Predicting House Prices:
A Machine Learning Project

Sale Price Of Houses in Ames, Iowa

Data Cleanup:

Interpreting ‘NA’ Values

Two Types of ‘NA’ Values

NA Values With Meaning

- NA has been encoded to mean ‘other’ or ‘none’ for multiple features
- Replace NAs with appropriate values

NA Values With No Meaning

- Interpolate all other values
- MSZoning Correlated with Neighborhood
- Utilities Feature has 2 NA values, all of other observations are ‘AllPub’
- Kitchen Quality and Overall Quality are highly correlated
- Functional feature is very highly skewed to ‘Typical’ (over 95%)
- Garage Cars and Garage Area NA value should correspond to their being no Garage

Lot Frontage: A Highly Spurious Feature

- Lot frontage is described as “Linear feet of street connected to property”
- Approximately 15% of observations have ‘NA’ recorded for this variable
- High correlation with another feature ‘Lot Area’ (~0.65 Pearson)
- Compute mean ratio of Lot Area and Lot Frontage for entire dataset
- For incomplete observations, multiply ratio by recorded Lot Area to impute Lot Frontage

Feature Engineering

The Process of Dimensionality Reduction

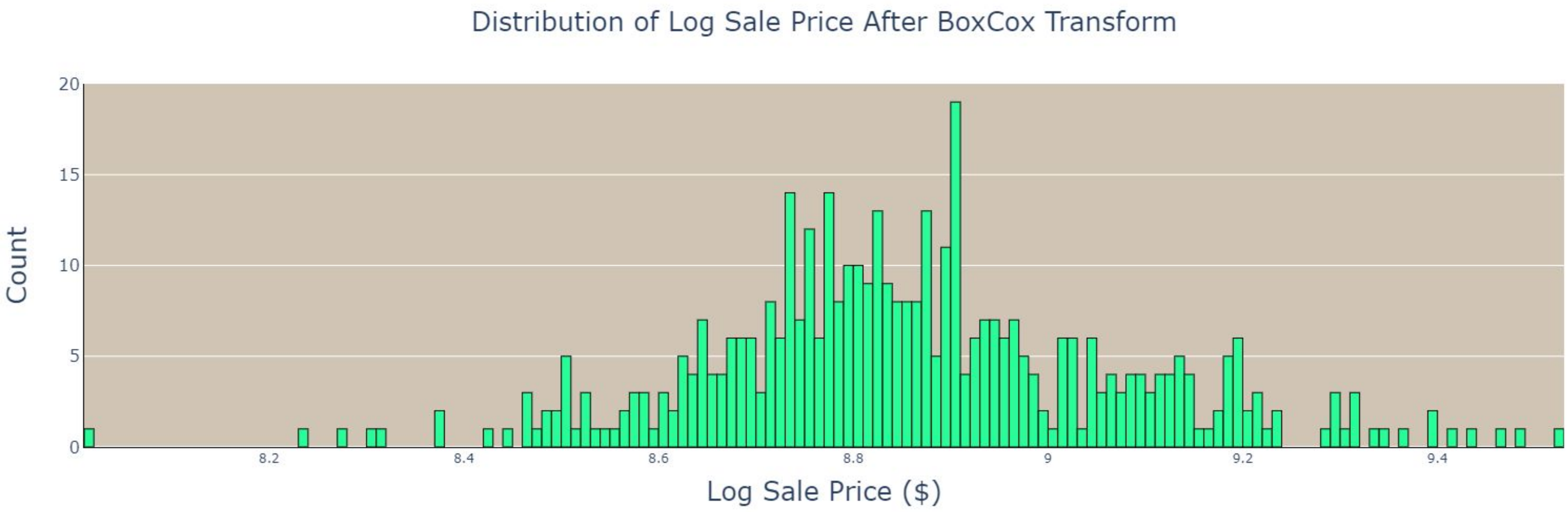
Low Count Categorical Values

- Choosing what data gets fed to training model is extremely important
- For select features there is only one observation of a particular class
    - For Example: ‘Roof Material’ has only one observation for class ‘Metal’
- Training Model will learn no real information for this feature, but it may affect overall predictions
- Drop values from training dataset that meet this criteria

Distribution of Target Variable

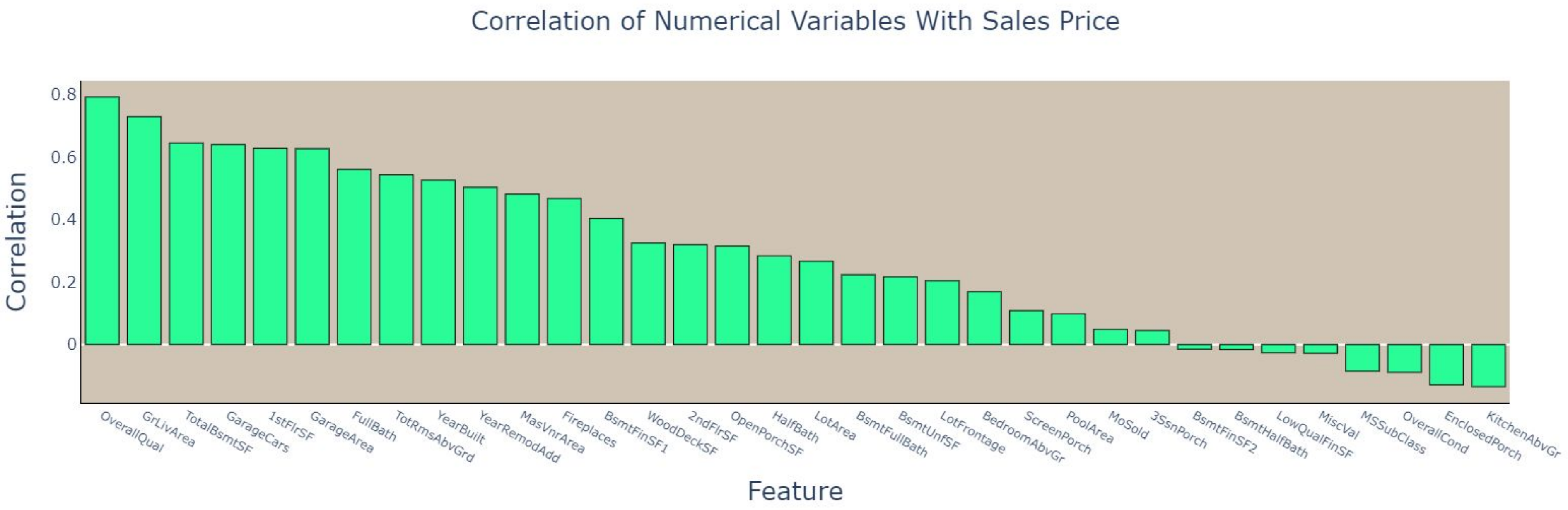
Not Normally Distributed

Target Variable Transformed with Box Cox



Much closer to normal distribution

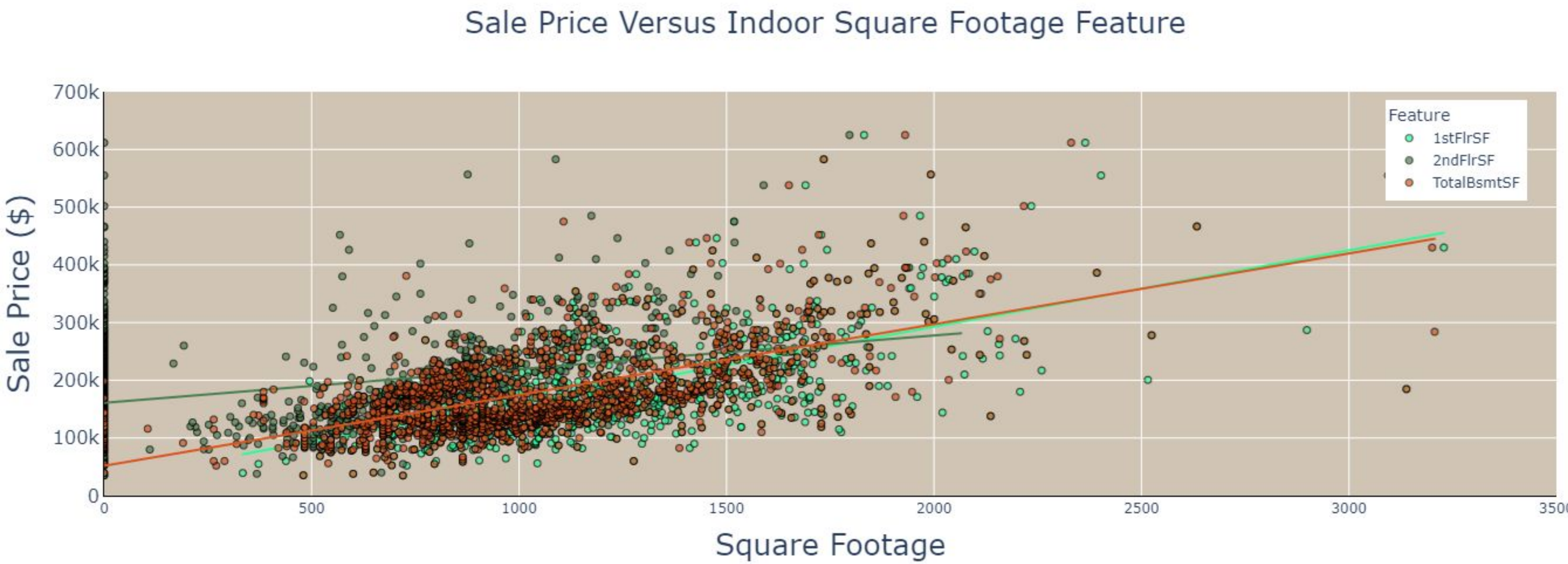
Correlation of Numerical Features to Sale Price



Combining Area Features

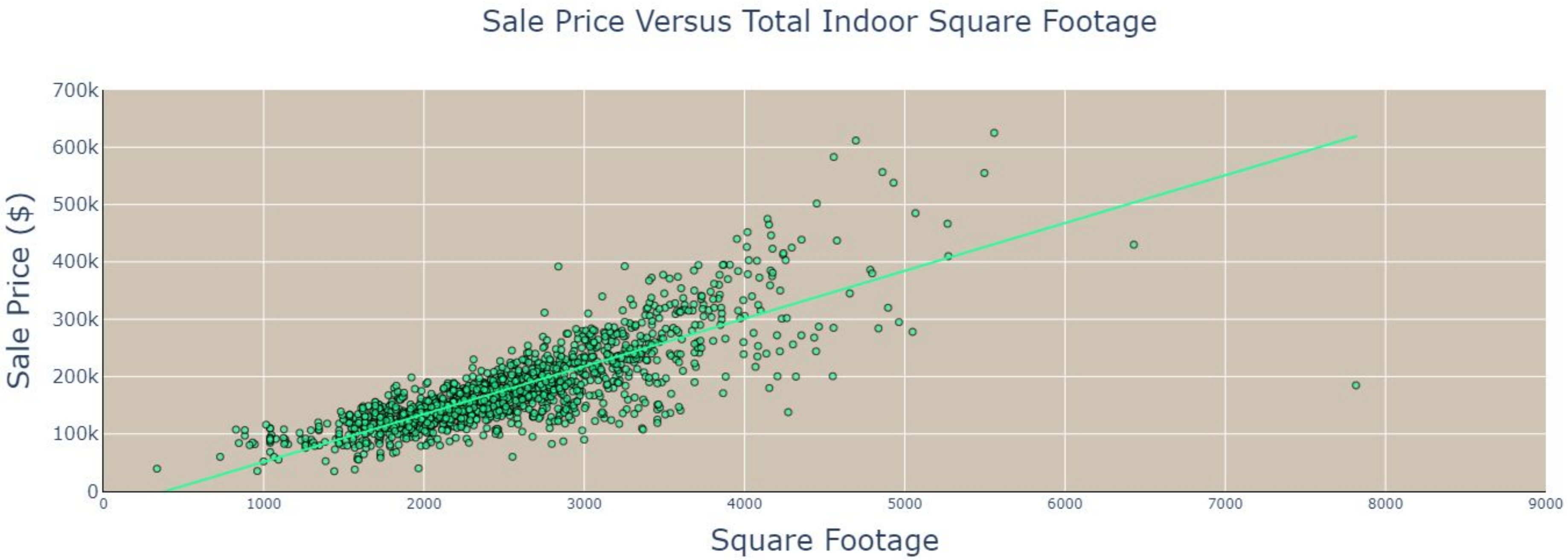
- Correlation data shows high importance with multiple features concerning area
- Some area features contain redundant information
    - Above ground living area would be highly correlated with first and second floor area
- Investigate linear relationship between sale price and size variables
- Combine where possible and see if relationship is maintained

Indoor Area

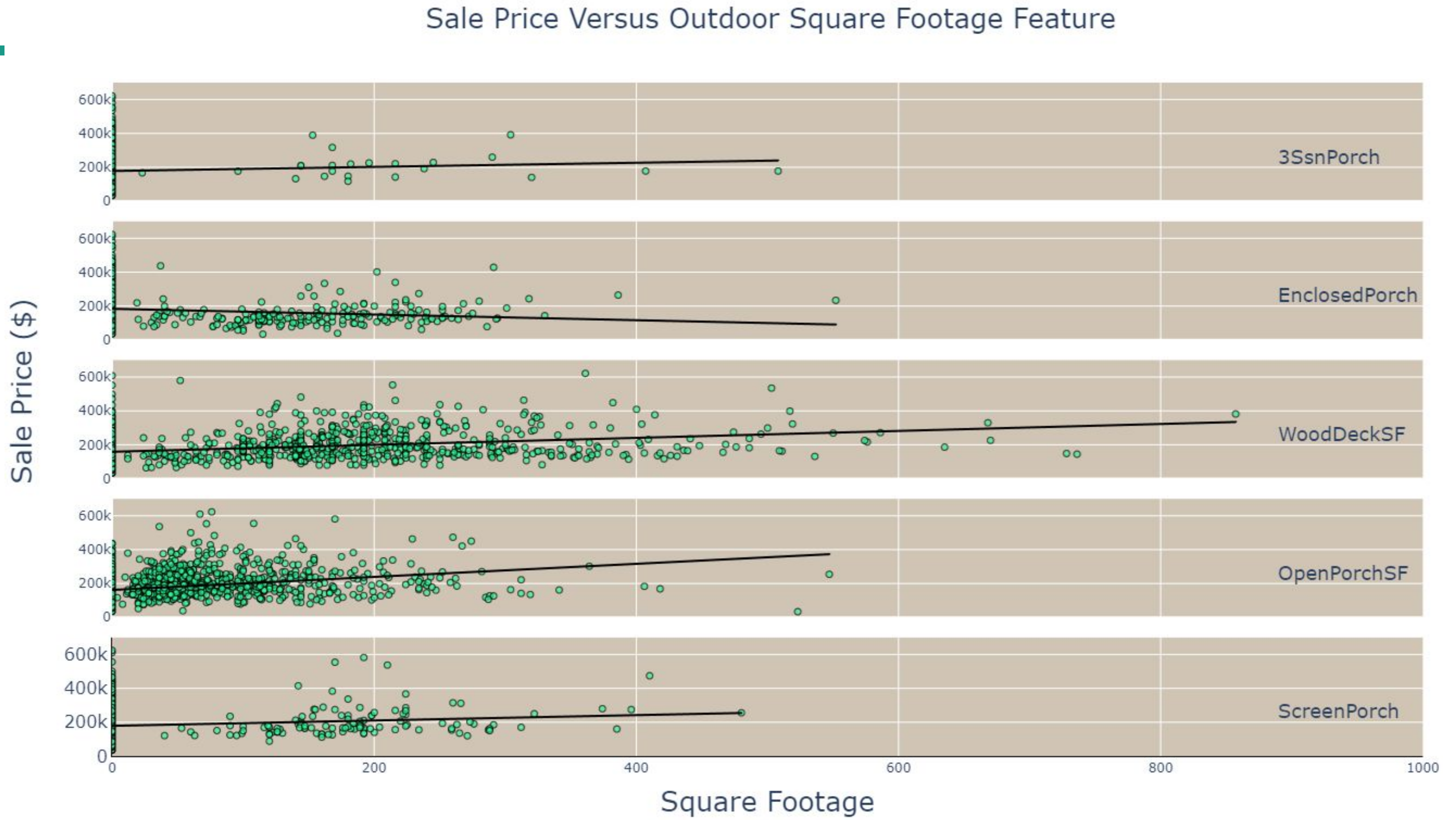


Sale Price Versus Indoor Square Footage Feature
Sale Price ($)
700k
600k
500k
400k
300k
200k
100k
0
0
500
1000
1500
2000
2500
3000
3500
Square Footage
Feature
1stFlrSF
2ndFlrSF
TotalBsmtSF

Combined Indoor Area



Outdoor Attached Area



Outdoor Attached Features Combined

Bathrooms and Sale Price

Total Number of Bathrooms

Total Living Area

Correlation between these two variables is 0.87

Quality and Condition Variables

Potential Problems

- There are 12 variables related to quality and condition.
- All nominally encoded
- Possibly subjective in nature
- Are there strong relationships between these variables and sale price?

Possible Solutions

- Encode values to ordinal with a replacement dictionary
- Could combine variables in a weighted manor
- Look at individual relationships and drop uncorrelated features

Usefulness of Quality Variables

	Feature	Correlation
0	OverallQual	0.792129
1	ExterQual	0.684333
3	KitchenQual	0.660592
2	BsmtQual	0.586866
4	FireplaceQu	0.518023
5	GarageQual	0.268863
6	PoolQC	0.122663

Usefulness of Condition Variables

	Feature	Correlation
4	GarageCond	0.257851
3	BsmtCond	0.206717
2	ExterCond	0.004743
1	OverallCond	-0.088405

Eliminating More Not Useful Features

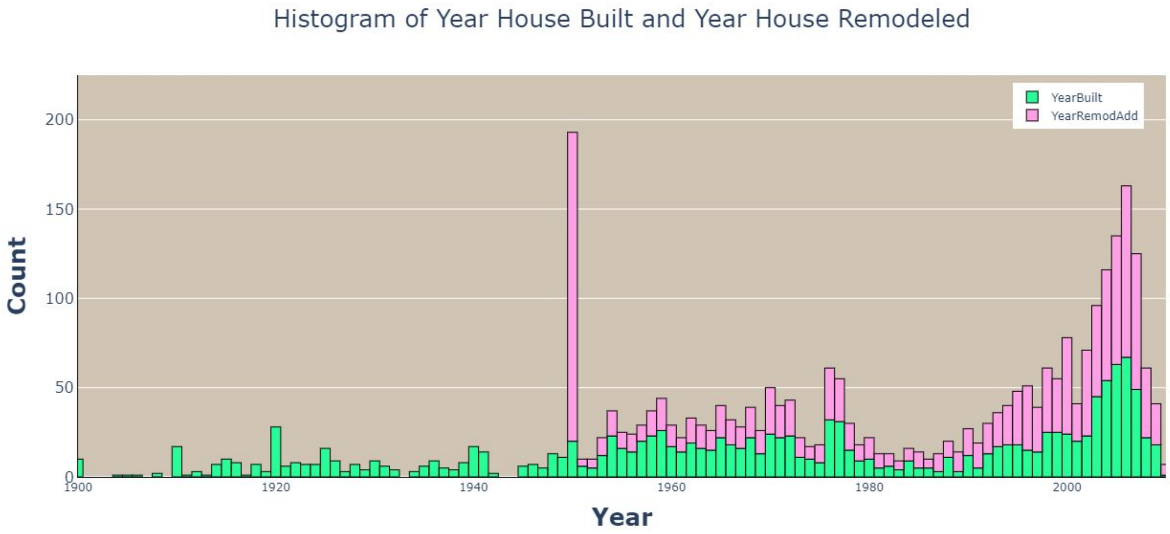
	Feature	Correlation
0	GarageYrBlt	0.259369
1	BsmtUnfSF	0.217009
2	BedroomAbvGr	0.168823
3	PoolArea	0.098241
4	MoSold	0.049140
5	BsmtFinSF2	-0.014524
6	LowQualFinSF	-0.026075
7	KitchenAbvGr	-0.134535

- Number Of Bedrooms
- Number Of Kitchens
- Finished Square Feet
- Garage Year Built
- Month Sold
- Lot Frontage
- Pool Area

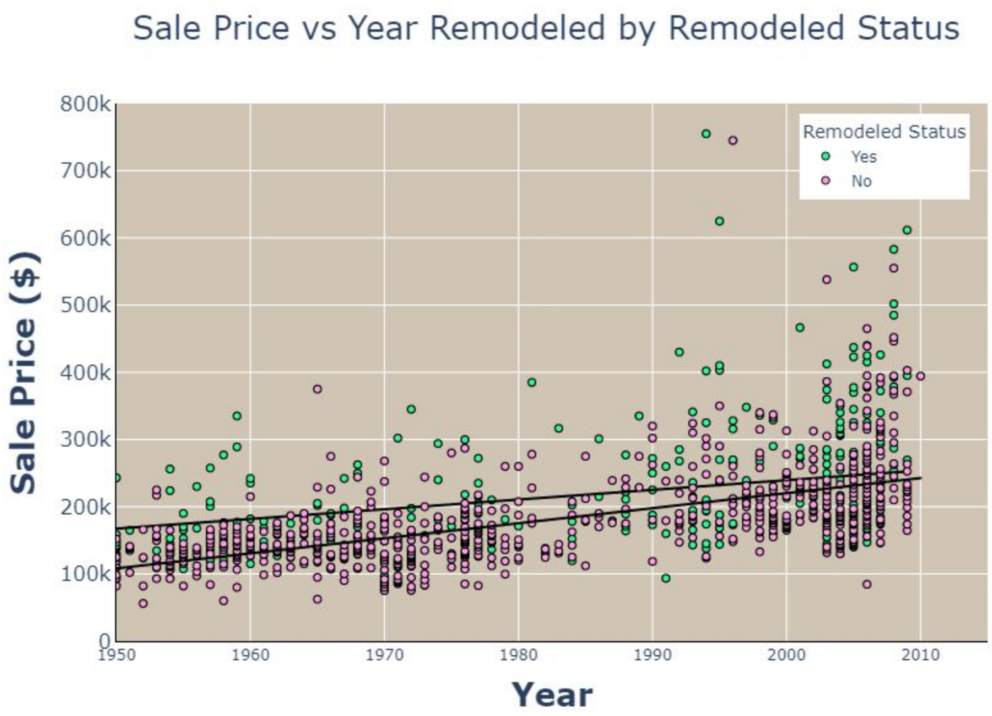
A Variance Inflation Factor Mystery

- Age related Variables have an extremely high VIF
- This level of information redundancy is very suspicious

	Feature	VIF
4	YearBuilt	11648.194337
5	YearRemodAdd	19672.494145
20	YrSold	22378.343953



House Remodeling Feature Solution



- Create binary identifier if house was remodeled
- Drop year house was built

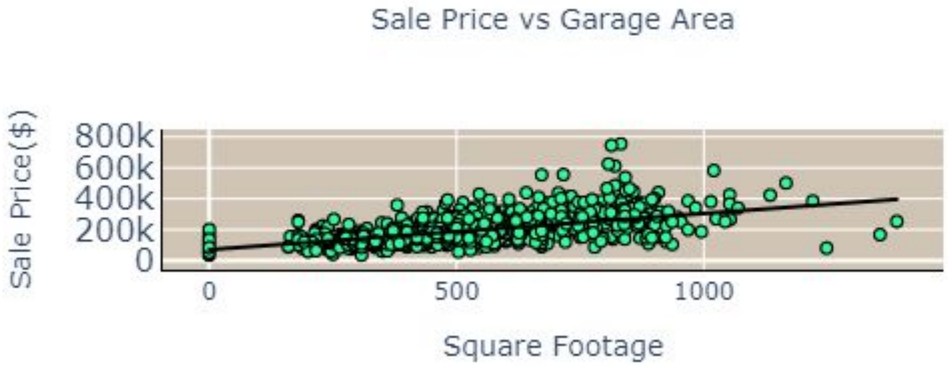
Garage Size Redundancy

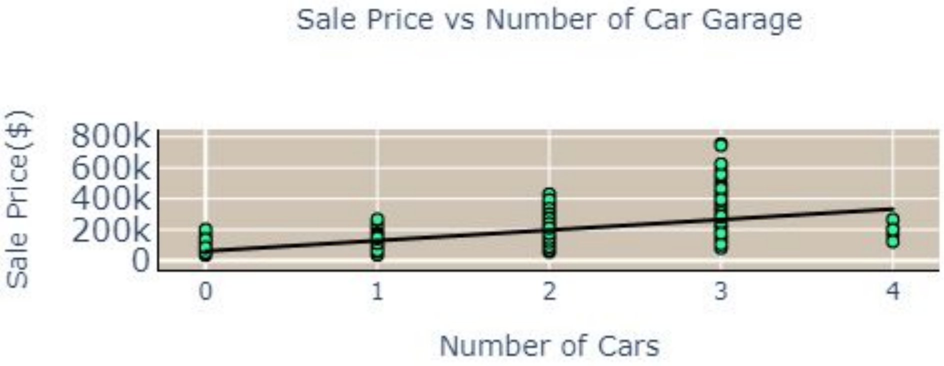
- Two variables exist describing garage size; ‘Garage Area’ and Number of Cars
- These variables are using different metrics to measure the same thing
- VIF data confirms

	Feature	VIF
14	GarageCars	36.230931
15	GarageArea	30.655436

- Which variable to keep?

Does it Matter?






```
==============================================
R-squared:                               0.392
Adj. R-squared:                          0.392
F-statistic:                             929.2
Prob (F-statistic):                 6.90e-158
Log-Likelihood:                        -17956.
AIC:                                 3.592e+04
BIC:                                 3.593e+04
```

```
==============================================
R-squared:                               0.409
Adj. R-squared:                          0.409
F-statistic:                             997.3
Prob (F-statistic):                 9.21e-167
Log-Likelihood:                        -17935.
AIC:                                 3.587e+04
BIC:                                 3.589e+04
```

Eliminating Even More Not Useful Features

	Feature	VIF	P-Value
5	MasVnrArea	1.767059	0.011679
20	TotBr	21.467268	0.013739
11	KitchenQual	76.701825	0.019655
2	LotArea	5.232557	0.022662
17	YrSold	17209.731119	0.324359
18	TotSq	43.581884	0.343021
0	MSSubClass	3.593098	0.515769

- Lot Area
- Year Sold
- MS Sub Class
- Total Indoor Square Footage?

How to Find Usefulness of Categorical Features

- Temporarily encode all features to ordinal
- Use Chi-Squared measurements to find useful features

Modeling

Data Preparation

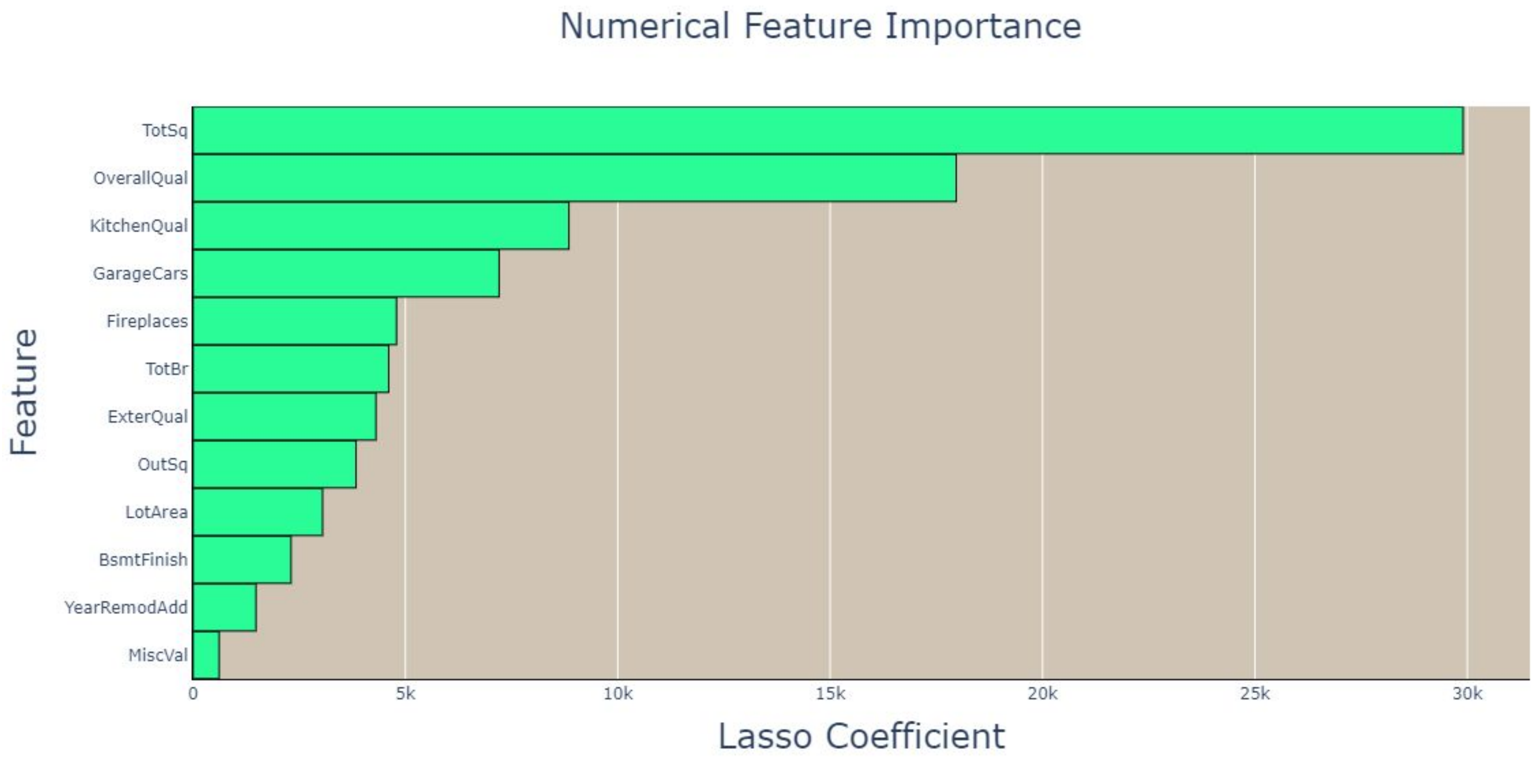
1. Use Standard Scaler to scale all numerical values to a normal distribution
2. Create a train test split (70/30)
3. Create two copies of Data
   a. Numerically encode nominal categories for tree based models
   b. Dummify nominal and ordinal categories for penalized linear regression models

Using Lasso for Dimensionality Reduction

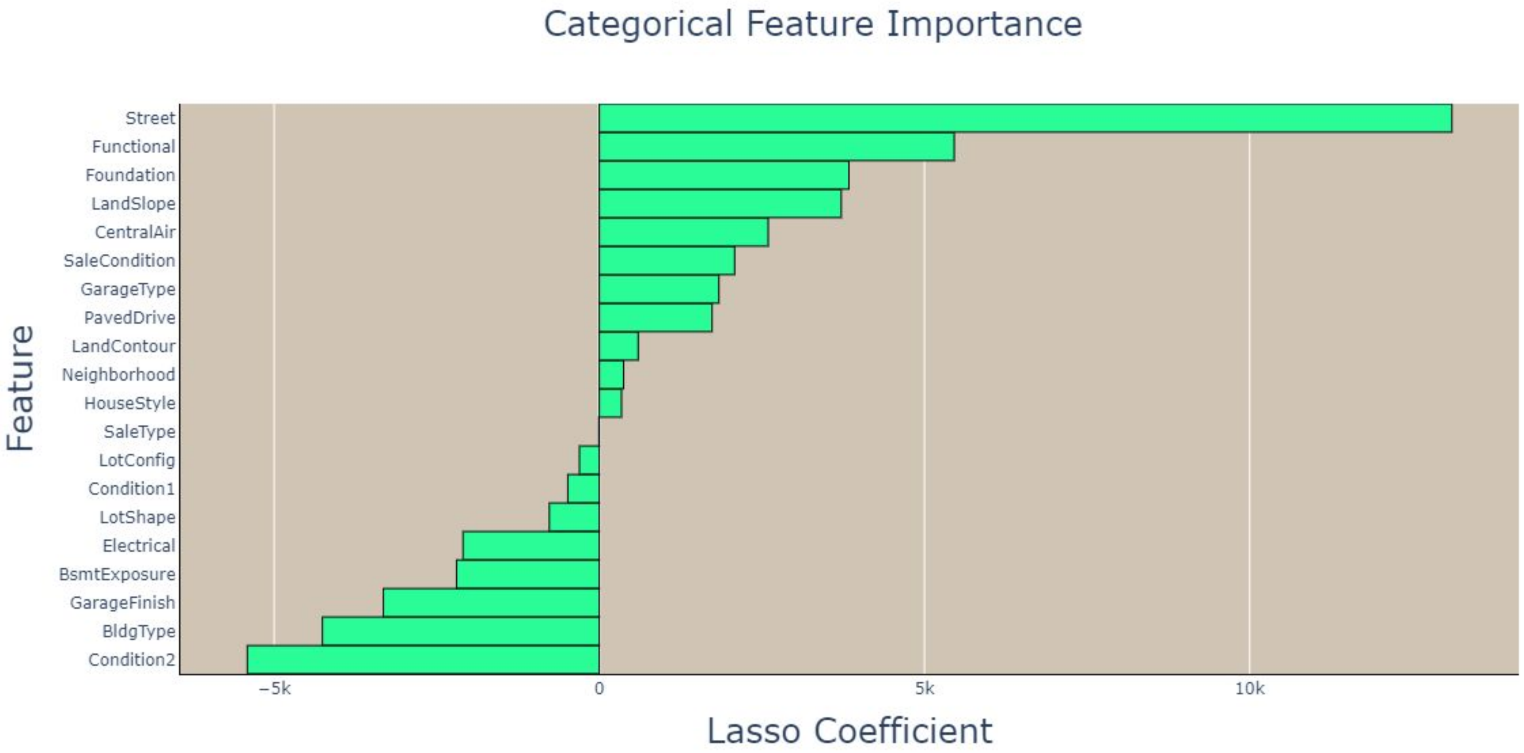
Lasso Regression coefficients can be used to eliminate non useful features

- Masonry Type
- Heating Quality Index
- Remodeled Status
- Secondary Exterior Material
- Roofstyle
- Basement Quality
- Pool Area
- Utilities
- Alley
- MS Zoning
- Miscellaneous Feature

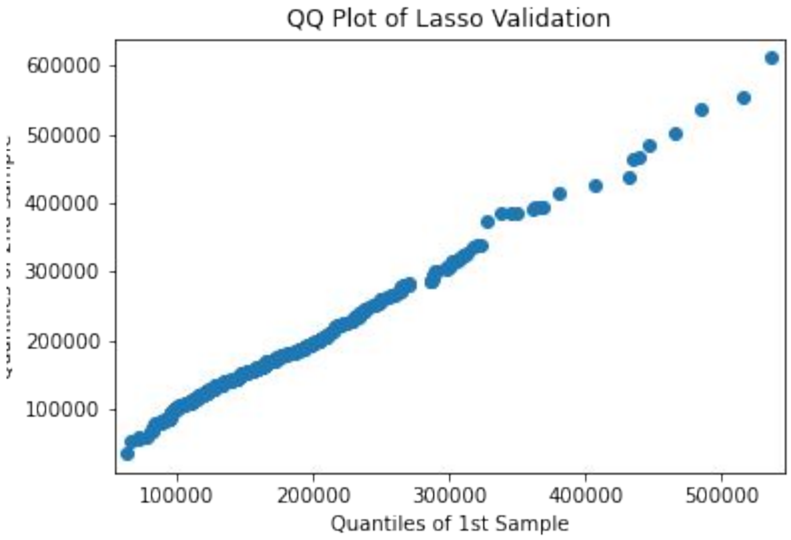
Lasso Feature Importance: Numerical



Lasso Feature Importance: Categorical

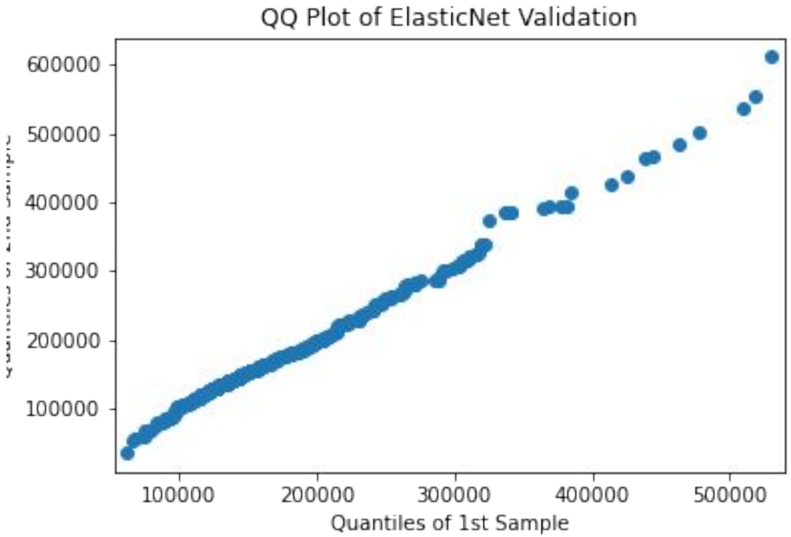


Lasso Validation Predictions



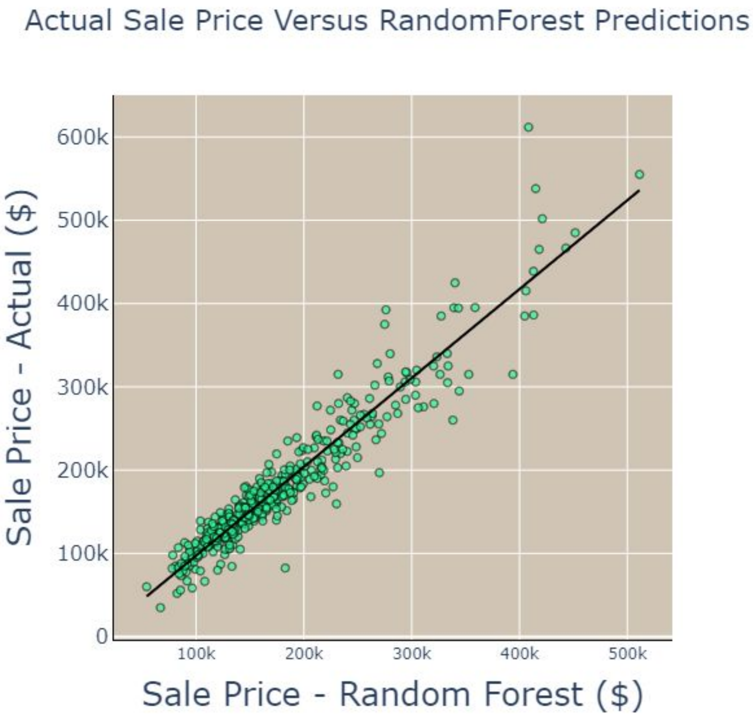
R Squared = 0.921
Mean Average Error = $15,710

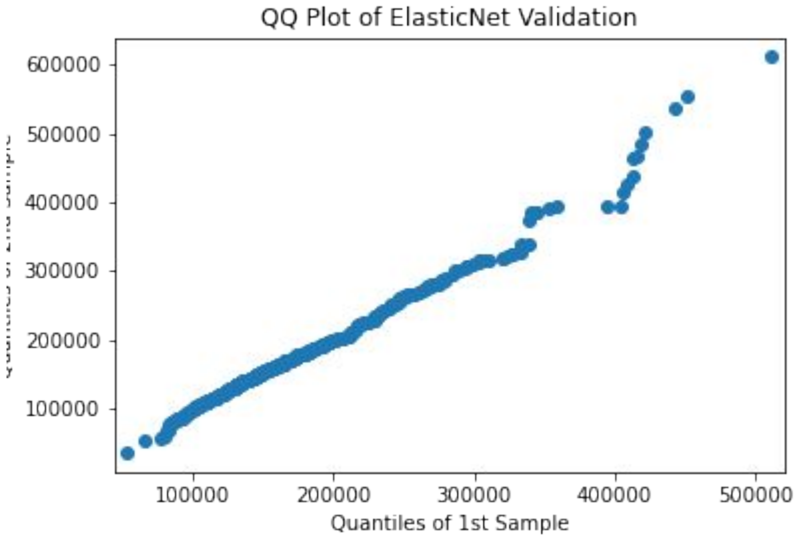
ElasticNet Validation Predictions



R Squared = 0.926
Mean Average Error = $15,268

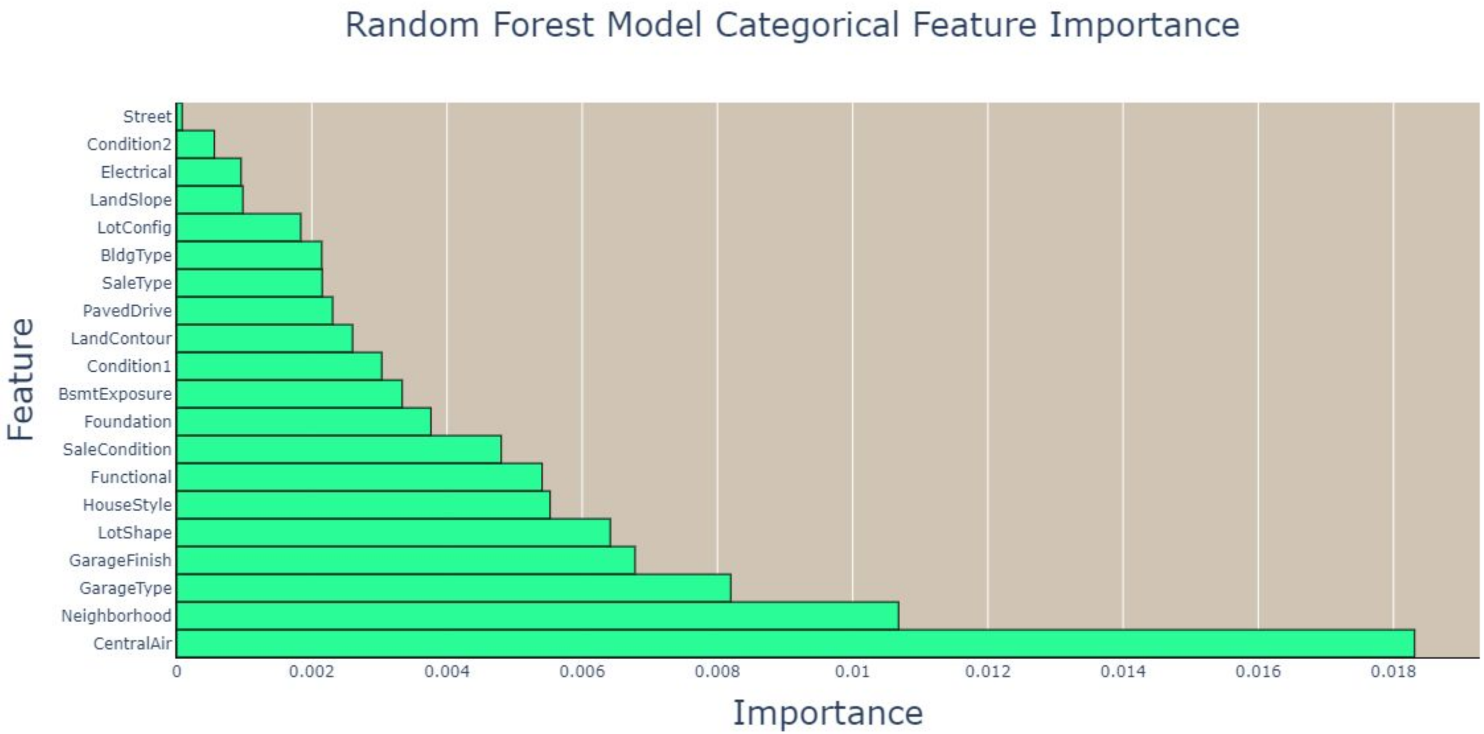
Random Forest Validation Predictions



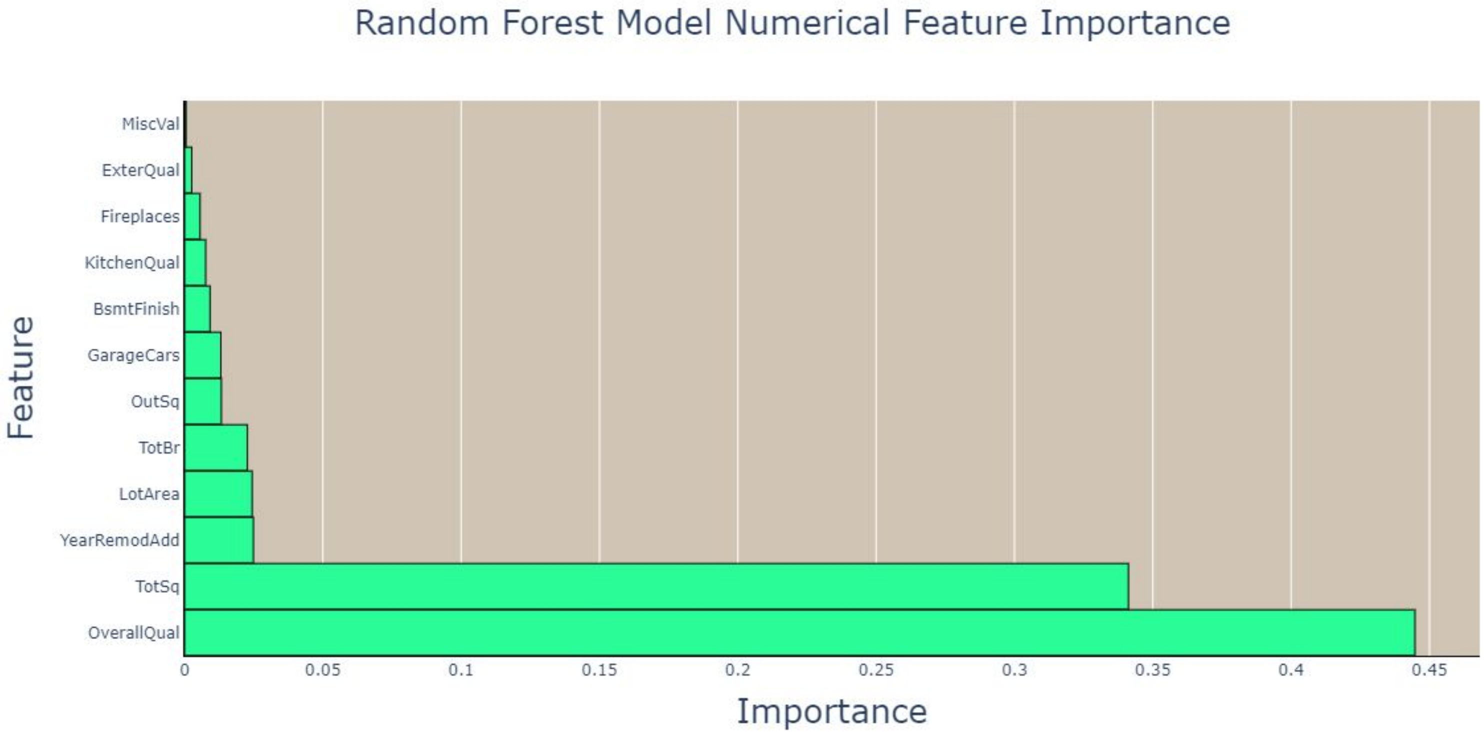


R Squared = 0.914
Mean Average Error = $16,418

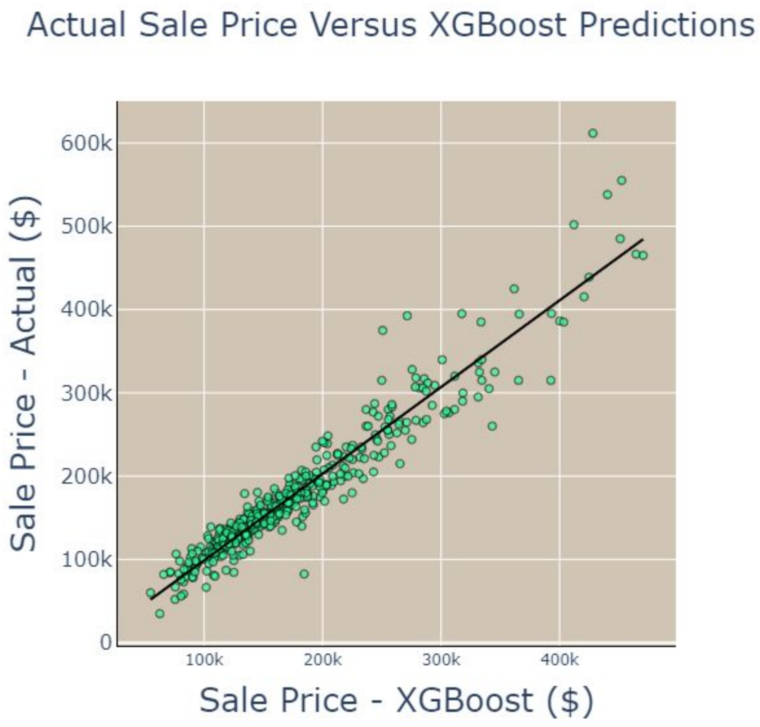
Random Forest Model Feature Importance

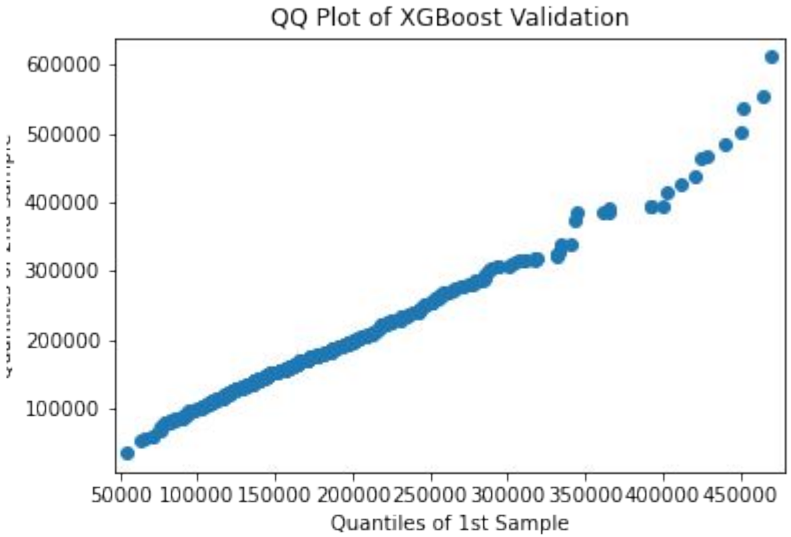


Random Forest Model Feature Importance



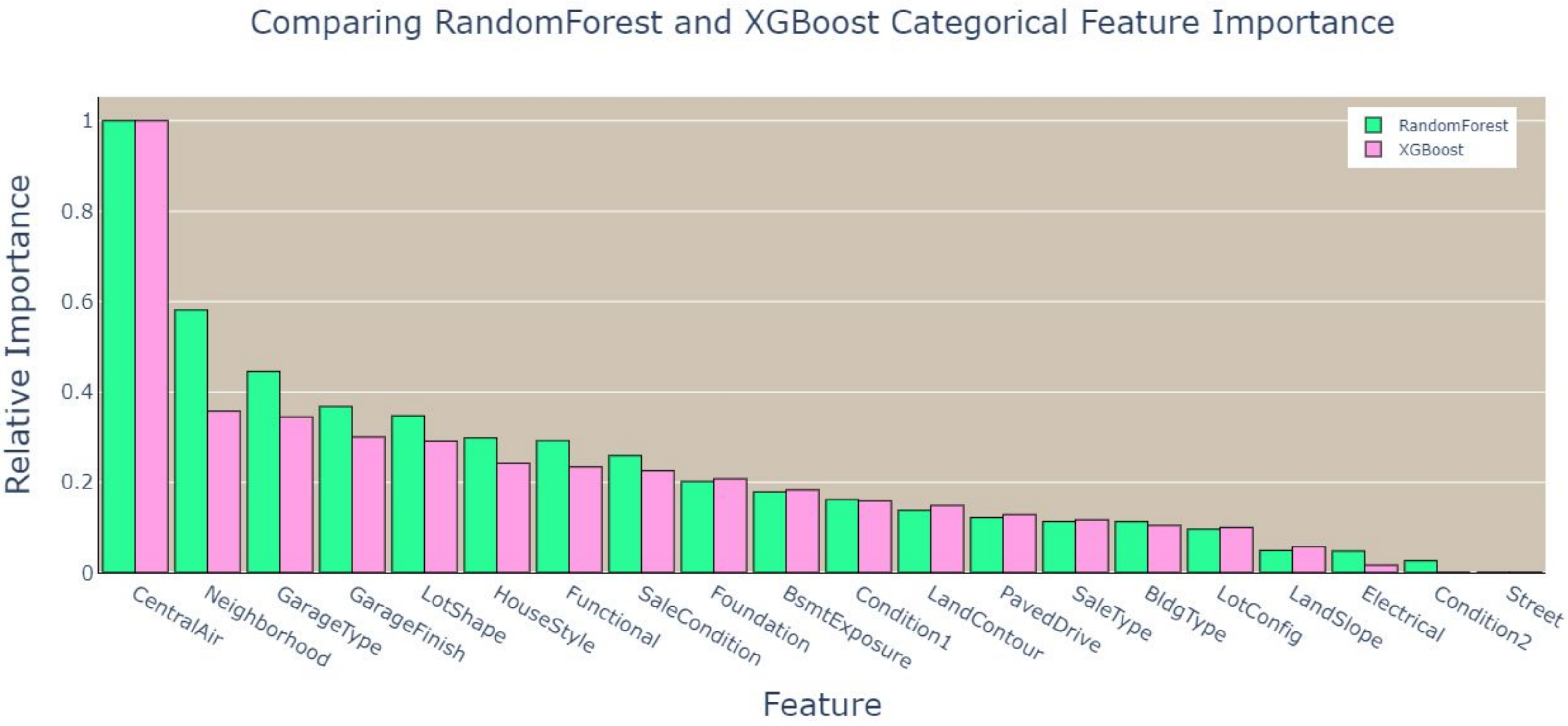
XGBoost Validation Predictions



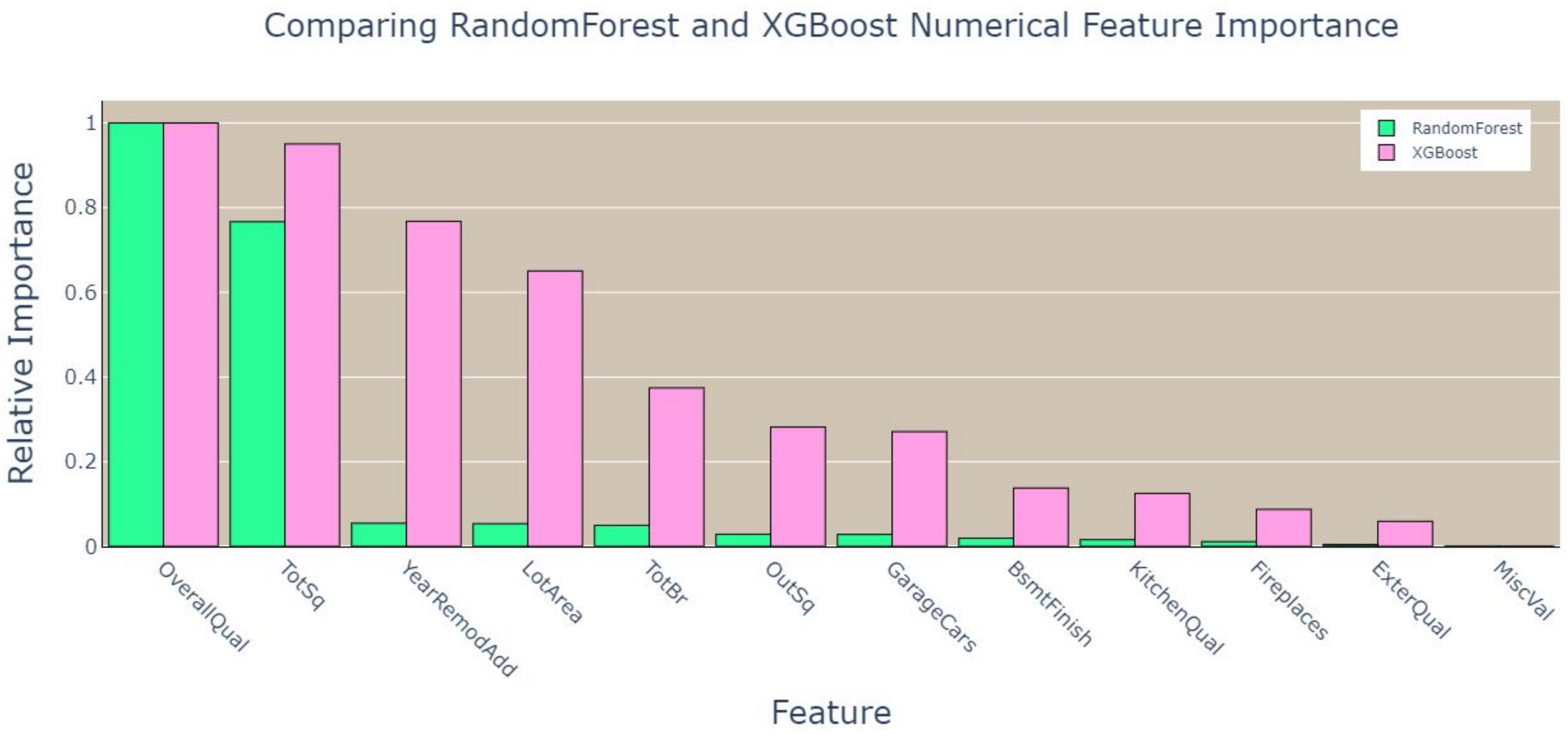


R Squared = 0.928
Mean Average Error = $15,015

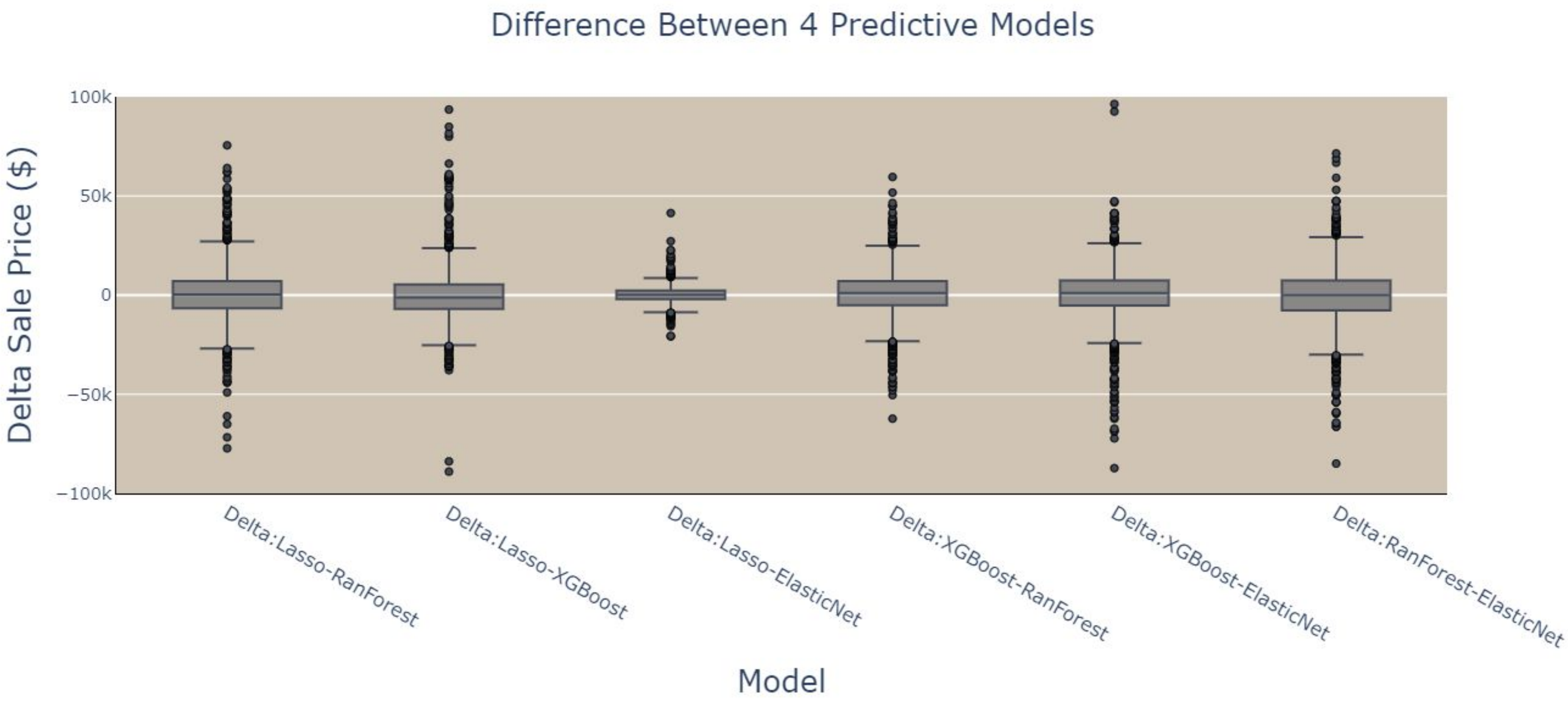
Comparing Feature Importance



Comparing Feature Importance

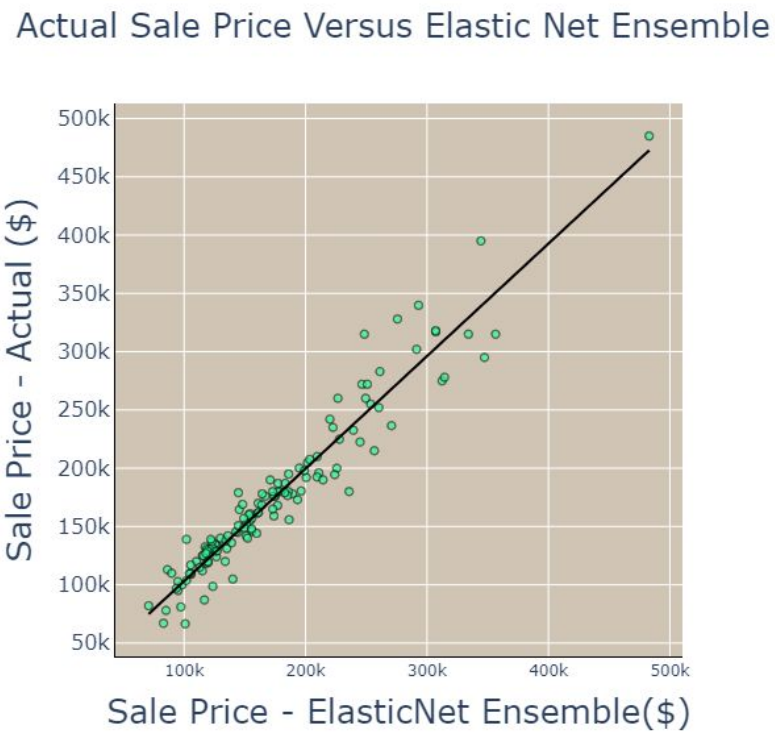


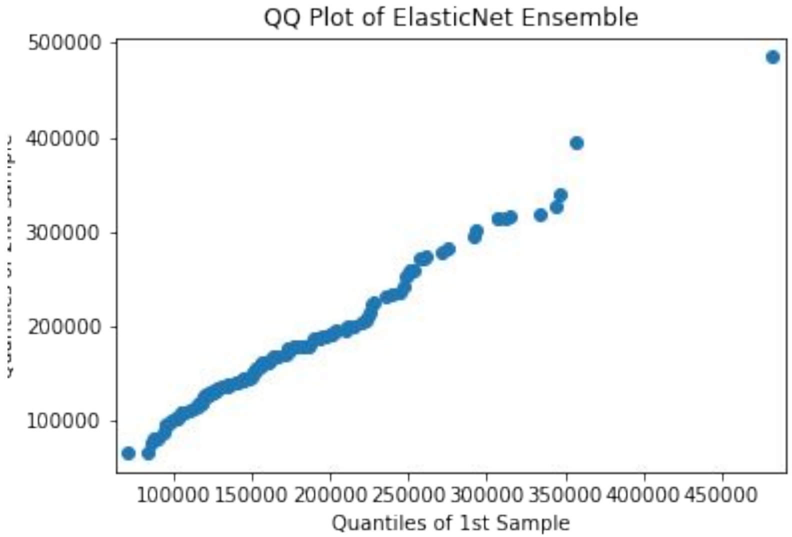
Ensembling With Averaging Models



50-50 XGBoost and Elastic Net Ensemble

Ensembling: Meta Modeling With Elastic Net

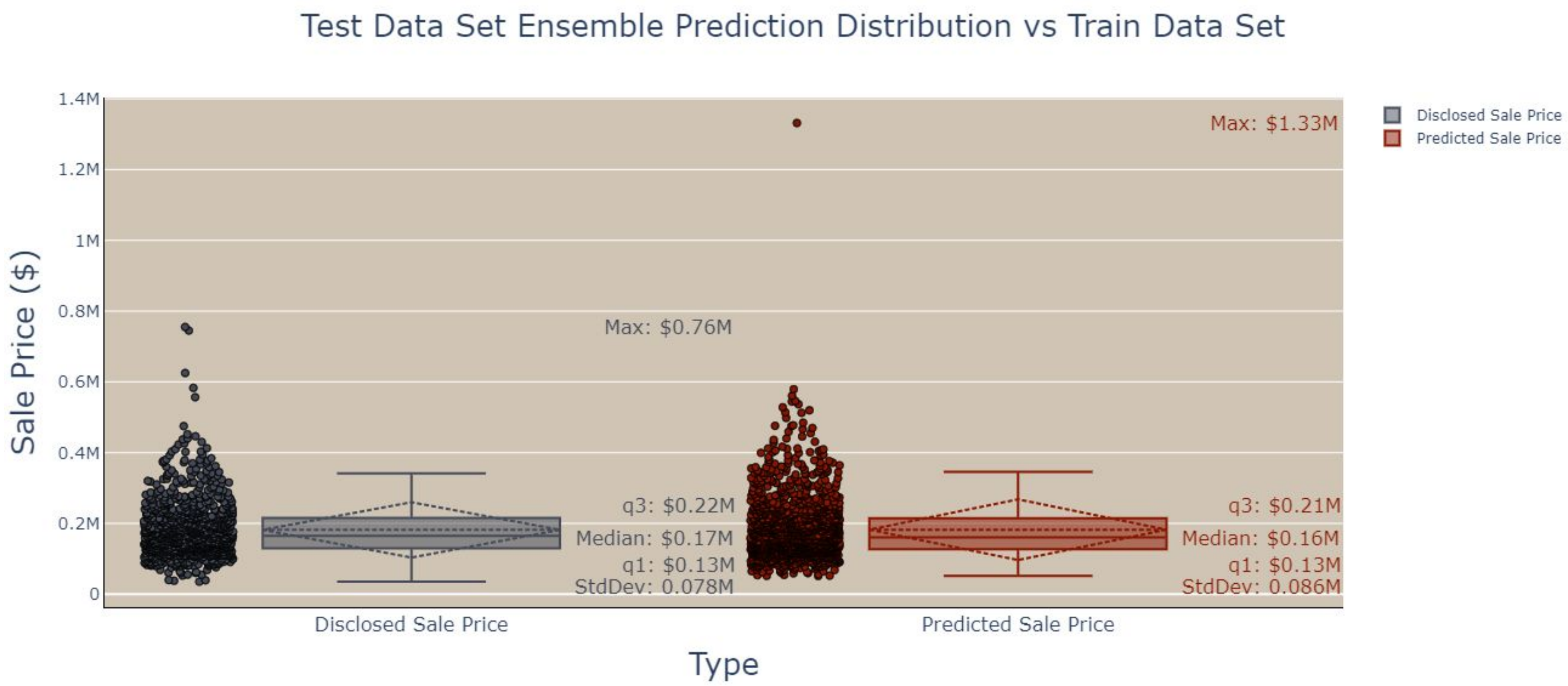




R Squared = 0.962

Mean Average Error = $13,661

Elastic Net Ensemble Distribution Comparison



With More Time

- Look into algorithmic imputation methods and feature selection
- Try other models like Pytorch or other boosting methods
- Try Different ensembling methods
- Dip deeper into differences between models
- Restyle my Matplotlib graphs
- Find methods for validating final 96% R Squared