gov/wp-content/uploads/documents/Internet_Inequality.pdf>

[12] "Open data & analytics for preventing fire deaths - enigma blog." 2015. 30 Apr. 2016 <http://blog.enigma.io/smoke-signals-open-data-analytics-for-preventing-fire-deaths/>

[13] "GitHub - enigma-io/smoke-signals-model: The Machine ..." 2015. 30 Apr. 2016 <https://github.com/enigma-io/smoke-signals-model>

[14] "US Census + Machine Learning to map entirely new ..." 2016. 30 Apr. 2016 <http://blog.cartodb.com/creating-segments-from-the-census/>

[15] Interactive map of Internet access in NYC