# Linear Regression

# Data Cleaning

The first step in preparing our data for linear regression was to one-hot encode the categorical variables. The resulting dataset has 364 features.

Next, we randomly split the data into training and test sets, assigning 80% of the observations to the training set and 20% of the observations to the test set.

Subsequently, we imputed missing values using the median value of the feature in the training set. Only three features (lot frontage, masonry veneer area, and bike score) were missing values, and no one feature was missing more than 18% of its observations.

Finally, we normalized the features in the training and test sets to have a mean of zero and a standard deviation of one. We normalized the test set based on the normalization parameters of the training set.

# Results

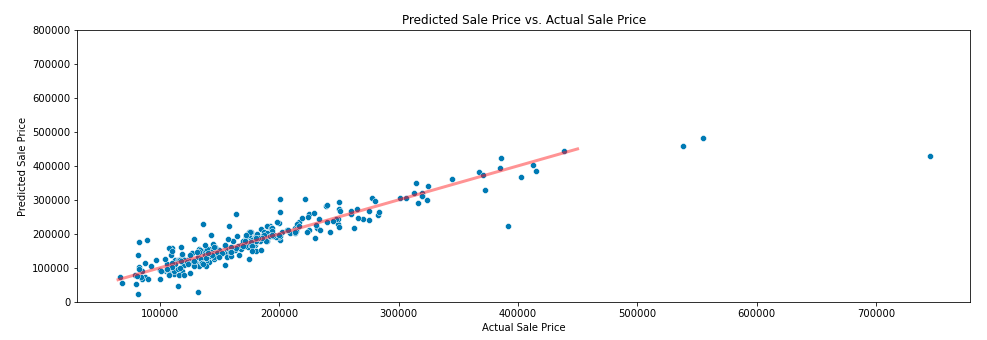
# Basic Linear Model

Our first linear model was a basic OLS linear regression model using all 364 features in our dataset.

The model has a mean absolute error of $24,133 and a mean squared error of 3,886,295,539 squared dollars. The R-squared is 43.7 percent.

In terms of fairness, among low-priced homes, the 25th percentile of the predictive error is negative $12,021, the median predictive error is $3,828, and the 75th percentile of the predictive error is $13,994. Among the high-priced homes, the 25th percentile of the predictive error is negative $6,313, the median is $7,967, and the 75th percentile is $35,116. Thus, we can see that the predictive error is systematically less negative for high-priced homes. This is evidence that low-priced homes tend to be over-valued relative to high-priced homes, which is a concern from a fairness perspective.

The predicted sale prices for the test set are plotted below versus the actual observed sale prices.



The five most important features in the basic linear model are proximity to a positive off-site feature (such as a park), the existence of a pool of excellent quality, the existence of a garage in excellent condition, the existence of a membrane roof, and the existence of a garage of excellent quality. Note that garage quality and garage condition are recorded as two separate features.

# Regularized Linear Model

For our second linear model, we applied Lasso, Ridge, and Elastic Net regression to our dataset. We performed a grid search and used 5-fold cross validation based on the minimum mean absolute error (MAE) to determine the best alpha hyper-parameter ranging from 0.1 to 10,000. Once we narrowed down the approximate location of the best alpha hyper-parameter, we performed a finer-grained grid search.

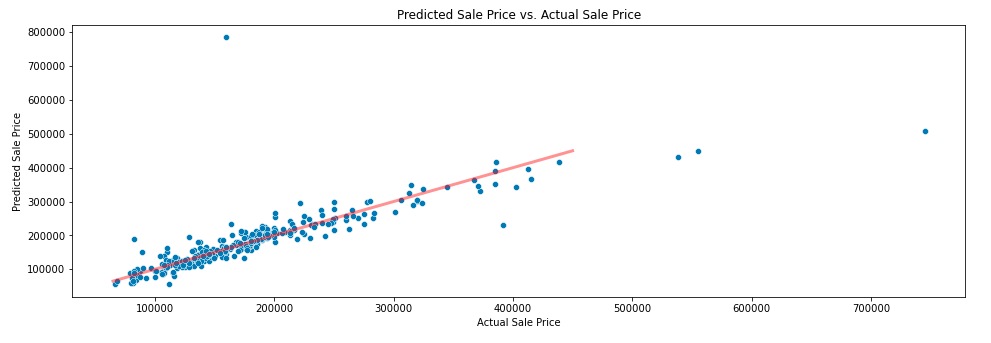
For the Lasso regression model, the optimal alpha hyper-parameter was 150. This yielded a mean absolute error of $20,449, a mean squared error of 2,466,065,314 squared dollars, and an R-squared of 64.3 percent. This model dropped 267 features. In terms of fairness, among low-priced homes, the 25th percentile of the predictive error is negative $6,415, the median predictive error is $3,233, and the 75th percentile of the predictive error is $11,956. Among the high-priced homes, the 25th percentile of the predictive error is $484, the median is $18,757, and the 75th percentile is $45,744. This is again indicative of systematic over-valuation of low-priced homes relative to high-priced homes—which is a concern from a fairness perspective.

For the Ridge regression model, the optimal alpha hyper-parameter was 31. This yielded a mean absolute error of $19,474, a mean squared error of 2,152,970,587 squared dollars, and an R-squared of 68.8 percent. This model dropped five features. In terms of fairness, among low-priced homes, the 25th percentile of the predictive error is negative $7,609, the median predictive error is $2,059, and the 75th percentile of the predictive error is $11,441. Among the high-priced homes, the 25th percentile of the predictive error is $2,302, the median is $23,020, and the 75th percentile is $41,785. This is again indicative of systematic over-valuation of low-priced homes relative to high-priced homes.

Finally, for the Elastic Net regression model, the optimal alpha hyper-parameter was 0.1. This yielded a mean absolute error of $19,790, a mean squared error of 2,162,823,546 squared dollars, and an R-squared of 68.7 percent. This model dropped five features. In terms of fairness, among low-priced homes, the 25th percentile of the predictive error is negative $8,882, the median predictive error is $2,673, and the 75th percentile of the predictive error is $11,820. Among the high-priced homes, the 25th percentile of the predictive error is $5,594, the median is $26,476, and the 75th percentile is $42,931. This is again indicative of systematic over-valuation of low-priced homes relative to high-priced homes.

Of these three models, the Ridge regression model performed the best in terms of mean absolute error—performing even better than the basic linear model. Of these three models, the Lasso regression model performed the best on fairness—although not better than the basic linear model—because it had lower gaps in predictive error between the low- and high-priced homes, while the Elastic Net regression model performed the worst based on fairness. The Ridge regression model was in the middle on fairness. Because the Ridge model had the lowest MAE and had fairness metrics that were similar to the other two models, we will choose the Ridge model as our overall best regularized model.

The Ridge model’s predicted sale prices for the test set are plotted below versus the actual observed sale prices.



The Ridge regression model dropped five features: the existence of severe damages, adjacency to a North-South railroad, proximity to a North-South railroad, the existence of a tennis court, and the existence of a clay or tile roof. None of the homes is the training set has these features, so they are natural candidates to drop.

The five most important predictors for the Ridge model were the existence of a kitchen of excellent quality, the overall quality of the home, the existence of a basement of excellent quality, above-ground square footage, and proximity to a positive off-site feature (such as a park).

# Linear Model with Polynomial Features

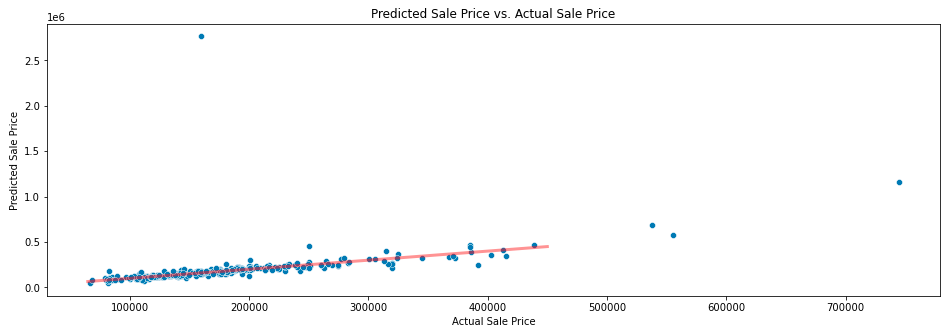
For our third linear model, we applied an OLS linear regression model with polynomial features. We performed a grid search and used 5-fold cross validation to determine whether the optimal number of degrees for the polynomial features was two or three degrees.

When we first attempted to run this model, it ran unsuccessfully due to inadequate memory. As a result, we decided to exclude the features that were dropped by the Lasso regression model (i.e. the features that the Lasso model determined to be the worst predictors of a home’s sale price). We chose to drop features based on the Lasso model, because the Elastic Net and Ridge models only dropped five features each. The Lasso model, on the other hand, dropped a total of 267 features. We opted to drop these same features in order to ensure that our truncated dataset was small enough for Scikit-Learn’s PolynomialFeatures() function to handle.

Based on mean absolute error, the grid search chose the model with three-degree polynomial features. This model has a mean absolute error of $29,169, a mean squared error of 24,774,630,374 squared dollars, and an R-squared of negative 259 percent. These statistics indicate that the fit with polynomial features is much worse than the fit of any of the above models.

In terms of fairness of the model with polynomial features, among low-priced homes, the 25th percentile of the predictive error is negative $11,854, the median predictive error is $710, and the 75th percentile of the predictive error is $8,474. Among the high-priced homes, the 25th percentile of the predictive error is negative $30,939, the median is $2,648, and the 75th percentile is $48,173. For this model, low-priced homes tend to be over-valued relative to high-priced homes, but not systematically so, because the 25th percentile of the predictive error for high-priced homes is actually lower than the 25th percentile for low-priced homes. Therefore, this model appears to have the best fairness metrics of all the linear models we’ve considered so far.

The model’s predicted sale prices for the test set are plotted below versus the actual observed sale prices.



The five most important predictors for the model with polynomial features are above-ground square footage; the interaction of above-ground square footage and the square of the bike score; the interaction of above-ground square footage and the square of the number of fireplaces; the interaction of finished basement square footage and the square of the unemployment rate; and the interaction of lot frontage, masonry veneer area, and second floor square footage.

# Regularized Linear Model with Polynomial Features

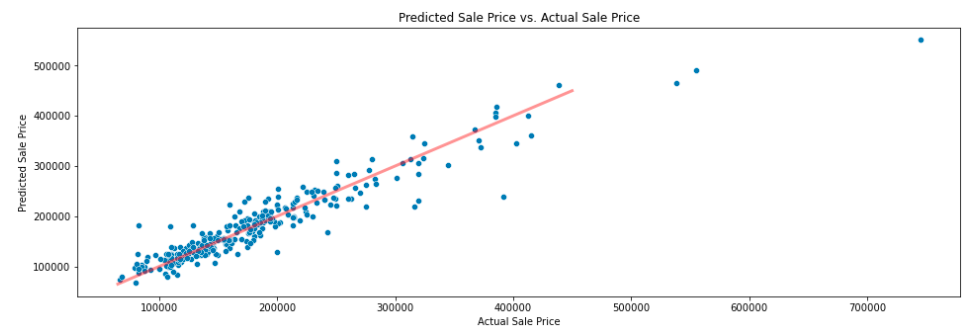
For our fourth and final linear model, we regularized the linear model with two-degree polynomial features. Because the model with three-degree polynomial features performed best above, we first tried to regularize this model. Unfortunately, the model ran for more than three days without stopping, so we opted to regularize the model with two-degree polynomial features instead. We felt that this was a reasonable decision given that the two un-regularized models performed very similarly on MAE, and the two-degree un-regularized model actually had a smaller MSE compared to the three-degree un-regularized model. Consequently, we applied Lasso, Ridge, and Elastic Net regression to the truncated dataset with two-degree polynomial features and used 5-fold cross validation and grid search based on mean absolute error to determine the best alpha hyper-parameter ranging from 0.1 to 10,000. Once we narrowed down the approximate location of the best alpha hyper-parameter for each model, we performed a finer-grained grid search.

For the Lasso regression model, the optimal alpha hyper-parameter was 942. This yielded a mean absolute error of $18,375, a mean squared error of 1,089,150,107 squared dollars, and an R-squared of 84.2 percent. In terms of fairness, among low-priced homes, the 25th percentile of the predictive error is negative $10,960, the median predictive error is negative $1,298, and the 75th percentile of the predictive error is $5,581. Among the high-priced homes, the 25th percentile of the predictive error is negative $18,420, the median is $29,871, and the 75th percentile is $49,304. For this model, low-priced homes tend to be over-valued relative to high-priced homes, but not systematically so (the 25th percentile of the predictive error for high-priced homes is actually lower than the 25th percentile for low-priced homes). The gaps between the fairness metrics for the high-priced and low-priced homes are smaller in this model than in any of the other linear models we’ve considered so far, so this model seems to perform best on fairness.

For the Ridge regression model, the optimal alpha hyper-parameter was 478. This yielded a mean absolute error of $17,659, a mean squared error of 751,730,689 squared dollars, and an R-squared of 89.1 percent. In terms of fairness, among low-priced homes, the 25th percentile of the predictive error is negative $13,593, the median predictive error is negative $2,005, and the 75th percentile of the predictive error is $6,309. Among the high-priced homes, the 25th percentile of the predictive error is negative $5,568, the median is $18,868, and the 75th percentile is $56,979. This is again indicative of systematic over-valuation of low-priced homes relative to high-priced homes.

Finally, for the Elastic Net regression model, the optimal alpha hyper-parameter was 1. This yielded a mean absolute error of $17,788, a mean squared error of 765,864,872 squared dollars, and an R-squared of 88.9 percent. In terms of fairness, among low-priced homes, the 25th percentile of the predictive error is negative $14,047, the median predictive error is negative $1,912, and the 75th percentile of the predictive error is $5,905. Among the high-priced homes, the 25th percentile of the predictive error is negative $6,171, the median is $21,040, and the 75th percentile is $57,056. This is again indicative of systematic over-valuation of low-priced homes relative to high-priced homes.

Of these three models, the Ridge regression model performed the best in terms of mean absolute error—performing even better than all other linear models considered in this paper. The Ridge model’s predicted sale prices for the test set are plotted below versus the actual observed sale prices.



The five most important predictors for the Ridge model were above-ground square footage, overall home quality, basement square footage, the interaction between above-ground square footage and the existence of a garage of average quality, and the number of cars the garage can accommodate.