Prediction Model of Real Estate Prices in Ames, Iowa

DSI-21 Project 2
Table 5

Problem Statement

"PropertyWizard is a real estate listing portal that wants to attract more users in Ames, Iowa by offering a tool to estimate the prices of listings on its website. We have been hired to build a prediction regression model that takes in housing feature data from the Ames City Assessor's Office and estimates the sale price to within an average error of $15,000.

Overview

Targets for Regression Model

- Mean Absolute Error (MAE) of <$15,000
- Low Bias as measured by high R^2 > 0.90 on train set
- Low Variance as measured by high R^2 > 0.90 on test set close to train score

Data Used

Housing sale prices and feature assessments in Ames, Iowa from 2006-2010

- Consisting of 80 explanatory variables, with 2929 observations split into train set (2051) and test set (878)

Issues

Data Cleaning

- 29 out of 80 variables had missing values
- Missing values not consistent across similar features (e.g. some houses had basement quality and condition values but missing basement exposure data
- Data had many categorical features which needed to be interpreted and dummified or converted to numerical data based on their order

Data Exploration

- Some variables had gross outliers that did not conform well to the relationship with sale price
- Many variables (e.g. those measuring area) were right skewed and not normally distributed, and would require transformation in order for the model to work well

Data Cleaning - Handling Null Values

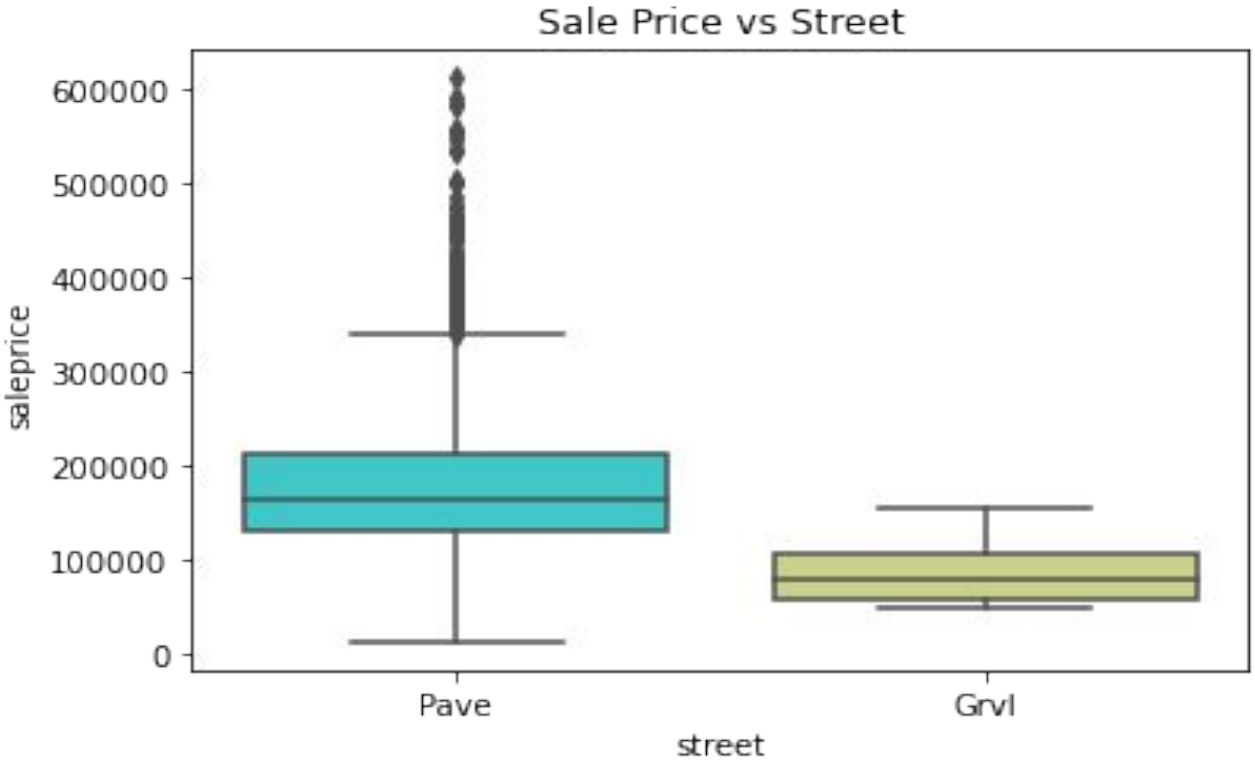
1. Drop variables with >90% null values

2. Investigate correlation with sale price
    a. Many nulls had some correlation meaning they were not randomly distributed

3. Examine null values and theorise reason
    a. If null value only in 1 variable across the features, likely mistake in data entry and impute mean/mode of similar houses
    b. If null value across all features except 1 variable, likely feature does not exist, impute 0/NA

Data Cleaning - Dummifying Categoricals

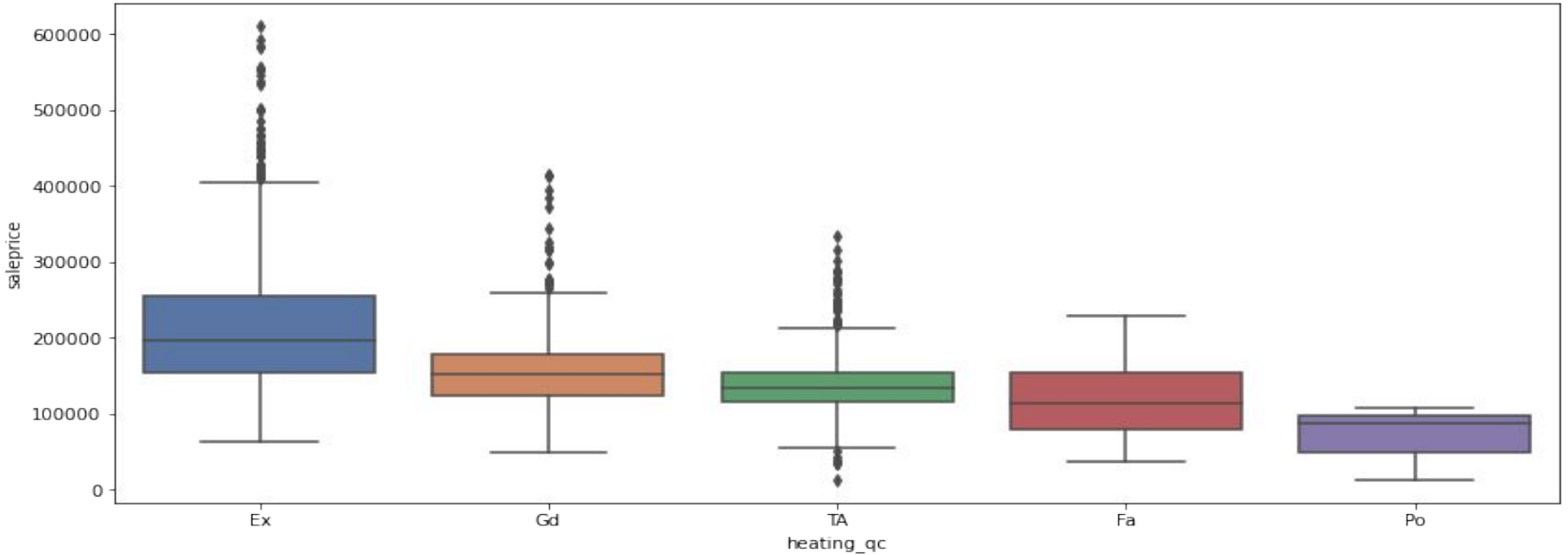
- Many features had to be dummified for the regression model as they were categorical (**nominal**/ordinal) data



Data Cleaning - Dummifying Categoricals

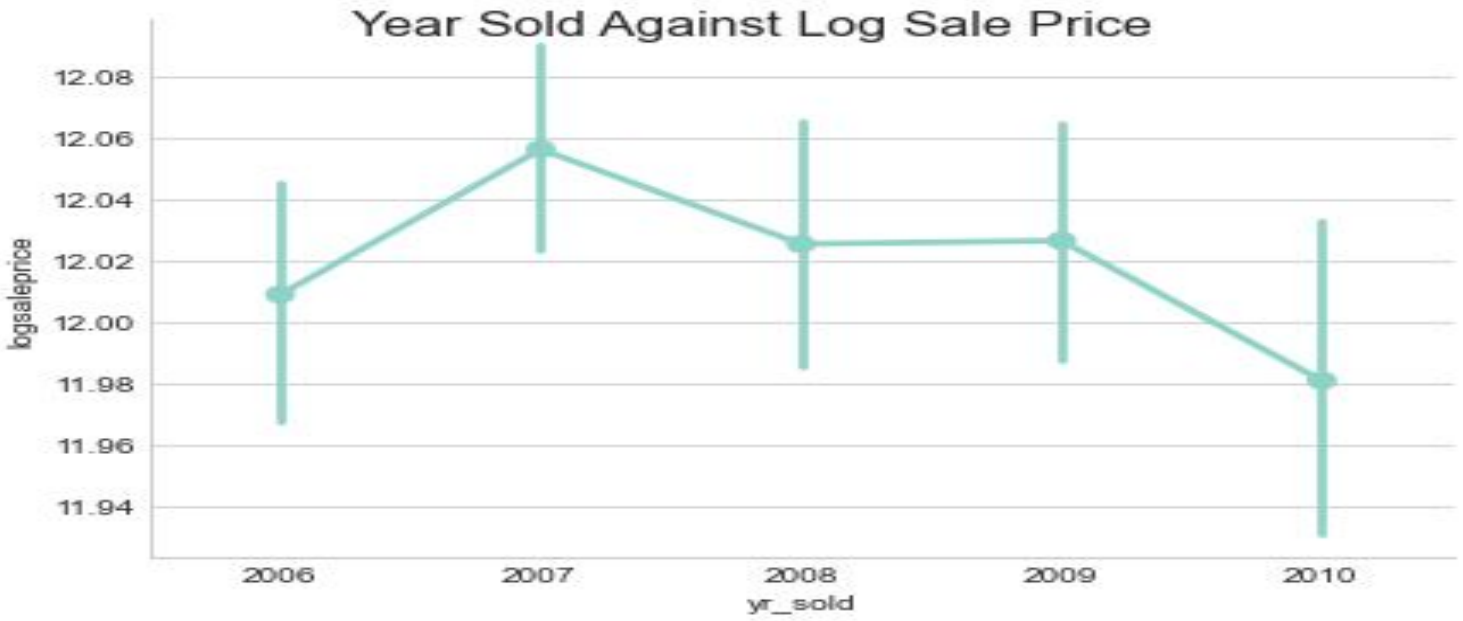
- Many features had to be dummified for the regression model as they were categorical (nominal/**ordinal**) data

Sale Price vs Heating Quality and Condition



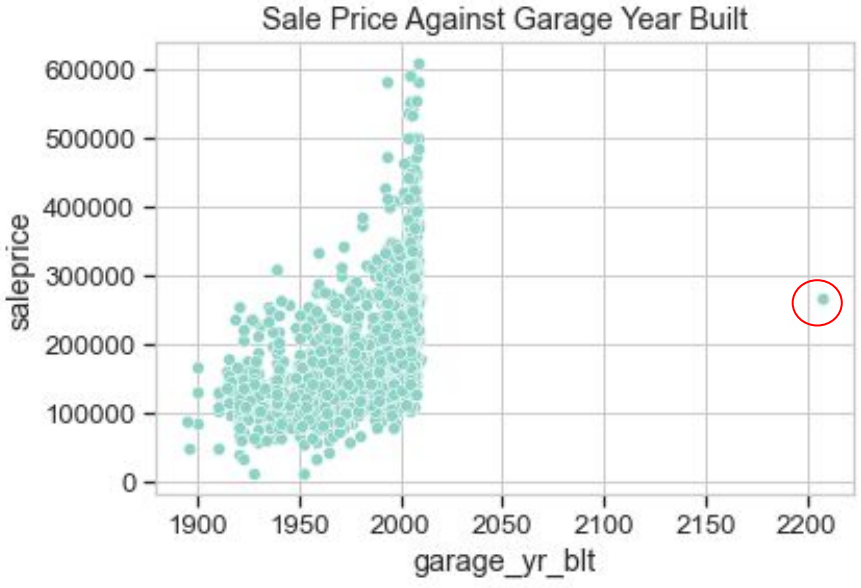
Data Cleaning - Dummifying Categoricals

- Chose to dummify a few **numerical** features (e.g. year sold) as they showed no clear relationship with sale price



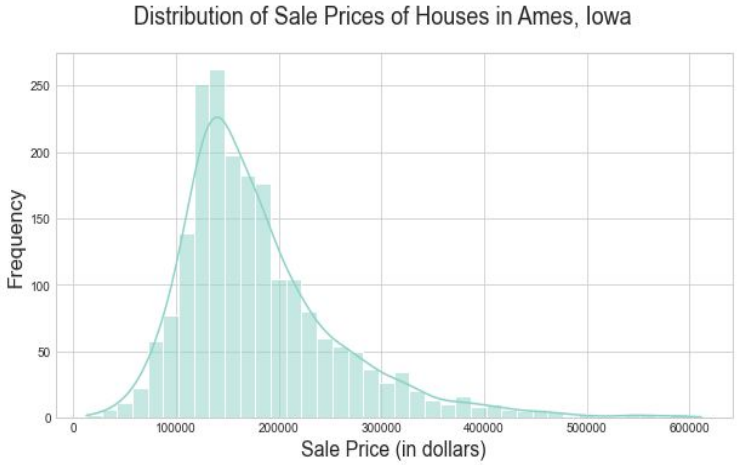
Data Exploration - Handling Outliers

- Many variables exhibited outliers which would bias the relationship and result in an inaccurate coefficient assigned to the variable
- Inspected each variable's relationship and removed clear outliers



Data Exploration - Transforming Variables

- Many variables were right skewed and not normal
- Taking a log transformation helped to improve the normality of variables and make the linear regression model more accurate

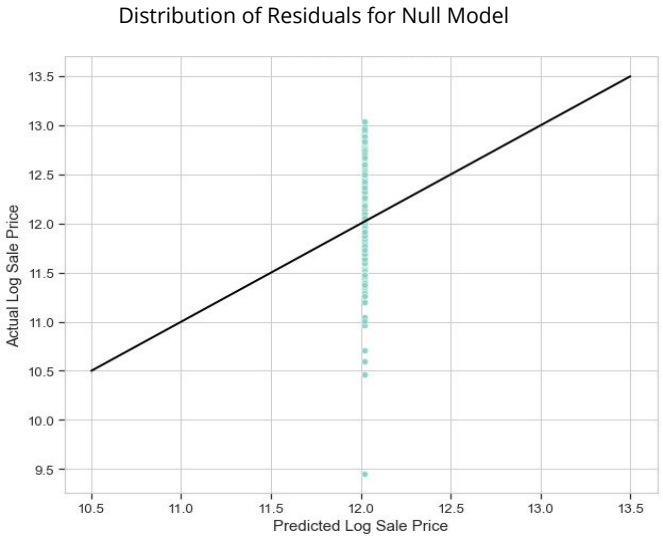


Log

Data Modeling - Choosing Model

- Train set was further split into train and test bloc for evaluation
- Calculated metrics on 5 models using RidgeCV, LassoCV and ElasticNetCV
- Lasso had the best performance, cross checked with cross-validation score

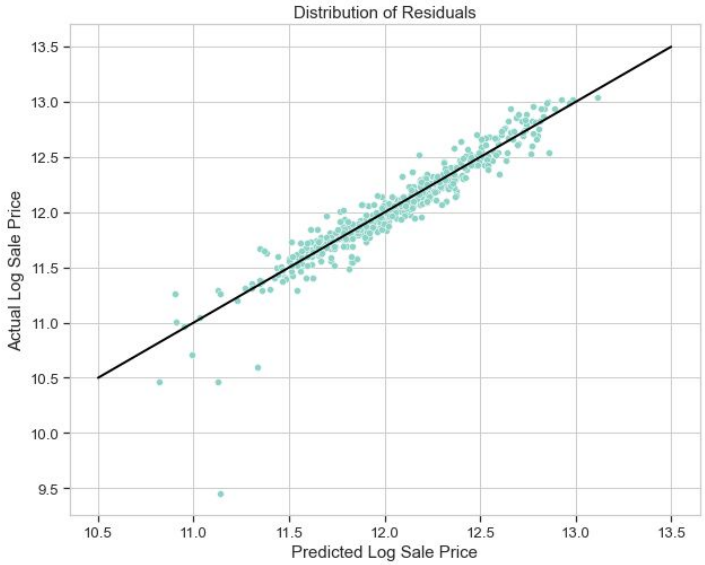
Model	Train R^2	Test R^2	CV R^2	RMSE	MAE	
Null Model	0	0	0	80,011	57,584	
OLS	0.95	0.87	0.73	23,966	15,342	
Ridge	0.92	0.87	0.85	24,961	16,757	
Lasso	0.92	0.89	0.84	22,838	15,259	★
ElasticNet	0.92	0.88	0.86	24,411	16,283	



Data Modeling - Final Model

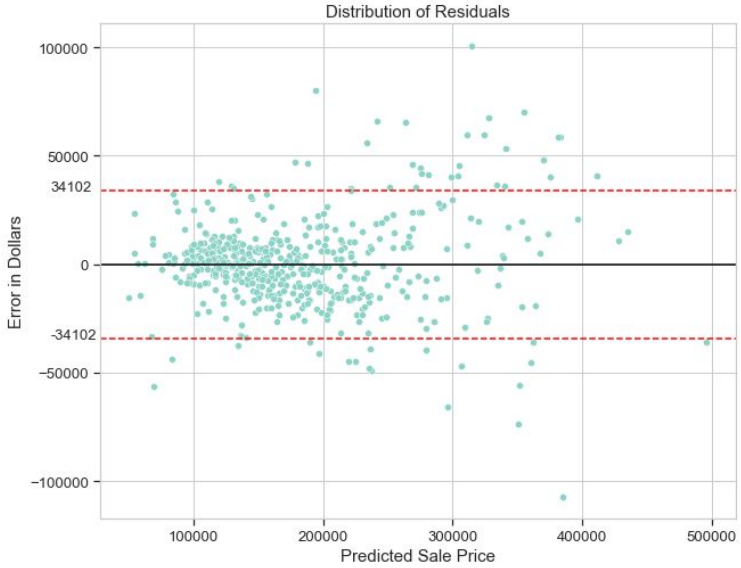
- Feature selection was doing using LassoCV
  - Removed ~200 variables with 0 coefficients
- After techniques applied, mean average error brought down to $14,398!
- Errors of final model seemed to be relatively homoskedastic other than for low sale prices

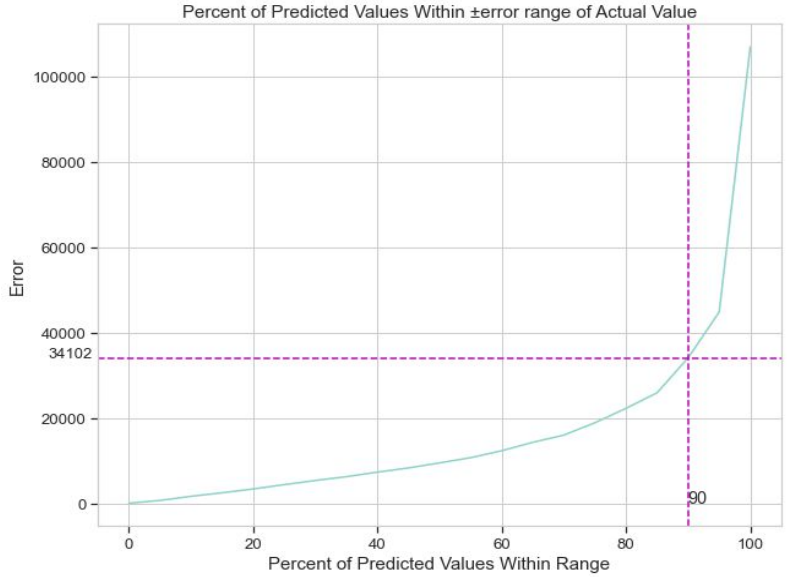
Model	Train R^2	Test R^2	CV R^2	RMSE	MAE
Lasso	0.92	0.89	0.84	22,838	15,259
After feature selection	0.93	0.89	0.86	22,704	15,138
After transforming variables	0.94	0.89	0.89	22,111	14,851
After dropping outliers **Final Model**	0.94	0.90	0.92	20,925	14,398



Data Modeling - Final Model Results

- Model able to predict sale price well with 90% of observations within an error of $34,102
- In percentage terms, the model is able to predict sale price correctly 90% of the time within an error of 12% of the predicted price





Findings and Recommendation

Summary of Model

- Model fulfils target of achieving mean absolute error of below $15,000 on the test set and $R^2 > 0.90$ on both the train set and test set indicating low bias and low variance
- Would be a useful guide for users of PropertyWizard, as the model is able to correctly estimate the potential value of a house in Ames 90% of the time to within an error of ~$35,000

Limitations

- The model performs less well at both ends of the sale price range, likely due to insufficient data
- As the data used is from 2006-2010, it is not clear if the model can be extrapolated to predict housing prices with sale date beyond this range
- While feature selection with LassoCV worked well at keeping down variance, there are still 90+ variables in the model. More work could be done to analyse multicollinearity with different combinations of variables to see if it improves the model's ability to generalise

Recommendation

We recommend for PropertyWizard to adopt the model as a draft, and procure more recent data to train the model on in order to improve its predictions for up-to-date listings, before launching the estimator tool