Paradigm Academic Press
Art and Society
ISSN 2709-9830
OCT. 2024 VOL.3, NO.5

Predictive Modelling Using Museum Data

Xinrui Li[1]

[1] Fettes College Guangzhou, Guangzhou, Guangdong, China

Correspondence: Xinrui Li, Fettes College Guangzhou, Guangzhou, Guangdong, China.

doi:10.56397/AS.2024.10.05

Abstract

Museums face challenges in maintaining and preserving their vast collections, particularly when identifying artworks that require restoration and detecting potential forgeries. This project leverages machine learning models to enhance museum collection management. Using data from a museum collection, Random Forest and Isolation Forest algorithms predict restoration needs and detect forgeries, respectively. The results show high accuracy in restoration prediction, with Age at Acquisition being the most significant feature. Forgery detection flagged 1,303 potential cases, providing museums with valuable insights for further investigation. This approach streamlines operational processes and ensures the long-term preservation and authenticity of art collections.

Keywords: museum collections, machine learning, artwork restoration, forgery detection, Random Forest, Isolation Forest, data-driven analysis, art preservation

1. Introduction

The museum has vast and different artwork tools of actual and aesthetic importance. Organizing such a collection requires systematizing and flourishing works and establishing their stability through restoration and authenticity. It covers a vast collection of different mediums, styles, and ages. However, museums face challenges in identifying a piece's selection for restoration and detecting potential forgeries**.** This leads the project toward machine learning techniques to help museums track their collections, predict restoration needs, and detect forgery risks. It consists of essential features such as artwork dimensions and artist details and collecting data. The main objective of this system is to organize collection management and protect the authenticity and stability of precious pieces. This work is necessary for improving museum efficiency. Operational processes are observed, and valuable works are preserved for future recommendation. The system can manage restoration efforts, identify potential forgeries, and be helpful for collections that have spread over the centuries of artistic heritage.

2. Rationale

The difficult task is selecting and managing museum collections, which requires skilled people in art history and preservation techniques. It was observed that the mentioned works at risk of degradation may be fraudulent so that they can be more demanding and energetic. The present situation of museum practices depends on expert keepers and restorers evaluating each piece, which may not be favorable for more comprehensive collections. When museum collections continue to grow, there is a clear need for data-driven methods to help in artwork restoration and authentication decision-making.

A well-organized approach is required for these tasks in this project. The data can be handled very well when machine learning models are applied to a museum's dataset and provide data awareness. Otherwise, it won't be easy to extract the data. The restoration prediction model identifies those pieces that have independently matured and have some significant physical characteristics. It also required restoration. Forgery detection techniques are used to select artworks that diverge from known patterns such as artist, medium, or dimensions. This evaluation can reduce the load on the museum staff. Using this technique, the accuracy and timeliness of decisions increase about the care of collections.

3. Data Description

The data is collected from the Museum Collection dataset available on Kaggle in this project. The dataset contains two main files: one contains details about the artists and the other about the artworks. These datasets were preprocessed and cleaned to handle missing values, which confirms data analysis compatibility.

3.1 Artists Dataset

This dataset or set of data gives information on the name, nationality, gender, and birth/death years of individual artists. This information is very sensitive for linking artists with their artworks and providing demographic awareness that may be correlated with predicting restoration needs or forgery risk detection.

3.2 Artworks Dataset

The artwork dataset provides complete information about each artwork, including its title, dimensions, medium, data collection, and classification. It demands the basic features used in the models to guess restoration needs and detect forgery risks.

3.3 Data Dictionary

3.3.1 Artists Dataset

Column Name	Description	Data Type
Artist ID	Unique identifier for the artist	int64
Name	Name of the artist	object
Nationality	Nationality of the artist	object
Gender	Gender of the artist	object
Birth Year	Year the artist was born	int64
Death Year	Year the artist passed away (if applicable)	int64

3.3.2 Artworks Dataset

Column Name	Description	Data Type
Artwork ID	Unique identifier for the artwork	int64
Title	Title of the artwork	Object
Artist ID	Unique identifier for the artist	Object
Name	Name of the artist	Object
Date	Date the artwork was created (year)	float64
Medium	The medium used to create the artwork	Object
Dimensions	Physical dimensions of the artwork (height, width, etc.)	Object
Acquisition Date	Date when the museum acquired the artwork	Object
Credit	Information on how the artwork was acquired (gift, purchase, etc.)	Object
Catalog	Catalog number of the artwork	int64
Department	The department or collection within the museum that the artwork belongs to	int64
Classification	Classification of the artwork (e.g., Architecture, Painting)	Object
Object Number	Internal museum number used to catalog the artwork	Object
Diameter (cm)	Diameter of the artwork in centimeters	float64
Circumference (cm)	Circumference of the artwork in centimeters	float64
Height (cm)	Height of the artwork in centimeters	float64
Length (cm)	Length of the artwork in centimeters	float64

Width (cm)	Width of the artwork in centimeters	float64
Depth (cm)	Depth of the artwork in centimeters	float64
Weight (kg)	Weight of the artwork in kilograms	float64
Duration (s)	Duration of the artwork (if applicable) in seconds	float64

4. Methodology

This project's methodology has two main objectives: predicting artworks' restoration needs and detecting potential forgeries. This approach includes data preprocessing, feature engineering, model training, and evaluation. Each step of the model is designed very carefully to ensure that the models can make accurate predictions on the available data. The two main models used in this project are the Random Forest Classifier for restoration prediction and an Isolation Forest for forgery detection. A detailed explanation of each step is given below.

4.1 Data Preprocessing

In data preprocessing, cleaned and preprocessed data is supportable for making machine learning models consistent and compatible. Every step or phase of preprocessing was applied to both data sets, the Artists Dataset and the Artworks Dataset.

4.1.1 Artists Dataset

Handling Missing Values

The missing values in the Birth Year and Death Year columns were filled using the median in the relevant columns. This assessment verified that no data were missing during the analysis. After imputing missing values, the Birth Year and Death Year columns were converted to integers for numerical analysis.

4.1.2 Artworks Dataset

Handling Missing Values

Missing values in categorical columns like Artist ID and Name were filled with the "Unknown." Missing numerical values in fields such as Height (cm), Width (cm), Depth (cm), Weight (kg), and Duration (s) were filled using the median values of each column. Meanwhile, the columns Diameter (cm) and Circumference (cm), which contained many missing values, were filled with 0, as they were not critical to the model's performance.

Label Encoding

The Catalogue and Department Categorical columns were label-encoded using Label Encoder. This conversion was necessary for the machine learning algorithms to process categorical data.

Type Conversion

Several columns initially treated as objects or strings were converted to appropriate numerical types. For example, the Date, Height (cm), Width (cm), Depth (cm), Length (cm), Weight (kg), and Duration (s) columns were converted to numeric types using pandas to numeric function.

4.2 Feature Engineering

Feature engineering was crucial because raw data was transformed into a format suitable for machine learning. The features engineered in this project improved the models' ability to predict restoration needs and detect forgeries.

4.2.1 Calculating Age at Acquisition

The new feature Age at Acquisition was derived by subtracting the artist's birth year from the acquisition year of the artwork. This feature is exceptional because older artworks are more suitable to require restoration.

4.2.2 Encoding Categorical Features

To transform the Artist ID column with a label-encoded label into a numerical feature for the models. This step was necessary for both the restoration prediction and forgery detection models. Converting categorical data into numerical format is required for machine learning algorithms to process it efficiently.

4.3 Model Training

Two well-defined machine learning models were used in this project:

4.3.1 Restoration Prediction Model (Random Forest Classifier)

The restoration prediction task was formulated as a dual classification problem. Artworks were classified in the sense of whether restoration was necessary or not. It is based on the Age at Acquisition and other physical attributes such as dimensions and weight.

Model Selection

The RF classifier was chosen because it effectively handles numerical and categorical data. The data is collected from random forests and provides information on the feature's importance. It helps one understand the critical factors in restoration requirements.

Features Engineering

The features used for the restoration prediction model are Artist ID Encoded, Age at Acquisition, Height (cm), Width (cm), Depth (cm), and Weight (kg).

Train-Test Split

The data was divided into training and testing using an 80-20 split. This confirms that the model was trained on a specific portion of the data and secured for evaluation to check for overfitting.

Model Training and Evaluation

The random forest model was selected for training. It was trained with 50 decision trees (n_estimators=50) and a maximum tree depth of 10 (max depth=10). These hyperparameters were chosen to save overfitting and protect model complexity. Then, training was completed, and the model's performance was checked using accuracy, precision, recall, and F1-score metrics.

The model evaluation in the trained model was selected on the test set. The confusion matrix and precision-recall curve calculate its performance and accuracy. The 1.0 accuracy shows the best performance in the restoration prediction task.

4.3.2 Forgery Detection Model (Isolation Forest)

The forgery detection task was selected for different detection problems. An Isolation Forest algorithm was used to identify artworks extracted from dimensions, weight, and age.

Model Selection and Features Engineering

The Isolation Forest algorithm was chosen for forgery detection due to its accuracy. It shows irregular data arranged in large datasets and observed separately, making it perfect for detecting potential forgeries.

The features used for restoration prediction were artist ID encoded, age at acquisition, height (cm), width (cm), depth (cm), and weight (kg). These were also used for forgery detection.

Anomaly Detection and Evaluation

The individual Forest was trained on the dataset with a contamination level of 0.01 (indicating that 1% of the data is suspected to be anomalous). After fitting the model, the artworks were simplified as alleged forgeries. A score of -1 was selected as suspected forgeries in artwork with a Forgery Risk. The data of suspected forgeries was accumulated for further analysis to identify patterns. The results were seen using a bar plot that shows the distribution of hypothetical forgeries from artist and medium.

4.4 Summary of Results

The Random Forest Classifier obtained an accuracy of 1.0. It predicts an artwork requires restoration. The main features affecting the model's predictions are Age at Acquisition, Height (cm), and Weight (kg). The separate or Isolation Forest algorithm indicates 1,303 artworks that are considered potential forgeries. Further analysis shows the pattern obtained from the distribution of forgeries by artist and medium. It also provides insights into potential risks within the collection.

5. EDA

In this section, we analyze the idea of exploratory data analysis (EDA) for the museum dataset. This idea explains the distribution of artist demographics, artwork dimensions, and correlations between different numerical features. A detailed explanation of each graph, along with the findings, is given below:

5.1 Distribution of Artists' Birth and Death Years

The following histogram shows the distribution of artists' birth and death years over time. The blue bars and line represent the birth years, while the red bars and line represent the death years.

It is considered that a concentration of artists born between the late 19th and mid-20th centuries, especially around 1950-1960. This peak shows the artist's birth. Similarly, in another case, many artists passed away around the mid-20th century, peaking around the 1990s; it shows the older generation of artists born earlier in the 20th century. The central part of the data is concentrated from the 1800s onwards. This indicates that the dataset predominantly covers modern artists.

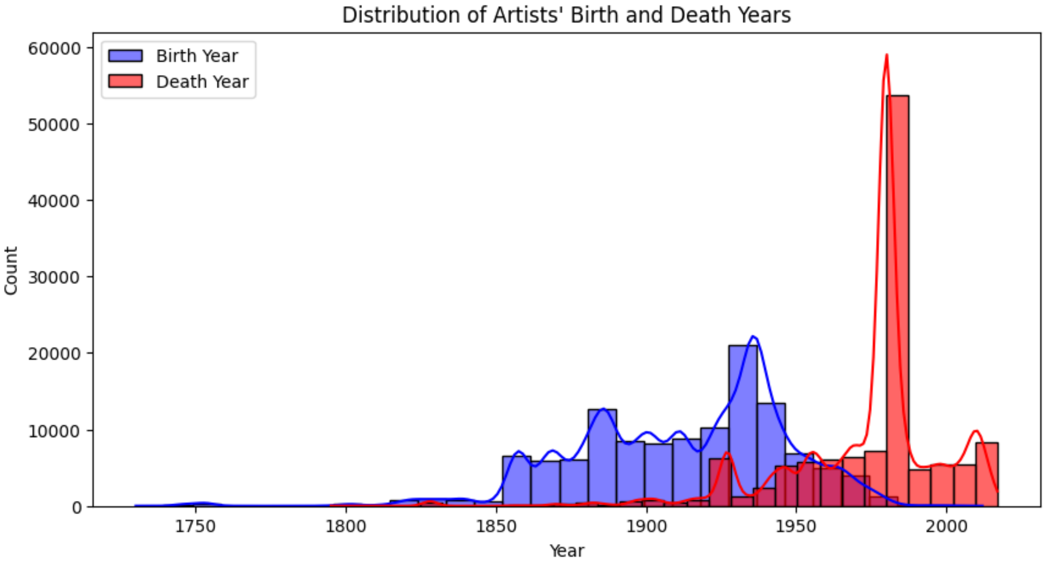


Figure 1. Distribution Graph of artists’ birth and death years

5.2 Distribution of Artwork Dimensions (Height, Width, Depth)

These histograms show the distribution of artworks’ height, width, and depth. The green, orange, and purple bars represent the artwork’s dimensions in centimeters. The majority of artworks are concentrated around smaller dimensions. This height, width, and depth are primarily below 200 cm. There are a few standards with considerably more oversized dimensions, and some large sculptures or installations are suggested in the collection. The depth of most artworks is close to 0, and the most expected artworks, such as paintings and drawings, have very little depth.

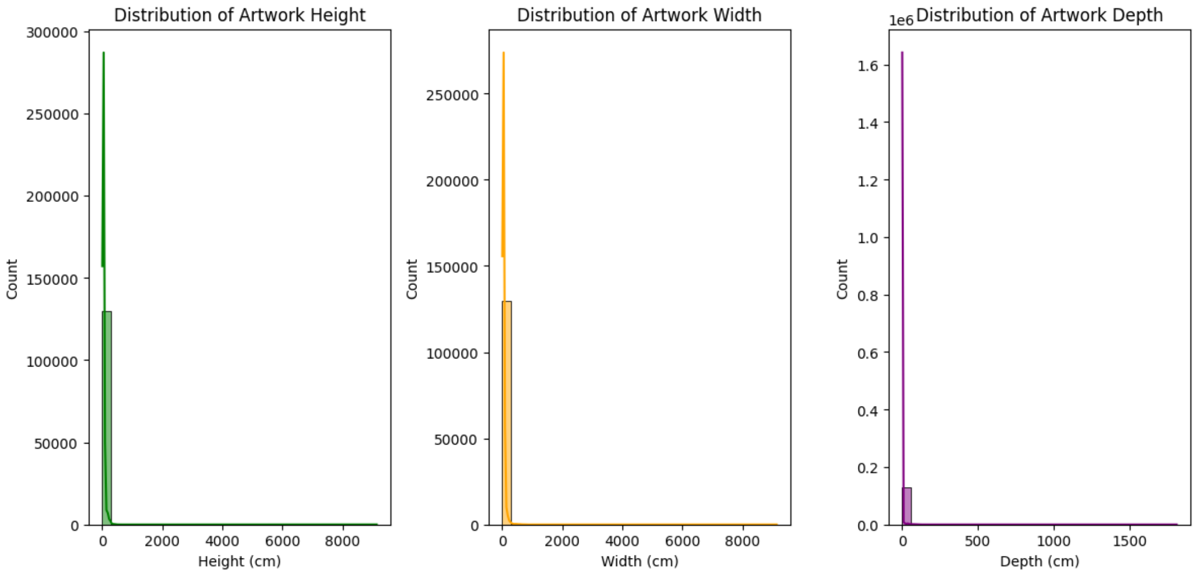


Figure 2. Distribution analysis of features

5.3 Top 10 Nationalities of Artists

This bar plot shows the dataset’s top 10 nationalities of artists. The number of artworks is indicated by artists from each nationality. The dataset is most prominent due to American artists, followed by French, German, British, and Spanish artists. This suggests that the museum collection should focus on Western art, particularly from American and European artists. Italian, Japanese, Swiss, and Russian nationalities appear in the top 10 but are distant and scattered artists compared to the dominant nationalities.

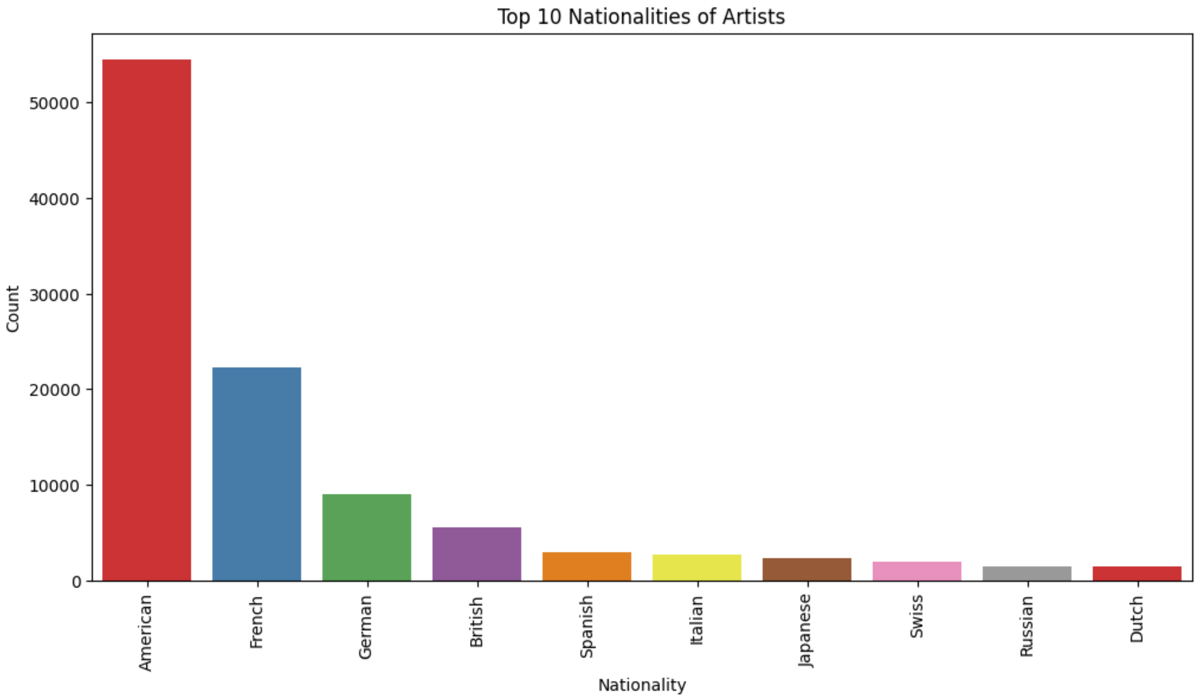
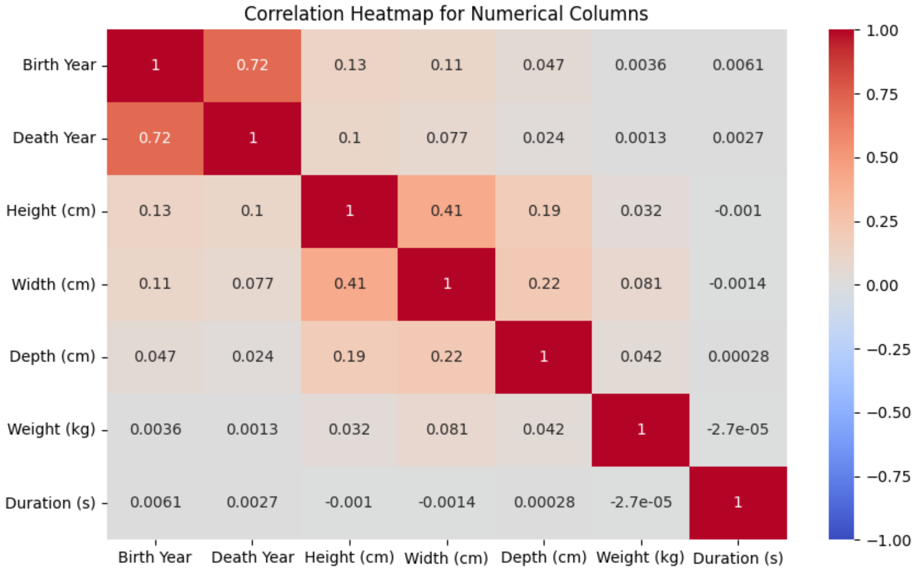


Figure 3. Top 10 Nationalities of Artists

5.4 Correlation Heatmap for Numerical Columns

The heatmap shows the correlation between various numerical features in the dataset, such as birth year, death year, and artwork dimensions. The color scale indicates the strength of the correlation. The darker red indicates a strong positive correlation, and the darker blue indicates a strong negative correlation. The highest correlation is between Birth Year and Death Year (0.72). It is observed that older artists tend to have closer birth and death years. The average correlation between Height (cm) and Width (cm) (0.41). It is suggested that taller artworks are also likely to be more comprehensive, but the correlation is not very strong. Other artwork dimensions, such as depth and weight, show very weak correlations with different features. It tells that each dimension is primarily independent. Interestingly, there is almost no correlation between Weight (kg) and any other features. It indicates that artwork weight is a unique characteristic of this dataset.



	Birth Year	Death Year	Height (cm)	Width (cm)	Depth (cm)	Weight (kg)	Duration (s)
Birth Year	1	0.72	0.13	0.11	0.047	0.0036	0.0061
Death Year	0.72	1	0.1	0.077	0.024	0.0013	0.0027
Height (cm)	0.13	0.1	1	0.41	0.19	0.032	-0.001
Width (cm)	0.11	0.077	0.41	1	0.22	0.081	-0.0014
Depth (cm)	0.047	0.024	0.19	0.22	1	0.042	0.00028
Weight (kg)	0.0036	0.0013	0.032	0.081	0.042	1	-2.7e-05
Duration (s)	0.0061	0.0027	-0.001	-0.0014	0.00028	-2.7e-05	1

Figure 4. Correlation Heatmap for Numerical Columns

5.5 EDA Insights

The dataset strongly focuses on modern or contemporary artists; most data is concentrated in the 19th and 20th centuries. The dimensions of most artworks are relatively small and have a few significant outliers that could represent installations or sculptures. The collection is dominated by American and European artists, mainly from Western countries. Artwork dimensions generally do not show strong correlations except for height and width. This suggests that each feature can provide unique insights for predicting restoration demand or forgery detection.

6. Restoration Prediction & Forgery Detection

The machine learning models provided acute results in both restoration prediction and forgery detection for the museum's artwork collection. The models used were a Random Forest Classifier for restoration prediction and an Isolation Forest for detecting potential forgeries. A detailed explanation is given below.

6.1 Confusion Matrix — Restoration Prediction

The confusion matrix shows the behavior of the restoration prediction model. The matrix exhibits the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values for the model's predictions. So, the matrix is perfectly diagonal and indicates 100% accuracy.

True Negatives (TN)

The model accurately predicted 25,885 artworks that do not require restoration.

True Positives (TP)

The model accurately predicted 168 artworks that do require restoration.

False Positives (FP) and False Negatives (FN)

No incorrect predictions exist (FP = 0, FN = 0), which shows the model's exceptional performance. This perfectly diagonal matrix results from the model's high accuracy. The details are given below.

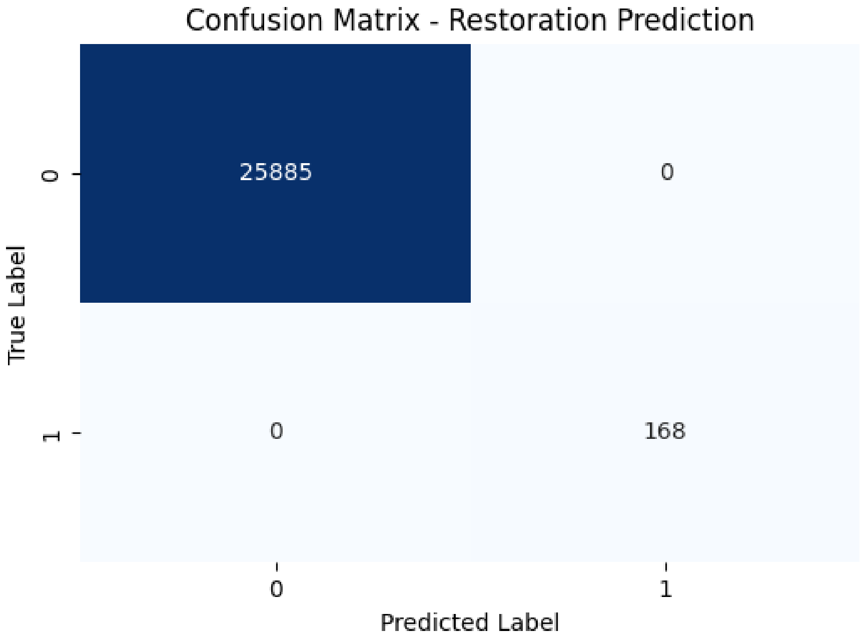


Figure 5. Confusion Matrix

6.2 Precision-Recall Curve

The precision-recall curve estimates the model's ability to handle different classes (in this case, restoration needs). Precision measures how many of the predicted positive instances are positive and recall measures how many actual positive instances were predicted correctly. The curve is nearly perfect, with precision and recall values of 1.0. It is almost the entire range of thresholds. The model shows excellent performance with high precision and recall, even for a small subset of artworks that require restoration. This indicates that the model can detect the true positives with no false positives or negatives.

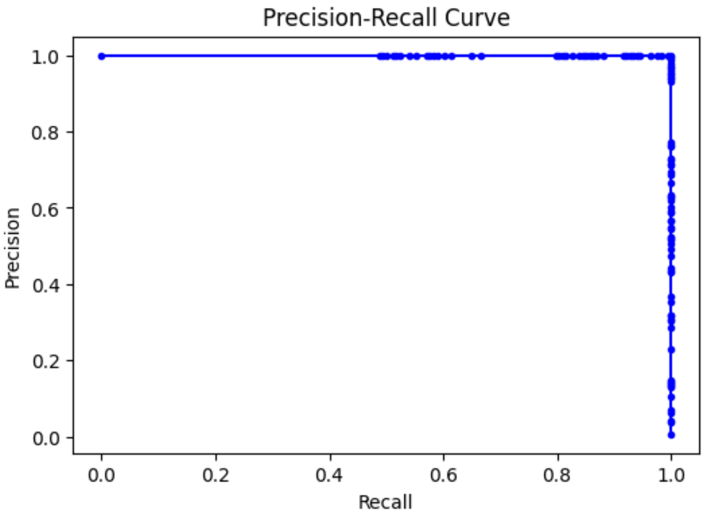


Figure 6. PR Curve

6.3 Feature Importances — Restoration Prediction

This plot shows the relative importance of each feature in the restoration prediction model. Feature importance determines the contribution of each input feature for making predictions.

Age at Acquisition

This feature shows an excellent result with a nearly 100% importance score. It makes the model's decisions dominant. This makes intuitive sense because older artworks are more likely to require restoration.

Meanwhile, the other features, such as Height (cm), Width (cm), and Weight (kg), provide very little to the prediction. Moreover, these may have some importance in other contexts, and their impact on restoration is minimal compared to the age of the artwork. This shows that the model strongly depends on the age of the artwork to predict restoration needs. It aligns with the supposition that older works are more likely to need restoration.

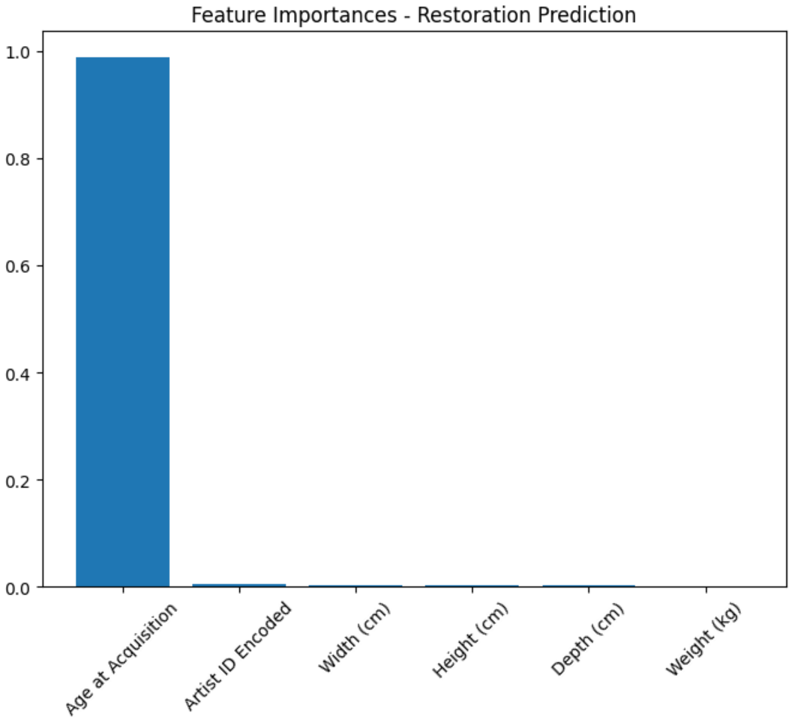


Figure 7. Top Features

6.4 Distribution of Suspected Forgeries

The histogram shows the distribution of forgery scores, which the isolation forest model calculates. The model classifies artworks with a forgery risk score of -1 as potential forgeries, while scores of 1 indicate regular artwork. A small percentage of artworks indicates supposed forgeries (scores of -1), which aligns with expectations. The majority of artworks have a forgery risk score of 1. It means that they are classified as standard models. One thousand three hundred three artworks indicated supposed forgeries. This represents a small but significant collection subset that may require further investigation.

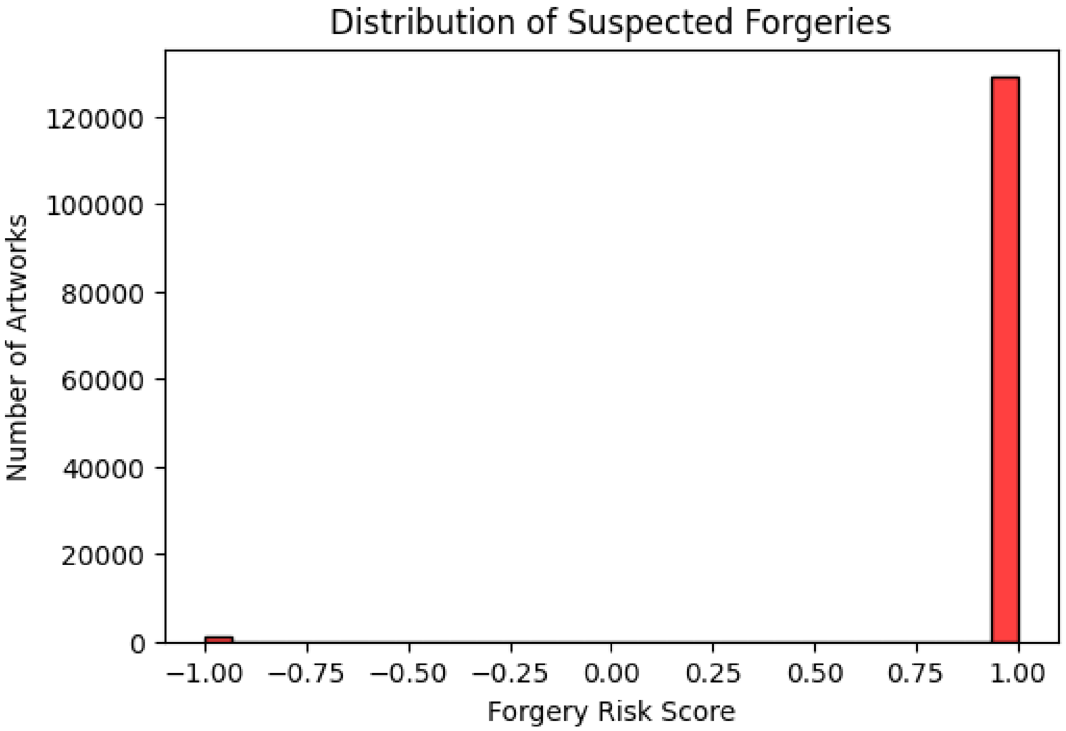


Figure 8. Distribution of Suspected Forgeries

6.5 Classification Report

```
Restoration Prediction Accuracy: 1.0
precision       recall    f1-score      support

0            1.00          1.00          1.00        25885
1            1.00          1.00          1.00          168

accuracy                                   1.00        26053
macro avg               1.00          1.00          1.00        26053
weighted avg            1.00          1.00          1.00        26053
```

Number of suspected forgeries: 1303

6.6 Top 10 Artists with Suspected Forgeries

This bar plot shows the dataset’s top 10 artists with the most suspected forgeries. The Isolation Forest model identifies them. Ludwig Mies van der Rohe has the highest number of alleged forgeries (more than 30 artworks), followed by Franz Erhard Walther and Frank Lloyd Wright. The classification of suspected forgeries across these top artists suggests that certain famous artists may be at higher risk of forgery attempts due to the high value of their works.

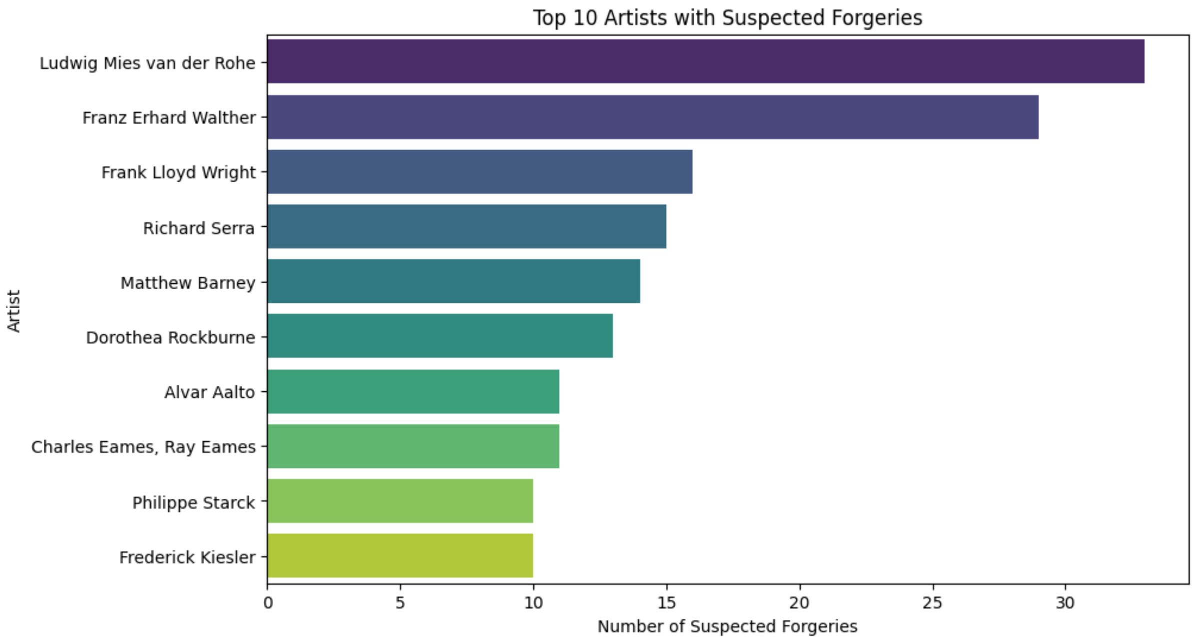


Figure 9. Top 10 Artists with Suspected Forgeries

6.7 Top 10 Mediums in Suspected Forgeries

This bar plot shows the top 10 mediums artists that are most commonly associated with suspected forgeries. Oil on canvas is the most common medium in alleged forgeries. It is followed by synthetic polymer paint on canvas and fabric. This recommends that paintings on canvas are more acceptable in forged than other mediums. Mediums like bronze, painted steel, and wood are the top mediums in suspected forgeries. It indicates that sculptures and installations could also be targets for forgery. This insight intuition allows museums to prioritize investigations into artworks with these mediums, especially if they come from the artists highlighted in the previous graph.

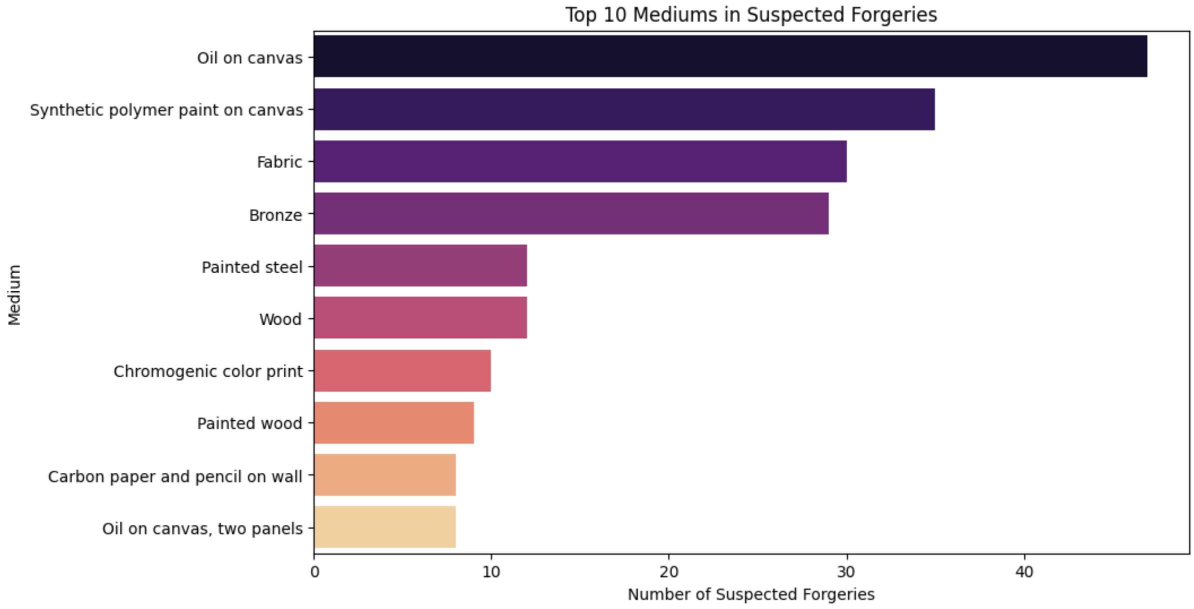


Figure 10. Top 10 Mediums in Suspected Forgeries

6.8 Summary

Restoration Prediction

The Random Forest Classifier accomplished extraordinary feats with an accuracy of 1.0, correctly predicting whether artworks need restoration. The model's precision, recall, and F1-score are all 1.0, with no false positives or negatives. The feature importance analysis imparts that Age at Acquisition was the dominant factor in moving restoration predictions. This correlates with expectations that older artworks are more likely to require maintenance.

Forgery Detection

The Isolation Forest identified 1,303 artworks as potential forgeries. It represents a small fraction of the total collection. The distribution of forgery risk scores shows that most artworks are classified as usual. Ludwig Mies van der Rohe, Franz Erhard Walther, and Frank Lloyd Wright are the artists who have the most suspected forgeries. It required further analysis of their works. The mediums associated with suspected forgeries include oil on canvas, synthetic polymer paint on canvas, and fabric. It indicates that paintings and sculptures in these mediums should be organized for verification.

These results provide actionable decisions for museums, curators, and collection managers. The high accuracy of the restoration prediction model ensures that valuable artworks are identified for restoration promptly, and the forgery detection model offers a data-driven approach to protecting the authenticity of the museum's collection.

7. Discussion and Conclusion

This project aims to develop a comprehensive data-driven solution for predicting artwork restoration needs and detecting potential forgeries in a museum collection. By utilizing machine learning techniques, the project successfully provided an actionable understanding of how museums can manage their collections more efficiently and protect against risks of forgery.

7.1 Discussion

The methodology used in this project imposed on two key machine learning models, i.e., Random Forest Classifier for restoration prediction and Isolation Forest for forgery detection. The results indicate the effectiveness of these models and describe the unique challenges museums face in managing and preserving their collections.

The Random Forest Classifier appears near-perfectly, correctly predicting whether an artwork needs restoration. The feature importance analysis revealed that Age at Acquisition was the most significant predictor of restoration needs. This result aligns with ordinary knowledge in the art preservation field: older artworks are generally more susceptible to wear and tear due to the passage of time, environmental conditions, and the materials used. Other features, such as artwork dimensions and weight, were also included in the model. Their contribution to restoration prediction was negligible, further emphasizing the critical role of age in determining restoration needs.

The restoration model's perfect performance is even more impressive. It suggests the potential for overfitting. The simplicity of the classification problem essentially drives the model's high accuracy. It predicts restoration needs are straightforward when based almost exclusively on the age of the artwork. However, the model provides a reliable first-line tool for museums, curators, and conservation teams to prioritize restoration efforts based on data-driven awareness.

7.1.1 Forgery Detection

The forgery detection task was more complex, and it was difficult to distinguish between authentic and counterfeit works based only on available data. The Isolation Forest model identified 1,303 artworks as potential forgeries, representing a small subset of the total collection. This aligns with expectations, as only a tiny percentage of museum collections are typically subject to forgery risks.

A deeper analysis exhibits patterns in the suspected forgeries. Some famous artists, including Ludwig Mies van der Rohe, Franz Erhard Walther, and Frank Lloyd Wright, were frequently indicated for alleged forgeries. This suggests that artworks by high-profile artists that are similar to those of higher market value are at greater risk of fraud. Identifying specific mediums — such as oil on canvas, synthetic polymer paint, and fabric — commonly associated with forgery attempts also provides valuable insight. These materials, particularly in paintings, are widely forged due to their historical value and facilitate counterfeiters that can replicate their appearance.

The forgery detection model provided meaningful results, but it should be noted that machine learning models alone cannot conclusively determine whether an artwork is a forgery. The model's results should be used as a starting point, and experts, such as art historians and forensic analysts, require further investigation. It indicates artworks can be prioritized for further authentication processes, including provenance research, material analysis, and expressive comparisons.

7.2 Conclusion

This project successfully indicated the utility of machine learning models having two critical challenges in museum collection management. It requires predicting restoration needs and detecting potential forgeries. The high accuracy of the restoration prediction model indicates that data-driven approaches can effectively complement traditional conservation efforts. By focusing on features such as Age at Acquisition, museums can better prioritize restoration efforts. It ensures that older and more fragile artworks receive the required care.

The forgery detection model is valuable for identifying potentially suspicious artworks within extensive collections. By highlighting artists and mediums more frequently associated with forgeries, the model provides curators and experts with a clear starting point for further authentication efforts. Although forgery detection remains challenging, combining machine learning with expert knowledge can significantly enhance a museum’s ability to protect its collection from counterfeit works.

References

Ł. Gałka, P. Karczmarek and M. Tokovarov, (2022). Isolation Forest Based on Minimal Spanning Tree. *IEEE Access,* (10), pp. 74175-74186. doi: 10.1109/ACCESS.2022.3190505.

https://www.analyticsvidhya.com/blog/2021/07/anomaly-detection-using-isolation-forest-a-complete-guide/

https://www.kaggle.com/datasets/momanyc/museum-collection

Liao, L. & Luo, B., (2018). Entropy isolation forest based on dimension entropy for anomaly detection. In *Proceedings of the International Symposium on Intelligent Computing Applications*, pp. 365-376.

Liu, F.T., Ting, K.M. & Zhou, Z., (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data*, *6*(1), pp. 1-39.

Liu, F.T., Ting, K.M. & Zhou, Z.H., (2008, December). Isolation forest. In *Proceedings of the 8th IEEE International Conference on Data Mining*, pp. 413-422.

Yang, Q., Singh, J. & Lee, J., (2019). Isolation-based feature selection for unsupervised outlier detection. In *Proceedings of the Annual Conference of the Prognostics and Health Management Society*, *11*, pp. 1-8.

Copyrights