AMES, IOWA
HOUSING

Predicting Sale Prices with Machine Learning

Welcome!

Chris Landschoot

I am a Data Scientist leading the data science team at Zillow

Background

Real Estate Today:

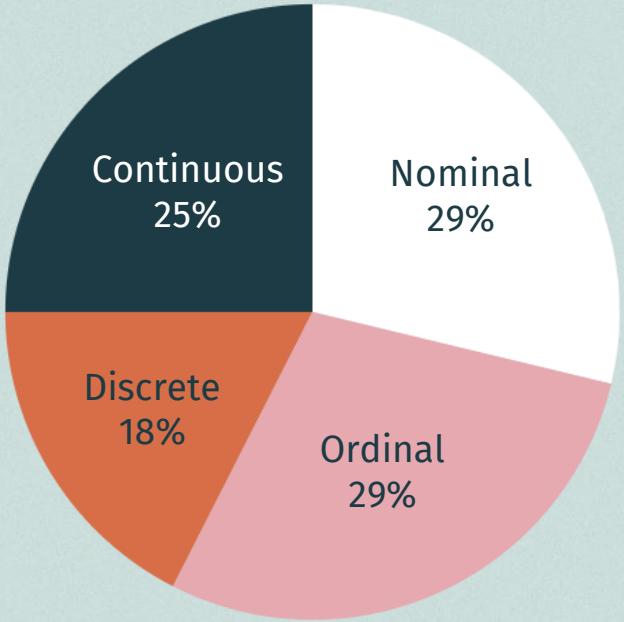
- Currently housing prices are predicted manually by assessors.
- This is time consuming and costly.
- Digital real estate companies are looking for a way to more efficiently and accurately predict home prices.
- Top digital real estate companies (Zillow, Trulia, Redfin, etc...) have agreed to enter a competition to focused on developing the best house price prediction algorithm.
- The platform Kaggle is being leveraged to host this competition.

PROBLEM STATEMENT

- Zillow is seeking to create a proof-of-concept linear regression machine learning model focused on accurately predicting the price of houses at sale.
- This model will compete in the real estate Kaggle competition.
- Housing data from Ames, IA will be used to prototype this technology.

DATASET

From the Ames Assessor's Office used in computing assessed values for individual residential properties sold in Ames, IA from 2006 to 2010.



81 Features

Numrical

20 Continuous Features
14 Discrete Features

Categorical

23 Nominal Features
23 Ordinal Features

Ames, IA Housing Data:
https://jse.amstat.org/v19n3/decock/DataDocumentation.txt

DATA SCIENCE PROCESS

01 Data Cleaning & Exploratory Data Analysis (EDA)

Fix data errors and visually explore the data

02 Pre-processing & Feature Engineering

Manipulate the data and create new features

03 Model Evaluation

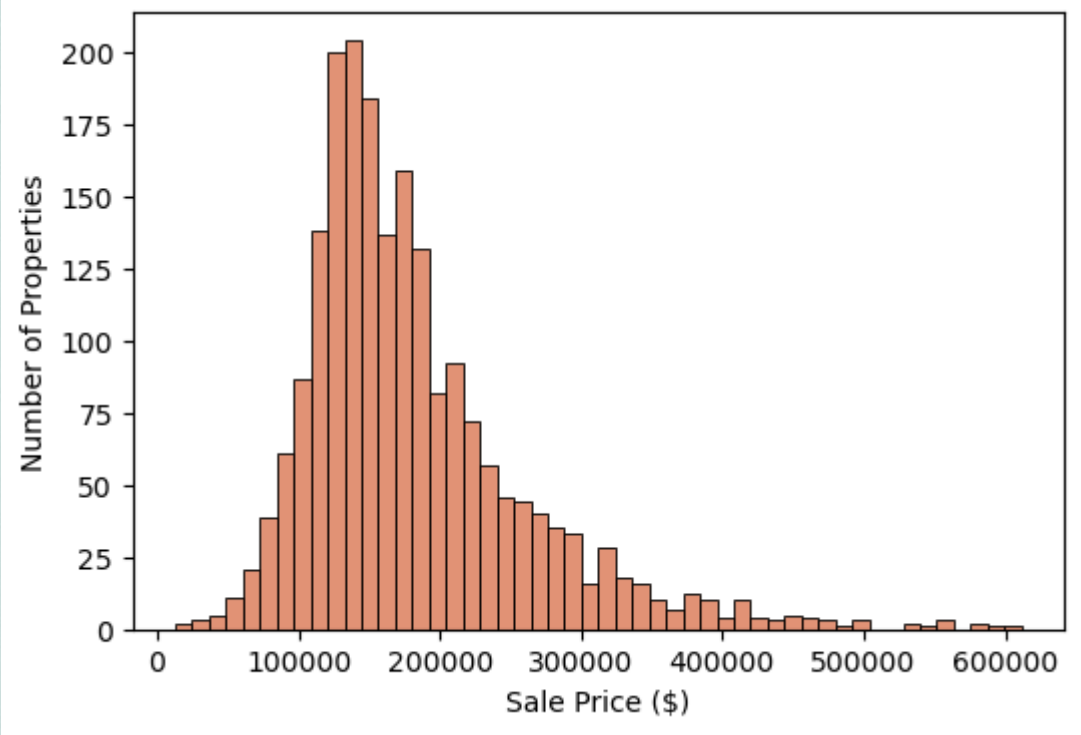
Test and tune different models

04 Conclusions & Recommendations

Interpret the model findings and provide insight

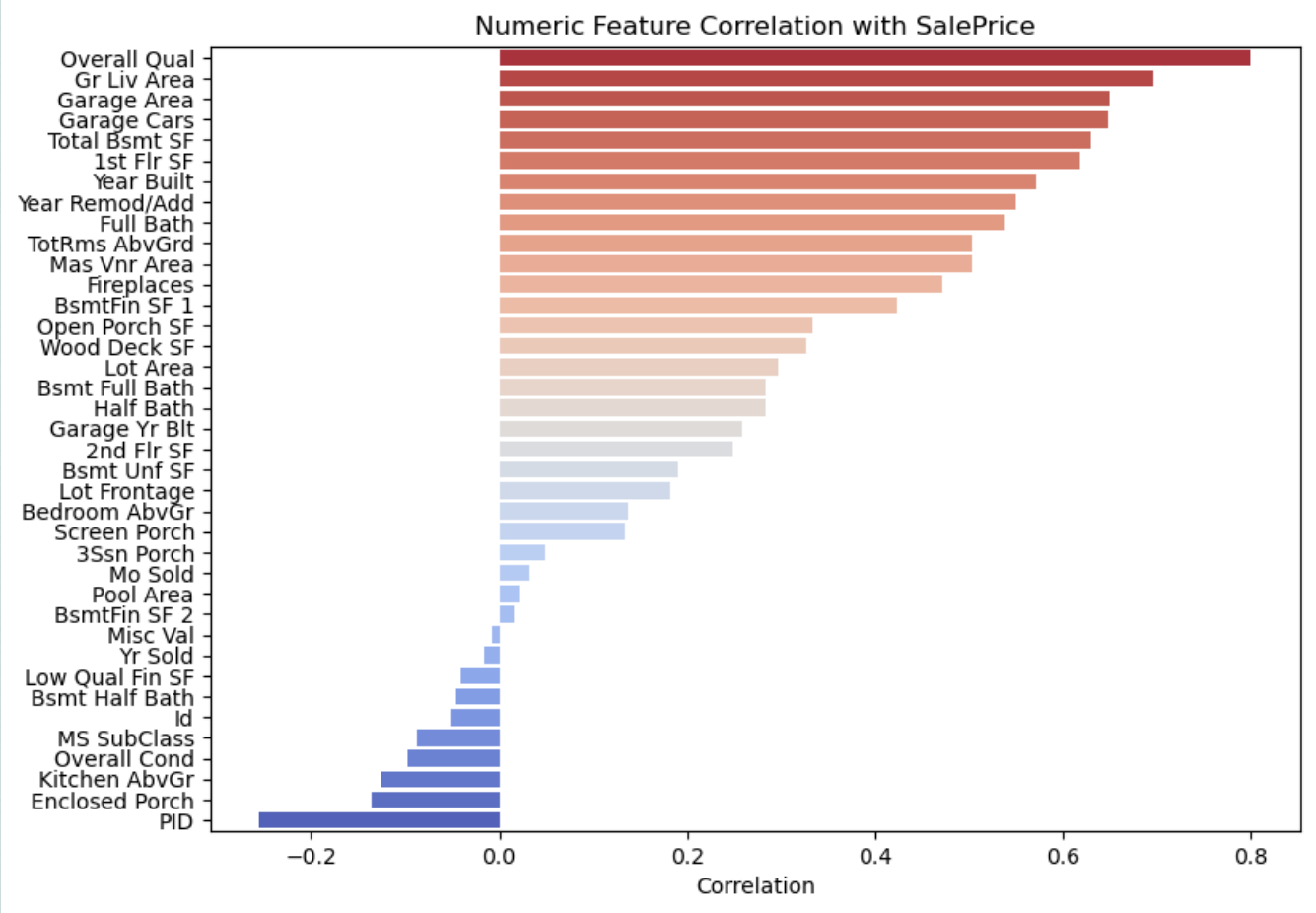
EDA

Sale Price Distribution



DA

Range of Feature Correlation



Feature Engineering

One-Hot Features

Convert nominal categories to columns of 1s and 0s

Mapping Ordinal Features

Converting categorical features with a natural scale to numbers

Engineered Features

Adding multiple features together to create new features

Polynomial Features

Multiplying and squaring features to determine interaction between features

Models

Ordinary Least Squares (OLS)

No Regularization

Ridge Regression

L2 Regularization

LASSO Regression

L1 Regularization

NOTE: Regularization helps models generalize better to new unseen data

Conclusions

- The OLS model performed the best on the training data.
- The Ridge model performed the best on unseen data.
- The Ridge model is selected because it generalizes best to new data.
- Final metric:
  - R^2 ~ 95% of the variability in Sale Price can be explained by our model, all else held constant.

Recommendations

- The prototype proved successful.
- Zillow should allocate financial resources toward developing this predictive technology.
- Resources should be distributed to collecting larger and better data sets as well as continuing to refine and improve the model.

- NOTE: This model is specifically targeted to predict home price.
- If there is interest in predicting how a single factor affects home price, a less complex model should be developed.

THANKS!

Any questions?