HOUSE PRICE PREDICTION USING MACHINE LEARNING

A Kaggle Competition

Yutian Zhou

BACKGROUND

- Predict sale price using the Ames housing dataset
- 4 files in total, training and testing datasets are separate
- The goal of the competition is to build a model with the highest predictive accuracy (metric they use: RMSE)
- Personally speaking, I want to achieve the competition goal and model interpretability at the same time as much as possible.

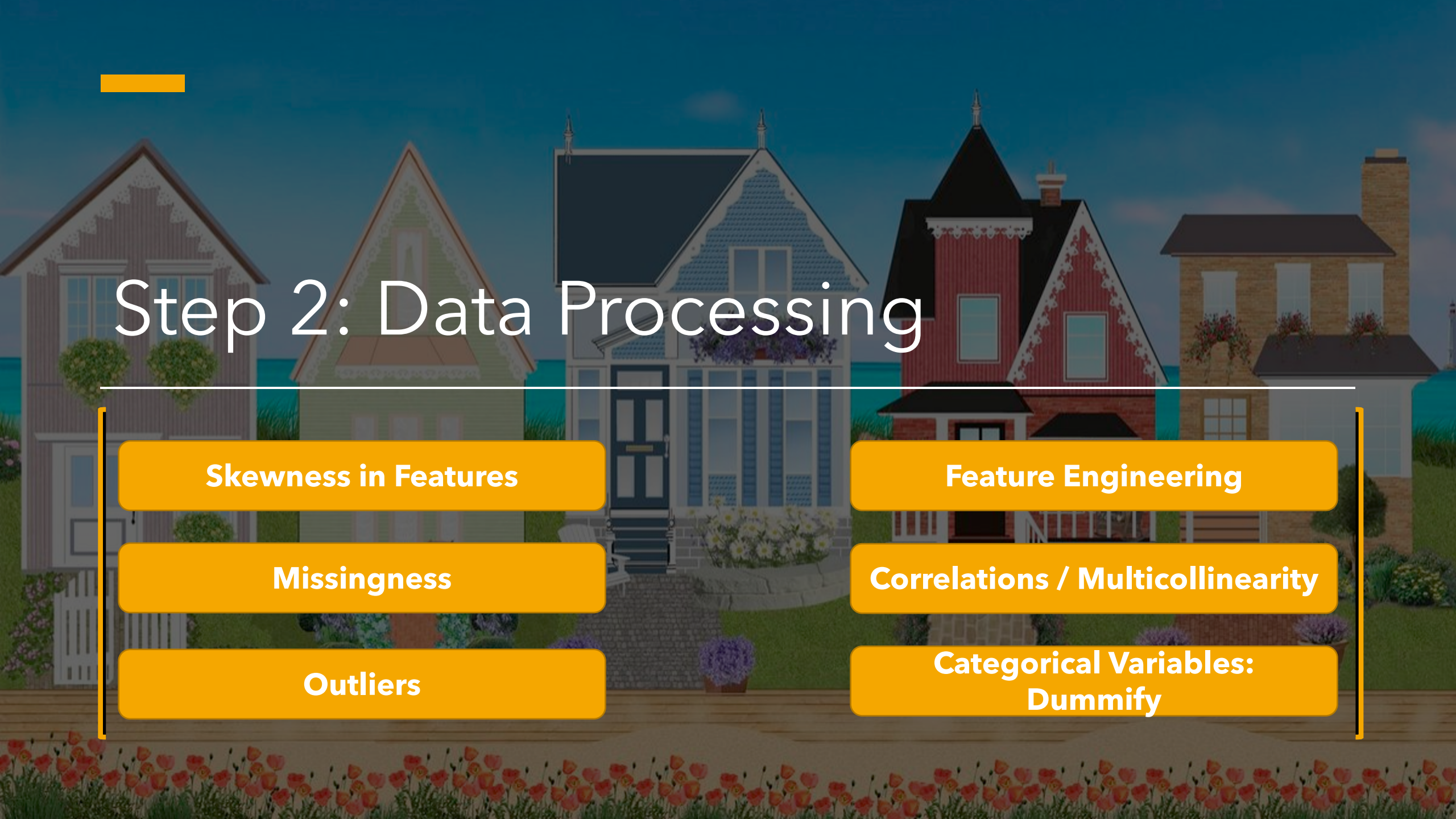
	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	I
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	

DATASET

- 79 predictive variables describing houses in Ames, Iowa
- Contains both categorical and numerical features, mostly categorical variables
- A lot of variables are correlated with each other, and house features are in very detailed descriptions (eg. house square feet are divided into basement square feet, first floor square feet, and second floor square feet ...)

Step 1: Exploratory Data Analysis

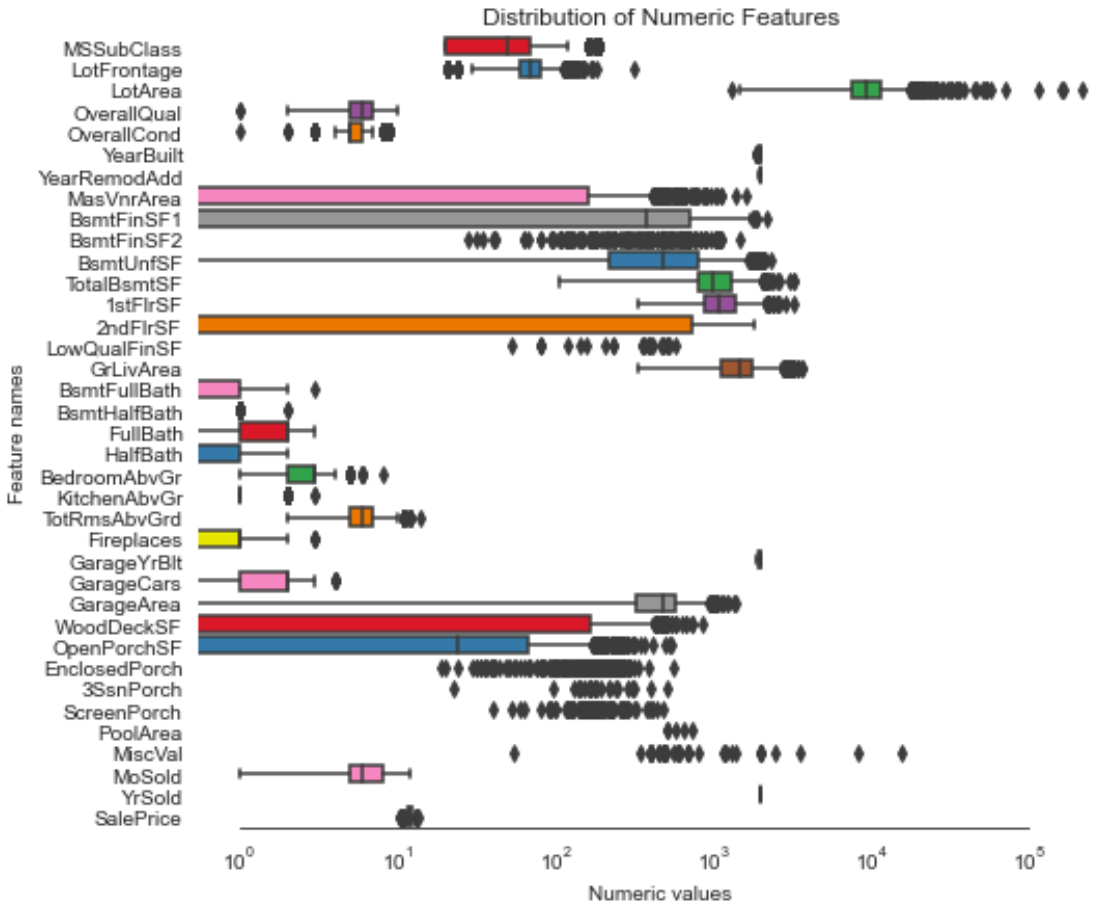
- Univariate analysis on the outcome SalePrice: positively skewed, log transformation to normal distribution
- Bivariate analysis based on some domain knowledge: The price is associated with grand living area, year built, and the overall quality



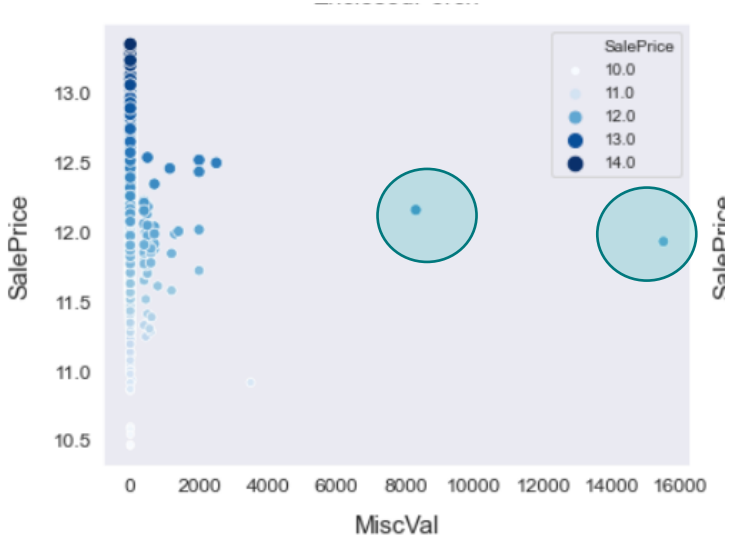
Step 2: Data Processing

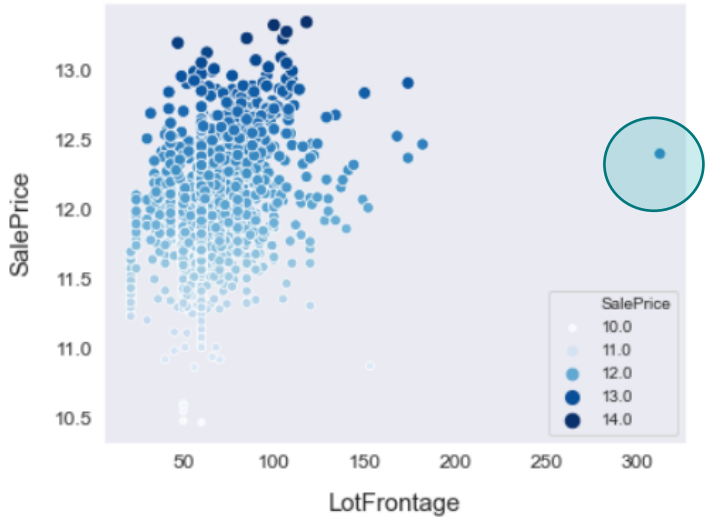
- Skewness in Features
- Missingness
- Outliers
- Feature Engineering
- Correlations / Multicollinearity
- Categorical Variables: Dummify

Skewness in Features



- Found 13 numeric features highly skewed (|skewness| >1)
- Did not transform them in order to keep more result interpretability





Missingness and Outliers

- PoolQC, MiscFeature, Alley, Fence, FireplaceQu, and LotFrontage are the top columns with a lot of missing values, but the null values all represent 'None' according to the documentation
- Other missing values: Grouped by and then impute mode/median
- Removed 7 observations in total which were considered as outliers

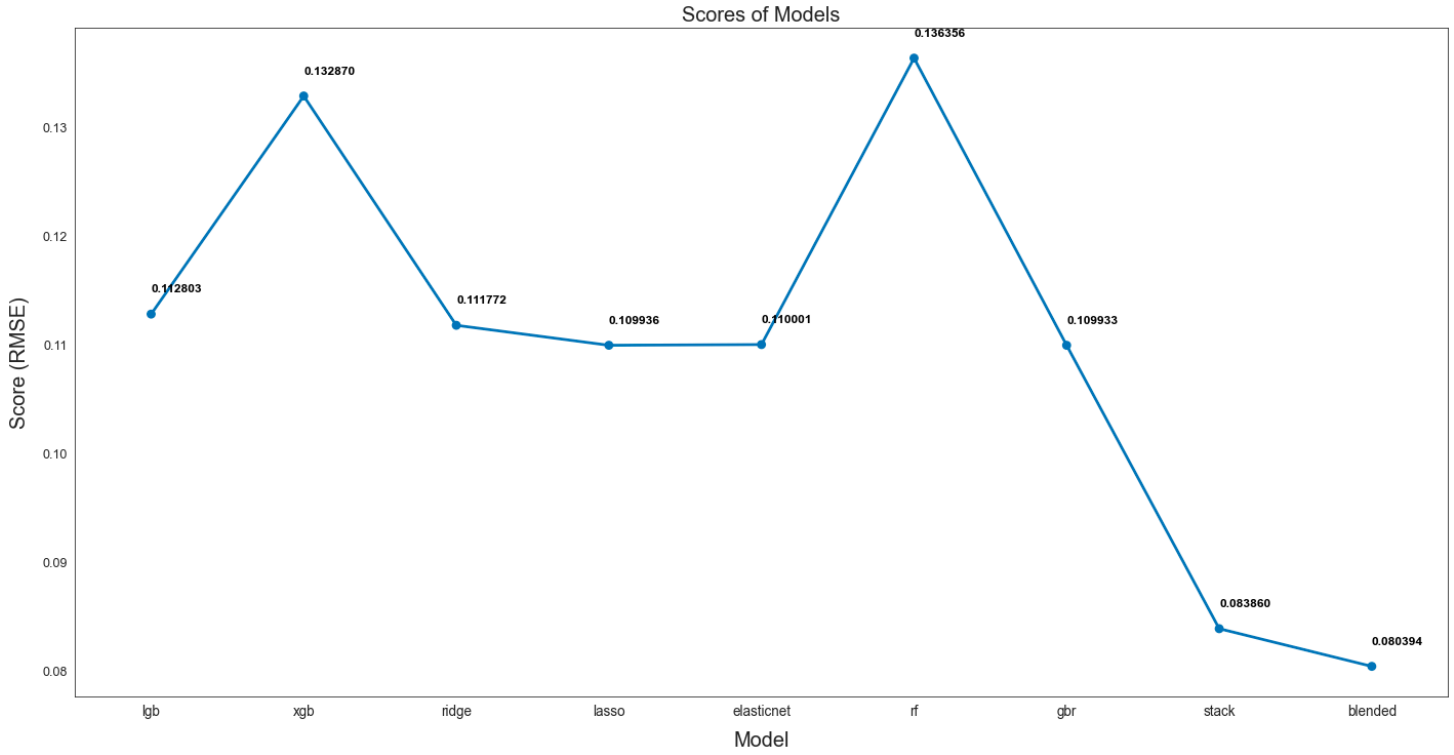
Feature Engineering and Correlations / Multicollinearity

- When combined, they can give us a better picture and make some significant difference on result interpretations
- TotalSF = TotalBsmtSF + 1stFlrSF + 2ndFlrSF (controversial)
- Total_Bathrooms = FullBath+ 0.5 * HalfBath +BsmtFullBath + 0.5 * BsmtHalfBath
- Total_porch_sf = OpenPorchSF+ 3SsnPorch+EnclosedPorch+ ScreenPorch +WoodDeckSF
- Multicollinearity exists among features that describe the same aspect of the house, so for regression models, only considered regularized regressions


Step 3: Model Training and Selection
Penalized Regressions
Gradient Boosting
Light GBM，XGBoost
Random Forest
Stacked Model
Blended Model

Level 1 Models

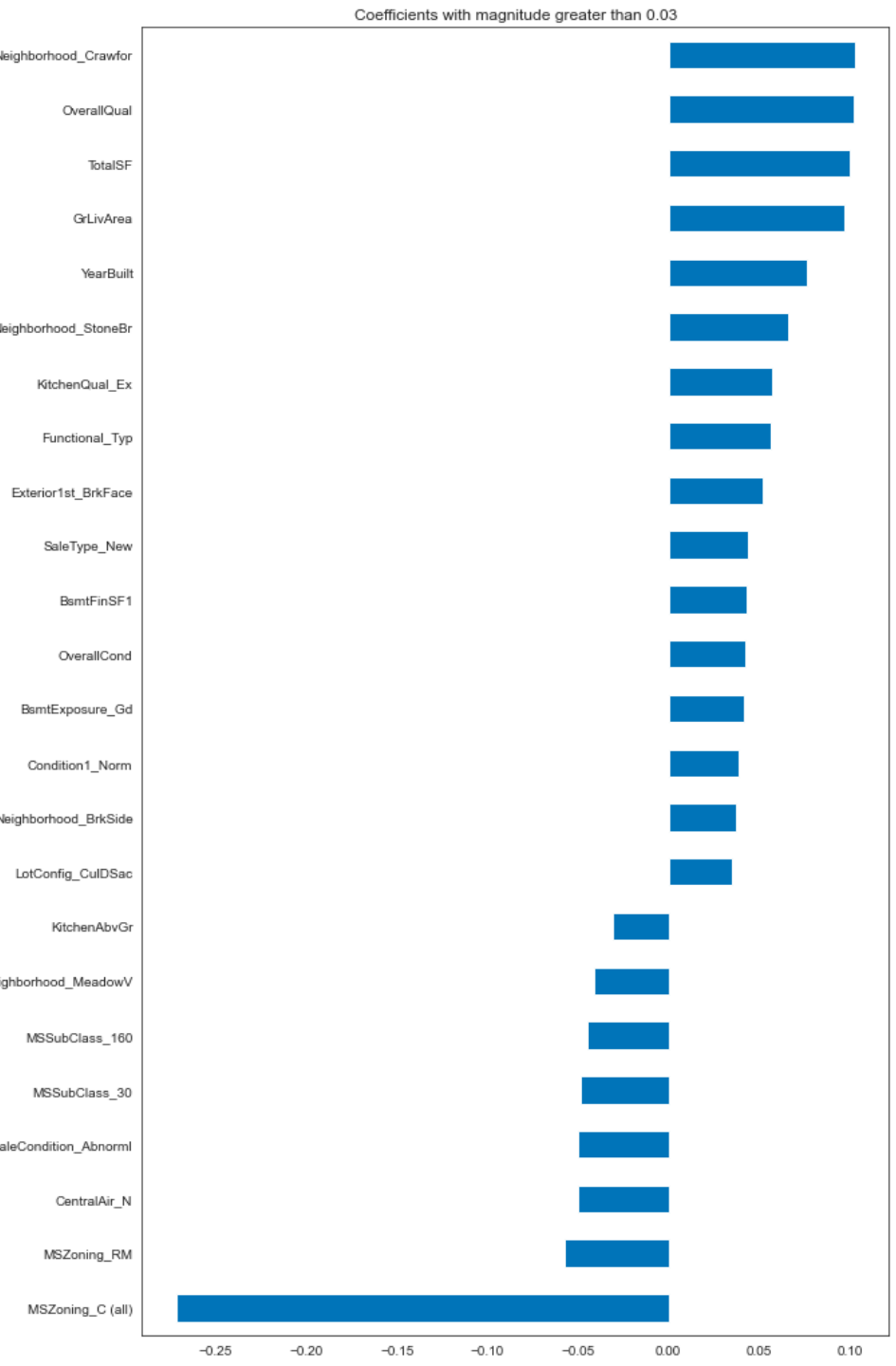
- Used 10-fold cross-validation to find the best hyperparameters
- Added RobustScaler to the regularized regression pipeline, and found this important since results can be different to some extent
- Gradient Boosting performed the best (with lowest average cross-validation RMSE)
- Among regularized regressions, elastic-net performed the best and it is more like a lasso regression



Level 2 Models

- Improved level 1 model performances
- Chose to submit my predicted result from the blended model and achieved top 12%

678	**Yutian Zhou**		0.11840

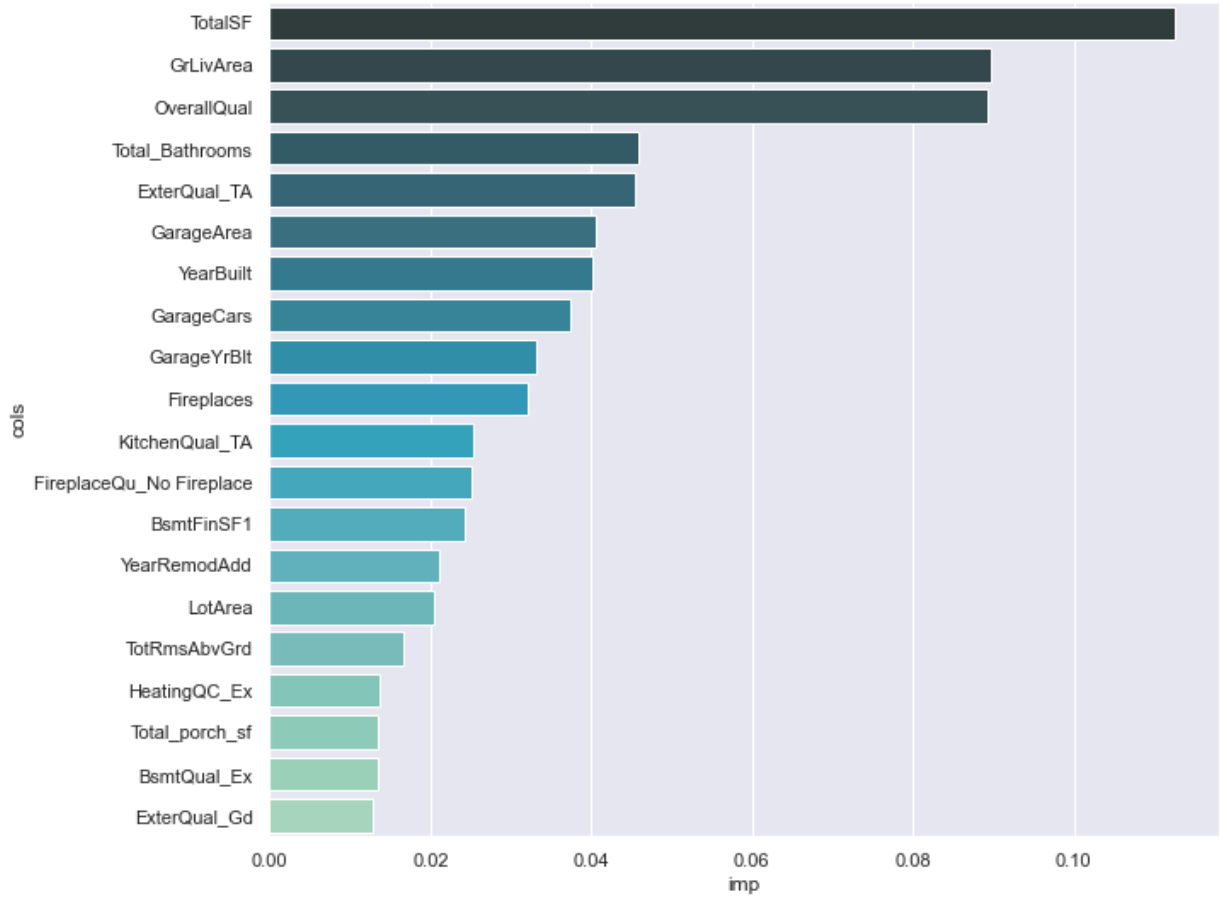


Step 4: Model Explain

- Coefficients from elastic-net:
- Living zone is important: houses in commercial area tend to obtain much lower sales prices;
- Other features such as the overall quality, total square feet, year built, ground living area are positively related to sale price, which is in accordance with the expectation that the larger the living area of the house, the newer the house, the better the quality, the higher the price;
- Neighborhood seems to be important, but need more data to explain why Crawford has higher average sale price (eg. school district, crime rate, newly developed area, etc.);

Step 4: Model Explain

- Feature Importance from gradient boosting:
- Similar to findings from regression coefficients, the overall quality, total square feet, year built, ground living area are important;
- Here, total number of bathrooms, and features of garage are also important features to sale price of houses in Ames;



Conclusion / Future Plans

1. Outliers affect the result a lot, will be more careful on the outliers if I do another round of data processing and see how this would affect model performances;
2. Try other ways of feature engineering, for example, rebuild total square feet, but not do too much to make model hard to explain;
3. Try gridsearch and focus on tuning models;
4. Do more analysis on why neighborhood_Crawford had a higher coefficient;


THANK YOU!