Real Estate Investment Using SARIMA Modeling

David Richter, Flatiron School, June 30, 2022

Business Problem and Data

- Our business objective is to develop a method for determining the most profitable zip codes for real estate investment.
- To measure profitability, I'll use year over year Return on Investment, measured as:
  - (Current Median Home Value - Median Home Value 1 ya)/(Median Home Value 1 ya)
- The Data for these calculations comes from Zillow's database of median housing values for US zipcodes since 1996.

Methodology

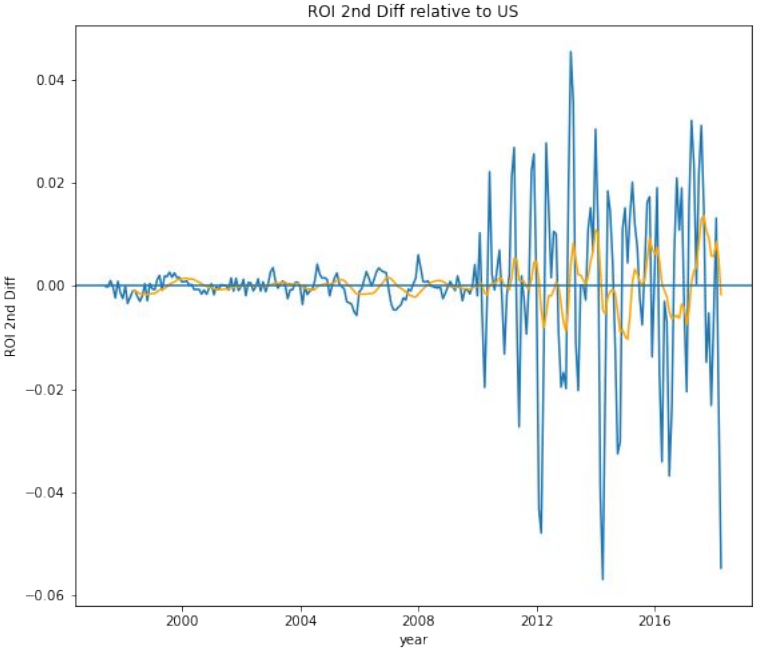
- Predictions will be made using ARIMA modeling
- ARIMA modeling is a form of regression in which **values in a time series are regressed against past values, differences, or noise of that same time series.**
- ARIMA isn't for modeling long term shifts in the real estate market, but for modeling **shorter term patterns.**
- This means that to the extent possible we want to eliminate the effects of long-term trends or exogenous factors from our data.

Detrending the Data: Subtracting US ROI

- The main trend we want to eliminate from our data is the overall trend of the US housing market, since this is not something that ARIMA modeling can help us with.
- This means that we'll be using ROI values from which we've subtracted the ROI for the US as a whole.

Detrending the Data: Limiting Our Training Set to Post-2010

- This sample second differencing from one zip code in the data set shows the much greater volatility of ROI values after 2010.
- This indicates that we want to train our SARIMA model on the post-2010 period, since we don't want to incorporate patterns from an earlier period into our model.



Selecting 20 Zip Codes for Modeling

Since the dataset contains over 10,000 zip codes, I began by looking at the top five zip codes according to four metrics:

- High average value over the past 12 months
- Highest average ROI over the past 12 months
- High average first difference (ROI increase)
- High average second difference (ROI acceleration)

Selecting a Universal Model

- Based on these twenty zip codes I tested 216 ARIMA models based on different combinations of features
- I trained each model on the 2010-2017 period for each zip code and calculated standard error based on predicted and true values for the 2017-2018 period.
- I then calculated the average standard error for each model across all zip codes.

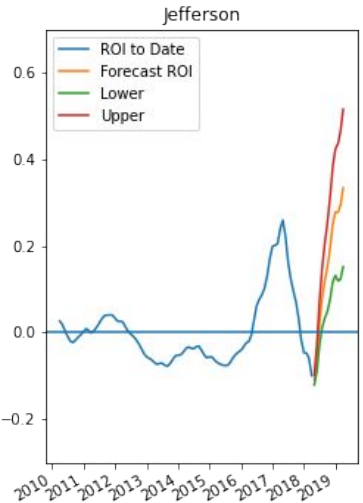
Model Features

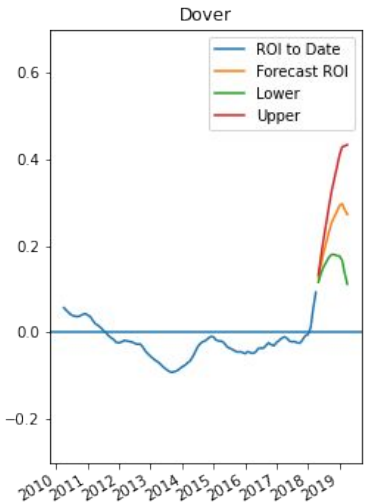
- The best performing model had non seasonal features (1, 1, 2) and 1 year seasonal features (1, 0, 1)
- This means that the model regresses on five features:
  - The most recent difference (change) in ROI
  - The two most recent differences (changes) in the model's noise
  - The previous year's ROI value
  - The previous year's noise
- This model had a standard error of 0.11, which is significant, since US ROI was 0.064 from 2018-2019

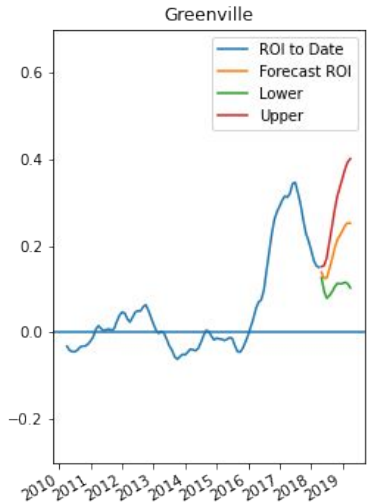
Using the Universal Model for Zipcode Selection

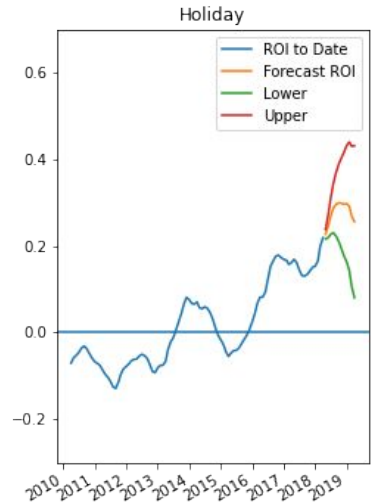
- I trained and generated forecasts on all of the zipcodes in the data set using this model.
- I then selected the top 5 zip codes by highest lower confidence interval (alpha=0.05) to investigate further.
- Standard error for these 5 zipcodes was 0.32, with most of the error coming from a single zipcode.

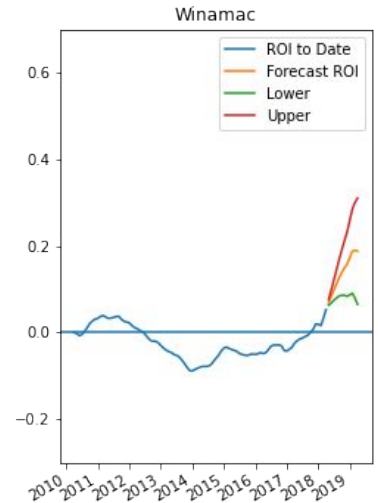
Forecasts for the Top Five Zip Codes



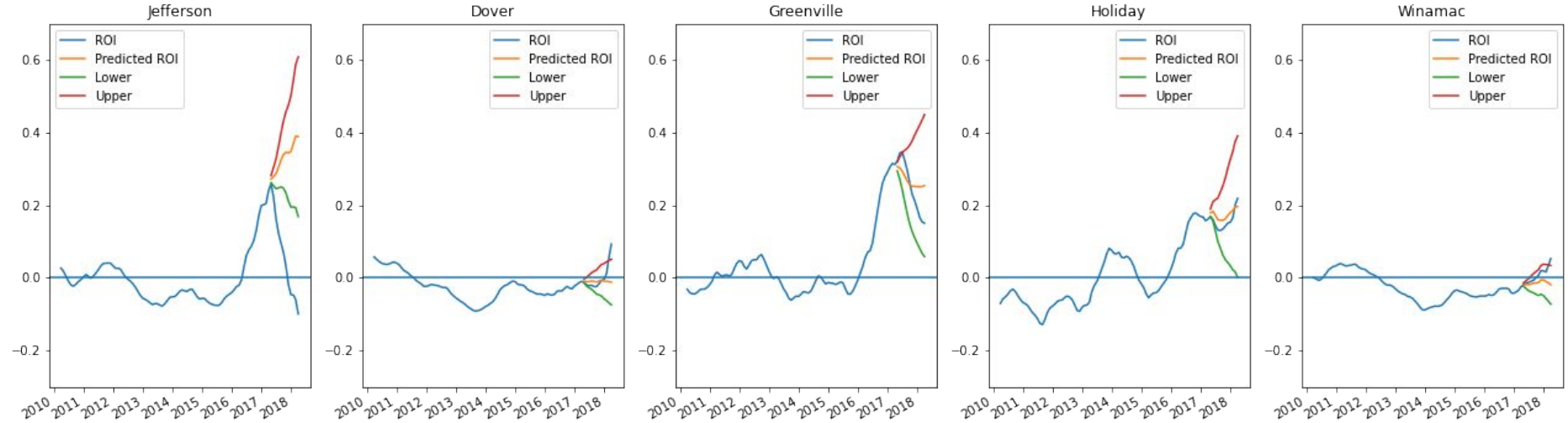








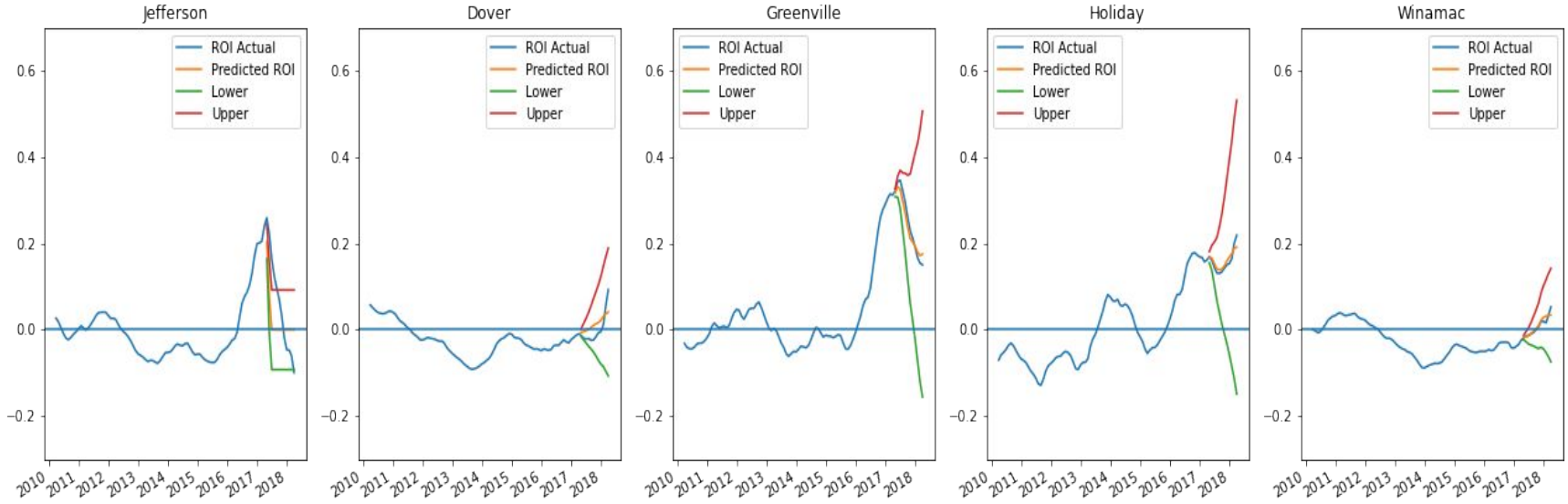
Performance of Model on Test Data



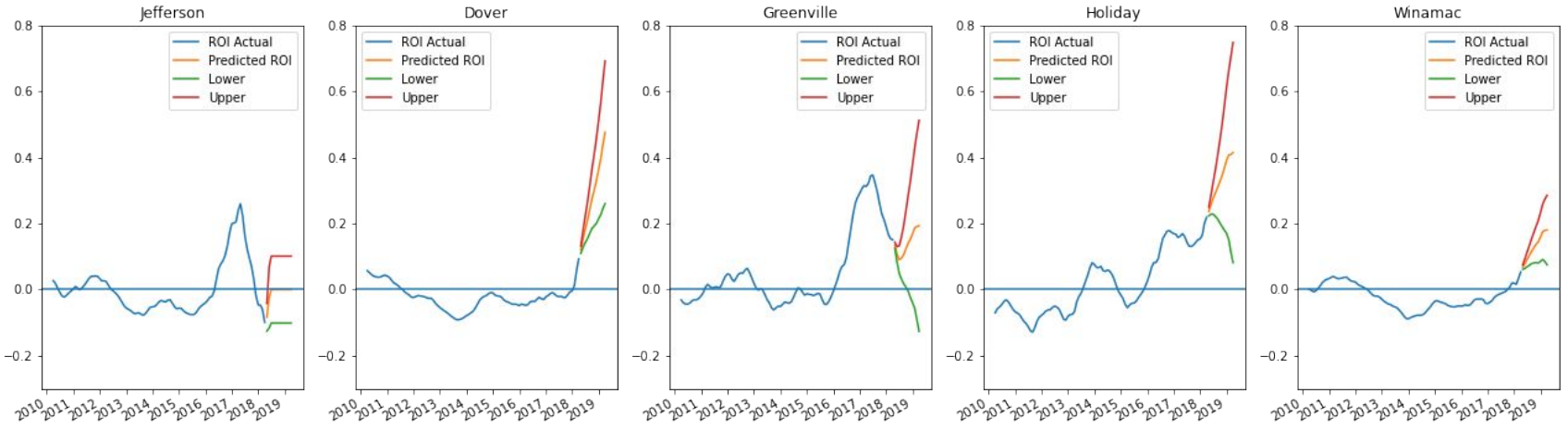
Custom Models for Each Zip Code?

- We should be suspicious about applying our ARIMA model to these zip codes since they were selected specifically for their high performance based on these zip codes.
- I therefore ran an individual grid search on each zip code to determine model features that would yield the smallest error on the test data.

Performance of Custom Models



Custom Models Forecasts



Methods for Forecasting

Which forecasting method is preferable?

- The ARIMA model validated on the twenty initial zipcodes?
- Or using the custom ARIMA models validated only a single zipcode?
- A third possibility is averaging the results of the two models and taking this average as a prediction.

Evaluating the Three Forecasting Methods

- Using the ROI forecasts for each of these models, I calculated median prices for each of the five zip codes for April 2019.
- I used the formula:
  - April 2018 Median Price * (1 + ROI + US ROI [2Q 2018-2Q 2019])
  - Where US ROI (2Q 2018-2Q 2019) = 0.064
- I then compared these to the actual median price for each zip code and calculated the standard error based on each method.

Results and Conclusion

- Universal Model: SE= $51,572
- Custom Models: SE = $74,866
- Average Model: SE = $55,294

- Based on these results, the best 1-year forecasting method is a (1, 1, 2)x(1, 0, 1, 12) ARIMA model fit on a zip code's data for post-2010 ROI values.

A Better Universal Model?

- The features for our ARIMA model were selected based on the model's performance for a specific set of zip codes–chosen for high earning potential.
- If we want a single model that applies to all zip codes, we might want to perform feature selection based on a randomly chosen group of zip codes.

Modeling by Category?

Another possible strategy is seeing if different models are optimal for different categories of zip code. Three intuitive ways of categorizing zip codes are:

- By Value - high value vs. low value
- By ROI variance - high variance zip codes vs. low variance zip codes
- By covariance of a zip code ROI and US ROI - high positive covariance vs. high negative covariance vs. low absolute covariance.

Thank you for viewing my presentation.

You can email any questions to d.richte@gmail.com.

And you can view the github for this project at https://github.com/DavidKRichter/dsc-real-estate-investment.