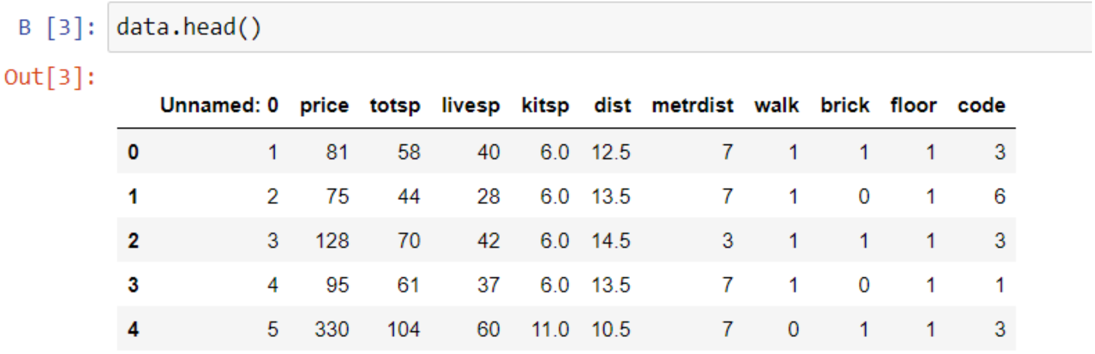
Machine learning for prediction of apartment prices in Moscow

Used data

The data have been downloaded from https://www.kaggle.com/hugoncosta/price-of-flats-in-moscow. It covers a dataset from 2017 with over 2000 apartment prices with their attributes.
The head of the dataframe of this dataset looks as follows:



```
В [3]: data.head()
```

Out[3]:

	Unnamed: 0	price	totsp	livesp	kitsp	dist	metrdist	walk	brick	floor	code
0	1	81	58	40	6.0	12.5	7	1	1	1	3
1	2	75	44	28	6.0	13.5	7	1	0	1	6
2	3	128	70	42	6.0	14.5	3	1	1	1	3
3	4	95	61	37	6.0	13.5	7	1	0	1	1
4	5	330	104	60	11.0	10.5	7	0	1	1	3

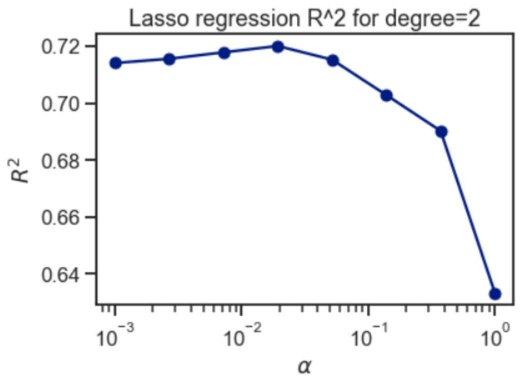
Attribute target variable and features

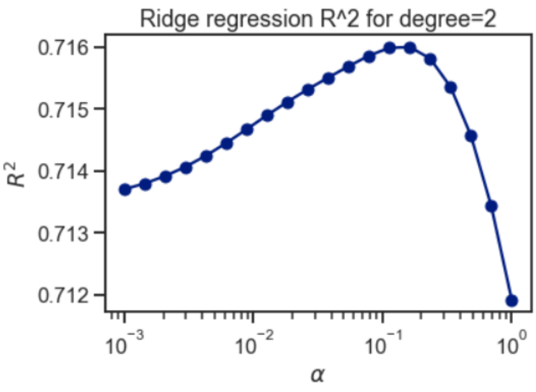
Target variable: “price” - Price in 1000 USD for an apartment.

Features: “totsp” – total space in sq meters, “livesp” – living space in sq meters, “kitsp” – kitchen stapce in sq meter, “dist” – distance from the city center in kilometers, “metrdist” – distance to the closest metro station in minutes, “walk” – walking distance to the metro (one-hot encoded, 1 – it’s close to a metro station, 0 – it is not), “brick” – the material of the house (one-hot encoded, 1 – concrete, monolithic, 0 – otherwise), “floor” – one-hot encoded floor number (1 – the apartment is on the first or the last floor of the house, 0 -otherwise).

Data exploration

- The columns “Unnamed: 0” and “code” were dropped from the feature list as they donot influence prediction.
- Before doing training-test-split procedures and determine the accuracy of each regression model. The cross-validation has been performed to find best alpha parameters for Lasso and Ridge regression. The figures below represent alpha value for





- The alpha parameters have been extracted, then five regression models will be analyzed: Linear regression, Polynomial featureswith degree 2 and 3, Lasso regression with polynomial features degree=2, and, finally, Ridge regression with polynomial features degree=2.

Summary of training and test data

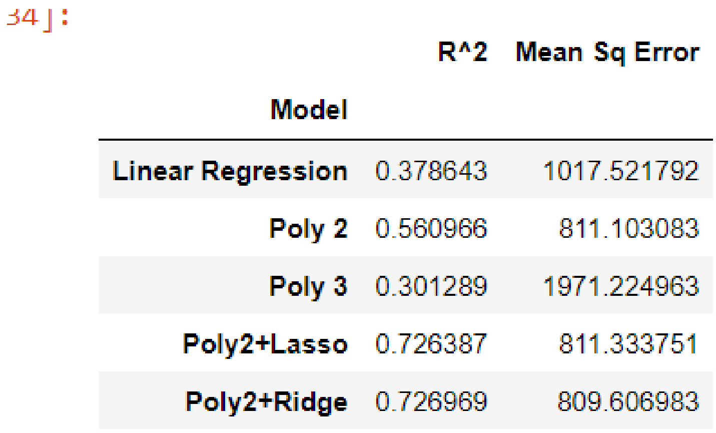
The X and Y data have been splitted as 70 % for the training set and 30 % for the test set.

As discussed before, five models were used:

1) Linear regression
2) Linear regression with polynomial feature transformation (degree=2)
3) Linear regression with polynomial feature transformation (degree=3)
4) Ridge regression with polynomial feature transformation (degree=2)
5) Lasso regression with polynomial feature transformation (degree=2)

The X data were scaled for all these models with a StandardScaler.

The predictability of these models has been characterized based on R^2 score and the mean squared error. The following dataframe represents the results:



34]:

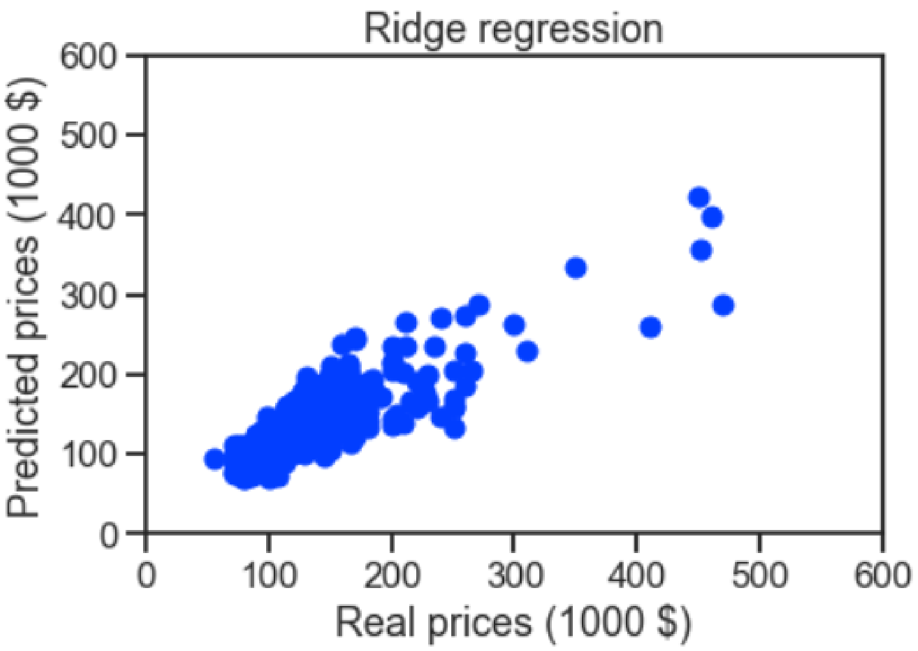
Model	R^2	Mean Sq Error
Linear Regression	0.378643	1017.521792
Poly 2	0.560966	811.103083
Poly 3	0.301289	1971.224963
Poly2+Lasso	0.726387	811.333751
Poly2+Ridge	0.726969	809.606983

As you can see, linear regression doesn’t perform well as R^2 equal to 0.38, meaning that the model is very much biased and underfit. Polynomial features of degree 2 improve a lot the score. However, Polynomial features of degree 3 seem to overfit the data too much and R^2 decreases back.

Finally, Lasso regression and Ridge regression perform equally well, reaching the R^2 of 0.73 on the test data. I choose Ridge regression as it yields slightly lower mean squared error.

Main result

I plot the predicted values of apartment price for the test dataset versus the real known values (shown below).



Even though there is a certain fluctuation, the Ridge model performs quite well. The predicted prices go along with real prices, indicating that the model works rather correctly especially in the range between 0 and 200, where we have the largest part of data.

I also determine features, that affect most the apartment price.

The parameters that are very important to boost the apartment price:

Feature	Coefficient
dist	3,008536
brick^2	5,300965
brick	5,300965
kitsp	7,930983
walk	
brick	21,37667

The parameters that are very negative for the apartment price:

Feature	Coefficient
floor^2	-22,4425
floor	-22,4425
walk^2	-10,0171
walk	-10,0171
livesp	-4,81171

Conclusion

I have performed the optimization of the model. The Ridge regression seem to perform rather well with R^2 score of 0.73. Based on this model we can predict the appratment price in Moscow and see the most important features to reduce or increase the apartment price.

I believe the model can be improved by adding extra features that will improve the model accuracy and will help avoid a significant value of the irreducible error.