Project 2

Ames-housing Price Prediction

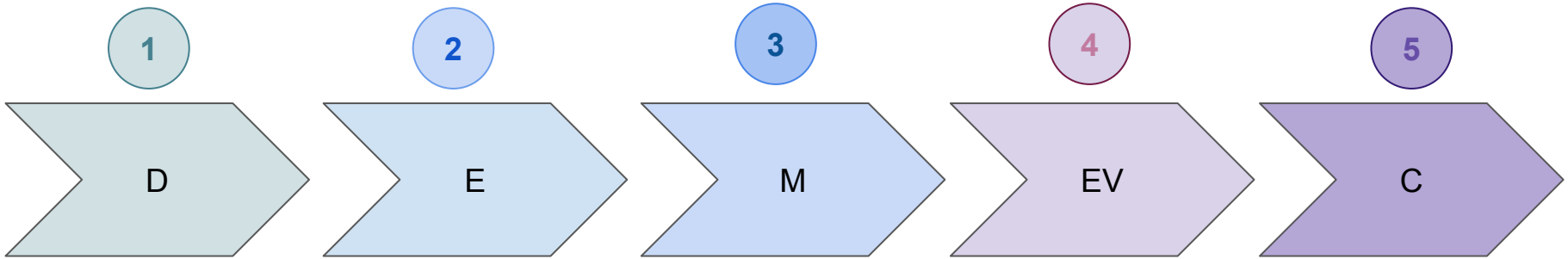
Problem Statement

If want to build a house for selling, and what main factors to think about? And what price should it be?

Objective

To create the linear regression model , for predicting the price of housing in ames city .

Outline



Data Preparation

- Cleaning
- Categorized type of data

EDA (Exploratory Data Analysis)

- Find correlation of variables

Modelling

- Feature selection
- Modeling Workflow

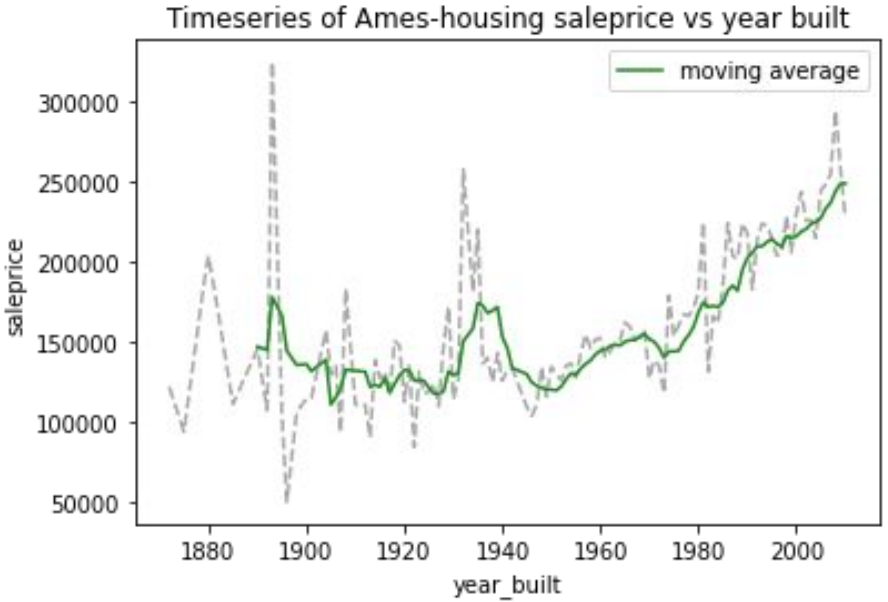
Evaluation

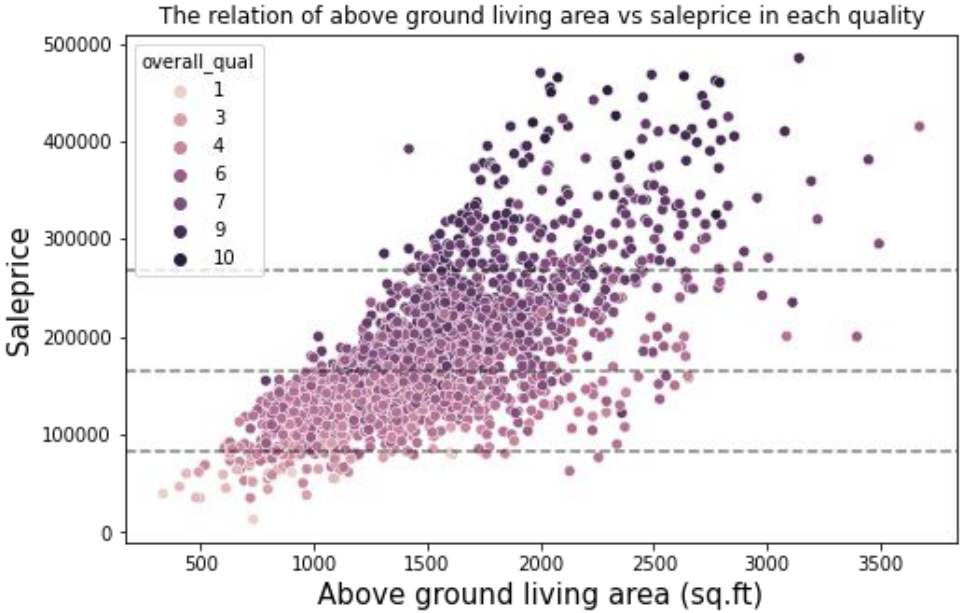
- Evaluating the performance of
- the model
- Actual price vs Predicted price

Conclusion

- Findings
- Recommendation

EDA - What affect to the Sale price?





- Sale prices of houses trend to increase as the newer of the house

- The larger of living area, the higher of the sale price, and overall quality

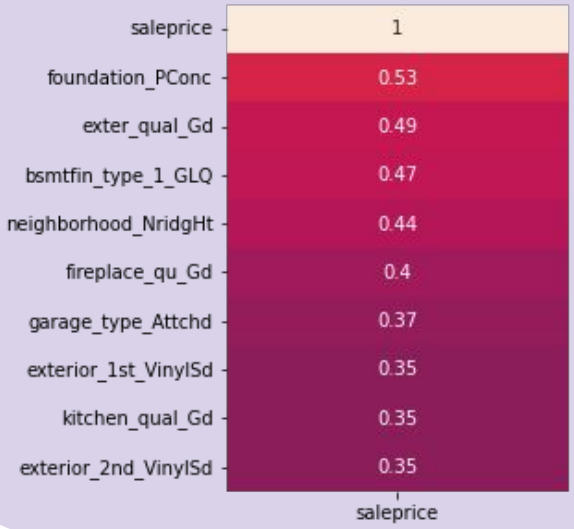
Feature engineering (I) - Imputation & Filtering

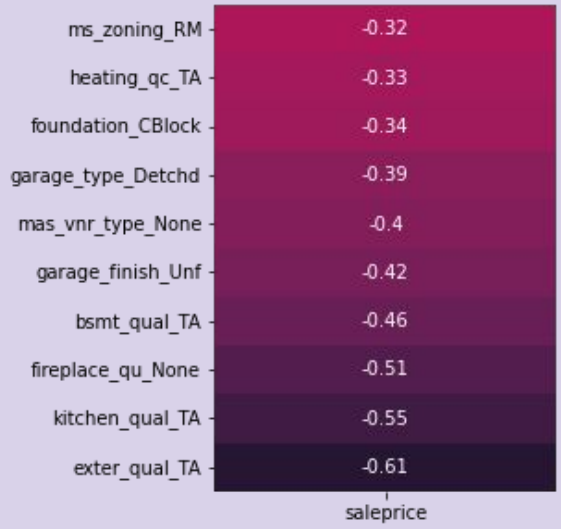
NUM

Numeric type : Impute with ‘0’ ; And use all of numeric features

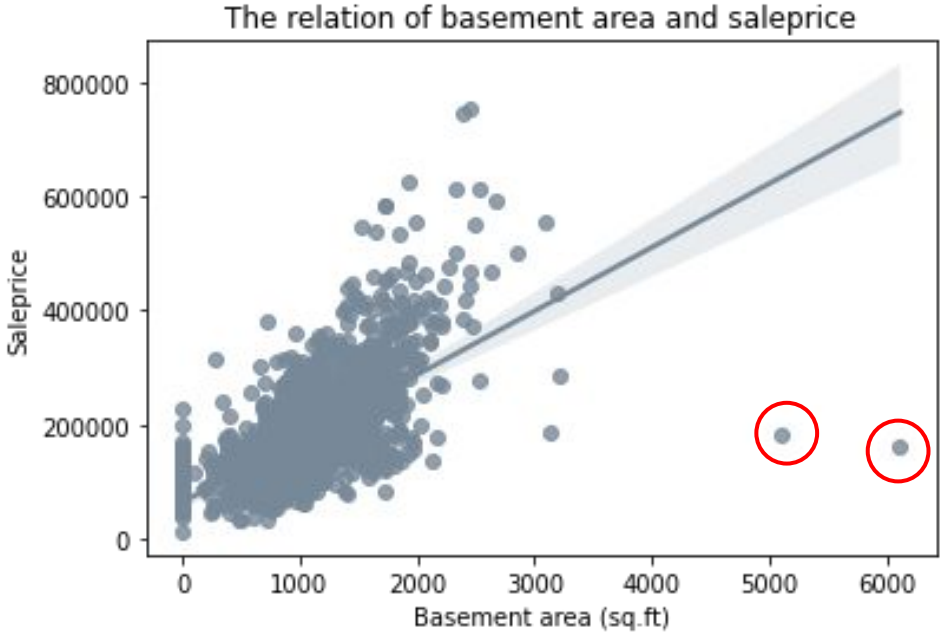
Object type : Impute with ‘NONE’ ;
Dummy and use high correlated features to sale price

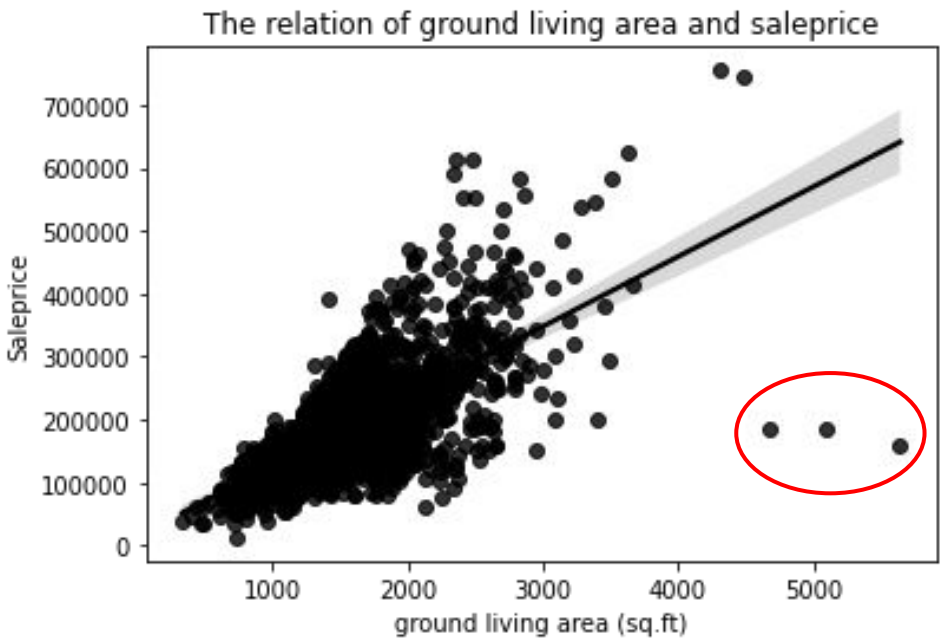
OBJ





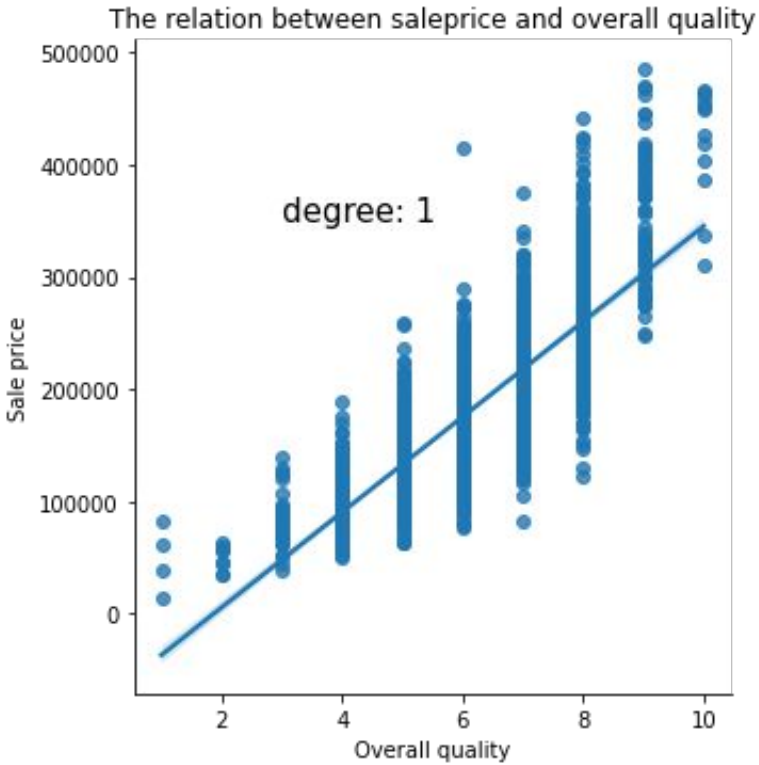
Feature engineering (II) - Outliers

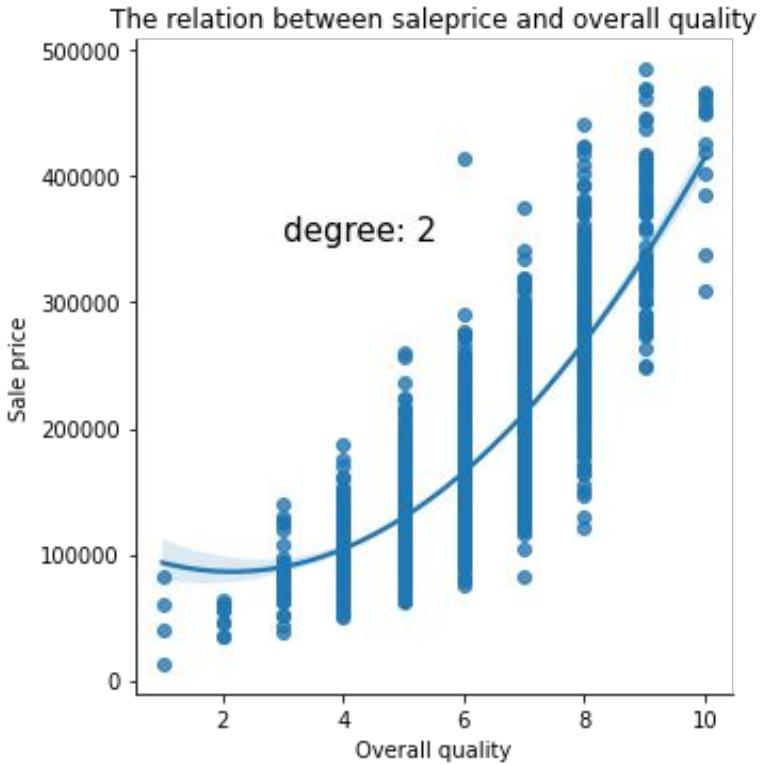




Which are : 'total_bsmt_sf', 'gr_liv_area', '1st_flr_sf', 'saleprice','mas_vnr_area', 'garage_area'

Feature engineering (III) - Polynomial term

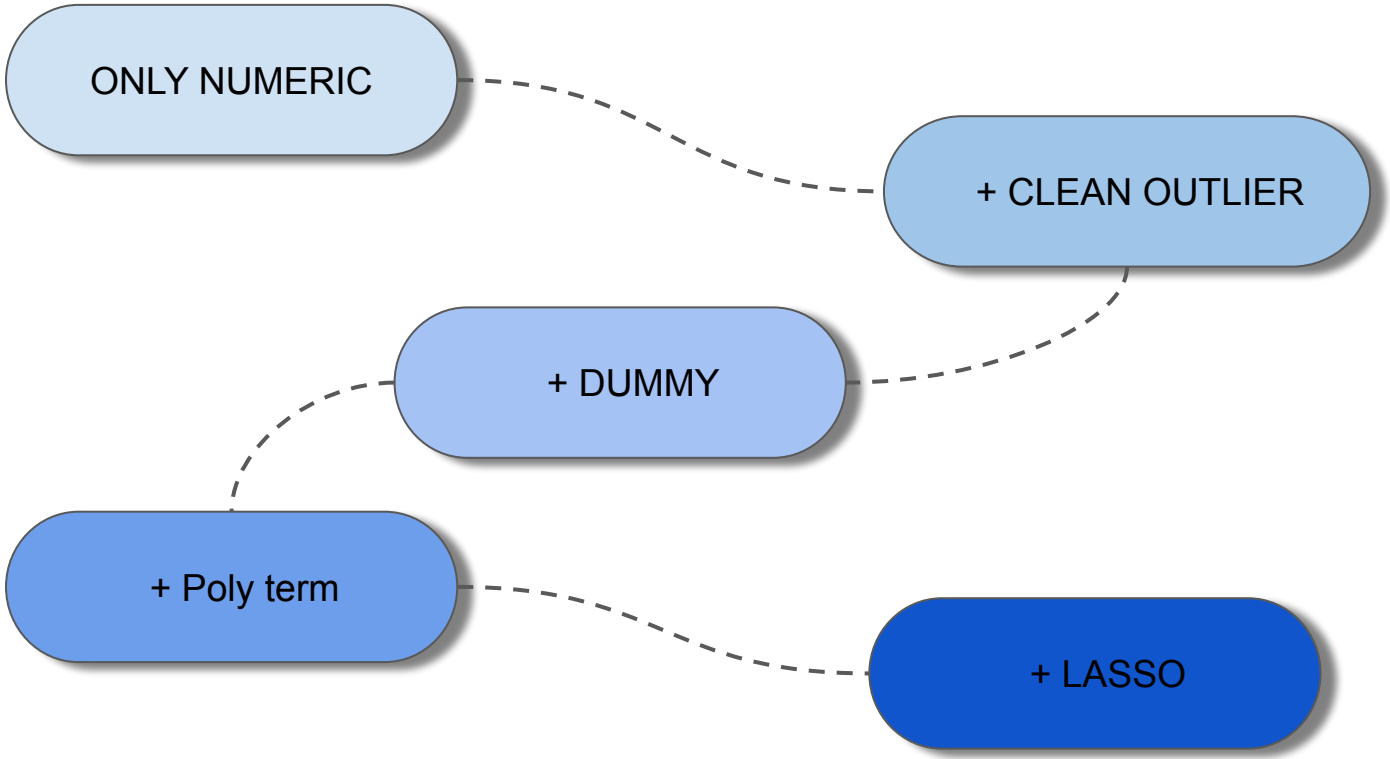




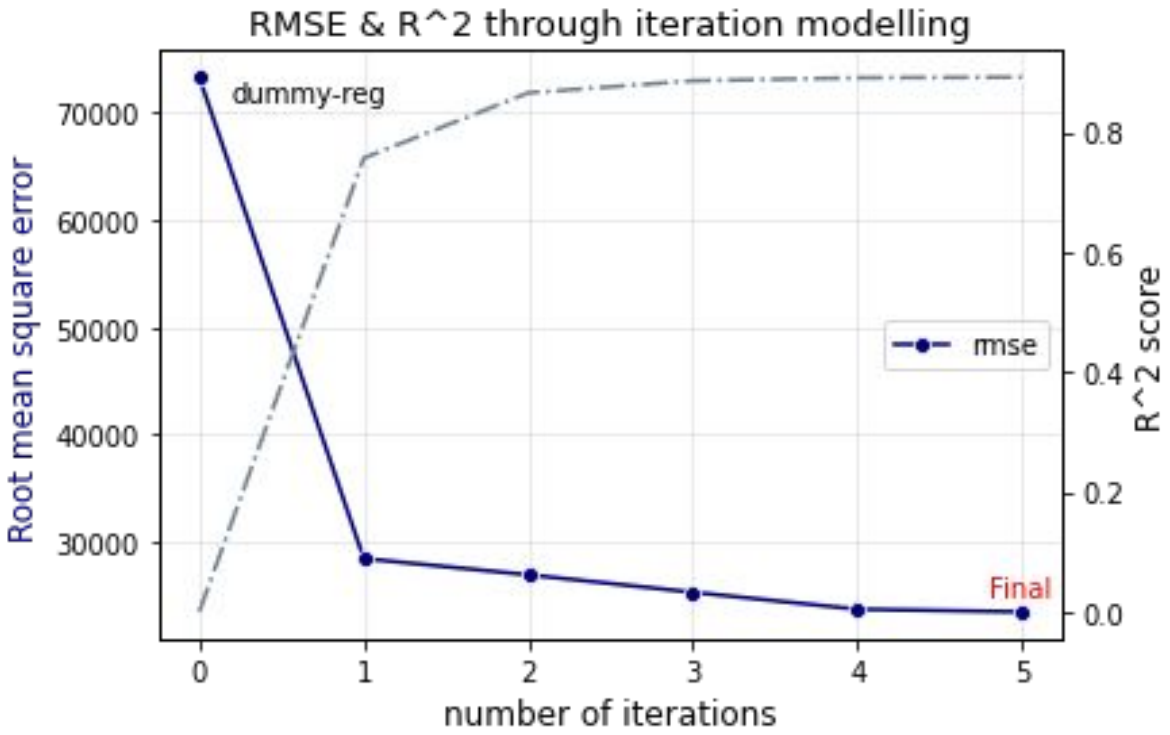
Which are : 'year_built' and ‘ overall_qual’

MODELING FLOW - 5 STEPS

- Using Train-Test split to split the data
- Evaluate the model by cross-validation



EVALUATION

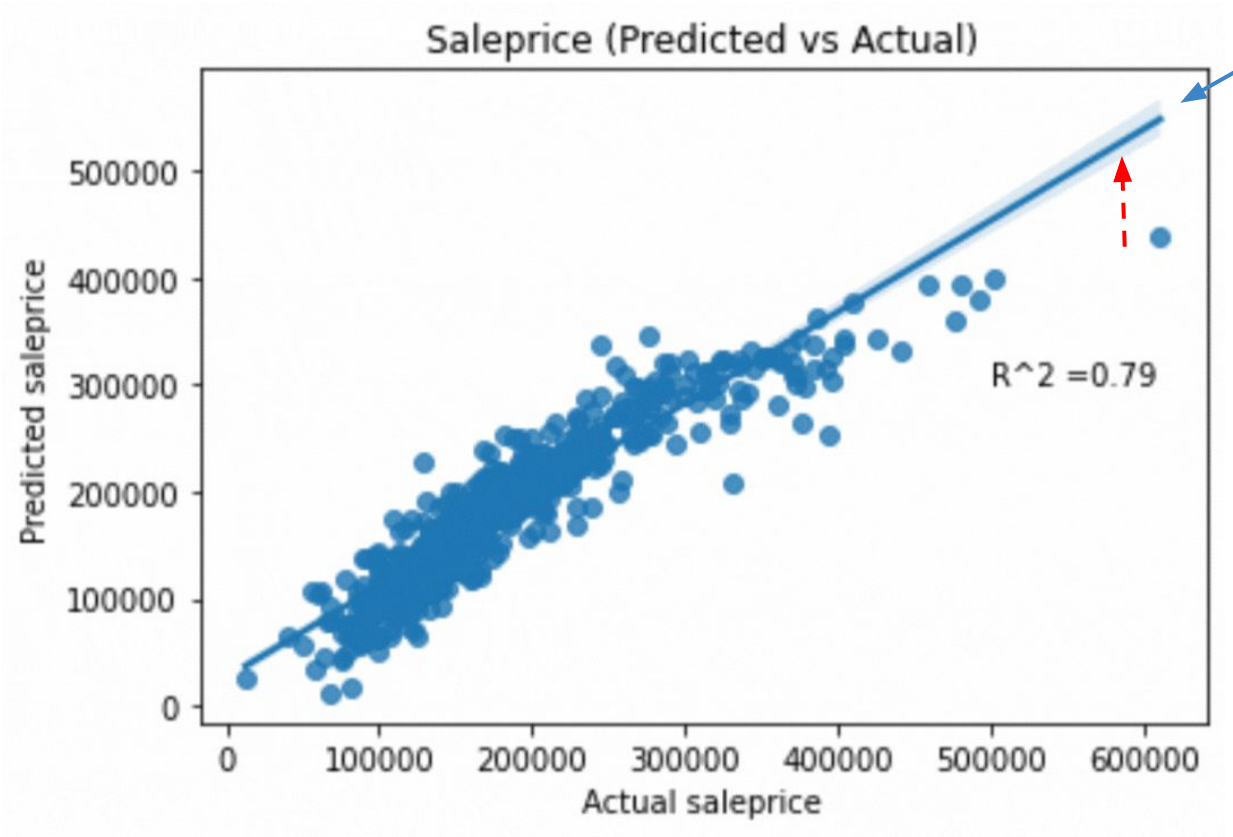


The development of modeling performance

Iterations	RMSE	R^2
Dummyreg	73220	0.000
1	28503	0.757
2	26997	0.865
3	25373	0.886
4	23797	0.891
5	23555	0.892

Final model : RMSE decreased from first model around 5000

EVALUATION (II) - Actual vs Predicted of Sale price

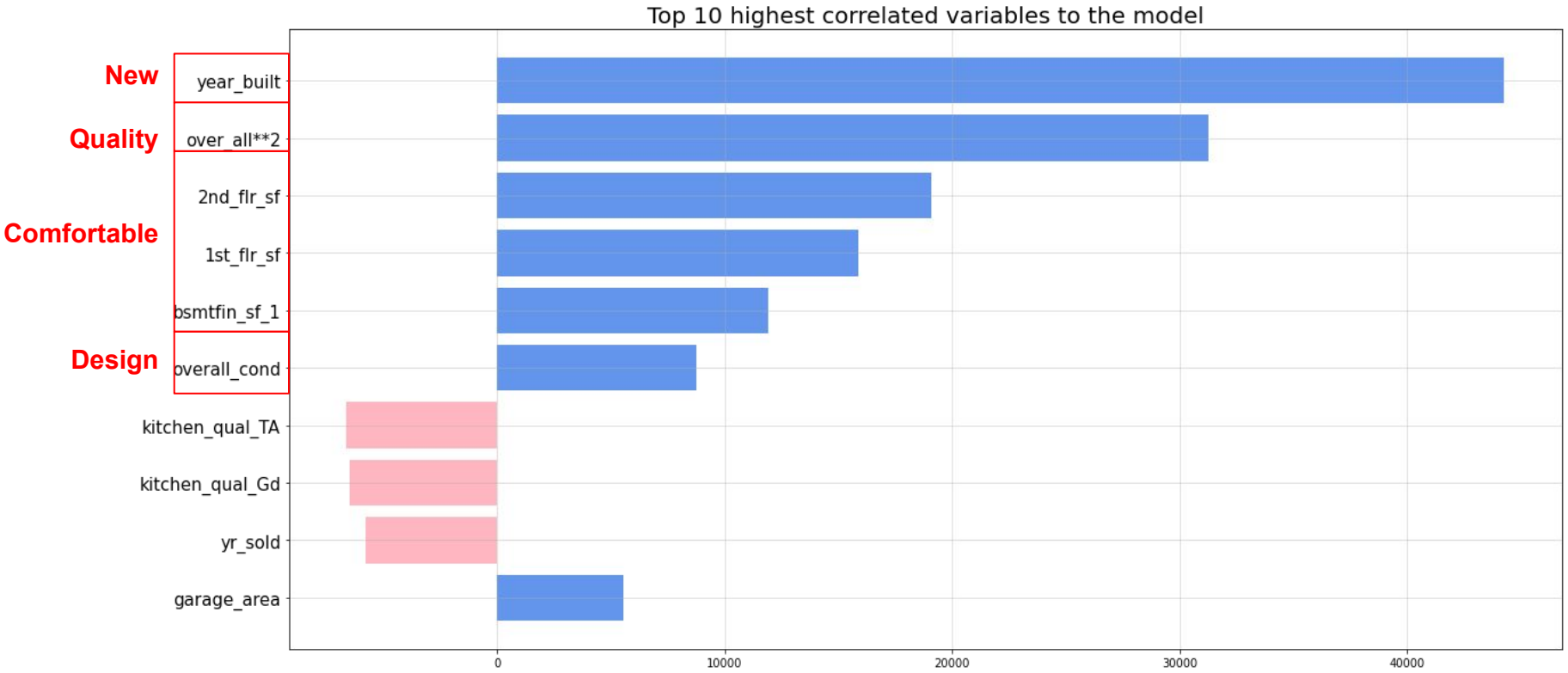


The reference line :
Actual price = Prediction price

- **This developing route of the model will increase the accuracy of sale price prediction, as the increased of R^2**

List of **high correlated variables** to the model

Top 4 things that sale price rely on : New, Quality, Comfortable, and Design.



Conclusion :

- The top 4 things that influence the sale price are: new (year built); quality (overall quality); comfortable (size of living area); and design (overall condition).
- The model has 60 features , including 41 numeric features and 19 dummied features.
- This model includes the use of train-test split to split the data for training and cross validation to test the performance of the model.
- This model includes cleaning outliers, adding polynomial terms & dummy variables, and regularization by LASSO.
- R^2 : 89.2% of the variability in sale price of ames houses can be explained by this model.
- RMSE : The final model can reduce RMSE by around 5000. As a result, the price predicted by this model may differ from the actual value of around 23555 dollars.

Recommendation :

- should add more the interaction term, and other variables

THANK YOU

FOR LISTENING

(KHOB KHUN KRUB)