Predicting Broadband Subscriptions by U.S. Region

by dehankerson
**Project Background**

**Problem:** Predict subscription rates for broadband Internet by geography, using demographic data (age, gender, and education etc.).

**Hypothesis:** Educational attainment, income and poverty will have the most impact on predicting broadband subscriptions. Disability status may also have an impact on broadband subs.

**Goal:** Predict which factors are related to higher rates of Broadband subscriptions for traditionally low-subscriber group. The context for this question is lower than average home broadband adoption (subscription) rates for a specific subsets of Americans.

**Algorithms/Models Applied:**

- Grid Search
- Random Forests Regression

*Agenda – Project Background --> Problem Statement and Data --> Modeling --> Results --> Conclusion --> Next Steps --> Appendix*
Several groups are shifting their home internet connectivity away from broadband and toward smartphones

<table>
<thead>
<tr>
<th>% of each group who have ...</th>
<th>Broadband at home</th>
<th>Smartphone, but no broadband at home</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2013</td>
<td>2015</td>
</tr>
<tr>
<td>All adults</td>
<td>70%</td>
<td>67%</td>
</tr>
<tr>
<td>African Americans</td>
<td>62%</td>
<td>54%</td>
</tr>
<tr>
<td>Rural residents</td>
<td>60%</td>
<td>55%</td>
</tr>
<tr>
<td>Household income &lt; $20K</td>
<td>46%</td>
<td>41%</td>
</tr>
<tr>
<td>$20K-$50K</td>
<td>67%</td>
<td>63%</td>
</tr>
<tr>
<td>$50K-$75K</td>
<td>85%</td>
<td>80%</td>
</tr>
<tr>
<td>Parents</td>
<td>77%</td>
<td>73%</td>
</tr>
<tr>
<td>High school degree or less</td>
<td>50%</td>
<td>47%</td>
</tr>
</tbody>
</table>

Source: Pew Research Center surveys
PEW RESEARCH CENTER

67%

Home broadband subscriptions in the U.S. has reached a plateau (2015).

→ Additionally, the CDC reports that the number of wireless only homes is nearly fifty percent of all Americans, with particularly high rates for traditionally low-broadband subscriber groups.
Project Background

History: There is an uneven distribution of broadband subscribers in the U.S. Lower than average-subscription rates are typically concentrated within a subset of Americans who tend to have:

1. Low-educational attainment,
2. Fall into lower-income groups and are also part of
3. Non-white ethnic groups.

Why is this important? Broadband-derived benefits are distributed like a patchwork pattern across the US producing disparate levels of economic activity (productivity) grouped by a number of factors. This phenomena has broad societal and economic costs.
Problem Statement and Data

Data Acquisition:

➔ Source: US Census, American Community Survey, (http://factfinder.census.gov)
➔ 2 Years, 2013 and 2014

The full data set included 308 features, (817 rows) and included some of the following categories:
➔ Female Fertility, Place of Birth, Household By Type, Occupation, Housing Tenure, Poverty status etc.

➔ Data columns corresponding to factors known to have an impact on broadband subscriptions were broken out into a separate .csv file. About 70% of the features from the full data set were removed.
Problem Statement and Data: Features at Glance

Feature Selection: The .csv file read into my .ipy notebook included 95 features (817 rows).

Baseline Model v1 MSE from Random Forest Regressor:

- Mean Squared Error on Train Set: 23.364
- Mean Squared Error on Test Set: 107.034
- Mean Squared Error on Entire Set: 56.893
Modeling: Insights

Exploratory Data Analysis, Outcome Variable: 'Hseholds_wbbandInternet_subs%

```python
data['Hseholds_wbbandInternet_subs%'].describe()
```

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>789.000000</td>
</tr>
<tr>
<td>mean</td>
<td>74.238023</td>
</tr>
<tr>
<td>std</td>
<td>10.142104</td>
</tr>
<tr>
<td>min</td>
<td>44.100000</td>
</tr>
<tr>
<td>25%</td>
<td>66.700000</td>
</tr>
<tr>
<td>50%</td>
<td>75.500000</td>
</tr>
<tr>
<td>75%</td>
<td>81.800000</td>
</tr>
<tr>
<td>max</td>
<td>96.600000</td>
</tr>
</tbody>
</table>

Histogram Hseholds_wbbandInternet_subs%
Exploratory Data Analysis, Outcome Variable: 'Hseholds_wbbandInternet_subs%' (For all USRegions (Midwest =1, West = 2, South = 3, Northeast = 4))
Feature Importances: Top 4 feature importances are
1. Hseholds_w_comp%
2. Edu_Pop_25yrsandover_BAdegree%
3. Poverty_Allpeople_18_64yrs%
4. Edu_Pop_25yrsandover_HSgrad%
Running a pair plot for top 3 features where hue='Pop_Group' helped to prove some initial assumptions about low-subscriber groups.
Modeling: Approach
(Random Forest Regressor)

What steps did your project take?
1. Removed missing values
2. Exploratory Data Analysis
3. Created a base model
4. Hyperparameter tuning using Grid search
5. Plot feature importances

Best Params:
{'n_estimators': 300}

Accuracy of current model: 0.909
Accuracy using best params: 0.912

RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=None, max_features='auto', max_leaf_nodes=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=1, oob_score=False, random_state=None, verbose=0, warm_start=False)

Accuracy of current model: 0.909
RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=None, max_features='auto', max_leaf_nodes=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=300, n_jobs=1, oob_score=False, random_state=None, verbose=0, warm_start=False)

Accuracy of updated model: 0.917
Modeling: Approach

- Feature selection reduced feature set from 95 features to 43 features.
- In order to interpret the visualizations it helps to remember that each feature has distinct values. For example, some columns measure percent of population, others represent population figures in the thousands.
- ‘Hseholds_w_comp%’ had the highest importance ranking and because it changed the scale of the barplot, it was removed from the visualization in order to show relative importance of other features in the model.

Agenda – Project Background --> Problem Statement and Data Cleansing --> **Modeling** --> Results --> Conclusion --> Next Steps --> Appendix
Features with the highest impact on my MSE

- 'Tot_PopMale%
- 'Tot_PopFem%
- 'Occhsing_Mobilehomeboateetc%
- 'Total_Pop'
- '18_24yrs%
- '25_34yrs'
- '35_44yrs'
- '45_54yrs'
- '55_64yrs'
- '65_74yrs'
- '75_yrsandover%
- 'Median_ageyrs'
- '18_yrsandover%
- '21_yrsandover%
- '62_yrsandover%
- 'Hseholds_thous'
- 'TotPop_18yrsandover_thous'
- '18yrsandover_Male%
- '18yrsandover_Fem%
- 'TotPop_18_34yrs_thous'
- '18_34yrsMale%
- 'TotPop_35_64yrs_thous'
- '65_yrsandover_Fem%
- Hseholds_thous'
- 'Hseholds_Fmlyhseholds'
- 'TotPop_18yrsandover_thous'
- Fmlyhseholds_MarriedcoupleFmly%
- 'Fmlyhseholds_Marriedcouplew_own_childunder18yrs'
- Pop_3yrsandover_inschool_College_gradschool%
- 'Edu_Pop_25yrsandover_LessthanHS%'
- 'Edu_Pop_25yrsandover_HSgrad%'
- 'Edu_Pop_25yrsandover_BAdegree%'
- 'Edu_Hsgradorhigher_Fem%'
- 'Poverty_Allpeople_18yrsandover%
- 'Poverty_Allpeople_18_64yrs%
- 'OccHsingunits_thous'
- 'Ownerocchsingunits%'
- 'Occhsing_2_4units%'
- 'OccHsing built1940to1959%'
- 'OccHsing built1939or%
- Hseholds_w_comp%
- 'Pop_Group'
- 'USRegion'
Results

Agenda – Project Background --> Problem Statement and Data Cleansing --> Modeling --> Results --> Conclusion --> Next Steps --> Appendix
Histograms by Region for 'Hseholds_wbbandInternet_subb%'

Midwest=1

South=3

West=2

Northeast=4
Barplots by Region for 'Hseholds_wbbandInternet_subs%' GroupBy 'Pop_Group'

Midwest=1

West=2
Barplots by Region for ‘Hseholds_wbbandInternet_subs%’ GroupBy ‘Pop_Group’

South=3

Northeast=4
Feature Importances by Region for 'Hseholds_wbbandInternet_subs%'

Midwest=1

West=2

Agenda – Project Background-->Problem Statement and Data Cleansing-->Modeling -->Results -->Conclusion-->Next Steps-->Appendix
Feature Importances by Region for 'Hseholds_wbbandInternet_subs%'

South = 3

- Poverty_Allpeople_18_64yrs%
- Poverty_Allpeople_18yrsandover%
- 18_24yrs%
- Edu_Pop_25yrsandover_HSgrad%
- Hseholds_Fmlyhseholds
- 75_yrsandover%
- Median_ageyrs
- 21_yrsandover%
- 18yrsandover_Male%

Northeast = 4

- Poverty_Allpeople_18_64yrs%
- Edu_Pop_25yrsandover_HSgrad%
- Fmlyhseholds_MarriecoupleFmlly%
- 13_24yrs%
- Edu_Pop_25yrsandover_BAdegree%
- tb_yrsandover_rem%
- Fmlyhseholds_Marriedcoupleow_childunder18yrs%
- Edu_Pop_25yrsandover_LessthanHS%
- 75_yrsandover%
- Pop_3yrsandover_inschool_College_gradschool%

Agenda – Project Background→Problem Statement and Data Cleansing→Modeling →Results →Conclusion→Next Steps→Appendix
Conclusion

- Ranked feature importances provide additional insight into this problem beyond what is provided in general research available for this topic.
- `Pred_for_new_test` Model predicts Broadband subscription rate and probability of broadband subscription uptake at that rate for any locale with specific demographic features.
Next Steps

What should this project do moving forward?

1. Improve MSE measure, tuning the model further.
2. ‘NAN’ errors for some columns, troubleshoot and fix, add these features to the model.
3. Create a more user centric means of interacting with the model; a way to input demographic values so that the model returns a broadband subscription rate
Acknowledgements:
Austin Brown
Jim Simpson
Kate Riesbeck
Rajashree Baradur
Matt Gordon

Thank you for all your help and support!

Citations:
DeVan Hankerson
@dvhankerson
@TelecomRick