Ames Housing Price Prediction

Introduction

PROBLEM STATEMENT

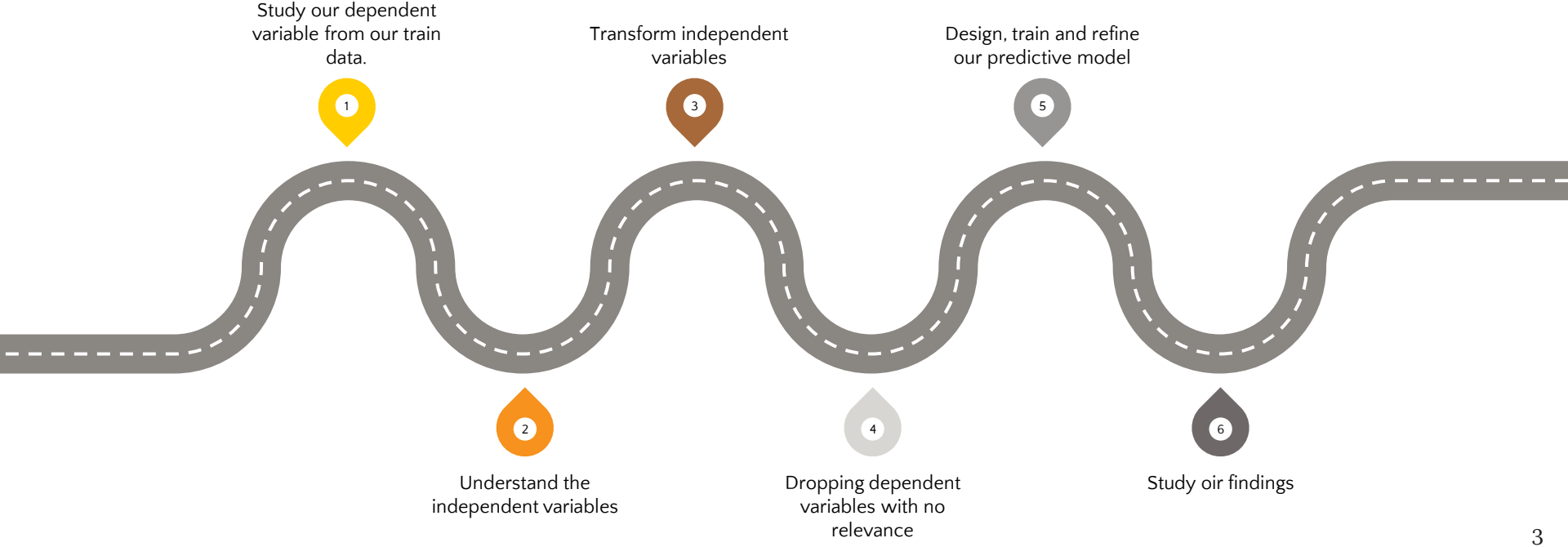
We seek to understand what the factors are that drive house prices and what exactly their relationship is.

CASE STUDY

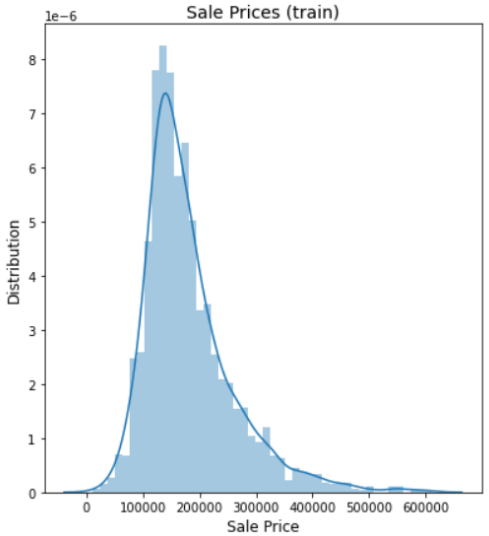
We use the data from Ames, a city in Story County Iowa. It is best known as the home of Iowa State University.

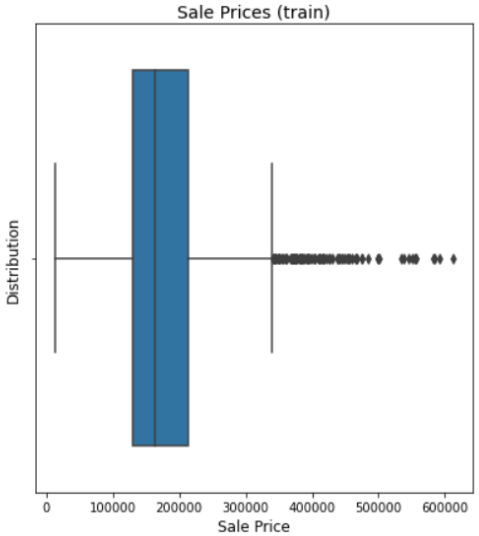
The data we have is obtained from the Ames Assessor's Office which is used for tax assessment purposes. The type of information obtained contains the same data a home owner would look at when making a house purchase.

Roadmap


Study our dependent variable from our train data.
1
Understand the independent variables
2
Transform independent variables
3
Dropping dependent variables with no relevance
4
Design, train and refine our predictive model
5
Study oir findings
6

Sales Prices from Data





From the training data we have, we have 2051 unique observations.

The mean house sale price is **181, 470**, with a standard deviation of **79, 259**.

The 25% - 75% ranges from **129,825** to **214,000**. The minimum is **12,789** and maximum is **611,657**.

We see that the data is normally distributed with a larger mass on the left of the distribution.

Understanding Independent Variables

80

Total variables from our training dataset that cover the size of various areas on the property, surrounding neighborhood characteristics and quality/ condition of house and features

Numeric and Categorical

Variables exist in our data

Numeric variables include things like lot area, first floor square foot etc.

Categorical variables include neighborhood, house style etc.

Ordinal/ Nominal/ Continuous

Variables exist.

Ordinal variables include quality and condition of certain features (good, bad, poor)

Nominal includes things like neighborhood.

Continuous includes things like square footage

Transforming Data

Modifying Variables

We change some variables to make them more useful – for example, introducing house age from 2 existing variables, year the house was built and the year it was sold

Handling Irregular Data

Understand why certain values are missing and construct meaningful ways to assign values to these fields

Identify outliers in our data and remove them as well

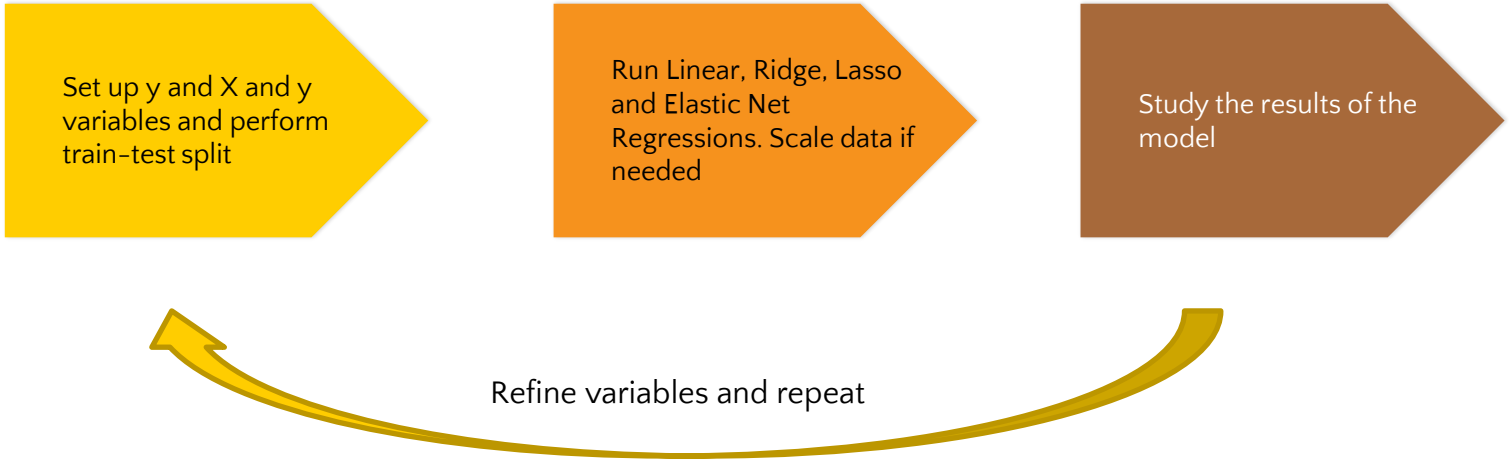
Converting Categorical Data

Our model needs numeric data to work with and we have to convert categorical data to numeric form

We plot distribution charts, boxplots and correlation maps to understand how best to reassign values

Dropping Variables

Model Workflow



Optimal Model

Model	StandardScaler	Variables	Mean Squared Error
Linear Regression	No	Original	734410767.92
Ridge Regression	No	Original	734143633.95
Lasso Regression	No	Original	1169563685.70
Elastic Net Regression	No	Original	1169563685.70
Ridge Regression	Yes	Original	733657123.27
Lasso Regression	Yes	Original	734691265.93
Elastic Net Regression	Yes	Original	734691265.93
Linear Regression	No	Refined	691132691.90
Ridge Regression	Yes	Refined	690447400.87
Lasso Regression	Yes	Refined	694845573.86
Elastic Net Regression	Yes	Refined	695424166.38

We see that the ridge regression with scaled data and a refined list of independent variables is the best model for our use case

Findings

Variables	Coef
Neighborhood_StoneBr	43368.555546
Neighborhood_NridgHt	27271.587655
Gr Liv Area	22746.082235
Neighborhood_Crawfor	18457.100056
Neighborhood_NoRidge	16388.453824
BsmtFin SF 1	11851.846853
Exter Qual	11652.203904
Overall Qual	11090.268480
Neighborhood_BrkSide	10864.290452
Kitchen Qual	9885.802937

Out of the 41 variables selected for our final model, we see that out of the top 10 ones with positive coefficients, neighborhoods have a large effect on house price.

Other notable variables include the total living area, basement area, quality of the exterior, kitchen and overall quality of the house.

Findings

Variables	Coef
Neighborhood_Mitchel	-2618.787855
Neighborhood_Veenker	-3157.646598
Remod Age	-3469.851026
House Style	-3569.375556
Neighborhood_Gilbert	-4066.250386
Mas Vnr Type	-4161.474781
Neighborhood_CollgCr	-4910.731472
House Age	-5429.274316
Exterior 2nd	-5559.735998
Neighborhood_NWAmes	-6609.657700
Neighborhood_SawyerW	-10653.482919

The top 10 variables that have a negative impact on house prices also include the neighborhood in which the house is located.

The type of exterior and mansonry veneer wall. The house age and age since last remodification also plays an important part.

Findings

LOCATION

The location of the house is by far, the most important factor in predicting the price of a house.

Certain neighborhoods fetch a higher premium while others, a discount.

SIZE and QUALITY

The size of the house is of course, an important factor in the price of a house.

The larger the house and living area, the more expensive the house will be.

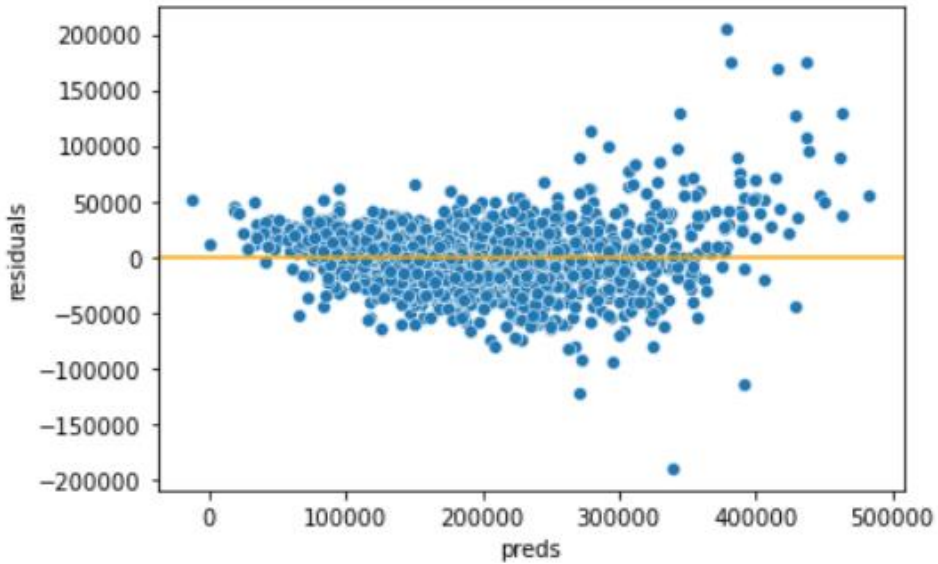
The quality of the house's exterior, kitchen and overall quality are also important factors.

AGE

The age of the house and time since last remodification is also another important factor that contributes to the price of a house.

The older the house, the cheaper it will be.

Shortcomings



Overall, the model is decently accurate in predicting house prices as seen from the plot of residuals clustering around 0.

However, the model becomes less accurate as we move above the 400,000 house sale price.

Thanks!