Springboard Data Science

Capstone 1

Predicting Housing Satisfaction with Machine Learning

Final Report

Table of Contents

01

Executive Summary

Springboard Data Science

March 2018

01 Executive Summary

This project aimed to predict whether a person was satisfied or dissatisfied with their housing using their responses to the American Housing Survey.

The vast majority of people tended to be 'satisfied' with their housing. Surprisingly, the factor correlated with housing satisfaction the most was whether or not a person was satisfied with their neighborhood.

Outside of neighborhood satisfaction, people tended to be more strongly affected by factors that can be considered "temporary," such as whether or not there is peeling paint, rather than factors that are permanent. Nine out the fifteen top factors related to housing satisfaction are features of a home that can be repaired or changed, such as whether or not there is peeling paint or damage on their roof. This indicates that property managers can actually affect housing satisfaction levels within their own buildings.

A logistic regression model was used for predicting a binary response variable, where 1 equaled satisfaction with housing and 0 equaled dissatisfaction. The greatest accuracy achieved was 0.93 with a tuned logistic regression model. While a random forest model was also attempted, tuning the hyperparameters took considerably more time, with no increase in model performance.

The entire project folder can be found here: https://github.com/kellywong1314/Springboard-Data-Science/tree/master/Capstone%201%20-%20American%20Housing%20Satisfaction

02

Background

Springboard Data Science

March 2018

02 Background

02.01 Purpose

Housing preferences in the United States have changed notably over the last several decades for a multitude of reasons, such as increased density in urban cores, changes in the average family size, and shifting trends in architectural styles.

The intent of this study is to elucidate which factors affect housing satisfaction the most and examine whether or not there are variations in housing preferences from region to region. Possible factors influencing housing satisfaction include, but are not limited to: home size, cost of housing, quality of the home, or age of the home. These can all be considered "intrinsic" to each home after construction. Other "temporary" factors may also affect housing satisfaction, such as the presence of peeling paint, whether or not the home has a dishwasher, or if the home's exterior is damaged.

Finally, this study will use a machine learning model to predict whether a survey respondent is satisfied or is satisfied with their housing using the other survey questions.

02.02 Intended Audience

There are several groups who would be able to leverage this data to make informed decisions:

Housing Developers
This data could help developers create high level strategies for new housing developments, as well as aid in decision-making for purchasing existing housing developments.

Large-Scale Property Owners
Property owners could tailor their renovation schedules to reduce the risk of tenant churn over time, and prioritize changing housing units which have characteristics that make it likely for tenants to be dissatisfied.

Municipalities
City redevelopment agencies or neighborhood community development groups are inclined to incentivize housing developments that are appealing to residents, and as a result, could use this data to guide their policies and development initiatives to promote resident retention.

02.03 Data Source

The data used for this project is from the 2015 American Housing Survey (AHS). This data is available at the following URL: https://catalog.data.gov/dataset/american-housing-survey-ahs.

The AHS national survey was conducted annually from 1973-1981 and every two years from 1983 – 2015. The survey questions have changed over time, but the questions have always pertained to the characteristics of the housing unit, such as the size and composition, as well as characteristics of the tenants and the survey respondents, such as their age, race, or income level.

For the purposes of this project, only characteristics directly related to the housing unit and the surrounding neighborhood were used, and questions pertaining to the socioeconomic background of the respondent, their family, and their fellow tenants were disregarded.

The survey question used as the target variable in the logistic regression model is "Overall Opinion of Present Home (1 – 10)." The responses were split into two categories, where 1 through 5 are classified as 0, or dissatisfied, and 6 through 10 are classified as 1, or satisfied.

The original dataset contained many more fields than were needed, had quite a few missing values, and was stored as an encoded csv file. The data wrangling steps taken to clean the data up and make it usable for the machine learning model is contained here: https://github.com/kellywong1314/Springboard-Data-Science/blob/master/Capstone%201%20-%20American%20Housing%20Satisfaction/Data%20Wrangling_American%20Housing%20Survey%202015.ipynb

03

Exploratory Data Analysis

Springboard Data Science

March 2018

03 Exploratory Data Analysis

A few simple plots were created using seaborn to examine and understand the data and explore how the ratings for homes and neighborhoods were distributed.

Figure 1 is a bar plot depicting the distribution of home ratings. The most common rating that respondents gave their homes was a 10, followed by 8. Very few people gave their homes a rating lower than 5. As the model is trying to predict one of two categories, the uneven distribution of the target variable may make it tricky.

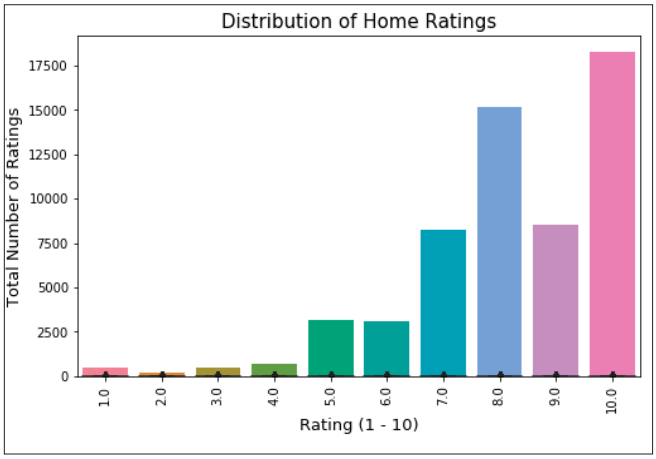

Figure 1

When home ratings are divided by metropolitan region, it is possible to see if there is a significant difference in housing satisfaction by region. The 2015 survey only included these sixteen regions and a category for "non-metropolitan" regions. The median rating for almost every region is 8, but the range of ratings varies slightly by region. The range of ratings for DC, Boston, and Atlanta were smallest, with 95% of ratings falling between 5 and 10. Seattle's distribution of ratings is also slightly different.

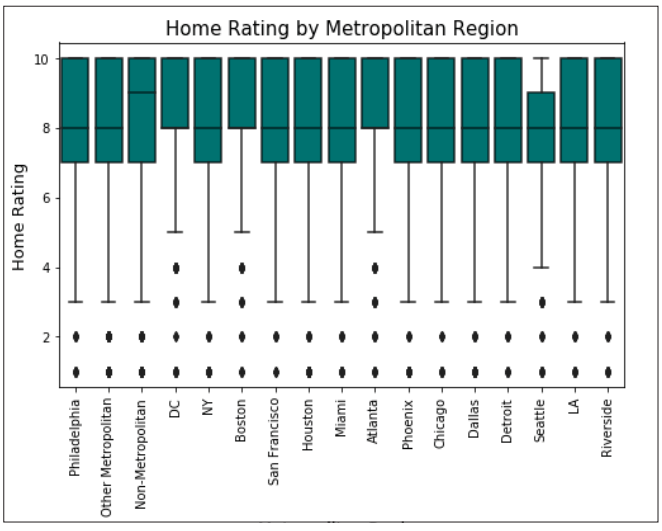

Figure 2

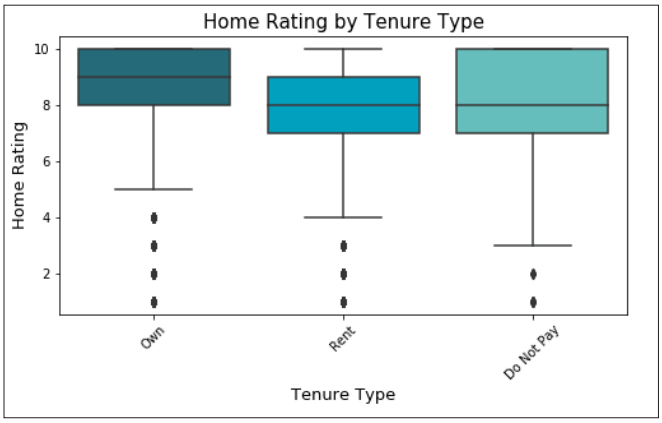

Figure 3

It may be worth examining the differences between ratings in metropolitan regions more closely.

Do home ratings vary between renters and homeowners? Figure 3 shows home ratings divided by renters, owners, and survey respondents who do not pay for their housing, such as people who live in a friend's home.

Homeowners tended to rate their homes higher than respondents who either rent their housing or do not pay for their housing, as shown in Figure 3.

If we split the ratings for every tenure type into their respective metropolitan regions, it can be seen that renters and homeowners tend to rate their homes similarly regardless of metro region, as shown in Figure 4.

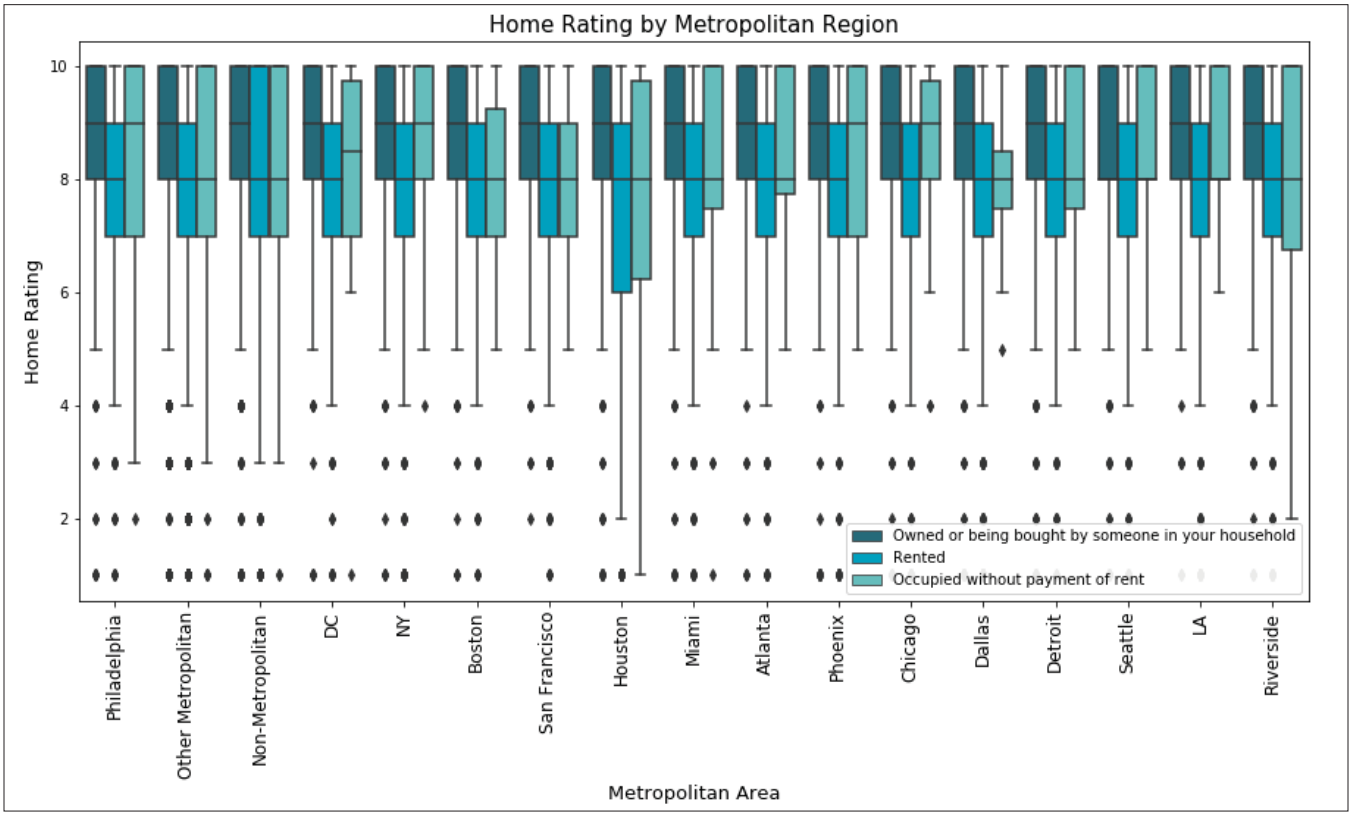

Figure 4

The number of rooms also may affect home ratings. Figures 5 and 6 depict home ratings by the number of rooms in the home and the number of rooms per resident. Homes with more than six rooms tend to have the highest average and median ratings, as did homes with more than two rooms per resident.

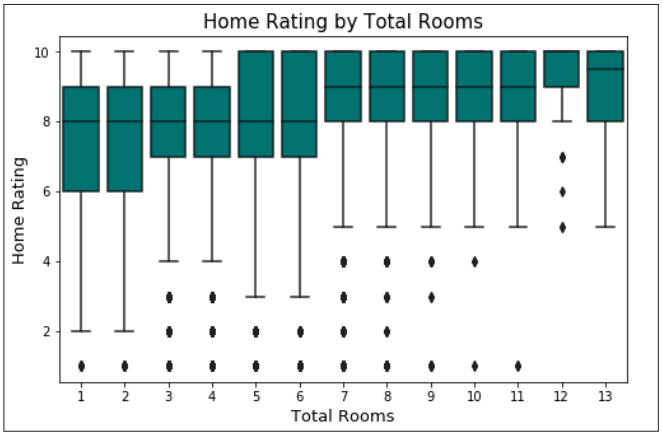

Figure 5

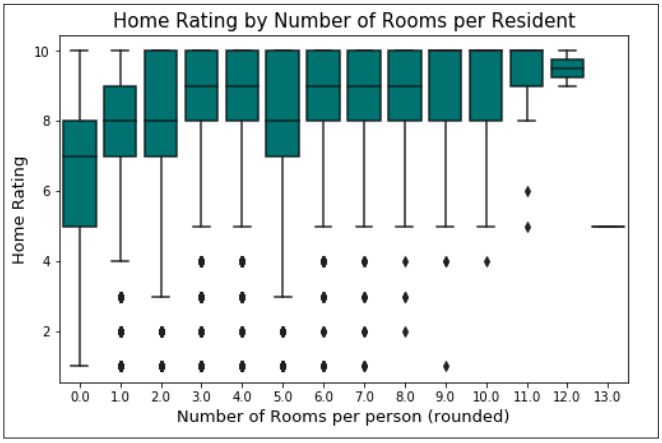

Figure 6

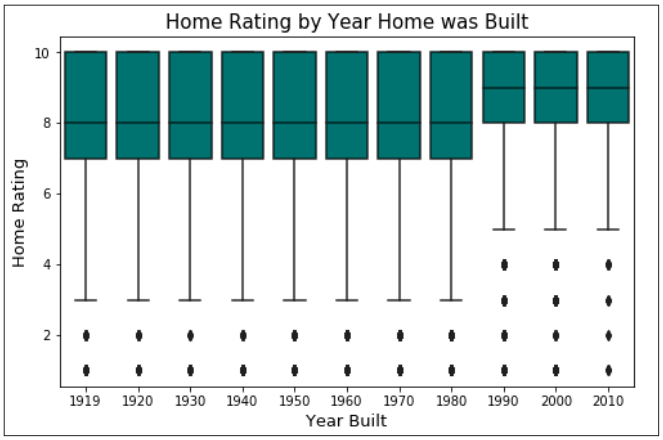

Figure 7

Finally, Figure 7 shows home rating by when the home was built. There was a distinct gap in ratings in homes constructed prior to 1990.

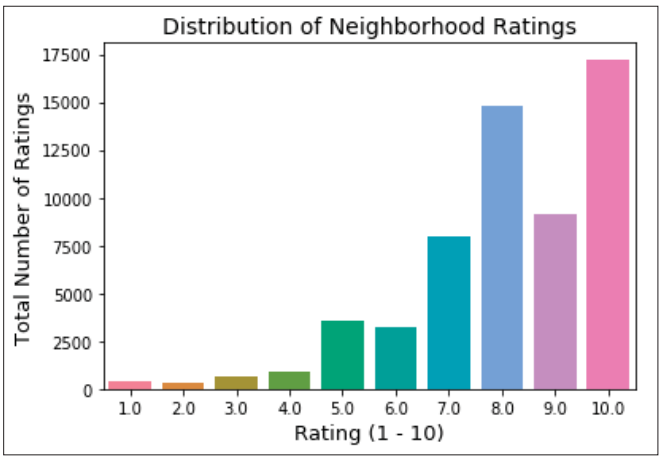

Figure 8

The neighborhood ratings were also charted onto a bar plot, and it shows a very similar distribution as the home ratings—10 is the most common rating, followed by 8, as depicted in Figure 8.

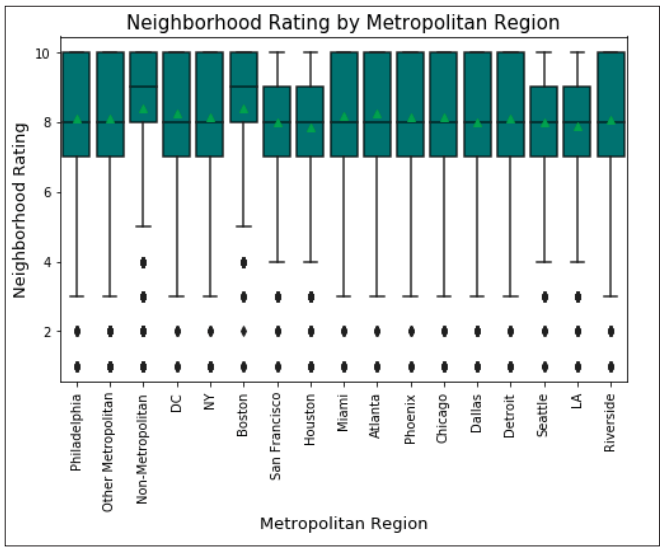

Figure 9

Neighborhood ratings are generally similar amongst the different metropolitan regions, but Boston and Non-Metropolitan neighborhoods had the highest median and average ratings.

The code used to generate these graphics can be found here: https://github.com/kellywong1314/Springboard-Data-Science/blob/master/Capstone%201%20-%20American%20Housing%20Satisfaction/EDA_American%20Housing%20Survey%202015.ipynb

04

Statistical Analysis

Springboard Data Science

March 2018

04 Statistical Analysis

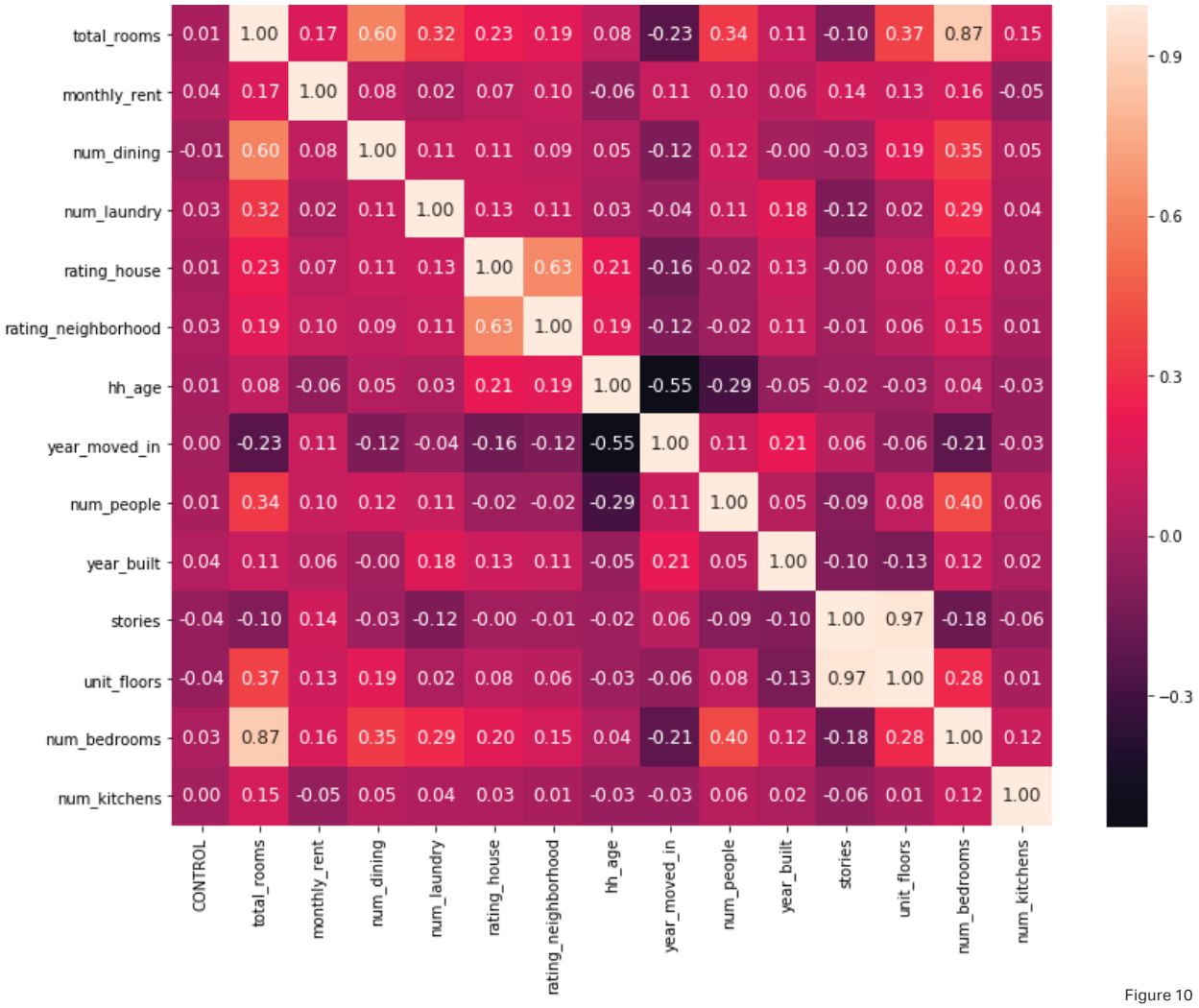

Figure 10

Statistical methods were applied to determine which features are most strongly related to the target variable. These methods would also guide feature selection if it was needed for the machine learning model.

04.01 Correlation Matrix

For the numerical features, a correlation matrix was omputed. The majority of features had very low correlations with the target feature, rating_house. Very few features had a correlation higher than 0.20. The entire correlation matrix is depicted above.

The values with the highest correlation with the target variable were:

- rating_neighborhood (0.630, p = 0.0)
- total_rooms (0.230, p = 0.0)
- hh_age (0.214, p = 0.0)
- num_bedrooms (0.196, p = 0.0)

All of the other correlations were under 0.20, but still have p-values under 0.05 (most of the p-values are 0). This indicates that the correlations are all statistically significant. However, the low Pearson correlations may make it difficult to use any of the numerical values in the machine learning model.

04.02 ANOVA

For the categorical variables, an analysis of variance (ANOVA) was performed to determine if any of the categorical variables contain a statistically significant difference between groups when looking at the home ratings.

The entire table of f-values and their p-values is attached in Appendix B, but the top 15 values are as follows:

Feature	f-value	p-value
petty_crime	2,642.07	0.00E+00
serious_crime	2,449.85	0.00E+00
rating_neighborhood	2,106.82	0.00E+00
dishwash	1,746.21	0.00E+00
wall_crack	1,611.99	0.00E+00
garage	1,492.11	2.47E-322
paint_peeling	1,332.31	1.98E-288
good_schools	1,309.70	2.49E-283
washer	1,168.80	1.21E-253
near_trash	1,107.91	0.00E+00
number_upkeep_probs	1,056.57	0.00E+00
adequacy	1,028.01	0.00E+00
floor_hole	986.54	9.52E-215
tenure	875.32	0.00E+00
home_better_than_last	804.17	0.00E+00

04.03 Chi-Squared Test

While an ANOVA analysis shows the strength of the relationship between home ratings and the categorical variables when the rating is treated as a continuous variable, the machine learning model will be predicting a binary response variable.

To examine the relationship between the categorical variables and a binary category for home ratings, a X^2 test for independence was performed. The 15 features with the highest X^2 statistics are as follows:

Feature	X^2 Statistic	p-value
num_bathrooms	3855.34	0.00E+00
rent_subsidy	3295.41	0.00E+00
adequacy	2571.49	0.00E+00
roach	2409.25	0.00E+00
unit_size	2078.07	0.00E+00
wall_crack	1470.20	0.00E+00
hud_subsidized	1385.21	3.43E-303
roof_sag	1326.01	2.52E-290
rodent	1305.28	8.06E-286
roof_hole	1300.68	8.02E-285
missing_siding	1276.19	1.69E-279
paint_peeling	1272.18	1.26E-278
missing_shingle	1250.17	7.62E-274
tenure	875.32	0.00E+00
home_better_than_last	804.17	0.00E+00

The top three features are characteristics of housing that are difficult to change, such as the number of restrooms, whether or not the building is subsidized, and whether the resident thought the housing was adequate for his or her needs.

However, the majority of the top fifteen features are characteristics that can be changed, such as the presence of pests, missing siding, or peeling paint.

The code used to perform this analysis can be found here: https://github.com/kellywong1314/Springboard-Data-Science/blob/master/Capstone%201%20-%20American%20Housing%20Satisfaction/Statistical_Analysis_AHS.ipynb

05

Machine Learning

Springboard Data Science

March 2018

05 Machine Learning

To predict whether or not a survey respondent is satisfied with their housing, a machine learning model was used. First, a binary response variable was created, where home ratings of 1 to 5 are classified as 0 and ratings of 6 to 10 are classified as 1. Out of the 58,233 samples, 53,285 are classified as 1, and 4,950 are classified as 0.

Naïve Model

As the majority of respondents rated their homes above 5, a dummy model predicting the most frequent class was created. This model attained a fairly high accuracy overall, but has a recall of 0 for predicting the negative class.

Negative	Positive
0	1,449
0	16,021

	Precision	Recall	f1-Score	Support
0	0.00	0.00	0.00	1,449
1	0.92	1.00	0.96	16,021
avg	0.84	0.92	0.88	17,470

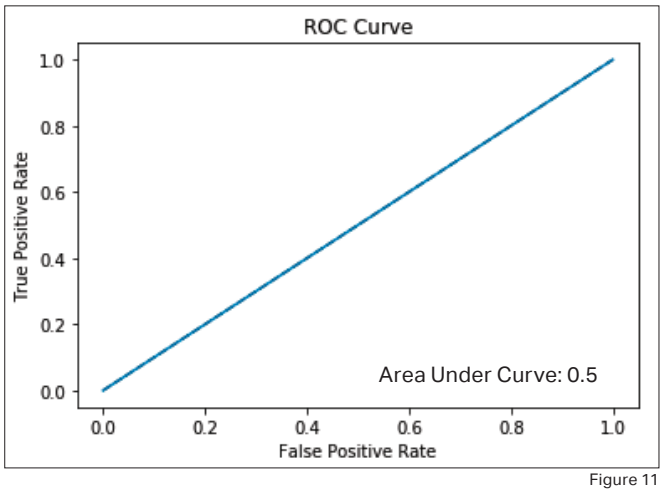

Figure 11

Out of the Box Logistic Regression

First, a logistic regression model was trained with the default parameters. This model already has a slightly higher overall accuracy than the naïve model, but the recall and precision for the negative class is low.

		PREDICTED	
		Negative	**Positive**
ACTUAL	**Negative**	396	1,053
	Positive	227	15,794

	Precision	Recall	f1-Score	Support
0	0.64	0.27	0.38	1,449
1	0.94	0.99	0.96	16,021
avg	0.91	0.93	0.91	17,470

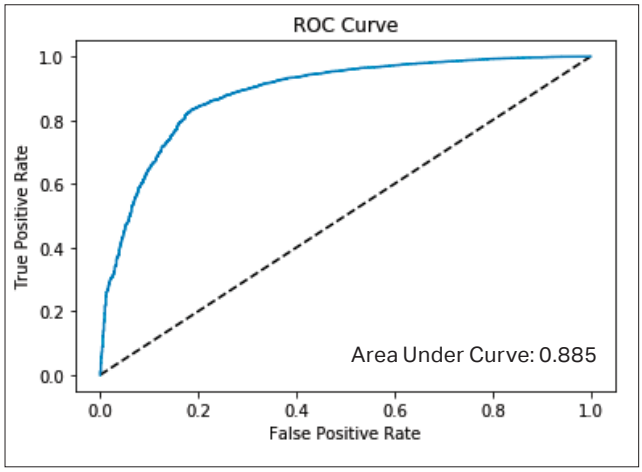

Figure 11

Tuned Logistic Regression

Tuning the hyperparameter C may increase the model's performance. C values of 0.0001, 0.001, 0.01, 0.1, 1, 10, and 100 were tested using cross validation. A C value of 100 resulted in the highest accuracy. Because lower C values result in stronger regularization, it seems that less regularization results in better model performance in this case.

		PREDICTED	
		Negative	**Positive**
ACTUAL	**Negative**	396	1,053
	Positive	227	15,794

	Precision	Recall	f1-Score	Support
0	0.64	0.27	0.38	1,449
1	0.94	0.99	0.96	16,021
avg	0.91	0.93	0.91	17,470

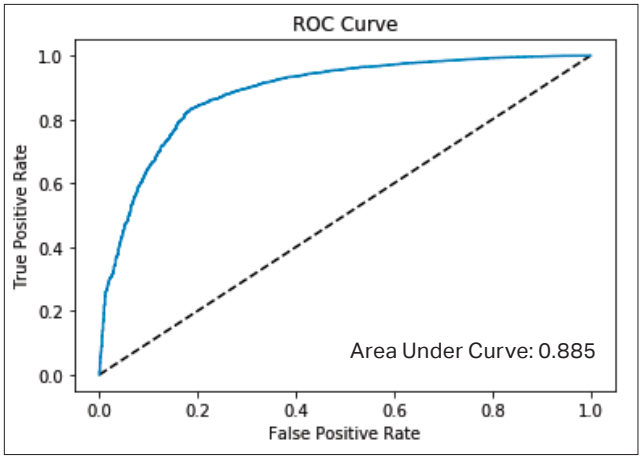

Figure 12

Although the precision and recall didn't increase, the tuned model is more likely to perform better for new data than an untuned model.

Tuned Logistic Regression with Feature Selection

Feature selection was performed to examine whether removing features could improve model performance. Features were ranked by their X^2 statistic and removed if they fell below a certain threshold.

However, the model's performance declined when features were removed. The more features removed, the lower the model's accuracy, although only slightly. If model complexity were a concern, features can be removed with a small loss in accuracy, but for the final model, all of the features were left in the model.

		PREDICTED	
		Negative	**Positive**
ACTUAL	**Negative**	388	1,061
	Positive	235	15,786

	Precision	Recall	f1-Score	Support
0	0.62	0.27	0.37	1,449
1	0.94	0.99	0.96	16,021
avg	0.91	0.93	0.91	17,470

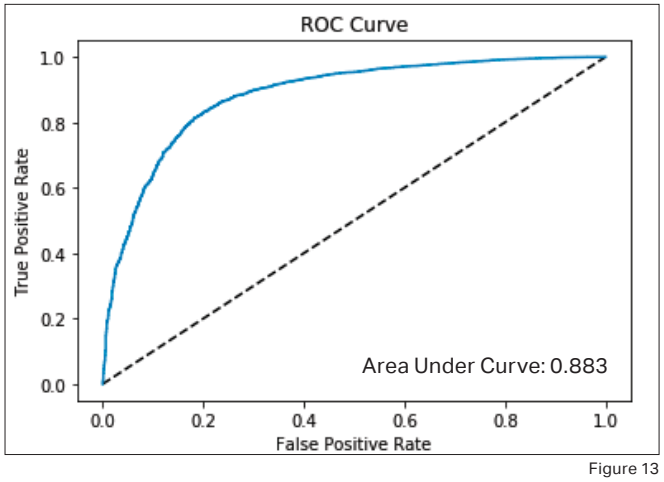

Figure 13

Tuned Logistic Regression with Undersampling

Because the response variable has only two classes and the positive class is present with an 11 to 1 ratio, every model is much better at predicting the modal class, 1, than it is at predicting the negative class. If the goal of the model is to perform equally as well for both classes, the dataset should be undersampled so that both response variables are present in equal amounts.

The dataset was undersampled so that both classes had 9,900 records, and then split into training and test sets. The hyperparameter C was tuned as well, and had the best results with a C value of 0.01.

The overall accuracy for the model decreased to 0.80, but the model performed equally well in predicting the positive class as it did in predicting the negative class.

		PREDICTED	
		Negative	**Positive**
ACTUAL	**Negative**	1,199	311
	Positive	270	1,190

	Precision	Recall	f1-Score	Support
0	0.82	0.79	0.80	1,510
1	0.79	0.82	0.80	1,460
avg	0.80	0.8	0.80	2,970

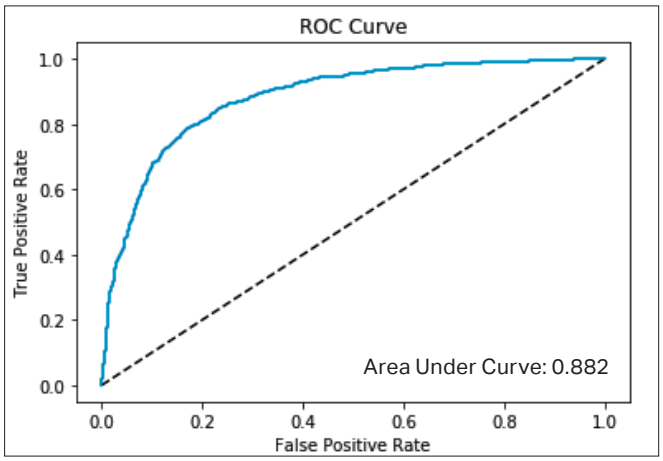

Figure 14

Random Forest Models

A few random forest models were also trained to try to improve model performance, but tuning the model hyperparameters took much longer. A linear model seemed to work very well for the features in this dataset.

While the model performance is comparable to the logistic regression model's, it is debateable whether or not it was worth the extra time it took, as the model's memory consumption was significantly higher.

		PREDICTED	
		Negative	**Positive**
ACTUAL	**Negative**	335	1,114
	Positive	157	15,864

	Precision	Recall	f1-Score	Support
0	0.68	0.23	0.35	1,449
1	0.93	0.99	0.96	16,021
avg	0.91	0.93	0.91	17,470

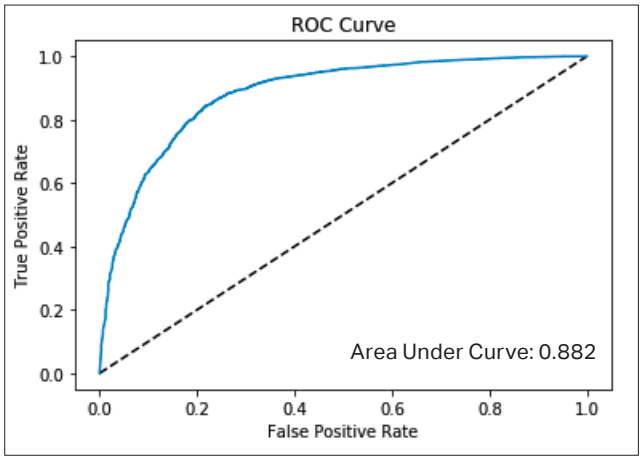

Figure 15

The Jupyter notebook with the code used to generate these models can be found here: https://github.com/kellywong1314/Springboard-Data-Science/blob/master/Capstone%201%20-%20American%20Housing%20Satisfaction/Predicting%20Housing%20Satisfaction%20with%20Machine%20Learning.ipynb

06

Conclusion

Springboard Data Science

March 2018

06 Conclusion

A simple logistic regression model resulted in the highest success in predicting the binary response variable. A tuned logistic regression model resulted in an overall accuracy of 0.93, and a precision of 0.91, with an area under the ROC curve of 0.885. However, the model was much more likely to predict the positive class ("Satisfied with Housing") than the negative class ("Dissatisfied with Housing"), as it was present in the dataset with an 11 to 1 ratio. Undersampling can be used to even out the model's performance between the two classes, but it drops the model's overall accuracy and precision.

The X^2 statistic was used to attempt feature selection, but feature selection did not improve model performance. Nonetheless, the X^2 statistic can be used to rank features on their relative importance to the response variable, and guide decision makers in how to increase the likelihood of their residents being satisfied with their housing housing.

This project also illustrated many of the difficulties in making predictions of a qualitative measurement using survey results. It is difficult to determine which features, if any, have a truly causal relationship to the response variable.

For a future project, it may be useful to consolidate similar surveys geared toward housing satisfaction to obtain more samples. Analyzing the historical survey data may also provide some insight into how housing preferences have changed over time.

Appendices

Springboard Data Science

March 2018

Appendix A: Data Wrangling

Overview

The data used for this project is from the 2015 American Housing Survey. As the survey has far more attributes than are needed for an initial examination, we pull out the 130 most relevant attributes and examine those. The chosen attributes are characteristics that are directly related to the respondents' homes and neighborhoods, such as the cost of the housing, the room composition, and whether the home is rented or owned. Socioeconomic characteristics of the survey respondent, such as their race, are excluded.

Initial Impression

The dataset has almost 70,000 entries and 1,091 attributes for each. The column names are abbreviated and make it difficult to understand what each attribute is. In addition, the values in the DataFrame are all encoded with numbers stored as strings. We will have to pull out the relevant columns, map all of the original values back into the DataFrame, and rename the columns so that it is clear what each attribute is when the analysis is performed.

Files

For this project, we need the original csv with all of the survey answers as well as two other files which are used to subset the data and rename values:

household.csv: This file holds all of the survey answers to the 2015 American Housing Survey

Column Values.xlsx: This excel file has the 129 columns that will be used in the analysis, and the new names for the columns.

AHS 2015 Value Labels.csv: This csv has all of the values and what they should be mapped to.

Data Wrangling

The household.csv file is read in as df_household. It has 69,493 entries and 1,091 columns. To select the 129 columns that will be used in the analysis, 'Column Values.xlsx' is read in as df_colnames and then used to select the columns from df_household. The new DataFrame has 129 columns and is named df_household_clean.

The .strip() method is used to remove all of the single

quotation marks which surround all of the values in the DataFrame.

The 'AHS 2015 Value Labels.csv' is read in as df_valuelabels and used to create a dictionary for every attribute so that the values can be mapped into df_household_clean. After that, the columns of df_household_clean are renamed using the new names in df_colnames.

Finally, the null values in df_household_clean, which are stored as -9 or -6, are changed in NaN values. Any record which has a null value in the rating_house or rating_neighborhood columns is dropped, as these two columns are necessary for this project.

Final DataFrame

The resulting DataFrame, df_household_clean, has 58,233 entries and 129 columns. All of the values are either remapped to the correct category, or are integers.

The data wrangling steps and the data can be viewed here: https://github.com/kellywong1314/Springboard-Data-Science/blob/master/Capstone%201%20-%20American%20Housing%20Satisfaction/Data%20Wrangling_American%20Housing%20Survey%202015.ipynb

Appendix B: ANOVA Table

The following table contains all of the f-values and their corresponding p-values for an ANOVA analysis of every categorical feature against the home ratings

Feature	f-value	p-value
petty_crime	2,642.07	0.00E+00
serious_crime	2,449.85	0.00E+00
rating_neighborhood	2,106.82	0.00E+00
dishwash	1,746.21	0.00E+00
wall_crack	1,611.99	0.00E+00
garage	1,492.11	2.47E-322
paint_peeling	1,332.31	1.98E-288
good_schools	1,309.70	2.49E-283
washer	1,168.80	1.21E-253
near_trash	1,107.91	0.00E+00
number_upkeep_probs	1,056.57	0.00E+00
adequacy	1,028.01	0.00E+00
floor_hole	986.54	9.52E-215
tenure	875.32	0.00E+00
home_better_than_last	804.17	0.00E+00
wall_slope	743.26	2.99E-162
roof_hole	730.77	1.41E-159
roof_sag	729.04	3.30E-159
toilet_broke	689.17	5.17E-151
missing_siding	677.6	3.29E-148
in_water_leaks	611.95	2.09E-134
too_cold	585.86	1.31E-252
out_water_leaks	513.66	3.14E-113
windows_broken	500.06	4.02E-110
porch	484.33	6.68E-107
roach	443.47	0.00E+00
hoa	438.23	6.04E-97
foundation_crumb	421.03	4.15E-93
near_abandoned	417.99	1.13E-268
missing_shingle	409.64	1.19E-90
nh_better_than_last	372.62	3.63E-233
dryer	349.9	3.24E-298
risk_of_flood	348.58	1.46E-77
windows_boarded	348.32	2.01E-77
fireplace	348.1	4.49E-224
no_running_water	310.31	2.84E-69
musty	305.53	6.90E-258
rodent	260.25	4.56E-222

Feature	f-value	p-value
near_bar_windows	230.88	1.35E-100
unit_size	206.45	0.00E+00
num_bedrooms	203.81	3.64E-216
num_bathrooms	203.23	0.00E+00
subdivision	199.27	3.89E-45
num_laundry	196.61	7.93E-86
num_dining	183.29	4.45E-80
stairs	181.54	2.98E-41
household_type	150.5	2.65E-190
bldg_type	141.71	5.91E-266
unit_floors	131.3	1.42E-57
total_rooms	118.24	5.27E-293
sewerbreakdowns	83.95	3.46E-88
entry_sys	82.38	1.26E-19
windows_barred	73.33	1.13E-17
fuse_blow	68.31	1.90E-71
housing_cost	66.37	6.33E-214
num_kitchens	55.28	1.06E-13
fridge	54.94	1.26E-13
year_built	54.09	2.63E-109
stairs_broken	50.14	1.50E-12
near_transit	49.54	1.97E-12
is_condo	21.27	3.99E-06
hud_subsidized	20.95	8.15E-10
partner_household	12.53	3.47E-10
rent_subsidy	11.65	7.06E-15
num_people	10.86	6.91E-27
interview_lang	7.75	4.29E-04
stories	6.82	3.02E-07
hh_age	5.18	1.04E-38
metro_area	5.01	1.63E-10
year_moved_in	5	2.07E-45
gut_rehab	4.41	3.57E-02
manager_onsite	3.79	9.88E-03
monthly_rent	3.59	6.80E-43
rent_control	3.12	7.76E-02
bath_exclusive	0.15	6.97E-01
kitchen_exclusive	0	9.55E-01