Documentation

Group 1: Housing Prices

Klement Florian
Minaeva Anastasiia
Pollek Patrick
Trush Maria

Group 1: Housing Prices

Klement Florian, Minaeva Anastasiia, Pollek Patrick, Trush Maria

Workflow 3

Phase 1: Project/Data Understanding 4

Phase 2: Data Preparation/ Modeling 5

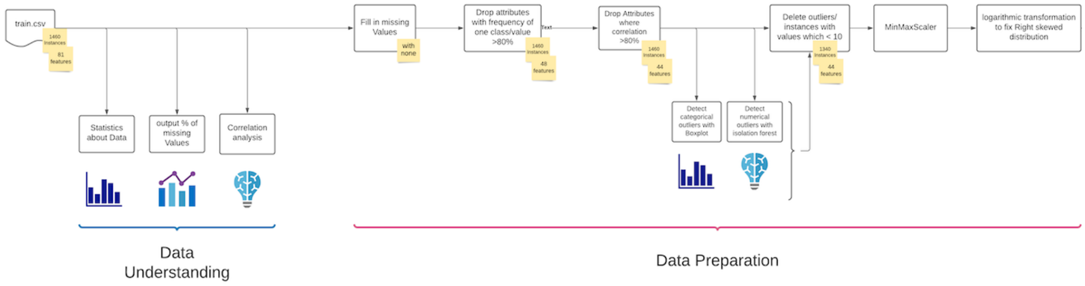
- General description of the dataset. 5
- Data Understanding & Preparation Process. 5
- Data Modeling Process: 6
  - Regression 6
  - Data Modeling: Clustering 10
  - Data Modeling: Classification 12

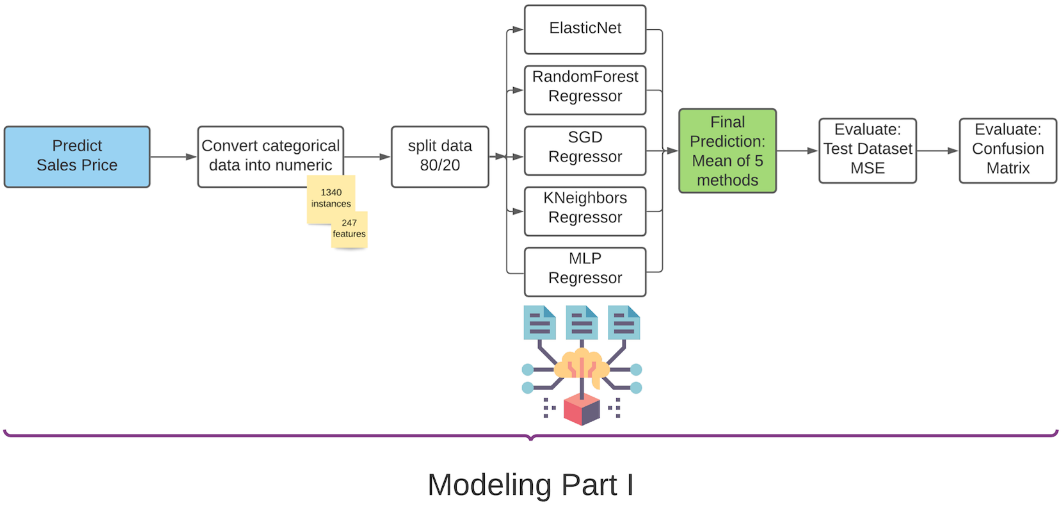
Phase 3: Evaluation 15

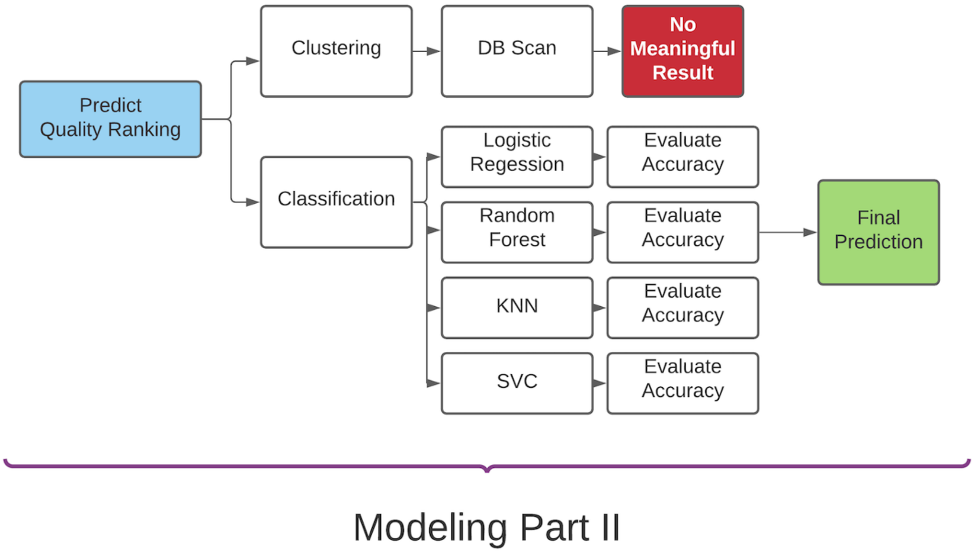
Group 1: Housing Prices

Klement Florian, Minaeva Anastasiia, Pollek Patrick, Trush Maria

Workflow


train.csv
1460 instances
81 features
Statistics about Data
output % of missing Values
Correlation analysis
Data Understanding
Fill in missing Values
with none
Drop attributes with frequency of one class/value >80%
Text
1460 instances
48 features
Drop Attributes where correlation >80%
1460 instances
44 features
Detect categorical outliers with Boxplot
Detect numerical outliers with isolation forest
Delete outliers/ instances with values which < 10
1340 instances
44 features
MinMaxScaler
logarithmic transformation to fix Right skewed distribution
Data Preparation





Group 1: Housing Prices

Klement Florian, Minaeva Anastasiia, Pollek Patrick, Trush Maria

Phase 1: Project/Data Understanding

The domain. The dataset contains sales transactions for the houses in a certain area with detailed characteristics of each house.

The Problem. We have decided to concentrate on two potential problems:

1. Predict a home price for a potential home buyer or a seller. A person enters several parameters (such as size, number of bedrooms, location) and a model produces a price estimate. *Potentially, a regression problem.*
2. Determine if houses can be broken down into classes based on certain parameters (quality, location, etc). A potential buyer or seller can then determine a desired class and filter for it. *Potentially, a clustering or a classification problem*.

Group 1: Housing Prices

Klement Florian, Minaeva Anastasiia, Pollek Patrick, Trush Maria

Phase 2: Data Preparation/ Modeling

General description of the dataset.

We have a dataset with 81 different attributes, which, however, contains just 1 460 instances. 43 out of 81 attributes are objects/categorical attributes (for example, different house features such as garage type, quality of basement, access to alley), the rest are integers or floats. The dataset covers a period from 2006 to 2010.

Data Understanding & Preparation Process.

We went through several steps during the data understanding stage: (i) scanned for Missing Values, (ii) looked at correlation between the attributes, (iii) visualized the data and considered its distribution, (iv) considered scaling, and (iv) scanned the data for outliers.

As a result, we have uncovered the following:

(i) Missing values. Certain attributes contained a lot of missing values (more than 90% of all values). However, in many cases it was just an indication that a house did not have a certain feature (no pool, garage finish etc.)

(ii) Correlation. In contrast to our expectations, only 4 pairs of attributes had a correlation of over 80%. Sales price (our target attribute) has the highest correlation with the overall quality and the living area.

(iii) Visualization. When looking at histograms, it became obvious that many categorical attributes contain more than 80% of the same value. For example, almost all houses had paved road access and all houses had heating. Sales price, a target parameter, has a skewed distribution.

(iv) Attributes reduction. As described above, we dropped the attributes in cases where 80% of the values for an attribute were the same. Thereby we were able to reduce the number of attributes from 81 to 44.

(v) Scaling. For certain numerical attributes (such as linear feet, square feet, number of bedrooms) scaling would be required to make sure larger values do not have a disproportionate effect on the modeling results. Therefore, we used Min-Max scaling.

(vi) Outliers. Visualization via boxplots has identified a fair number of outliers. It is an issue for us, given that the dataset is already quite small, and we will have to further reduce it by splitting it into Test and Training sets. Therefore, we have also considered methods for an automatic outlier detection (Isolation Forest and One Class SVM). In case of automated methods, certain hyper parameters must be chosen manually. For categorical attributes, we have decided to drop instances which are found less than 10 times.

Data Modeling Process:

Regression

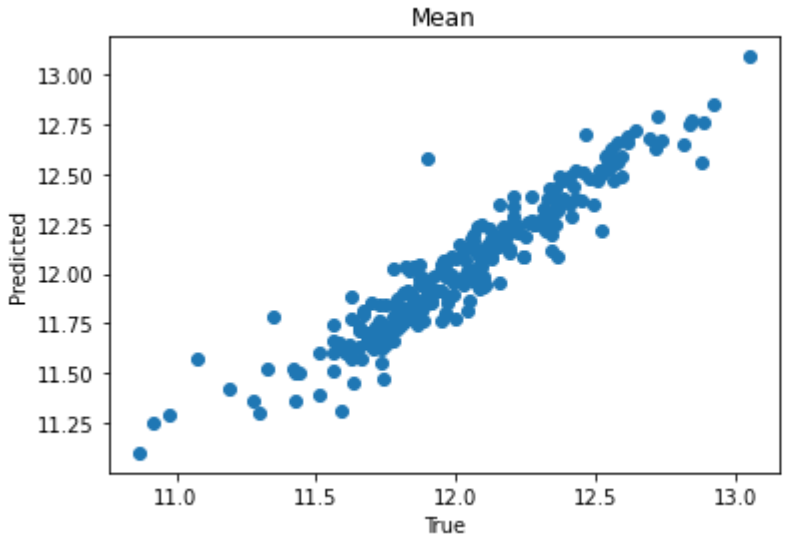
We will use an ensemble of different models to predict the price of the houses.
We will report MSE of the logarithmic transformed Data.

The models we trained:

ElasticNet :

ElasticNet(alpha=0.001, l1_ratio=0.32, max_iter=100000)

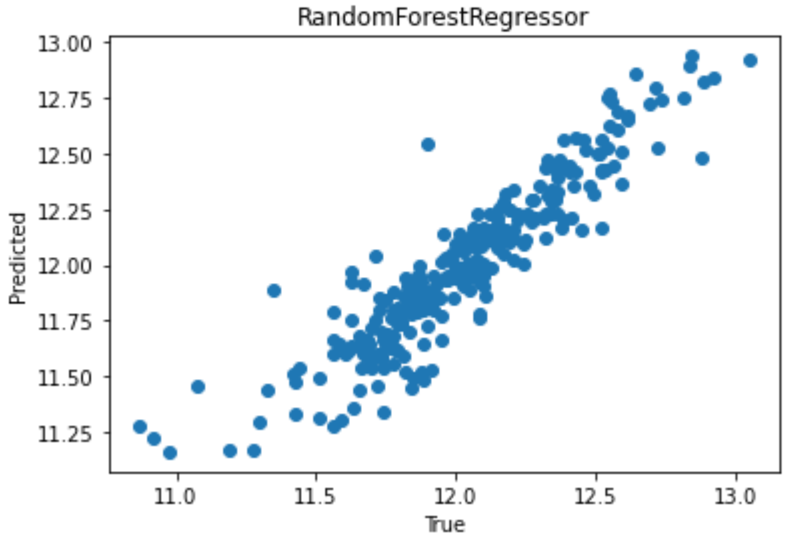
	ElasticNet
MSE Train	0.0084
MSE Test	0.0139



RandomForestRegressor:

RandomForestRegressor(max_features='auto', n_estimators=200, n_jobs=-1)

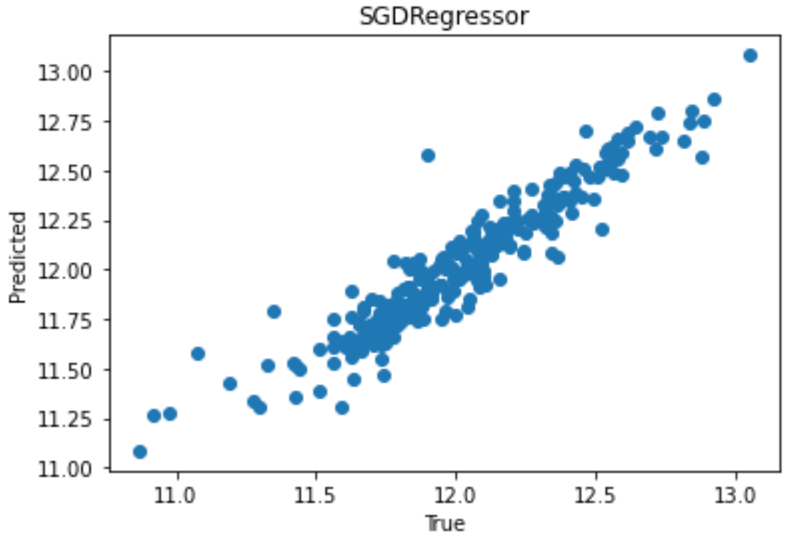
	RandomForestRegressor
MSE Train	0.002295
MSE Test	0.021447



SGDRegressor:

SGDRegressor(max_iter=10000, tol=1e-20,n_iter_no_change=200)

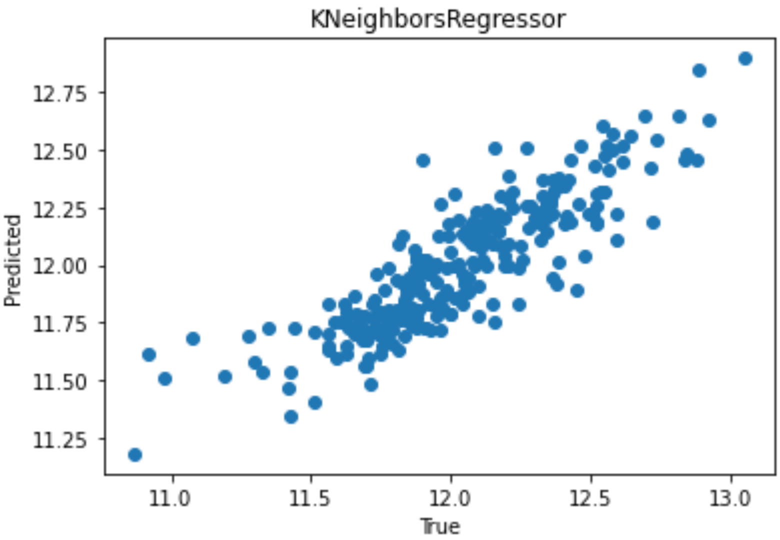
	SGDRegressor
MSE Train	0.00844
MSE Test	0.014587



KNeighborsRegressor:

KNeighborsRegressor(n_neighbors=7)

	KNeighborsRegressor
MSE Train	0.0300146398
MSE Test	0.0341187



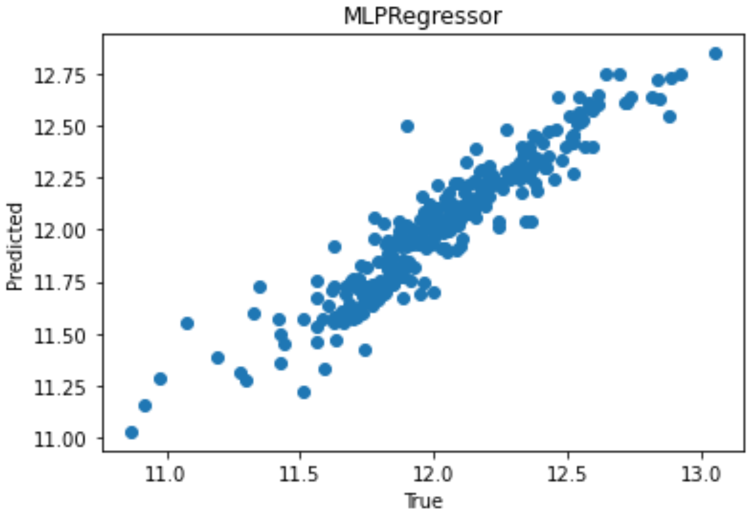
MLPRegressor (NN):

MLPRegressor(random_state=1, max_iter=10000,hidden_layer_sizes=[256,256,256],activation='tanh',tol=1e-8,n_iter_no_change=10,verbose=1,alpha=1,batch_size=20)

	MLPRegressor
MSE Train	0.013460109157314326
MSE Test	0.015903740902729474

Group 1: Housing Prices

Klement Florian, Minaeva Anastasiia, Pollek Patrick, Trush Maria

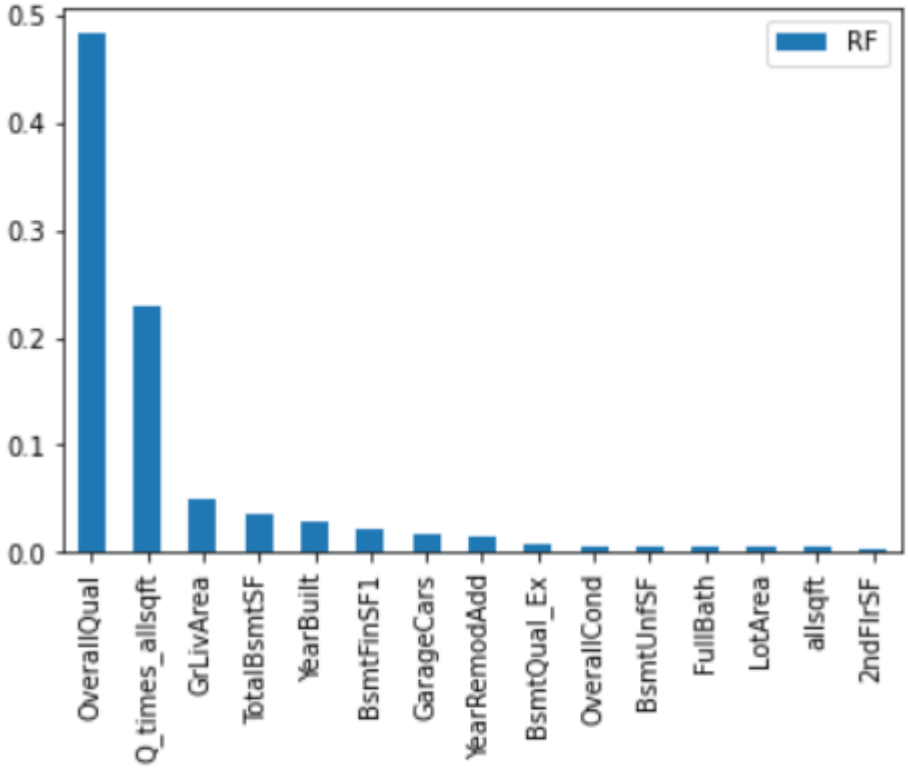


<u>All Hyperparameters of the Models have been selected via Cross Validation to ensure the best generalisation.</u>

Data Modeling: Clustering

We had intended to improve the predictive quality of our regression and subsequent clustering through the help of unsupervised learning, in order to potentially use the created categories as an additional predictive variable.
Unfortunately this did not prove to be fruitful and of the most important features, including the newly generated ones (listed below), none proved to carry potential for the grouping of our data.

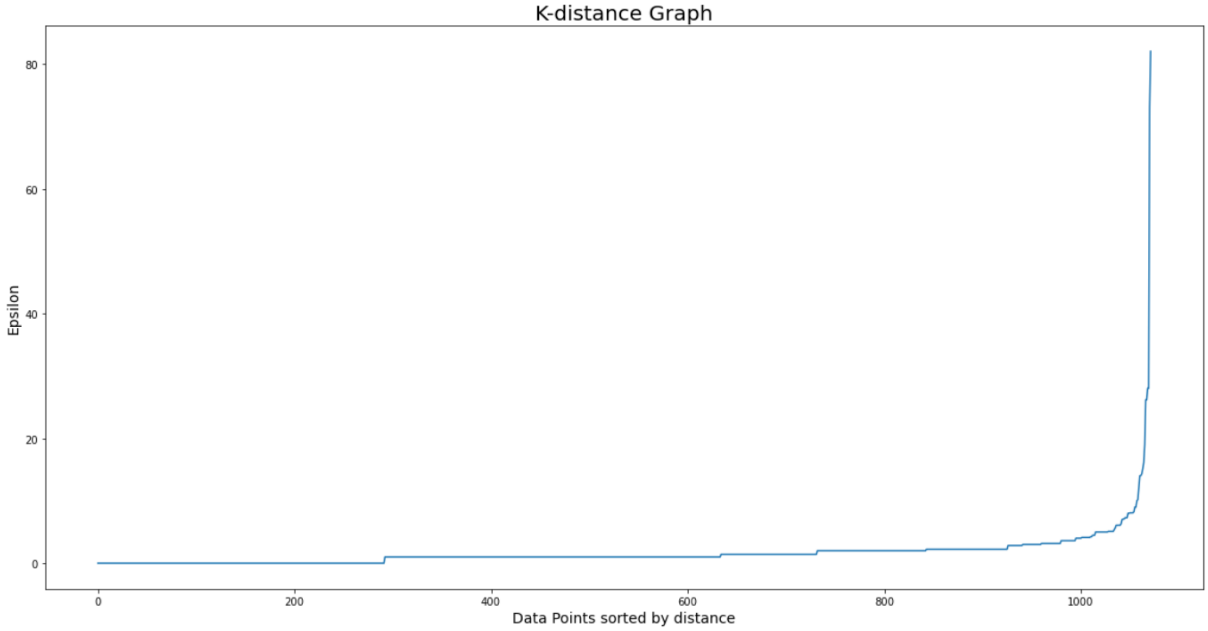
```
data['allsqft']=data['TotalBsmtSF']+data['2ndFlrSF']+data['OpenPorchSF']+data['GrLivArea']+
data['MasVnrArea']+data['LotArea'] data['Q_times_allsqft']=data['allsqft']*data['OverallQual']
```



```
clustering_set=train[['GrLivArea','OverallQual']]
DB = DBSCAN(eps=9, min_samples=10)
DB.fit(clustering_set)
```

```
Estimated no. of clusters: 18
Estimated no. of noise points: 193
Silhouette Coefficient: 0.328
```

The best clustering attempt was optimized with the help of a KNN, but still displayed a very weak silhouette coefficient score and does not seem adequate enough for improving the other predictive measures.


K-distance Graph
Epsilon
80
60
40
20
0
0
200
400
600
800
1000
Data Points sorted by distance

Group 1: Housing Prices

Klement Florian, Minaeva Anastasiia, Pollek Patrick, Trush Maria

Data Modeling: Classification

Our second task we set out to address during project understanding was a multiclass classification on the 'OverallQual' attribute, an attribute which also proved it’s strong impact in the regression analysis. It was the single most important feature alone, but also carried a lot of weight within our created features.

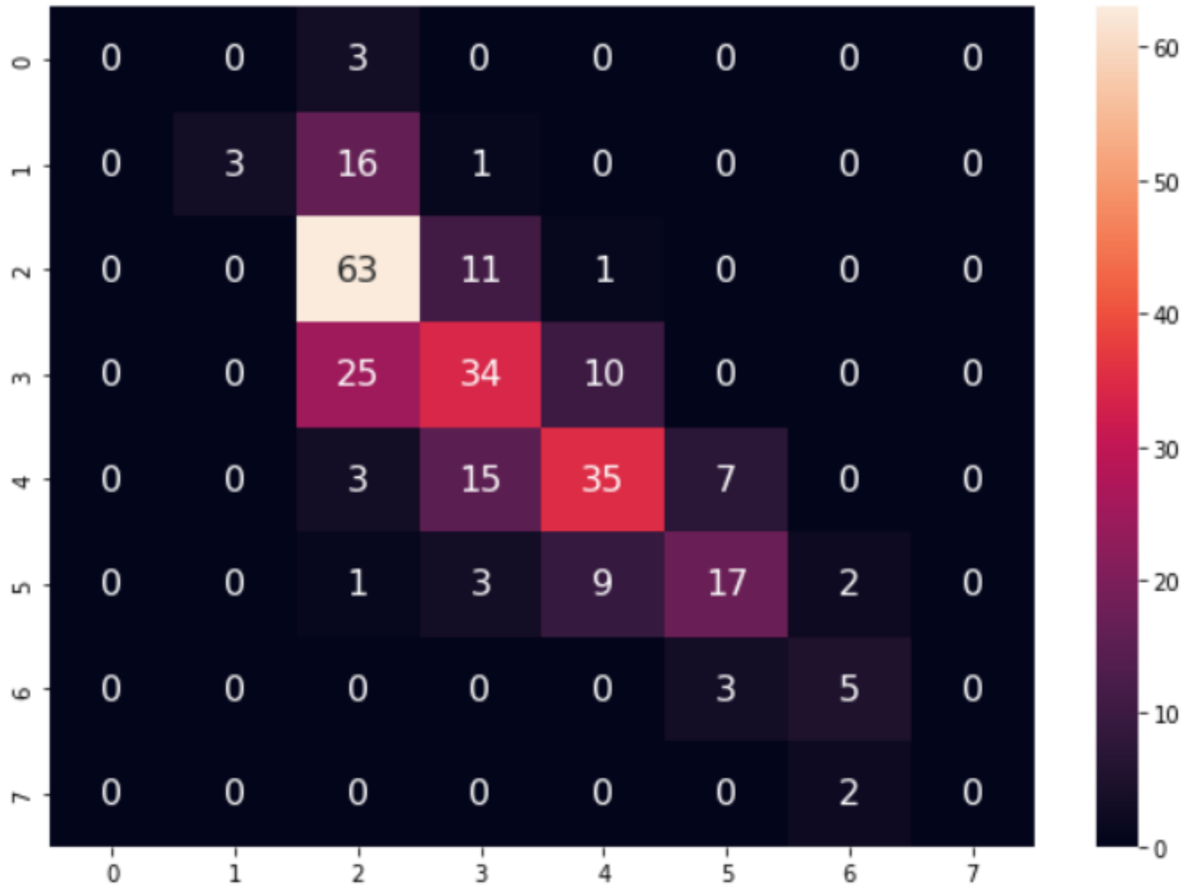
For this we chose a selection of models that are better equipped for this and can handle this scenario natively according to a One-against-the-Rest strategy, like Logistic Regression classifiers and Random Forest classifiers, but also other less intuitive approaches as a point of comparison, like a KNN or a Support Vector Machine classifier, that operates strictly pairwise.

In order to retain the underlying, heavily skewed distribution of values during the train-test split, stratified sampling was implemented.

While the pure accuracy of the predictions of these models for the adherence to one specific class is not wholly sufficient, ranging among the classifiers between 51 to 58 percent, when looking at the corresponding confusion matrices, the classifier which proved to provide the best prediction, the Random Forest classifier could achieve satisfactory results.

The amount of predictions falling within a range of a correct assessment and one category above or below is 95.54 %.

```
([ 0,  0,  3,  0,  0,  0,  0,  0],
 [ 0,  3, 16,  1,  0,  0,  0,  0],
 [ 0,  0, 62, 13,  0,  0,  0,  0],
 [ 0,  0, 25, 34, 10,  0,  0,  0],
 [ 0,  0,  4, 13, 36,  7,  0,  0],
 [ 0,  0,  1,  3,  9, 17,  2,  0],
 [ 0,  0,  0,  0,  0,  4,  4,  0],
 [ 0,  0,  0,  0,  0,  0,  2,  0])
```



The results for all the classifiers, their parameters and their confusion matrices are as follows:

LogisticRegression(C=0.01, penalty='l2', solver='newton-cg')

accuracy : 0.5541044776119403

```
    ([ 0,  0,  3,  0,  0,  0,  0,  0],
     [ 0,  5, 13,  2,  0,  0,  0,  0],
     [ 0,  3, 52, 19,  1,  0,  0,  0],
     [ 0,  2, 21, 27, 19,  0,  0,  0],
     [ 0,  0,  2, 18, 34,  6,  0,  0],
     [ 0,  0,  1,  1, 13, 15,  2,  0],
     [ 0,  0,  0,  0,  0,  3,  5,  0],
     [ 0,  0,  0,  0,  0,  0,  2,  0])
```

Group 1: Housing Prices

Klement Florian, Minaeva Anastasiia, Pollek Patrick, Trush Maria

```
RandomForestClassifier(criterion='gini', max_depth=13, max_features='auto',
min_samples_split=4, n_estimators=500)

accuracy : 0.5836431226765799

    ([ 0,  0,  3,  0,  0,  0,  0,  0],
     [ 0,  3, 16,  1,  0,  0,  0,  0],
     [ 0,  0, 63, 11,  1,  0,  0,  0],
     [ 0,  0, 25, 34, 10,  0,  0,  0],
     [ 0,  0,  3, 15, 35,  7,  0,  0],
     [ 0,  0,  1,  3,  9, 17,  2,  0],
     [ 0,  0,  0,  0,  0,  3,  5,  0],
     [ 0,  0,  0,  0,  0,  0,  2,  0])
```

```
KNeighborsClassifier(algorithm='auto', n_neighbors=20, p=1)

accuracy : 0.516728624535316

    ([ 0,  0,  3,  0,  0,  0,  0,  0],
     [ 0,  3, 15,  2,  0,  0,  0,  0],
     [ 0,  2, 64,  8,  1,  0,  0,  0],
     [ 0,  1, 32, 22, 14,  0,  0,  0],
     [ 0,  0,  5, 15, 37,  3,  0,  0],
     [ 0,  0,  2,  2, 18,  9,  1,  0],
     [ 0,  0,  0,  0,  1,  3,  4,  0],
     [ 0,  0,  0,  0,  0,  0,  2,  0])
```

```
SVC(C=1.0, gamma=0.001, kernel='rbf')
accuracy : 0.5130111524163569
    ([ 0,  0,  3,  0,  0,  0,  0,  0],
     [ 0,  5, 13,  2,  0,  0,  0,  0],
     [ 0,  3, 52, 19,  1,  0,  0,  0],
     [ 0,  2, 21, 27, 19,  0,  0,  0],
     [ 0,  0,  2, 18, 34,  6,  0,  0],
     [ 0,  0,  1,  1, 13, 15,  2,  0],
     [ 0,  0,  0,  0,  0,  3,  5,  0],
     [ 0,  0,  0,  0,  0,  0,  2,  0])
```

Phase 3: Evaluation

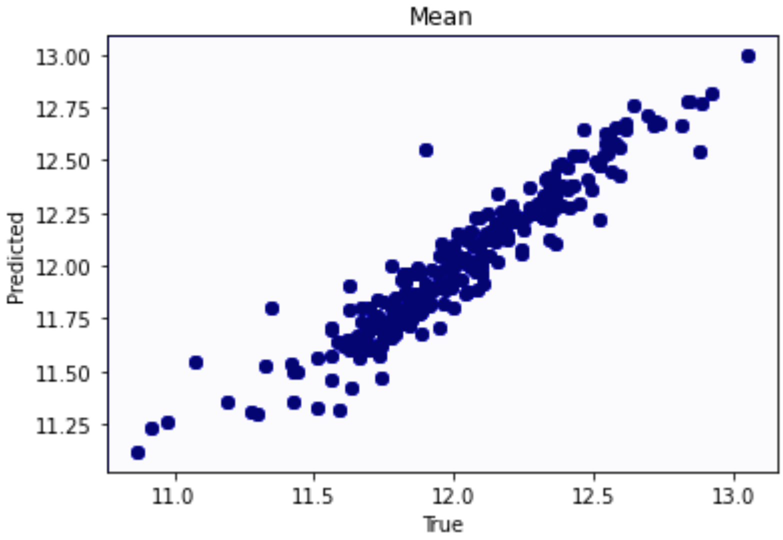
Both of the tasks, the prediction of house prices according to their features, as well as the classification of houses within certain quality categories proved to yield satisfactory results, capable of serving the problems set out to be solved in project understanding and provide support for adequate quality for decision making.

The final ensemble of the Regressors yields:

Ensemble

This now calculates the mean of all predictions.

	Ensemble
ME Test	20608
MSE Test	0.013460100278418997


Mean
Predicted
True