Herald College, Kathmandu

Concepts and Technologies of AI

5CS037

Coursework Report

By: Prajwal Dahal

Student id:2406799

ANALYSIS REPORT
Regression Analysis Report

1. Overview

Objective:
The goal of this analysis is to predict a target variable using regression. The World risk repot dataset contains a variety of features to predict Vulnerability. This analysis includes data preprocessing, EDA's, model building, hyperparameter tuning, cross- validation, feature selection and model building.

2. Dataset Description

Dataset Name: World Risk Report
Source: Kaggle
Description:
The dataset chosen for this analysis is World risk report dataset which contains over 1500 rows and 12 columns. It contains features like exposure, lack of coping abilities, lack of adaptive capacities and WRI.

3.Data Preprocessing

Data Cleaning:
Handled missing values by dropping them. They were dropped because there were only few missing values. Outliers were handled by using IQR.
Feature Transformation:
Standardized numerical features using Standard Scaler and Encoded categorical variables using Label Encoder.

3. Exploratory Data Analysis(EDA)

Histograms of each column, box plots, bar charts, line graph, scatter plot, heat map were created to understand the distribution of features and detect patters.
Insights:
The correlation heatmap showed that World Risk Index(WRI) had strong relationships with Exposure.
The country which is at the highest risk is Vanuatu with WRI of 51.3 and Qatar with lowest risk with WRI of 0.31.

4. Model Building

Linear Regression from scratch: A linear regression from scratch was made.

Linear Regression from sklearn: A linear regression from sklearn was made.
Ridge Regression: A Ridge Regression was made.
Training Process:
Data was split into training and testing sets(80%/20%)
Models were trained and evaluated based on R squared, Mean squared error and Mean absolute error.

5. Model Evaluation

Performance metrices:
R-squared: Measures the proportion of variance in target variable explained by the model.
Mean Absolute Error: Quantifies the difference between predicted and actual values , with lower values indicating better model performance.
Mean squared error: The MSE measures the average squared difference between the actual and predicted values.

Model Performance:

Linear Regression from scratch:
Train MAE: 1.82
Train R2: 0.97
Train MSE: 5.49
Test MAE: 1.99
Test R2: 0.97
Test MSE: 6.66

Linear Regression using sklearn
Training MSE: 5.15
Training R2: 0.97
Training MAE: 1.71
Test MSE: 6.97
Test R2: 0.97
Test MAE: 2.00

Ridge Regression
Training MSE: 5.15
Training R2: 0.97
Training MAE: 1.71
Test MSE: 6.88
Test R2: 0.97
Test MAE: 1.98

6. Hyperparameter Tuning

GridSearchCv was applied to fine-tune the parameters of Linear Regression and Ridge Regression.

7. Feature Selection

Techniques used:
Select K best and Recursive Feature Elimination were used to select features.

8. Conclusion

Final Model:
The Ridge Regression model after hyperparameter tuning provided the best results.
Final Model Training MSE: 5.15
Final Model Training R2: 0.97
Final Model Training MAE: 1.71
Final Model Test MSE: 6.96
Final Model Test R2: 0.97
Final Model Test MAE: 1.99

Challenges:
There were issues with data quality and there were also imbalanced data. Some countries were having different names and had to fix that. The distribution of WRI was not good and also Exposure.

Classification Analysis Report

Overview

Objective:
The goal of this analysis is to predict a target variable using regression. The World risk repot dataset contains a variety of features to Exposure Category. This analysis includes data preprocessing, EDA's, model building, hyperparameter tuning, cross- validation, feature selection and model building.

2. Dataset Description

Dataset Name: World Risk Report
Source: Kaggle
Description:
The dataset chosen for this analysis is World risk report dataset which contains over 1500 rows and 12 columns. It contains features like exposure, lack of coping abilities, lack of adaptive capacities and WRI.

3.Data Preprocessing

Data Cleaning:
Handled missing values by dropping them. They were dropped because there were only few missing values. Outliers were handled by using IQR.
Feature Transformation:
Standardized numerical features using Standard Scaler and Encoded categorical variables using Label Encoder and also one hot encoder.

4. Exploratory Data Analysis(EDA)

Histograms of each column, box plots, bar charts, line graph, scatter plot, heat map were created to understand the distribution of features and detect patters.
Insights:
The correlation heatmap showed that World Risk Index(WRI) had strong relationships with Exposure.
The country which is at the highest risk is Vanuatu with WRI of 51.3 and Qatar with lowest risk with WRI of 0.31.

5. Model Building

Logistic Regression from scratch: A Logistic regression from scratch was made.
Logistic Regression from sklearn: A Logistic regression from sklearn was made.
Random forest classifier: A Random forest classifier was made.
Training Process:
Data was split into training and testing sets(80%/20%)
Models were trained and evaluated based on Accuracy score, precision score, recall score and f1 score.

6. Model Evaluation

Performance metrices:
Accuracy score: Accuracy is the proportion of correctly predicted instances
Recall score: Proportion of true positives out of all actual positive instances
Precision score: Proportion of true positives out of predicted positive instances.
F1: Harmonic mean of precision and recall

Model Performance:
Logistic Regression from scratch:
Train Accuracy: 0.88
Train Precision: 0.88
Train Recall: 0.88
Train F1 Score: 0.88
Test Accuracy: 0.87
Test Precision: 0.87
Test Recall: 0.87
Test F1 Score: 0.87
Logistic Regression using sklearn
Training Accuracy: 0.88
Training Precision: 0.88
Training Recall: 0.88
Training F1 Score: 0.88
Test Accuracy: 0.87
Test Precision: 0.87
Test Recall: 0.87
Test F1 Score: 0.87
Random Forest Classifier
Training Accuracy: 0.96
Training Precision: 0.96
Training Recall: 0.96
Training F1 Score: 0.96
Test Accuracy: 0.98
Test Precision: 0.98
Test Recall: 0.98

Test F1 Score: 0.98

7. Hyperparameter Tuning

GridSearchCv was applied to fine-tune the parameters of Logistic Regression and Random Forest Classifier.

8. Feature Selection

Techniques used:
Select K best and Recursive Feature Elimination were used to select features.

9. Conclusion

Final Model:
The Random Forest Classifier after hyperparameter tuning provided the best results.
Final Model Training Accuracy: 0.98
Final Model Training Precision: 0.98
Final Model Training Recall: 0.98
Final Model Training F1 Score: 0.98
Final Model Test Accuracy: 0.98
Final Model Test Precision: 0.98
Final Model Test Recall: 0.98
Final Model Test F1 Score: 0.98
Challenges:
Very hard to code and run all cells and running all cells takes a long time. Fixing error was the hardest part.

Limitations

The dataset has limitations in terms of geographical coverage if features. For example, some countries with limited data may not have accurate representation of exposure which could affect model predictions. While there are key features like exposure, coping abilities and adaptive capacities, there is lack of additional external features which could be economic, political could limit the predictive power of models. Logistic Regression and Random Forest Classifiers may introduce biases because of dataset having imbalanced classes and underrepresented countries.

Future Work

Exploring more on advance machine learning like neural networks and handling imbalanced data Synthetic Minority Over-sampling Technique and focusing more on feature engineering can be done.

By addressing these limitations and exploring these future work, this analysis could become useful, scalable and robust for understanding and mitigating global risk factors.