Predicting Ames, IA Housing Prices

Group: Glengarry Glen Ross

- Liam McDermott
- Jan Ruffner
- David Wasserman
- Ethan Zien

Agenda

01 Setting the Stage

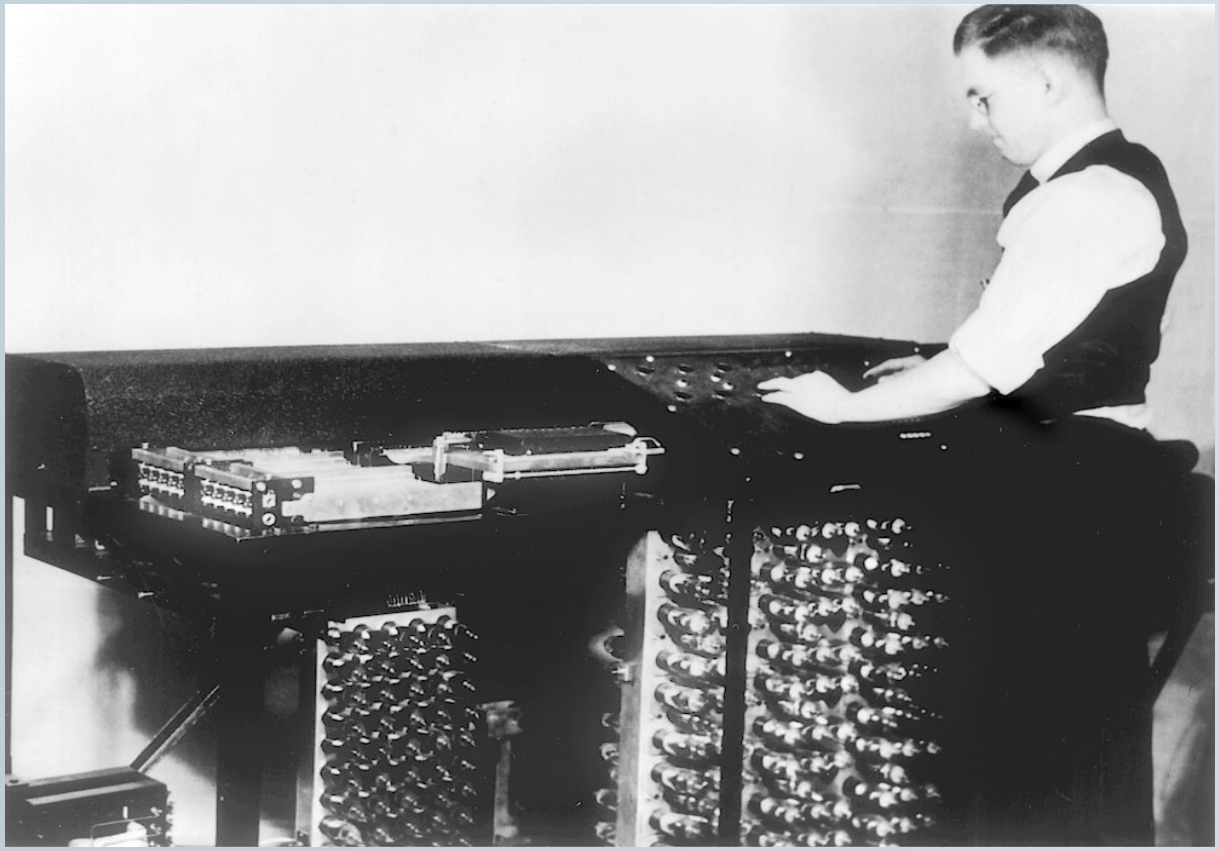
02 Feature Analysis and Selection

03 Model Methodology

04 Results & Conclusion

Objective: Deliver a model to describe housing market in Ames, IA

Only US City with a population greater than 65,000 that has 45% of residents enrolled in college or graduate school



The Atanasoff–Berry Computer (ABC)
First Automatic Electronic Computer
Successfully tested 1942 - Iowa State College

Green Hill : $198.65 Per Square Foot

South West of ISU : $89.89 per square foot

Model Preparation (Data Cleaning)

A. Removed duplicate records from the training set

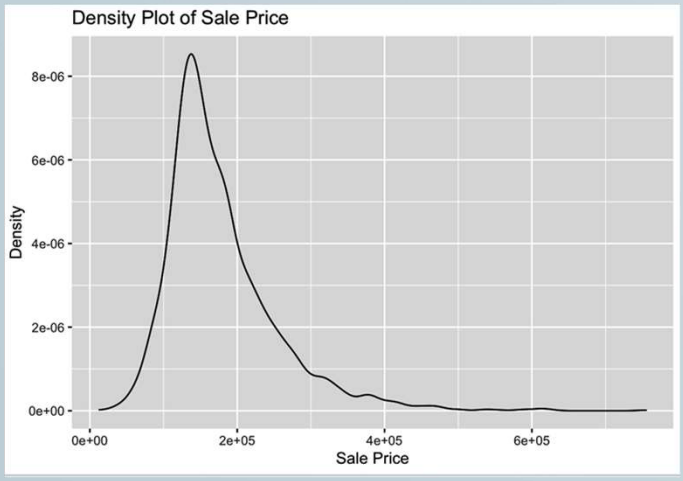
B. Imputed **zero (0)** for **numerical** variables where NA (except Year)

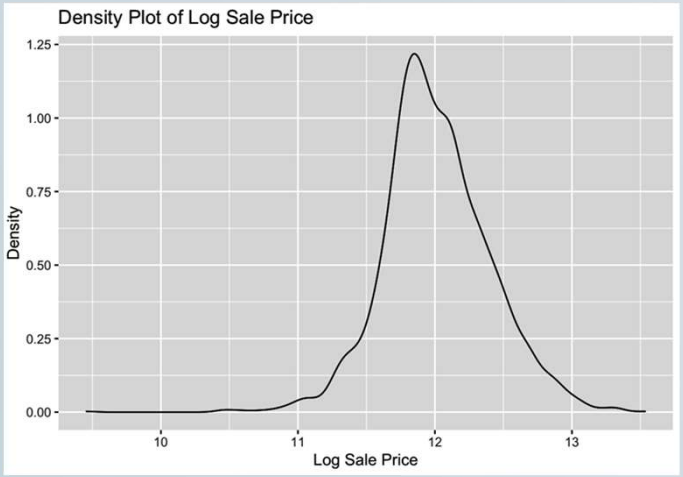
C. Imputed **None** for **categorical** variables where NA

Feature Analysis & Selection

Dependent Variable

- Chose to analyze log sale price to due to having a more normal distribution



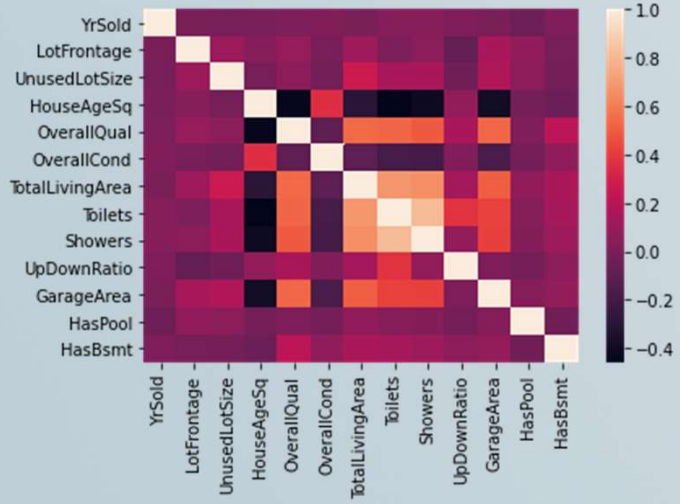


Feature Creation

Feature	Code Name	Calculation
Total Living Area	TotalLivingArea	GrLivArea + TotalBsmttSF - BsmtUnSF
Unused Lot Size	UnusedLotSize	LotArea - 1stFlrSF
Has Pool	HasPool	PoolArea > 0
Has Basement	HasBsmt	BsmtQual != None
Toilets	Toilets	HalfBath + FullBath + BsmtHalfBath + BsmtFullBath
Showers	Showers	FullBath + BsmtFullBath
House Age	HouseAge	max(YearBuilt) - YearBuilt + 1
House Age Squared	HouseAgeSq	HouseAge ^ 2
Up/Down Ratio	UpDownRatio	2ndFlrSF / 1stFlrSF

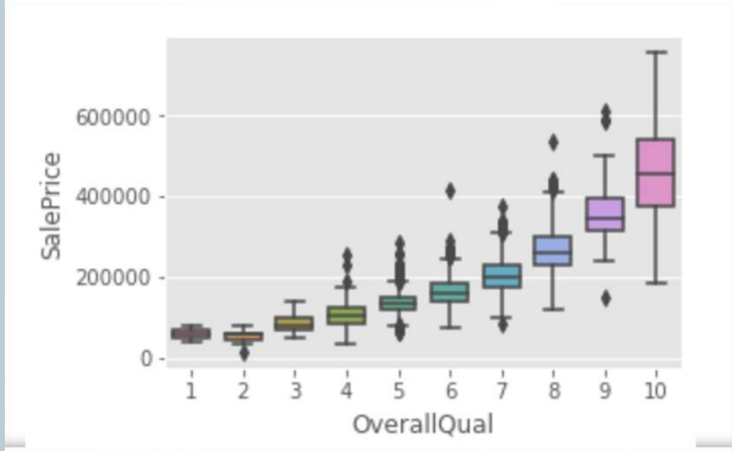
Feature Selection

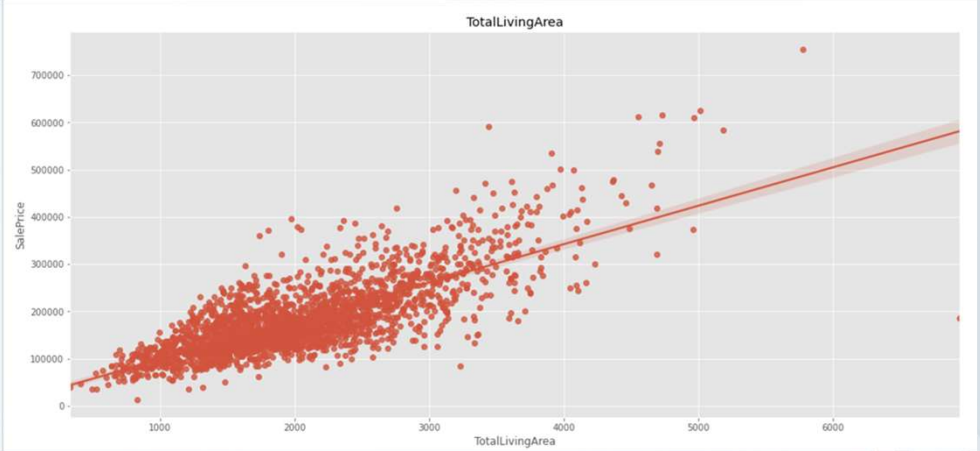
- Looked at correlation across metrics



Feature Selection

Looked at different variables and how they correlated with sale price





Model Variable Selection

- Basement Quality
- Building Type
- Garage Area
- Has Basement
- Has Pool
- House Age (Squared)
- Lot Frontage
- Neighborhood
- Overall Condition
- Overall Quality
- Sale Condition
- Showers
- Toilets
- Total Living Area
- Unused Lot Size
- Up/Down Ratio
- Year Sold

These variables showed a strong correlation to the sale price on a house and limited multi-collinearity.

Model Methodology

Regression Models Tested

1) Multiple Linear Regression with Lasso Regression
2) Random Forest
3) XGBoost
4) LightGBM
5) CatBoost
6) Support Vector Machine

Multiple Linear Regression

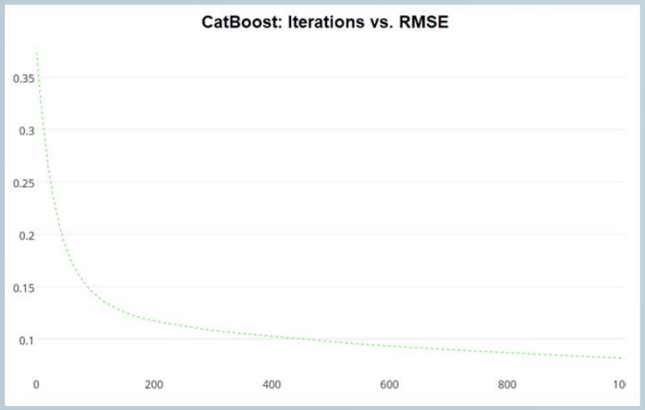
	coef
const	10.4162
BQ_Ex	0.1175
BQ_Gd	0.0893
BT_1Fam	0.0970
GarageArea	0.0003
LotFrontage	0.0002
Nbhd_Blmngtn	0.1708
Nbhd_ClearCr	0.0941
Nbhd_Crawfor	0.1362
Nbhd_Gilbert	0.0811
Nbhd_GrnHill	0.6054
Nbhd_NoRidge	0.1008
Nbhd_NridgHt	0.1361
Nbhd_Somerst	0.1168
Nbhd_StoneBr	0.1774
OverallCond	0.0378
OverallQual	0.1064
SC_Partial	0.0375
Toilets	0.0308
TotalLivingArea	0.0002
UnusedLotSize	2.736e-06

The Power of Boosting Models

CatBoost

XGBoost

LightGBM



Cross-Model Comparison

Model	Training R^2	Testing R^2	Best RMSE
CatBoost	94.65%	91.62%	0.113
LightGBM	97.40%	91.00%	0.117
XGBoost	99.35%	90.89%	0.118
Random Forest	97.86%	89.66%	0.126
Support Vector Machine	90.38%	88.30%	0.134
Linear Regression	87.00%	88.24%	0.138

Modeling Lessons Learned

1) Boosting models perform well but take long to train (especially with CV)

2) Linear regression performs poorly when you have many features compared to other models

3) While the models are not in complete agreement, some variables always show high importance

Results & Conclusion

Conclusion

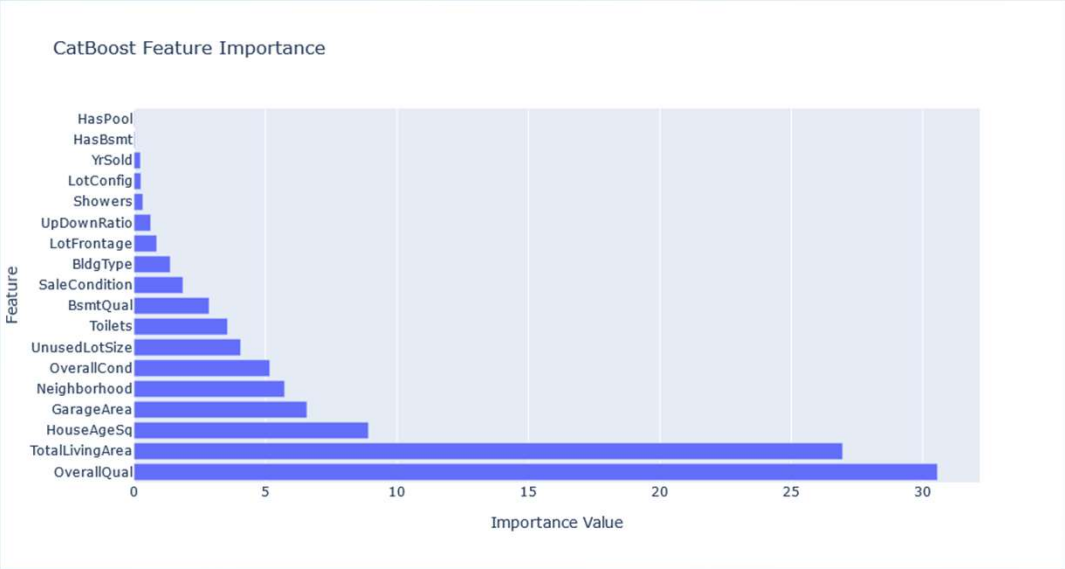
Most successful model:

- CatBoost Model

Final Performance:

- R^2: 91.6%,
- Root mean squared error: 0.1135

Feature Importances from CatBoost



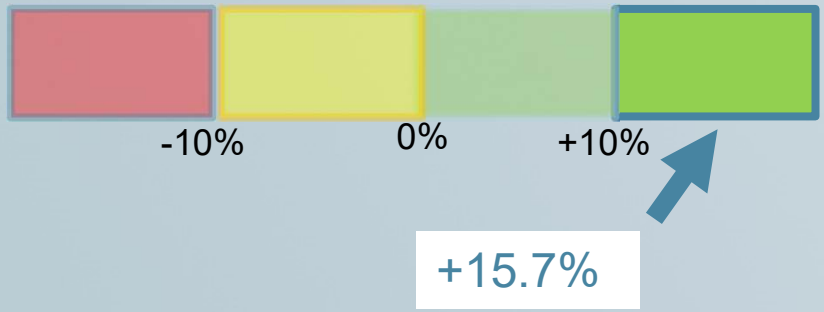
Feature	Importance
Overall Quality (OverallQual)	31%
Total Living Area (TotalLivingArea)	27%
House Age Squared (HouseAgeSq)	9%
Garage Area (GarageArea)	7%
Neighborhood (Neighborhood)	6%
Overall Condition (OverallCond)	5%
Unused Lot Size (UnusedLotSize)	4%
# of Toilets (Toilets)	4%
Basement Quality (BsmtQual)	3%
Sale Condition (SaleCondition)	2%
Building Type (BldgType)	1%
Lot Frontage (LotFrontage)	1%
Upstairs/Downstairs Ratio (UpDownRatio)	1%

Next steps

Compare offering prices to what the model predicts to find bargains

Sale Price: $280,000
4 bedrooms
2 full baths + 2 half baths
2,241 sqft

Our estimate: $324,000



Thank you