BIG DATA ANALYTICS PROJECT - NO POVERTY

Shubham Bagi, Anurag Pradhan, GuruSangama Prasad, Anirudh M

INTRODUCTION

Eradicating poverty in all its forms remains one of the greatest challenges faced by mankind. Poverty has many dimensions, but its causes and its effects include unemployment, social exclusion, and high vulnerability of certain populations to disasters, diseases and other phenomena which prevent them from being productive. So that is why the United Nations came up with the idea of **Sustainable Development Goals (SGDs)** goals which needs to be met for the peace and harmony of mankind. One such Goal (SGD 1) is based on No Poverty and this is to be taken very seriously because over 10% of the world population, still live in extreme poverty and are struggling to fulfil the most basic needs like health, education, and access to water and sanitation[1]. Ultimately, SGD's goal is to **end poverty** 2030.

The academic and education community play a crucial role when it comes to the impact of poverty. Science provides the base for new, sustainable methods, solutions, and technologies to tackle the problem. For example, it has enabled access to safe drinking water, and improved hygiene to reduce health risks related to unsafe drinking water and lack of sanitation. So, we as engineers should try to find solutions which get us closer to the SGD Goal. In this project we are trying to analyze how Poverty and its factors shape up in each county and then we try to deduce some inference based on it.

BACKGROUND

The 2016 Census states that there are almost 40 million Americans who live below the Poverty line, This accounts to almost 14 percent of the whole population. According to the research conducted by the Centre for Social Justice, there are five Pathways to Poverty[10]. They are family breakdown, education, worklessness and dependency, addiction, and serious personal debt. These five factors are intimately interconnected and contribute to the complexity of the problem, which requires integrated, innovative solutions. **Unemployment** – One of the most prominent ways to defeat Poverty is by providing employment to the people. **Education** - Each year, 1.2 million students in the United States drop out of high school, putting them at great risk for drug use, unemployment. This needs to be stopped, as this will lead straight to kids being unemployed in the future and further leading to poverty. **Health and Addiction** – According to a report released by the surgeon in 2016, 1-in-7 Americans are expected to develop a substance use disorder at some point, and while substance use disorders affect millions of Americans, only 10 percent receive treatment. **Population Density** – According to a paper on Population Growth and Poverty by Dennis A Ahlburg[9] there is some direct evidence on the impact of population growth on Povertys. This is mainly due to diminutive per capita resource availability.

Looking at the targets of the SGD set by the United Nations, we notice that emphasis is given to the No Poverty goal and targets to achieve the goal have been mentioned. Few targets to be met by 2030 are:

1. To eradicate Extreme poverty for all the people around the world.
2. To reduce the percentage of men, women and children who are living in poverty by **half**.
3. To provide a few social protection systems for **specially-challenged/disabled** people.

4. To create sound policy frameworks for **allocation of resources**[11] at the national, regional, and international levels to support accelerated investment in poverty eradication actions.

By looking at these factors we have identified that **Disability, Unemployment, Literacy/Education, Population Density, Hunger** will be few of our main features in identifying the Poverty rate of each county of the United States and how other counties in the US can improve on these features in order to have a minimum Poverty percentage.

DATA

Data Source	Category	Details
US Department of Agriculture [2]	Poverty	2013-18, 34 Features
	Population	2012-19, 175 Features
	Unemployment	2002-18, 88 Features
	Education	1970-2018, 47 Features
	Area [8]	3k rows
National Center for Education Statistics [3]	Schools	2015-19, 102k rows, 25 Features
Country Health Rankings & Roadmaps [4]	Health Stats	2014-20, 270 Features
Social Security Website [5]	Disability	1998-2020, 10 Features
Homeland Infrastructure Foundation-Level Data [6]	Education	7k rows, 32 Features
U.S. Department of Housing and Urban Development [7]	Poverty Index	73k rows, 12 Features

Table 1: Dataset sources and details

Some of the data sets we used in our computation is from the US Department of Agriculture [2]. USDA has datasets for many socioeconomic indicators like poverty rates, population change, unemployment rates, and education levels vary geographically across U.S. States and counties. Each county has a **Federal Information Processing Standards (FIPS)** county code associated with it. The dataset has almost **150** features and there are many different categories like Poverty, Population, Unemployment, Education, and the County area. All these factors totally contribute towards a county's Poverty Index.

Another dataset which we considered getting information about Schools comes from the National Center for Education Statistics [3]. This ranges from the year 2015 to 2019 and has over 100,000 rows. For Health Statistics, Disability and Education, we considered the dataset from the federal sites [4][5][6].

In addition to the percentage of people below the poverty line, we have taken into consideration another Poverty Index, obtained from the U.S. Department of Housing and Urban Development [7]. Poverty Index is an indication of the poverty of community in a country, developed by the United Nations which

concentrates on the deprivation in the three essential elements of human life already reflected in the HDI: longevity, knowledge and a decent standard of living.

METHOD

After obtaining the data files (around 12 GB) from various data sources as mentioned before, we dumped it into the **HDFS** and then proceeded to processing in a distributed fashion using **Spark** to transform the data. All the individual data sources had to be processed separately as each of them have a different structure/format. After processing each dataset, we finally dumped it back to HDFS using a sequence number. The next step was to merge the features county wise. Each county has a designated unique FIPS county code. Almost all the datasets had the FIPS county code, but for the data sets with FIPS county code missing, we made use of broadcast variables and tagged the FIPS county code using the combination of state and county name. This broadcast variable was being created from another dataset which had both FIPS county code, county name and state name.

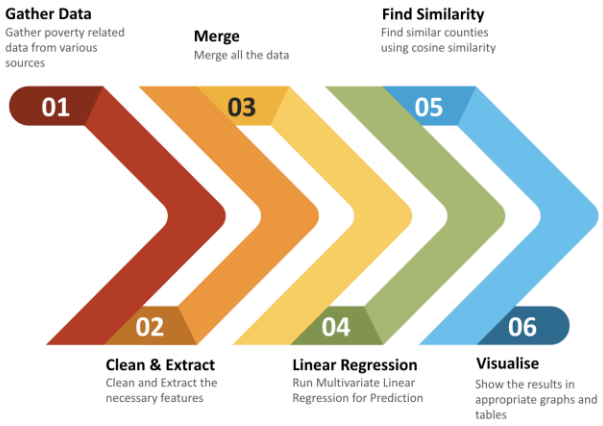

Figure 1. All the methods involved in a glance.

The next challenge was to keep the same sequence of features while merging. So, we identified each dataset from the sequence, and used it while dumping back into the HDFS. During merging, we used Spark again to basically group the data county wise and year wise. The sequence number allowed us to maintain the correct sequence of features while merging data of the same county. Also for some of the features like schools and hospitals, we converted it in terms of per million, in a step toward normalising the data. Like for FIPS county code, we made use of a broadcast variable to store the population county wise. Now we have the data on which we can run regression and cosine similarity. After normalising and clubbing together some features, we finally took into consideration 33 features for analysis. And we picked the features which were relevant to resource allocation over a vast pool of features.

The degree of impact each parameter has on poverty is obtained in the form of the beta values by using **Multivariate Linear Regression** with the percentage of people below poverty being the Y and the feature vector being X. The set of beta values obtained from this are then used to get a standardised poverty score by vector multiplication of beta with the feature vector of a county. The feature vector will change when a new resource is allocated to a county. The updated feature vector results in an updated poverty score. The

difference between the updated and the old poverty score gives the change in poverty in the county because of resource allocation. By repeating the same process for all the counties and computing the change in poverty scores, a county with the best poverty change score is selected as the ideal county for the allocation of that resource. So this method of using multivariate linear regression provides a way to find out an ideal county for resource allocation.

If a county is known of its resource requirement and if there is a need to know the counties which may have similar resource requirements, a cosine similarity method is used to get the results. A **Cosine Similarity** (cosine distance) of the given county with the rest of the counties is computed which gives similarity scores with respect to each county. These scores are sorted and the counties with best scores are recommended as the counties with similar resource requirements. The methods described can be used to help the policy/decision makers to allocate resources/formulate policies that can help in alleviating poverty to a greater extent. Given the set of resources (budget for the resources), the first method provides a way to find out the counties which may be best fit for the resource allocations. Given a county and its resource requirement is known, the second method helps to find out the counties which may be in need of similar resources.

RESULTS

The results have been generated for the above analysis wherein an experiment is carried out to increase the number of schools and study the impact the change has on poverty. Number of schools per million is one of the features in the feature set. An increase in the number of schools changes the primary feature (Number of schools per million) and also a few other parameters (called as secondary features) such as unemployment rate and literacy rate which are also a part of the feature set. So a correlation between the primary feature and secondary features is considered in changing the secondary features. The values of the primary and secondary features are changed to obtain an updated feature vector and thereby an updated poverty score. A plot of change in the poverty score for various counties is obtained.

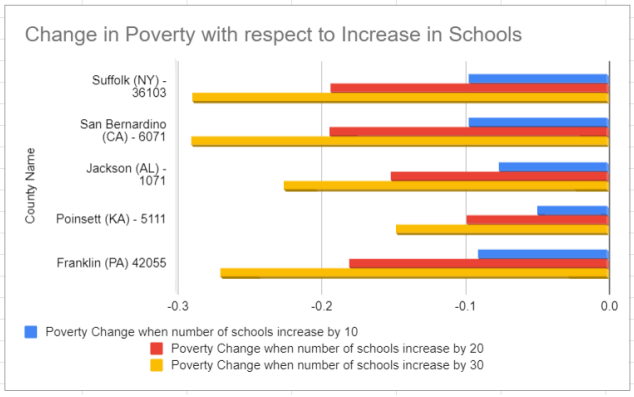

Figure 2. Change in poverty among different counties with respect to increase in schools.

The figure 2 shows the change in poverty scores in 5 counties when the 10 schools, 20 schools and 30 schools are allocated to the counties represented by blue ,red, yellow bars respectively. Each county's

poverty score was changed on the resource allocation, and the county where the change was highest was given as the ideal county for that resource allocation.

The earlier method provided a way to find the county to allocate resources for. Here given a county and its resource demands, we find the counties which may be in similar condition and may need similar resources. A cosine similarity between a given county and all the other counties is computed to get similarity scores which are sorted to get the similar counties.

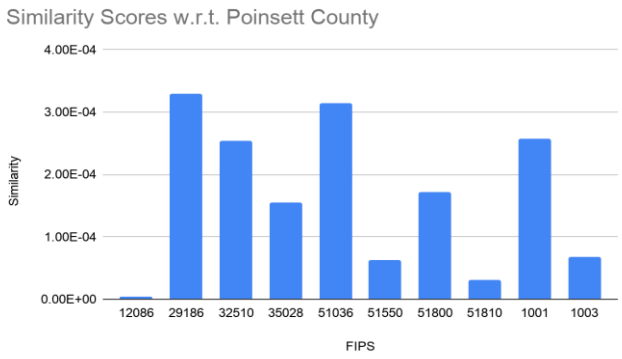

Figure 3. Similarity Scores with respect to Poinsett County

Figure 3 shows the similarity scores of a few counties with respect to poinsett county (FIPS-5111). The horizontal axis provides the county details in the form of FIPS code and the vertical axis provides a linear measure of similarity score.

Counties similar to San Bernardino (CA) - 6071

FIPS_CODE	% Literacy	% Unemployment	Median Income	Population	Hospitals	Poverty Index	Public Schools	Food Insecure %	Disability	Population Density	% Below Poverty
1005	25.1	3.8	34382	24686	40.50879041	21.11111111	364.5791137	22	7.577178781	12562.57033	30.9
1053	24.4	3.5	38418	36633	54.59558322	20.11111111	436.7646657	18.4	10.8419751	12640.17694	23.6
1087	30.9	4.3	32495	18068	55.3464689	23.33333333	442.7717512	25.6	8.088721919	9264.289317	30.2
1093	31.7	3.4	37887	29709	67.31966744	29.125	504.8975058	13.3	10.54203724	9361.297583	20.2
1109	24.2	3.3	37259	33114	30.1987075	31.125	332.1857824	21	13.69880215	13746.07117	23.6
1119	24.6	4.5	27859	12427	80.46994448	21	563.2896113	27.7	3.718717152	9167.281052	34.7
1131	24.1	7.1	25385	10373	96.4041261	7.5	578.4247566	29.3	3.165403522	12514.0662	33.4
5005	34.8	4	39686	41932	23.84813508	38.88888889	238.4813508	13.9	17.89177403	9438.904195	15.1
5107	34.3	6	29945	17782	56.2366438	9.5	506.1297942	29.5	6.406850593	15065.38357	35.4
5123	31.2	5.2	33257	24994	40.0096023	19	400.096023	25.7	10.05139177	16394.3968	35.6
6071	32.8	3.8	63310	2180085	14.67832676	40.91598916	267.8794634	10.1	28.13216118	25659.68072	14.9

Figure 4. Counties similar to San Bernardino (CA)

Figure 4 shows a sample of the feature dataset used with a sample set of features. The percentage below poverty was the Y value used for multivariate linear regression. The entire table is an extract from the counties similar to county **San Bernardino** (FIPS-6071).

CONCLUSION

By 2030, the UN, under the SGD goal of "No Poverty", wants to ensure that all men and women, in particular the poor and the vulnerable, have equal rights to economic resources, access to basic services, natural resources, appropriate new technology and financial services. Through our project, we have utilized big data concepts like **HDFS**, **Spark**, **Multivariate Linear Regression** and **Cosine Similarity** to aid in realising this goal. We have used HDFS and spark to clean and explore the data to gain insights about various features affecting poverty. Then we have used linear regression and cosine similarity to develop a methodology that will help governments and organisations to allocate aforementioned resources among different subjects(like counties, states) in an economical and efficient manner.

REFERENCES

1. https://www.dosomething.org/us/facts/11-facts-about-global-poverty#fn1
2. https://www.ers.usda.gov/data-products/county-level-data-sets/
3. http://hudgis-hud.opendata.arcgis.com/datasets/low-poverty-index/data
4. https://nces.ed.gov/programs/edge/Geographic/SchoolLocations
5. https://www.countyhealthrankings.org/
6. https://www.ssa.gov/policy/docs/statcomps/ssi_sc/
7. https://hifld-geoplatform.opendata.arcgis.com/datasets/hospitals
8. https://www.nass.usda.gov/Publications/AgCensus/2012/Online_Resources/Ag_Atlas_Maps/mapfiles/ag_co_metadata_faq_12.html
9. http://www.ilo.org/wcmsp5/groups/public/---ed_emp/---ifp_skills/documents/publication/wcms_107921.pdf
10. https://www.centreforsocialjustice.org.uk/policy/breakthrough-britain
11. https://www.un.org/sustainabledevelopment/poverty/