Gauging Migration in Bangladesh

Ishaan Babbar, Liz Kovalchuk, Tiana Le, Sheeba Moghal

Abstract

This project uses retrospective migration data from the research of Amanda R. Carrico and Katharine Donato [1] to explore the relationship between extreme weather occurrences and internal migration trends in southwest Bangladesh. Carrico and Donato issued extensive surveys with questions over various climatic and socio economic factors in the country. Their findings showed a relationship between extreme weather conditions in the region and patterns of migration (particularly after dry spells, albeit also after warm spells and above average rainfall but to a lesser extent). While our work showed similar conclusions to Carrico and Donato's publication, our analysis investigates beyond the primary author's conclusion associating weather with increased migration.

Using regression and classification analysis showed that internal migration could be predicted by socio-economic factors such as occupation, wages in taka, and education attainment. The most effective approaches for the two analyzed datasets were ensemble models: After utilizing 10 models for each task, this dataset demonstrates the power of ensemble modeling, where bagging and boosting techniques.

Overall, this project reveals the rise of "climate displacement" in Bangladesh. The survey data and subsequent analysis shed light on the intricate relationships between migration and climate change in vulnerable areas of Bangladesh; increasing climate effects are forcing rural residents to migrate to metropolitan areas in search of alternate means of subsistence. These migratory movements put an additional burden on the resources and infrastructure of cities, aggravating pre-existing socio economic issues.

Introduction

Bangladesh is a nation renowned for its rich history, charming scenery, and vibrant culture. Located in the center of South Asia and surrounded by the Bay of Bengal, India, and Myanmar, this country is home to the eighth largest population in the world [2] and is one of the most densely populated areas. The nation's economy is expanding quickly due to its strong textile and agricultural sectors, and it is a major producer of textiles worldwide [3]. Bangladesh is noteworthy for its remarkable progress in social development metrics including health and education, even in the face of obstacles like urbanization and natural disasters.

In Bangladesh, the intertwined challenges of climate change and migration have become their own term: *climate displacement*. As climate change intensifies, rising sea levels and erratic weather patterns wreak havoc on the livelihoods of millions, particularly affecting those dependent on agriculture. Migratory workers, often from rural areas, find themselves compelled to leave their homes in search of alternative means of sustenance as their traditional occupations become untenable. These migrations, spurred by environmental degradation, exacerbate existing socio-economic strains, leading to overcrowded urban centers and strained resources.

Bangladesh has a long and complex history of migration due to its partition from India and Pakistan, with a rush of external migration [4]. Due to markers of socio-economic growth, limited job opportunities with a greater percentage of lower economic class, there has been a greater movement for employment within the Middle East; particularly within Saudi Arabia for socioeconomic reasons. Labour migration, like other types of human mobility, is caused by a variety of complex and partially interlinked factors or 'drivers' that motivate individuals to relocate. The effects of climate change have become increasingly important in this region [5]. Hence, as a nation it is grappling with ongoing efforts to mitigate the impacts of climate change. Despite knowledge of climate related impacts, recent surges in internal migration are complicated by socio economic factors that make it difficult to isolate environmental pressures and ongoing socio-economic ones [7].

Bangladesh has had several recent natural disasters, including storm surges, droughts, and tropical cyclones, these occurrences have had a significant influence on millions of people's lives. Such environmental problems result in a great deal of ongoing displacement [8]. According to estimates from the International Labour Organisation, seven people from Bangladesh will be relocated internationally for every 45 persons displaced by climate change by the year 2050 [6].

This high rate of displacement serves as a stark reminder of how seriously threatened by climate change the country is Bangladesh is among the world's most climate-vulnerable nations due to its location and high population, which further exacerbates the problem. The ensuing migration presents significant obstacles for communities and policymakers alike, both domestically and abroad. [8]

This project takes a quantitative approach to understand the internal migration through the migration data collected through surveys in southwest Bangladesh. Using [1] as a base, this project delves deeper into the relationship between the extreme weather conditions and the increased migration patterns, extending it to understand the socio-economic factors' prevalent association. The project explores three datasets collected through root level survey analysis with each exploring climatic, socio-economic and migration based aspects.

Literature Review

Though internal migration is still undervalued in public policy, it is becoming more widely acknowledged as an adaptive approach for communities at risk from environmental change. [9]. The traditional legal and policy frameworks used to manage migration are partly to blame for this omission; they are primarily intended for economic migrants and refugees, not for people who are uprooted due to natural disasters. This has added to the controversial nature of the connection between migration and climate change in academic and political debate [10]. Furthermore, not many regulations currently in place successfully handle this intersection [11]. The National Adaptation Programme of Action (NAPA), for example, frames internal displacement (IDP) as a result of the consequences on livelihoods through changes in employment, income, and consumption and sees migration as an unwanted outcome of climate change [12].

In Bangladesh, communities experiencing hardship have historically turned to migration as an adaptive option, especially when it comes to ensuring livelihoods. People who live in impoverished rural areas frequently migrate in circles to urban centers, which is an essential livelihood strategy [13]. Communities can better handle temporary vulnerabilities brought on by climatic stress and other disruptions when they have this kind of short-term migration. The influx of remittances back to the communities of origin is a major benefit of migration, since it enhances adaptive ability in several ways. Remittances support a variety of livelihood shocks, including droughts, by preserving access to basic necessities like food, financing the acquisition of human, social, physical, and natural capital, and increasing demand locally, which in turn boosts production [14].

Data collection and analysis can aid in addressing some of these issues by clarifying the difficulties faced by these migrants and the extent of their migration, which in turn opens up possibilities for suggesting areas for support and more research.

Data and Methodology

Data

The data for this project was collected through surveys, leveraged from the work of Carrico and Donato [1]. These extensive surveys are broken down into three individual datasets trying to focus on socioeconomic, household data and on the climatic aspects for internal migration. Each of the three cleaned sets are below; the donut charts are meant to show how much of the information provided was categorical, driving our methodology and choice of instruments.

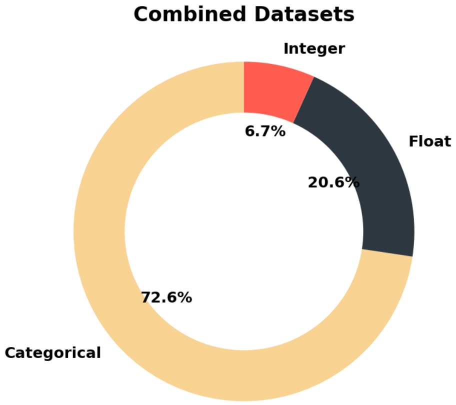

Figure 1: The data type representation of combined datasets

Each dataset held survey question responses, often categorically or binary encodings. The integer and float values are often of identifying information, not of numeric values. Each set was used to investigate climate change's impact on Bangladesh migratory workers.

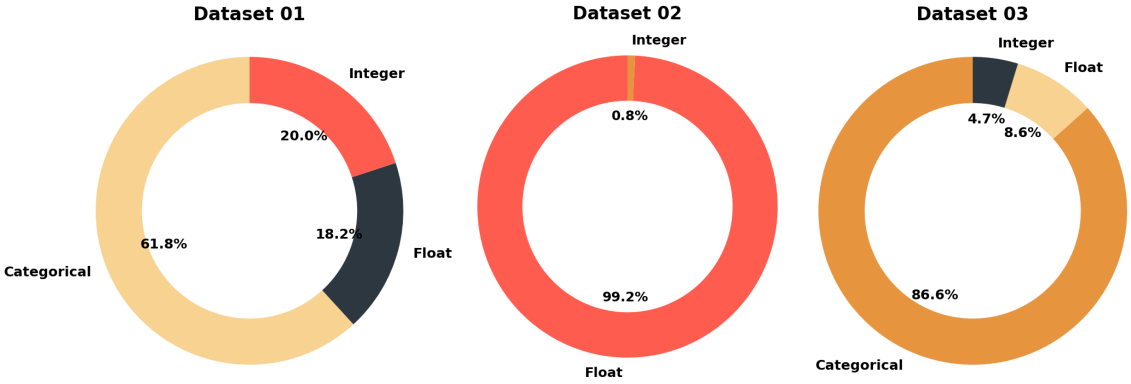

Figure 2: Individual data types of each dataset

Methodology

Data Exploration

For the data exploration of [1] in terms of the three datasets, there is a presence of more than thousand columns for each of the datasets, with more than 70% missing values for the majority of the columns. Acknowledging that data from the primary source can be subject to no response bias, sensitivity of the questions, the survey design and cultural design, most of the data was unanswered. To better aid our objective, the variables with more than 35% of missing data have been dropped and have been subset to variables that are required for the analysis. In terms of all the three datasets, the first and the second have been used for modeling, whereas the third dataset was not used.

Dataset 1

The first dataset primarily focused on understanding the impact of environmental change on migratory workers. Due to a huge set of columns, feature selection was performed to choose the relevant columns in terms of both numerical and categorical features. The most relevant values based on understanding the environmental change have been identified using ANOVA using f-test for numerical features and Chi-squared test based on the p-values of the categorical variables. We performed analysis with a p value less than 0.05 for significance level.

Below is what the first dataset looked specifically for with and without feature selection.

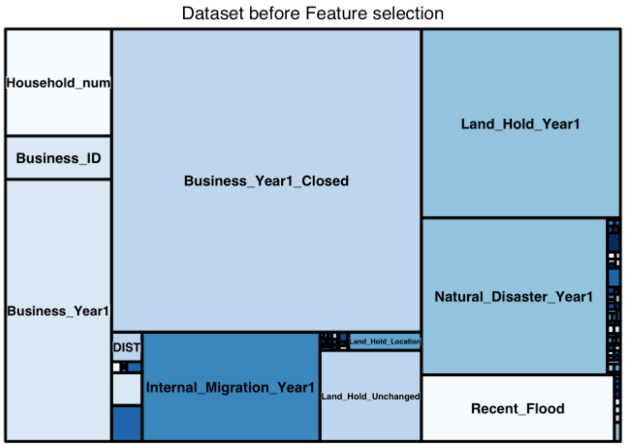

Figure 3: Data set before feature selection

As seen, it was difficult to prioritize the important factors due to potential multicollinearity and columns that do not contribute to the target variable. Using feature selection helped reduce the number of variables being looked at and streamline the dataset to have a specific amount to focus on in the EDA.

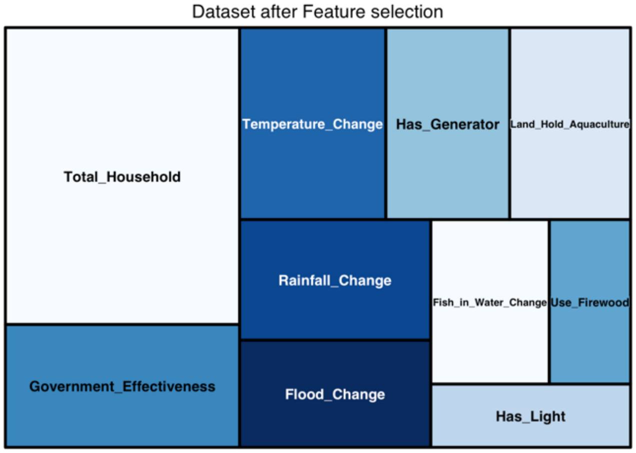

Figure 4: Dataset after feature selection

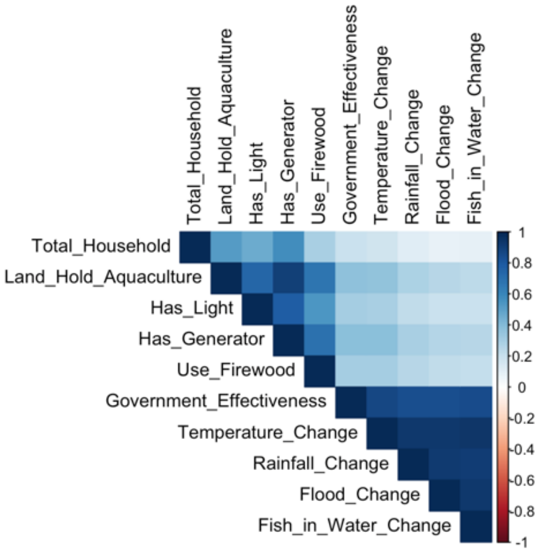

Figure 5: Heat map after feature selection

In the first EDA analysis, a correlation matrix is best to help identify which areas in our subset data are mostly correlated. Based on the correlation matrix above, it was determined that questions such as *Temperature_Change*, *Rainfall_Change*, *Flood_change* have a high correlation with each other. The *Temperature_change* questions in the survey focus on whether certain environmental changes were a reason for migration such as temperature, flooding, and amount of rainfall.

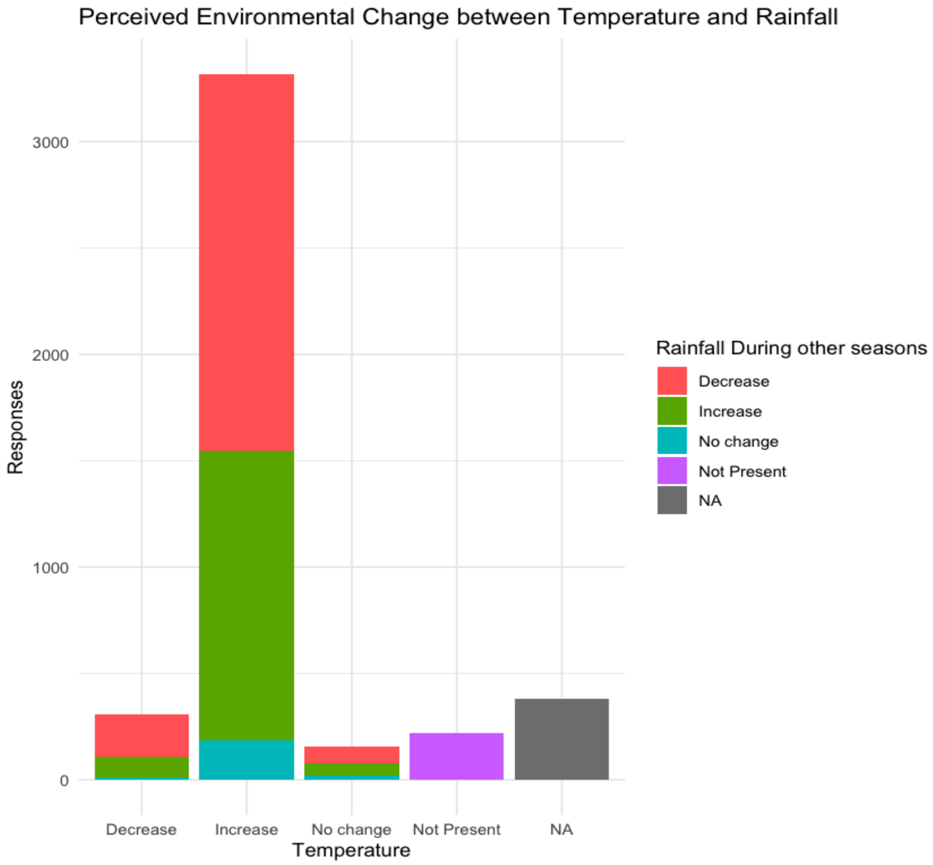

Figure 6: Perceived Impacts of Temperature and Rainfall on Environmental Change

Following the correlation analysis, a bar plot is used to compare the perceived impacts of temperature and rainfall in the environment to see if they have a similar amount of impact on migration. Temperature increase can lead to both decreased/increased amounts of rainfall depending on the environment. Increased temperature, for example, can lead to drought which can prevent rainfall from occurring, causing a decrease and can make people consider migration for better circumstances.

This plot shows that an increase in temperature had the highest frequency response; when comparing it to the amount of rainfall during other seasons, most survey results were split between decrease and increase being a reason for migration. Thus, while temperature increase is a major factor for migration, most respondents were split between rainfall decrease or increase as a reason for the same. Contextually, as far as shifting agricultural or other livelihood factors, this result seems logical. This points out areas for future research better identifying climate fluctuations, as Carrico and Donato were largely unable to find meteoric and climate information to support more than perceived climate changes.

Dataset 2

For dataset 2, the target variable is the primary purpose of the trip was to work or earn money as a classification analysis, around 154 identifiers have been used to find patterns in source of work and the purpose of migration in the survey. The value of 1 denotes “yes” that the primary purpose of migration was to work or earn money while 0 denotes “no”. This question was answered by 2000 participants. Therefore, the survey was subset to these 2000 participants and the questions that these participants answered. The questions that had a mean value of 0 were dropped. The remaining variables included categorical and numerical variables. The categorical variables were converted to binary or dummy variables in preparation for classification methods.

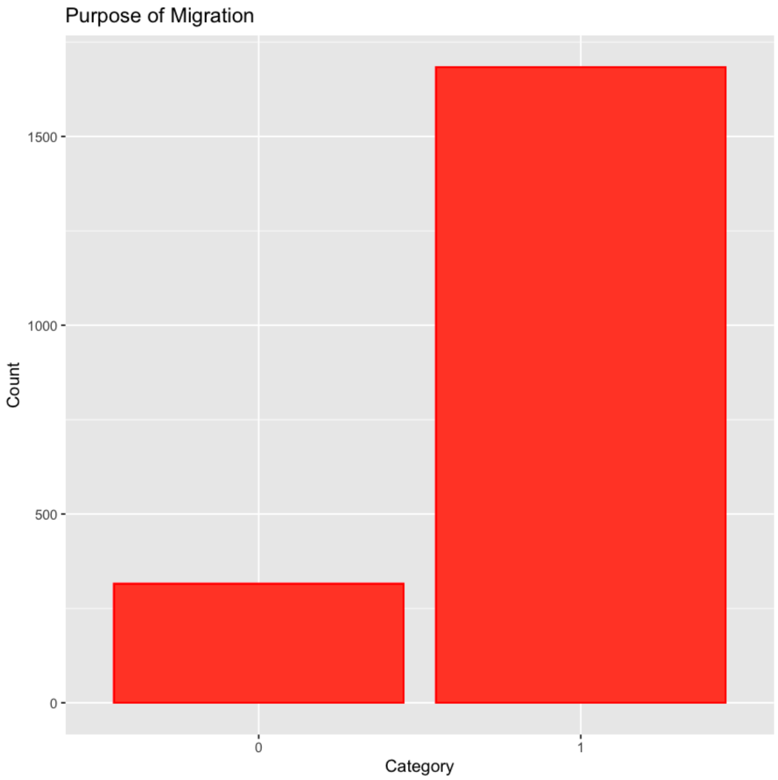

Figure 7 : Purpose of Migration

After performing logistic regression as one of the methods to find the significant features, 13 statistically significant features were identified as the predictors for the logistic regression model for classification. The logistic regression base model showed that the total number of trips of the head of household, rent per month, and average monthly remittances sent home were predictors of the target question (did survey respondents migrate for work or to earn money). There was also a relationship to “wage” in taka. Duration of stay, level of education (at five different levels), respondents’ livelihood/occupation, were also predictors of migration. These features were used for further modeling. Beginning with logistic regression, these statistically significant features from the base logistic regression were repeated in another logistic regression model. The most statistically significant feature in logistic regression was average monthly remittances, duration of stay, and education level 8 (at the university level).

Machine learning models were also created using ensemble methods, including Decision Trees, Random Forest, and XGBoost. Ensemble methods balance different combinations of features and the order of these features in their methods. Ensemble methods handle complex relationships compared to logistic regression that tries to assume linear relationships. The ensemble methods such random forests and gradient boosting, for instance capture intricate non-linear correlations, are frequently used to pick features that are distinct from those selected by logistic regression. These techniques ignore unnecessary information and rank features according to how well they contribute to prediction accuracy. Logistic regression, on the other hand, calculates coefficients for every feature, possibly giving significance to unimportant or the non-linear features. The feature importance metrics of ensemble methods further improve their capacity to identify significant features, resulting in different feature selections. Leveraging the approach of ensemble methods, dataset 2 without the feature selection was performed.

Each of the ensemble methods showed different variables of importance. In the Decision Tree Model, the most important features were paid in taka, wage of the last head of household, and duration of stay.

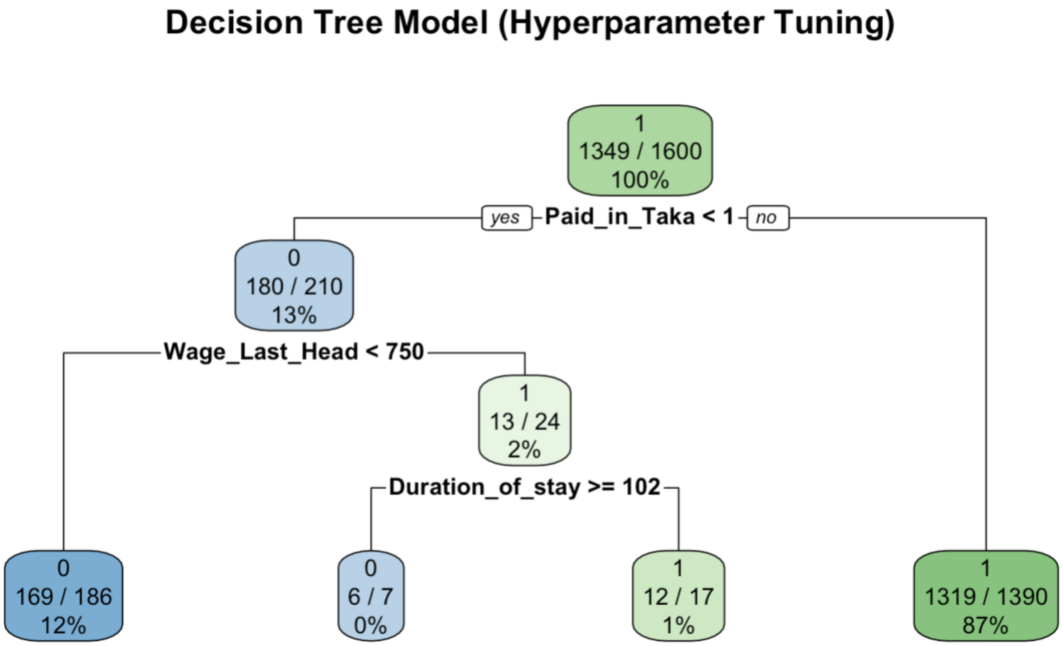

Figure 8: Decision Tree Plot Analysis for Feature Selection

In the Variance Importance Plot for Random Forest, the most important features in the model according to mean decrease accuracy were paid in taka, wage of last head of household, and livelihood occupation 17, which was homemaker.

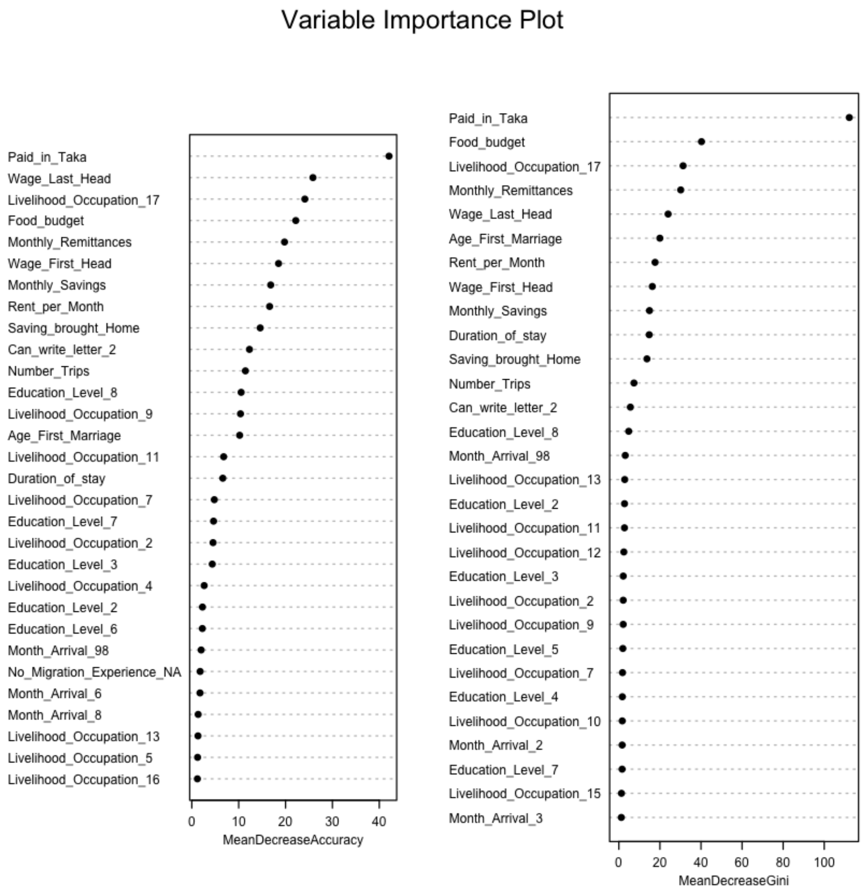

Figure 9: Variable Importance Plot

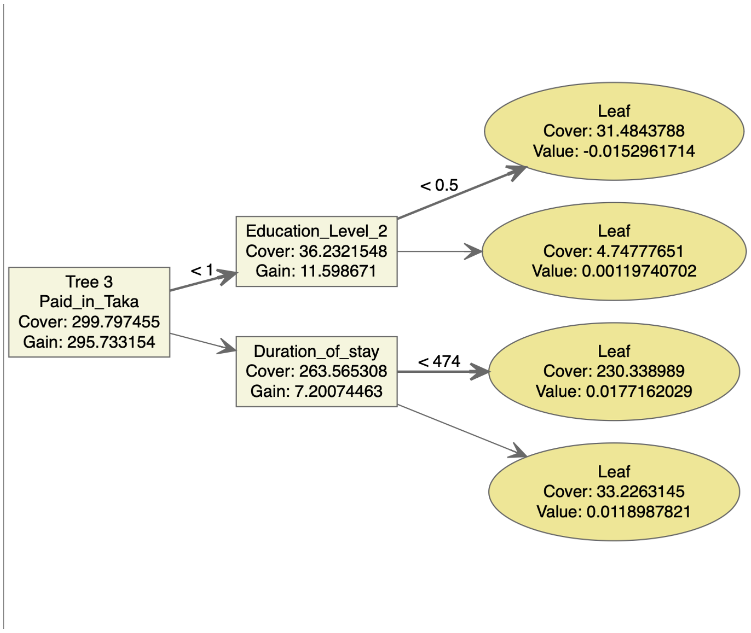

Figure 10: XG Boost Tree Plot

Lastly, in the XGBoost model, the most important features are paid in taka, duration of stay, and education level 2, which is Class I-IV.

In comparison to the logistic regression model with feature selection, the ensemble methods had different results in variable importance. The ensemble methods were leveraged to handle the complexity and non-linear relationships in the data providing insights about the other important variables collected in the survey. Among all the important variables, variables related to income remained the most important in all the models.

Data Analysis

Dataset 1

As the first dataset focuses on the regression analysis, with perceived environmental change as the target variable, using ANOVA and Chi-Square test, we subsetted the data based on relevant variables that aid to the analysis. The metrics used for this analysis are MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), R-Squared (R-Squared), and AIC (Akaike information criterion) and BIC (Bayesian information criterion).

The modeling was performed using the regression methods, bagging and the boosting techniques (ensemble methods). All the variables selected during the feature selection were used. The following is the performance metric table below.

Model	RMSE	MAE	R-Squared	AIC	BIC
Base Linear Regression	0.389	0.268	0.953	854.37	902.12
Linear Regression	0.389	0.269	0.953	847.80	876.46
Ridge Regression	0.407	0.283	0.950	-	-
Lasso Regression	0.404	0.302	0.952	-	-
Elastic Net Regression	0.406	0.300	0.952	-	-
Decision Tree (Base)	0.315	0.127	0.969	-	-
Random Forest (Base)	0.320	0.142	0.968	-	-
Decision Tree (Hyperparameter Tuned)	0.315	0.127	0.969	-	-
Random Forest (Hyperparameter Tuned)	0.317	0.134	0.969	-	-
Gradient Boosting	0.317	0.136	0.969	-	-
XG Boost	0.334	0.163	0.965	-	-

Stacked Model (GLM + Random Forest)	0.315	0.150	0.969	-	-

Table 1: Performance Metrics for Dataset 1

Model performance metrics offer a way to mathematically compare scores in order for the contextually informed analyst to make a judgment call about the “best” or most effective model for their purposes. When comparing the quality of fit and model complexity, the model with the lowest RMSE and MAE, in addition to highest R-squared are frequently regarded as the best model - these metrics take a balanced approach to complexity management and model accuracy. With MAE and R-Squared values of 0.127 and 0.969, the "Decision Tree (Base)" and "Decision Tree (Hyperparameter Tuned)" models have the best values in this instance, showing goodness of fit performance. Though AIC is computable for the linear regression model and not for the other non-parametric and regularization models, it is still added for the performance. It is seen that in comparison to base Linear Regression and Linear Regression with significant variables taken as inputs, the latter performs best. However, these error metrics cannot be applied to the other model types.

The following metrics are shown visually below, where we see that Decision Tree’s based and tuned model both have higher R-Squared values. RMSE, MAE, and R-Squared can often be the closest common metric across many different model types. Though AIC and BIC are better indicators for the best suited models where parameters are estimated via likelihood methods, typically in more traditional statistical models like linear and logistic regression models. For models incorporating regularization (like Ridge, Lasso and Elastic Net) or non-parametric models (like Decision Trees and Random Forests), these criteria are generally not applicable unless approximations or modifications to the original criteria are used. Hence, although we decided to use this as a method of comparison, it will lead to biased outcome.

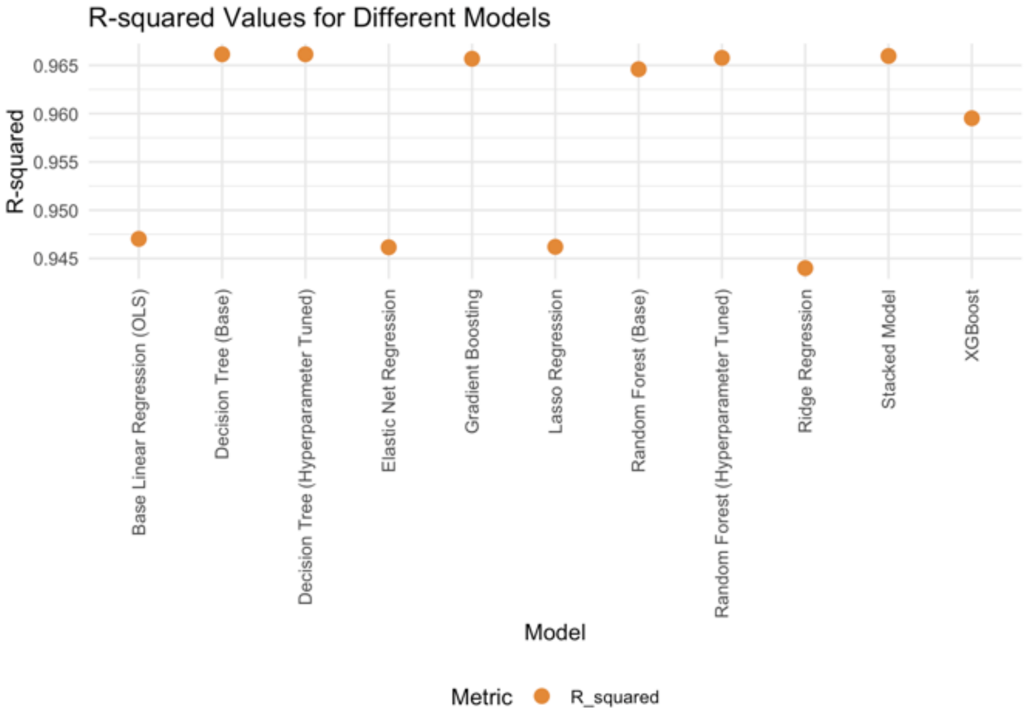

Figure 11: R-Squared Values for Different Models

It is also seen from Figure 8, the ensemble methods perform better than the regression models.

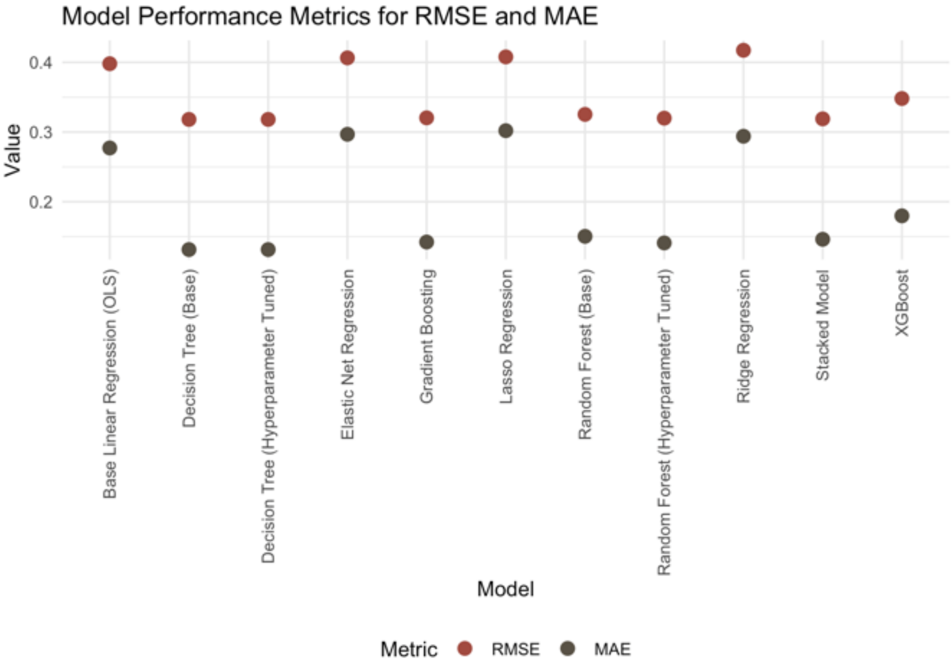

Figure 12: RMSE and MAE Performance Metrics

The Decision Trees, Gradient Boosting, Random Forests, and Stacked Model all performed well. The R squared values are all above 0.90, indicating a strong positive relationship. RMSE for these values were approximately 0.3 or below. Their MAE are also closer to zero, showing a lower absolute error for the model's predictions after training. Therefore, the Decision Tree (Base) is the best choice for predictive modeling for the first dataset. For future prediction of migration of workers as a relationship to temperature, using a decision tree or similar model should be effective.

Dataset 2

For the classification model, the survey data was subset to include the heads of households with the primary purpose of migration for work or to earn money.

The features that were selected for the final logistic regression model were repeated in other modeling. These features predicted the primary purpose of the trip was to work or earn money, which is a binary classification target variable. These predictors include total number of trips, rent per month, average monthly remittances, monthly wage, levels of education, and occupations, including rickshaw driver, construction worker, non-agricultural workers (factory worker, blue collar service), homemaker, and household education level. Feature selection was not performed on ensemble techniques to understand if similar features have been mapped or not.

The following table showcases the performance metrics. While it is challenging to declare a single "best" model, balanced performance across all metrics is a positive indicator of a useful model. From the provided list of models below, "XG Boost" appears to perform well across most metrics, achieving the highest values in specificity, accuracy, sensitivities, precision, AUC, and F1 score. Based on these metrics, "XG Boost" was chosen as the best performing model for the second data set.

	Specificity	Accuracy	Sensitivity	Precision	AUC	F1 Score	AIC	BIC
Logistic Regression	0.45	0.89	0.97	0.88	0.82	0.92	918	993
Ridge	0.67	0.89	0.91	0.95	0.88	0.93	-	-
Lasso	0.67	0.89	0.91	0.96	0.88	0.93	-	-
Elastic	0.67	0.89	0.91	0.95	0.89	0.93	-	-
Decision Tree (Base)	0.66	0.94	0.99	0.93	0.83	0.95	-	-
Decision Tree (w/ tuning)	0.66	0.94	0.99	0.93	0.89	0.95	-	-
Random Forest	0.68	0.94	0.98	0.94	0.94	0.95	-	-
XG Boost	0.78	0.94	0.97	0.95	0.98	0.96	-	-

Table 2: Performance metrics for dataset 2

The performance of XG Boost as the best metric is also shown through the ROC curve below as XG boost's value is closer to 1.

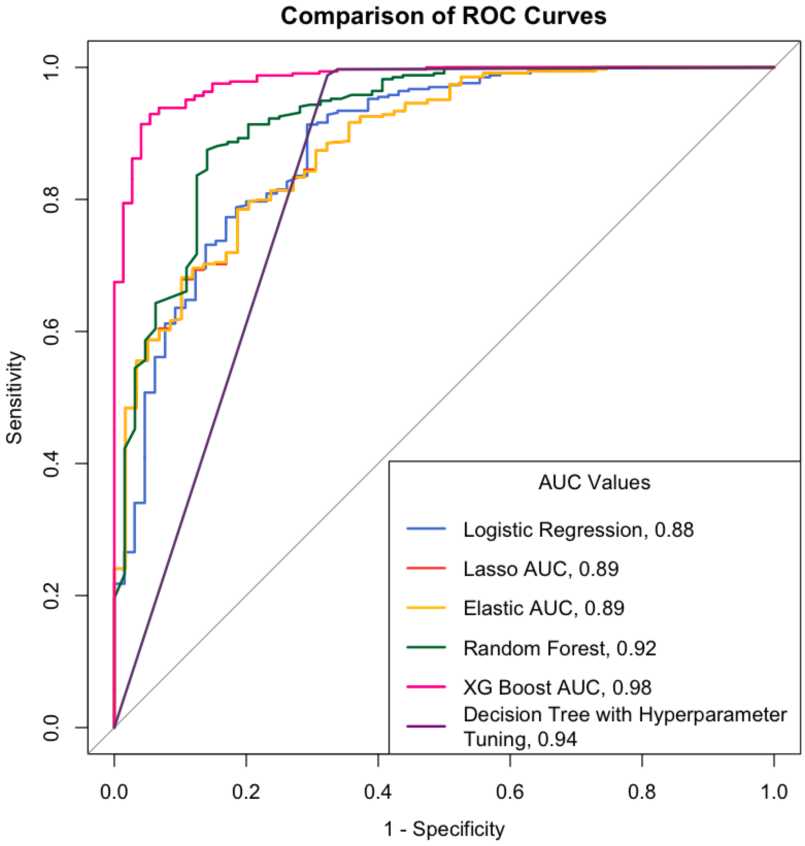

Figure 13: ROC Curve for all the models

The ROC curve provides additional insight into the discriminative power of different modeling techniques, especially between logistic feature selection and ensemble methods. While the logistic regression demonstrated a reasonable performance, its ROC curve suggests limitations in distinguishing between true positive and false positive rates, as indicated by its lower AUC value compared to ensemble methods like XG Boost. The steep rise and high AUC value of XG Boost's ROC curve signify its ability to achieve higher true positive rates while minimizing false positive rates across various decision thresholds. This superior discrimination capability stems from the ensemble nature of XG Boost, which integrates multiple weak learners to form a robust predictive model.

Unlike logistic feature selection, which relies on linear relationships between features and the target variable, ensemble methods can capture complex nonlinear patterns and interactions, thereby enhancing predictive performance. Ensemble methods are less susceptible to overfitting due to their built-in regularization techniques and ability to generalize well to unseen data: thus, the ROC curve analysis underscores the advantage of ensemble methods over logistic feature selection in terms of discriminative power and predictive accuracy.

Dataset 3

The third survey's unique questionnaire data contained health information about respondents' diagnosis, their visits to physicians and other forms of seeking medical care, or recurring symptoms. Many respondents did not report such data, and gauging the status of "health" before and after migration experiences was ephemeral to categorize within the survey. As a result, much of this data was discarded to focus on the primary modeling from the first and second datasets.

Results and Discussion

Our models validated many of Carrico and Donato's findings: that socioeconomic factors like wage and education level are good predictors of migratory patterns as climate change has large impacts on the region of Bangladesh. Our models used primary factors, such as variations in temperature and altered precipitation patterns, in order to better understand and ultimately predict likelihood of migration. In the broader scope of understanding this region and climate change, these conclusions are unsurprising but are now validated with data and with instruments that can aid in our understanding of how to potentially mitigate the hardships of a shifting world order to provide resources to citizens.

Rainfall variations and increasing flood levels have a direct effect on agriculture and living situations, forcing people to relocate in pursuit of better living conditions. When taken as a whole, these components highlight the vital roles that resource dependence, household adaptation, and economic stability play in reaction to changes in the environment. These points draw attention to the complex ways that migratory dynamics in places like Bangladesh are impacted by climate change. The degree to which local governance is able to mitigate the effects of climate change can also have a substantial impact on the migration decisions made by families, since well-managed communities may reduce the need for relocation.

In our analysis of the second survey dataset, the selected variables provide a comprehensive perspective on internal migration in Bangladesh, highlighting migration motives, educational attainment, and patterns of migration. A migration trip's main objective, to work to earn money, emphasizes economic incentives and may suggest destinations with job prospects or economic disparities. Rent and remittances highlight domestic economic issues, which stimulate migration as a way to improve financial circumstances—key reasons for importance based on the second dataset. Wage levels and earning potential are closely associated, which is one of the main factors influencing migration. Education affects employment prospects, and specific jobs—like building or rickshaw driving—often correlate with job opportunities in cities, which promotes migration. The decision to relocate for economic development is influenced by each of these variables.

Conclusion and Further Research

In terms of our analysis, which attempts to understand the association between climatic and socioeconomics factors corresponding to internal migration, it was interesting to find interesting patterns within the primary survey data used. Our work showed patterns in Bangladesh for migration; further research would be to follow migration patterns in similar demographic or neighboring countries. As climate change and other economic challenges occur as a result of rising temperatures, following migratory patterns of individuals and understanding why they choose to relocate or are relocated is likely to become more and more relevant. Additionally, our project also found prominent patterns in understanding the economic motivations as primary players for internal migration in Bangladesh. Both the individual analysis do highlight the importance of government intervention and how it would aid the people less vulnerable to these changes.

Therefore, it is crucial that academics and policymakers take into account a comprehensive strategy that integrates socioeconomic development with strong environmental and educational regulations in order to handle these complex issues. To increase resilience and make it possible for communities to adjust to environmental changes, it is imperative to strengthen local infrastructure and governance. Targeted initiatives should also improve employment and education prospects in order to lessen economic inequality and better control migration.

References

[1] Donato, K. & A.R. (2019). Extreme weather and migration: evidence from Bangladesh. Population and Environment, 41, 1–31. https://doi.org/10.1007/s11111-019-00322-9

[2] Central Intelligence Agency. (n.d.). Bangladesh. In The World Factbook. Retrieved from https://www.cia.gov/the-world-factbook/countries/bangladesh/

[3] Trade.gov. (n.d.). Bangladesh: Textiles and apparel. Retrieved from https://www.trade.gov/country-commercial-guides/bangladesh-textiles-and-apparel

[4] Sarwar, S. (2022). Tracing the Impact of Migration in Bangladesh: From Partition to the Pandemic. International Journal on Responsibility, 5(2). https://doi.org/10.62365/2576-0955.1066. Available at: https://commons.lib.jmu.edu/ijr/vol5/iss2/4

[5] Momtaz, S., & Shameem, M. (2015). Experiencing climate change in Bangladesh: Vulnerability and adaptation in coastal regions. London: Academic Press.

[6] International Labour Organization. (n.d.). [PDF file]. Retrieved from https://www.ilo.org/wcmsp5/groups/public/---asia/---ro-bangkok/---ilo-dhaka/documents/briefingnote/wcms_882280.pdf

[7] Kartiki, K. (2011). Climate change and migration: A case study from rural Bangladesh. Gender and Development, 19(1), 23–38.

[8] Domínguez Mujica, J. A. (2016). Global change and human mobility. Singapore: Springer.

[9] Martin, S. F. (2017). Environmental change and human mobility: Trends, law and policy. Comparative Population Studies, 42. https://doi.org/10.12765/CPoS-2017-13en

[10] Kothari, U. (2014). Political discourses of climate change and migration: Resettlement policies in the Maldives. The Geographical Journal, 180(2), 130–140.

[11] Vlassopoulos, C. A. (2013). Defining environmental migration in the climate change era: Problem, consequence, or solution? In T. Faist & J. Schade (Eds.), Disentangling migration and climate change: Methodologies, political discourses and human rights (pp. 145–164). Dordrecht: Springer.

[12] Government of the People's Republic of Bangladesh (2005). Retrieved from [source needed]

[13] Kniveton, D. R., Smith, C. D., & Black, R. (2012). Emerging migration flows in a changing climate in dryland Africa. Nature Climate Change, 2(6), 444–447.

[14] Pairama, J., & Le Dé, L. (2018). Remittances for disaster risk management: Perspectives from Pacific Island migrants living in New Zealand. International Journal of Disaster Risk Science, 9(3), 331–343.

[15] Naser, M. M., Swapan, M. S. H., Ahsan, R., Afroz, T., & Ahmed, S. (2019). Climate change, migration and human rights in Bangladesh: Perspectives on governance. *Asia Pacific Viewpoint*, *60*(2), 175-190.