Predicting harmful algal blooms and uncovering mortgage bias: a data-intensive thesis

by

Kavya Kompella

B.Tech, Sagi Rama Krishnam Raju Engineering College, 2020

A THESIS

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Computer Science
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2024

Approved by:

Major Professor

Dr. Lior Shamir

Copyright

Abstract

This thesis presents two data science approaches for important environmental and social problems: predicting harmful algal blooms (HABs) resulting from cyanobacteria and identifying racial biases inherent in home mortgage systems.

In the first chapter, a machine learning model is developed to forecast HABs in Marion Reservoir, Kansas. HABs are a threat to water resources as they emit toxic chemicals that are harmful to agriculture and aquatic species. Early prediction of algae growth will help manage and prevent further growth. Various models are utilized for the prediction, including Random Forest, Support Vector Machine, Gaussian Bayes, Decision Tree, Long Short-Term Memory models, and XGBoost. In addition, using feature analysis, several factors were found that do not significantly affect the accuracy of predictions. Furthermore, the research extends its scope by comparing the algal bloom trends observed in Owasco Lake, New York, with those in Marion Reservoir. The findings of this research highlight the capacity of data science methodologies to tackle environmental issues, hence offering insights into the topic of proactive regulation of the water ecosystem.

The second chapter examines an extensive dataset of federal home mortgage data in the United States. This dataset covers 13 years and includes a vast number of loans. By utilizing machine learning methodologies, we reveal a significant correlation between the qualities of borrowers and mortgage data, particularly concerning the borrower's racial background. The results of our study indicate an association between the personal attributes of borrowers and loan data, suggesting that borrower race plays a significant role in the observed racial discrepancies in mortgage lending. Although other historical and present prejudices may be at play, this study offers quantitative evidence of racial biases across the home mortgage system. By identifying and examining these biases, our study makes a

valuable contribution to enhancing comprehension of the social concerns about equality and discrimination within the financial industry.

Together, these chapters emphasize the significance of employing data-driven research methodologies to address complex environmental challenges and uncover disparities in social equity. This highlights the multidisciplinary capacity of data science in the pursuit of achieving a more sustainable and equitable future.

Table of Contents

List of Figures vii

List of Tables x

List of Nomenclature xiii

Acknowledgements xiv

Dedication xv

Introduction 1

Related Work 3

1 Data Science Approaches for Prediction of algal blooms in Marion Reservoir . . . 5

1.1 Summary 5

1.2 Introduction 6

1.3 Data 9

1.4 Methodology 11

1.4.1 Data preprocessing 11

1.4.2 Feature Selection 17

1.4.3 Method 20

1.4.4 Cross-Validation 24

1.5 Impact of dataset size on Phycocyanin prediction 29

1.6 Comparison of Marion and Owasco Lake data 31

1.7 Conclusion 34

2 Quantitative analysis of racial bias in home mortgage loans 41
2.1 Summary 41
2.2 Introduction 42
2.3 Data 43
2.4 Methods 50
2.5 Results 50
2.5.1 Feature selection 53
2.5.2 Clustering 61
2.5.3 T-test 62
2.5.4 Unbalanced dataset 64
2.6 Conclusion 69

List of Figures

1.1 Eerie algae bloom in Marion Reservoir, Kansas. a. The green appearance on the lake's surface indicates algae bloom. b. Marion reservoir map where the data is collected. 10

1.2 The percentage missing values of each column in the Marion reservoir dataset were collected from the sensors. It is observed that about 10 % of data is missing for the features Chlorophyll, Dissolved Oxygen, Specific Conductivity, Phycocyanin, Turbidity, and Water Temperature. Approximately 1 % of the entire data is missing for features such as Precipitation, Storage, Air Temperature, Wind Direction, Wind Speed, Relative Humidity, and Solar Radiation. 11

1.3 Correlation Matrix: Integrating In-Lake Water Quality and Weather Data . 12

1.4 The figure represents the trend in each feature over time. 14

1.5 Optional: Short caption to appear in List of Figures 15

1.6 Distribution of features of Marion reservoir data. 16

1.7 Comparison of Missing Value Imputation Methods: Time Interpolation vs. Iterative Imputation. The left side of the graph showcases values filled using time interpolation, depicted in red, while the right side illustrates those filled via iterative imputation in green. This figure specifically focuses on features with a significant percentage of missing values. 18

1.8 Optional: Short caption to appear in List of Figures 20

1.9 The prediction of Phycocyanin ahead of different time spans using RF is illustrated in the above subplots. 24

1.10 Predictions of Phycocyanin at various future intervals using an ensemble of RF, XgBoost, and Long Short-Term Memory model. 25
1.11 The prediction of Phycocyanin ahead of different time spans using XgBoost is illustrated in the above subplots. 27
1.12 Optional: Short caption to appear in List of Figures 29
1.13 Optional: Short caption to appear in List of Figures 30
1.16 The effect of the dataset size on the performance of the ensemble model (RF, XGB, LSTM) prediction of Phycocyanin for a prediction duration of seven days 31
1.17 Owasco Finger Lake, New York. 34
1.14 Optional: Short caption to appear in List of Figures 38
1.15 Cross-validation results of the ensemble model combining RF, XGBoost, and LSTM. The bar chart displays the average values of MSE, Pearson Correlation, and RMSE, with error bars representing one standard deviation. 39
1.18 Optional: Short caption to appear in List of Figures 40

2.1 Mortgage information 44
2.2 Feature importance for the prediction of the borrower's race as determined by XGBoost. 54
2.3 Confusion matrix for the prediction of the race based on the loan information. The labels in the figure represent the race of the borrower, which is specified in Table 2.2. 57
2.4 Similarity Matrix for classification of Borrower Race. The labels in the figure represent the race of the borrower, which is specified in Table 2.2 58
2.5 Confusion Matrix for classification of borrower race using top ten best features 2.2. 59
2.6 Similarity Matrix for classification of borrower race using top ten best features 2.2. 60
2.7 Clustering with 3 and seven clusters on borrower's race. 62

2.8 Year-wise Clustering with K = 3 (Part one). 65
2.8 Year-wise Clustering with K = 3 (Part Two). 66
2.8 Year-wise Clustering with K = 3 (Part Three). 67
2.9 Determining the optimal number of Clusters using the elbow method. 68
2.10 Principle Component Analysis based on borrower's race. 69
2.11 This graph represents the demographics of BoRace over 13 years and Table 2.2 gives more details of all races present in the data. 70
2.12 The prediction accuracy of borrower race over the years for sampled data . . 71

List of Tables

1.1 Concise description of in-lake water quality and climate data at Marion reservoir 9

1.2 The table above illustrates the performance of the XgBoost model for predicting the different amounts of Phycocyanin ahead of time. The correlation coefficient and the RMSE are the values between observed and predicted values. 23

1.3 The table above illustrates the performance of the RF model for predicting different time periods ahead of time. The correlation coefficient and RMSE values represent the accuracy between observed and predicted values. 26

1.4 The table above illustrates the performance of the LSTM model for predicting the different amounts of Phycocyanin ahead of time. The correlation coefficient and the RMSE are the values between observed and predicted values. . 26

1.5 The table above illustrates the performance of the ensemble model created using RF, XgBoost, and LSTM for predicting the different amounts of Phycocyanin ahead of time. The correlation coefficient and the RMSE are the values between observed and predicted values. 26

1.6 The table above illustrates the performance of the ensemble model created using RF[1], SVM, K Neighbors Regressor, and Gaussian Process Regressor for the prediction of the different amounts of Phycocyanin ahead of time. The correlation coefficient and the RMSE are the values between observed and predicted values. 28

1.7 The table above illustrates the performance of the ensemble model created using RF[1], SVM, K Neighbors, and Gaussian Process Regressor for the prediction of the different amounts of Phycocyanin ahead of time by using the mean of n records above and below the predicting record (for the result above n = 5). The correlation coefficient and the RMSE are the values between observed and predicted values. 28
1.8 Comparison of predicted value correlation coefficients between Marion and Owasco Lake using RF, XGBoost, and LSTM ensemble model. 35
1.9 Comparison of predicted value RMSE between Marion and Owasco Lake using RF, XgBoost, and LSTM ensemble model. 35
1.10 Top five important features based on feature importance of ensemble of RF and XgBoost for a prediction duration of three days. 35

2.1 Features, Data Type, and Description of features in the dataset. 49
2.2 The different borrower races used in the dataset. 50
2.3 Precision, recall, accuracy, and F1 score for the classification of borrower race using the XGBoost classification algorithm. 51
2.4 Precision, recall, accuracy, and F1 score when predicting the race of the borrower using a "dummy" ZeroR classifier. 51
2.5 Precision, recall, accuracy, and F1 score for the classification of borrower race after removing the borrower's ethnicity column from the data set. 51
2.6 Precision, recall, accuracy, and F1 score for the classification of borrower race after removing the co-borrower race and borrower ethnicity from the data. . 52
2.7 Accuracy metrics for the predictability of borrower's race between two distinct races. 53
2.8 Mean value of each feature of the FHLB dataset. 56
2.9 The above table shows the model accuracy of predicting BoRace using the top ten features. 61

2.10 Top ten features influencing the prediction of BoRace. 61
2.11 Precision, recall, accuracy, and F1 score for the classification of borrower race using the XGBoost classification algorithm using the balanced dataset. . . . 67
2.12 Precision, recall, accuracy, and F1 score when predicting the borrower's race using a "dummy" ZeroR classifier using the balanced dataset. 68
2.13 Precision, recall, accuracy, and F1 score for the classification of borrower race after removing the co-borrower race and borrower ethnicity from the balanced data. 68
2.14 The results of the XgBoost classifier on under-sampled data. 72
2.15 The year-wise classification accuracy of the borrower race over the years ZeroR for under-sampled data. 72
2.16 Year-wise ZeroR results for the entire data. 73
2.17 Year-wise classification results of Borrower Race XgBoost. 73

Nomenclature

Machine Learning Models

- **RF** – Random Forest
- **SVM** – Support Vector Machine
- **GB** – Gaussian Bayes
- **DT** – Decision Tree
- **LSTM** – Long Short-Term Memory
- **XGBoost** – Extreme Gradient Boosting

Model Evaluation Metrics

- **MSE** – Mean Squared Error
- **RMSE** – Root Mean Squared Error
- **Pearson's Correlation** – Measure of linear correlation between variables

Acknowledgments

First and foremost, I express my deepest gratitude to my major advisor, Dr. Lior Shamir, for his invaluable guidance, unwavering support, and belief in my capabilities. His expertise and mentorship have been the cornerstone of my research journey, shaping this thesis into what it has become.

I am equally thankful to my committee members, Dr. Hande McGinty and Dr. Daniel Andresen, for their insightful critiques and constructive feedback. Their profound knowledge and rigorous review process have significantly contributed to the refinement and depth of this work.

Special thanks are extended to Dr. Trisha Moore, Dr. Aleksey Sheshukov, and Dr. Daniel Flippo for their guidance through the intricate paths of the algal bloom project. Their meticulous reviews, thoughtful feedback, and encouragement were instrumental in overcoming the challenges encountered during this research.

My sincere appreciation goes to Laura Krueger, whose assistance in unraveling the technical jargon of algal bloom prediction and organizing pivotal data sets has been indispensable. Her contributions have been vital to the progress and success of this work.

I am also grateful to Ishitaa Sayal and Abdullah Alaklabi for collaborating on the Federal Home Loan project. Their perspectives and partnership enriched the research experience, highlighting the power of collaborative effort.

This thesis has benefited immensely from the collective wisdom, encouragement, and support of all those mentioned. I extend my heartfelt thanks to them.

Dedication

This thesis is wholeheartedly dedicated to my uncles, Dr. Uday Kompella and Arun Kompella, whose unwavering encouragement and insightful advice inspired me to embark on this journey towards acquiring my master's degree. Their belief in my potential and relentless support have been my guiding light through the challenges and triumphs of this academic endeavor.

I also extend my deepest gratitude to my family, whose unconditional love, patience, and support have been the backbone of my strength and perseverance. Their sacrifices have not gone unnoticed, and this achievement is as much theirs as it is mine.

Additionally, I extend thanks to my friends, Dr. Nitin Mishra, Samatha, Anika, and Mana, for their consistent support and companionship throughout this journey.

Introduction

In a contemporary period characterized by the increasing significance of data science and machine learning in tackling intricate environmental and social issues, this thesis introduces two noteworthy data-centric initiatives that seek to enhance our comprehension and facilitate constructive transformation. The study has two discrete but interrelated sections, each addressing a crucial matter.

The initial portion of this study is on the prediction of Harmful Algal Blooms (HABs), a periodic ecological hazard capable of causing significant harm to water supplies, agriculture, and wildlife. In light of the urgent requirement for early prediction and mitigation, this study utilizes a range of machine learning algorithms including Random Forest (RF), Support Vector Machine (SVM), Gaussian Bayes (GB), Decision Tree (DT), Long Short-Term Memory (LSTM), and XGBoost, to anticipate the incidence of algal blooms. The results of this study go beyond mere prediction and examine the various aspects that impact the accuracy of predictions. Furthermore, this chapter expands its range by conducting a comparative analysis of algal bloom patterns in various geographical regions, illuminating the potential of data science approaches in the proactive management of water ecosystems.

In the second chapter, a detailed analysis is conducted on a comprehensive dataset of federal home mortgage data in the United States. This analysis uncovers a significant relationship between borrower characteristics and mortgage data. This chapter reveals a notable correlation between borrowers' race and differences in mortgage lending, indicating an inherent bias inside the financial system. Using machine learning approaches highlights the significance of individual characteristics in the observed racial disparities. This study presents empirical data that supports the existence of racial differences within the house mortgage system, contributing to the understanding of historical and contemporary lending prejudices. The research addresses significant concerns of equality and discrimination by

examining these disparities.

Collectively, these chapters highlight the capabilities of data science in tackling environmental issues and revealing inequalities in social justice. They emphasize the significance of utilizing data-driven research approaches to promote a sustainable and fair future. As society becomes more reliant on data, the aforementioned efforts serve as significant illustrations of the crucial influence that data science has on defining our comprehension of the environment and society and instigating substantial transformations.

Related Work

In the past few decades, the vast amounts of data combined with sophisticated computing power have led to the emergence of data science, which has emerged as a cornerstone for innovation, offering unprecedented insights across various domains. This field makes it possible to make better decisions by revealing patterns and trends in large, complex datasets that are frequently invisible to humans through the careful use of ML algorithms. The predictive modeling of HABs and the quantitative examination of racial prejudice in federal home loan data are two different applications of the research conducted for this thesis that highlight the revolutionary potential of data science. These case studies demonstrate the methodological breakthroughs that enable such broad applications while also highlighting the adaptability of data science in tackling issues ranging from environmental sustainability to social equality[2–7].

Alongside escalating ecological concerns, HABs have emerged as a significant threat to aquatic ecosystems, characterized by their detrimental impacts on water quality, marine life, and public health[2]. In the face of escalating environmental challenges, Marion Reservoir, Kansas, stands as a critical example of the pervasive threat of harmful algal blooms, propelled by a confluence of eutrophication and climate change, underscoring an urgent need for innovative predictive modeling and management strategies[3].

Recent developments in HAB forecasting highlight a paradigm shift away from conventional process-based models and toward innovative, data-driven strategies Lin et al.[4]. Yu et al.[5] further contribute to this body of work by focusing on the prediction of coastal algal blooms using environmental factors through ML methods. Recognizing HABs as a major type of marine disaster, their study proposes a method based on ML to predict the occurrence of algal blooms by analyzing environmental parameters. Their work validates the prediction performance on two real datasets from the United States and China, employing

ML algorithms to select models and feature subsets for accurate phytoplankton concentration predictions. This research not only demonstrates the efficiency of ML methods in short-term prediction but also reveals the crucial environmental factors contributing to the outbreak of harmful algal blooms, thereby enriching our understanding and management strategies against this ecological menace.

Ozgur et al.'s[6] work explores the complicated dynamics of bank lending in an emerging economy by using ML techniques to analyze the complex relationships between macroeconomic indicators, bank-specific features, and external influences. This work is noteworthy because it uses a variety of ML techniques to improve our understanding of lending behaviors while addressing the nonlinear and nonparametric interactions present in bank lending operations. This work highlights the important role of ML in financial analysis and decision-making by identifying important determinants of bank lending and highlighting the nonlinearities involved. Policymakers, bank management, and regulatory agencies may benefit greatly from these insights.

Using data from New Jersey, Samuel et al.[7] investigate the effect of racial differences on mortgage lending practices. The authors conduct an in-depth investigation that, beyond what can be explained by variations in creditworthiness, connects a sizable percentage of the racial disparity in loan denial rates to discriminatory lending practices using a three-step estimator. This paper is particularly significant for its robust approach to examining how racial factors influence loan approvals, even when accounting for credit risk and other variables. To maintain justice and equity, the lending industry is challenged by the findings, which offer crucial information to regulators, bank management, and policymakers.

Chapter 1

Data Science Approaches for Prediction of algal blooms in Marion Reservoir

1.1 Summary

Harmful algal blooms (HABs) caused by cyanobacteria can have a detrimental impact on water ecosystems, leading to the need for accurate prediction and prevention strategies. If an algal bloom could be predicted, a local management strategy could be implemented and applied to treat the bloom effectively to preserve the water's quality. In this study, we applied a data science approach to predicting cyanobacteria blooms in Marion Reservoir, Kansas. We collected data and developed models using Random Forest, Support Vector Machine, Gaussian Bayes, Decision Tree, Long Short-Term Memory, and XgBoost. Our results showed that algal bloom can be predicted before it grows and affects the water quality. XgBoost and RF performed better than other models, indicating their effectiveness in predicting cyanobacteria blooms. Furthermore, we identified that parameters such as Specific Conductivity, Phycocyanin value on the day of prediction, Storage, dam release, and Wa-

ter Temperature had no impact on the prediction, which helped us understand the features that were affecting the HABs in the Marion reservoir. Additionally, a comparative analysis is conducted by applying the ensemble models of RF, XgBoost, and LSTM that demonstrated the highest performance on Marion Reservoir data to predict HABs in Owasco Lake, New York. Our findings demonstrate the potential of data science approaches for predicting HABs caused by cyanobacteria,which can significantly impact environmental management and public health.

1.2 Introduction

Cyanobacteria are abundant in aquatic ecosystems, and certain species can produce harmful toxins, leading to the formation of HABs that render water unfit for use. These blooms not only contaminate drinking water for humans but also have detrimental effects on animals, aquatic life, and their reproductive capabilities. In recreational areas, contact with algal blooms and accidental consumption of toxins in water when boating and swimming pose a threat to human health. In addition, the escalation of algal blooms releases toxic gases that pose are dangerous to human health if inhaled[8]. Efficient management techniques are essential to mitigate the impact of HABs, and one crucial aspect is the ability to predict their recurrence. Preventive measures can be implemented by predicting these blooms in advance, such as early spraying with small amounts of chemicals. Treating blooms at an early stage is more manageable, more effective, efficient, and cost-effective than dealing with them after they have escalated. This proactive approach significantly reduces the expenses associated with cleaning lakes and prevents the contamination of water bodies. Ultimately, it ensures the availability of fresh, non-toxic water for various purposes, including irrigation and drinking, while allowing the areas open to visitors. Closure of the recreational regions due to algal bloom can reduce revenue and increase drinking water treatment costs when the toxin is present. Machine learning represents a powerful tool for predicting the levels of Phycocyanin, a pigment indicative of cyanobacterial presence.

In this chapter, we applied a data science approach to predict cyanobacteria blooms in Marion Reservoir, located in central Kansas within Marion County. The U.S. Army Corps of Engineers constructed the reservoir as a multipurpose reservoir for flood control, water supply, recreation, and wildlife habitat. Construction began in 1964, and closure was completed in 1967. The entire flood control operation started in 1968. The reservoir has a surface area of 6402 acres (26 sq. kilometer) and a depth of nine meters max; 3.4 m mean. It is the 12th largest federal lake in Kansas by volume (80,659 acre-feet)[9]

The first noted bloom in the reservoir was in 2003, and there have been nearly annual blooms since then[9]. Marion Reservoir has been closed when a harmful algal bloom is documented at a hazardous level. Levels are based on high toxin levels and high cell counts. When at a hazardous level, the public is directed to avoid contact with the water by closing all parks, boat ramps, and recreation areas near the water[10;11].

Marion Reservoir is a vital water source for many communities and serves as a popular destination for recreational activities, making it imperative to monitor and prevent the occurrence of HABs. We collected a large environmental and water quality variable dataset from the reservoir. Water quality data is compiled from a stationary sensor, and the environmental data is collected from the USACE website[12]. We developed predictive models using ML and recurrent neural network algorithms. The study's main objectives are to assess the extent to which different machine learning algorithms predict cyanobacteria blooms and to determine the essential parameters that influence their occurrence.

This research is significant because it addresses an important problem that has significant environmental and public health implications. Furthermore, it adds to the expanding corpus of research on data science methods for HAB prediction, which has the potential to improve our ability to monitor and manage freshwater bodies. The study's findings will have practical applications for water management agencies and policymakers seeking to mitigate the impact of HABs.

Abbrev. Name	**Full Name**	**Type**	**Description**
Time1	DataTime	DateTime	Timestamp of the sample recording.
TAL-PC RFU	Phycocyanin	float64	Optical pigment measurement indicating cyanobacteria presence.
Chlorophyll RFU	Chlorophyll	float64	Optical pigment measurement from green plants and algae.
ODO % sat	Dissolved Oxygen	float64	Optically measured dissolved oxygen percentage in water.
SpCond µS/cm	Specific Conductivity	float64	Conductivity measurement reflecting water's ion content.
Turbidity FNU	Turbidity	float64	Optical measurement of water cloudiness in FNU.
Temp °C	Water Temperature	float64	Thermistor-measured Water Temperature in °C.
PRECIP	Precipitation	float64	Precipitation in inches measured at the gage.
Storage	Reservoir Storage	float64	Stored water in the reservoir, measured in ac-ft.
INFLOW	Drainage Basin Inflow	float64	Inflow into the reservoir in cubic feet per second.
RELEASE	Gate Release	float64	Water released by gates, measured in cubic feet per second.
AIR-TEMP	Air Temperature	float64	Gage-measured air temperature in °C.
WIND-DIR	Wind Direction	float64	Gage-measured wind direction in degrees.

WIND-SPEED	Wind Speed	float64	Wind Speed in mph measured at the gage.
REL-HUMID	Relative Humidity	float64	Relative Humidity percentage measured at the gage.
SOLAR-RAD	Solar Radiation	float64	Solar Radiation in W/m2 measured at the gage.

Table 1.1: Concise description of in-lake water quality and climate data at Marion reservoir

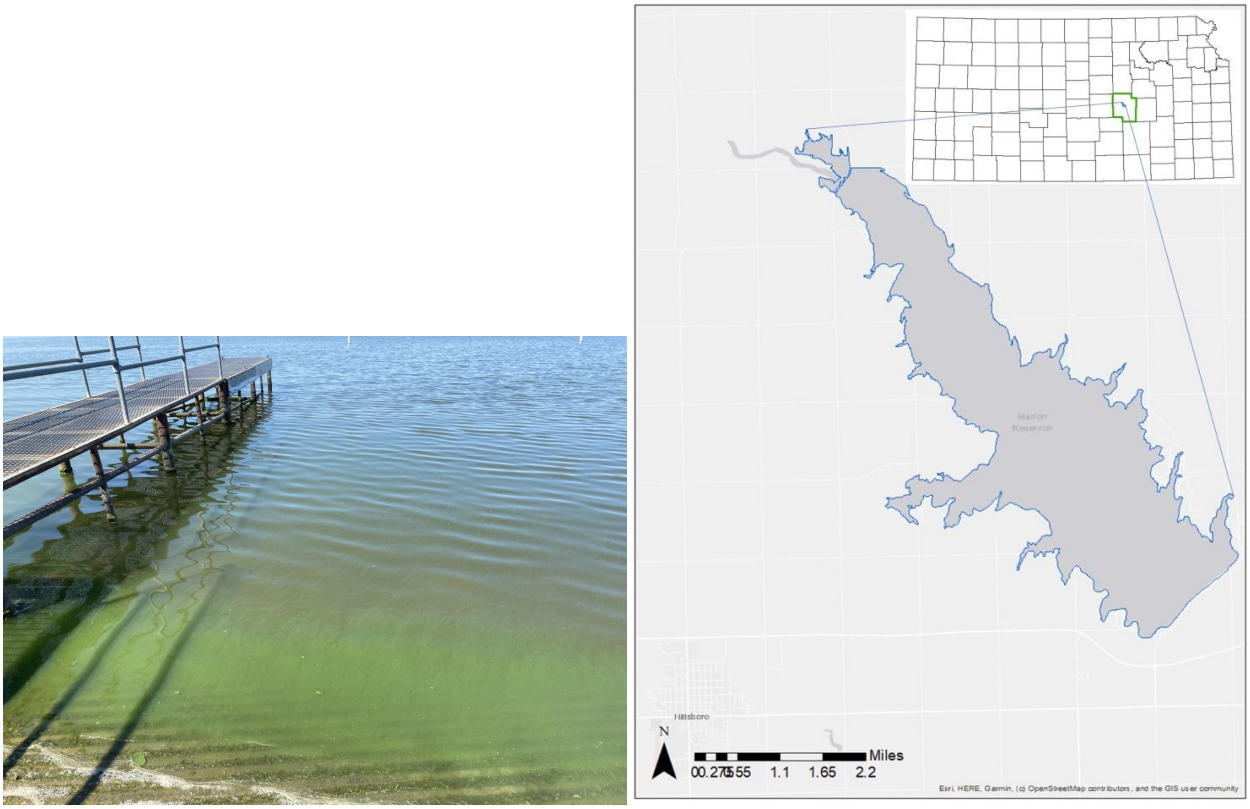
1.3 Data

The data was collected hourly using a multi-parameter water quality sensor[13] located in-lake in a buoy and from the US Army Corps of Engineers (USACE) weather station[12] located on the reservoir dam. The data collection period spanned from May 12th, 2022, to September 5th, 2023, resulting in a dataset comprising 7943 records and 15 features[14]. The data was not collected in December, January, and February as the lake was frozen during that time. The distribution of each parameter in the dataset is represented in Figure 1.6 and the trend of each parameter with time is represented in Figure 1.4.

The parameters such as Air Temperature, Chlorophyll, Phycocyanin, Turbidity, etc were collected from Marion Lake and are described in Table 1.1. It is important to note that the dataset contains missing values, as the sensors did not always capture specific parameters because of situations like damage to the sensor caused by internal or external reasons and the draining of batteries. A feature like light value was incorrectly noted because the algal blooms or other biofilms covered the sensor and hindered the measurement of light values accurately so it had to be excluded from the dataset used for training the model. All 15 columns in the dataset exhibit missing values in some records specifically, the columns related to Chlorophyll, Dissolved Oxygen, Specific Conductivity, Water Temperature, and Turbidity show 782 records with missing values. In contrast, the variables like Solar Ra-

diation, Relative Humidity, and Air Temperature have eight missing records. The inflow, release, and Precipitation variables exhibit 110, 87, and seven missing records, respectively. The percentage of missing values for each column is shown in Figure 1.2

In total, 921 records have at least one missing value across the features. The correlation matrix is depicted in Figure 1.3 to understand the interrelationships between the features. This matrix provides insight into the degree of association between different variables in the dataset.



(a) Algae Bloom in Marion Reservoir, Kansas. (b) Marion reservoir map.

Figure 1.1: Eerie algae bloom in Marion Reservoir, Kansas. a. The green appearance on the lake's surface indicates algae bloom. b. Marion reservoir map where the data is collected.

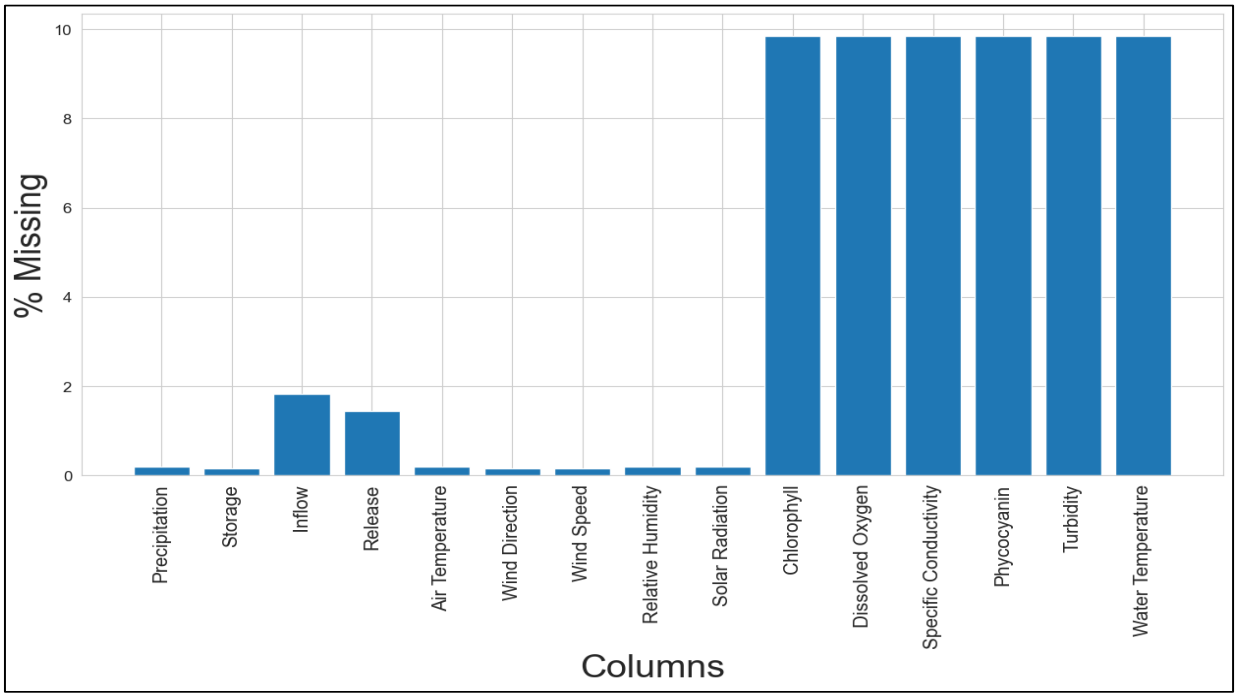


Figure 1.2: The percentage missing values of each column in the Marion reservoir dataset were collected from the sensors. It is observed that about 10 % of data is missing for the features Chlorophyll, Dissolved Oxygen, Specific Conductivity, Phycocyanin, Turbidity, and Water Temperature. Approximately 1 % of the entire data is missing for features such as Precipitation, Storage, Air Temperature, Wind Direction, Wind Speed, Relative Humidity, and Solar Radiation.

1.4 Methodology

1.4.1 Data preprocessing

As discussed in the text referenced in section 1.3, The data for this study was collected from two sensors: one for water quality and the other for the weather. The two datasets were combined into one, and columns with special characters were removed to avoid errors during data processing. During the initial days of sensor installation, there were minor portions of negative values in the dataset, which were nearly zero. These values have been adjusted to zero for clarity and accuracy. There were 921 rows of missing data, which was a significant challenge.

There are various methods for handling missing data, such as deleting rows with

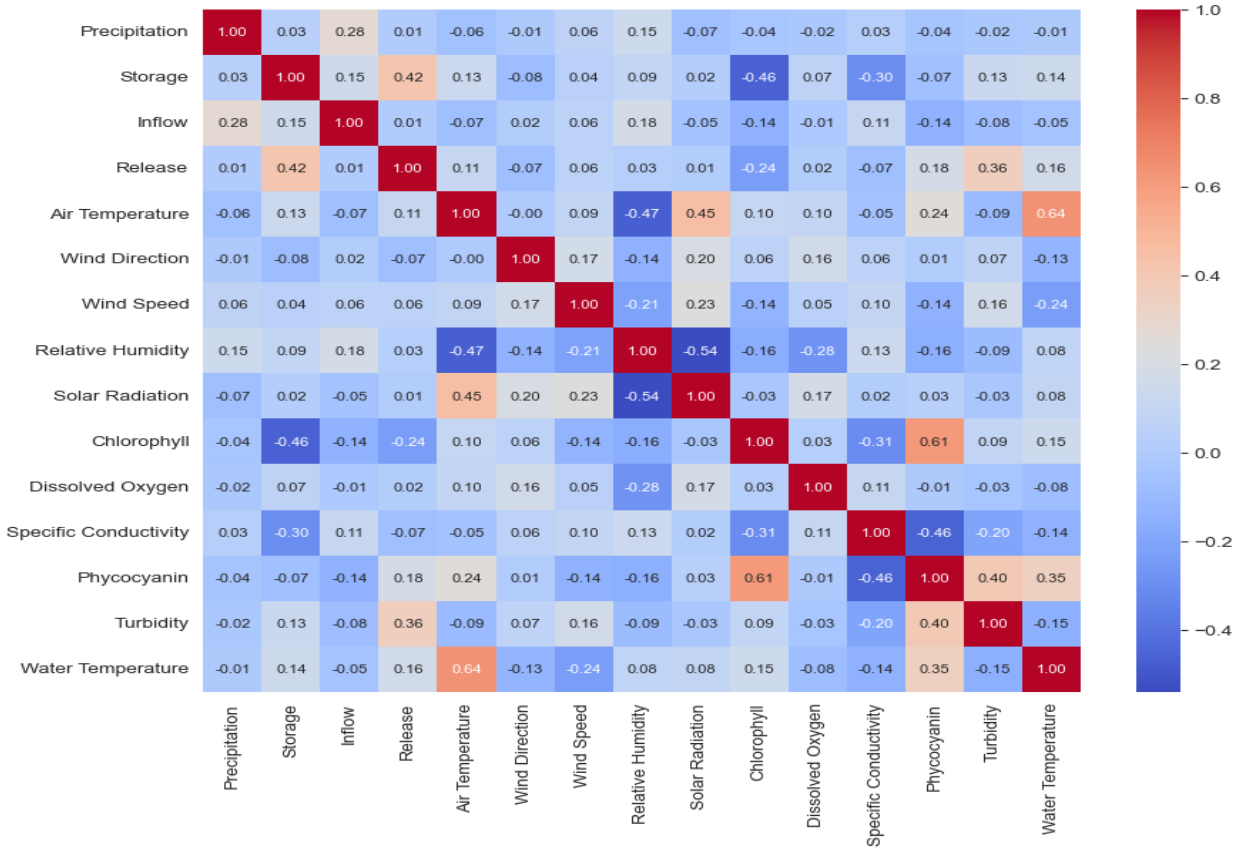
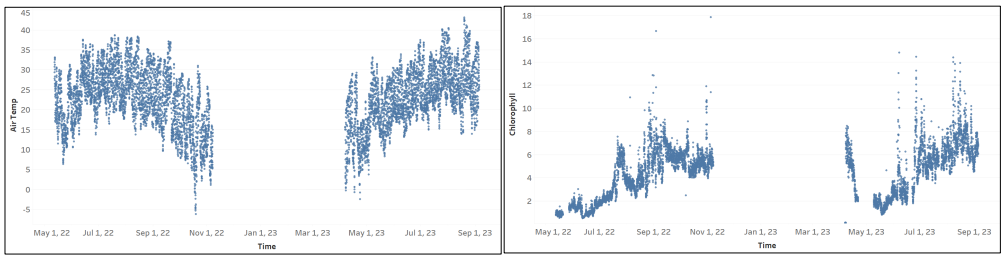


Figure 1.3: The correlation matrix integrates in-lake water quality data and weather variables. Notably, strong correlations emerge between parameters: Phycocyanin and Chlorophyll, Turbidity and Water Temperature; Air Temperature and Solar Radiation; Storage with Release, and Turbidity and negatively correlates with Chlorophyll and Specific Conductivity; and Air Temperature and Water Temperature; Specific conductive strong negative correlation with Phycocyanin; Relative Humidity negatively correlates with Solar Radiation.

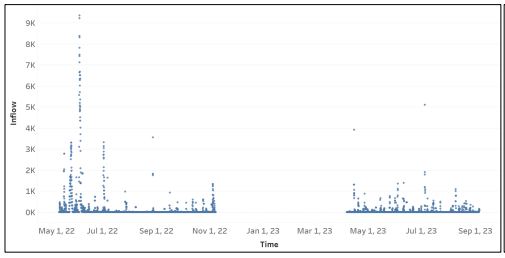
missing values or imputing placeholder values such as the mean, median, or zero. However, these methods are unsuitable for time series data, where the order of records is important, and the missing values may not be random. In particular, continuous missing values can cause problems for these methods, as they can disrupt the temporal trends in the data.

Craig D. N. et al.[15] described several methods for dealing with missing data in the time series. It is important to carefully consider the different approaches and select one appropriate for the specific dataset. Time interpolation and iterative imputation were employed in this study to fill in the missing values. Based on Figure 1.7, it can be observed that time interpolation was utilized to fill gaps in the data points in a linear manner. While this

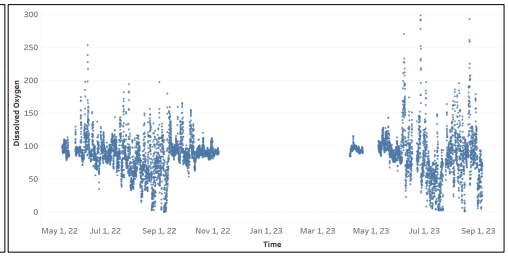


(a) Air Temperature

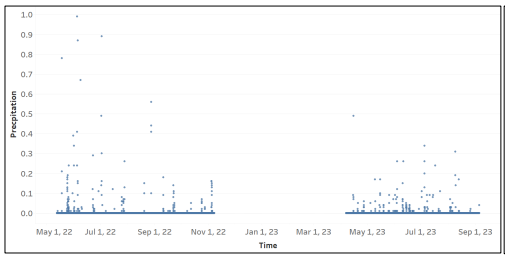
(b) Chlorophyll



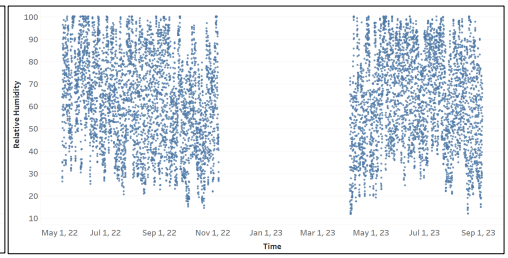
(c) Inflow



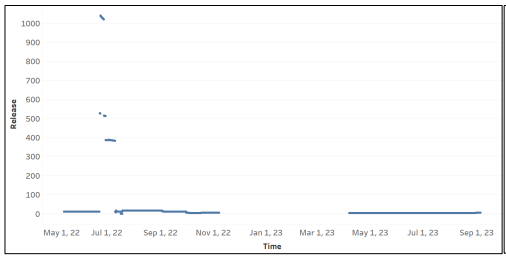
(d) Dissolved Oxygen



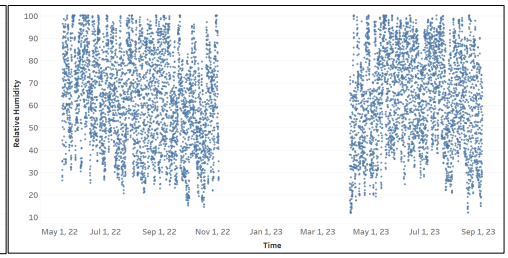
(e) Precipitation



(f) Relative Humidity

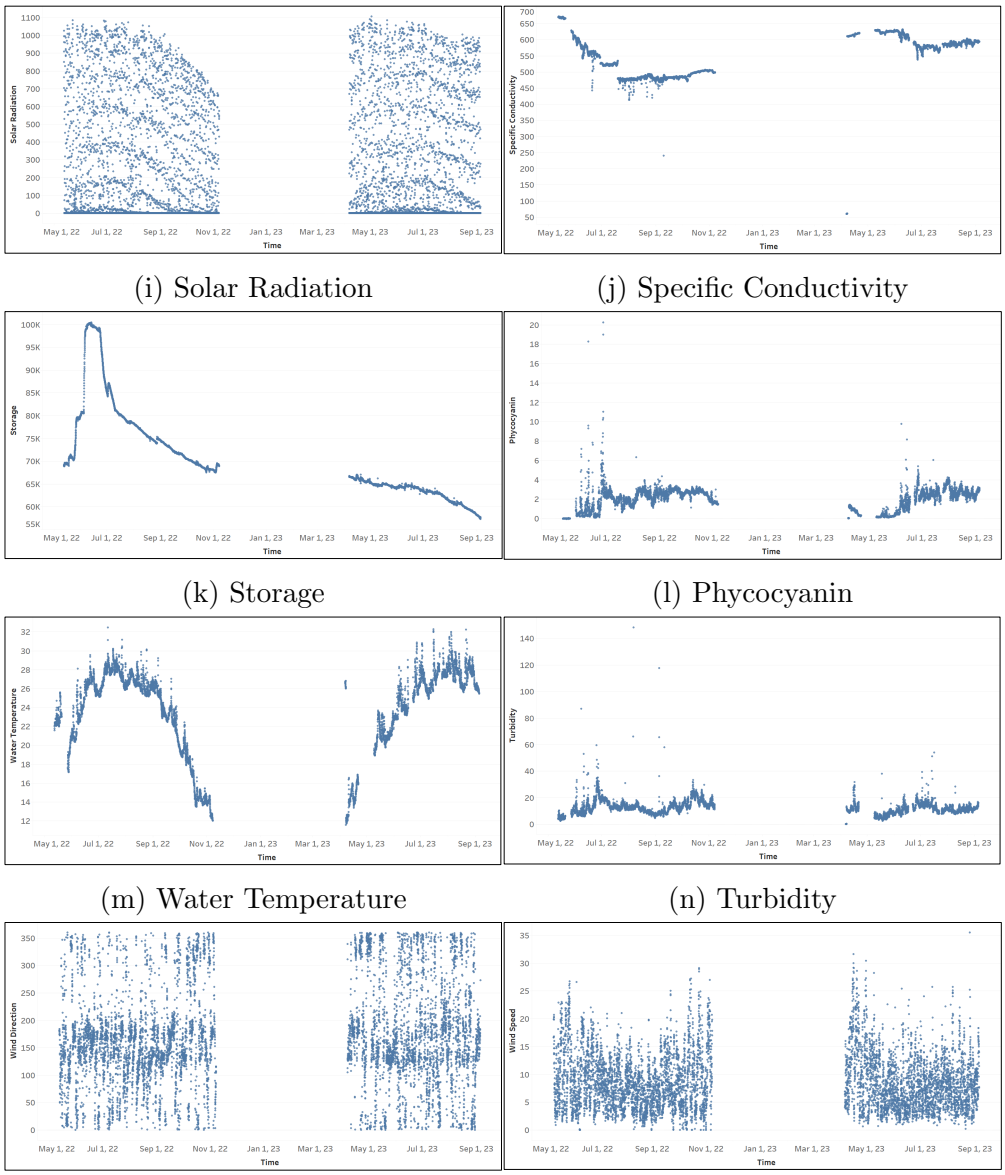


(g) Release



(h) Relative Humidity

approach is simpler, it may not fully capture all the inherent patterns, particularly in cases where the data displays non-linear attributes or is subject to external influences. In contrast, using iterative imputation, specifically employing the 'IterativeImputer' module from the 'sklearn' library, exhibited a more reliable methodology. Instead of solely focusing on the sequence over time, this method treats each feature with missing values as dependent on other features. This process is performed repeatedly, addressing all the features in the data until the imputation reaches a state of agreement. This approach has several advantages, as it takes the overall structure of the dataset and the interrelationships among its components



(i) Solar Radiation

(j) Specific Conductivity

(k) Storage

(l) Phycocyanin

(m) Water Temperature

(n) Turbidity

(o) Wind Direction

(p) Wind Speed

Figure 1.4: The figure represents the trend in each feature over time.

into account, and a more dependable technique was used to fill in the missing data. Based on Figure 1.7, it can be observed that the iterative imputation method exhibits a higher degree of agreement with the original data patterns. This suggests that iterative imputation may be considered a more suitable option than time interpolation in the given dataset. Iterative imputation has been extensively studied and provides the benefit of doing multi-feature imputation by considering correlations and patterns within a dataset. The research by Buuren

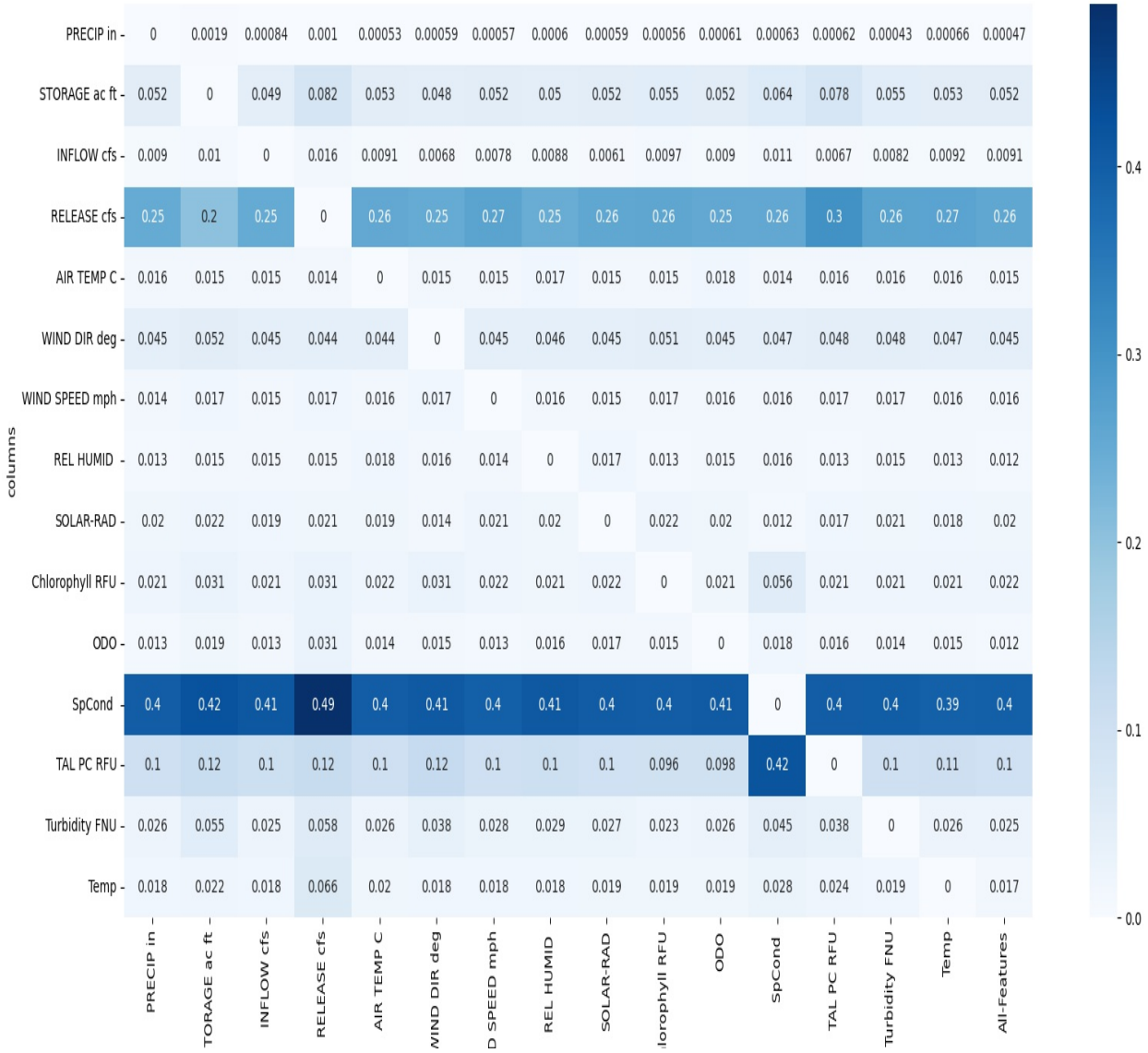
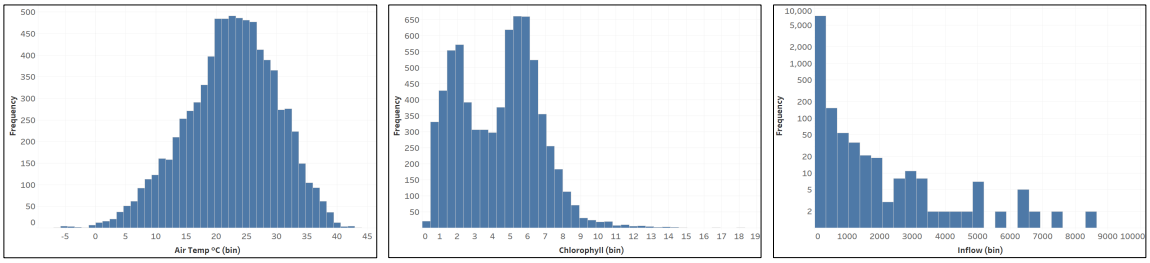


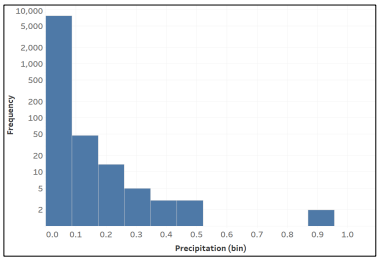
Figure 1.5: Feature importance for predicting Phycocyanin levels. Each feature's contribution is evaluated by systematically removing them one by one and observing the impact on the prediction accuracy.



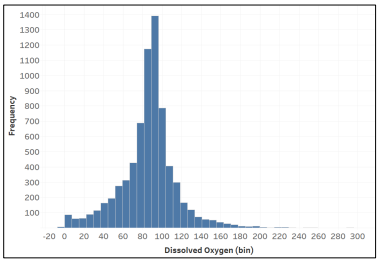
(a) Air Temperature

(b) Chlorophyll

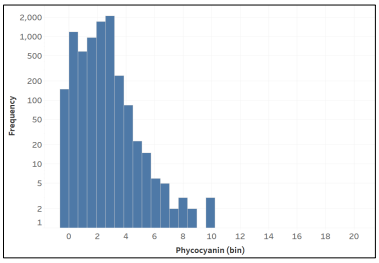
(c) Inflow



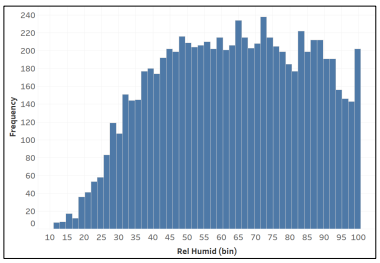
(d) Precipitation



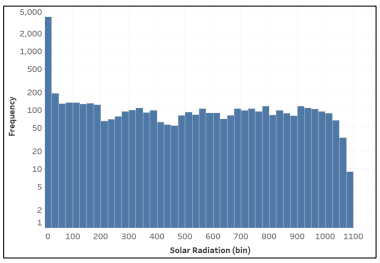
(e) Dissolved Oxygen



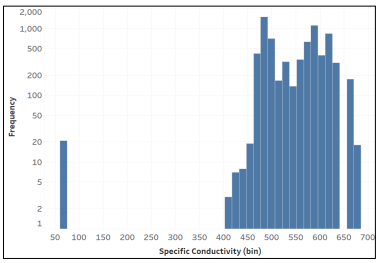
(f) Phycocyanin



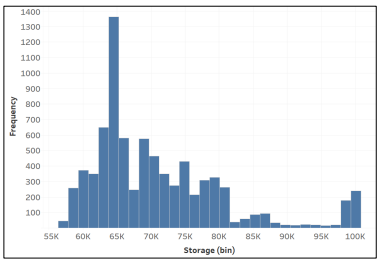
(g) Relative Humidity



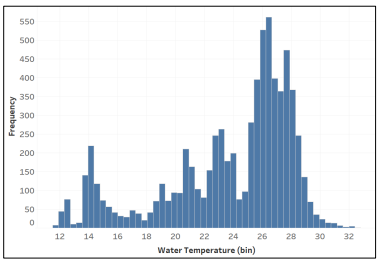
(h) Solar Radiation



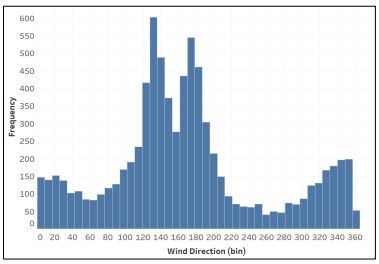
(i) Specific Conductivity



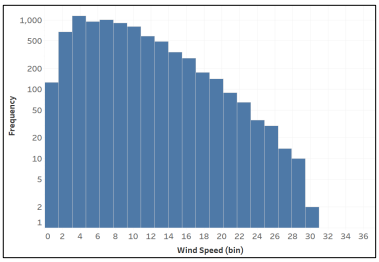
(j) Storage



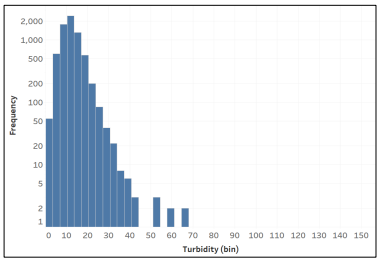
(k) Water Temperature



(l) Wind Direction



(m) Wind Speed



(n) Turbidity

Figure 1.6: Distribution of features of Marion reservoir data.

S.V. et al.[16] provides an introductory overview of this technology, elucidating its underlying algorithmic principles and possible advantages.

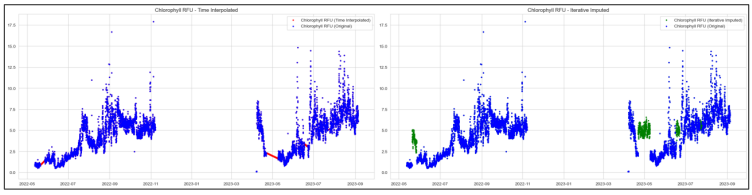
The next step in the data preparation phase was to normalize the data. Normalization is essential in machine learning to prevent features of varying sizes and magnitudes from adversely affecting the model's learning process. The min-max normalization approach was used in this analysis. To ensure all the features in this study had an equal scale and to prevent the dominance of characteristics with greater magnitudes, the feature values were re-scaled between zero and one using the min-max normalization technique 1.1. This makes it possible to compare characteristics fairly.

$$normalizedvalue = (value - minvalue)/(maxvalue - minvalue) \tag{1.1}$$

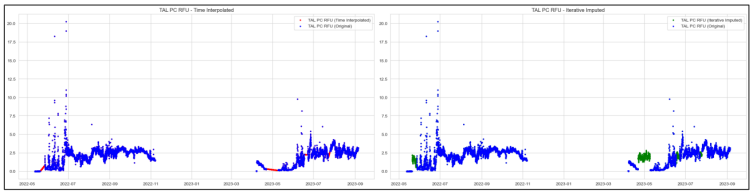
1.4.2 Feature Selection

The feature selection step identifies the most important features for Phycocyanin predictions to understand HABs better. The XgBoost and RF models' built-in feature importance functions were used in feature importance approaches. These techniques analyze each feature's influence on the prediction of the target variable and rank each one according to its significance. Ensemble learning methods like RF and XgBoost models are known for handling complicated connections and capturing feature interactions. For each variable, these models offer a feature importance score that indicates the feature's importance in the prediction. These models provide a feature importance score for each variable that denotes the significance of the feature in the prediction. Figure 1.8 shows the relative importance of each feature in contributing to the prediction. The feature importance values were derived from the RF model[1], which quantifies the significance of each feature based on its impact on the prediction accuracy.

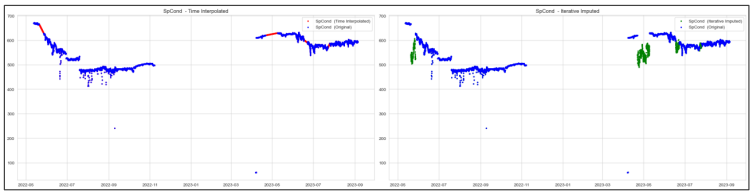
We examined the partial effect of each feature on the prediction of the target variable. The partial effect[17] refers to the influence of a specific feature while the rest of the features



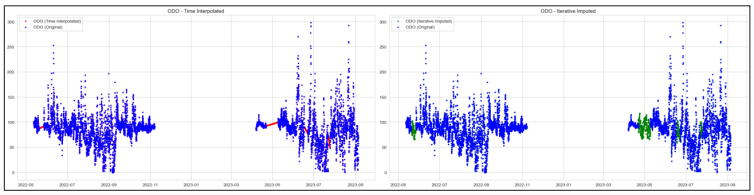
(a) Chlororphyll



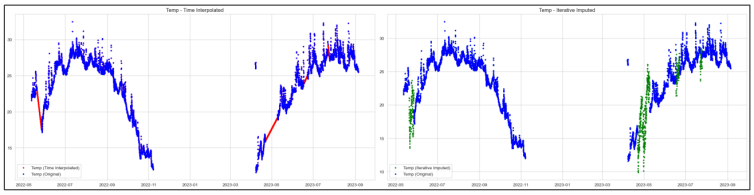
(b) Phycocyanin



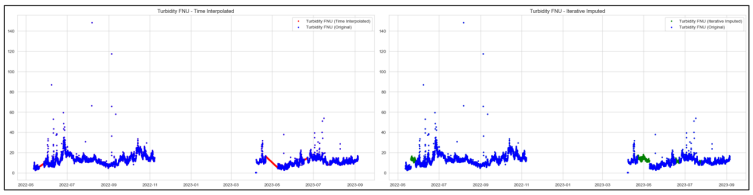
(c) Specific Conductivity



(d) Dissolved Oxygen



(e) Water Temperature



(f) Turbidity

Figure 1.7: Comparison of Missing Value Imputation Methods: Time Interpolation vs. Iterative Imputation. The left side of the graph showcases values filled using time interpolation, depicted in red, while the right side illustrates those filled via iterative imputation in green. This figure specifically focuses on features with a significant percentage of missing values.

remain unchanged. By analyzing the partial effects, we better understood how individual features contribute to the overall prediction and their relative importance.

The permutation importance procedure was utilized to determine a variable's effect on the model. This method involves arbitrarily shuffling the values of a single feature while leaving the values of the other features unchanged. The resulting decrease in model performance is then measured to determine the significance of this feature. The permutation importance values were computed using the permutation importance function of the scikit-learn library. This method comprehensively measures feature importance by considering feature interactions and dependencies.

By analyzing the partial effects and outcomes of permutation importance, we can identify the essential characteristics that have a significant impact on the prediction of the target variable. This information facilitates feature selection and model interpretation, allowing us to prioritize the most influential features and enhance the model's overall predictive performance. The partial effect of each feature on the target variable using RF, when a prediction is made three days ahead, is represented in Figure 1.14. TAL-PC after three days while holding all other features constant. Notably, the partial dependence plot for inflow demonstrates a positive, non-linear correlation, signifying that as inflow increases, we can expect a rise in TAL-PC, although at a decreasing rate. Conversely, the plot for Precipitation reveals a negative, linear connection, indicating that increasing Precipitation leads to a decrease in TAL-PC. The plot shows a positive, non-linear relationship regarding Relative Humidity, implying that higher humidity levels contribute to an increase in TAL-PC, albeit with diminishing returns. The Wind Direction plot portrays a complex, non-linear relationship that varies depending on other feature values. We observe a positive, non-linear correlation for Wind Speed, and similarly, the Storage plot displays a positive, non-linear connection. These partial dependence plots collectively emphasize the importance of all features in predicting TAL-PC after three days. However, these relationships are intricate and non-linear.

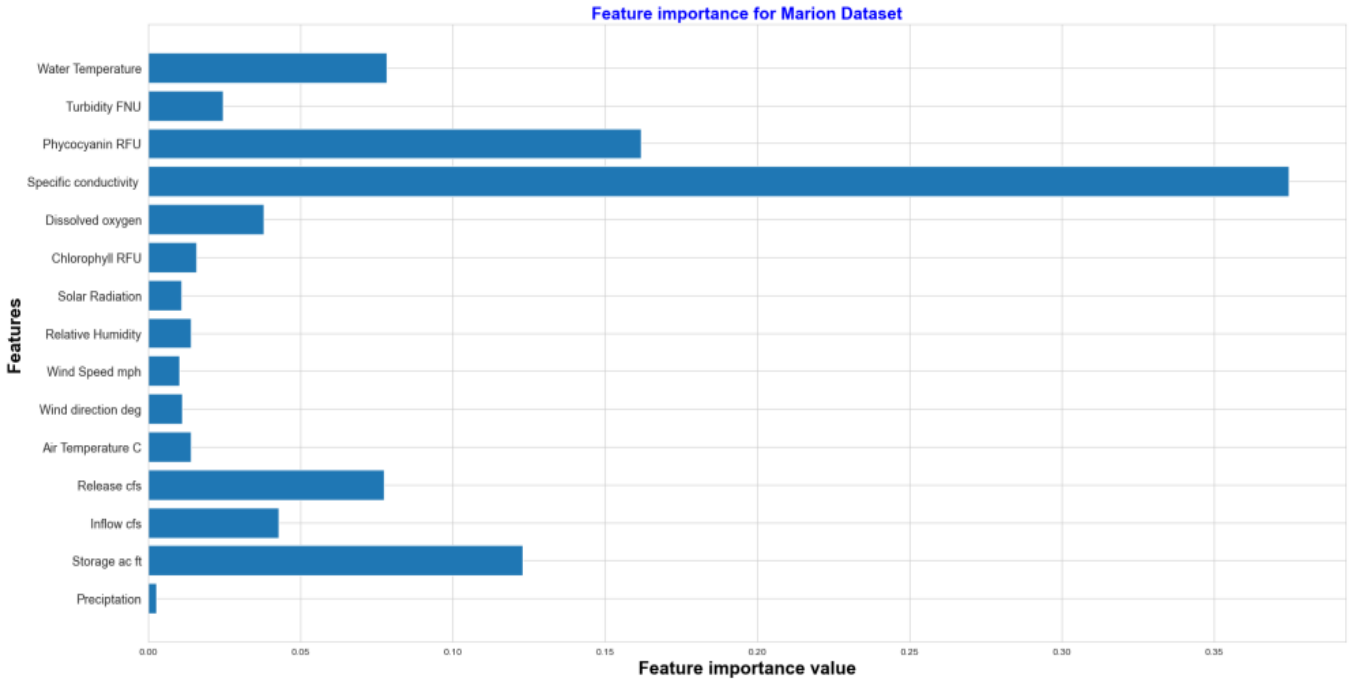


Figure 1.8: The figure illustrates the feature importance of the dataset for predicting Phycocyanin using the RF model[1].

1.4.3 Method

This project used different predefined machine learning, neural networks, and ensemble models. These models utilize a set of 15 features from the dataset as mentioned in Table 1.1, with the Phycocyanin level after a specified duration serving as the dependent variable. The parameters were tuned to identify the best model for the data.

As a part of this process, we first started with the RF model, the most commonly used model for predicting algal blooms. We employed a RF algorithm[1] for analysis. RF is an ensemble technique which combines the prediction outputs of various decision trees to produce precise predictions. RF is a popular machine learning technique for both classification and regression tasks, valued for its flexibility and strength. As an ensemble learning approach, RF creates a "forest" consisting of multiple decision trees. Each tree is developed from a randomly chosen subset from the training dataset and a random set of attributes. These individual trees are typically known a "base" or "weak" learners. The collective decision

from all these trees results in a more robust and accurate prediction. These decision trees are built using a technique known as bootstrap aggregating, often known as bagging. Each tree is trained in bagging using a unique, replacement-selected random subset of the training data. To provide even more randomness and avoid over-fitting, only a random subset of features is taken into account at each branch in the tree.

Each DT in the RF individually predicts something during the prediction phase. The final result is determined through a majority vote (in classification) or by averaging the outcomes from the individual trees (in regression). Using an ensemble technique increases the model's overall accuracy and robustness while the variance decreases. In addition to its ability to manage large datasets with high dimensionality, handle missing values and outliers, and offer estimates of feature relevance, RF has several benefits. Furthermore, it is resistant to over-fitting and typically only needs a small amount of hyper-parameter adjustment. The following parameter settings has been used $n_{\text{estimators}} = 400$, $min_{\text{samples\_split}} = 5$, $min_{\text{samples\_leaf}} = 1$, $max_{\text{features}} =$ 'sqrt', $max_{\text{depth}} = 100$, and $bootstrap =$ True. These values were selected based on prior studies, empirical experiments, and tuning procedures to optimize the RF model's performance. The results of this model are described in Table 1.3 and Figure 1.9.

We used the XgBoost supervised learning algorithm for our experiment. XgBoost[18] framework was used for gradient boosting because of its remarkable performance in handling complicated and non-linear data relationships. Extreme Gradient Boosting (XgBoost) is a powerful and well-known machine learning method that excels in resolving a variety of problems, particularly in the area of structured data analysis. It is well known for its ability to produce highly accurate predictions and handle large datasets faster.

The foundation of XgBoost's architecture is the gradient boosting principle, which combines several weak learners (Decision Trees) to create a powerful prediction model. It uses a gradient-boosting framework to build decision trees sequentially and repeatedly, each time optimizing a given objective function. The approach uses gradient descent optimization to reduce the loss function, enabling the model to improve its predictions with each iteration by learning from previous failures[18].

To avoid over-fitting and promote generalization, XgBoost employs a regularized boosting method. In order to manage the model's complexity and promote simplicity, regularization terms are introduced to the objective function. This decreases the possibility of the training data getting over-fitted. The model prediction of the Phycocyanin using XgBoost is represented in Table 1.2, and the results of XgBoost predicted at different times are represented in Figure 1.11

Ensemble models are used as the major modeling approach because of their ability to combine the predictions of numerous individual models, resulting in better accuracy and reductions in errors. In this study, we implemented two ensemble models and compared their performance, and the best performance is used in further study.

In one of the ensemble models, we used the XgBoost, Randomforest, and Long short-term memory models. XgBoost has been widely utilized, and the next model used in this ensemble is RF. RF[1] itself is a form of ensemble learning that involves constructing a large number of decision trees during training, after which the approach outputs either the class or an average forecast of the individual trees. A random subset of the training data is used to train each DT that makes up the RF. The features used to train the trees are also chosen randomly. The presence of this randomness contributes to a reduction in over-fitting and an improvement in generalization. RF is not only able to handle high-dimensional datasets, but it is also resistant to outliers and noisy data. It is well-known for its capacity to comprehend intricate linkages and non-linear trends present in the data.

We made use of the RandomForestRegressor class that is included in the scikit-learn library. The training data was used to train the RF model, given independent parameters, and the target feature, which in this case was the presence of Phycocyanin in the water. The model was configured with a maximum depth of 20 and a total of 100 estimators.

The Keras library with a TensorFlow backend was utilized to implement the LSTM model. The proposed model architecture comprises three LSTM layers, each having 50 units. Subsequently, dropout layers with a dropout rate of 0.2 are employed. A univariate output

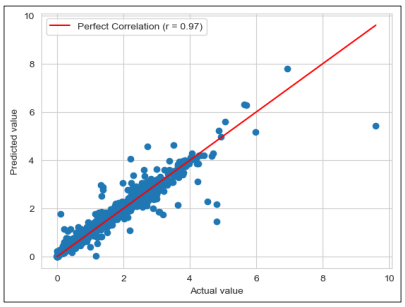
layer was appended in the form of a dense layer. Using the scaled training data, the LSTM model underwent 50 epochs of training with a batch size of 32. The results of predicting Phycocyanin using LSTM are represented in Table 1.4.

The ensemble model was constructed by averaging the RF, XgBoost, and LSTM predictions. To evaluate their efficacy, we generated predictions using the test data the results are represented in the Figure 1.10. We calculated Pearson's correlation coefficient and the MSE between the observed and predicted values represented in the Table 1.5. This ensemble model yielded the most favorable results in our analysis.

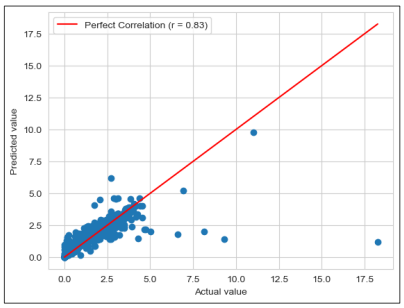
An ensemble model was made using a combination of several algorithms: RF, SVM, K Neighbors Regressor, and Gaussian Process Regressor, all of which were sourced from the Scikit Learn library. The ensemble's predictions were generated by averaging the outcomes of all these individual models. The corresponding results can be found in Table 1.6. Subsequently, a modification was introduced to the ensemble technique. For each record's prediction, we considered the predictions of 2*n surrounding records: n records preceding and n records following the target record. The final prediction for the record was then determined by calculating the mean of these 2*n predictions. Unfortunately, this augmented ensemble approach did not yield improved performance, and the results are documented in Table 1.7.

Prediction Time	**Correlation Coefficient**	**RMSE**
1 Hour	0.9745	0.2541
10 Hours	0.8832	0.5717
1 Day	0.9151	0.4985
7 Days	0.9469	0.3582
14 Days	0.8989	0.5241
30 Days	0.9107	0.4797
60 Days	0.9731	0.2307

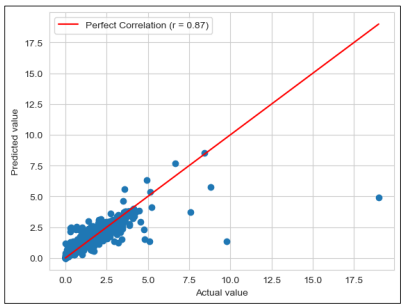
Table 1.2: The table above illustrates the performance of the XgBoost model for predicting the different amounts of Phycocyanin ahead of time. The correlation coefficient and the RMSE are the values between observed and predicted values.



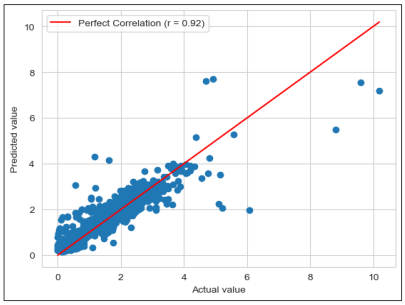
(a) One hour



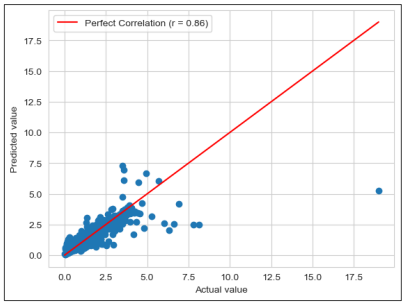
(b) 10 Hours



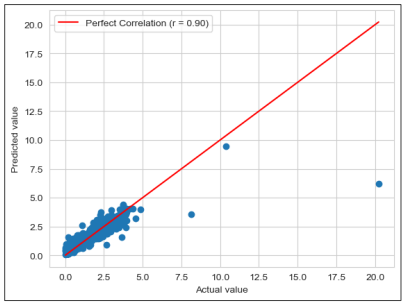
(c) One day



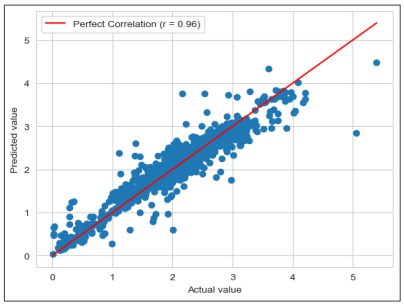
(d) Seven days



(e) 14 days



(f) 30 days

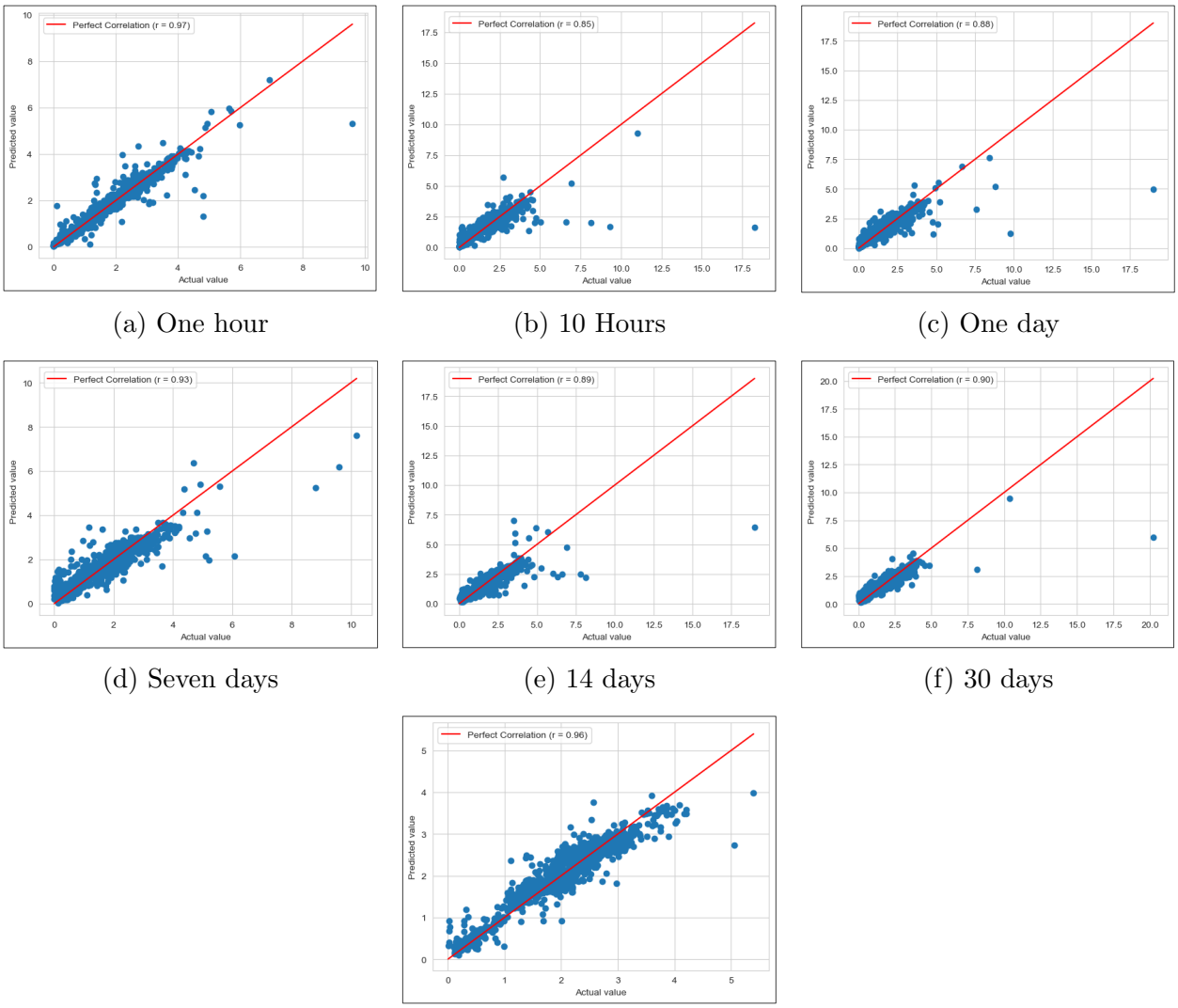


(g) 60 days

Figure 1.9: The prediction of Phycocyanin ahead of different time spans using RF is illustrated in the above subplots.

1.4.4 Cross-Validation

Cross-validation using the k-fold method is used in the study to validate the performance of different methods used for forecasting Phycocyanin. A dataset is split into k subsets, or folds, for testing purposes, with the remaining k-1 folds being used for training in k-fold cross-validation[21]. The results of cross-validation for Marion reservoir data using 10-fold cross-validation are averaged and are specified in Figure 1.15 which is calculated for the RF, XgBoost, and LSTM model.



(a) One hour (b) 10 Hours (c) One day

(d) Seven days (e) 14 days (f) 30 days

(g) 60 days

Figure 1.10: Predictions of Phycocyanin at various future intervals using an ensemble of RF, XgBoost, and Long Short-Term Memory model.

Over 30 days, the Mean Squared Error (MSE) is calculated to be 0.1497, with a standard deviation of 0.0736, which is relatively high. This observation suggests that the model's predictions exhibit less variability, with a corresponding moderate average error. Over 60 days, the MSE exhibited a lower value of 0.0753, with a reduced standard deviation of 0.0274. These findings indicate a rise in prediction accuracy. For one hour, the MSE is calculated to be 0.1702, with a standard deviation of 0.0836. Notably, these values indicate a bigger magnitude compared to the prior findings. This suggests that short-term forecasts may exhibit slightly lower levels of accuracy and a slightly greater level of variability. For 10

Prediction Time	Correlation Coefficient	RMSE
1 Hour	0.97	0.2
10 Hours	0.84	0.6
1 Day	0.87	0.5
7 Days	0.92	0.4
14 Days	0.86	0.6
30 Days	0.90	0.5
60 Days	0.96	0.2

Table 1.3: The table above illustrates the performance of the RF model for predicting different time periods ahead of time. The correlation coefficient and RMSE values represent the accuracy between observed and predicted values.

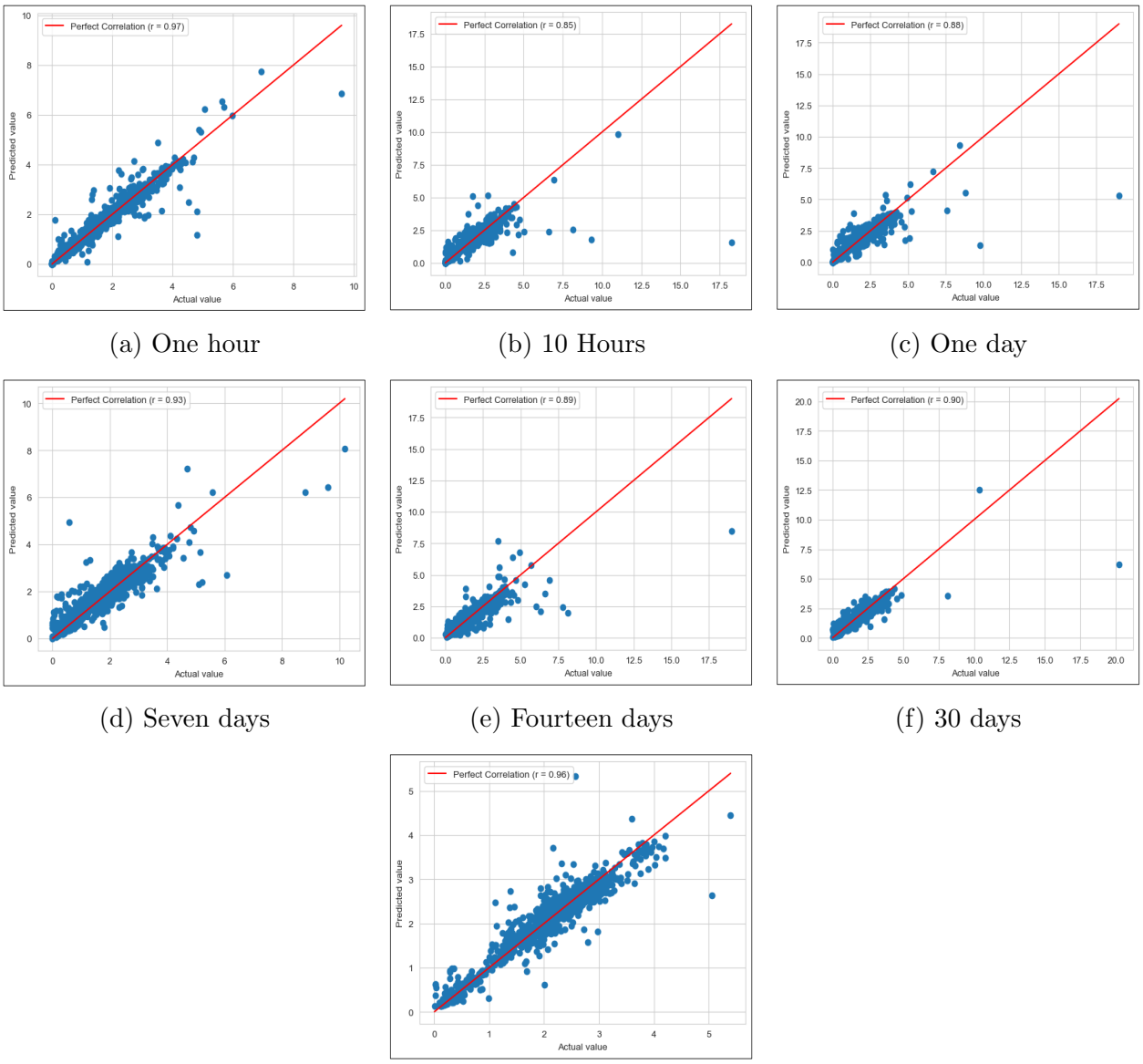
Prediction Time	Correlation Coefficient	RMSE
1 Hour	0.9608	0.2605
10 Hours	0.8229	0.6418
1 Day	0.8204	0.6476
7 Days	0.8409	0.5493
14 Days	0.8003	0.6751
30 Days	0.8975	0.4857
60 Days	0.9602	0.2522

Table 1.4: The table above illustrates the performance of the LSTM model for predicting the different amounts of Phycocyanin ahead of time. The correlation coefficient and the RMSE are the values between observed and predicted values.

Prediction Time	Correlation Coefficient	RMSE
1 Hour	0.9738	0.23581
10 Hours	0.8804	0.65783
1 Day	0.9120	0.5088
7 Days	0.9445	0.43670
14 Days	0.9004	0.5219
30 Days	0.9030	0.5018
60 Days	0.9718	0.2375

Table 1.5: The table above illustrates the performance of the ensemble model created using RF, XgBoost, and LSTM for predicting the different amounts of Phycocyanin ahead of time. The correlation coefficient and the RMSE are the values between observed and predicted values.

hours, the MSE is calculated to be 0.1935, accompanied by the largest standard deviation of 0.1130. This observation implies that long-term projections exhibit a notable degree of



(a) One hour (b) 10 Hours (c) One day

(d) Seven days (e) Fourteen days (f) 30 days

(g) 60 days

Figure 1.11: The prediction of Phycocyanin ahead of different time spans using XgBoost is illustrated in the above subplots.

uncertainty and unpredictability.

The correlation coefficients observed during cross-validation in the presented findings indicate the model's capacity to accurately depict the association between predicted and real Phycocyanin levels across various periods. The model exhibits a high and significant correlation across several time frames, with coefficients of 0.9407 for 30 days, 0.9640 for 60 days, 0.9424 for one hour, and 0.9323 for 10 hours. These findings suggest that the model

Prediction Time	Correlation Coefficient	RMSE
1 Hour	0.96	0.30278
10 Hours	0.85	0.59982
1 Day	0.86	0.59389
7 Days	0.89	0.47080
14 Days	0.85	0.61580
30 Days	0.87	0.6184
60 Days	0.94	0.1451

Table 1.6: The table above illustrates the performance of the ensemble model created using RF[1], SVM, K Neighbors Regressor, and Gaussian Process Regressor for the prediction of the different amounts of Phycocyanin ahead of time. The correlation coefficient and the RMSE are the values between observed and predicted values.

Prediction Time	Correlation Coefficient	RMSE
1 Hour	0.26564	0.66564
10 Hours	0.279407	0.59744
1 Day	0.2672	0.56851
7 Days	0.238041	0.47875
14 Days	0.2285692	0.42733
30 Days	0.23952	0.36157
60 Days	0.189146	0.25356

Table 1.7: The table above illustrates the performance of the ensemble model created using RF[1], SVM, K Neighbors, and Gaussian Process Regressor for the prediction of the different amounts of Phycocyanin ahead of time by using the mean of n records above and below the predicting record (for the result above n = 5). The correlation coefficient and the RMSE are the values between observed and predicted values.

is successful at comprehending and forecasting ecological trends throughout these time intervals. In addition, the correlation coefficients for various time frames 30 days: 0.0234, 60 days: 0.0115, one hour: 0.0238, and 10 hours: 0.0313 exhibit low standard deviations. This indicates that the model consistently maintains robust correlations and can generate dependable predictions across diverse time intervals. Longer time frames demonstrate slightly higher correlations and consistent performance, while even the shortest time frame of one hour consistently exhibits strong correlations with minimal variability. This highlights the model's effectiveness in capturing both short-term fluctuations and long-term trends.

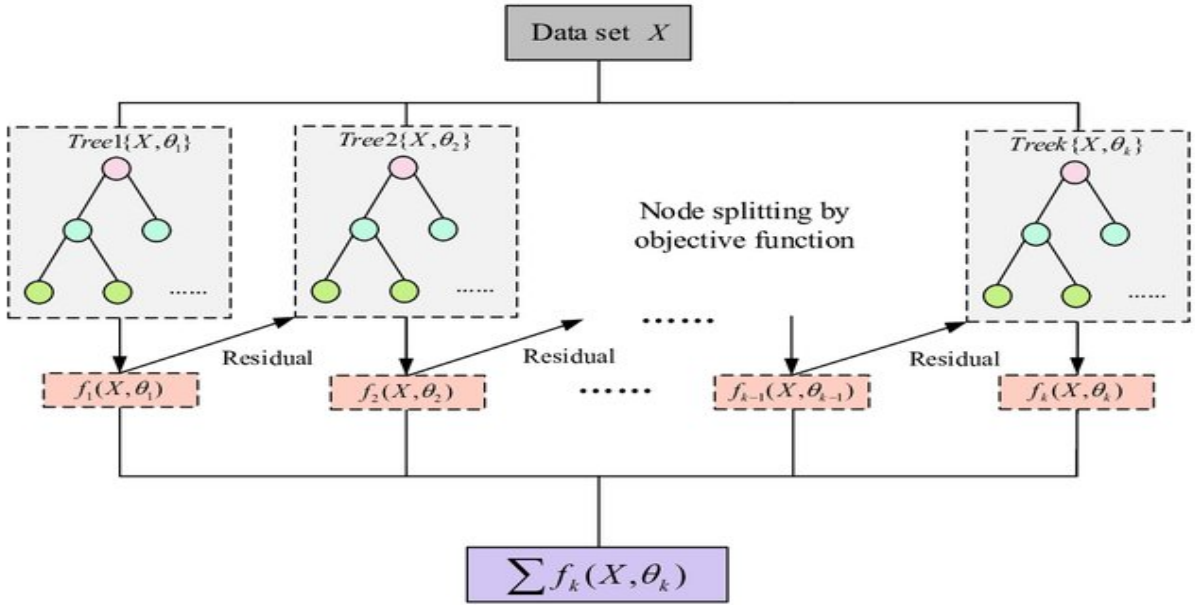


Figure 1.12: Figure that represents the architecture of XgBoost[19].

1.5 Impact of dataset size on Phycocyanin prediction

The prediction of Phycocyanin concentration exhibited significant variability with the increase in the number of records, as illustrated in Figure 1.16. This section explores the influence of dataset size on prediction accuracy and performance, shedding light on the intricate relationships between various data parameters and Phycocyanin concentration.

Our findings indicate a noteworthy trend: as the dataset size expanded from 1,000 records to 7,943 records, the overall predictive performance improved consistently. This observation underscores the importance of data volume in enabling the model to comprehend the complex interdependencies within the dataset and enhance its predictive capabilities.

An interesting insight emerged when we examined the performance at various dataset sizes. Initially, as we transitioned from 1,000 to 2,000 records, there was a sharp increase in performance. This can be attributed to the model grasping the fundamental relationships within the data. However, as the dataset continued to grow, the performance gains began to reduce, likely because the model started grappling with more intricate and nuanced data relationships. Remarkably, it was only around the 8,000-record mark that the model appeared

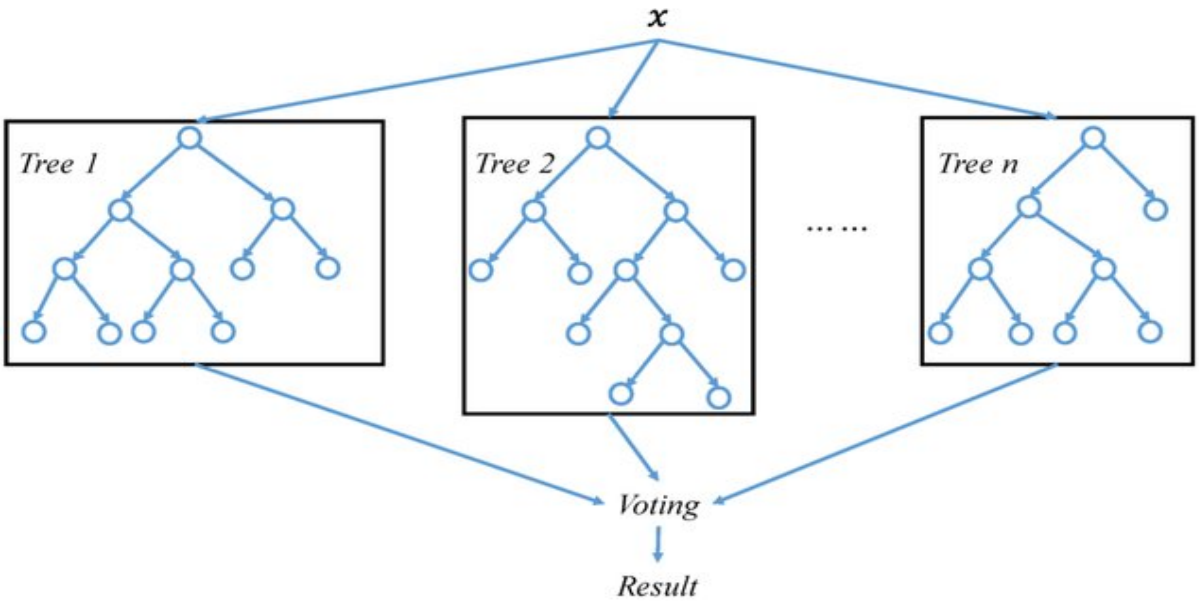


Figure 1.13: The figure represents the architecture of RF[20].

to fully comprehend the dataset's complexities, resulting in optimal predictive performance.

It is essential to note that the dataset size required for accurate Phycocyanin prediction may vary depending on the specific lake, its environmental conditions, and geographical factors. Therefore, the recommended dataset size of approximately 8,000 records serves as a valuable guideline but may necessitate adjustment based on the unique characteristics of the study area.

This analysis highlights dataset size's critical role in improving Phycocyanin predictions' accuracy. Understanding this relationship is fundamental to achieving precise predictions and contributes significantly to our knowledge of lake conditions and their impact on Phycocyanin concentrations.

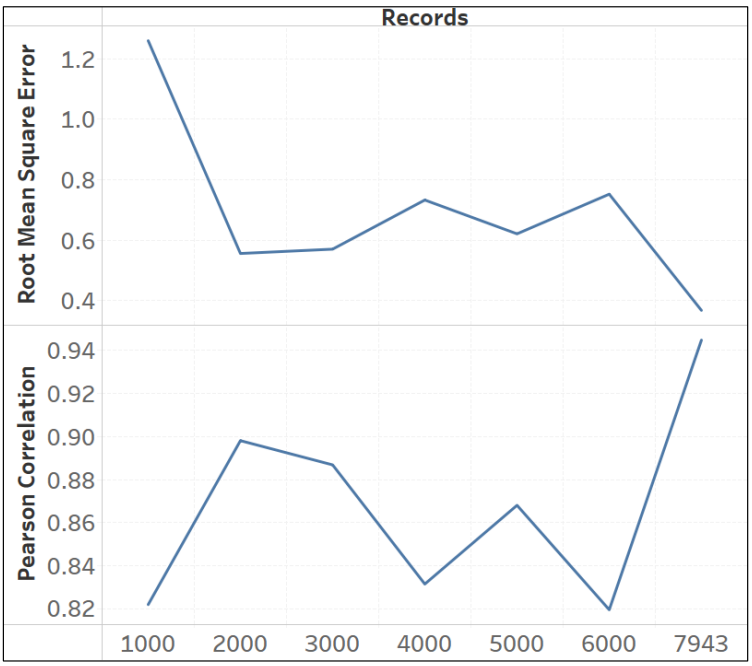


Figure 1.16: The effect of the dataset size on the performance of the ensemble model (RF, XGB, LSTM) prediction of Phycocyanin for a prediction duration of seven days

1.6 Comparison of Marion and Owasco Lake data

In this analysis, we collected data from Lake Owasco and compared it with Marion Reservoir, which has similar bloom occurrences. Owasco Lake, a picturesque body of water in the Finger Lakes region of New York State, is situated at approximately 42.8272° N latitude and 76.4897° W longitude[22]. These coordinates place the lake within the heart of the Finger Lakes, a region renowned for its natural beauty and pristine freshwater resources. Unlike Marion Lake, Owasco Lake is a glacier-formed lake. Owasco Lake's unique geographical location plays a significant role in its environmental characteristics and ecosystem dynamics. The lake is seven miles long with a width of one mile, average and maximum depths of 95ft and 177ft, respectively, and a volume of 212 billion gallons[23]. The bird's-eye view of the Owasco Finger Lake is represented in Figure 1.17a.

Table 1.8 compares the correlation coefficients between Marion Lake and Owasco Lake for various forecast periods. The linear relationship between two variables' strength

and direction is measured by the correlation coefficient. Within this particular context, it signifies the level of effectiveness exhibited by the prediction models in accurately collecting the instances of bloom occurrences in both lakes. The percentage of missing values in Owasco Lake is specified in Figure 1.18. Compared to Marion, the percentage of missing records is less in Owasco Lake data.

When considering a forecast length of 10 hours, it is shown that Owasco Lake has a significantly high correlation value of 0.9968, but Marion Lake demonstrates a correlation coefficient of 0.8804. This implies that the model's forecasts for Owasco Lake exhibit a high correlation with the observed events compared to Marion Lake. When the forecast length is increased to one day, Owasco Lake has a strong correlation value of 0.9964, indicating a high level of correlation. Similarly, Marion Lake also demonstrates an improved correlation coefficient of 0.9120. In the case of extended periods, specifically 14 days, one week, one month, and two months, it is noteworthy that Owasco Lake exhibits consistently strong correlation coefficients, ranging from 0.9954 to 0.9982. Conversely, Marion Lake displays a range of correlation coefficients between 0.9004 and 0.97188.

The analysis of the correlation coefficients between Marion Lake and Owasco Lake suggests that Owasco Lake constantly performs better than Marion Lake in terms of the prediction models' capacity to capture bloom occurrences accurately.This is partly because Marion Lake exhibits a wider range of cyanobacteria concentration values than Owasco Lake.

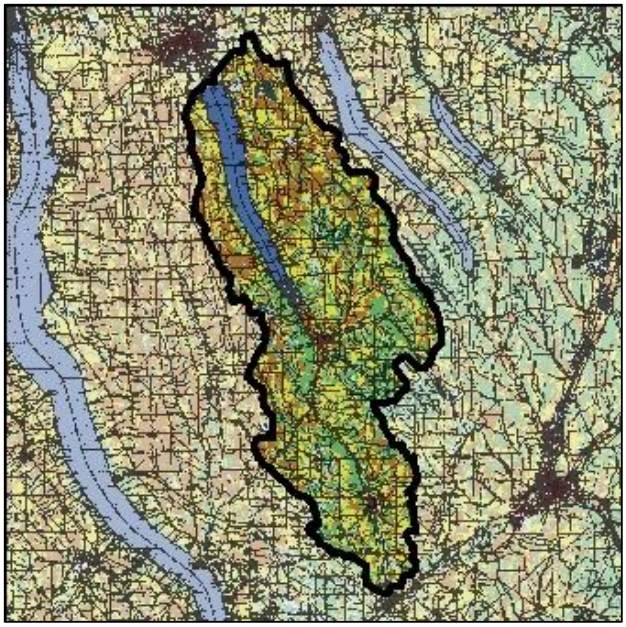
Table 1.9 presents a comparative analysis of the Root Mean Square Error (RMSE) values for Marion Lake and Owasco Lake over various forecast periods. By calculating the average amount of the prediction errors, the RMSE gives important information about how accurate the models are.

In the context of a 10-hour forecast time, it is seen that Owasco Lake has a very low RMSE value of 0.0323, but Marion Lake displays a substantially larger RMSE value of 0.3860. This finding suggests that the predictive models for Owasco Lake exhibit a higher degree of accuracy and a lower margin of error in comparison to those for Marion Lake. When the

forecast length is extended to one day, Owasco Lake's RMSE is consistently low at 0.0338. Similarly, Marion Lake's RMSE also shows improvement, reaching 0.4948. Owasco Lake exhibits consistently low RMSE values throughout extended periods, such as 14 days, one week, one month, and two months, with values ranging from 0.0232 to 0.0338. In contrast, Marion Lake's RMSE values fluctuate between 0.0232 and 0.5219.

The comparative analysis of RMSE values for Marion Lake and Owasco Lake, as shown in Table 1.9, indicates that Owasco Lake consistently reports lower RMSE values. The notable discrepancy in RMSE between the lakes can be attributed to Marion Lake's wider range of phycocyanin concentrations, which introduces larger prediction errors, thereby increasing its RMSE compared to the more stable data from Owasco Lake. Thus, the predictive models for Owasco Lake demonstrate remarkable precision with minimal errors, highlighting the reliability of Owasco Lake data for forecasting algal bloom occurrences. We identified the top five critical features influencing the water ecosystem by averaging the feature importance values from both XGBoost and RF methodologies. As outlined in Table 1.10, it becomes evident that while blooms might appear similar across different ecosystems, the underlying factors driving these phenomena can differ substantially.

The analysis conducted reveals our efforts to compare these two lakes just to solidify that similar models could be applied to contrasting lake ecosystems, although the factors influencing the bloom might vary.



(a) Owasco lake map[24].

(b) Owasco Lake algal bloom in the year 2018[25].

Figure 1.17: Owasco Finger Lake, New York.

1.7 Conclusion

This chapter primarily used data science approaches to predict cyanobacteria blooms in Marion Reservoir. HABs are a major threat to the watery ecosystem because they make the water unsafe for people, animals, and aquatic life to use. It is important to be able to predict the future algal bloom to handle HABs well and lessen their effects.

Prediction Duration	Marion Lake	Owasco Lake
10 Hours	0.8804	0.9968
1 Day	0.9120	0.9964
1 Week	0.9004	0.9982
14 Days	0.9030	0.9975
30 Days	0.9078	0.9974
60 Days	0.97188	0.9954

Table 1.8: Comparison of predicted value correlation coefficients between Marion and Owasco Lake using RF, XGBoost, and LSTM ensemble model.

Prediction Duration	Marion Lake	Owasco Lake
10 Hours	0.3860	0.0323
1 Day	0.4948	0.0338
14 Days	0.5219	0.0232
1 Week	0.3846	0.0278
30 Days	0.5071	0.0257
60 Days	0.3233	0.0288

Table 1.9: Comparison of predicted value RMSE between Marion and Owasco Lake using RF, XgBoost, and LSTM ensemble model.

Owasco Lake	Marion Lake
Specific Conductivity	Phycocyanin
Inflow	Release
Turbidity	Storage
Chlorophyll	Water Temperature
Storage	Specific Conductivity

Table 1.10: Top five important features based on feature importance of ensemble of RF and XgBoost for a prediction duration of three days.

Machine learning and recurrent neural network methods were used to make predictions based on a large set of collected water quality and environmental factors. The main goal was to see how well different machine learning systems could predict cyanobacteria blooms and figure out what external factors were most likely to cause them.

This study explores an important problem that affects the environment and general health. By applying data science approaches to predict HABs, the study contributes to the growing body of literature on freshwater management. Water management organizations

and lawmakers can use the findings to lessen the effects of HABs and ensure the safety of water supplies and water recreational areas.

The data used in this study was obtained from Marion Reservoir, an important water source for many places and a popular place for outdoor recreation. The data set consists of hourly readings from two devices. One sensor measured the state of the lake water, and the weather data was collected from the USACE webpage[11]. The dataset contained missing values, which were handled using appropriate methods for time series data. Min-max normalization was used to normalize the data so that features could be compared fairly and bigger features did not dominate the remaining features.

Techniques like XgBoost and RF models were used to determine the most important features for predicting cyanobacteria blooms, which gave us insight into how blooms occur. These models ranked the features by how important they were to the prediction and gave useful information about their relative importance and how they interacted with each other.

Ensemble models like XgBoost, RF, and LSTM were used to improve the accuracy of predictions. The ensemble model put together what these models said would happen. This made the data more accurate and reliable. With test data, Pearson's correlation coefficient and the MSE between actual and projected values were used to evaluate the ensemble models. The model's accuracy at predicting cyanobacteria blooms was checked by looking at how well it worked.

The impact of the size of the data on prediction is analyzed as part of this study. It is observed that the more data there is, the better it is for a machine learning model to understand the trends in data. The Marion reservoir data is compared with Owasco Lake, New York data. The comparative study provided insights into how similar methods can be applied to contrasting water bodies, and the parameters that significantly influence the algal bloom were not always the same.

Overall, this study shows how data science methods could be used to predict and control HABs. By knowing the most important natural factors and using machine learning

methods, water management bodies can make smart decisions to protect water supplies and public health. Future studies can be based on these results and explore other ways to predict and stop HABs.

In conclusion, this study contributes to the fields of data science and environmental management by shedding light on the capabilities of predicting cyanobacteria blooms and giving useful ways to lessen their effects.

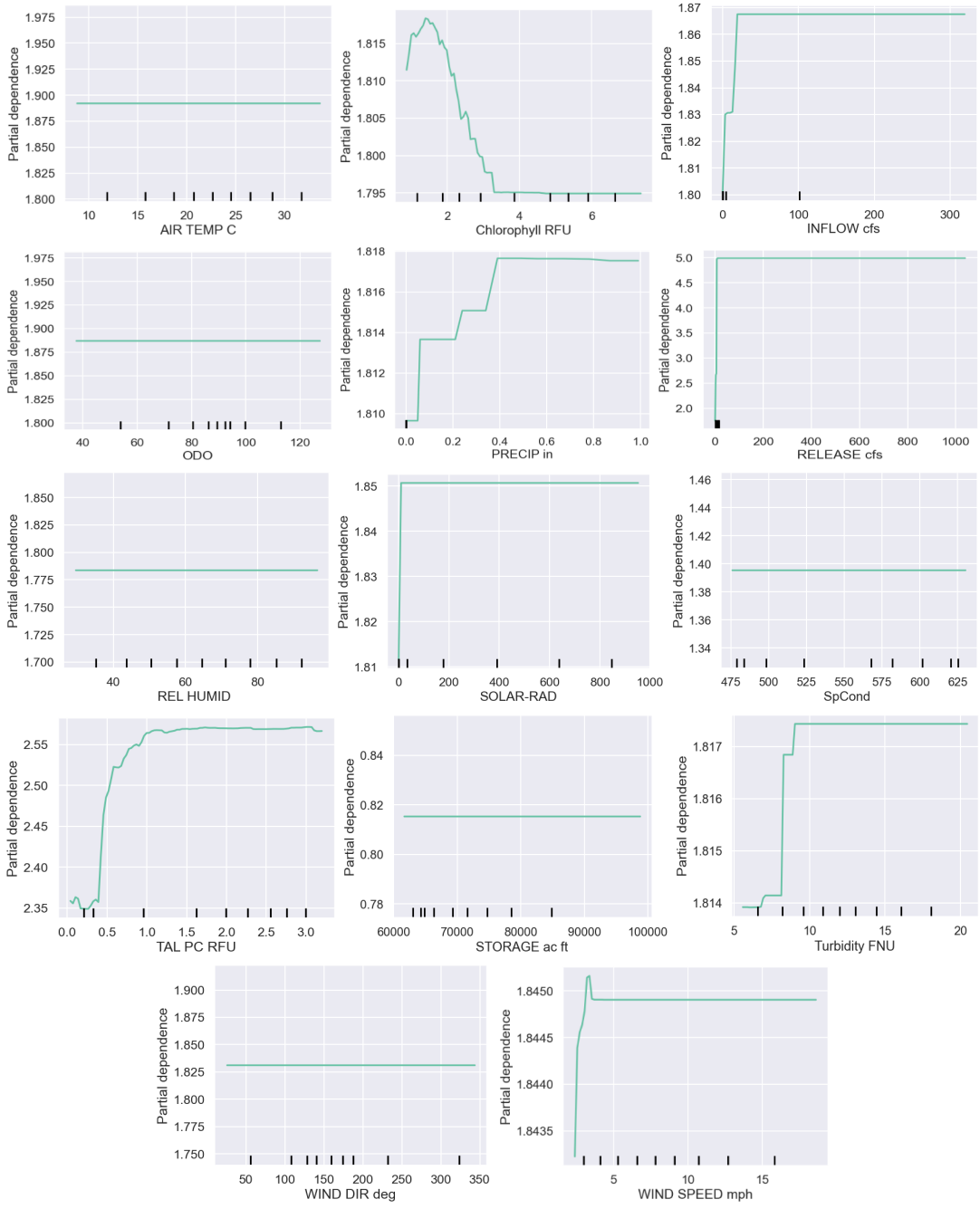
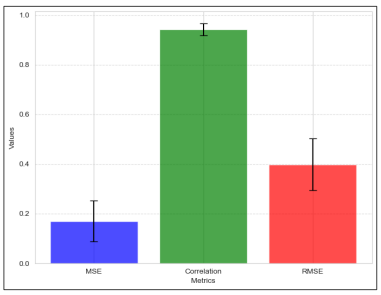
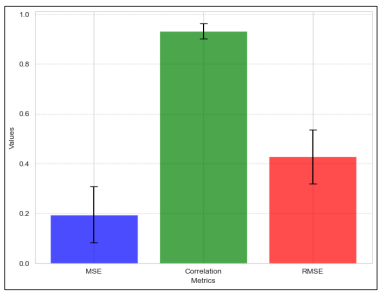


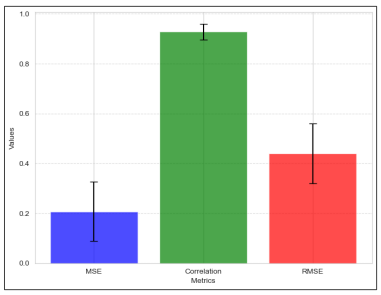
Figure 1.14: Partial dependency plot for all the features used in predicting Phycocyanin three days ahead using RF.



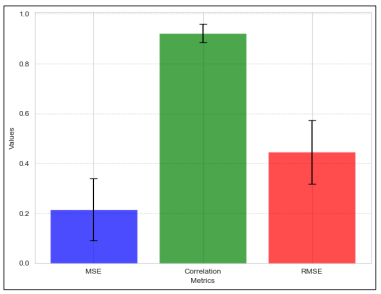
(a) One Hour



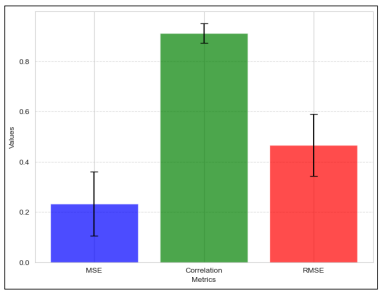
(b) Ten Hours



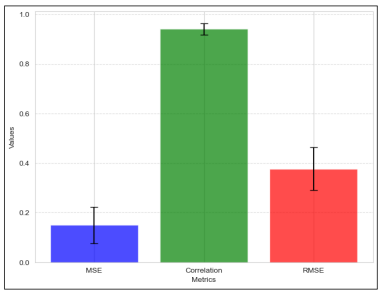
(c) One Day



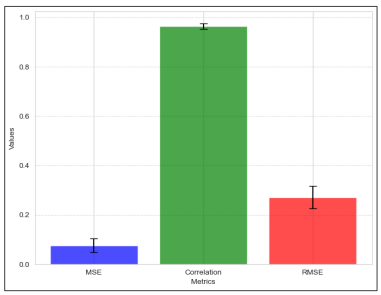
(d) Seven Days



(e) Fourteen Days



(f) Thirty Days



(g) Sixty Days

Figure 1.15: Cross-validation results of the ensemble model combining RF, XGBoost, and LSTM. The bar chart displays the average values of MSE, Pearson Correlation, and RMSE, with error bars representing one standard deviation.

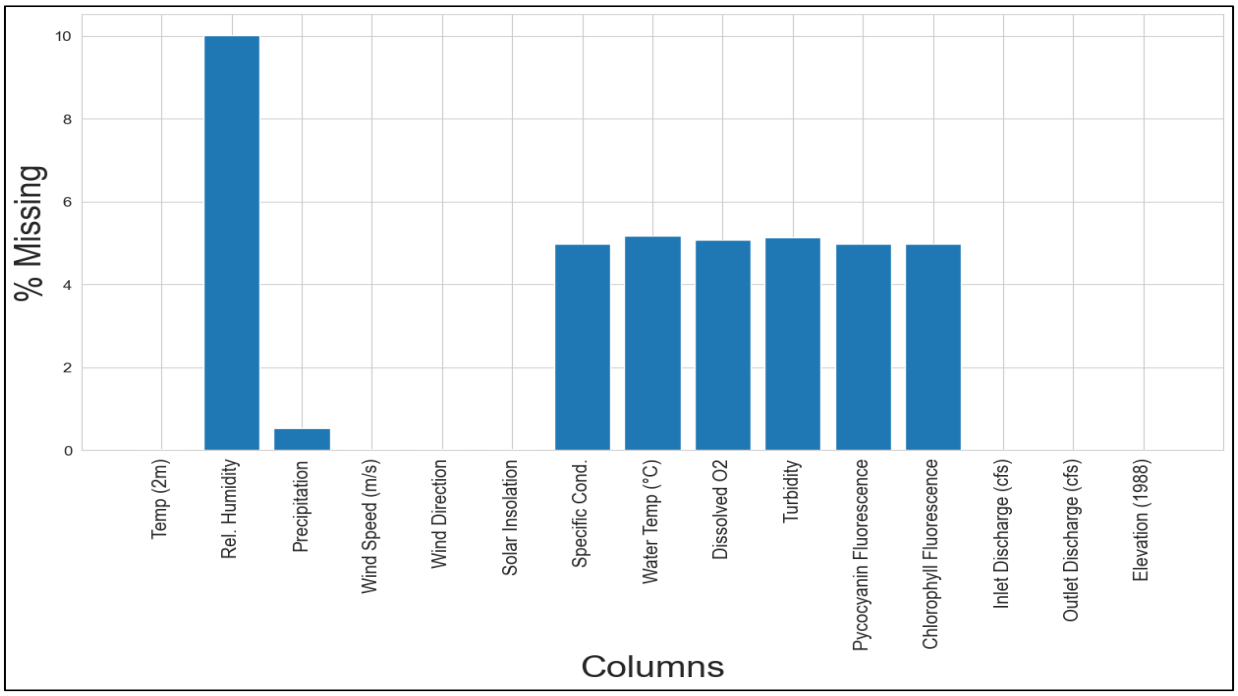


Figure 1.18: Figure represents the percentage missing values of each column in the dataset collected from Owasco Lake. The parameter Relative Humidity has the highest percentage of missing values with approximately 10%. Features like Specific Conductivity, Water Temperature, Dissolved Oxygen, Turbidity, Phycocyanin, and Chlorophyll concentration have approximately 5% of data that is missing.

Chapter 2

Quantitative analysis of racial bias in home mortgage loans

2.1 Summary

A home mortgage is one of the most common forms of financial loans and a primary source for household financing. Here we use 13 years of federal home mortgage data collected in the United States, which includes $\sim 7 \cdot 10^6$ loans. By applying machine learning, we show that the race and ethnicity of the borrower can be identified by using the mortgage information, indicating a link between the borrower and the loan information that is not directly related to the borrower's race. The analysis also shows that the information that correlates with the borrower's race is the borrower's gender. While reasons can include complex current and historical sets of biases, the results provide quantitative confirmation of racial biases embedded in home mortgage systems.

2.2 Introduction

A home mortgage is a critical and common financial tool that has substantial economic and social implications[26–28]. It is also a non-static field that changes consistently over time and responds to the economic and social ecosystem[29]. Naturally, the decision of a lender about a mortgage depends on a set of financial indicators related to the person and the property being purchased. Taking all information into account, the underwriter decides whether to approve or deny a mortgage application[26;30].

While the process of approving a loan is driven by economic factors, it might also include indicators that may directly or indirectly lead to racial bias[31–33]. For instance, racial biases played an essential role in approving high-cost, high-risk loans to black and Latino borrowers compared to white borrowers in the United States[34]. These activities also used community information to identify specific potential borrowers and earn their trust. Racial inequality is also reflected by higher rates of predatory lending, targeting specifically borrowers from underrepresented minorities[35]. Racial discrimination in mortgage approval has also been linked to differences in credit scores[36] and a higher chance of being denied a mortgage loan[37]. More recent studies have shown a decline in racial discrimination in mortgages[38]. Mortgage discrimination can also vary by geographical location[39].

In this context, it is crucial to understand the dynamics of the housing market and the factors that contribute to the availability and affordability of housing loans. By analyzing data from FHLB databases, decision-makers and industry stakeholders can get insight into housing market trends and make well-informed decisions about housing financing practices and policies.

The purpose of this study is to apply quantitative analysis to identify and profile patterns of racial discrimination in mortgages. Since a loan mortgage is a high-dimensional space, this study also uses machine learning to perform a comprehensive analysis of mortgage data using the data from the Federal Home Loan Bank.

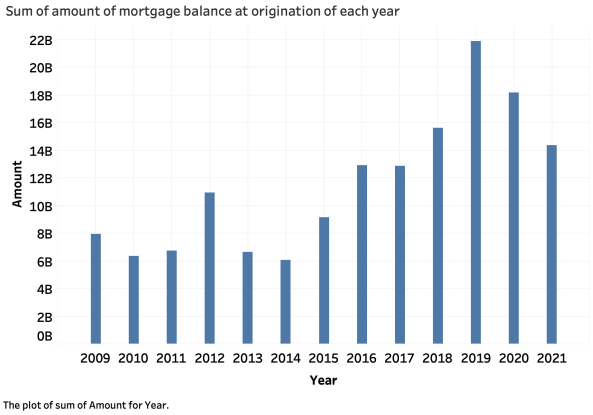
2.3 Data

The Federal Home Loan Bank (FHLB) is a US government-sponsored institution that was established in 1932 to provide funding for member banks and thrifts to make housing loans available to consumers. FHL Banks operate in a cooperative structure and provide low-cost funding to their members, which can then be used to make loans for the purchase, construction, or renovation of homes. Housing loans and mortgages are a critical part of the FHLB system. These loans provide individuals and families with the means to purchase homes and build wealth through home ownership. Typically, mortgages are long-term loans that the borrower repays in regular installments over a predetermined time period.

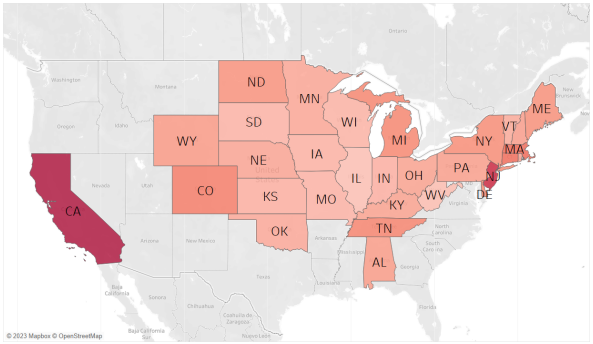
FHL banks play a significant role in the housing market by providing liquidity to their members, enabling them to provide mortgage loans to consumers. Through their lending activities, FHL banks help to promote home ownership and affordable housing, particularly for low and moderate-income households. However, housing loans and mortgages are not without risks. Mortgage borrowers who are not able to make their payments may default, leading to foreclosure and the loss of the property. Default rates can also have an impact on the financial stability of FHL banks and the broader housing market.

The FHLB data set covers the period from 2009 to 2021, i.e., 13 years of mortgage data, and contains 89 features. It provides a rich source of data for analyzing trends in the housing market. Table 2.1 shows the columns in the dataset. The dataset contains information about a very large number of $\sim 7 \cdot 10^6$ loans.

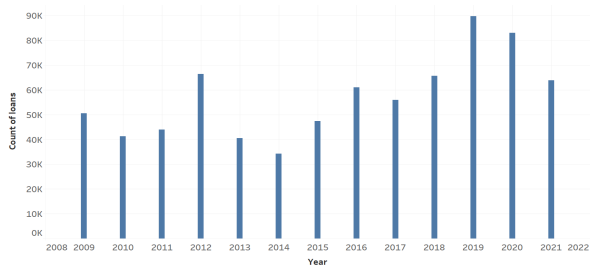
Figure 2.1a displays the sum of mortgage balance at the origination of loan assignment in each year. The average amount of loans assigned in each state is shown in Figure 2.1b. As the figure shows, the states of California and New Jersey have the highest average amount of loans assigned, and Illinois has the least. The sum of the number of loans assigned in each year is described in Figure 2.1c. The number of loans assigned to each race per year is represented in Figure 2.1d.



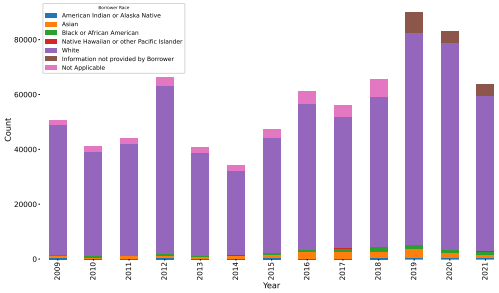
(a) Total mortgage balance at origination in each year.



(b) The distribution of loans across states state.



(c) Loans per Year.



(d) Distribution of Race per Year.

Figure 2.1: Mortgage information

Feature	**Data Type**	**Description**
Year	int64	Year Loan Was Reported
AssignedID	int64	Unique Record ID (not actual loan number)
FHLBankID	object	Name of Federal Home Loan Bank District
Program	float64	AMA Program
FIPSStateCode	int64	Two Digit FIPS State Code
FIPSCountyCode	int64	Three Digit FIPS County Code
MSA	int64	Core Based Statistical Area Code
FeatureID	float64	Geographic Names Information System (GNIS) Feature ID
Tract	float64	The property's Census Tract of Block Numbering Area (BNA)

MinPer	float64	The percentage of the property's census
TraMedY	int64	Tract population that is a minority.
LocMedY	int64	The property's median income for the area based on the most recent decennial census
TractRat	float64	Tract Income Ratio
Income	int64	Total Monthly Income Amount
CurAreY	int64	The current median income for a family of four for the area as established by HUD
IncRat	float64	Borrower Income Ratio
UPB	int64	The Amount of unpaid principal balance in whole dollars when acquired by the FHLBank.
LTV	int64	The loan-to-value ratio of the mortgage at time of origination
MortDate	int64	The loan-to-value ratio of the mortgage at time of origination
AcquDate	int64	Year the mortgage was acquired.
Purpose	int64	Purpose of Loan: 1 = Purchase, 2 = No-Cash Out Refinancing, 3 = Second Mortgage, 4 = New Construction, 5 = Rehabilitation or Home Improvement, 6 = Cash-out Refinancing, 7 = Other
Coop	float64	Cooperative Unit Mortgage 1 = yes; 2 = no
Product	int64	Purpose of Loan: 1 = Purchase, 2 = No-Cash Out Refinancing, 3 = Second Mortgage, 4 = New Construction, 5 = Rehabilitation or Home Improvement, 6 = Cash-out Refinancing, 7 = Other
FedGaur	int64	Type of mortgage, and whether the mortgage is guaranteed: 0 = conventional, 1 = FHA, 2 = VA, 3 = USDA Rural Housing-FSA Guaranteed, 4 = HECMs, 5 Title1-FHA
Term	int64	Term of the Mortgage in Months

AmorTerm	int64	For Amortizing Mortgages, the term of amortization in months; 998 if non-amortizing loan
SellType	int64	Type of institution from which the FHLBank acquired the mortgage 01 = Insured depository institution, 02 = Housing Associate, 03=Insurance Company, 04 = non-federally Federally Insured CU; 05 = Non-Depository CDFI. 06 = Other FHLBank, 09 = Other
FHFBID	float64	Acquiring Lender Institution Federal Housing Finance Agency Membership ID
Seller	object	Acquiring Lender Name
SellCity	object	Acquiring Lender City
SellSt	object	Acquiring Lender State
NumBor	int64	Number of Borrowers
First	int64	Numeric codes indicate whether the borrower is a first-time homebuyer. 0 = no, 1 = yes.
CICA	float64	Code indicate whether the mortgage is on a project funded under an AHP, CIP, or other CICA program.
BoRace	int64	Numerical code indicates the borrower's race. 1 = American Indian or Alaska Native; 2 = Asian; 3 = Black or African American, 4 = Native Hawaiian or other Pacific Islander; 5 = white; 6 = information not provided by the borrower; 7 = not applicable (the first or primary borrower is an institution, corporation, or partnership).

CoRace	int64	Numeric codes indicate the race of the co-borrower. 1 = American Indian or Alaska Native; 2 = Asian; 3=Black or African American; 4=Native Hawaiian or other Pacific Islander; 5 = White. 6 = information not provided by the borrower; 7 = not applicable (the first or primary borrower is an institution, corporation, or partnership); 8 = no co-borrower.
BoGender	int64	A numerical code indicating the sex of the first or primary borrower. 1 = male, 2 = female, 3 = information not provided by the borrower, 4 = not applicable (the first or primary borrower is an institution, corporation, or partnership), and 6 = borrower selected both male and female.
CoGender	int64	A numerical code indicating the sex of the co-borrower: 1 = male, 2 = female, and 3 = information not provided by the borrower. 4 = no co-borrower; 5 = not applicable (first or primary borrower is an institution, corporation, or partnership); 6 = borrower selected both male and female.
LienStatus	int64	Lien Priority Type
BoAge	int64	Age in years of the borrower at the time application submitted
CoAge	int64	Borrower2 Age at the time of Application
Occup	int64	Numerical code indicate whether the property is owner-occupied, a second home, or a rental investment property. 1 = Principal Residence, 2 = Second Home, 3 = Investment Property
NumUnits	int64	Total number of units in the property
Bed1	float64	Unit1–Number of Bedrooms 98 = no non-owner-occupied dwelling units

Bed2	float64	Unit2–Number of Bedrooms 98 = no non-owner-occupied dwelling units
Bed3	float64	Unit3–Number of Bedrooms 98 = no non-owner-occupied dwelling units
Bed4	float64	Unit4–Number of Bedrooms 98 = no non-owner-occupied dwelling units
Bed5	float64	Unit5–Number of Bedrooms 98 = no non-owner-occupied dwelling units
Bath1	float64	Unit1–Number of Bathrooms 98 = no non-owner-occupied dwelling units
Bath2	float64	Unit2–Number of Bathrooms 98 = no non-owner-occupied dwelling units
Bath3	float64	Unit3–Number of Bathrooms 98 = no non-owner-occupied dwelling units
Bath4	float64	Unit4–Number of Bathrooms 98 = no non-owner-occupied dwelling units
Bath5	float64	Unit5–Number of Bathrooms 98 = no non-owner-occupied dwelling units
Aff1	float64	Unit1–Affordable category meets the housing goals implemented by HERA Section 1205 1 = yes; 2 = no
Aff2	float64	Unit2–Affordable Category meets the housing goals implemented by HERA Section 1205 1 = yes; 2 = no
Aff3	float64	Unit3–Affordable Category meets the housing goals implemented by HERA Section 1205 1 = yes; 2 = no
Aff4	float64	Unit4–Affordable Category meets the housing goals implemented by HERA Section 1205 1 = yes; 2 = no

Rent1	float64	Unit1–Amount of Rent (whole dollars) 99999 = Not Applicable
Rent2	float64	Unit2–Amount of Rent (whole dollars) 99999 = Not Applicable
Rent3	float64	Unit3–Amount of Rent (whole dollars) 99999 = Not Applicable
Rent4	float64	Unit4–Amount of Rent (whole dollars) 99999 = Not Applicable
RentUt1	float64	Unit1–Affordable Rental Unit 1–Utilities Included 1 = yes; 2 = no
RentUt2	float64	Unit2–Affordable Rental Unit 2–Utilities Included 1 = yes; 2 = no
RentUt3	float64	Unit3–Affordable Rental Unit 3–Utilities Included 1 = yes; 2 = no
RentUt4	float64	Unit4–Affordable Rental Unit 4–Utilities Included 1 = yes; 2 = no
Geog	object	Geographic Division and Region
RateSpread	float64	Rate Spread
HOEPA	int64	Home Ownership and Equity Protection Act (HOEPA) Status: 1 = yes, 2 = no
Lien	float64	Lien Status1 = First Lien, 2 = Subordinate Lien, 3 = Not Applicable, 4 = No Lien, 5 = Not secured by a lien
MMIF	int64	Code indicate if the mortgage is insured by the Mutual Mortgage Insurance Fund (MMIF).
GEOID	int64	FIPS code, which is a unique identifier for the Census Tract
APPR	float64	Appraisal: 1 = Original Appraisal, 2 = Updated Appraisal, 3 = No Appraisal, 4 = Not Disclosed

Table 2.1: Features, Data Type, and Description of features in the dataset.

One of the main goals of this study is to identify patterns that correlate with the race of the borrower. In the dataset, each borrower is assigned one of seven races. Table 2.2 shows the races used in the dataset.

Race Number	Description
1	American Indian or Alaska Native
2	Asian
3	Black or African American
4	Native Hawaiian or other Pacific Islander
5	White
6	Information not provided by Borrower
7	Not Applicable

Table 2.2: The different borrower races used in the dataset.

2.4 Methods

This study is focused on quantitative analysis of federal home loan data to reveal trends and patterns in a multivariate fashion. For the purpose of classification, the common XGBoost classifier[18] is used. XGBoost is a widely used supervised machine learning that has shown good performance in the environment of high-dimensional datasets[18]. It uses a gradient-boosted tree classification algorithm. Gradient boosting, a type of supervised learning that combines the predictions of several tree-based classifiers to make a more accurate prediction of a target variable. When using the XGBoost classifier, 70% of the data was used for training, and the other 30% of the loans were used for testing.

The standard metrics of classification accuracy, precision, recall, and F1 are used to evaluate the classifier's performance. The informative features are identified by applying the Chi-square feature selection and analyzed further by applying the Student t-test.

2.5 Results

If the race of the borrower is not linked to the mortgage decision directly or indirectly, it is expected that the mortgage information cannot predict the race, as the information for all races is distributed regardless of race. Table 2.3 shows the classification accuracy of the race of the borrower based on the loan information.

As the table shows, the classification accuracy of the XGBoost classifier is far higher

Accuracy	0.951
Precision	0.947
Recall	0.951
F1 score	0.946

Table 2.3: Precision, recall, accuracy, and F1 score for the classification of borrower race using the XGBoost classification algorithm.

Accuracy	0.888
Precision	0.789
Recall	0.888
F1 score	0.836

Table 2.4: Precision, recall, accuracy, and F1 score when predicting the race of the borrower using a "dummy" ZeroR classifier.

than the mere chance accuracy of ∼14%. Table 2.4 shows the same analysis, but when using a "dummy" ZeroR classifier instead of the XGBoost classifier. As the differences between the tables show, the race can be identified from the data, and the lower accuracy when using a ZeroR classifier shows that the ability to predict the race is not driven merely by the uneven distribution of borrower race in the dataset, but by the ability of the classifier to identify the race of the borrower from the loan data.

The dataset also includes the ethnicity of each borrower, which can be either "Hispanic or Latino" or "not Hispanic or Latino." That information can correlate with the race and increase the ability to predict the race correctly. To avoid using that information, the analysis was also done after removing the "ethnicity" field.

Another piece of information that can directly help identify the race and is unrelated to the loan itself is the co-borrower race. For instance, if the borrower has a co-borrower, it is possible that the race of the co-borrower can provide certain information about the

Accuracy	0.754
Precision	0.756
Recall	0.754
F1 score	0.752

Table 2.5: Precision, recall, accuracy, and F1 score for the classification of borrower race after removing the borrower's ethnicity column from the data set.

Accuracy	0.922
Precision	0.910
Recall	0.922
F1 score	0.904

Table 2.6: Precision, recall, accuracy, and F1 score for the classification of borrower race after removing the co-borrower race and borrower ethnicity from the data.

borrower. When removing the ethnicity and the co-borrower race, the prediction accuracy of the race of the borrower is shown in Table 2.6.

Figure 2.3 shows the confusion matrix after applying the classification. As Table 2.6 and Figure 2.3 show, even when removing all race and ethnicity information of the borrower, a machine learning system can identify the race of the borrower with accuracy higher than mere chance or from a "dummy" classifier that merely uses the statistical distribution of the races in the dataset. This shows in a quantitative manner that the loan information is sensitive to the borrower's race. That is, the loan information of a borrower from a certain race has different patterns than that of a borrower of a different race.

The fact that the race of the borrower can be predicted from loan information indicates a certain racial bias in mortgages. That bias is not necessarily direct and can be indirect by correlating other variables with race rather than with the mortgage decision directly. A separate analysis for each year was performed to test if that bias changes over time. Figure 2.12a shows the change in the accuracy of predicting the borrower's race from the loan data using XgBoost. The race identification accuracy was determined using the XGBoost classifier as done above, and without using ethnicity or co-borrower race information.

To profile the ability to predict the race of the borrower based on the loan information, each pair of races was tested as a two-way classification problem to test how well a classifier can identify between the two races. That is, the dataset is divided into smaller datasets that contain just two races. Table 2.7 contains the classification results. From the results, it is evident that the race of American Indian or Alaska Native is the most predictable one. The next most predictable races are Asians, Blacks, and African Americans.

Race	American Indian or Alaska Native	Asian	Black or African American	Native Hawaiian or other Pacific Islander	White
American Indian or Alaska Native	1	0.83	0.80	0.75	0.81
Asian	0.83	1	0.77	0.70	0.79
Black or African American	0.80	0.77	1	0.70	0.79
Native Hawaiian or other Pacific Islander	0.75	0.70	0.70	1	0.64
White	0.81	0.79	0.79	0.64	1

Table 2.7: Accuracy metrics for the predictability of borrower's race between two distinct races.

2.5.1 Feature selection

As described above, a machine learning classifier is able to identify the race of the borrower from the mortgage information alone, which shows evidence of racial bias in mortgages. We used XGBoost embedded feature selection to figure out which features were most important for identifying the race in order to make a profile of the signs that were linked to this possible bias. Figure 2.2 displays the features that have the strongest predictive information for the prediction of the borrower's race. Table 2.8 shows the mean value of each feature and for each race.

In the next step, the column BoEth which is the borrower's ethnicity has been removed from the dataset and predicted as the borrower's race. The importance of each column in the dataset for making predictions used by the xgboost model is illustrated in Figure 2.9 The results of the model are specified in Table 2.5. The classification output for this data is 75.4% accurate, which is much better than happening by chance. The accuracy of the dummy classifier is 0.3346. The results of the dummy classifier are mentioned in Table 2.4. The confusion matrix and similarity matrices are represented in the figures 2.5 and 2.6 respectively.

In the next step, the second important feature 2.10 co-borrower race is removed from

the dataset, and predictions are made.

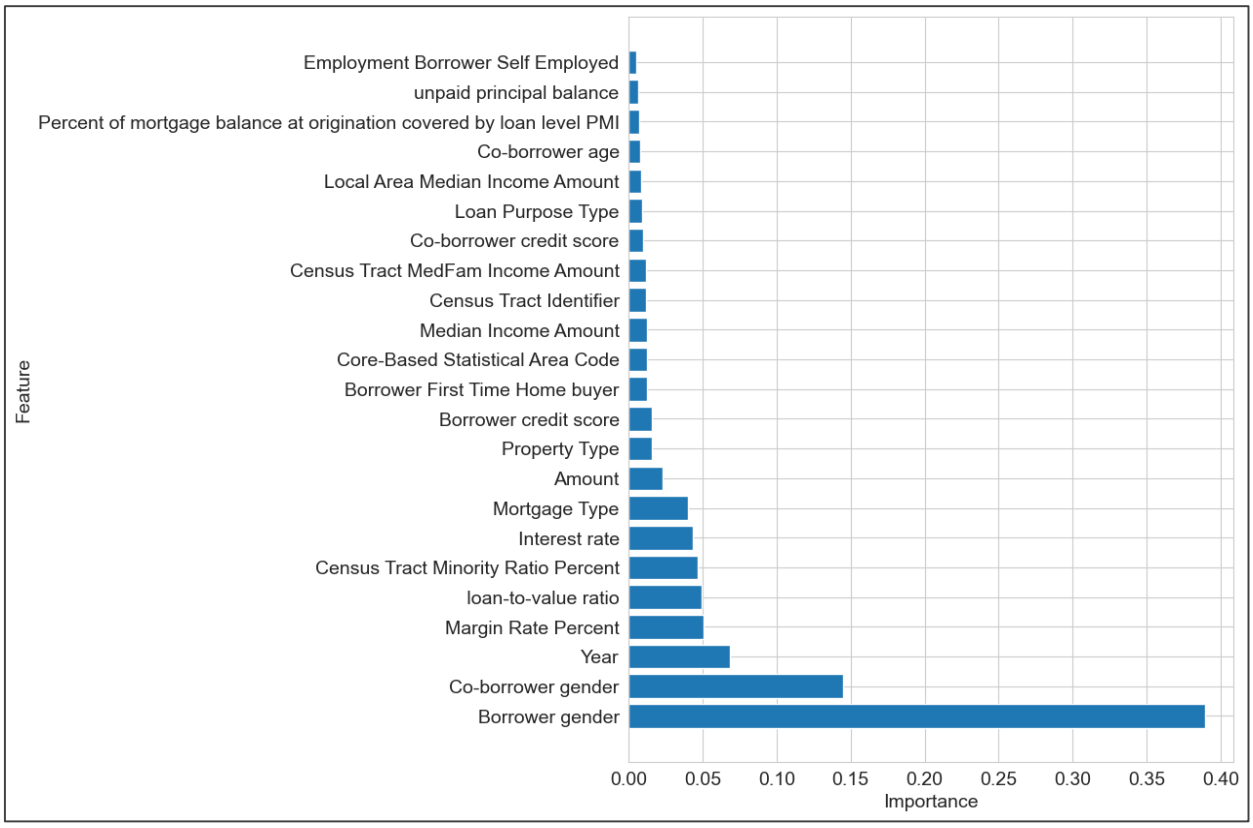


Figure 2.2: Feature importance for the prediction of the borrower's race as determined by XGBoost.

Borrower's Race	**American Indian or Alaska Native**	**Asian**	**Black or African American**	**Native Hawaiian or other Pacific Islander**	**White**
Year	2015	2016	2016.811	2015	2015
Core Based Statistical Area Code	**47492.260**	31314.700	**30960.800**	45268.950	40805.180

Census Tract Identifier	**3948.610**	**2264.340**	2280.930	3347.480	3188.070
Census Tract Minority Ratio Percent	24.950	33.080	**39.260**	33.200	**12.280**
Census Tract Median Family Income Amount	**65104.110**	**97845.920**	78782.250	75348.640	75096.090
Local Area Median Income Amount	**62312.710**	**75337.830**	72921.960	65114.250	63864.730
HUD Median Income Amount	**68144.880**	**80585.340**	78202.310	70883.090	69643.950
Unpaid principal balance	**169037.620**	**314914.350**	226968.130	249500.650	193836.550
Loan-to-value ratio	26.160	23.650	**31.200**	**16.900**	23.530
Loan Purpose Type	1.870	**1.800**	2.010	1.910	**2.050**
Mortgage Type	**0.580**	**0.050**	0.370	0.470	0.150
Borrower First-time Homebuyer	1.260	1.280	**1.160**	**1.440**	1.331

Borrower's Gender	**1.416**	1.270	**1.416**	1.276	**1.258**
Co-borrower's Gender	2.794	2.929	**3.175**	2.718	**2.679**
Interest rate	1.117	1.142	**1.410**	**0.824**	1.086
Amount	**169208.258**	**315350.082**	227278.459	249764.725	193976.470
Co-Borrower's Credit Score	6.269	6.764	**7.082**	**6.132**	6.153
PMI Coverage Percent	1.151	1.356	**2.685**	**0.645**	1.228
Employment Borrower Self-Employed	1.390	1.395	**1.257**	**1.560**	1.366
Property Type	**0.684**	**2.190**	1.950	1.116	0.776
Margin Rate Percent	71443.020	71398.882	**64453.290**	**79979.801**	71239.879
Borrower's Credit Score	**3.818**	**4.472**	3.875	4.176	4.440

Table 2.8: Mean value of each feature of the FHLB dataset.

XGBoost Classifier

Table 2.10 presents a ranked list of the top ten features utilized by the XGBoost classifier model for making predictions.

In descending order of importance, the first feature is BoEth, which represents the

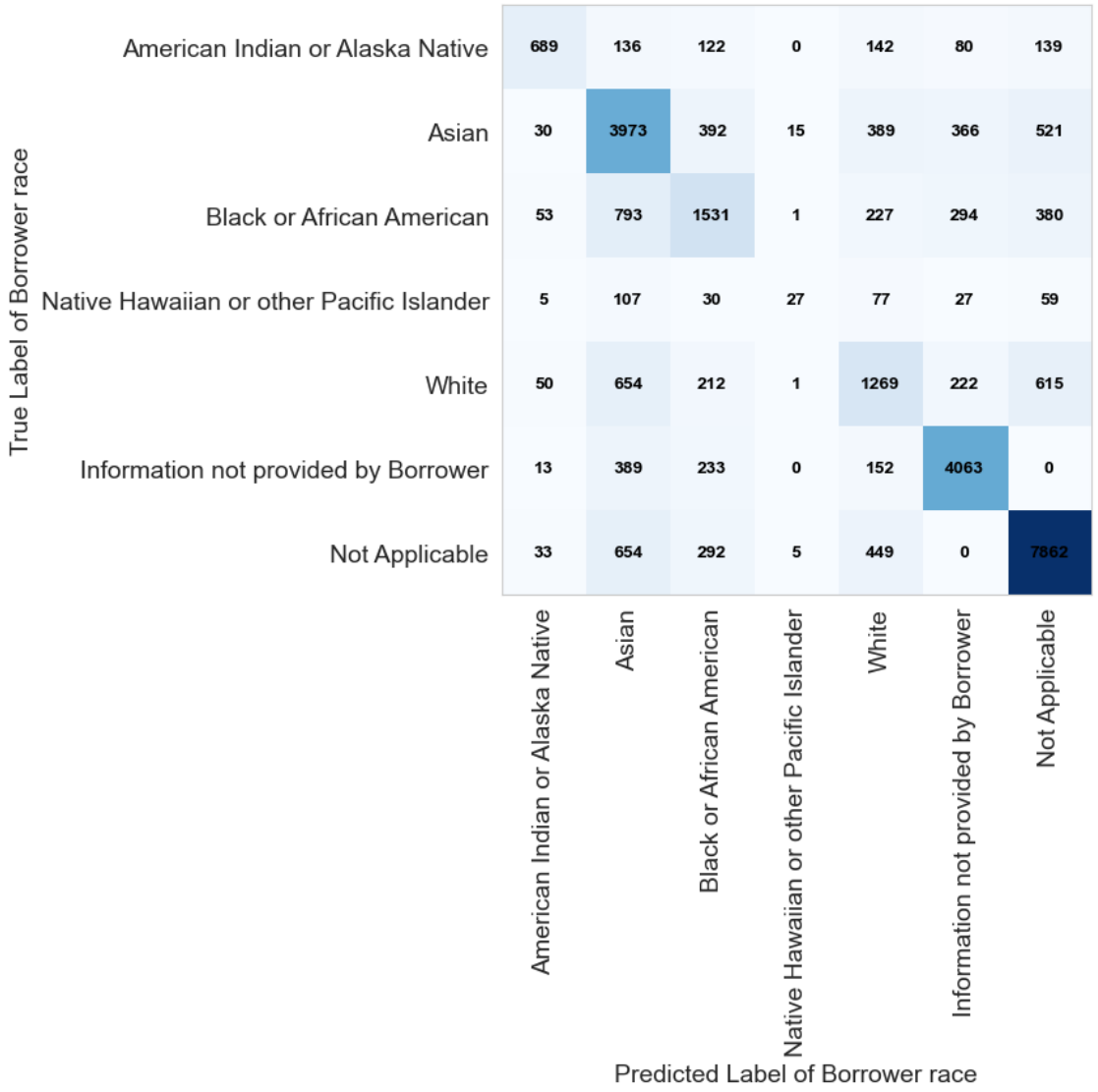


Figure 2.3: Confusion matrix for the prediction of the race based on the loan information. The labels in the figure represent the race of the borrower, which is specified in Table 2.2.

borrower's ethnicity. The second feature is CoRace, indicating the co-borrower's race. The third feature is Borrower's Gender, reflecting the borrower's gender. The fourth feature is the Co-borrower's Gender, denoting the co-borrower's gender.

The fifth feature is the Year, referring to the loan's year. The sixth feature is the Loan-to-value ratio, or the loan-to-value ratio, which measures the mortgage amount relative to the property value. The seventh feature is NumBor, representing the number of borrowers.

The eighth feature is Mortgage Type, indicating if the loan is federally guaranteed.

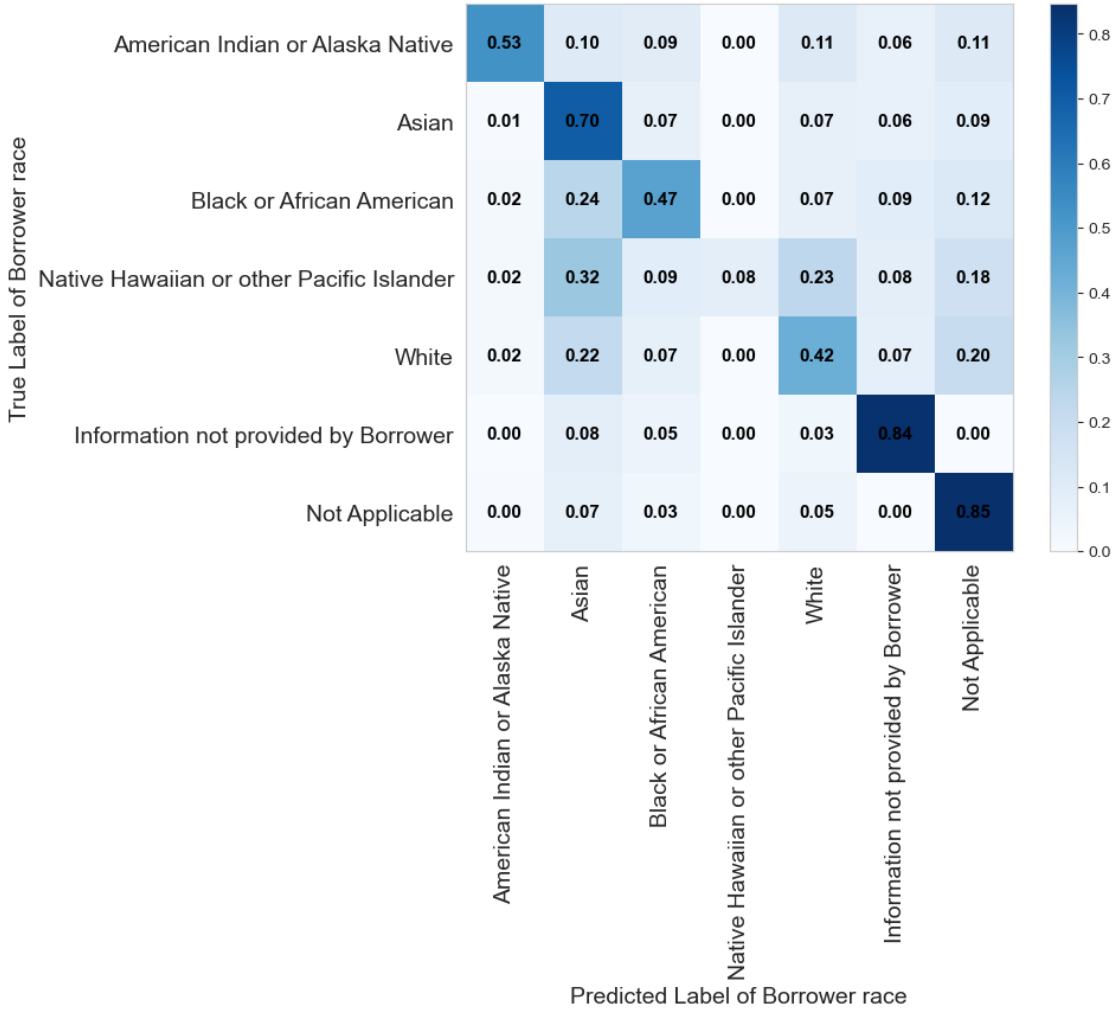


Figure 2.4: Similarity Matrix for classification of Borrower Race. The labels in the figure represent the race of the borrower, which is specified in Table 2.2

.

The ninth feature is the Census Tract Minority Ratio percentage, standing for the minority percentage in the area. The tenth feature is Interest rate, referring to the interest rate of the loan. These features have been identified as the most significant factors influencing the performance of the XGBoost classifier model when predicting the borrower race.

Chi-square-based k-best feature selection

Table 2.10 also shows the ranked list of the top ten features ascertained by the k-best feature selection method utilizing chi-square.

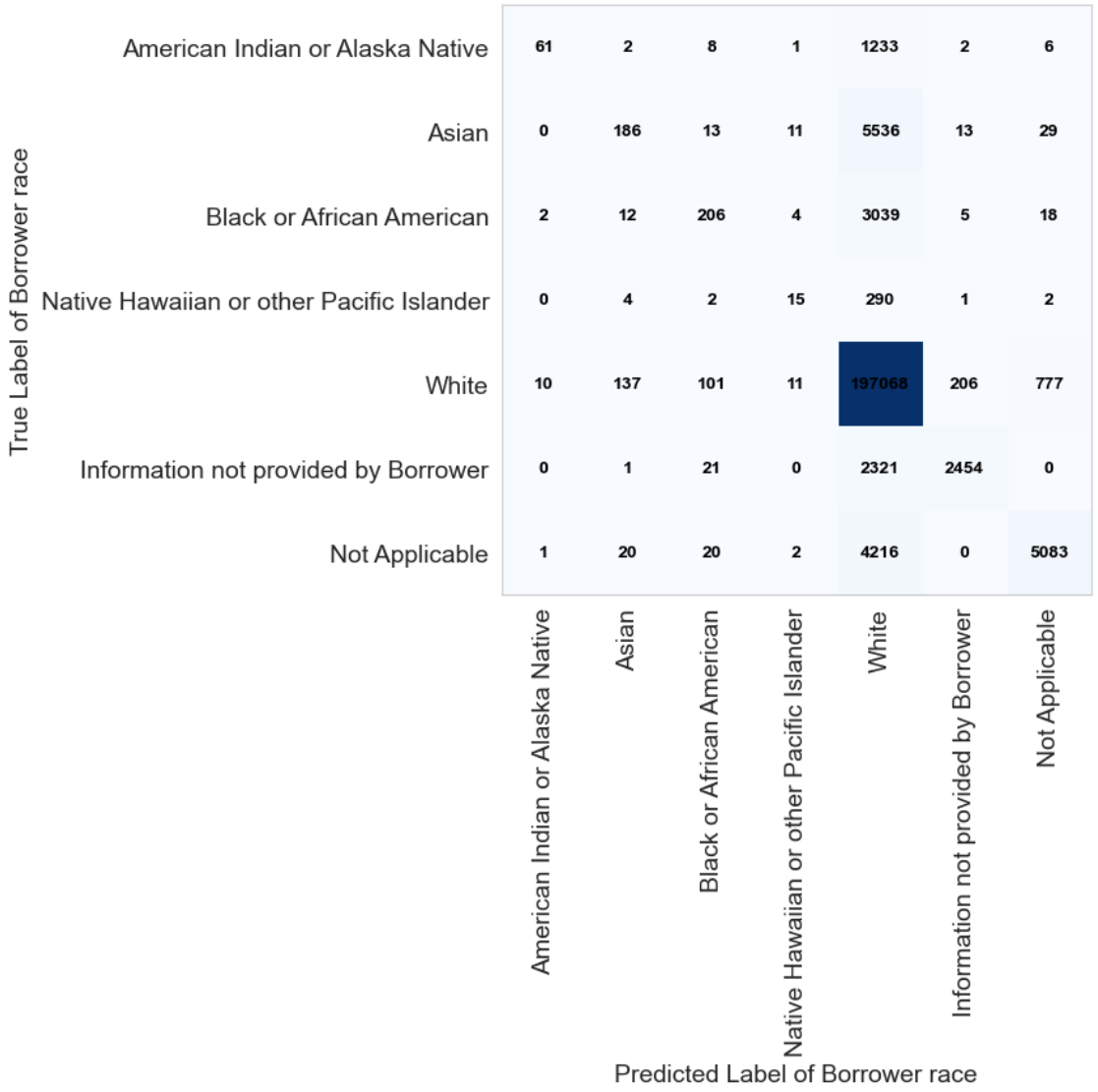


Figure 2.5: Confusion Matrix for classification of borrower race using top ten best features 2.2.

The attributes are listed in decreasing order of significance, include Census Tract Minority Ratio Percent, denoting the minority percentage in the area; Loan-to-value ratio, or the loan-to-value ratio, a measure of the mortgage amount in relation to the property value; and First, signifying if the borrower is a first-time home buyer. Borrower's Gender, representing the borrower's gender; CoAge, indicating the co-borrowers age; and Interest rate, pertaining to the loan's interest rate, are also on the list.

Other features include Employment Borrower Self Employed, identifying if the bor-

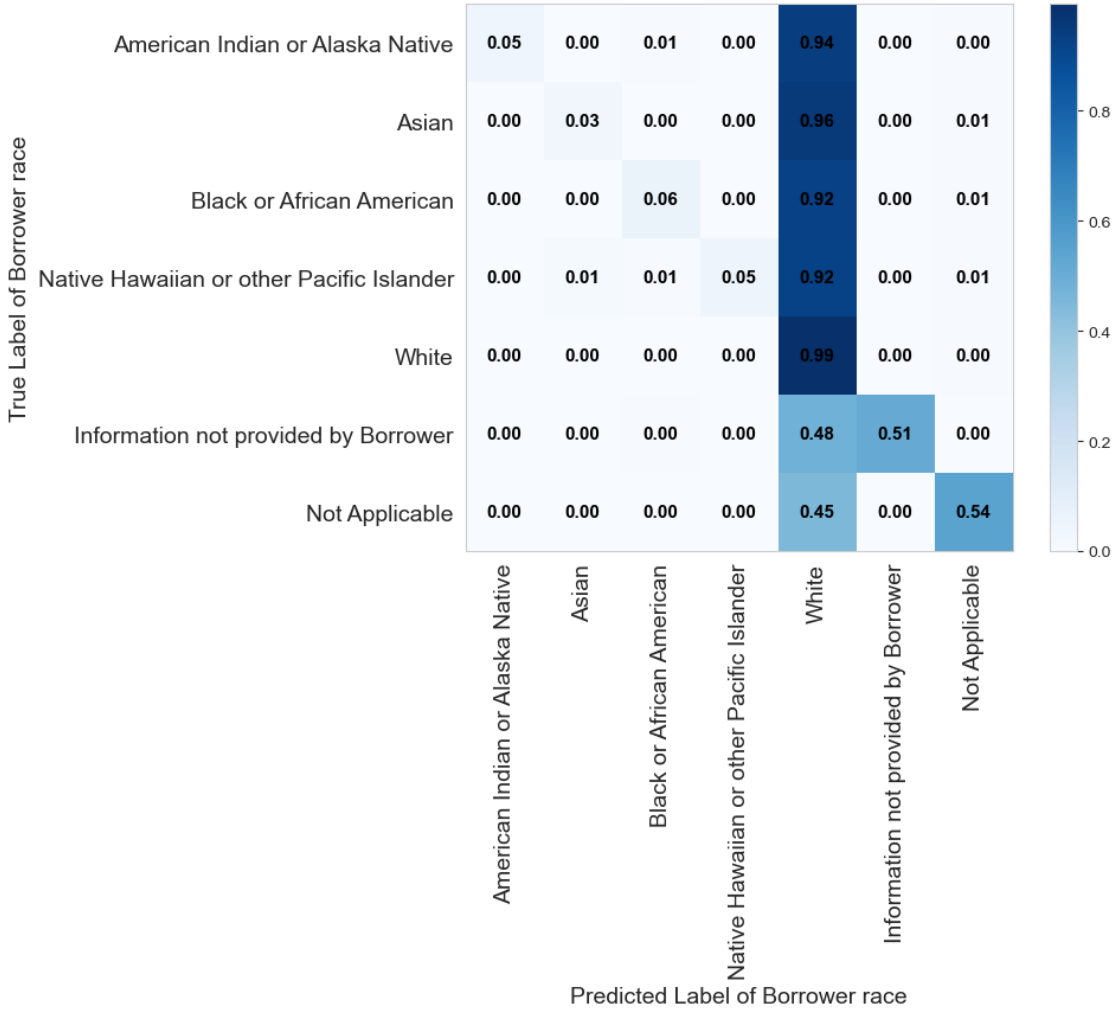


Figure 2.6: Similarity Matrix for classification of borrower race using top ten best features 2.2.

rower is self-employed; PropertyType, reflecting the property type; Margin Rate Percent, referring to the adjustable rate mortgage margin; and BoEth, alluding to the borrower's ethnicity. The k-best feature selection method employing chi-square has determined these features as the most impactful factors.

Looking at the best top ten features, we observed that BoEth, CoRace, Borrower's Gender, and CoAge are part of the top ten best features that influence the prediction of BoRace. In Table 2.9, we can see that when predicting BoRace using just the best top ten features, the XGBoost classifier outperforms the dummy classifier.

Task 1: BoRace Prediction		
	XGBoost	Dummy Classifier
Accuracy	0.918	0.888
Precision	0.901	0.789
Recall	0.918	0.888
F1 score	0.895	0.836

Table 2.9: The above table shows the model accuracy of predicting BoRace using the top ten features.

Feature Importance (Top ten)	
XGBoost	**K-Best (Chi-Square)**
Borrower Gender	Census Tract Minority Ratio Percent
Margin Rate Percent	Loan-to-value ratio
Loan-to-value ratio	Loan Purpose Type
Year	First-time homebuyer
Interest rate	Borrower Gender
Census Tract Minority Ratio Percent	Co-Borrower Age
First-time homebuyer	Interest rate
Property Type	Employment Borrower Self Employed
Co-Borrower Age	Property Type
Loan Purpose Type	Margin Rate Percent

Table 2.10: Top ten features influencing the prediction of BoRace.

2.5.2 Clustering

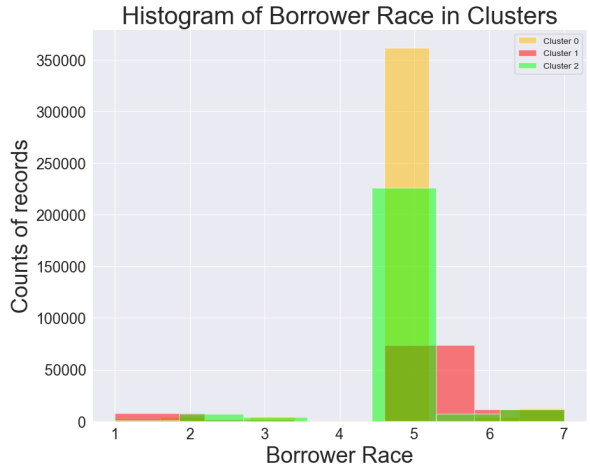
We employed clustering techniques to find patterns and groups comparable together based on their BoRace attribute. For this investigation, we used the KMeans and Kprototype clustering algorithms. To start with, we first normalized the data and performed clustering tasks with BoRace as the target variable, in a sequential manner.

We conducted experiments with different numbers of clusters and used the elbow method to determine the optimal number of clusters for the data. To perform the elbow method, we first applied a clustering algorithm to our dataset and calculated the within-cluster sum of squares (WCSS) for each number of clusters. We then plotted the WCSS values against the number of clusters and observed the resulting curve.

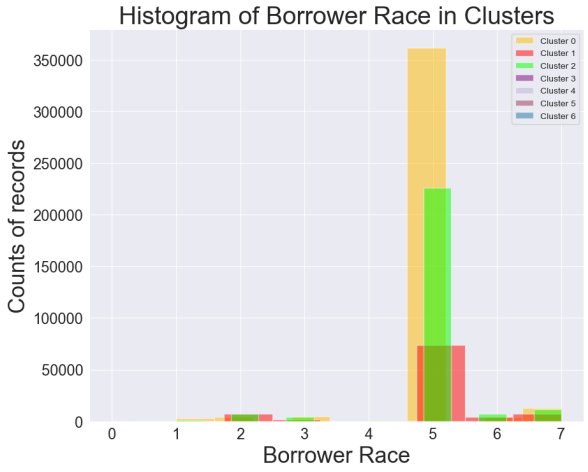
From Figure 2.9, we can observe that the curve initially decreased rapidly as we increased the number of clusters, indicating that adding more clusters led to a significant

reduction in WCSS. However, at a certain point, n=7, the curve began to flatten out, suggesting that adding additional clusters would not lead to a significant reduction in WCSS. An optimal number of clusters is important because it provides a good balance between capturing the underlying patterns in the data and avoiding over-fitting or under-fitting.

For the experiment, we started with three clusters and then increased the number of clusters to seven to see if there was any improvement in the quality of the clusters formed. Figure 2.7a shows the clusters formed using the Kprototype algorithm when taking BoRace as the target variable with the number of clusters as three, whereas Figure 2.7b shows the clusters formed with the number of clusters as seven.



(a) The figure represents the distribution of seven different races in three different clusters using the Kprototype algorithm. The borrower race description is mentioned in the Table



(b) The figure represents the distribution of seven different races in seven different clusters using the Kprototype algorithm. The borrower's race description is mentioned in Table 2.2

Figure 2.7: Clustering with 3 and seven clusters on borrower's race.

2.5.3 T-test

In order to calculate the two races' statistically significant differences, we used a T-test. It allows the determination of the level of support for a valid distinction between the compared variables. Based on race, the data set was split into two groups. Firstly, we grouped all the loans assigned to Black or African. The American race is in one group, and

all the other races are in another group. The interest rates between these two groups were compared using a two-sample independent T-test.

The statistical value of the T-test was 20.470. This figure suggests that the interest rates between the two categories are significantly different. In addition, the test's p-value, which was 4.181e-93, was very close to zero. The p-value indicates the likelihood that such severe results were just a coincidence. Given the extremely low p-value in this instance, it is clear that there is a significant difference in interest rates between black or African Americans and all other racial groups in the dataset, and the null hypothesis must be rejected.

A T-test on the 'Amount' variable (mortgage balance at origination) between the same two groups was also done in addition to the earlier findings on interest rates. With a T-statistic of 22.884 and a p-value of 7.384e-116, the analysis showed a statistically significant gap, showing a significant difference in loan amounts between blacks and all other racial groups. These findings demonstrate the existence of systemic imbalances and underline the necessity of addressing and redressing them in the lending practices of the housing sector. Continuing the previous findings, an additional analysis was conducted on the Asian racial group. Group one consisted of all the loan records borrowed by Asians, while Group 2 included individuals of all other races.

With a p-value of 0.0002 and a T-statistic of 3.674, the interest rate comparison T-test result showed a statistically significant difference. This implies that there is a significant difference in interest rates between Asians and all other racial groups.

The T-test also showed a significant discrepancy in loan amounts, with a t-statistic of 135.006 and a p-value of 0.0. This indicates that Asians receive loans at significantly higher rates than all other racial groupings.

These additional results prove discrepancies in lending procedures, especially for Asian borrowers. It underlines how critical it is to address these disparities and implement policies that give people of all races fairness and equal opportunity in the housing sector.

As we continue our examination of various racial groups, we now turn our attention

to the American Indian or Alaska Native class. All the loans borrowed by American Indians or Alaska Natives are grouped in Group One, whereas people of all other races are in Group Two.

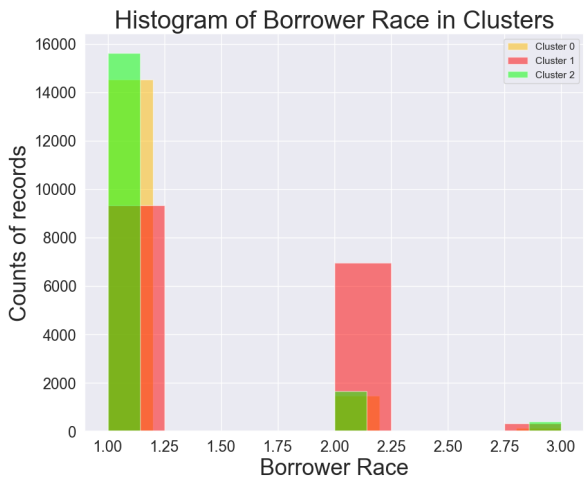
The T-test resulted in a statistic of 0.696 and a p-value of 0.487 for the interest rate comparison. These findings suggest that the interest rates for the American Indian or Alaska Native group and the other racial groups are not significantly different.

However, the T-test showed a significant discrepancy when looking at loan amounts. The American Indian or Alaska Native group received significantly more loans than all other racial groups, according to the T-statistic of -17.808 and the accompanying p-value of 6.352e-71.

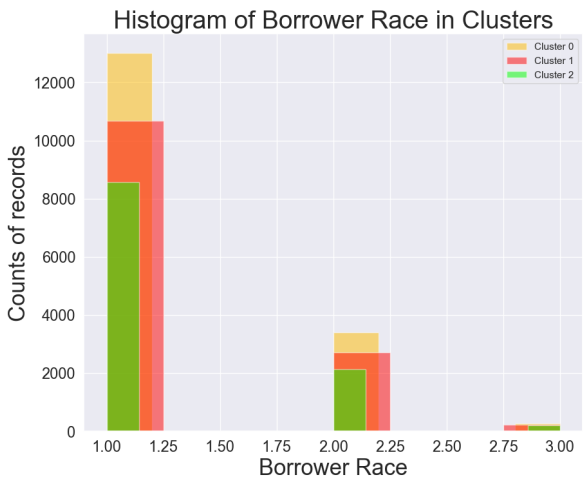
These findings highlight the fact that American Indian or Alaska Native populations experience considerable inequalities in loan amounts but not in interest rates. It highlights the importance of resolving these disparities and putting policies in place to support fair lending practices and equality in the housing sector for people from this racial background.

2.5.4 Unbalanced dataset

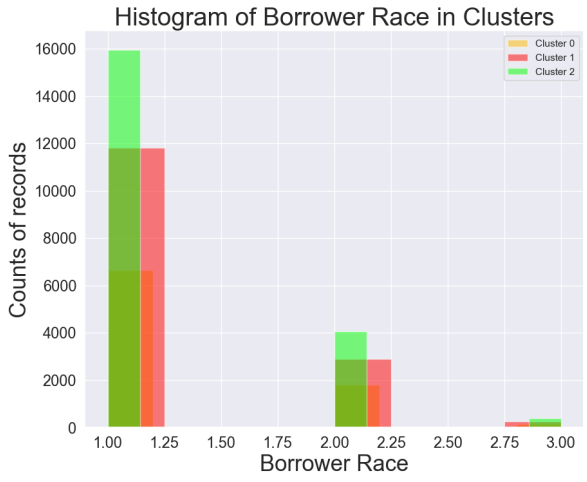
The federal home loan dataset as represented in Figure 2.11 is unbalanced. The whites are 88.90% of the dataset and loans assigned to American Indians or Alaska Natives are just 0.58% of the loans assigned to all the races. So, the dataset is under-sampled, which is implemented by taking the same number of records from the class with the lowest samples: American Indians or Alaska Natives. These samples are randomly chosen from the dataset. After this process, there are 7287 records. In the next step, a student T-test is performed on this data. Firstly, all the records of race black or African American are made into one group and whites into another group. The student T-test value for this is statistic = 2.876, p-value = 0.004. This indicates that not only are more loans assigned to whites, but there are also significant differences in the interest rates assigned to these groups. In the next step, all the races except whites are in one group and whites are in another group, and a



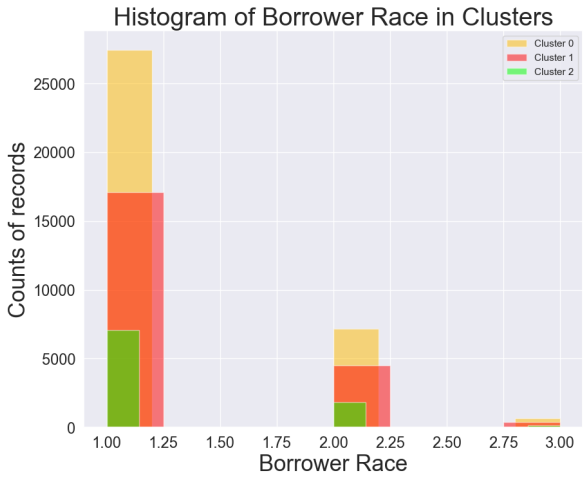
(a) 2009



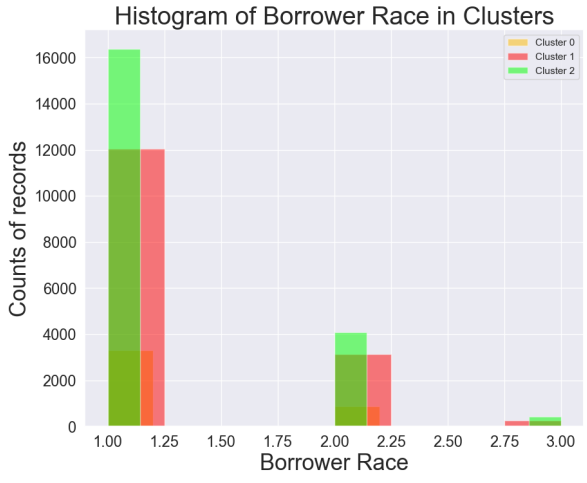
(b) 2010



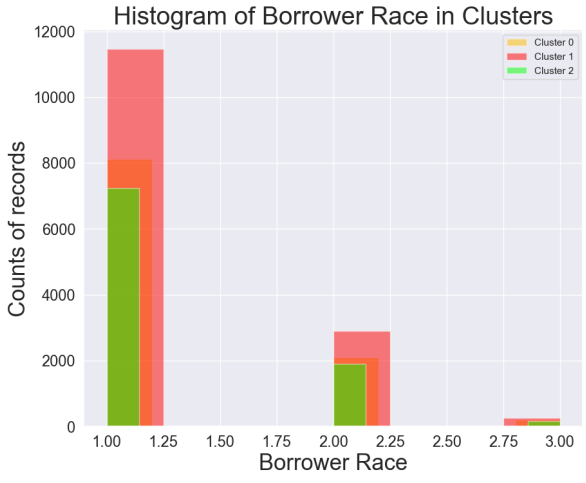
(c) 2011



(d) 2012

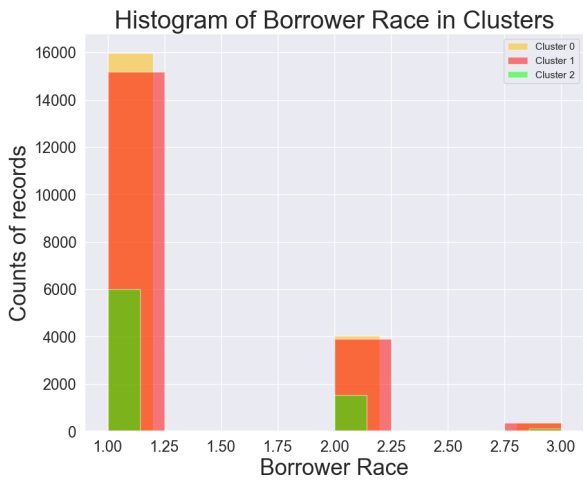


(e) 2013

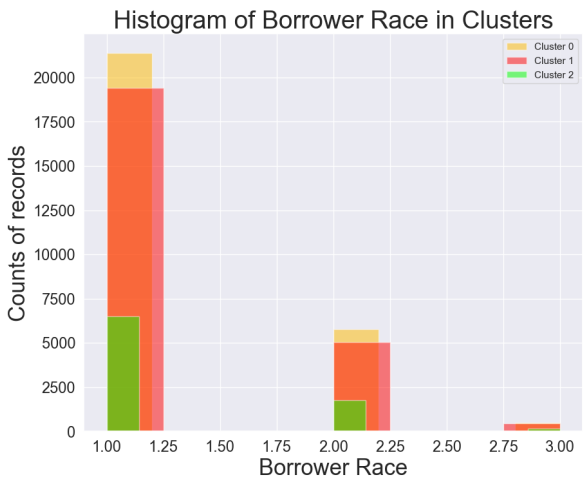


(f) 2014

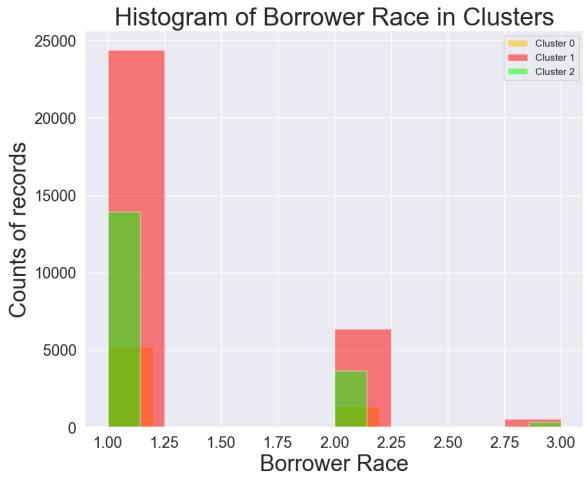
Figure 2.8: Year-wise Clustering with K = 3 (Part one).



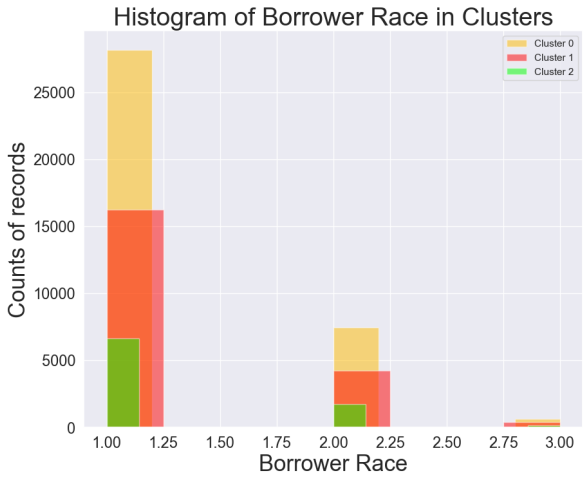
(g) 2015



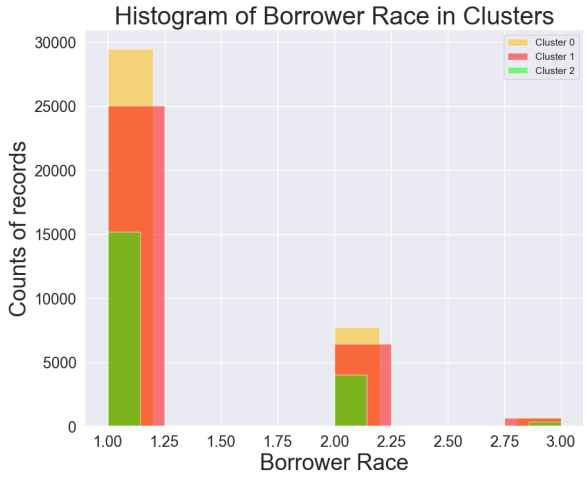
(h) 2016



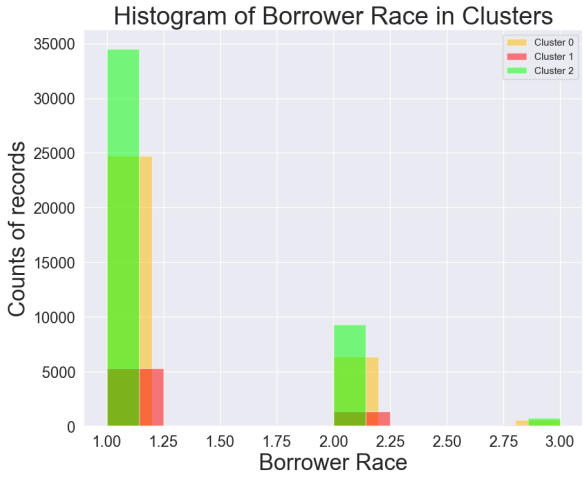
(i) 2017



(j) 2018

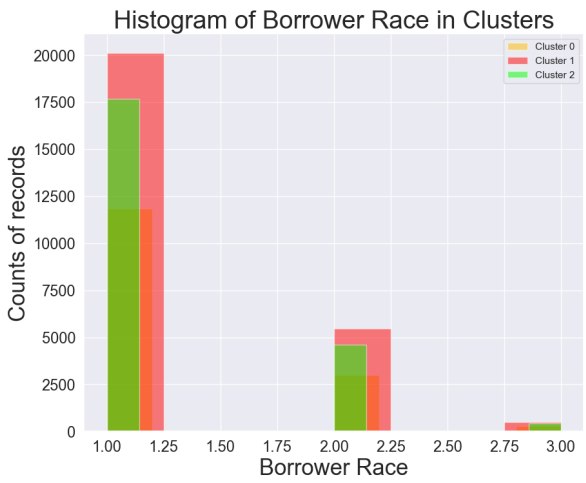


(k) 2019



(l) 2020

Figure 2.8: Year-wise Clustering with K = 3 (Part Two).



(m) 2021

Figure 2.8: Year-wise Clustering with K = 3 (Part Three).

Accuracy	0.6913
Precision	0.6986
Recall	0.6913
F1 score	0.6905

Table 2.11: Precision, recall, accuracy, and F1 score for the classification of borrower race using the XGBoost classification algorithm using the balanced dataset.

T-test is performed on these groups: statistic = -4.541, the p-value is 5.677e-06 for the rate of interest, and the statistic value is -11.50, the p-value is 2.305e-30 for the amount of loan assigned to each group. These findings suggest substantial disparities in both the interest rates and loan amounts allocated among racial groups, with whites generally receiving more favorable terms than other races.

In the next step, the prediction of the borrower race is made using the balanced dataset 2.11. Which is more than five times better than what happens by chance2.12. As borrower race, borrower ethnicity, and co-borrower ethnicity may have a direct influence on the borrower race, these columns are removed from the data, and a prediction of borrower race is made 2.13

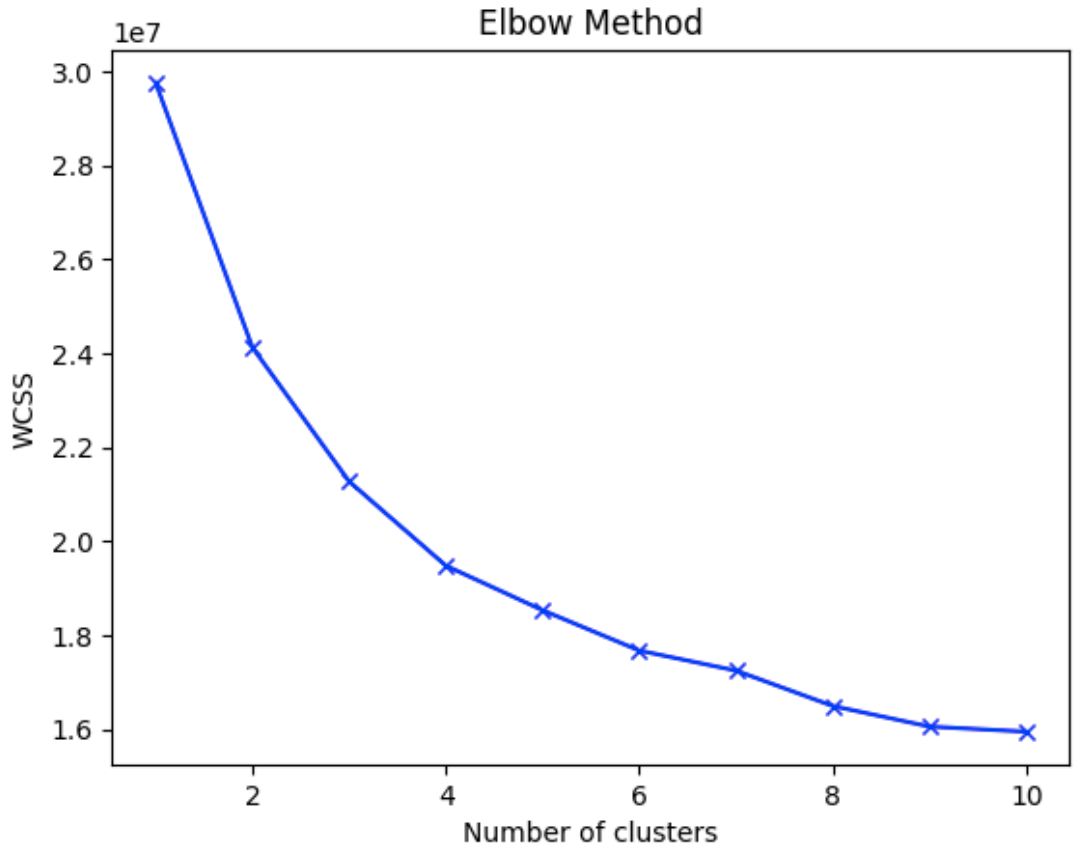


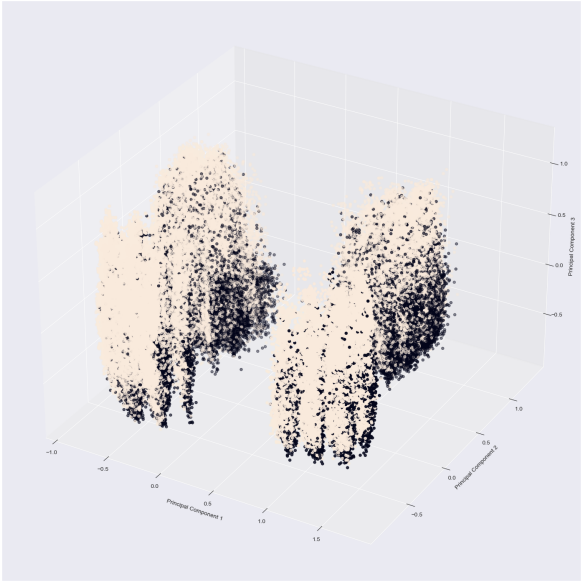
Figure 2.9: Determining the optimal number of Clusters using the elbow method.

Accuracy	0.1385
Precision	0.0191
Recall	0.1385
F1 score	0.0337

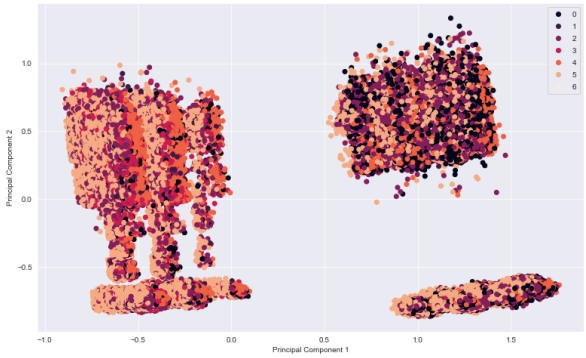
Table 2.12: Precision, recall, accuracy, and F1 score when predicting the borrower's race using a "dummy" ZeroR classifier using the balanced dataset.

Accuracy	0.5189
Precision	0.5366
Recall	0.5189
F1 score	0.5201

Table 2.13: Precision, recall, accuracy, and F1 score for the classification of borrower race after removing the co-borrower race and borrower ethnicity from the balanced data.



(a) 3D PCA visualization of Federal Home Loan data showing distinct clusters based on borrower's gender



(b) 2D PCA visualization of Federal Home Loan data showing distinct clusters based on borrower's gender

Figure 2.10: Principle Component Analysis based on borrower's race.

2.6 Conclusion

In conclusion, the study's findings prove that the Federal Home Loan Project contains biases, particularly when determining a borrower's race. It is evident that the data utilized in the research has inherent biases that influence the predictions, given the excellent accuracy rate, 90%, of the borrower's race.

The ability to correctly predict a borrower's race suggests that some patterns or factors correlate highly with particular racial groups. This raises serious questions regarding fairness and justice in the loan process since it implies that people of different racial backgrounds can experience unequal treatment or opportunity.

Furthermore, the T-test results showed that there were significant disparities in loan amounts between blacks and other racial groups, Asians and other racial groups, and American Indian or Alaska Native people and other racial groups. These results reveal racial biases in lending since various racial groups regularly receive different loan amounts.

These T-test results prove the claim that the Federal Home Loan Project has biases

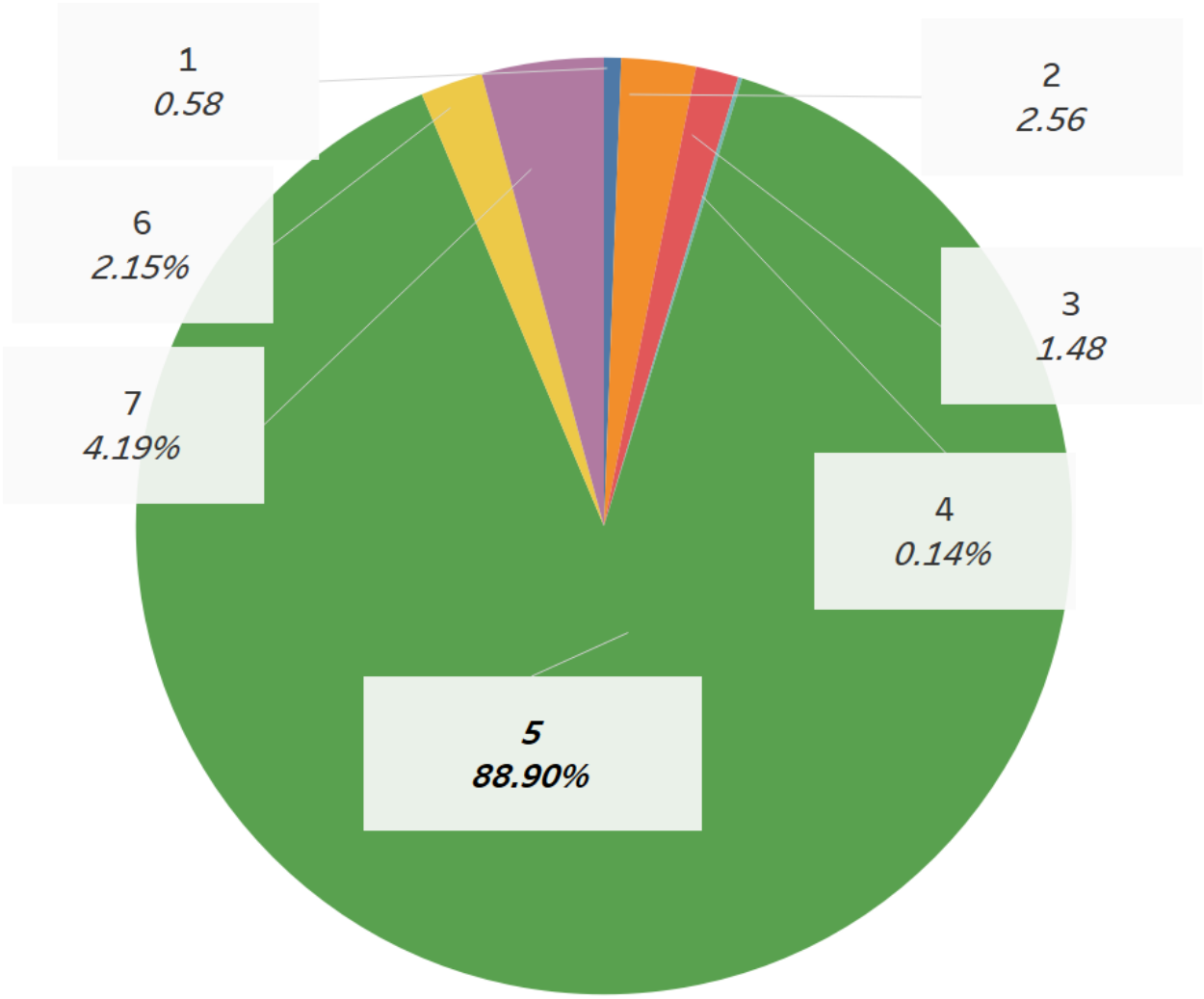
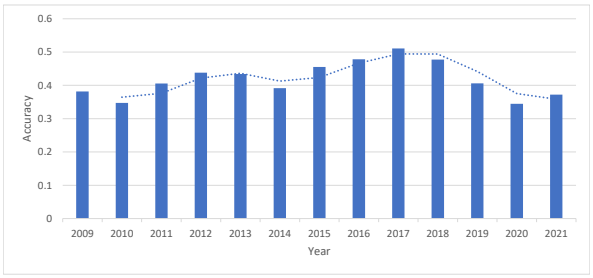
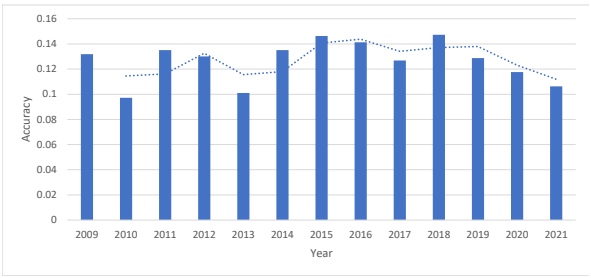


Figure 2.11: This graph represents the demographics of BoRace over 13 years and Table 2.2 gives more details of all races present in the data.



(a) XgBoost



(b) ZeroR

Figure 2.12: The prediction accuracy of borrower race over the years for sampled data

influencing judgments about granting loans based on a borrower's race. Such biases have significant ramifications for the loan sector and society at large. It highlights the urgent requirement for all-encompassing measures to address and eliminate these prejudices in order to guarantee a just and impartial lending system.

As machine learning algorithms are increasingly used in decision-making, it is crucial to thoroughly examine the data and models to identify any potential biases. Policymakers, regulators, and business stakeholders must work together in the future to develop effective initiatives that support equity, openness, and equitable access to economic opportunities. We may work towards a more equal and inclusive society where loan decisions are based on objective and fair criteria rather than maintaining systemic disadvantages by recognizing and correcting the biases found in this study.

Year	Accuracy	Precision	F1-Score	Recall
2009	0.5277	0.5305	0.5258	0.5277
2010	0.2361	0.2231	0.2261	0.2361
2011	0.3918	0.3918	0.4035	0.3918
2012	0.4520	0.4520	0.4463	0.452
2013	0.4343	0.4771	0.4491	0.4343
2014	0.4054	0.3974	0.3966	0.4054
2015	0.4471	0.5081	0.4646	0.4471
2016	0.5040	0.5158	0.5008	0.5040
2017	0.5121	0.5136	0.5254	0.5121
2018	0.4715	0.4936	0.4756	0.4715
2019	0.4878	0.5040	0.4942	0.4878
2020	0.3739	0.4016	0.3818	0.3739
2021	0.4634	0.4643	0.4574	0.4634

Table 2.14: The results of the XgBoost classifier on under-sampled data.

Year	Accuracy	Precision	F1-Score	Recall
2009	0.1319	0.0174	0.0307	0.1319
2010	0.09722	0.0094	0.0972	0.09722
2011	0.1351	0.0182	0.0321	0.1351
2012	0.1301	0.0169	0.0299	0.1301
2013	0.1010	0.01020	0.0185	0.1010
2014	0.1351	0.0182	0.0321	0.1351
2015	0.1463	0.0214	0.0373	0.1463
2016	0.1413	0.0199	0.0349	0.1413
2017	0.1269	0.0161	0.0285	0.1269
2018	0.1473	0.0217	0.0378	0.1473
2019	0.12871	0.0165	0.0293	0.12871
2020	0.11764	0.0138	0.0247	0.11764
2021	0.1063	0.0113	0.0204	0.1063

Table 2.15: The year-wise classification accuracy of the borrower race over the years ZeroR for under-sampled data.

Year	Accuracy	Precision	F1-Score	Recall
2009	0.9305	0.8659	0.8970	0.9305
2010	0.9230	0.8519	0.8860	0.9230
2011	0.9215	0.8491	0.8838	0.9215
2012	0.9263	0.8581	0.8909	0.9263
2013	0.9219	0.8499	0.8844	0.9219
2014	0.8965	0.8038	0.8476	0.8965
2015	0.8840	0.7816	0.8297	0.8840
2016	0.8689	0.8080	0.7550	0.8689
2017	0.8516	0.7252	0.7834	0.8516
2018	0.8325	0.6931	0.7565	0.8325
2019	0.8531	0.7277	0.7854	0.8531
2020	0.9055	0.8200	0.8606	0.9055
2021	0.8856	0.7844	0.8319	0.8856

Table 2.16: Year-wise ZeroR results for the entire data.

Year	Accuracy	Precision	F1-Score	Recall
2009	0.9458	0.9318	0.9301	0.9458
2010	0.9408	0.9252	0.9246	0.9408
2011	0.9367	0.9205	0.9180	0.9367
2012	0.9455	0.9354	0.9317	0.9455
2013	0.9433	0.9330	0.9292	0.9433
2014	0.9307	0.9131	0.9140	0.9307
2015	0.9258	0.9162	0.9086	0.9258
2016	0.9179	0.9053	0.9030	0.9179
2017	0.9028	0.8879	0.8818	0.9028
2018	0.8853	0.8695	0.8617	0.8853
2019	0.8991	0.8848	0.8718	0.8991
2020	0.9320	0.9203	0.9125	0.9320
2021	0.9182	0.9060	0.8965	0.9182

Table 2.17: Year-wise classification results of Borrower Race XgBoost.

Bibliography

[1] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[2] Kevin G. Sellner, Gregory J. Doucette, and Gary J. Kirkpatrick. Harmful algal blooms: causes, impacts and detection. *Journal of Industrial Microbiology and Biotechnology*, 30:383–406, 2003. doi: 10.1007/s10295-003-0074-9.

[3] Md Atiqul Islam. Cyanobacterial harmful algal bloom modeling in eutrophic water bodies. Master's thesis, Kansas State University, Manhattan, Kansas, 2020. Master of Science in Biological and Agricultural Engineering.

[4] Shuqi Lin, Donald C. Pierson, and Jorrit P. Mesman. Prediction of algal blooms via data-driven machine learning models: an evaluation using data from a well-monitored mesotrophic lake. *Geoscientific Model Development*, 16:35–46, 2023. doi: 10.5194/gmd-16-35-2023.

[5] Peixuan Yu, Rui Gao, Dezhen Zhang, and Zhi-Ping Liu. Predicting coastal algal blooms with environmental factors by machine learning methods. *Ecological Indicators*, 123: 107334, 2021.

[6] Onder Ozgur, Erdal Tanas Karagol, and Fatih Cemil Ozbugday. Machine learning approach to drivers of bank lending: evidence from an emerging economy. *Financial Innovation*, 7(20), 2021. doi: 10.1186/s40854-021-00237-1.

[7] Samuel L. Jr. Myers and Tsze Chan. Racial discrimination in housing markets: Accounting for credit risk. *Social Science Quarterly*, 76(3):543–561, 1995.

[8] Jiabao Wen, Jiachen Yang, Yang Li, and Liqing Gao. Harmful algal bloom warning based on machine learning in maritime site monitoring. *Knowledge-Based Systems*, 245: 108569, 2022.

[9] Harmful Algal Blooms in Kansas: 2019 Statistics. https://www.kdhe.ks.gov/DocumentCenter/View/6636/Marion-Reservoir---what-we-know-what-we-need-to-learn-PDF?bidId=. [Online; accessed 2023-08-14].

[10] Kansas lake, campgrounds close due to hazardous algae presence. https://www.ksnt.com/news/local-news/kansas-lake-campgrounds-close-due-to-hazardous-algae-presence/, jun 2 2022.

[11] Army Corps of Engineers: Marion Reservoir. https://salinapost.com/posts/a7064f8b-4183-44fa-a904-82a49bc3cf34, jun 3 2022.

[12] Marion Lake Page. https://www.swt-wc.usace.army.mil/MARI.lakepage.html.

[13] Ysi EXO3 Multiparameter Water Quality Sonde | ysi.com. https://www.ysi.com/exo3.

[14] Laura Krueger, Trisha Moore, Aleksey Sharupov, Lior Shamir, Daniel Flippo, and Kavya Kompella. Investigation marion reservoir cyanobacteria habs through integrated modeling and data approaches. In *Governor's Water Conference*, 2022.

[15] Craig D Newgard and Roger J Lewis. Missing data: how to best account for what is not known. *Journal of the American Medical Association*, 314(9):940–941, 2015.

[16] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2010.

[17] Mridul K Thomas, Simone Fontana, Marta Reyes, Michael Kehoe, and Francesco Pomati. The predictability of a lake phytoplankton community, over time-scales of hours to years. *Ecology Letters*, 21(5):619–628, 2018.

[18] Xgboost: A scalable tree boosting system. Chen, tianqi and guestrin, carlos. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[19] Rui Guo, Zhiqian Zhao, Tao Wang, Guangheng Liu, Jingyi Zhao, and Dianrong Gao. Degradation state recognition of piston pump based on iceemdan and xgboost. *Applied Sciences*, 10(18):6593, 2020.

[20] Yuanchao Wang, Zhichen Pan, Jianhua Zheng, Lei Qian, and Mingtao Li. A hybrid ensemble method for pulsar candidate classification. *Astrophysics and Space Science*, 364:1–13, 2019.

[21] Cross-validation: evaluating estimator performance. https://scikit-learn.org/stable/modules/cross_validation.html.

[22] Owasco lake coordinates. https://www.latlong.net/place/owasco-lake-ny-usa-21464.html, .

[23] Owasco lake. https://www.visitfingerlakes.com/plan-your-trip/finger-lakes-facts/owasco-lake/, .

[24] Owasco lake 2022. https://www.dec.ny.gov/data/IF/CSLAP/2022_CSLAPreport_Owasco%20Lake(0706OWA0212).html, .

[25] Finger Lakes Land Trust. Toxic algae facts for everyone to know. https://www.fllt.org/toxic-algae-facts, 2018.

[26] Robert D Dietz and Donald R Haurin. The social and private micro-level consequences of homeownership. *Journal of urban Economics*, 54(3):401–450, 2003.

[27] Karin Kurz. *Home ownership and social inequality in comparative perspective*. Stanford University Press, 2004.

[28] Dalton Conley and Brian Gifford. Home ownership, social insurance, and the welfare state. In *Sociological Forum*, volume 21, pages 55–82. Springer, 2006.

[29] Richard K Green and Susan M Wachter. The american mortgage in historical and international context. *Journal of Economic Perspectives*, 19(4):93–114, 2005.

[30] Susan Wharton Gates, Vanessa Gail Perry, and Peter M Zorn. Automated underwriting

in mortgage lending: Good news for the underserved? *Housing Policy Debate*, 13(2): 369–391, 2002.

[31] John Yinger. Discrimination in mortgage lending: A literature review. *Mortgage Lending, Racial Discrimination, and Federal Policy*, pages 29–74, 1996.

[32] Stephen L Ross and John Yinger. *The color of credit: Mortgage discrimination, research methodology, and fair-lending enforcement*. MIT press, 2002.

[33] John Goering and Ron Wienk. *Mortgage lending, racial discrimination and federal policy*. Routledge, 2018.

[34] Justin P Steil, Len Albright, Jacob S Rugh, and Douglas S Massey. The social structure of mortgage discrimination. *Housing Studies*, 33(5):759–776, 2018.

[35] Richard Williams, Reynold Nesiba, and Eileen Diaz McConnell. The changing face of inequality in home mortgage lending. *Social Problems*, 52(2):181–208, 2005.

[36] Eric Rosenblatt. A reconsideration of discrimination in mortgage underwriting with data from a national mortgage bank. *Journal of Financial Services Research*, 11(1-2): 109–131, 1997.

[37] Dara D Mendez, Vijaya K Hogan, and Jennifer Culhane. Institutional racism and pregnancy health: using home mortgage disclosure act data to develop an index for mortgage discrimination at the community level. *Public Health Reports*, 126(3_suppl): 102–114, 2011.

[38] Lincoln Quillian, John J Lee, and Brandon Honoré. Racial discrimination in the us housing and mortgage lending markets: a quantitative review of trends, 1976–2016. *Race and Social Problems*, 12:13–28, 2020.

[39] Michael Reibel. Geographic variation in mortgage discrimination: Evidence from los angeles. *Urban Geography*, 21(1):45–60, 2000.