Analyzing and Preventing Poverty in America

Matthew Matero and **Kiranmayi Kasarapu** and **Dibyajyoti Pati** and **Neha Indraniya**

Abstract

With the advancement of computers and applications being able to work with larger sets of data, we are increasingly able to perform more complex analysis. We aim to try and build accurate predictive models of the poverty level across America. First we examine poverty levels through linear models and then compare to deep learning. As well as perform a clustering to compare job markets.

1 Introduction

Our problem arises from the need to accurately know the percentage of people living below the poverty line and how we can improve their lives. While there already exist a few variations on predicting poverty we believe they are not looking at the correct predictors and can improve upon them.

To begin we are going to analyze our dataset using common linear models. With features selected from a mix of the world bank's world development indicators and the bureau of labor statistics, we believe we can not only compete with widely accepted poverty models but add insight into how we can improve them.

We also believe that there is a possibility that using a deep learning recursive neural network may lead to a more precise estimation the nature of the data over time. Although we build this model more for exploratory purposes to see if there is any promise continuing NNs in this area.

Lastly, we explore a clustering around job markets to see which industries seem to hold promise for growth and advancement.

The application of this work is to improve the lives of the many people still living in poverty. With an accurate prediction we can direct discussion around how we can pro-actively help reduce this number. Our clustering of job markets also aims to help direct people to how they can better their skill set to improve their economic standing.

2 General Information

The UN's sustainable development goals include various measurements of world health, and prosperity(UnitedNations, 2015). These goals are related to not only the well-being of people by trying to amplify quality of life, freedom, and education but also improve the health of our planet. This is accomplished through tracking aquatic life, climate change, and advancements in renewable energy.

2.1 Sustainable Development Goal

Our focus is on goal #1, which is to eradicate extreme poverty by 2030. Specifically one of our main focuses will be on goal 1.2. This aims to reduce the amount of people under the poverty line by 50 percent defined by their national definition. This is the reason we will be monitoring poverty rates with our predictive models. We want to firstly be able to build accurate models, then we can use them to see how close we are to reaching this goal if we continue on our current path.

2.2 Background

As previously mentioned we will be using a combination of linear models and neural networks.

Our clustering approach will be using K-Means. To create clusters around industries

We will be using a RNN. Due to having data over time we feel this may capture a good representation of the data.

2.3 Data

There were two core data repositories used to create our feature set. The main collection we worked with was compiled of many surveys from the bureau of labor statistics. We supplemented this with a version of the World Bank's World Development Indicators found on Kaggle.com. The total raw size of our available data is roughly 16GB.

Each survey in the labor statistic set varied from being of the size 10-50MB to a few hundred MB. We planned on originally using the more specific sets for state and county level data, but time constraints only allowed us to work on the national level.

3 Methods

3.1 Data Processing

We worked with a handful of surveys from the labor stats collection to first build a feature set that included job openings, closings, gross job gain, gross job loss, industry, year, quarter, region, # strikes/work stoppages and CPI. Data from the WDIs were then joined in to grab the average unemployment rate for men and woman across America. We had information from the years 2000 - 2007, spread across 4 quarters per year, giving a total of 512 observations at the national level. This data was also replicated at the state level for each of the 50 states, giving a total of roughly 25,000 observations across all states.

The data files were processed using Python/PySpark running on a 3 machine cluster(4core/8gb ram), to assist with the quick loading and processing of the total data set.

3.2 Linear Models

We used Linear Regression to create a model for poverty prediction. With the preprocessing of the data, we arrived at certain feature variables which affect the poverty rates. Some of the features are unemployment percentages, consumer product index, job openings, gains and losses etc. We used the data

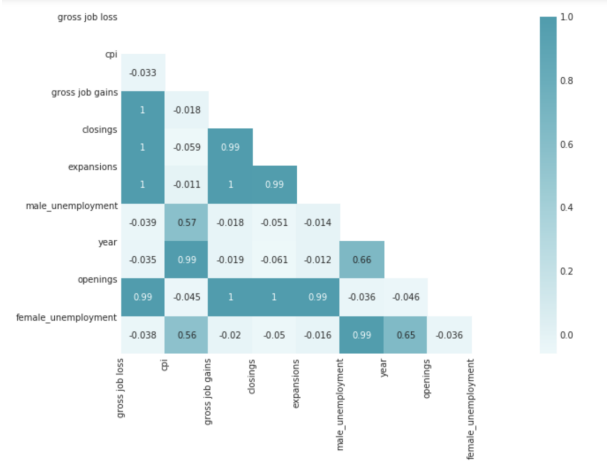

Figure 1: Correlation Matrix

Figure 2: Comparison of Unemployment rates between men and women

to calculate the correlation among the features which is shown in figure 1

After analyzing the data carefully, we see that unemployment rate is highly and positively correlated to poverty and job openings and gains in employment are negatively correlated which is expected. The results are shown in figure 3

To train the model, we divided the data set into training and testing data, and we repeated the process several number of times with different training and testing points, so that each data observation is used in the training process. We combined the results and took the average of the predictions. The results are shown in section 4.

3.3 Auto regression

We analyzed the data to forecast the Job gains per sector(We determined that to be one of the most influential factor in poverty).We started off by check-

ing for stationarity in the time series i.e.In a stationary time series, the statistical properties over time remains constant and auto co-variance should be time independent.When running a regression, we expect the observations to be independent of each other ,while in a time series we know that the observations are time dependent. So, to use regression methods on time dependent variables, the data must be stationary. The techniques that apply to independent random variables also apply to stationary random variables.

We employed Dickey-Fuller test, a statistical test with the null hypothesis that the time series is non-stationary to understand if the time-series is indeed stationary.We also analyzed the trend and seasonality by applying decomposition on the Job gains data.In the next steps we plotted the auto correlation function(ACF) and partial auto correlation function(PACF) to determine the optimal values for p,d,q for the ARIMA model.And Finally a model was created for each sector based on the quarterly job gain reports for each sector to successfully predict the Job gains in the upcoming quarter.

3.4 Clustering

We analyzed the data from multiple perspectives to perform our clustering. For all of them we used K-means approach to create a predefined amount of groupings.

The results we are focusing on are finding emerging industries and employment trends by industry. We also go deeper into emerging industries by looking at a yearly change in job growth.

4 Results and Discussion

4.1 Linear Models

We approached the problem by training a multiple linear regression model to try and predict poverty to the best of our abilities. Since our data spanned years 2000 - 2007 we compared against those values.

The trained model was able to come very close to the actual label values for each year. We also ran analysis to see which features most highly correlated with the poverty rate. By representing the correlations visually it is easy to see which features are most correlated.

```
Mean Square Error 0.0052
R^2 Score         0.9755
```

year	ActualPercentage	PredictedPercentage
2000	11.3	11.359941
2001	11.7	11.610465
2002	12.1	12.109128
2003	12.5	12.547251
2004	12.7	12.634235
2005	12.6	12.587305
2006	12.3	12.423442
2007	12.5	12.431480

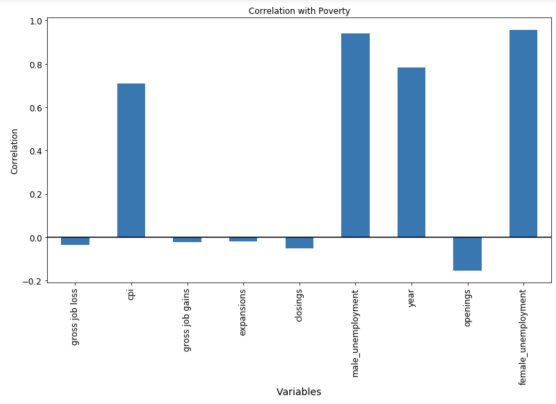

Figure 3: Correlation of poverty with various features

4.2 Auto regression

In this section we analyze the most important and correlated feature of poverty that is sector wise job gain.These numbers are normalized with respect to the total employment in the particular sector before modelling. Here is how the gross job loss and job gains figure look for ”Information” sector

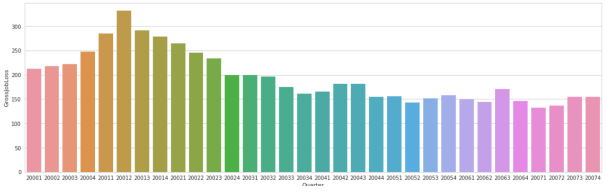

Figure 4: Quarterly Job loss in Information sector

Next, we checked for stationarity in the time series by plotting the Rolling mean & rolling standard deviation along with the original time series. We also performed Dickey-Fuller test to determine the

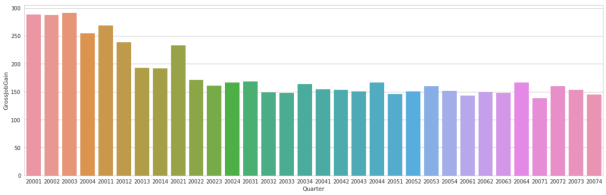

Figure 5: Quarterly Job Gains in Information sector

p-value to check if the time series is stationary.

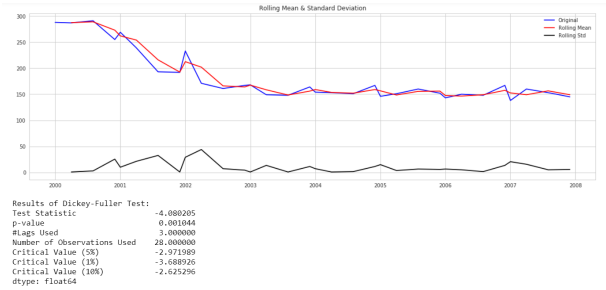

Figure 6: Job Gains with rolling mean and standard deviation

The p-value computed was well within the 5% significance we were testing for, So, the series is indeed stationary And we proceed with the decomposition of the data for analyzing the trend and seasonality in it.

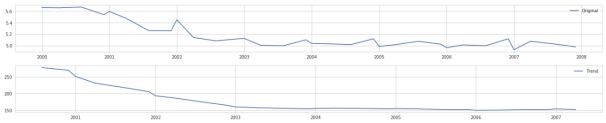

Figure 7: Trend from time series by decomposition

The trend shows a relative drop in the job gains in the recent years in the Information sector (normalized with total population in that sector). We determined the auto correlation and the partial auto correlation to determine the optimal parameters for ARIMA model

By looking at the ACF and PACF, the model parameters were initially set to (4,1,4), but later changed to (4,1,2) due to non-convergence

This process was followed for each of the sector using spark. The forecast accuracy result was aggregated from the model in the action step.

4.3 Clustering

Firstly, we examined job growth by industry to see which industries were growing the fastest in our data

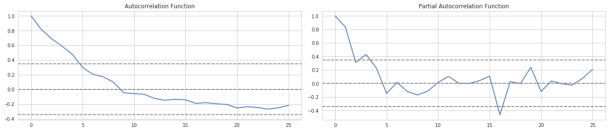

Figure 8: ACF and PACF

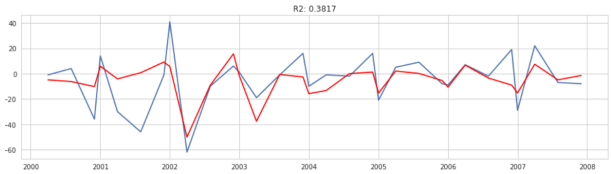

Figure 9: Auto reg model

set. After getting a better understanding of the data through visualizations the next step was to look at the yearly changes of job gains by each industry.

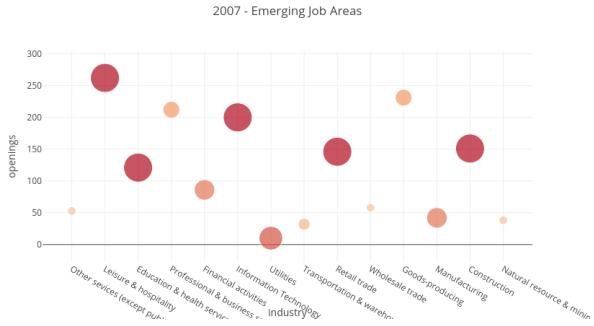

Figure 10: Growing Industries

We can see that many industries seemed to have a steady decline or growth over the years. While there were industries such as leisure and hospitality that had a checkered growth.

5 Conclusion

We believe the work we have done to have been a success. Our first main goal was to find a model that can accurately predict the poverty level, and gain insights into what predictors are the best telling. From the work of our linear models we see that we were able to complete both of these tasks.

As well as knowing where the poverty level is headed, our analysis of industries and job openings offer great insight for those looking for work. Given the information of which industries are growing, it is possible for someone to better improve their skills to fill these roles. For example, it is clear the infor-

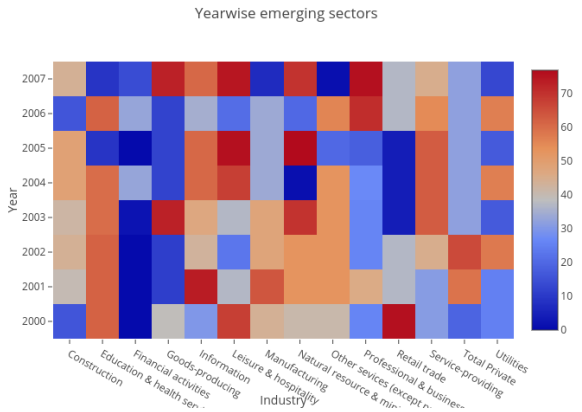

Figure 11: Industry Trends by Year

mation industry has been growing so someone may want to redefine their skill set to better market themselves to these jobs.

Acknowledgments

Distribution of work:

- Matt - Data Processing/Feature Generation
- Kiran - Linear Modeling + Visualizations
- Dibya - Auto Regression + Visualizations
- Neha - Kmeans Cluster

Cluster Information:

- AWS Elastic Map Reduce Instance
- 3 M4.large instances(4Core,8Gb ram each)

References

UnitedNations. 2015. Transforming our world: the 2030 agenda for sustainable development.