Texas State University
College of Science and Engineering
San Marcos, Texas

Final Project Proposal for Introduction to Machine Learning CS4347

ACREAGE/SQUARE FOOTAGE REGRESSION ANALYSIS BY ZIP CODE

Beckett Allen
ra1205@txstate.edu

Kent Kenyon
kck46@txstate.edu

Supervisor(s): Supervisor Dr. Yan Yan
Texas State University, San Marcos, Texas

10/10/2020

1. Description

In this proposal, we are aiming to describe some of the high level design for our CS4347 Machine Learning Project. For our project, we will observe the relationship between our target parameter, the value of a given home in each ZIP Code,from the given input parameters that we collect data on. Our project will focus on creating the regression algorithm needed to calculate house valuations in a ZIP Code geographical area. The project is aimed at possibly using two different regressions in order to determine the cost of a house in a given ZIP Code. One regression will take in the parameters of the house that will calculate the houses value based on the number of bedrooms, number of bathrooms, square footage, if it has a pool, if it has a garage, etc. From this regression, we will get the price of a house very generally, and then apply this price to a second regression for determining the price of this house in constraints to a specific ZIP Code. The ZIP Code will take the houses market value, but adjust this value based on the population of the ZIP Code, the number of schools in the ZIP Code, the square acreage of the ZIP Code, the average income of the ZIP Code, or any other parameters that we can find. Our analysis and application will feature all the typical parameters that are used in housing cost analysis, yet will also be more centered toward understanding how ZIP Codes can effect these trends and allow the user to have a better understanding of housing prices as a result of varying ZIP Codes.

2. Existing Research

There are many existing projects on estimating housing values based on parameters like the number of bathrooms, number of bedrooms, the square footage, etc. We believe that these projects will be beneficial as a starting point for our project in understanding machine learning better, yet we plan to take this idea farther by understanding the relation of ZIP Codes and the different parameters that can be addressed when ZIP Codes are taken into consideration. In one article researched [1], the author describes how he went through the process of using a regression algorithm for his houses where his parameters described the characteristics of a home. This article only took into account the specifics of the home, while we plan to branch farther into community factors that will effect a homes value. "Hedonic" factors are included another article [2], these features point out the features of a home, as parameters, that would bring the most satisfaction to the consumer, yet still doesn't appropriately apply all the outside factors from the community that may be involved in the home appraisal. The last article that we analyzed for background information [3] where the author tried to generate values close to the X-Value seen in the Singapore Real Estate Exchange(SRX). In this machine learning property valuation, the author used many housing characteristic features, yet did not include the extent of parameters outside of a home that we are looking to analyze.

3. Differences in Proposed Work

Our Proposed research aims at taking in many more factors that influence a houses value aside from the physical aspects of the house itself. We will be working with the main features of house, including all the amenities that a house can have, yet look to aim at a broader image of the local factors that can effect a houses value. These factors would include parameters like how big the population of the ZIP Code is in relation to how many acres are in that ZIP Code, the number of schools within the ZIP Code, average revenue for businesses within that ZIP Code, the government, average income per ZIP Code, and any other parameters that we have access to and deem important.

4. Preliminary Plan

1. We will continually do research on what is already published in terms of machine learning for housing appraisal and keep in mind that this is a topic many people have researched and focus on the specific aspects of our project that make it unique. This will help to further establish our problem statement and start thinking into the technical requirements for this

project. This research will also include research into the different algorithms we could use and how we could create an algorithm that would best fit our data.

2. We plan to start our research off with finding as much data as we can find to work with through websites like Kaggle, drivendata.org, kdnuggets, dataquest, and many more to find as much usable data as possible for our project. We are aiming to find diverse, abundant, and unbiased data to best use for our project. Tools for collecting this data may vary depending on where the data is coming from.

3. Once we have found some ideal data to work with, we will run through this data and clean the data to get rid of unwanted aspects of the data, bad data, or outliers. We will also need to make sure that the data is formatted appropriately and is normalized. From this data, we can appropriately define our parameters.

4. We will need to compare different algorithms to see what will work best with the data defined in our parameters. During this stage of the project, we will try to determine what algorithm will work best and compare algorithms more once we have the code done in Python programming language.

5. After the data is properly cleaned, we will need to have a model coded into the Python programming language to be manipulated by training. In this training, we will need to focus on the training accuracy, the validation accuracy, the training loss, and the validation loss. To visualize these training metrics, we will be taking advantage of Matplotlib in Python. This will allow us to identify trends of under-fitting or over-fitting and correct these issues within our parameters. During this stage we will be able to compare algorithms and adjust these algorithms as needed.

6. We will need to then conduct evaluations on the performance of the model after training. We should receive good evaluation results, yet still be able to fine tune any parameters to refine the model for improvement. We will also begin to analyze how we have derived theoretical properties of the algorithms.

7. Once evaluations and refinements to the model are complete, we will display the results and research graphically so that the research is easily understood. The viewer should be able to understand how different factors can effect the price of housing within a specific ZIP Code.

References

[1] I. Bershadskiy, *Predicting house price using regression algorithm for machine learning*. [Online]. Available: https://yalantis.com/blog/predictive-algorithm-for-house-price/.

[2] F. Bergadano, R. Bertillone, D. Paolotti, and G. Ruffo, *Learning real estate automated valuation models from heterogenous data sources*, 2019. [Online]. Available: https://arxiv.org/pdf/1909.00704.pdf.

[3] C. Chow, *Machine learning for property valuation*, 2019. [Online]. Available: https://chrischow.github.io/dataandstuff/2019-09-15-machine-learning-for-property-valuation/.