PREDICTING HOUSING PRICES

STEPHEN GODFREY

PROBLEM STATEMENT

Find a simple and explainable model to predict housing prices in Ames, IA

Data

- Housing prices and characteristics from Ames, Iowa
- Some 80 variables from lot and home size to many characteristics such as overall quality or type of heating
- Split into a training set of 2051 and a testing set of 879 observations
- Evaluate models based on a subset of the training data and then on the testing set
- Testing set evaluation produced a Root Mean Square Error score known here as the Kaggle RMSE

Question – which variables should we include?

Process for selecting model variables

Step 1: Evaluate combinations

- Select correlated numeric variables
- Compare all possible 5 and 10 variable models
- Pick the best model
- Add up to 4 categorical variables
- Pick the best model

Step 2: Eliminate features

- Start with full dataset
- Consider higher order terms
- Use an approach to pick the best variables at points up to 30 (SelectKBest)
- Use another technique to pick the best 5, 7, 9 models (RFE)

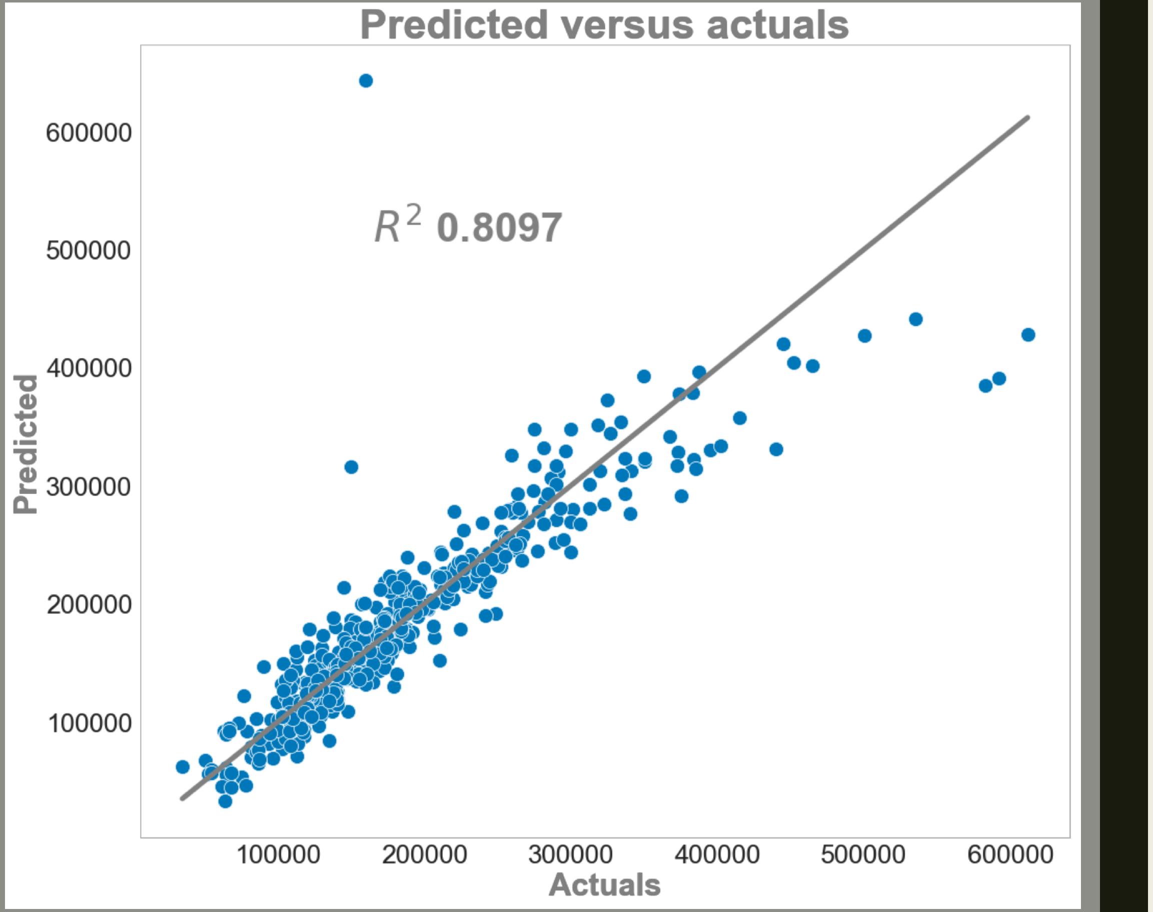
Step 3: Combine 1 and 2

- Combine the outputs of approaches 1 and 2 to select variables

Step 3b: Simplify

- Use judgment to select a model

Evaluation

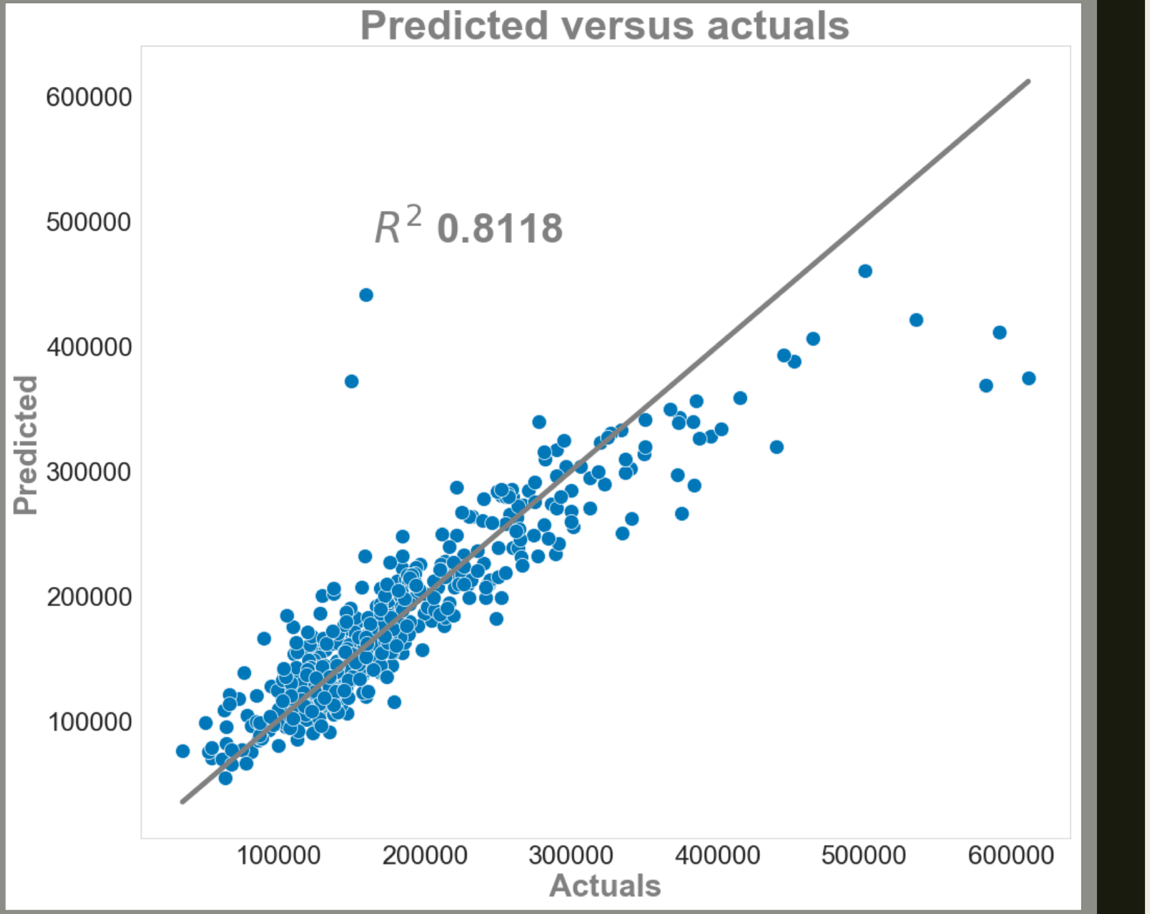


Kaggle RMSE = $33,891

- Results
  - *Selected 10 numerical variables over 5*
  - *Added 3 categorical*
- Observations
  - *Good place to begin further analysis*
  - *Could examine more models – need the computing power*
  - *High degree of multicollinearity*

- *Numeric*
  - Over quality
  - Living Area
  - Garage Area
  - Garage Cars
  - 1st Floor SF
  - Age at sale
  - Remodel at sale
  - Fireplaces
  - Basement Fin SF
  - Open Porch SF
- *Categorical*
  - Neighborhood
  - Building Type
  - Kitchen Quality

Feature elimination

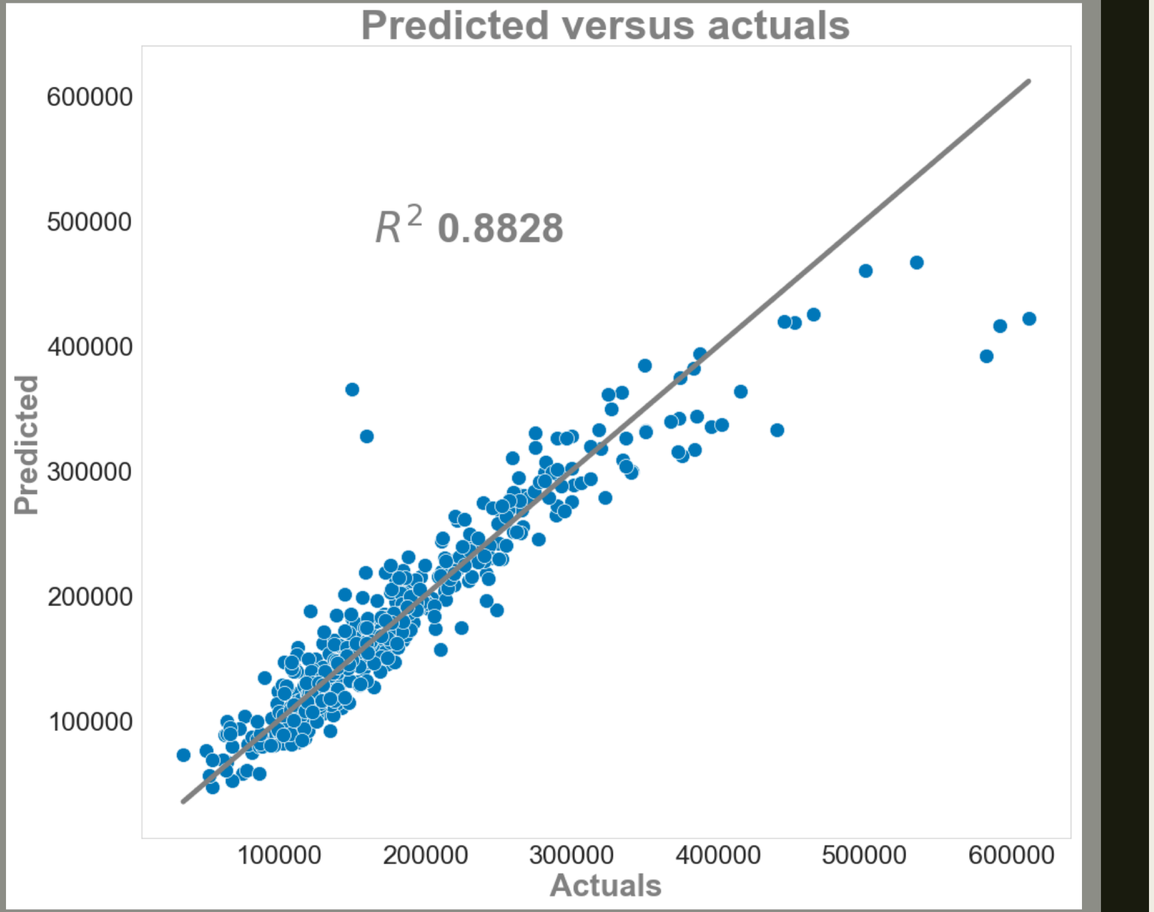


Kaggle RMSE = $36,679

- Results
  - *1 numerical*
  - *Several interaction terms*
- Observations
  - *Good place to begin further analysis*
  - *Interaction terms are hard to understand*
  - *High degree of multicollinearity*

- *Numeric*
  - Over quality
- *Interaction*
  - Over qual x Qual tot bsmt SF
  - Over_qual x Gr_liv_area
  - Over_qual x yr_sold
  - Tot_bsmt_x Sf gr_liv_area
  - Tot_bsmt_sf x Yr_sold
  - Tot_bsmt_sf x Tot_rms_abv_grd
  - Tot_rms_abv_x Grd 1st_flr_sf

Combination

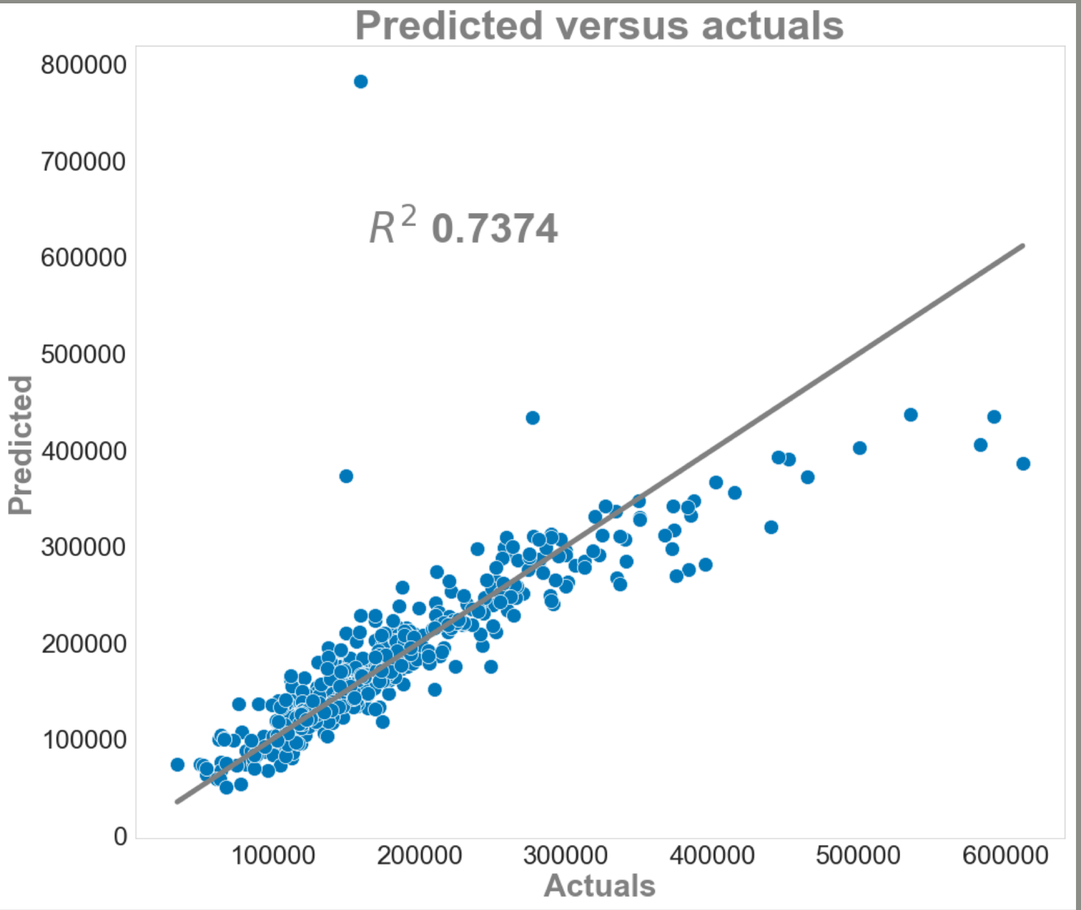


Kaggle RMSE = $30,612

- Results
  - *Best Kaggle score*
  - *Numeric, categorical and interaction variables*
- Observations
  - *Hard to interpret*
  - *Questionable use*
  - *High degree of multicollinearity*

- *Numeric*
  - Over quality
  - Living Area
  - Garage Area
  - Garage Cars
  - 1st Floor SF
  - Age at sale
  - Remodel at sale
  - Fireplaces
  - Basement Fin SF
  - Open Porch SF
- *Categorical*
  - Neighborhood
  - Building Type
  - Kitchen Quality
- *Interaction*
  - Over qual x Qual tot bsmt SF
  - Over_qual x Gr_liv_area
  - Over_qual x yr_sold
  - Tot_bsmt_x Sf gr_liv_area
  - Tot_bsmt_sf x Yr_sold
  - Tot_bsmt_sf x Tot_rms_abv_grd
  - Tot_rms_abv_x Grd 1st_flr_sf

Simplification



Kaggle RMSE = $35,296

- Results
  - *Worst Kaggle score*
  - *8 numeric, 1 categorical variables, 3 interaction*
- Observations
  - *Easy to interpret*
  - *Multicollinearity*

- *Numerical*
  - Over quality
  - Living Area
  - Age at sale
  - Age remodel at sale
  - Lot area
  - Fireplaces
  - Basement Fin SF
  - Open Porch SF

- *Categorical*
  - Neighborhood
- *Interaction*
  - Over_qual x Gr_liv_area
  - Over_qual x yr_sold
  - 'lot_area over_qual

CONCLUSION

For our purposes simpler is better

Balance and judgment are needed in modeling