Revenue Classification and Profitability Analysis of Museums

Alejandra Sevilla
Georgia Institute of Technology, Student ID: #04, Group: #27, asevilla8@gatech.edu

***Abstract* - This report analyzes the financial performance of U.S. museums using the Museum File 2018 dataset provided by the Institute of Museum and Library Services (IMLS). This study investigates factors contributing to museum revenue and profitability, classifying institutions into high- and low-revenue groups. Predictive models, including ridge regression and gradient boosting machines, were employed to classify museums based on geographic region, discipline, and financial indicators. The analysis also identified fundamental factors influencing museum profitability. Logistic regression was found to be an effective model for this. These findings can support policymakers and museum administrators seeking to improve financial health and sustainability.**

INTRODUCTION

Museums are important cultural and educational institutions, yet many face significant financial challenges due to funding constraints, high operational costs, and reliance on external support. The Institute of Museum and Library Services (IMLS) report 'Exhibiting Public Value' highlights the difficulties museums encounter with inconsistent public funding and emphasizes the need for sustainable financial strategies [1]. The recent economic crisis further emphasizes this urgency [2].

This study investigates the factors driving museum revenue and profitability to support long-term sustainability. Using the November 2018 Museum Data Files, which include data on approximately 30,000 museums across the United States, the analysis identifies relevant determinants of financial success. The dataset consists of museum disciplines, geographic locations, finances, and institutional affiliations. Cleaning and preprocessing addressed missing data, high revenue variance, and categorical variable consolidation. Feature engineering, such as creating a Revenue Category, improved classification performance.

Predictive models such as logistic regression, random forest, gradient-boosting machine (GBM), ridge regression, and naive Bayes were employed to achieve the study's objectives. The hyperparameters for each model were tuned to optimize performance, including cross-validation for ridge regression and tree depth for GBM.

The findings revealed patterns in revenue generation across museum disciplines and highlighted the impact of location type and institutional affiliations on financial performance. This research provides recommendations for policymakers and museum administrators.

This report is structured as follows:

- **Abstract:** A summary of the project.
- **Introduction:** Background, motivation, and overview.
- **Problem Statement or Data Sources:** Description of the museum data.
- **Proposed Methodology:** Explanation of data preprocessing and modeling techniques.
- **Analysis and Results:** Findings from the data analysis.
- **Conclusions:** Summary of findings, future work, and lessons learned.

PROBLEM STATEMENT OR DATA SOURCES

I. Data Overview

This project used the Museum Data Files from November 2018 from the Institute of Museum and Library Services (IMLS) [3]. The dataset consists of three separate files, each focusing on different subsets of museums:

- **File 1:** Museums by specific disciplines (e.g., Art, History, Science, etc.).
- **File 2:** General museums and related organizations (e.g., multidisciplinary museums).
- **File 3:** Historical societies and historic preservation organizations.

The three files were combined into a single dataset, resulting in 30,178 rows and 58 variables. The primary variables of interest for this analysis include:

- **Revenue15:** Total revenue.
- **Income15:** Total income.
- **Discipl:** Museum discipline (e.g., art, history, science).
- **Beareg:** Geographic region, based on Bureau of Economic Analysis (BEA) classifications.
- **Locale4:** Urban-rural classification of location.
- **CoLocationResource:** Presence of shared resources, such as being part of a museum district.
- **NonprofitStatus:** Whether a museum operates as a nonprofit organization.
- **UniversityAffiliationFlag:** Museums affiliated with academic institutions.
- **PrivateFoundationFlag:** Whether a museum is linked to private foundations.

II. Data Cleaning

The combined dataset was evaluated for duplicates, missing values, and overall structure. No duplicate rows were

identified. However, several relevant variables contained missing values, including Revenue (9,827 missing values), Income (9,219), Locale (78), and Region (8).

To address these missing values, rows with missing financial data (Revenue or Income) were removed. For geographic variables, missing Locale values were imputed using the most common value within each city, and the remaining rows with missing Locale were excluded. Missing Region values were imputed based on a predefined mapping of State to Bureau of Economic Analysis regions.

Variables considered irrelevant or redundant were removed. Identifiers were excluded because they did not contribute to the analysis, and administrative details like phone numbers or web URLs were removed because they were irrelevant to the research objectives.

After these steps, the final cleaned dataset contained 20,334 rows and 19 variables.

III. Feature Engineering

Some examples of feature engineering performed include numeric region codes were mapped to descriptive labels, such as 'New England' and 'Southeast' in the Region variable, urban-rural classifications in Locale were converted into categories like 'City,' 'Suburb,' 'Town,' and 'Rural.' Income categories were recoded into ordered levels, such as '$1-$9K,' '$10K-$24K,' and '$50M+.' Museum discipline codes were replaced with full descriptive names, such as 'Art Museums' and 'History Museums.'

Two new variables were derived to facilitate segmentation and analysis. The Revenue Category variable classified museums into 'High' or 'Low' revenue groups based on a threshold defined by the 75th percentile of revenue values. Similarly, the Profitability variable identified whether a museum was financially profitable, categorizing institutions as 'Profitable' if their income exceeded their revenue (Income2015 > Revenue2015) and 'NonProfitable' otherwise.

To reduce sparsity in categorical variables, low-frequency or ambiguous values in some variables were consolidated. For example, categories in Colocation Resource were grouped into broader classes, such as 'Museum District' and 'Library/Archives.'

IV. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to examine museums' financial and operational characteristics in the dataset. The goal was to uncover patterns in revenue and profitability, evaluate data distributions, and identify relationships between variables.

Summary statistics (see Appendix A—A1) were used to describe the central tendencies and variability within the data. Visualizations, including boxplots, scatterplots, histograms, and bar charts, were employed to highlight trends, detect outliers, and explore relationships among variables.

Emphasis was placed on analyzing Income and Revenue, focusing on their distributions and associations with categorical variables such as region and discipline.

IVa. Descriptive Statistics

The financial metrics displayed variability, reflecting the diverse economic landscapes of U.S. museums. Income ranged from $0 to $83.18 billion, with a mean of $113.8 million and a median of $1,455, highlighting the presence of extreme outliers. Revenue ranged from $2.13 million to $5.84 billion. Both income and revenue distributions were heavily right-skewed, with most museums operating at lower financial levels (see Appendix A—A2).

These boxplots show the clustering of financial values near zero and the influence of outliers.

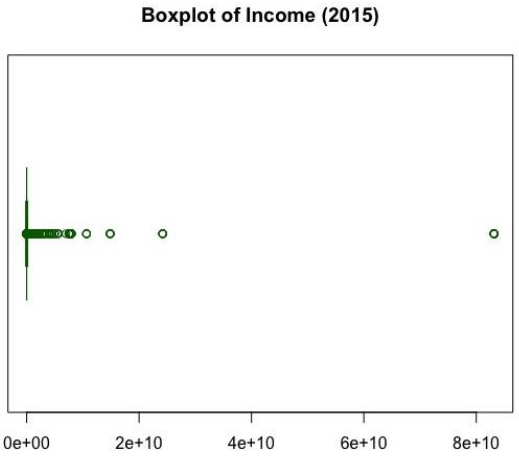

FIGURE I
BOXPLOT OF INCOME (2015)

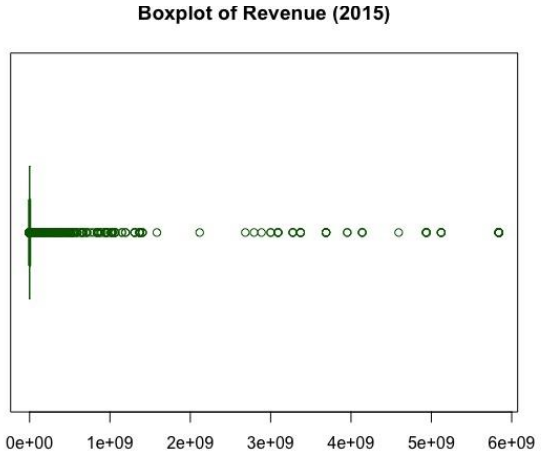

FIGURE II
BOXPLOT OF REVENUE (2015)

IVb. Outliers Analysis

Outliers were identified in Income and Revenue, with income exceeding $80 billion attributed to high-performing institutions such as prominent universities and large museums. These outliers were retained in the dataset because they reflect realistic financial extremes and preserve critical variability necessary for analysis.

The scatterplot shows a strong positive linear relationship between income and revenue, indicating that higher-income museums tend to generate higher revenue.

While most museums align closely with this trend, a few outliers exhibit high incomes or revenue that deviate from the general pattern. The clustering of points along the trend line highlights a stable relationship, reflecting predictable financial behavior for most institutions.

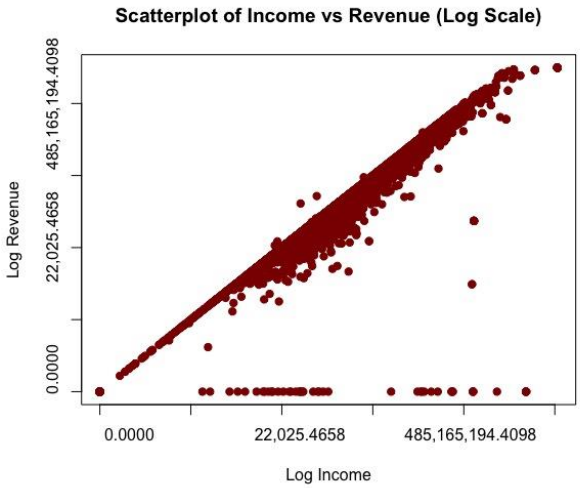

FIGURE V
SCATTERPLOT OF INCOME VS. REVENUE

IVc. Revenue Analysis

The analysis of revenue patterns showed differences across museum disciplines, geographic regions, and organizational characteristics.

Art Museums, History Museums, and General Museums (including Uncategorized institutions) account for many high-revenue museums (see Appendix A—A3). However, despite their large numbers, Historical Societies and Historic Preservation organizations are mainly in the low-revenue category.

There is a concentration of high-revenue museums within a few disciplines, such as Art and General Museums. Most museum disciplines, however, remain in the low-revenue category, emphasizing that only a small fraction of institutions achieve significant income.

Museums in urban areas, particularly those in the Mid-East, South-East, and Great Lakes regions, generate more revenue.

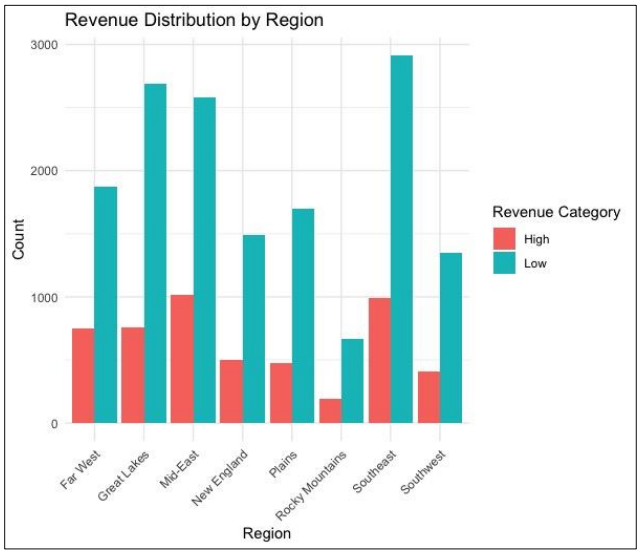

FIGURE VII
REVENUE DISTRIBUTION BY REGION

Despite having many of them, rural museums typically have low revenue due to limited funding and a lower visitor count.

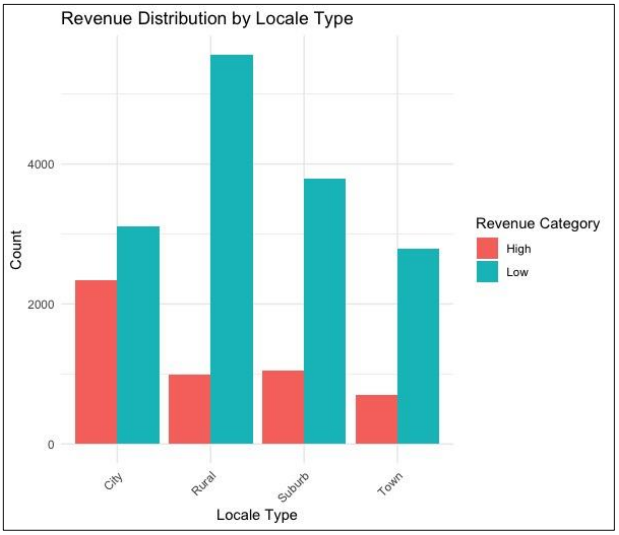

FIGURE VIII
REVENUE DISTRIBUTION BY LOCALE TYPE

Museums affiliated with private foundations have higher revenue, emphasizing the importance of external financial support (see Appendix A—A4). On the other hand, nonprofit museums are disproportionately represented in the low-revenue category, reflecting their reliance on fluctuating visitors and donor funding (see Appendix A—A5).

IVd. Profitability Analysis

Profitability, defined as the difference between income and revenue, was also analyzed to identify patterns within the dataset.

Most museums were found to be non-profitable (see Appendix A—A6), suggesting they face challenges in generating a profit. Non-profitability was more noticeable among rural museums and Historical Societies. Urban museums, especially those categorized as art and children's museums, displayed a higher likelihood of profitability.

Profitability was correlated with organizational characteristics, including nonprofit status and affiliations with private foundations. Museums affiliated with private foundations showed higher profitability rates.

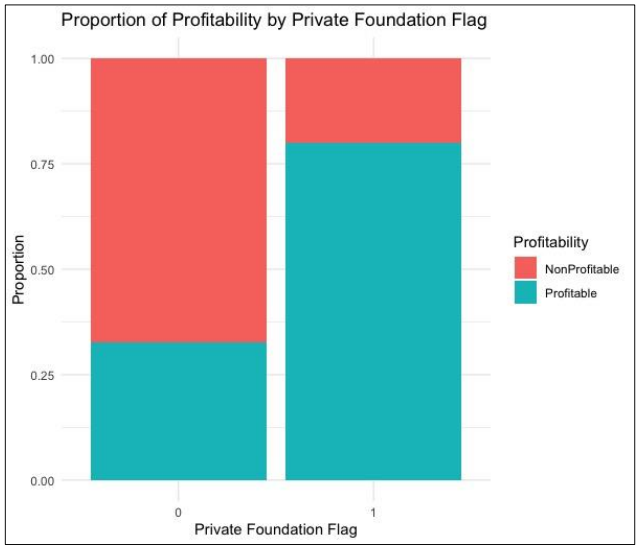

FIGURE XII
PROPORTION OF PROFITABILITY BY PRIVATE FOUNDATION FLAG

Geographic trends show regions like New England and the Southeast have more profitable institutions.

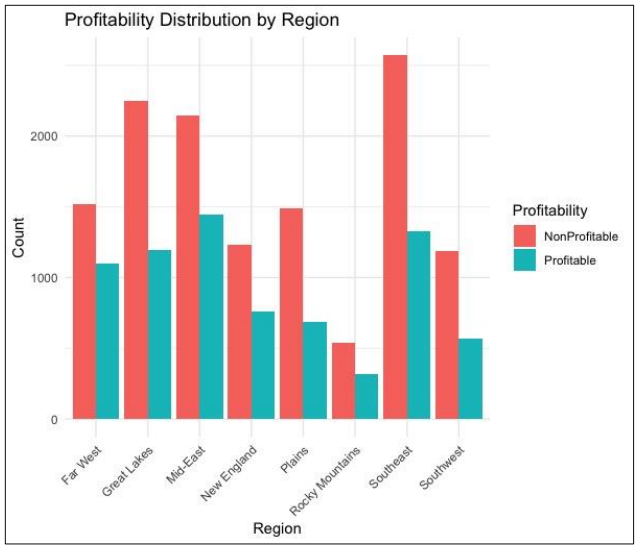

FIGURE XIII
PROPORTION OF PROFITABILITY BY REGION

This analysis highlights the critical impact of organizational features and location on profitability. Museums in urban areas, particularly those with private foundation affiliations, are better positioned to achieve profitability. On the other hand, rural museums and nonprofits face challenges in sustaining financial stability.

IVe. Top Performers and Geographic Distribution

The top 10 museums by revenue and income were identified (see Appendix A—A7). These include standalone museums and universities with affiliated museum collections or exhibits.

Institutions such as Harvard University, Johns Hopkins University, and Stanford University lead in revenue, emphasizing the scale and financial impact of their educational and cultural outreach [4]. In addition to these universities, the list includes dedicated museums such as botanical gardens and art collections.

The geographic distribution (Appendix A—A8) showed that museums with high revenue are predominantly located in urban and densely populated areas, particularly along the East Coast, Midwest, and West Coast regions. This could suggest higher financial sustainability in areas with larger populations and tourist attractions. Although high-revenue museums are concentrated in urban centers, low-revenue museums are also present in these areas.

In contrast, rural and less populated areas, especially in the western U.S., host fewer museums, and those present often exhibit lower revenue levels. This shows gaps in accessibility to cultural institutions, suggesting that geographic location influences the availability of resources and financial performance.

PROPOSED METHODOLOGY

This research applies the following methodology to classify U.S. museums into high- and low-revenue categories and determine their profitability based on financial and organizational characteristics.

I. Data Preprocessing for Model Building

To prepare the dataset for modeling, a preprocessing strategy was implemented:

- Categorical variables such as Revenue Category, Museum Discipline, Region, Locale Type, and institutional flags were encoded because most machine learning models can only process numerical data.
- Low-frequency levels across categorical variables like Museum Discipline and Locale Type were combined into an ‘Other’ category to reduce sparsity.
- The dataset was split into training (70%) and testing (30%) sets for model development and evaluation. The split preserved the distribution of the target variable (Revenue Category).
- Factor levels in the testing set were aligned to those in the training set to prevent prediction mismatches.

II. Model Development

This study employed different machine learning models to classify museums into revenue and profitability categories. Each model was selected based on specific characteristics [5], as detailed below:

- **Logistic Regression:** Estimates the probability of outcomes, provides feature importance through odds ratios, and sets a benchmark for more complex models.
- **Random Forest:** Combines multiple decision trees and handles complex patterns and interactions.
- **Ridge Regression:** Reduces overfitting by shrinking coefficients, making it ideal for datasets with highly correlated features while retaining all features and regularizing their impact.
- **Gradient Boosting Machine:** Improves predictions by correcting errors from previous models, adjusting based on residuals, and performing well with nonlinear relationships.
- **Naive Bayes:** Handles high-dimensional data efficiently, is quick to implement, and provides probability distributions as output.

III. Hyperparameter Optimization

Hyperparameter tuning was applied in the methodology to improve model performance and obtain accurate predictions. Each model's parameters were optimized based on its characteristics and requirements, as described below:

- **Logistic Regression:** The classification threshold was tuned from 0.1 to 0.9 to maximize the F1 Score while balancing precision and recall. The final threshold was selected based on the highest observed F1 score.
- **Random Forest:** Hyperparameter tuning focused on:
  - mtry: Number of features considered at each split, optimized using grid search over values of 3, 6, and 9.
  - Cross-validation (5-fold) was employed to select the optimal mtry value for the model.
- **Ridge Regression:** The lambda penalty term, which controls the regularization level, was optimized using 10-fold cross-validation. The optimal value minimized cross-validation error.
- **Gradient Boosting Machine:** The following hyperparameters were tuned:
  - Internal 5-fold cross-validation (cv.folds) set to 5.
- **Naive Bayes:** This model was evaluated using default hyperparameters due to its. Predictions were assessed based on class probabilities.

TABLE II
SUMMARY OF HYPERPARAMETER TUNING

Model	Parameter	Range/Value	Tuning Method
Logistic Regression	Classification Threshold	0.1 to 0.9	F1-Score Maximization
Random Forest	mtry	3, 6, 9	Grid Search with 5-Fold CV
	ntree	300	Fixed Value
Ridge Regression	lambda	Automatically Determined	10-Fold Cross-Validation
Gradient Boosting Machine	n.trees, interaction.depth	500, 3	5-Fold Cross-Validation
	shrinkage	0.01	Fixed Value
Naive Bayes	Default Parameters	Not Tuned	Evaluated with AUC/ROC

IV. Performance Evaluation

To assess the implementation of the classification models, performance evaluation was completed in two stages:

1. **Single Training-Test Split Evaluation**: Each model was evaluated on a single 70%/30% dataset split. This evaluation provided an initial assessment of model performance and facilitated a preliminary comparison across the different models.
2. **Cross-Validation Evaluation**: The models were further evaluated using cross-validation. For each model, fifty iterations of 70%/30% random splits were performed, and the metrics were averaged across all iterations to account for variability due to data partitioning. This approach divides the dataset into multiple subsets, iteratively training on a combination of subsets and validating on the remaining ones. Cross-validation reduces the risk of overfitting and shows the models' ability to maintain consistent performance on unseen data [6].

Multiple evaluation metrics were used to capture diverse aspects of model performance [7]:

- **Accuracy**: The proportion of all predictions (both positive and negative) that are correct, with higher values indicating better performance.

$$Accuracy = \frac{True\ Positives\ (TP) + True\ Negatives\ (TN)}{Total\ Predictions}$$

- **Area Under the Curve (AUC)**: A measure of the model's ability to distinguish between positive and negative classes. The AUC quantifies the likelihood that a randomly chosen positive instance is ranked higher than a negative instance. AUC values closer to 1 indicate better model performance.

- **Precision**: The proportion of correct positive predictions. Higher values indicate better precision in identifying positive cases.

$$Precision = \frac{TP}{TP + False\ Positives\ (FP)}$$

- **Sensitivity (Recall)**: The proportion of actual positive instances correctly identified by the model. High sensitivity means that most true positive cases are detected.

$$Sensitivity = \frac{TP}{TP + False\ Negatives\ (FN)}$$

- **Specificity**: The proportion of actual negative instances correctly identified. Higher values indicate better identification of negative cases, reducing false positives.

$$Specificity = \frac{TN}{TN + FP}$$

- **F1 Score**: The harmonic mean of Precision and Recall, balancing the trade-off between the two metrics. It is valuable when dealing with imbalanced datasets, as it emphasizes the model's performance on the positive class. Higher values represent better performance.

$$F1\ Score = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity}$$

Given the challenges presented by the dataset's imbalanced revenue and profitability distribution, these performance metrics were selected to evaluate model performance.

V. Profitability Model

This study also included a separate analysis to predict museum profitability, defined as the difference between income and revenue. A binary variable, Profitability, was created, and museums were categorized as 'Profitable' if their income exceeded revenue and 'Non-Profitable' otherwise.

This section outlines the methodology followed to develop predictive models for profitability.

Va. Feature Engineering

A separate binary classification model was developed to investigate museum profitability. Profitability was defined as whether the museum's income exceeded its revenue, and a new binary variable, Profitability, was created to represent this.

Vb. Dataset Preparation

The dataset was split into training (70%) and testing (30%) sets, maintaining the distribution of the target variable across both subsets to get a consistent representation of profitable and non-profitable museums in both data partitions. Factor levels in the test set were aligned to those in the training set to prevent mismatches during prediction.

Vc. Model Development

Four machine learning algorithms, Logistic Regression, Random Forest, Ridge Regression, and Naive Bayes, were trained to predict profitability. These models were selected for their complementary behavior in handling categorical variables, regularization, and probabilistic classification. Hyperparameter tuning for each model mirrored the approach used in revenue classification:

- **Logistic Regression**: A classification threshold was optimized to balance sensitivity and precision.
- **Random Forest**: The model was trained using 300 trees (ntree = 300) to reduce prediction variance. The mtry parameter, which determines the number of features considered at each split, was set to 4.
- **Ridge Regression**: The regularization parameter (lambda) was optimized using cross-validation to mitigate overfitting and account for multicollinearity among predictors.
- **Naive Bayes**: Given its adaptability for categorical data, this probabilistic model was evaluated using its default parameters.

Vc. Model Evaluation and Cross-Validation

A two-step evaluation approach to evaluate model performance for profitability. Initially, each model was assessed using a single 70%/30% training-test split, providing baseline performance metrics. This initial evaluation allowed for a starting comparison of models.

Cross-validation was used further to validate the results and account for the variability.

ANALYSIS AND RESULTS

This section presents the results of the predictive models for revenue classification and profitability prediction. Each subsection analyzes the model's performance, highlights important factors found, and examines the findings.

I. Revenue Classification Analysis

The goal of the revenue classification model was to categorize museums into high- and low-revenue groups based on their financial data and institutional characteristics. The evaluated models included logistic regression, random forest, ridge regression, and GBM.

Ia. Single Training-Test Split Results

A separate binary classification model was developed to investigate museum profitability. Profitability was defined as whether the museum's income exceeded its revenue.

- **Logistic Regression**: The classification threshold was fine-tuned to maximize the F1 Score while balancing precision and recall for high-revenue museums. Due to the imbalanced distribution of high-revenue and low-revenue museums, the analysis identified an optimal threshold of 0.1, deviating from the default of 0.5.

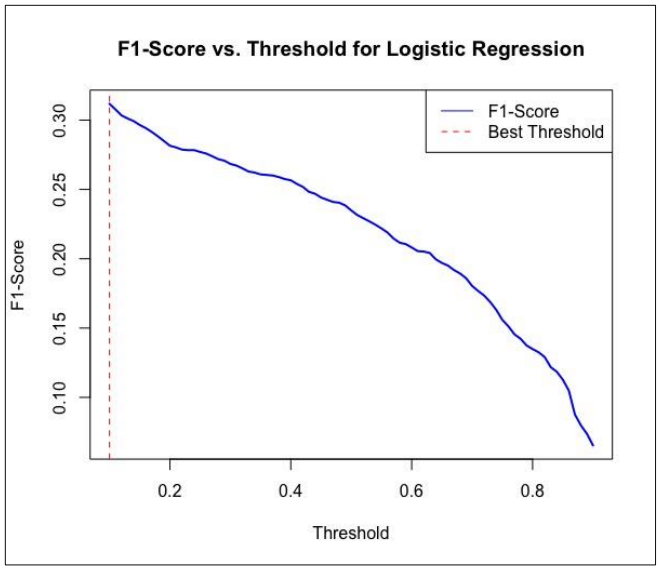

FIGURE XVI
F1-SCORE VS. THRESHOLD FOR LOGISTIC REGRESSION

- **Random Forest**: The optimal mtry value, representing the number of features considered at each split, was determined to be 6 after cross-validation. The ROC curve identified the optimal classification threshold as 0.045, deviating from the standard 0.5 due to the imbalanced dataset. This adjustment improved the model's ability to identify high-revenue museums.

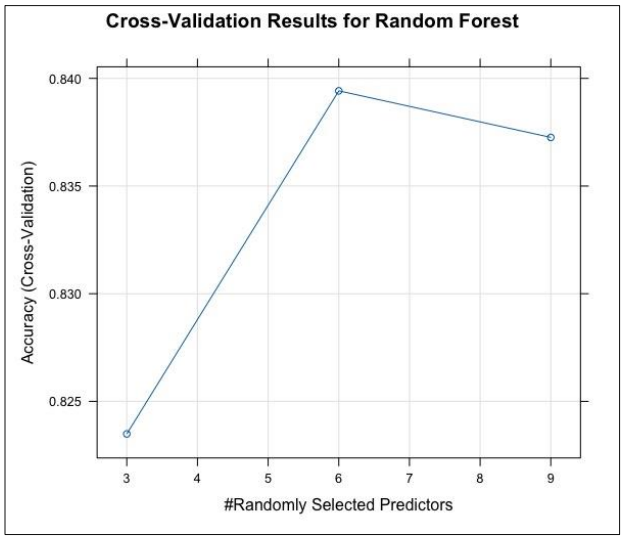

FIGURE XVII
CROSS-VALIDATION RESULTS FOR RANDOM FOREST

- **Ridge Regression**: This model addressed the challenge of multicollinearity by regularizing the coefficients. Cross-validation determined the optimal value for the regularization parameter (lambda), with the best-performing lambda identified as 0.0158.

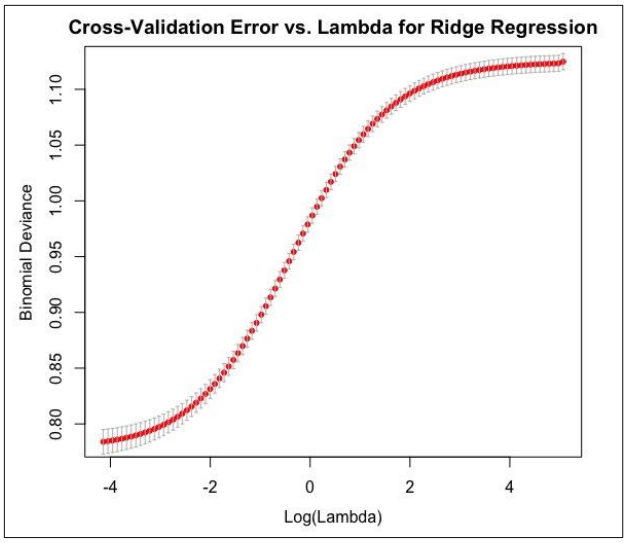

FIGURE XVIII
CROSS-VALIDATION ERRORS VS. LAMBDA REGRESSION

- **GBM**: The model was trained with 500 trees, using a shrinkage parameter of 0.01 and a tree depth of 3. Cross-validation identified the optimal number of trees at approximately 500 iterations. The cross-validation error decreased consistently before stabilizing.

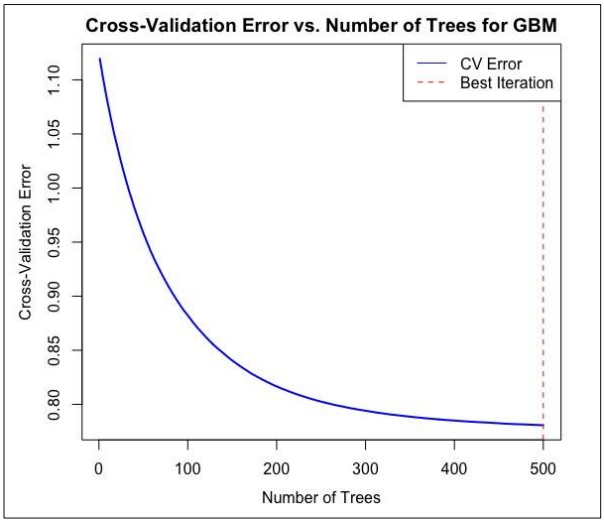

FIGURE XIX
CROSS-VALIDATION ERRORS VS. NUMBER OF TREES FOR GBM

- **Naive Bayes**: The model's initial estimate of the likelihood of each target class (‘low revenue’ and ‘high revenue’), based only on their relative frequencies in the dataset without considering input features, reflected the dataset's imbalance: approximately 75% of museums were classified as low revenue, and 25% as high revenue. These prior probabilities were incorporated to optimize classification performance.

The predictive models were evaluated, with Ridge Regression achieving the highest accuracy (0.8426), precision (0.9709), and specificity (0.9025), effectively minimizing false positives while maintaining high classification accuracy. Naive Bayes had the best AUC (0.8570) and F1 Score (0.8953), making it fit for balanced classification of both positive and negative classes. GBM also performed well, with accuracy (0.8398), precision (0.9657), and F1 Score (0.8469).

Random Forest achieved the highest sensitivity (0.8081), effectively identifying true positives. However, it had lower precision (0.7193) and specificity (0.6269. Logistic Regression performed poorly, with an accuracy of 0.1898, sensitivity of 0.0070, and specificity of 0.3130.

TABLE II

SINGLE TRAINING-TEST SPLIT MODELS PERFORMANCE RESULTS

Metric	Logistic Regression	Random Forest	Ridge Regression	GBM	Naive Bayes
Accuracy	0.1898	0.8298	0.8412	0.8398	0.8356
AUC	0.1898	0.8137	0.8375	0.8435	0.8570
Precision	0.7384	0.7797	0.8242	0.9657	0.9373
Sensitivity	0.0070	0.4453	0.4635	0.4623	0.5305
Specificity	0.3130	0.9580	0.9670	0.9004	0.8953
F1 Score	0.8432	0.5666	0.5932	0.8469	0.8953

Ib. Cross-Validation Results

Cross-validation, which averaged model performance over 50 iterations with different data splits, showed Ridge Regression to be the most accurate model (0.8412). It also exhibited the highest precision (0.8242) and specificity (0.9670), indicating its success in minimizing false positives while maintaining accuracy. GBM exhibited similar performance, with an accuracy of 0.8396 and the highest specificity (0.9707).

Random Forest, with an accuracy of 0.8298, had the highest sensitivity (0.4453) but lower specificity (0.9580) and precision (0.7797), suggesting a tendency to over-classify some museums as high revenue. Naive Bayes also performed well, achieving an accuracy of 0.8321 and the highest AUC (0.8354), showing a balanced classification ability.

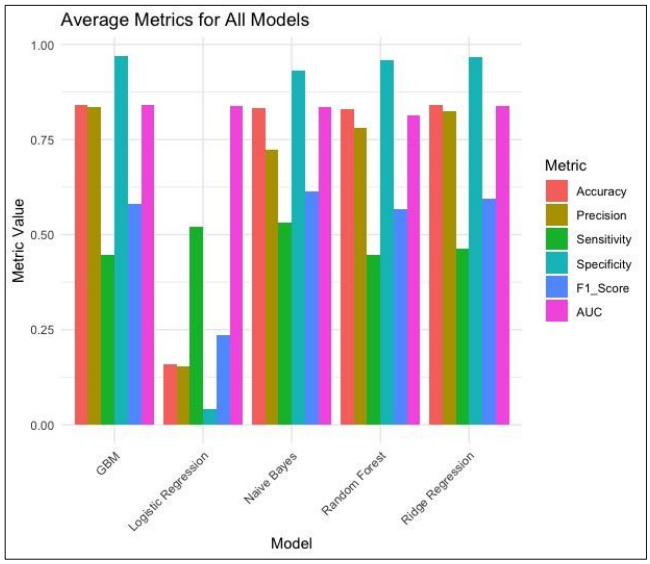

FIGURE XX

AVERAGE METRICS FOR ALL MODELS – CROSS VALIDATION

In contrast, logistic regression underperformed, with significantly lower accuracy (0.1603), sensitivity (0.5193), and specificity (0.0407).

TABLE III

CROSS-VALIDATION MODELS PERFORMANCE RESULTS

Metric	Logistic Regression	Random Forest	Ridge Regression	GBM	Naive Bayes
Accuracy	0.1603	0.7859	0.8426	0.8396	0.8321
AUC	0.8384	0.5554	0.8430	0.8398	0.8354
Precision	0.1529	0.7193	0.9709	0.8354	0.7233
Sensitivity	0.5193	0.8081	0.4577	0.4466	0.5321
Specificity	0.0407	0.6269	0.9025	0.9707	0.9321
F1 Score	0.2362	0.8265	0.8425	0.5820	0.6131

Ic. Model Selection and Variable Importance Analysis

This section discusses the models selected for revenue classification and interprets the importance of features contributing to their predictions. Ridge Regression and GBM were identified as the best models due to their consistent performance across single training-test splits and cross-validation. The importance of relevant variables was analyzed to understand the factors influencing museum revenue.

Ridge Regression was the best-performing model, showing high accuracy, precision, and specificity across evaluation methods. Its regularization of coefficients prevents overfitting while retaining all variables.

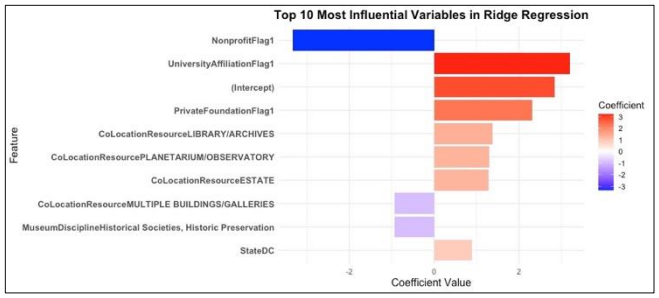

FIGURE XXI
TOP 10 MOST INFLUENTIAL VARIABLES IN RIDGE REGRESSION

GBM was also identified as a proper model with high precision, specificity, and F1 Score. By iteratively correcting errors, GBM captures complex relationships within the data.

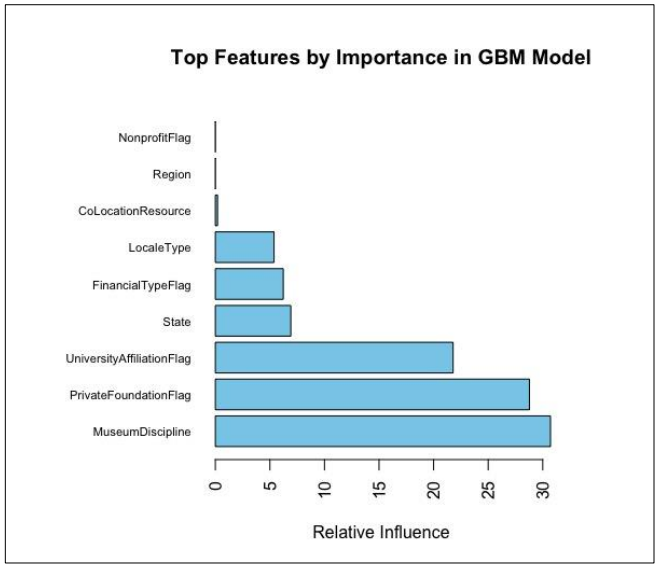

FIGURE XXII
TOP FEATURES BY IMPORTANCE IN GBM MODEL

Both Ridge Regression and GBM emphasized similar influential variables:

- **Museum Discipline:** Certain museum types, such as art, history, and science, are more likely to generate higher revenue.
- **Private Foundation Affiliation:** This was strongly correlated with high revenue, likely due to consistent financial sponsorship, larger donor bases, and grant access.
- **University Affiliation:** This was associated with higher revenue, possibly due to shared resources, partnerships, and established infrastructure.
- **State:** Geographic location was a key predictor, with certain states, like Washington D.C., likely due to tourism and cultural concentration.
- **Financial Type:** Museums categorized by third-party aggregators showed distinct financial profiles that influenced revenue.
- **Locale Type:** Urban museums typically generate higher revenue due to larger visitor bases [8].
- **Nonprofit Status**: Nonprofit museums exhibited a strong negative correlation with revenue, likely due to their reliance on limited funding sources like grants and donations.
- **Co-located Resources**: Museums sharing facilities with libraries, planetariums, or other institutions showed higher revenue, emphasizing the advantages of collaborative environments.

II. Profitability Prediction Analysis

The profitability model intended to identify whether a museum's income exceeded its revenue, categorizing institutions as 'Profitable' or 'NonProfitable.' Models evaluated include Logistic Regression, Random Forest, Ridge Regression, and Naive Bayes.

IIa. Model Selection and Variable Importance Analysis

The profitability classification models performed consistently during the single training-test split evaluation, with accuracy ranging from 0.6586 (Ridge Regression) to 0.6640 (Random Forest). F1 scores were closely aligned, with Ridge Regression achieving the highest at 0.8067, followed by Logistic Regression (0.8043), Random Forest (0.8039), and Naive Bayes (0.7973).

Precision values were also similar, going from 0.7293 (Ridge Regression) to 0.7362 (Naive Bayes), highlighting consistent identification of profitable museums. Sensitivity, or the ability to detect all profitable cases, was highest for Ridge Regression at 0.9025, while specificity, indicating the accurate classification of non-profitable museums, was highest for Naive Bayes at 0.4557.

TABLE IV
SINGLE TRAINING – TEST SPLIT MODELS PERFORMANCE RESULTS

Metric	Logistic Regression	Random Forest	Ridge Regression	Naive Bayes
Accuracy	0.6591	0.6640	0.6586	0.6626
Precision	0.7306	0.7348	0.7293	0.7362
Sensitivity	0.8945	0.8874	0.9025	0.8694
Specificity	0.4238	0.4405	0.4146	0.4557
F1 Score	0.8043	0.8039	0.8067	0.7973

IIb. Cross-Validation Results

The cross-validation evaluation over 50 iterations showed Ridge Regression with the highest sensitivity (0.9651). Logistic Regression had the best overall accuracy (0.6763) and the highest F1 score (0.8147).

Naive Bayes and Random Forest also performed well, with F1 scores of 0.7877 and 0.8117, respectively. Naive Bayes achieved high sensitivity (0.9571) and precision

(0.6693). Compared to logistic regression, Random Forest had high sensitivity (0.9443) but lower accuracy (0.6383).

Despite their high sensitivity, Ridge Regression and Naive Bayes struggled with specificity (0.1962 and 0.1739, respectively), making it harder to identify non-profitable museums correctly. Logistic Regression, with the highest specificity (0.4484), provided the most balanced classification performance.

TABLE V
CROSS-VALIDATION MODELS PERFORMANCE RESULTS

Metric	Logistic Regression	Random Forest	Ridge Regression	Naive Bayes
Accuracy	0.6763	0.6383	0.5807	0.5654
Precision	0.7411	0.7118	0.6771	0.6693
Sensitivity	0.9043	0.9443	0.9651	0.9571
Specificity	0.4484	0.3322	0.1962	0.1739
F1 Score	0.8147	0.8117	0.7959	0.7877

IIc. Model Selection and Variable Importance Analysis

Logistic Regression was chosen as the optimal model for predicting museum profitability due to its balanced performance across metrics during cross-validation (Appendix C—C1).

The model's coefficients show how institutional, geographic, and operational features contribute to profitability. Relevant findings include:

- **Nonprofit:** Museums classified as nonprofits were less likely to be profitable. This highlights the financial challenges faced by nonprofits, which often rely on external funding that may only partially cover operating costs.
- **University Affiliation:** University-affiliated museums showed a strong tendency toward profitability. This suggests that institutional support, resource sharing, and stable partnerships contribute to financial sustainability.
- **Co-Located Resources:** Museums near planetariums, observatories, or estates were more profitable. These specialized or shared facilities likely attract more visitors or offer cost-sharing benefits.
- **State-Level Influence:** Museums in certain states, such as Virginia, Florida, and the Carolinas, were more likely to be profitable. This may reflect variations in regional policies, funding structures, or visitor engagement.
- **Regional Variation:** Museums in New England and the Southwest also showed increased profitability, suggesting that geographic factors and regional support play a role.
- **Locale Type:** Urban museums were more likely to be profitable than those in rural or suburban areas, likely due to differences in visitor numbers, donor bases, and funding availability.

CONCLUSION

This study analyzes the Museum File 2018 dataset to help us understand the financial health of U.S. museums. It identifies factors influencing museum revenue and profitability and makes practical recommendations for administrators and policymakers.

Ridge Regression and GBM were the best-performing models for distinguishing between high- and low-revenue museums. These models had the best accuracy, specificity, and F1 Score, managing the dataset's complexity and addressing multicollinearity.

The primary revenue drivers included museum discipline, affiliation with private foundations, geographic location, and locale type. Museums focused on art, history, and science consistently exhibited higher revenue levels, reflecting the importance of their appeal and operational scale. Urban institutions also showed advantages, likely benefiting from larger visitor bases and stronger funding networks.

Logistic Regression was shown to be the most balanced model for profitability prediction, with a trade-off between sensitivity, specificity, and interpretability.

University-affiliated museums and those co-located with specialized resources, such as planetariums, had significantly higher odds of profitability. These findings show the value of institutional backing and resource sharing in boosting financial performance. In contrast, nonprofit museums and those in rural areas faced financial challenges, struggling to sustain profitability.

Ia. Future Work

This research establishes a foundation for further exploration, addressing current limitations and extending the scope of the study. The following areas highlight opportunities for future work:

- **Increased Feature Engineering:** Incorporating additional variables such as visitor statistics, membership fees, and event revenues offers a deeper understanding of museum finances. Data on seasonal variations, marketing expenses, or exhibit popularity could also contribute.
- **Time-Series Analysis:** Examining revenue and profitability trends over multiple years could explain how external factors such as economic recessions, policy changes, or demographic shifts affect museum sustainability. Time-series models like ARIMA could be employed to forecast future financial performance, helping museums anticipate challenges and optimize long-term strategies.
- **Clustering and Segmentation:** Applying clustering techniques, such as k-means or hierarchical clustering, could group museums with similar financial, geographic, or operational characteristics. These clusters could help identify common challenges and strengths among

museums in each group, enabling more customized recommendations.

- **Qualitative Data:** Future studies could integrate qualitative data, such as museum mission statements, visitor reviews, or employee feedback, to provide context for the quantitative findings. Natural language processing (NLP) could analyze this unstructured data, showing sentiment trends or recurring themes that influence financial performance.

Ib. Lessons Learned

This project has been both challenging and rewarding, providing perspective into the data mining process and its applications. The most important lessons learned are outlined below:

i. Addressing missing values, consolidating categorical variables, and mitigating outliers were critical steps in preparing the data for analysis, and these tasks formed a considerable portion of the project.
ii. The imbalance in target variables for revenue classification and profitability prediction highlighted the importance of using techniques like threshold optimization, cross-validation, and alternative evaluation metrics to achieve better results.
iii. Cross-validation was necessary to guarantee accurate model evaluation. It highlighted the importance of averaging performance across multiple iterations to reduce the influence of any single data split and increase confidence in the results.
iv. This project reinforced the iterative nature of data science, highlighting the need for adaptability and a problem-solving mindset throughout the research process. As new information was obtained from the data analysis, initial assumptions often required refinement.
v. The project emphasized the role of effective communication in data science. Clear presentation of complex analyses helps understand and translate findings into actionable information.

APPENDIX A – EXPLORATORY DATA ANALYSIS

A-1. Summary of Dataset Statistics

TABLE I
SUMMARY STATISTICS OF MUSEUM DATASET

Variable	Description	Summary Statistics
CoLocationResource	Shared resources among museums	Character data; categorical
NTEECode	Classification by nonprofit purpose	Character data; low-frequency and unclassified values identified
MuseumName	Names of museums	Character data; total: 20,334
City	Cities where museums are located	Character data; categorical
State	States where museums are located	Character data; categorical
IncomeCategoryCod	Categorized income levels	Most frequent: '$0' (9,999); less frequent: '$50M+' (837)
Income2015	Total income (USD, 2015)	Min: $0; Max: $83.18B; Mean: $113.8M; Median: $1,455; 1st Quartile: $0; 3rd Quartile: $185K
Revenue2015	Total revenue (USD, 2015)	Min: -$2.13M; Max: $5.84B; Mean: $21.69M; Median: $460; 1st Quartile: $0; 3rd Quartile: $151.6
TaxExemptionDate	Tax-exemption dates for nonprofits	Character data; categorical
InstitutionName	Institutional names	Character data; total: 20,334
Longitude	Geographic longitude	Range: -166.54 to -66.98; Mean: -90.15; Median: -86.13; 23 missing values
Latitude	Geographic latitude	Range: 19.35 to 68.24; Mean: 39.16; Median: 40.04; 23 missing values
Region	Bureau of Economic Analysis regions	Character data; categorical
LocaleType	Urban-rural classification	Character data; categorical
NonprofitFlag	Indicates nonprofit status	Min: 0; Max: 1; Mean: 0.9999; Median: 1
PrivateFoundationFlag	Indicates affiliation with private foundations	Min: 0; Max: 1; Mean: 0.0786; Median: 0
FinancialTypeFlag	Categorizes financial characteristics	Min: 0; Max: 1; Mean: 0.2878; Median: 0
UniversityAffiliationFlag	Indicates university affiliation	Min: 0; Max: 1; Mean 0.0461; Median: 0
MuseumDiscipline	Type of museum (Art, History, Science, etc.)	Character data; categorical

A-2. Revenue and Income Distributions

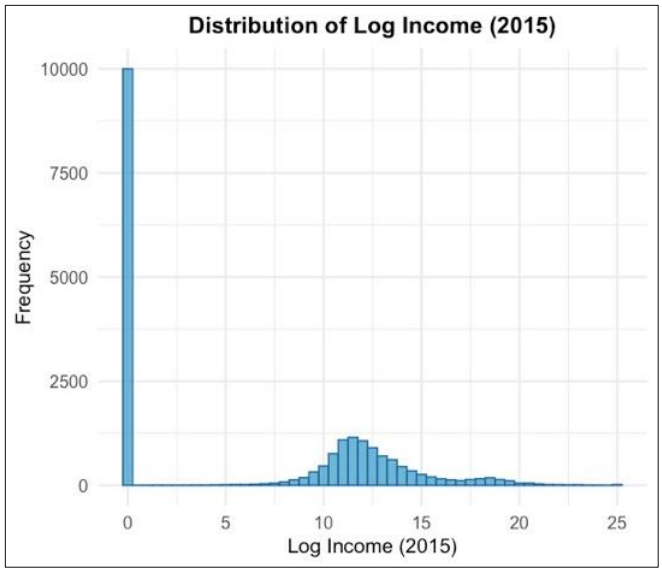

FIGURE III
DISTRIBUTION OF LOG INCOME (2015)

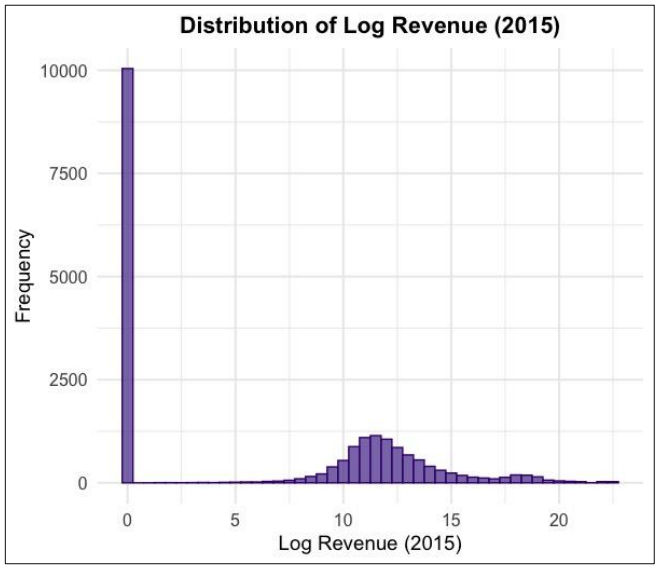

FIGURE IV
DISTRIBUTION OF LOG REVENUE (2015)

A-3. Revenue Distribution by Museum Discipline

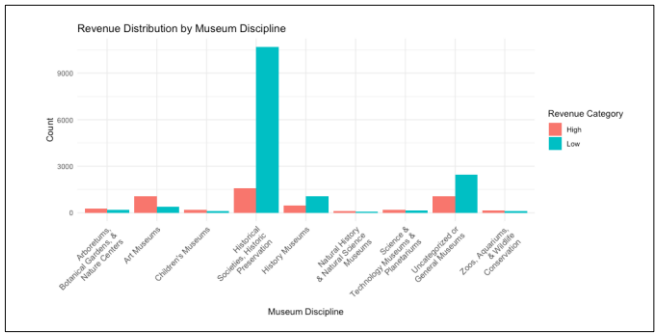

FIGURE VI
REVENUE CATEGORY BY MUSEUM DISCIPLINE

A-4. Proportion of High-Low Revenue by Non-Profit Status

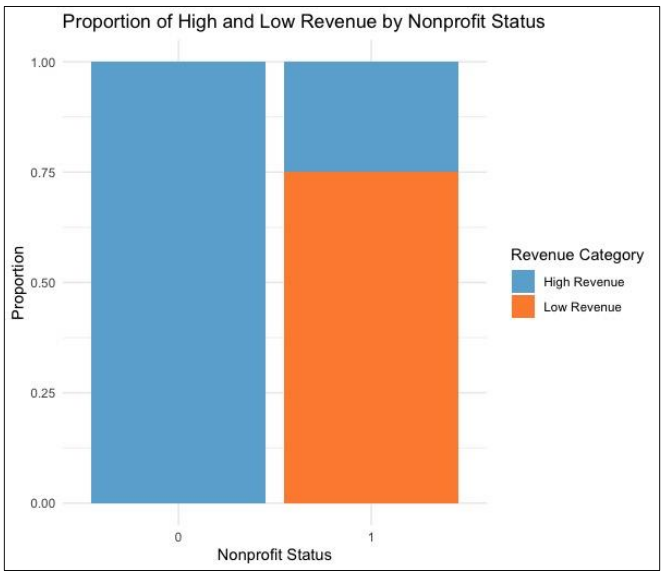

FIGURE IX
PROPORTION OF HIGH AND LOW REVENUE BY NONPROFIT STATUS

A-5. Proportion of High-Low Revenue by Private Foundation Status

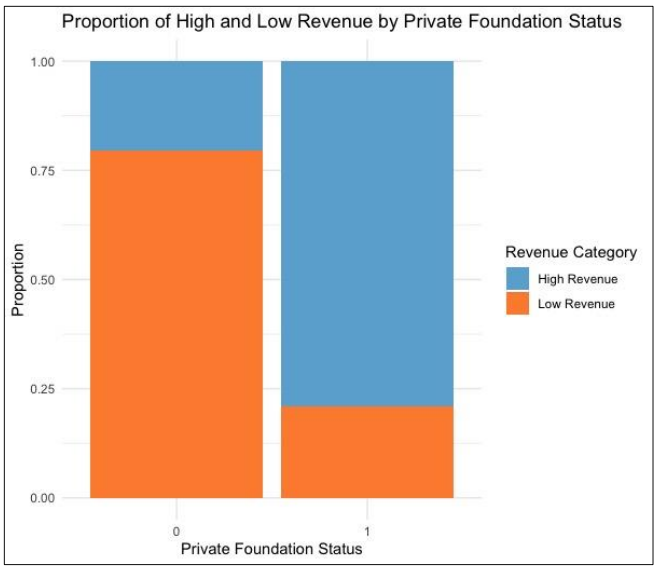

FIGURE X
PROPORTION OF HIGH AND LOW REVENUE BY PRIVATE FOUNDATION STATUS

A-6. Profitability Distribution by Museum Discipline

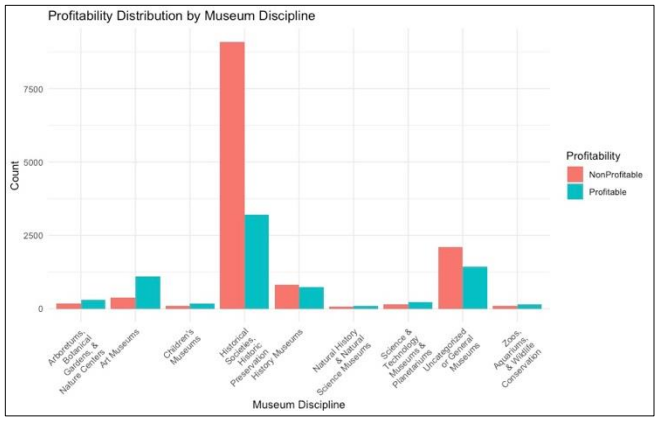

FIGURE XI
PROFITABILITY DISTRIBUTION BY MUSEUM DISCIPLINE

A-7. Top 10 Institutions by Revenue

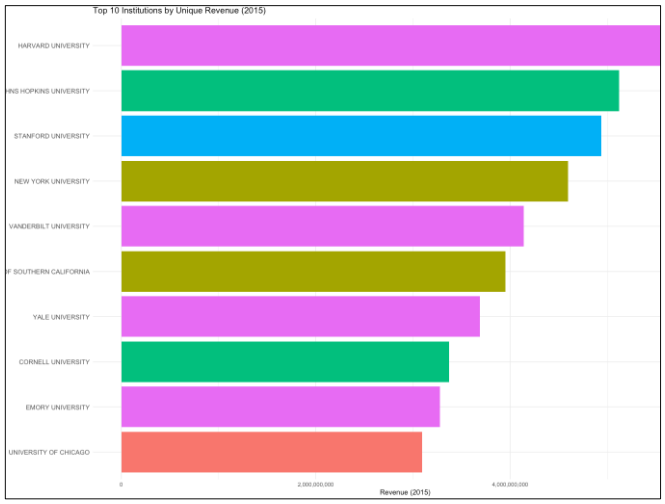

FIGURE XIV
PROFITABILITY DISTRIBUTION BY MUSEUM DISCIPLINE

A-8. Geographic Distribution of Museums

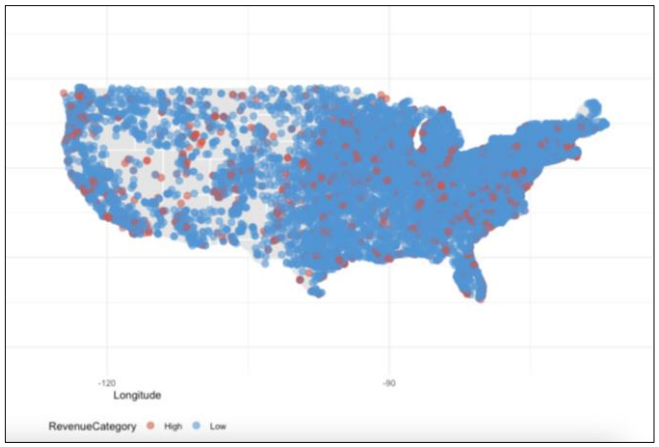

FIGURE XV
GEOGRAPHIC DISTRIBUTION OF MUSEUMS

APPENDIX B – REVENUE MODEL

B-1. Summary of Dataset Statistics

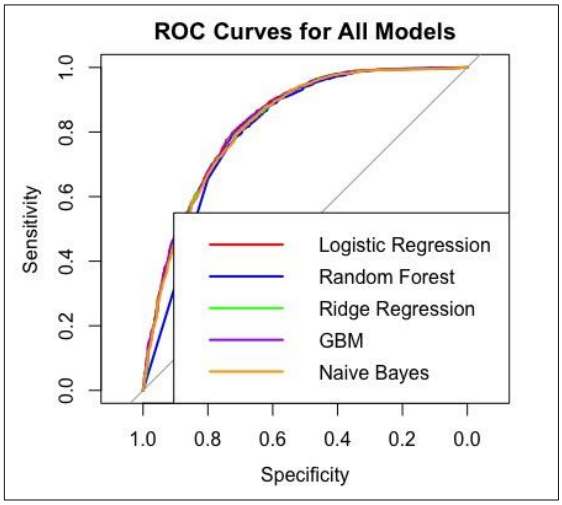

FIGURE XXIII
REVENUE MODELS – ROC CURVES COMPARISON

APPENDIX C – PROFITABILITY MODEL

C-1. Top 25 Features Profitability Model (Logistic Regression)

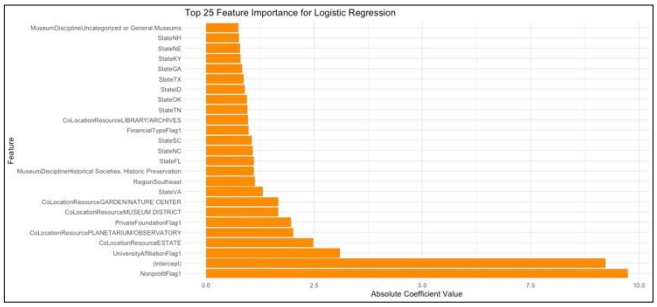

FIGURE XXIV
TOP 25 FEATURES LOGISTIC REGRESSION - PROFITABILITY

REFERENCES

[1] Manjarrez, C., Rosenstein, C., Colgan, C., and Pastore, E. "Exhibiting Public Value: Government Funding for Museums in the United States (IMLS-2008-RES-02)." Institute of Museum and Library Services, 2008. Available: https://www.imls.gov/sites/default/files/publications/documents/museumpublicfinance_0.pdf.

[2] Lindqvist, K. "Museum Finances: Challenges Beyond Economic Crises." *Museum Management and Curatorship*, vol. 27, no. 1, 2012, pp. 1–15. Available: https://doi.org/10.1080/09647775.2012.644693.

[3] Institute of Museum and Library Services. "Federal Support for Libraries and Museums." Accessed: [Month Day, Year]. Available: https://www.imls.gov/.

[4] Plaza, C. (2022). Museums in universities: Predicaments and potentialities. *Museum International, 74*(1–2), 74–85. https://doi.org/10.1080/13500775.2022.2157563

[5] Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random forest Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications, and research directions. *SN Computer Science*, 2, Article 160. https://doi.org/10.1007/s42979-021-00592-x

[6] Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics, 19*(1), 270. https://doi.org/10.1186/s12859-018-2264-5

[7] Trevethan, R. (2017). Sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice. *Frontiers in Public Health, 5,* Article 307. https://doi.org/10.3389/fpubh.2017.00307

[8] Villeneuve, P., & Martin-Hamon, A. (2007). At the heart of it: Museums and place-based study in rural communities. *The Journal of Museum Education, 32*(3), 253–262. https://www.jstor.org/stable/40479616