Machine Learning Engineer Nanodegree
Capstone Proposal

Xiao Wei

Domain Background

Real-estate in the United States is estimated to be worth $30 trillion dollars at the end of 2016, which is more than the combined market capitalization of every publicly listed company in United States[1]. In addition, much of middle class households in United States (and around the world) have their net worths tied closely to the real estate market. This means that fluctuations in the price of real-estate affects many more people than fluctuations in the securities market (which are mostly owned by wealthy individuals).

The price of a house is made up of the attributes of the house and the surrounding areas. There is a rich history in academia of predicting housing prices. Recently, tech firms have been moving into this space and accurate prediction of housing prices is something that would disrupt the real estate broker industry that is plagued by inefficiencies and misaligned incentives (as detailed in Freakonomics). To help guide my prediction algorithm, I will be using an academic paper that predicted housing prices in Singapore[2].

Problem Statement

At the writing of this proposal there is a competition on Kaggle that is sponsored by Zillow that seeks to predict housing prices in the Los Angeles area. The competition provides the training data and the metric measurement criteria. Zillow predicts prices for almost every listing that they have on their website. Any model that is created can be trained on more data or used to predict more test cases.

Datasets and Inputs

The training data is provided by Zillow includes features for the houses. The actual log errors (y-label) for the training data is also provided alongside a data dictionary. In addition to the provided data, I also plan to use the S&P Case Shiller Index for the greater Los Angeles Area[3].

---

[1] https://www.zillow.com/research/2016-total-home-value-rents-14028/
[2] Jiang, Liang and Phillips, Peter C. B. and Yu, Jun, A New Hedonic Regression for Real Estate Prices Applied to the Singapore Residential Market (December 2, 2014). Cowles Foundation Discussion Paper No. 1969. Available at SSRN: https://ssrn.com/abstract=2533017 or http://dx.doi.org/10.2139/ssrn.2533017
[3] https://fred.stlouisfed.org/series/LXXRSA

Another source of data would be Chinese foreign exchange reserves, since Chinese buyers have been snatching up American real estate, especially in Los Angeles[4].

- Properties_2016.csv - provided by Zillow. Contains features of properties in greater Los Angeles area
- Train_2016.csv - provided by Zillow. The y-labels for the properties
- LXXRSA.csv -S&P Shiller Case price index from the internet
- Chinese foreign reserves - instrument variable for Chinese buyers from the internet

Solution Statement

I plan to use a generalized linear model. I will most likely use elasticnet model of sklearn library since there will be many features that may be correlated and the elasticnet should help filter them out (or shrink them to 0). Regression is the most used type of model for prediction of continuous response variable. Predictions will be compared to actual using the mean absolute error metric.

Benchmark Model

The benchmark model will be taking the average of the response variable as the predictor for all observations. This is equivalent to a regression model with only the bias term. In regression, such a model should have a coefficient of determination value of 0. The R^2 value will serve as a comparison between the benchmark model and the actual model.

Evaluation Metrics

Zillow provides the attributes for all the observations for the training set as well as what the evaluation metric is going to be. The evaluation metric is pretty weird (difference between mean absolute error of the predicted log error and actual log error (actual log error is the log error of what Zillow predicted and the actual selling price). This means that the goal isn't to predict housing sales but rather predict what is inaccurate about Zillow's prediction algorithm.

Project Design

First step to any model is to conduct exploratory data analysis on the data. In this case the Properties_2016.csv. Any features with too many missing values will need to be discarded. Important features with some missing values will need to have those values imputed. One important step for regression models is to account for heteroscedasticity. Heteroscedasticity is more prevalent than the homoscedasticity assumption that linear regression models make. Certain features will need to be transformed (most likely log-transformation).

---

[4]http://www.safe.gov.cn/wps/portal/!ut/p/c5/04_SB8K8xLLM9MSSzPy8xBz9CP0os3gPZxdnX293QwP30F
AnA8_AEBc3C1NjIxMjA6B8JE55dzMDArrDQfbhVhFsjFcebD5I3gAHcDTQ9_PIz03VL8iNMMgMSFcEAP
5jfwo!/dl3/d3/L2dJQSEvUUt3QS9ZQnZ3LzZfSENEQ01LRzEwT085RTBJNkE1U1NDRzNMTDQ!/?WCM_
GLOBAL_CONTEXT=/wps/wcm/connect/safe_web_store/state+administration+of+foreign+exchange/dat
a+and+statistics/forex+reserves/foreign+exchange+reserves/bb4353804c420bf6aa0aaefd3fd7c3dc

Some of the data (such as chinese foreign reserves will need to be interpolated for months in 2017 of which there is no data). The foreign reserve and S&P Shiller Case Price index will also need to be manually merged into the Properties_2016 data by date.

It is possible that PCA will be done as part of the EDA if there are features that could be combined so that number of dimensions can shrink before the model.

After preprocessing is finished I will run elasticnet regression alongside maybe Huber regression and use cross-validation to select the best hyperparameters. The winner of elasticnet vs Huber will be used as part of the final model. I will then use a neural network. Tuning the neural network parameters will require much time since I don't have much experience with the method. The two models' outputs will be averaged to produce the final prediction. I may also weight the model predictions depending on how they do in cross validation.