Machine Learning Engineer Nanodegree

Capstone Proposal - Zillow House Price Predictions

Ann Chong

3 July 2017

Project Overview

The aim of this project is to build and compare the performance/predictive power of a linear model versus a neural network. In particular, this project aims to replicate the `hedonic pricing approach`[1] to house price predictions and to compare this to the predictive power of a neural network using the Zillow[2] dataset.

This project is inspired by the NZARES Paper[3] and Kaggle's Zillow[2] competition.

> [1] *The hedonic pricing approach is one which identifies price factors according to the premise that price is determined both by the internal characteristics of the good being sold and external factors affecting it.* Source: http://www.investopedia.com/terms/h/hedonicpricing.asp *It breaks down the item being valued into its constituent characteristics, and obtains estimates of the contributory value of each characteristic.*
>
> *The approach measures the relative importance – through use of regression analyses – of independent 'explanatory' variables on property prices. If, for example, through regression analyses increased distance from an open cast mining site is found to be correlated with increased house prices, it can be ascertained that the open cast site has a negative impact on house prices. The regression analysis can also be used to provide a value for the size of the relative impact. It may be found that a 1km movement away from the open cast site equates to an increase of £5,000 on a house price.* Source: http://www.cbabuilder.co.uk/Quant5.html
>
> *Under this approach, the change in a house price resulting from the marginal change in one of these characteristics is called the hedonic price (sometimes referred to as the implicit price or rent differential) and can be interpreted as the additional cost of purchasing a house that is marginally 'better' in terms of a particular characteristic.*

[2] https://www.kaggle.com/c/zillow-prize-1 (https://www.kaggle.com/c/zillow-prize-1)
[3] http://ageconsearch.umn.edu/bitstream/97781/2/2004-9-house%20price%20prediction.pdf (http://ageconsearch.umn.edu/bitstream/97781/2/2004-9-house%20price%20prediction.pdf)

Domain Background

Why estimate house prices?

Individuals, corporations and governments alike are interested in house prices.

For many individuals, buying a house constitutes the largest financial transaction they will ever make. This, together with the obvious financial gain that it is possible to achieve from accurately predicting house prices, makes house price predictions important to many stakeholders, including the following:

- prospective homeowners
- property developers
- investors
- government agencies
- other real estate market participants such as mortgage lenders and insurers.

Whilst individuals and corporations may be interested in house prices as a direct input into their investment decisions and financial planning, government agencies have an interest as the housing market is often thought to be an important indicator of social sentiment and how the economy of a country is doing.

Whether implicitly or explicitly, these stakeholders employ predictive models in their investment/decision making processes, albeit with varying degrees of sophistication, conscious realisation and conscious application. Nonetheless, it is difficult to imagine that any of these stakeholders would argue against more accurate predictive models.

The interest in accurate house price predictions can be seen in the number of competitions on the subject, including the following:

- Kaggle's Zillow competition - predicting the log error between Zillow's house price estimate and the actual sale price
- Kaggle's Sberbank competition - predicting the sale price of property in Russia

Common method for estimating house prices

Property prices are commonly estimated using the hedonic pricing method: the price of a property is determined by the characteristics of the property (size, appearance, features, condition) as well as the characteristics of the surrounding neighborhood (accessibility to schools and shopping, level of water and air pollution, value of other homes, etc.) The hedonic pricing model is used to estimate the extent to which each factor affects the price.[Source: http://www.investopedia.com/terms/h/hedonicpricing.asp#ixzz4kj51Z87i]

The price of a property is then the sum of the base price plus the contribution from each property characteristic.

Advantages of the hedonic pricing approach

The hedonic pricing approach uses statistically familiar multi-variate methods which makes it possible to test whether a proposed explanatory variable is statistically significant and to estimate the covariance between these variables. It also allows for the calculation of confidence intervals for the marginal contribution of each explanatory variable to property prices.

This results in some intuitively appealing output such as people's marginal willingness to pay for a specific characteristic (e.g. extra bathroom) which can also be expressed as the changes in property values for a unit change in each characteristic.

An alternative to the hedonic pricing approach?

The literature to date suggests the hedonic pricing approach to be the most popular in the valuation of property prices.

The results from the NZARES Paper, however, suggests an alternative in the form of a neural network. More importantly, the paper suggests that hedonic pricing models do not outperform neural network models, whilst also acknowledging that other papers have reached a different conclusion and commented on the black box nature of neural networks.

It is the apparent ambiguity as to the performance of these two types of models on house price predictions that is the area of interest and further investigation of this project.

Problem Statement - house price predictions

The variable we are trying to predict is the price of a house at a point in time. The potential solution to this is an algorithm that can make predictions about the sale price of a home at a point in time.

The accuracy of a prediction is classified as: `|predicted price - actual price|` i.e. the absolute difference between predicted and actual sale price. The smaller this number is, the more accurate the prediction is.

In assessing the predictive quality of the two sets of algorithms, the R^2 metric and residual plots is proposed:

> R^2 = explained variance/total variance. This is the proportion of the variation in the target variable that is explained by the linear model, and takes values between 0% and 100%. The closer the R^2 value is to 100%, the better the model explains the variability of the target variable around its mean.
>
> Residual plots are used as the R^2 variable on its own is not able to determine if a model is biased.[Ref:http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit]

Datasets and Inputs

The Zillow dataset provided by Kaggle is comprised of the following:

1. Training data for transactions predominantly pre 15 October, 2016.
2. Test data which will be used to determine the public leaderboard is comprised of transactions made between 15 Oct - 31 Dec 2016.
3. The test data which will be used for determining the private leaderboard test is comprised of all transactions made between 15 Oct - 31 Oct 2017. This is for information purpose only and does not affect this capstone project.

For the purpose of this project, the following data approach is suggested:

- i. Assume Kaggle's training data comprises the project's entire dataset.
- ii. Split the project's dataset (i.e. the Kaggle training data from i. above) into a training set and test set.
- iii. the project's success should be measured based on performance on the test set from part ii. above.

Solution Statement

The proposed solution is the following:

1. build the best hedonic i.e. regression model to use to predict future house price
2. build the best neural network model to predict future house price
3. compare the predictions from the two models on test dataset to determine which model type has better predictive capability on the test data

As previously mentioned, in assessing the predictive quality of the two sets of algorithms, the R^2 metric and residual plots is proposed.

Benchmark Model

In this instance, the benchmark model is one which reflects a hedonic pricing approach. As previously described, the regression based model would serve as the benchmark model in this instance.

Evaluation Metrics

As previously mentioned, the evaluation metrics proposed are:

- R^2 = explained variance/total variance. This is the proportion of the variation in the target variable that is explained by the linear model, and takes values between 0% and 100%. The closer the R^2 value is to 100%, the better the model explains the variability of the target variable around its mean.
- Residual plots: these are proposed as the R^2 variable on its own is not able to determine if a model is biased.

Project Design

The types of regression models to consider include:

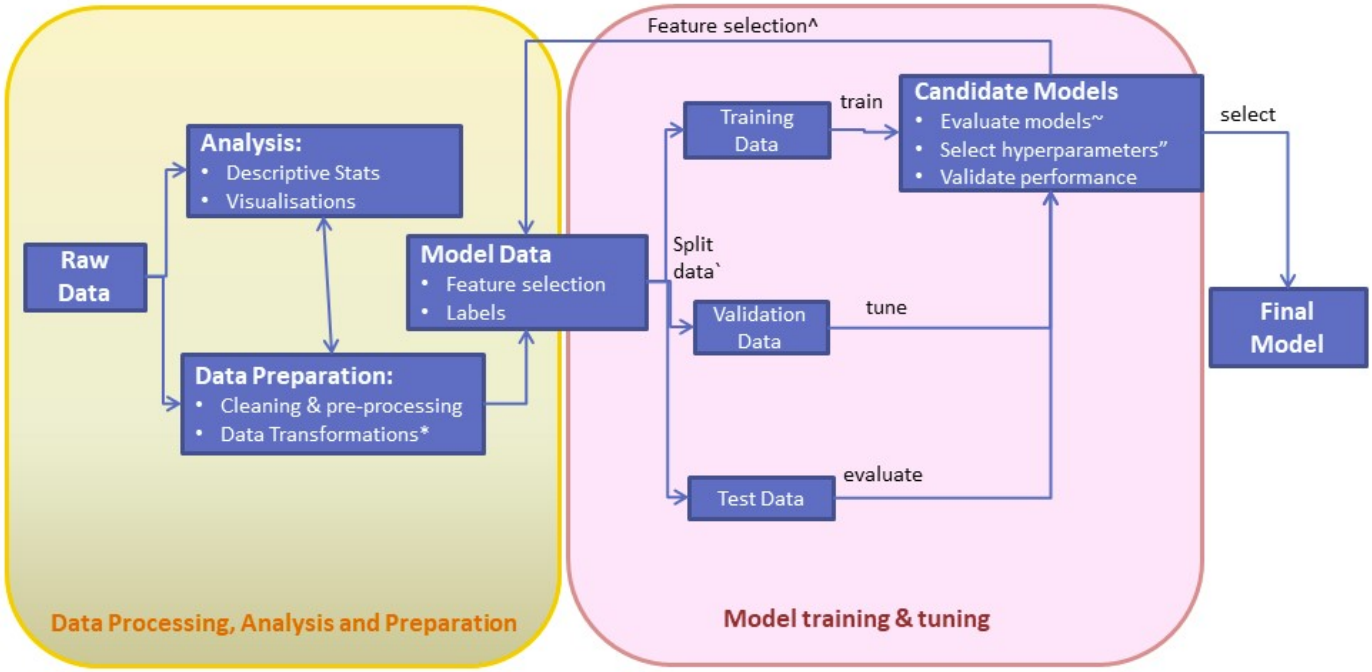
```
- linear regression models eg lasso, ridge
```

The items to consider in relation to the neural networks include:

```
- the number of hidden layers
```

The following schematic sets out the proposed workflow for each of the regression based model and neural network model.

Proposed Workflow



*:transformations to better align the structure of prediction problems to modelling algorithms as different algorithms make different assumptions about data.
`:data can be split a variety of ways, including k-fold cross validation and repeated random train-test split
~:evaluation metrics include mean absolute error, mean squared error and R-squared.
^:features can be selected using Recursive Feature Elimination or PCA (a data reduction technique that can aid in feature selection)
":hyperparameters may be selected using grid search approaches or random search approach

References:

The following documents/links have been referenced in writing this proposal:

[1] https://www.diva-portal.org/smash/get/diva2:131529/FULLTEXT01.pdf (https://www.diva-portal.org/smash/get/diva2:131529/FULLTEXT01.pdf)
[2] http://ageconsearch.umn.edu/bitstream/97781/2/2004-9-house%20price%20prediction.pdf (http://ageconsearch.umn.edu/bitstream/97781/2/2004-9-house%20price%20prediction.pdf)
[3] http://www.doc.ic.ac.uk/~mpd37/theses/2015_beng_aaron-ng.pdf (http://www.doc.ic.ac.uk/~mpd37/theses/2015_beng_aaron-ng.pdf)
[4] http://www.ecosystemvaluation.org/hedonic_pricing.htm (http://www.ecosystemvaluation.org/hedonic_pricing.htm)
[5] http://oppla.eu/sites/default/files/uploads/methodfactsheethedonic-property-pricing-method_0.pdf (http://oppla.eu/sites/default/files/uploads/methodfactsheethedonic-property-pricing-method_0.pdf)
[6] https://en.wikipedia.org/wiki/Hedonic_regression (https://en.wikipedia.org/wiki/Hedonic_regression)
[7] http://www.cbabuilder.co.uk/Quant5.html (http://www.cbabuilder.co.uk/Quant5.html)
[8] http://www.investopedia.com/terms/h/hedonicpricing.asp (http://www.investopedia.com/terms/h/hedonicpricing.asp)
[9] http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit (http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit)