Team: Yingjie Lian, Shibo Tang

Course Name: CS-5350-Machine Learning

Instructor: Shandian Zhe

Version: March 9th, 2019

Prediction of Housing Prices with Machine Learning

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.

Introduction

In this article, we will mainly focus on supervised learning methods. First, because that is what our project was relying on, so that is the category we had to bring an answer to. Second, since that is generally what people think about when they hear 'Machine Learning'.

In fact, using Machine Learning algorithms will benefit our project in these several aspects. People generally associate Machine Learning with inferential statistics such as a discipline which aims to understand the underlying probability distribution of a phenomenon within a specific population. This is also what we want to do in ML in order to generate a prediction for housing prices.

In contrast, traditional statistics relies on assumptions. On the one hand, the first step of the statistical method is to choose a model with unknown parameters for the underlying law governing the observed property(Teboul 2018). On the other hand, correlations and other statistical tools help us determine the values for the parameters of this model. Hence, if the assumptions about the data are wrong, computation of the parameters will make no sense and the model will never fit your data with enough accuracy.

Motivation

The collapse of house prices often causes social disasters and the breakdown of many families. Many articles have shown that individuals' perceptions about constraints on their ability to purchase and own homes, meanwhile, are not generally predictive of future tenure intentions(Henderson and Ioannides 1983;). These findings suggest that future research on tenure decisions should do more to account for behavioral factors. The analysis finds that such beliefs are strong indicators of expectations to own, more so than even some economic and socio-demographic characteristics that are commonly assumed to drive tenure preferences, such as family composition and income while people buying houses.

As a result, we hope people will be able to rationally predict housing prices by using our researches in order to avoid excessive investment and reduce the social risks.

What have we done

For this project, we finished loading data and packages, test variable and multivariable analysis. The package we are using right now includes pandas, numpy, seaborn, matplotlib, warnings, xgboost, and lightgbm. Plus, we are using LinearRegreesion, LassoCV, Ridge, LassoLarsCV, and ElasticNetCV from sklearn.linear_model.

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib
import matplotlib.pyplot as plt
import warnings
import xgboost as xgb
import lightgbm as lgb

from scipy.stats import skew
from scipy import stats
from scipy.stats.stats import pearsonr
from scipy.stats import norm
from collections import Counter
from sklearn.linear_model import LinearRegression,LassoCV,
Ridge, LassoLarsCV,ElasticNetCV
from sklearn.model_selection import GridSearchCV
```

The data has been searched from Zillow.com. As we can see that the price is extremely correlated with the House location. Below is our analysis result from the Zillow.com. We can see the average house prices $250000.

```
count      1460.000000
mean     180921.195890
std       79442.502883
min       34900.000000
25%      129975.000000
50%      163000.000000
75%      214000.000000
max      755000.000000
```

There are two types of feature in our house data, numerical, and categorical. The categorical data is in those categories. There is no need to be linear. But we need to follow some kind of rule, such as if the feature is “Downtown”. We will have the results such as “Near” or “Far”. Then, we can predict the house prices based on the “Near” or “Far”. We will still work on how to predict the correct house price using the data from Zillows.com

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.

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.