



Project 2: Ames Housing Data & Kaggle Challenge

General Assembly
DSI8



Problem Statement

Create and tune a Linear Regression model to predict housing prices at sale in Ames, IA. The Ames Housing Dataset includes over 70 features. Detailed documentation can be found here:

<http://jse.amstat.org/v19n3/decock/DataDocumentation.txt>

Top 10 Features Correlated to Sale Price:

- Overall Quality
- Gross Living Area
- Garage Cars
- Year Built
- Garage Area
- Full Bath
- Total Basement SF
- Year of Remodel
- Foundation
- 1st Floor SF

Modeling Process

- Train / Test Split data
- Standardize features
- Cross Validate models (Linear Regression, Lasso, Ridge)
- Fit Train data to model with best CV score (Ridge)
- Evaluate & Compare Train / Test Scores
- Predict Sale Price & model evaluation score (R^2)

Conclusions: *Model is Overfit*

- Training Score: 92%
- Testing Score: 90%

Recommendations:

Apply feature engineering to reduce model variance.

Remove some less important noisy features and combine others to create new high-value features.

Kaggle Competition

The screenshot shows the Kaggle website interface. At the top, the browser address bar displays 'kaggle.com'. The page header includes the Kaggle logo, a search bar, and navigation links for Competitions, Datasets, Kernels, Discussion, and Learn. The main content area features a banner for the 'DSI-US-8 Project 2 Regression Challenge' with the subtitle 'Predict the price of homes at sale for the Ames Iowa Housing dataset'. Below the banner, a navigation bar includes links for Overview, Data, Kernels, Discussion, Leaderboard (selected), Rules, Team, My Submissions, and Late Submission. The Leaderboard section is divided into Public and Private tabs. A message states: 'The private leaderboard is calculated with approximately 70% of the test data. This competition has completed. This leaderboard reflects the final standings.' A 'Refresh' button is present. Below the message is a table listing the top 10 teams on the public leaderboard.

#	△ pub	Team Name	Kernel	Team Members	Score 📊	Entries	Last
1	▲ 9	Tonya			24769.451...	23	14d
2	▲ 33	Joseph H			24933.062...	5	16d
3	▲ 26	Jerel Novick			26439.733...	21	15d
4	▲ 11	Ari Mello			27846.063...	7	17d
5	▲ 26	Patrick Wales-Dinan			28131.803...	32	15d
6	▲ 22	Bobby Kelsey			28393.195...	36	16d
7	▲ 42	Brandon S			28936.874...	7	16d
8	▲ 17	Gaurav Munjal			29006.449...	39	16d
9	▲ 23	Teng Mao			29017.950...	21	16d
10	▲ 29	Temple Moore			29538.137...	22	16d

Kaggle Leaderboard

Kaggle Conclusions:

- R2 Score: 90.7%
- Model evaluation metric is not aligned with Kaggle

Recommendations:

Re-evaluate model based on RMSE evaluation metric used by Kaggle in order to improve Leaderboard position.