Report on Diamond Price Prediction Model Performance Stage 3

Camille Dupre La Tour

December 2024

1 Report on Feedback from Stage 2 and Responses

1.1 Feedback from Stage 2 and Action Plan

During Stage 2, several key pieces of feedback were provided regarding the diamond price prediction model. The analysis of the results needed to be more in-depth, particularly by color group, to better evaluate the model's performance. The feedback suggested displaying performance curves for each group with distinct colors, identifying areas where the model performed poorly, and addressing those specific cases. Additionally, a more detailed and clear presentation of results for all color groups was recommended.

To address these points, the following steps were taken:

- Performance curves were displayed for each color group, with distinct colors representing different categories.
- Colors where the model underperformed were identified.
- The model have to be optimized to improve predictions for these specific colors.
- Presented the results with detailed numerical values for each color group.

2 Analysis of Model Performance by Color Category: Identifying Areas for Improvement

In this section, I focus on evaluating the model's performance across different diamond color categories to gain insights into areas where the model is underperforming and to suggest improvements. Here's a breakdown of the process:

2.1 Initial Comparison: Reality vs Prediction by Color

I started by visualizing the actual vs. predicted values for each color category, before and after optimizing the model. Scatter plots were used, color-coded by diamond category, to show the predictions for each color. Comparing these plots before and after optimization allowed me to identify any significant performance differences for specific color groups.

Purpose: This step provided an initial overview of model performance across all categories and helped pinpoint any obvious issues, such as large discrepancies between predicted and actual values.

Results: At this stage, there were no clear performance issues specific to any color category, indicating that the model was relatively consistent across all groups. However, further analysis was needed to examine the metrics in more detail.

2.2 Metrics Comparison: Mean Absolute Error (MAE) and R-Squared by Color

To gain deeper insights into the model's performance, I analyzed MAE and R^2 metrics for each color category. This comparison, before and after optimization, allowed me to determine whether the adjustments had led to measurable improvements.

Purpose: This analysis helped identify how accurate the model was (MAE) and how well it fit the data (R^2) for each color group, providing a more granular understanding of the model's performance.

Results: The analysis revealed that, while overall model performance improved, there was a slight decline in performance for colors 4 and above.

This suggested that the model had more difficulty handling higher color categories.

2.3 Group Comparison: Separating Colors into Two Categories

To better assess model performance, I grouped colors into two categories: colors 0-3 (lower color categories) and colors 4-6 (higher color categories). I re-evaluated MAE and R^2 for these groups to understand whether the model's accuracy differed significantly between the two.

Purpose: This grouping allowed me to investigate whether the model faced particular challenges with higher color categories, which might involve outliers or more complex data patterns.

Results: The analysis showed that the higher color categories (4-6) contained outliers, which appeared to affect the model's performance. Despite improvements from the optimization, the model still struggled to predict values accurately for these categories.

2.4 Insights and Next Steps

The outliers present in higher color categories likely explain the model's reduced performance in these areas. Although optimization improved the model overall, the remaining discrepancies suggest that further fine-tuning is necessary. Future steps could include:

- Investigating Data Distribution: Exploring whether the higher color categories have skewed distributions that may influence predictions.
- Feature Importance Review: Analyzing the role of features related to color and considering if certain features need better representation in the model.
- Model Comparaison: Testing alternative algorithms to better capture the complexities of higher color categories.

By addressing these specific challenges, I can work toward improving the model's performance across all color categories.

3 Model Comparison and Stacking Approach

In my testing, I explored four models—Linear Regression, Decision Tree, Random Forest, and K-Nearest Neighbors—to evaluate how well each handled outliers and predicted diamond prices. These models were selected for their different strengths in managing extreme values, which were critical for this task.

3.1 Linear Regression

Performance: Linear Regression demonstrated limited effectiveness in handling outliers. While the model fit the data, its performance metrics (MAE and R^2) were weaker compared to the other models.

Reason: Linear Regression assumes a linear relationship between features and target variables, which is sensitive to outliers. Extreme data points distort the regression line, leading to suboptimal predictions.

3.2 Decision Tree

Performance: The Decision Tree performed better than Linear Regression but was still outperformed by Random Forest. Decision Trees are non-linear and can handle outliers by partitioning the feature space into regions.

Reason: Although the Decision Tree can capture non-linear relationships and manage some outliers, it is still prone to overfitting, especially with noisy data or irregular outliers.

3.3 Random Forest

Performance: Random Forest performed best in terms of MAE and R^2, demonstrating its strength in handling outliers. By averaging over many individual decision trees, Random Forest reduces the likelihood of overfitting and is less sensitive to noise.

Reason: As an ensemble method, Random Forest benefits from the diversity of multiple trees, each making decisions based on different data subsets. This improves robustness and smoothens predictions, particularly in the presence of outliers.

3.4 K-Nearest Neighbors (KNN)

Performance: KNN performed reasonably well but was not as effective as Random Forest. KNN relies on the proximity of data points, and while it can manage some outliers, it struggles when the data is noisy or contains significant outliers.

Reason: Outliers in KNN can distort distances between neighbors, negatively affecting predictions, especially when outliers are far from the majority of the data points.

4 Stacking Approach

After evaluating the individual models, I applied a **stacking method** to combine Decision Tree and Random Forest, using Linear Regression as a meta-model to improve overall performance.

Stacking Model Performance: After stacking, the results were nearly identical to Random Forest:

- **Random Forest**: MAE = 358.5338, $R^2 = 0.9764$
- **Stacking**: MAE = 358.5026, $R^2 = 0.9762$

Analysis: The stacking approach did not offer a significant improvement over Random Forest. The metrics were nearly the same, indicating that combining the models didn't add substantial value. This suggests that Random Forest was already capturing the data patterns effectively, and stacking did not improve its performance.

5 Insights and Conclusions

The analysis reveals the following key insights:

- **Random Forest Effectiveness**: Random Forest outperformed all other models, including Decision Tree, KNN, and even the stacking model. Its ensemble nature allows it to handle my dataset and outliers effectively.
- **Decision Tree's Role**: The Decision Tree model showed promise but was not as robust as Random Forest. It can handle outliers but is prone to overfitting, making it less reliable for this task.

- **Limited Improvement with Stacking**: The stacking method did not significantly improve model performance. The slight drop in R^2 and MAE after stacking suggests that combining Decision Tree and Random Forest didn't provide extra value.
- **Outlier Handling**: Random Forest and Decision Tree are the most capable of handling outliers, with Random Forest showing the most stability in predictions.

6 Conclusion

Based on the comprehensive testing and analysis, **Random Forest** remains the most reliable and accurate model for predicting diamond prices while handling outliers effectively. Although stacking and other models were valuable for comparison, Random Forest provides the best solution for this dataset.