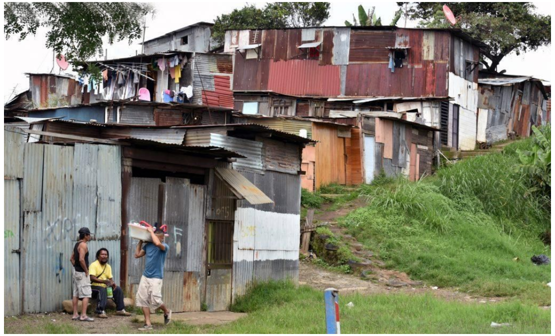
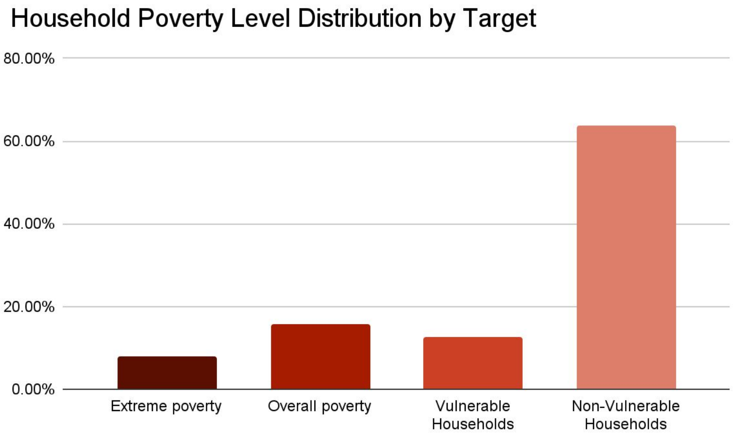
Costa Rica Poverty

Yueyue Wang, Gregory Ho, Sarah Walker, Jonathan Juarez

Background: multi-class classification problem with imbalanced dataset



Housing situation in Costa Rica



- **Target 1 - Extreme poverty ($0 - $2.5)**
  - Using food poverty line index (FPLI), the amount a household requires per member to meet a minimum caloric intake
- **Target 2 - Overall poverty ($2.5 - $4.0)**
  - Considers other basic resources beyond food
- **Target 3 - Vulnerable Households ($4.0 - $10.00)**
  - Households that were able to fulfill basic needs, but remain at risk of falling back into poverty due to unexpected circumstances.
- **Target 4 - Non-Vulnerable Households (Above $10.0)**
  - Non-poor & Non-vulnerable households.

Note: Income thresholds are harmonized per capita income for individuals

Data Processing

Dataset: Inter-American Development Bank Data on Costa Rican Households

Collection of observable Costa Rican household characteristics and associated poverty level for

9,557 individuals belonging to **2,973** households

143 variables that are a mix of and observations

observations

Aggregate individual observations

observations

Key Discrepancies:

- Large numbers of Null values in the following variables:
  - monthly rent
  - tablets owned
  - years behind in school

Main Cause & Solutions:

- Aggregation errors:
  - Build contingency table - Replace each missing value with appropriate value
  - Monthly rent: Impute based on median regional rent (for households in precarity and assigned housing)

Feature Engineering Overview: 25 new features were generated based on 4 dimensions

Housing

- Living conditions
  - physical conditions and materials of wall, floor, and roof
  - Overcrowding
- Housing stability
- Ownership status

Technology

- Technology usage, such as ownership of mobile phones, computers, and tablets in household

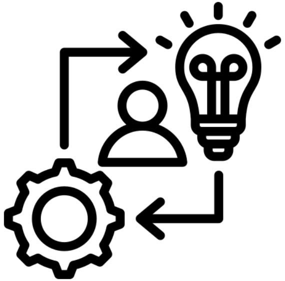
Education

- Educational attainment
- Years of education lost

Access to Basic Amenities

- Toilet System
- Water Provision
- Electricity Access
- Cooking Energy
- Rubbish Disposal

Feature Engineering and Selection

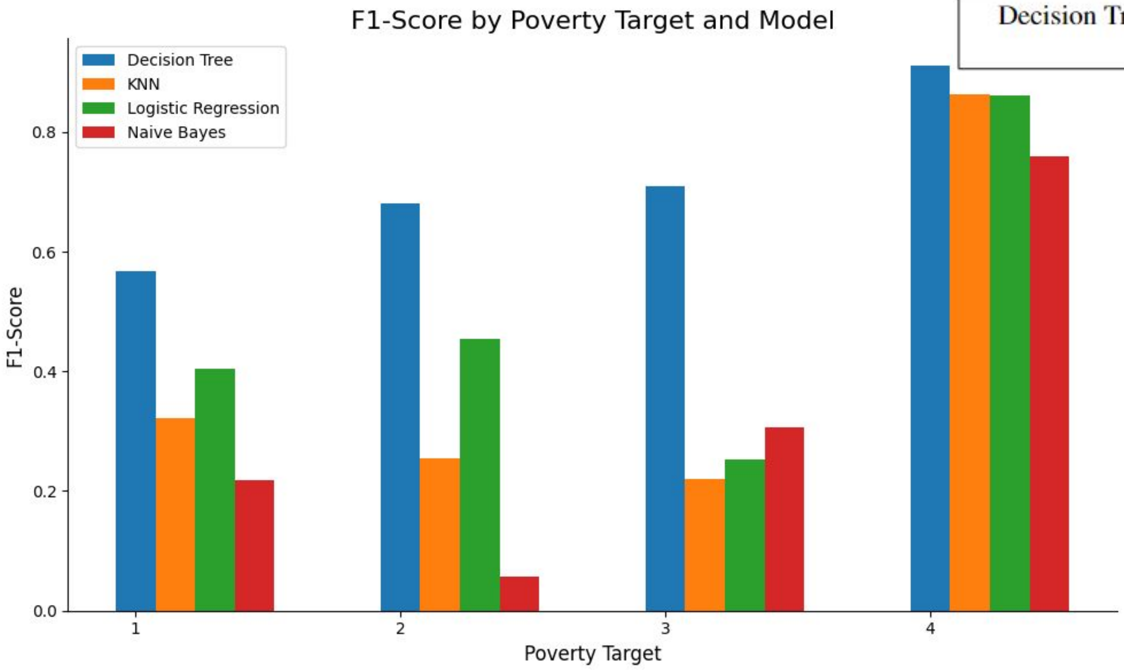


Technique we used:
- Encoding Categorical Variables
- Domain-Specific Feature Engineering: “monthly rent”
- One-hot encoding

Feature selection:
removed variables with
- Low correlation with target labels
- No observable difference in characteristics among target labels

Models Tried & Performance

Model Accuracies			
Model	Parameters	Accuracy Score	F1 Score
Logistic Regression	C = 0.01, penalty = ’l2’	0.66	0.65
Weighted KNN	K = 5	0.63	0.65
Naive Bayes	$\alpha = 10, \sigma = 0.01$	0.60	0.54
Decision Tree 1	max depth=none, use all variables	0.74	0.82
Decision Tree 2	max depth=26, use correlated variables	0.74	0.74
Decision Tree 3	max depth=29, use household level variables	0.74	0.8



Decision trees result in best F1 score for each target variable and overall.

Prone to overfitting, so random forest models created to address issue.

Random Forest - Hypertuning

Decision Trees:

- ➔ Full dataframe with cleaned and engineered features
- ➔ Label encoding for categorical variables
- ➔ Avoiding overfitting:
  - ◆ Limiting depth of tree
  - ◆ train/valid/test data
  - ◆ SMOTE
  - ◆ Random Forest
  - ◆ Accuracy, Precision, Recall, F1
- ➔ Final model + test on testing data
  - ◆ select_smote_rf

v2a1

asset_owned
mean_per_capita_income
tablet_per_capita
computer_per_capita
yrs_edu_lost
wall_material

Varname	Imp
* v2a1	0.393229
* mean_per_capita_income	0.120004
SQBescolari	0.049391
* asset_owned	0.048275
cielorazo	0.047837
escolari	0.045338
SQBedjefe	0.038075
meaneduc	0.036185
qmobilephone	0.027671
SQBmeaned	0.027423
instlevel8	0.020179
SQBhogar_nin	0.019539
overcrowding	0.019345
SQBovercrowding	0.018991
* wall_material	0.016949
v18q1	0.016651
rooms	0.012289
r4t1	0.011516
* yrs_edu_lost	0.009684
hogar_nin	0.008063

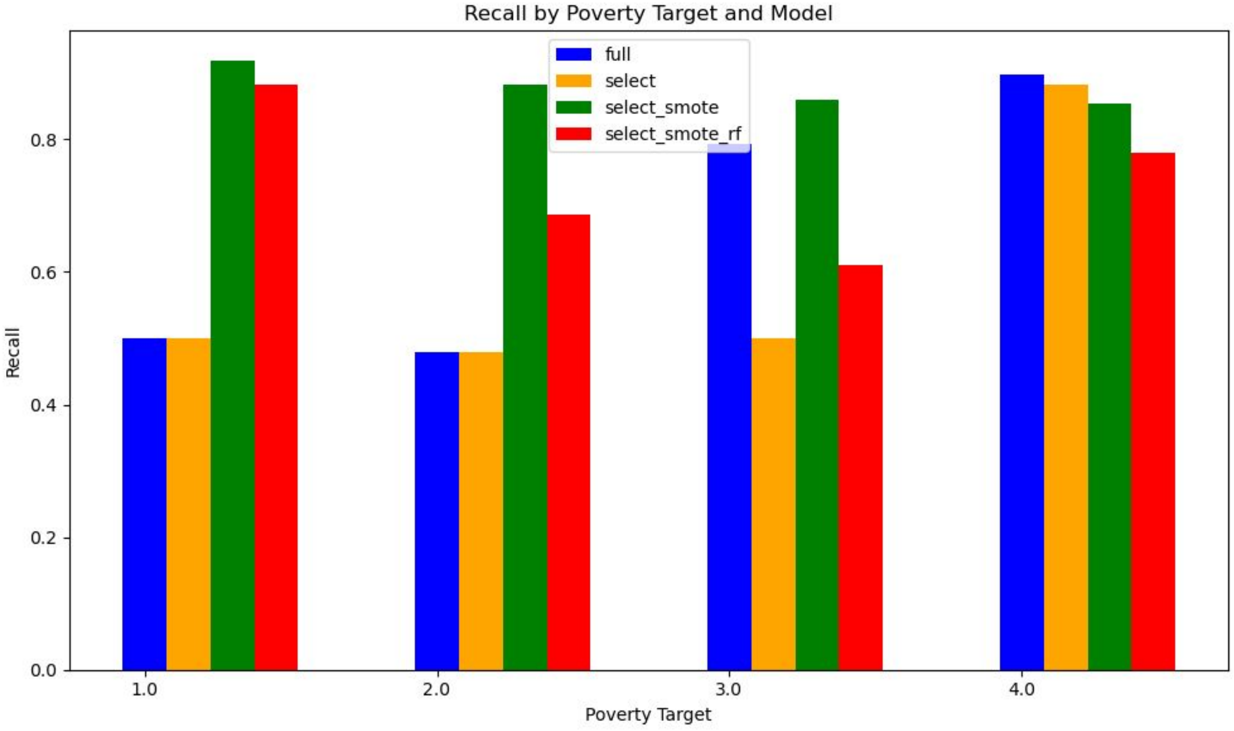
Final Model Selection

➔ Final model + test on testing data
  - select_smote_rf (red)

**note: select_smote (green)*

```
Accuracy: 0.7378516624040921
Classification Report:
              precision    recall

         1.0       0.67      0.87
         2.0       0.71      0.62
         3.0       0.82      0.63
         4.0       0.79      0.83
```



Recap & Reflections

What worked:

- ➔ Appropriately filling v2a1(rent) and engineered variables (ex: mean_per_capita_income, asset_owned) worked as they are high predicting features
- ➔ Achieving high recall scores with Random Forest models

What didn’t:

- ➔ Logistic Regression & Bayesian Classifier - conditional independence assumption violated
- ➔ KNN suffers from curse of dimensionality
- ➔ How to best handle overfitting; Random forest tree can be ambiguous to understand

A comment about target features:

- ➔ Targets 1 and 2 have concrete meaning - it is based on the Cost of Basic Needs approach.
- ➔ Targets 3 and 4 are loosely defined:
  - ◆ What does it mean for a household to be vulnerable?
- ➔ Target features had a direct mapping to an income-based classification of poverty.
  - ◆ features which had a closer relationship with monetary deprivation had better explanatory power.
- ➔ Dataset contained mostly indicator variables, while target measure was income based.