Exploratory Data Analysis of Ames Housing Prices

For this assignment we continued our Analysis of the Housing Prices Dataset for Ames Iowa.

Management/Research Question:

Our primary research question was: "Given a set of home features, how accurately can we predict the sale price of a home in Ames, Iowa?"

This question is important because accurate housing price prediction supports multiple stakeholders:

- Buyers and sellers benefit from realistic pricing expectations.
- Real estate agents and developers use predictions to guide investment and renovation strategies.
- Lenders and financial institutions rely on price forecasts for loan underwriting and risk management.
- Local governments apply these insights in property tax assessments and planning.

From an analytical perspective, this problem also provides a valuable case study in predictive modeling. Housing data is multivariate, heterogeneous, and often noisy, which makes it a strong testbed for modern regression approaches. By applying cross-validation, feature engineering, and regularized regression methods (Ridge, Lasso, and ElasticNet), as well as comparing them with a novel academic model (the professor's BCR algorithm), we aim to answer not only how well can we predict prices, but also which modeling approach generalizes best to new, unseen data.

Exploratory Data Analysis Findings:

The Ames Housing dataset consists of 1,460 residential property sales with 81 descriptive variables. These features span structural characteristics (e.g., square footage, number of bathrooms, garage size), quality ratings (e.g., OverallQual, KitchenQual), and location-based variables (e.g., Neighborhood). Our target variable is SalePrice, which represents the transaction price of each home.

Target Variable Distribution:

SalePrice is right-skewed, ranging from $34,900 to $755,000, with a median around $163,000. Both histograms and box-plots reveal a long right tail corresponding to luxury properties. To stabilize variance and improve model fit, we applied a log1p transformation, which produced a distribution close to normal. This transformation is critical since regularized regression models like Ridge, Lasso, and ElasticNet assume approximately normally distributed residuals.

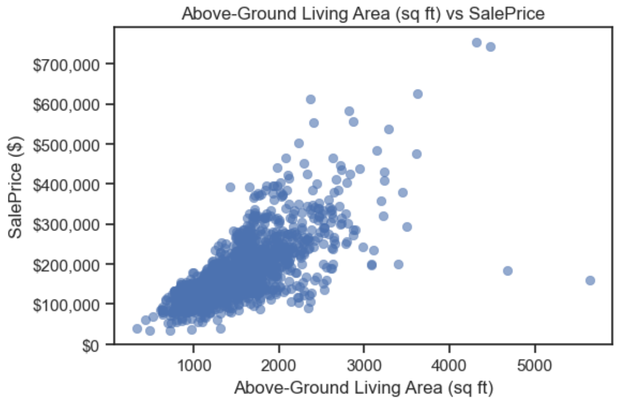
Missing Data:
Several predictors contained missing values. For example:

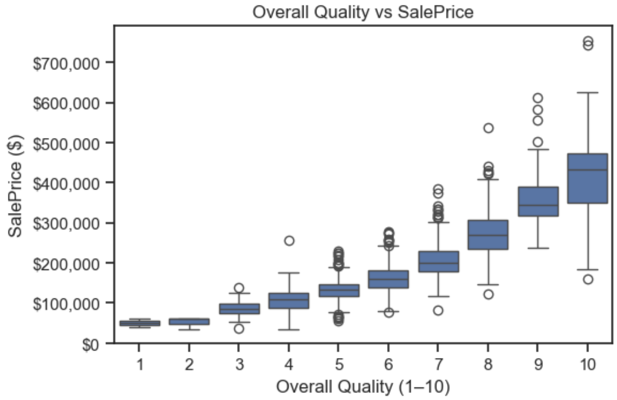
- LotFrontage had missing entries that were imputed using median values within neighborhoods.
- GarageYrBlt and PoolQC were missing primarily when the property lacked a garage or pool; these were treated as meaningful "No Garage" / "No Pool" categories.
- Other numeric features with low missing values were filled using median imputation.

Feature Engineering & Key Predictors:

To improve predictive performance, we expanded the dataset with new features designed to capture important housing characteristics and non-linear effects. Specifically, we created:

- **TotalSF** – Total square footage of the home (sum of basement, first and second floors).
- **TotalBaths** – Combined count of full and half bathrooms, including basement bathrooms.
- **HouseAge** – Difference between the year sold and the year built to reflect property age.
- **SinceRemodel** – Number of years since the last remodel or update.
- **IsRemodeled** – Indicator variable flagging homes that have been remodeled since construction.
- **Piecewise GrLivArea** – Split Above-Ground Living Area into two segments at 2,000 sq ft (GrLivArea_low and GrLivArea_high) to capture non-linear price increases for larger homes.
- **OverallQual_sq** – Squared term for the overall quality score to model non-linear effects of quality.
- **HasGarage** – Indicator variable for whether a property includes a garage.





These features allowed us to evaluate polynomial, indicator, dichotomous, and piecewise components which was the main objective for this assignment. We combined these additional variables with key original features such as Neighborhood, HouseStyle, Exterior Quality, and Kitchen Quality to form a comprehensive predictor set.

Key Relationships:

Correlation analysis revealed that OverallQual, GrLivArea, GarageCars, and TotalSF are strongly associated with SalePrice. Neighborhood effects also play a significant role, as certain locations (e.g., StoneBr, NridgHt) consistently yield higher property values.

This enriched feature set provides a stronger foundation for predictive modeling. Compared to the initial EDA, we improved handling of missing data, applied transformations for skewness, and engineered additional features that better capture home value drivers.

Modeling Approach & Cross-Validation Design:

Our goal was to evaluate multiple regression approaches for predicting home sale prices in Ames, with a consistent design for training, validation, and comparison. To ensure robust results, we adopted a cross-validation framework, specifically 10-fold cross-validation, which divides the training data into ten equal partitions. Each model was trained on nine folds and validated on the remaining fold, rotating until each partition had been used for validation. This design reduces overfitting risk and provides a stable estimate of generalization error.

Models Considered:

1. Ridge Regression-
   - Adds an L2 penalty on large coefficients, which reduces variance in the presence of multicollinearity.
   - Effective when many correlated predictors are present.

2. Lasso Regression-
   - Adds an L1 penalty, shrinking some coefficients all the way to zero.
   - Provides feature selection by keeping only the most predictive variables.

3. ElasticNet Regression-
   - Combines L1 and L2 penalties, balancing feature selection with coefficient shrinkage.
   - Hyperparameters include both the penalty strength (alpha) and the mix ratio (l1_ratio).
   - We performed hyperparameter tuning across a grid of alpha values and multiple l1_ratios (0.1, 0.5, 0.9) to find the best balance.

4. Professor's BCR (Box-Constrained Regression) Algorithm-
   - A novel academic method developed by our professor.
   - Performs regression with a constraint that coefficients remain within a bounded region, offering robustness against outliers and unstable estimates.
   - We implemented this using manual cross-validation folds and compared its RMSE to standard regression models.

Preprocessing Pipeline:

- Standardization was applied to all numeric predictors so that penalty-based methods (Ridge, Lasso, ElasticNet) operated on comparable scales.
- Log1p transformation of SalePrice ensured approximately normal residuals.
- Feature engineering from the EDA section was incorporated into the pipeline, meaning models had access to both original and engineered predictors.

This modeling design allowed us to test how regularized linear regression methods (Ridge, Lasso, ElasticNet) perform relative to each other, while also benchmarking against the professor's BCR algorithm. By holding the cross-validation framework constant, we ensured that differences in performance reflected modeling approach rather than sampling variation.

Results & Model Comparison:

```
Training data shape: (1460, 37)
Target shape: (1460,)

Ridge best alpha: 104.81131341546852
Ridge CV RMSE: 34508.5354040932
Lasso best alpha: 568.9866029018293
Lasso CV RMSE: 34793.04884222548
ElasticNet best alpha: 0.21209508879201905
ElasticNet best l1_ratio: 0.5
ElasticNet CV RMSE: 34862.68448694951

Running BCR cross-validation...
Fold 1 RMSE: 35865.74
Fold 2 RMSE: 35608.93
Fold 3 RMSE: 55606.41
Fold 4 RMSE: 30795.25
Fold 5 RMSE: 25981.28
BCR fold RMSEs (in $): [35865.743880328555, 35608.93499668994,
55606.4133299498, 30795.251671284866, 25981.278880372836]
BCR CV RMSE (in $): 36771.524551725204

--- Model Comparison ---
Ridge CV RMSE:      34508.5354040932
Lasso CV RMSE:      34793.04884222548
ElasticNet CV RMSE: 34862.68448694951
BCR CV RMSE:        36771.524551725204
```

These results show that regularized regression models (Ridge, Lasso, ElasticNet) performed consistently well, with Ridge achieving the lowest error on cross-validation.

Management Implications and Recommendations:

Our analysis demonstrates that housing prices in Ames, Iowa can be predicted with reasonable accuracy using regularized regression models. Ridge, Lasso, and ElasticNet regression each achieved cross-validation errors in the range of $34K–35K, which is a relatively small fraction of the median home price (~$163K). On Kaggle's evaluation metric (RMSLE), our models ranked competitively, with Ridge achieving 0.1499 and ElasticNet improving further to 0.1342.

Key Findings for Stakeholders:

- **Buyers and Sellers:**

Predictions from Ridge and ElasticNet models provide a reliable benchmark for estimating fair market value. For buyers, this reduces the risk of overpaying, while sellers can use these models to set realistic listing prices and expectations.

- **Real Estate Agents and Developers:**

The models identified several important features driving price variation, including total square footage, bathroom count, remodeling status, and overall quality scores. These insights help agents highlight value-adding characteristics when marketing homes and guide developers in prioritizing renovations or upgrades.

- **Lenders and Financial Institutions:**

With RMSE values around $34K, these models can support loan-to-value assessments and risk management. A predictive framework provides a consistent way to evaluate collateral, especially for mid-range homes where prediction error is lowest.

- **Local Governments and Policy Makers:**

Improved price prediction models enhance property tax assessment accuracy and revenue forecasting. Predictive analytics could also be applied to assess the impact of zoning changes, infrastructure investments, or housing subsidies.

Recommendations:

1. ElasticNet for Generalization:

While Ridge showed slightly stronger stability in cross-validation, ElasticNet achieved a lower Kaggle RMSLE. We recommend ElasticNet as the preferred model for deployment, as it balances predictive accuracy with robustness on unseen data.

2. Feature Engineering Matters:

Features like total square footage, total bathrooms, remodeling indicators, and quality scores consistently improved performance. We recommend that housing datasets include these engineered features for practical applications.

3. Limitations:

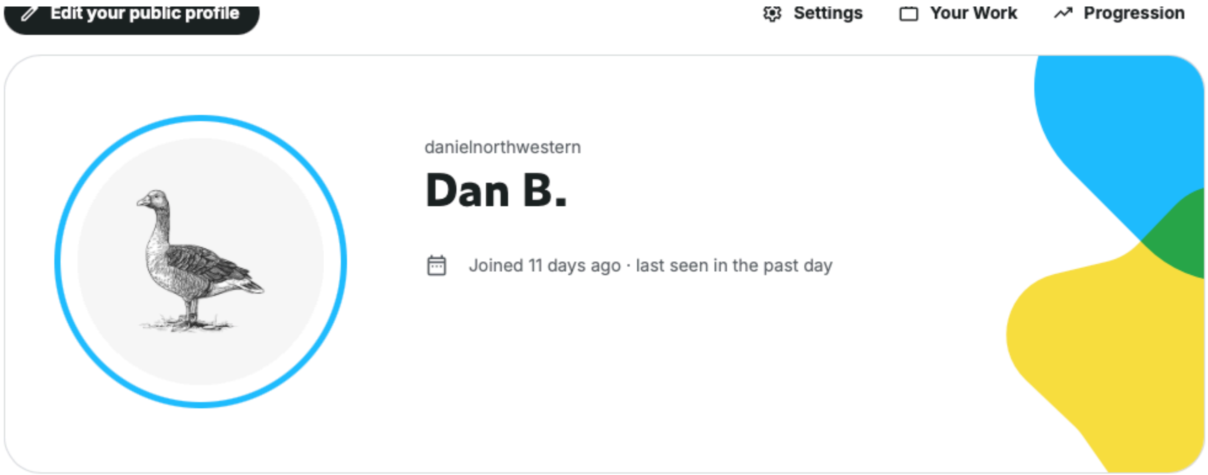
All models showed greater prediction uncertainty for high-priced homes. Stakeholders should treat predictions for the upper end of the housing market with caution and consider combining regression with non-linear approaches (e.g., gradient boosting) in future iterations.

Conclusion:

Accurate predictive modeling of housing prices benefits a broad set of stakeholders and provides a foundation for data-driven decision-making. Our results suggest that regularized regression models remain a strong baseline, with ElasticNet standing out as the most reliable method for balancing accuracy, interpretability, and generalization to unseen data.

Kaggle Results:

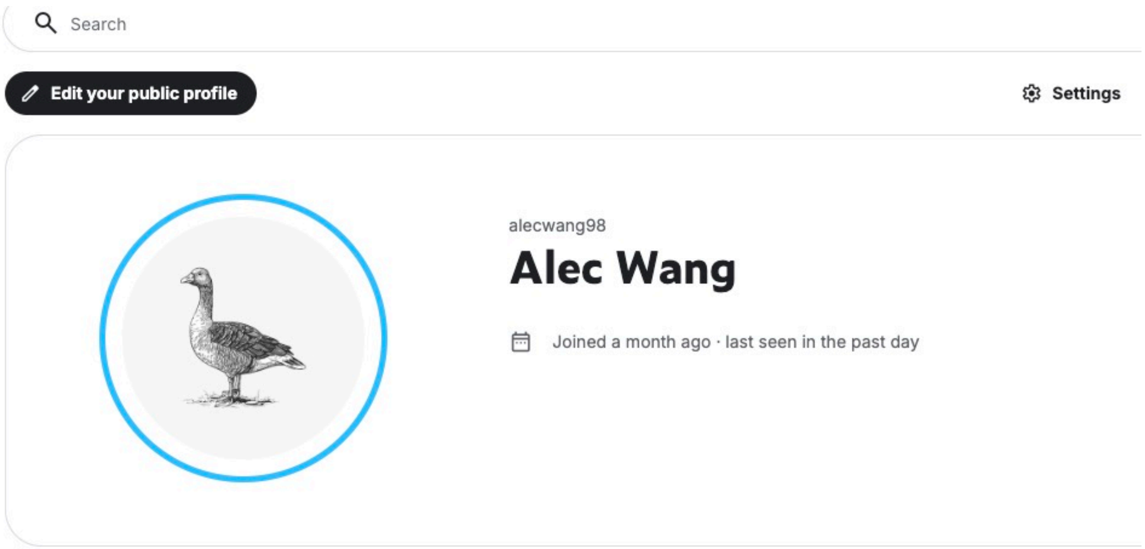
User Name: Danielnorthwestern (Daniel Balette) - **0.13472**
User Name: alecwang98 (Alec Wang) - **0.13771**



danielnorthwestern

Dan B.

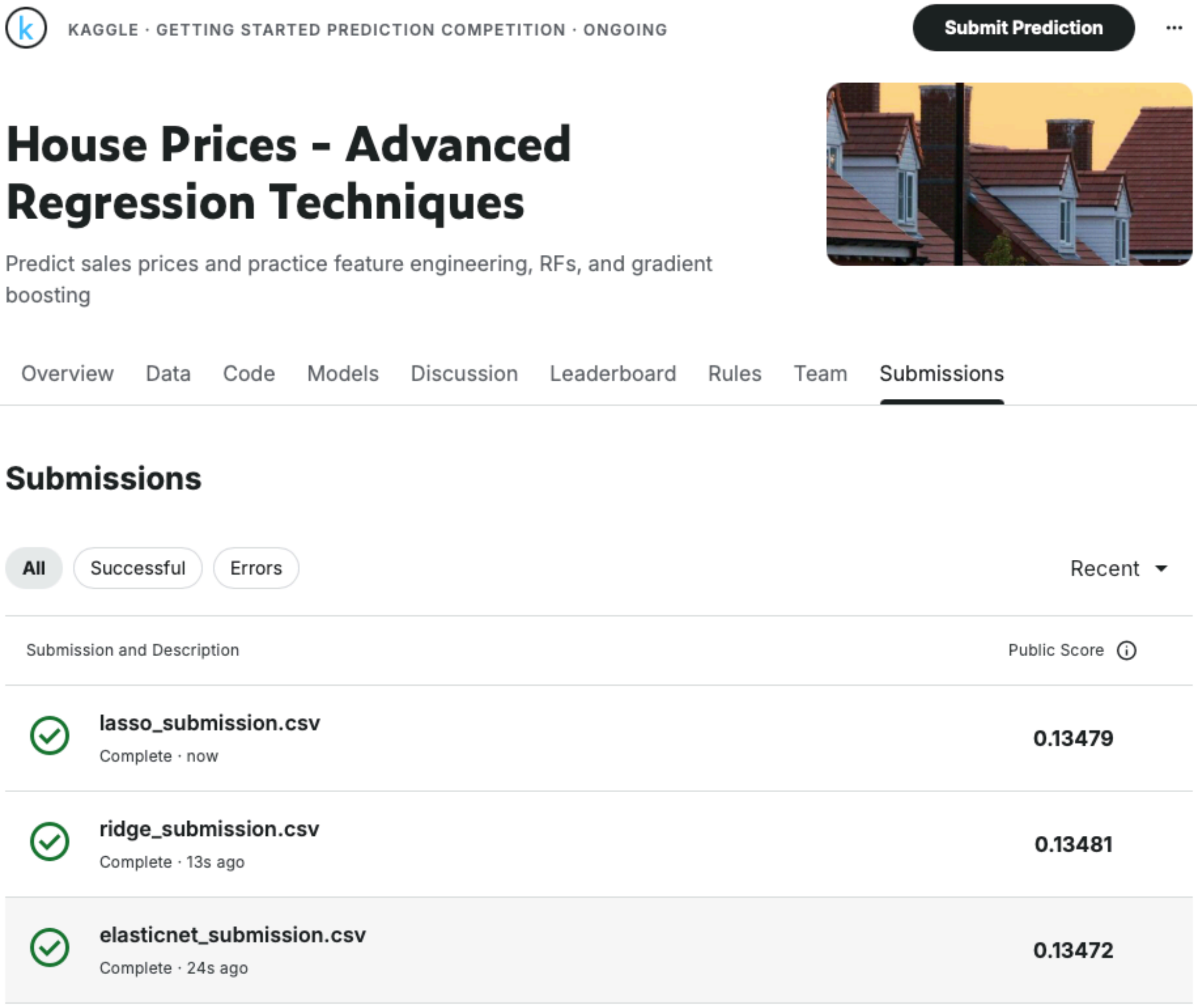
Joined 11 days ago · last seen in the past day



alecwang98

Alec Wang

Joined a month ago · last seen in the past day



KAGGLE · GETTING STARTED PREDICTION COMPETITION · ONGOING

House Prices - Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

Submissions

Submission and Description	Public Score
✓ lasso_submission.csv Complete · now	0.13479
✓ ridge_submission.csv Complete · 13s ago	0.13481
✓ elasticnet_submission.csv Complete · 24s ago	0.13472