MEASURING HOUSEHOLD POVERTY RISK IN NEW YORK CITY

Machine Learning for Cities

Professor Daniel Neill

Spring 2017

Anastasia Shegay, anastasia.shegay@nyu.edu

Shalmali Kulkarni, sck408@nyu.edu

Vishwajeet Shelar, vys217@nyu.edu

Background and Motivation

New York City is one of the wealthiest cities in the world, yet 1.8 million of its residents live in poverty. The latest poverty threshold, which takes into account the high cost of housing in New York City, amounts to $31,581 for a family of two adults and two children.[1] Based on this threshold, 20.7 percent of the New York City population was living in poverty in 2014, a rate that has not changed significantly since the American Community Survey was first implemented in 2005 (Figure 1). In comparison, the national poverty rate was 14.8 percent in 2014.

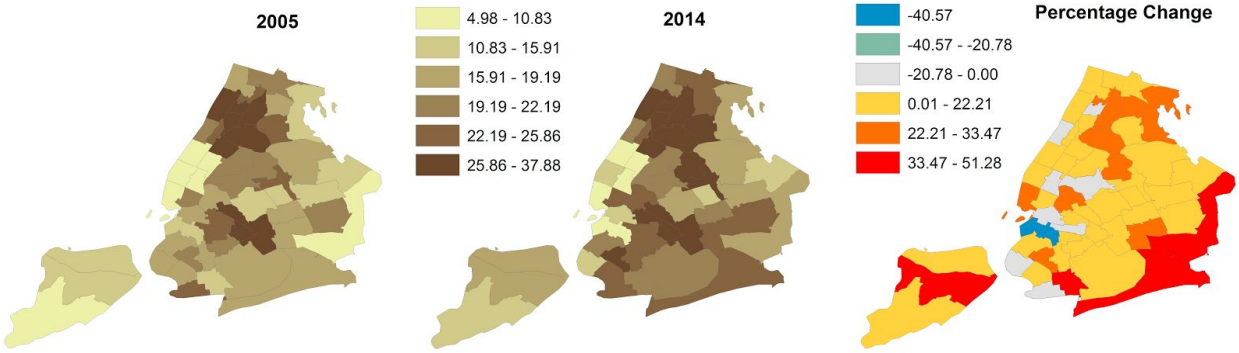
The focus of this study is on the New York City households whose incomes put them above the poverty line, but below 150 percent of the threshold. This segment constitutes 24.5 percent of the population, or 2.1 million people. We classify these households as "at risk of poverty" because while their income levels are in the "not poor" category, they are not sufficiently high to live comfortably. These individuals might still live in precarious conditions and slip into poverty easily.

The reason we chose to focus on this segment of the population is twofold. First, much of the urban poverty research is aimed at analyzing the poor and the extreme poor. The segments of the population close to the poverty line receive less attention from researchers. Second, similarly to poverty research, most policy interventions are targeted at the poor and the extreme poor populations. Understanding the factors that put households "at risk of poverty" can inform policy makers in designing targeted interventions that can prevent these groups from falling into poverty and even elevate them above 150 percent of the poverty line into greater prosperity.

[1] New York City Center for Economic Opportunity. CEO Poverty Measure 2005-2014. http://www.nyc.gov/html/ceo/downloads/pdf/CEO-Poverty-Measure-2016.pdf

The project aims to answer two questions: (1) What are the most important attributes in predicting that a household is at risk of poverty? and (2) What is the probability of a household being at risk of poverty conditional on each attribute? The project aims to answer the above questions using machine learning classification techniques, namely a Decision Tree for identifying the features and a Bayesian Network for calculating probabilities.

Figure 1. Percent of population at risk of poverty, 2005-2014



Data

American Community Survey (ACS) 1-year Estimates[2]

The primary data source for the project is the ACS 1-year estimates, a survey conducted annually since 2005 by the U.S. Census Bureau. Data is sampled from Public Use Microdata Areas (PUMAs), geographic areas with a population of at least 100,000 people. New York City is divided into 55 PUMAs, which are closely aligned with community districts/neighborhoods. The 1-year estimates are available for ten years from 2005 to 2014. Each year's dataset contains responses from approximately 24,000 households. The ACS dataset has around 500 variables with detailed information about each individual in the household along with household details

[2] United States Census Bureau. American Community Survey. https://www.census.gov/programs-surveys/acs/data/pums.html

such as type of household, duration of stay, type of internet connection, electrical appliances, etc.

NYC Center for Economic Opportunity Poverty Measure[3]

In addition to the ACS data, we obtained poverty thresholds available for each year from 2005 to 2014 and for each type of household from the CEO Poverty Measure datasets. The CEO poverty measure takes into account regional variations by adjusting the threshold to the higher cost of housing in New York City. Other features which we included in our model from the CEO dataset are benefits received and expenses incurred by a given household, e.g. housing adjustment, medical out-of-pocket expenses (MOOP), Housing and Energy Assistance Program (HEAP), food stamps, school lunch, etc. (Refer to Appendix 1 for detailed variable description).

Methodology

Traditional poverty prediction methods are based on statistical projections--typically time-series--that use a set of assumptions about the rates of economic and population growth. These forecasts, however, provide only national or subnational level of analysis without the granularity which household poverty analysis allows. Some of the recent efforts in poverty prediction included application of machine learning approaches to mobile phone usage and satellite data. Researchers at Stanford University were able to train a transfer learning model on daytime satellite imagery to accurately predict spatial distribution of poverty across five African countries.[4] Another study used anonymized data from Rwanda's largest mobile phone network to

[3] New York City Center for Economic Opportunity. CEO Poverty Measure 2005-2014. http://www.nyc.gov/html/ceo/downloads/pdf/CEO-Poverty-Measure-2016.pdf

[4] Jean, Neil at al. Combining Satellite Imagery and Machine Learning to Predict Poverty. Science Magazine. 19 August 2016. Vol 353, Issue 6301: 790-794. https://web.stanford.edu/~mburke/papers/JeanBurkeEtAl2016.pdf

predict the poverty and wealth of individual subscribers.[5]

The approach we take in this project is different from these earlier methods and is unique in four major ways: (1) While most predictive models use a binary classifier, i.e. poor and not poor, we introduce a third class--"at risk of poverty"--to capture vulnerable populations living close to the poverty line; (2) The approach allows us to conduct analysis at the household level using publicly available datasets without having to rely on longitudinal survey data or proprietary mobile phone data which is often difficult and expensive to obtain; (3) While most approaches are purely income-based, our model also incorporates benefits and expenses as predictors of household poverty risk;* (4) Implementation of the model in a Bayesian Network allows us to not only explore conditional dependencies among variables, but also report conditional probabilities for each attribute, which in turn ensures greater interpretability.

Preprocessing

We manually reduced large number of initial features by removing repetitive features common to encoding survey responses. We also aggregated the features like age, education attainment, child care expenses, food stamps, etc. to household level from individual level by, for instance, only including the highest level of education achieved per household and counting the number of individuals in each age category.

Multi-class classification

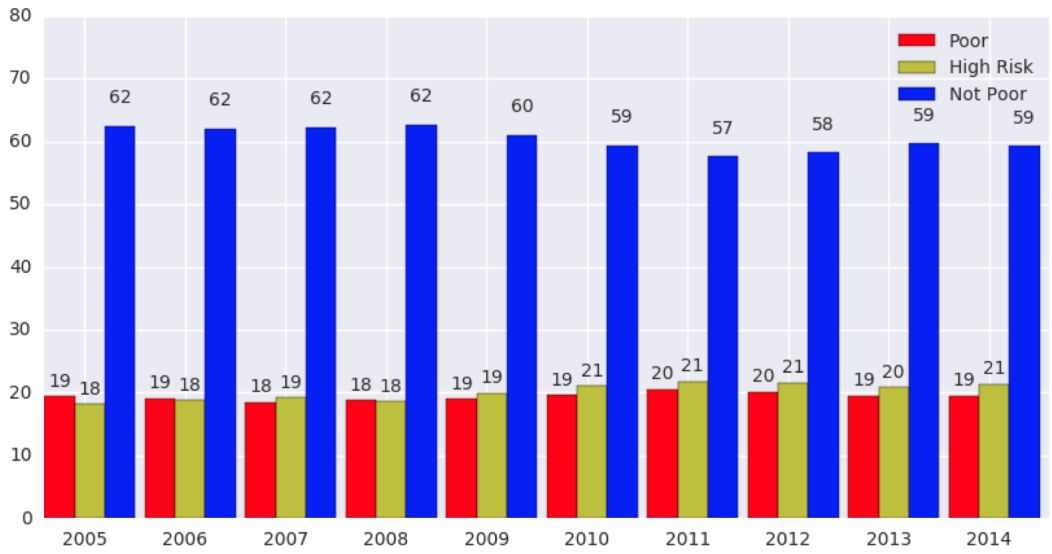
We labeled the data into three classes: *poor*, *at risk of poverty* and *not poor* based on the

[5] Blumenstock, Joshua. Predicting Poverty and Wealth from Mobile Phone Metadata. Science Magazine. 27 November 2015. Vol 350, Issue 6264: 1073-1076. http://science.sciencemag.org/content/sci/350/6264/1073.full.pdf?ijkey=jl1FOo2RaNJQk&keytype=ref&siteid=sci

* Careful measures must be taken while interpreting the results having expense and benefits as predictors of poverty

poverty threshold calculated by the CEO. The threshold for "at risk of poverty" was taken at 150 percent of the CEO threshold and used to divide the segment above the poverty line into two subclasses (Figure 2).

Figure 2 : Percentage Distribution of three-classes, 2005-2014



Discretization of variables

The Bayesian Network package in Python requires the features to be discretized in order to process the network. The features for benefits and expenses were continuous variables in absolute dollar amounts, which we categorized into "yes, received benefit/incurred expense" (1) and "no, did not receive benefit/did not incur expense" (0).

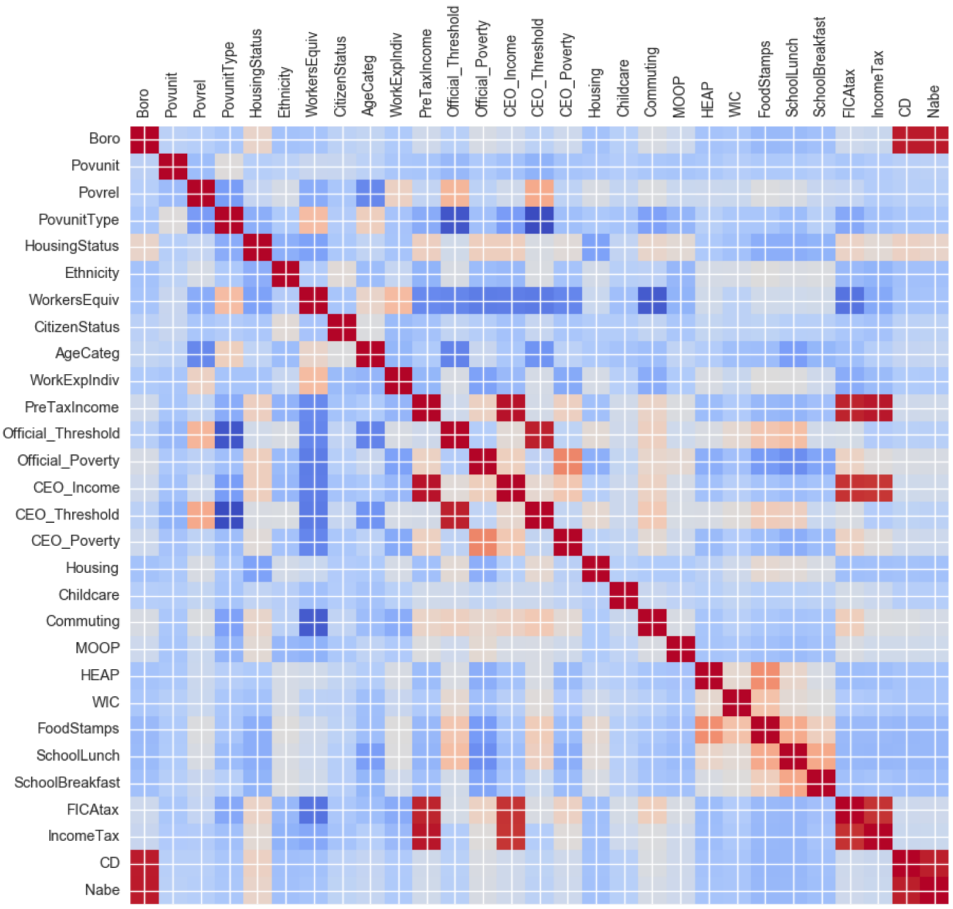
Exploratory Analysis

Correlation coefficients matrix

The initial exploration about the features was to understand the correlations between various features. It was observed that WorkersEquiv (number of working hours per household)

has a negative correlation with commuting and medical out of pocket expenses, as well as HEAP and Women, Infants, and Children (WIC) program benefits whereas all the benefits have a positive correlation among themselves (Figure 3).

Figure 3: Matrix of correlation coefficients

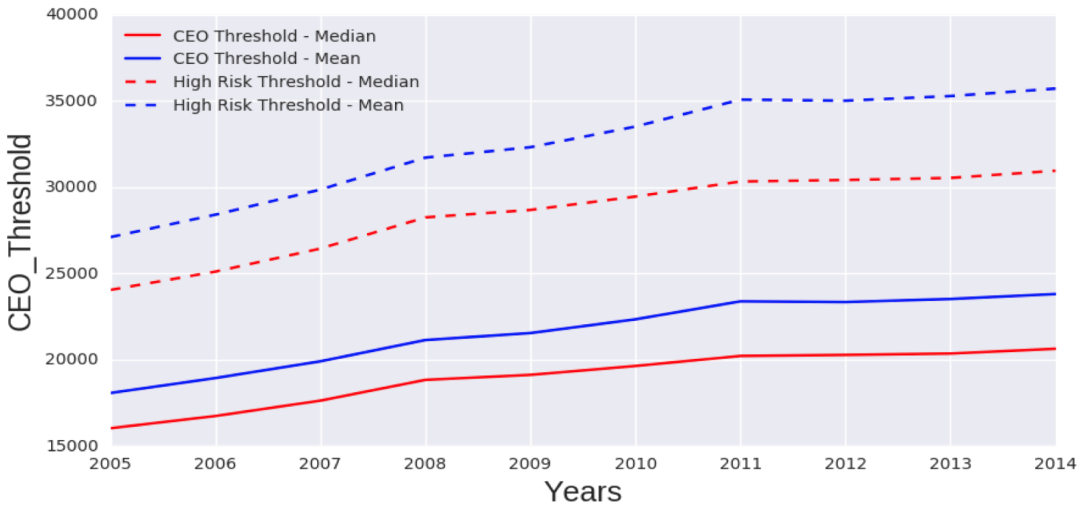


Threshold variation

The official threshold values, unadjusted for regional variation, increase from $19,806 in 2005 to $23, 283 in 2012 whereas the CEO New York City-adjusted threshold increased from $24,532 in 2005 to $31,039 in 2012. An interesting trend is seen for both threshold values during 2008-09 where they drop slightly than the previous year indicating income drop due to economic

recession. This is also observed in the CEO thresholds which drops from 6.8 percent increase to 1.5 percent increase (Figure 4).

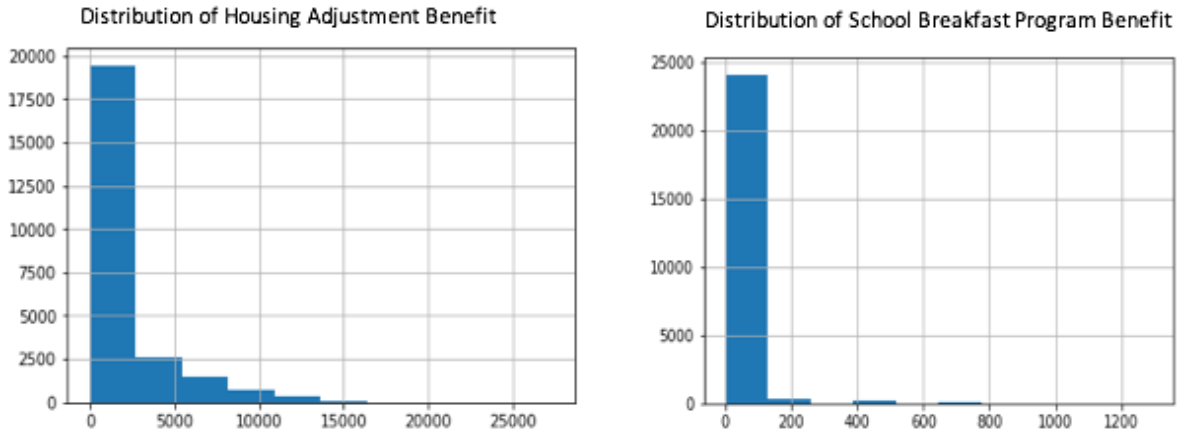
Figure 4: CEO Threshold (USD) and the 'At risk of poverty' Thresholds, 2005 -2014



Skewed Data

The data is skewed for certain variables such as food stamps, commuting, housing adjustments, etc. towards '0'. This values indicates the dollar amounts received or paid, but the value '0' indicating no amount is paid or received by the household. These variables are discretized for study as a way to deal with the skewness. Below are the figures of the distributions of some of these variables (Figure 5).

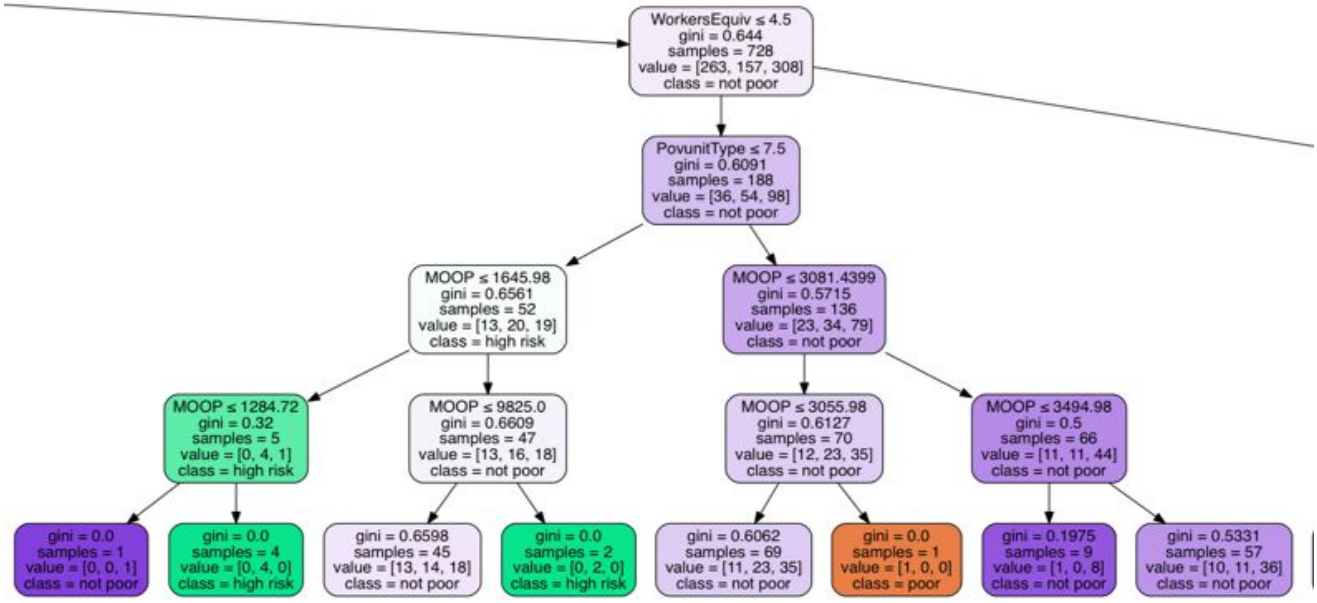
Figure 5: Distribution of Housing Adjustment & School Breakfast program



Classification Model - Decision Trees

The simplest and effective supervised learning classification model of Decision trees was trained on all the features for 10 years data. As the problem is defined with three classes, the approach of 'One versus rest' classifier was used to have interpretability of the model. The initial decision tree with 18 features resulted into a complex tree structure, which was further tuned using the hyper-parameters such as *max-depth* and *max-leaf-node* to get a pruned decision tree. The prediction accuracy increased from 60 to 68 percent. The trained Decision Tree Classifier presented the feature importance based on the 'gini importance' of each variable. This importance was used to further reduce the number of features for the Bayesian Network model. The top ten features were selected for further analysis as well as a network structure between the features. Thus, we used the Decision tree classifier to classify the data in three classes and also as a feature selection tool based on the gini importance.

Figure 6 : Decision Tree with max-depth 8 (only a part of the tree is shown)



Probabilistic Model - Bayesian Network

The next step was to get the structure of the network with all the selected features from Decision Tree. The Bayesian Network was used to get conditional dependencies of the selected features. *Hill Climb searc*h was used to estimate the Bayesian model structure with nodes and edges, using '*BIC (Bayesian Information criteria) Score*'. This method scores the measure of how much a given variable is "influenced" by a given list of potential parents and the lowest scoring model is selected to avoid overfitting.[6] A different network structure was created for each year (10 years) and a network structure for all the combined years. The year-wise network and the combined network appeared to be largely similar, as all the models used the same variables. The minor variations were the result of the Hill Climb search method as it gets the local maximum score rather than the global maxima.

The best models derived from the Hill Climb search were then trained on the data to get

[6] https://en.wikipedia.org/wiki/Bayesian_network

the conditional probabilities (CPD). The probabilities for each path in the network was calculated using the Bayesian Model estimators (Table 1). The trained models produced an average accuracy of 69.6 percent. The 10 years of data was merged together to derive a Bayesian Network using Hill Climb search. The figure 7 shows the Bayesian Network Structure obtained derived from the search.

The CPD table shows that the Pov_risk is conditionally dependent on Workers Equivalent which is the Worker equivalent hours in the household. The probability that the household is at risk given the Worker Equivalent is (4) i.e 'No hours worked' is 26 percent. The 'pgmpy' package builds the models as per the data it receives and it has its own limitations and assumptions, the model is not the accurate representation of the reality.

Table 1. Conditional Probability Distributions for "At Risk of Poverty" Class

WorkersEquiv	WorkersEquiv(0)	WorkersEquiv(1)	WorkersEquiv(2)	WorkersEquiv(3)	WorkersEquiv(4)
Pov_risk(0.0)	0.0391342132421	0.0952445492577	0.111580878585	0.355201618	0.410812173837
Pov_risk(1.0)	0.118684482103	0.175372493743	0.213573375971	0.249428740896	0.263002675087
Pov_risk(2.0)	0.842181304655	0.729382956999	0.674845745444	0.395369641103	0.326185151077

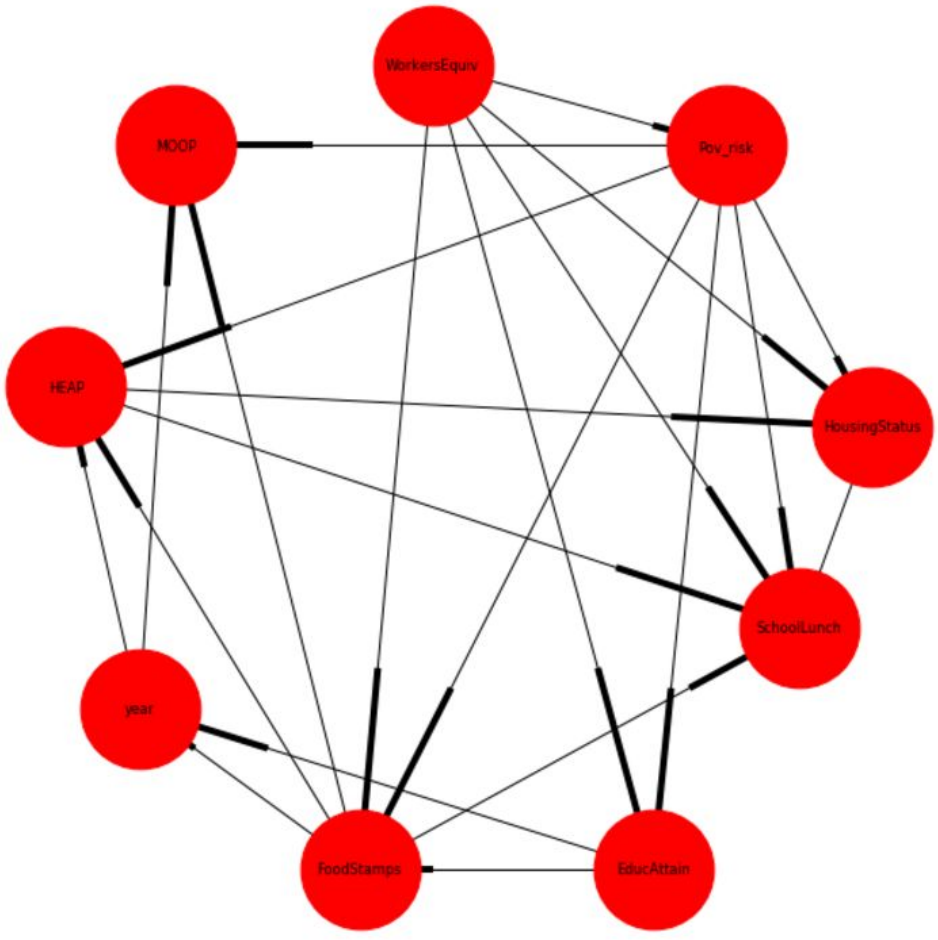
The classification report of the 10 Bayesian network models is given in Appendix II. The precision and the recall for the 'At_Risk' label is low for all years. This is because the data is skewed, the number of actual households with true label 'At_Risk' is less in the datasets. A resampling of the dataset would have helped in improving the metrics.

Calculating the Bayesian Network is also computationally expensive. The total entries for all the 10 years is 257,394. So, testing the network was not possible for all the 10 year data at once.

This study aims to presents an approximate method for classification of households at risk of poverty considering the accuracy of the data, hence Bayesian Model is selected as it gives a probabilistic estimation as opposed to a deterministic approach. The limitation of the package used for modelling Bayesian Network was 'pgmpy' which converges at local maxima.

The individual Bayesian network models created from the 10 years do show some plausible dependencies, however the structure is quite random. The structures can be found here: https://github.com/vishelar/Predicting_Poverty/blob/master/Notebooks/Data_munging_vish_bayes_net_indi_years.ipynb

Figure 7: Bayesian Network created from 10 years of merged data



Conclusion

New York City spends millions of dollars in poverty-alleviation measures each year, yet the poverty rate in the city has not improved in the last ten years. The motivation behind this project was that understanding the factors that put households at risk of poverty would allow us to design early, more targeted interventions. Understanding poverty risk is utmost important since it informs how the government should direct scarce resources to reach the poor and those at risk of poverty and help lift them into more prosperous well being. At the same time, as the results of our project have demonstrated, a lot of care should be exercised in interpreting results of predictive models. For instance, while some of our models have identified conditional dependencies between certain benefit programs, such as food stamps and HEAP, and the target variable we should not assume causality between a household being a recipient of social assistance and being at a greater risk of poverty. The model simply indicates that a household that benefits from social assistance might also be at risk of poverty.

Further Research

Poverty measurement is difficult task given various approaches and indicators. The study presents a simple yet effective data-driven approach to identify, understand and respond to the population at risk of poverty. A different machine learning approach using unsupervised learning could be used to evaluate the results of supervised learning methods used in the study and also understand the spatial autocorrelation among these geographic areas which is present and can be seen in figure 1. A different level of aggregation can also be done and the results could be compared to further understand how these variables interact at different levels.

The Bayesian network could be further improved by using prior knowledge and

providing a structure to the 'pgmpy' package and then using the data, the structure can be further optimised.

Contributions

We worked closely as a team from the beginning to the end of the project, meeting every week to discuss the progress. The code written for the project is available on Github: https://github.com/vishelar/Predicting_Poverty. Our main contributions to the project were the following:

Anastasia: project idea, domain knowledge, defining research questions, framing the problem, interpretation, writing the report, presentation.

Shalmali: data acquisition and cleaning, implementation of Decision Tree in Python, interpretation, visualization, writing the report, presentation.

Vishwajeet: data acquisition and cleaning, implementation of Bayesian Network in Python, interpretation, visualization, writing the report, presentation.

APPENDIX I

Variable Descriptions:

PovunitType - 1 Husband/Wife + child
2 Husband/Wife no child
3 Single Male + child
4 Single Female + child
5 Male unit head, no child
6 Female unit head, no child
7 Unrelated Indiv w/others
8 Unrelated Indiv Alone
The household configuration is represented from this variable.

HousingStatus - 1 Renter - Public Housing
2 Renter - Mitchell Lama Rental
3 Renter - Tenant-Based Subsidy
4 Renter - Rent Regulated
5 Renter - Other Regulated
6 Renter - Market Rate
7 Renter - No Cash Rent
8 Owner - Owned Free & Clear
9 Owner - Paying Mortgage

WorkersEquiv - 1 Two FT Workers (3500+ Hrs)
2 One FT + One PT Worker (> 2340 & < 3500 Hrs)
3 One FT Worker (>= 1750 & <= 2340 Hrs)
4 Less than One FT Worker (< 1750 Hrs)
5 No Hrs Worked

EducAttain - 1 less than High School
2 High School Degree
3 Some College
4 Bachelors Degree or higher
The highest education attainment is considered within the household.
AgeCateg - 1 Under 18
2 18 to 64
3 65+
Expenses
Childcare - Childcare Expenses in DOllar Amount
Commuting - Commuting Expenses in Dollar Amount
MOOP - Medical Out-of-Pocket Expenses in Dollar Amount

Benefits
Housing - Housing Adjustment Benefit in Dollar Amounts

HEAP - Home Energy Assistance Program Benefits
Food Stamps - Food stamps program benefits
School Lunch - School lunch program benefits
School Breakfast - School breakfast program
WIC - Women Infant Children benefits
HEAP - Home Energy Assistance Program Benefit

APPENDIX II

Classification Report for Bayes Net

2014

	precision	recall	f1-score	support
Poor	0.72	0.60	0.66	1060
At_Risk	0.48	0.22	0.31	1131
Not_Poor	0.75	0.94	0.83	3130
avg / total	0.69	0.72	0.69	5321

2013

	precision	recall	f1-score	support
Poor	0.71	0.65	0.68	1041
At_Risk	0.49	0.20	0.28	1085
Not_Poor	0.76	0.94	0.84	3123
avg / total	0.69	0.73	0.69	5249

2012

	precision	recall	f1-score	support
Poor	0.51	0.40	0.45	1090
At_Risk	0.45	0.19	0.27	1151
Not_Poor	0.71	0.93	0.81	3048
avg / total	0.61	0.66	0.62	5289

2011

	precision	recall	f1-score	support
Poor	0.49	0.49	0.49	1040
At_Risk	0.41	0.11	0.18	1121
Not_Poor	0.72	0.91	0.81	3047
avg / total	0.61	0.66	0.61	5208

2010

	precision	recall	f1-score	support
Poor	0.53	0.48	0.50	1050
At_Risk	0.46	0.23	0.31	1126
Not_Poor	0.74	0.90	0.81	3057
avg / total	0.64	0.67	0.64	5233

2009

	precision	recall	f1-score	support
Poor	0.55	0.49	0.52	1025
At_Risk	0.48	0.13	0.20	1016
Not_Poor	0.74	0.94	0.83	3107
avg / total	0.65	0.69	0.64	5148

2008

	precision	recall	f1-score	support
Poor	0.54	0.48	0.51	932
At_Risk	0.48	0.12	0.19	945
Not_Poor	0.75	0.95	0.84	3177
avg / total	0.66	0.71	0.66	5054

2007

	precision	recall	f1-score	support
Poor	0.53	0.43	0.47	916
At_Risk	0.48	0.11	0.18	996
Not_Poor	0.73	0.95	0.83	3102
avg / total	0.65	0.69	0.63	5014

2006

	precision	recall	f1-score	support
Poor	0.53	0.47	0.50	946
At_Risk	0.49	0.12	0.19	926
Not_Poor	0.75	0.95	0.84	3154
avg / total	0.66	0.71	0.66	5026

2005

	precision	recall	f1-score	support
Poor	0.57	0.52	0.54	976
At_Risk	0.47	0.12	0.19	865
Not_Poor	0.77	0.95	0.85	3099
avg / total	0.68	0.72	0.67	4940