House Prices: Advanced Regression Technique

Members:
Kurt Maxwell Kusterer
Kanishk Gupta
Carlos Montenegro

0. Summary: Purpose

- The aim of this project is to identify the undervalued properties having a maximum budget of EUR 70’000
- Have a key focus on all categories of customers.
- Make consumer aware about important facets that determine the real worth of a property.
- To ensure that Customers get best return on investment by making highest possible accurate predictions.
- To ensure that customers get the best investment price for the property along with all necessary utilities.

0. Summary: problem

Needs

- Prediction of sale price using different machine learning models
- Consideration of only highly influential values that might affect purchase decision.
- Use top machine learning algorithms to make high accurate predictions.

Data

- We divided dataset in train, test
- Description of all types of data i.e numerical and categorical data.
- All the preprocessing steps used to clean the necessary data.
- Calculating mean for numeric variables and mode for categorical variables
- Final variables selected for machine learning models

Topics

Data Summary

Methodology

Results

1. Data Summary: Missings

Methodology

Data cleaning on missing values can be broken down into three categories according to 'Little and Rubin 1987', Ignoring and discarding data,Parameter Estimation and Imputation

•Let's consider numerical values first, in this case we have considered numerical values which are missing to be 0, based on the other categorical variables within the table indicating this.

•Categorical values. In that instances where values of NA occurred, were understood to be instances in which these particular attributes did not exist within an observation. These instances were replaced by 'None'.

List of missing values

```
[1] "LotFrontage has 259 number of missing values"
[1] "Alley has 1369 number of missing values"
[1] "MasVnrType has 8 number of missing values"
[1] "MasVnrArea has 8 number of missing values"
[1] "BsmtQual has 37 number of missing values"
[1] "BsmtCond has 37 number of missing values"
[1] "BsmtExposure has 38 number of missing values"
[1] "BsmtFinType1 has 37 number of missing values"
[1] "BsmtFinType2 has 38 number of missing values"
[1] "Electrical has 1 number of missing values"
[1] "FireplaceQu has 690 number of missing values"
[1] "GarageType has 81 number of missing values"
[1] "GarageYrBlt has 81 number of missing values"
[1] "GarageFinish has 81 number of missing values"
[1] "GarageQual has 81 number of missing values"
[1] "GarageCond has 81 number of missing values"
[1] "PoolQC has 1453 number of missing values"
[1] "Fence has 1179 number of missing values"
[1] "MiscFeature has 1406 number of missing values"
```

1. Data Summary: categorical features

One-hot encoding

For categorical variables dummy encoding was implemented. A dummy variable is a numeric stand in for a qualitative fact or a logical proposition(Susan Gravagila and Asha Sharma,1998).

•The most important variables are selected by means of the function corcat using the ‘lsr’ package.

•Then from these the most important categorical variables will be then dummy encode for use in the models.

Below we see the 43 Original Categorical variables :

Example

Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

```
There are 43 categorical variables
```

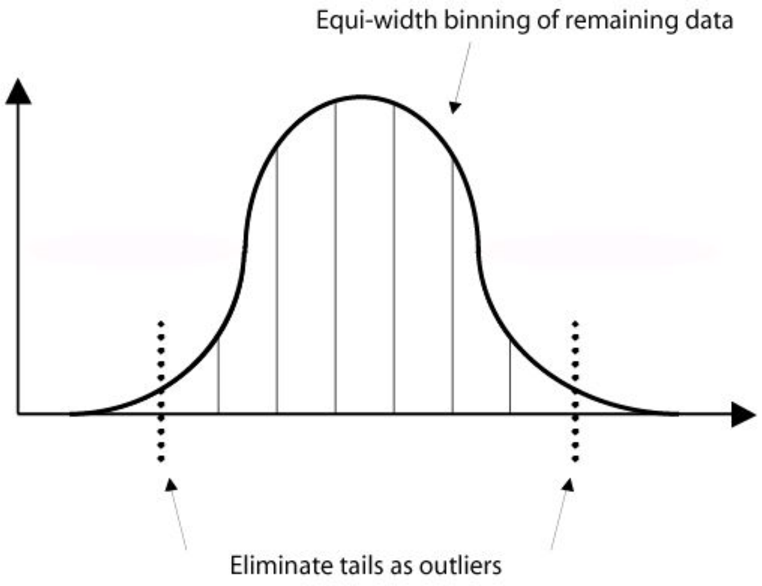
'MSZoning' 'Street' 'Alley' 'LotShape' 'LandContour' 'Utilities' 'LotConfig' 'LandSlope' 'Neighborhood' 'Condition1' 'Condition2' 'BldgType' 'HouseStyle' 'RoofStyle' 'RoofMatl' 'Exterior1st' 'Exterior2nd' 'MasVnrType' 'ExterQual' 'ExterCond' 'Foundation' 'BsmtQual' 'BsmtCond' 'BsmtExposure' 'BsmtFinType1' 'BsmtFinType2' 'Heating' 'HeatingQC' 'CentralAir' 'Electrical' 'KitchenQual' 'Functional' 'FireplaceQu' 'GarageType' 'GarageFinish' 'GarageQual' 'GarageCond' 'PavedDrive' 'PoolQC' 'Fence' 'MiscFeature' 'SaleType' 'SaleCondition'

1. Data Summary: numerical variables

Winzorization

```
install.packages("robustHD")
library(robustHD)

train_clean$GrLivArea <- winsorize(train_clean$GrLivArea,probs = c(0.0, 0.95))
train_clean$GarageArea <- winsorize(train_clean$GarageArea,probs = c(0.0, 0.95))
train_clean$TotalBsmtSF <- winsorize(train_clean$TotalBsmtSF,probs = c(0.0, 0.95))
train_clean$X1stFlrSF <- winsorize(train_clean$X1stFlrSF,probs = c(0.0, 0.95))
train_clean$YearBuilt <- winsorize(train_clean$YearBuilt,probs = c(0.0, 0.95))
train_clean$YearRemodAdd <- winsorize(train_clean$YearRemodAdd,probs = c(0.05, 0.95))

test_clean$GrLivArea <- winsorize(test_clean$GrLivArea,probs = c(0.0, 0.95))
test_clean$GarageArea <- winsorize(test_clean$GarageArea,probs = c(0.0, 0.95))
test_clean$TotalBsmtSF <- winsorize(test_clean$TotalBsmtSF,probs = c(0.0, 0.95))
test_clean$X1stFlrSF <- winsorize(test_clean$X1stFlrSF,probs = c(0.0, 0.95))
test_clean$YearBuilt <- winsorize(test_clean$YearBuilt,probs = c(0.0, 0.95))
test_clean$YearRemodAdd <- winsorize(test_clean$YearRemodAdd,probs = c(0.0, 0.95))
```

Example




```
There are 37 numeric variables
```

'MSSubClass' 'LotFrontage' 'LotArea' 'OverallQual' 'OverallCond' 'YearBuilt' 'YearRemodAdd' 'MasVnrArea' 'BsmtFinSF1' 'BsmtFinSF2' 'BsmtUnfSF' 'TotalBsmtSF' 'X1stFlrSF' 'X2ndFlrSF' 'LowQualFinSF' 'GrLivArea' 'BsmtFullBath' 'BsmtHalfBath' 'FullBath' 'HalfBath' 'BedroomAbvGr' 'KitchenAbvGr' 'TotRmsAbvGrd' 'Fireplaces' 'GarageYrBlt' 'GarageCars' 'GarageArea' 'WoodDeckSF' 'OpenPorchSF' 'EnclosedPorch' 'X3SsnPorch' 'ScreenPorch' 'PoolArea' 'MiscVal' 'MoSold' 'YrSold' 'SalePrice'

1. Data Summary: Selected categorical features

ETS-squared

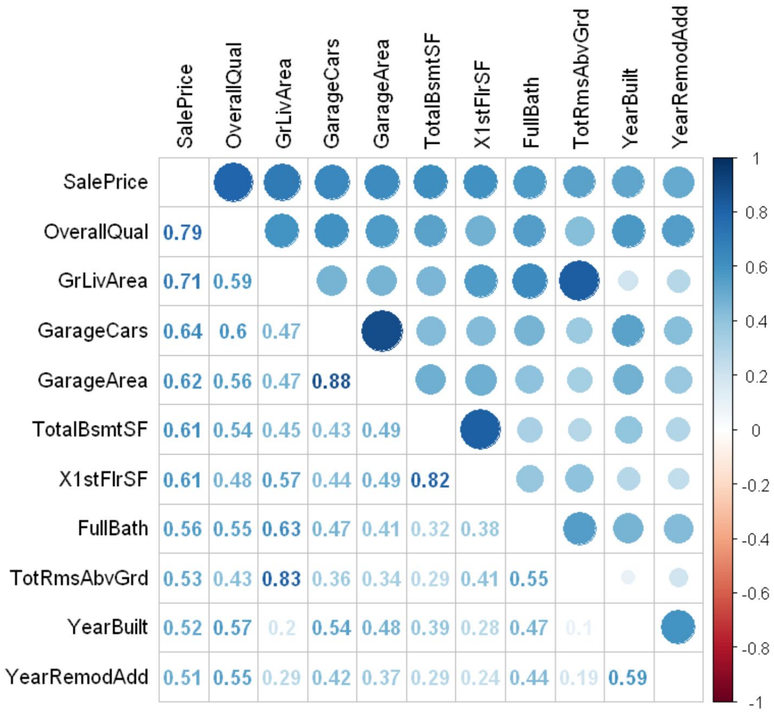
ETS squared measures the proportion of the total variance in a dependent variable that is associated with the membership of different groups defined by an independent variable.

Top variables with their ETS-squared

Neighborhood	0.545574990809563
ExterQual	0.477387777727006
KitchenQual	0.456598624444538
BsmtQual	0.453756066322398
PoolQC	0.448651398739346
Alley	0.285496725954748
GarageFinish	0.267276356592116
Foundation	0.256368401530418
GarageType	0.206638403932996
HeatingQC	0.195500485840093

1. Data Summary: Selected numeric features

Correlation



RFE

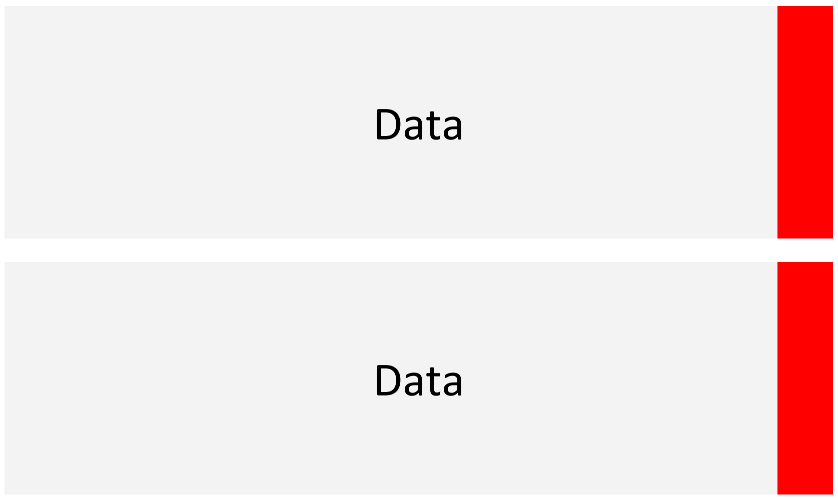
```
Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

 Variables  RMSE Rsquared  MAE RMSESD RsquaredSD  MAESD Selected
         4 15140   0.9688 8375   4061   0.010989 1347.4
         8 13146   0.9750 6579   4035   0.014445  936.1
        16 10100   0.9853 4791   3584   0.009545  961.6        *
        37 10527   0.9842 4573   4393   0.010371  993.3

The top 5 variables (out of 16):
   SalePrice, OverallQual, GrLivArea, YearBuilt, GarageCars
```

1. Data Summary: Selected features

Area	Quality	Utilities
TotalBsmtSF Mean: 1034.9	**KitchenQual** Mode: TA	**FullBath** Mean: 1.571
GrLivArea Mean: 1463.7	**BsmtQual** Mode: TA	**GarageArea** Mean: 472.11
X1stFlrSF Mean: 1137.1	**PoolQC** Mode: Ex	**Neighborhood** Mode: Names

2. Methodology: Data

Train dataset



Test dataset

2. Methodology: Models and performance

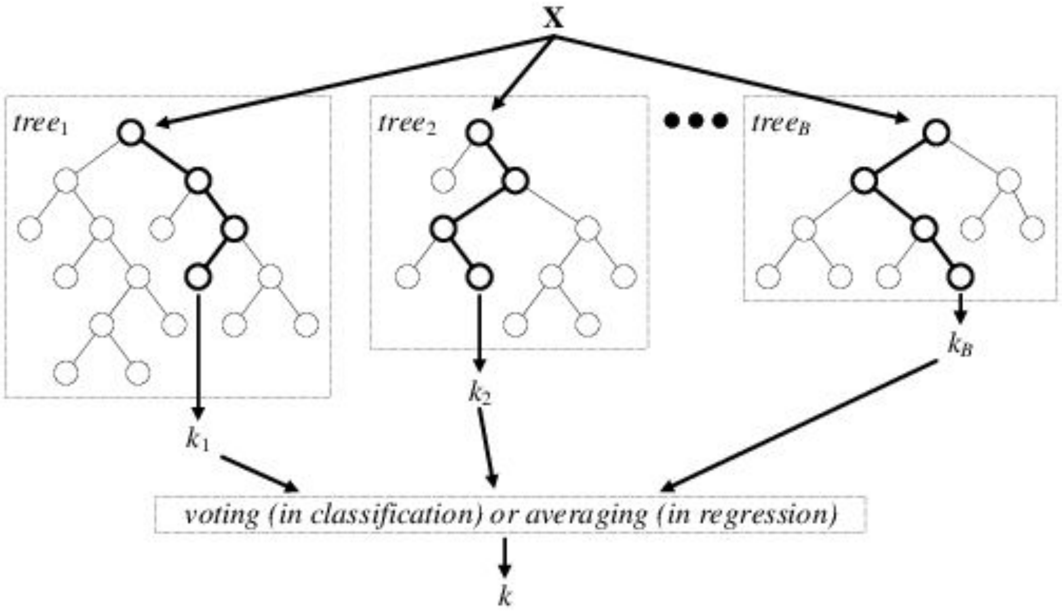
Models	Performance (R2)
Linear Regression	0.8192
Regression tree	0.7319
Random Forest	0.8469
Lasso	0.8217

2. Methodology: Random Forest

Description

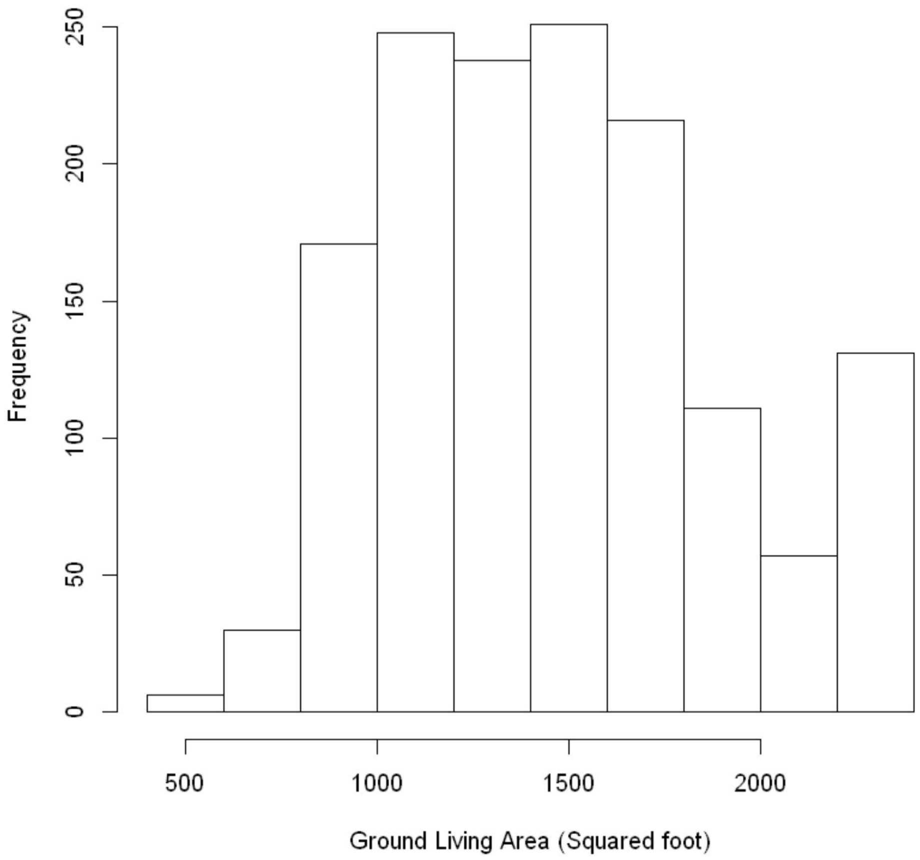
The **random forest** is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

Example

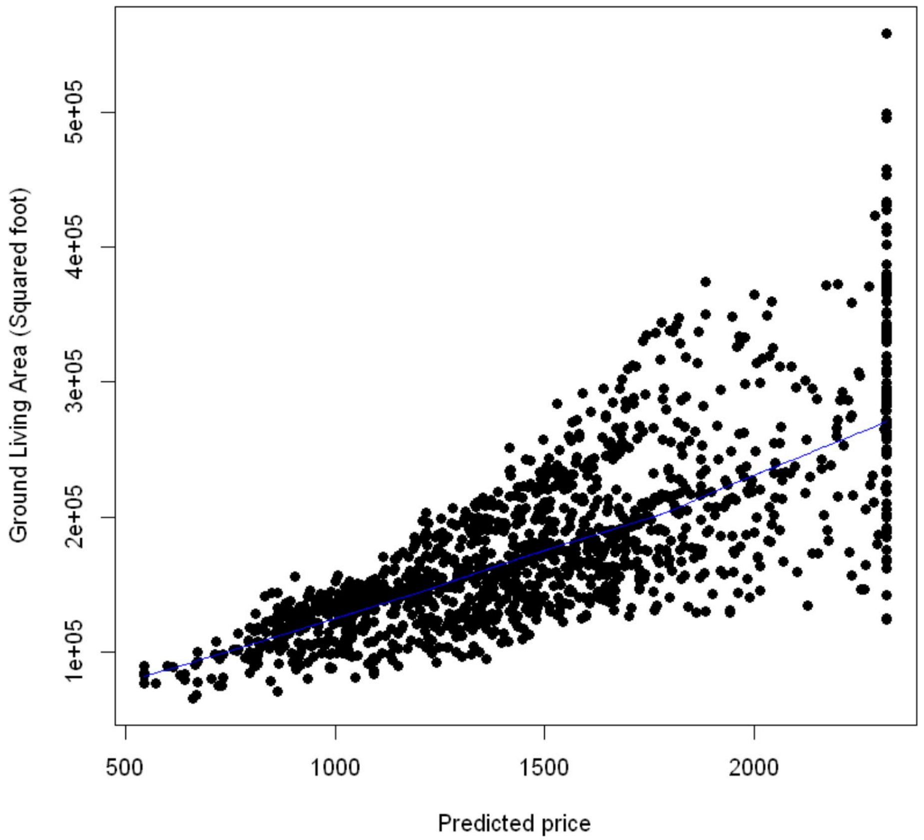


2. Methodology

Distribution of ground living area

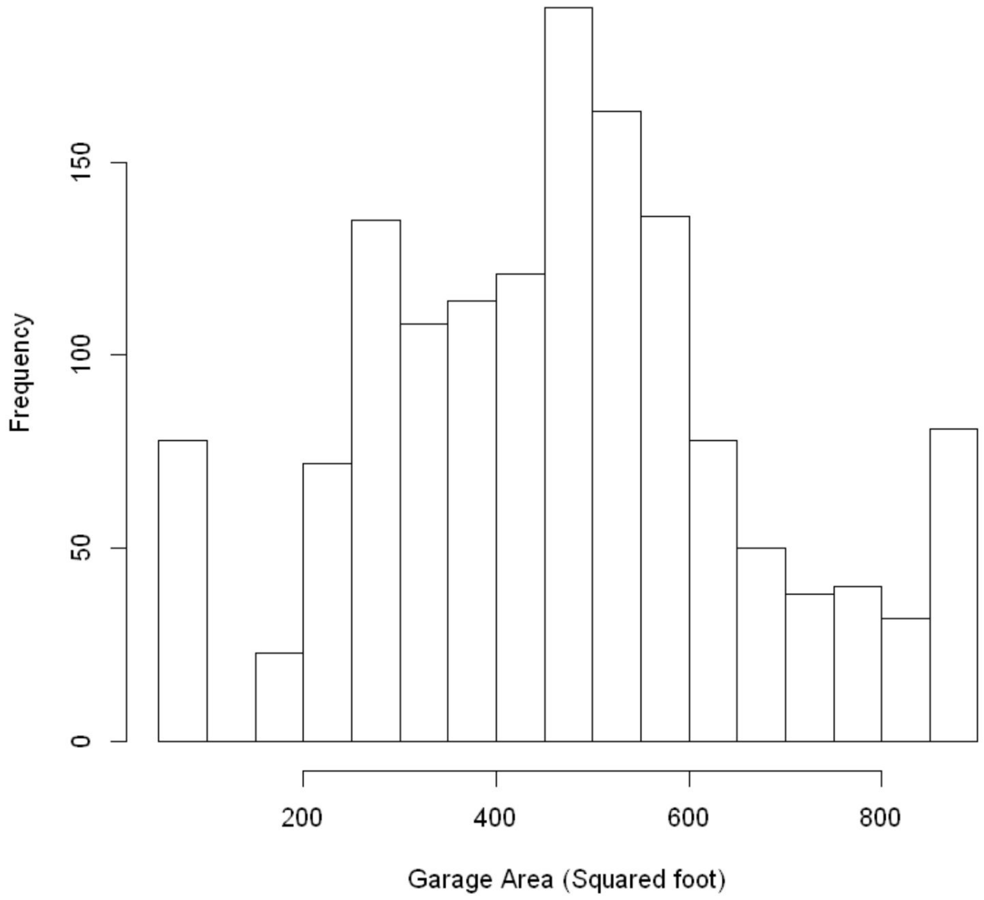


Ground living area vs. Predicted price

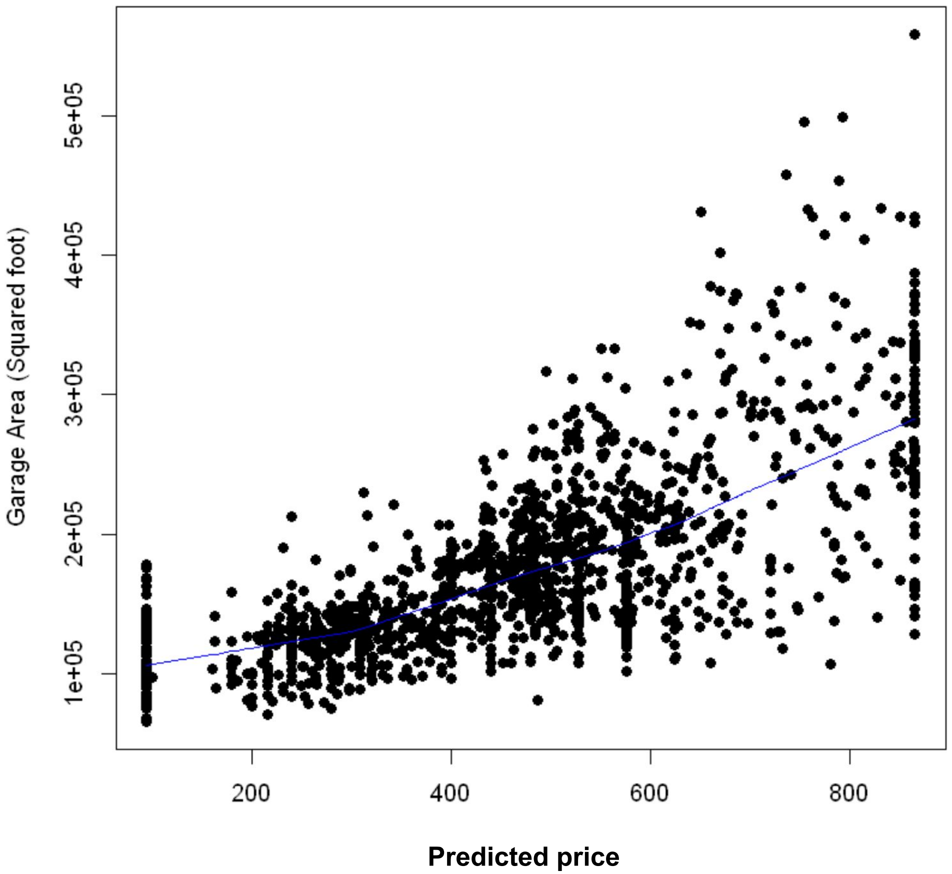


2. Methodology

Distribution of garage area

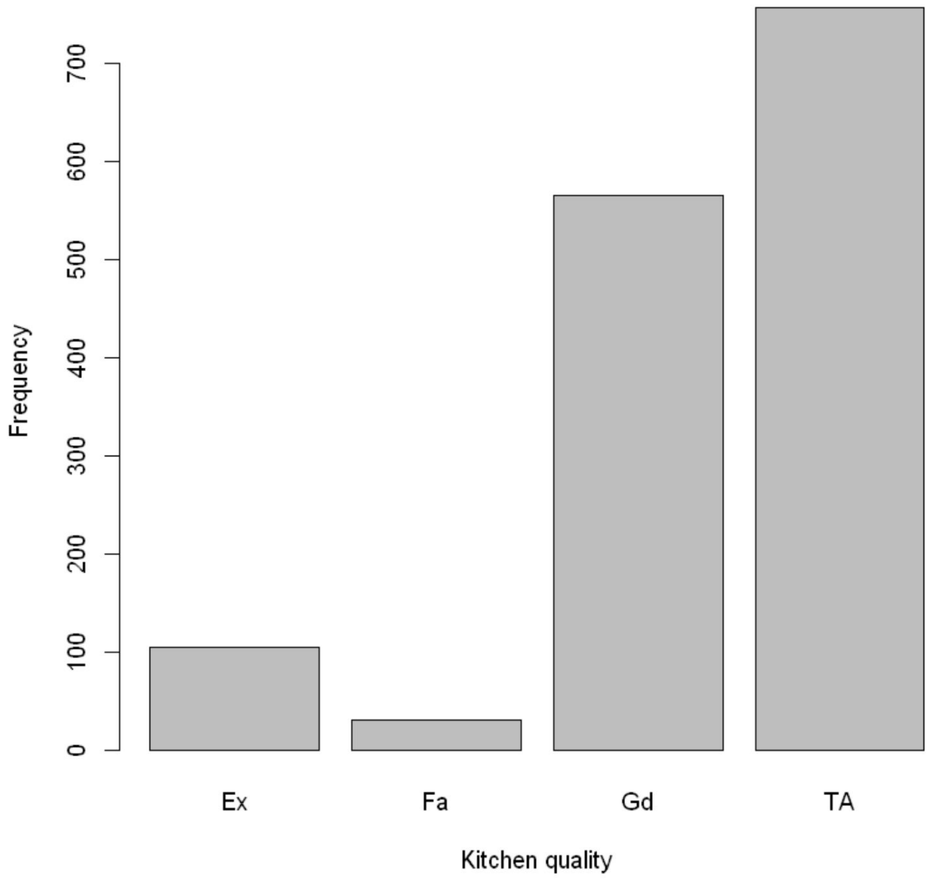


Garage Area vs. predicted price

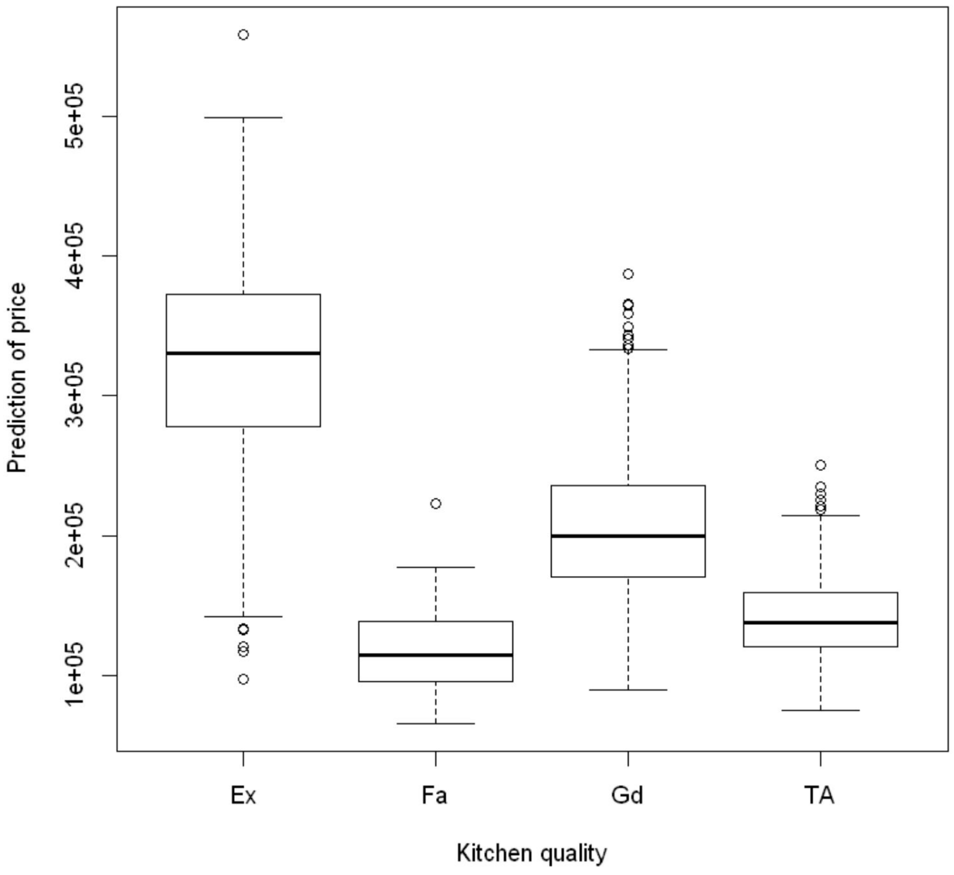


2. Methodology

Distribution of the kitchen quality

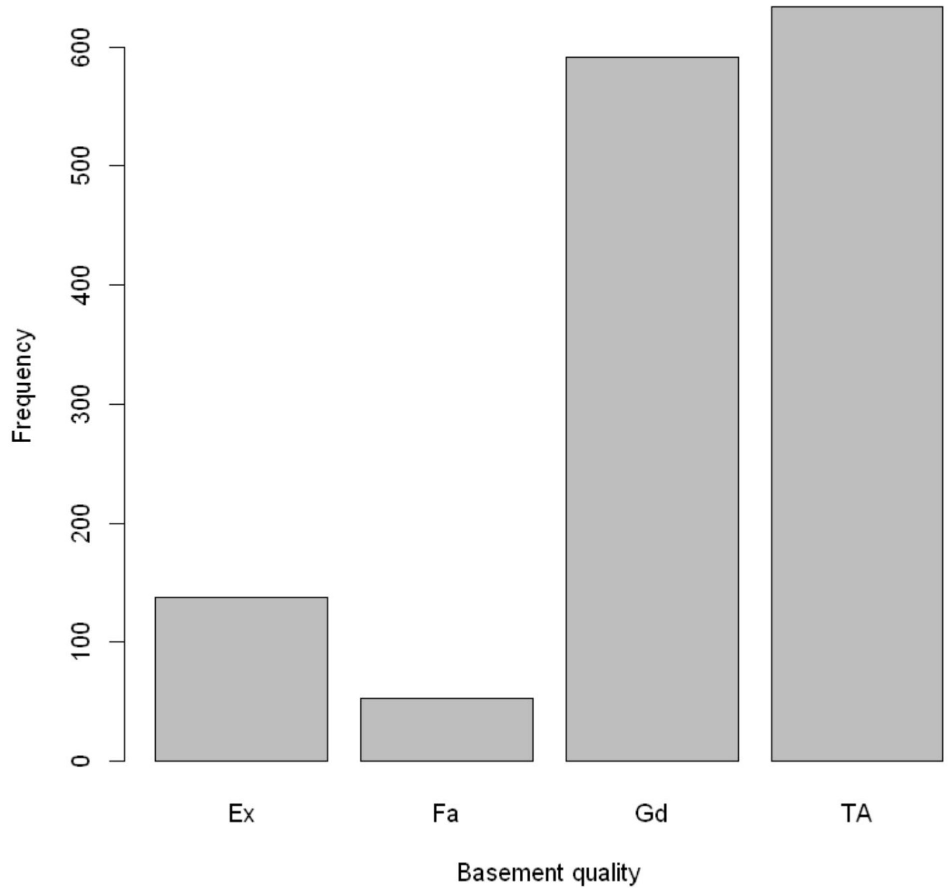


kitchen quality vs. predicted price

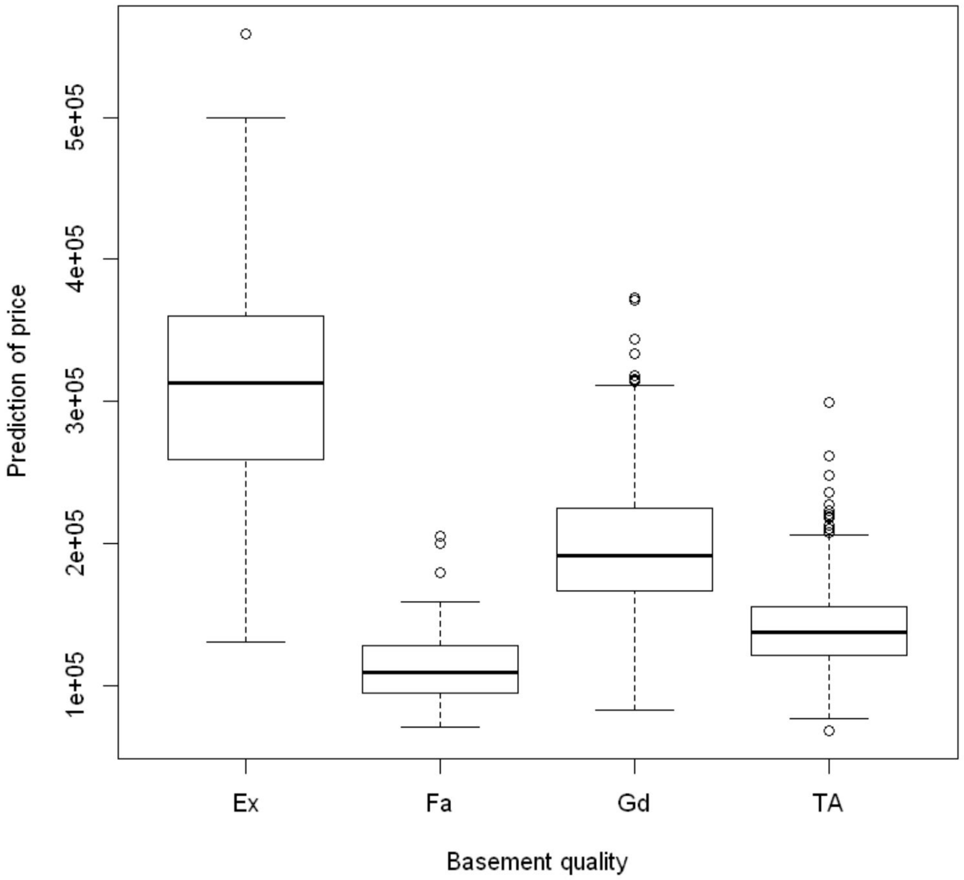


2. Methodology

Distribution of basement quality



Basement quality vs. predicted price



3. Results: Undervalued properties

Methodology

The methodology used consists on identifying the properties whose predicted value was higher than the actual value that we found in the market. If this difference was significant therefore we conclude that we identified an undervalued property.

Rule of thumb

Undervalued
Predicted Price > Market Price (observed)

Overvalued
Predicted Price < Market Price (observed)

3. Results: Properties under EUR 70 000

List of undervalued properties under EUR 70 000

TotalBsmtSF	GrLivArea	FullBath	GarageArea	X1stFlrSF	KitchenQual	BsmtQual	PoolQC	Neighborhood	pred_random_test	difference	MarketPrice
1108	2318.595	2	670	1148	Ex	Ex	NA	NridgHt	329707.0	262430.87	67276.17
1129	2318.595	2	596	1129	Gd	Gd	NA	CollgCr	263488.1	194866.16	68621.99
1554	1554.000	2	627	1554	Gd	Gd	NA	NridgHt	221386.2	151606.03	69780.18
850	1764.000	2	560	886	Gd	Gd	NA	CollgCr	214376.4	147308.84	67067.52
858	1716.000	2	615	858	Gd	Gd	NA	Somerst	205129.4	137591.06	67538.37
1348	1384.000	2	404	1384	Gd	Gd	NA	Gilbert	190382.4	124903.75	65478.69
756	1573.000	2	440	769	Gd	Gd	NA	Somerst	180431.0	113258.77	67172.22
744	2140.000	2	549	825	TA	TA	NA	NAmes	173228.8	103325.28	69903.56
600	1223.000	2	480	520	Gd	Gd	NA	Somerst	155711.8	89892.29	65819.51
960	1040.000	1	616	1040	TA	TA	NA	Sawyer	138005.0	70366.14	67638.86
1169	1144.000	1	286	1144	TA	TA	NA	NAmes	137440.6	68136.48	69304.15
864	874.000	1	576	874	TA	TA	NA	NAmes	125992.0	59519.47	66472.53
644	1316.000	1	369	672	TA	TA	NA	Crawfor	124592.2	58754.99	65837.18
896	936.000	1	288	936	TA	TA	NA	NAmes	122327.1	53623.65	68703.41
827	1251.000	1	240	827	Fa	Gd	NA	OldTown	115624.4	50067.02	65557.39
864	864.000	1	732	864	TA	TA	NA	Sawyer	118186.9	49827.87	68359.04

3. Results: Properties under EUR 70 000

Property

Neighborhood	:	NridgHt
Bathrooms	:	2
Ground living area	:	2319 sf
Kitchen and Basement quality	:	Excellent

Investment

Buy	:	EUR 67 276
Sell	:	EUR 329 707
Profit	:	EUR 262 431
ROI	:	390%

Thank you for your attention