Real Estate Investment Using ARIMA Modeling

David Richter, Flatiron School, July 5, 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 home values for US zip codes between 1996 and 2018.

Methodology

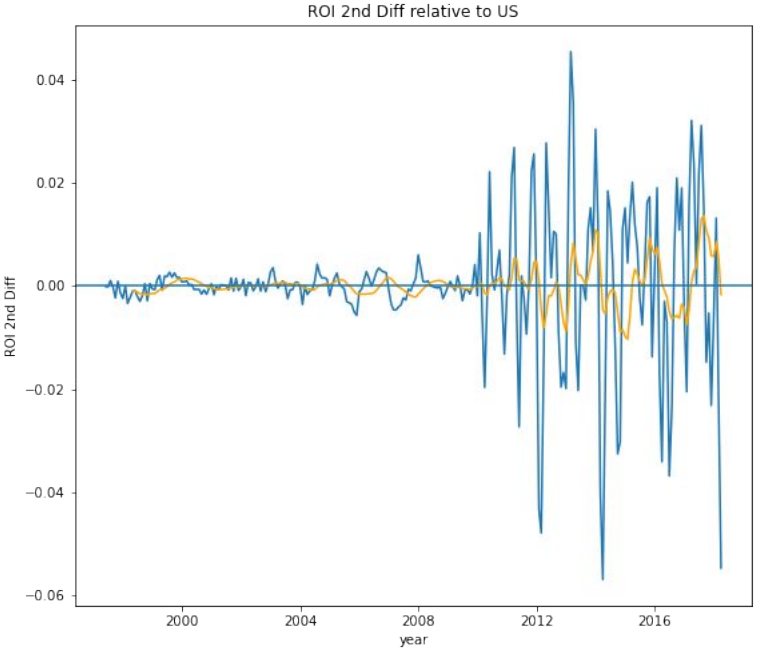
- Predictions will be made using ARIMA modeling
- ARIMA modeling is a form of regression in which **values in a time series are regressed against past values, differences, or noise of that same time series.**
- ARIMA isn't for modeling long term shifts in the real estate market, but for modeling **shorter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

Detrending the Data: Subtracting US ROI

- The main trend I wanted to eliminate from my data was the overall trend of the US housing market, since this is not something that ARIMA modeling can predict.
- This means that for a given month ROI values are calculated as:
  - Zip code ROI - US ROI

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 ARIMA 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.

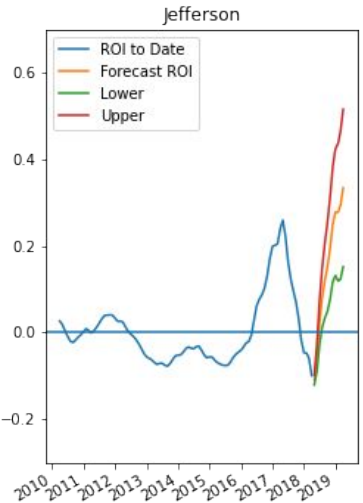
Model Features

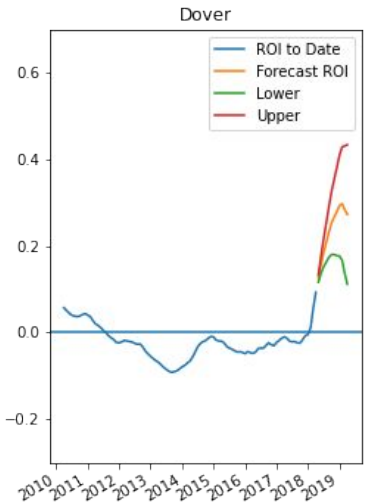
- The best performing model had non seasonal features (1, 1, 2) and 1 year seasonal features (1, 0, 1)
- This means that the model regresses on five features:
  - The most recent difference (change) in ROI
  - The two most recent differences (changes) in the model's noise
  - The previous year's ROI 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

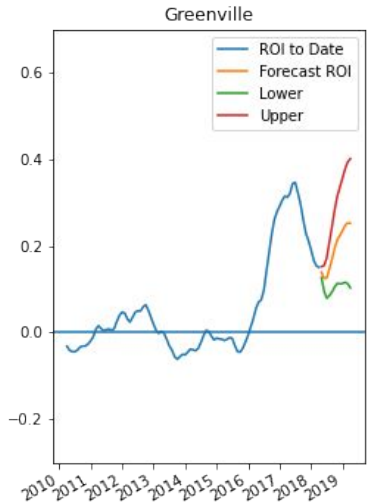
Using the Universal Model for Zipcode Selection

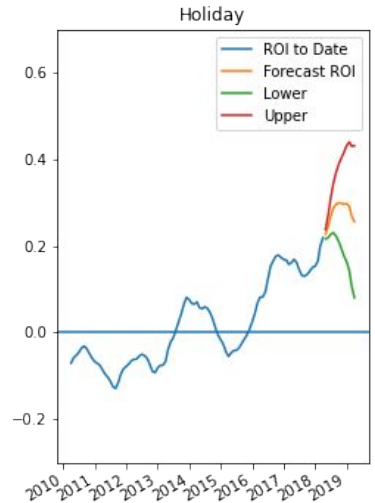
- I trained the universal ARIMA model on data from 2010-2018 from all zip codes and made forecasts for each zip code through April 2019.
- I then selected the top 5 zip codes by highest lower confidence interval (alpha=0.05) for April 2019 as the top zip codes for investment.

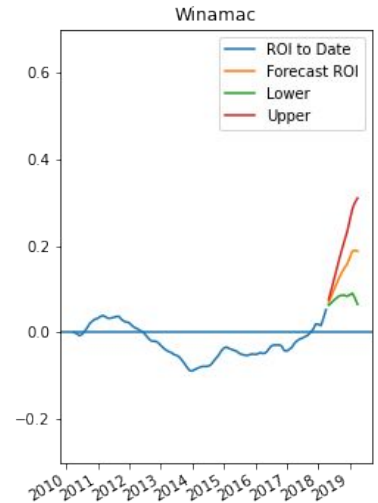
Forecasts for the Top Five Zip Codes











Forecasted ROIs for TOP 5 Zip Codes

	zipcode	predicted mean	lower interval
6715	70121	0.333956	0.152095
8816	72837	0.273015	0.112034
6514	29601	0.252431	0.103501
3936	34691	0.256219	0.080849
9190	46996	0.188035	0.065623

Forecasting Prices

Based on the price information, our model's ROI forecasts, and US ROI (0.064% from 2Q 2018- 2Q 2019), we get the following forecasted median home prices for each of our five zip codes. In all cases this is below the US median of 358K for 2019.

RegionName	City	State	Median Price 2018-04	Forecast 2019-04	Lower Interval 2019-04
29601	Greenville	SC	261900	344773.0	305769.0
34691	Holiday	FL	104800	138359.0	119980.0
46996	Winamac	IN	90800	113685.0	102570.0
70121	Jefferson	LA	171100	239190.0	208074.0
72837	Dover	AR	112000	149746.0	131716.0

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.