Art prediction model improvement report

This study utilizes a dataset of observed art auction results for prints and multiples, which includes information on artwork characteristics, historical auction data, artist names, and relevant market indicators. The dataset has been carefully curated to ensure accuracy and relevance, providing a robust foundation for the predictive models discussed in this work.

1. Datasets

From the original dataset, four distinct datasets were generated by further extracting image-related features of the artworks. The core features common to all datasets include:

- ARTIST: Name of the artist.
- TECHNIQUE: Medium or method used (e.g., lithograph, etching).
- SIGNATURE: Whether the artwork is hand signed, plate signed or unsigned.
- CONDITION: Physical state of the artwork.
- TOTAL DIMENSIONS: Area of the artwork.
- YEAR: Year of creation.
- PRICE: Final auction price.

Key Differences across the datasets are as follows:
AuctionResultsNoImg (NoImg) Dataset contains only the core features without any image-related features.
AuctionResultsColor (Color) Dataset includes an additional Colorfulness Score [18], a measure of color intensity and variety of the image of the artwork.
AuctionResultsSVD (SVD) Dataset adds SVD Entropy [19] of the image of the artwork to the core features, excluding the Colorfulness Score.
AuctionResultsColorSVD (ColorSVD) Dataset adds both Colorfulness Score and SVD Entropy, which quantifies the complexity of the artwork's visual representation.

2.. Dataset Description

The cardinality of the data plays a crucial role in the analysis of the model. Table 1 offers an overview of the number of unique values for the core features, with the exception of CONDITION and SIGNATURE, which have only three distinct values each.

Table 1. Number of unique values for each feature and its ratio to total number of values.

Calculated Field	ARTIST	TECHNIQUE	TOTAL-DIMENSIONS	YEAR	PRICE
Unique Values	395	10	2534	123	795
Unique Values/Total Count (%)*	1.55	0.04	9.97	0.48	3.13

* Total size of the data frame after preprocessing resulted in 25408 rows.

The four datasets under consideration share identical characteristics for the features presented in Table 1, differing only in image-related attributes. The high cardinality of the ARTIST feature is particularly significant, as it presents challenges that the models must address. In contrast, the TECHNIQUE feature exhibits relatively low cardinality. The TOTAL-DIMENSIONS feature demonstrates a substantial number of unique values, as it is derived from the product of two dimensions of the artworks. The YEAR feature reflects a temporal span of 123 years, indicating the breadth of the data. Lastly, the unique value distribution of the PRICE feature suggests considerable variation between individual price points, highlighting the feature's complexity.
The distribution of the core features is shown in Table 2.

Table 2. The distribution of the numerical core features.

Calculated Field	TOTAL-DIMENSIONS	YEAR	PRICE
Count	25408	25408	25408
Mean	2259.03	1973.6	225.68
Standard deviation	1674.0	21.7	505.40
Minimal value	10.64	1900	1
25% Quantile	875	1963	50
50% Quantile	1750	1974	99
75% Quantile	3401	1985	200
Maximal value	10000	2023	10000

3. Model Selection and Training

The analysis involved a comprehensive comparison of various machine learning models, including Linear Regression, KNN, Decision Trees, Random Forests, Gradient Boosting Machines (GBMs such as XGBoost, CatBoost), Model Trees and neural models such as MLP and other models like, VIME, DeepGBM, DeepFM, and SAINT.
Each model was trained using 5-fold cross-validation to ensure robust performance across different data subsets, thereby reducing the risk of overfitting. Hyperparameter tuning was conducted using the Optuna library [21], with a maximum of 5 trials to opti-mize the model settings and maximize the models' performance. Neural network models were trained for up to 1,000 epochs, with early stopping implemented after 20 epochs without improvement. The models were trained on their mean squared error.

4. sMAPE Score Analysis Across Datasets

Method	AuctionResultsNoImg	AuctionResultsColor	AuctionResultsSVD	AuctionResults ColorSVD	CaliforniaHousing
LinearModel	101.33 ± 0.78	101.01 ± 0.84	102.75 ± 1.16	101.13 ± 0.93	28.70 ± 0.41
KNN	61.38 ± 0.35	62.68 ± 0.32	60.18 ± 0.65	60.93 ± 0.53	22.75 ± 0.51
DecisionTree	58.92 ± 0.62	58.39 ± 0.69	56.70 ± 0.78	59.21 ± 0.69	21.70 ± 0.56
RandomForest	61.20 ± 0.60	**55.95 ± 0.29**	56.94 ± 0.30	**55.51 ± 0.80**	17.50 ± 0.36
XGBoost	**55.11 ± 0.60**	57.50 ± 0.61	**54.83 ± 0.59**	58.76 ± 1.07	14.84 ± 0.25
CatBoost	58.91 ± 0.79	60.45 ± 1.02	58.20 ± 0.64	58.25 ± 0.48	14.92 ± 0.46
LightGBM	57.53 ± 2.32	58.29 ± 2.34	58.04 ± 0.59	57.52 ± 1.22	**14.71 ± 0.31**
MLP	62.98 ± 0.89	61.86 ± 1.14	63.68 ± 0.32	65.02 ± 0.75	17.52 ± 0.63
VIME	75.29 ± 3.37	66.29 ± 1.59	86.44 ± 2.28	74.09 ± 2.91	19.40 ± 1.69
ModelTree	68.50 ± 0.37	66.59 ± 0.66	71.68 ± 3.62	68.70 ± 1.35	23.86 ± 0.33
DeepGBM	76.92 ± 7.66	77.14 ± 7.40	87.96 ± 10.62	76.18 ± 4.23	35.13 ± 2.11
DeepFM	63.10 ± 0.76	63.45 ± 0.87	63.28 ± 0.25	63.06 ± 0.81	17.75 ± 0.36
SAINT	60.57 ± 0.95	60.75 ± 0.78	62.41 ± 1.19	60.63 ± 1.43	16.64 ± 0.30

The Table 3 provides a comparative evaluation of the models based on their sMAPE scores across the datasets. In the NoImg dataset, XGBoost demonstrates the best perfor-mance, achieving a sMAPE score of 55.11%, with LightGBM following at 57.53%. The LinearModel, however, exhibits significantly inferior performance with sMAPE over 100% for each dataset, indicating its limitations in handling the complexity of auction data. Similarly, DeepGBM, with a sMAPE of 76.92% and VIME with 75.29% struggle to gener-alize effectively, while other models present moderate performance ranging from 60.57% for SAINT to 68.50% for ModelTree.

In the Color dataset, RandomForest emerges as the top performer with a sMAPE of 55.95%, outperforming XGBoost, which records 57.50%. Other models result in similar range of error with only improvement for VIME with almost 10 percentage points of de-crease.

In the SVD dataset, XGBoost once again leads with a sMAPE score of 54.83%, fol-lowed closely by RandomForest at 56.94%. On the contrary to Color dataset, providing SVD Entropy resulted in the worst noted error for VIME (86.44%) and DeepGBM (87.96%).

For the ColorSVD dataset, RandomForest delivers the best performance, achieving a sMAPE of 55.51%, with CatBoost in second place at 58.25%. VIME (74.09%) and DeepGBM (76.18%) still exhibit higher error rates than other moderately-performing models, sug-gesting that complex neural network-based models are less effective compared to ensem-ble methods when Colorfulness Score and SVD Entropy enhance the auction data.

The Housing dataset is simpler compared to the auction datasets, as it displays higher linearity between features. In this setting, LightGBM achieves the best performance with a sMAPE of 14.71%, slightly outperforming XGBoost (14.84%). However, DeepGBM performs poorly in this benchmark, with a sMAPE of 35.13%, further demonstrating DeepGBM may not be suitable for this kind of tabular datasets, while this time VIME is placed in moderate performers.

5. Possible improvements

Below there are identified possible improvements:

1. **Data Augmentation and Feature Engineering**:
    a. **Expantion of the feature set**: While features like ARTIST, TOTAL DIMENSIONS, and YEAR were influential, there may be additional features that could improve the models. For example:
        i. **Market trends**: Incorporating real-time market sentiment data (e.g., economic indicators, social media trends, or cultural shifts) could provide dynamic context that could improve neural network performance, as these models can benefit from larger datasets with richer feature interactions.
        ii. **Temporal features**: More detailed temporal features like auction seasonality or market cycles might improve predictions by considering the influence of timing on auction outcomes.
        iii. **Artist-specific metrics**: Metrics such as the number of works available by the artist in the market, changes in artistic periods, and thematic variations could provide additional context to help predict auction prices.
2. **Handling Data Imbalance**:
    a. **Addressing the imbalance between high-value and low-value art**: In art auctions, the price range is often skewed, with a few artworks selling for disproportionately higher prices. Using techniques like **SMOTE** (Synthetic Minority Over-sampling Technique) or focusing on specific strata of the market through segmentation (e.g., high-end vs. mid-range artworks) could improve model accuracy for outliers.
    b. **Outlier detection**: Implementing more robust outlier detection methods (e.g., Isolation Forest or RobustScaler) can help deal with extremely high or low-priced artworks, which may otherwise disproportionately impact the model's performance.
3. **Model Complexity and Regularization**:
    a. **Regularization techniques**: Investigating advanced regularization techniques (such as **Elastic Net regularization**) can help balance model flexibility with the need to avoid overfitting, especially for smaller datasets.
4. **Neural Networks Improvements**:
    a. **Hybrid models**: While neural networks (MLP, VIME, DeepGBM) underperformed, their potential could be tapped through hybrid models. For example, using **stacking** or **ensemble learning** that combines tree-based models and neural networks might harness the strengths of both approaches. Tree-based models could handle structured data, while neural networks could handle image or text data more effectively.
    b. **Transfer learning**: If larger image datasets related to artworks can be acquired, **transfer learning** (using pre-trained models on related tasks, such as object or style recognition) could be used to boost neural network performance on auction datasets.
    c. **Neural networks with embeddings**: Using **entity embeddings** for categorical features like ARTIST or TECHNIQUE might allow neural networks to better

capture relationships between categorical variables and auction prices, especially in smaller datasets.

5. **Incorporating Economic and Sentiment Data**:
    a. **Social media and sentiment analysis**: Incorporating **text mining** on artist mentions in news articles, social media platforms, or reviews could enhance the models. For instance, models like **BERT** or **RoBERTa** could be used to analyze text data, which could be combined with existing features for auction predictions.
    b. **Macro-economic indicators**: Adding data on broader economic trends, inflation, or GDP fluctuations could provide important context for art auction prices, especially in times of economic uncertainty.
6. **Dataset Size and Quality**:
    a. **Increase dataset size**: Art auction datasets are typically small due to the niche nature of the market. Expanding the dataset with more historical data or incorporating auction data from multiple platforms or regions (e.g., online auction houses, private sales) could provide a richer dataset for model training.
    b. **Cross-market data fusion**: Integrating data from related markets, such as private galleries or art fairs, could provide additional context, helping models to better generalize predictions across different market segments.
7. **Advanced Interpretability Techniques**:
    a. **Model interpretability with SHAP and LIME**: As noted, tree-based models are relatively interpretable, but employing more advanced interpretability techniques like **SHAP (Shapley Additive Explanations)** or **LIME (Local Interpretable Model-Agnostic Explanations)** could deepen the understanding of how specific features influence model predictions.
    b. **Interactive dashboards**: Building **interactive dashboards** for visualizing feature importance, model predictions, and explanations could improve user trust and allow auction houses or investors to explore different auction scenarios.
8. **Cross-validation and Robust Evaluation**:
    a. **More robust cross-validation**: Use **nested cross-validation** or **time-series cross-validation** (for temporal data) to ensure the models are not overfitting and to provide a better estimate of out-of-sample performance.
    b. **Ensemble stacking**: Beyond simple ensemble methods like bagging and boosting, applying **stacking techniques** that combine several models (e.g., tree-based and linear models) might yield improved performance. The first layer could consist of multiple models, whose predictions are fed into a final meta-model to make the final prediction.
9. **Explore Auction Timing Impact**:
    a. **Temporal modeling**: Explore using **temporal features**, such as the time since an artwork's previous sale, market seasonality, or the timing of high-profile art exhibitions. Using **Time-series models** (like **Prophet** or **LSTM**) could help capture how auction timing affects prices.
10. In summary, further improvements in the dataset can focus on **adding richer features** and handling **imbalances**. As for model improvement, enhancing **regularization**,

hyperparameter tuning, and exploring **hybrid models** that combine tree-based and neural network approaches may yield better results. Additionally, **advanced interpretability methods** like SHAP or LIME could make the models more transparent and trustworthy for art market stakeholders.