5CS037

Concepts and Technologies of AI

Evaluation of Regression Models for predicting Rainfall

Author

Narayan Lohani

2408869, L5CG15

Supervisor

Siman Giri

Module Leader

Table of Contents

Table of Figures

Abstraction

The objective of this report is to predict the total rainfall in mm before the landslide. For this the dataset chosen is “Landslide dataset” which can be found in Kaggle, under Raju Mavinmar. It contains data about the rainfall, steepness of slope where landslide occurred, saturation measure of soil, vegetation cover, last earthquake activity, proximity to water, and type of soil along with a landslide occurrence representation. The data was already cleaned so nothing was done to clean data further.

Foremost, to understand the data, Exploratory Data Analysis was performed with help of histograms, bar charts and boxplots.

Linear Regression and Random Forest Regressor were used as regression models, imported from scikit-learn library. Optimizing the hyper-parameters and selecting features with most impact on performance were performed to check if they improve their performance. They were evaluated using Mean Absolute Error, Root Mean Square Error, and r^2 score. Both models performed similarly with Linear Regression showing MAE of 36.47, RMSE of 1836.27 and r^2 score of 0.5803 and Random Forest Regressor showing MAE of 36.96, RMSE OF 1885.82 and r^2 score of 0.5690.

Linear Regression performed better than Random Forest Regressor, however it was only 0.02 difference in r^2 score metric. So, the relationship between features and target may be close to linear. In such case, Linear Regression is a sufficient model to predict the rainfall, as it is not only simpler and faster but also yields better predictions.

Introduction

The goal of the project is to predict the amount of rainfall based on slope of the land, soil saturation, vegetation cover, last earthquake activity and other geological metrics.

The dataset used on for this project is “Landslide dataset” which can be found in Kaggle, under Raju Mavinmar. It has 2000 rows and ten columns. It provides information on numerous factors influencing landslides, with Rainfall_mm as the target variable. The independent variables include slope angle, soil saturation, measured as a fraction from 0 to 1, and vegetation cover, also represented as a fraction. Other features include earthquake activity, which records the magnitude of recent earthquakes in the area, proximity to water, measured as a fraction where lower values indicate closer proximity, and soil type indicators for gravel, sand, and silt. The dataset aims to provide insights into the environmental conditions leading to rainfall events associated with landslides. Preprocessing challenges involve handling, normalizing continuous features, and engineering additional features to enhance model performance. This analysis aligns with **United Nations Sustainable Development Goals (UNSDG) 13: Climate Action** by improving rainfall prediction models, which are essential for early warning systems and preparing against disasters. By accurately predicting rainfall, this study contributes to climate adaptation strategies which helps to reduce the risk of extreme weather events, including landslides, hence supporting long-term resilience against climate-related hazards.

The objective of this analysis is to train a predictive regression model which can estimate the rainfall based on the features mentioned above in the dataset.

Methodology

Data Preprocessing

The dataset obtained was clean. None of the values were missing, all of columns had numeric values. Additionally, when scaled, the dataset showed an increase by 0.001 value only in r^2 score metric. Thus, this step was skipped for this project.

Exploratory Data Analysis (EDA)

To understand the relationship between features and target variables different plots were made, which are summarized below.

1. Histogram of Rainfall

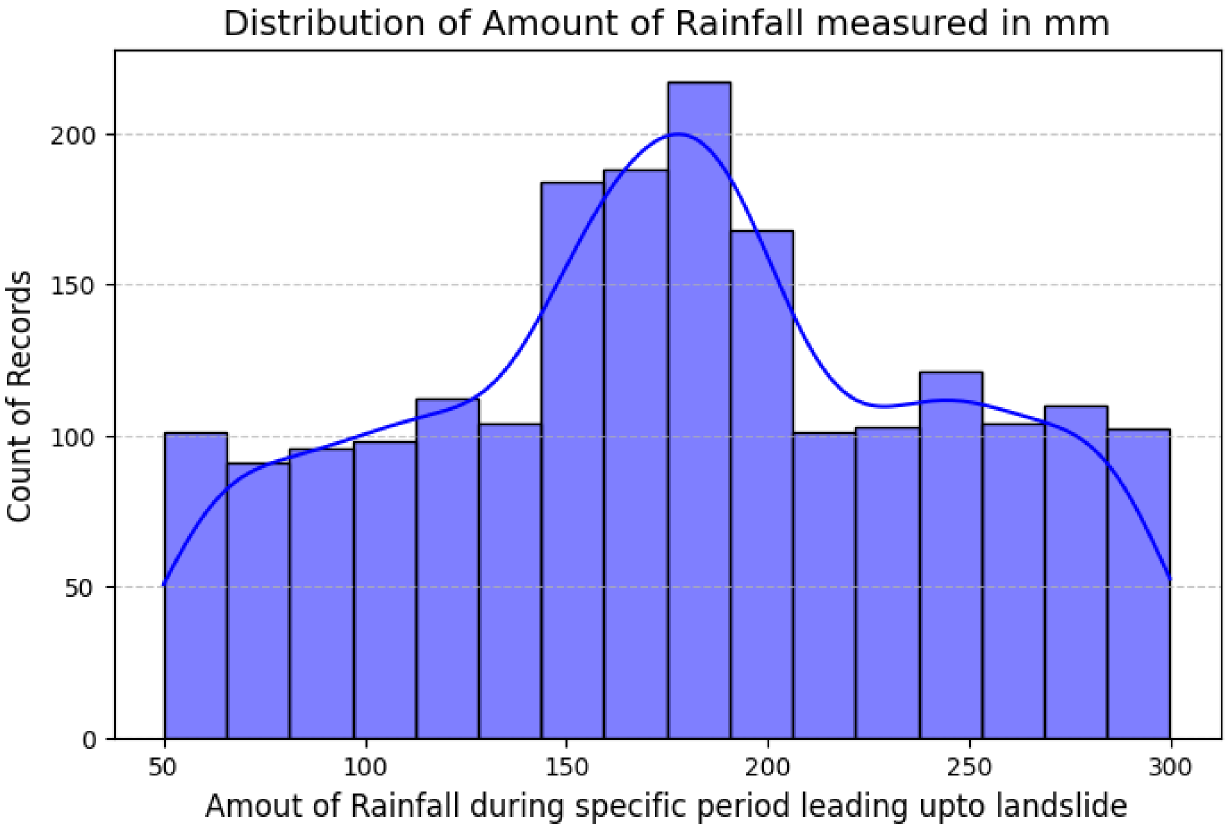

Figure 1: Distribution of Amount of Rainfall

The histogram of Rainfall distribution is unimodal and shows close bell shape with peak at 180mm, suggesting a most of the rainfall amount are of 180mm. The histogram also looks symmetrical with their mean and median values closely aligned showing low skewness.

2. Box Plot of Rainfall

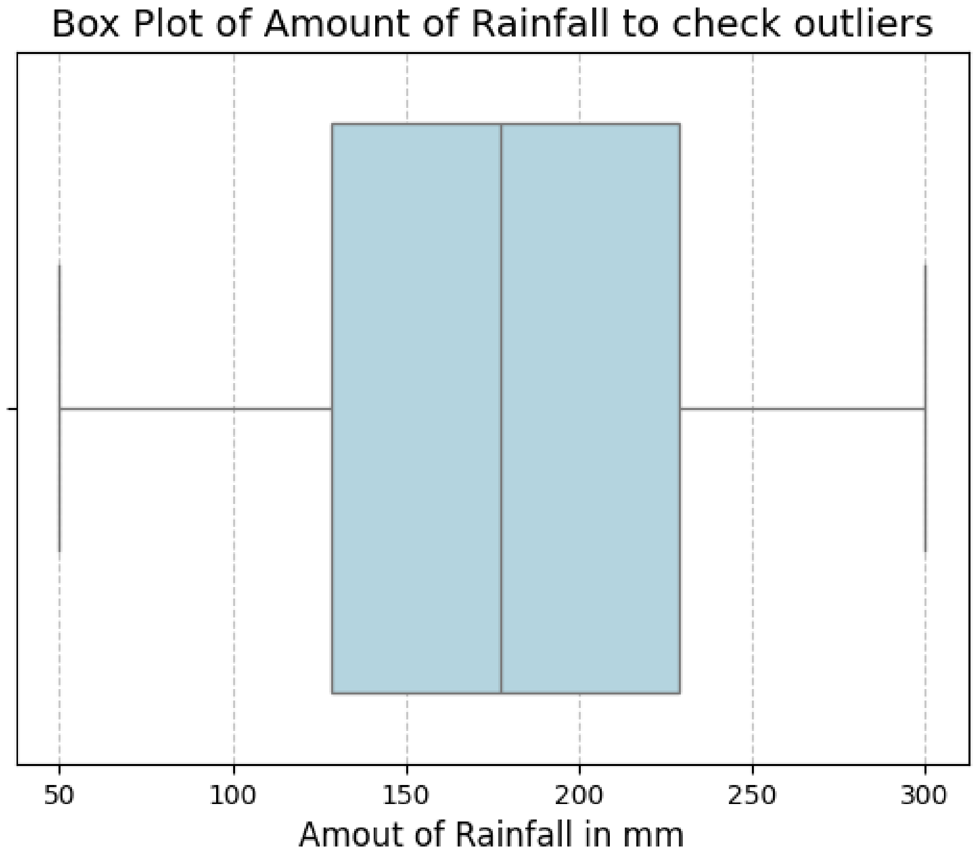

Figure 2: Box Plot of Rainfall

The Box plot shows that there are no outliers in the data at all, which might be great for predicting rainfall.

3. Landslide occurrence categorized by type of soil

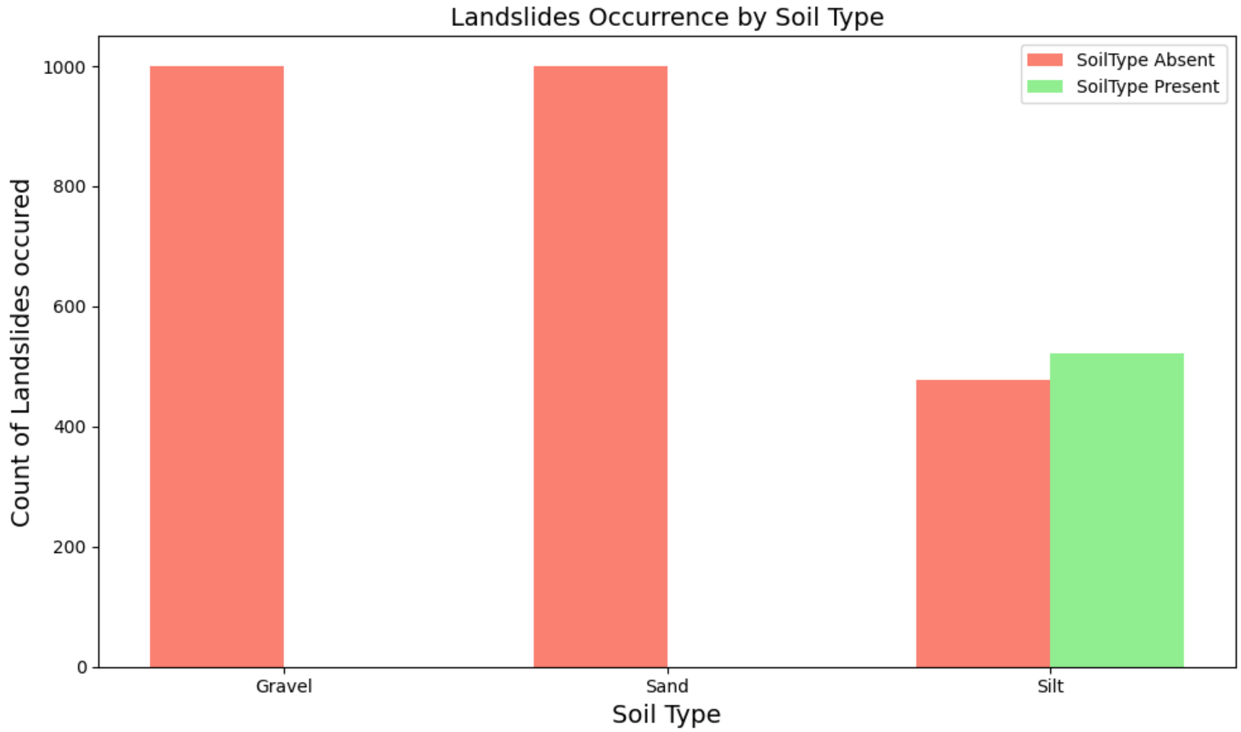

Figure 3: Categorizing type soil and landslide occurrence

The above chart shows that landslide does not occur when soil is either Gravel or Sand suggesting a correlation between them.

Model Building

Linear Regression a linear regression model and Random Forest Regressor a non-linear regression model was chosen to predict rainfall. The dataset was loaded directly from Kaggle. After removing Rainfall column from dataset and storing it in a new series, Feature Matrix and Label Vector were created. The Feature matrix and label vector were then divided into training and test parts with test part having the 20% of whole dataset, both feature and labels. The training features and labels were then used to train both Linear Regression and Random Forest Regressor.

Model Evaluation

Upon accomplishment of model training, the trained model was given test features to make predictions about rainfall. The predicted rainfall and actual rainfall separated label vector for test part were then evaluated on following metrices:

Mean Absolute Error: The average of the absolute differences between predicted and actual values.

Root Mean Squared Error: The square root of the average of the squared differences between predicted and actual values.

r^2 score: A statistical measure that represents the proportion of variance in the target variable explained by the model (features).

For Linear Regression:

MAE	RMSE	r^2 score
36.4742	1836.2768	0.5803

For Random Forest Regressor:

MAE	RMSE	r^2 score
36.9608	1885.8232	0.569

Hyper-parameter optimization

To improve the prediction values of the rainfall, hyper-parameters optimization was conducted using GridSearchCV. The optimal parameters for both modes are listed below:

Linear Regression has limited hyperparameters so optimizing only one parameter:

- fit_intercept : True

Random Forest Regression on the other hand has lots of parameters, optimizing three of them:

- max_depth : 5
- min_samples_split : 10
- n_estimators : 100

Feature Selection

Another way to improve performance of models is to avoid training them on features that do not have any significant impact on models. For this analysis, Recursive Feature Elimination was used to find six prominent features and only train on the selected features. The selected features found for both models are listed below:

Linear Regression:

Soil_Saturation

Vegetation_Cover

Proximity_to_Water

Landslide

Soil_Type_Gravel

Soil_Type_Sand

Random Forest Regression:

Slope_Angle

Soil_Saturation

Vegetation_Cover

Earthquake_Activity

Proximity_to_Water

Landslide

Conclusion

Key Findings

Despite being a non-linear model that often finds complex patterns in data, this time Random Forest Regressor did not perform significantly higher than Linear Regression, which is quite simple, faster. This could mean the feature and label have more linear relationship. Thus, using Linear Regression to build the final data with optimized parameters and selected features.

Final Model

The final model Linear Regression is then trained on Soil_Saturation, Vegetation_Cover, Proximity_to_Water, Landslide, Soil_Type_Gravel, Soil_Type_Sand features with fit_intercept set to True, the metrics obtained from this model are as follows:

MAE	RMSE	r^2 score
36.4156	1825.6453	0.5827

Challenges

Working on this project, quite a bit of challenges was faced. First challenge was finding the correct dataset. There are bunch of datasets, however most of them have large proportion of missing values and they are significantly skewed. Another challenge faced was lack of computation power. Considerable time was required to optimize the parameters. Finally, the last challenge was model not good enough to capture complex patterns within the data.

Future Task

Since both linear and non-liner models evaluate similar performance, to predict target more accurately feature transformation could be applied, applying log, square root on features with low correlation, advance linear models such as Elastic Net could be used, Optimization of even more parameters could be done.

Discussion

Model Performance

Linear Regression was evaluated using r^2 score, whose standard value for good model is 0.5 to 0.9, so this model is one the edge of capturing data patterns. The model does not underfit, as the r^2 score is not below 0.4, but also is not high to be considered a well performing model. Hence, the model performed averagely.

Impact of hyperparameter tuning and feature selection

Linear Regression does not have a wide range of hyperparameters, the one hyperparameter chosen to be optimized – fit_intercept was found to have default value True, so the optimization of hyperparameter did not affect the model's performance at all.

On the other hand, selecting six instead of nine (total number of features) decreased RMSE and increased r^2 score, thus it improved the performance of the model.

Result Interpretation

Since both Linear Regression and Random Forest Regressor showed similar performance, it shows that the relationship between feature and target is not highly non- linear. Linear Regression captures the linear relationship but its performance (r^2 score of 0.58) shows that non-linear pattens might be missing. On the other hand, Random Forest Regressor, which oversees non-linear patterns better, did not perform better than Linear Regression suggesting that features and targets do not have strong complex relationships that benefit from tree-like models.

Limitations and suggestions for future research

Despite successfully building the model, some limitations remain, these include limited data size 2000 rows may not be enough to train model well, Future research could explore more advanced regression models such as Gradient Boosting or Neural Networks to capture the complex patterns, which this model failed to. Further feature engineering, including polynomial transformation could help improve prediction accuracy.