Model Evaluation Summary

This analysis compares the performance of three models: **XGBoost**, **Random Forest**, and **Gradient Boosting**, evaluated with and without an outlier in the dataset. The models were tested using both **Train-Test Split** and **Cross-Validation** to assess stability and robustness. Below are the main findings:

---

1. Train-Test Split vs. Cross-Validation

- **Train-Test Split** results provide a one-time assessment of the model's performance on a specific train-test partition.
- **Cross-Validation** results show average performance across multiple folds, which better captures how the model might perform on unseen data.

2. Impact of Outliers

- Models trained with the **outlier** generally performed **worse** than those trained without it.
- The **RMSE** and **MAE** values increased significantly for models with the outlier, indicating higher error rates.
- **XGBoost** was particularly affected by the outlier, with its **R²** dropping significantly from **0.95** to **0.76** in cross-validation. The standard deviation of RMSE also became very high, indicating instability.

3. Best Model Selection

- Across all metrics, the **Gradient Boosting Regressor (No Outlier)** showed the best performance:
    - **Train-Test Split**: RMSE = **10.68**, R² = **0.96**, MAE = **2.06**
    - **Cross-Validation**: Mean RMSE = **10.26**, Mean R² = **0.96**, Mean MAE = **1.88**
    - The low RMSE and high R² indicate that Gradient Boosting effectively captures the variance in the target variable without overfitting.
- **XGBoost (No Outlier)** also performed well, but **Gradient Boosting** had a slight edge in stability, as shown by the lower **standard deviation** of RMSE during cross-validation.

4. Observations on Models with Outliers

- The **outlier** had a major impact on model performance, leading to a large increase in RMSE and a decrease in R².
- The **Random Forest** and **XGBoost** models had especially high **standard deviation** for RMSE during cross-validation, indicating sensitivity to outliers.
- **Gradient Boosting (With Outlier)** still managed to provide relatively stable performance, with a smaller drop in R² and lower standard deviation compared to other models.

5. Recommendation

- For the dataset **without the outlier**, **Gradient Boosting** is the preferred model due to its strong generalization and lower error rates.

- The inclusion of the outlier significantly degrades performance for all models.
Thus, **outlier removal or handling** is recommended for better predictive accuracy and model stability.

Summary of Results

Model	Evaluation Strategy	Best Metric (RMSE)	R²	Impact of Outlier
XGBoost (No Outlier)	Train-Test / CV	10.50 / 10.94	0.95	Sensitive
XGBoost (With Outlier)	Train-Test / CV	63.72 / 33.01	-0.29 / 0.76	Highly Impacted
Random Forest (No Outlier)	Train-Test / CV	13.75 / 14.19	0.93	Impacted
Random Forest (With Outlier)	Train-Test / CV	16.70 / 35.34	0.91 / 0.74	Highly Impacted
Gradient Boosting (No Outlier)	Train-Test / CV	10.68 / 10.26	0.96	Least Impacted
Gradient Boosting (With Outlier)	Train-Test / CV	10.77 / 29.60	0.96 / 0.80	Stable Compared to Others

Based on the metrics above, **Gradient Boosting without the outlier** emerges as the most effective and stable model for this task.

Model Evaluation Summary Without `agent_avg_revenue`

This section compares the performance of the models—XGBoost, Random Forest, and Gradient Boosting—when the feature `agent_avg_revenue` was excluded from the dataset. Both Train-Test Split and Cross-Validation methods were employed for evaluation, and the results indicate a notable degradation in performance.

Results Without `agent_avg_revenue`

Train-Test Split Results

Model	RMSE	MAE	R²
XGBoost (No Outlier)	16.17	3.92	0.91
XGBoost (With Outlier)	87.72	6.44	-1.44

Model	RMSE	MAE	R²
Random Forest (No Outlier)	20.77	5.96	0.85
Random Forest (With Outlier)	45.30	8.05	0.35
Gradient Boosting (No Outlier)	12.54	3.09	0.94
Gradient Boosting (With Outlier)	43.47	4.06	0.40

Cross-Validation Results

Model	Mean RMSE	Std RMSE	Mean R²	Mean MAE
XGBoost (No Outlier)	16.91	1.79	0.88	4.09
XGBoost (With Outlier)	39.00	37.01	0.67	4.91
Random Forest (No Outlier)	21.41	3.29	0.81	5.98
Random Forest (With Outlier)	51.03	32.90	0.35	7.62
Gradient Boosting (No Outlier)	14.12	1.87	0.91	3.03
Gradient Boosting (With Outlier)	55.45	36.64	-0.02	4.56

Observations¶

1. **Performance Decline**: The exclusion of `agent_avg_revenue` resulted in an increase in RMSE and MAE values across all models, indicating a decrease in predictive accuracy.
2. **Impact on XGBoost**: The performance of XGBoost, particularly with the outlier, dropped significantly, as evidenced by an RMSE of 87.72 and a negative R² value, suggesting a poor fit.
3. **Gradient Boosting Resilience**: Although Gradient Boosting (No Outlier) performed best among the models without `agent_avg_revenue`, its overall performance still degraded compared to previous results with the feature included.

Conclusion

The removal of the `agent_avg_revenue` feature negatively impacted model performance, demonstrating its importance in accurately predicting the target variable. It is recommended to retain this feature for future modeling efforts to ensure optimal predictive capability.