Predicting House Prices with Machine Learning

Introduction

This notebook is going to be focused on solving the problem of predicting house prices for house buyers and house sellers.

A house value is simply more than location and square footage. Like the features that make up a person, an educated party would want to know all aspects that give a house its value.

We are going to take advantage of all of the feature variables available to use and use it to analyze and predict house prices.

We are going to break everything into logical steps that allow us to ensure the cleanest, most realistic data for our model to make accurate predictions from.

1. Load Data and Packages
2. Analyzing the Test Variable (Sale Price)
3. Multivariable Analysis
4. Impute Missing Data and Clean Data
5. Feature Transformation/Engineering
6. Modeling and Predictions

Understanding the Client and their Problem

A benefit to this study is that we can have two clients at the same time! (Think of being a divorce lawyer for both interested parties) However, in this case, we can have both clients with no conflict of interest!

Client Housebuyer: This client wants to find their next dream home with a reasonable price tag. They have their locations of interest ready. Now, they want to know if the house price matches the house value. With this study, they can understand which features (ex. Number of bathrooms, location, etc.) influence the final price of the house. If all matches, they can ensure that they are getting a fair price.

Client Houseseller: Think of the average house-flipper. This client wants to take advantage of the features that influence a house price the most. They typically want to buy a house at a low price and invest on the features that will give the highest return. For example, buying a house at a good location but small square footage. The client will invest on making rooms at a small cost to get a large return.

Importing Libraries and Dataset

Here we are using

- **Pandas** – To load the Dataframe
- **Matplotlib** – To visualize the data features i.e. barplot
- **Seaborn** – To see the correlation between features using heatmap

Practice Skills

- Creative feature engineering
- Advanced regression techniques like random forest and gradient boosting

Acknowledgments

The Ames Housing dataset was compiled by Dean De Cock for use in data science education. It's an incredible alternative for data scientists looking for a modernized and expanded version of the often cited Boston Housing dataset.
Photo by Tom Thain on Unsplash.

Evaluation

link

keyboard_arrow_up

Goal

It is your job to predict the sales price for each house. For each Id in the test set, you must predict the value of the SalePrice variable.

Metric

Submissions are evaluated on Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price. (Taking logs means that errors in predicting expensive houses and cheap houses will affect the result equally.)

Submission File Format

The file should contain a header and have the following format:

```
Id,SalePrice
1461,169000.1
1462,187724.1233
1463,175221
etc.
```

You can download an example submission file (sample_submission.csv) on the Data page.

more_horiz

Overview

all_inclusive

This competition runs indefinitely with a rolling leaderboard. Learn more.

Description

link

keyboard_arrow_up

Start here if...

You have some experience with R or Python and machine learning basics. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.

Getting Started Notebook

To get started quickly, feel free to take advantage of this starter notebook.

Competition Description

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. But this playground competition's dataset proves that 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, this competition challenges you to predict the final price of each home.

Practice Skills

- Creative feature engineering

- Advanced regression techniques like random forest and gradient boosting

Acknowledgments

The Ames Housing dataset was compiled by Dean De Cock for use in data science education. It's an incredible alternative for data scientists looking for a modernized and expanded version of the often cited Boston Housing dataset.
Photo by Tom Thain on Unsplash.

Evaluation
link

keyboard_arrow_up

Goal

It is your job to predict the sales price for each house. For each Id in the test set, you must predict the value of the SalePrice variable.

Metric

Submissions are evaluated on Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price. (Taking logs means that errors in predicting expensive houses and cheap houses will affect the result equally.)

Submission File Format

The file should contain a header and have the following format:

```
Id,SalePrice
1461,169000.1
1462,187724.1233
1463,175221
etc.
```

You can download an example submission file (sample_submission.csv) on the Data page.

Tutorials
link

keyboard_arrow_up

Kaggle Learn

Kaggle Learn offers hands-on courses for most data science topics. These short courses prepare you with the key ideas to build your own projects.
The **Machine Learning Course** will give you everything you need to succeed in this competition and others like it.

Other R Tutorials

Fun with Real Estate Data

- Use Rmarkdown to learn advanced regression techniques like random forests and XGBoost

XGBoost with Parameter Tuning

- Implement LASSO regression to avoid multicollinearity
- Includes linear regression, random forest, and XGBoost models as well

Ensemble Modeling: Stack Model Example

- Use "ensembling" to combine the predictions of several models
- Includes GBM (gradient boosting machine), XGBoost, ranger, and neural net using the caret package

A Clear Example of Overfitting

- Learn about the dreaded consequences of overfitting data

Other Python Tutorials

Comprehensive Data Exploration with Python

- Understand how variables are distributed and how they interact
- Apply different transformations before training machine learning models

House Prices EDA

- Learn to use visualization techniques to study missing data and distributions
- Includes correlation heatmaps, pairplots, and t-SNE to help inform appropriate inputs to a linear model

A Study on Regression Applied to the Ames Dataset

- Demonstrate effective tactics for feature engineering
- Explore linear regression with different regularization methods including ridge, LASSO, and ElasticNet using scikit-learn

Regularized Linear Models

- Build a basic linear model