IT362: DATA SCIENCE PROJECT PHASE 3

NATURAL DISASTER PREDICTION MODEL: PREDINA

Technical Presentation

Supervised By: Dr. Mashael AlDayel

Section: 75148

TEAM MEMBERS: Nouf AlMansour - Sara AlOqiel - Shahad AlMutairi

OUTLINE

1. OVERVIEW
2. MODEL BUILDING
3. MODEL EVALUATION
4. TECHNICAL HURDLES
5. FUTURE WORK

OVERVIEW

The following slides outline the key phases and segments involved in the development of the Natural Disaster Prediction Model PREDINA, from data collection to model evaluation. Each phase consists of detailed steps designed to ensure a thorough and effective analysis, prediction, and optimization process:

Phase 1: Data Collection Research and Assessment

- Data from multiple sources (GDACS, EOSDIS, EMDAT) were integrated into a unified dataset.
- Columns from different datasets were standardized to ensure consistency across all sources (e.g., 'Disaster Group,' 'Latitude,' 'Longitude,' 'Magnitude').
- Only the relevant columns were retained for analysis, ensuring the dataset was compact and focused on the essential information.

Phase 2: Exploratory Data Analysis

- Identifying patterns, trends, and anomalies through statistical summaries and visualizations.
- Conducted multiple EDA processes relevant to our research question.
- Data Cleaning & Preprocessing:
- Comparison of Primary and Secondary Data
- Key findings that will guide further analysis and hypothesis development.

Phase 3: Modelling and Communication

- Apply models based on data characteristics and research questions (classification, regression).
- Develop a baseline model for performance comparison.
- Build and evaluate multiple models, selecting the best-performing one.
- Explore regression/classification models.
- Experiment with different algorithms and configurations to optimize performance.
- Document model selection, evaluation metrics, and challenges encountered.

MODEL BUILDING

Phase Objective:

This phase of the data science project focuses on applying suitable machine learning models answer our research questions, for instance predict the “Total Affected” and “Total Damages” from “Earthquake Magnitude” . The objective is to explore the relationship between the magnitude of earthquakes and the impact they have, evaluating model performance and refining our approach.

Task:

- Modelling Task: Selecting the right models for regression or classification based on the data.
- Baseline Model Creation: Developing a simple model to compare with more advanced models.
- Model Evaluation: Building and testing different models to find the best one.
- Clustering/Classification: Applying various algorithms for both clustering and regression tasks.

Models:

- Linear Regression
- Random Forest Regressor
- Gradient Boosting Regressor

MODEL EVALUATION

Linear Regression

Rationale:

- Baseline Model: Linear Regression serves as a simple, interpretable model that assumes a linear relationship between the input feature (Magnitude) and the target variable.
- Simplicity: It establishes a performance benchmark against more complex models.
- Interpretability: Coefficients directly show how features influence the outcome.

Model Evaluation:

- Total Affected: $R^2 = 0.1668$
- Total Damages: $R^2 = 0.0278$
- Linear Regression performed well for Total Affected but struggled for Total Damages, showing that Magnitude alone is not a strong predictor for damages.

Random Forest Regressor

Rationale:

- Ensemble Learning: It uses multiple decision trees to capture non-linear relationships.
- Robustness: Handles outliers, missing values, and complex feature interactions effectively.
- Feature Importance: Provides insight into how each feature contributes to predictions.

Model Evaluation:

- Total Affected: $R^2 = 0.2683$
- Total Damages: $R^2 = -0.0430$
- Random Forest performed better than Linear Regression for Total Affected, but its predictions for Total Damages were worse than random guessing.

Gradient Boosting Regressor

Rationale:

- Boosting Technique: Builds trees sequentially, correcting previous errors, and models complex relationships.
- Versatility: Effective on both small and large datasets and helps prevent overfitting.
- Improvement Over Random Forest: Often outperforms Random Forest with fine-tuning.

Model Evaluation:

- Total Affected: $R^2 = 0.2444$
- Total Damages: $R^2 = -0.0548$
- Gradient Boosting improved predictions for Total Affected compared to Linear Regression but still underperformed for Total Damages, similar to Random Forest.

Upgraded Models: Adding Additional Features

- To improve predictions, we included features like Development Level and Region alongside Earthquake Magnitude.
- Development Level: Countries classified as High, Medium, or Low Development.
- Region: Earthquake locations classified by region (e.g., Asia-Pacific, Americas).

- **Results from Upgraded Models:**
- Linear Regression with Magnitude + Development Level: $R^2 = 0.0515$
- Random Forest with Magnitude + Development Level: $R^2 = 0.1204$
- Gradient Boosting with Magnitude + Development Level: $R^2 = 0.1026$
- Adding Development Level showed minimal improvement, especially for Total Damages.
- Linear Regression with Magnitude + Region:
- Random Forest with Magnitude + Region: $R^2 = 0.4856$ (Best-performing model)
- Gradient Boosting with Magnitude + Region: $R^2 = 0.4794$
- Incorporating Region significantly improved predictions, particularly for Random Forest.

Conclusion

- Magnitude Alone: Limited predictive power, especially for Total Damages.
- Additional Features: Development Level and Region improved predictions, with Region offering the greatest boost.
- Model Selection: Random Forest performed best overall, especially with Region, while Gradient Boosting showed competitive results.

Next Steps

- Feature Engineering: Consider adding features like infrastructure and preparedness data to improve performance.
- Hyperparameter Tuning: Fine-tuning Random Forest and Gradient Boosting could further optimize performance.

This phase highlighted the importance of selecting the right model and features. Random Forest emerged as the best model for predicting Total Affected with Region incorporated.

TECHNICAL HURDLES

Defining Development Level:

- Development Level of regions had to be classified into High, Medium, and Low based on ISO country codes.
- This required us to manually categorize countries, which introduced subjectivity and complexity.
- Data for development levels wasn’t available directly, so we had to define and map it based on external knowledge, which could introduce inaccuracies.

Region Classification Based on Latitude/Longitude:

- Mapping earthquakes to specific regions based on geographic coordinates (latitude/longitude) was another challenge.
- Creating meaningful and geographically accurate regions required a deep understanding of the relationships between coordinates and disaster impacts, which wasn't straightforward.
- This region classification involved creating custom rules for assigning earthquakes to regions, which could potentially oversimplify complex geographical factors.

Handling Missing Data:

One of the main technical hurdles faced during feature engineering was dealing with missing or incomplete data for certain regions or countries. To address this, we split the dataset further into more manageable parts to isolate rows with missing values. By splitting the dataset, we ensured that important features weren't compromised due to missing values in unrelated columns, making the dataset more reliable and less prone to bias.

FUTURE WORK

Real-Time Data Prediction:

One of the major goals for future work is to read data in real-time to predict earthquakes based on magnitude and historical data. This would allow us to provide timely predictions and responses to natural disasters, enhancing preparedness and response strategies.

Improving Model Accuracy:

Improving accuracy is another key focus, particularly by working on datasets tailored to specific disasters (e.g., earthquakes, storms, floods). By segmenting the data based on disaster type, we can develop more specialized models that capture the unique patterns and characteristics of each disaster type.

Collaboration and Integration:

We aim to collaborate with geospatial data providers and disaster response teams to integrate external data sources, enhancing our ability to predict and understand disaster impacts across different regions and types of disasters.