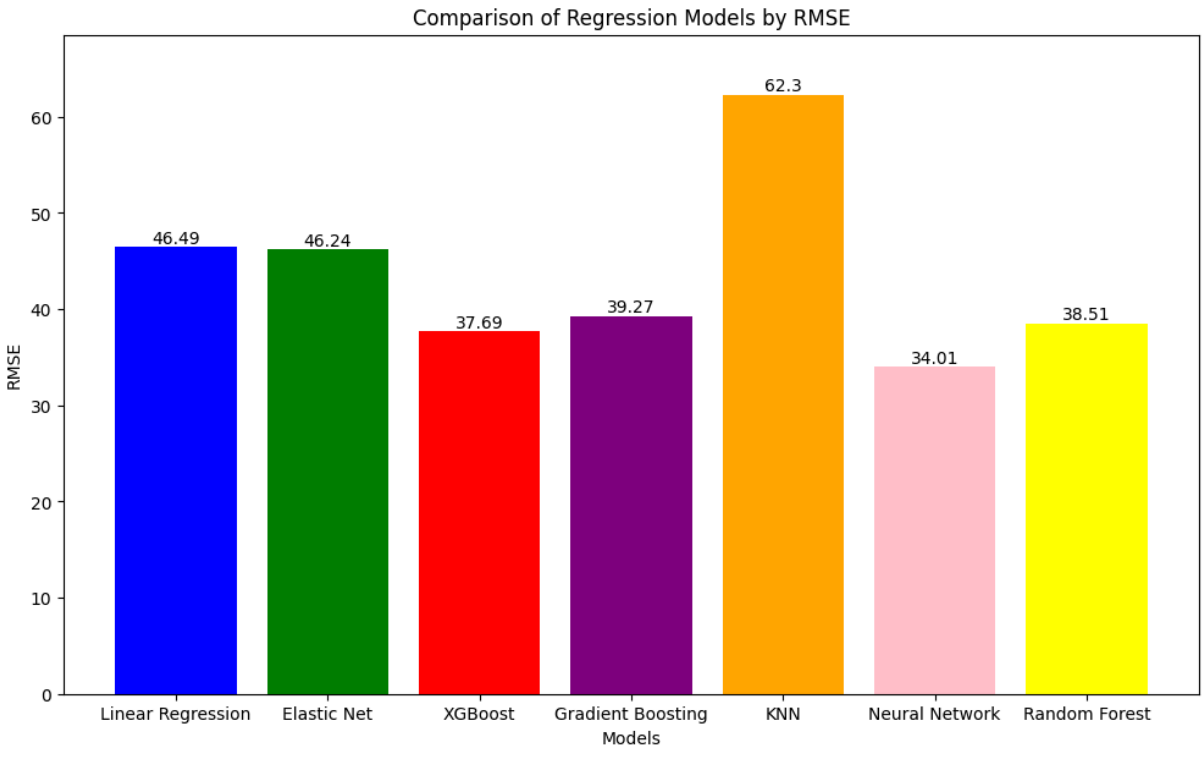
Key Insights

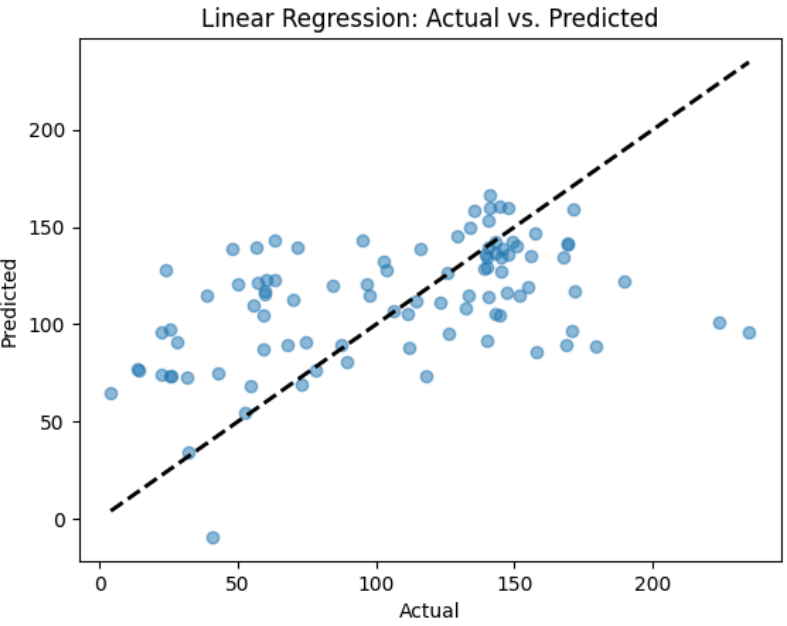
By: Kermi Kotecha
AIML

- **Best Performing Model:** The Neural Network stands out with the lowest RMSE, indicating its superior capability to model the complexity and nuances of the dataset. It should be considered for deployment if prediction accuracy is the primary criterion.

- **Competitive Models**: Both XGBoost and Random Forest show strong performances, making them excellent alternatives, particularly when model interpretability or computational efficiency is also a consideration.

- **Underperforming Model:** KNN's higher RMSE suggests it might not be suitable without adjustments or reconsideration of its parameters and the metric used for measuring distances.





Linear Regression:

Description: Linear regression predicts a target variable by fitting a linear equation to observed data. The coefficients of the equation are derived by minimizing the sum of the squared difference between the observed and predicted values.

Plot Observations: Points are clustered around the line but show variance, indicating decent predictions but potential underfitting as linear regression cannot capture more complex patterns.

RMSE: 46.49

Use-case Suitability: Best for scenarios where data is expected to have a linear distribution.

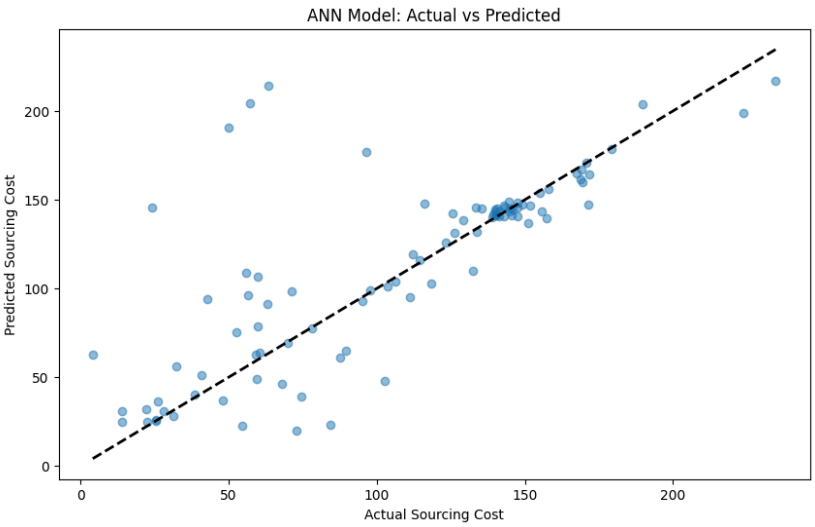
Neural Network (ANN):

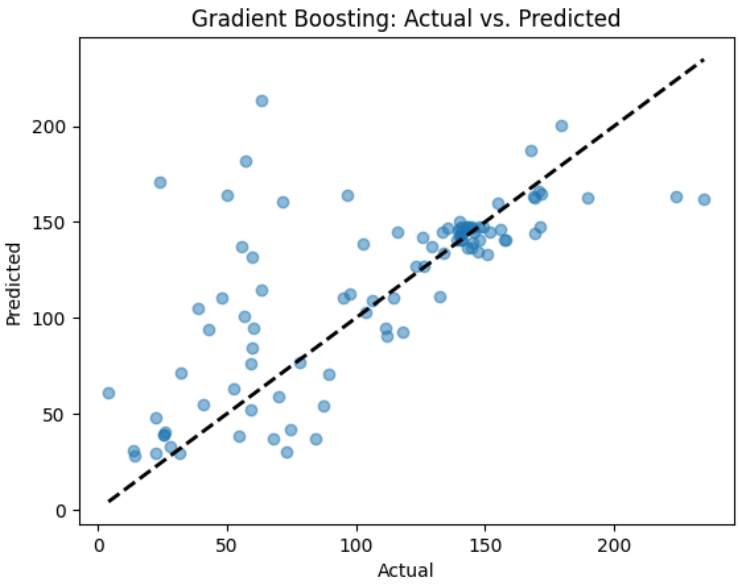
Description: Uses layers of neurons with each neuron performing a weighted sum of its inputs followed by a non-linear activation function. This structure allows it to approximate any function and capture complex patterns in the data.

Plot Observations: The dense clustering of points along the ideal line for lower to mid-range actual values indicates that the ANN has learned a good representation of the data's underlying patterns within this range.

RMSE: 34.01

Use-case Suitability: Effective in complex scenarios where relationships between data points are nonlinear and intricate.





Gradient Boosting:

Description: Builds an ensemble of weak prediction models, typically decision trees. Each subsequent model corrects the errors of the previous ones in a greedy manner, focusing on instances hardest to predict.

Plot Observations: Data points are closely aligned with the ideal line, suggesting good performance and ability to handle non-linearities and interactions.

RMSE: 39.27

Use-case Suitability: Works well for varied data types, and complex relationships, and is robust against overfitting in many cases.

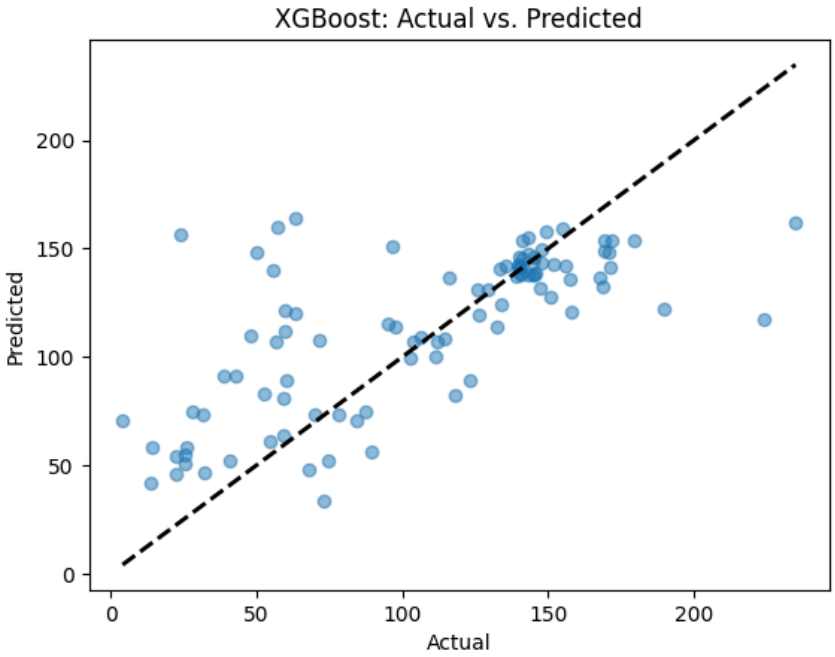
XGBoost:

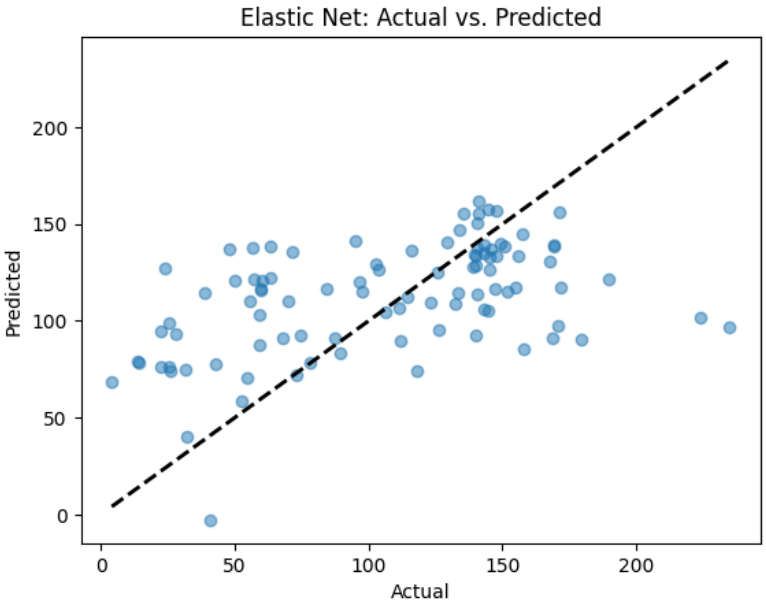
Description: An optimized distributed gradient boosting library that enhances the speed and performance of traditional gradient boosting. It includes built-in regularization which helps to prevent overfitting

Plot Observations: Similar to Gradient Boosting but slightly more scattered, still maintains a strong alignment with the ideal line.

RMSE: 37.69

Use-case Suitability: Often preferred for its speed and performance, especially in data science competitions for regression tasks.





Elastic Net:

Description: Combines the properties of both Ridge and Lasso regression. It can shrink less important feature's coefficients like Ridge and perform feature selection like Lasso by reducing some coefficients to zero.

Plot Observations: Shows a moderate fit; some data points are distant from the line, indicating a balance between bias and variance.

RMSE:46.24

Use-case Suitability: Useful when there are correlations among features, and you want to maintain a balance between Ridge and Lasso's properties.

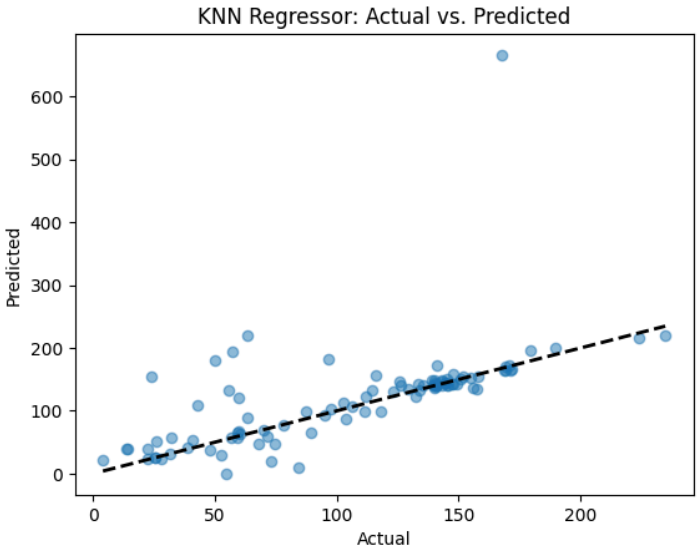
KNN Regressor:

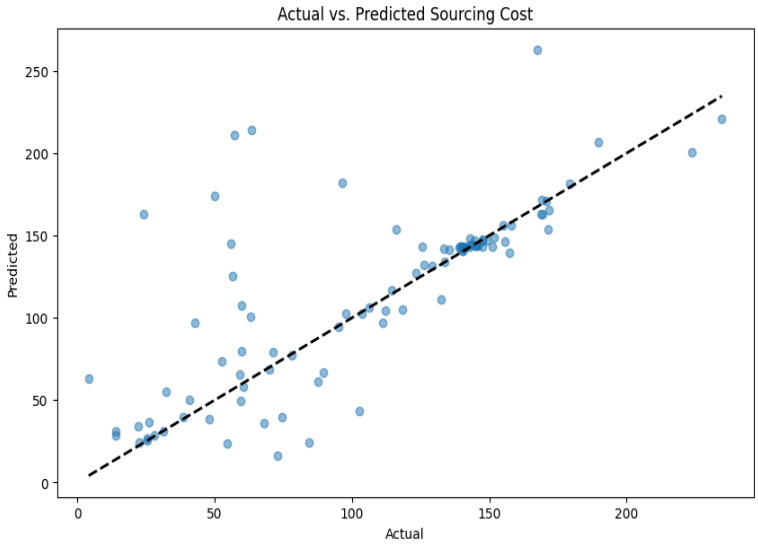
Description: A non-parametric method that predicts the target value based on the average or median of the k-nearest neighbors.

Plot Observations: Shows considerable scatter away from the ideal line, especially for higher values, suggesting potential sensitivity to outliers or noise.

RMSE: 62.3

Use-case Suitability: Good for scenarios where data forms distinct clusters; however, it can perform poorly if the dimensionality is high (curse of dimensionality).





Random Forest:

Description: It operates by constructing a multitude of decision trees at training time and outputting the mean prediction of the individual trees for regression tasks.

Plot Observations: The concentration of points along the line, especially for lower values, suggests that the model predicts more accurately in this range. As values increase, there's more spread and some data points fall farther from the line, indicating less accuracy with higher values.

RMSE: 38.51