Yingjie Lian & Shibo Tang

Course Name: CS-5350-Machine Learning

Machine Learning Final Report

May 4th, 2019

Machine Learning Final Report

Abstract

Using Machine Learning algorithms to predict future things by using old data has been more and more popular. For this project, Shibo Tang and Yingjie Lian will use Machine Learning algorithms extensively to apply and study onto housing prices prediction. Little work has been done to adapt it to the end-to-end training of the previous data on large-scale datasets upon greater Salt Lake City's housing prices. In this work, we are going to be focused on solving the problem of predicting house prices for house buyers and house sellers. In addition, we will take advantage of all of the feature variables available to use and use it to analyze and predict house prices.

Problem Definition and Motivation

As we know today, a house value is simply more than location and square footage. Like the features that make up a person, an educated party would want to know all the aspects that give a house its value. We are planning to use a machine learning algorithm to solve the house value problem.

The benefit of this study is that we can assume we have two clients at the same time. We can imagine that being a divorce lawyer for both interested parties. However, we cannot have both clients with no conflict of interest (Teboul 2018). Client Housebuyer: This client wants to find their next dream home with a reasonable price tag. They have their locations of interest ready. Now, they want to know if the house price matches the house value. With this study, they can understand which features (ex. Number of bathrooms, location, etc.) influence the final price of the house. If all matches, they can ensure that they are getting a fair price. Client House seller: Think of the average house-flipper. This client wants to take advantage of the features that influence a house price the most (Henderson and Ioannides 1983). They typically want to buy a house at a low price and invest on the features that will give the highest return. For example, buying a house at a good location but small square footage. The client will invest on making rooms at a small cost to get a large return.

Regression takes in a feature data as a vector. Every data point is defined as some number of input features. For instance, features about the house. The regression model uses these features to predict the Targets. The model trains itself by learning from the dataset to minimize a loss function.

Our solution

Normally, in Machine learning, we randomly split up (specific numbers can fluctuate depending on the problem) 70% of our data to train a model on, 20% of our to validate the performance on, and reserve 10% of our data for the very end to test our release performance on.

We repeatedly change the model, train, and validate, until our validation accuracy seems to be maximized. We believe the model that generalizes to new unseen data performs best, so we utilize the model that does best on the validation set.

To ensure the correctness and cleanest of our learning model, we are going to break everything into logical steps that allow us to ensure the cleanest accurate predictions:

- Load the experiment data and load the packages
- Analyzing the test variables and also the sale price
- Multivariable analysis for the test variables and clean data
- Feature the transformation/engineering
- Modeling and predictions
- Impute missing data and also clean data

Experimental evaluation

We evaluated once on our test set to approximate how well our model will generalize to the real world. In this way, we can build complicated models to represent our data and then calculate metrics that don't over-estimate our performance estimates for how well it will perform.

In time series modeling, we use Walk Forward Optimization, where we split up the time series into many temporal cutoffs T1, T2, T3, etc. We first train on data until T1 and validate our performance on the time period between T1 and T2. We then extend our training dataset until T2 and then validate the period between T2 and T3. Our overall validation performance metric,

therefore, is our average validation performance from the different data cuts, and our test set can be the very last time slice of the data or some data points throughout the different time slices that were excluded from the validation data points.

Future plan

Our future plan for this project is to keep doing multivariable analysis, then we will focus on imputing missing data. Also, we will import more data to train our model in the future. In the end, our final goal for our project is modeling and predicting the greater Salt Lake City's house prices in 3 years.

Our Github Repository

https://github.com/shibot/ML5350finalproject

Reference

1. Teboul, W. (2018, July 20). Why use Machine Learning Instead of Traditional Statistics. Retrieved from https://towardsdatascience.com/why-use-machine-learning-instead-of-traditional-statistics-334c2213700a
2. Henderson, J.V., and Y.M. Ioannides. 1983. “A Model of Housing Tenure Choice.” American Economic Review 73, no. 1: 98-113.