UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

Master in Innovation and Research in Informatics

Machine Learning

House Price Prediction

Authors:
Asaf Badouh
Pau Rodríguez Esmerats

Professors:
Jaume Baixeries Juvillà
Marta Arias Vicente

June 20, 2018

Contents

1 Introduction

The goal of this project is to make the most out of data on house prices to build a model that is able to predict the price of a house based on it's attributes. This includes pre-processing the data, performing feature engineering (feature selection and/or extraction) and training a regression model to predict the price.
An important part of our effort will be focused on the decision about what are the most important variables or features that we could extract upon them, as well as what are the best parameter configurations for the model we will train. Different approaches to each phase will be considered and compared appropriately before choosing the final implementation.

This report is organized in different sections. First the different analysis approaches over the data are presented, where we discuss preprocessing, manual exploration, unsupervised analysis and testing assumptions over the data. Secondly, the different feature extraction techniques and resulting feature sets are explained. Then the different types of models and their configuration are presented. Finally, we explain the approach to perform feature selection and model selection throught different experiments. In the end, the conclusion of our work are presented.

2 Multivariate data analysis

2.1 Preprocessing

Missing data

Outliers

2.2 Exploratory data analysis

Histograms

Plots

Ratios
While exploring the variables in our dataset we had intuition about the ration relationship between that variable. For example, the ratio between number of bedrooms to number of bathroom may be good feature to help us predict the house price. Exploring on the Tax policy in USA[1], we learn that tax of house also base on the characteristics of the house. Therefore we though it will be useful to check those features.

need to add a bit about the trees here

Logs

Gaussianization

Correlation analysis

2.3 Unsupervised analysis

PCA 2 **Clustering**
We cluster the individuals using hierarchical clustering. We did it using only continuance variables in order to see in we have some pattern in the data. As we can see in Fig.1, more than 85% of the individuals are grouped in a single cluster, the rest are divided into 2-3 clusters with smaller size. therefore, we deduce that we will not split our data into cluster and we will explore it as a whole.

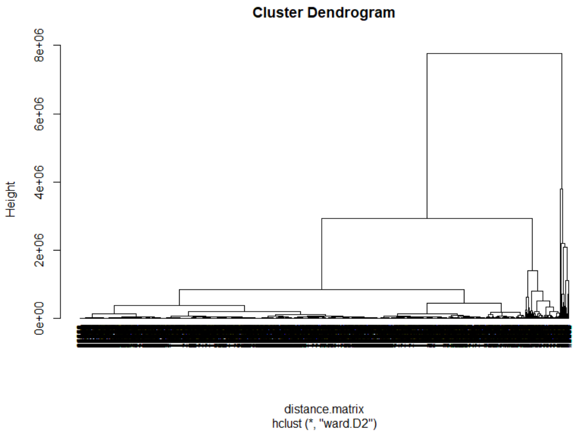

Figure 1: Hierarchical Clustering

2.4 Assumptions testing

Normality

Independence

Homoscedasticity

Skewness

3 Feature extraction

3.1 Intuitive transformations

3.2 Uncorrelated subsets

3.3 PCA

4 Models

In this section we will present the different types of models that we will later train during the experiments.

4.1 Methodology

4.2 Model types

Linear

Polynomial

Ridge regression

Lasso regression

PCR

Decision trees

Random Forest

5 Experiments

This section exposed the selected approach to perform a sound feature and model selection. We are usually faced with problems where feature selection is performed before model selection. In our case, with many different models that are so different between them, the selection of feature set can affect a lot the model performance. So in the end, to avoid performing space exploration without a strategy, we decide to choose between different strategical approaches:

- Perform feature selection based on data analysis and then perform model selection over selected feature set
- pick a model based on tested assumptions on the data and then perform feature selection using forward selection while trainig the model
- perform PCA and select significant components, then perform model selection with the extracted PCA features and finally perform feature selection over all feature sets with the selected model.
- pick a baseline model type, and then perform feature selection while training the baseline model. Once the feature set is selected, we then could perform model selection over all the different types
- perform all the previous approaches and choose the pair of feature set and model that yields the lower validation error.

5.1 Approach 1

We concluded that this approach does not make sense, unless we limit our feature space to features extracted from a PCA analysis, where we can use the percentage of exaplined inertia to select features.

	Model	Feature set	Validation.NRMSE	Testing.NRMSE
1	simple linear regression	featureset_nocorrelation01	1.37	0.60
2	ridge regression MASS	featureset_nocorrelation01	0.64	0.60
3	ridge regression GLMNET	featureset_nocorrelation01	0.63	0.61
4	lasso regression GLMNET	featureset_nocorrelation01	0.64	0.60
5	Lasso regression LARS	featureset_nocorrelation01	0.64	0.61

5.2 Approach 2

This approach turns out to be exploration of the space of feature sets and models without any real guidance or strategy. Even if this if doable to certain extent, we decide to drop this approach as it won't be suitable in a real case with bigger data sets.

	Model	Feature set	Validation.NRMSE	Testing.NRMSE
1	simple linear regression	featureset_allmanual	1.42	0.51
2	simple linear regression	featureset_logratios	1.39	0.56
3	simple linear regression	featureset_logs	1.37	0.61
4	simple linear regression	featureset_nocorrelation01	1.37	0.60
5	simple linear regression	featureset_nocorrelation02	1.37	0.59
6	simple linear regression	featureset_nocorrelation03_logs	1.33	0.67
7	simple linear regression	featureset_nocorrelation04_ratios	1.30	0.72
8	simple linear regression	featureset_original_nooutliers	1.40	0.55
9	simple linear regression	featureset_pca_nooutliers	1.32	0.68
10	simple linear regression	featureset_pca	1.37	0.68
11	simple linear regression	featureset_ratios	1.41	0.54
12	simple linear regression	raw_continuous_dataset	1.38	0.66
13	ridge regression MASS	featureset_allmanual	0.56	0.51
14	ridge regression MASS	featureset_logratios	0.62	0.56
15	ridge regression MASS	featureset_logs	0.62	0.61
16	ridge regression MASS	featureset_nocorrelation01	0.64	0.60
17	ridge regression MASS	featureset_nocorrelation02	0.63	0.59
18	ridge regression MASS	featureset_nocorrelation03_logs	0.68	0.67
19	ridge regression MASS	featureset_nocorrelation04_ratios	0.74	0.72
20	ridge regression MASS	featureset_original_nooutliers	0.58	0.55
21	ridge regression MASS	featureset_pca_nooutliers	0.72	0.68
22	ridge regression MASS	featureset_pca	0.71	0.68
23	ridge regression MASS	featureset_ratios	0.57	0.54
24	ridge regression MASS	raw_continuous_dataset	0.70	0.66
25	ridge regression GLMNET	featureset_allmanual	0.56	0.53
26	ridge regression GLMNET	featureset_logratios	0.62	0.58
27	ridge regression GLMNET	featureset_logs	0.62	0.62
28	ridge regression GLMNET	featureset_nocorrelation01	0.63	0.61
29	ridge regression GLMNET	featureset_nocorrelation02	0.63	0.60
30	ridge regression GLMNET	featureset_nocorrelation03_logs	0.68	0.67
31	ridge regression GLMNET	featureset_nocorrelation04_ratios	0.74	0.73
32	ridge regression GLMNET	featureset_original_nooutliers	0.58	0.56
33	ridge regression GLMNET	featureset_pca_nooutliers	0.71	0.69
34	ridge regression GLMNET	featureset_pca	0.69	0.68
35	ridge regression GLMNET	featureset_ratios	0.57	0.55
36	ridge regression GLMNET	raw_continuous_dataset	0.69	0.67
37	lasso regression GLMNET	featureset_allmanual	0.55	0.51
38	lasso regression GLMNET	featureset_logratios	0.61	0.57
39	lasso regression GLMNET	featureset_logs	0.62	0.61
40	lasso regression GLMNET	featureset_nocorrelation01	0.64	0.60
41	lasso regression GLMNET	featureset_nocorrelation02	0.63	0.59
42	lasso regression GLMNET	featureset_nocorrelation03_logs	0.69	0.67
43	lasso regression GLMNET	featureset_nocorrelation04_ratios	0.74	0.72
44	lasso regression GLMNET	featureset_original_nooutliers	0.58	0.55
45	lasso regression GLMNET	featureset_pca_nooutliers	0.71	0.68
46	lasso regression GLMNET	featureset_pca	0.70	0.68
47	lasso regression GLMNET	featureset_ratios	0.57	0.54
48	lasso regression GLMNET	raw_continuous_dataset	0.69	0.66
49	Lasso regression LARS	featureset_allmanual	0.55	0.51
50	Lasso regression LARS	featureset_logratios	0.62	0.58
51	Lasso regression LARS	featureset_logs	0.61	0.62
52	Lasso regression LARS	featureset_nocorrelation01	0.64	0.61
53	Lasso regression LARS	featureset_nocorrelation02	0.63	0.60
54	Lasso regression LARS	featureset_nocorrelation03_logs	0.69	0.67
55	Lasso regression LARS	featureset_nocorrelation04_ratios	0.74	0.72
56	Lasso regression LARS	featureset_original_nooutliers	0.58	0.55
57	Lasso regression LARS	featureset_pca_nooutliers	0.72	0.68

	Model	Feature set	Validation.NRMSE	Testing.NRMSE
58	Lasso regression LARS	featureset_pca	0.71	0.68
59	Lasso regression LARS	featureset_ratios	0.58	0.54
60	Lasso regression LARS	raw_continuous_dataset	0.69	0.66
61	PCR	featureset_allmanual	0.75	0.74
62	PCR	featureset_logratios	0.74	0.73
63	PCR	featureset_logs	0.79	0.78
64	PCR	featureset_nocorrelation01	0.81	0.79
65	PCR	featureset_nocorrelation02	0.77	0.73
66	PCR	featureset_nocorrelation03_logs	0.82	0.81
67	PCR	featureset_nocorrelation04_ratios	0.85	0.83
68	PCR	featureset_original_nooutliers	0.74	0.71

5.3 Approach 3

In this experimental approach, we combine PCA for feature selection, then we perform model selection on that new dataset. Once the model has been selected, we perform again feature selection over all the feature sets (including again PCA) to finally select the features.
We begin by doing a Principal Components Analysis over all continuous variables of the dataset. Within this analysis we select the significant components (by the percentage of explained inertia) and then build a new preprocessed dataset using the projections of each individual on the principal components and its target value (the price). The next step is to train all different types of models over this dataset. For each model type, we perform cross validation to select the best hyper parameters of the model. We finally compare all the validation errors of the best models of each type to select our best model. Once the model is selected, we perform again crossvalidation over each of the feature data sets that we have availabe. In the same manner, the selected feature data set will be the one that yields the minimum validation error with the chosen model.

	Model	Feature set	Validation.NRMSE	Testing.NRMSE
1	simple linear regression	featureset_pca	1.37	0.68
2	ridge regression MASS	featureset_pca	0.71	0.68
3	ridge regression GLMNET	featureset_pca	0.69	0.68
4	lasso regression GLMNET	featureset_pca	0.70	0.68
5	Lasso regression LARS	featureset_pca	0.71	0.68
6	ridge regression GLMNET	featureset_allmanual	0.56	0.53
7	ridge regression GLMNET	featureset_logratios	0.62	0.58
8	ridge regression GLMNET	featureset_logs	0.62	0.62
9	ridge regression GLMNET	featureset_nocorrelation01	0.63	0.61
10	ridge regression GLMNET	featureset_nocorrelation02	0.63	0.60
11	ridge regression GLMNET	featureset_nocorrelation03_logs	0.68	0.67
12	ridge regression GLMNET	featureset_nocorrelation04_ratios	0.74	0.73
13	ridge regression GLMNET	featureset_original_nooutliers	0.58	0.56
14	ridge regression GLMNET	featureset_pca_nooutliers	0.71	0.69
15	ridge regression GLMNET	featureset_pca	0.69	0.68
16	ridge regression GLMNET	featureset_ratios	0.57	0.55
17	ridge regression GLMNET	raw_continuous_dataset	0.69	0.67

5.4 Approach 4

In this experimental approach, we choose a baseline model and the train it over different feature sets in order to do feature selection. After that, all the models are trained over the selected feature set to do model selection.

	Model	Feature set	Validation.NRMSE	Testing.NRMSE
1	ridge regression MASS	featureset_allmanual	0.56	0.51
2	ridge regression MASS	featureset_logratios	0.62	0.56
3	ridge regression MASS	featureset_logs	0.62	0.61
4	ridge regression MASS	featureset_nocorrelation01	0.64	0.60
5	ridge regression MASS	featureset_nocorrelation02	0.63	0.59
6	ridge regression MASS	featureset_nocorrelation03_logs	0.68	0.67
7	ridge regression MASS	featureset_nocorrelation04_ratios	0.74	0.72
8	ridge regression MASS	featureset_original_nooutliers	0.58	0.55
9	ridge regression MASS	featureset_pca_nooutliers	0.72	0.68
10	ridge regression MASS	featureset_pca	0.71	0.68
11	ridge regression MASS	featureset_ratios	0.57	0.54
12	ridge regression MASS	raw_continuous_dataset	0.70	0.66
13	simple linear regression	featureset_ratios	1.41	0.54
14	ridge regression MASS	featureset_ratios	0.57	0.54
15	ridge regression GLMNET	featureset_ratios	0.57	0.55
16	lasso regression GLMNET	featureset_ratios	0.57	0.54
17	Lasso regression LARS	featureset_ratios	0.58	0.54

5.5 Approach 5

The last approach consists of gathering the results of previous approaches and compare. We can select the pair feature set and model that obtain the lower validation error.

	Model	Feature set	Validation.NRMSE	Testing.NRMSE
1	simple linear regression	featureset_nocorrelation01	1.37	0.60
2	ridge regression MASS	featureset_nocorrelation01	0.64	0.60
3	ridge regression GLMNET	featureset_nocorrelation01	0.63	0.61
4	lasso regression GLMNET	featureset_nocorrelation01	0.64	0.60
5	Lasso regression LARS	featureset_nocorrelation01	0.64	0.61
6	simple linear regression	featureset_allmanual	1.42	0.51
7	simple linear regression	featureset_logratios	1.39	0.56
8	simple linear regression	featureset_logs	1.37	0.61
9	simple linear regression	featureset_nocorrelation01	1.37	0.60
10	simple linear regression	featureset_nocorrelation02	1.37	0.59
11	simple linear regression	featureset_nocorrelation03_logs	1.33	0.67
12	simple linear regression	featureset_nocorrelation04_ratios	1.30	0.72
13	simple linear regression	featureset_original_nooutliers	1.40	0.55
14	simple linear regression	featureset_pca_nooutliers	1.32	0.68
15	simple linear regression	featureset_pca	1.37	0.68
16	simple linear regression	featureset_ratios	1.41	0.54
17	simple linear regression	raw_continuous_dataset	1.38	0.66
18	ridge regression MASS	featureset_allmanual	0.56	0.51
19	ridge regression MASS	featureset_logratios	0.62	0.56
20	ridge regression MASS	featureset_logs	0.62	0.61
21	ridge regression MASS	featureset_nocorrelation01	0.64	0.60
22	ridge regression MASS	featureset_nocorrelation02	0.63	0.59
23	ridge regression MASS	featureset_nocorrelation03_logs	0.68	0.67
24	ridge regression MASS	featureset_nocorrelation04_ratios	0.74	0.72
25	ridge regression MASS	featureset_original_nooutliers	0.58	0.55
26	ridge regression MASS	featureset_pca_nooutliers	0.72	0.68
27	ridge regression MASS	featureset_pca	0.71	0.68
28	ridge regression MASS	featureset_ratios	0.57	0.54
29	ridge regression MASS	raw_continuous_dataset	0.70	0.66
30	ridge regression GLMNET	featureset_allmanual	0.56	0.53

	Model	Feature set	Validation.NRMSE	Testing.NRMSE
31	ridge regression GLMNET	featureset_logratios	0.62	0.58
32	ridge regression GLMNET	featureset_logs	0.62	0.62
33	ridge regression GLMNET	featureset_nocorrelation01	0.63	0.61
34	ridge regression GLMNET	featureset_nocorrelation02	0.63	0.60
35	ridge regression GLMNET	featureset_nocorrelation03_logs	0.68	0.67
36	ridge regression GLMNET	featureset_nocorrelation04_ratios	0.74	0.73
37	ridge regression GLMNET	featureset_original_nooutliers	0.58	0.56
38	ridge regression GLMNET	featureset_pca_nooutliers	0.71	0.69
39	ridge regression GLMNET	featureset_pca	0.69	0.68
40	ridge regression GLMNET	featureset_ratios	0.57	0.55
41	ridge regression GLMNET	raw_continuous_dataset	0.69	0.67
42	lasso regression GLMNET	featureset_allmanual	0.55	0.51
43	lasso regression GLMNET	featureset_logratios	0.61	0.57
44	lasso regression GLMNET	featureset_logs	0.62	0.61
45	lasso regression GLMNET	featureset_nocorrelation01	0.64	0.60
46	lasso regression GLMNET	featureset_nocorrelation02	0.63	0.59
47	lasso regression GLMNET	featureset_nocorrelation03_logs	0.69	0.67
48	lasso regression GLMNET	featureset_nocorrelation04_ratios	0.74	0.72
49	lasso regression GLMNET	featureset_original_nooutliers	0.58	0.55
50	lasso regression GLMNET	featureset_pca_nooutliers	0.71	0.68
51	lasso regression GLMNET	featureset_pca	0.70	0.68
52	lasso regression GLMNET	featureset_ratios	0.57	0.54
53	lasso regression GLMNET	raw_continuous_dataset	0.69	0.66
54	Lasso regression LARS	featureset_allmanual	0.55	0.51
55	Lasso regression LARS	featureset_logratios	0.62	0.58
56	Lasso regression LARS	featureset_logs	0.61	0.62
57	Lasso regression LARS	featureset_nocorrelation01	0.64	0.61
58	Lasso regression LARS	featureset_nocorrelation02	0.63	0.60
59	Lasso regression LARS	featureset_nocorrelation03_logs	0.69	0.67
60	Lasso regression LARS	featureset_nocorrelation04_ratios	0.74	0.72
61	Lasso regression LARS	featureset_original_nooutliers	0.58	0.55
62	Lasso regression LARS	featureset_pca_nooutliers	0.72	0.68
63	Lasso regression LARS	featureset_pca	0.71	0.68
64	Lasso regression LARS	featureset_ratios	0.58	0.54
65	Lasso regression LARS	raw_continuous_dataset	0.69	0.66
66	PCR	featureset_allmanual	0.75	0.74
67	PCR	featureset_logratios	0.74	0.73
68	PCR	featureset_logs	0.79	0.78
69	PCR	featureset_nocorrelation01	0.81	0.79
70	PCR	featureset_nocorrelation02	0.77	0.73
71	PCR	featureset_nocorrelation03_logs	0.82	0.81
72	PCR	featureset_nocorrelation04_ratios	0.85	0.83
73	PCR	featureset_original_nooutliers	0.74	0.71
74	simple linear regression	featureset_pca	1.37	0.68
75	ridge regression MASS	featureset_pca	0.71	0.68
76	ridge regression GLMNET	featureset_pca	0.69	0.68
77	lasso regression GLMNET	featureset_pca	0.70	0.68
78	Lasso regression LARS	featureset_pca	0.71	0.68
79	ridge regression GLMNET	featureset_allmanual	0.56	0.53
80	ridge regression GLMNET	featureset_logratios	0.62	0.58
81	ridge regression GLMNET	featureset_logs	0.62	0.62
82	ridge regression GLMNET	featureset_nocorrelation01	0.63	0.61
83	ridge regression GLMNET	featureset_nocorrelation02	0.63	0.60
84	ridge regression GLMNET	featureset_nocorrelation03_logs	0.68	0.67
85	ridge regression GLMNET	featureset_nocorrelation04_ratios	0.74	0.73
86	ridge regression GLMNET	featureset_original_nooutliers	0.58	0.56
87	ridge regression GLMNET	featureset_pca_nooutliers	0.71	0.69

	Model	Feature set	Validation.NRMSE	Testing.NRMSE
88	ridge regression GLMNET	featureset_pca	0.69	0.68
89	ridge regression GLMNET	featureset_ratios	0.57	0.55
90	ridge regression GLMNET	raw_continuous_dataset	0.69	0.67
91	ridge regression MASS	featureset_allmanual	0.56	0.51
92	ridge regression MASS	featureset_logratios	0.62	0.56
93	ridge regression MASS	featureset_logs	0.62	0.61
94	ridge regression MASS	featureset_nocorrelation01	0.64	0.60
95	ridge regression MASS	featureset_nocorrelation02	0.63	0.59
96	ridge regression MASS	featureset_nocorrelation03_logs	0.68	0.67
97	ridge regression MASS	featureset_nocorrelation04_ratios	0.74	0.72
98	ridge regression MASS	featureset_original_nooutliers	0.58	0.55
99	ridge regression MASS	featureset_pca_nooutliers	0.72	0.68
100	ridge regression MASS	featureset_pca	0.71	0.68
101	ridge regression MASS	featureset_ratios	0.57	0.54
102	ridge regression MASS	raw_continuous_dataset	0.70	0.66
103	simple linear regression	featureset_ratios	1.41	0.54
104	ridge regression MASS	featureset_ratios	0.57	0.54
105	ridge regression GLMNET	featureset_ratios	0.57	0.55
106	lasso regression GLMNET	featureset_ratios	0.57	0.54
107	Lasso regression LARS	featureset_ratios	0.58	0.54

5.6 Results

The following table summarizes the results of the experiments performed. The models are trained over the selected feature set and then they are ranked according to their validation error. The model with minimal validation error will be our selection.

6 Conclusion

w

We proposed a series of improvements over the current work:

-

References

[1] Property tax in the united states. https://en.wikipedia.org/wiki/Property_tax_in_the_United_States#Tax_rates, 2018. Accessed: 20th May, 2018.

[2] Kaggle: Your Home for Data Science. House sales in king county, usa. https://www.kaggle.com/harlfoxem/housesalesprediction, 2016. This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015, Accessed: 4th April, 2018.

[3] The statistic portal. Number of existing homes sold in the united states from 2005 to 2018. https://www.statista.com/statistics/226144/us-existing-home-sales, 2018. Accessed: 5th April, 2018.