AMES HOUSE PRICE PREDICTION

FEB 2023

Alice Wang

INTRODUCTION

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. There are much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, the purpose of this project is to predict the final price of each home.

Problem Statement

How much are the residential homes in Ames, Iowa given all aspects/attributes of the properties?

Goal

1. Build regression models to leverage attributes of properties to predict sales price.
2. Provide useful house process prediction for setting data-driven budgets for buyers.
3. Provide insightful tool for setting more reasonable prices when selling/buying properties.

Data

Ames House Dataset (Kaggle Dataset) with 79 attributes and sales prices of 1,460 properties.

Features examples:

- Exterior features: Exterior material quality, type of foundation, masonry veneer type
- Interior features: heat condition, central air conditioning, kitchen quality
- Location feature: Zoning classification, slope of property, physical location within city limits

Data Cleaning

[Data Cleaning Notebook]

Major steps:

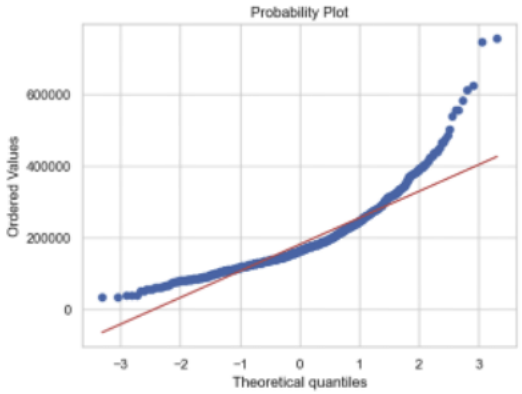
1. Check for duplicates: There are 0 duplicates for this dataset
2. Impute missing values
3. Transform numerical features to categorical when they are true categorical
4. Encode some categorical features as ordered numbers when there is information in the order

EDA

[EDA Notebook]

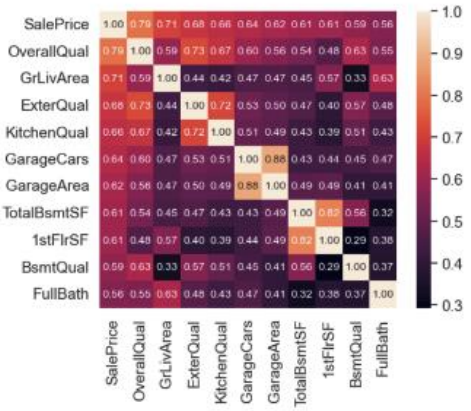
Sale Price is what we are predicting, EDA is all related to SalePrice (our target feature). Below are some take away:

Price of properties in Ames, Iowa

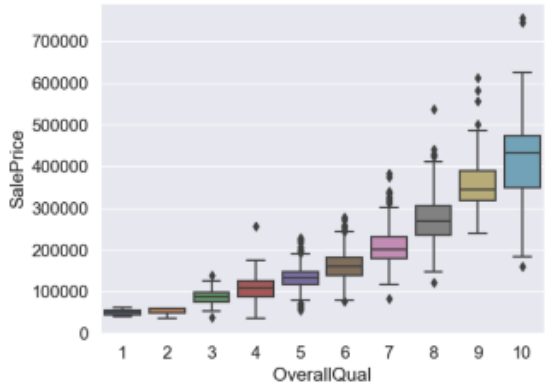


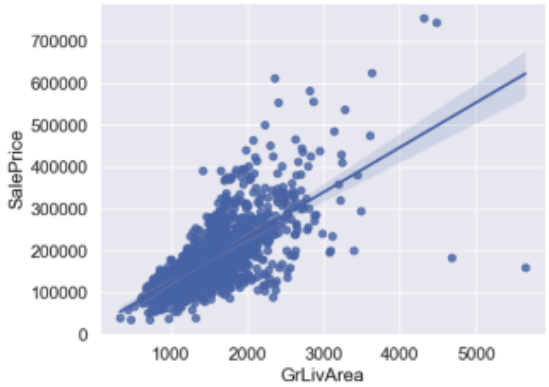
1. Average house sale price is around $180,000 in Ames.
2. There are a few properties over $500,000, but majority of houses are priced under $214,000.
3. The target variable is right skewed (positive skewness) and shows weakness. As (linear) models fits better on normally distributed data, we require proper transformation.

The most important numeric predictors

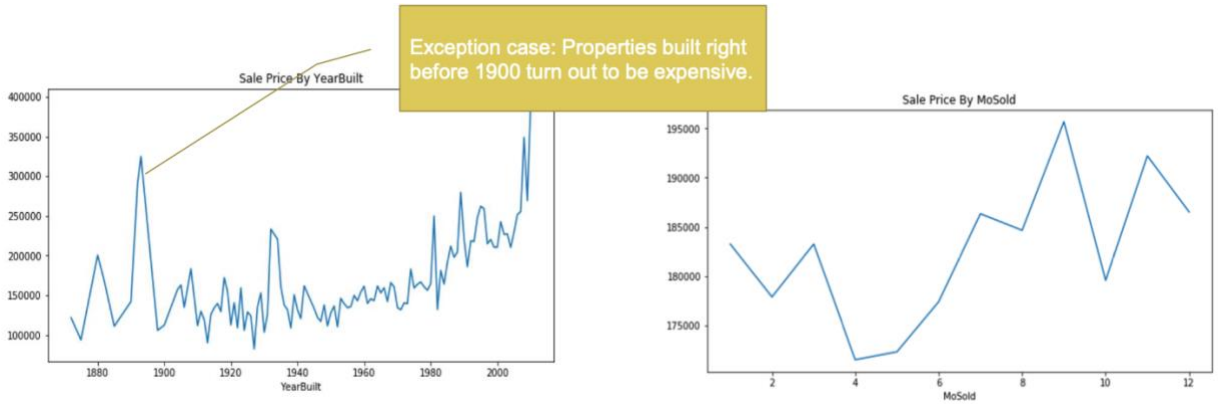


Attributes most correlated with sale price





Time vs Sale price



Feature Engineer & Modeling

Feature engineering & ML Notebook

Feature Engineering Major Steps:

1. Encode some categorical features as ordered numbers when there is information in the order.
2. Create new features.
3. Log transform of the skewed numerical features to lessen impact of outliers.
4. Transformation of categorical features vis one-hot encoding.
5. Split into testing and training datasets.
6. Standardize the magnitude of numeric features using a scaler.

Modeling Major Steps:

Created below models with RandomizedSearchCV hyperparameter tuning for model optimization.

1. Ridge
2. Lasso
3. ElasticNet
4. XGBoost
5. Gradient Boosting Regression
6. Light GBM
7. Stacking Model with ElasticNet as the meta_mode

Model Validation Metric:

1. RMSE
2. R Squared

	Model	R2	RMSE
0	Ridge	0.903	0.124
1	Lasso	0.898	0.127
2	ElasticNet	0.904	0.123
3	XGBoost	0.898	0.127
4	Gradient Boosting	0.890	0.132
5	Light GBM	0.891	0.131
6	Stacking Model	0.910	0.119

Prediction Tool

The users can use the prediction tool (created by IPyWidget) to estimate the house price by adjusting the top 5
features.

Future improvements

I would love to spend more time testing new hyperparameters or different modeling methods to predict the sale price even more accurately.