Housing Price Prediction

Group 4

Clement, Gilbert, Kah Beng

Introduction

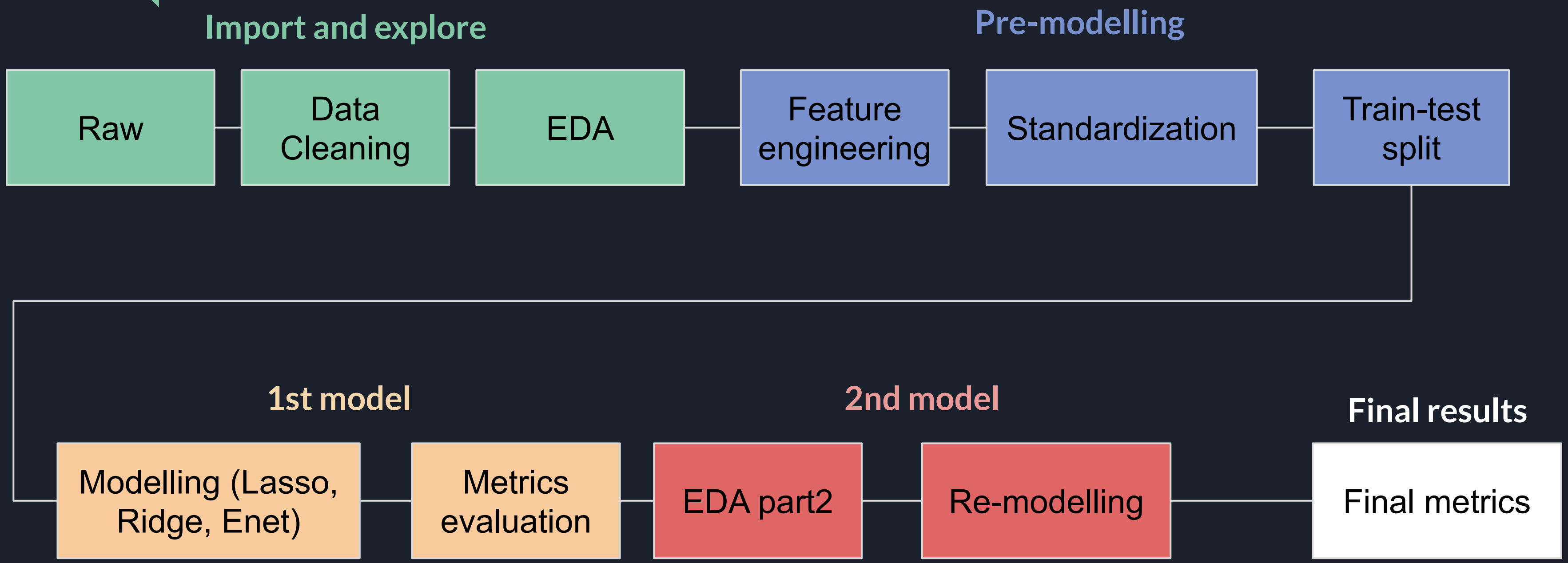
Context

We are tasked to assist property agents in **creating models** for predicting the **house prices** in Ames (Iowa, USA) in order to generate baseline pricing for valuation of the house.

Problem Statement

- Which **regularization** (L1, L2, ENet) resulted in the **best model**?
- Which **features** can best predict house prices?
- Investigate **trade-off** between **complexity** versus the **accuracy** (**R2**)
- How to **improve** upon the initial model

Process



Import and explore

Raw

Data Cleaning

EDA

Pre-modelling

Feature engineering

Standardization

Train-test split

1st model

Modelling (Lasso, Ridge, Enet)

Metrics evaluation

2nd model

EDA part2

Re-modelling

Final results

Final metrics

Data Cleaning

Dealing with 2 types of missing values

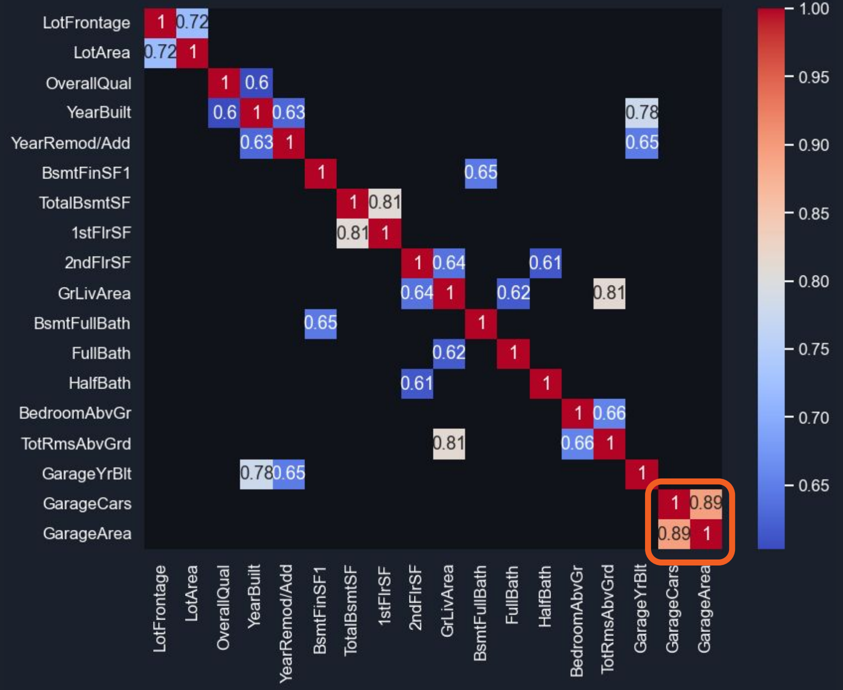
- Missing value because **the feature does not exist** (Garage, Fireplace, Fence)
  - These were imputed using NA or 0 depending on the feature type
- “Genuine” missing values like Lot Frontage was imputed using linear regression based on Lot area

Removing non-informational dimensions like PID

KEY FACT - NO FEATURE WAS DROPPED DUE TO MISSING VALUES

EDA

Checking for multicollinearity

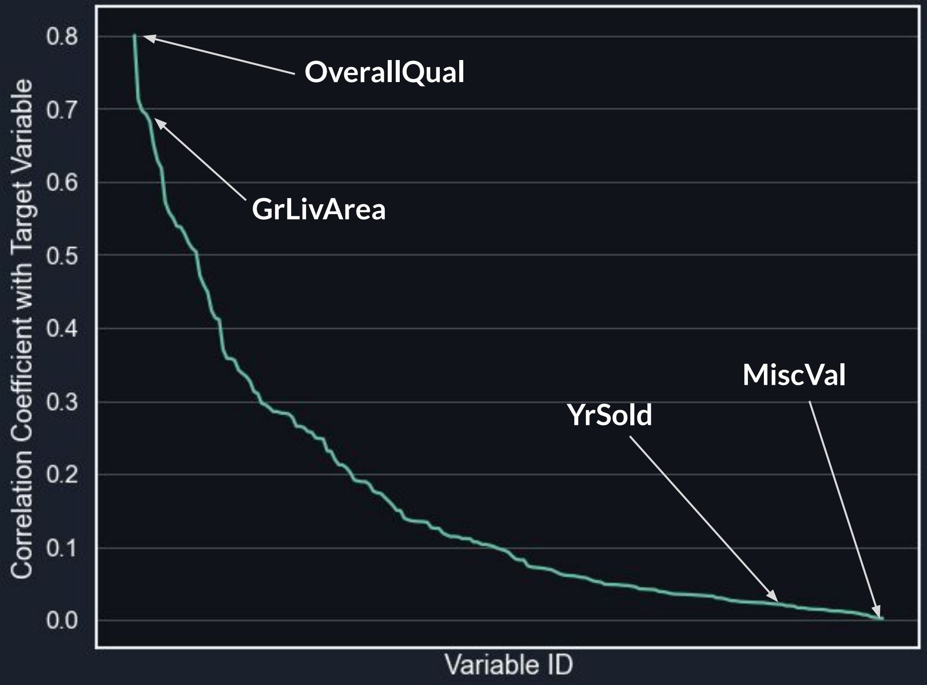


	LotFrontage	LotArea	OverallQual	YearBuilt	YearRemod/Add	BsmtFinSF1	TotalBsmtSF	1stFlrSF	2ndFlrSF	GrLivArea	BsmtFullBath	FullBath	HalfBath	BedroomAbvGr	TotRmsAbvGrd	GarageYrBlt	GarageCars	GarageArea
LotFrontage	1	0.72																
LotArea	0.72	1																
OverallQual			1	0.6														
YearBuilt			0.6	1	0.63											0.78		
YearRemod/Add				0.63	1											0.65		
BsmtFinSF1						1					0.65							
TotalBsmtSF							1	0.81										
1stFlrSF							0.81	1										
2ndFlrSF									1	0.64			0.61					
GrLivArea									0.64	1		0.62			0.81			
BsmtFullBath						0.65					1							
FullBath										0.62		1						
HalfBath									0.61				1					
BedroomAbvGr														1	0.66			
TotRmsAbvGrd										0.81				0.66	1			
GarageYrBlt				0.78	0.65											1		
GarageCars																	1	0.89
GarageArea																	0.89	1

GarageCars was removed

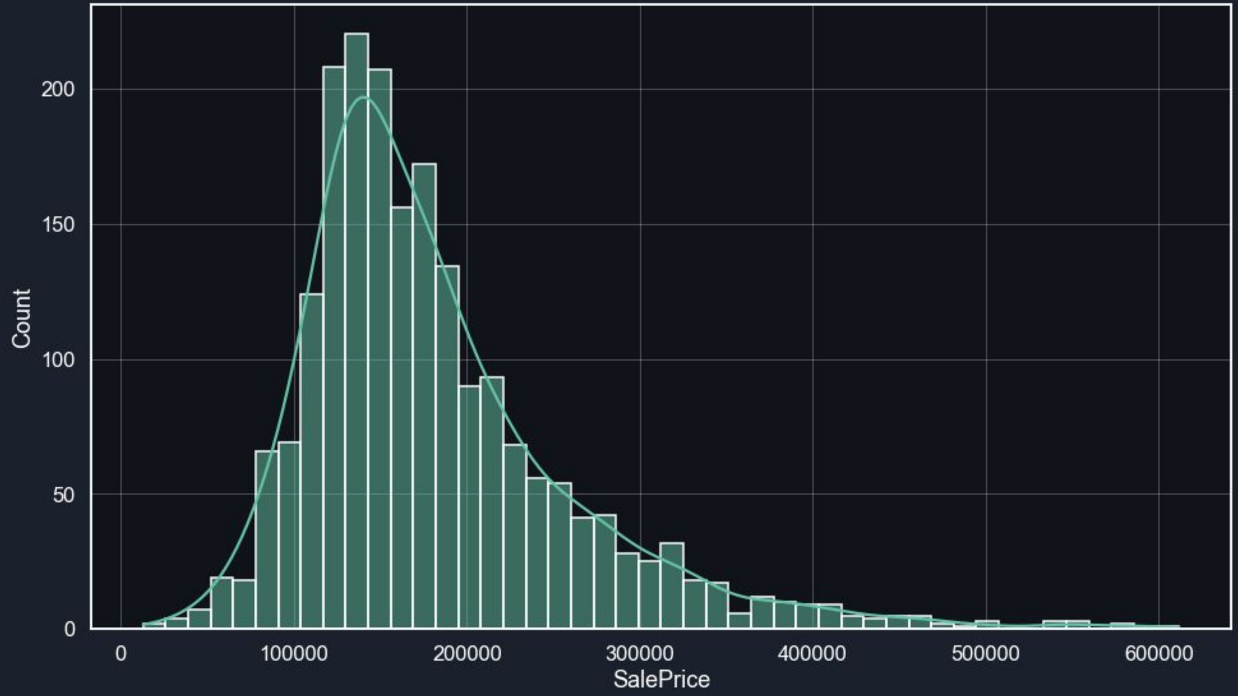
The burning question: What **correlates** to SalePrice?

Top 10 correlated features: 'OverallQual', 'ExterQual', 'GrLivArea', 'KitchenQual', 'BsmtQual', 'GarageArea', 'TotalBsmtSF', '1stFlrSF', 'YearBuilt', 'GarageFinish'



Distribution of SalePrice

Our target variable exhibits a **normal distribution shape**, slightly **skewed** by extremely high value houses



The main benefit of having a normally distributed variable is that it gets easier to describe we see here that the **mode, mean and median almost overlap**

Feature Engineering

Categorical variables

Mapping ordinal variables to numbers

Ex: quality features like Bsmt Qu expressed in **“Good / fair / poor” terms got assigned numerical values instead**

One-Hot Encoding the remaining (nominal) variables

Ex: **features not following any order**(neighborhood, lot shape ...)

Modelling

Focused on models with regularisation techniques applied. **Best models** from cross-validated **grid search**:

Lasso: penalty (α) = 0.00681

Ridge: α = 464.16

Elasticnet: α = 0.1; l1_ratio = 0.05 (penalty combination is mostly Ridge)

Standardisation of features was done using **standard scalar** to ensure penalty is applied equally to all features

Train-Test split ratio of **70/30** was used for all models

Models are evaluated on **MAE, RMSE, R2, R2_adj** when predicting **test dataset**

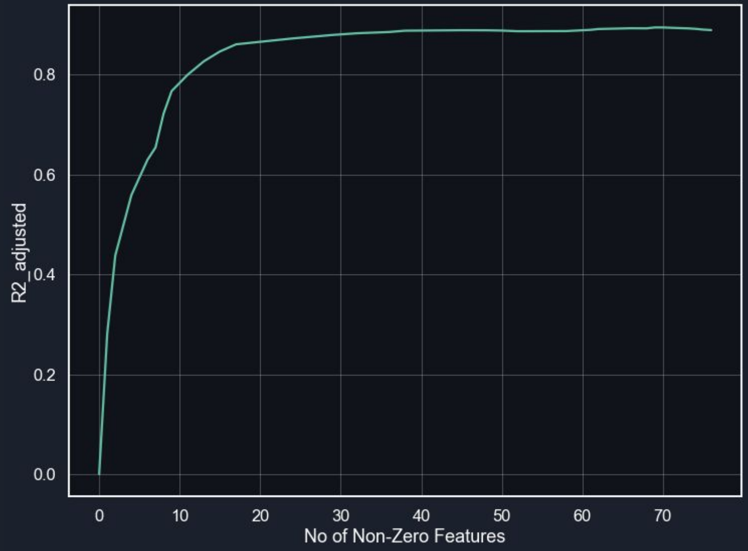
Model Comparison

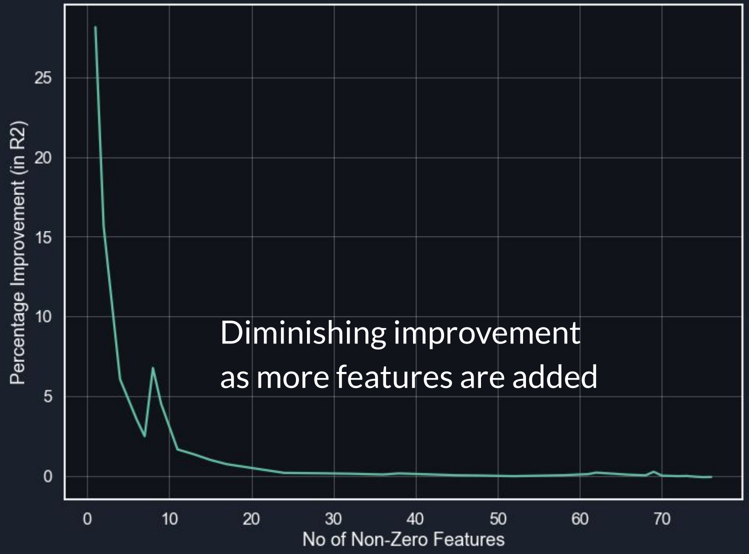
Lasso model has the smallest residual and highest R2 with least features

Regression Type	Non-Zero Variables	Non-Zero Features	MAE	RMSE	R2 (test)	R2_adj (test)
Baseline (mean)	0	0	0.725	0.966	-0.001	-0.001
Simple Linear Regression	195	77	9.4e+07	1.7e+09	-2.7e+18	-3.0e+18
Ridge	191	76	0.233	0.328	0.884	0.872
ElasticNet	126	59	0.230	0.323	0.887	0.880
Lasso	**103**	**57**	**0.226**	**0.316**	**0.892**	**0.887**

Accuracy-Complexity Trade-off

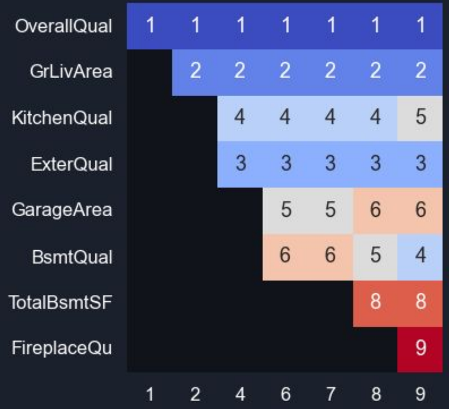
- A property for LASSO is that it is able to **reduce the number less relevant features** to 0 based on the α
- By varying α, we can examine the variations in **number of features selected**, and the **resulting R2**

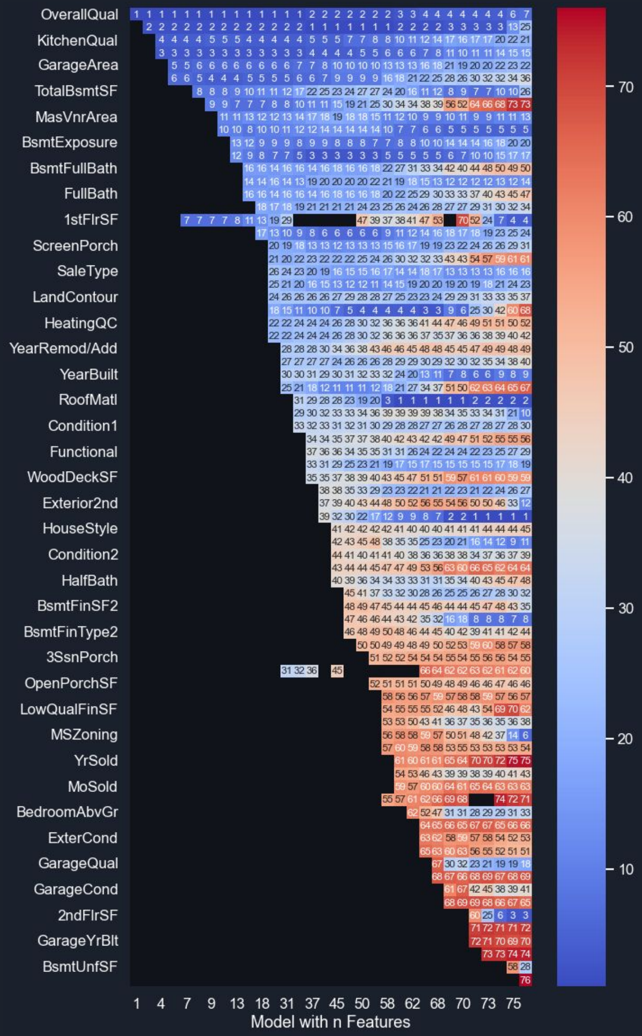




Diminishing improvement
as more features are added

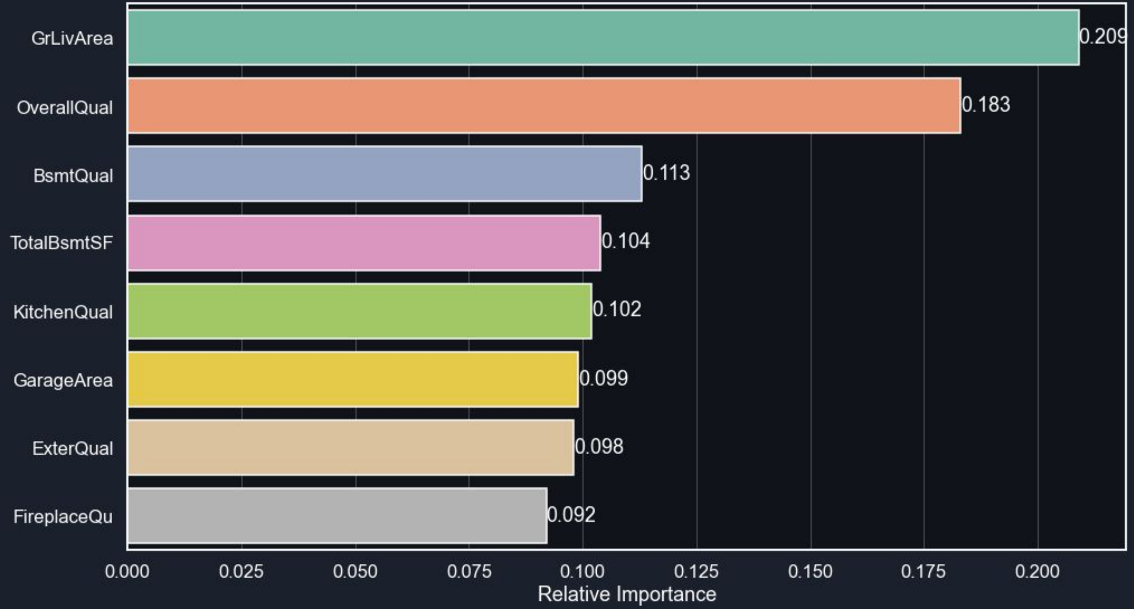
Selected Features





EDA Part 2: Feature Importance

Features with highest regression coeff. can be broadly categorised into house size (square area of various rooms) and house quality (quality of rooms and facilities)

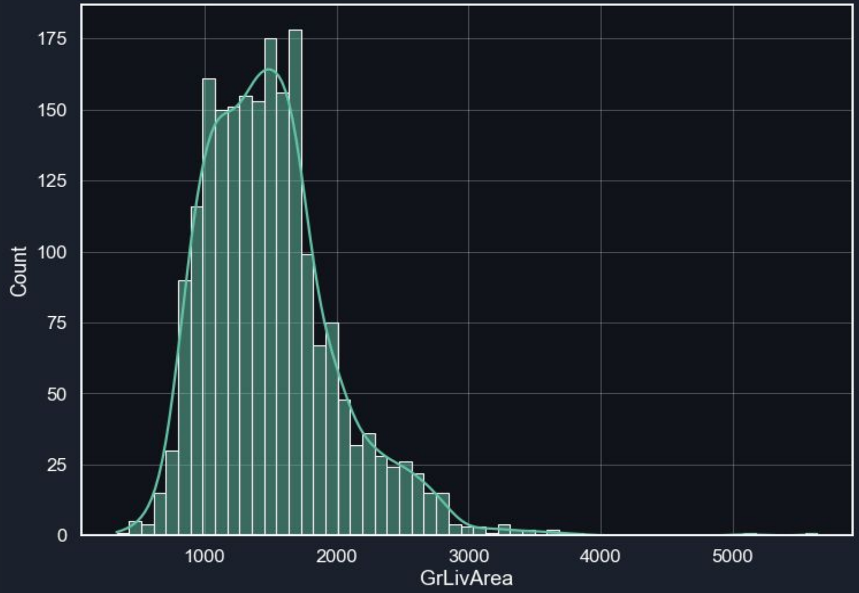


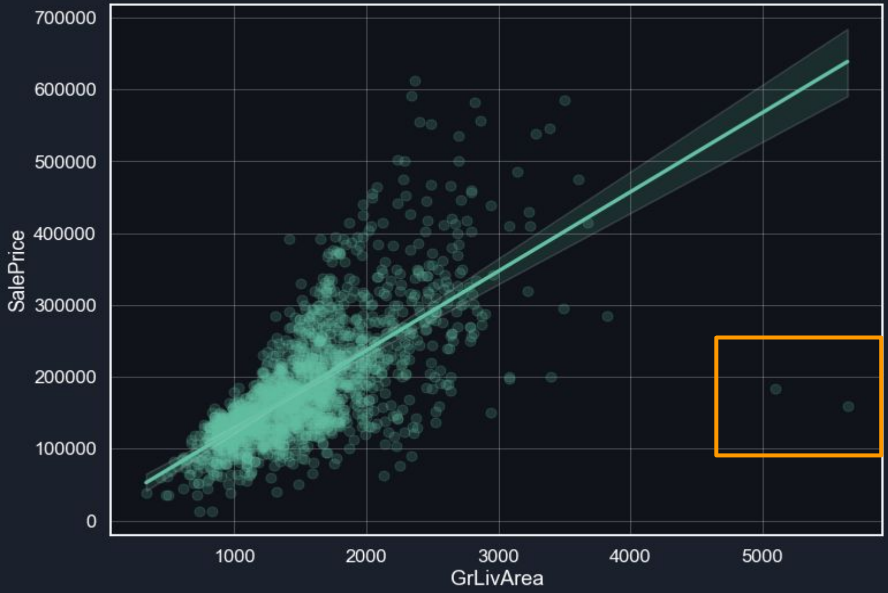
1. Above-Ground Living Area

Distribution of above-ground living area is skewed to the right similar to sales price

Scatter plot for above-ground living area and sale price shows positive relationship

2 outliers in the above-ground living area feature

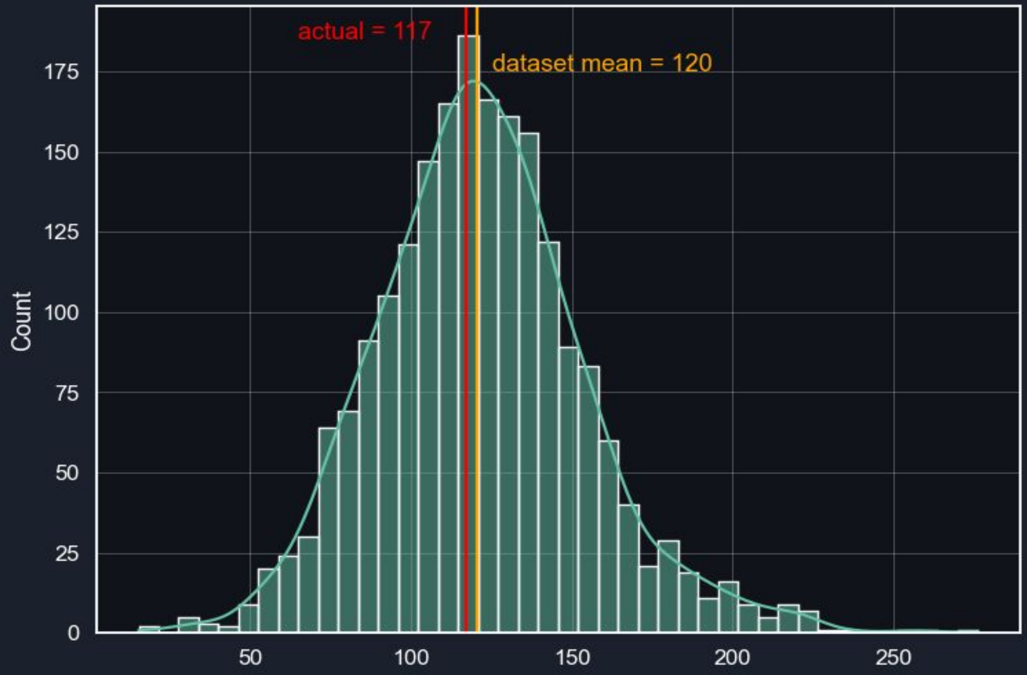




1. Above-Ground Living Area

Distribution of sale price per square feet resembles a normal distribution

Dataset median is also close to the actual median prices

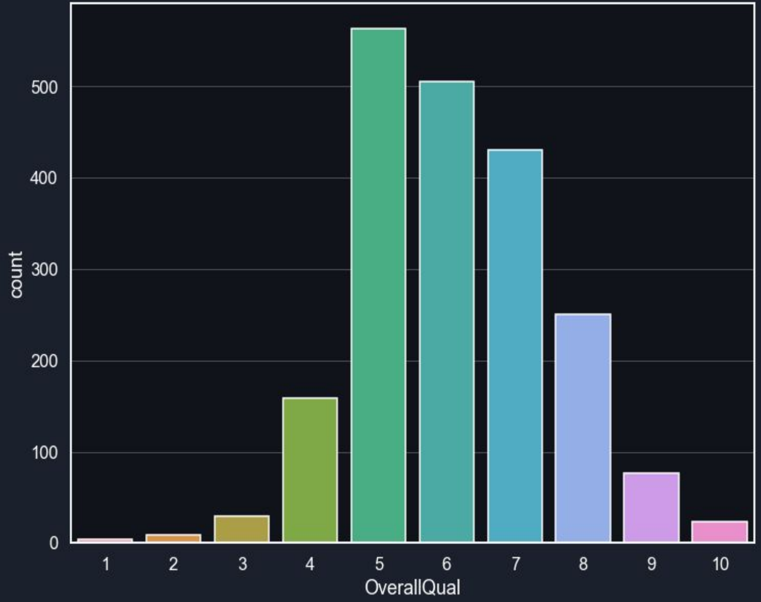


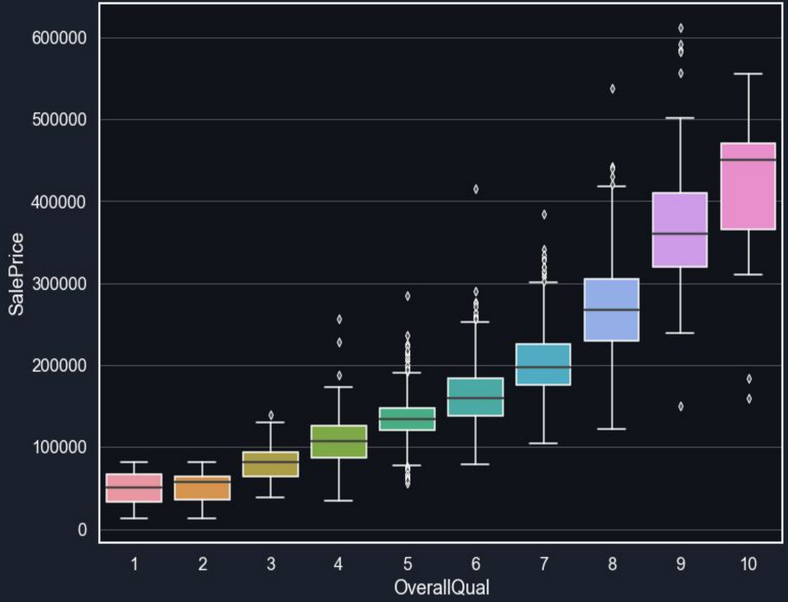
2. Overall Quality

Distribution of overall quality is also skewed to the right

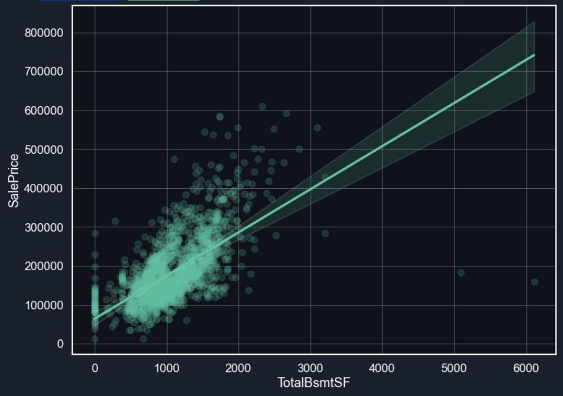
Skew is reasonable since poor quality houses are unlikely to be purchased

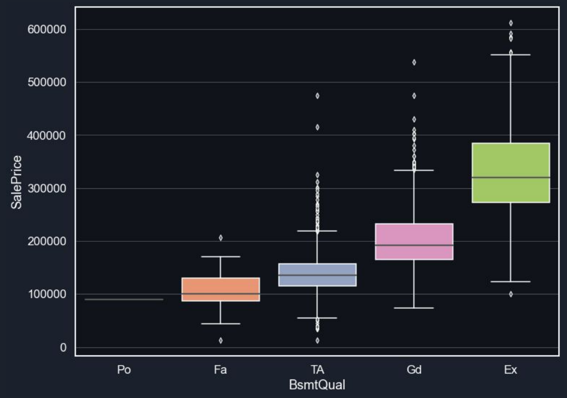
Relationship with sale price does not seem linear

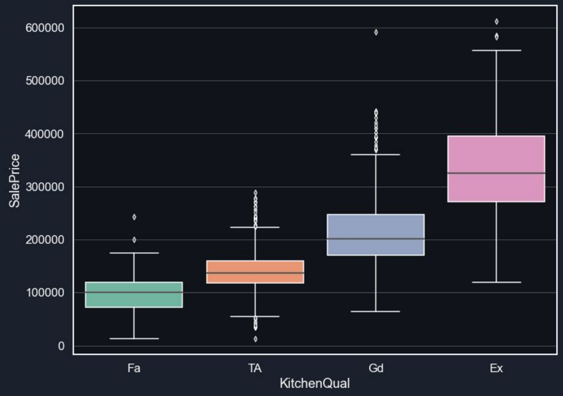


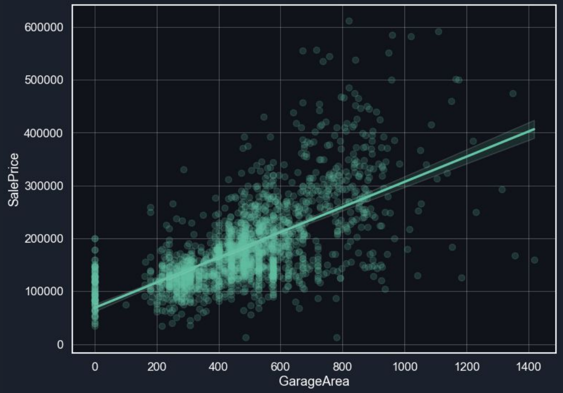


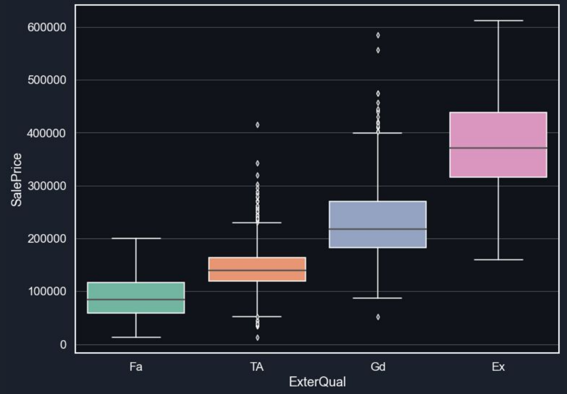
The rest of the top 8 features:

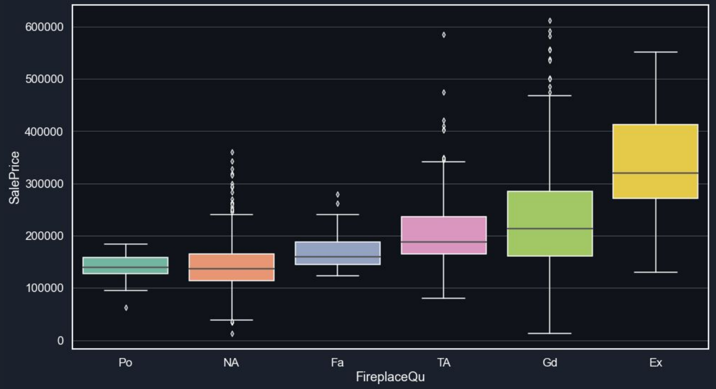








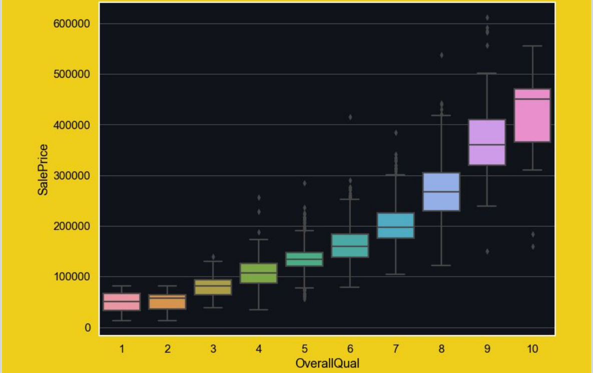




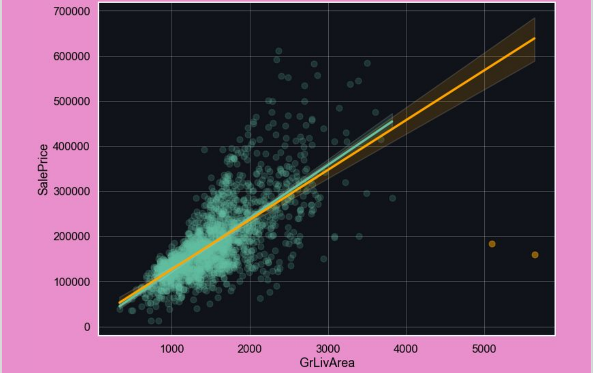
Improving the Model

Based on the 2nd EDA, we observed potential methods for **improving our model**:

Adding **Non-linearity** for Ordinal Categories



Removing **Outliers**



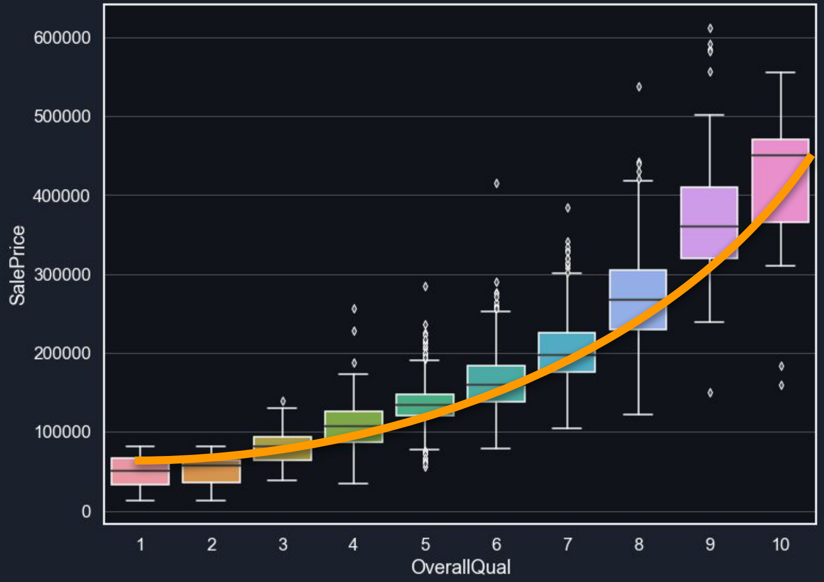
Combining Both Methods

1. Non-Linearity for Ordinal Variables

The ordinal variables (e.g.: Quality and Condition) are scored on a linear scale:

- Overall: 1, 2 ,3 ,4, 5, 6, 7, 8, 9, 10
- Others: NA, Po, Fa, TA, Gd, Ex ⇒ 0, 1, 2, 3, 4, 5

However, the response of these scores (against Sale Price) are **NOT LINEAR**



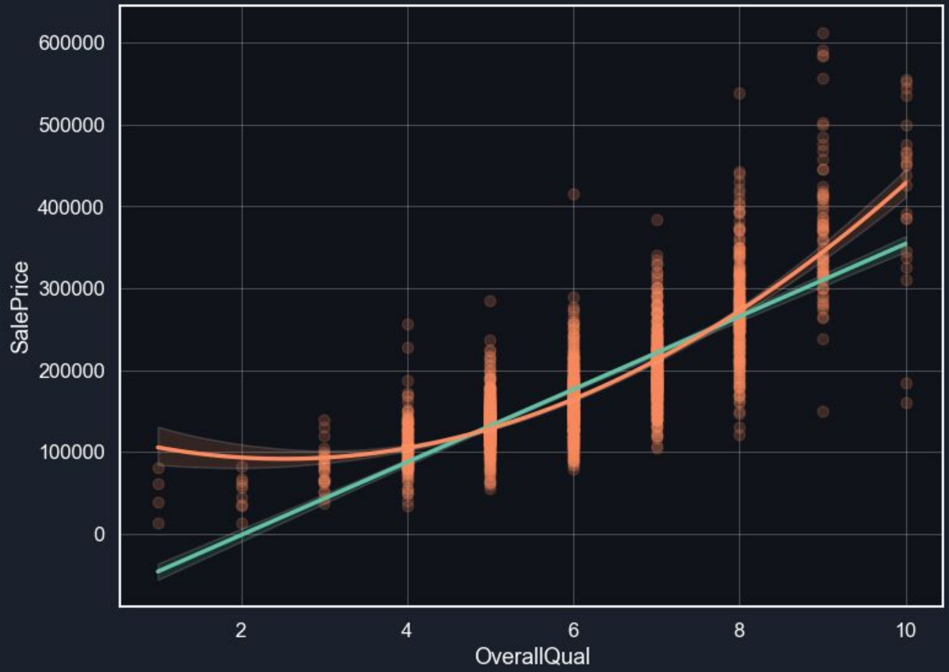
1. Non-Linearity for Ordinal Variables

Looking at the distribution of the data, we can estimate the response of OverallQuality (and other quality-related columns) to be **quadratic (i.e.: *y~x*^{2})**

Therefore, we will now **square** ALL quality-related columns and re-run the regression model

Result:

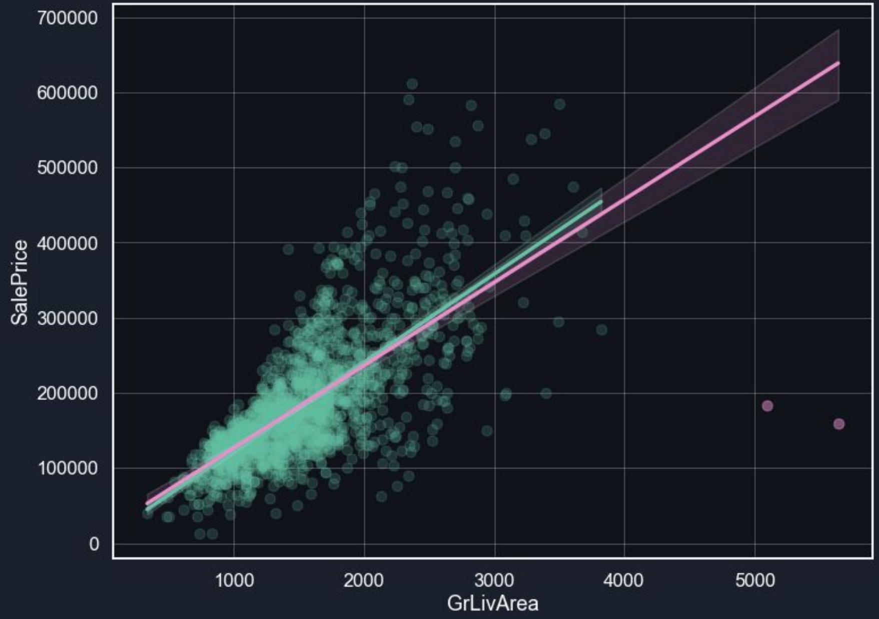
Regression Type	Non-Zero Features	R2_adj
Lasso	57	0.887
Lasso *(non-linear ordinal variables)*	56 **(-1)**	0.897 **(+0.010)**



2. Removing Outliers

Observing the scatter plot between SalePrice and GrLivArea, we can see that there are **two outliers**Regression-based models are typically very sensitive to outliers

We will now re-run the regression model after **removing the outliers**



Result:

Regression Type	Non-Zero Features	R2_adj
Lasso	57	0.887
Lasso *(removed outliers)*	55 **(-2)**	0.910 **(+0.023)**

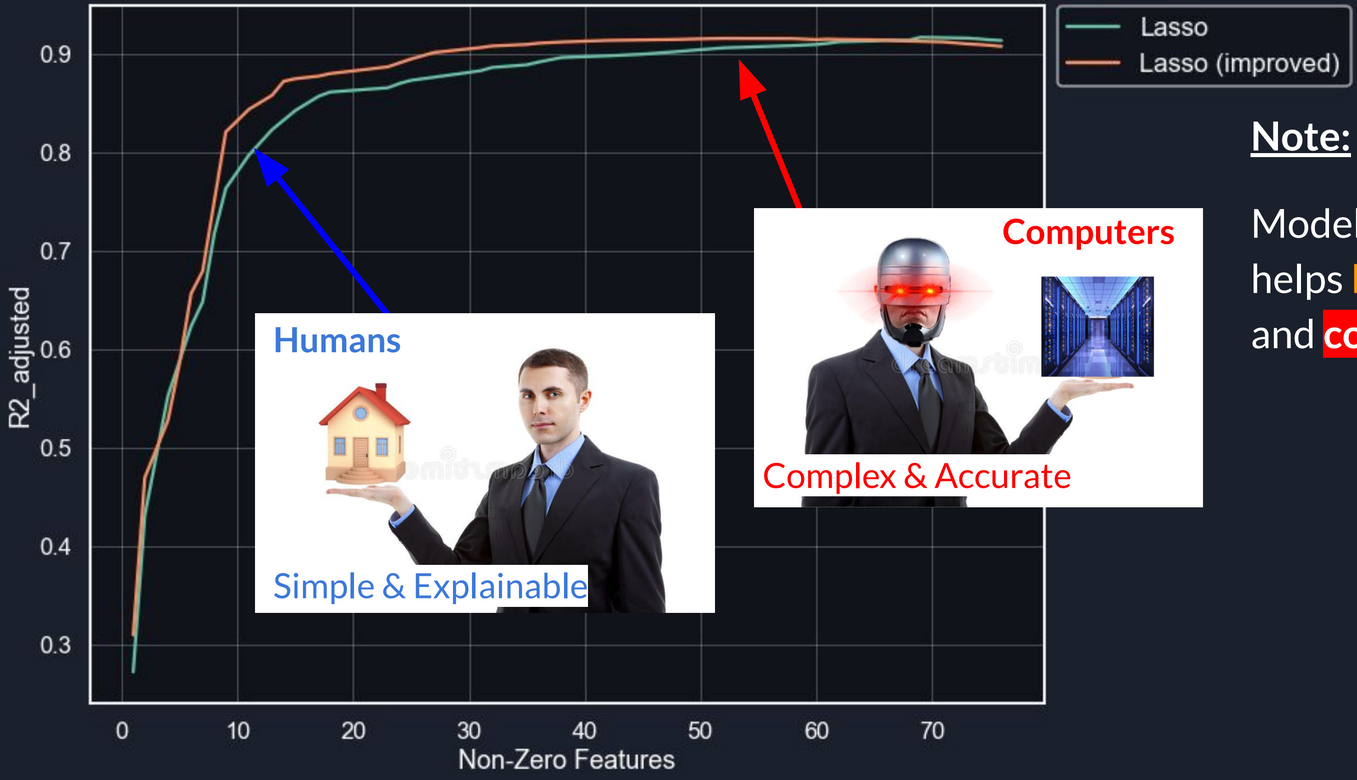
Model Summary

Regression Type	Non-Zero Variables	Non-Zero Features	MAE	RMSE	R2	R2_adj
Baseline (mean)	0	0	0.725	0.966	-0.001	**-0.001**
Simple Linear Regression	195	77	9.4e+07	1.7e+09	-2.7e+18	-3.0e+18
Ridge	191	76	0.233	0.328	0.884	0.872
ElasticNet	126	59	0.230	0.323	0.887	0.880
Lasso	103	57	0.226	0.316	0.892	0.887
Lasso ***(non-linear ordinal variables)***	98 **(-5)**	56 **(-1)**	0.218	0.302	0.902 **(+0.010)**	0.897 **(+0.010)**
Lasso ***(removed outliers)***	92 **(-11)**	55 **(-2)**	0.219	0.307	0.914 **(+0.022)**	0.910 **(+0.023)**
Lasso ***(non-linear + outliers)***	91 **(-12)**	52 **(- 5)**	0.215	0.298	0.919 **(+0.027)**	0.915 **(+0.028)**

Which model to choose?

Short answer: Lasso...

Long answer:



Humans

Simple & Explainable

Computers

Complex & Accurate

Note:

Model improvement helps BOTH the simple and complex models

Summary

- **Lasso** resulted in the **highest accuracy** and the **lowest complexity** out of all the tested regularization methods (**L1**, L2, ENet).

- **Main predictor** of house pricing:
  - Above-Ground **Square Footage**
  - Overall **Quality**
  - Other **area- and quality- related** metrics for specific locations / sections

- **Diminishing return** between number of features included (**complexity**) and R2_adjusted (**accuracy**)

- Methods for **improving** predictions:
  - adding **non-linearity**
  - removing **outliers**

Future Works

The following items may be investigated for future works:

- Comparing selected variables from **Lasso** against those from **SelectKBest**

- Add **interaction terms** for several variables

- Use of more complex models to improve accuracy, such as using **Tree-Based models**

- Adjusting SalePrice for **inflation** (depending on the YearSold)

- See impact of h**ousing bubble**, if any
    - (compare prices of house sold before and after 2008)

Summary and Conclusion

In this project, we are tasked to assist property agents in creating models for predicting the house prices in Ames (Iowa, USA).

- **Data Pre-Processing:** The dataset were processed through the following processes:
  - Data cleaning were conducted on the provided data set. This is done by imputing values for all of the Null values in the data set.
  - All of the ordinal categorical variables were converted to its equivalent numerical scoring
  - All of the nominal categorical variable over one-hot encoded
- **Modelling:** Three types of regularization technique (Lasso, Ridge, and ElasticNet) were applied on the data set. Out of the three techniques, Lasso resulted in the high accuracy and the lowest number of features included. The Lasso model achieved an R2 of 0.892 in predicting the test dataset.
- **Most important features:** The two features that can best predict house price is the above ground living area as well as the overall quality of the house. It is followed by a series of features which measures the area and quality of various specific location / aspect of the house.
- **Accuracy-Complexity Tradeoff:** Assuming no overfitting, the accuracy of the model increases as the number of features included increases. This is the trade-off between accuracy and complexity. There is a diminishing returns to adding more features. It was found that using around 10 features already resulted in a model with an R2 of aroudn 0.8. This simple model can be used as a heuristic model for humans when comparing the prices of different houess.
- **Model Improvement:** Two methods for further improving the model were attempted:
  - Converting ordinal variables to non-linear
  - Removing outliers
  - A combination of both
- The techniques mentioned above resulted in significant improvement in accuracy as well as a reduction in number of features selected.
- **Final Model:** The final model has an R2 of 0.919 in predicting the test dataset, while using only ~ 50% of the potential variables from about ~68% of available features.