Predicting House Prices in Ames, Iowa

Friday, April 14th 2023
Martijn de Vries
General Assembly
DSIR-0320 Project 2

CITY OF
AMES
IOWA
Caring People
Quality Programs
Exceptional Service

Problem Statement

A real estate company in Ames, Iowa is looking for a better way to evaluate the market value of a house. Using the **Ames dataset**, I will build a **predictive model** to predict the sale price of a house.

Benchmark model:

An model using only the **total living area** to predict sale price.

How much can we improve on this?

Data

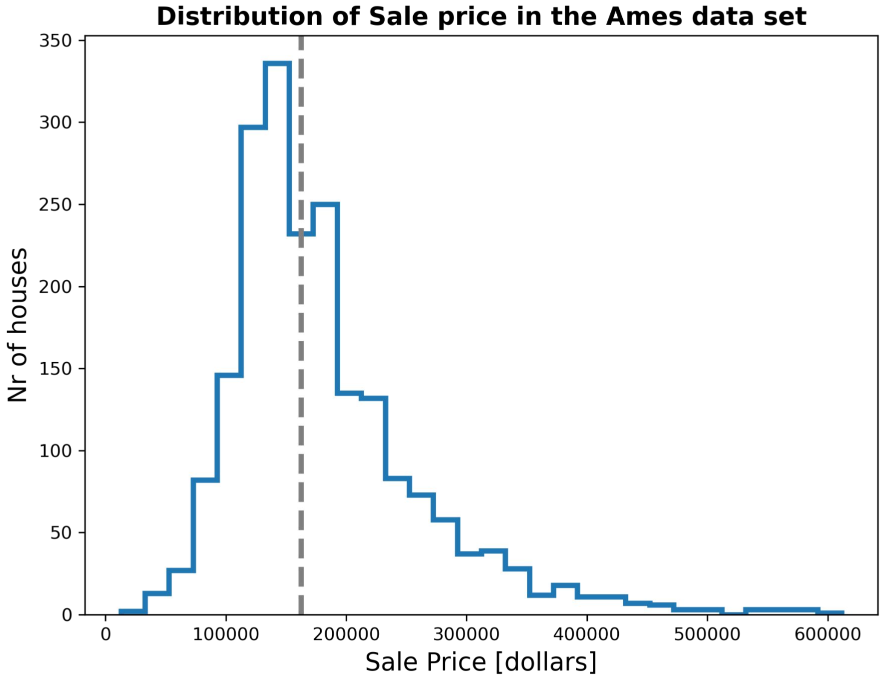
The ‘famous’ Ames data set[1]

Training data: 2051 houses

Houses sold in Ames, IA from 2006-2010

78 columns (!) + Sale price

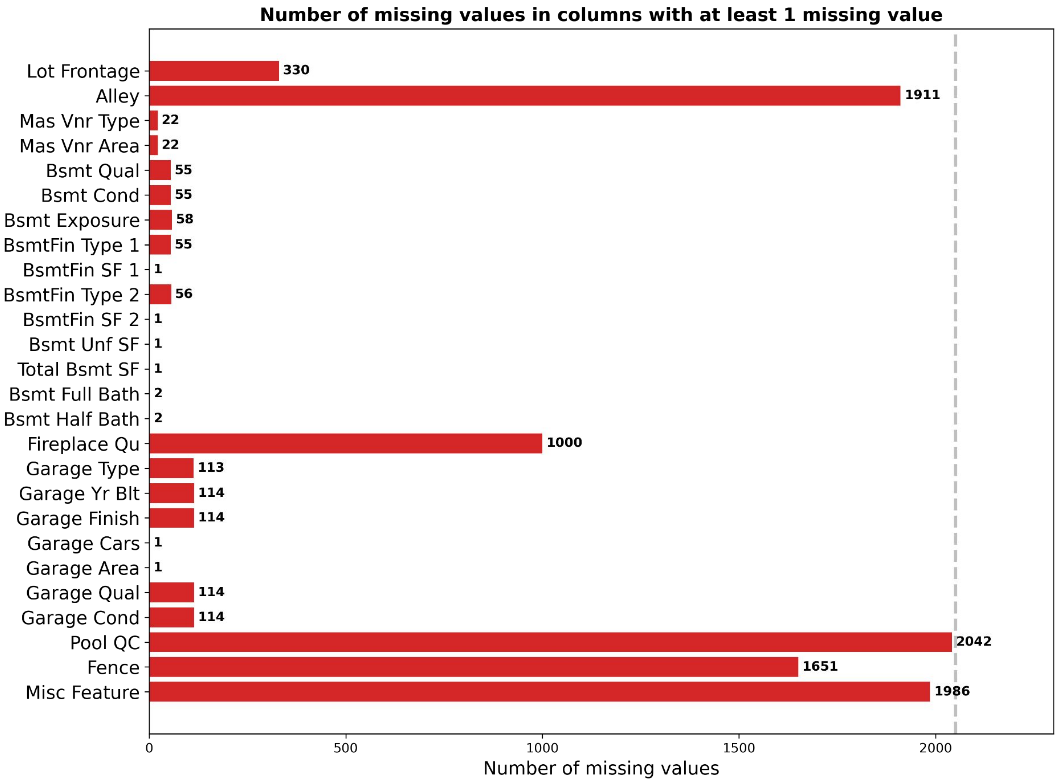
- Numerical (eg: Square footage)
- Ordinal (eg: Quality of the kitchen)
- Nominal (eg: What kind of garage)



[1] https://jse.amstat.org/v19n3/decock.pdf

Data Cleaning

1. ‘NA’ values interpreted as missing -> ‘NP’
2. Related columns set to zero (eg. Garage Area)
3. Lot Frontage: imputation based on sqrt(Lot Area)
4. Everything else: mode or median imputation



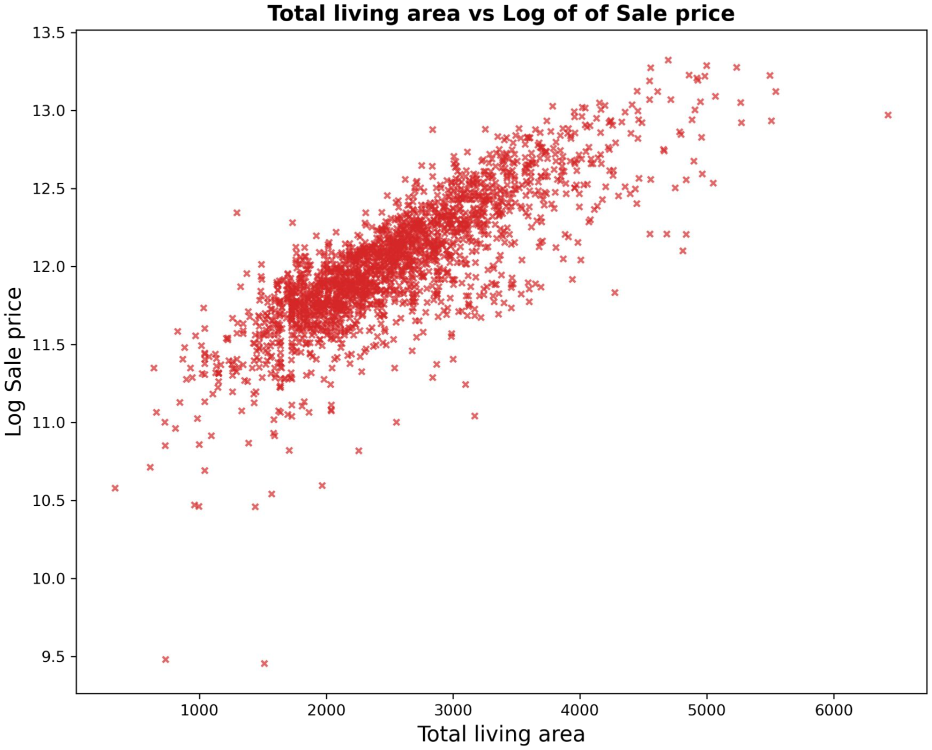
Building a model

Start simple, then increase complexity (= add more columns)

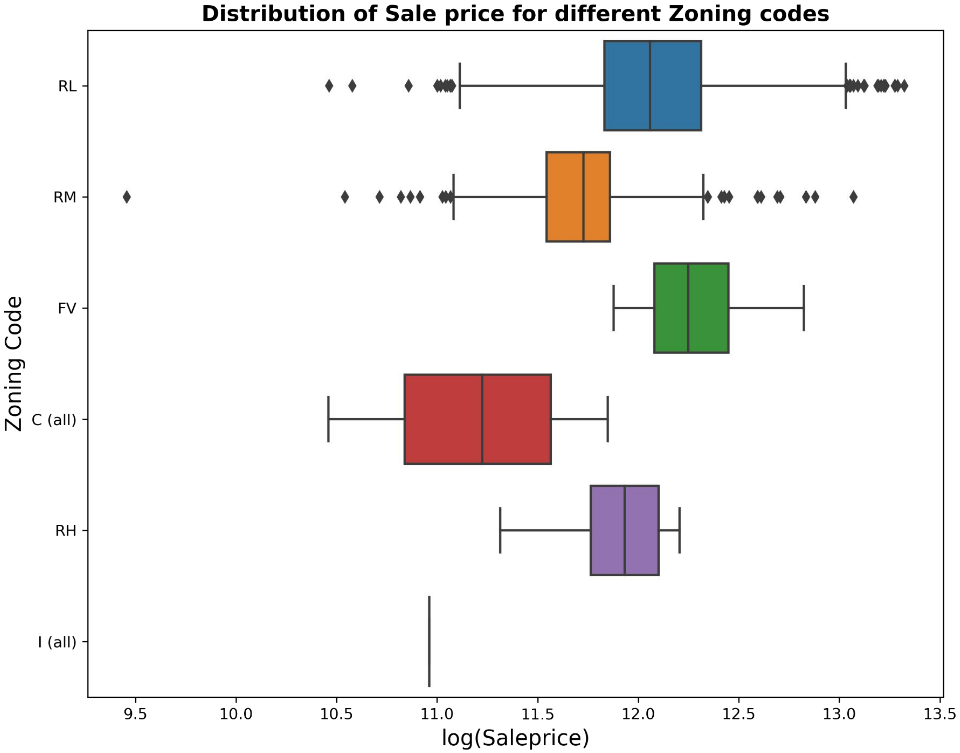
Most important:
SF above + below ground:
‘total living area’

Linear relationships to log(Saleprice)

Outliers



Building a model



For **Categorical** data, I checked

1) Difference in sale price

2) How many are there in each category?

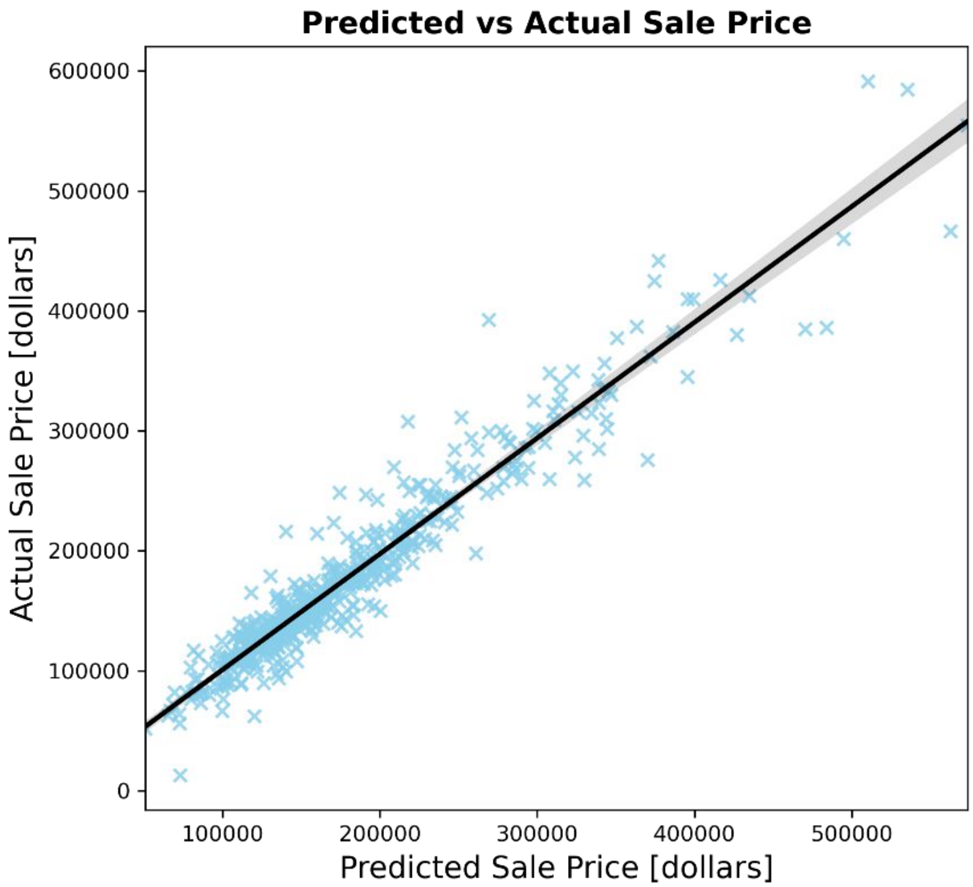
Model

Best model:
12 numerical features and
9 categorical features

How much does this improve over the benchmark model?

Benchmark versus best model R2-score:
0.64 vs **0.88**

Benchmark versus best model RMSE:
$44700 vs **$22100**



Conclusions

1) The new model I built