Predicting Housing Sale Prices

A Machine Learning Approach

Andrea Pena
03/06/2025

Why predict home prices?

Why predict home prices?

- Buyers
- Sellers
- Investors
- Real estate agents

Why predict home prices?

- Buyers
- Sellers
- Investors
- Real estate agents

Image Source: Topp_Yimgrimm/Getty Images

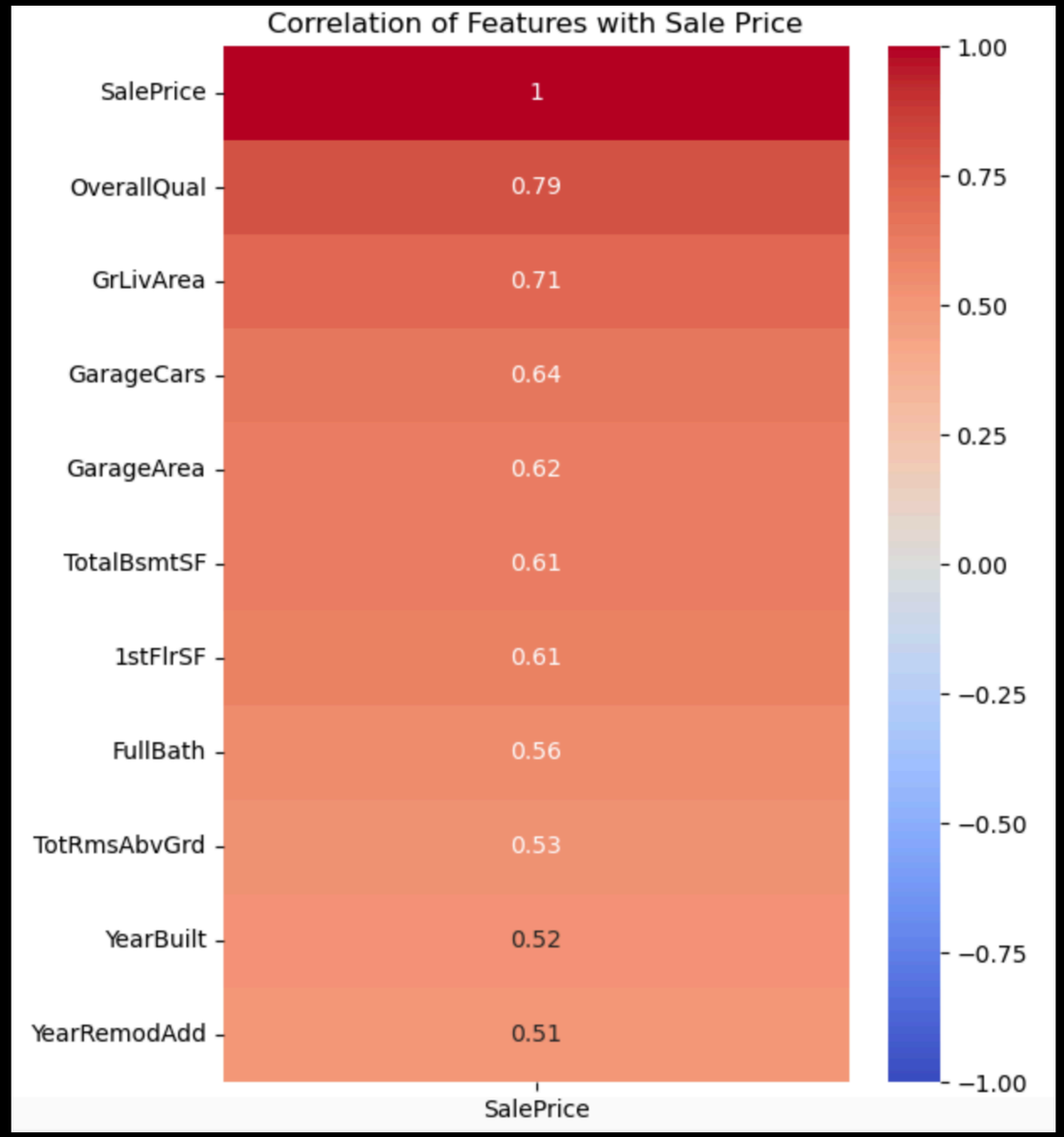
Data Overview

Data Overview

- Dataset source: Kaggle
- Size: 1,460 rows, 80 columns
- Data Cleaning
- Exploratory Data Analysis
- Preprocessing

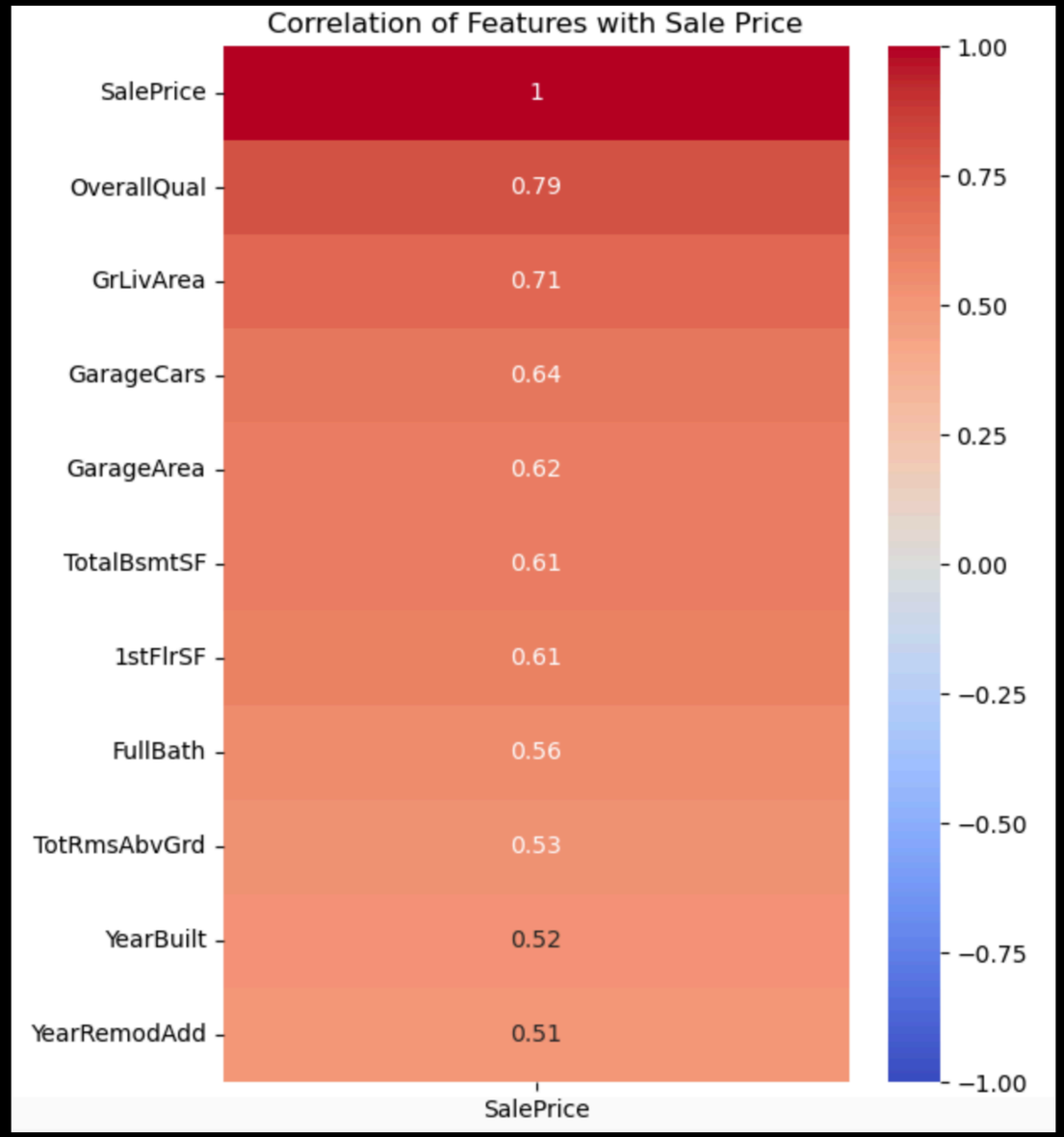
Data Overview

- Dataset source: Kaggle
- Size: 1,460 rows, 80 columns
- Data Cleaning
- Exploratory Data Analysis
- Preprocessing



Data Overview

- Dataset source: Kaggle
- Size: 1,460 rows, 80 columns
- Data Cleaning
- Exploratory Data Analysis
- Preprocessing



Since the dataset was small, all missing values were imputed to median for numerical values and mode for categorical. In EDA, histograms, scatterplots and bar plots were used for visualization. There was a one-hot encoded to prepare for modeling.

Model Selection

Linear Models Tested

Model Selection

Linear Models Tested

- Ridge
- OLS
- Lasso

Model Selection

Linear Models Tested

- Ridge
- OLS
- Lasso

Lasso Residual Plot

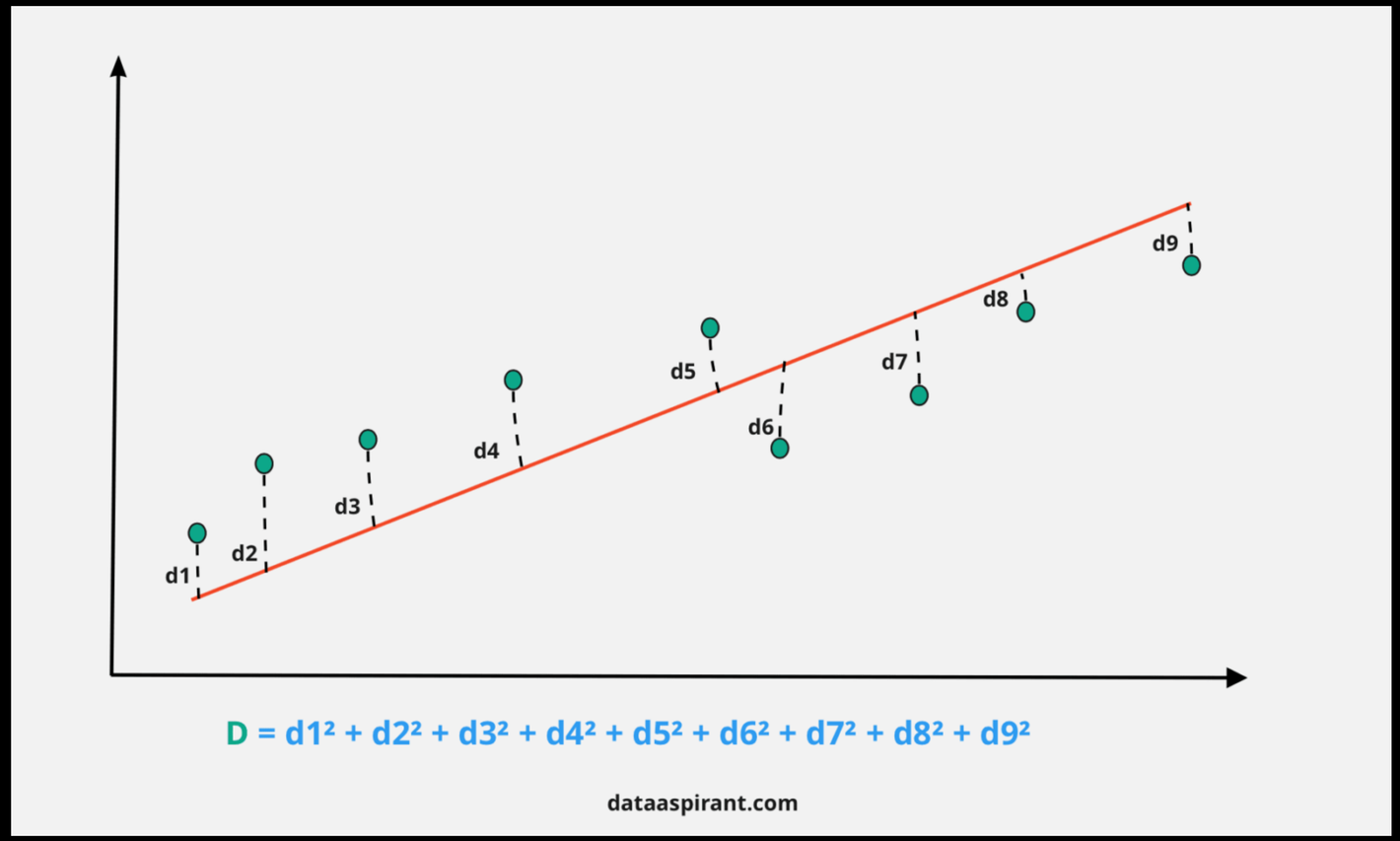


Image Source: dataaspirant.com

Model Selection

Linear Models Tested

- Ridge
- OLS
- Lasso

Lasso Residual Plot

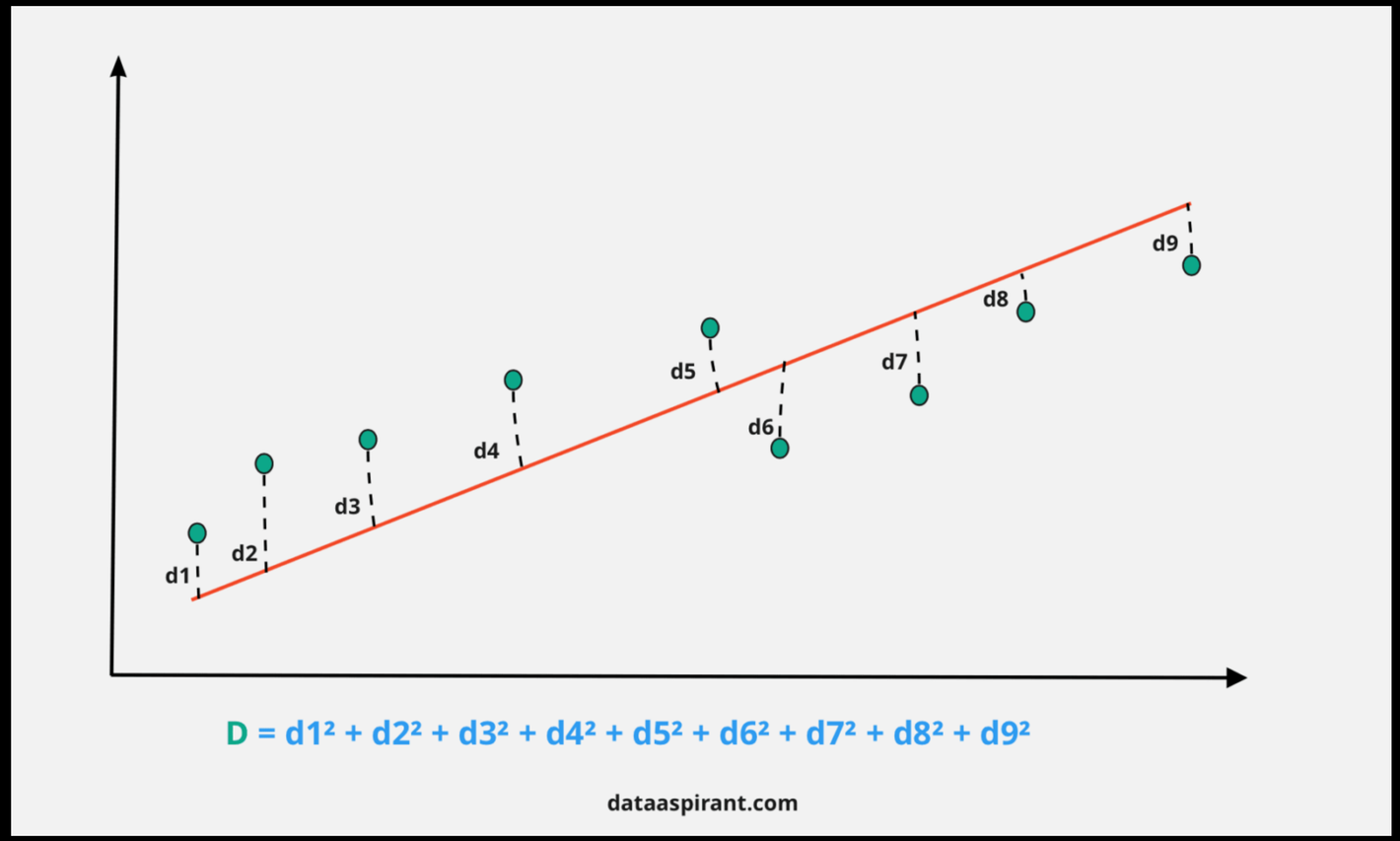


Image Source: dataaspirant.com

Three linear models were evaluated, with Lasso having the lowest score. This may be because Lasso performs well when dealing with many features by applying regularization, which helps prevent overfitting.

Model Selection

Nonlinear models

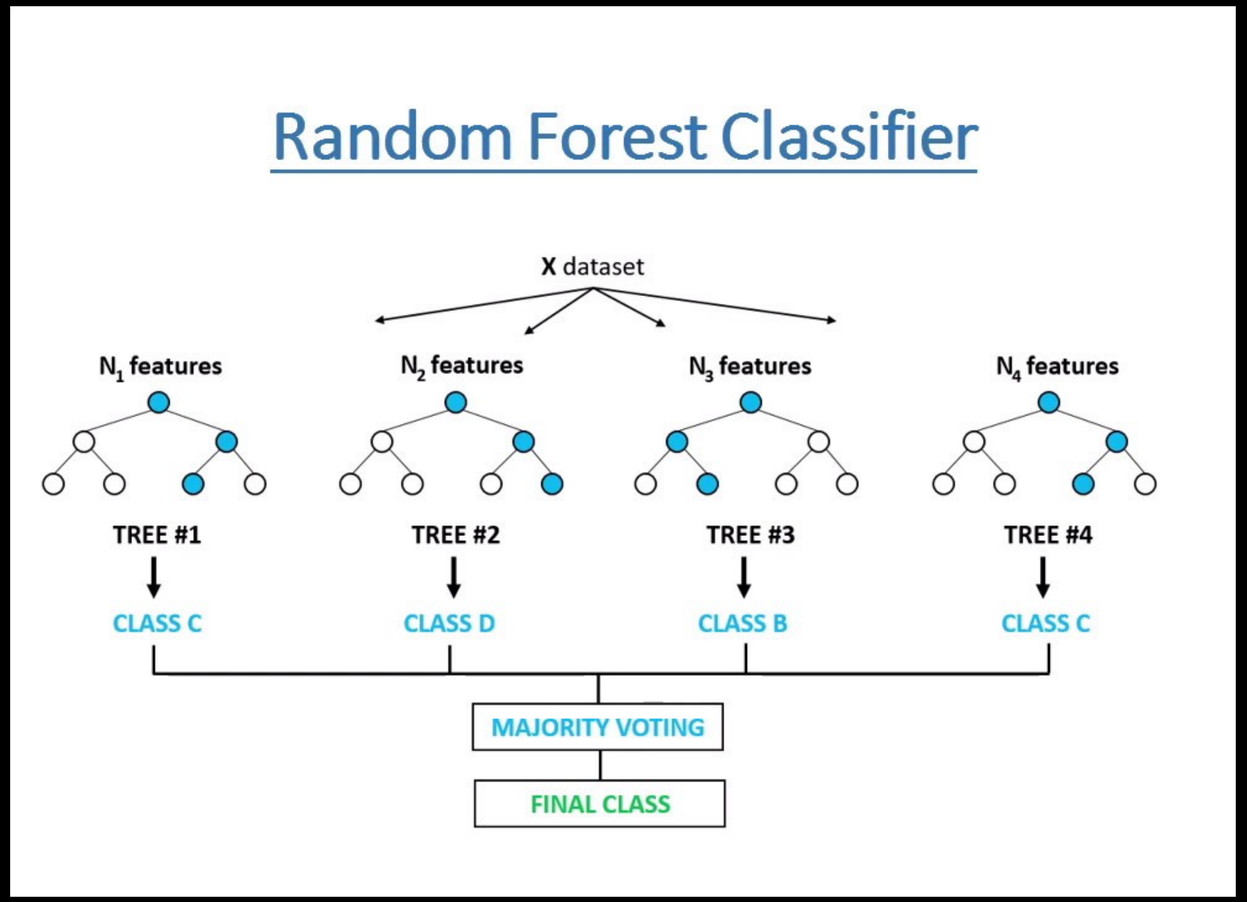


Image Source: Kaggle (Prashant Banerjee)

Model Selection

Nonlinear models

- Random Forest
- Gradient Boosting
- SVR

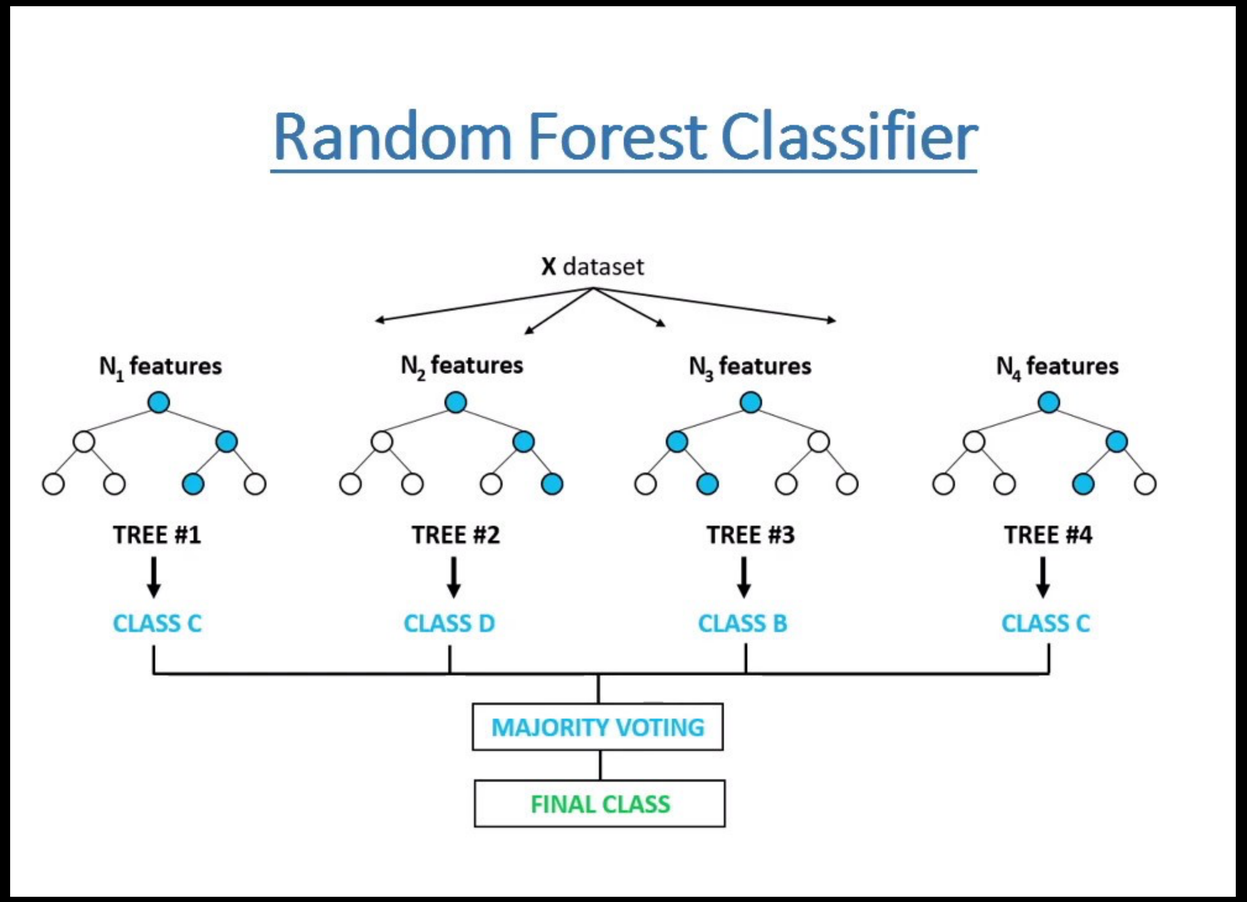


Image Source: Kaggle (Prashant Banerjee)

Model Selection

Nonlinear models

- Random Forest
- Gradient Boosting
- SVR

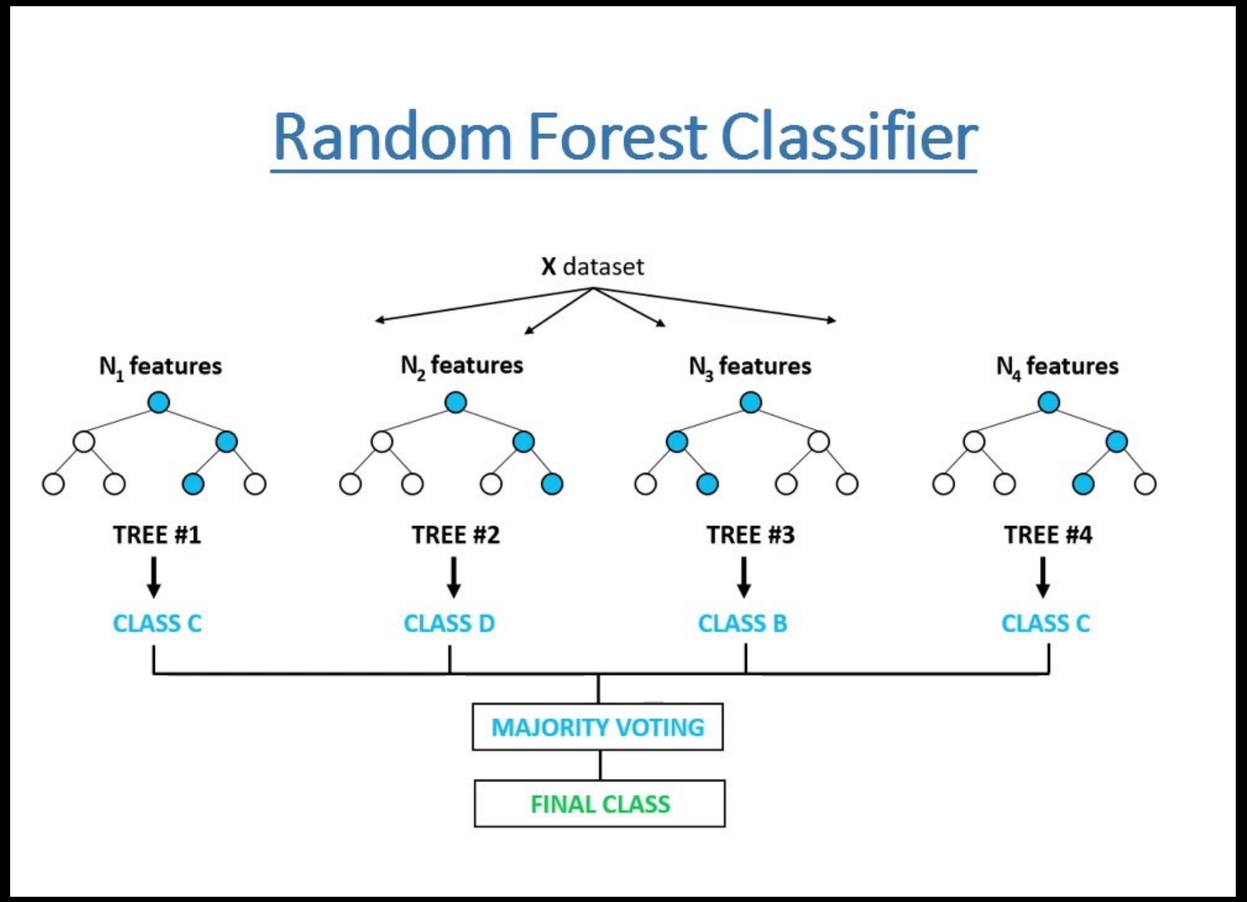


Image Source: Kaggle (Prashant Banerjee)

Model Selection

Nonlinear models

- Random Forest
- Gradient Boosting
- SVR

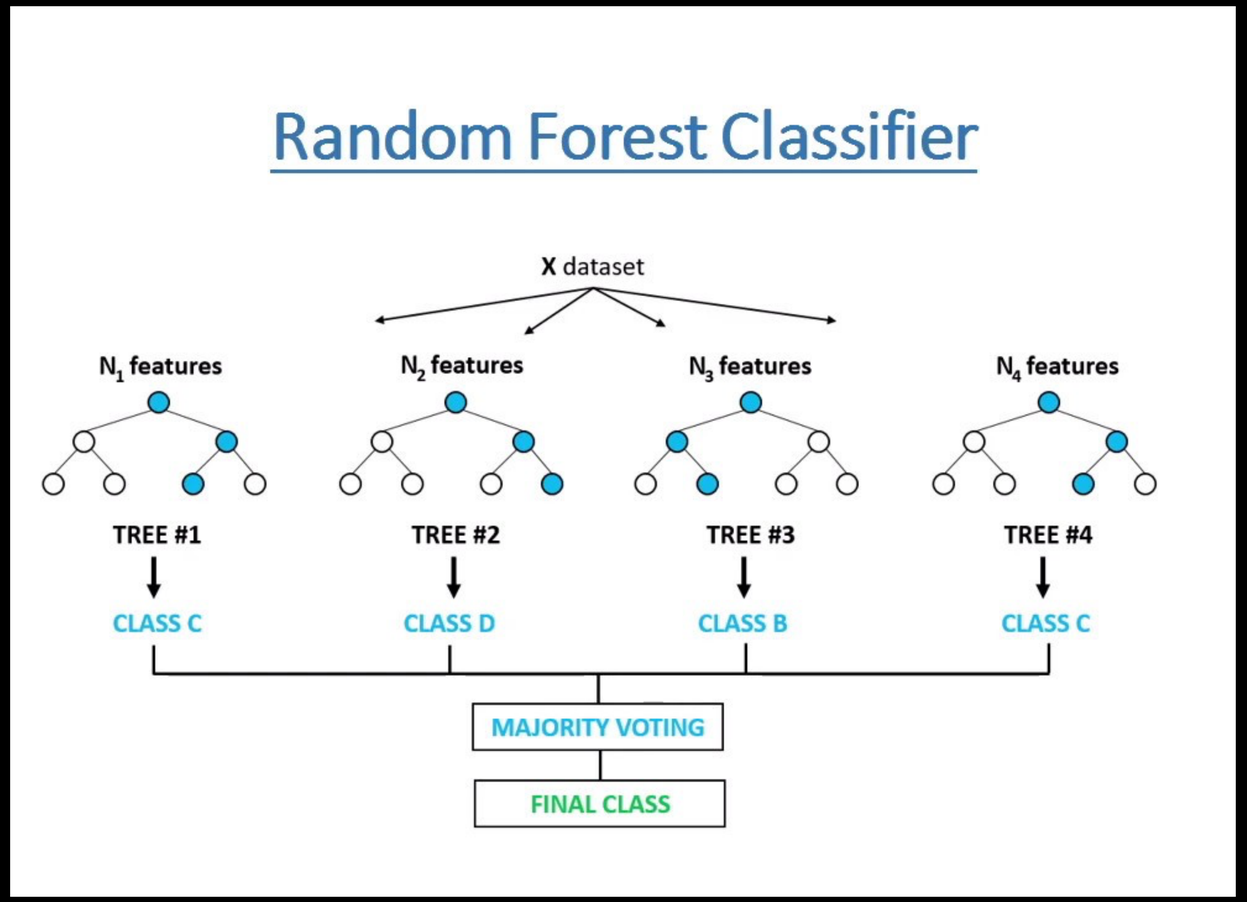
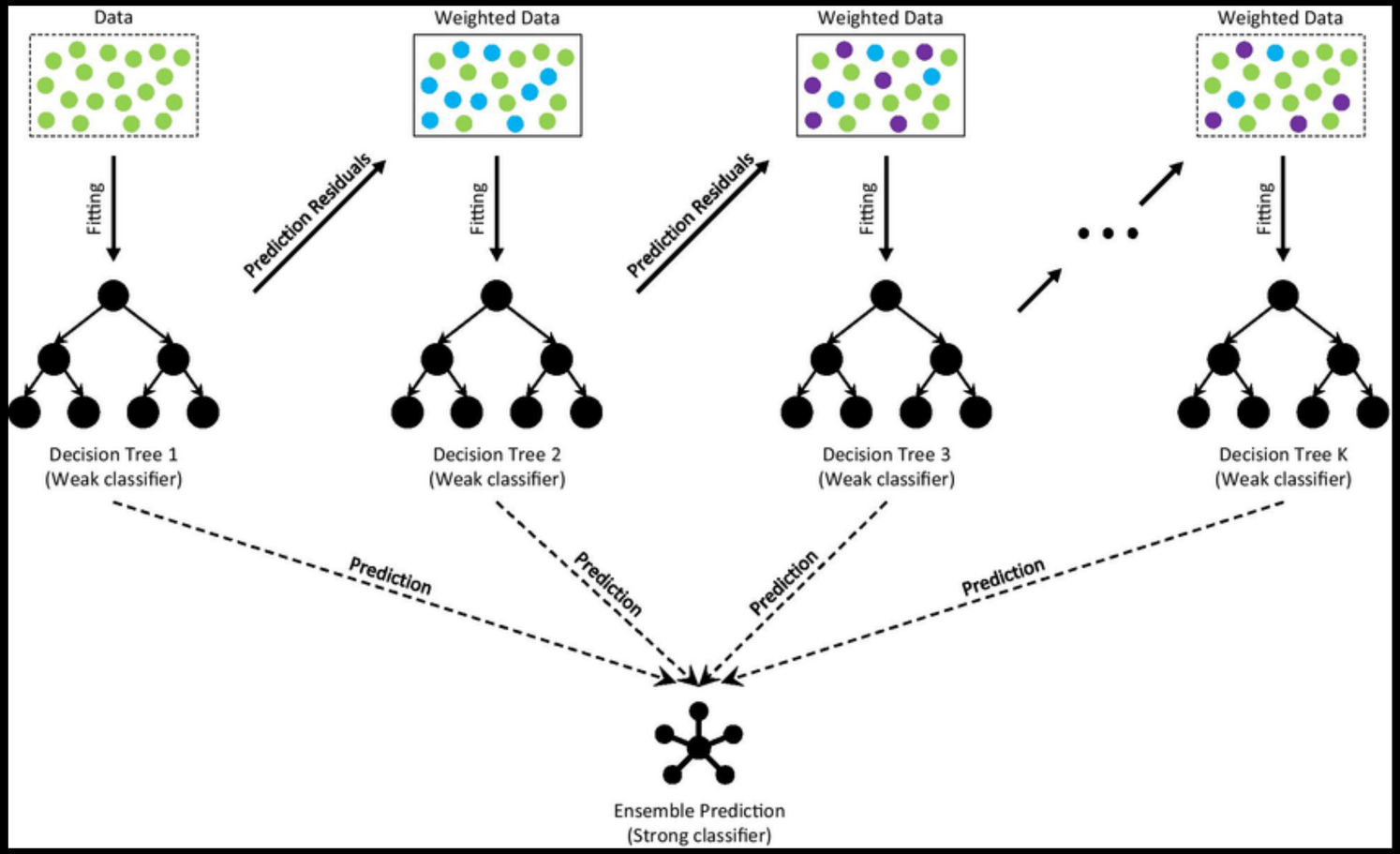


Image Source: Kaggle (Prashant Banerjee)

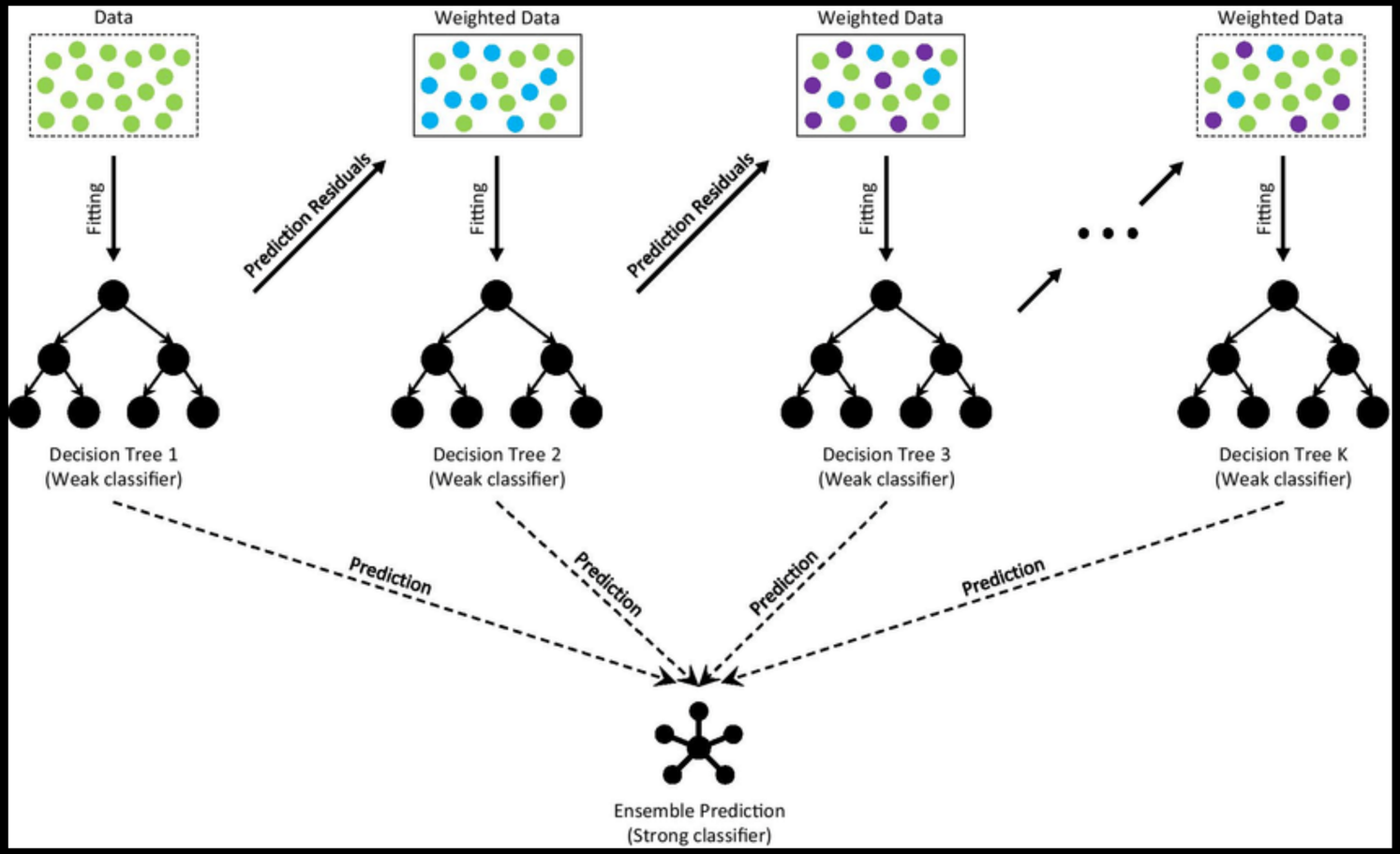
The Random Forest (RF) model was compared to Lasso and performed better, likely because RF can handle complex features and the dataset exhibits nonlinearity. RF was then compared to other non-linear models, and Gradient Boosting (GB) achieved a lower error score.

Winner: Gradient Boosting!



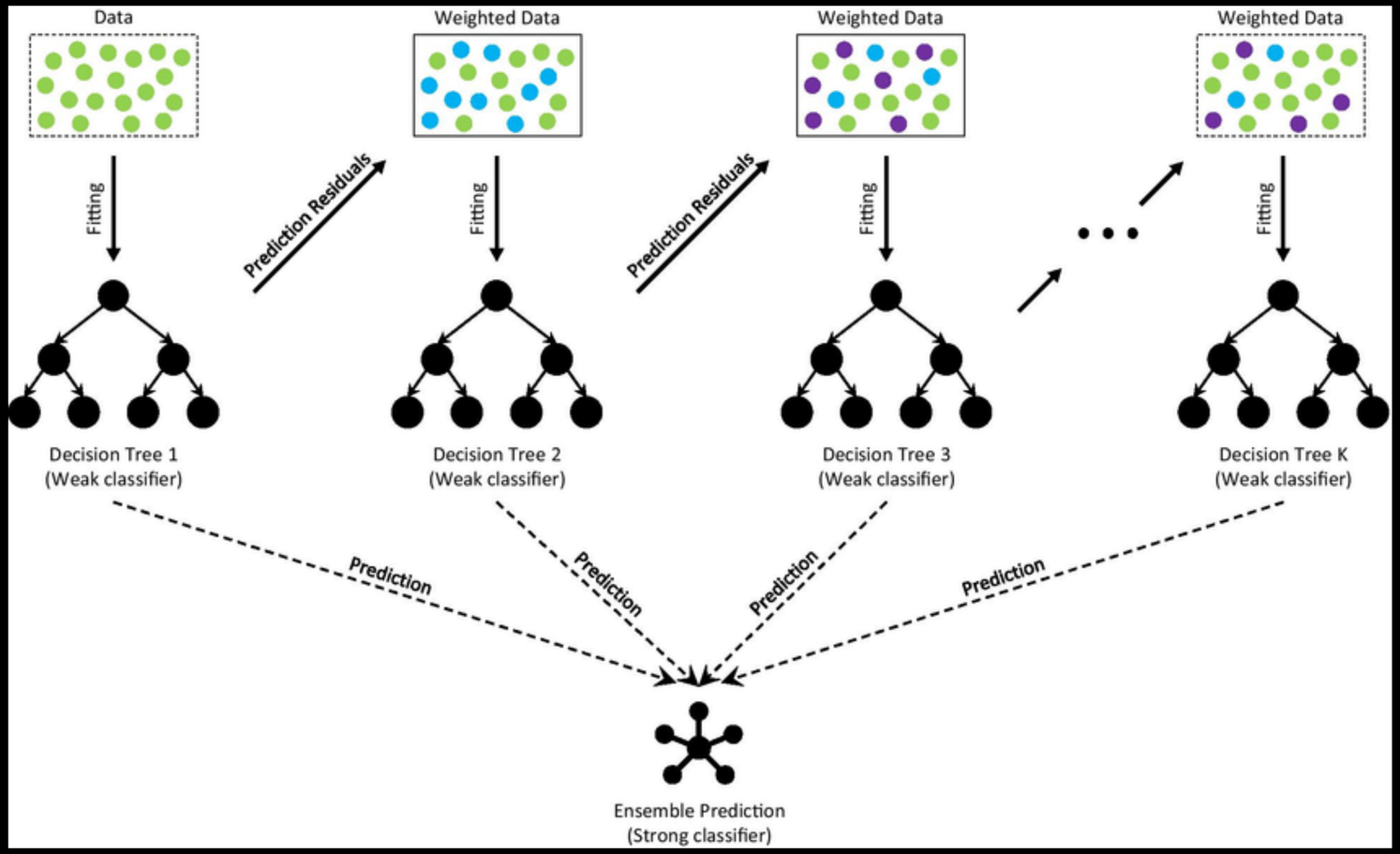
Ensemble learning for the early prediction of neonatal jaundice with genetic features - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/The-architecture-of-Gradient-Boosting-Decision-Tree_fig2_356698772 [accessed 5 Mar 2025]

Winner: Gradient Boosting!



Ensemble learning for the early prediction of neonatal jaundice with genetic features - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/The-architecture-of-Gradient-Boosting-Decision-Tree_fig2_356698772 [accessed 5 Mar 2025]

Winner: Gradient Boosting!



Ensemble learning for the early prediction of neonatal jaundice with genetic features - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/The-architecture-of-Gradient-Boosting-Decision-Tree_fig2_356698772 [accessed 5 Mar 2025]

GB model corrects mistakes over many iterations.

Model Performance

Model Performance

Model	Train	Test
Lasso Regression	--	34, 262.05
Random Forest	--	27,245.97
Tuned RF	11,054.71	27,173.45
Modified Tuned RF	26,085.00	28,653.73
Gradient Boosting	4,143.35	25,844.64
Tuned GB	8,782.17	24, 327.50
Tuned GB	17, 055.26	25,405.89
Tuned GB	**20,307.08**	**25,492.86**
GB(with outlier removal)		26,930.21

Model Performance

Model	Train	Test
Lasso Regression	--	34, 262.05
Random Forest	--	27,245.97
Tuned RF	11,054.71	27,173.45
Modified Tuned RF	26,085.00	28,653.73
Gradient Boosting	4,143.35	25,844.64
Tuned GB	8,782.17	24, 327.50
Tuned GB	17, 055.26	25,405.89
Tuned GB	**20,307.08**	**25,492.86**
GB(with outlier removal)		26,930.21

This figure shows the train and test scores of each model. After RF outperformed the linear models, hyperparameter tuning was performed, improving its performance. However, after GB achieved a better score, it was further optimized, ultimately yielding the lowest test error

Key Features Impacting Price

Top Predictors

Key Features Impacting Price

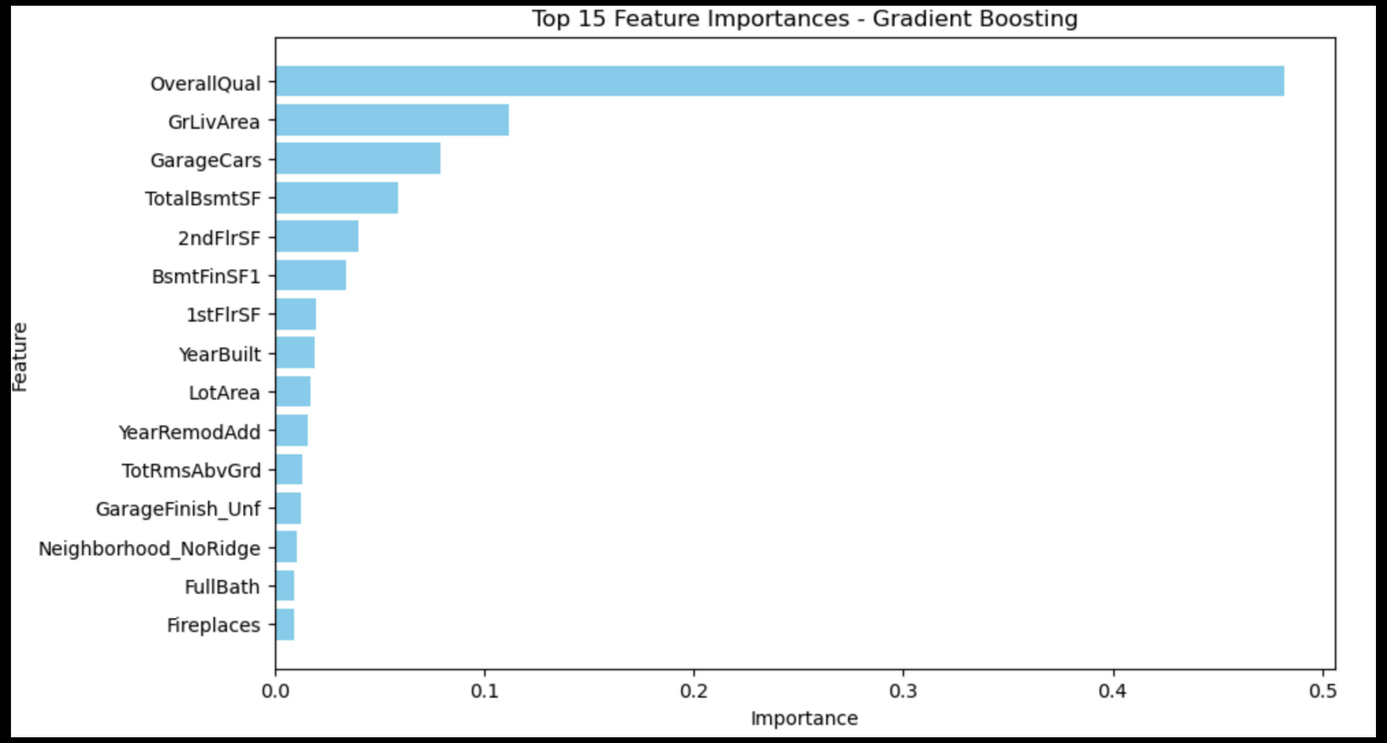
Top Predictors

- Overall Quality
- Ground Living Area
- Garage Cars
- Total Basement (sq footage)
- Second Floor (sq footage)

Key Features Impacting Price

Top Predictors

- Overall Quality
- Ground Living Area
- Garage Cars
- Total Basement (sq footage)
- Second Floor (sq footage)



Business Impact

Why does this matter?

Image Source: iStock(GelatoPlus)

Image Source:iStock(iraanamwong)

Business Impact

Why does this matter?

Image Source: iStock(GelatoPlus)

Image Source:iStock(iraanamwong)

Business Impact

Why does this matter?

Image Source: iStock(GelatoPlus)

Image Source:iStock(iraanamwong)

Business Impact

Why does this matter?

Image Source: iStock(GelatoPlus)

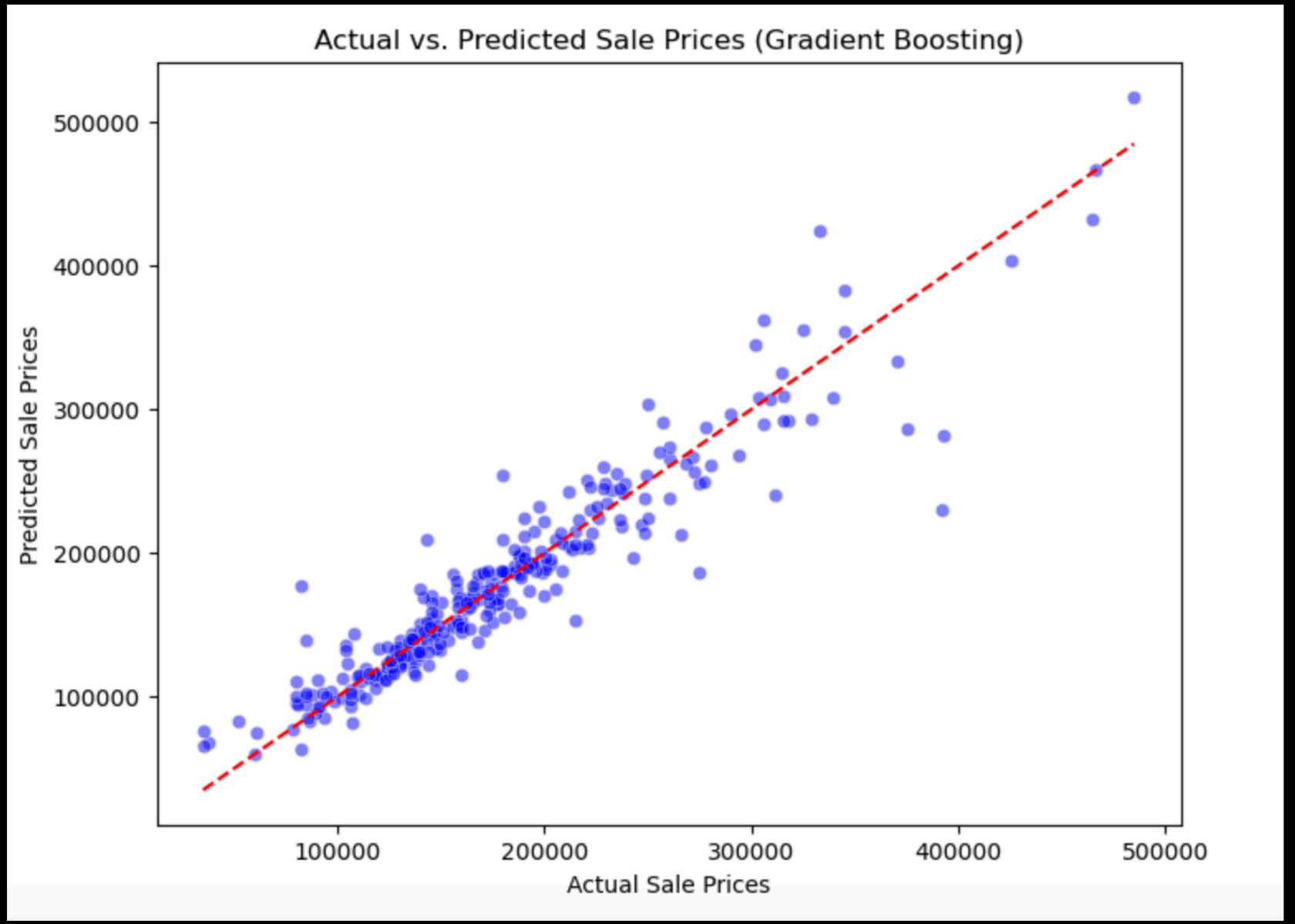
Image Source:iStock(iraanamwong)

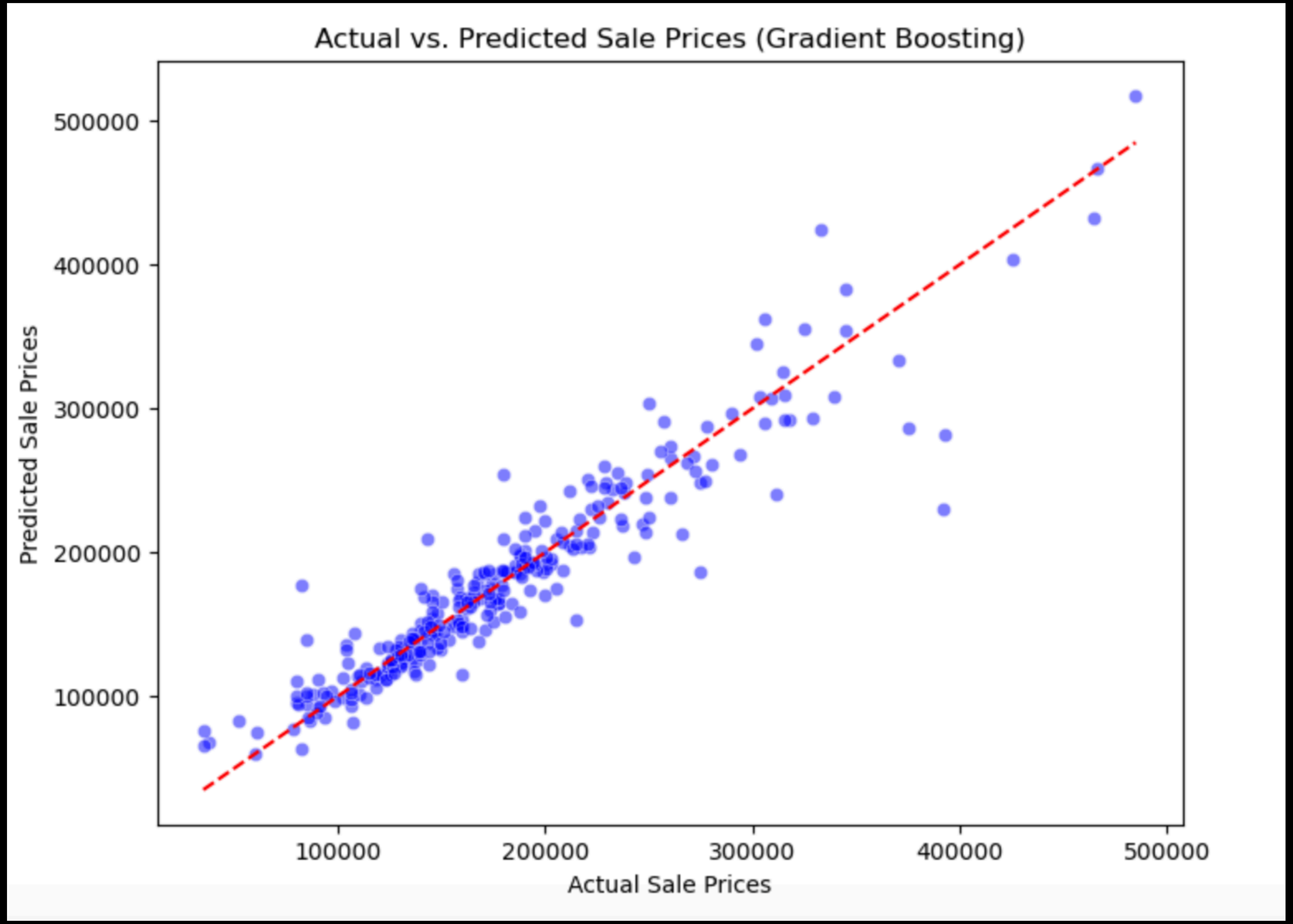
Sellers can use model insights to set competitive prices, while buyers ensure they aren't overpaying. For real estate agents, understanding key property features that drive sale prices can help guide clients in their buying and selling decisions. Lastly, identifying undervalued properties based on predictive modeling can help investors make data-driven purchasing decisions to maximize returns.

Conclusion

Key Takeaways:

- Final Model: Gradient Boosting
- Predicting home prices is possible with strong feature selection.
- Property characteristics (like size, quality and garage space) matter most.


Actual vs. Predicted Sale Prices (Gradient Boosting)
Predicted Sale Prices
Actual Sale Prices
100000
200000
300000
400000
500000



Most predictions align well, but some high-value homes are harder to predict.

Future Steps

- Fine-tuning hyperparameters.
- Trying different ensemble models.
- Testing additional datasets.
- Incorporating real-time market trends.

Image source: depositphotos (lemono)

Future Steps

- Fine-tuning hyperparameters.
- Trying different ensemble models.
- Testing additional datasets.
- Incorporating real-time market trends.

Image source: depositphotos (lemono)