House Prices: Advanced Regression Techniques

Date:27.06.2018

Objectives

1. Data presentation and analysis
    a. General exploration of the predictors and target variable
2. Data preprocessing & Feature engineering
    a. missing values
    b. outliers
    c. Multicollinear
    d. near zero variance predictors
    e. Hot encoding of categorical variables
3. Principal component analysis Clustering
4. Prediction
    a. Evaluation of different prediction algorithm
    b. Tuning of the algorithm
    c. Results

Introduction 1/2

- Kaggle Data set
- Predicting house price based on 79 variables
- Features describe every aspects of the house
- 1460 Individuals (training)
- Complex dataset
    - Many NAs, outliers, incorrect values...

Introduction 2/2

Early research conclusion about house pricing:

- **Location** - location is key for high valuations, therefore having a safe, well facilitated and well positioned house within a good neighbourhood, is a large contributing factor.
- **Size** - The more space, rooms and land that the house contains, the higher the valuation.
- **Features** - the latest utilities and extras are highly desirable.

Preprocessing - Missing Data 1/6

35/79 features with missing values

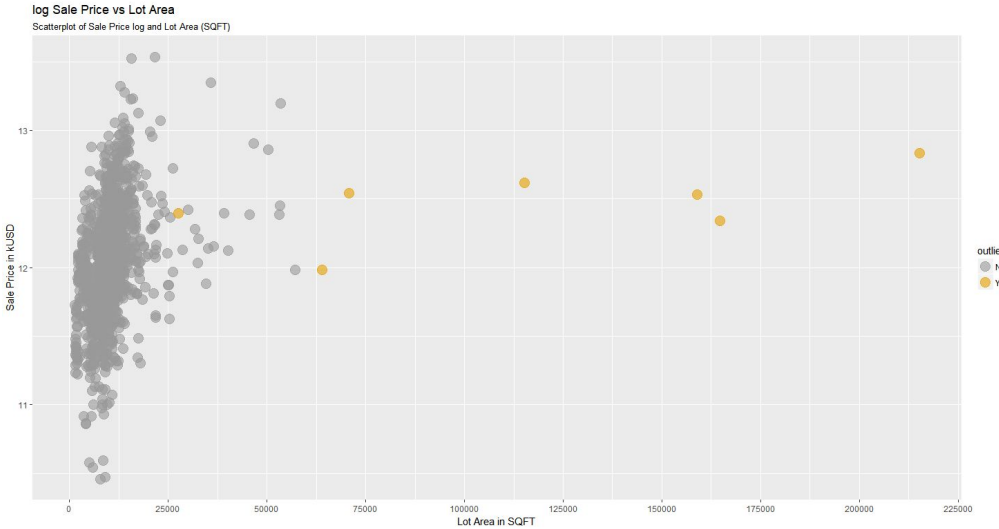
Identified on:

- Missing for a reason (many): N/A represent e.g. No Pool
  - Solution: Impute value “none” or ”0”
- Missing at random (some):
  - Solution: Impute value based on random forest or reasoning

Preprocessing - Outliers 2/6

Mahalanobis distance with a threshold of 30

7 outliers detected in that case



Preprocessing - One-Hot encoding of categorical values 3/6

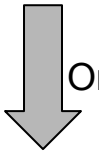
All categorical non-ordinal predictor levels were converted into single numeric columns

Categorical feature “Street” with levels: “gravel” and “paved”

Encoded into new features “HasGarvel” and “hasPaved” - The numerical values are ‘1’ for existing and ‘0’ otherwise.

Increasement of dimensions : 78->110

Street
gravel
paved



One-Hot Encoder

Street = gravel	Street = paved
1	0
0	1

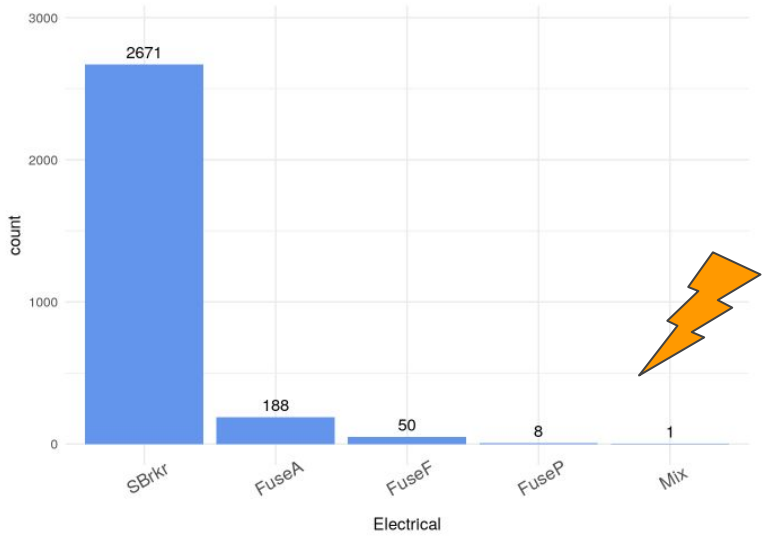
Preprocessing - Near Zero Variance 5/6

Some of these features have

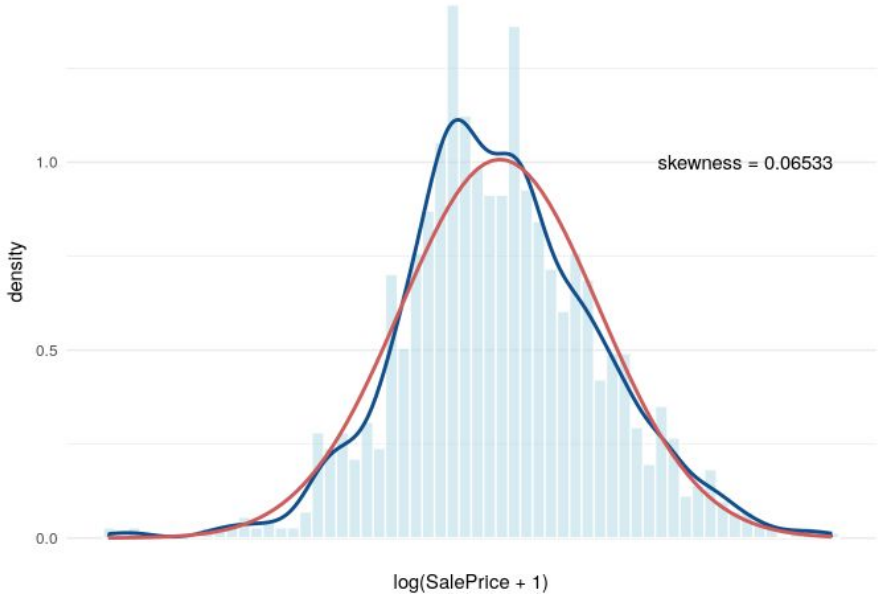
become zero-variance predictors

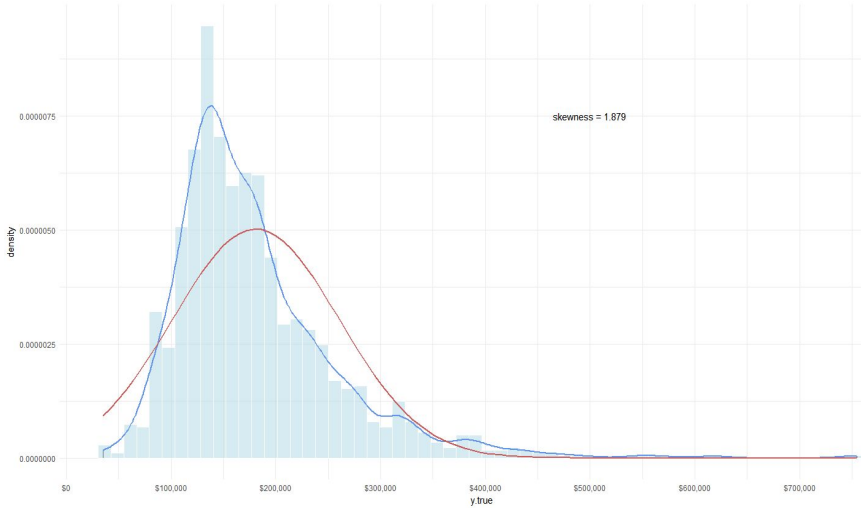
Decision to remove them->

From 110 to 68 features



Preprocessing - Normality 4/6





Standardization of predictor features.
Useful in distance-based algorithm (K-means) & optimization (Regression).

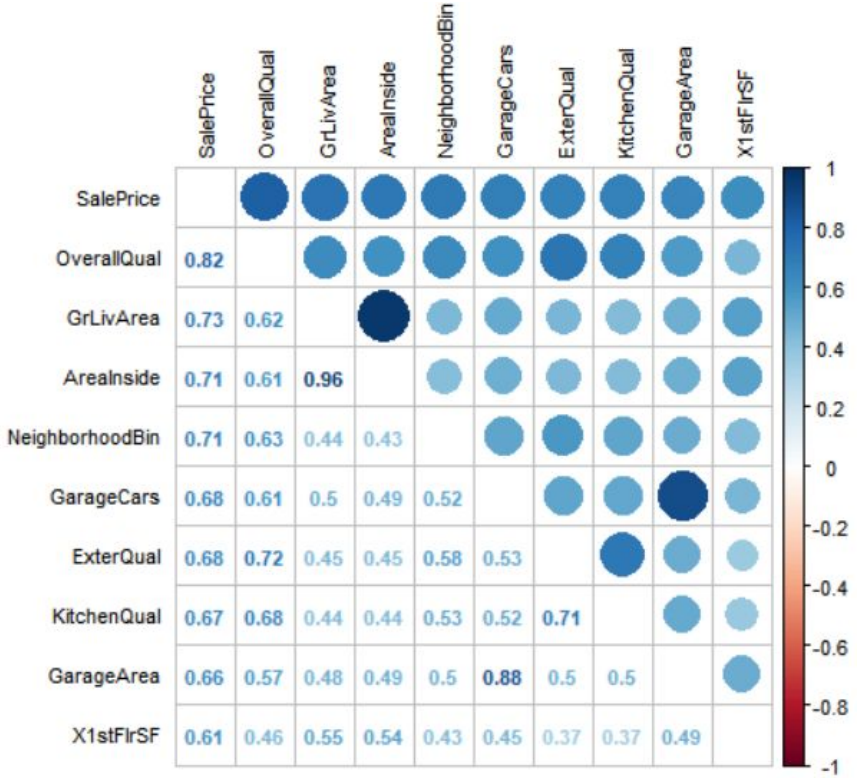
Data was skewed in many cases (positive above) so we applied **Log transformation** of the continuous variables.

Preprocessing - Multicollinearity 6/6

Creating new feature - geometric mean

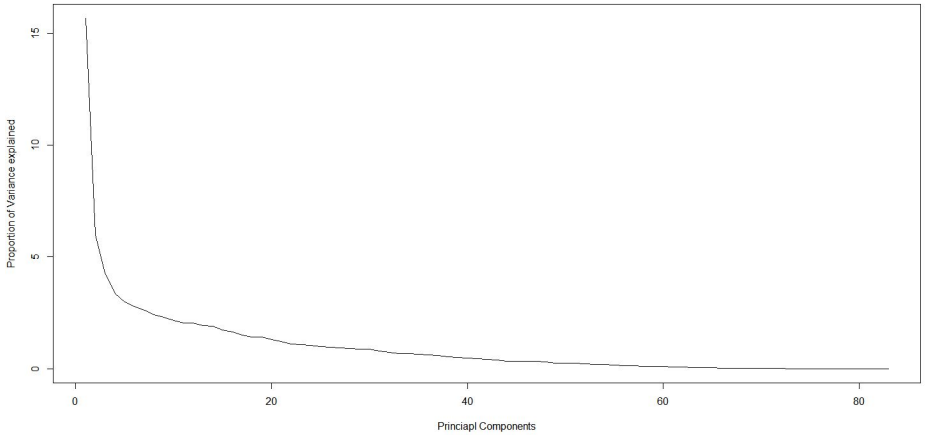
-AreaInside and Greater Living area 96% correlated

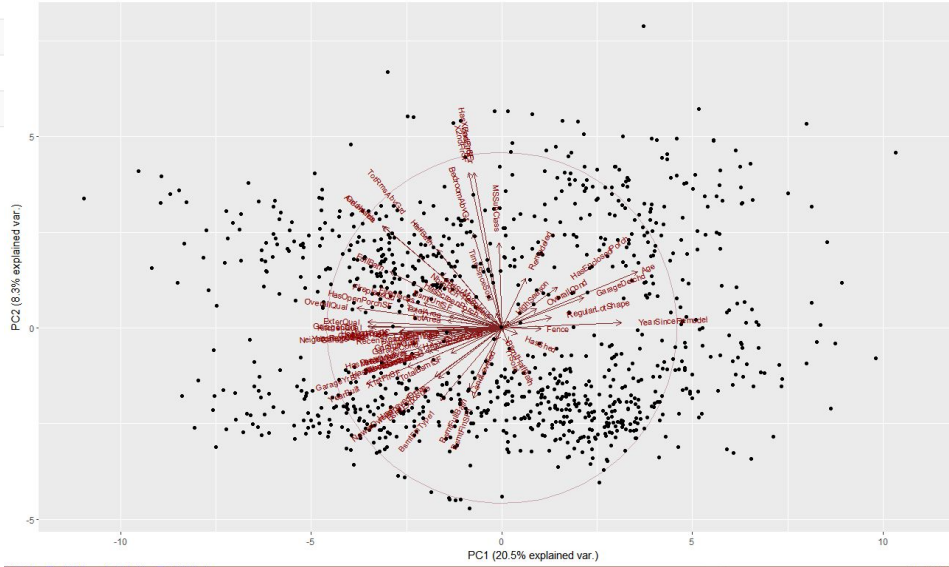
-GarageArea and GarageCars 88% correlated



PCA and Cluster Analysis 1/3

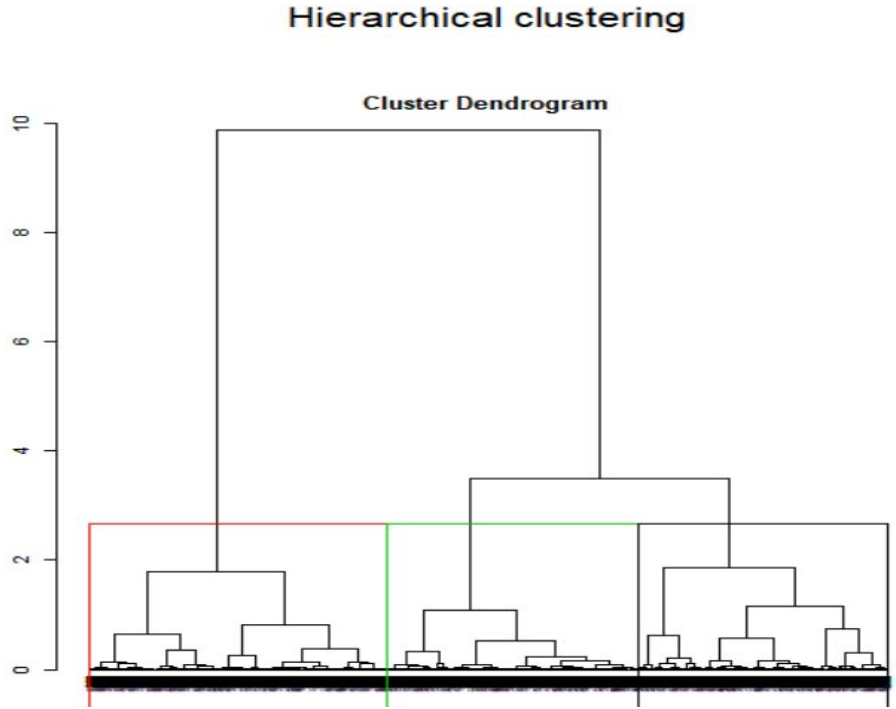
	coord.Dim.1	coord.Dim.2	coord.Dim.3	coord.Dim.4	coord.Dim.5	coord.Dim.6	coord.Dim.7	coord.Dim.8	coord.Dim.9
1	OverallQual	X2ndFlrSF	LotFrontage	WoodDeckSF	GarageQual	OverallCond	WoodDeckSF	YrSold	BsmtUnfSF
2	YearBuilt	GrLivArea	LotArea	HasWoodDeck	GarageCond		HasWoodDeck	TimeSinceSold	TotalBsmtSF
3	YearRemodAdd	BedroomAbvGr	X1stFlrSF	HasWoodDeckSF			HasWoodDeckSF		
4	X1stFlrSF	TotRmsAbvGrd	TotalArea						
5	GrLivArea	Has2ndFlr							
6	FullBath	HasX2ndFlrSF							
7	TotRmsAbvGrd	AreaInside							
8	GarageYrBlt								
9	GarageCars								

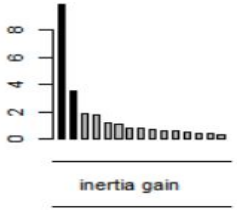




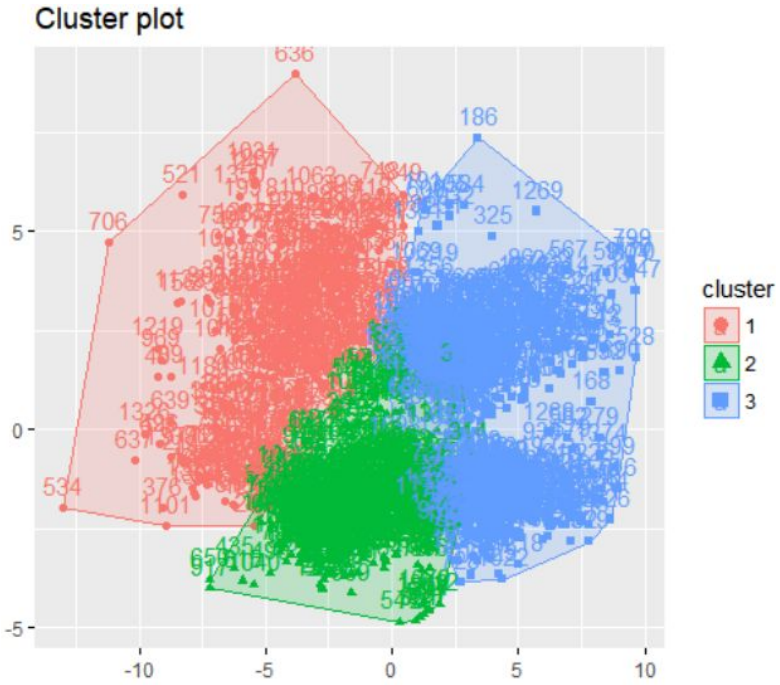
PCA and Cluster Analysis 2/3

agglomerative clustering





PCA and Cluster Analysis 3/3



c1 = 157470€
Age 70 years

c2 = 170673 €
Age 30 years
Overall Quality 6/10

c3 = 203424 €
remodeled - median 1999

Predictive Analysis

Regression

Lasso - imposes sparsity among the coefficients
Ridge - limits size of coefficient vector

elastic net - penalty term as a mix of lasso and ridge

Trees:

Random forest - fully grown decision trees
XGBoost (boosted trees) - shallow trees

Predictive Analysis - results

Initial evaluation:

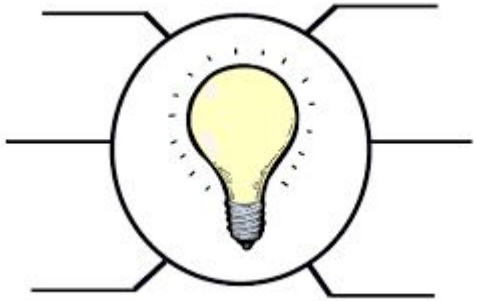
Method	test-rmse
XgBoost	6540
RandomForest	10321.12
lasso	21133.27
ridge	20179.57
net	21184.43

Winner - full evaluation:

Method	train-rmsle	test-rmsle	cross-validation-rmsle
xgBoost	0.055603	0.191394	0.202694
0.5*xgboost+ 0.25*regression+ 0.25*randomforest			0.195643

Conclusion

- Analysis and Pre-Processing was ~90% of the time
- First evaluation result in upper 15% percentage
- Managed to apply the course knowledge to compete on Kaggle
- XGBoost good, but does not extrapolate like regression
- Optimization is needed on a further step



Future Work

- Identify ordinal categorical distribution (One-hot encoding)
- Improve feature selection/engineering
  - Using the PCA analysis to create new features
  - Reasoning on correlation
  - Cluster Analysis results
- Evaluation of other prediction algorithm
- test prediction isolated on clusters
- Outlier detection on all relevant variables