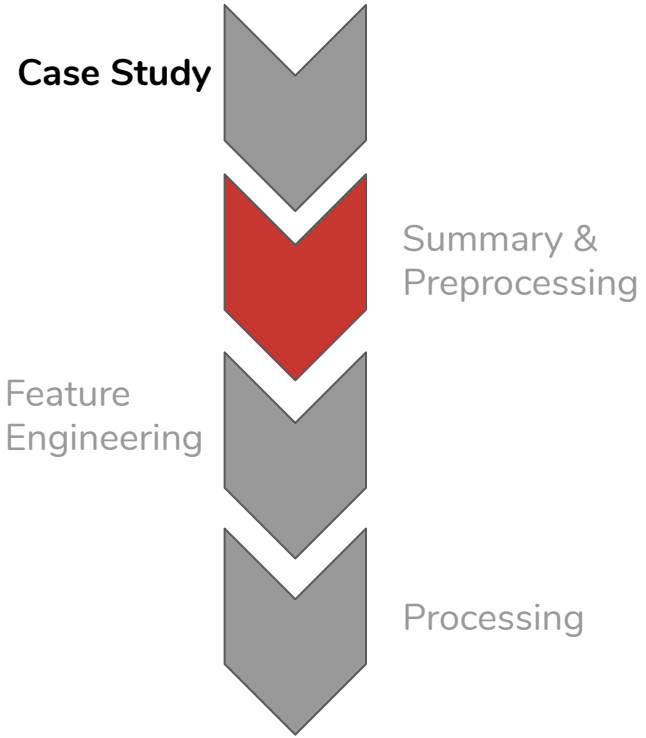
House Prices:
Advanced Regression Techniques

Andres Olivera
Eduardo Alfonso Razo
Clarie Hutin

Table of contents

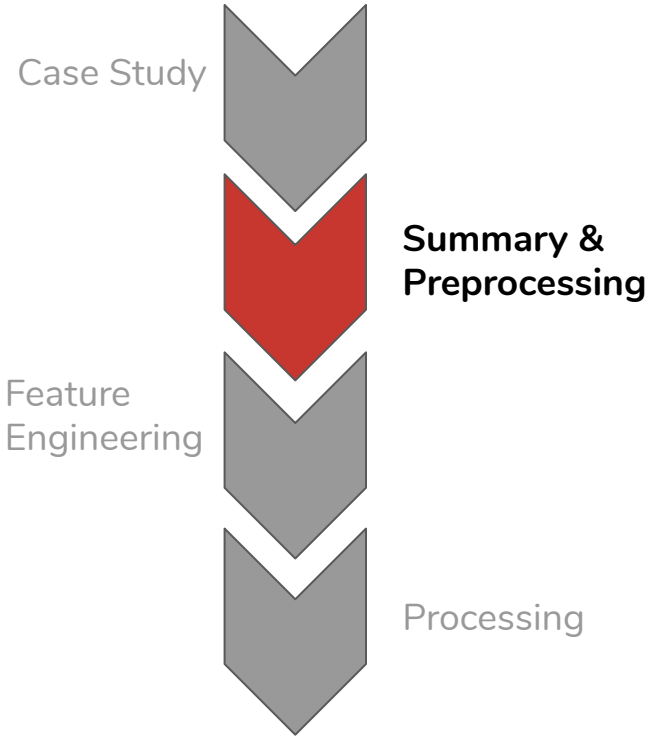


1. Case Study

Explanation of the problem

Case Study

- Kaggle competition: using data from residential homes of Aimes, Iowa, USA, try to predict the sale price of the houses.
- Regression problem
- Target = Sale price
- 81 dependant variables



2. Summary & Preprocessing

Exploration of the dataset and steps to prepare de data

Data Summary

Train

Variables:81

Observations:1460

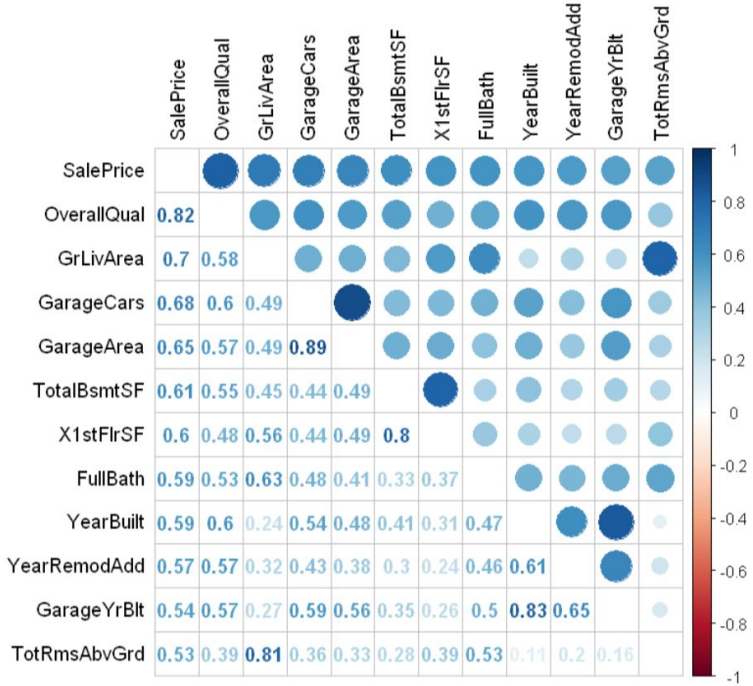
Test

Variables:80

Observations:1459

Categorical	Numerical	Dates
29	49	3

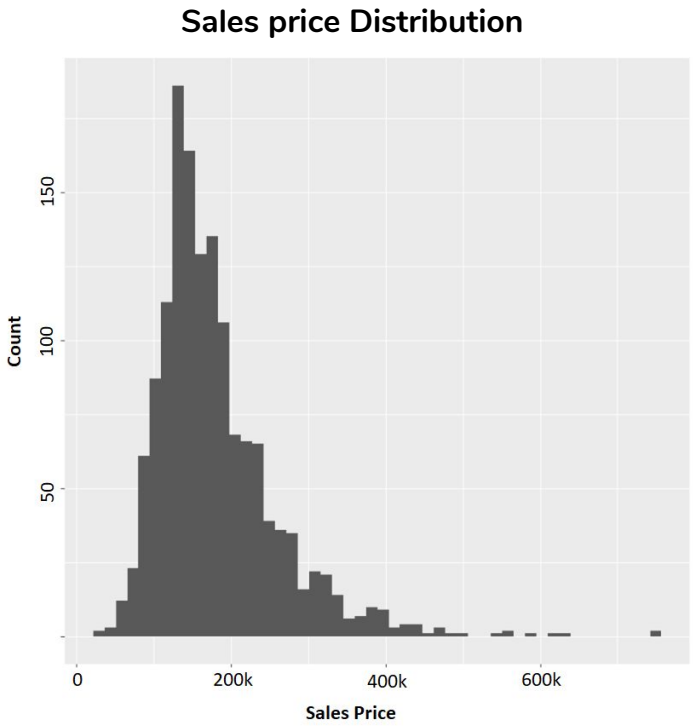
Processing

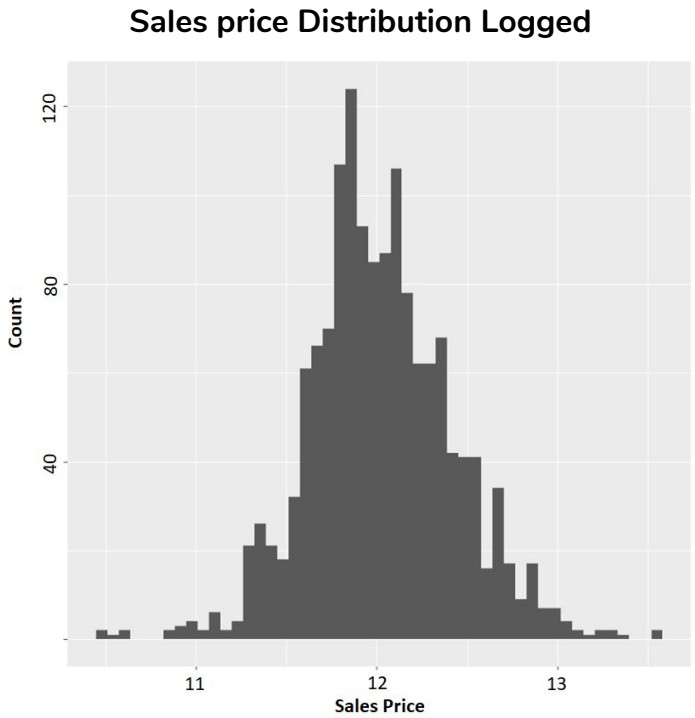


Top 11 correlated variables with the Sale Price

Overall Quality, Living Area above ground and the amount of garages are the most relevant variables at first glance

Processing


Sales price Distribution
Count
150
100
50
0
200k
400k
600k
Sales Price


Sales price Distribution Logged
Count
120
80
40
11
12
13
Sales Price

Handling Missing Values

1

Flag Missing

19 and 33 new columns flagged for train and test respectively

2

Replace Numerical

10 replaced by 0
1 replaced by **median**

3

Replace Categorical

7 with specific string for both.
1 and 15 replaced by ‘None’ respectively.

T

Other Pre-Processing

Transform numerical that are actually categorical

5

Dates and factors

Change character variables into integers

17

Mostly describing qualities For better easiness of reading in data

Removing variable

1

Utilities = 0 variance

Log of IVariables

13

Log of variables' which skewness is >0.5 or <-0.5

Feature Engineering

Basement area

Addition of the following features:

- TotalBsmtSF
- X1stFlrSF
- X2ndFlrSF

Number of bathroom

Addition of the following features:

- FullBath
- HalfBath * 0.5
- BsmtFullBath
- BsmtHalfBath * 0.5

House remodeled

Boolean:

If YearBuilt==YearRemodAdd

0 = No remodeling
1 = Remodeling

House age

Subtraction of the following features:

- YrSold
- YearRemodAdd

Feature Engineering

House new

Boolean:

If
YrSold==YearBuilt

0 = no new
1 = New

Total area

Addition of the following features:

GrLivArea
+
TotalBsmtSF

Sale price per neighborhood

Grouping

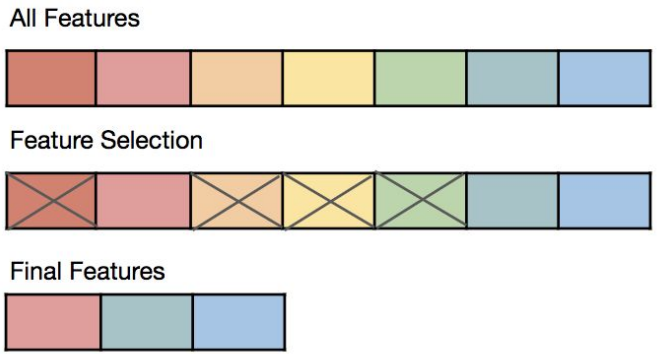
3 groups:
0, 1, 2

Dummy encoding

Encoding of categorical variables

0 = No
1 = Yes

Feature Selection



Boruta

Wrapper of Random Forest.

Selects variables by measuring their importance

Variables

256 ---> 101

4. Modeling

The models selected to conduct the benchmark the sale price

Modeling

1. Gradient Boosting Machine
2. Linear Regression
3. XGBoost
4. Random Forest Regression
5. Lasso
6. Generalized Linear Regression

Hyperparameter Tunning
GBM

Resampling method	Iterations	Performance metrics	n.trees
Cross Validation	12	RMSE Root Mean Square Error	500

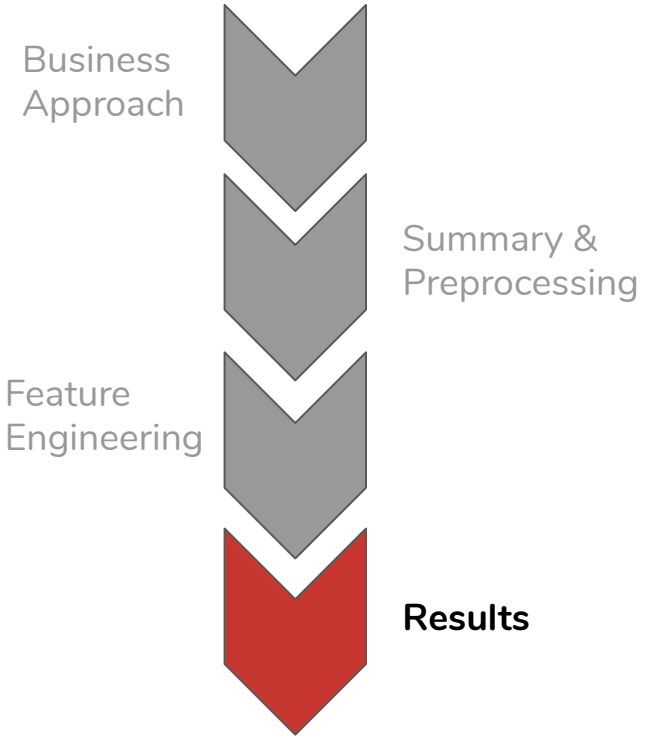
Hyperparameter Tunning
Linear Regression

Resampling method	Iterations	Performance metrics
Cross Validation	10	RMSE Root Mean Square Error

Hyperparameter Tunning
XGBoost

Resampling method	Iterations	Performance metrics	nrounds
Cross Validation	100	RMSE Root Mean Square Error	200 - 600

Max_depth	Lambda	Gamma
3 - 20	0.00055 - 0.0060	0.5 - 0.60



4. Results

The metrics to evaluate the performance of the model

Results

Evaluation			
Metric	GBM	Linear R	XGboost
RMSE	**0.1340**	**0.1436**	**0.1447**

Performance of best models