For this assignment, I was asked to perform data analysis for the CEO of PA-VA Realty. My goal was to identify the significant predictors of housing prices in Pittsburgh and Richmond, as well as develop a model that can accurately predict the price of a house. I was given a training set, *train.csv,* and a test set, *test.csv.*

The first thing I did was classify all factor variables as factors. I treat yearbuilt as numeric, because it is ordinal. I found that for the variable fireplace, almost half of the values in *train.csv* and over half of the values in *test.csv* were NA. I did not feel comfortable making a prediction for half of the observations, and intuitively, it does not seem like fireplace would contain much signal about the price of a home, so I removed it. I performed a permutation test for the importance of zipcode and found that zipcode significantly improved a model containing AvgIncome and state. After closer inspection however, I found that most zipcode values were insignificant in the final model, which was a red flag for collinearity issues. I found that some levels of zipcode had perfect multicollinearity with AvgIncome by aliasing. Intuitively, most of the signal from zipcode is contained by AvgIncome, which is the average income in the given zipcode where the house is located, and state, which accounts for the different real estate markets in Pittsburgh and Richmond. I removed zipcode.

I examined the linear correlations between the numeric predictors, and found that sqft, bathrooms, totalrooms, and bedrooms had a high correlation with price. I also found that totalrooms was very highly correlated with some of the other numeric predictors. The VIF of totalrooms was 3.71, which meant it was a concern for collinearity. Intuitively, much of the signal captured by totalrooms was already captured by sqft, bathrooms, bedrooms, and numstories, so I chose to remove totalrooms as a predictor of interest.

Next, I removed observations where desc was “MOBILE HOME” and exteriorfinish was “concrete.” This was because there are not enough observations at these factor levels to make accurate predictions. This leaves test.csv with 597 observations and train.csv with 1394 observations. I will amend this later. I then split train.csv into a training set called trainSet, and a test set called testSet by a random 70/30 split. These contain price and 12 predictor variables: desc, numstories, yearbuilt, exteriorfinish, basement, bedrooms, bathrooms, sqft, lotarea, state, and AvgIncome.

After cleaning and splitting my data, I started creating models. Below is a table that lists each of the models I created and provides some insight into how they were developed and what their results were. All models were constructed on trainSet using all 12 predictors unless otherwise stated, and were evaluated on test MSE from the predictions of testSet. I will refer to the first model, (simple linear regression with all 12 predictors), as the comparison model, and evaluate the results of all models by referencing the comparison. I use this model as the comparison model because it is simple and interpretable, and any more complicated model must justify the reduction in interpretability with an increase in predictive accuracy.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Name** | **How the Model Was Created** | **Test MSE** | **How Variables Were Deemed “Important”** | **Important Predictors Identified by the Given Model** |
| Simple Linear Regression (Comparison) | Simple linear model with all 12 predictors in trainSet. | 13366467975 | Summary() of the linear model, significant at alpha= 0.05 | descROWHOUSE, exteriorfinishStone, exteriorfinishStucco, rooftypeSHINGLE, rooftypeSLATE, bedrooms, bathrooms, sqft, stateVA, AvgIncome |
| Best Subset Regression | Model with best number of predictors determined by minimizing mallow’s CP. | 13523018752 | Inclusion in the model that has the lowest Mallow’s CP. | Bedrooms, bathrooms, sqft, lotarea, state, AvgIncome, MultiFamily, ShingleRoof, SlateRoof, Frame, Stone, & Stucco finish |
| Ridge Regression | Glmnet package, lambda selected by 10-fold-CV to minimize test MSE. | 13191435309 | Difficult to determine, glmnet does not return coefficient st. error | Difficult to determine, glmnet does not return coefficient st. error |
| LASSO Regression | Glmnet package, lambda selected by 10-fold-CV to minimize test MSE. | 13363189278 | Coefficients that were not shrunk to zero. | All variables except rooftypeROLL |
| Ridge Regression on Best Subset | Same as earlier ridge regression, but only on significant predictors from best subset model. | 13149081707 | Difficult to determine, glmnet does not return coefficient st. error | Difficult to determine, glmnet does not return coefficient st. error |
| LASSO Regression on Best Subset | Same as earlier LASSO regression, but only on significant predictors from best subset model. | 13501812955 | Coefficients that were not shrunk to zero. | All variables identified as significant in BSS are important in this model given this very loose definition of variable importance. |
| Principal Components Regression | ncomp=17 determined by minimizing the adjusted MSEP by 10-fold-CV, looking at variance explained in training, and by inspection of the validation plot. | 13295246547 | Too difficult to determine individual variable importance | Too difficult to determine individual variable importance |
| Partial Least Squares Regression | ncomp=13 determined by minimizing the adjusted MSEP by 10-fold-CV, looking at variance explained in training, and by inspection of the validation plot. | 13298938633 | Too difficult to determine individual variable importance. | Too difficult to determine individual variable importance. |
| Polynomial Regression | Identified bathrooms, sqft, AvgIncome, and bedrooms as variables that could benefit from a polynomial transformation. Used repeated 10-fold-CV to find the degree that best minimized training RMSE. Put 4 variables w their degrees into a model with other variables. | 13030805771 | Summary() of the linear model, significant at alpha= 0.05. Highest significant degree included for all variables that were raised to a degree. | descSINGLE FAMILY, exteriorfinishStone, exteriorfinishStucco, rooftypeSHINGLE, rooftypeSLATE, stateVA, bathrooms^2, sqft^1, AvgIncome^2, bedrooms^3 |
| Cubic Splines | For the 4 variables identified in polynomial regression, made cubic spline with df selected by 10-fold-CV to minimize test MSE | 13884630712 | Significance at alpha=0.05 | descSINGLE FAMILY, exteriorfinishStone, exteriorfinishStucco, rooftypeSHINGLE, rooftypeSLATE, stateVA, bathrooms, bedrooms, lotarea, AvgIncome |
| GAM | For the 4 variables identified in polynomial regression, made a GAM with df=4. | 15987579360 | For numerical variables: is ANOVA for nonparametric effects significant at 0.05? | Bathrooms, sqft, lotarea, AvgIncome |
| Single Decision Tree | Tree was built on training data and pruned to the best size using 10-fold-CV. | 30685005388 | All predictors/factor levels featured on the leaves of the final printed tree. | Sqft, bathrooms |
| Bagging | RandomForest() with mtry=p=12, ntree=1000 | 10506444010 | Relatively high %IncMSE, listed from highest to lowest | Sqft, state, AvgIncome, rooftype |
| Random Forest | RandomForest() with mtry=which mtry minimizes the test MSE, ntree=1000 | 8719232762 | Relatively high %IncMSE, listed from highest to lowest | Sqft, state, bathrooms, AvgIncome, |

There are a few important considerations regarding the table above. Every model except best subsets, LASSO on best subsets, cubic spline, GAM, and single decision tree had a lower test MSE than the comparison model. In terms of minimizing test MSE, ridge regression outperformed LASSO. This makes sense, given that we have 12 predictors on over 1000 observations. Ridge also outperformed PCR and PLS, which is reasonable considering the large number of components chosen by cross-validation. Surprisingly, polynomial regression was the best of the non-tree methods.

Overall, four variables were most consistently identified as significant: sqft, state, AvgIncome, and bathrooms. Not all roof types and finishes were important. The only rooftypes that seem to have predicted price were shingle and slate. The only exterior finishes that seemed to matter were frame, slate, and stucco. From the more interpretable models that give out the estimates of individual coefficients, I found that price increases as sqft, AvgIncome, and bathrooms increases. Price is higher when the state is Virginia, and price is lower when the exterior finish is stone and stucco. Different “desc” levels were considered significant in different models.

A few variables that intuitively might seem significant were not identified as significant. Whether a home has a basement, the year the home was built, and the lot area of a home were consistently not identified as significant predictors. One might expect that all of these factors would affect the price of a home, but it seems that they do not significantly affect home price in Pittsburgh and Richmond.

The three models with the lowest test MSE are random forest, bagging, and polynomial. To determine which of these three models was the most appropriate to predict price on a new dataset, I cross-validated each model by building it 10 times with a new random train test split. In the end, I found that random forests consistently did better than bagging and polynomial models and improved upon the comparison model. My final model is a random forest model with M=5 selected through 10-fold-CV to reduce test MSE. I used this model to make my predictions on test.csv.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **MIN** | **Q1** | **Q2** | **Q3** | **MAX** |
| Known Prices | 35846.97 | 139612.20 | 267904.62 | 426895.91 | 3990701.13 |
| Predicted Prices | 47708.8 | 153046.0 | 269109.6 | 432268.6 | 2252374.0 |

One issue I faced is factor levels like desc=”MOBILE HOME” with only a few observations. I decided to remove these observations from trainSet and testSet. When I made the final prediction on test.csv, I replaced the 3 missing values with the mean of all predictions. This likely added some error to my predictions, and it means that the model is not scalable to future homes with those factor levels.

Above is a table of the five-number summary of the given prices and the prices that my model predicted. Quartiles 1-3 are incredibly close between my predictions and the known prices, which indicates that the dispersion of predicted and known prices are very similar. 75% of homes are worth around $430,000 or less. I believe that my final model test MSE of 8719232762 could be drastically reduced if homes with very high “true” prices were not considered. A high outlier in the known prices using the 1.5\*IQR rule is 426895.91+1.5\*(42689.91-139612.2)=857821. There are 82 such observations in train.csv that violate this rule. A plot of predicted price vs actual price in testSet where true price values are known reveals that the model tends to systematically underestimate the price of homes with a very high true price. It underpredicted the value of a 2-million-dollar home by almost one million dollars, which is certainly a cause for concern. This is confirmed when examining the test MSE of a subsample of my predictions and observations. The test MSE of the model while selecting only observations where the predicted price is less than $500,000 is 3444478723, which is less than half of the test MSE calculated using all observations.

Based on my investigations, I believe that the model I created is relatively accurate in predicting the theoretical “true” price of a home, but only when that “true” price is not unusually high. The model tends to vastly underestimate the price of a home when the home’s “true” price is incredibly high, as evidenced by the smaller maximum in the predictions and the reduction in test MSE from the full model when examining a subset. In a real application of this model, the “true” price will not be known. With that in mind, I believe that this model is still useful so long as the systematic underestimation from homes with a high “true” price is acknowledged and mitigated by guessing homes with high “true” prices using methods outside the model, like examining photos of the home in question.