ML 1 – Linear Regression and Ridge LR

Baseline LR

All features included
Target Log Transformation
For **Categorical** features:
- Label Encoding (for binary)
- One-Hot Encoding (for multilevel)

For **Numerical** Feature:
- MinMaxScale

MSE (log): 0.1751
R^2 (log): 0.5422

MSE (original): 94412233726.71
R^2 (original): 0.4582

LR – VIF-based correction and control of multicollinearity

Feature exclusion based on:
- Extremely high VIF
- Low correlation with Target

Result:
- All features VIF < 200
- Decreased multicollinearity risk
- Excluded perfect multicollinear features

MSE (log): 0.1758
R^2 (log): 0.5405

MSE (original): 94654649372.42
R^2 (original): 0.4568

many features

multicollinearity risk

Ridge LR

Generated key features:
sqft*grade,
sqft*beds,
sqft*bathrooms_total,
condition*grade,
sqft^2,
sqft^3,
Cyclical Time Transformation
for day/week/month

ML 1 – Linear Regression and Ridge LR

Ridge LR

Previously selected features
Tuned hyperparameter:
alpha: 0.02976

- Stability of coefficients
- Robustness to multicollinearity
- Reduction of overtraining

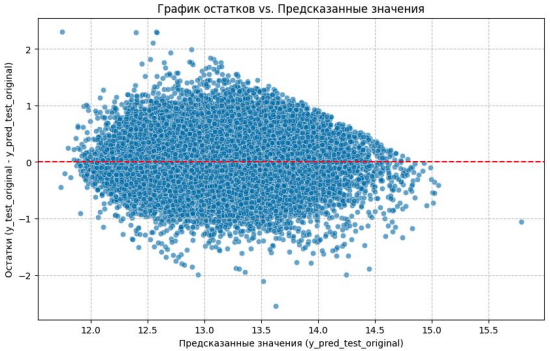
MSE (log): 0.1758
R^2 (log): 0.5405

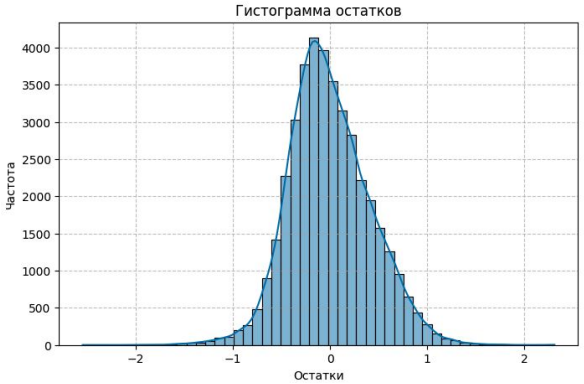
MSE (original): 94647706973.3386
R^2 (original): 0.4568

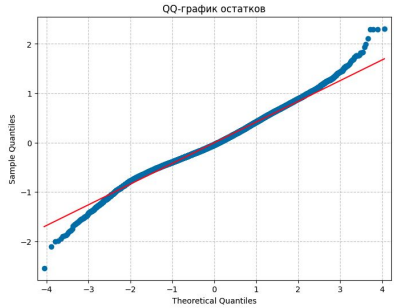
Coefficients (top 10)

sqft	6.406928
sqft*grade	-4.742752
sqft^2	-2.765552
sqft*beds	2.637684
grade	2.269869
sqft_lot	2.203943
sqft*bathrooms_total	-2.043469
imp_val	1.789285
garb_sqft	1.330570
bathrooms_total	1.316744

Linear Regression Assumptions:







ML 1 - Linear Regression and Ridge LR

Linear Regression Assumptions on Restored Target - **Failed**

- prediction errors in dollars increase with increasing house value.
- there is a heavy tail of large negative errors (overestimation).
- presence of significant outliers: the model is highly erroneous on individual properties, especially underestimating them