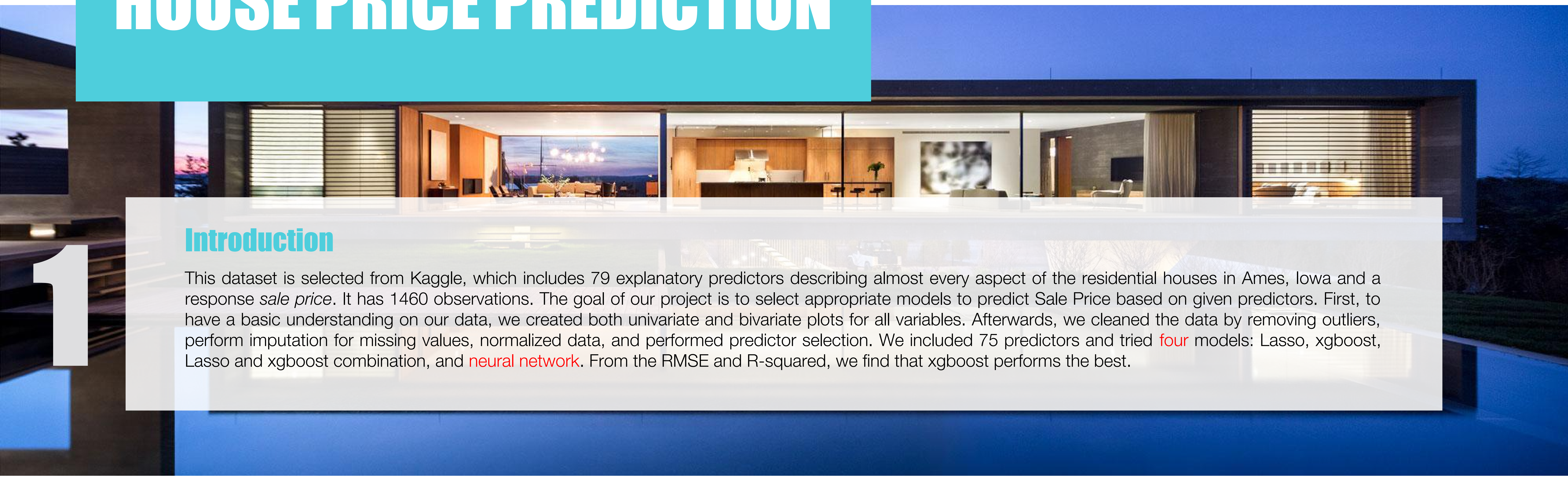
HOUSE PRICE PREDICTION

Na Song, Mai Ghazy, Yifang Li
CPSC 6300



1

Introduction

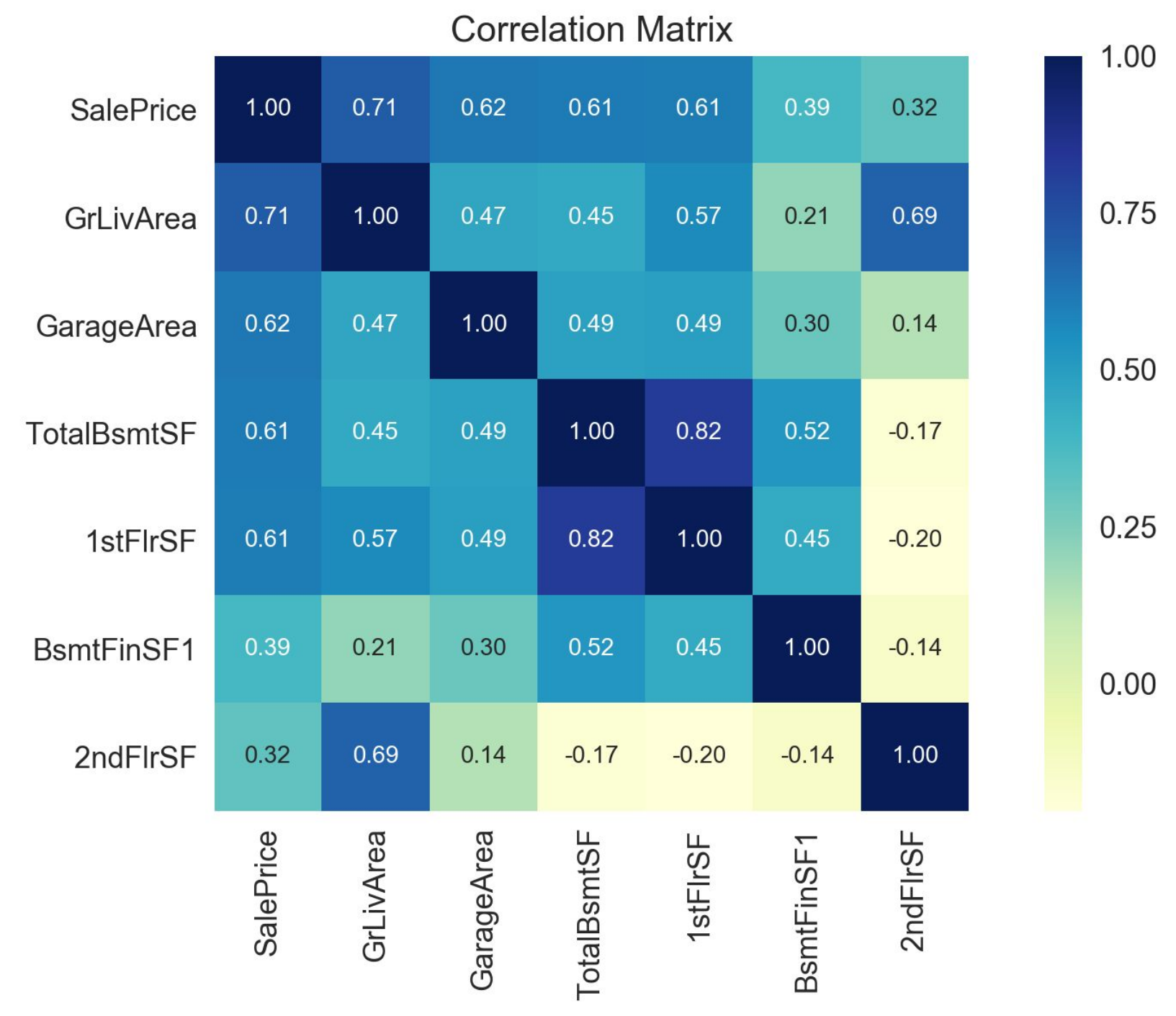
This dataset is selected from Kaggle, which includes 79 explanatory predictors describing almost every aspect of the residential houses in Ames, Iowa and a response *sale price*. It has 1460 observations. The goal of our project is to select appropriate models to predict Sale Price based on given predictors. First, to have a basic understanding on our data, we created both univariate and bivariate plots for all variables. Afterwards, we cleaned the data by removing outliers, perform imputation for missing values, normalized data, and performed predictor selection. We included 75 predictors and tried four models: Lasso, xgboost, Lasso and xgboost combination, and neural network. From the RMSE and R-squared, we find that xgboost performs the best.

2

Exploratory Data Analysis (EDA)

We generated univariate plot for each predictor. For bivariate plots, we created a correlation matrix for all quantitative predictors, and bar plots for all qualitative predictors to show their relationships with the response Sale Price. Due to the limited space, below we just show bivariate plots for **the top 10 important predictors**.

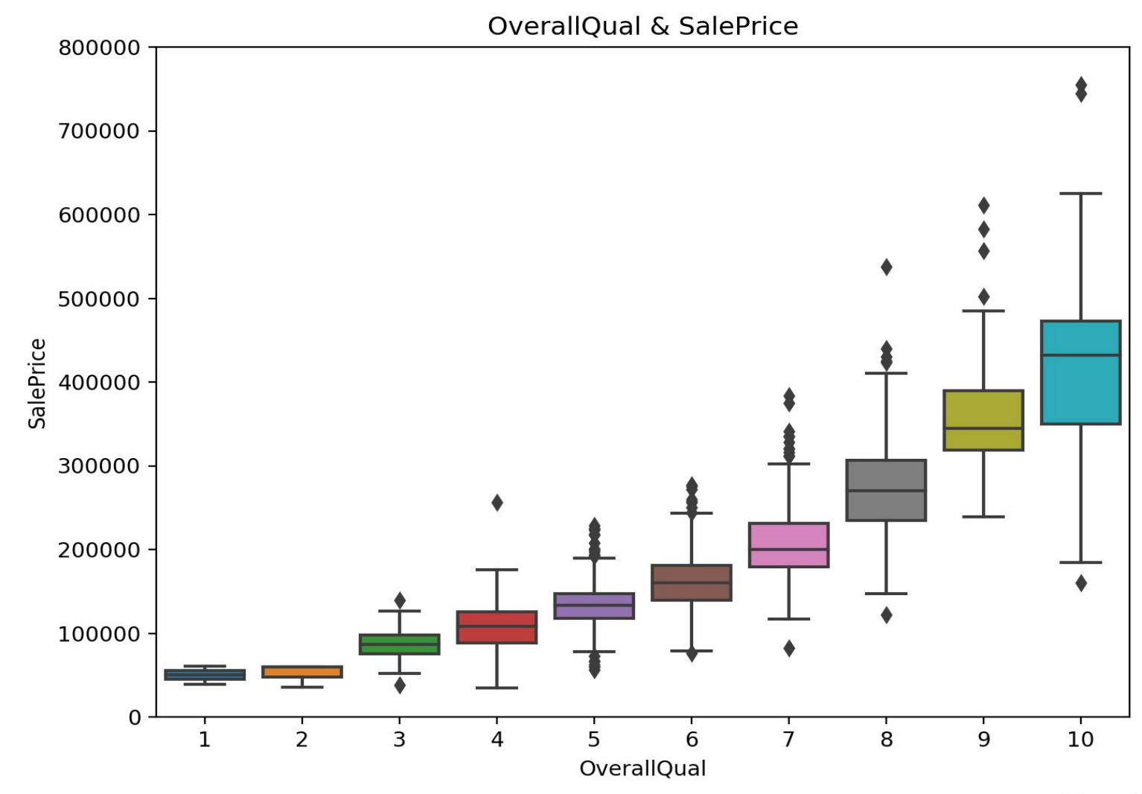
Bivariate Plots For 7 Quantitative Predictors In Top 10

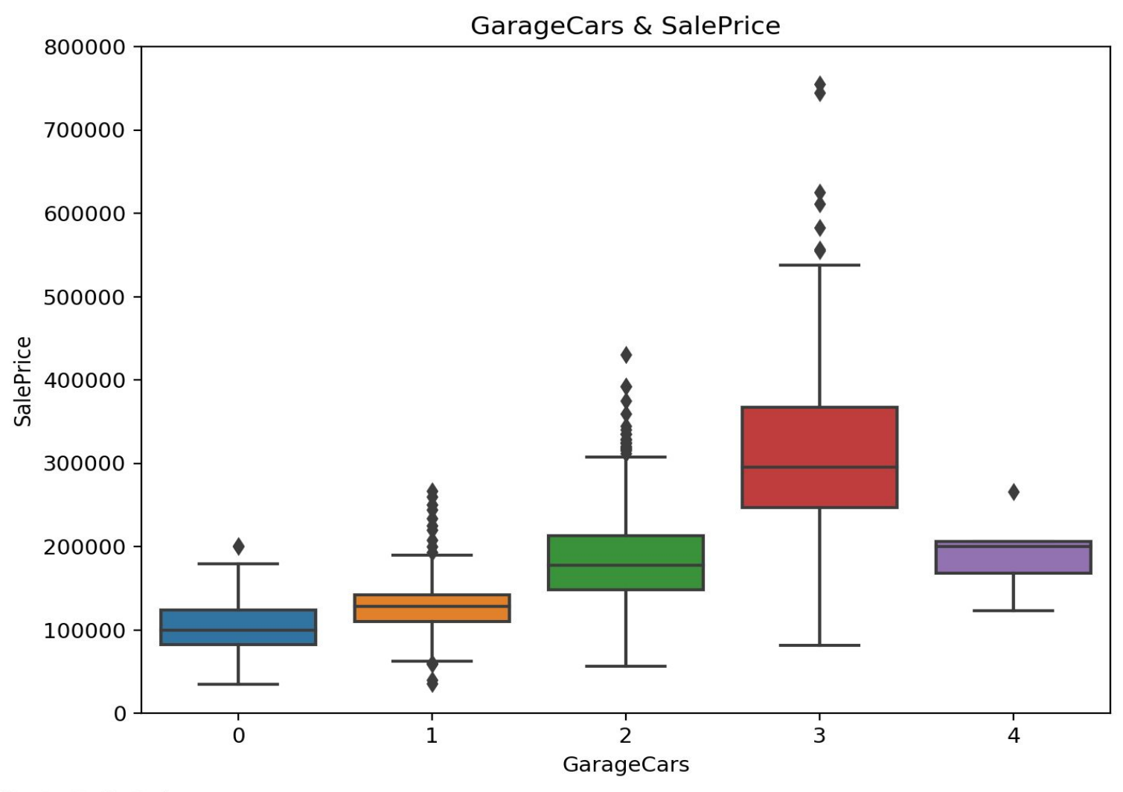


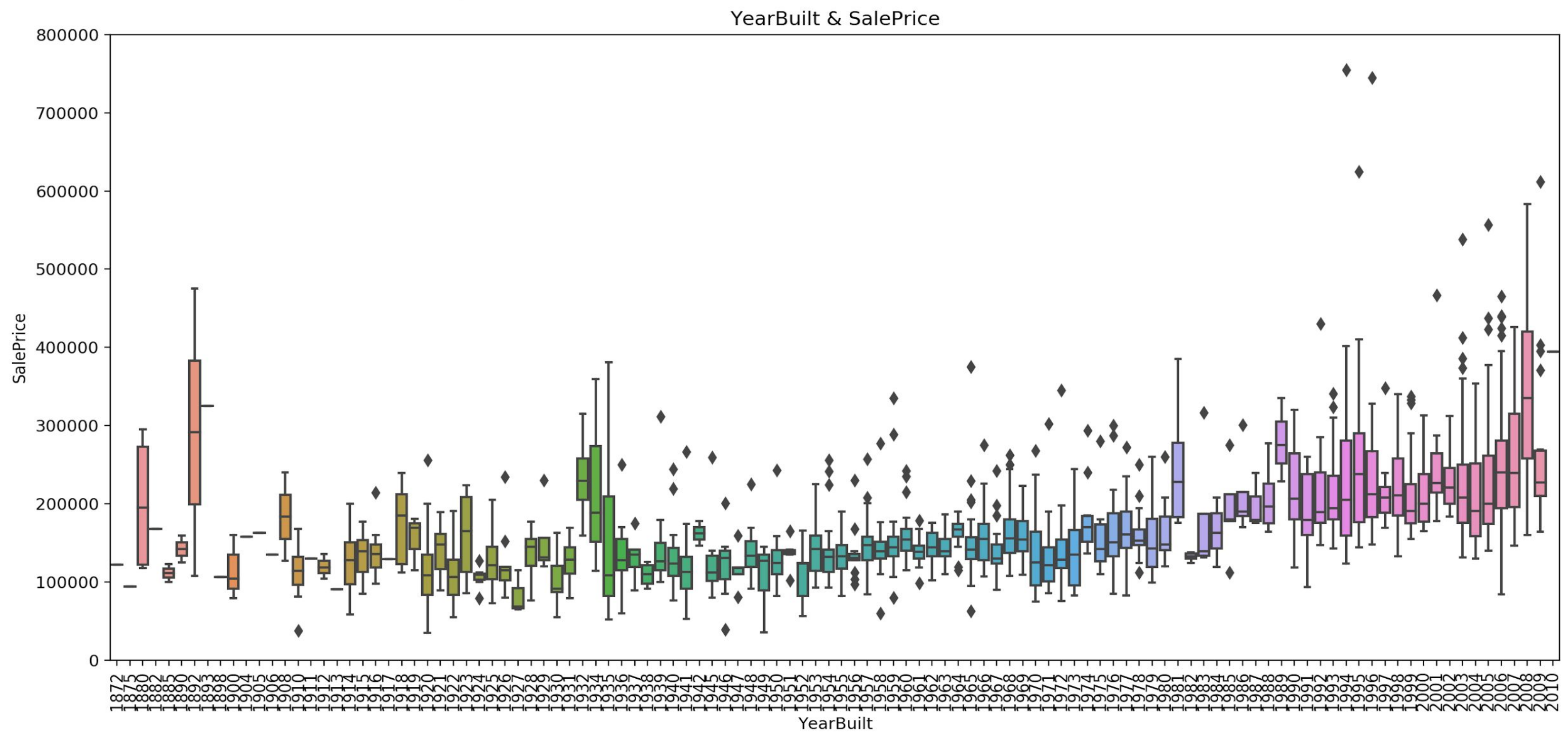
	SalePrice	GrLivArea	GarageArea	TotalBsmtSF	1stFlrSF	BsmtFinSF1	2ndFlrSF
SalePrice	1.00	0.71	0.62	0.61	0.61	0.39	0.32
GrLivArea	0.71	1.00	0.47	0.45	0.57	0.21	0.69
GarageArea	0.62	0.47	1.00	0.49	0.49	0.30	0.14
TotalBsmtSF	0.61	0.45	0.49	1.00	0.82	0.52	-0.17
1stFlrSF	0.61	0.57	0.49	0.82	1.00	0.45	-0.20
BsmtFinSF1	0.39	0.21	0.30	0.52	0.45	1.00	-0.14
2ndFlrSF	0.32	0.69	0.14	-0.17	-0.20	-0.14	1.00

From the correlation matrix, we find that *Living Area Square Feet* has the highest correlation with *Sale Price* (r = 0.71). *Garage area* (r = 0.62), *total basement square feet* (r = 0.61), and *1st floor square feet* (r = 0.61) also have high correlations with *Sale Price*. These findings are what we expected, because they all describe the house area, which indeed are the most important concerns when purchasing a house.

Bivariate Plots For 3 Qualitative Predictors In Top 10







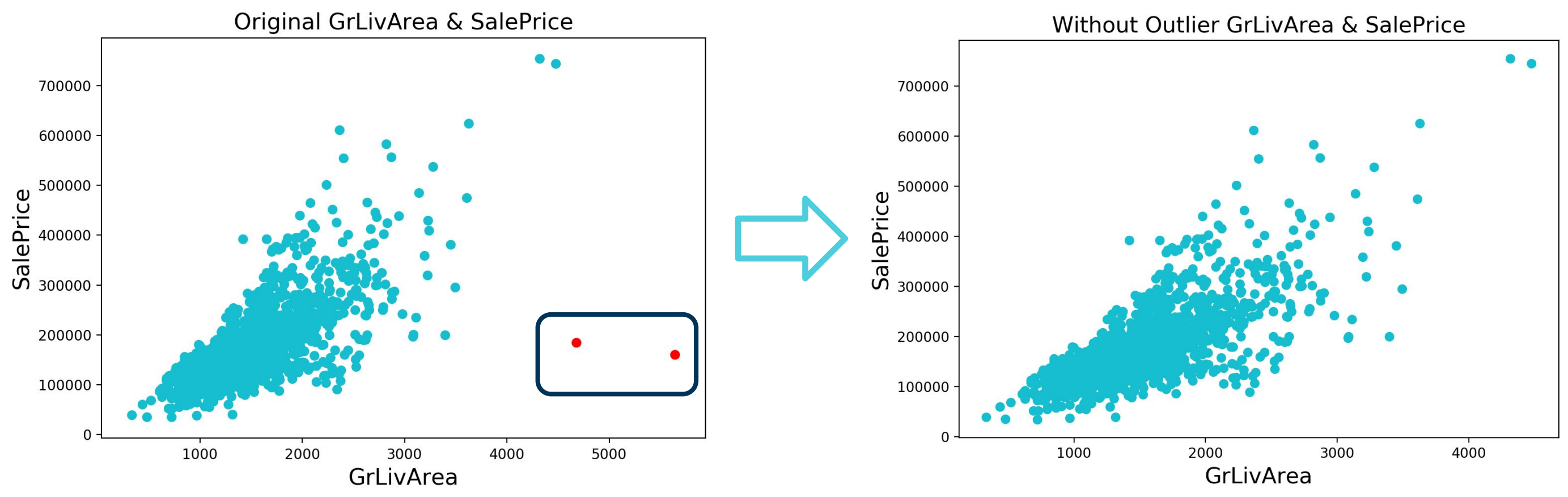
From the boxplot between *overall quality* and *sale price*, we find that with the increasing of house *quality*, the *sale price* increases. From the boxplot between *number of car spaces in garage* and *scale price*, basically the *scale price* increases with *the number of car spaces in garage*. However, when there are four cars spaces, the price goes down. The possible reasons may be that other 78 predictors influence the price, or there are fewer observations in this category. Similarly, the *year built* and *sale price* has a positive relationship with several exceptions.

3

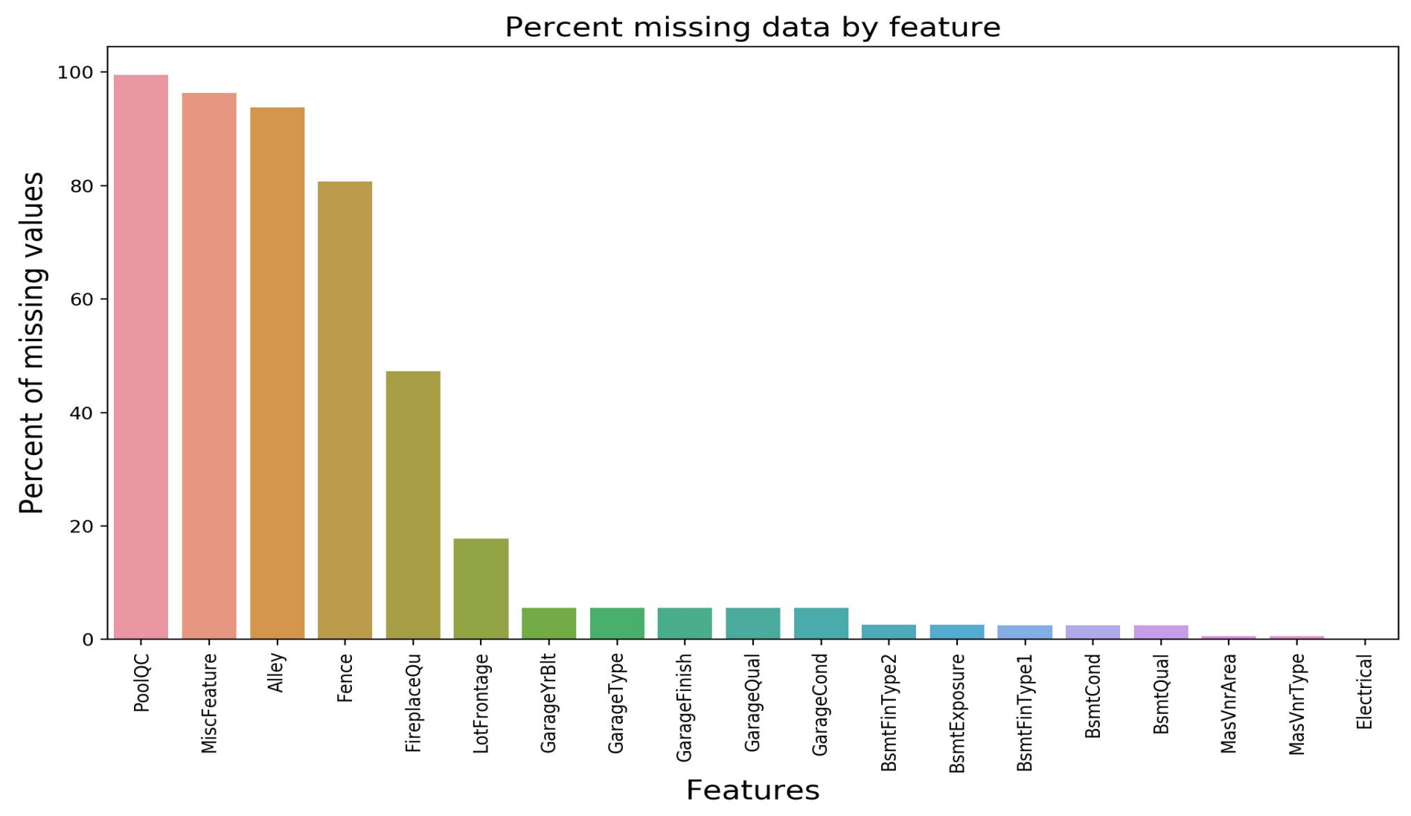
Data Preparation

To prepare our data, we 1) checked outliers in each variable and removed them, 2) checked missing values in each variable and performed imputations (e.g., replace the NA by the value with highest frequency or zero), 3) performed predictor selection, and 4) normalized the variables.

Below is an example of before and after removing outliers in one variable. The two red data points in the left chart (in the dark blue box) are removed.



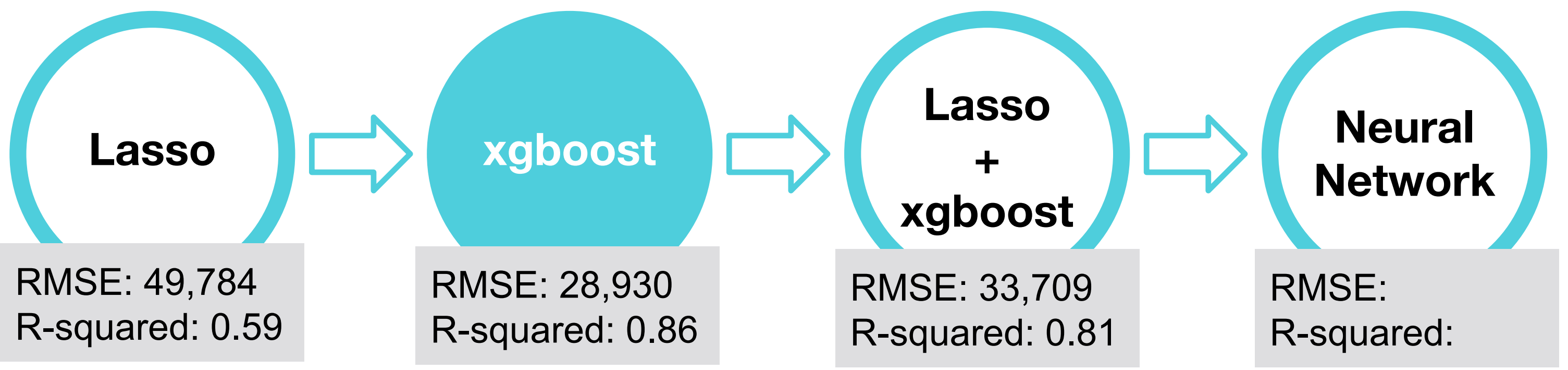
From the percentage of missing values chart, we decided to drop the top four predictors *pool quality*, *miscellaneous feature*, *alley*, and *fence*, and kept in total 75 predictors.



4

Models

We created four models to fit the data: Lasso, xgboost, Lasso and xgboost combination, and neural network. We compared models using root mean square error (**RMSE**) and **R-squared**. We began with Lasso, because it both built the model and did the predictor selection. Its RMSE is **49,784** (the unit is dollar) and R-squared is **0.59**. Next we moved on to xgboost, which used a more regularized model formalization to control overfitting and achieve a better performance. The RMSE significantly reduces to **28,930**, and the R-squared increases to **0.86**. Additionally, we combined Lasso and xgboost to see if we could get better result, however the RMSE increased to **33,709**, and R-squared was **0.81**. The last model is neural network, which is nonlinear black box learning approach. The RMSE (XX) shows that it is the best approach among these four. Hence, XX is the best model.



Lasso
RMSE: 49,784
R-squared: 0.59

xgboost
RMSE: 28,930
R-squared: 0.86

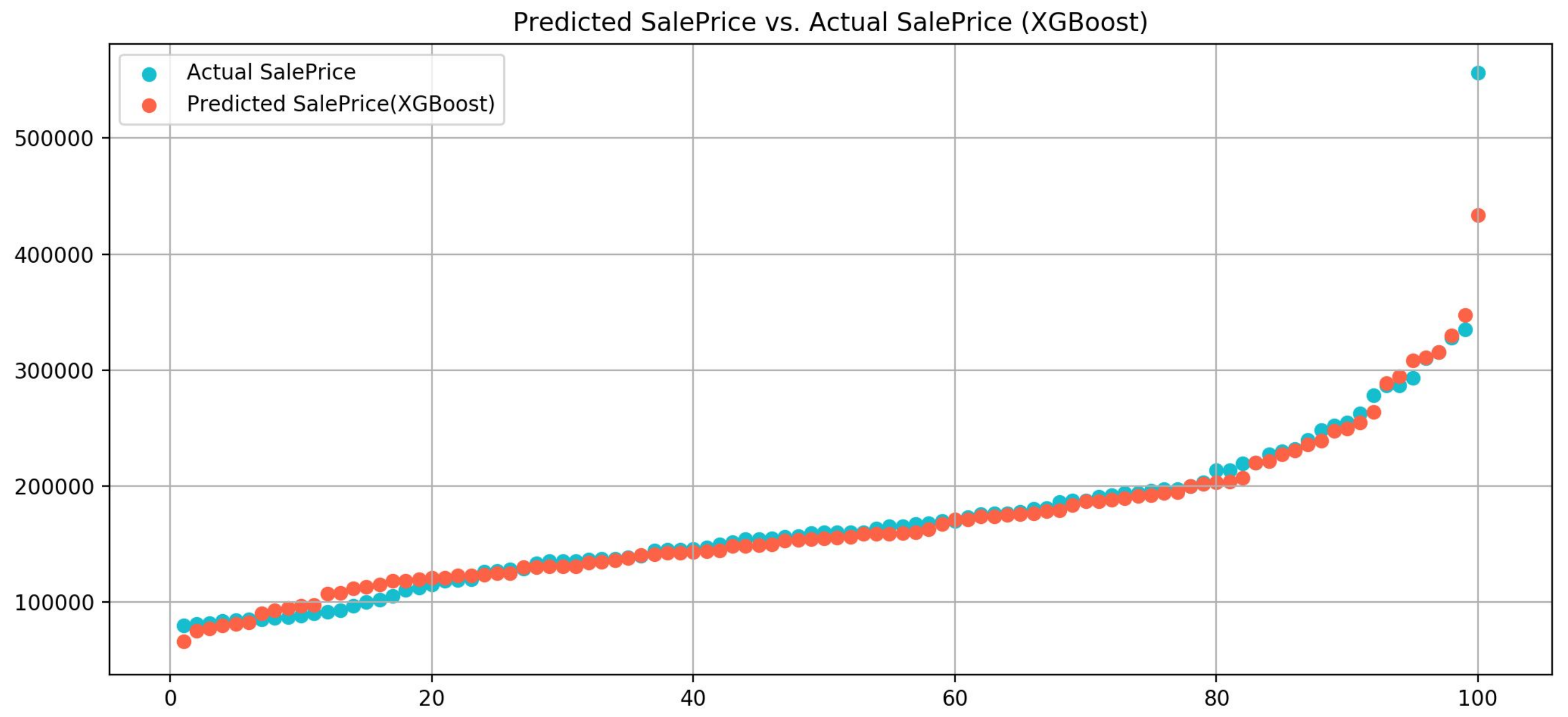
Lasso + xgboost
RMSE: 33,709
R-squared: 0.81

Neural Network
RMSE:
R-squared:

5

Performance

To visualize the performance of our best model xgboost, we tested it on 460 observations and generated below plot using 50 observations. We ordered the *sale price* from the lowest to highest to make the plot easier to understand. The plot shows that our model is accurate, but it does not perform very well in predicting expensive house.


Predicted SalePrice vs. Actual SalePrice (XGBoost)
Actual SalePrice
Predicted SalePrice(XGBoost)