Feature Selection with Permutation Testing on the Housing dataset

Tram Ngo

May 2023

Abstract

This project is focused on analyzing a housing dataset and performing feature selection using permutation testing. The objective is to identify the most important features that have a statistically significant effect on home prices. The dataset includes various features such as the number of bedrooms, bathrooms, square footage, location, and other characteristics of the houses.

The project starts by discussing a couple of mathematical concepts that are the baseline for the other concepts of this paper. After that, the paper goes on and explores the dataset and visualizes the relationships between different features and the target variable (price). Hypothesis tests are then conducted to determine the significance of different features on home prices.

Based on the analysis, the top 7 features by importance were 'sqft_living', 'grade', 'lat', 'yr_built', 'long', 'sqft_living15', and 'waterfront'. Using these features, three machine learning models were trained and evaluated: linear regression, random forest regressor, and support vector machines. The results showed that the random forest regressor had the best performance in terms of RMSE.

Contents

1 Overview of mathematical concepts

1.1 What is permutation testing?

Permutation testing is a statistical approach that includes random shuffling or reassignment of group labels multiple times to evaluate the performance of a model or the variation between the groups. This generates a null distribution of the test statistic, which can be used to estimate the probability or p-value of obtaining the observed result by chance.

To explain this mathematically, let H_0 be the null hypothesis that there is no difference between two groups, and H_1 be the alternative hypothesis that there is a significant difference between them. To test the hypothesis, we need to take the following steps:

1. Calculate the observed test statistic, $t_{observed}$, based on the original data.

2. Randomly permute or reassign the labels of the samples or variables and recalculate the test statistic t_i for the i^{th} iteration to generate a null distribution of test statistics.

3. Calculate the p-value as the proportion of t_i that are equal to or more significant than the observed $t_{observed}$. If the p-value is smaller than the selected significance level, reject the H_0 and accept the alternative hypothesis.

1.2 What is Feature Selection?

Feature selection is the process of selecting a subset of features from a larger set of features that are most relevant to the target variable. One interesting way to look at why we have to perform feature selection in a predictive model is to think about the Principle of Parsimony (part of the Occam's Razor). The concept simply says: given a set of equally good explanations for a given problem, the correct explanation is the simplest explanation as possibly obtained. The goal of feature selection is to improve the model's accuracy by reducing the number of features - to avoid the curse of dimensionality - and removing irrelevant or redundant features that do not contribute much to the target variable. This way, we will have the "simplest" and thus presumably effective solution to the problem.

Feature selection can be mathematically represented as an optimization problem. The cost function represents the model's performance, while the constraints impose limitations on the selected features. The general form of the feature selection problem can be written as:

minimize C(F) subject to: $|F| <= k$

where C(F) is the cost function of the model, F is the set of selected features, k is the maximum number of allowed features, and $|F|$ represents the cardinality of the set F.

Different cost functions can be used, depending on the type of the predictive model. For instance, in linear regression, the cost function can be the mean squared error, while in classification models, the cost function can be the misclassification rate or log-likelihood.

The constraint $|F| <= k$ ensures that the number of selected features does not exceed a pre-defined maximum number k.

2 Overview of the Housing dataset

The dataset contains information about home prices and their associated features. The dataset has 21,613 rows and 21 columns, with each row representing a single home sale. The dataset was created for the purpose of exploring the relationships between various features of homes (such as size, location, and condition) and their sale prices, and to build models that can predict the sale prices of homes based on these features.

2.1 Exploration Data Analysis

Let us look at the interconnections between prices and a few variables such as bedrooms/bathrooms, sqft living, sqft lot. Further exploration work can be found under the Appendix section with a link to the Github repository that hosts several in-depth charts about the dataset.

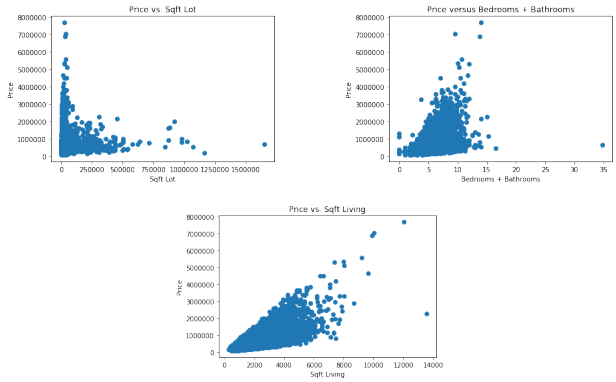

Figure 1: Scatter Plots

2.2 Conduct hypothesis Testing

The main goal of this paper is to choose the most important and relevant features from a the set of 21 given features. Before we take the main approach - which is to apply permutation testing - based Feature selection technique on it, let us create some hypothesis tests for some "reasonably relevant" features from a common-sense point of view. These are: Lattitude, Longtitude, Bedrooms, Bathrooms, Sqft Living, and Sqft Lot size. The below are the summaries of the OLS models conducted:

The multiple linear regression analysis shows that the variables "bedrooms" and "bathrooms" are statistically significant predictors of home prices at the 95% confidence level. The coefficient for the "bedrooms" variable is negative, which suggests that, all other things being equal, as the number of bedrooms in a home increases, the price of the home tends to decrease. The coefficient for the "bathrooms" variable is positive, which suggests that, all other things being equal, as the number of bathrooms in a home increases, the price of the home tends to increase.

No of Bedrooms/No of Bathrooms Effect on Prices

OLS Regression Results

Dep. Variable:	price	R-squared:	0.278
Model:	OLS	Adj. R-squared:	0.278
Method:	Least Squares	F-statistic:	4154.
Date:	Wed, 10 May 2023	Prob (F-statistic):	0.00
Time:	15:53:41	Log-Likelihood:	-3.0409e+05
No. Observations:	21613	AIC:	6.082e+05
Df Residuals:	21610	BIC:	6.082e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.064e+04	8270.147	-3.705	0.000	-4.69e+04	-1.44e+04
bedrooms	2.014e+04	2664.002	7.559	0.000	1.49e+04	2.54e+04
bathrooms	2.378e+05	3217.093	73.912	0.000	2.31e+05	2.44e+05

Omnibus:	17367.076	Durbin-Watson:	1.961
Prob(Omnibus):	0.000	Jarque-Bera (JB):	907583.417
Skew:	3.478	Prob(JB):	0.00
Kurtosis:	33.975	Cond. No.	16.9

Similarly, I also found that sqft_living, sqft_lot, lattitude, and longtitude are also statistically significant, with each coefficient having a p value less than 0.05.

SQFT LIVING/SQFT LOT Effect on Prices

OLS Regression Results

Dep. Variable:	price	R-squared:	0.494
Model:	OLS	Adj. R-squared:	0.494
Method:	Least Squares	F-statistic:	1.054e+04
Date:	Wed, 10 May 2023	Prob (F-statistic):	0.00
Time:	15:52:58	Log-Likelihood:	-3.0025e+05
No. Observations:	21613	AIC:	6.005e+05
Df Residuals:	21610	BIC:	6.005e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.39e+04	4398.565	-9.981	0.000	-5.25e+04	-3.53e+04
sqft_living	282.8787	1.964	144.030	0.000	279.029	286.728
sqft_lot	-0.2893	0.044	-6.644	0.000	-0.375	-0.204

Omnibus:	14768.444	Durbin-Watson:	1.982
Prob(Omnibus):	0.000	Jarque-Bera (JB):	538142.185
Skew:	2.809	Prob(JB):	0.00
Kurtosis:	26.791	Cond. No.	1.09e+05

Longtitude/Lattitude Effect on Prices

OLS Regression Results

Dep. Variable:	price	R-squared:	0.098
Model:	OLS	Adj. R-squared:	0.098
Method:	Least Squares	F-statistic:	1178.
Date:	Wed, 10 May 2023	Prob (F-statistic):	0.00
Time:	15:51:41	Log-Likelihood:	-3.0649e+05
No. Observations:	21613	AIC:	6.130e+05
Df Residuals:	21610	BIC:	6.130e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.872e+07	2.13e+06	-8.800	0.000	-2.29e+07	-1.46e+07
lat	8.365e+05	1.73e+04	48.428	0.000	8.03e+05	8.7e+05
long	1.679e+05	1.7e+04	9.880	0.000	1.35e+05	2.01e+05

Omnibus:	20459.631	Durbin-Watson:	1.973
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1556719.862
Skew:	4.408	Prob(JB):	0.00
Kurtosis:	43.632	Cond. No.	1.18e+05

3 Applying Permutation testing-based Feature Selection on the Housing dataset

Given the scope of this final project, I performed the following steps to apply feature selection using permutation/randomization technique under the hood to permute the values for each of the features while keeping all of the other features constant. Finally, I ranked the features by the changes in RMSE (Root Mean Squared Error) that the permutation has caused. The features with the biggest changes are considered to be the most impactful features and will be selected for the final treatment model.

	Features	Changes in RMSE
0	lat	122009.679837
1	sqft_living	99708.746095
2	grade	71443.255218
3	long	66560.019545
4	yr_built	20544.235310
5	waterfront	15868.488646
6	sqft_living15	11852.301252
7	bathrooms	5865.225501
8	zipcode	5761.097187
9	sqft_above	4053.559176
10	sqft_basement	3739.843569
11	view	2819.059234
12	sqft_lot15	1752.043104
13	sqft_lot	1465.562212
14	condition	561.229505
15	bedrooms	420.099383
16	floors	248.694452
17	yr_renovated	90.472518

Figure 2: Ranked Features

1. Define the loss metric: I chose RMSE as the loss metric for this problem because the price unit (dollars) can be quite big. Using MSE (Mean Squared Error) will magnify such magnitude and might make impact or the changes in the model's results harder to grasp. Using MAE (Mean Absolute Error), on the other hand, might not take into account the positive/negative impact of a feature on the home prices, which is an important impact.

2. Perform permutation testing: For each feature in the dataset, I randomly permuted its values in the train set, fit the model, evaluate the loss metric on the test set, and record the difference in the RMSE compared to the non-permutated set. I then repeated this process multiple times to obtain a distribution of the differences.

3. Rank the features: Lastly, I ranked the features by the magnitude of their average difference in RMSE compared to the non-permutated data.

I decided to choose the top 7 features and stopped at the 8th - bathrooms, as the changes in RMSE dropped almost 45% from the 7th to the 8th feature in the ranking chart. That was a good sign as to where to stop.

4 Train and Evaluate

4.1 Train models

We are going to train a baseline model that looks at all of the available features and a model that only takes in the seven most important features by ranking specified above. For each model, we will look at three algorithms: Linear Regression (LR), Random Forest Regressor (RFR), and Support Vector Machines (SVM). These three algorithms satisfy the below criteria for this particular regression problem: a high level of interpretability, simplicity, and accuracy.

4.2 Evaluate models

The below is the results of the baseline model trained on a whole set of features and the treatment model trained on the 7 most important features.

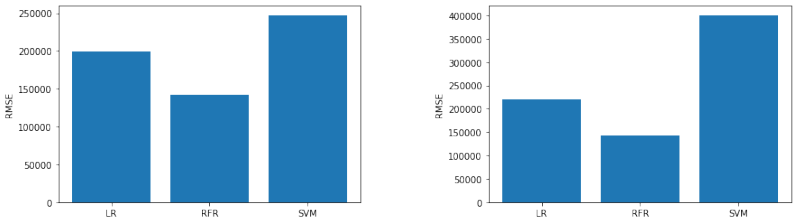

Figure 3: Baseline Model (Left) VS Treatment Model (Right)

We can see that the model without feature selection performed slightly better with a lower RMSE value. This suggests that including all available features in the model may be more effective in predicting home prices in this dataset. However, it's worth noting that the difference in RMSE between the two models is relatively small, so further analysis and experimentation may be needed to draw more conclusive insights.

5 Advantages and Disadvantages

There are many advantages as to why we should use permutation testing-based feature selection. These include:

- Reduces overfitting: Overfitting occurs when the model is too complex and captures noise in the training data rather than the underlying patterns. By randomly permuting the features and evaluating their importance, we add a randomization factor to the model that could separate the features by how important they are.

- Identifies relevant features: Permutation testing can identify the most relevant features for the prediction task, by measuring the contribution of each feature to the model's performance. This can help reduce the number of features in the model and reduces the model complexity that can cause computational exhaustion.

- Robust to noise: Permutation testing is less sensitive to noise in the data than traditional statistical methods, as it focuses on the overall performance of the model rather than individual data points.

Meanwhile, there are also some disadvantages to the technique that we have to consider before applying it to a problem:

- Computationally expensive: Permutation testing can be computationally expensive, especially when applied to large and complex datasets. This can make it challenging to apply in practice.

- May require a large number of permutations: Permutation testing may require a large number of permutations to obtain a reliable estimate of the model's performance, which can further increase computational complexity and time requirements.

- May not always identify the optimal feature subset: While permutation testing can identify the most relevant features, it may not always identify the optimal subset of features because the optimal subset may depend on the specific problem and the particular model used.

6 Alternatives

It is worth noting that there are other alternatives when it comes to feature selection. In particular, Bayesian regression using Horseshoe priors is useful in in high-dimensional or highly sparse settings where many features are likely irrelevant or noisy. Furthermore, correlation-based feature selection can also be a method worth considering due to its high interpretability and simplicity. This technique selects features based on their correlation with the target variable, and can be applied to both linear and nonlinear relationships. It can be faster and less computationally intensive than permutation testing, but may not capture complex relationships between features.

7 Conclusion

All in all, permutation testing was used in this project to perform feature selection and identify the most important features for predicting home prices in the Housing dataset. We trained and evaluated several machine learning models using the selected features, and compared their performance using the RMSE

loss metric.

Another interesting application of this dataset is to use it to analyze trends and patterns in the housing market, to identify factors that affect the supply and demand of housing, or to explore the relationship between housing and other socioeconomic indicators such as unemployment, educational levels, etc.

8 Appendix

This is the Github repository where I hosted all of the code:
https://github.com/tramngo1603/Theory-of-Stats-II-Final-Project

9 Bibliography

Next page.