Boston Housing Price Prediction Model

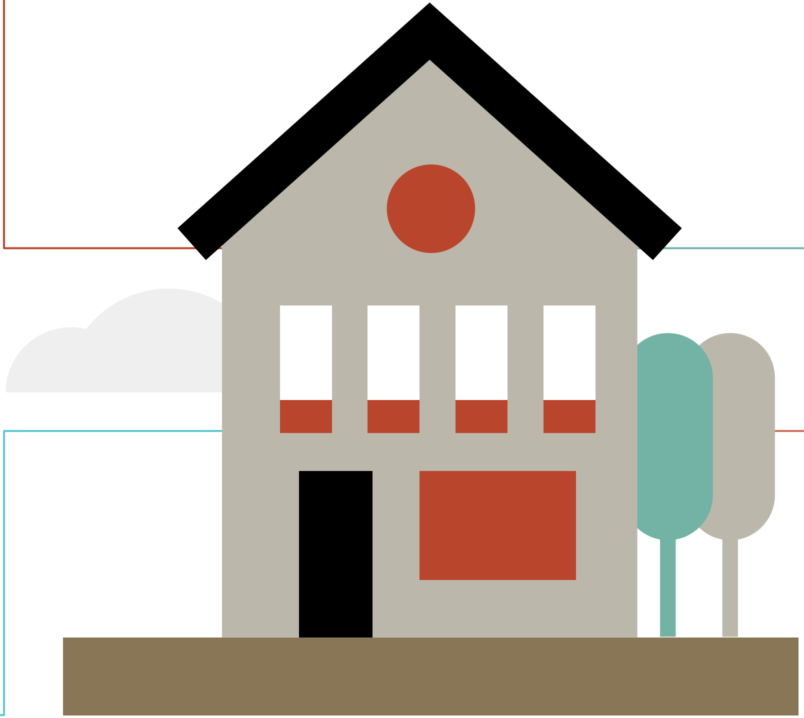
Regression Model

By Suci Rahma Nura

Suci Rahma Nura

- Data Science Project Based Intern IDX Partners X Rakamin
- Data Science Intern 360DigiTMG
- Credit Analyst PT. BFI Finance Indonesia, Tbk
- Bachelor's degree in Mathematics Andalas University

Table of Contents



Flowchart : Regression Model

Data Understanding

Check missing and duplicate values

Data Preprocessing

VIF Score and Heatmap correlation

Model Training

- Using Ridge and Lasso
- Find the best alpha

Evaluation

- Diagnostic Study
- Matrix Evaluation using MAE, MAPE, and RSME

01

Background Business and Objective

Background Business

Creating a Boston Housing Price Prediction Model aims to **support real estate** professionals and investors by providing insights into market trends, optimizing pricing and marketing strategies, and enabling informed decision-making in a dynamic real estate market.

Objective

To **accurately predict** home prices based on relevant features, enabling informed real estate decisions

02

Data Understanding

```
boston.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 506 entries, 0 to 505
Data columns (total 14 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   crim     506 non-null    float64
 1   zn       506 non-null    float64
 2   indus    506 non-null    float64
 3   chas     506 non-null    int64  
 4   nox      506 non-null    float64
 5   rm       506 non-null    float64
 6   age      506 non-null    float64
 7   dis      506 non-null    float64
 8   rad      506 non-null    int64  
 9   tax      506 non-null    int64  
 10  ptratio  506 non-null    float64
 11  black    506 non-null    float64
 12  lstat    506 non-null    float64
 13  medv     506 non-null    float64
dtypes: float64(11), int64(3)
memory usage: 55.5 KB
```

```
[ ] boston.duplicated().sum()
```

```
0
```

- The dataset was sourced from www.Kaggle.com .
- The target feature defined as 'medv'.
- The dataset consists of 506 rows and 14 columns.
- No missing values (non-null) or duplicate values.

03

Data Preprocessing

VIF Score

	feature	vif_score
1	crim	1.713187
2	zn	2.465631
3	indus	3.877855
4	chas	1.096674
5	nox	4.469150
6	rm	1.947809
7	age	2.989948
8	dis	4.168578
9	rad	7.658316
10	tax	8.943301
11	ptratio	1.851448
12	black	1.325121
13	lstat	2.818045

```
# calculate VIF scores
from statsmodels.stats.outliers_influence import variance_inflation_factor as vif
from statsmodels.tools.tools import add_constant

X = add_constant(feature_boston_train)

vif_df = pd.DataFrame([vif(X.values, i)
                for i in range(X.shape[1])],
               index=X.columns).reset_index()
vif_df.columns = ['feature','vif_score']
vif_df = vif_df.loc[vif_df.feature!='const']
vif_df
```

Features with VIF values **greater than 4** will be considered for **dropping** as it can lead to a reduction in model efficiency.

Heatmap correlation

Based on the heatmap shown, we will drop the features "**indus**," "**tax**," and "**rad**" as they exhibit high correlations with other non-target features (>=0,70).

Recheck VIF Score

	feature	vif_score
1	crim	1.575252
2	zn	2.363346
3	chas	1.062361
4	rm	1.798318
5	age	2.780238
6	dis	3.586339
7	tax	2.381965
8	ptratio	1.578882
9	black	1.308853
10	lstat	2.742745

After dropping several features, there are no longer any VIF values > 4, indicating that the data is now **ready for modeling**.

04 Model and Evaluation

Modeling with Ridge after find the best alpha

	feature	coefficient
0	intercept	12.875802
1	crim	-0.066220
2	zn	0.034787
3	chas	1.841424
4	rm	4.885661
5	age	-0.014556
6	dis	-1.153516
7	tax	-0.003419
8	ptratio	-0.679689
9	black	0.012965
10	lstat	-0.535040

Interpretation

Intercept : The intercept is the constant value of the regression model when all other features are zero. In this case, the intercept is approximately 12.876.

crim: The coefficient for the "crim" feature is -0.066220. This means that each unit increase in "crim" (crime rate) will result in a decrease of approximately 0.066220 units in the target variable, with all other features remaining constant.

And so on for other features such as "zn", "chas", "rm", "age", "dis", "tax", "ptratio", "black", and "lstat." The coefficient for each feature provides information on how that feature contributes to the target variable.

Modeling with Lasso after find the best alpha

	feature	coefficient
0	intercept	25.823535
1	crim	-0.041070
2	zn	0.025267
3	chas	0.000000
4	rm	2.504144
5	age	0.022054
6	dis	-0.599318
7	tax	-0.002994
8	ptratio	-0.666247
9	black	0.011401
10	lstat	-0.712430

Interpretation

Intercept : The intercept is the constant value of the regression model when all other features are zero. In this case, the intercept is approximately 25.8235.

crim: The coefficient for the "crim" feature is -0.041070. This means that each unit increase in "crim" (crime rate) will result in a decrease of approximately 0.041070 units in the target variable, with all other features remaining constant.

And so on for other features such as "zn", "chas", "rm", "age", "dis", "tax", "ptratio", "black", and "lstat." The coefficient for each feature provides information on how that feature contributes to the target variable.

Diagnostic Study (Ridge)

```
from sklearn.metrics import r2_score

y_predict_train = ridge_best.predict(X_admit_train)

print('R-squared for training data is {}'.format(r2_score(y_admit_train, y_predict_train)))
```

```
R-squared for training data is 0.746036188189175
```

74.6% of the variability of the target variable has been **successfully** modelled with the existing features.

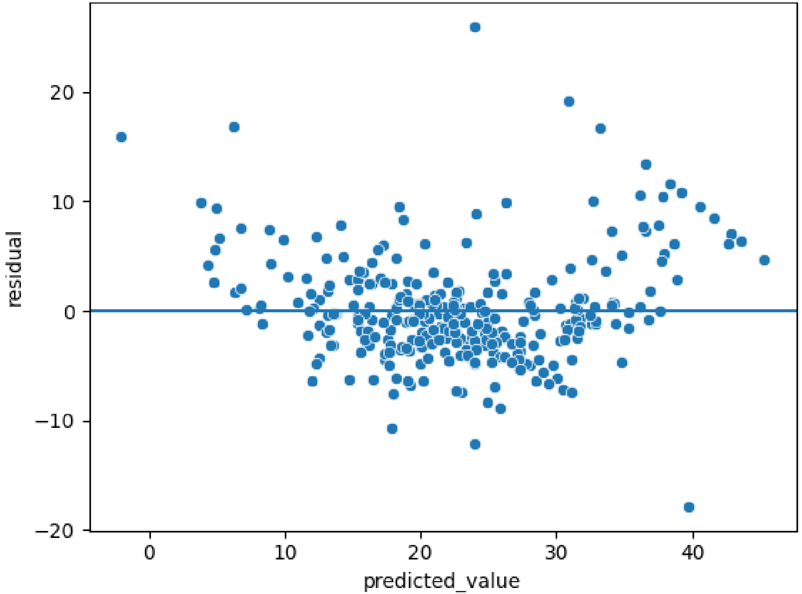
Diagnostic Study (Lasso)

```
from sklearn.metrics import r2_score

y_predict_train = Lasso_best.predict(X_admit_train)

print('R-squared for training data is {}'.format(r2_score(y_admit_train, y_predict_train)))
```

```
R-squared for training data is 0.7056813361226921
```

70.56% of the variability of the target variable has been **successfully** modelled with the existing features.

Plot Residual



Assumptions :

- **Linear relationship** : The horizontal line y = 0 does **not over-represent all** residual points. Because the residuals are closer to the centre only.
- **Variance stable** : NO. The variation is close to the middle, but at the ends of the scatter plot there are quite a lot of residuals that widen, especially at the top of y> 0.
- **Independent residuals** : OK. There is **no noticeable pattern** in nearby residuals.

Evaluation

In the regression model, we will perform evaluation using evaluation metrics such as **RMSE**, **MAE**, and **MAPE** .

RMSE (Root Mean Square Error) : It measures the average of the squared differences between actual and predicted values. RMSE emphasizes larger errors, making it sensitive to outliers.

MAE (Mean Absolute Error) : It calculates the average of the absolute differences between actual and predicted values. MAE considers all errors equally and is more robust to outliers.

MAPE (Mean Absolute Percentage Error) : It computes the average percentage difference between actual and predicted values. MAPE provides insights into the average relative error in percentage terms.

Evaluation

Matrix Evaluation	Ridge		Lasso	
	Training	Testing	Training	Testing
RMSE	4.80	5.21	5.17	5.12
MAE	3.38	3.30	3.68	3.39
MAPE	**16.92%**	**18.02%**	**17.49%**	**17.85%**

Interpretation :

- Based on the results of the Training and Testing error check above, this model can be said to be **quite good** with the **MAPE** value of both which is still **below 20%**. So it can be concluded that this model is **not underfitting or overfitting**.

Thank You

sucirahma.srn@gmail.com

https://www.linkedin.com/in/sucisrn/

https://github.com/eseren