

Diamond Price Prediction

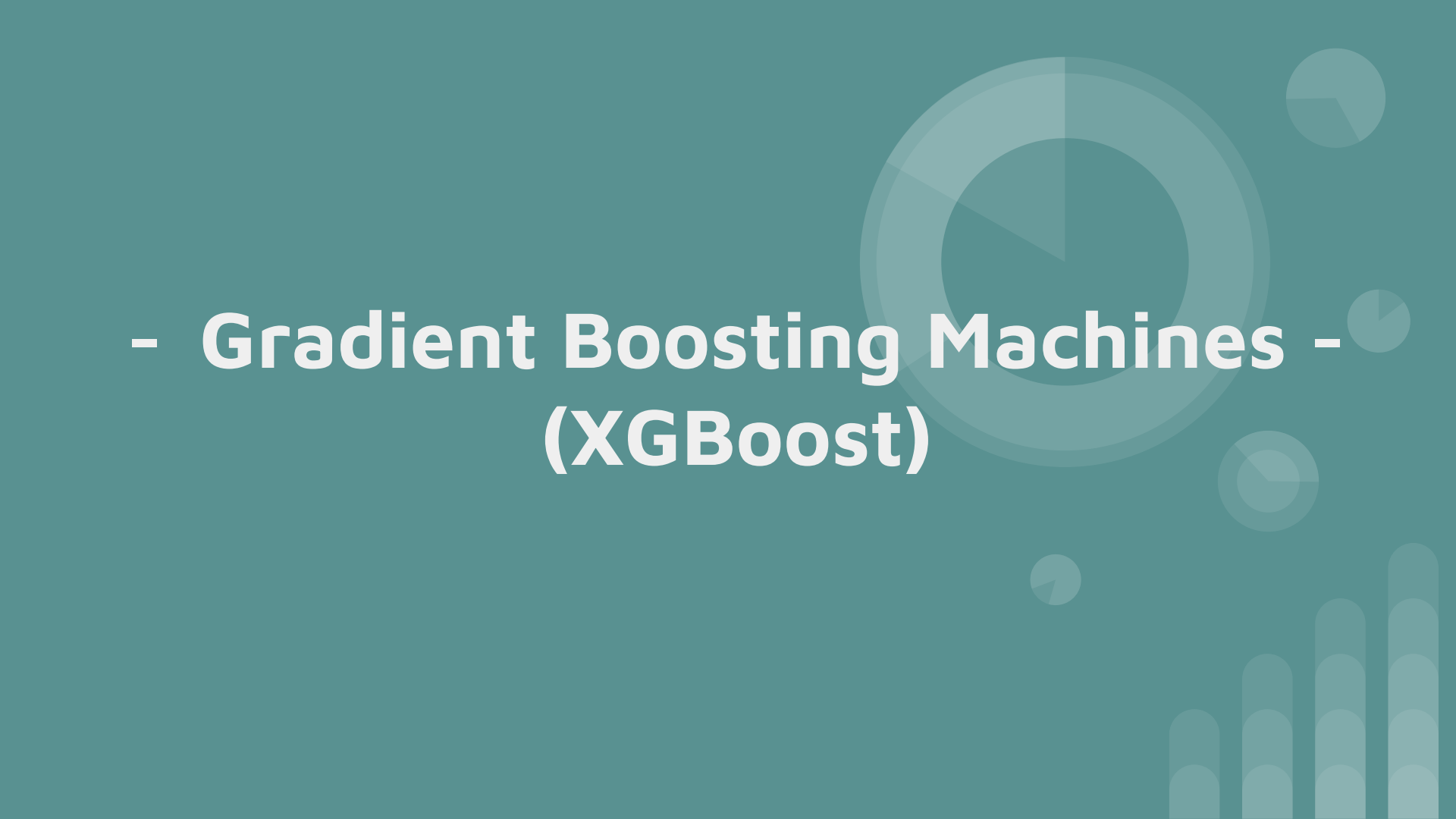
- Final Report -

Presented by:

- Arda Yalçındağ
- Eray Eroğlu

OUTLINE

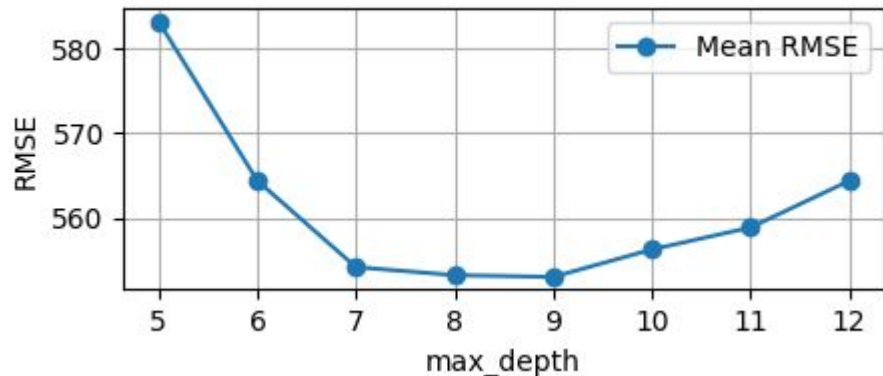
- ❖ In this presentation, we will report the results of the remaining two models that are planned to be used. Those two models are “Gradient Boosting Machines” and “Neural Networks”.
- ❖ After getting results of all the models are planned to be used, we will compare the results that are previously shown and the results with only 4 parameters which are determined to be as the “most important features” according to feature importance results of random forest which have been already presented in the progress report.
- ❖ Finally results of all the models used in this project is compared (with all of our features) and we will comment on the “best model” for the diamond price predicting task.



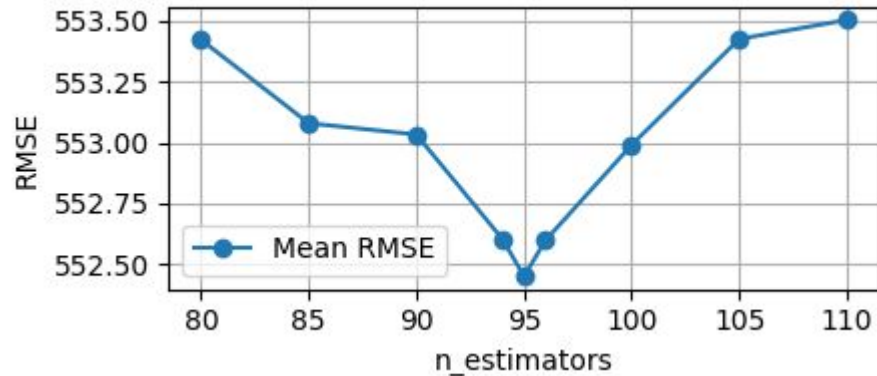
- Gradient Boosting Machines - (XGBoost)

5-Fold Cross Validation Results

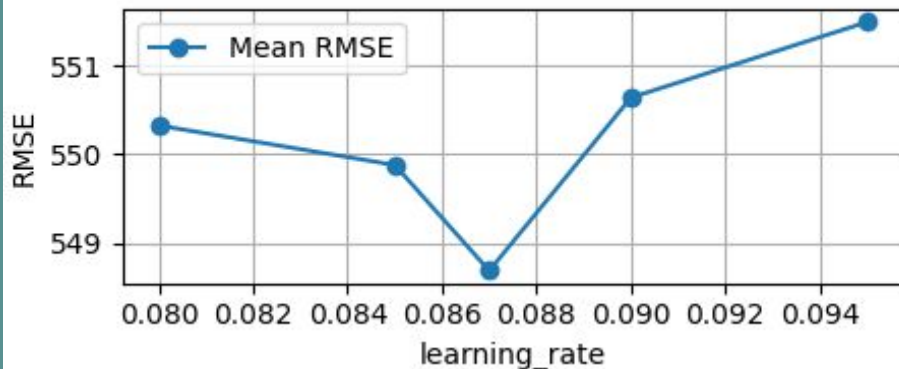
MAX DEPTH



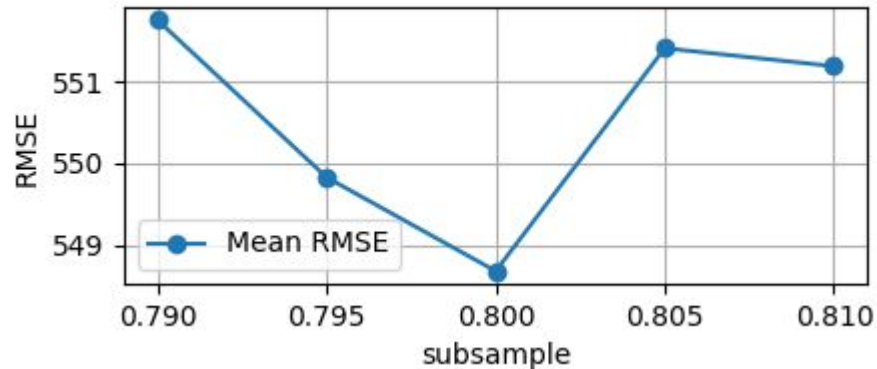
NUMBER OF TREES



LEARNING RATE



SUBSAMPLE



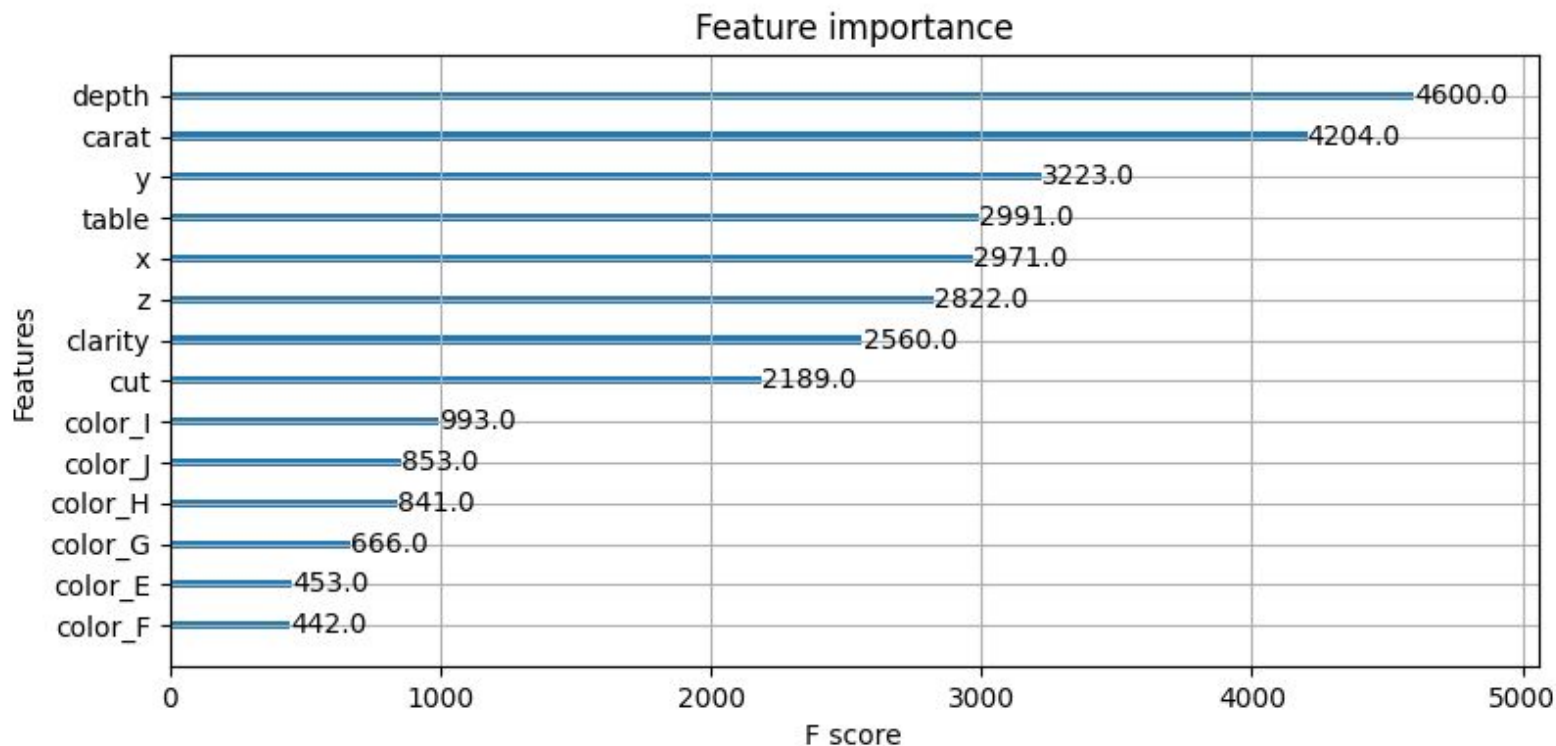
Train Results:

Mean Price of Train Set	3003.50
Standard Deviation of Train Set	2598.43
Mean Absolute Error (MAE)	151.23
Mean Squared Error (MSE)	65185.57
Root Mean Squared Error (RMSE)	255.31
R ² Score	0.990

Test Results:

Mean Price of Test Set	2979.21
Standard Deviation of Test Set	2596.52
Mean Absolute Error (MAE)	192.94
Mean Squared Error (MSE)	115367.14
Root Mean Squared Error (RMSE)	339.66
R ² Score	0.983
Execution Time (seconds)	0.33

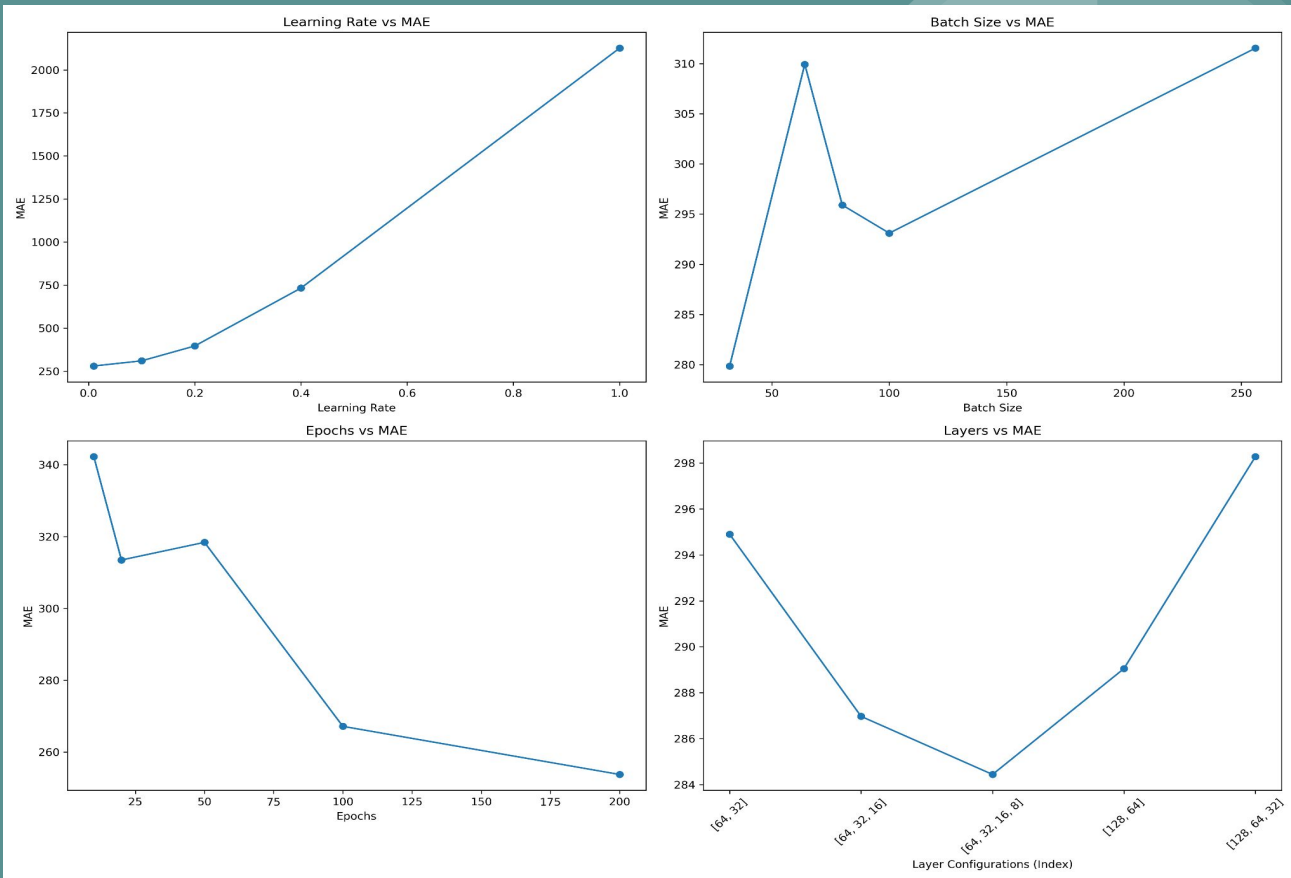
Feature Importance



- **Neural Network** -



5-Fold Cross Validation Results



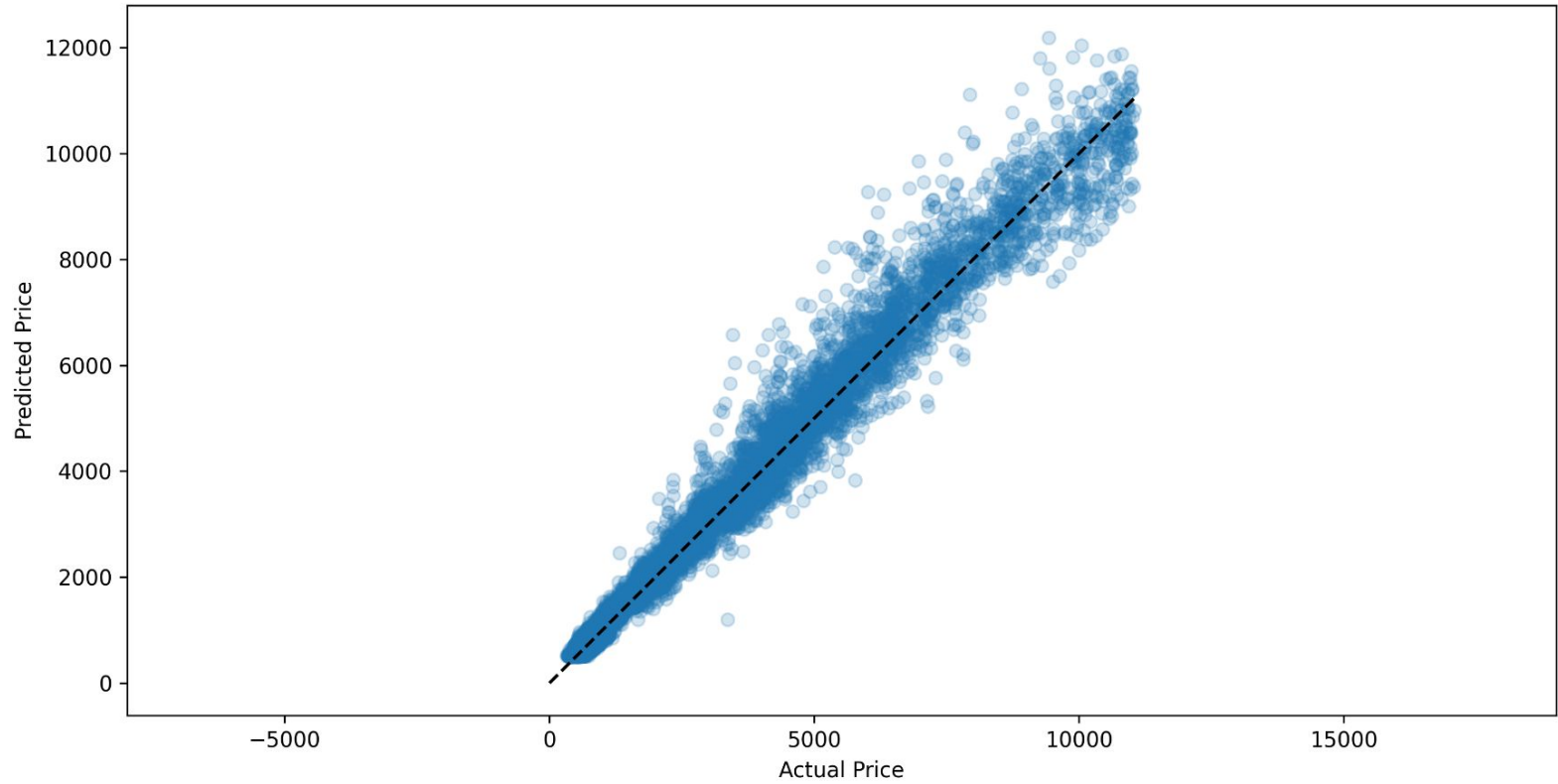
Train Results:

Mean Price of Train Set	3003.50
Standard Deviation of Train Set	2598.43
Mean Absolute Error (MAE)	239.64
Mean Squared Error (MSE)	156073.66
Root Mean Squared Error (RMSE)	395.06
R ² Score	0.9769
Execution time (seconds)	111.25

Test Results:

Mean Price of Test Set	2979.21
Standard Deviation of Test Set	2596.52
Mean Absolute Error (MAE)	232.61
Mean Squared Error (MSE)	142378.00
Root Mean Squared Error (RMSE)	377.33
R ² Score	0.981
Execution Time (seconds)	0.1

Actual vs Predicted Prices





Models' Performances with the 4 Most Important Features*

*Features are selected by evaluating the random forest model results

LINEAR REGRESSION

Metrics	Test results with all features	Test results with selected 4 features
Mean Squared Error (MSE)	542885.77	961193.35
Mean Absolute Error (MAE)	512.23	750.27
R ² Score	0.92	0.86
Execution Time (seconds)	0.1	< 0.1

RANDOM FOREST

Metrics	Test results with all features	Test results with selected 4 features
Mean Squared Error (MSE)	140212.39	948707.65
Mean Absolute Error (MAE)	212.56	586.36
R ² Score	0.979	0.859
Execution Time (seconds)	2.94	2.77

GRADIENT BOOST

Metrics	Test results with all features	Test results with selected 4 features
Mean Squared Error (MSE)	115367.14	814484.05
Mean Absolute Error (MAE)	192.94	548.4
R ² Score	0.983	0.879
Execution Time (seconds)	0.33	0.14

NEURAL NETWORK

Metrics	Test results with all features	Test results with selected 4 features
Mean Squared Error (MSE)	142378.00	420339.81
Mean Absolute Error (MAE)	232.61	394.17
R ² Score	0.981	0.9377
Execution Time (seconds)	0.1	0.08

Comparison of All Models

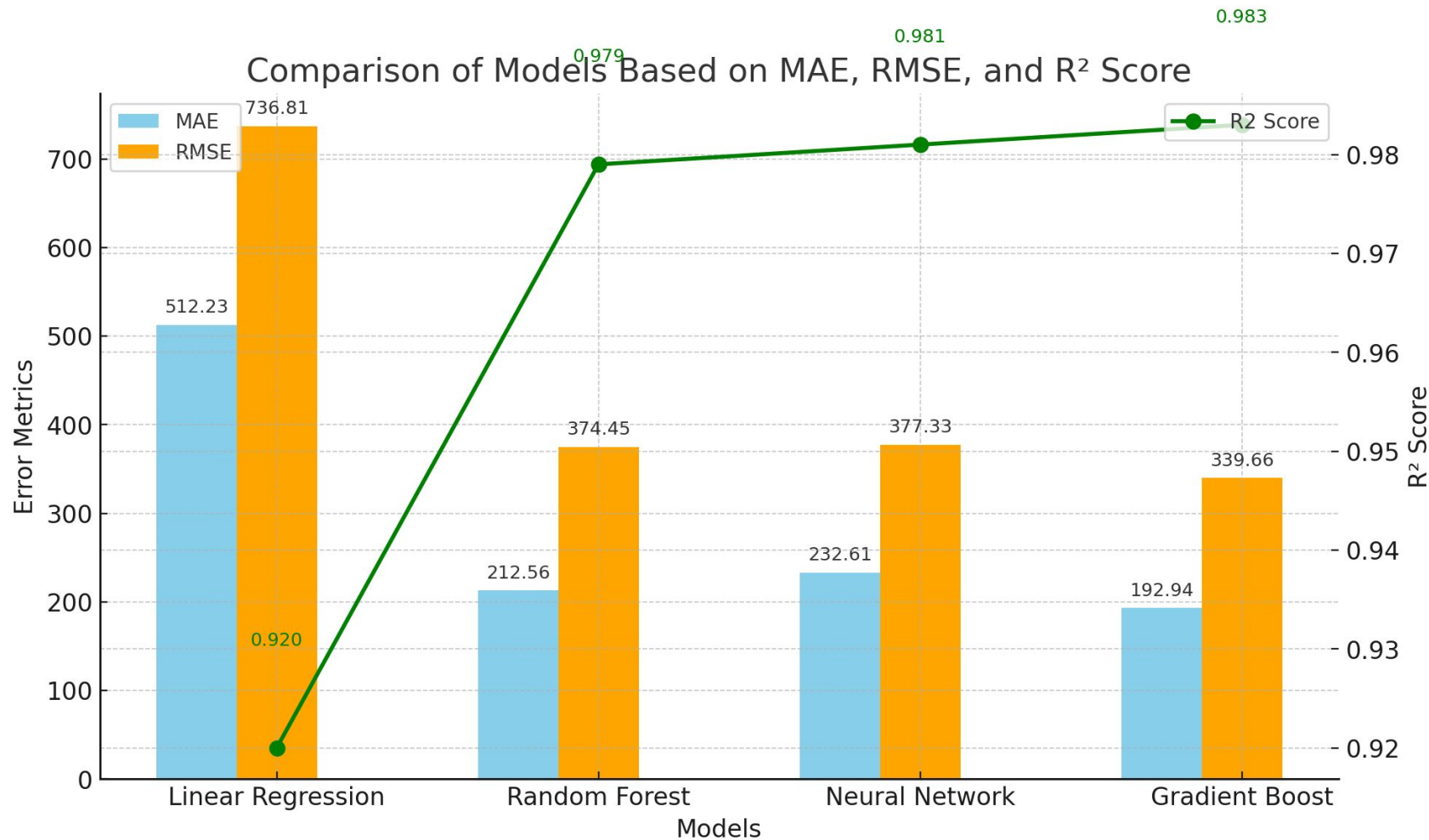


LINEAR REGRESSION	
Mean Absolute Error (MAE)	512.23
Mean Squared Error (MSE)	542885.77
Root Mean Squared Error (RMSE)	736.81
R ² Score	0.92
Execution Time (seconds)	0.1

NEURAL NETWORK	
Mean Absolute Error (MAE)	232.61
Mean Squared Error (MSE)	142378.00
Root Mean Squared Error (RMSE)	377.33
R ² Score	0.981
Execution Time (seconds)	0.1

RANDOM FOREST	
Mean Absolute Error (MAE)	212.56
Mean Squared Error (MSE)	140212.39
Root Mean Squared Error (RMSE)	374.45
R ² Score	0.979
Execution Time (seconds)	2.94

GRADIENT BOOST	
Mean Absolute Error (MAE)	192.94
Mean Squared Error (MSE)	115367.14
Root Mean Squared Error (RMSE)	339.66
R ² Score	0.983
Execution Time (seconds)	0.33



Final Evaluation

- ❖ Based on the results, the **Gradient Boosting Machine (GBM)** is the best-performing model for this dataset. It has the **lowest MAE and RMSE**, the **highest R^2 score**, and a remarkably **low execution time**. While the Neural Network is faster, the difference in execution time is negligible, making GBM the clear winner.
- ❖ The **Random Forest model** and the **Neural Network** deliver very similar results, with both being strong contenders as runner-ups to the GBM. Their MAE, RMSE, and R^2 scores are comparable, making it difficult to declare one superior to the other. But they both fall short compared to the GBM.
- ❖ Among the four models, **Linear Regression** demonstrates the weakest performance. However, it is important to note that its results are still very good, with an R^2 score of 0.92 indicating a decent fit. As the simplest model with the **lowest computational cost**, Linear Regression showed us a baseline for performance metrics.