Detecting Housing Price Shifts Using Machine Learning Methods

Nicole Lange

Nicholas Colonna

Abha Jhaveri

QF 301: Advanced Time Series Analysis & Machine Learning

Professor German Creamer

December 10, 2018

1 Problem Understanding

In the aftermath of the 2008 Financial Crisis that was brought on by a collapse in the housing market, forecasting the health of the housing market has been an important subject of research and analysis over the past decade. Unlike the stock market, where most people understand and accept the risk that stock prices might fall, most people who buy a house do not think that the value of their home might decrease. The key to making a good "risk-based" decision is to understand and measure the risks by making financially sound estimates. This is especially applicable to the largest and most important financial decision most people make - the purchase and financing of a home, which our group members and classmates will likely be looking to to do in the next five to seven years.

The total value of all homes in the United States has a total value greater than $32 trillion, which is larger than the U.S. equity market (Ramirez 2017). Given its size, the housing market has the potential to cause significant disruptions to the economy if large price fluctuations occur. The housing bubble in 2008 triggered housing prices to decline more than 40% and led to mortgage defaults that caused millions of foreclosures. This project examines advanced machine learning methods that can be used as a tool to predict the directional changes of United States housing prices in the future. By doing so, homeowners can be aware of future contractions in the housing market, and therefore real estate investors and homeowners can time their purchases or sales of property accordingly to avoid suffering from huge losses, like what happened in 2008.

2 Data Understanding

The dataset used in our models contained 371 monthly observations from February 2, 1987 through December 1, 2017 and was extracted from datasets collected by the St. Louis Federal Reserve (FRED). We selected the direction of monthly returns for the Case Shiller

Home Price Index as our dependent variable. This index is computed using the changes in the prices of single-family, detached residences (houses) using a combination of the National Home Price Index, the Ten City Composite Index, the Twenty City Composite Index, and twenty individual metro area indexes. Not only is this index an indicator of how the housing market is performing, but it is also a good indicator of the health of the economy in general. When choosing variables for our data-set, we referred to past research done by Karl Case and Robert Shiller on predicting housing prices (Case and Shiller 1990).

The independent variables used encompass various macroeconomic indicators, housing market trends, and consumer preferences. The factors we selected were:

- Workers Over 55 Years of Age: shows that the aging population is still working and not moving into retirement homes at the same rate as they have historically

- Housing Starts: count of newly constructed homes in the US; an increase in demand for housing and increased posterity in housing market causes an increase in this value

- Monthly Supply of Houses in the US: ratio of houses for sale to houses sold; indication of the size of the for-sale inventory in relation to the number of houses currently being sold

- University of Michigan Consumer Sentiment Index: gauges consumer confidence in the market; helps to estimate trends in future consumer spending and saving

- Unemployment Rate in the US: when the unemployment rate is high, less people are employed and can afford to make large investments, such as purchasing a home

- Population Growth: as more children are born and the population grows, we would expect the demand for houses to increase accordingly

- 10 Year Treasury Yield: yield of 10 year bonds issued by the US government, which can be used as a general proxy for interest rates Median Household Income in U.S.: as median income median income rises, people are likely more willing to make larger purchases, such as a house

- Change in Consumer Price Index: used to assess price changes associated with the cost of living; indicator of inflation/deflation in the economy

- Gini Income Inequality Ratio: gauge of economic inequality by measuring income distribution and inequality within the United States

3 Data Preparation

In order to compute the direction of the monthly returns for the Case Shiller Home Price Index, we first calculated the median of the monthly returns, 0.0031417. To verify that existing data was not skewed, we computed the number of months where the index was above the threshold (186) and the number of months it was below (185). Since these counts were nearly equal, we were able to confirm that the median value was a valid threshold to use for determining the direction of our dependent variable. This method classified the observation as "up", a value of one, if the monthly return was greater than the median, and "down", a value of zero, if less than the median.

To train and test our data, we created two different splits of our data. The first split used all of the data to test/train the model. The training set contained the first 323 observations from 2/1/87 - 12/1/13 and the testing set contained the remaining 48 observations from 1/1/14 - 12/1/17. This split was the one ultimately used in the final models discussed below.

An alternative train-test split that we created was based on the value of the Case Shiller Index before and after the Subprime Mortgage Crisis. The pre-crisis portion contained

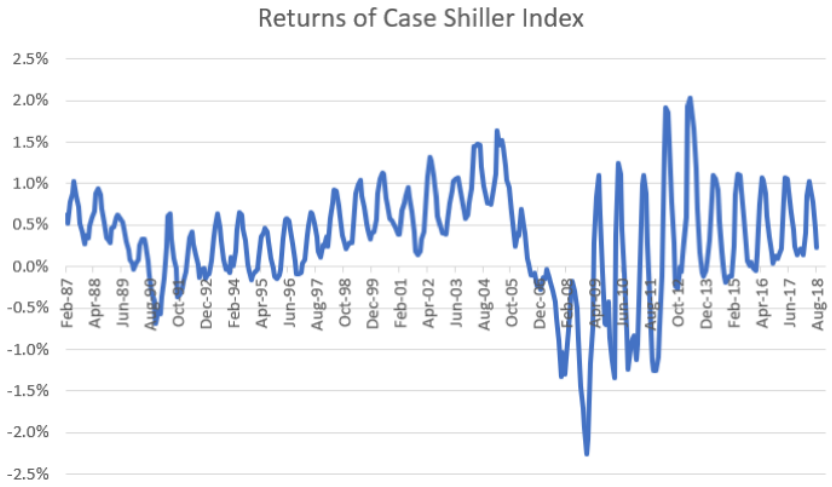


Figure 1: Returns of Case Shiller Home Price Index

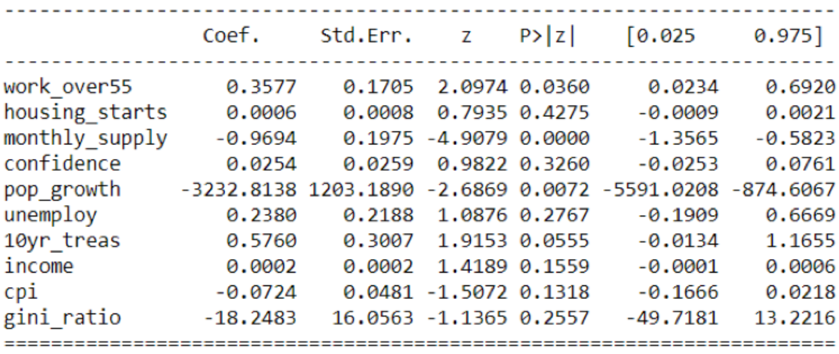
the observations from 2/1/87 - 7/1/06, where the final observation in this subset, July 2006, is the peak of the Case Shiller Home Price Index before the crisis. The post-crisis set contained all values after this peak through the end of our data-set (8/1/06 - 12/1/18).

4 Modeling

When creating our models, we created three models for each of the six different algorithms we chose to use. One of the models was trained on the pre-crisis data and tested on the post-crisis data, the next was trained on the post-crisis data and tested on the pre-crisis data, and the last was trained on data encompassing both pre-crisis and post-crisis data and tested on the last four years. We ran a logistic regression as a baseline model, then we tested LDA, QDA, KNN, SVM and Random Forest for comparison to find the optimal performers.

First, we ran a logistic regression on our largest data split, with the direction of housing price returns as our target variable and the other ten variables as the features (Figure 2).

From there, we used backwards step-wise selection to remove variables that were not significant, which narrowed our number of features down to six. The six features were



```
------------------------------------------------------------------------
                  Coef.    Std.Err.     z    P>|z|    [0.025     0.975]
------------------------------------------------------------------------
work_over55       0.3577    0.1705  2.0974 0.0360     0.0234    0.6920
housing_starts    0.0006    0.0008  0.7935 0.4275    -0.0009    0.0021
monthly_supply   -0.9694    0.1975 -4.9079 0.0000    -1.3565   -0.5823
confidence        0.0254    0.0259  0.9822 0.3260    -0.0253    0.0761
pop_growth    -3232.8138 1203.1890 -2.6869 0.0072 -5591.0208 -874.6067
unemploy          0.2380    0.2188  1.0876 0.2767    -0.1909    0.6669
10yr_treas        0.5760    0.3007  1.9153 0.0555    -0.0134    1.1655
income            0.0002    0.0002  1.4189 0.1559    -0.0001    0.0006
cpi              -0.0724    0.0481 -1.5072 0.1318    -0.1666    0.0218
gini_ratio      -18.2483   16.0563 -1.1365 0.2557   -49.7181   13.2216
========================================================================
```

Figure 2: Logistic Regression

workers over 55, monthly supply of houses, population growth, 10 year Treasury yield, median household income, and CPI inflation. These are the six variables that every model going forward utilized. After rerunning the logistic regression with these features, we received a test Mean Squared Error of 0.479 and an F1-score of 0.496. The results of the logistic regression are useful to give us a baseline to compare other algorithms to.

Next, we created a pre-crisis model and post-crisis model for each of the six algorithms. Given that neither of these data-sets fully covered both the rise and fall of the housing market from 2006-2009, we were not expecting the resulting models to produce very strong results. After running all of the algorithms, we found this hypothesis to mainly be true. Our best performing pre-crisis model was KNN with K=2 and the weighting method as distance. This gave us a test Mean Squared Error of 0.394 and an F1-score of 0.612 for predicting the post-crisis split. However, this model performed very poorly on our test set of he last four years, with a test Mean Squared Error of 0.542 and an F1-score of 0.288. The post-crisis models performed even worse, with random forest classifier being our best model using a max depth of 6 and number of trees as 3. This gave a test Mean Squared Error of 0.598 and an F1-score of 0.240 for predicting the pre-crisis data. Overall, the pre and post crisis models proved to be ineffective at predicting the future of the housing market.

After running utilizing various algorithms on our pre and post crisis splits, we moved on to the data split that would encompass the most amount of data into our data set, which included data from both before and after the housing crash. Although it was useful to test the algorithms with pre-crisis and post-crisis splits, we felt that the models incorporating data from both time periods would be the most effective at predicting the market. We started off by exploring how linear discriminant analysis and quadratic discriminant analysis performed, since they were some of the more basic algorithms learned in the beginning of the semester. The LDA model performed slightly better than random, yielding an F1-score of 0.5009 and a test Mean Squared Error of 0.458. The QDA model performed worse than the LDA model, giving an F1-score of 0.395 and a test Mean Squared Error of 0.563. We attributed the QDA model underperformance to it being too flexible and overfitting on our training data.

The next algorithm we examined on the whole data split was K-Nearest Neighbors for classification. We felt that this algorithm could be useful for our problem due to its classification based on similar data points. In order to find the optimal value for K to minimize test Mean Squared Error, we ran a loop to test for various possible values. In addition, we also tested two different weighting methods, uniform and distance, to determine which was optimal for our current problem. The results of this test gave an optimal value of K as 9 and a weighting method using distance. The resulting test Mean Squared Error was 0.417 and F1-score was 0.583. This was an improvement over the LDA model earlier, in all metrics.

From there, we decided to explore one of the more advanced algorithms we learned in the latter part of the semester, support vector machines for classification. We felt that this powerful algorithm would be able to separate the data nicely to yield quality prediction results on our largest data split. First, we ran various tests to determine which kernel fit best for our problem. After running various cross validation tests, we determined that radial basis function performed best on our data. From there, we had two main parameters to optimize in order to minimize test Mean Squared Error, which were the penalty parameter C and kernel coefficient gamma. We ran loops to test multiple values for C and gamma with K-fold cross validation to find the optimal pair of parameters. The results proved that C=0.1 and gamma=0.0000007 gave the optimal test accuracy. Once we fit the SVC model, however, the results were rather surprising. The test Mean Squared Error was 0.417 and the F1-score was 0.583, which are identical to the results from KNN.

We felt that the last algorithm we should run on the whole data-set was Random Forest for classification, which is a powerful, tree based ensemble method. Given the size and variety of our data, it was possible that a decision tree algorithm could perform very well. To ensure the best results, we once again ran a loop to test for optimal parameters for this model. The two parameters that we decided to focus on were the number of estimators, or trees used by the algorithm, and the maximum depth of the tree, to help avoid overfitting. After running the loops and cross-validation, we discovered that using four decision trees and a maximum depth of three provided the best results in terms of mean squared error (Figure 3).

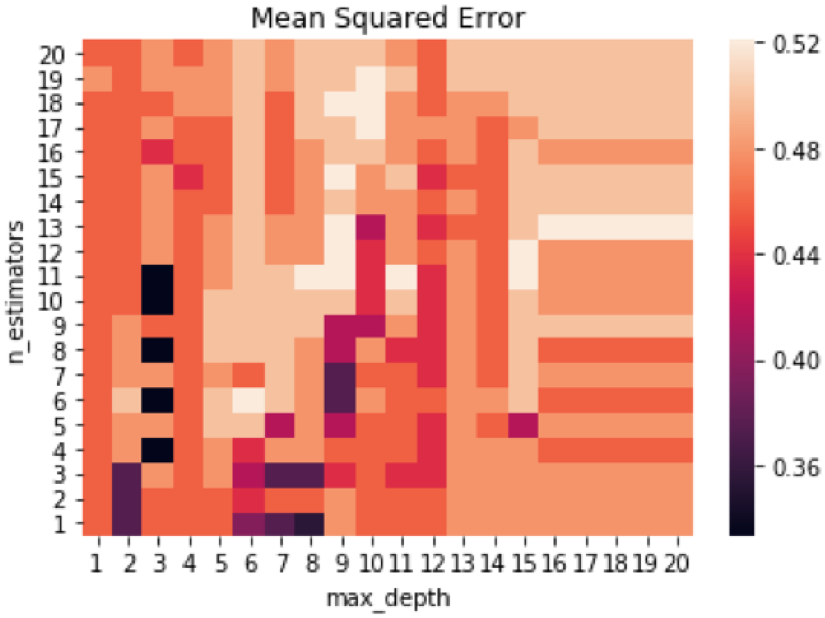
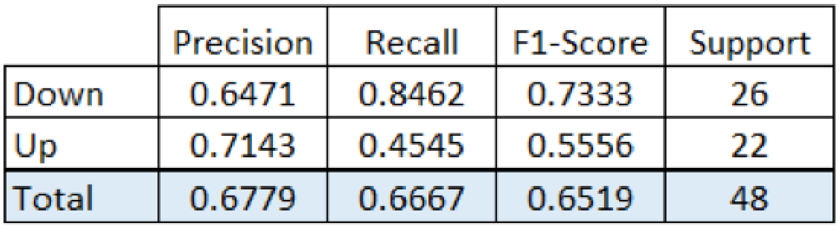


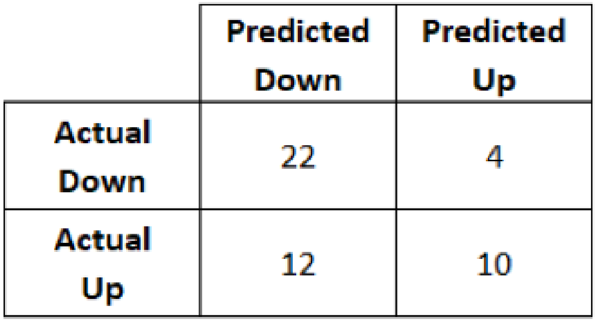
Figure 3: Heat map showing optimal mean squared error

With this model, we received a test Mean Squared Error of 0.333 and an F1-score of 0.652, which is our best performing algorithm. A visualization of the four decision trees used by the model can be found in the Appendix of this report. Below is the classification report (Figure 4) and confusion matrix (Figure 5) generated from deploying our model on the test set.



	Precision	Recall	F1-Score	Support
Down	0.6471	0.8462	0.7333	26
Up	0.7143	0.4545	0.5556	22
Total	0.6779	0.6667	0.6519	48

Figure 4: Classification Report



	Predicted Down	Predicted Up
Actual Down	22	4
Actual Up	12	10

Figure 5: Confusion Matrix

Based on the results from all of our models, across all data splits, the Random Forest Classifier was the one that produced the best results. Taking this into consideration, it is important to consider the pros and cons of deploying this model. One pro of this model is that it outperforms all of the other algorithms we tested. Another advantage is that by utilizing random forest, we were able to reduce both the variance and the bias of our results. One of the major cons of this model is that it correctly predicts the direction of housing price movements 65% of the time. Although this is better than a completely random, 50/50 approach, it still has a pretty large misclassification error.

5 Evaluation

For this project we used cross-validation to determine the parameters that produced the model with the smallest Mean Squared Error and F1 score. We split the data into two different categories, training and testing. Training is for the machine learning algorithm to learn the proper way to evaluate the classifier that we are looking at. We then took the results from the training and ran it on the test set to get a matrix of the accuracy on how well the training predicted the test set. As covered in the modeling section, Random Forest Classifier is the model that produces the most accurate prediction results. After utilizing loops and cross-validation to find the optimal parameters, which was 4 estimators with a max depth of 3, we received a test mean squared error of 0.333 and an f1-score of 0.652. This model was our best performing, which predicts the monthly direction of the housing market 65% of the time. We consider this result good, but not great, since it does outperform random by correctly predicting more than 50% of the time. However, we noticed that the model tends to correctly predict down movements more often than up movements, which we hope can be improved with further development of the model (Figure 6).

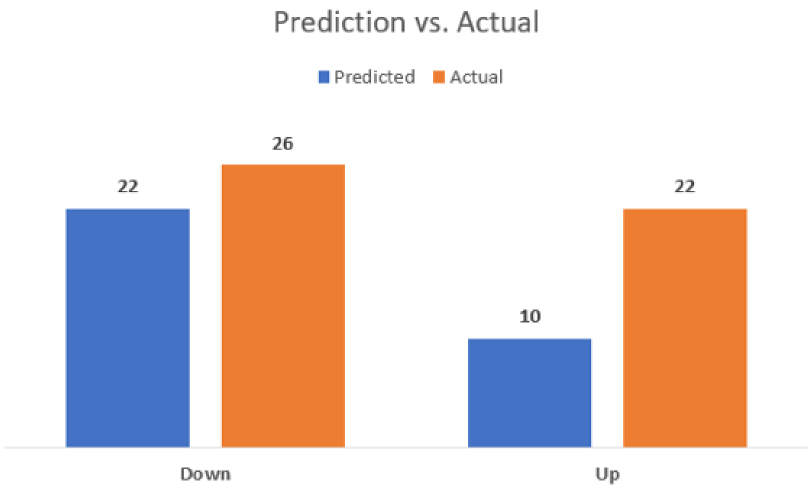


Figure 6: Predicted values vs actual values

A business case to project expected improvement could be fitting a model using regression techniques to predict the magnitude of the price movement, in tandem with predicting the direction. In addition, since our model tends to correctly predict up movements more often than down, finding additional variables that could help predict upward movements could also make a case for improving this project.

The ROI for this model has the potential to be high if we are able to improve the accuracy to a more confident degree. Given that property value of homes tends to be on a large scale, timing of purchases based on our model could result in tens of thousands of dollars in additional profits or savings. In addition, the cost of running this model is very low, since it does not require very much computing power. This means it is possible to run on a typical PC, making it possible for this model to be deployed anytime and anywhere. Once this model is deployed maintenance should be rather low. As new data is made available and we begin to see fundamental changes in the market, the cross-validation may need to be rerun to ensure that the optimal parameters have remained the same. Otherwise, the overall structure of the model would not need to be modified, and therefore the cost of maintenance is low, allowing for more profit in the long-run.

6 Deployment

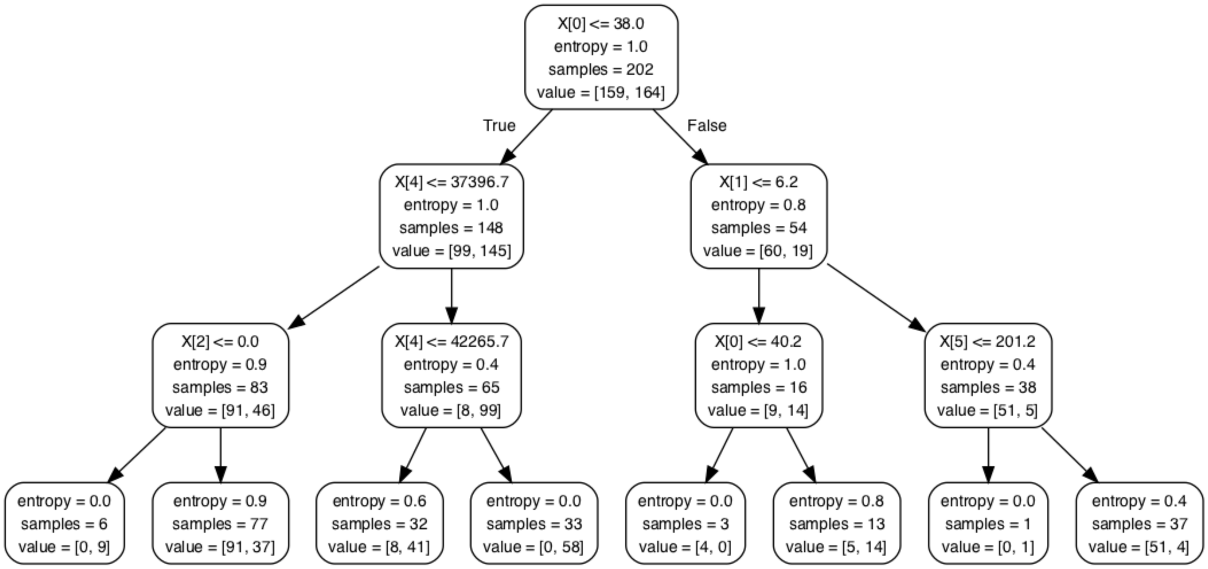
A model that predicts the directional changes of the housing market can be utilized by investors and homeowners looking to make transactions or buy/sell homes in the near future. By forecasting whether the the market will be heading up or down, they can better time their investments and sales. Although not guaranteed to be correct 100% of the time, our model increases the likelihood of making an accurate prediction to 65%, compared to a random, 50-50 approach, and thus could lower the risk involved with buying a home or making an investment.

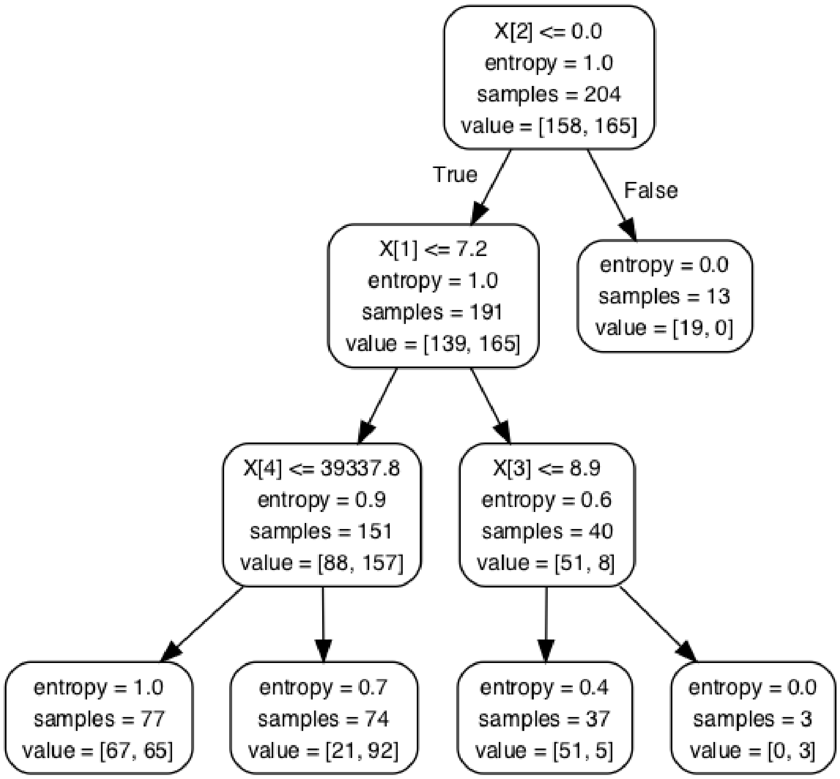
Usage of this model does come with associated risks. Some potential risks are formation of another housing bubble and changes in consumer preferences, which could cause a fundamental shift in the market. One current example of the change in consumer preferences is a growing focus on the community outside the property lines. These risks exist regardless of our model, but would hamper the effectiveness of its ability to predict fluctuations in the housing market.

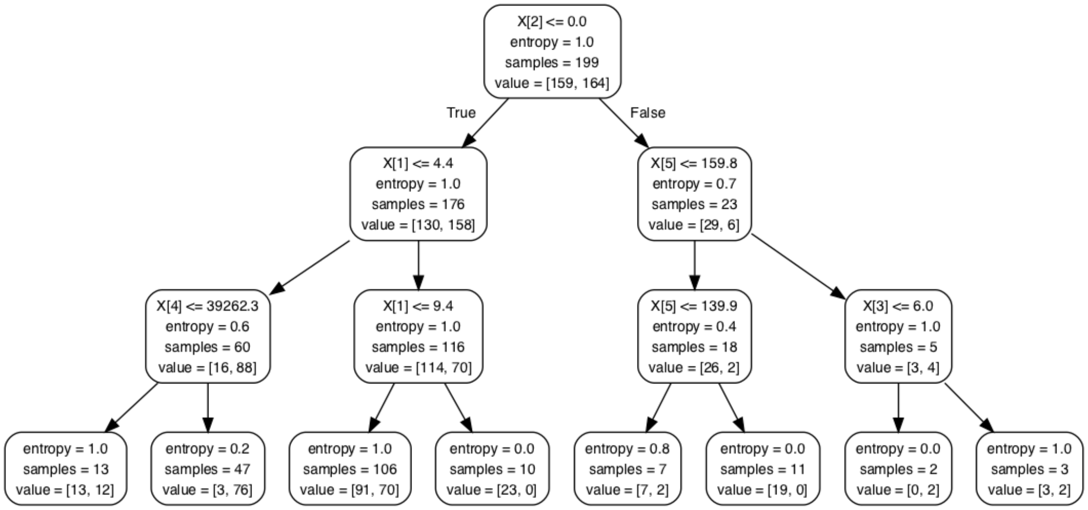
There are thousands of analysts predicting the next crash based on patterns, thousands of articles printed daily about where different companies see the market going, when they think the next crash will be. All of these things affect consumer preferences. Our model is meant to solely show the predicted directional change of the housing market. Leaving the analysis up to the user allows each user to create their own preferences to lower the risk of everyone making the same decisions and in turn causing a market shift or housing bubble.

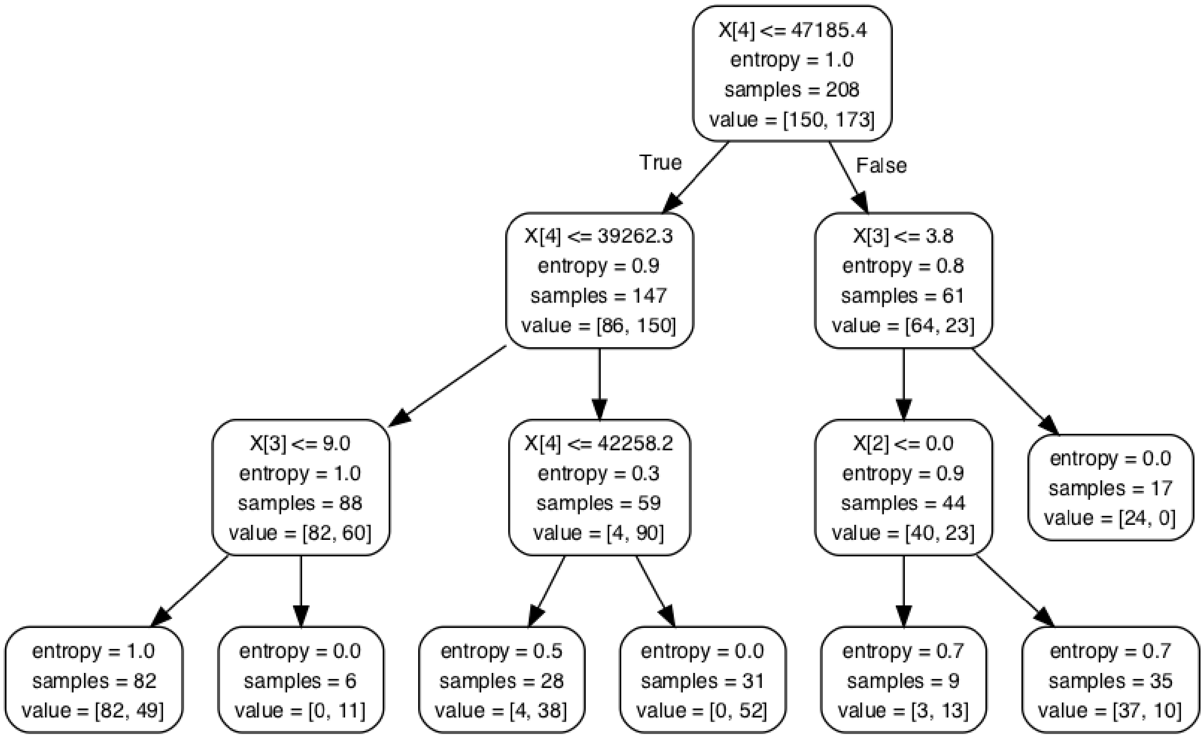
7 Appendix

Four Trees Created by Optimal Random Forest Model

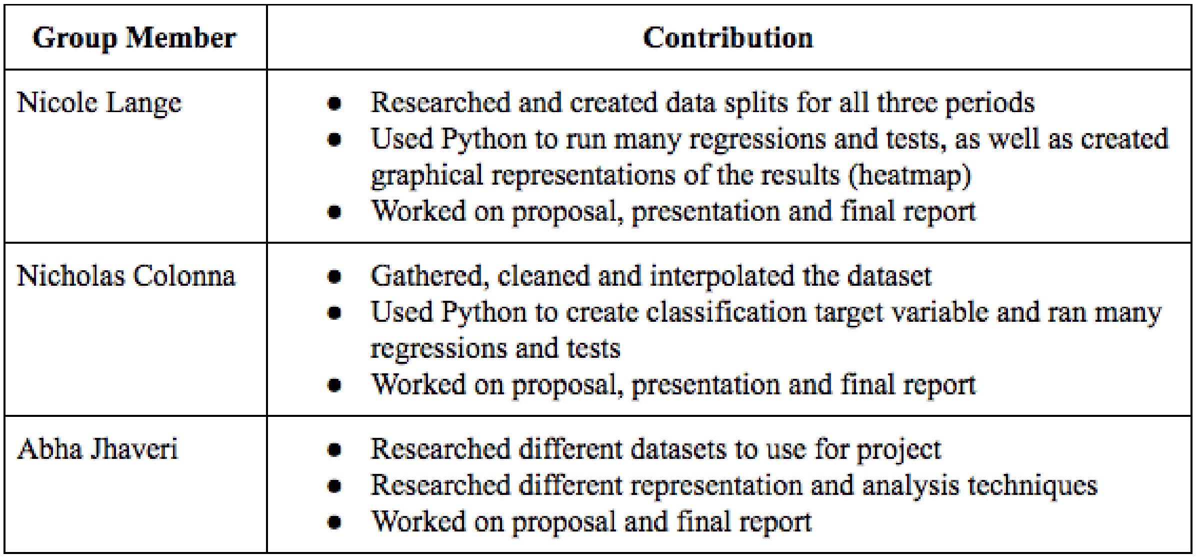



X[2] <= 0.0
entropy = 1.0
samples = 204
value = [158, 165]
True
False
X[1] <= 7.2
entropy = 1.0
samples = 191
value = [139, 165]
entropy = 0.0
samples = 13
value = [19, 0]
X[4] <= 39337.8
entropy = 0.9
samples = 151
value = [88, 157]
X[3] <= 8.9
entropy = 0.6
samples = 40
value = [51, 8]
entropy = 1.0
samples = 77
value = [67, 65]
entropy = 0.7
samples = 74
value = [21, 92]
entropy = 0.4
samples = 37
value = [51, 5]
entropy = 0.0
samples = 3
value = [0, 3]


X[2] <= 0.0
entropy = 1.0
samples = 199
value = [159, 164]
True
False
X[1] <= 4.4
entropy = 1.0
samples = 176
value = [130, 158]
X[5] <= 159.8
entropy = 0.7
samples = 23
value = [29, 6]
X[4] <= 39262.3
entropy = 0.6
samples = 60
value = [16, 88]
X[1] <= 9.4
entropy = 1.0
samples = 116
value = [114, 70]
X[5] <= 139.9
entropy = 0.4
samples = 18
value = [26, 2]
X[3] <= 6.0
entropy = 1.0
samples = 5
value = [3, 4]
entropy = 1.0
samples = 13
value = [13, 12]
entropy = 0.2
samples = 47
value = [3, 76]
entropy = 1.0
samples = 106
value = [91, 70]
entropy = 0.0
samples = 10
value = [23, 0]
entropy = 0.8
samples = 7
value = [7, 2]
entropy = 0.0
samples = 11
value = [19, 0]
entropy = 0.0
samples = 2
value = [0, 2]
entropy = 1.0
samples = 3
value = [3, 2]



Team Member Contributions



Group Member	Contribution
Nicole Lange	• Researched and created data splits for all three periods • Used Python to run many regressions and tests, as well as created graphical representations of the results (heatmap) • Worked on proposal, presentation and final report
Nicholas Colonna	• Gathered, cleaned and interpolated the dataset • Used Python to create classification target variable and ran many regressions and tests • Worked on proposal, presentation and final report
Abha Jhaveri	• Researched different datasets to use for project • Researched different representation and analysis techniques • Worked on proposal and final report

References

Case, K. E. and R. J. Shiller (1990). Forecasting prices and excess returns in the housing market. *Real Estate Economics 18*(3), 253–273.

Ramirez, K. (2017, Dec). Value of u.s. housing market climbs to record 31.8*trillion*.