Rensselaer

IOWA HOUSE PRICE PREDICTION

Baiting Gai Contact: gaib@rpi.edu

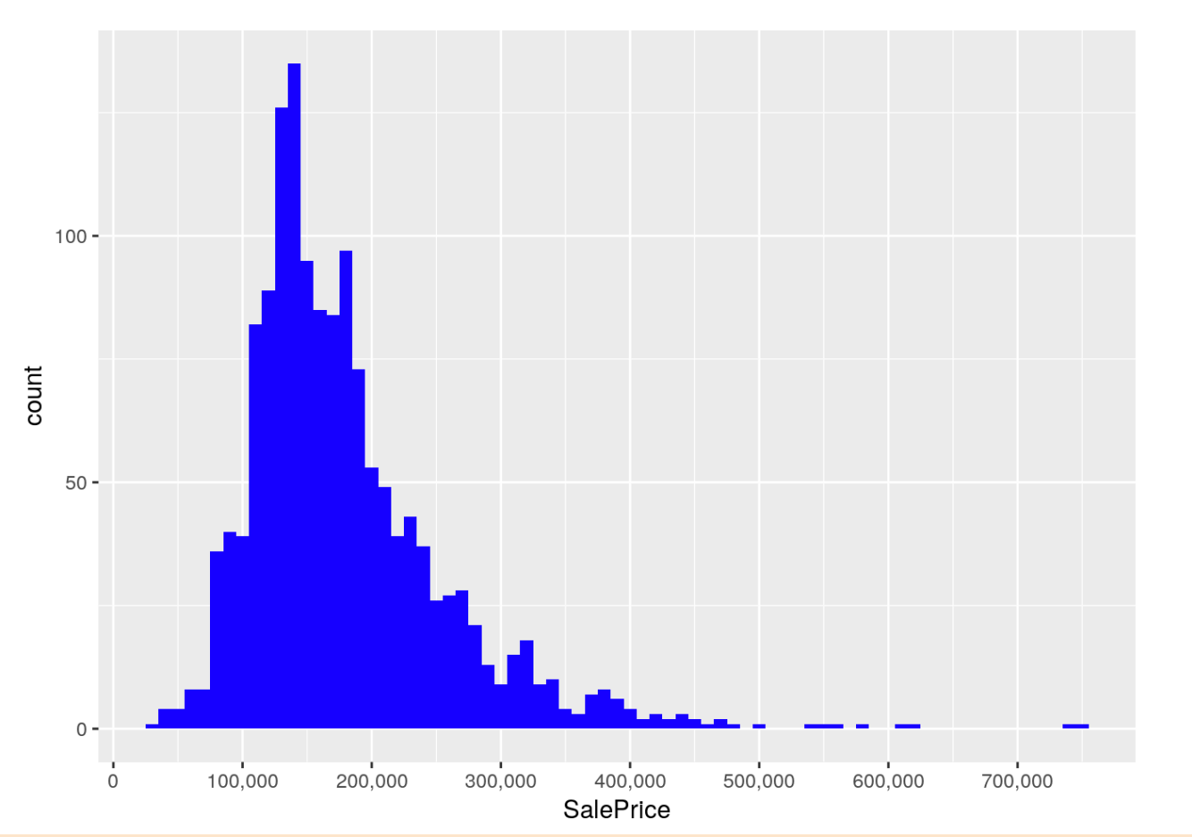
Abstract

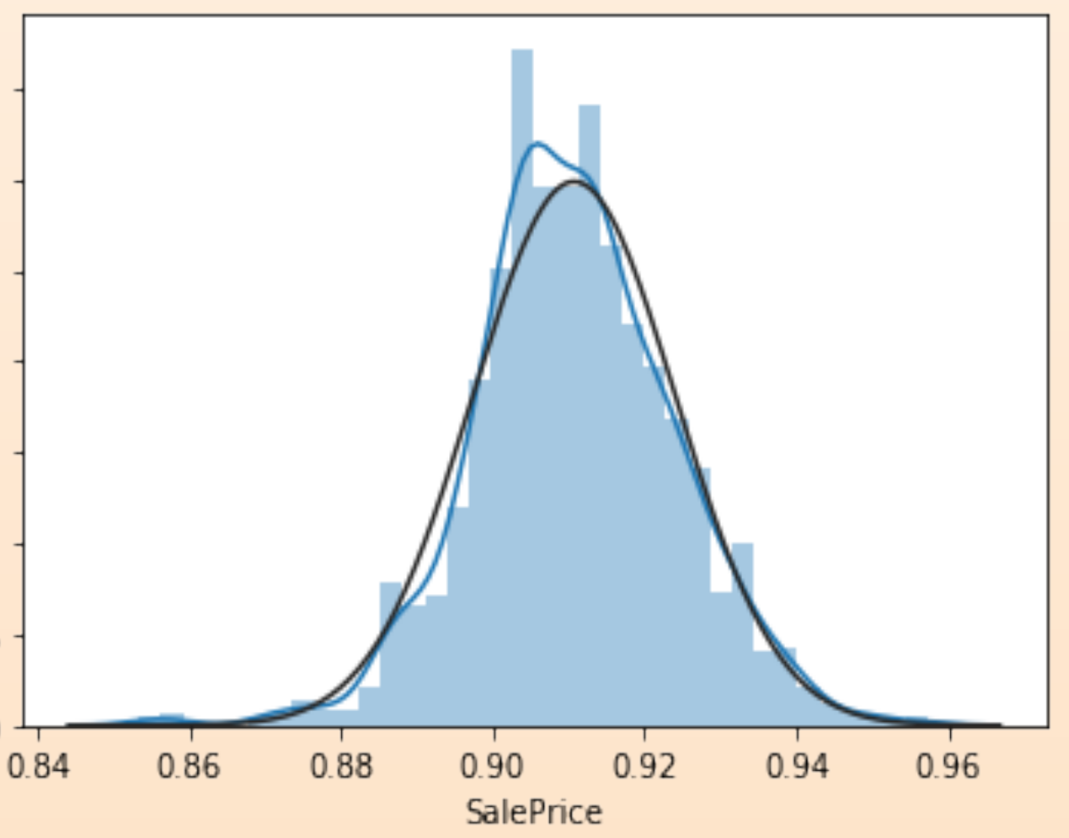
The project aims to use 79 independent variables(36 quantitative features and 43 categorical features) describing every aspect of one house to predict every house price. To improve the model accuracy, I did a lot of feature engineering work where missing values imputation and feature creation are most important tasks. Finally, using 254 features to do modeling with Lasso, Ridge, Elastic Net, Xboost and XGBoost and get 0.1108 RMSE.

Reference

- Dataset: https://www.kaggle.com/c/house-prices-advanced-regression-techniques
- Other solutions:
  - https://www.kaggle.com/serigne/stacked-regressions-top-4-on-leaderboard
  - https://www.kaggle.com/serigne/stacked-regressions-top-4-on-leaderboard
  - https://www.kaggle.com/erikbruin/house-prices-lasso-xgboost-and-a-detailed-eda
- Spearman's rank correlation coefficient: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient
- XGBoost and Boost: https://shirinsplayground.netlify.com/2018/11/ml_basics_gbm/

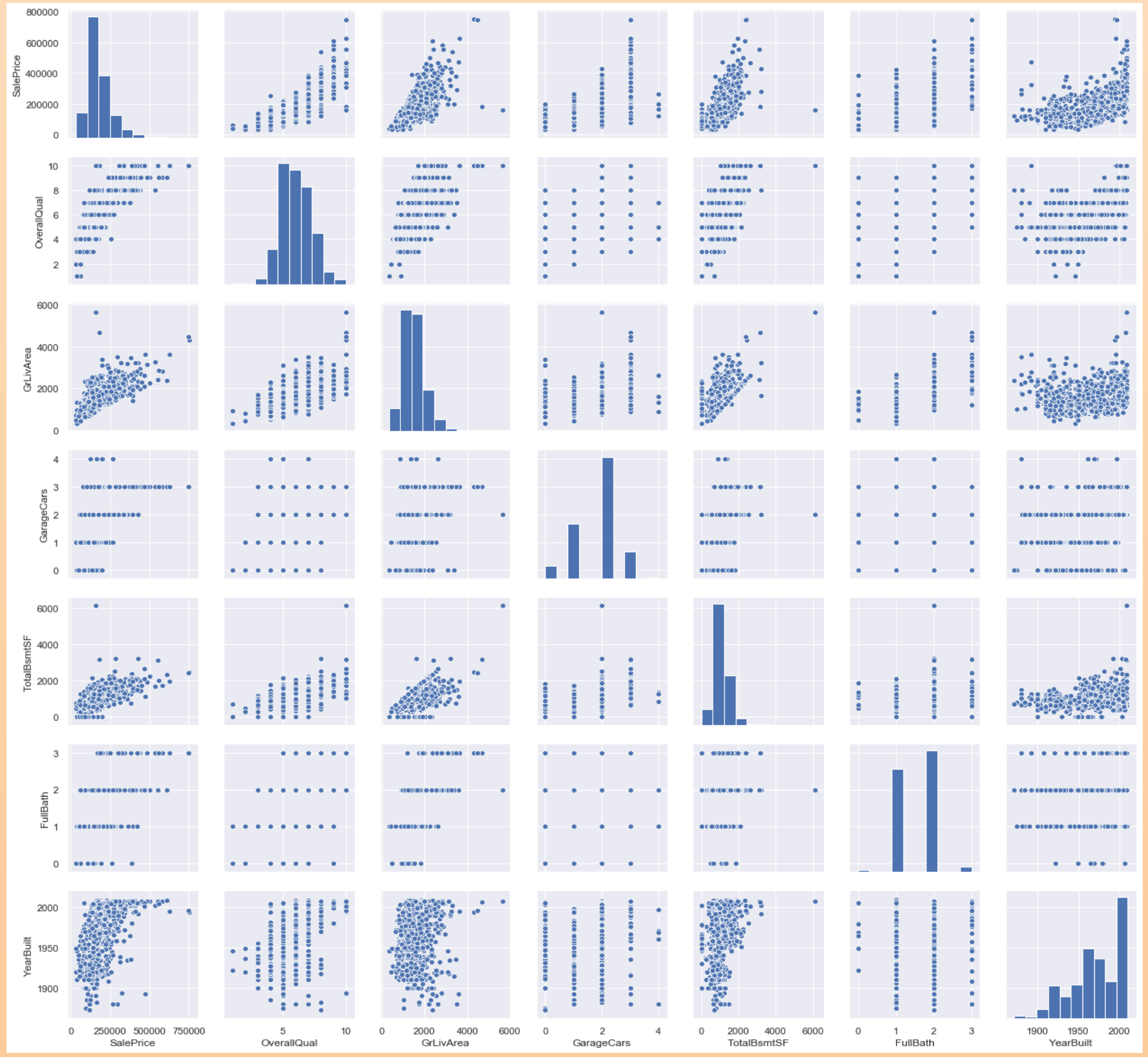
Data Visualization

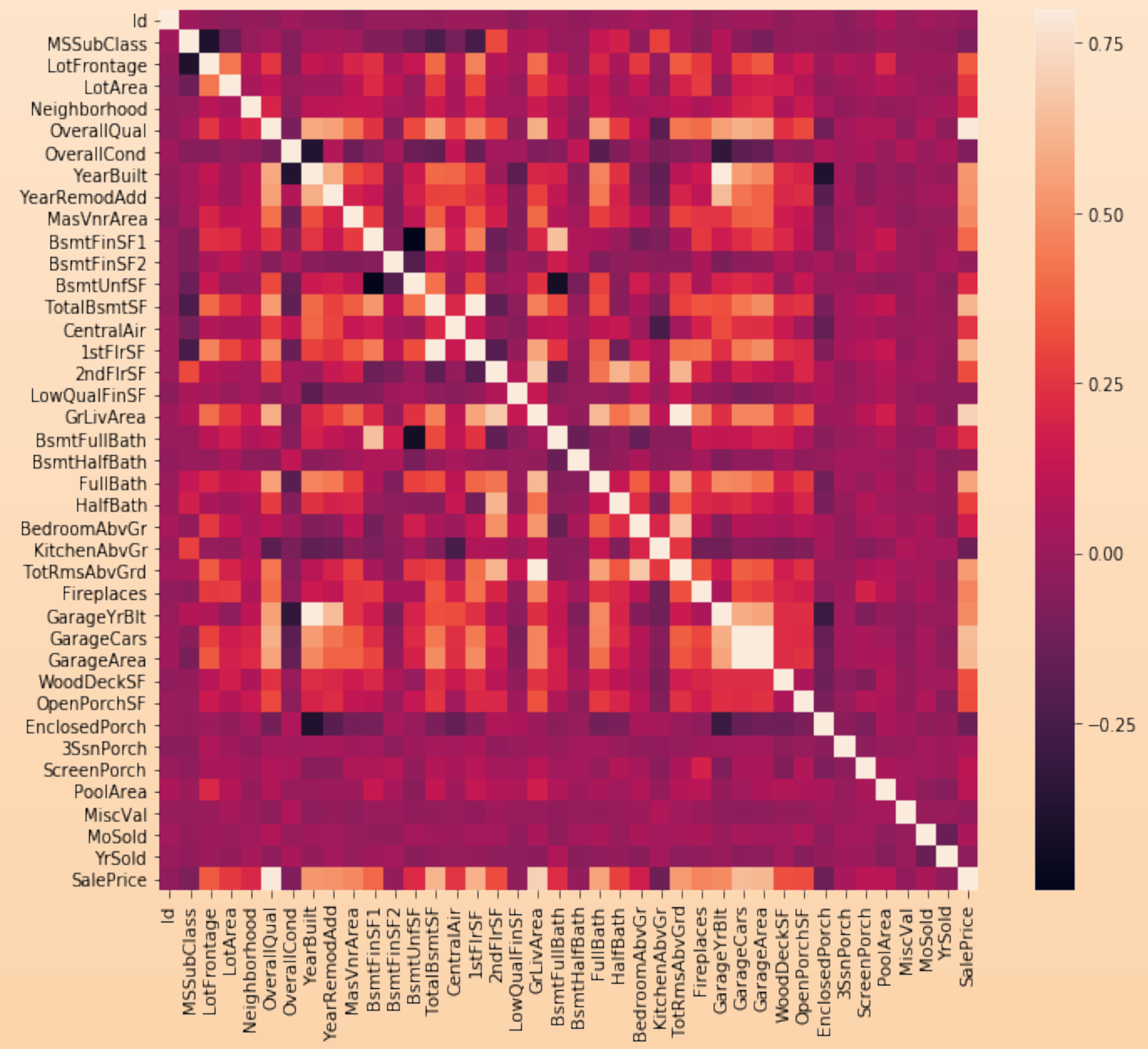




- Target variable---- Sale Price
- Not normal distribution
- Log transformation

- Explore relationship between independent variables
- Top features: Overall quality, Grliv area, Garagecars, Garage areas, Total basement area, Bathroom area, Year built





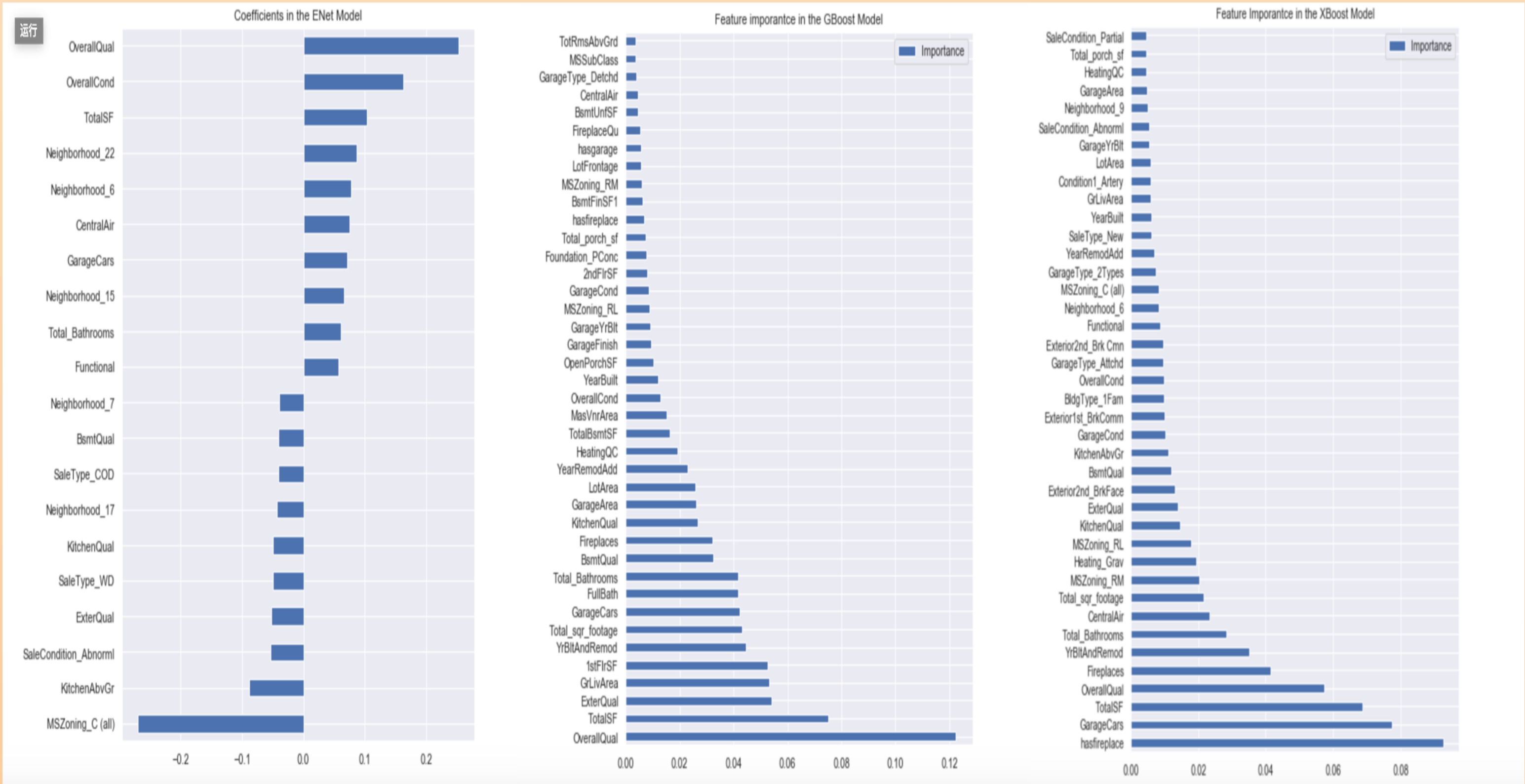
- More data visualization for relationship between independent variables
  - Linear/non-linear
  - Useful for feature creation
- Scatter plots with Sale price
  - useful for finding outliers

Models

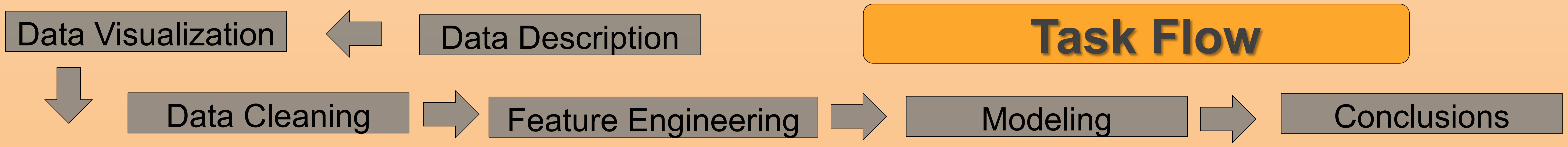
- Define loss function (RMSE)
- Define validation method (10-fold)
- Model performance

Models	RMSE mean (standard deviation)	Algorithm	Feature Performance
Lasso	0.1111 (0.0141)	L1 regularization	More Concentrated
Ridge	0.1204 (0.0166)	L2 Regularization	More dispersive
Elastic Net	0.1110 (0.0142)	Balance L1 and L2 (L1 ratio=0.9)	Good balance
GBoost	0.1120 (0.0176)	Gradient Descent	More emphasis on overall quality
XGBoost	0.1108 (0.0161)	Gradient Descent + Regularization	Distribute more even/pay attention to binary features

- Feature importance
  - Elastic Net: overall quality/commercial house?/overall condition/total square feet/in Somerset or College Creek?
  - XGBoost: has a fireplace? Quantity/central air/bathroom quality (measured in area) and condition/in medium density region?



Task Flow



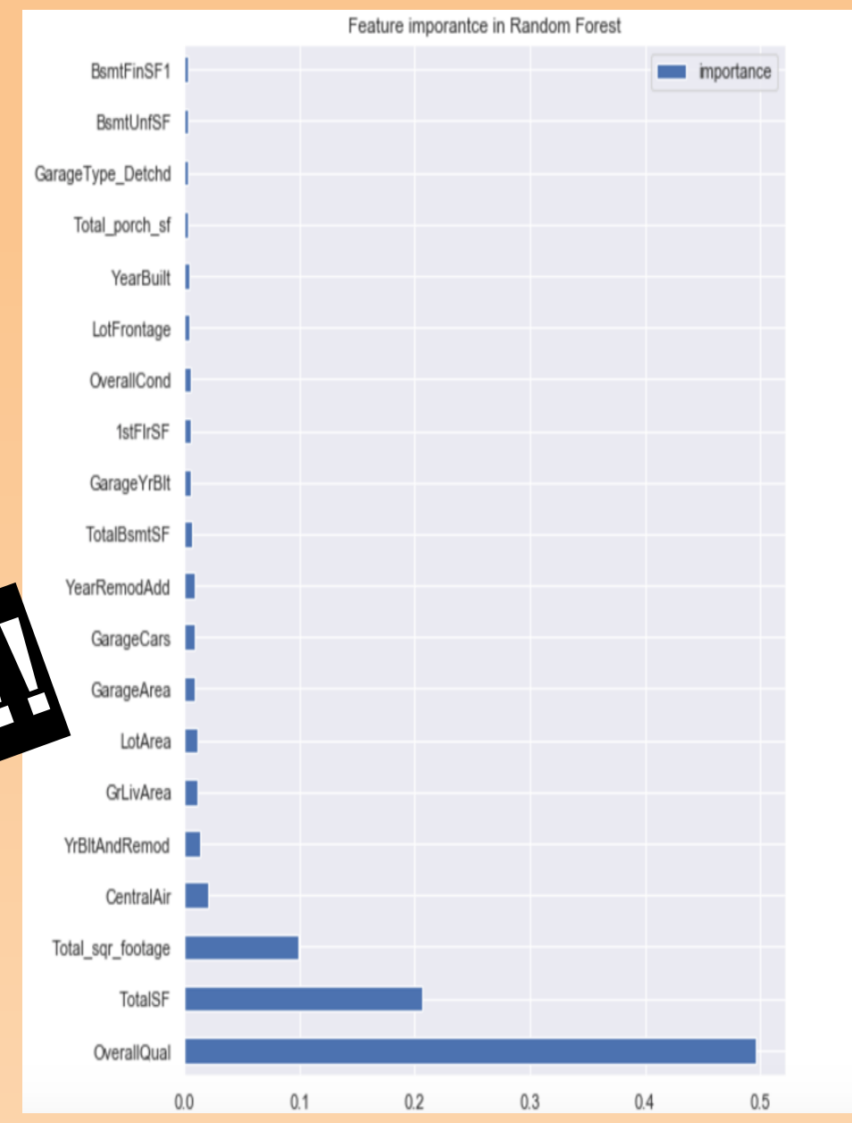
Feature Engineering

- Increase data put (combine train and test dataset)
- Advanced independent relationship exploration
  - Non-linear relationship(Spearman correlation)
  - Clustering(useful more feature generation)
- Feature generation
  - Combine related features
    - Built year/Area(total/footage)/Bathroom/Garage
  - Create binary features
    - Haspool/Has2ndfloor/Hasgarage/ Hasbsmt/Hasfireplace

- Data cleaning
  - Missing value based on different features(look for each one)
    - Missing ratio(high/low)/Interpretation/Distribution/Data type/Method/Grouping
    - Exclude outliers based on scatter plots (GrLivArea >4000 or Sale Price <300000)
    - Label Encoding (for categorical variables in numeric
    - Transforming data(skewness>0.5)
    - Dummy variables

GOT 254 FEATURES!!!

- Feature Selection



Conclusions & Future Works

- Recommendations
  - For property developers
    - improve their internal quality (bathroom, garage, fireplace) to raise prices even though the houses are not in an expensive district
  - For individual house purchasers
    - choose the ones which facilities are not perfect to decrease house purchasing expenses if you want to buy houses in commercial or medium resident density
- Future works
  - Adjust hype parameters
  - More feature engineering
  - Stacking models (adding a meta-model)