House prices - Advanced Regression Techniques

Lamba,Anchal
EstradaBerlanga,Alex
Guenthner,Toby
Kim,Jeffrey
Govindarajan,Krithika

(Where are they?)
Lin,Jinbing
Lan,XingYang
Li,Derek

July 26, 2021

Abstract

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetuer a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetuer. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.

Contents

1 Introduction

There are so many characteristics of a house that contribute to its price, like its location, available utilities, neighborhood, and area. For our project, we use machine learning to create a regression model that predicts the final price of a house after considering all these factors. We decided this project would tremendously help people make educated decisions on buying or selling a house. Since many people are not aware of everything that goes into a houses final price, or even what trends will cause a house to appreciate/depreciate, we thought that this project would be ideal for people to analyze the potential outcomes of their purchase or sale.

We developed our project after reading various research studies about regression analysis for housing price prediction. Building on these, our project focuses on finding the sales price of a house based on variables including the lot area, land slope, condition, building type, number of bedrooms and bathrooms, included amenities, and many others. Since the sales price of a house is a continuous variable, we used regression to accurately fit the data. We started out by formatting the categorical and numerical data, converting all the categorical data into strings. We then extracted all the missing values and replaced them with None, taking into consideration that many of these missing values had to do with the absence of a specific amenity, for example, No Garage. We normalized some of the values that didnt make sense when equal to zero and scaled all the numerical data.

After preprocessing and scaling the data, we dropped some of the features that we felt were irrelevant. With the rest of the dataset, we applied five machine learning models, namely, Ridge Regression, Lasso Regression, XGBoost Regression, KNN Regression, Linear Regression, Support Vector Regression, and Neural Networks. After analyzing the results, we trained each of these models again with extra hyperparameters to see if they increased or decreased the accuracy. By comparing the accuracy of all the algorithms used, we were able to decide on an efficient regression algorithm for the problem.

We obtained our dataset from a research study conducted in Ames, Iowa. The study was a machine learning project undertaken by Dean De Cock to create a dataset to describe housing prices in Boston. The complete dataset contains 80 attributes, the last one being the sales price, our target variable.

2 Literature Review

3 Dataset Description and Exploratory Data Analysis

The data comprises a set of information of various homes in Ames, Iowa. The data set contains roughly 1500 entries with 75 features that can be used to describe each house that a sample represents. These features are used to predict the price that a house was sold for. The features in the data provide information about the age of the house; when it was sold; materials used for some parts of the house, such as the roof; quality, size, and quantity of rooms and amenities; and more. Together the 75 features detail effectively every part of the house. A full list of the features and their descriptions can be found at https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data.

Not every feature provided was of significance however. Some features proved either to have no correlation to the house price, or had too many null values to be effective in training. For example, a feature titled Id provided the numerical value in the data for each entry, effectively the row number of the matrix. This clearly has no particular relation to houses so it was removed. Removal based on null value frequency was based on Figure 1. PoolQC and MiscFeature, as an example, had 99% and 96% of their total data filled in as null values; with so few data points to work with these were removed. The threshold for removing these types of features was about

80%, fewer than that and we were willing to keep the feature for training.

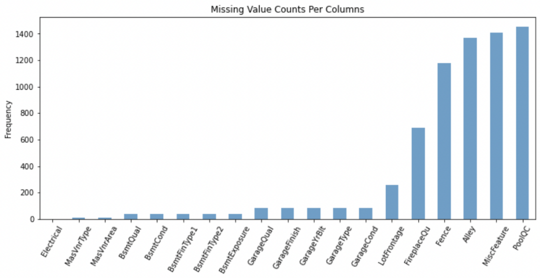


Figure 1: Features with NaN values and the total number of NaN values.Those close to 1400 are effectively filled only with NaN values and thus a negligible feature.

Many categorical features that had null values also had data descriptions that allowed for the null values to make logical sense. The null values for these features meant that this feature was not present for the particular house, so we replaced those null values with an alternative categorical string value to be trained with later. The rest of the features with null values were split into numerical and categorical values.

The categorical values that were left over with null values did not have any information to explain what the null values meant. Thus these were treated as missing values from the data. Given that the quantity of missing data was low, only up to 8 null values total for each of these features, we decided to remove entries with missing values from the dataset. As a result, 9 entries in total were removed from the initial dataset.

Many of the numerical values indicated a null value due to lack of necessary prerequisite, such as basement bathroom quantity being marked null due to the home not having the basement feature. In these cases we filled the null values to zero in accordance with the lack of the feature. The LotFrontage and GarageYrBlt features are trickier because their null values do not indicate a value of zero. Since LotFrontage contains an appreciable number of entries in the data with null values, removing the entries as with the categorical variables could reduce the effectiveness of training later. Therefore, we decided to populate LotFrontage null values with the median value of this feature. GarageYrBlt was found to have a strong linear correlation with the year the house was built, generally they followed a trend of being built together in the same year or the garage being built a bit earlier than the house as a whole. Thus, we used the correlation to fill the null values in this feature with the year the house was built subtracted against the median difference in the year of the house and garage being built.

At this point every null value in the data has been processed and the data has been cleaned, but not fully processed. Some values in the data presented themselves as obvious outliers when plotted against the sale price that we want to predict. As an example, in figure 2, we noticed clear outliers on the right side of the graph that do not trend with the linear relationship between these two variables and thus these entries in the dataset containing the outliers were also removed.

Plotting the q-q graph of the sale price dependent variable we additionally found that the distribution is not normal, but rather skewed right. Our goal with applying machine learning to the dataset is to produce predictions on the test data, where these predictions are expected to form a normal distribution themselves. Thus, for the best possible scores for the algorithms we decided to transform the dependent variable to approximate a normal distribution. In testing we found that applying a logarithm provided the best approximation. Thus we will later use algorithms to form predictions on the log of the sale price and exponentiate the results to score the predictions against the true sale price values, rather than the transformed values.

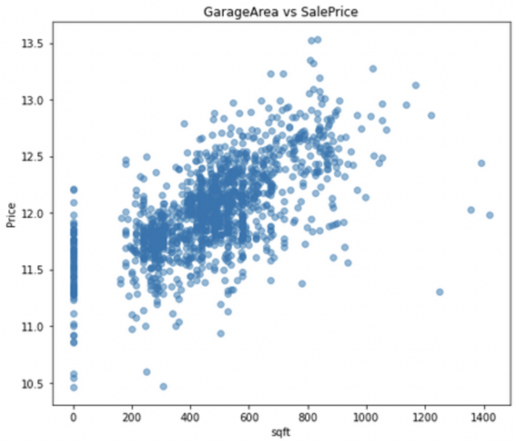


Figure 2: Garage Area vs Sale Price. Square footage greater than 1200 and price less than 13 (hundred thousand dollars) appear to be outliers from the linear trend between these two variables and are marked for removal.

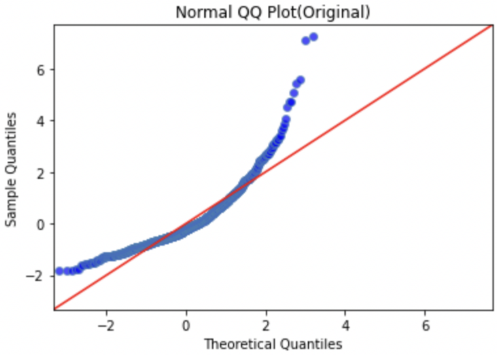


Figure 3: Normal QQ plot. It assess if a set of data plausibly came from some theoretical distribution.

4 Proposed Methodology

4.1 Feature Seletion

There are 282 features left after extensive data preprocessing and cleaning. In order to identify the best set of features that would enable our model to accurately predict the SalePrice, we need to perform the feature selection process. From Scikit-Learn, we attempted three separate methods:

VarianceThreshold
SelectFromModel using Lasso Regression
RFE

After comparing each method's impact on the R^2 scores and MSE values for the ML regression methods below, we came to the conclusion that the dataframe(df) should be left as is, without any feature selection. A visual of this is in feature$_{s}election_{a}ttempt.ipynbwiththeVarianceThreshold()method, wherethe$R values dropped significantly after the removal of some features.

4.2 Regression Methods

Each regression method has been researched and chosen for its capabilities in supervised learning, especially regression. In order to avoid using biased models, 10-fold cross validation has been used to generalize each model based on the given data set. Specifically, the R^2 score and MSE has been obtained for each iteration of the cross validation, as well as the average of both metrics.

4.2.1 Ridge Regression

This model solves a regression model where the loss function is the linear least squares function and regularization is given by the L2-norm.

4.2.2 Lasso Regression

This model solves a regression model where the loss function is the linear least squares function and regularization is given by the L1-norm.

4.2.3 XGBoost Regressionn

This model is an implementation of the gradient boosting trees algorithm for regression-based data.

4.2.4 KNN Regression

This model performs regression based on k-nearest neighbors, where the target variable is predicted by local interpolation of the nearest neighbors in the training set.

4.2.5 Linear Regression

This model perfoms ordinary least squares linear regression by minimizing the residual sum of squares between the observed targets (predicted by linear approximation) in the dataset.

4.2.6 Random Forest Regression

This model uses a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset, and uses averaging to improve the predictive accuracy and control over-fitting.

4.2.7 Support Vector Regression

This model identifies the data points within the epsilon-defined error decision boundary to produce a line of best fit for predicting discrete values. For our dataset, the linear kernel was the best fit.

4.2.8 Neural Network

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

4.3 Finding Optimal Hyperparameters

We will use GridSearch() to populate optimal hyperparameters for the above regression methods. In this section and the following section, we will exclude lasso regression and linear regression, as their R^2 scores were negative and indicate a worse fit than the mean line. Additionally, we will exclude the neural network because optimization has already been attempted through trial-and-error of the number of hidden layers, neurons, activations, and epochs.

5 Experimental Results

5.1 Ridge Regression

This model solves a regression model where the loss function is the linear least squares function and regularization is given by the L2-norm.

5.2 Lasso Regression

This model solves a regression model where the loss function is the linear least squares function and regularization is given by the L1-norm.

5.3 XGBoost Regressionn

This model is an implementation of the gradient boosting trees algorithm for regression-based data.

5.4 KNN Regression

This model performs regression based on k-nearest neighbors, where the target variable is predicted by local interpolation of the nearest neighbors in the training set.

5.5 Linear Regression

This model perfoms ordinary least squares linear regression by minimizing the residual sum of squares between the observed targets (predicted by linear approximation) in the dataset.

5.6 Random Forest Regression

This model uses a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset, and uses averaging to improve the predictive accuracy and control over-fitting.

5.7 Support Vector Regression

This model identifies the data points within the epsilon-defined error decision boundary to produce a line of best fit for predicting discrete values. For our dataset, the linear kernel was the best fit.

5.8 Neural Network

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

6 Conclusion and Discussion

6.1 Conclusion

As we can see, ridge regression had the best results, followed by support vector regression.

ML Algorithms	**Average R^2 score**	**Average MSE**
Ridge	0.9108344112750121	0.013836693343777774
SVR	0.9102385055200841	0.013900808872506202
XGBoost 3	0.9015706563418072	0.015280968941103323
Neural Network 3	-	0.019019585102796555
Random Forest	0.8726704143296444	0.019821473835094878
KNN	0.780005248889875	0.03405818328993641

Table 1: Table of the R^2 scores and MSE values for each regression method after implementing the optimal hyperparameters.

6.2 Discussion

A Data Fields

SalePrice: the property's sale price in dollars. This is the target variable that you're trying to predict.
MSSubClass: The building class
MSZoning: The general zoning classification
LotFrontage: Linear feet of street connected to property
LotArea: Lot size in square feet
Street: Type of road access
Alley: Type of alley access
LotShape: General shape of property
LandContour: Flatness of the property
Utilities: Type of utilities available
LotConfig: Lot configuration
LandSlope: Slope of property
Neighborhood: Physical locations within Ames city limits Condition1: Proximity to main road or railroad Condition2: Proximity to main road or railroad (if a second is present)
BldgType: Type of dwelling
HouseStyle: Style of dwelling
OverallQual: Overall material and finish quality
OverallCond: Overall condition rating
YearBuilt: Original construction date
YearRemodAdd: Remodel date
RoofStyle: Type of roof
RoofMatl: Roof material
Exterior1st: Exterior covering on house
Exterior2nd: Exterior covering on house (if more than one material)
MasVnrType: Masonry veneer type
MasVnrArea: Masonry veneer area in square feet
ExterQual: Exterior material quality
ExterCond: Present condition of the material on the exterior Foundation: Type of foundation
BsmtQual: Height of the basement
BsmtCond: General condition of the basement
BsmtExposure: Walkout or garden level basement walls
BsmtFinType1: Quality of basement finished area
BsmtFinSF1: Type 1 finished square feet
BsmtFinType2: Quality of second finished area (if present)
BsmtFinSF2: Type 2 finished square feet
BsmtUnfSF: Unfinished square feet of basement area
TotalBsmtSF: Total square feet of basement area Heating: Type of heating
HeatingQC: Heating quality and condition
CentralAir: Central air conditioning
Electrical: Electrical system
1stFlrSF: First Floor square feet
2ndFlrSF: Second floor square feet
LowQualFinSF: Low quality finished square feet (all floors)
GrLivArea: Above grade (ground) living area square feet
BsmtFullBath: Basement full bathrooms
BsmtHalfBath: Basement half bathrooms
FullBath: Full bathrooms above grade
HalfBath: Half baths above grade
Bedroom: Number of bedrooms above basement level
Kitchen: Number of kitchens
KitchenQual: Kitchen quality
TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
Functional: Home functionality rating

Fireplaces: Number of fireplaces

FireplaceQu: Fireplace quality

GarageType: Garage location

GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

GarageQual: Garage quality

GarageCond: Garage condition

PavedDrive: Paved driveway

WoodDeckSF: Wood deck area in square feet

OpenPorchSF: Open porch area in square feet

EnclosedPorch: Enclosed porch area in square feet

3SsnPorch: Three season porch area in square feet

ScreenPorch: Screen porch area in square feet

PoolArea: Pool area in square feet

PoolQC: Pool quality

Fence: Fence quality

MiscFeature: Miscellaneous feature not covered in other categories

MiscVal: Value of miscellaneous feature

MoSold: Month Sold

YrSold: Year Sold

SaleType: Type of sale

SaleCondition: Condition of sale

B Report Instructions

Your technical report should focus on the work that you have done in the project. Additionally, you should also provide brief descriptions of the components/parts that you rely on in your work. Make sure to describe everything in your own words!

This template includes the basic structure for your final technical report. You should keep the overall structure by not changing the *sections*. You can still adjust the structure a bit by adding *subsections*. In the appendix, you find these instructions as well as some LaTeX examples. Before you hand in the report, makes sure you have deleted all *lipsum* fillings and the instructions/examples from the appendix.

Abstract

- Maximum length: 10 lines!
- Important: Do not put references into the abstract
- Content of the abstract:
  - Punchline
  - Introduction. Why should I care?
  - What is the problem? How did you tackled the problem?
  - How did you go about doing the research that follows from your idea?
  - Whats the key impact of your research?

Main Part

- Main Part consists of 6 sections: Introduction, Dataset, Methodology, Methods, Evaluation, Conclusion
- Maximum(!) length without figures/tables: 3 pages
- Suggested length with figures/tables: 4 pages
- Additional content should be put in appendix.
- Content in appendix must not be required to understand the main part.

Introduction

- General introduction of the problem that you were trying to solve.
- What is the brief problem description?
- Why is it important/interesting/challenging?

Dataset

- General description of the dataset that you were using.
- Important details about the dataset.
- What modifications/preprocessing did you do (if any)? Only mention things here that you used for all of your methods later, e.g., resizing, cropping, or combining images.
- What are the details of the final modified dataset that you used.

Methodology

- What do you want to do? What kind of problem is it?
- What is the dataset?
- What is the input of each part?
- What is the output of each part?
- What is/are the metric/s you want to evaluate on?

Remember: show your reader the wall before you begin to examine the bricks.

Methods

- Methods that you tried

Evaluation

- Evaluation of the methods with the described metric
- Comparison of your methods

Conclusion

- Conclusion of your work
- What is working? What does not work? Where could your methods be improved?

Appendix

- Additional content can be put in the appendix.
- Appendix must not be required to understand the main part.
- You should still not dump all images and data into the appendix. You should make a selection of useful additions.
- Appendix is not required.

C LaTeX Examples

The following examples should help you to write your technical report using LaTeX. You'll find here the examples of tables, figures, citations and references. For other features of LaTeX, see tutorials on **Overleaf** or use this **cheatsheet**. To work with this template, download its entire folder (including /sections, /bibliography and /figures), and run your LaTeX editor like **Overleaf**.

Example Citation

Example of citation: [1] and [2].

Example References

Example of table reference: see Table 2.
Example of equation reference: see Equation (1).
Example of reference to Section **??**.
Example of reference to Subsection **??**.
Example of figure reference: see Figure 4.
Example of subfigure reference: see Figure 5a.

Example list

- Bullet point one
- Bullet point two
- Nested list items:
  - Nested item one
  - Nested item two

Enumerations

1. Numbered list item one
2. Numbered list item two
3. Nested list items:
   (a) Nested item one
   (b) Nested item two

Example Table

Treatments	**Response 1**	**Response 2**
Treatment 1	0.0003262	0.562
Treatment 2	0.0015681	0.910
Treatment 3	0.0009271	0.296

Table 2: Table caption

Example Equation

Equations within the text: $e = mc^2$. Equation with label on its own line:

$$e = mc^2 \tag{1}$$

Example Figures

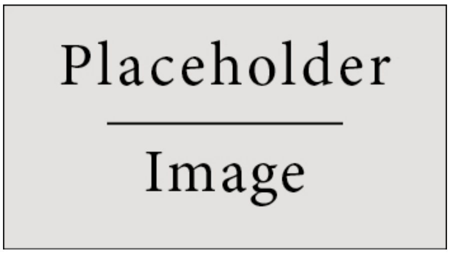
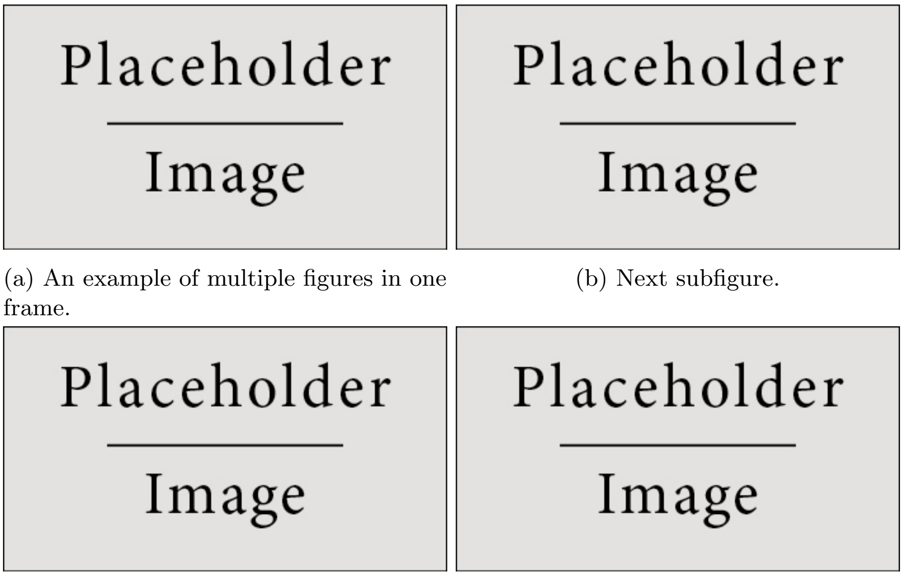


Figure 4: An example of simple figure.



(a) An example of multiple figures in one frame.

(b) Next subfigure.

(c) Subfigure on another line.

(d) Yet another subfigure.

Figure 5: More figures in appendix.

References

[1] A. B. Jones and J. M. Smith. Article Title. *Journal Title*, 13(52):123–456, March 2013.

[2] J. M. Smith and A. B. Jones. *Book Title*. Publisher, 7th edition, 2012.