Summary of Ames Housing Analysis Project

MSDS 7335 - Machine Learning II

May 27, 2025

Challenges and Solutions

The most significant challenge I encountered was with the neural network implementation, which initially performed extremely poorly with a negative R^2 value (-3.61) and high RMSE (187,990). This underperformance compared to traditional models highlighted the difficulty of applying deep learning to tabular data with limited samples. I addressed this by implementing several key optimizations: adding batch normalization layers to reduce internal covariate shift, normalizing the target variable to stabilize training, implementing an adaptive learning rate scheduler, and redesigning the architecture with appropriate regularization. These modifications dramatically improved performance, bringing the neural network's R^2 to 0.8849, making it competitive with the other models.

Another challenge was ensuring proper preprocessing across different model types. Each model had different sensitivities to missing values, feature scaling, and categorical encoding. I developed a robust preprocessing pipeline using scikit-learn's ColumnTransformer to apply appropriate transformations to numerical and categorical features, ensuring consistency across all models while respecting their unique requirements.

Learning Moments

This project significantly deepened my understanding of regularization techniques in machine learning. Working with both Ridge (L2) and Lasso (L1) regression allowed me to observe how different penalty approaches affect model coefficients and feature selection. The clear visualization of lambda's impact on model performance through cross-validation created an intuitive understanding of the bias-variance tradeoff that theoretical discussions hadn't fully conveyed.

I also gained valuable experience in neural network optimization for structured data. Prior to this project, my neural network experience was primarily with image and text data. Learning

to properly normalize features and targets, implement batch normalization, and use learning rate scheduling specifically for tabular data expanded my deep learning toolkit considerably.

Key Accomplishment

My proudest achievement was developing a comprehensive modeling approach that effectively balanced performance, interpretability, and implementation complexity. The final analysis demonstrated that:

1. Lasso Regression achieved the best overall performance (R^2: 0.8960) while providing valuable feature selection
2. Random Forest captured important non-linear relationships (R^2: 0.8901)
3. The optimized Neural Network performed competitively (R^2: 0.8849) after addressing initial challenges
4. Ridge Regression provided a solid baseline (R^2: 0.8755) with stable coefficients

This comparative framework provides not just performance metrics but actionable insights for model selection based on specific application requirements. The small performance differences between models ($\sim$1.1% variation in R^2) with their distinct characteristics offers a nuanced perspective on model selection that goes beyond simply choosing the "best" performer.