Capstone: Predicting a Climate Refugee Crisis

Robert Meyer, Felicia Liu, Brendan Foo, Leanna Chraghchian

Roadmap

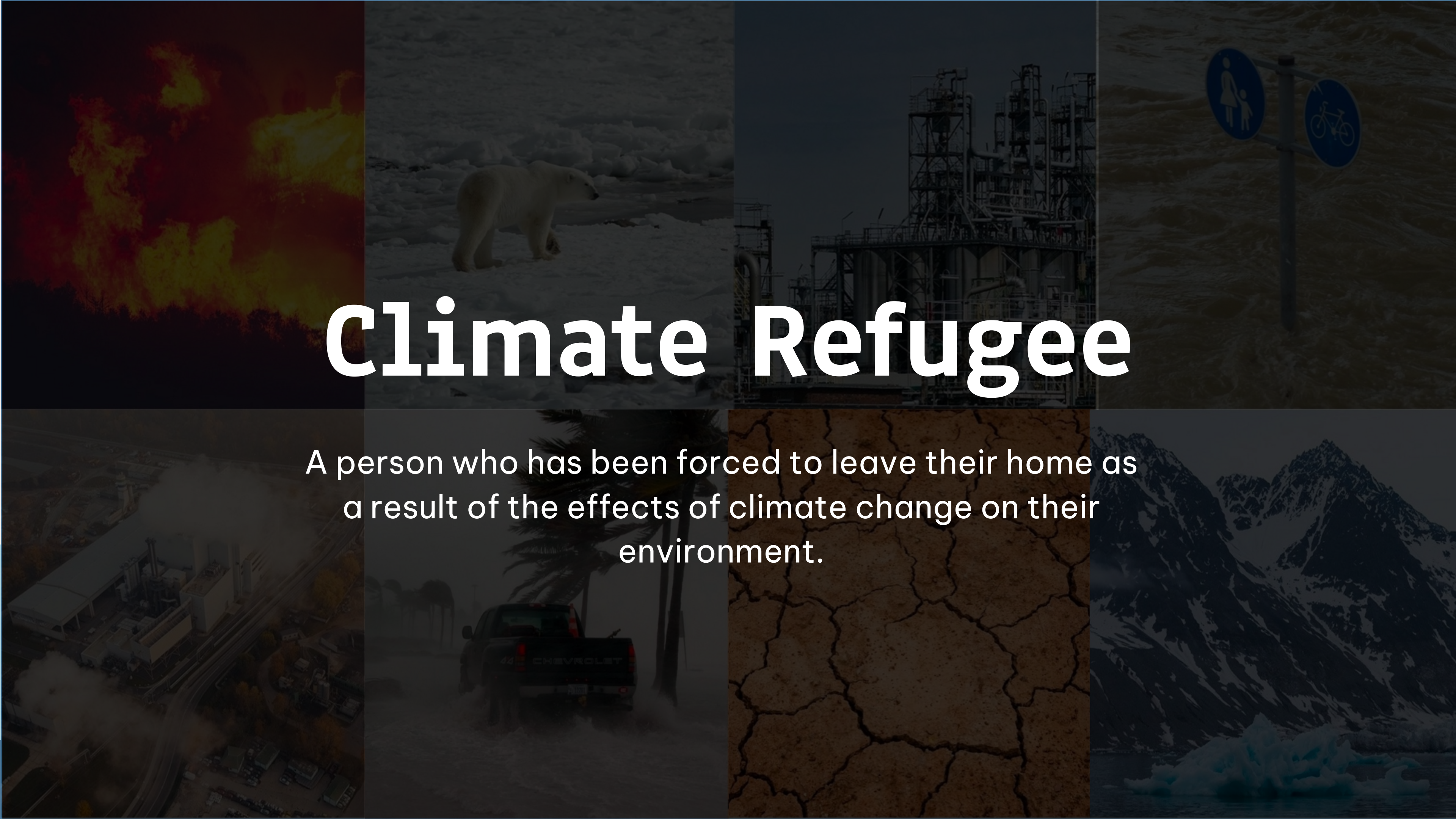
01 Introduction

02 Demo

03 Data and Modeling

04 Wrap-Up


CHEVROLET



Climate Refugee

A person who has been forced to leave their home as a result of the effects of climate change on their environment.

Internally Displaced Person (IDP)

Persons who have been forced to leave their homes as a result of violence, violations of human rights or natural or human-made disasters, and who have not crossed an internationally recognized state border.

The Problem

~3.5 Billion

people were living in countries with a high vulnerability to climate change in 2022

31.8 Million

people were internally displaced within the borders of their country due to weather-related hazards in 2022

60%

of the 31.8 million displacements were a result of floods in 2022

The Problem

31.8 Million

people were internally displaced within the borders of their country due to weather-related hazards in 2022

60%

of the 31.8 million displacements were a result of floods in 2022

1.2 Billion

people are predicted to be displaced globally by 2050 due to climate change and natural disasters. This is about 15% of the current world poplation.

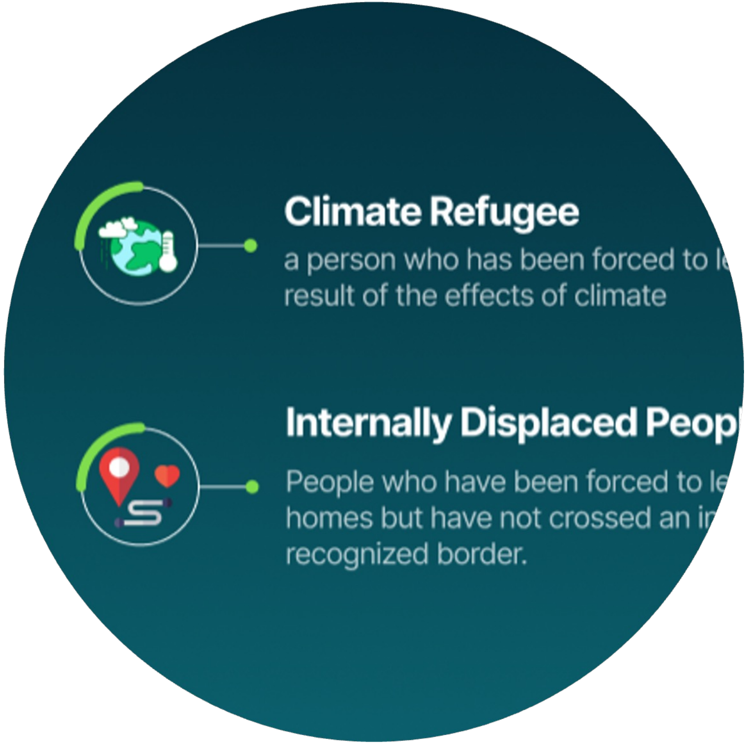
Yet, no predictive technologies exist to help mitigate a climate refugee crisis.

The Mission

Bring awareness to the neglect that the data in this field is experiencing and show the potential life-changing impact of complete, quality data.

Our Model

A model that **predicts the relative level of internally displaced individuals** in a specific region *if a flood were to take place.*



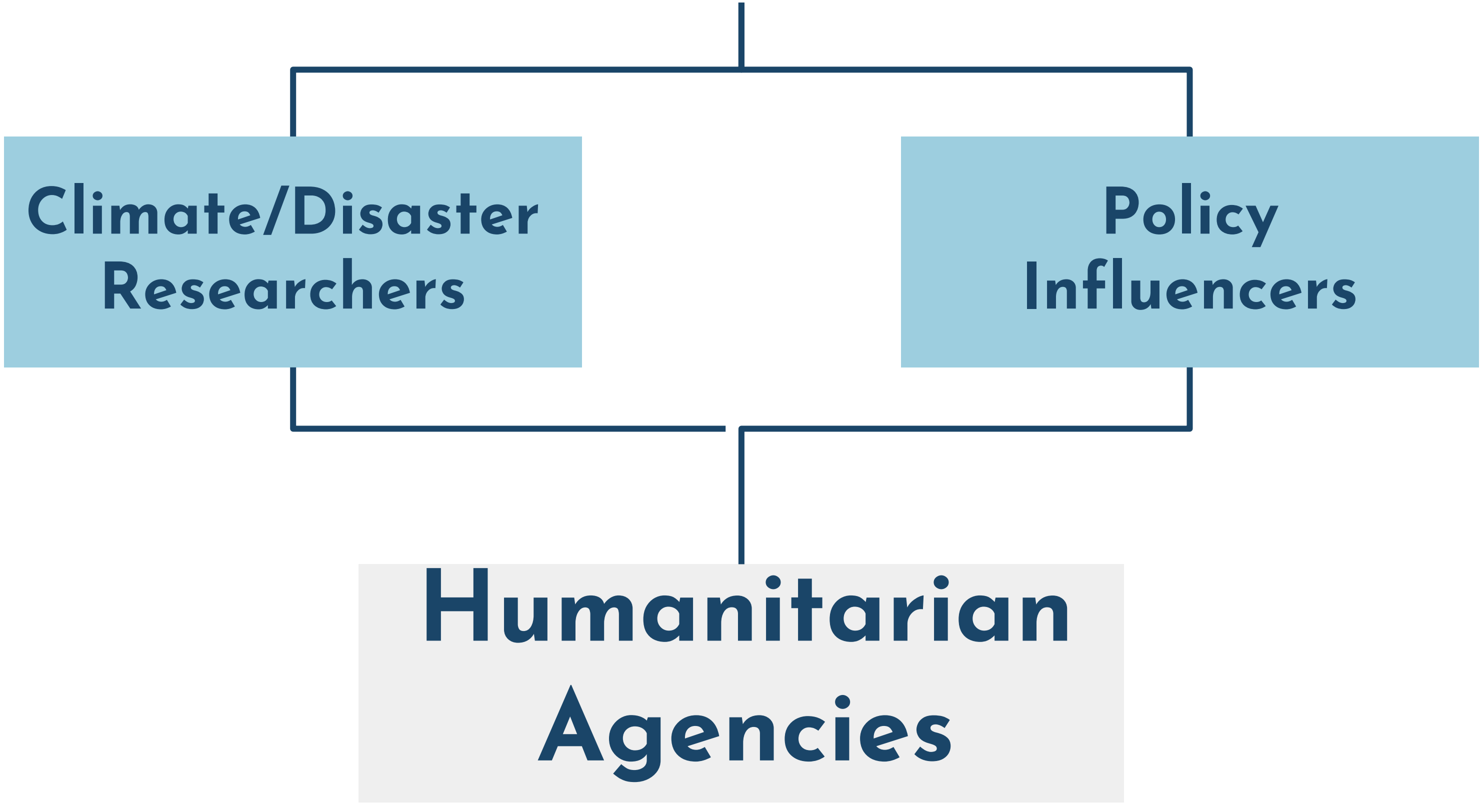
MVP Demo

https://www.figma.com/proto/7XYiW1g9OGeivnKblfI3pY/Diving-Landing-Page?type=design&node-id=67-25&t=zTs43Q4hjzlEi1gQ-0&scaling=min-zoom&page-id=0%3A1&starting-point-node-id=67%3A25

Who Are We Targeting?

- Climate/Disaster Researchers
- Policy Influencers

Who Are We Targeting?



Data and Modeling

Solution

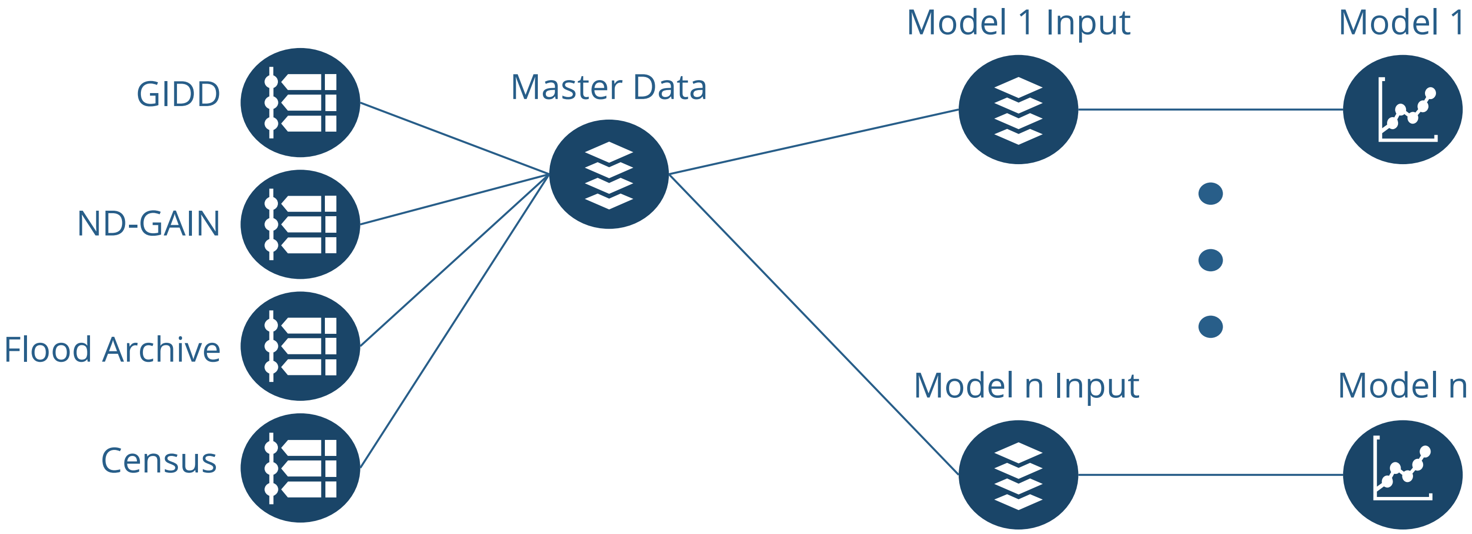
The Model

- **Random Forest model** predicting a “low”, “medium”, or “high” level of displacement (relative to global events) with an accuracy of ~50%.
- This model outputs the most **impactful features** in creating accurate predictions.
- Here we will see that specific strategies for measuring the magnitude of floods are effective, as well as measures of a countries preparedness for such an event

The Data

- Currently, data in the wild is cripplingly decentralized and unorganized.
- We have **specific recommendations** to address these issues in data collection
- Our data **combines four datasets** into one that encodes, for each event, the time/place, magnitude, causes, and features of the country such as population measures and various factors related to preparedness for natural disasters

Data Pipeline



Raw Data Sources

GIDD

	ISO3	Country / Territory	Year	Event Name	Date of Event (start)	Disaster Internal Displacements	Disaster Internal Displacements (Raw)	Hazard Category	Hazard Type	Hazard Sub Type
0	TLS	Timor-Leste	2013	Babulu gale	2013-01-17	5	5	Weather related	Storm	Storm

ND-GAIN

	ISO3	Name	1995	1996	1997	1998	1999	2000	2001	2002	...	2012	2013	2014	2015	2016	2017	2018	2019	2020
0	AFG	Afghanistan	0.496497	0.496497	0.496497	0.496497	0.496497	0.496497	0.496497	0.496497	...	0.175065	0.178628	0.201495	0.200231	0.261156	0.238742	0.21024	0.224049	0.213706

Flood Archive

	Country	Area	Began	Ended	MainCause	Severity	ISO3	Year Month
0	Indonesia	2178.65	2008-01-02	2008-01-06	Heavy rain	1.0	IDN	2008-01

Census

	Name	Region	GENC	Year	Population	Population Density (People per Sq. Km.)	Net international migrants, both sexes	ISO3
0	Afghanistan	2008,Afghanistan	AF	2008	27,703,539	42.5	222,570	AFG

Master Data

	IDPs from Event	Economics	Governance	Social	Capacity	Ecosystem	Exposure	Food	Habitat
0	270	0.178628	0.172592	0.335777	0.757451	0.507907	0.480512	0.580916	0.537736
1	740	0.201495	0.193780	0.341216	0.732208	0.503280	0.480512	0.576083	0.539343
2	244	0.201495	0.193780	0.341216	0.732208	0.503280	0.480512	0.576083	0.539343

	Health	Infrastructure	Sensitivity	Area	Began	Ended	MainCause	Severity	Duration	Magnitude
0	0.832165	NaN	0.437181	14653.47	2013-04-23	2013-04-29	Torrential Rain	1.0	6.0	11.384192
1	0.828587	NaN	0.436659	83722.34	2014-04-20	2014-05-16	Torrential Rain	1.5	26.0	14.998823
2	0.828587	NaN	0.436659	83722.34	2014-04-20	2014-05-16	Torrential Rain	1.5	26.0	14.998823

	Population	Population Density (People per Sq. Km.)	Net international migrants, both sexes	Scaled_IDP
0	31,098,161	47.7	-67,219	270
1	31,809,829	48.8	-58,115	740
2	31,809,829	48.8	-58,115	245

Modeling Approach

Feature Engineering

Combined a few features (i.e. duration, severity, and area)

Skewed Data

Performed logarithmic transformation on IDP counts for flood events. Normalized all inputs

Data Binning

Binned IDP counts two or three quantiles. Our problem then becomes a multiclass classification task.

Model Selection

PROS		CONS
• Able to capture nonlinear complex patterns • High accuracy on categorical models	**Neural Network**	• Requires large amounts of labeled data • Black box: difficult to understand model's decision making process
• Performs well in high dimensional spaces • High accuracy for regression model	**Support Vector Machine**	• Can be sensitive to noisy data and outliers • Black box: might lack interpretability when using complex kernel functions
• Good accuracy and robustness to overfitting • Provides a measure of feature importances for interpretability	**Random Forest**	• Slow and computationally intensive: not suitable for real time model • May not perform well beyond range of training data

Model Results Summary

Model Type	Output	Recall	Precision	Test Accuracy
NN	3 Classes	**H** 0% | **M** 0% | **L** 100%	**H** 0% | **M** 0% | **L** 32%	33.90%
NN [Log + Norm]	3 Classes	**H** 60% | **M** 8% | **L** 76%	**H** 52% | **M** 41% | **L** 45%	47.20%
RF [Log]	3 Classes	**H** 52% | **M** 41% | **L** 45%	**H** 52% | **M** 41% | **L** 45%	52.00%
SVM [Norm]	3 Classes	**H** 38% | **M** 30% | **L** 53%	**H** 48% | **M** 34% | **L** 40%	40.59%
NN [Log + Norm]	2 Classes	**H** 56% | **L** 67%	**H** 67% | **L** 68%	66.30%
NN [Log + Norm]	Continuous	-	-	**RMSE** 167168
SVM [Log + Norm]	Continuous	-	-	**RMSE** 225414
RF [Log]	Continuous	-	-	**RMSE** 258013

Norm = Normalized variables
Log = Log-transformed variable(s)

Model Results Summary

Model Type	Output	Recall	Precision	Test Accuracy
NN	3 Classes	**H** 0% | **M** 0% | **L** 100%	**H** 0% | **M** 0% | **L** 32%	33.90%
NN [Log + Norm]	3 Classes	**H** 60% | **M** 8% | **L** 76%	**H** 52% | **M** 41% | **L** 45%	47.20%
RF [Log]	3 Classes	**H** 52% | **M** 41% | **L** 45%	**H** 52% | **M** 41% | **L** 45%	52.00%
SVM [Norm]	3 Classes	**H** 38% | **M** 30% | **L** 53%	**H** 48% | **M** 34% | **L** 40%	40.59%
NN [Log + Norm]	2 Classes	**H** 56% | **L** 67%	**H** 67% | **L** 68%	66.30%
NN [Log + Norm]	Continuous	-	-	**RMSE** 167168
SVM [Log + Norm]	Continuous	-	-	**RMSE** 225414
RF [Log]	Continuous	-	-	**RMSE** 258013

Norm = Normalized variables
Log = Log-transformed variable(s)

Main Models

Random Forest Classification

Data Preprocessing

- Binned logged IDP counts into 3 quantiles
- Dropped null values
- One hot encoded month
- Simplified Main Cause
- Dropped country, area, severity, and duration as it was simplified into magnitude
- Logged skewed features

Model Training

- Hyperparameter tuning: Performed a random search with 100 estimators and 3 fold cross validation
- Suggested hyperparameters created large trees

Model Evaluation

- Achieved overall test accuracy of ~52%
- Top features included magnitude and other ND GAIN indicators (i.e. infrastructure, social, economic, etc)
- Confusion matrix and class recall scores showed model did best among all models for 'Medium' class

Support Vector Machine Regression

Data Preprocessing

- SVM is sensitive to noisy data and outliers
- Dropped missing values
- Normalized all input variables and logged IDP counts and skewed features
- Performed PCA Analysis for Dimensionality Reduction

Model Training

- Hyperparameter tuning: Performed a grid search with 3 fold cross validation
- Trained model with and without PCA

Model Evaluation

- SVM with PCA achieved 72% accuracy within 3000 IDP counts of the predicted value
- RMSE was fairly high (above 200,000)
- Data was still right-skewed with many outliers
- Need better labelled data

Wrap-Up

Challenges

Data Availability

- Data access
- Lack of observations
- Lack of spatial information
- Limits of coverage and sharing

Event Definition

- Neglected field
- Varying definitions of what counts as a distinct flood event

Our Contribution

Current Industry Steps

- Reactive: only makes short term predictions after the event has occurred to allocate resources
- Have monitoring stations to predict IDP counts at those regions only

Our Improvements

- Takes other factors into account (environment, economic, social, and population density) in predicting IDP counts
- Makes IDP count predictions in the event of a flood, meaning one would only have to look at flood data
- Focuses on root causes

Next Steps

Use a unified global system better assess the impacts of climate change on flood displacement risk

Create a publicly available, up to date, centralized database where all the scattered information can come together

Thank you for listening!

Any questions?

Robert Meyer
calrobert@berkeley.edu

Leanna Chraghchian
theleanna@berkeley.edu

Brendan Foo
bfoo@berkeley.edu

Felicia Liu
felicialiu@berkeley.edu