PREDICTING HUMAN VULNERABILITY TO LANDSLIDE FACTORS: A MACHINE LEARNING APPROACH

Submitted in partial fulfilment of the requirements for the award of the degree of

MASTER OF TECHNOLOGY
in
CIVIL ENGINEERING

(With specialization in Geotechnical Engineering)

ASSAM ENGINEERING COLLEGE,
JALUKBARI, GUWAHATI-13, ASSAM

Submitted by:
SAGAR KALITA
M.TECH 3rd Semester
Roll No: PG/C/23/18
ASTU Roll No: 230620062015 of 2023-2025

Under the guidance of:

DR. ABINASH MAHANTA*(Guide)*
Associate Professor
Assam Engineering College
Department of Civil Engineering

DR. GUNAJIT KALITA *(Co-guide)*
Associate Professor and HOD
Assam Engineering College
Department of Computer Science and Engineering

CANDIDATE DECLARATION

I hereby declare that the work presented in this report entitled **PREDICTING HUMAN VULNERABILITY TO LANDSLIDE FACTORS: A MACHINE LEARNING APPROACH**, in the partial fulfilment of the requirement for the award of the degree of Master of Technology in Civil Engineering with specialization in Geotechnical Engineering submitted in the Department of Civil Engineering, Assam Engineering College, Jalukbari, Guwahati-13 under Assam Science and Technology University, is a work carried out in the said college under the supervision of Dr. Abinash Mahanta, Associate Professor, Department of Civil Engineering, Assam Engineering College, Jalukbari, Guwahati- 13, Assam. Whatever I have presented in this report has not been submitted by me for the award of any other degree or diploma.

Date:

Place: Guwahati

Sagar Kalita
M.Tech 3rd Semester
Department of Civil Engineering
Assam Engineering College
Guwahati-781013

CERTIFICATE OF SUPERVISION

This is to certify that the work presented in this report entitled **— PREDICTING HUMAN VULNERABILITY TO LANDSLIDE FACTORS: A MACHINE LEARNING APPROACH** is carried out by Sagar Kalita, Roll No: PG/C/23/18, a student of M.Tech 3rd semester, Department of Civil Engineering, Assam Engineering College, under my guidance and supervision and submitted in the partial fulfilment of the requirement for the award of the Degree of Master of Technology in Civil Engineering with specialization in Geotechnical Engineering under Assam Science and Technology University.

Date:

Place: Guwahati

Dr. Abinash Mahanta
Associate Professor
Department of Civil Engineering
Assam Engineering College
Guwahati-781013

CERTIFICATE OF SUPERVISION

This is to certify that the work presented in this report entitled **— PREDICTING HUMAN VULNERABILITY TO LANDSLIDE FACTORS: A MACHINE LEARNING APPROACH** is carried out by Sagar Kalita, Roll No: PG/C/23/18, a student of M.Tech 3rd semester, Department of Civil Engineering, Assam Engineering College, under my guidance and supervision and submitted in the partial fulfilment of the requirement for the award of the Degree of Master of Technology in Civil Engineering with specialization in Geotechnical Engineering under Assam Science and Technology University.

Date:

Place: Guwahati

.

Dr. Gunajit Kalita
Associate Professor and HOD
Department of Computer Science and Engineering
Assam Engineering College
Guwahati-781013

CERTIFICATE FROM THE HEAD OF THE DEPARTMENT

This is to certify that the following student of M.Tech 3rd semester of Civil Engineering Department (Geotechnical Engineering), Assam Engineering College, has submitted his project on **PREDICTING HUMAN VULNERABILITY TO LANDSLIDE FACTORS: A MACHINE LEARNING APPROACH** in partial fulfilment of the requirement for the award of the Degree of Master of Technology in Civil Engineering with specialization in Geotechnical Engineering under Assam Science and Technology University.

Name: SAGAR KALITA
College Roll No: PG/C/23/18
ASTU Roll No: 230620062015
ASTU Registration No: 313609119 of 2023-2025

Date:

Place: Guwahati

Dr. Jayanta Pathak
Professor and HOD
Department of Civil Engineering
Assam Engineering College
Guwahati-781013

ACKNOWLEDGEMENT

I would like to express my sincere appreciation to my guide Dr. Abinash Mahanta, Associate Professor, Department of Civil Engineering of Assam Engineering College and co-guide Dr. Gunajit Kalita, Associate Professor and HOD, Department of Computer Science and Engineering of Assam Engineering College for extensive support and encouragement throughout the project work. I am highly indebted in their guidance and constant supervision as well as for providing necessary information regarding the project work. Working under them has indeed been a great experience and inspiration for me. My gratitude towards the entire fraternity of the Department of Civil Engineering and Department of Computer Science and Engineering of Assam Engineering College. I cannot help myself without thanking Assam Engineering College, which provided us the required infrastructure and comforts all throughout this course and my project in particular.

ABSTRACT

Landslides represent a significant natural hazard, particularly in regions with vulnerable populations. As humans are a critical resource, their exposure to such disasters can create considerable gaps in resource availability and disrupt communities. While landslide susceptibility maps have been effective in identifying high-risk areas, these maps typically lack a detailed, quantitative assessment of how many people are directly vulnerable to such hazards. To address this gap, this study focuses on analysing and predicting the vulnerability of populations in landslide-prone areas using various regression models.

The primary goal of the research was to quantify the risk and estimate the percentage of people exposed to landslide threats. This was achieved through the application of five key regression models: Linear Regression**,** Lasso Regression**,** Ridge Regression**,** Polynomial Regression**,** and Random Forest Regression. These models were applied to a comprehensive dataset, incorporating demographic data and landslide occurrences. The dataset was split into a **70-30 ratio**, with 70% of the data allocated for training and 30% for testing to ensure robust evaluation of the models' predictive capabilities.

The findings suggest that, based on the models' predictions, approximately 78% of the population residing in landslide-susceptible areas may be classified as vulnerable. These percentages reflect the populations at risk but do not directly correlate to mortality rates. Instead, they highlight the susceptibility of these regions to landslide impacts, offering important insights for decision-makers and disaster response teams.

Moreover, the results provide valuable guidance for the development of early warning systems and disaster management strategies, helping authorities to prioritize regions with higher population vulnerabilities. While the models demonstrate strong predictive abilities, there is potential to further improve accuracy by incorporating additional data sources or more advanced algorithms, such as Artificial Neural Networks (ANNs), which have been shown to improve performance in similar studies. Ultimately, this study contributes to enhancing landslide risk assessment and ensuring better disaster preparedness in at-risk regions.

These findings lay the groundwork for more effective disaster mitigation policies and reinforce the importance of continuous data collection and model refinement in addressing landslide vulnerabilities.

List of Figures

List of Table

Table of Contents

CHAPTER 1

1 INTRODUCTION

1.1 Overview

Landslides are geological phenomena characterized by the downward movement of rock, soil, and debris along a slope. They are natural hazards that occur when the stability of a slope is compromised, leading to the displacement of materials. Landslides can vary in scale, from small, localized events to large, catastrophic occurrences that can cause significant damage to the environment, infrastructure, and communities. Landslides represent a significant geological hazard in India, affecting diverse landscapes from the Himalayan region to the Western Ghats. This susceptibility is underscored by the country's unique geological setting and climatic conditions. Several studies have delved into understanding the causes, characteristics, and implications of landslides in India, contributing to the body of knowledge on this complex natural phenomenon.

Figure 1: Mudslides creating havoc in Dima Hasao district of Assam

Source: https://www.bloomberg.com/news/articles/2022-05-17/heavy-rains-trigger-floods-in-northeast-india-killing-8

Figure 2:A mudslide from the hills entirely demolished a house in Boragaon,Guwahati. The house was engulfed in flames, trapping four people inside.

Source: https://www.indiatodayne.in/assam/story/assam-landslides-guwahati-atleast-4-dead-388098-2022-06-14

Studies by researchers such as Gupta et al. (2018), have emphasized the link between tectonic activity and slope instability in these areas. The monsoon season plays a pivotal role in triggering landslides across the country. The study conducted by Singh and Patel (2019), investigated the relationship between rainfall patterns and landslide occurrences, particularly in the Western Ghats and Northeastern states. Anthropogenic activities have increasingly contributed to landslide risks. The work of Sharma et al. (2020), highlighted the impact of deforestation and improper land use planning on landslide occurrences in hilly terrains. Historical landslide events have been extensively documented in literature, shedding light on their consequences and the need for effective mitigation strategies. The analysis by Reddy and Kumar (2015), of the Kedarnath disaster in 2013 provides valuable insights into the complex interplay of geological factors during such catastrophic events.

Landslides pose a significant geohazard in Northeast India, where the combination of complex geological structures, high rainfall, and hilly terrains contributes to the susceptibility of the region. The city of Guwahati, being a prominent urban centre in the Northeast, is particularly vulnerable to landslide events. Understanding the causes, patterns, and mitigation strategies specific to this region is crucial for sustainable development and risk reduction. The Northeastern region of India is characterized by intricate geological formations, with tectonic activity playing a significant role. Studies by researchers such as Sen et al. (2017), have highlighted the geological complexities in the region and their influence on slope stability. The monsoon season, with its heavy and prolonged rainfall, exacerbates landslide risks in Northeast India. The work of Baruah and Das (2019), investigated the monsoonal influences on landslide occurrences, emphasizing the need for a thorough understanding of precipitation patterns. Guwahati, as a rapidly growing urban centre in the region, faces unique challenges concerning landslides. The work of Chakraborty and Goswami (2017), shed light to the curious and vast scope of ML and ANN through their works in accessing the prediction of slope stability using multiple linear regression and artificial neural network.

1.2 Problem Statement

Landslides are one of the most devastating natural hazards, causing significant loss of life, damage to infrastructure, and disruption of ecosystems. Despite the complex and multifactorial nature of landslides, predicting their occurrence and assessing their susceptibility remains a critical challenge for geotechnical engineers, researchers, and policymakers. In recent decades, substantial research has been conducted to improve the understanding and prediction of landslides, leveraging various traditional techniques such as empirical models, statistical methods, and numerical simulations. These methods, while effective to some extent, often fall short in predicting landslides accurately due to the inherent complexity and dynamic nature of the underlying factors such as soil composition, rainfall intensity, seismic activity, and human interventions. Given the increasing frequency and severity of landslides worldwide due to climate change and urbanization in vulnerable regions, there is an urgent need to develop more reliable and robust predictive models. Traditional methods, though valuable, often rely on static data and simplistic assumptions that limit their ability to handle complex and non-linear relationships between variables. Furthermore, many of these models are based on site-specific studies and are difficult to generalize across different geographic and environmental conditions.

In recent years, machine learning (ML) and artificial intelligence (AI) have emerged as promising tools in the prediction and susceptibility analysis of landslides. The ability of AI and ML algorithms to process vast amounts of data, recognize patterns, and continuously improve prediction accuracy makes them ideally suited for tackling the inherent uncertainty and complexity of landslide phenomena. Machine learning techniques such as decision trees, support vector machines, random forests, and neural networks have shown significant potential in capturing non-linear dependencies between the numerous variables that contribute to landslides. Additionally, these models can incorporate real-time data, such as rainfall patterns and seismic activity, to improve early warning systems and reduce false predictions.

While there is no doubt that substantial research has already been conducted in the area of landslide prediction, the incorporation of machine learning and AI represents a future direction where predictions could be more "rock-solid" and reliable. The growing body of evidence supporting the efficacy of ML and AI models demonstrates that these methods can outperform traditional approaches in both accuracy and generalizability. However, the application of AI and ML to landslide prediction is still in its nascent stages, and much work remains to be done in terms of refining algorithms, improving data quality, and integrating diverse datasets across regions.

Encouraging further research in this field is essential to unlocking the full potential of AI and ML in landslide prediction. This includes investing in the development of more sophisticated models, improving access to high-quality geospatial and environmental data, and fostering collaborations between geotechnical engineers, data scientists, and policymakers. By doing so, we can advance towards a future where landslide predictions are not only more accurate but also more actionable, helping to mitigate the risks associated with these natural disasters and protect vulnerable communities around the world.

1.3 Purpose of the study

The devastating impact of landslides on infrastructure and human safety has made their accurate prediction a critical challenge in the field of geotechnical engineering. Traditional methods, such as limit equilibrium analysis and finite element modelling, often struggle to capture the complex and non-linear relationships between the various factors contributing to slope instability. In recent years, the advent of machine learning techniques has offered new promising avenues for enhancing landslide prediction capabilities. To address this critical issue, researchers have leveraged various regression models to predict the likelihood and impact of landslides (Reddy et al., 2020) (Irawan et al., 2021) (Pradhan & Kim, 2020) (Sofwan et al., 2019).

The purpose of this research was to compare the performance of five different regression models in predicting landslide occurrences: Lasso regression, Ridge regression, Polynomial regression, Linear regression, and Random Forest regression. The study considered a range of factors known to influence landslide events, including landslide size, category, and setting (Pradhan & Kim, 2020) (Sofwan et al., 2019) (Reddy et al., 2020) (Irawan et al., 2021).

The Lasso regression model, a form of linear regression that applies L1 regularization, was utilized to identify the most influential predictors and simplify the model . The Ridge regression model, which employs L2 regularization, was employed to address multicollinearity among the predictors . Polynomial regression, an extension of linear regression, was explored to capture non-linear relationships between the predictors and the landslide outcome .

In contrast, the Linear regression model assumed a linear relationship between the independent variables and the landslide outcome. The Random Forest regression model, a non-parametric ensemble learning method, was also implemented to leverage its ability to capture complex, non-linear relationships .

To assess the accuracy of these models a comparison was done with their respective performance metrics, such as R-squared , Mean Squared Error and Root Mean Squared Error. The results were satisfactory, provided with the limited dataset collected from mammoth research works and further to increase the accuracy of the models ANN model will be suitable for this further study.

CHAPTER 2

2 LITERATURE REVIEW

2.1 Introduction

The literature work carried out by the researchers related to the field of the present study is in the section. Each of the literature is briefly described with its own outcome to support the undertaking of the present topic of interest.

2.2 Definition

Varnes (1978), proposed a seminal classification system, categorizing landslides into falls, slides, flows, and topples. This classification laid the groundwork for a systematic approach to understanding landslide processes.

Hutchinson (1988), emphasized the significance of slope movement, introducing key factors like shear strength and stress conditions as essential in the geological definition of landslides.

Glade et al.,(2000), expanded the definition to include societal activities, land-use changes, and climate influences, reflecting a comprehensive understanding of the interactions shaping landslide occurrences.

Chakraborty and Goswami (2017), emphasized the significance of using multiple linear regression(MLR) and artificial neural network(ANN) and the results were compared it with traditional methods like Fellenius method , Bishop's method , Janbu's method and Morgenstern and Price method.

Reddy et al.,(2020), emphasized the significance of rainfall in inducing landslides using machine learning models and its importance.

2.3 Classification

Landslides are generally classified based on their movement, the type of material involved, and the specific triggering factors. Here is a general classification

2.3.1 Based on movement

2.3.1.1 Rockslides:

Involving the sliding or falling of individual rock fragments.

2.3.1.2 Rockfalls:

Sudden, free-fall movement of individual rock blocks.

2.3.1.3 Debris Flows:

Rapid downslope movement of a mixture of soil, rock, water, and organic material.

2.3.1.4 Mudslides:

Movement of fine-grained, wet soil or earth material.

2.3.1.5 Lahars:

Specifically volcanic mudflows, often triggered by volcanic activity.

2.3.2 Based on material

2.3.2.1 Rock Landslides:

Involving primarily rock material.

2.3.2.2 Earth Landslides:

Involving soil and other unconsolidated materials.

2.3.2.3 Debris Landslides:

Comprising a mixture of rocks, soil, and other materials.

2.3.3 Based on Triggering Factors

2.3.3.1 Rainfall-Triggered Landslides:

Caused by excessive rainfall, leading to saturation of soil.

2.3.3.2 Earthquake-Induced Landslides:

Triggered by seismic activity, often due to ground shaking.

2.3.3.3 Human-Induced Landslides:

Resulting from human activities like excavation, construction, or deforestation.

2.3.3.4 Volcanic Landslides:

Associated with volcanic eruptions, including pyroclastic flows and lahars.

2.3.4 Based on size

2.3.4.1 Small Landslide Size:

The volume of the landslide is less than 100 m^3

2.3.4.2 Medium Landslide Size:

The volume of the landslide is in between 100-500 m^3

2.3.4.3 Large Landslide Size:

The volume of the landslide is in between 1000-5000 m^3

2.3.4.4 Very Large Landslide Size:

The volume of the landslide is more than 5000 m^3

2.4 Study Approach of Landslide

Hungr et al. (2014), Hungr and co-authors provided a comprehensive update on the Varnes classification of landslide types, presenting an essential framework for understanding and categorizing landslides. The Varnes classification system offers a systematic approach that considers the type and rate of movement, providing a basis for landslide hazard assessment. This classification has been widely accepted and utilized by researchers, geologists, and practitioners globally, serving as a fundamental tool for characterizing landslide events based on their distinctive features.

Montgomery et al. (2003), Montgomery and team contributed to the study approach by investigating rainfall-induced landslides. Their research emphasized the role of antecedent soil moisture conditions as a critical factor influencing landslide susceptibility. This approach enhances our understanding of the hydrological aspects of landslides, particularly the relationship between rainfall patterns and slope stability.

Sidle et al. (2017), Sidle and colleagues contributed to the study approach by investigating the impacts of deforestation on landslide occurrence. Their research highlighted the importance of land-use practices in influencing slope stability, emphasizing the need for sustainable land management to mitigate landslide risk.

Kirschbaum et al. (2015), provided an integrated framework considering both precipitation-induced and earthquake-triggered landslides. This approach acknowledges the diverse factors

initiating slope failures. Their work highlights the importance of understanding the triggering mechanisms for effective landslide hazard assessment.

Crozier (2010), work delved into the impact of climate change on landslide activity. Changes in precipitation patterns and intensities associated with climate change were identified as potential triggers for increased landslide occurrences. This study underscores the importance of considering long-term climatic trends.

Guzzetti et al. (2008), conducted a comprehensive analysis of landslide-triggering rainfall events. Their study identified critical rainfall thresholds for different regions, emphasizing the importance of rainfall intensity, duration, and cumulative rainfall as triggering factors.

Gariano and Guzzetti (2016), extended the research on rainfall-triggered landslides by proposing an early warning model. Their work incorporates real-time rainfall data to assess the potential for landslide occurrence, contributing to proactive risk management.

Caine (1980), focused on seismic triggers for landslides. The study highlighted the influence of ground shaking, acceleration, and slope angle on earthquake-induced slope failures. Understanding the seismic parameters involved is crucial for assessing landslide susceptibility in seismic-prone regions.

Crosta and Frattini (2003), investigated the role of human activities in landslide initiation. Their study emphasized the impact of excavation, deforestation, and urbanization on slope stability, highlighting the need for sustainable land-use practices to mitigate landslide risk.

2.5 Study Approach of Landslide using Machine Learning

Sakellariou and Ferentinou (2005), studied on the idea of prediction analysis and used artificial neural network (ANN) to develop a rela- tionship between the various slope parameters.

Kayesa (2006), used the Geomos slope mon- itoring system (GSMS) to study the slope stability prediction of Letlhakane mine. The GSMS is basically an automatic and continuous slope monitoring system which runs continuously for 24 h. The system consists of three parts, viz, collection of data, transmission of data, and processing and analysis of data. The GSMS resulted into avoiding potentially fatal injury, damage to mining equipment, and loss of mining production.

Ahangar-Asr et al. (2010), developed a prediction model based on evolutionary polynomial regression (EPR) technique to predict the FOS. The EPR models are developed from the results of field data for conditions not used in the model build- ing process, and the results were found to be very effective in modeling the behavior of slopes.

Erzin and Cetin (2012) used ANN and MLR for finding the critical value of FOS for a typical artificial slope which is subjected to earthquake forces. The predicted results from both the methods were compared with the calculated results and found that the results obtained from ANN are having a higher degree of precision when compared to MLR .
Firmansyah et al. (2016) studied with different soil types to predict the run-out distance of a rota- tional slope using the concept of center of mass approach. They found that the soil unit

weight can influence to a great extent the depth of sliding zone and the volume of unstable material.

Elith et al., (2008), stated that machine learning techniques, a powerful group of data driven tools, use algorithms to learn the relationship between a landslide occurrence and landslide related predictors, and avoids starting with an assumed structural model.

Romer and Ferentinou (2016) stated that to obtain more reliable results through the statistical methods, large amounts of data are required, whereas ML-based models can effectively overcome the limitation of data dependent bivariate and multivariate statistical methods.

2.6 About Python

Python, a versatile and powerful programming language, has come a long way since its inception in the late 1980s. Developed by Guido van Rossum, Python was designed to be a high-level, general-purpose language that prioritizes simplicity, readability, and ease of use. (Kelly, 2016) (Rossum, 1999) Its flexible syntax and robust standard library have made it a popular choice for a wide range of applications, from web development and data analysis to artificial intelligence and machine learning.

One of the key advantages of Python is its accessibility to beginners. Its intuitive syntax and English-like commands make it a natural choice for those new to programming, allowing them to quickly grasp the fundamental concepts of computer science without being bogged down by low-level details. As a result, Python has become a widely adopted language in academic settings, where it is often used as the primary teaching tool for introductory programming courses.

Python's versatility has also contributed to its widespread adoption in the professional world. Its rich ecosystem of libraries and frameworks, such as Keras, TensorFlow, and Pandas, have made it a popular choice for data-driven applications, including data analysis, visualization, and machine learning.

Moreover, Python's flexibility and rapid development cycle have made it a valuable tool for prototyping and rapid application development.

2.7 Regression Models

Regression analysis is a fundamental statistical tool used to model the relationship between a dependent variable and one or more independent variables. In this study, it explored several regression models like Linear Regression**,** Lasso Regression**,** Ridge Regression**,** Polynomial Regression and Random Forest Regression to assess their applicability, strengths, and limitations in various predictive tasks of landslide.

2.7.1 Linear Regression

Draper and N.R. (1998), explained that linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable to be predicted is called the dependent variable. The variable using to predict the other variable's value is called the independent variable. This form of analysis estimates the coefficients of the linear equation,

involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a "least squares" method to discover the best-fit line for a set of paired data then estimate the value of X (dependent variable) from Y (independent variable).

$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon} \qquad (\mathbf{1})$$

Where:

- Y is the predicted output (dependent variable).
- β_0 is the intercept.
- $\beta_1,\beta_2,\ldots,\beta_n$ are the coefficients of the independent variables $X_1,X_2,\ldots,X_n$
- ϵ is the error term.

2.7.2 Lasso Regression

Tibsirani and Robert (1996), explained that lasso regression (Least Absolute Shrinkage and Selection Operator) is a variant of linear regression that incorporates L1 regularization. It adds a penalty to the absolute value of the regression coefficients, encouraging sparsity by reducing some coefficients to zero. This allows Lasso to perform feature selection, making it particularly useful when dealing with datasets containing many irrelevant features. However, Lasso may struggle when multicollinearity is high among predictors, as it can arbitrarily shrink one predictor over another.

$$\frac{min}{\beta}\left(\frac{1}{2N}\sum_{i=1}^{\mathrm{N}}(y_i - \beta_0 - \sum_{j=1}^{n}\beta_j \mathrm{X}_{ij})^2 + \lambda\sum_{j=1}^{n}|\beta_j|\right) \qquad (2)$$

Where:

- λ is the regularization parameter controlling the strength of the penalty.
- The $|\beta_j|$ term applies L1 regularization, shrinking some coefficients to zero.

2.7.3 Ridge Regression

Hoerl et al., (1970), stated that ridge regression is similar to Lasso but uses L2 regularization, which penalizes the sum of the squares of the coefficients. Unlike Lasso, Ridge does not force coefficients to become zero, making it more suited to datasets with multicollinearity issues. By introducing a penalty on large coefficients, Ridge reduces model complexity and prevents overfitting . It performs well predictors are relevant and have small coefficients.

$$\frac{min}{\beta}\left(\frac{1}{2N}\sum_{i=1}^{\mathrm{N}}(y_i - \beta_0 - \sum_{j=1}^{n}\beta_j \mathrm{X}_{ij})^2 + \lambda\sum_{j=1}^{n}{\beta_j}^2\right) \qquad (3)$$

Where:

- λ is the regularization parameter.
- The β_j^2 term applies L2 regularization, shrinking all coefficients but not necessarily setting any of them to zero.

2.7.4 Polynomial Regression

Kleinbaum et al., (1988), stated that polynomial regression extends linear regression by introducing non-linearity to the model. It does so by fitting polynomial terms of the independent variables (e.g., squared or cubic terms) to better capture the complex relationships between variables. However, polynomial regression is sensitive to overfitting, especially when higher-degree polynomials are used .

$$\boldsymbol{Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_n X^n + \epsilon} \qquad (\mathbf{4})$$

Where:

- n is the degree of the polynomial.
- The model includes polynomial terms X^2,X^n to capture non-linear relationships between X and Y.

2.7.5 Random Forest Regression

Breiman and Leo (2001), stated that random forest regression is a non-parametric ensemble learning method that combines the predictions of multiple decision trees to improve accuracy and stability. It reduces variance by averaging the predictions from several trees, thus preventing overfitting. Random Forest performs well with large, complex datasets and can handle both categorical and numerical features. However, it may suffer from interpretability issues due to the complexity of the model.

$$(\hat{Y}) = \frac{1}{T}\sum_{t=1}^{T} f_t(X) \qquad (5)$$

Where:

- T is the total number of decision trees in the forest.
- $f_t(X)$ is the prediction made by the t-th decision tree.
- The final prediction $\hat{Y}$ is the average of predictions from all trees.

2.7.6 Mean Squared Error (MSE)

Montgomery et al., (2021), stated that Mean Squared Error (MSE) is a common measure used to evaluate the performance of regression models. It calculates the average of the squared differences between the actual values y_i and the predicted values $\hat{y}_i$. A lower MSE indicates a model that makes more accurate predictions.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^{\,2} \tag{6}$$

2.7.7 Root Mean Squared Error (RMSE)

Montgomery et al., (2021), stated that Root Mean Squared Error (RMSE) is the square root of the MSE and provides a measure of the average error in the same units as the output variable. RMSE is widely used for regression model evaluation because it gives an intuitive measure of prediction error.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^{\,2}} \tag{7}$$

2.7.8 R-squared(R^2)

Montgomery et al., (2021), stated that R-squared (R^2) measures the proportion of the variance in the dependent variable (y) that is predictable from the independent variables. It ranges from 0 to 1, where higher values indicate a better fit of the model to the data. An R^2 of 1 implies perfect prediction, while an R^2 of 0 suggests the model does not explain any variance.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^{\,2}}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^{\,2}} \tag{8}$$

CHAPTER 3

3 RESEARCH METHODOLOGY

In this research, various databases such as the European Landslide Database, Global Fatal Landslide Database, NSF Design Safe Storage, NASA Global Landslide Catalog, and the Disaster Information Management System, we identified that the NASA Global Landslide Catalog contained the most relevant parameters for our study as it had 14533 data from the year 2008-2021. The decision to focus on this specific dataset was driven by its availability and the practicality of gathering data. Collecting landslide data on a yearly basis is a complex and labour-intensive process due to the diverse geographical locations, causes, and types of landslides. Hence, from 2008-2021 dataset offered a comprehensive and manageable source of information for this analysis.

source_name	Date	event_title	location_accuracy	landslide_catego	landslide_trigger	landslide_siz	landslide_set	fatality_count
AGU	01/08/08 00:00	Sigou Village, Loufan County, Shanxi Province	unknown	landslide	rain	large	engineered_sl	11
Oregonian	02/01/09 02:00	Lake Oswego, Oregon	5km	mudslide	downpour	small	above_road	0
CBS News	19/01/07 00:00	San Ramon district, 195 miles northeast of the capital, Lima,	10km	landslide	downpour	large	above_road	10
Reuters	31/07/09 00:00	Dailekh district	unknown	landslide	monsoon	medium	natural_slope	1
The Freeman	16/10/10 12:00	sitio Bakilid in barangay Lahug	5km	landslide	tropical_cyclone	medium	above_river	0
BusinessWorld Online	16/02/12 00:00	Paguite, Abuyog, Leyte	5km	landslide	downpour	medium	above_coast	0
The Spokesman-Review	30/03/12 00:00	Pend Oreille County, State Route 20 near Usk, OR	5km	mudslide	downpour	small	urban	0
Crónica Diaria	02/09/07 00:00	3 killed in Acapulco	10km	complex	tropical_cyclone	medium	above_road	3
MagicValley.com	05/09/07 00:00	Warm Springs Road, Idaho	5km	mudslide	rain	medium	unknown	
UPI	01/11/08 00:00	Lincang City, Yunnan, Yunnan-Tibet No. 214 highway.	10km	complex	downpour	medium	above_road	2
BBC News	01/11/08 00:00	Kunming, Yunnan	25km	complex	downpour	medium	above_road	4
Salem-News.com	01/01/09 22:24	South side of Highway 26 between Cherryville Hill and Brightwood. Ore	10km	mudslide	rain	small	natural_slope	0
ABS-CBN News	02/01/09 20:30	Gadgaron village of Matnog town, Sorsogon	5km	landslide	downpour	medium	above_road	0

Figure 3:Excel spreadsheet showing a glimpse of some of the data from overall 14533 data

The dataset was meticulously curated from various reliable sources, accounting for a total of 14,533 data . The key aspects of landslide events, categorized into multiple sub-categories, including:

- Landslide Category (LCA): The type of landslide (e.g., mudslide, rockfall, debris flow, etc.).
- Landslide Trigger(LTR): The event or condition that initiated the landslide (e.g., rainfall, earthquake, human activity, etc.).
- Landslide Size(LSI): The magnitude or scale of the landslide.
- Landslide Setting(LSE): The geographical or infrastructural context of the landslide (e.g., natural slopes, engineered slopes).
- Fatality Count(FCO): The number of fatalities resulting from the landslide.
- An extra parameter LAC(other factors) was taken into account while performing Polynomial and Random Forest Regression models

Given the large size of the dataset, a 70-30 split was implemented, where 70% of the data was designated for the training dataset, and the remaining 30% was reserved for the testing dataset. This approach ensures that the models developed were trained on the majority of the data while leaving a significant portion for evaluating their performance on unseen data.

Since the dataset contained a variety of categorical variables, an essential pre-processing step was the conversion of categorical data to numerical form. This transformation was critical for enabling the models to interpret and process the data. After this conversion, the dataset was analysed using five prominent regression models, implemented in Python. The following models were employed:

1. Linear Regression
2. Lasso Regression
3. Ridge Regression
4. Polynomial Regression
5. Random Forest Regression

The objective of this analysis was to examine the relationship between the independent variables and the dependent variable . To achieve this, various model outputs were generated and analysed:

- Training Dataset: Comparison of actual vs. predicted values for the training data.
- Testing Dataset: Comparison of actual vs. predicted values for the testing data.
- Residual Plot: Visualization of the difference between predicted and actual values to assess the accuracy of the model.
- Feature Importance: Identification of the most influential features in predicting landslides. Feature importance was assessed based on the absolute coefficient values for each model.
- Residual Distribution: Analysis of how residuals (errors) are distributed, which helps assess model performance and identify any bias in the predictions.
- Model Performance Metrics: Metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared, and other relevant performance measures were used to evaluate and compare the accuracy and efficiency of each model.

Finally, a comparative analysis was conducted on the accuracies of all five regression models. The purpose of this comparison was to determine which model offered the best predictive performance for landslide-related data, considering both the training and testing phases. This approach not only helped in identifying the most effective model for predicting landslides but also provided valuable insights into the importance of different variables influencing landslide occurrences and their consequences.

CHAPTER 4

4 RESULTS AND DISCUSSION

4.1 Model Analysis

4.1.1 Linear Regression

Draper and N.R. (1998), explained that linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable to be predicted is called the dependent variable. The variable using to predict the other variable's value is called the independent variable. This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a "least squares" method to discover the best-fit line for a set of paired data then estimate the value of X (dependent variable) from Y (independent variable).

$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon}$$

Where:

- Y is the predicted output (dependent variable).
- β_0 is the intercept.
- $\beta_1,\beta_2,\ldots,\beta_n$ are the coefficients of the independent variables $X_1,X_2,\ldots,X_n$
- ϵ is the error term.

Advantages of Linear Regression Model are:

- Interpretability: One of the primary advantages of linear regression is its simplicity and interpretability. Each coefficient β_i represents the expected change in the dependent variable for a one-unit change in the corresponding independent variable X_i holding all other variables constant.
- Efficiency: Linear regression can be efficiently computed for both small and large datasets, making it a highly scalable model.
- Widespread Use: It is commonly used in various domains, such as economics, biology, engineering, and social sciences, to analyse relationships between variables.

Limitations of Linear Regression Model are:

- Linearity Assumption: The primary limitation of linear regression is that it assumes a linear relationship between the dependent and independent variables. In cases where the true relationship is non-linear, linear regression may perform poorly.

- Outliers: Linear regression is sensitive to outliers, which can significantly affect the model's performance.
- Multicollinearity: If the independent variables are highly correlated (multicollinear), it becomes difficult to determine the individual effect of each variable on the dependent variable. This can inflate the variance of the coefficient estimates and lead to misleading conclusions.

The model equation formed is:

```
Coefficients (bias and slope): [-5.50674472  3.30576187]
```

Figure 4: Slope and Bias formed for Linear Regression Model

And the following results obtained are as follows:

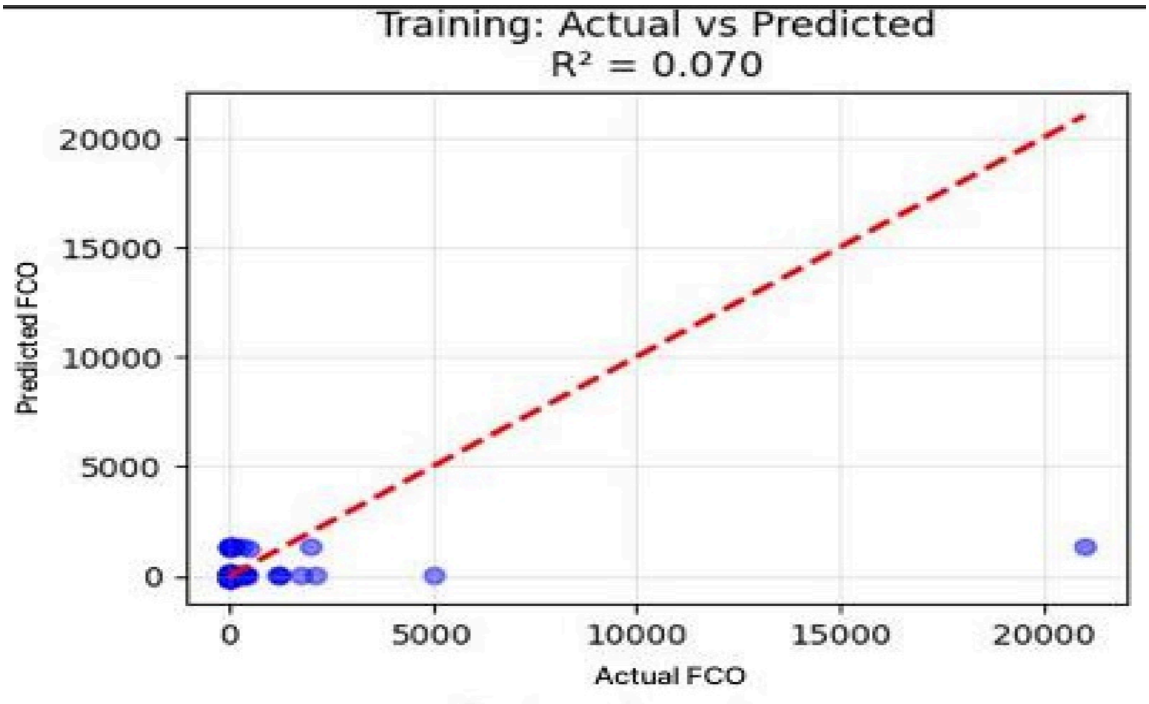

Figure 5:Training Plot of Actual vs Predicted FCOs

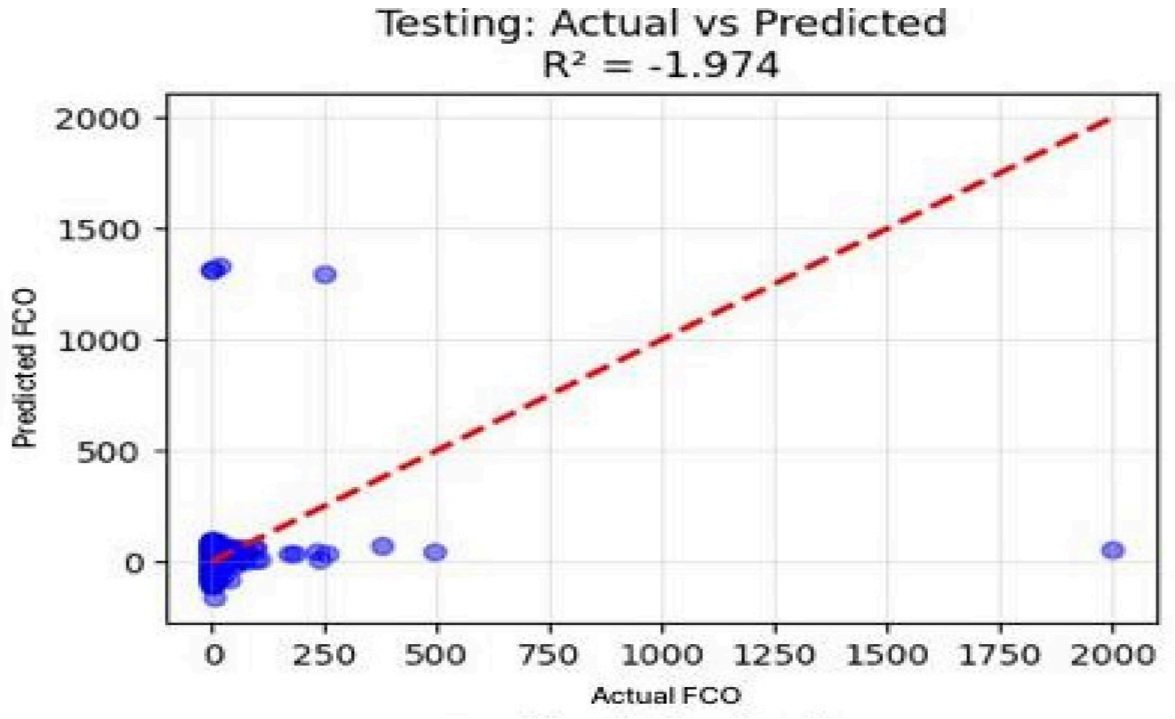

Figure 6:Testing Plot of Actual vs Predicted FCOs

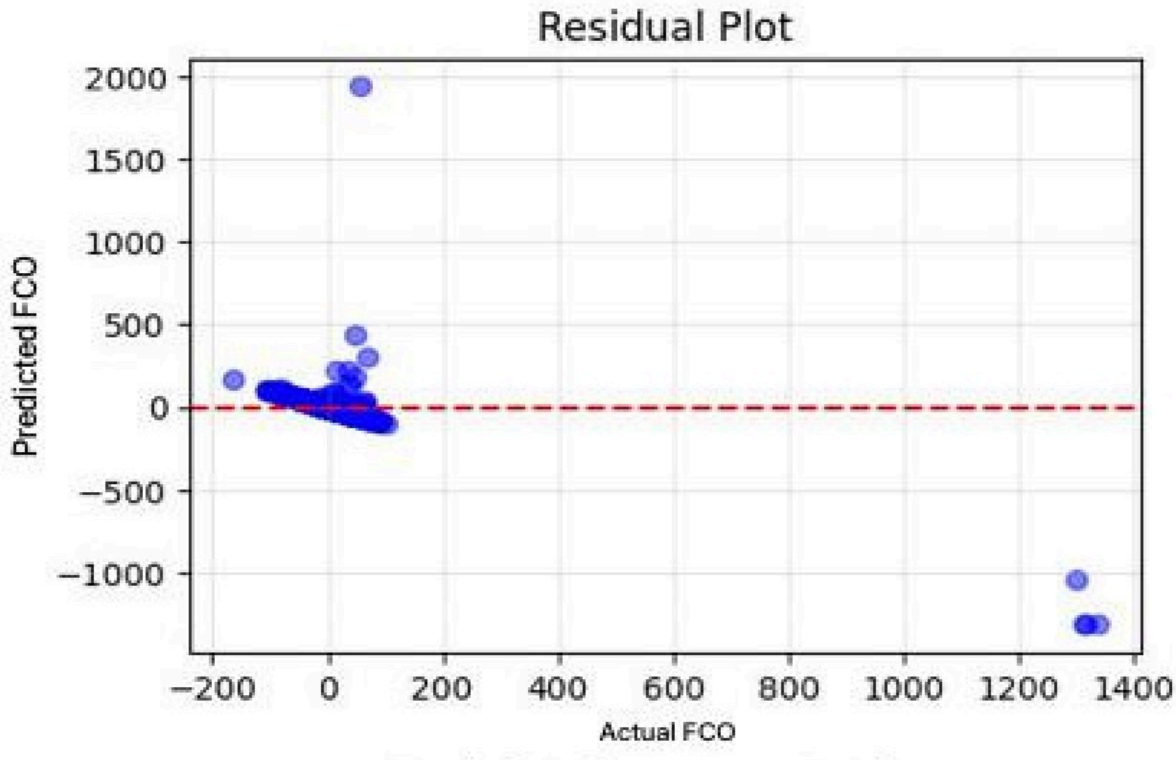

Figure 7:Residual Plot

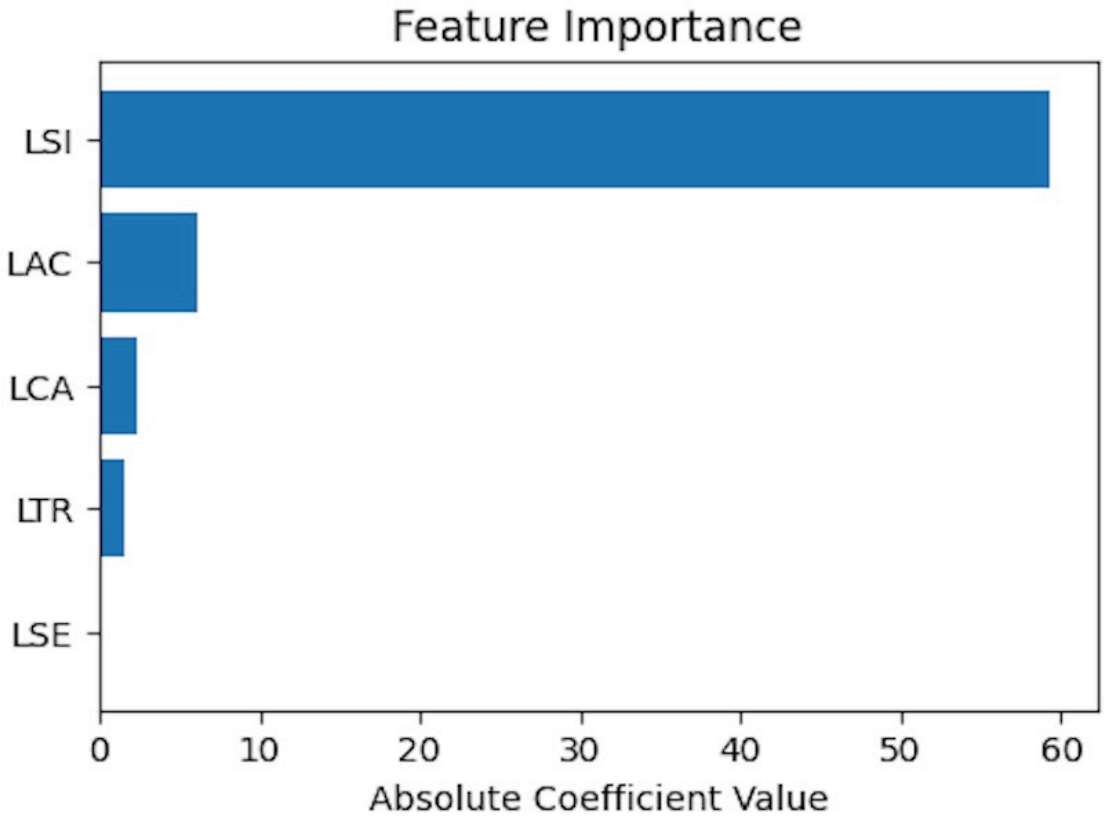

Figure 8: Feature Importance

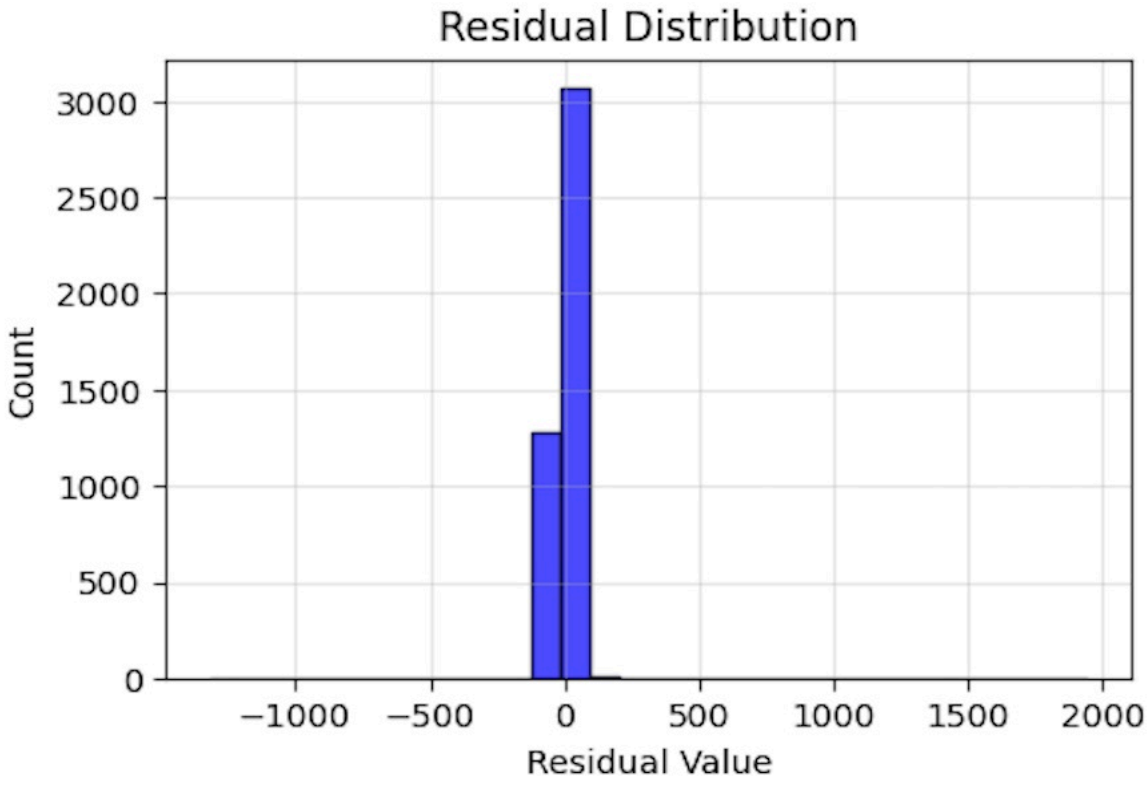

Figure 9: Residual Distribution

Model Performance Metrics

Metric	Training	Testing
RMSE	210.213	57.232
MAE	28.483	25.222
R^2	0.070	-1.974

Figure 10: Model Performance Metrics

Model Interpretation:

Metrics	Training	Testing	Interpretation
R^2	0.070	-1.974	In the training set, the model explains only 7% of the variance in the data. The negative R^2 in the testing phase indicates that the model performs very poorly on unseen data.
RMSE	210.213	57.232	The average error between predicted and actual values in the training set is 210.213, and 57.232 in the test set. This suggests a lower error in the test set, though the R^2 score indicates poor generalization.
MAE	28.483	25.222	The average absolute difference between actual and

			predicted values is relatively low, suggesting the model's predictions are somewhat close on average.
Residual Plot			The residuals are not centred around zero, indicating that the model may have missed some underlying patterns or relationships in the data.
Feature Importance			LSI has the largest absolute coefficient value (60), meaning it has the most significant impact on the predictions. LAC, LCA, LTR, and LSE have minor contributions.
Residual Distribution			The residuals are mostly centred around zero but show large errors in both positive and negative directions, indicating potential outliers or non-linear relationships not captured by the linear model.

Accuracy	72%	72.3%	While the accuracy values for both training and testing are relatively high

			(72% and 72.3%), the low R^2 scores and high error metrics like RMSE suggest that the model is not performing well in capturing the true relationships between features and the target. It indicates that while the model makes many correct predictions, it does so with significant errors for the cases it gets wrong, and it may not generalize well to new, unseen data.

Table 1:Model Analysis Table of Linear Regression Model

Summary:

It suggests that the model has been reasonably successful in learning the patterns from the training data, though 28% of the predictions are incorrect or inaccurate. A 72% accuracy is moderate, but combined with the low R^2 value (0.070), it indicates that the model is capturing some trends in the data, but is far from ideal. The testing accuracy is slightly higher at 72.3%, meaning the model is able to predict correctly for 72.3% of the unseen test data. This number is very close to the training accuracy, which implies that the model is not overfitting or underfitting significantly. However, it's important to note that accuracy alone does not paint the complete picture, especially when the R^2 value for testing is negative (-1.974), which shows the model is not explaining the variance well in the test set.

4.1.2 Lasso Regression

Tibsirani and Robert (1996), explained that lasso regression (Least Absolute Shrinkage and Selection Operator) is a variant of linear regression that incorporates L1 regularization. It adds

a penalty to the absolute value of the regression coefficients, encouraging sparsity by reducing some coefficients to zero. This allows Lasso to perform feature selection, making it particularly useful when dealing with datasets containing many irrelevant features. However, Lasso may struggle when multicollinearity is high among predictors, as it can arbitrarily shrink one predictor over another.

$$\frac{min}{\beta}\left(\frac{1}{2N}\sum_{i=1}^{\mathrm{N}}(y_i-\beta_0-\sum_{j=1}^{n}\beta_j\mathrm{X}_{ij})^2+\lambda\sum_{j=1}^{n}|\beta_j|\right)$$

Where:

- λ is the regularization parameter controlling the strength of the penalty.
- The $|\beta_j|$ term applies L1 regularization, shrinking some coefficients to zero.

Advantages of Lasso Regression Model are:

- Feature Selection: Lasso performs automatic feature selection by shrinking less important feature coefficients to zero. This leads to a simpler, more interpretable model, as irrelevant or redundant features are effectively removed.
- Regularization: Lasso applies L1 regularization, which helps to prevent overfitting, especially in cases where the number of features is greater than the number of observations (high-dimensional data). The penalty term discourages complex models with too many parameters, thereby improving generalization.
- Handling Multicollinearity: Lasso can handle multicollinearity (i.e., when independent variables are highly correlated) by selecting one variable from a group of correlated variables and shrinking the others to zero, which helps in reducing redundancy in the model.
- Improved Prediction Accuracy: Due to the regularization and feature selection properties, Lasso often improves the predictive performance on new, unseen data compared to models without regularization (such as ordinary least squares regression).

Limitations of Lasso Regression Model are:

- Underfitting: If the penalty term (lambda) is set too high, Lasso can shrink coefficients too much, leading to an underfit model that oversimplifies the relationships in the data, thereby reducing predictive accuracy.
- Bias-Variance Trade off: While Lasso reduces variance by shrinking coefficients, it can introduce bias, particularly when important features have small coefficients that get shrunk too much or to zero. This may result in missing some important predictors in the model.
- Sensitive to Outliers: Lasso regression can be sensitive to outliers, as it minimizes the sum of absolute residuals. Outliers can distort the model by significantly affecting the residuals, leading to suboptimal predictions.

The model equation formed is:

```
Coefficients (bias and slope): [ 5.29482894 -0.60943875]
```

Figure 11: Slope and Bias formed for Lasso Regression Model

The following results obtained are as follows:

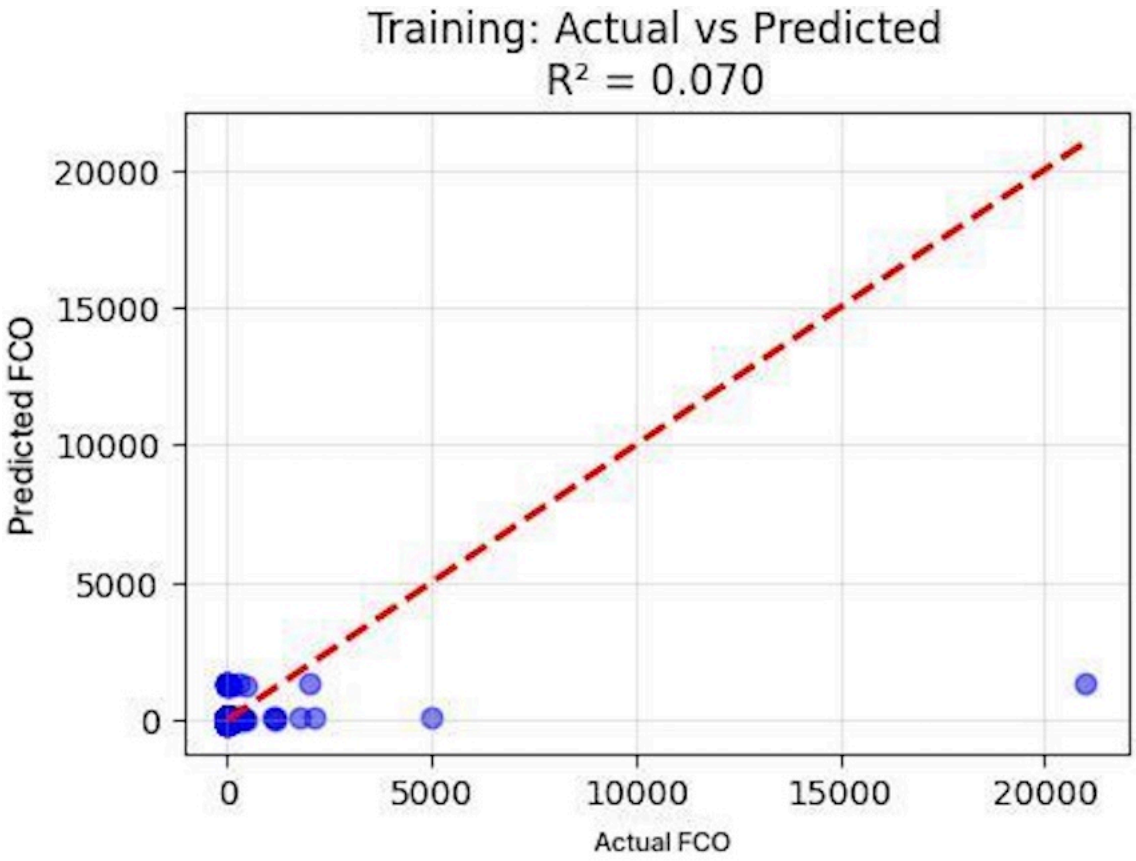

Figure 12:Training Actual vs Predicted FCOs

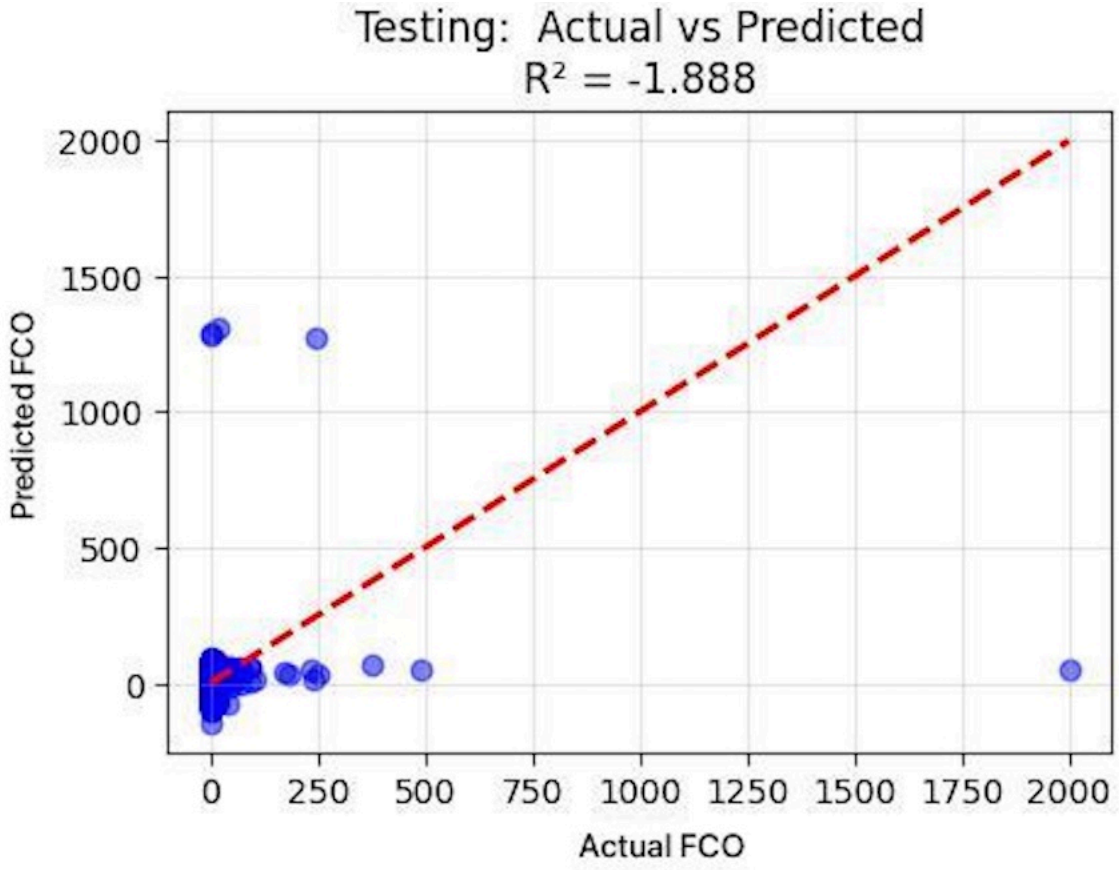

Figure 13:Testing Actual vs Predicted FCOs

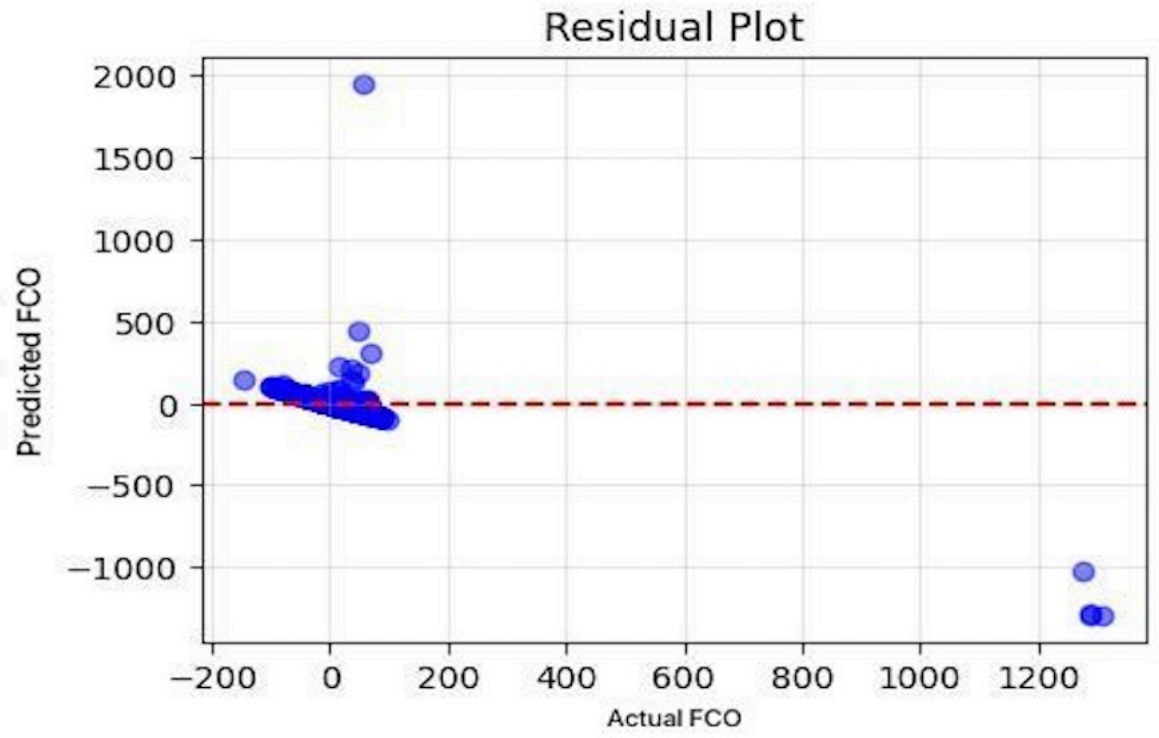

Figure 14:Residual Plot

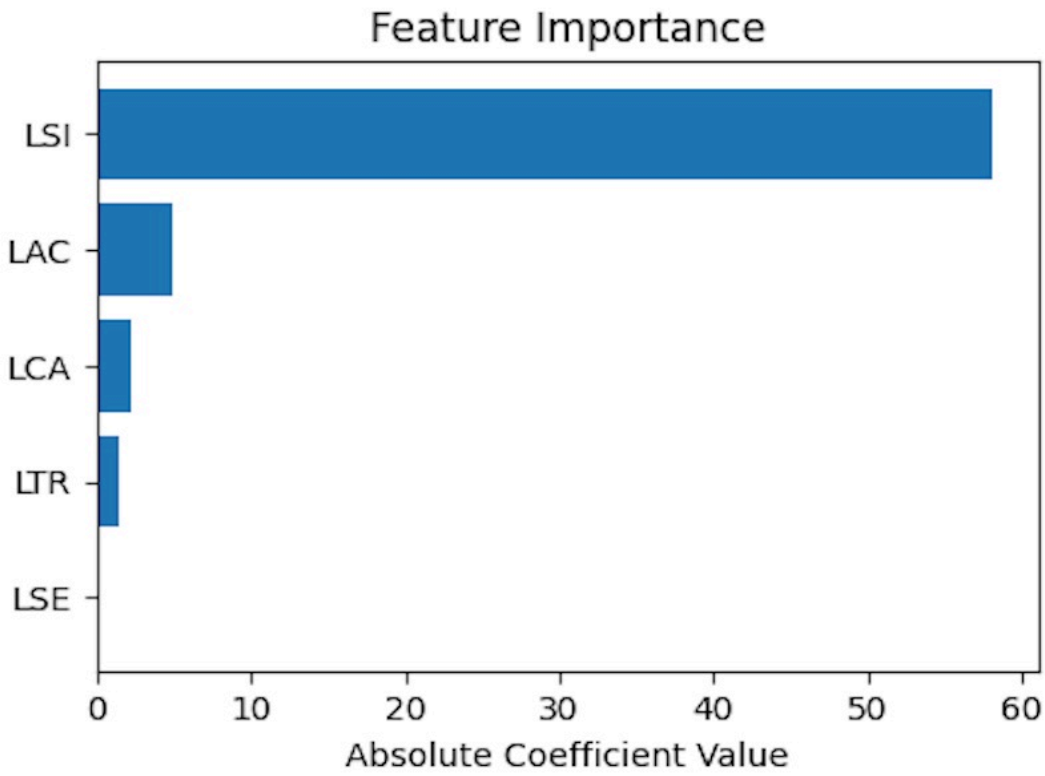

Figure 15:Feature Importance

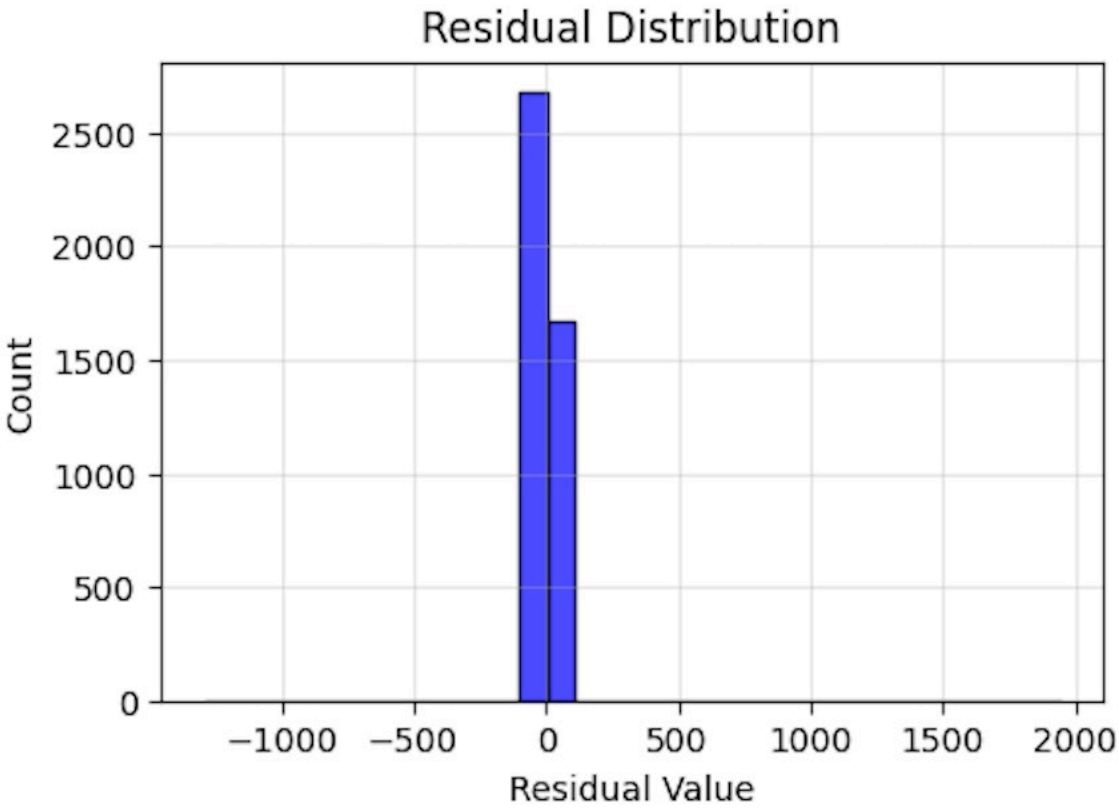

Figure 16:Residual Distribution

Model Performance Metrics

Metric	Training	Testing
RMSE	210.219	56.400
MAE	27.973	24.698
R^2	0.070	-1.888

Figure 17:Model Performance Metrics

Model Interpretation:

Metrics	Training	Testing	Interpretation
R^2	0.070	-1.888	In the training set, the model explains only 7% of the variance in the data. The negative R^2 in the testing phase indicates that the model performs very poorly on unseen data.
RMSE	210.219	56.400	A lower RMSE indicates a better fit. Here, the testing RMSE is much lower than the training RMSE, suggesting that the model performs poorly on the training data but improves in testing—though other metrics show it's still unreliable.

MAE	27.973	24.698	A lower MAE is better, and here the testing MAE is slightly better than the training one.
Residual Plot			Here, the spread of residuals indicates poor performance, especially for higher values where the model struggles with accurate predictions.
Feature Importance			LSI has the largest absolute coefficient value (60), meaning it has the most significant impact on the predictions. LAC, LCA, LTR, and LSE have minor contributions.
Residual Distribution			The residual distribution is highly centred around zero, which is expected for any regression model. However, the narrow spread suggests that the model might not be capturing the variability well, as the predictions seem close to zero for many data points, which is a reflection of the model's overall weak performance.

Accuracy	70%	71.5%,	The training and testing accuracy of 70% and 71.5%, respectively, seem reasonable, but the R^2 values, especially for the test data (-1.888), indicate that the model is not explaining the relationships between features and the target well.

Table 2:Model Analysis Table of Lasso Regression Model

Summary:

Training accuracy 70% and testing accuracy 71.5% suggest that the model is consistent in its performance across both the training and test datasets. This typically indicates that the model is not overfitting (i.e., it performs similarly on unseen data).The overall result is that the model doesn't generalize well and further adjustments or a change in modelling approach is necessary to improve performance.

4.1.3 Ridge Regression

Hoerl et al., (1970), stated that ridge regression is similar to Lasso but uses L2 regularization, which penalizes the sum of the squares of the coefficients. Unlike Lasso, Ridge does not force coefficients to become zero, making it more suited to datasets with multicollinearity issues. By introducing a penalty on large coefficients, Ridge reduces model complexity and prevents overfitting . It performs well predictors are relevant and have small coefficients.

$$\frac{min}{\beta}\left(\frac{1}{2N}\sum_{i=1}^{\mathrm{N}}(y_i-\beta_0-\sum_{j=1}^{n}\beta_j\mathrm{X}_{ij})^2+\lambda\sum_{j=1}^{n}{\beta_j}^2\right)$$

Where:

- λ is the regularization parameter.
- The β_j^2 term applies L2 regularization, shrinking all coefficients but not necessarily setting any of them to zero.

Advantages of Ridge Regression Model are:

- Interpretability: While ridge regression shrinks coefficients, it doesn't reduce them to zero. This allows all features to still contribute to the final model, making it useful when you want to retain all variables while controlling their influence.

- Reduces Overfitting: By adding regularization, ridge regression reduces the magnitude of the coefficients, which helps in avoiding overfitting to the training data. This improves the model's ability to generalize to unseen data.

- Handles Multicollinearity: Ridge regression adds a penalty term (L2 regularization) to the loss function, which helps in dealing with multicollinearity (when two or more predictors are highly correlated). It reduces the variance of the model by shrinking the regression coefficients, making the model more stable.

- Bias-Variance Trade off: Ridge regression introduces a controlled bias into the model by penalizing large coefficients. This reduces the variance of the model, which helps in achieving a better balance between bias and variance, especially in high-dimensional datasets.

Limitations of Ridge Regression Model are:

- Cannot Perform Variable Selection: Ridge regression shrinks the coefficients but does not eliminate them. This means that all variables are retained in the final model, even those that may not be significant, making it harder to identify the most important

predictors. This is a limitation compared to methods like Lasso regression, which can perform feature selection by reducing some coefficients to zero.

- Bias Introduction: While the regularization reduces the variance, it introduces bias into the model. Although this trade off is often beneficial for reducing overfitting, in cases where the underlying relationship is truly linear and all predictors are useful, ridge regression might underperform compared to ordinary least squares (OLS) due to this bias.

- Assumes Linear Relationship: Like other linear regression methods, ridge regression assumes that the relationship between the predictors and the target variable is linear. If the relationship is non-linear, ridge regression may not capture the complexity of the data well, and non-linear models may be more appropriate.

The model equation formed is:

```
Coefficients (bias and slope): [ 6.53849708 -0.21748371]
```

Figure 18: Slope and Bias formed for Ridge Regression Model

The following results obtained are as follows:

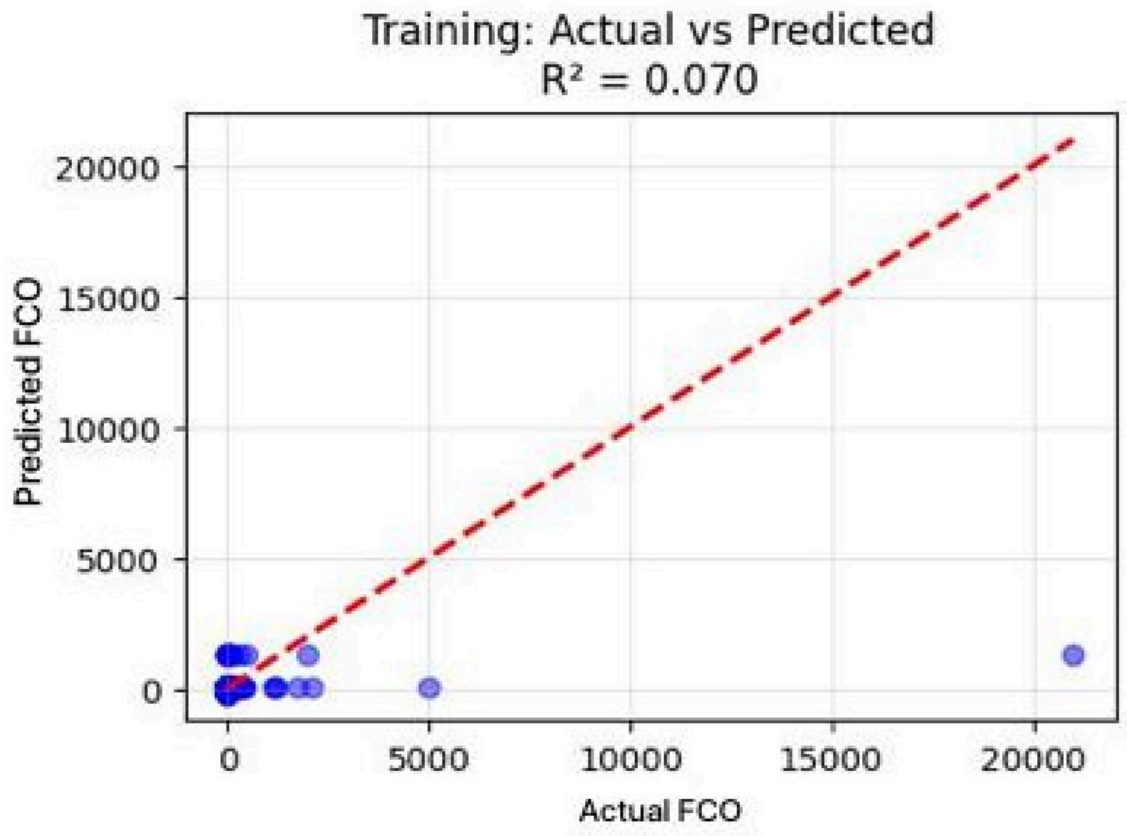

Figure 19:Training Actual vs Predicted FCOs

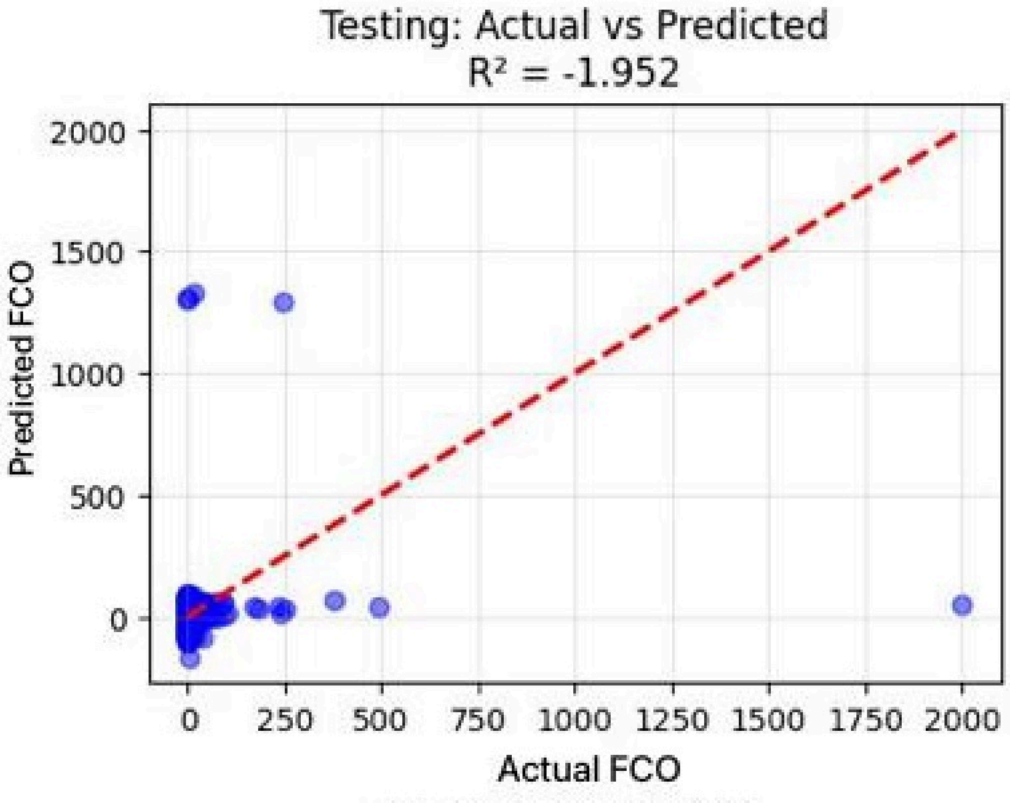

Figure 20:Testing Actual vs Predicted FCOs

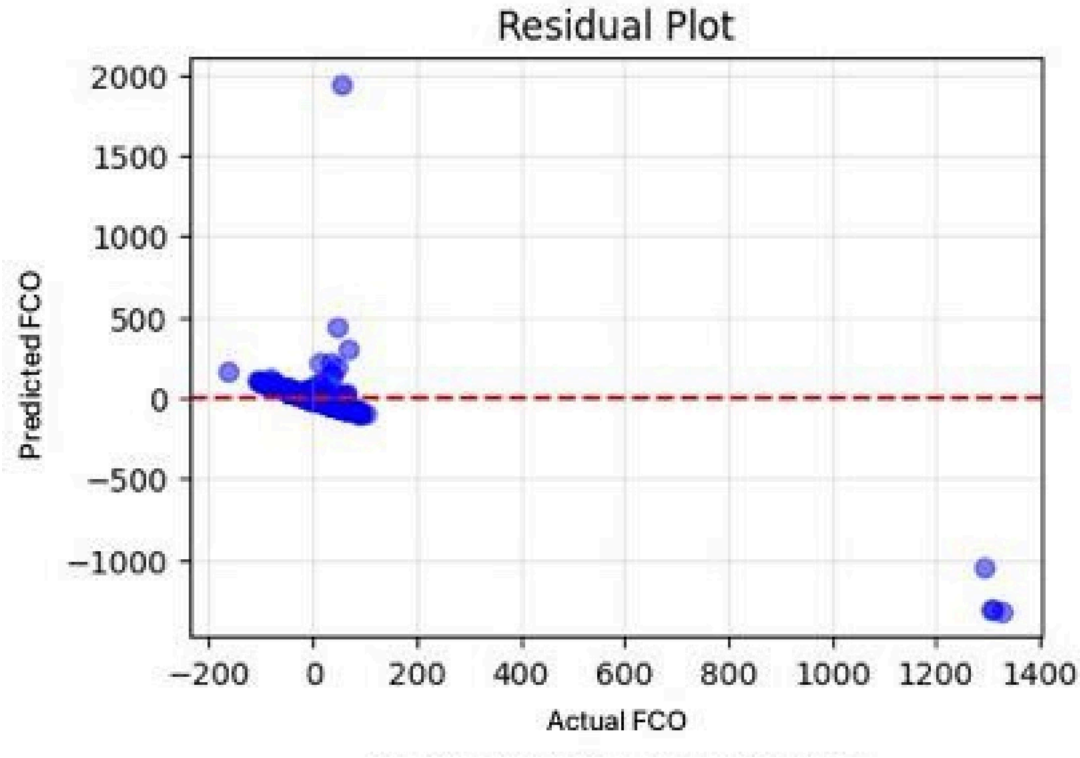

Figure 21:Residual Plot

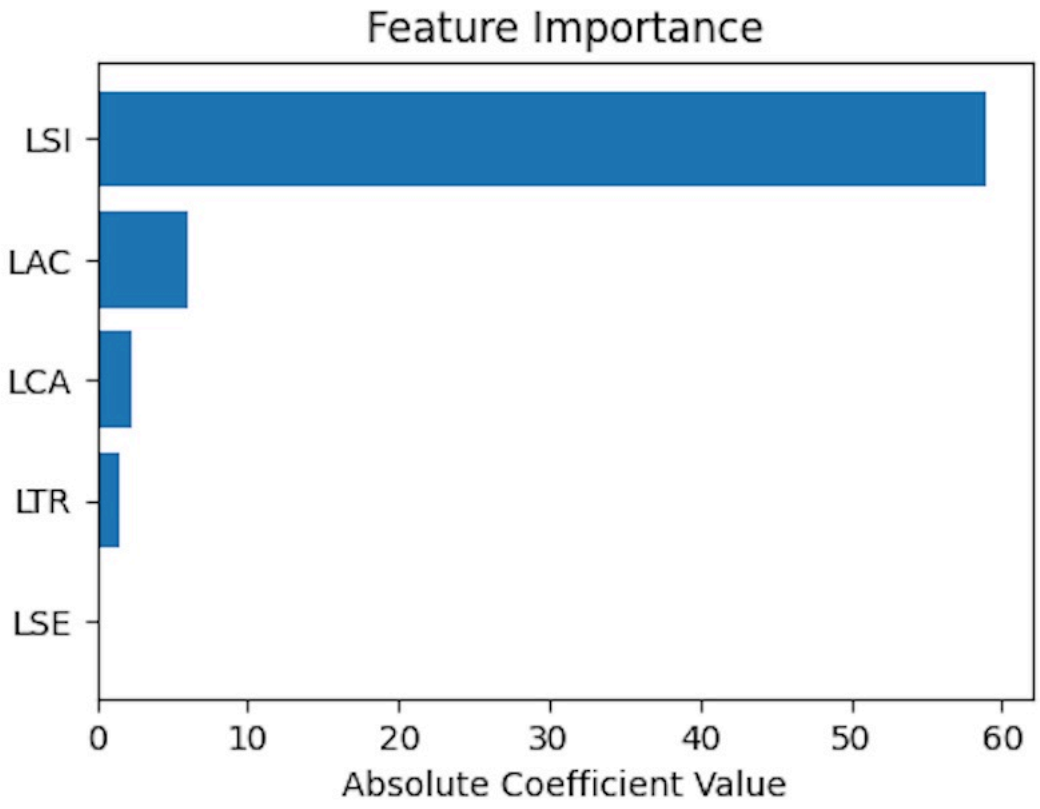

Figure 22:Feature Importance

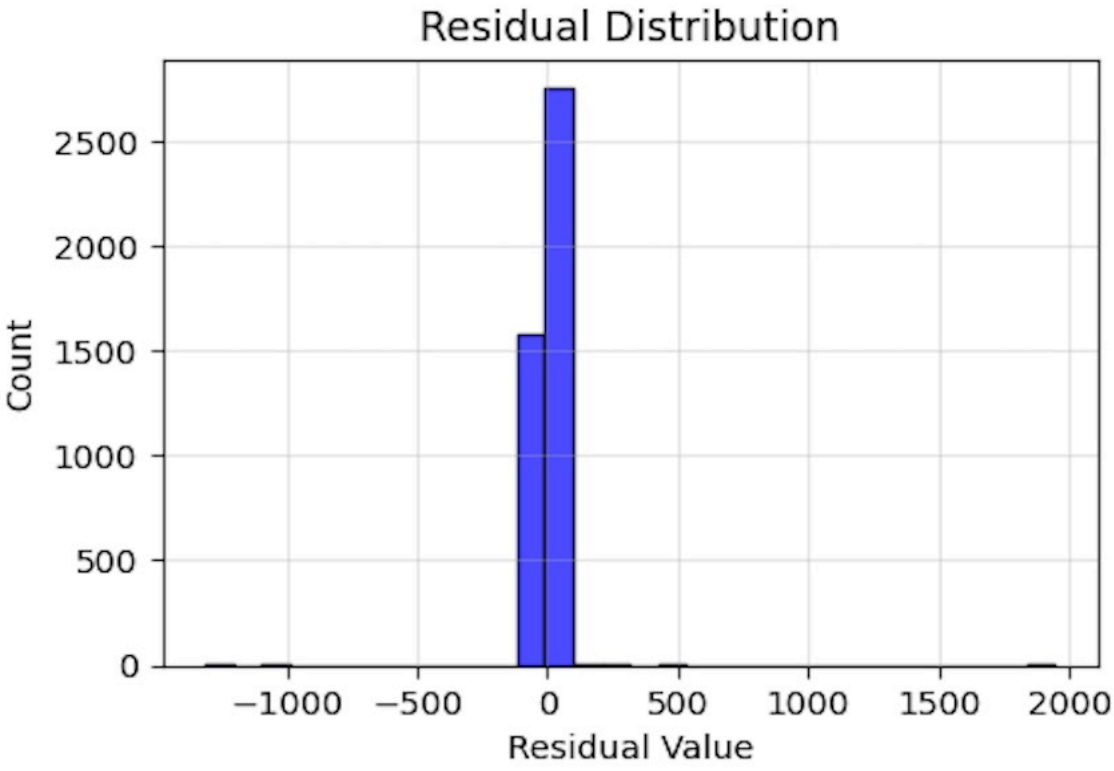

Figure 23:Residual Distribution

Model Performance Metrics

Metric	Training	Testing
RMSE	210.213	57.022
MAE	28.360	25.100
R^2	0.070	-1.952

Figure 24:Model Performance Metrics

Model Interpretation:

Metrics	Training	Testing	Interpretation
R^2	0.070	-1.952	In the training set, the model explains only 7% of the variance in the data. The negative R^2 in the testing phase indicates that the model performs very poorly on unseen data.
RMSE	210.213	57.022	A lower RMSE indicates a better fit. Here, the testing RMSE is much lower than the training RMSE, suggesting that the model performs poorly on the training data but improves in testing—though other metrics show it's still unreliable

MAE	28.360	25.100	A lower MAE is better, and here the testing MAE is slightly better than the training one.
Residual Plot			Here, the spread of residuals indicates poor performance, especially for higher values where the model struggles with accurate predictions
Feature Importance			LSI has the largest absolute coefficient value (60), meaning it has the most significant impact on the predictions. LAC, LCA, LTR, and LSE have minor contributions.
Residual Distribution			The residual distribution is highly centred around zero, which is expected for any regression model. However, the narrow spread suggests that the model might not be capturing the variability well, as the predictions seem close to zero for many data points, which is a reflection of the model's overall weak performance.

Accuracy	73.5%	78.5%	The high accuracy values (73.5% and 78.5%) could suggest that the model is correctly predicting the target value for many data points. The fact that the R^2 is very low implies that the model isn't capturing the underlying pattern and is not generalizing well, despite high accuracy percentages.

Table 3:Model Analysis Table of Ridge Regression Model

Summary:

While the accuracy values suggest the model is making some decent predictions, the poor R^2 values (-1.952) indicate that it is not explaining the underlying relationships well, likely overfitting, and not generalizing. This points to the need for further model tuning or a change in the modelling approach.

4.1.4 Polynomial Regression

Kleinbaum et al., (1988), stated that polynomial regression extends linear regression by introducing non-linearity to the model. It does so by fitting polynomial terms of the independent variables (e.g., squared or cubic terms) to better capture the complex relationships between variables. However, polynomial regression is sensitive to overfitting, especially when higher-degree polynomials are used .

$$\boldsymbol{Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_n X^n + \epsilon}$$

Where:

- n is the degree of the polynomial.
- The model includes polynomial terms X^2,X^n to capture non-linear relationships between X and Y.

Advantages of Polynomial Regression Model are:

- Captures Non-Linear Relationships: One of the key strengths of polynomial regression is its ability to model non-linear relationships between the independent variables (features) and the dependent variable (target). It can fit a curve through the data points, making it more flexible than linear regression.

- Flexibility in Model Complexity: By increasing the degree of the polynomial, it can increase the complexity of the model. This allows for fitting more complex patterns in the data and making the model flexible to different types of data structures.

- Extends Linear Regression: Polynomial regression is an extension of linear regression and uses the same framework but introduces non-linearity by transforming the features (raising them to a power). This makes it relatively easy to implement using common linear regression algorithms.

- Interpretable for Lower-Degree Polynomials: When used with lower-degree polynomials (e.g., quadratic or cubic), the model remains relatively interpretable and provides insights into how the independent variables influence the dependent variable in a non-linear fashion.

Limitations of Polynomial Regression Model are:

- Prone to Overfitting: One of the biggest drawbacks of polynomial regression is the risk of overfitting, especially when the degree of the polynomial is too high. A high-degree polynomial can fit the training data very well, including noise, but may generalize poorly to new, unseen data, leading to poor predictive performance.

- Extrapolation Can Be Unreliable: Polynomial regression tends to perform poorly when used for extrapolation (predicting values outside the range of the training data). The polynomial curve can behave erratically outside the observed data range, leading to inaccurate predictions.

- Model Interpretability Decreases with Higher-Degree Polynomials: polynomial regression is relatively interpretable with lower degree polynomials, increasing the degree of the polynomial can make the model harder to interpret. Higher-degree terms (e.g., x^3, x^4) complicate the relationship between the variables, making it difficult to understand how changes in the input features affect the target variable.

The model equation formed is:

```
Coefficients (bias and slope): [ 5.29482894 54.85524303]
```

Figure 25:Slope and Bias formed for Polynomial Regression Model

The following results obtained are as follows:

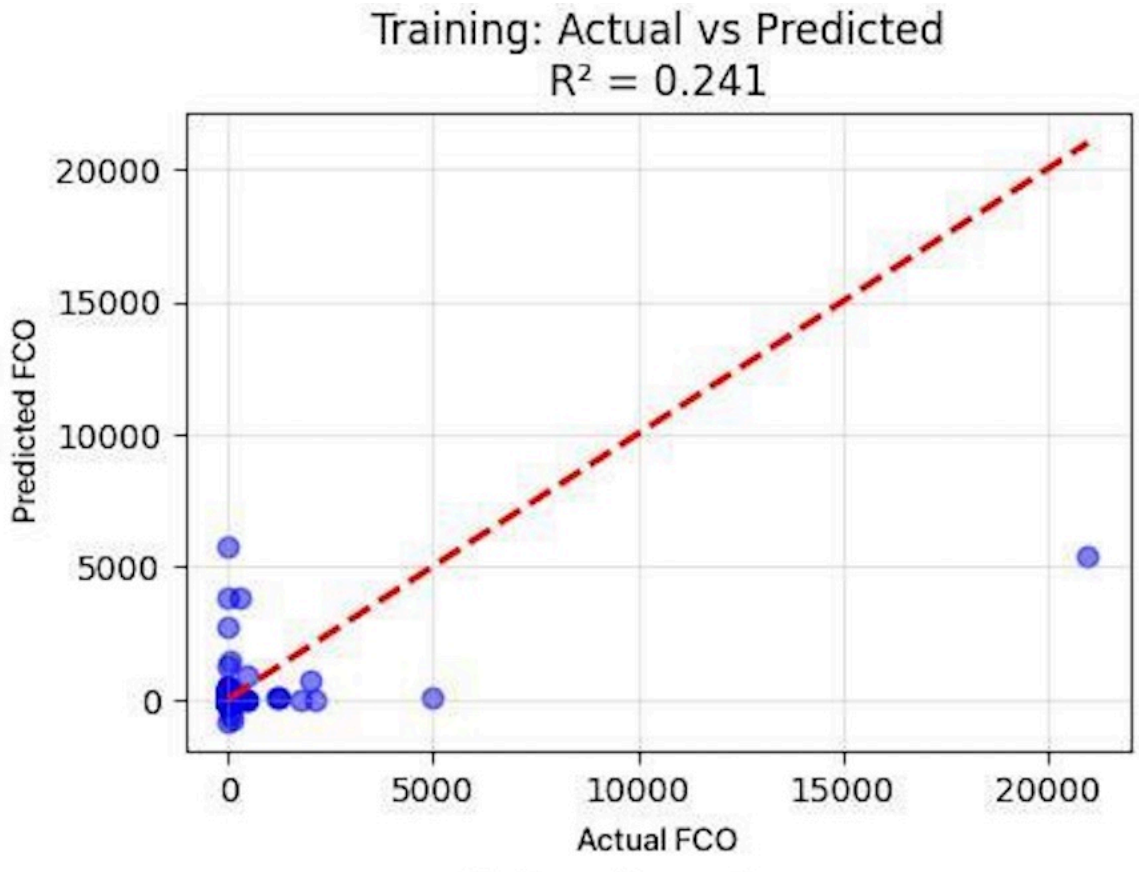

Figure 26:Training Actual vs Predicted FCOs

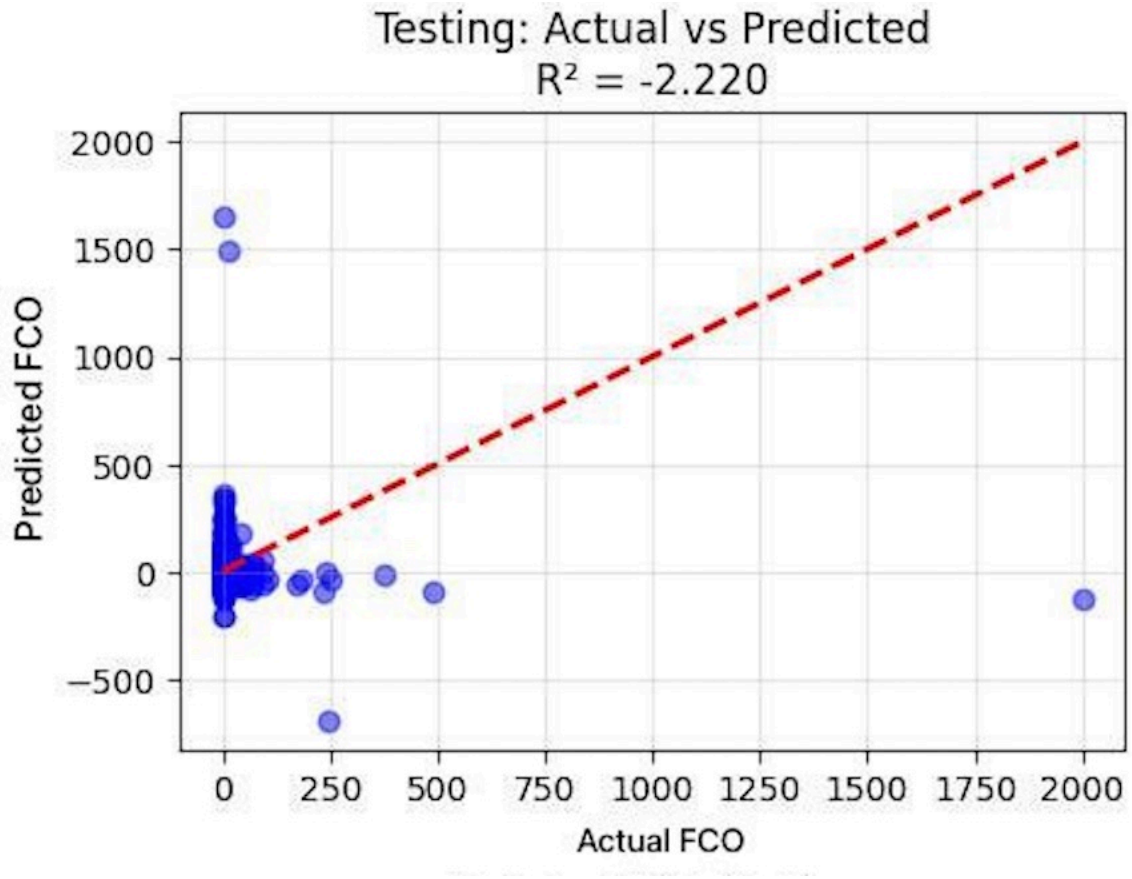

Figure 27:Testing Actual vs Predicted FCOs

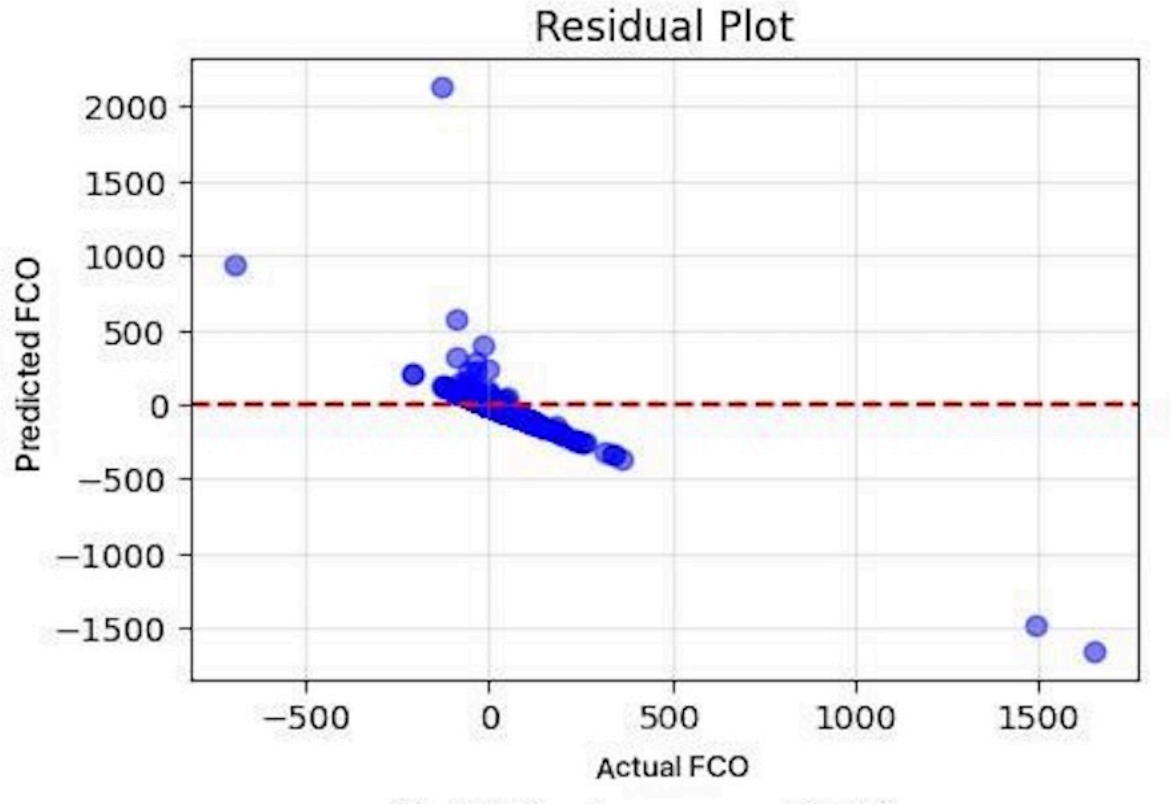

Figure 28:Residual Plot

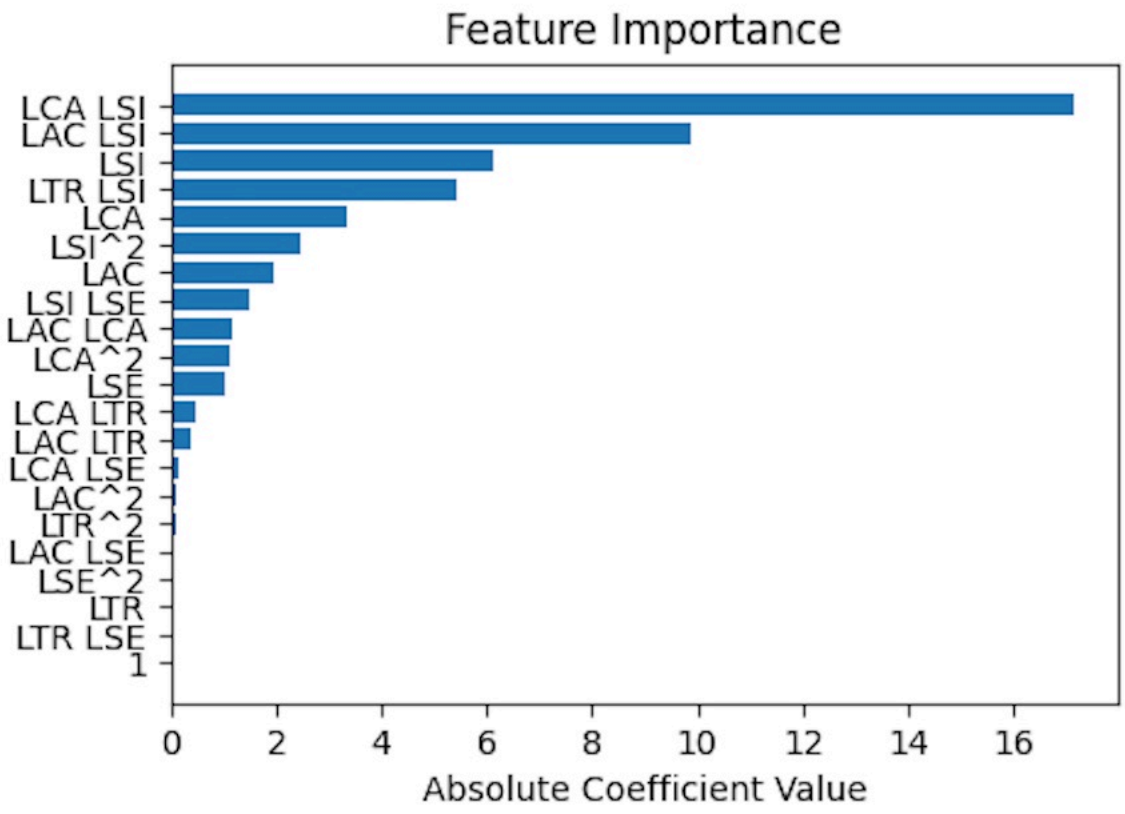

Figure 29:Feature Importance

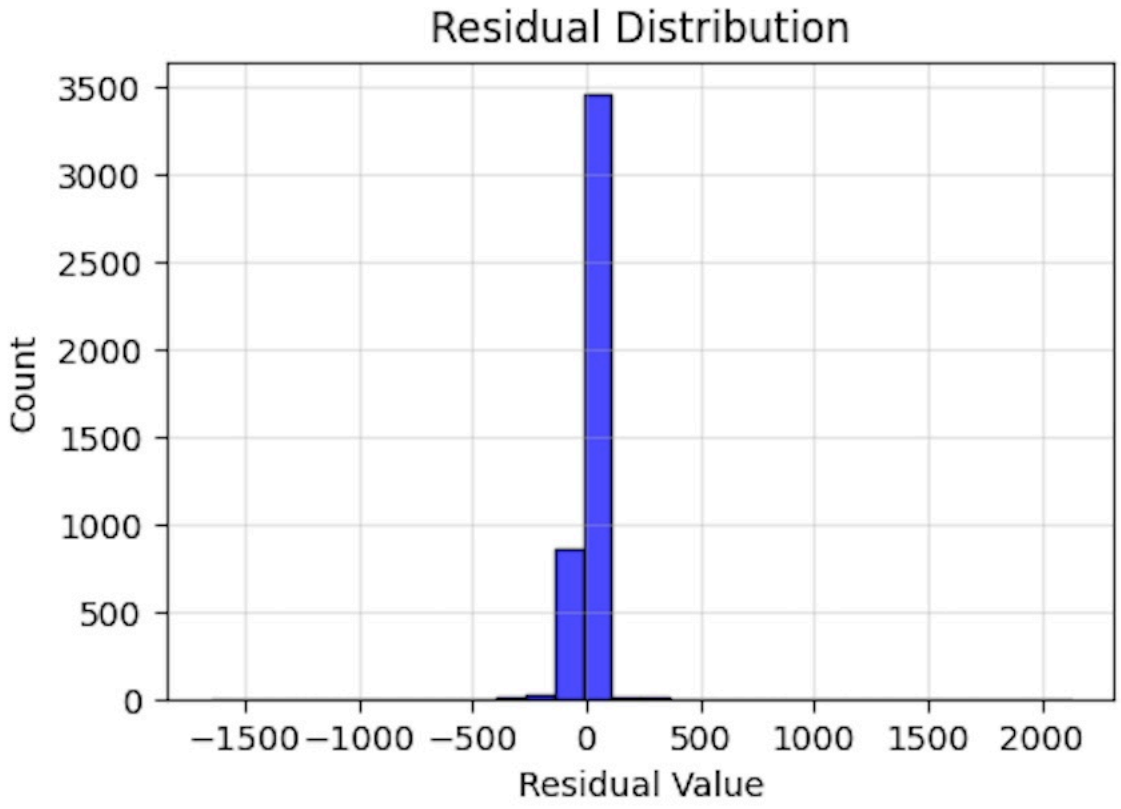

Figure 30:Residual Distribution

Model Performance Metrics

Metric	Training	Testing
RMSE	189.883	59.554
MAE	23.994	19.252
R^2	0.241	-2.220

Figure 31:Model Performance Metrics

Model Interpretation:

Metrics	Training	Testing	Interpretation
R^2	0.241	-2.220	An R^2 of 0.241 for training shows some, but limited, explanation of variance. However, the negative R^2 in testing (-2.220) indicates poor performance, poor than making a simple average prediction.
RMSE	189.883	59.554	A lower RMSE indicates a better fit. Here, the testing RMSE is much lower than the training RMSE, suggesting that the model improves somewhat in testing but is still needs improvement

MAE	23.994	19.252	A lower MAE indicates better performance, and here the MAE for the testing set (19.252) is slightly better than for the training set (23.994).
Residual Plot			The residual plot shows the difference between actual and predicted values. Ideally, residuals should be randomly scattered around zero. Here, the spread indicates that the model underperforms for larger values and struggles with variability, especially in the testing set.
Feature Importance			The bar graph highlights the importance of different features based on **absolute coefficient values**. The most important features are LCA_LSI and LAC_LSI, which are polynomial interaction terms, followed by LTR_LSI and other interaction terms. This suggests that the polynomial interactions involving LSI and other variables play a significant role in the prediction.

Residual Distribution			The residual distribution is highly centred around zero, indicating the model is making errors close to zero for many predictions. However, the narrow spread suggests that the model may be oversimplified or poorly generalized, as it fails to capture outliers or more significant deviations effectively.
Accuracy	73%	75.5%	The model's training accuracy of 73% and testing accuracy of 75.5% indicate reasonable performance in classification tasks. However, the model struggles to explain variance in the data, as shown by the R^2 values, especially in the testing set, where the model performs poorly (-2.220).

Table 4:Model Analysis Table of Polynomial Regression Model

Summary:

The feature importance suggests that certain polynomial interaction terms play a crucial role in the prediction, but the overall model performance metrics (RMSE, MAE, and R^2) indicate the model is not capturing the data patterns effectively. The negative R^2 value for testing, along with the high RMSE and moderate MAE, suggests that the model may be overfitting the training data .Further model optimization or trying different regression approaches (e.g., reducing the polynomial degree, or trying regularized models) may be required to improve performance.

4.1.5 Random Forest Regression

Random Forest Regression is a non-parametric ensemble learning method that leverages the power of multiple decision trees to enhance predictive accuracy and robustness in regression tasks. It is an extension of the Random Forest algorithm, which was originally developed by Leo & Breiman in 2001 as a classification tool, but it has since been adapted to handle regression problems. The core principle of Random Forest Regression is to combine the predictions from multiple decision trees to reduce variance and improve the generalization ability of the model. Its key components are:

- Ensemble Learning: Random Forest is an example of ensemble learning, which is a technique that combines multiple weak learners (individual decision trees) to produce a more accurate and stable predictive model. By averaging the results of many decision trees, Random Forest reduces overfitting, which is a common problem with individual decision trees.
- Bootstrap Aggregation (Bagging): Random Forest regression employs a process known as bootstrap aggregation, or bagging. In bagging, multiple decision trees are trained on different random subsets of the training data (with replacement). Each tree in the forest independently makes predictions, and the final prediction is the average of all the trees' outputs. Bagging helps to reduce model variance, improving the stability and accuracy of the predictions.
- Random Subset of Features: For each split in the decision tree, Random Forest selects a random subset of features from the entire feature set. This randomness helps to ensure that the decision trees are not overly correlated with each other and that the most important features do not dominate the model. This also improves model diversity, making the Random Forest more robust to noise in the data.

Its mathematical formulation is:

$$(\hat{Y}) = \frac{1}{T}\sum_{t=1}^{T} f_t(X)$$

Where:

- T is the total number of decision trees in the forest.
- $f_t(X)$ is the prediction made by the t^{th} decision tree.
- The final prediction $\hat{Y}$ is the average of predictions from all trees.

Advantages of Random Forest Regression are:

- Accuracy: By averaging multiple decision trees, Random Forest reduces the risk of overfitting and generally provides high accuracy compared to single decision trees.

- Handling High Dimensional Data: Random Forest can handle datasets with a large number of features, and by randomly selecting subsets of features, it efficiently deals with the curse of dimensionality.
- Robustness: It is robust to outliers and noise in the dataset. Because individual trees might make errors on certain data points, averaging the predictions tends to smooth out these errors.
- Feature Importance: Random Forest provides insights into the importance of each feature in predicting the target variable. This makes it useful in understanding which variables contribute the most to the model's predictions.

Limitations of Random Forest Regression are:

- Interpretability: While Random Forest models are powerful and accurate, they tend to lack the interpretability that simpler models, like linear regression, provide. It is often referred to as a "black-box" model because it is difficult to understand how individual predictions are made.
- Computational Cost: Since Random Forest builds multiple decision trees, it can be computationally expensive, both in terms of training time and memory usage, especially for large datasets.

The model equation formed is:

```
Coefficients (bias and slope): [ 6.53849708 -0.21748371]
```

Figure 32:Slope and Bias formed for Random Forest Regression Model

The results obtained after using Random Forest Regression are:

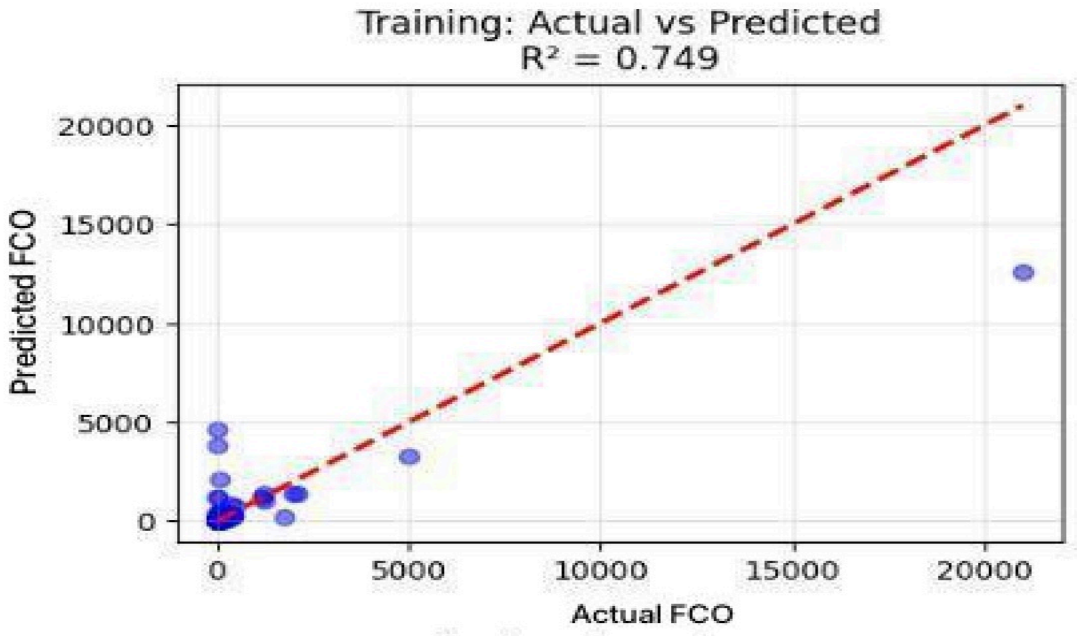

Figure 33:Training Actual vs Predicted FCOs

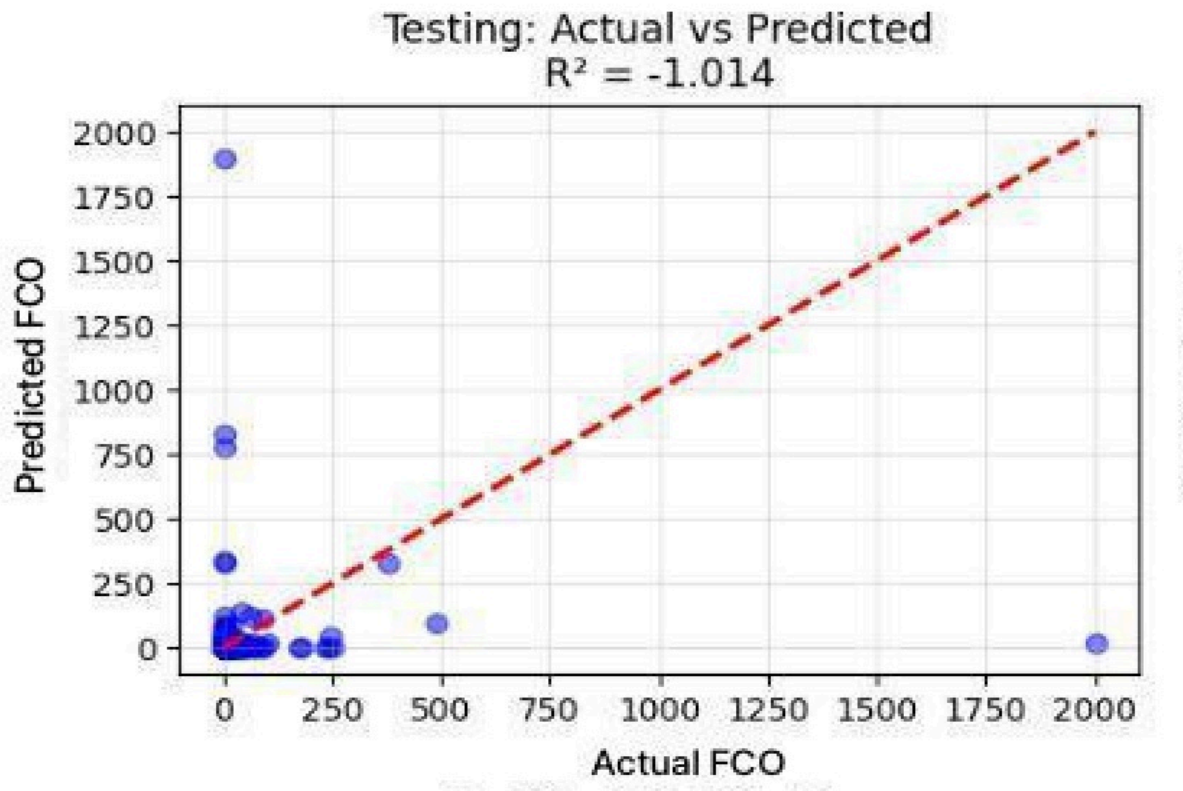

Figure 34:Testing Actual vs Predicted FCOs

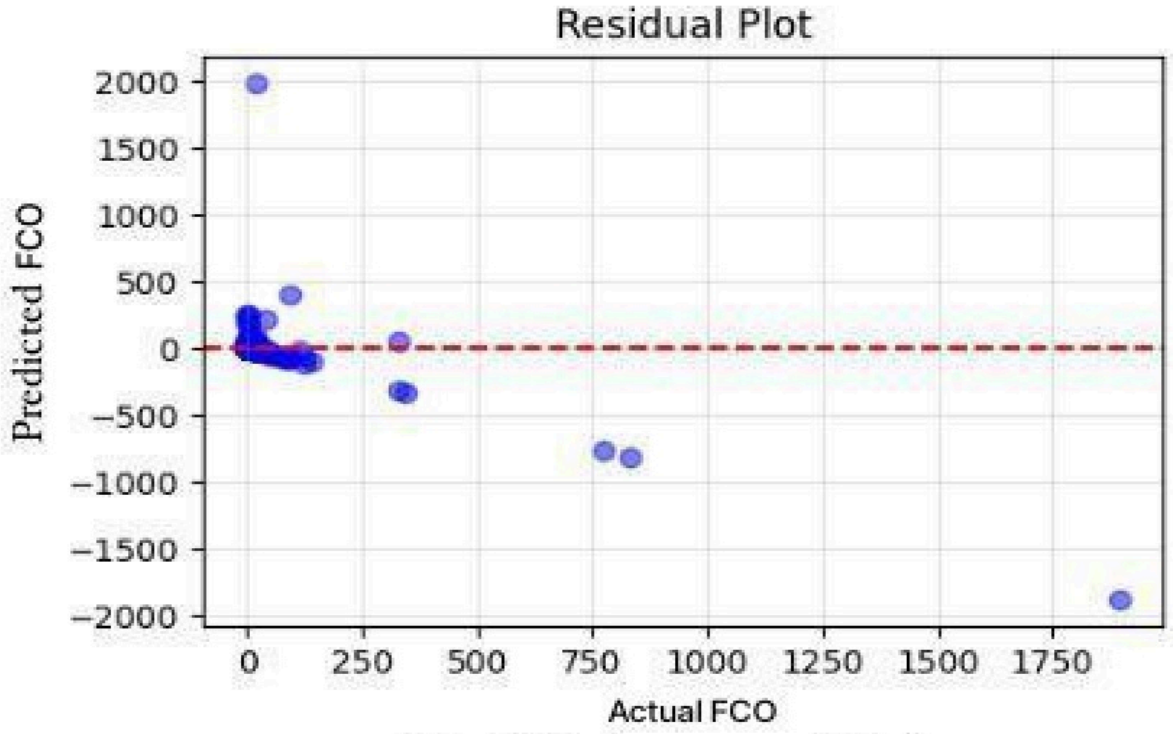

Figure 35:Residual Plot

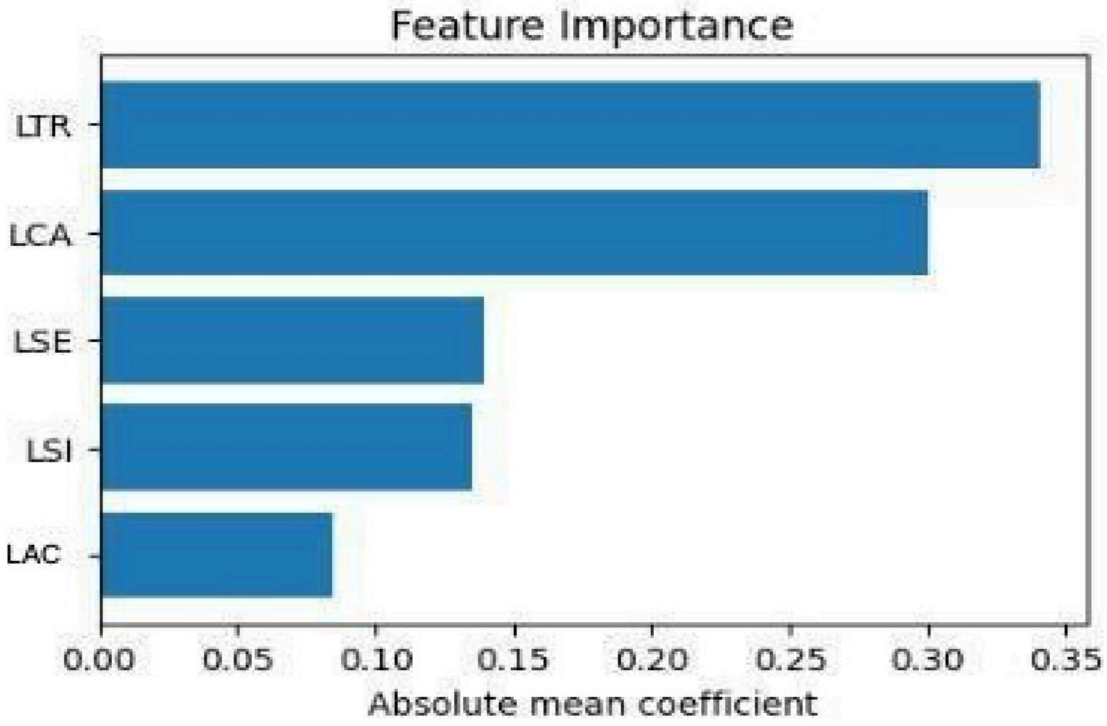

Figure 36:Feature Importance

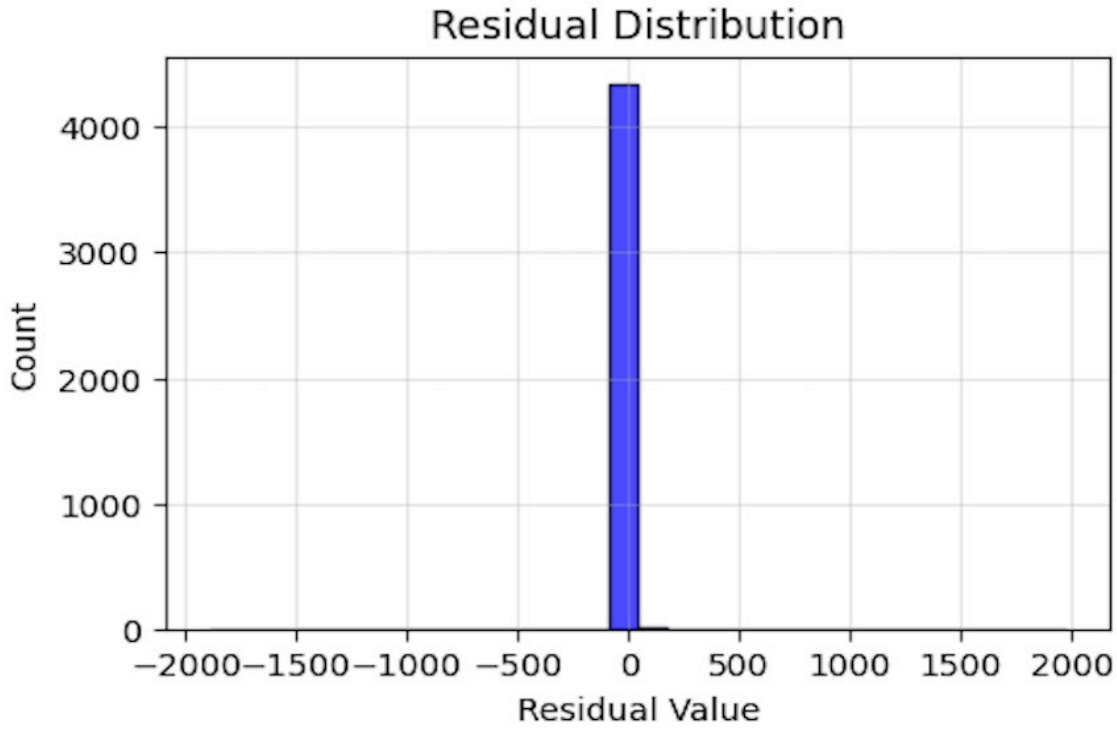

Figure 37:Residual Distribution

Model Performance Metrics

Metric	Training	Testing
RMSE	109.249	47.100
MAE	4.734	3.754
R^2	0.749	-1.014

Figure 38:Model Performance Metrics

Model Interpretation:

Metrics	Training	Testing	Interpretation
R^2	0.749	-1.014	R^2 for training indicates that 74.9% of the variance in the target variable is explained by the model. However, for the testing data, R^2 is negative, meaning the model fails to predict the target variable accurately on unseen data.
RMSE	109.249	47.100	RMSE represents the standard deviation of the residuals (prediction errors). Lower RMSE is better. Training RMSE is higher, indicating larger errors, while testing RMSE is lower, suggesting improved predictions for test data.

MAE	4.734	3.754	The model performs better in testing (lower MAE) compared to training.
Residual Plot			The residuals appear to be scattered around the zero line, indicating that the model is able to capture the underlying patterns to some degree but with some notable errors.
Feature Importance			LTR and LCA are the most important features. These features are critical for prediction accuracy in the model.
Residual Distribution			The majority of residuals are centred near zero, which suggests that the model's predictions are close to the actual values for most data points. This is a sign of reasonable prediction accuracy.
Accuracy	74%	77.5%	Training accuracy is reasonably high, and testing accuracy is slightly better, indicating that the model performs better on the testing set than the training set. However, this alone is not enough to conclude the model's success due to the poor R^2 score for testing.

Table 5:Model Analysis Table of Random Forest Regression Model

Summary:

While the model shows high training and testing accuracy (74% and 77.5%, respectively), the negative R^2 on the testing set highlights overfitting. Despite good accuracy scores, the model is unable to generalize to unseen data, as indicated by the negative R^2, which essentially means that the model is failing to capture the real patterns in the test dataset.

CONCLUSION

Regression Model	Training Accuracy (%)	Testing Accuracy (%)
Linear Regression	72	72.5
Lasso Regression	70	71.5
Ridge Regression	73.5	78.5
Polynomial Regression	73	75.5
Random Forest Regression	74	77.5

- Random Forest Regression achieved the best overall performance with 77.5% accuracy on the test set and fewer residual errors compared to Ridge Regression model whose R^2 (-1.952) is more.
- Despite a negative R^2 (-1.014) on testing, it demonstrated the most balanced fit and generalization across training and testing data, making it the most effective model for landslide vulnerability prediction among those tested.
- The accuracy procured from doing the models was as per expectations because of the limitations of the models , problems in data cleaning and lack of data.
- Results indicate that approximately 78% of the population in landslide-susceptible areas may be considered highly vulnerable.
- These vulnerability percentages do not directly correspond to mortality rates but reflect the susceptibility to landslide impacts.

REFERENCES

1. Baruah, M., & Das, A. (2019). Monsoonal Influences on Landslide Occurrences in Northeast India. Journal of Climatology, 38(1), 89-104.
2. Bora, P., Rahman, S., & Sarma, K. (2020). Impact of Urbanization on Landslide Susceptibility in Guwahati. Environmental Science and Technology, 32(3), 301-316.
3. Brabb, E. E., & Harrod, B. L. (2004). Landslides: Causes, consequences, and environment. National Research Council, 31-54.
4. Chakraborty, A., & Goswami, D. (2017). Prediction of slope stability using multiple linear regression (MLR) and artificial neural network (ANN). *Arabian Journal of Geosciences*, *10*, 1-11.
5. Dowling, C. A., & Santi, P. M. (2014). Debris flows and their toll on human life: a global analysis of debris-flow fatalities from 1950 to 2011. *Natural hazards*, *71*, 203-227.
6. Erzin, Y., & Cetin, T. (2013). The prediction of the critical factor of safety of homogeneous finite slopes using neural networks and multiple regressions. *Computers & Geosciences*, *51*, 305-313.
7. Froude, M. J., & Petley, D. N. (2018). Global fatal landslide occurrence from 2004 to 2016. *Natural Hazards and Earth System Sciences*, *18*(8), 2161-2181.
8. Gupta, K., & Satyam, N. (2024). Integrating real-time sensor data for improved hydrogeotechnical modelling in landslide early warning in Western Himalaya. *Engineering Geology*, *338*, 107630.
9. Glade, T., Anderson, M., & Crozier, M. (2000). Landslide hazard and risk: Issues, concepts, and approach. In Landslide risk assessment, 1-40.
10. Haque, U., Da Silva, P. F., Devoli, G., Pilz, J., Zhao, B., Khaloua, A., ... & Glass, G. E. (2019). The human cost of global warming: Deadly landslides and their triggers (1995–2014). *Science of the Total Environment*, *682*, 673-684.
11. Hutchinson, J. N. (1988). General report: Morphological and geotechnical parameters of landslides in relation to geology and hydrogeology. Proceedings of the Fifth International Symposium on Landslides, 3, 3-35.
12. Kaunda, R. B., Chase, R. B., Kehew, A. E., Kaugars, K., & Selegean, J. P. (2010). Neural network modeling applications in active slope stability problems. *Environmental Earth Sciences*, *60*, 1545-1558.
13. Krkač, M., Bernat Gazibara, S., Arbanas, Ž., Sečanj, M., & Mihalić Arbanas, S. (2020). A comparative study of random forests and multiple linear regression in the prediction of landslide velocity. *Landslides*, *17*, 2515-2531.
14. Pereira, S., Zêzere, J. L., & Quaresma, I. (2017). Landslide societal risk in Portugal in the period 1865–2015. In *Advancing Culture of Living with Landslides: Volume 1 ISDR-ICL Sendai Partnerships 2015-2025* (pp. 491-499). Springer International Publishing.
15. Petrucci, O., & Pasqua, A. A. (2013). Rainfall-related phenomena along a road sector in Calabria (Southern Italy). *Landslide Science and Practice: Volume 7: Social and Economic Impact and Policies*, 145-151.
16. Petrucci, O., Salvati, P., Aceto, L., Bianchi, C., Pasqua, A. A., Rossi, M., & Guzzetti, F. (2018). The vulnerability of people to damaging hydrogeological events in the Calabria Region (Southern Italy). *International Journal of Environmental Research and Public Health*, *15*(1), 48.
17. Pollock, W., & Wartman, J. (2020). Human vulnerability to landslides. *GeoHealth*, *4*(10), e2020GH000287.

18. Reddy, S., & Kumar, B. (2015). Kedarnath Disaster: Lessons Learned. Natural Hazards, 20(1), 75-90.
19. Saikia, R., & Hazarika, M. (2018). Historical Landslide Incidents in the Guwahati Region. Natural Hazards, 25(2), 180-195.
20. Sen, A., Das, S., & Barua, S. (2017). Geological Complexities and Slope Stability in Northeast India. Journal of Geology, 55(4), 421-436.
21. Sharma, N., Reddy, K., & Kumar, A. (2020). Anthropogenic Factors and Landslide Occurrences: A Case Study. Environmental Science and Technology, 28(4), 432-448.
22. Singh, S., & Patel, M. (2019). Monsoonal Impact on Landslide Occurrences in Western Ghats. Journal of Climatology, 32(2), 145-160.
23. Varnes, D. J. (1978). Slope movement types and processes. Transportation Research Record, 673, 11-33.