Predictive Analysis of Costa Rican Household Poverty Index Based on Data Mining Classification Methods

Aaron Carr and Anusia Edward

University of San Diego

Background

- Poverty in Costa Rica
- Inter-American Development Bank
- Proxy Means Test (PMT)

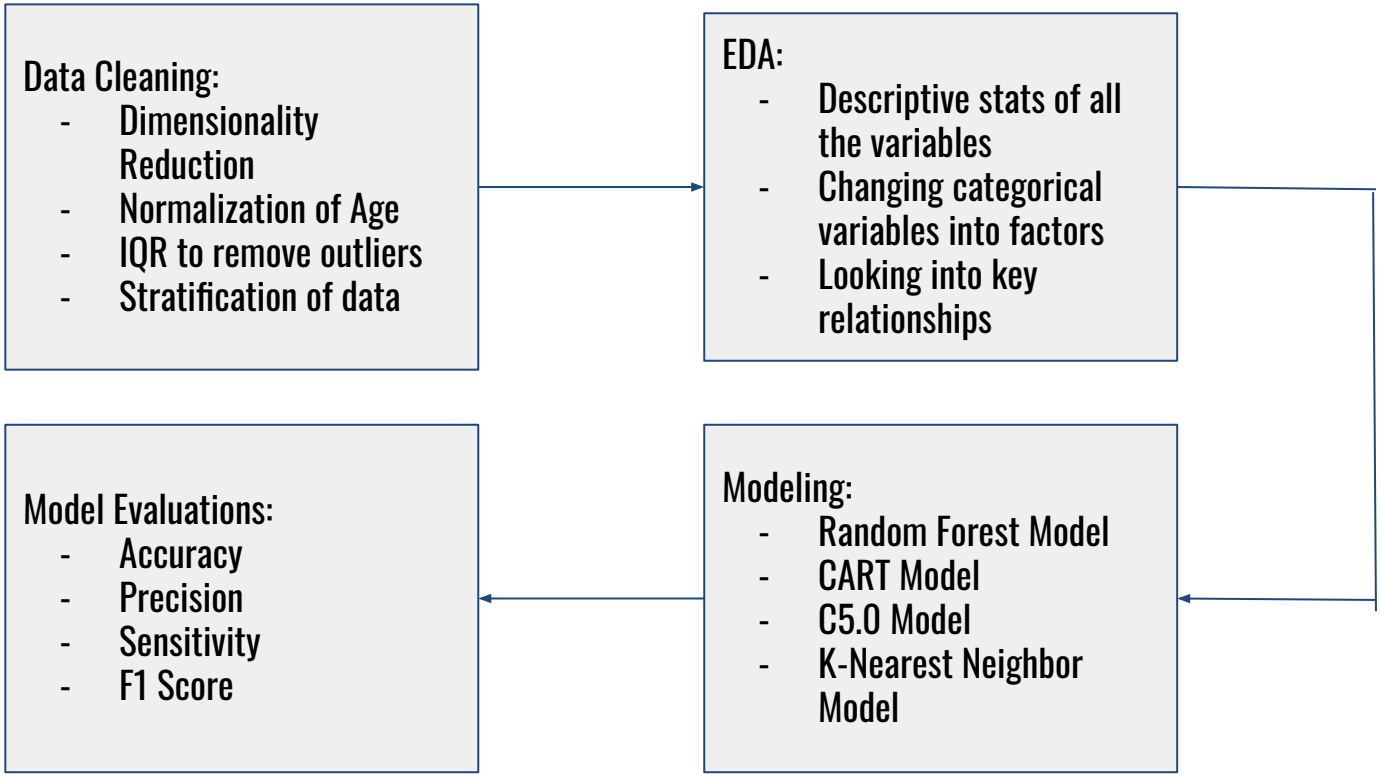
Purpose + Task

Purpose: The purpose of this study is to help the Inter-American Bank determine which Costa Rican families are in need of support due to their economic disposition, in order to be able to allocate resources and funds to those families.

Task: Create a predictive model using classification methods to predict whether or not a household is living in poverty.

University of San Diego

Overview of Project + Plan for Methods



University of San Diego

Data Cleaning Process

Exploratory Data Analysis (EDA)

- Examine data characteristics
- Descriptive Analyses
  - Univariate

Table 1

	total_persons	num_children	dependency
Mean	4.013	1.412	0.400
Median	4.000	1.000	0.400
Standard Deviation (SD)	1.766	1.368	0.254
Sample Variance	3.118	1.871	0.065
Range	12.000	9.000	1.000
Minimum	1.000	0.000	0.000
Maximum	13.000	9.000	1.000

Note. n = 6,690

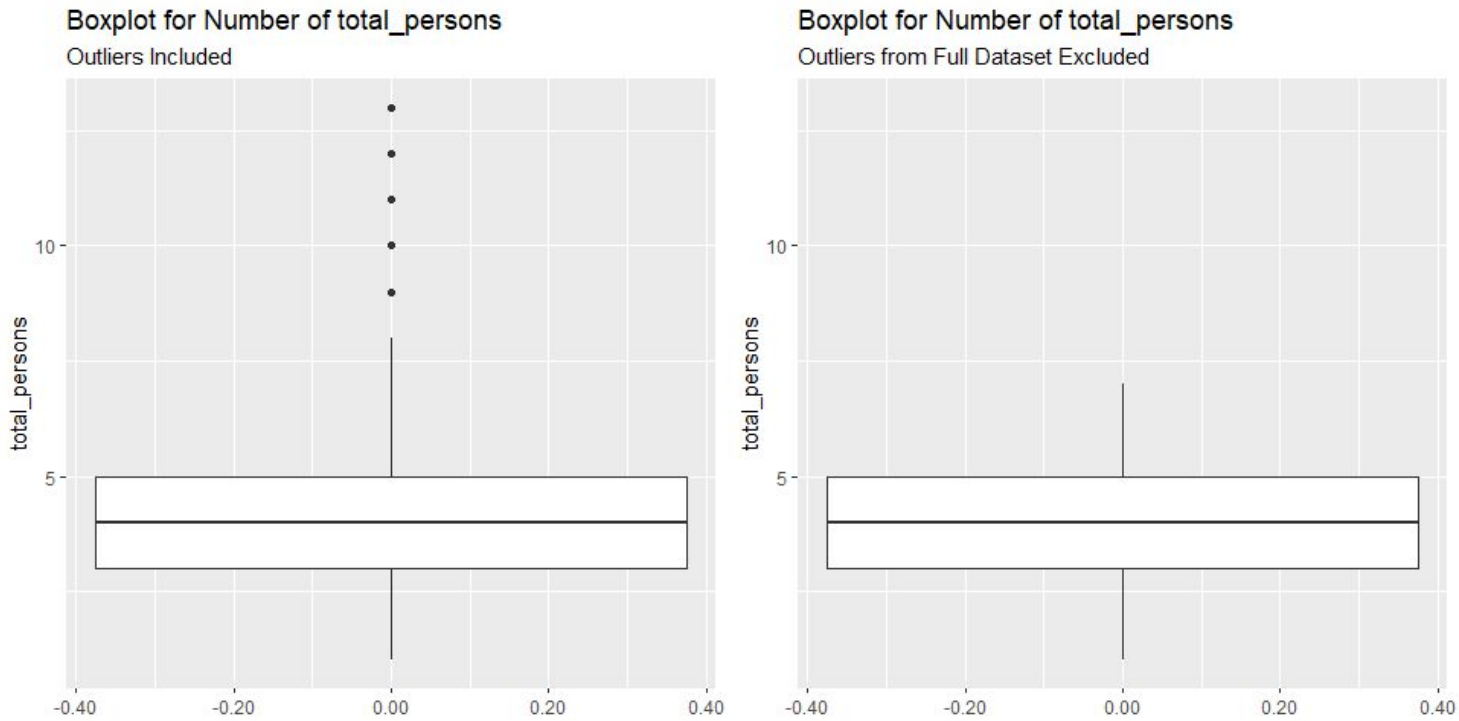
  - Multivariate
    - Contingency tables
    - Correlation matrix
  - Visualizations

EDA Cont'd - Several Types of Visualizations

- Boxplots
- Bar graphs
- Correlation Plot
- Scatterplot matrix

EDA Cont'd - Boxplot for total_persons

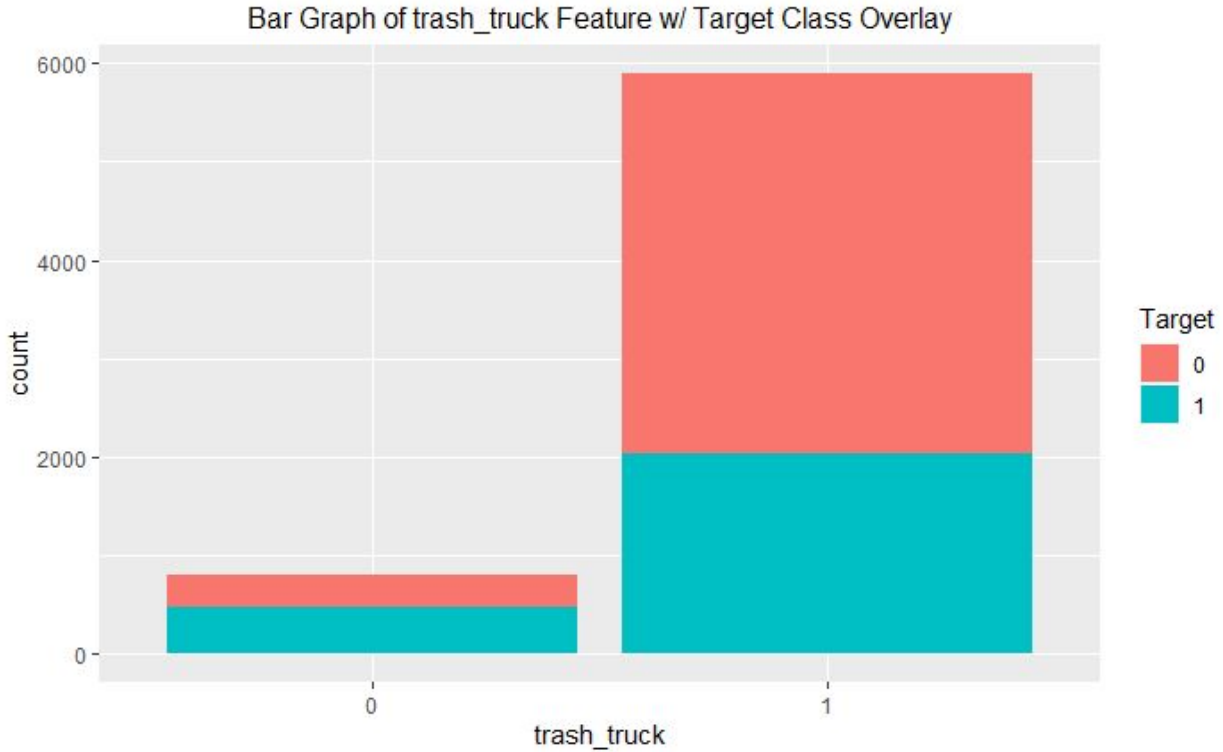
Figure 1



University of San Diego

EDA Cont'd - Bar Graph for trash_truck w/ Target Overlaid

Figure 2



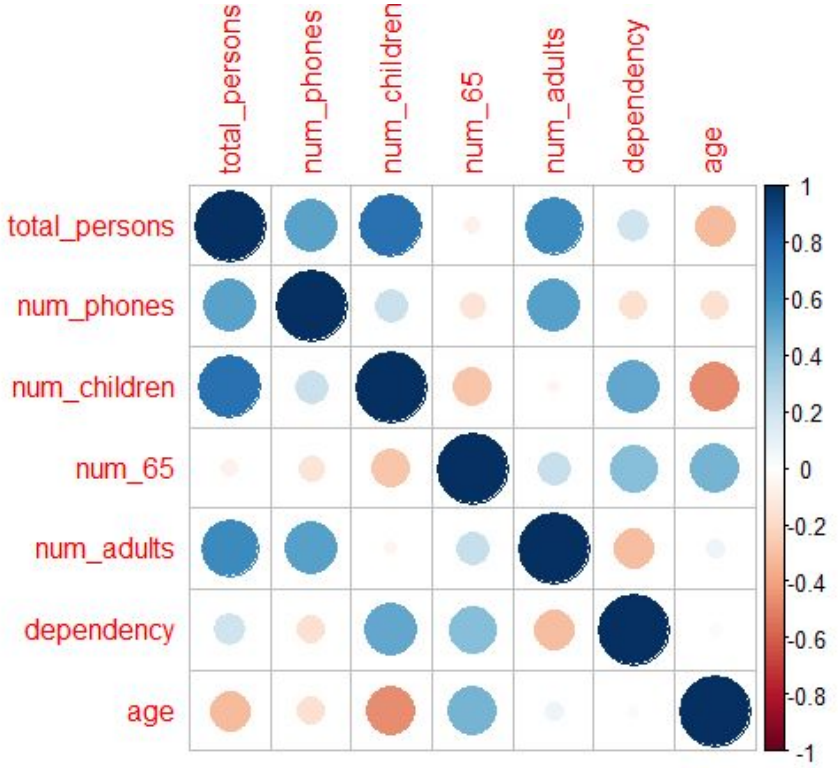
University of San Diego

EDA Cont'd - Correlation Matrix Plot

Table 2

	total_persons	num_phones	num_children	num_65	num_adults	dependency	age
total_persons	1.000	0.530	0.748	-0.063	0.633	0.199	-0.319
num_phones	0.530	1.000	0.220	-0.139	0.542	-0.154	-0.157
num_children	0.748	0.220	1.000	-0.274	-0.040	0.517	-0.469
num_65	-0.063	-0.139	-0.274	1.000	0.224	0.425	0.468
num_adults	0.633	0.542	-0.040	0.224	1.000	-0.304	0.068
dependency	0.199	-0.154	0.517	0.425	-0.304	1.000	-0.029
age	-0.319	-0.157	-0.469	0.468	0.068	-0.029	1.000

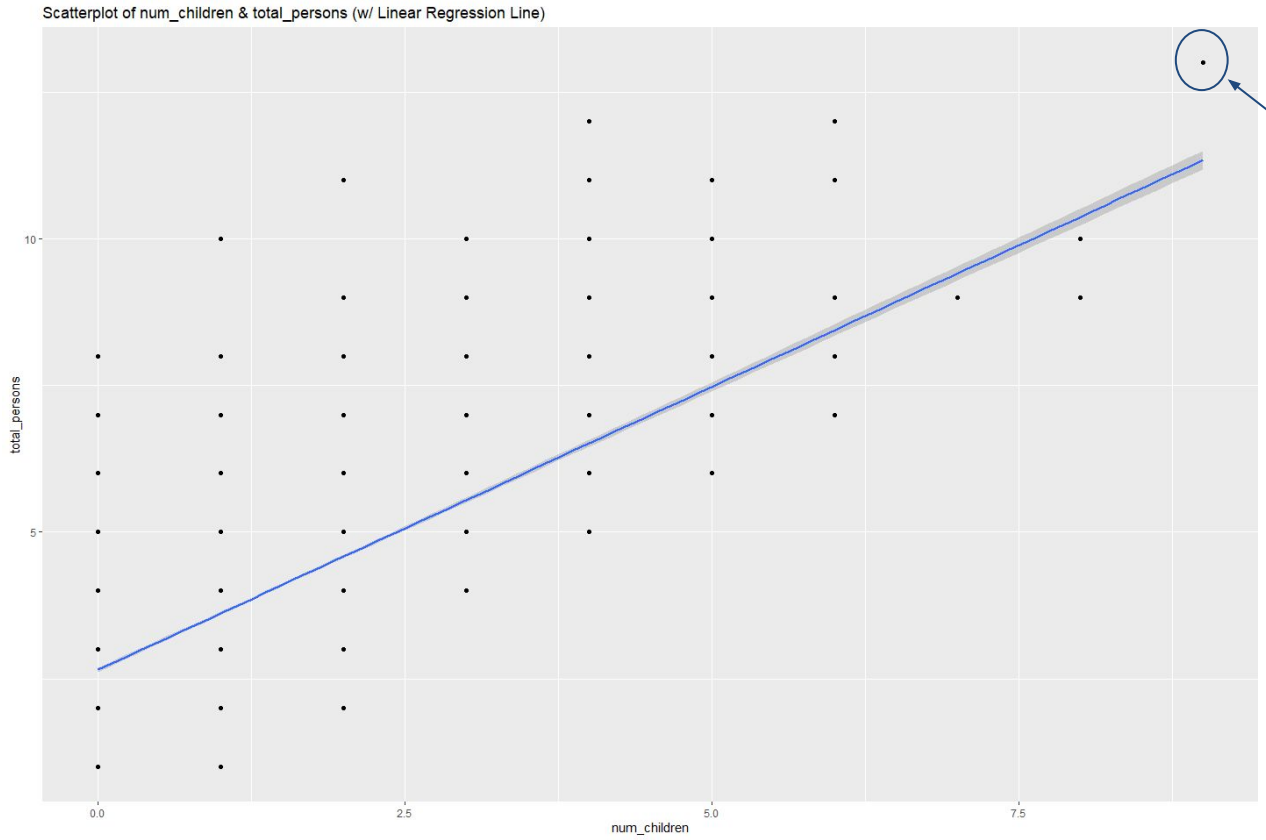
Figure 4



University of San Diego

EDA Cont'd - Scatterplot of num_children & total_persons

Figure 3



University of San Diego

Preliminary Models

Model Evaluations

- Focus on sensitivity and precision as class. eval. measures
- No stand-out “best” model

Table 3

Model	Classifier	Accuracy	Sensitivity	Precision	F_1
M_1	Random Forest	**.739**	.466	**.741**	.572
M_2	CART	.707	.464	.655	.543
M_3	C5.0	.722	**.497**	.677	.573
M_4	KNN-81	.709	.375	.712	.491
M_5	KNN-3	.732	**.490**	.705	**.578**

Conclusion & Next Steps

- Predictive models created
- Performance of all models was mixed
- Utilized broadly available features
- Future directions
  - Investigate other features to add to model
  - Refine parameters

University of San Diego

References

Inter-American Development Bank. (2018). Costa Rican household poverty level prediction. *Kaggle*. https://www.kaggle.com/competitions/costa-rican-household-poverty-prediction/overview/description

The World Bank. (2015, September 30). *FAQs: Global poverty line update*. https://www.worldbank.org/en/topic/poverty/brief/global-poverty-line-faq

The World Bank. (2021, October 6). *The World Bank in Costa Rica*. https://www.worldbank.org/en/country/costarica/overview#1

United Nations. (n.d.). *Ending poverty*. https://www.un.org/en/global-issues/ending-poverty