Use Machine Learning Method Predict House Price

Feng Yongxiang yfeng24@buffalo.edu
Ashutosh Anand aanand7@buffalo.edu

Abstract-In general, predict house price is a difficult work, because the price is influenced by many factors. People who want to predict it should first collect many related data like type of house and neighborhood. Next use some methods to calculate the weight of every factor. Last to create the math model to predict the house price. But now we can collect massive data and use some models to let computer deals with data by itself. Usually, we can use different type of machine learning methods to get the ideal result.
Index Term-machine learning, random-forest tree.

1. Introduction

In the recent years, much data appeared in the world. We are busy to deal with massive seeming unrelated data and easily to drop in. Fortunately, many AI scientists create some methods for us to deal with these massive data. We can directly use these models to get connections with data, like price and environment. And these models can be calculated by computer. It can enhance work efficiency.

In this passage, we use some technic methods to solve the problem about prediction of house price. The process of it is clear, get data from website, preprocess data, select the model, train the model and get the result. We will use different methods to process the data to see which one is better to predict the result.

2. Database

There are many websites provide data. We choose Kaggle to get the data because it provides relatively complete data frame and many methods to use it like download and using API. This dataset includes 80 features like street and house style and 1 label SalePrice. We should use these features to predict the SalePrice. These features also divide into number and text.

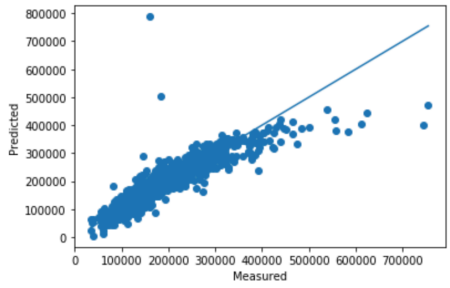
3. Method

Firstly, we load the data and show some information about it. It has the number data and text data and some of it misses some data. Then we visualize the data, get some information about the distribution of it. Some data are crowd and some data are discrete. Next we use the method to calculate the correlation of the features and the SalePrice. Some data have less correlation may drop off in the later. Next we preprocess the number data, we first get some miss data's medium number, and use the method in the sklearn to fill miss data with the medium number in the column. We preprocess the text data, because the model only receives number data, so we should transfer text data to number data. After transform the data, we also use medium data to fill in the miss data.

We select several models to train the data. First one is linear regression, we separate the train data and test data includes features and SalePrice. Use the MSE and RMSE to get result of train. Also, we use cross validation method try to get better result. Then we use sgd regression to reduce the error of model.

Figure1. Linear Regression

```
985538703.0200325 31393.290732575846
1261913335.9896653 35523.41954246051
```




```
/Users/fengyongxiang/anaconda3/lib/python3.7/site-packages/sklearn/utils/validation.py:760: DataConversionWarning:
A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for examp
le using ravel().
  y = column_or_1d(y, warn=True)
```

```
0.8288884500968096
```

We get the result of MSE and RMSE of linear regression. It has space to improve because the error is huge, and the sgd regression model can fit around 80% of data. Also, the figure shows accuracy of predict result and real result, the point next to the line is correct, the far point next to line, the less correct the predict.

We next use ridge regression to train the data, and get feature different result about linear regression.

Figure2. Ridge Regression

```
[[ 9.89776334e-01 -2.57512067e+02 -9.56628207e+01  1.66167595e-01
   1.18314965e+04  6.26953487e+03  2.29172464e+02  2.09459448e+01
   2.00574875e+01  7.68582345e-01  4.56314780e+00 -2.25191386e+00
   3.07918894e+00  2.15154792e+01  2.17538029e+01 -2.94736833e+01
   1.37964990e+01  1.35355587e+04  2.95034794e+03  1.03123835e+04
   3.20487771e+03 -3.38080521e+03 -7.07880786e+03  5.36873303e+03
   5.41149805e+03  6.39409227e+01  1.39097772e+04 -6.62136212e+00
   1.49614128e+01 -8.12309319e+00  4.18475190e+00 -1.17902608e+01
   2.66828171e+01  2.93513609e+01  7.83859924e-01 -8.18091288e+01
  -9.71557790e+02 -2.09249186e+04 -6.35583600e+03  3.22918461e+04
   5.20763364e+02 -5.53185484e+03 -6.03605330e+03  6.03605330e+03
   0.00000000e+00  0.00000000e+00 -6.31489089e+03  0.00000000e+00
  -1.77735981e+04 -8.88227854e+03  5.59021871e+03 -5.12673086e+03
   9.48097867e+03 -2.21264618e+04 -1.68336747e+04 -1.25262437e+04
  -4.43077920e+03 -1.32549686e+04 -1.62402148e+04 -1.17608655e+04
  -2.69459594e+04  5.56761574e+04  3.44430851e+04 -1.41196100e+04
  -1.95706508e+04 -1.84099701e+04 -1.43482712e+04  1.63824588e+04
   7.01115051e+04  1.29670311e+04  2.40137332e+04  7.88444987e+03
   1.67926804e+04 -3.40195911e+04  9.34246085e+03]]
1143685490.8507419 33818.41940201732
1261884103.4384894 35523.00808544357
```

We can see there are different MSE and RMSE between linear regression and ridge regression. It also has huge space to go up. In the rest time, we will search methods how to improve the performance of models.

Also, we use svm model and random-forest model to predict the result of house price. They have slight differences. So the next time is search how to get the best performance of every model to get the outstanding result.

4. Conclusion

In this experiment, we preprocess the data and select some models to train the data. Some works are progressive and get excellent result. But it also has many disadvantages. In the rest of experiment, we will pay attention to improve the performance of different models. Use some methods to reduce the error between real data and predict data. Hopefully we have time to build an input interface let someone enter some features of the house and get the SalePrice of the house.