🏠 Predicting House Prices

Anisha Prashanth, Aryan Dhingra, Arjun Gajula, Harish Kanthi, Rishi Magiawala

Introduction and Background

Literature Review

- Housing price indices are crucial to stakeholders including real estate professionals, policymakers, and homebuyers. The development of predictive ML models can help forecast future prices and guide policy decisions [1].
- A study found that macroeconomic uncertainty can be a useful predictor in understanding how national and regional housing markets evolve [2]. Interestingly, the Kaggle dataset we are using only has houses sold between 2006-2010. According to [1], 2006 was the peak of the housing market in that timeframe, and the market bottomed out in 2011 after the housing market crash. It would be interesting to see if that biases the model.
- Study [3] focused on the same Kaggle dataset - it found that using a hybrid model - combining well-performing individual models - produced the most accurate predictions. MSE was used as the performance metric.
- Another study ran ML algorithms including XGBoost, CatBoost, Random Forest, Lasso, and Voting Regressor to predict house prices in Volusia County, FL. Performance metrics included MSE, MAE, and computational time. They found XGBoost to be the best-performing overall [4].

Dataset Description

The dataset contains 79 numeric and categorical features that describe various aspects of 1460 homes in Ames, Iowa. The target variable is SalePrice (price each home sold for). Kaggle provides training and test data.

Dataset Link

Dataset

Problem Definition

Problem

Predict the final sale price of homes based on 79 features. Complexity comes from a diverse set of features influencing house prices, including location, size, and overall condition of property.

Motivation

Motivation is the broader economic benefits of accurate predictive models in the real estate industry [1]. If accurate, our model has applications in financial planning, real estate investments, and market trend analysis. Buyers and sellers could make better decisions, agents could provide more accurate advice, and banks could evaluate loans with higher accuracy.

Methods

Implemented Data Preprocessing Methods

- **Missing Data Handling**:
    - Numerical: We replaced N/A values with the median calculated within a related category (e.g., for missing values in lotSize, we used the median lotSize specific to each lotType). unlike mean, median is less influenced by outliers
    - Categorical: We replaced N/A values with the mode of a related column, applying a similar approach as with numerical data.
    - NA: In certain cases (e.g., Garage), "NA" indicates a missing category like "No garage" as specified in the data description, rather than missing data.
- **Encoding Features**:
    - We made categorical variables compatible with our ML models by applying one-hot encoding or label encoding (e.g., for Neighborhood).
- **Scaling**:
    - We normalized numerical data using StandardScaler before applying models sensitive to feature scales, such as Linear Regression.

- **Feature Reduction**:
    - P-Value:
        - For numerical features, we used Pearson correlation p-values with SalesPrice and removed features with p-values > 0.01.
        - For categorical features, we used ANOVA p-values with SalesPrice and removed those with p-values > 0.01.
    - PCA algorithm: Applied the PCA Alogrithm on all features.

ML Algorithms/Models

- **Linear Regression**: good baseline due to simplicity
- **Random Forest**: Effective for datasets with many features, as it helps reduce overfitting.
- **Neural Network**: Useful for exploring complex datasets and handling a large number of features.

Results and Discussion

Metric	Linear Regression	Random Forest	Neural Network
RMSE	36,032.174	31,484.796	38,445.834
R-Squared	0.795	0.843	0.767
MAE	21,646.056	18,541.535	None
MAPE	0.132	0.108	0.132

Overall Conclusion

- Regarding RMSE, Random Forest was the best performing model, with the lowest RMSE value of $31,500
- Random Forest also boasted the best R^2 value, showing the highest correlation factor of 0.843
- Its robustness and ability to handle complex relationships in tabular data make it the most suitable model for this task. Future steps include fine-tuning hyperparameters and exploring hybrid/ensemble models for further improvement.

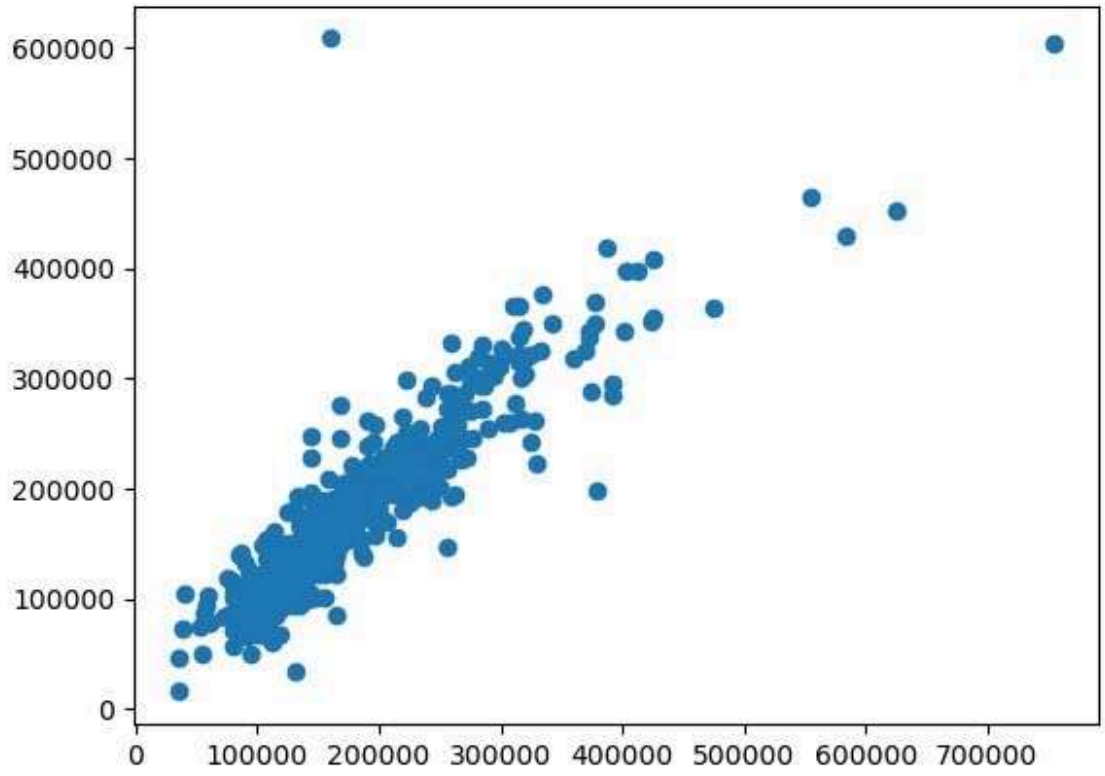
Linear Regression

Analysis

- We expected our RMSE value to be around 25,000 dollars (10-15% of the average house price) to show that our model fits the dataset well and minimizes error, however, we resulted in an RMSE value of 36,000 dollars, which is slightly worse than expected.
- The R2 value of 0.79 was between 0.75 and 1, which illustrates a strong, statistically significant linear relationship.
- Linear Regression succeeds in showing simplistic linear relationships, which may be the explanation for the performance of our model.
- Random Forest and Gradient Boosting algorithms will decipher non-linear relationships within our dataset and improve our prediction accuracy in upcoming iterations of our model, and may allow the metrics to score higher.
- Neural Network model could find complicated relationships between features.

Discussion and Next Steps

- Regarding feature selection, PCA and p-value methods were used. While running the linear regression model, results from the PCA feature-reduced dataset were discarded as the RMSE value was ~ 20,000 dollars worse than the p-value data. It may be interesting to consider how regularization could help these feature reduction methods, reducing multicollinearity if that was an issue (this is common among housing datasets, as multiple features are correlated with each other).- One of the largest reasons for a lack of excellent performance within our model may be the existence of non-linear relationships between the features and the target variable, which cannot be modeled by first-order linear regression models.
- Our linear regression model is highly interpretable as a positive or negative coefficient with each feature can be interpreted as a positive or negative influence on the final sales price. More complex, non-linear models like Random Forest, and famously, neural networks, are not explainable. In some applications, the slight increase in performance may not be worth the tradeoff in lack of explainability.
    - Mathematics research has shown that the black box of neural networks consists of a list of linear models, which provides an interesting rebuttal to the above point. [5]



Plot of y_predicted (linear regression) vs. y_actual (sales price of homes)

Random Forest

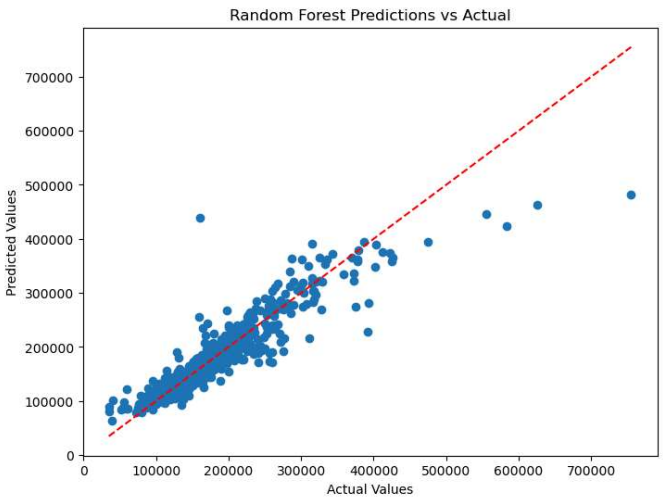
Analysis

- We expected our RMSE value to be around $25,000 (10-15% of the average house price) to show that our model fits the dataset well and minimizes error, however, we resulted in an RMSE value of 31,500 dollars, which is worse than expected.
- The R2 value of 0.843 was between 0.75 and 1, which illustrates a strong, statistically significant linear relationship.
- Linear Regression succeeded in showing simplistic linear relationships.
- The Random Forest algorithm will decipher non-linear relationships within our dataset and improve our prediction accuracy in upcoming iterations of our model, and allow the metrics to score higher.

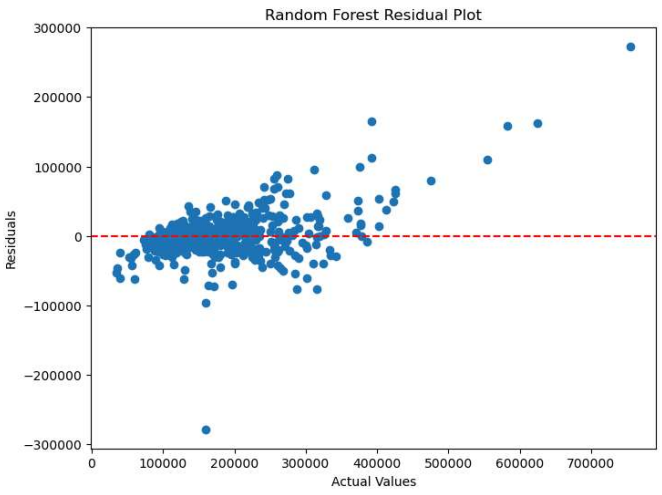
Discussion and Next Steps

- For random forest, hyperparameters like the number of estimators, maximum depth, the number of samples in a split and the max # features could be optimized using grid search or a randomized search.
- Beyond just hyperparameter tuning, the features could be engineered with specific attention paid to the results of random forest.

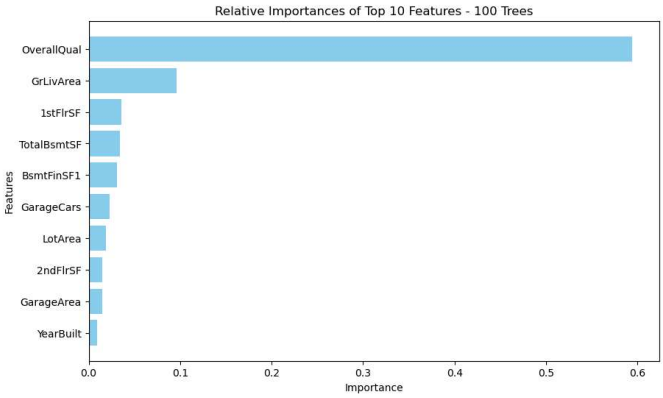
- Introducing interaction terms between main effects of the experiment could reveal more complex interactions between variables, although this would require more complex analysis.



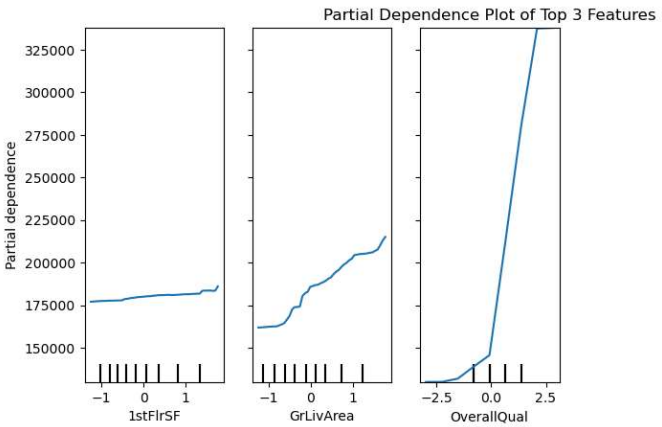
Prediction v Actual



Residual Plot



Top 10 Features



Top 3 Partial Dependence Plot

Neural Network

Analysis

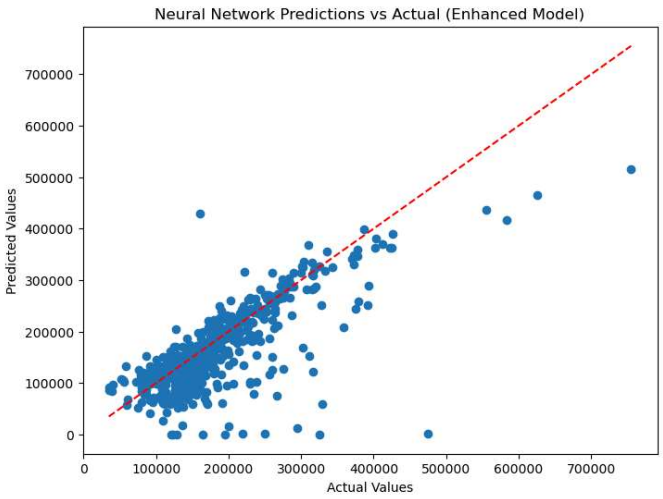
- We expected our RMSE value to be around $25,000 (10-15% of the average house price) to show that our model fits the dataset well and minimizes error, however, we resulted in an RMSE value of 38,500 dollars, which is worse than expected.
- The R2 value of 0.767 was between 0.75 and 1, which illustrates a strong, statistically significant linear relationship.
- The Neural Network model was implemented to find complicated relationships between features, but may have missed simple features
- We suspect that the higher RMSE compared to previous models indicates that the NN did not do well in the finer accuracy. The issue with the NN is that it´s end result cannot be adjusted very easily due

to its process being a black box. The training must occur again and the weights must be set in a different manner.

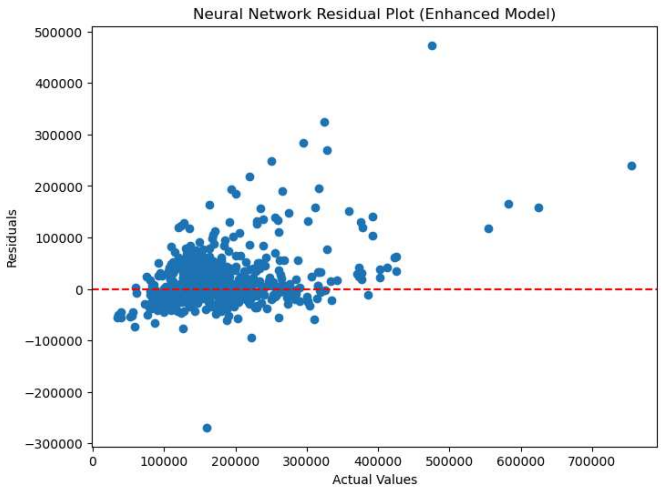
- Neural networks excel with capturing non-linear relationships and complex interactions within the data. They can also scale to incorporate more data/features without changing the codebase much.
- The NN here may have overfit here if the neurons learned the noise and other irrelevant details along with the patterns in the features. This may have occurred due to the excessive number of parameters. Data quality or lack of data is not an issue due to the high quality and large size of the data used for the project.

Discussion and Next Steps

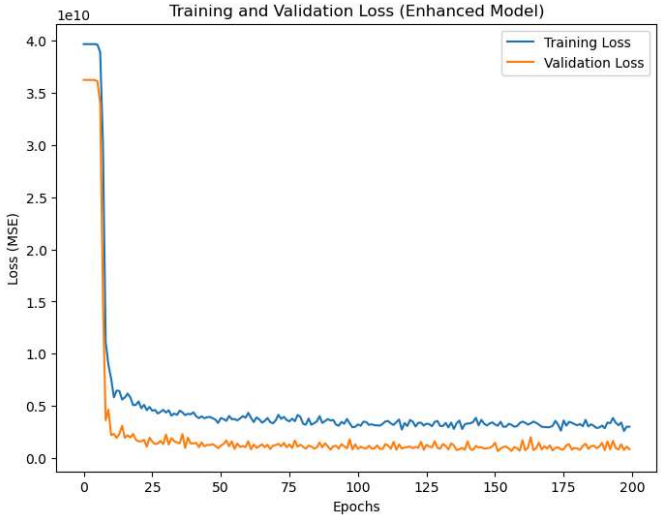
- Similarly to random forest, hyperparameter optimization is required here. Parameters could include the number of hidden layers, the number of neurons per layer, the learning rate, and more. The method to tune the hyperparameters could be the same as RF/GB, which are grid search or Bayesian optimization.
- Sometimes, the inputs to the neurons result in better model performance when the features are transformed (using log or polynomial transformations). This can help the model understand non-linear relationships.
- Another solution to the slightly lackluster performance from the NN is to combine or ensemble it with other methods like random forest or LR in an attempt to capture the NN's nonlinear power and LR's linear power.
- If a basic neural network is not capturing the complexity of the data, we could use something like residual networks, which would allow us to skip several connections and mitigate decreasing gradients. If this shows to improve training and produces lower validation loss, it could be a potential solution



Prediction v Actual



Residual Plot



Loss Curve

Gantt Chart

Gantt Chart

Contributions Table

Name	Contribution
Anisha Prashanth	Random Forest, Streamlit
Aryan Dhingra	Random Forest, Github and Readme
Harish Kanthi	Slides, Presentation, Results & Discussion
Arjun Gajula	Neural Network
Rishi Magiawala	Presentation & Neural Network

References

[1] “Bem-vindo(a),” Sciencedirect.com, 2024. https://www.sciencedirect.com/science/article/abs/pii/S0957417414007325.

[2] R. Gupta, H. A. Marfatia, C. Pierdzioch, and A. A. Salisu, “Machine Learning Predictions of Housing Market Synchronization across US States: The Role of Uncertainty,” The Journal of Real Estate Finance and Economics, Jan. 2021, doi: https://doi.org/10.1007/s11146-020-09813-1.

[3] X. Zhou, “Comparative Analysis of Machine Learning Performance in House Price Prediction for Ames Iowa,” Highlights in Science, Engineering and Technology, vol. 39, pp. 738–743, Apr. 2023, doi: https://doi.org/10.54097/hset.v39i.6638.

[4] S. B. Jha, R. F. Babiceanu, V. Pandey, and R. K. Jha, “Housing Market Prediction Problem using Different Machine Learning Algorithms: A Case Study,” arXiv:2006.10092 [cs, stat], Jun. 2020, Available: https://arxiv.org/abs/2006.10092.

[5] Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. Neural Networks, 2(5), 359–366. https://www.sciencedirect.com/science/article/abs/pii/0893608089900208?via%3Dihub