Predicting the Case-Shiller Index for Twenty American Cities

Project Category: Finance and Commerce

Name: Mattheus Wolff
SUNet ID: mbwolff
Department of Computer Science
Stanford University
mbwolff@stanford.edu

Name: Ari Webb
SUNet ID:
Department of Computer Science
Stanford University
arijwebb@stanford.edu

Name: Laywood Fayne
SUNet ID: lfayne
Department of Computer Science
Stanford University
lfayne@stanford.edu

This template is built on the NeurIPS 2019 template[1] and provided for your convenience. Your proposal should include a 300-500 word description of what you plan to do. Presenting pointers to one relevant dataset and one example of prior research on the topic are a valuable (optional) addition.

1 Key Information to include

- External collaborators or mentors (if you have any): N/A
- Sharing project with another class: N/A

2 Motivation

We are addressing the challenge of predicting the best locations to invest in real estate. This is an application problem because we plan to use historical data for both our training and testing sets and could theoretically use our results to inform any current real estate investment decisions.

3 Methodology

We are planning to apply linear regression and Unsupervised Learning in our project to make a prediction model. We think that it would be interesting to see what features and modifications of them help our prediction(s) the most.

4 Intended experiments

We plan to train our model on per-month per-city data on various macroeconomic factors including, but not limited to: unemployment rate, crime rate, CPI, GDP by industry and population movement. If possible, we would like to get more specific with per-county or per-zip code data, but we expect that this information will be more difficult to find.

We are planning to test our model on per-month city-data for the latest five years of data, attempting to predict the Case-Shiller Index for each city five years in the future (we may also use other house price indices as real estate investment measures if the data is more complete). These overall

[1] https://www.overleaf.com/latex/templates/neurips-2019/tprktwxmqmgk

predictions can be made by projecting the next month's performance, adding it to the historical dataset, and then repeating. An interesting challenge is how we can/should update the historical factors as we predict further into the "future". We could also train/test our model on different time periods to see how this might affect any predictions. For example, it might be interesting to see how the model performs around the 2008 time period (e.g. trying to predict what happened using data from 2007 and earlier). We could also use combinations of different factors or look at some weighted variance analysis to see which ones have the most impact on our prediction.

We can evaluate our algorithm by simply comparing our predictions to historical data (since we are essentially predicting the past). We can measure error at every predictive time step (monthly), or have a final measure at the end of our five year testing period.