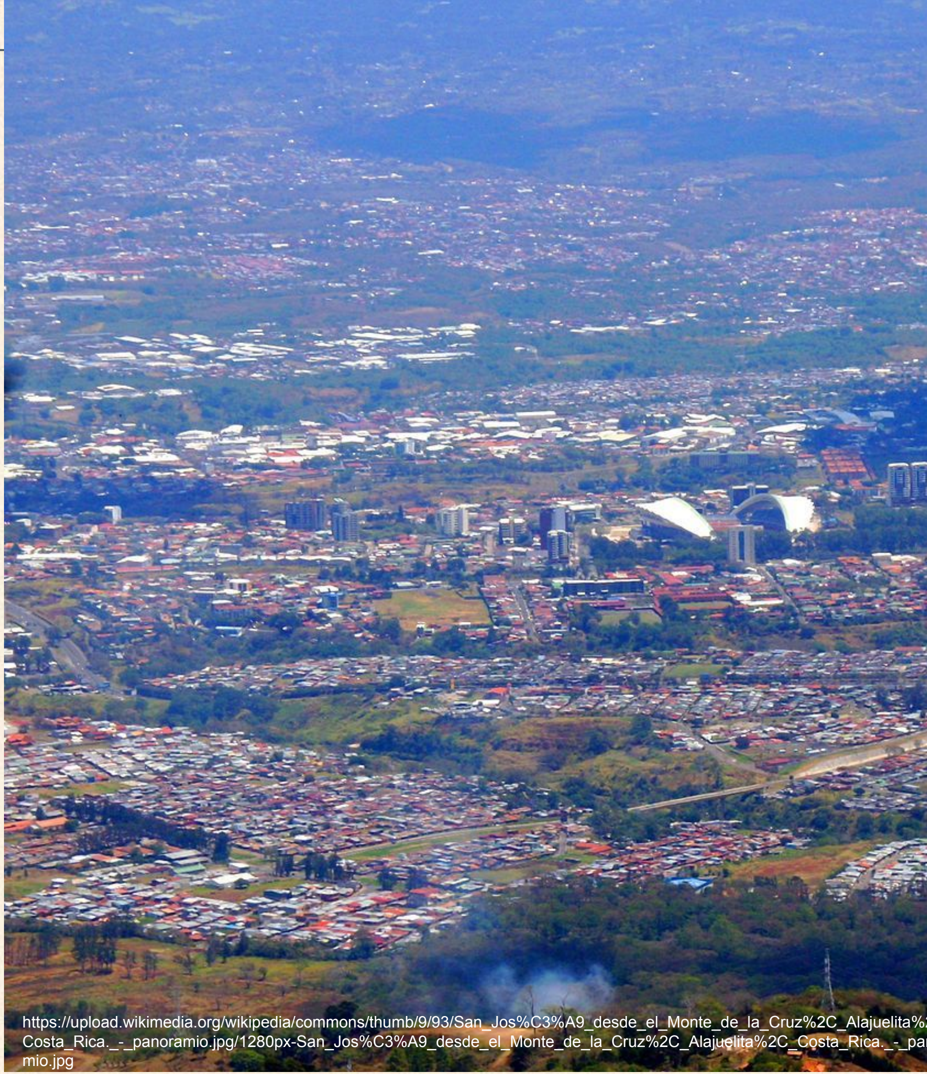
Predicting Poverty in Costa Rica

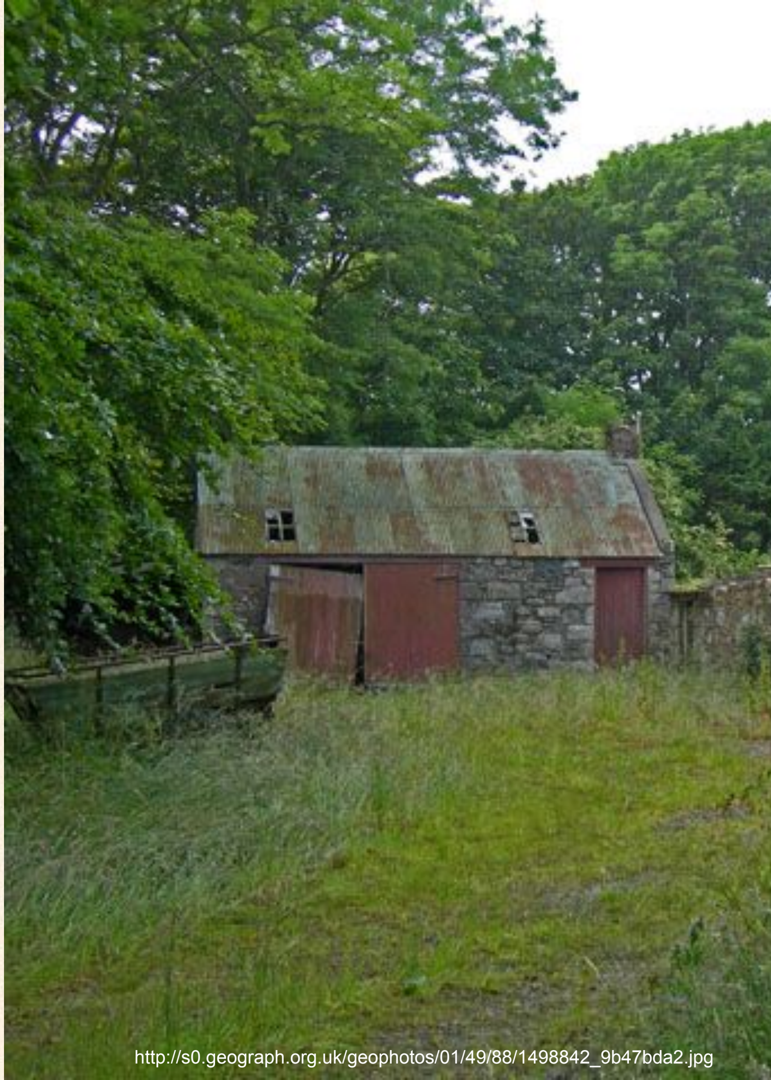
Sam Blass
Metis Classification
23 March 2022



https://upload.wikimedia.org/wikipedia/commons/thumb/9/93/San_Jos%C3%A9_desde_el_Monte_de_la_Cruz%2C_Alajuelita%2
Costa_Rica._-_panoramio.jpg/1280px-San_Jos%C3%A9_desde_el_Monte_de_la_Cruz%2C_Alajuelita%2C_Costa_Rica._-_par
mio.jpg

Objective

- Prioritize aid only to most vulnerable households
- Predict vulnerability using household observable attributes (e.g. size, dwelling quality)
- Kaggle competition 2019



http://s0.geograph.org.uk/geophotos/01/49/88/1498842_9b47bda2.jpg

Data Overview

https://upload.wikimedia.org/wikipedia/commons/thumb/8/89/CRI_orthographic.svg/1920px-CRI_orthographic.svg.png

Target variable - level of poverty

- 1 = extreme poverty
- 2 = moderate poverty
- 3 = vulnerable households
- 4 = non vulnerable households

Each row corresponds to one individual

Each individual in a household gets the same household ID number

Features (~140)

- Individual: Age, gender, years of education
- Household: Number of household members, quality of dwelling
- Geography: Region of Costa Rica located

Methodology

Exploratory data analysis and feature engineering

- Identify class imbalances
- Explore relationships not otherwise capturable in a model

Define classification metric

- Reflects objective of project (prioritize aid by identifying those most in need)

Test and optimize various classification models

- Recommend best performing model
- Train-test-split by household to avoid data leakage (Group Shuffle Split)

Data Prep and Model Setup

Exploratory Data Analysis

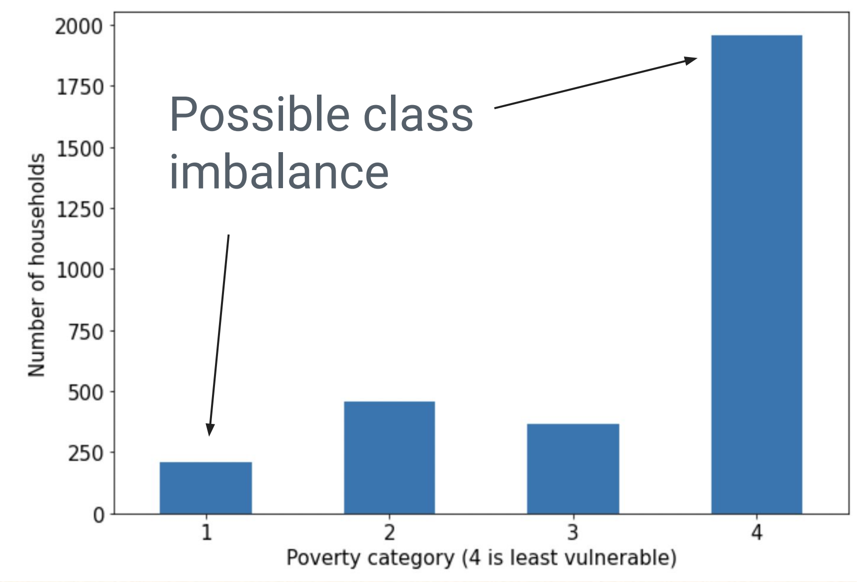
- Most households are not vulnerable
- Tuning model may require class balancing

Feature Engineering

- Scale features by geographical area to correct for cost of living

Classification Metric: **Recall**

- Minimize number of false negatives
- Prioritize category 1 (most vulnerable)
- Weighted average for each recall



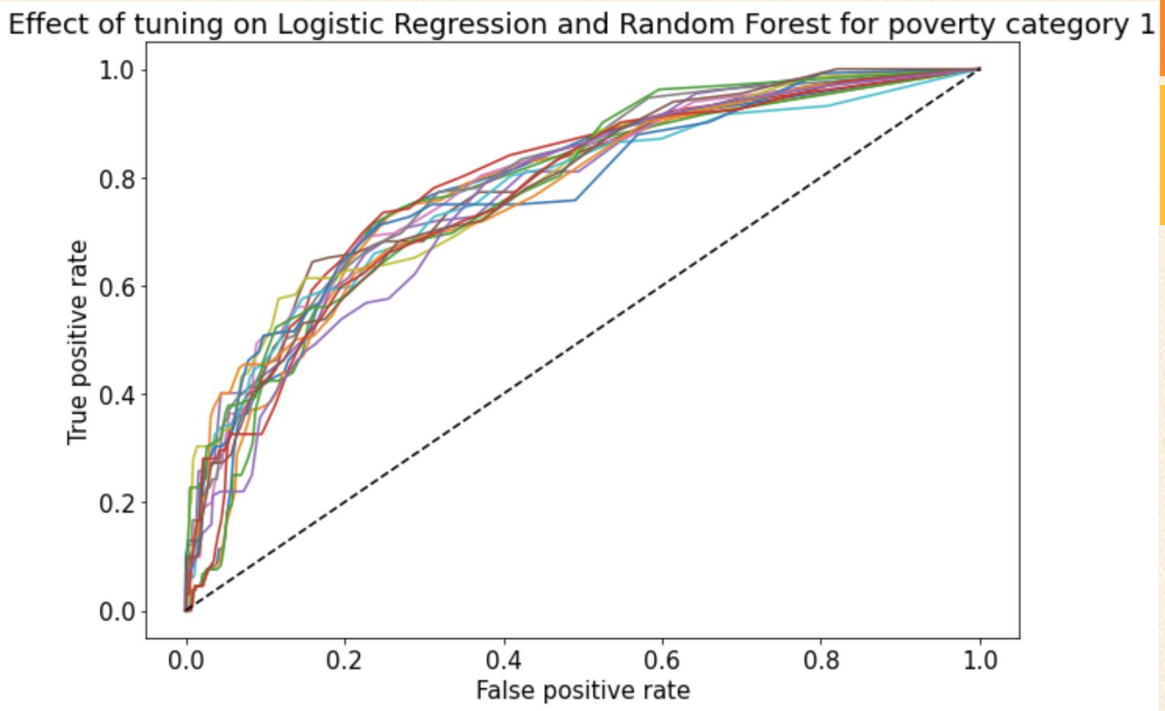
Possible class imbalance

Model Testing and Tuning

Model	Hyperparameters to tune
Random Forest	Max number of features, criterion, class weight
Logistic Regression	Regularization

*Max AUC is with **unbalanced** data

Metric	Max and Min
Recall	Max: **0.345** Min: **0.273**
AUC	Max: **0.782** Min: **0.739**



Model Testing and Tuning

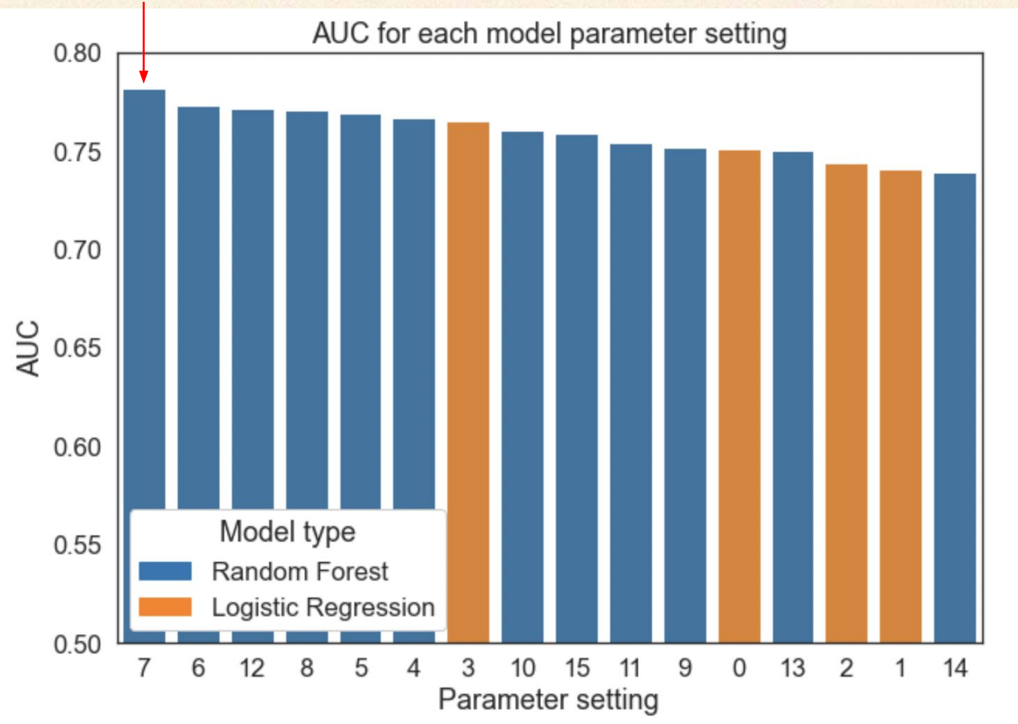
Model	Hyperparameters to tune
Random Forest	Max number of features, criterion, class weight
Logistic Regression	Regularization

*Max AUC is with **unbalanced** data

Metric	Max and Min
Recall	Max: **0.345** Min: **0.273**
AUC	Max: **0.782** Min: **0.739**

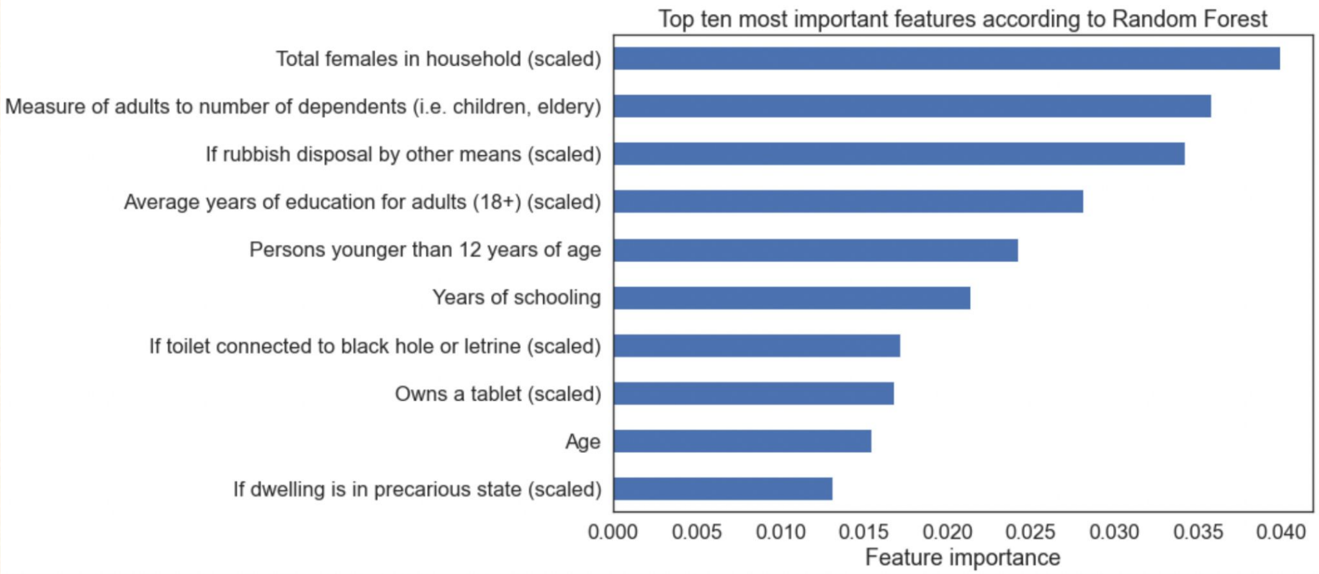
Best results with default settings

Default



Feature **Importance**

- Top two features are **number of females in household** and **dependency rate**
- Scaled features are important suggesting scaling relative to geographical area is important


Top ten most important features according to Random Forest

Scaled = relative to geographic area

Conclusions & Future Work

01 Data Cleaning

May improve recall scores

02 Hyperparameter tuning

No improvement in recall

03 Feature importance

Identified which features to examine further

04 Target variable subjectivity

Likely reduces model performance

Thanks!

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik**