MACHINE LEARNING PROJECT

PROJECT 3

Aron Berke
Ashish Sharma
Austin Cheng
Gabriel Corbal

The project is centered on a dataset containing ~1400 house prices and associated predictors

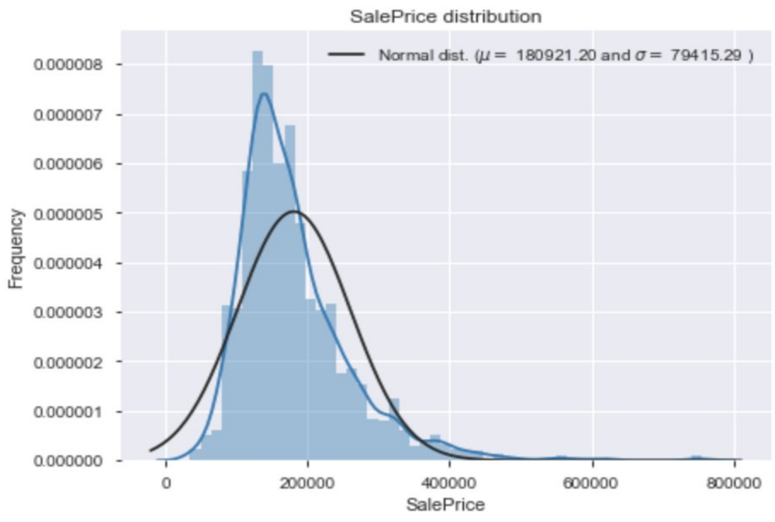
- Kaggle dataset
- House Prices: Advanced Regression Techniques
- 79 explanatory variables describing aspects of residential homes in Ames, Iowa.
- Predict the price

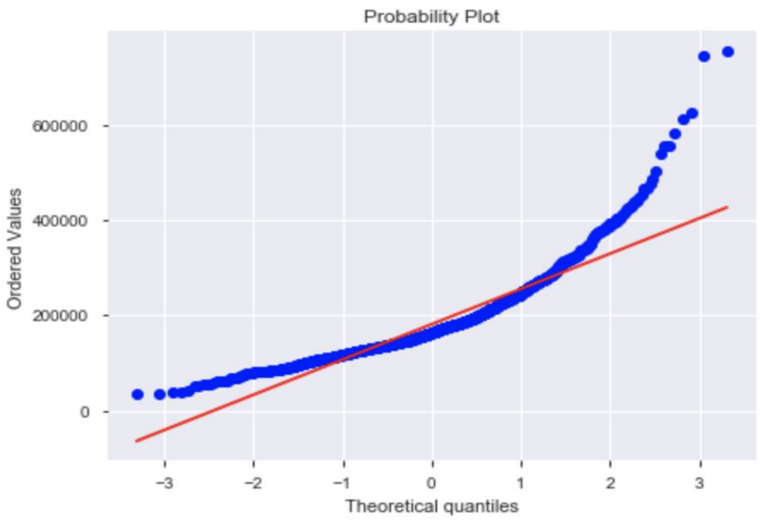
Our goal was to gain familiarity with feature engineering and regularized/tree-based models

- Improve our EDA skills
- Learn more about regressions technics
- Get more comfortable with machine learning methods.

The target variable shows right skew

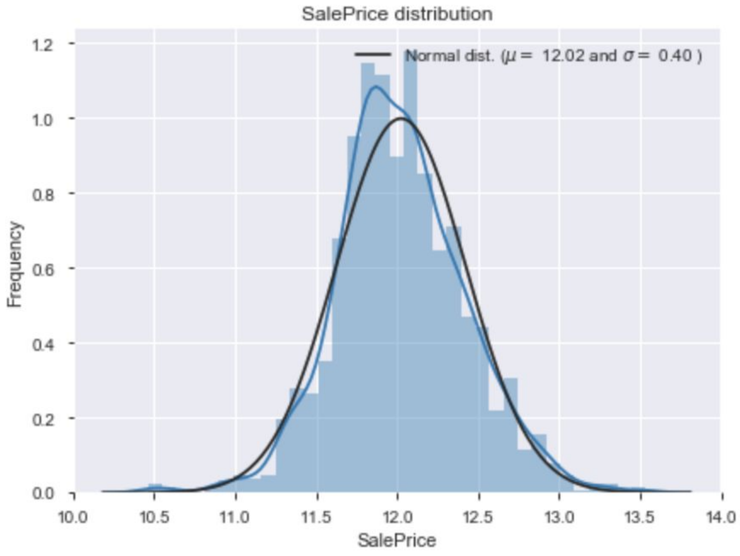
Sale Price

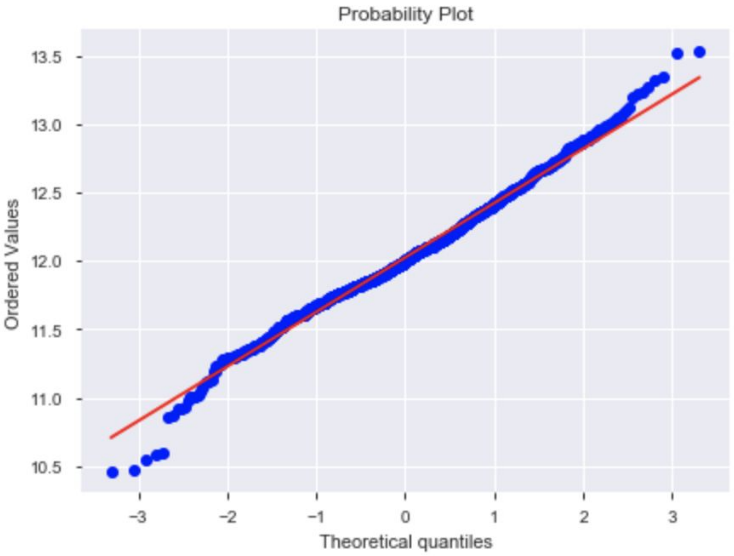




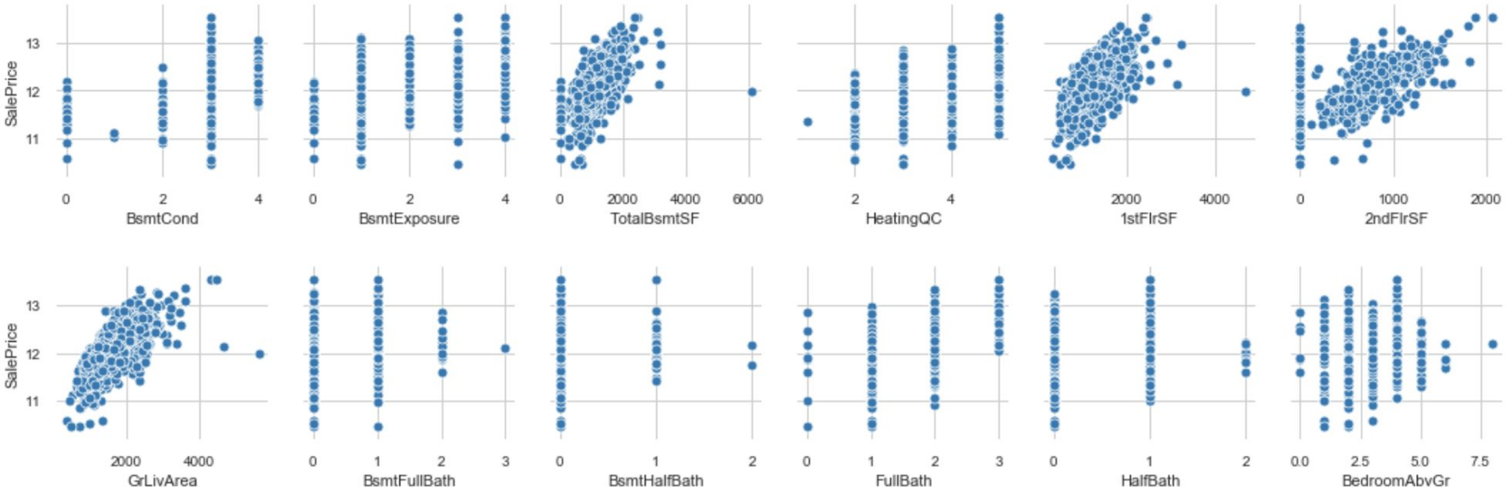
After log transformation, it is normally distributed

Sale Price - Log

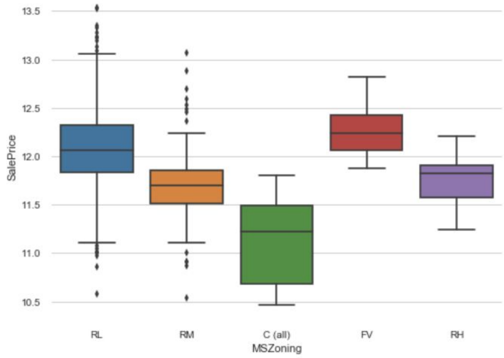


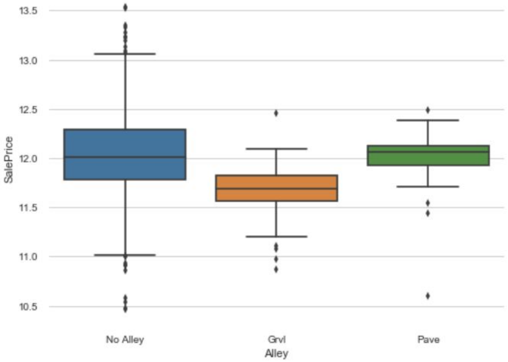


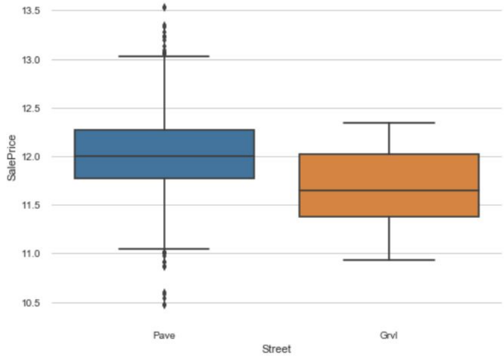
Size-related continuous and interval variables showed moderate correlation with SalePrice

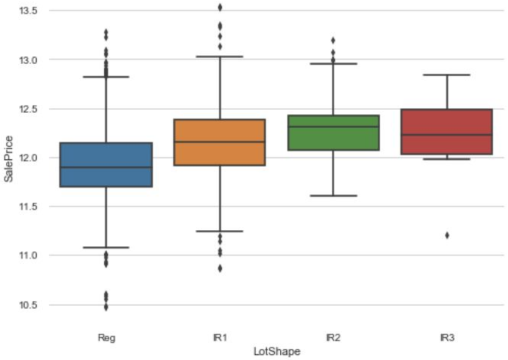


SalePrice distribution differed across several categorical variables, particularly MSZoning





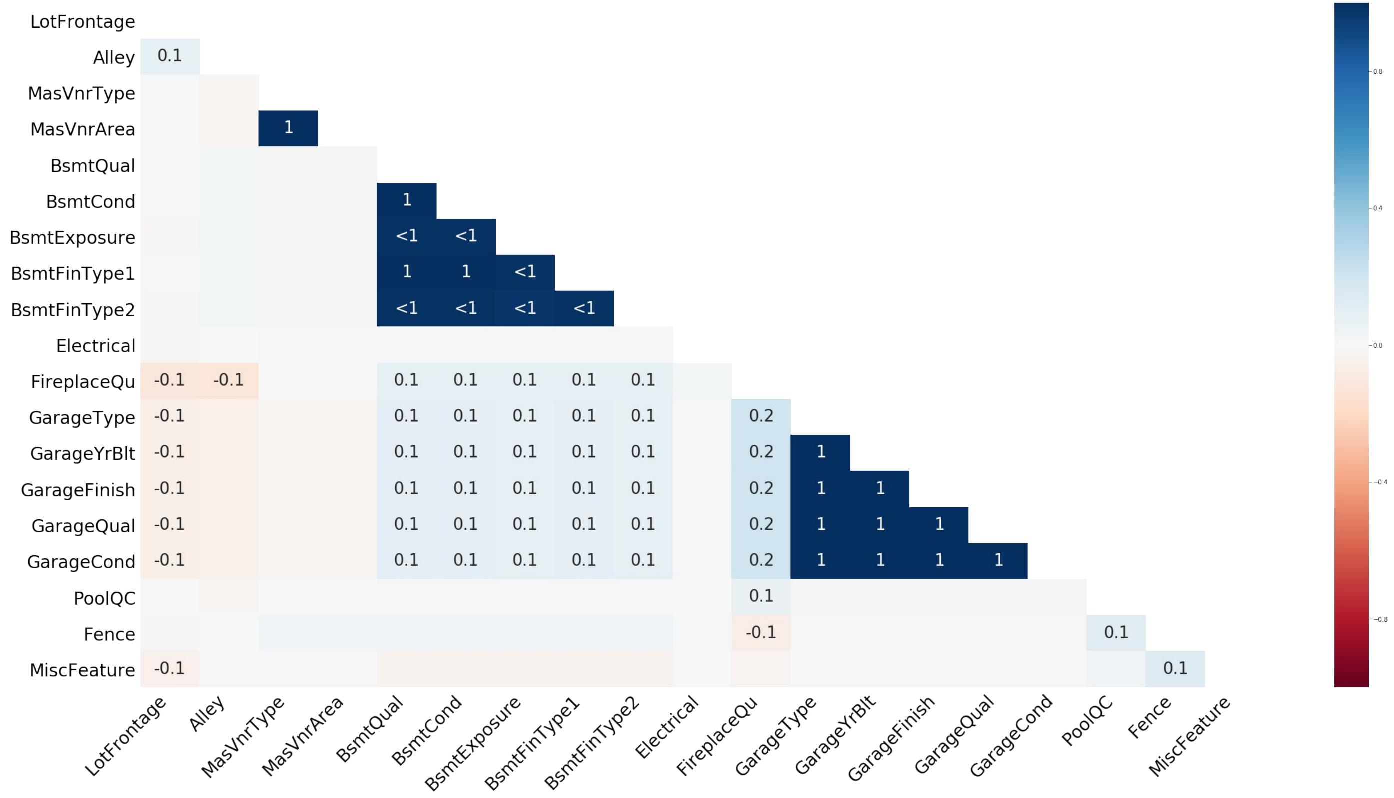




35 variables had at least one missing value

	Total	Percent
PoolQC	2909	99.66
MiscFeature	2814	96.40
Alley	2721	93.22
Fence	2348	80.44
SalePrice	1459	49.98
FireplaceQu	1420	48.65
LotFrontage	486	16.65
GarageCond	159	5.45
GarageYrBlt	159	5.45
GarageQual	159	5.45
GarageFinish	159	5.45
GarageType	157	5.38
BsmtCond	82	2.81
BsmtExposure	82	2.81
BsmtQual	81	2.77
BsmtFinType2	80	2.74
BsmtFinType1	79	2.71
MasVnrType	24	0.82
MasVnrArea	23	0.79
MSZoning	4	0.14
Utilities	2	0.07
Functional	2	0.07
BsmtFullBath	2	0.07
BsmtHalfBath	2	0.07
GarageCars	1	0.03
BsmtFinSF2	1	0.03
Exterior2nd	1	0.03
GarageArea	1	0.03
TotalBsmtSF	1	0.03
BsmtUnfSF	1	0.03
BsmtFinSF1	1	0.03
Exterior1st	1	0.03
KitchenQual	1	0.03
SaleType	1	0.03
Electrical	1	0.03

NAs were highly correlated among some variables



For 6 variables, ‘NAs’ were replaced with ‘No’ type

	Total	Percent
PoolQC	2909	99.66
MiscFeature	2814	96.40
Alley	2721	93.22
Fence	2348	80.44
SalePrice	1459	49.98
FireplaceQu	1420	48.65
LotFrontage	486	16.65
GarageCond	159	5.45
GarageYrBlt	159	5.45
GarageQual	159	5.45
GarageFinish	159	5.45
GarageType	157	5.38
BsmtCond	82	2.81
BsmtExposure	82	2.81
BsmtQual	81	2.77
BsmtFinType2	80	2.74
BsmtFinType1	79	2.71

MasVnrType	24	0.82
MasVnrArea	23	0.79
MSZoning	4	0.14
Utilities	2	0.07
Functional	2	0.07
BsmtFullBath	2	0.07
BsmtHalfBath	2	0.07
GarageCars	1	0.03
BsmtFinSF2	1	0.03
Exterior2nd	1	0.03
GarageArea	1	0.03
TotalBsmtSF	1	0.03
BsmtUnfSF	1	0.03
BsmtFinSF1	1	0.03
Exterior1st	1	0.03
KitchenQual	1	0.03
SaleType	1	0.03
Electrical	1	0.03

For 6 other variables, ‘NAs’ were replaced with 0

	Total	Percent
PoolQC	2909	99.66
MiscFeature	2814	96.40
Alley	2721	93.22
Fence	2348	80.44
SalePrice	1459	49.98
FireplaceQu	1420	48.65
LotFrontage	486	16.65
GarageCond	159	5.45
GarageYrBlt	159	5.45
GarageQual	159	5.45
GarageFinish	159	5.45
GarageType	157	5.38
BsmtCond	82	2.81
BsmtExposure	82	2.81
BsmtQual	81	2.77
BsmtFinType2	80	2.74
BsmtFinType1	79	2.71

MasVnrType	24	0.82
MasVnrArea	23	0.79
MSZoning	4	0.14
Utilities	2	0.07
Functional	2	0.07
BsmtFullBath	2	0.07
BsmtHalfBath	2	0.07
GarageCars	1	0.03
BsmtFinSF2	1	0.03
Exterior2nd	1	0.03
GarageArea	1	0.03
TotalBsmtSF	1	0.03
BsmtUnfSF	1	0.03
BsmtFinSF1	1	0.03
Exterior1st	1	0.03
KitchenQual	1	0.03
SaleType	1	0.03
Electrical	1	0.03

For variables that had a few observations MCAR, missing values were replaced with the mode

	Total	Percent
PoolQC	2909	99.66
MiscFeature	2814	96.40
Alley	2721	93.22
Fence	2348	80.44
SalePrice	1459	49.98
FireplaceQu	1420	48.65
LotFrontage	486	16.65
GarageCond	159	5.45
GarageYrBlt	159	5.45
GarageQual	159	5.45
GarageFinish	159	5.45
GarageType	157	5.38
BsmtCond	82	2.81
BsmtExposure	82	2.81
BsmtQual	81	2.77
BsmtFinType2	80	2.74
BsmtFinType1	79	2.71
MasVnrType	24	0.82
MasVnrArea	23	0.79
MSZoning	4	0.14
Utilities	2	0.07
Functional	2	0.07
BsmtFullBath	2	0.07
BsmtHalfBath	2	0.07
GarageCars	1	0.03
BsmtFinSF2	1	0.03
Exterior2nd	1	0.03
GarageArea	1	0.03
TotalBsmtSF	1	0.03
BsmtUnfSF	1	0.03
BsmtFinSF1	1	0.03
Exterior1st	1	0.03
KitchenQual	1	0.03
SaleType	1	0.03
Electrical	1	0.03

Other 'NA's were replaced with the mean to avoiding impacting the slope relationship with SalePrice

	Total	Percent
PoolQC	2909	99.66
MiscFeature	2814	96.40
Alley	2721	93.22
Fence	2348	80.44
SalePrice	1459	49.98
FireplaceQu	1420	48.65
LotFrontage	486	16.65
GarageCond	159	5.45
GarageYrBlt	159	5.45
GarageQual	159	5.45
GarageFinish	159	5.45
GarageType	157	5.38
BsmtCond	82	2.81
BsmtExposure	82	2.81
BsmtQual	81	2.77
BsmtFinType2	80	2.74
BsmtFinType1	79	2.71
MasVnrType	24	0.82
MasVnrArea	23	0.79
MSZoning	4	0.14
Utilities	2	0.07
Functional	2	0.07
BsmtFullBath	2	0.07
BsmtHalfBath	2	0.07
GarageCars	1	0.03
BsmtFinSF2	1	0.03
Exterior2nd	1	0.03
GarageArea	1	0.03
TotalBsmtSF	1	0.03
BsmtUnfSF	1	0.03
BsmtFinSF1	1	0.03
Exterior1st	1	0.03
KitchenQual	1	0.03
SaleType	1	0.03
Electrical	1	0.03

Lot Frontage 'NAs' were replaced with neighborhood median

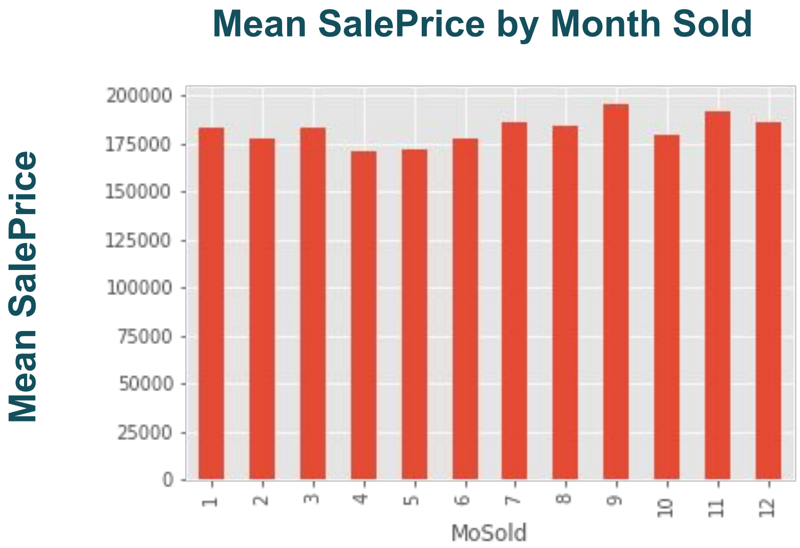
	Total	Percent
PoolQC	2909	99.66
MiscFeature	2814	96.40
Alley	2721	93.22
Fence	2348	80.44
SalePrice	1459	49.98
FireplaceQu	1420	48.65
LotFrontage	486	16.65
GarageCond	159	5.45
GarageYrBlt	159	5.45
GarageQual	159	5.45
GarageFinish	159	5.45
GarageType	157	5.38
BsmtCond	82	2.81
BsmtExposure	82	2.81
BsmtQual	81	2.77
BsmtFinType2	80	2.74
BsmtFinType1	79	2.71

MasVnrType	24	0.82
MasVnrArea	23	0.79
MSZoning	4	0.14
Utilities	2	0.07
Functional	2	0.07
BsmtFullBath	2	0.07
BsmtHalfBath	2	0.07
GarageCars	1	0.03
BsmtFinSF2	1	0.03
Exterior2nd	1	0.03
GarageArea	1	0.03
TotalBsmtSF	1	0.03
BsmtUnfSF	1	0.03
BsmtFinSF1	1	0.03
Exterior1st	1	0.03
KitchenQual	1	0.03
SaleType	1	0.03
Electrical	1	0.03

Functional 'NAs' were replaced with 'typical'

	Total	Percent
PoolQC	2909	99.66
MiscFeature	2814	96.40
Alley	2721	93.22
Fence	2348	80.44
SalePrice	1459	49.98
FireplaceQu	1420	48.65
LotFrontage	486	16.65
GarageCond	159	5.45
GarageYrBlt	159	5.45
GarageQual	159	5.45
GarageFinish	159	5.45
GarageType	157	5.38
BsmtCond	82	2.81
BsmtExposure	82	2.81
BsmtQual	81	2.77
BsmtFinType2	80	2.74
BsmtFinType1	79	2.71
MasVnrType	24	0.82
MasVnrArea	23	0.79
MSZoning	4	0.14
Utilities	2	0.07
Functional	2	0.07
BsmtFullBath	2	0.07
BsmtHalfBath	2	0.07
GarageCars	1	0.03
BsmtFinSF2	1	0.03
Exterior2nd	1	0.03
GarageArea	1	0.03
TotalBsmtSF	1	0.03
BsmtUnfSF	1	0.03
BsmtFinSF1	1	0.03
Exterior1st	1	0.03
KitchenQual	1	0.03
SaleType	1	0.03
Electrical	1	0.03

Quality scale variables were ordinalized, while other numerical data was converted to categorical

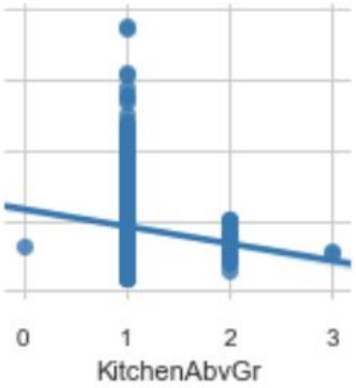


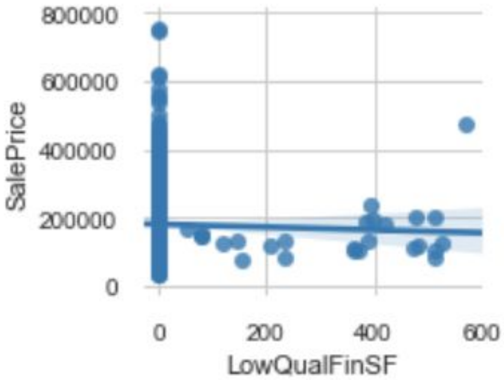
While 'month' has a natural (circular) order, it was categorized due to lack of any obvious linear relationship with SalePrice

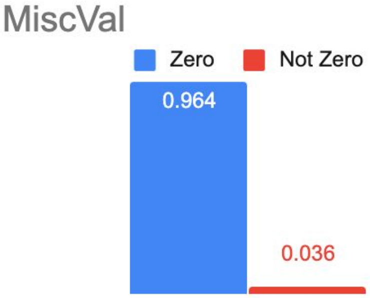
- MiscFeature -> Shed/ No Shed
- FireplaceQu, HeatingQC, ExterCond, ExterQual -> Ordinal
- MSSubClass, MoSold -> Converted to category

A few variables were removed due to major class imbalance or no relationship with SalePrice

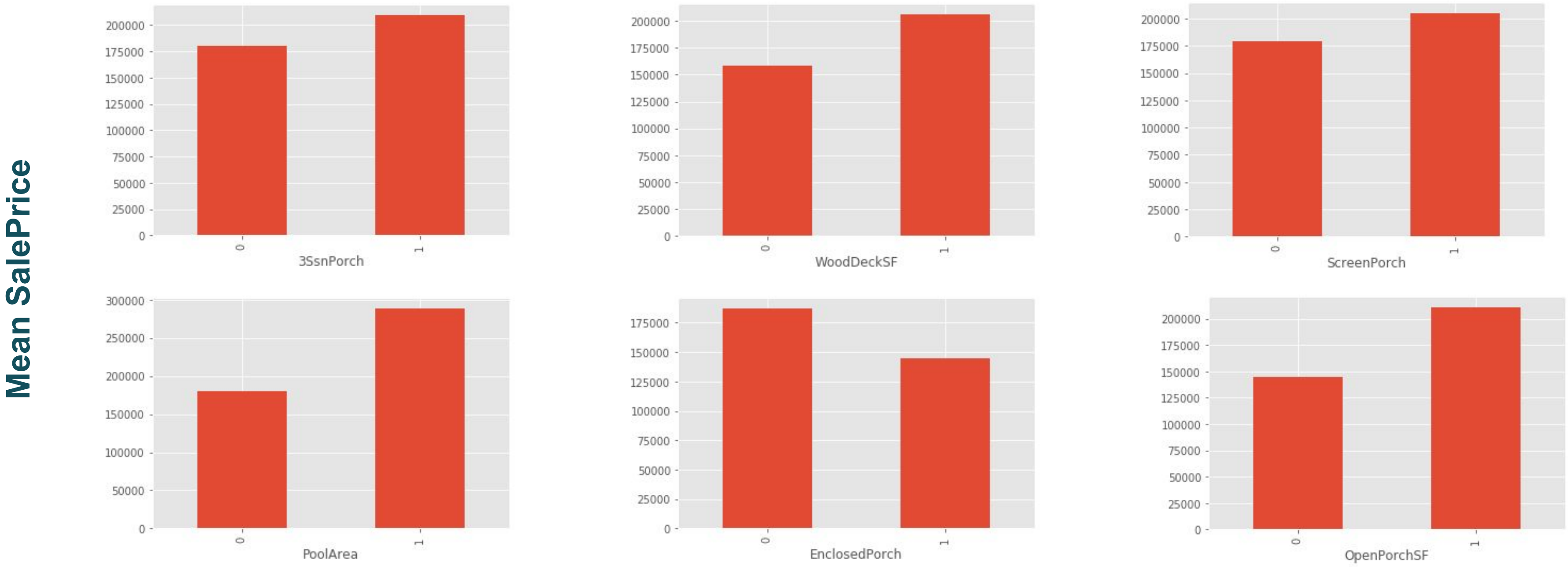
- Utilities, PoolQC,BsmtFinType2, KitchenAbvGr -> Dropped for variance
- BsmtFinSF1, BsmtFinSF2, BsmtUnfSF -> Part of TotalBsmSF
- LowQuaFinSFl -> No relationship with SalesPrice
- MiscVal -> Most values are 0







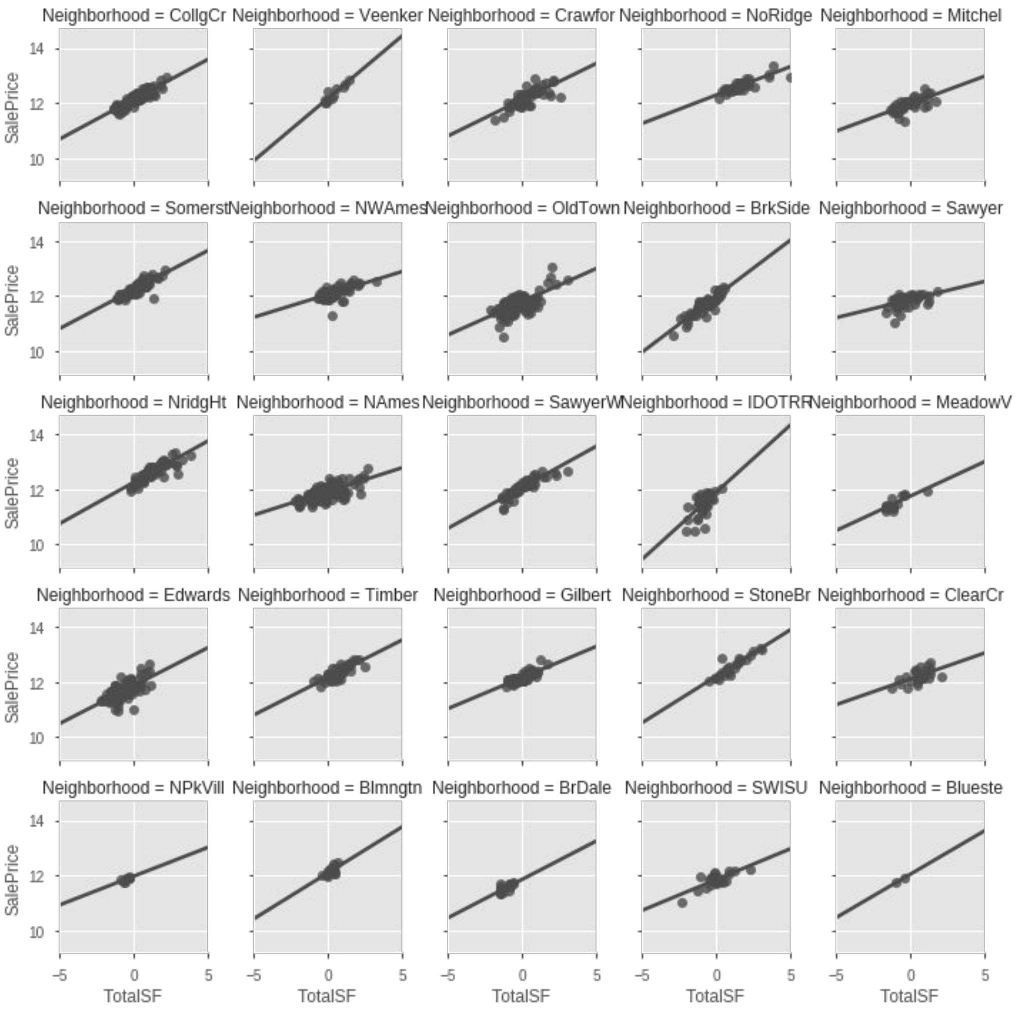
Continuous variables with a large proportion of '0s' were dummified to aid in interpretation



- These variables referenced porch, deck, and pool size - a '1' indicates presence of the feature
- T-test of means showed significant difference for all except three season porches

All continuous and ordinal variables were standardized

- Only numericals
- Applied to helps Lasso and Ridge treats variables more fairly on equal scale
- Standardization and Normalization gave us the same result



Relationship with SalePrice differed by neighborhood, justifying use of interaction terms

Tackling Multicollinearity

1. Intuition of related predictors
2. Pairwise correlation of all predictors
3. VIF of predictors
4. Remedy via combination

Intuition

1 Intuition of related predictors → 2 Pairwise correlation of all predictors → 3 VIF of predictors → 4 Remedy via combination

1stFlrSF

2ndFlrSF

TotalBsmtSF

GarageArea

GarageCars

HalfBath

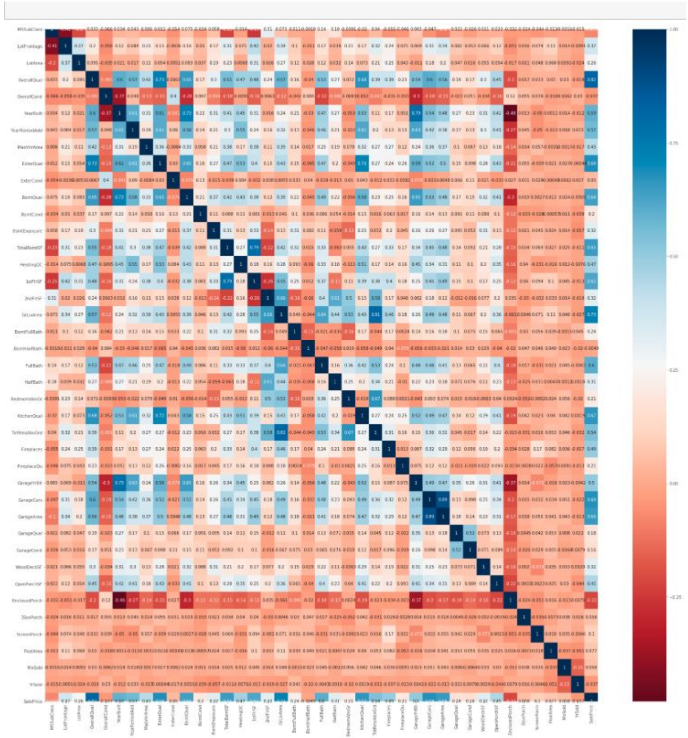
FullBath

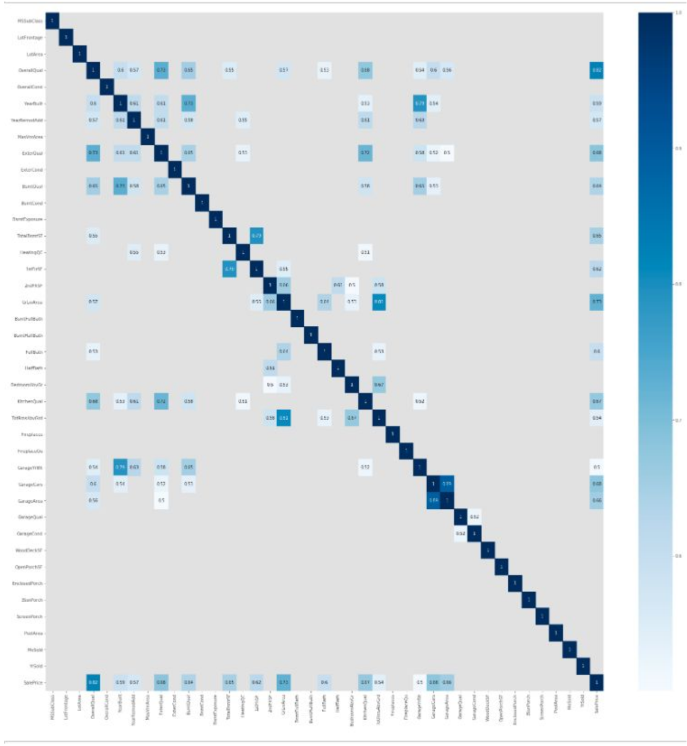
BsmtHalfBath

BsmtFullBath

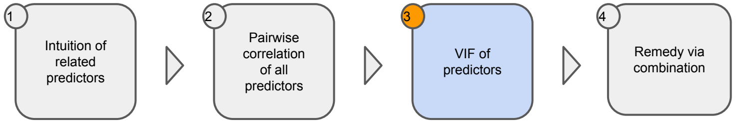
Pairwise correlation

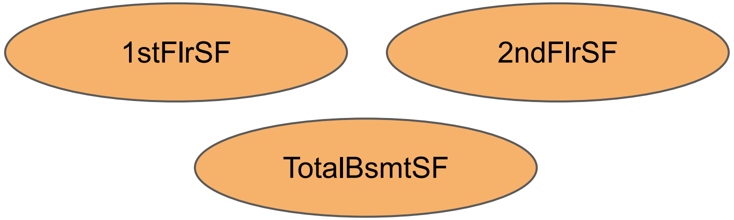
1 Intuition of related predictors → 2 Pairwise correlation of all predictors → 3 VIF of predictors → 4 Remedy via combination



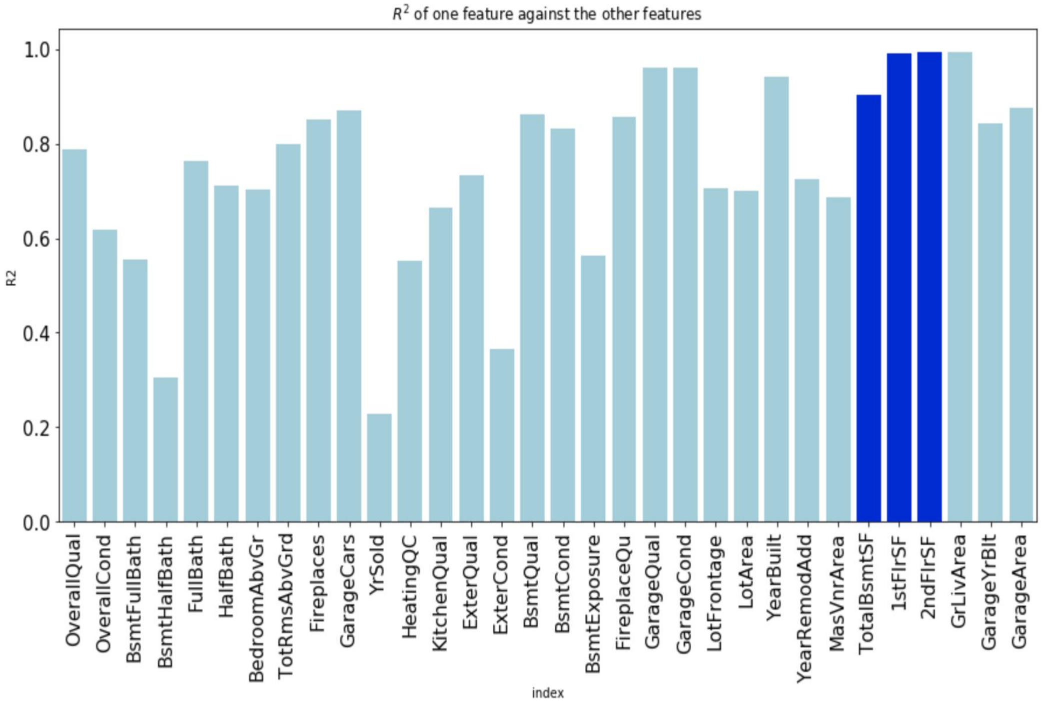


VIF among predictors

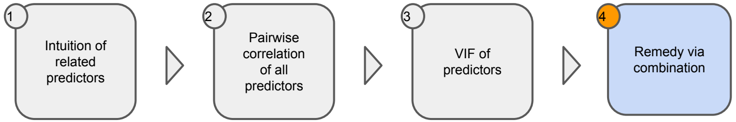


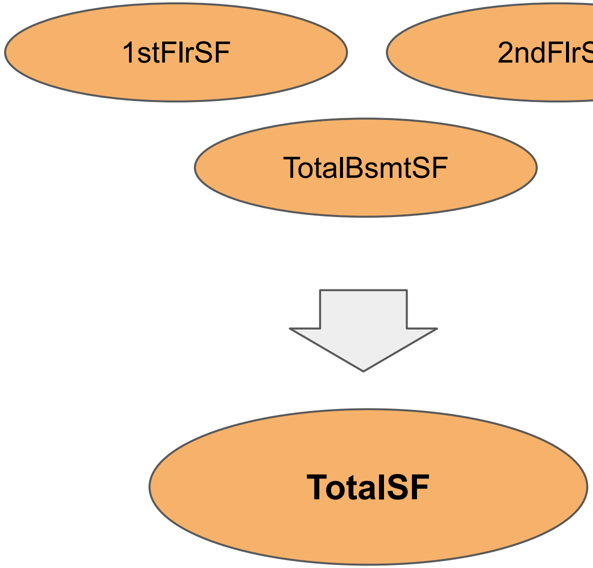


1stFlrSF	0.9905195556735192
2ndFlrSF	0.993118146028361
TotalBsmtSF	0.9022011227161376

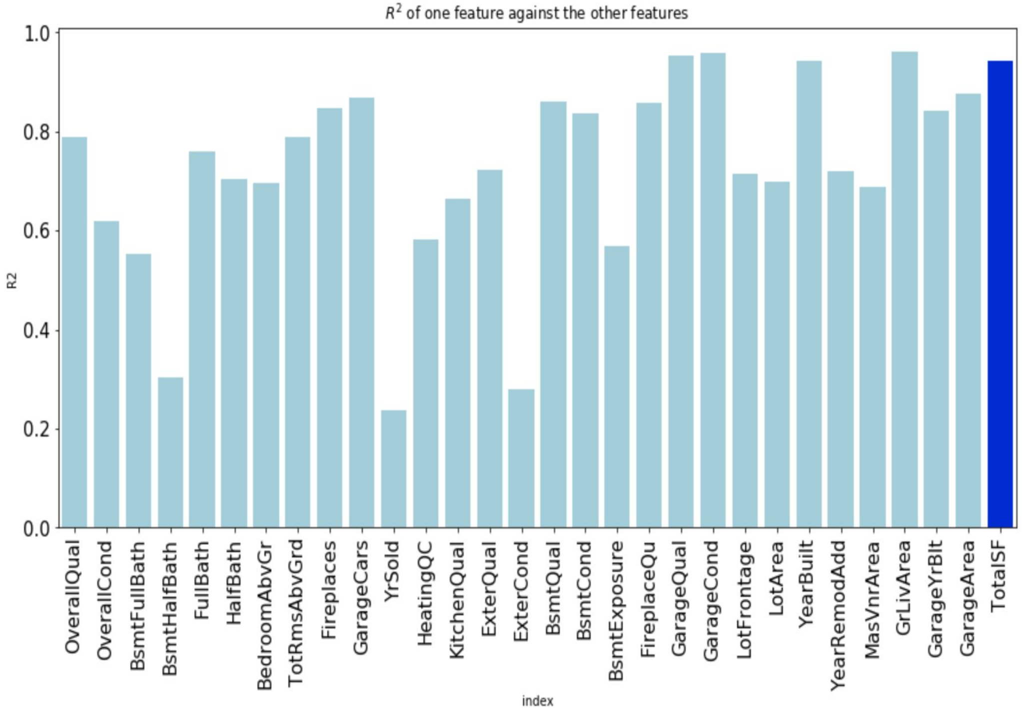


Remedy via combination

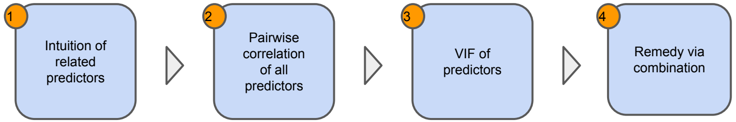




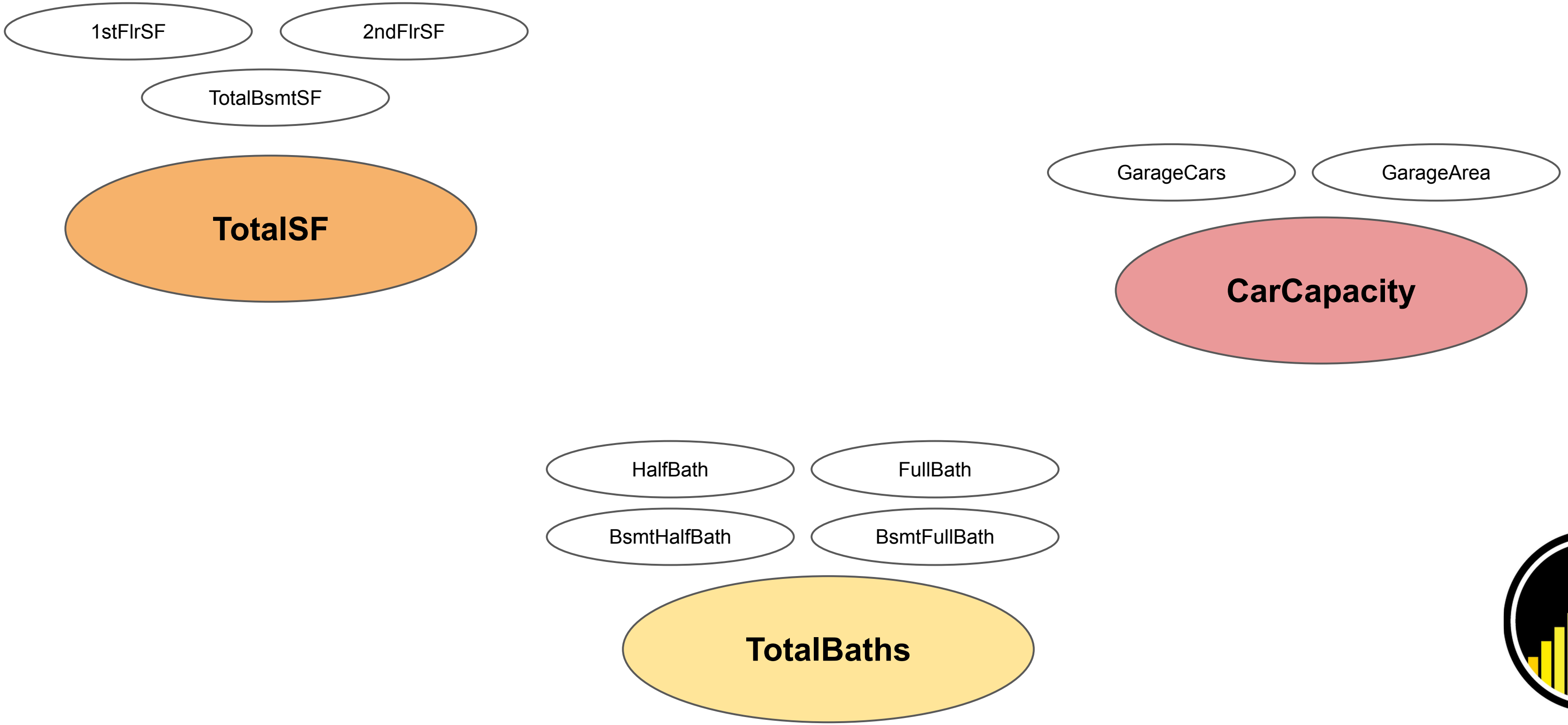
TotalSF	0.9437251375770624



Results



1
Intuition of related predictors
2
Pairwise correlation of all predictors
3
VIF of predictors
4
Remedy via combination



1stFlrSF
2ndFlrSF
TotalBsmtSF
TotalSF
GarageCars
GarageArea
CarCapacity
HalfBath
FullBath
BsmtHalfBath
BsmtFullBath
TotalBaths

The EDA previously described led us to generate multiple unique datasets to test in our ML models

Dataset	Description of Engineered Features
A	Basic imputation of NAs, transformation (ordinalization, dummification, standardization, etc.), and cleaning (dropping irrelevant columns)
B	“TotalSF“, “TotalBaths” to consolidate living spaces, improving correlation with target and solving for multicollinearity
C	“Car Capacity” to consolidate garage and car data, further reducing multicollinearity
C2	Neighborhood interaction with Total SF
D, E, F …..	numerous types of “Quality Factors” and “Time” related features

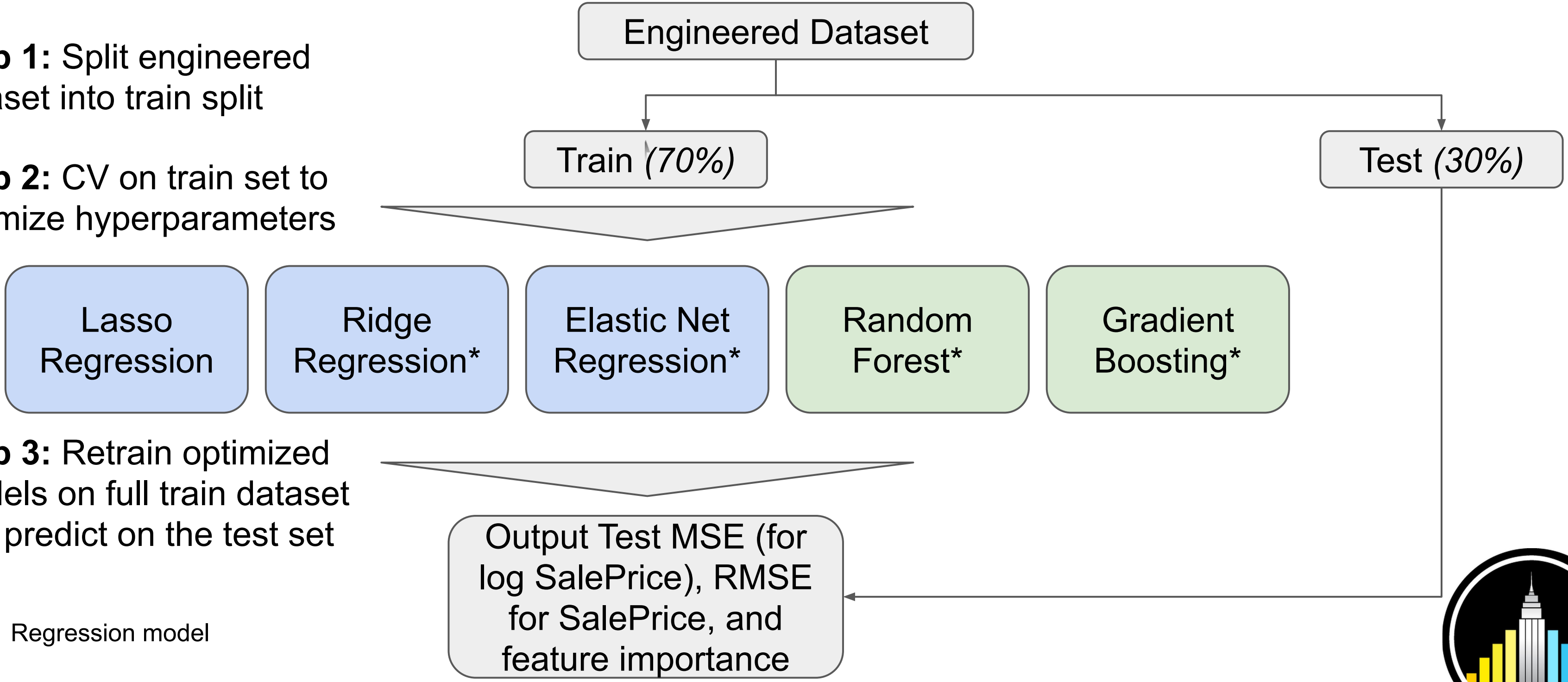
Model Complexity

No improvement in test MSE beyond dataset ‘C’

We created a pipeline to automate testing of engineered datasets across multiple models

Step 1: Split engineered dataset into train split

Step 2: CV on train set to optimize hyperparameters



Step 3: Retrain optimized models on full train dataset and predict on the test set

Regression model

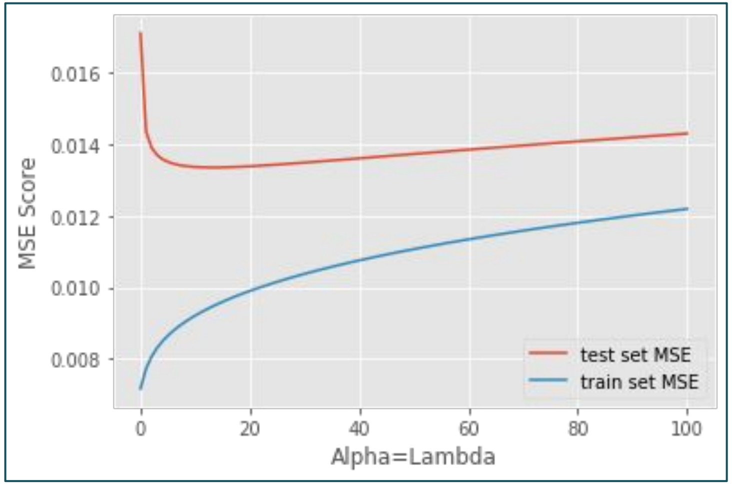
Tree-based model

---

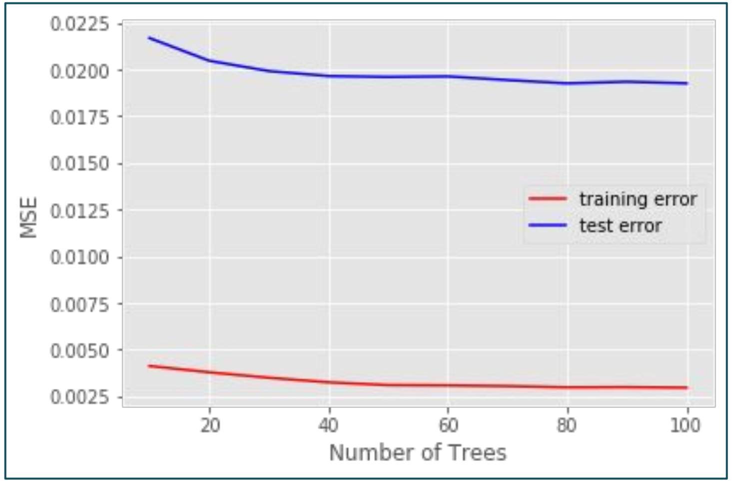
*We also ensembled models and used models for recursive feature selection, which will be covered later on

Both our regression and tree models tended to overfit to the training data

Ridge Cross-Validation Results



Random Forest Cross-Validation Results

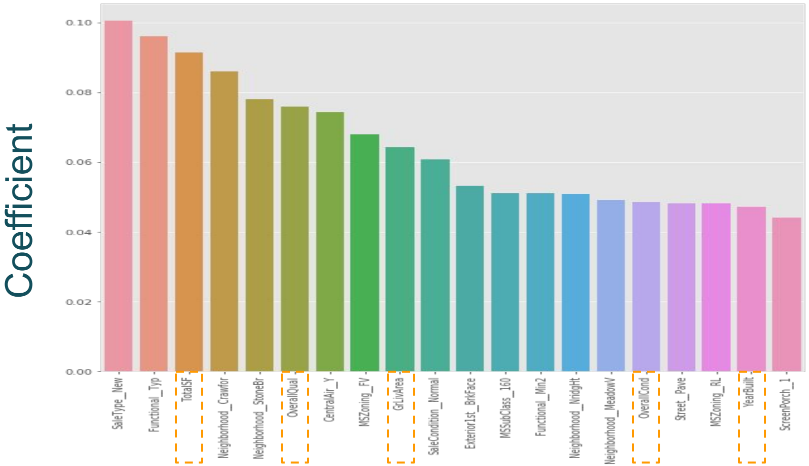


- In theory, for the ridge model, MSE on the test set should fall below train MSE as alpha increases - the fact that it doesn't suggests our regression models overfit the data at all tested alpha values

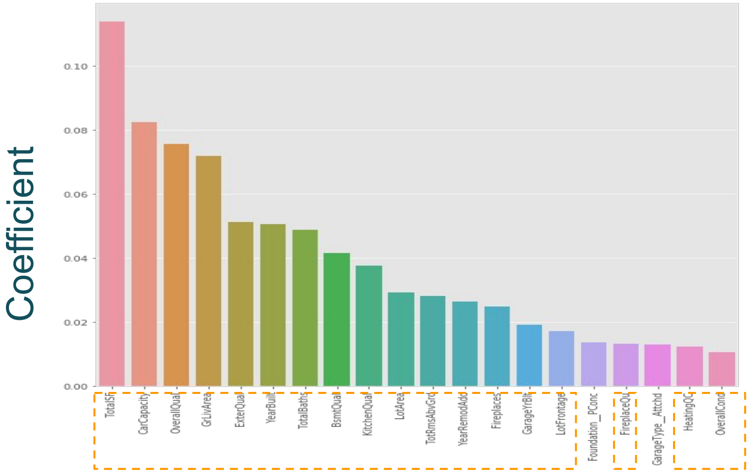
- As expected, the random forest model always overfits

Regression models emphasized categorical variables vs. continuous features for tree-based models

Lasso Model Important Features



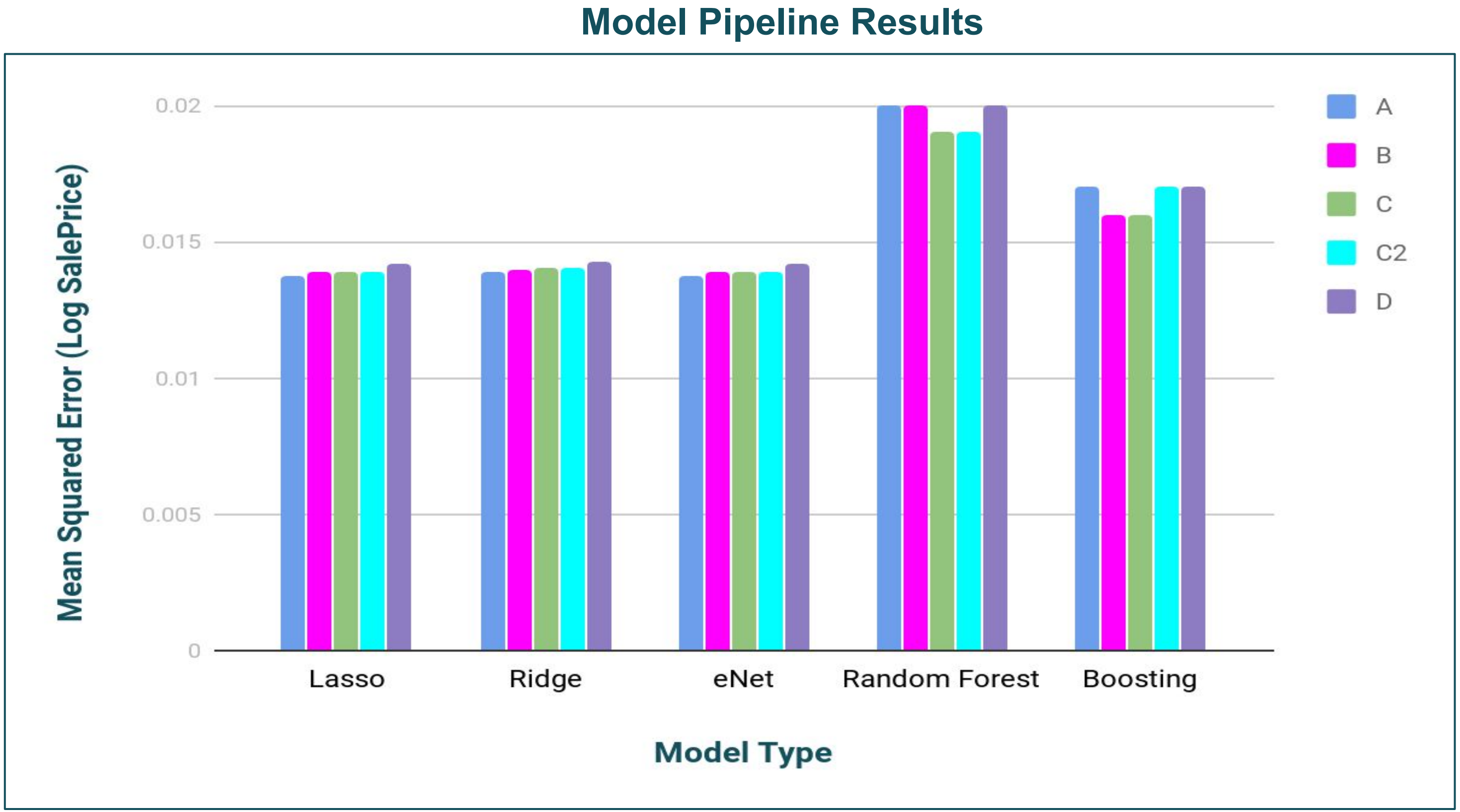
Random Forest Model Important Features



- Regularized models highlighted categorical features as most important, whereas continuous/ordinal variables are emphasized in our tree-based models
- This fits with expectation, as tree-based models tend to disadvantage dummified categorical variables (label encoding offered minimal improvement)
- Overall, location and aesthetic/quality variables were highly important to the model

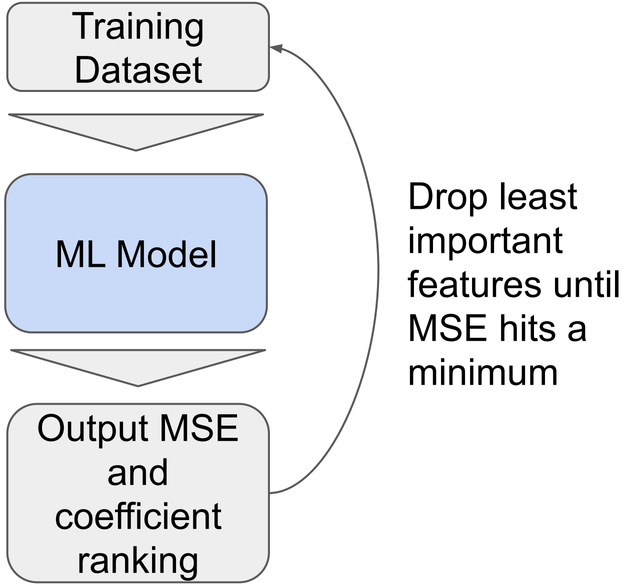
Continuous/ordinal variables

Dataset 'A' shows the best overall results in the linear models, but 'C2' performed best on Kaggle

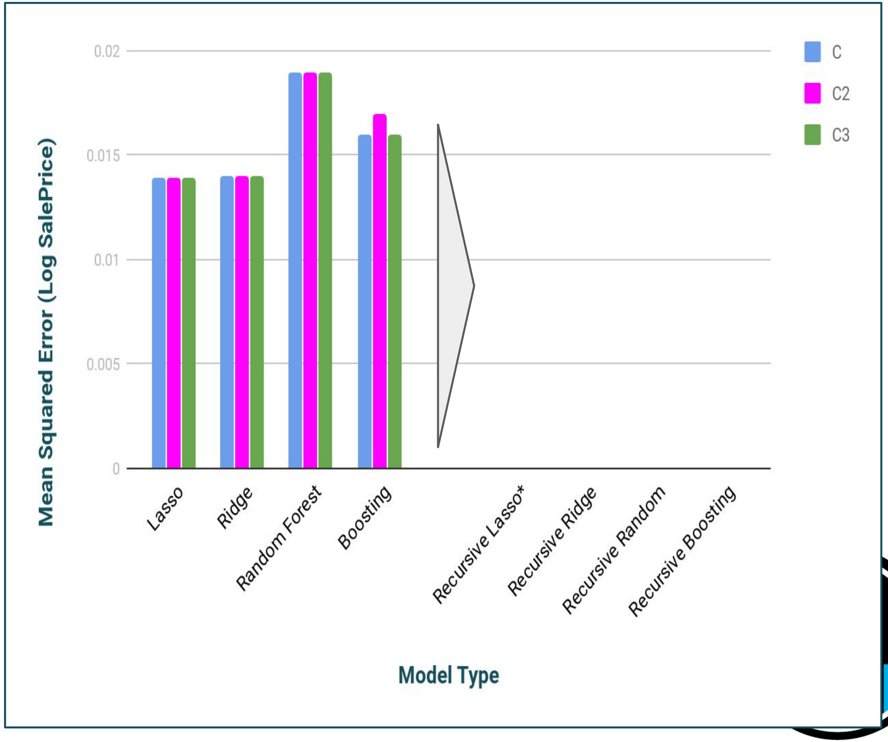


Recursive feature selection created some minor improvements in MSE for several datasets

Recursive Feature Selection

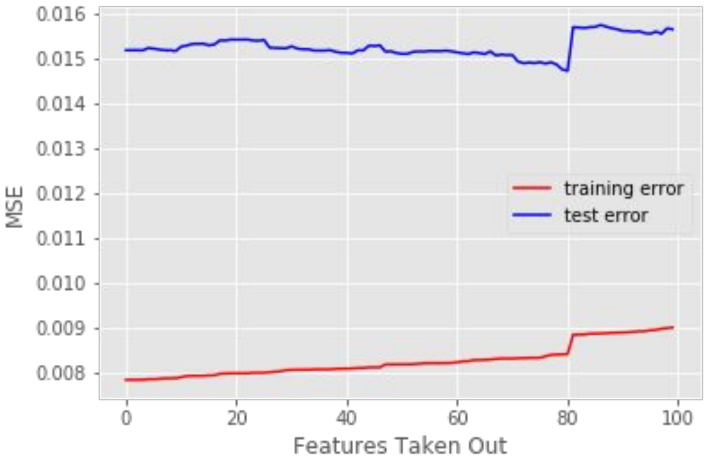


Results

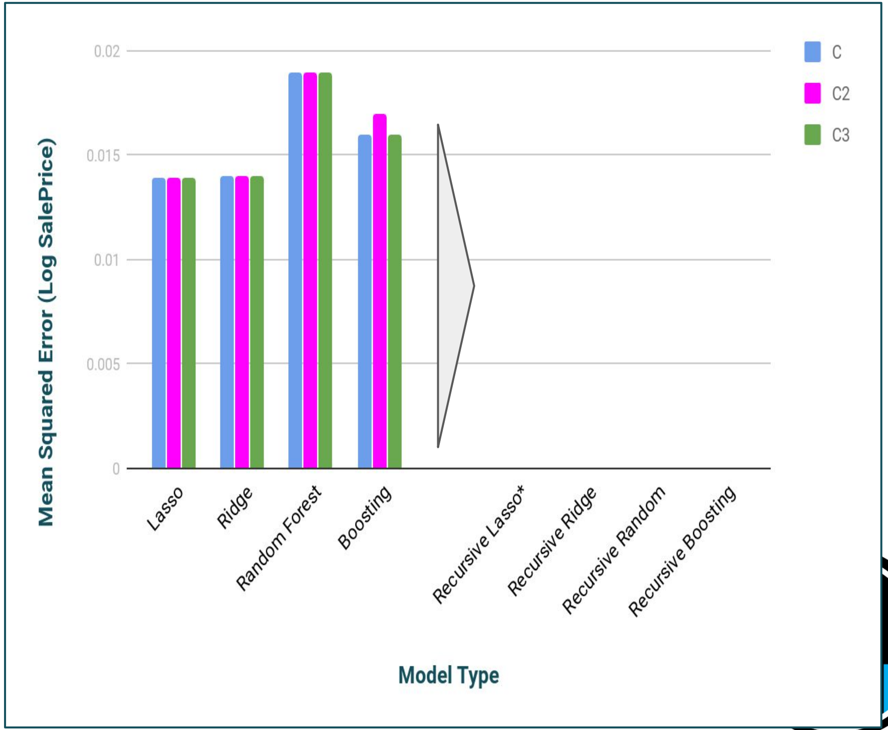


Recursive feature selection created some minor improvements in MSE for several datasets

Recursive Feature Selection



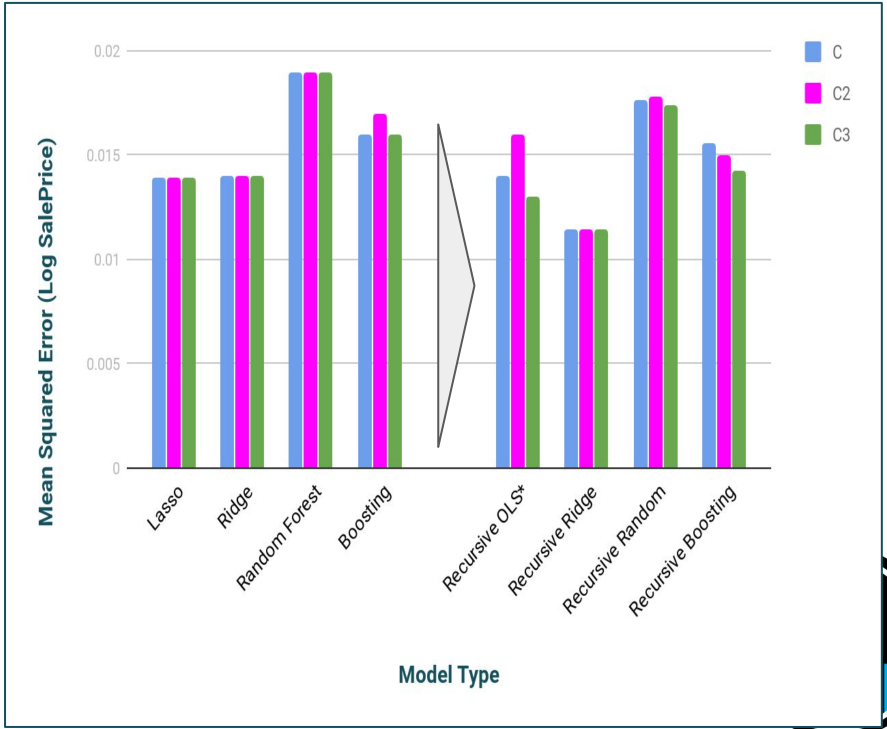
Results



Recursive feature selection created some minor improvements in MSE for several datasets

Recursive Feature Selection

Results



Ensembling across models improved our score for predictions on the Kaggle test sets for some datasets

Model Ensembling

Step 1: Train on optimized training dataset with CV grid search

Recursively Selected Training Dataset

Lasso Regression + Ridge Regression + Elastic Net + Random Forest + Gradient Boosting

Step 2: Minimize MSE to calculate weighted ensemble coefficients

Optimized weighted average predictions for log SalePrice

Regression model

Tree-based model

Input/Output

- Ensembling balances the strengths and weaknesses of the model types in theory
- Scores on the Kaggle test set were improved by ~0.002 (~100 ranks)

Attempts to reduce the dimensionality of our datasets proved futile

Clustering Categories

Nominal Categorical Variables + TotalSF, OverallQual ...

K-means clustering (*k selected via silhouette analysis*)

Regrouped Nominal Categorical Variables

Finding Principal Axes

All Continuous Variables

Principal Component Analysis

Principal component vectors

ML Model Pipeline

Results

- Attempted k-means clustering to regroup categorical variables, as several showed a large class imbalance
- We attempted to use PCA for dimensionality reduction for continuous variables
- Neither technique improved our model performance

ML Model

Input/Output