House Price Prediction

A Capstone Project
By Hope Frost

When a house goes on the market, how much will it really sell for?

Photo by melodi2 from https://freeimages.com FreeImages - Photo by Michael & Christa Richert from https://freeimages.com FreeImages

Who might care?

Realtors

Homeowners

Buyers

What factors might determine final sale price?

Realtors consider:

- Year Built
- Overall Condition
- Bedrooms & Bath

But what about:

- Basement Sq Ft
- Building Materials
- Heating & cooling

The Data Set:

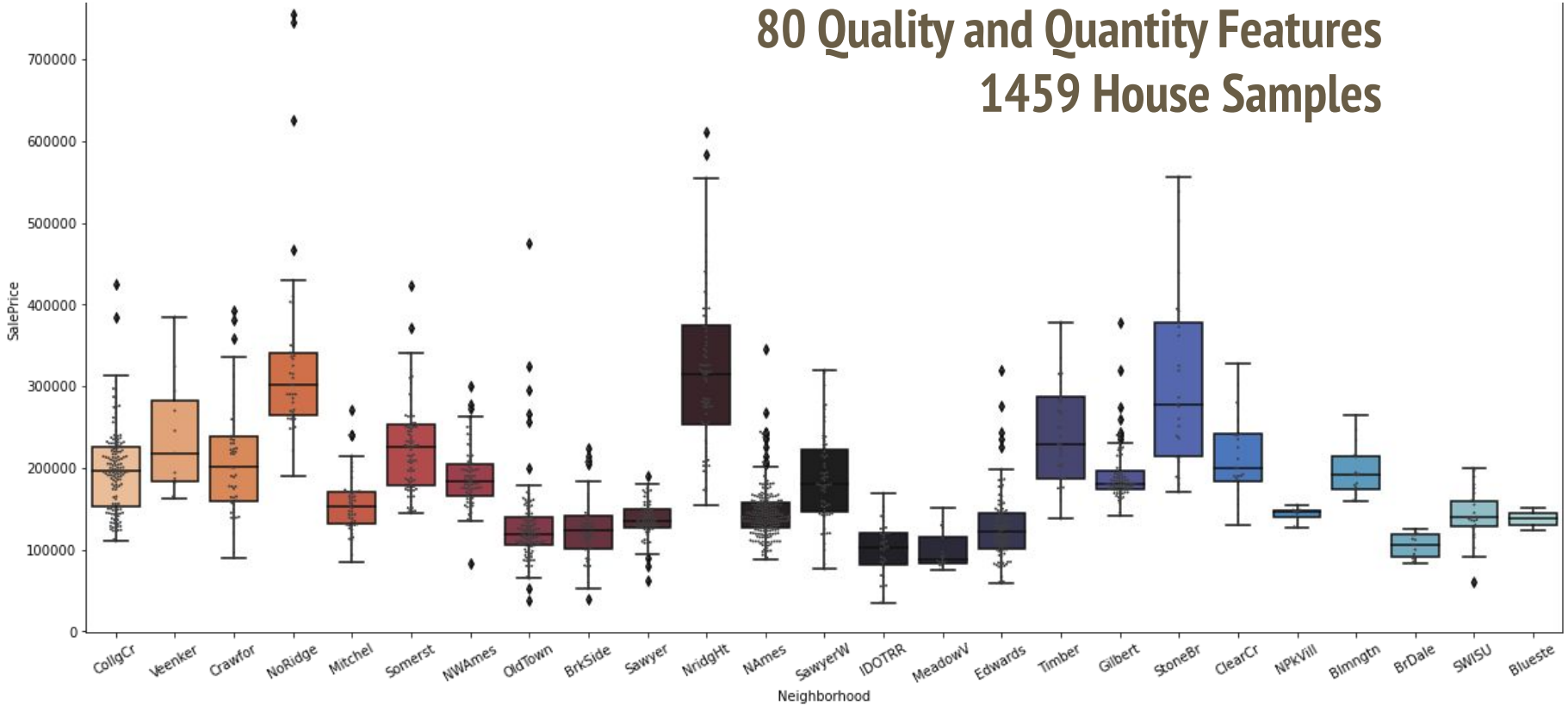
Kaggle: House Price Competition

"The Ames Housing dataset was compiled by Dean De Cock for use in data science education"

https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview

The Competition data contains:

80 Quality and Quantity Features
1459 House Samples



Sales Prices of House in Ames Iowa by Neighborhood

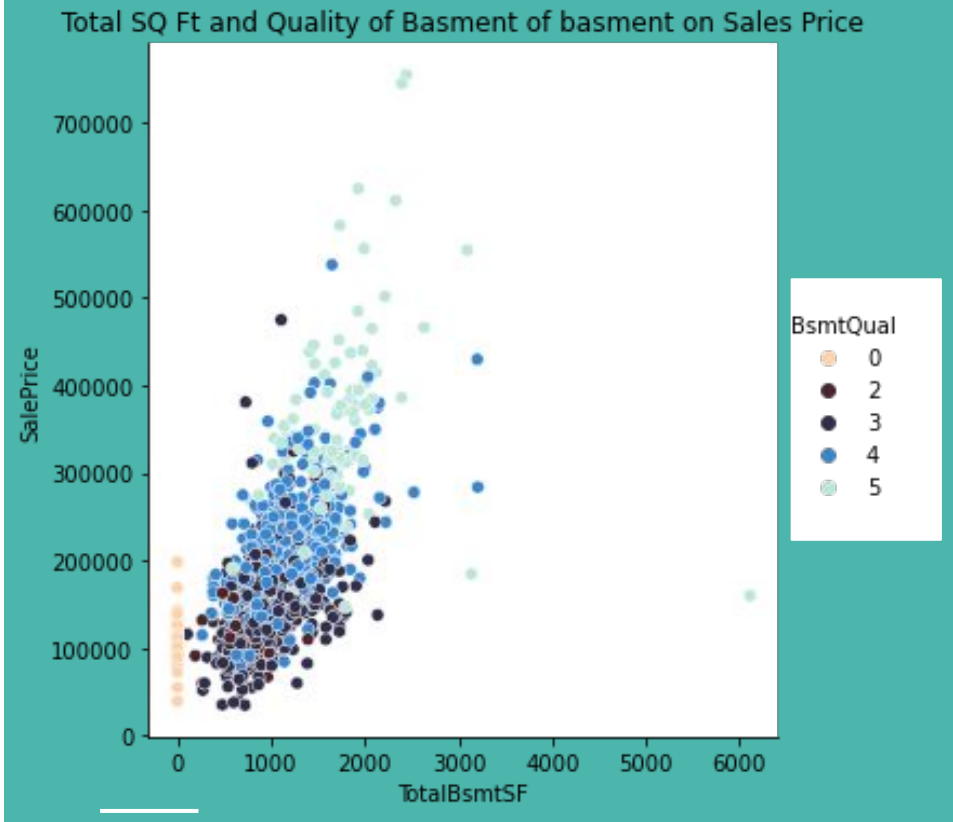
The Data Process

Impute Missing Data

Convert quality categorical feature to numeric features

Build A few new features

Assess and Drop outliers



Model Steps

1. One Hot Encode categorical features
2. Split Train and Test sets
3. Scale and Standardize the numeric features

4. Linear Models:
   Linear Regression
   Ridge
   Lasso
   ElasticNet

5. Non-Linear Models:
   Random Forest
   Decision Tree
   Gradient Boost
   SVR
   XGBoost
   LGBMLight

Model Comparison

	model type	cv_score	MAE	MSE	RMSE	R2 score
model1	LinearRegression	NA	19755.2	6.6996e+08	25883.6	0.88
model2	Ridge	0.84	19877.8	6.80661e+08	26089.5	0.88
model3	Lasso	0.83	19685	6.65856e+08	25804.2	0.88
model4	ElasticNet	0.57	34745.7	2.25335e+09	47469.5	0.59
model5	RandomForestRegressor	0.86	15464.3	5.24048e+08	22892.1	0.9
model6	DecisionTreeRegressor	0.77	20760.6	8.88974e+08	29815.7	0.84
model7	GradientBoostingRegressor	0.87	13804.9	4.36792e+08	20899.6	0.92
model8	SVR	-0	52181.4	5.29723e+09	72782.1	0.03
model9	GradientBoostingRegressor	0.875921	13797.7	4.31391e+08	20770	0.921332
model10	XGBRegressor	0.86	14481.4	4.57588e+08	21391.3	0.92
model11	LGBMRegressor	0.87	15216.2	5.09175e+08	22564.9	0.91
model12	CatBoostRegressor	0.83	17211.2	5.72056e+08	23917.7	0.9

Evaluation Metrics

model9 using a GradientBoostingRegressor is the best **RMSE** score

The final competition score is evaluated on Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price.

model9	GradientBoostingRegressor	0.875921	13797.7	4.31391e+08	20770	0.921332

Predictions and Residuals

Test Residuals:

STD $15546 .15

Mean $13797. 69

50% $8791.58

Train Residuals:

STD $12070.97

Mean $6824.612

50% $4281.90

Competition score:

0.14051

Rank at time of submission: 6352

Next Steps

1. Engineer More Features
2. Correct Skew of Particular Features
3. Explore other models

Thank You

To Springboard

And particularly for all the help from:

Silvia Seceleanu

DJ Sarkar