



Houses Price Prediction

Kaggle Competition
Roman Garayev



Dataset. Overall 2600 object, 81 columns.

SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.

MSSubClass: The building class

MSZoning: The general zoning classification

LotFrontage: Linear feet of street connected to property

LotArea: Lot size in square feet

Street: Type of road access

Alley: Type of alley access

LotShape: General shape of property

LandContour: Flatness of the property

Utilities: Type of utilities available

LotConfig: Lot configuration

...

EDA

Correlations

OverallQual and target 0.7909816005838048

GrLivArea and target 0.7086244776126522

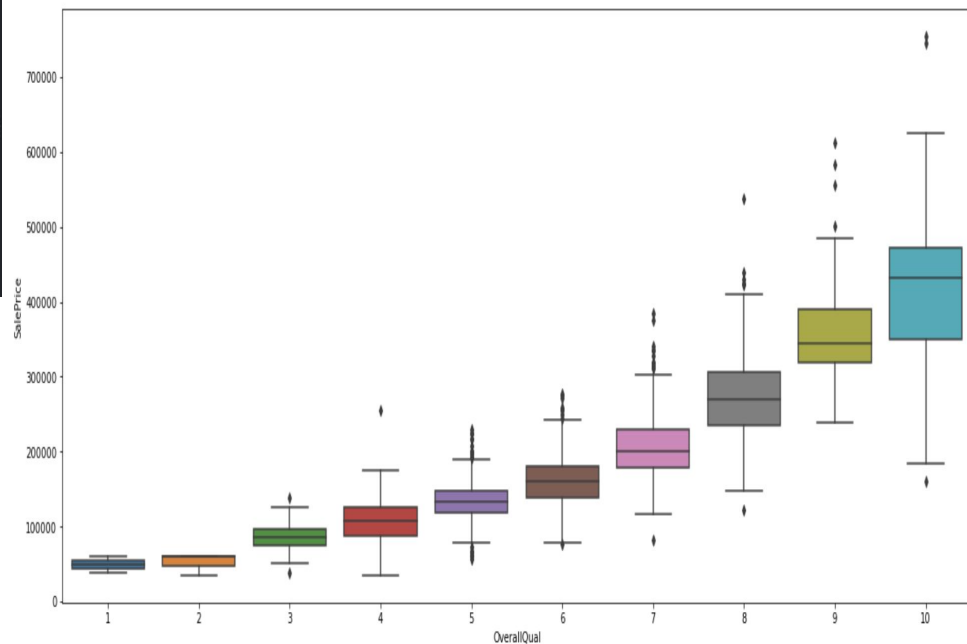
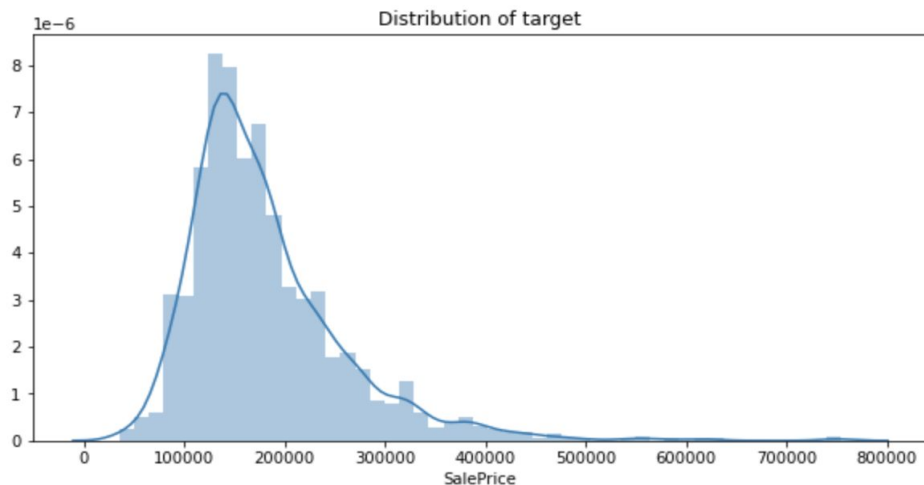
GarageCars and target 0.6404091972583531

TotalBsmtSF and target 0.6135805515591956

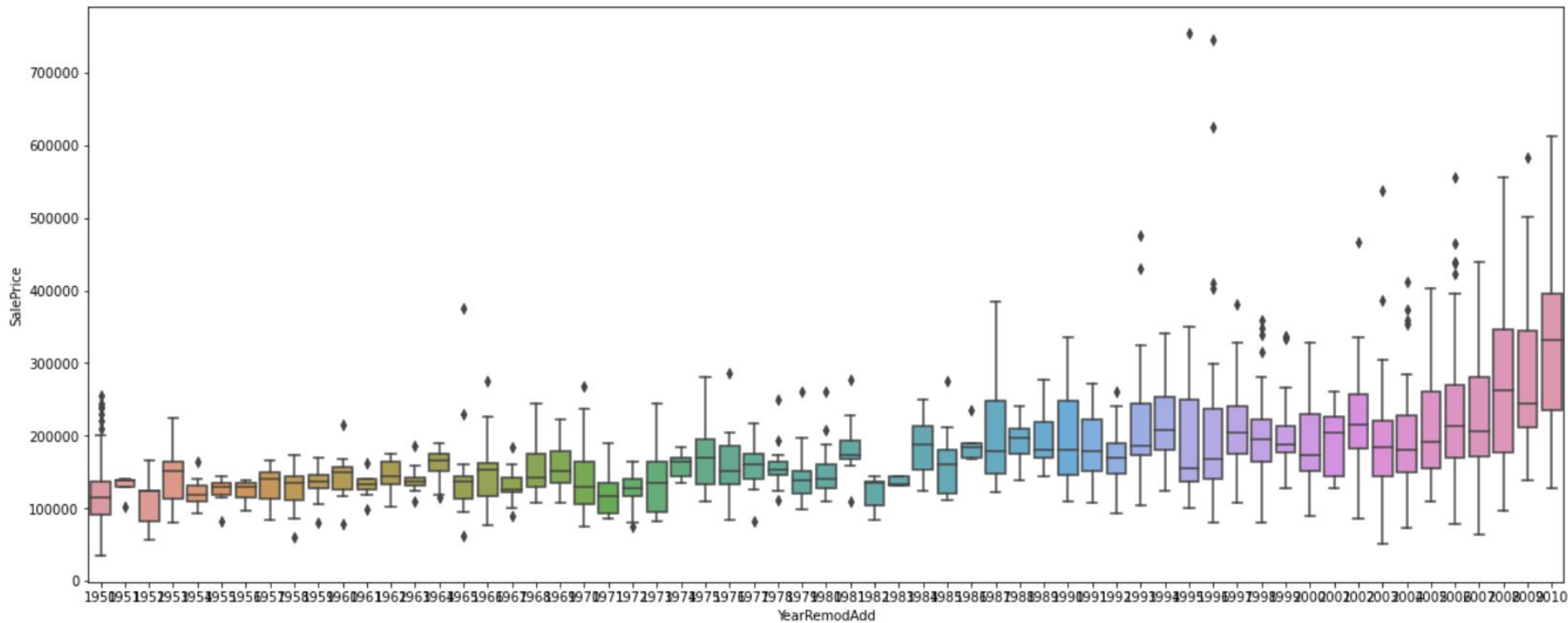
FullBath and target 0.5606637627484456

YearBuilt and target 0.5228973328794969

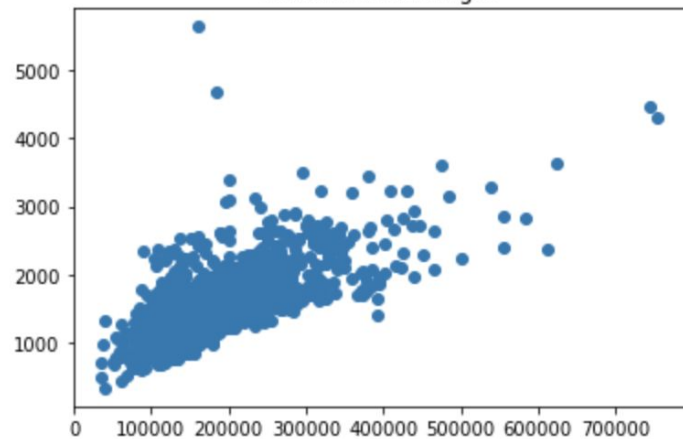
Shapiro-Wilk's test p-value ~0.00000000...3



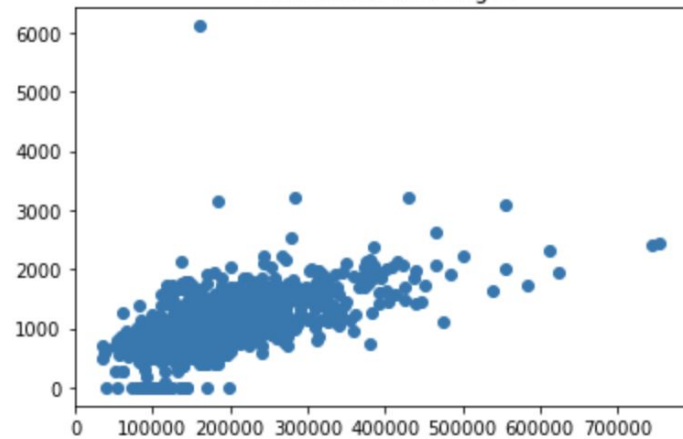
Improvement year and target variable



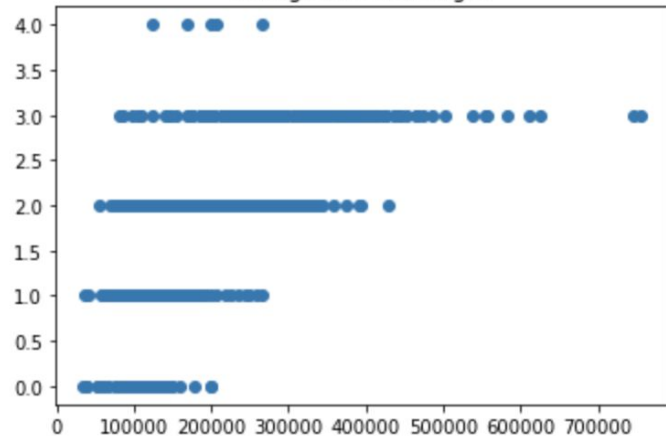
GrLivArea and target



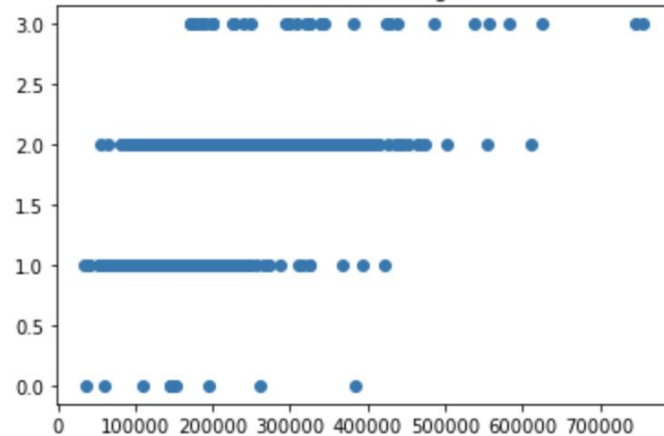
TotalBsmtSF and target



GarageCars and target



FullBath and target



NA's Ratio



PoolQC	99.657417
MiscFeature	96.402878
Alley	93.216855
Fence	80.438506
FireplaceQu	48.646797
LotFrontage	16.649538
GarageFinish	5.447071
GarageYrBlt	5.447071
GarageQual	5.447071
GarageCond	5.447071
GarageType	5.378554
BsmtExposure	2.809181
BsmtCond	2.809181
BsmtQual	2.774923
BsmtFinType2	2.740665
BsmtFinType1	2.706406
MasVnrType	0.822199
MasVnrArea	0.787941
MSZoning	0.137033
BsmtFullBath	0.068517
BsmtHalfBath	0.068517

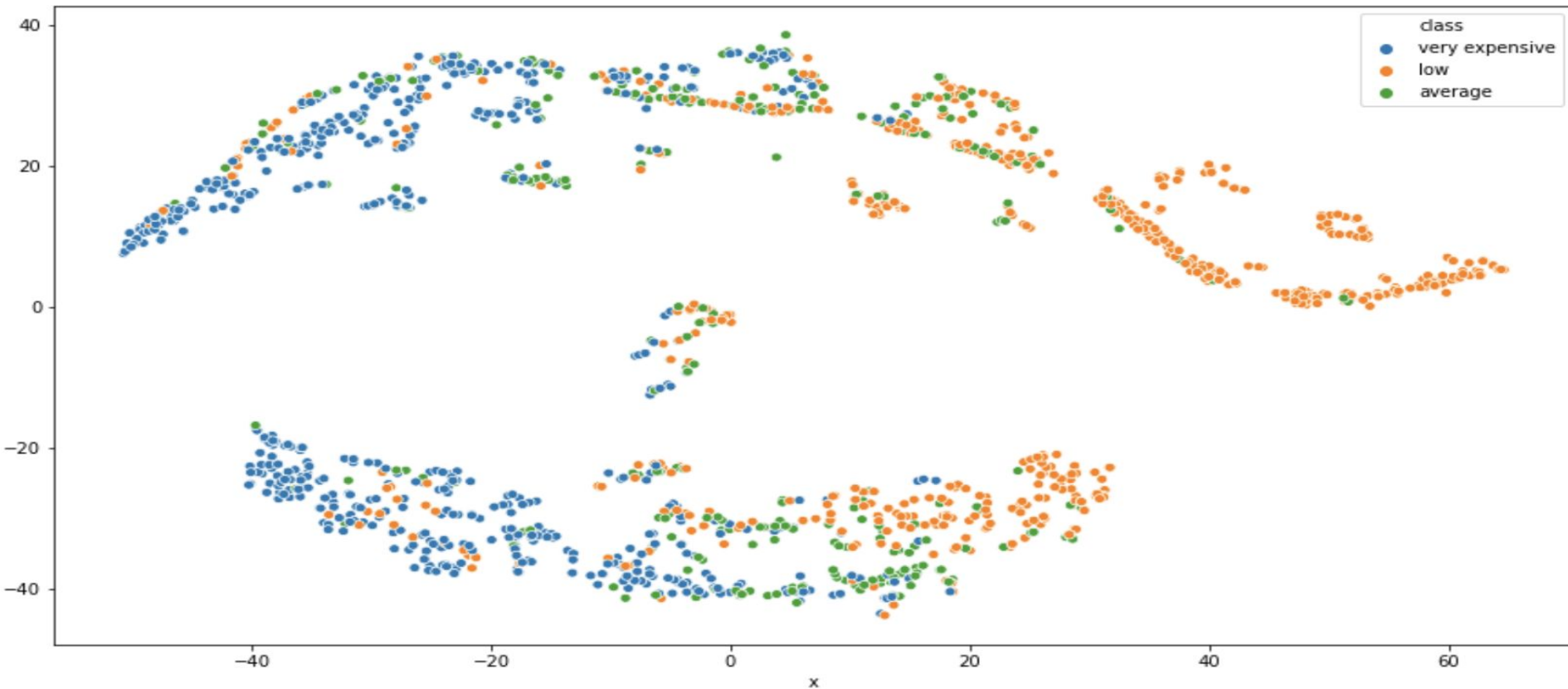
NA's Imputting Strategy

Because this is a dataset about houses some "NAs" can be replaced with "None" or mode. Not all houses have the same options and advantages.

Skewed feature was transformed with the help of 'box-cox' transformation.

T-SNE. Visualization.

Very expensive: > 60 percentile. Average: > 40 percentile. Low: others





Modeling

Models: XGBoost, LightBoost, CatBoost, ElasticNet

Metric: RMSE

Validation: Cross-Validation, 5 folds

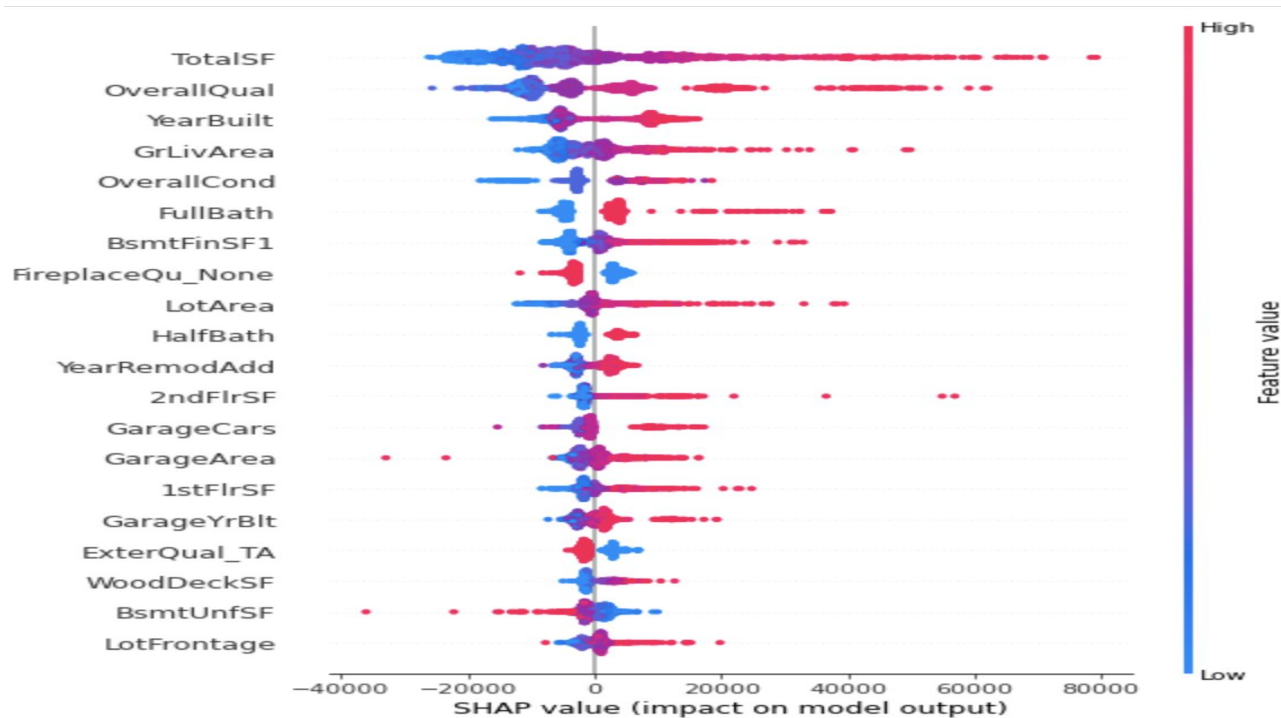


Results

Predictor	RMSE
XGBoost	25152
LBoost	26528
CatBoost	26286
ElasticNet	32393
Averaged XGBoost, CatBoost, LBoost	25203

CatBoost's Feature Selection


8]:



Features Interaction Values



	Feature-Pair	Interaction Score
0	OverallQual TotalSF	1.702891
1	OverallQual GarageCars	0.739821
2	GrLivArea Neighborhood_Edwards	0.664737
3	GrLivArea TotalSF	0.642716
4	YearBuilt TotalSF	0.619882
5	TotalSF Neighborhood_Edwards	0.599441
6	OverallQual GarageArea	0.538202
7	TotalSF GarageType_2Types	0.525051
8	LotArea TotalSF	0.491799
9	1stFlrSF GarageArea	0.485383
10	TotalSF PoolQC_None	0.442253
11	BsmtFinSF1 TotalSF	0.439642
12	OverallQual BsmtFinSF1	0.437740
13	TotRmsAbvGrd TotalSF	0.429132
14	OverallQual TotRmsAbvGrd	0.426386
15	OverallQual LandContour_Bnk	0.419083
16	GarageCars TotalSF	0.383272
17	TotalBsmtSF GarageArea	0.377477
18	OverallQual 1stFlrSF	0.371913
19	LotArea OpenPorchSF	0.353062



According to Interactions, 'Golden Features' were processed differently (by using "border_count parameter), therefore it increased performance of catboost model and averaged stacking to 24K rmse.