Regression and Regularization

Melbourne House Price Prediction Using Regression & Regularization

Objectives of Analysis:

1. The analysis will be focused on Prediction.
2. To use the tools and techniques to train a few linear regressions.
3. To implement regularization and compare results with Simple Linear Regression, simple linear regression with MinMax scaling and Cross Validation.
4. To evaluate model using r2 score.
5. To tune hyperparameters to find best parameters for model.
6. To select the regression model that gives best result.
7. To report present findings, insights, and next steps.

Dataset Description:

Dataset: Melbourne Housing Dataset

Link to dataset: https://www.kaggle.com/anthonypino/melbourne-housing-market

Melbourne housing dataset contains 21 attributes and 34857 rows. Target column is 'Price'. Description and data types of attributes are given below:

#	Attribute	Description	Data Type
1	Suburb	Suburb Name	object
2	Address	Address of house	object
3	Rooms	Number of rooms	int64
4	Type	Type of house: br - bedroom(s); h - house, cottage, villa, semi, terrace; u - unit, duplex; t - townhouse; dev site - development site; o res - other residential.	object
5	Price	Price of house in Australian dollars	float64
6	Method	S - property sold; SP - property sold prior; PI - property passed in; PN - sold prior not disclosed; SN - sold not disclosed; NB - no bid; VB - vendor bid; W - withdrawn prior to auction; SA - sold after auction; SS - sold after auction price not disclosed. N/A - price or highest bid not available.	object
7	SellerG	Real Estate Agent	object
8	Date	Date sold	object

9	Distance	Distance from CBD in Kilometres	float64
10	Postcode	Postcode	float64
11	Bedroom2	Scraped # of Bedrooms (from different source)	float64
12	Bathroom	Number of Bathrooms	float64
13	Car	Number of carspots	float64
14	Landsize	Land Size in Metres	float64
15	BuildingArea	Building Size in Metres	float64
16	YearBuilt	Year the house was built	float64
17	CouncilArea	Governing council for the area	object
18	Lattitude	Self-explanatory	float64
19	Longtitude	Self-explanatory	float64
20	Regionname	General Region (West, North West, North, North east …etc)	object
21	Propertycount	Number of properties that exist in the suburb.	float64

Data Cleaning and Feature Engineering:

I have not much focused on Data Cleaning and Feature Engineering because of time constraint.

In a simple way, I have dropped some categorical and very less useful features. Then I dropped rows where there is a null value in any column. Dummy encoding is performed to convert categorical columns to numeric form. A new feature 'SellYear' is created from 'Date' feature by extracting year from Date feature.

Variations of linear regression models and Results:

Below table gives regression models used in analysis and r2 score. I have used r2 score as model evaluation metric.

Regression Model	R2 score
Simple Linear Regression	0.67562885178392
Simple Linear Regression with MinMax Scaling	0.6756288517839144
Simple Linear Regression with MinMax Scaling & Cross Validation	0.6706223090110537
Lasso Regression	0.6706100396142524
Lasso Regression with Polynomial Features	0.8188907503683591
Ridge Regression	0.6583384462216006
Ridge Regression with Polynomial Features	0.8338163531066394
ElasticNet Regression	-0.40802889406223236
ElasticNet Regression with Polynomial Features	0.8050397351428221

```
In [49]: r2score_dict

Out[49]: {'Simple Linear Regression': 0.67562885178392,
          'Simple Linear Regression with MinMax Scaling': 0.6756288517839144,
          'Simple Linear Regression with MinMax Scaling & Cross Validation': 0.6706223090110537,
          'Lasso Regression': 0.6706100396142524,
          'Lasso Regression with Polynomial Features': 0.8188907503683591,
          'Ridge Regression': 0.6583384462216006,
          'Ridge Regression with Polynomial Features': 0.8338163531066394,
          'ElasticNet Regression': -0.40802889406223236,
          'ElasticNet Regression with Polynomial Features': 0.8050397351428221}
```

Key Findings and Conclusion:

1. Wherever the Cross Validation is used the R2 score is mean of R2 scores hence we can see that ElasticNet regression gives negative score. That is also because of hyperparameter tuning and score went negative and very low for some hyperparameter combinations.
2. Adding simple regularization with simple linear regression model has not improved R2 score much.
3. Adding regularization with Polynomial Features has improved R2 score very well.
4. From comparison of R2 scores it is clear that **Ridge Regression with Polynomial Features is best model** for predicting house price for Melbourne Housing Dataset.

Result Discussion and Further Steps:

1. In this assignment, I have not focussed on data preparation much because of time constraint. Hence by dropping all rows with null values the original data has reduced to around 8800 rows from 37000 rows. The model performance can be improved by appropriate data preparation and null value treatment further.
2. I have used dummy encoding for all categorical columns however One Hot encoding can be used wherever suitable.
3. I have used only MinMax scaling. Other scaling techniques can be used to observer impact on model performance.
4. The same data can be modelled using LassoCV, RidgeCV and ElasticNetCV to check if performance improvement further.
5. I have tuned hyperparameters only to limited range however hyperparameter range can be extended and more best parameters can be found.
6. We can also use GridSearchCV and RandomSearchCV for finding best parameters.
7. The models can be evaluated on different error metrics to understand effect of different techniques on model performance.