Income Category Prediction

Revanth Patil
Computer Science
PES University
Bangalore, India
revanthpatil@gmail.com

Rakeshgowda D S
Computer Science
PES University
Bangalore, India
abhirakeshgowdads@gmail.com

Rishab S
Computer Science
PES University
Bangalore, India
rishab.santosh189@gmail.com

Rakshitha N
Computer Science
PES University
Bangalore, India
rakshithanagesh29@gmail.com

***Abstract*—Income inequality is present in the United States and it is a major concern. People with certain set of features seem to be earning more compared to other people. Through various Machine Learning techniques, this study aims to show the attributes that best explain the income inequality that is present in the United States. The dataset used here is obtained from Kaggle.com which in turn was provided by the U.S. Census Bureau.**

***Index Terms*—income inequality, classification, Support Vector Machine, Gradient Boosting Classifier, Naïve Bayesian**

I. Introduction

A. How does income inequality affect our lives?

The big question is "is income inequality good or bad for the society and economy?". To put it in another way, does people like Mark Zuckerberg getting richer affect the life of a typical citizen? Some argue that it does in a bad way and some argue that it does in a good way. Let us look at both points of view.

B. How is it good?

One very obvious way is how it allows entrepreneurs to take bigger risks. For example, Steve Jobs and others like him wouldn't have taken the risk of starting a company and running it if it was not worth it. It allowed him to accumulate tremendous amount of wealth. If there was no permission for him, or others for that matter, to accumulate such massive amounts of capital, then very few people would have started a company. Which would mean lesser jobs and lesser economic growth. So some people argue that income inequality is necessary for economic growth.

C. How is it bad?

Numerous studies have shown that excessive inequality is bad for growth. One study points out that low-earning families invest less in education and skills. This means that there is reduced number of skilled workers which results in slow economic growth as less productive work is being done. It is also observed that people that go to universities tend to earn more than the people who do not. Now, this is a serious issue because the rich stay rich by sending their kids to universities and the poor stay poor because they cannot do so. This widening gap results in other serious problems. Some of the issues can be explained by how different economic groups behave and interact with each other as the income gap widens.

How income inequality affects the way different groups behave:
a. The poor invest less in education and health. This causes problems to the economy if they are more in number.
b. If inequality squeezes the middle class then it reduces their demand for goods and services.
c. The rich may use their economic power to lobby against the rules that are against their interests at the expense of greater social good.

How income inequality affects the way different groups interact:
a. Trust is lowered between the groups leading to high transaction costs. For example, if a customer trusts the business owner and vice versa, then they may agree to a deal without any legal advice or contracts.
b. People's network of social relationships might not extend beyond their own income group. Hence, the elite groups might use their network to exclude outsiders from economic opportunities.
c. Social unrest and volatility may be caused because high levels of inequality leads to societies not coming to a political consensus.

As can be inferred from the above explanations, extreme income inequality on a large scale is definitely not good. Our study is focused on finding the factors that cause this inequality with the dataset we have. Before taking a look at our approach, we study what others have done with this dataset.

II. Literature survey

The first study has used various classification algorithms like Naïve Bayesian, K Star, Random Forest and Zero R. They have used the first 100 instances of the dataset and the instances with missing values are removed. The less sensitive attributes like final weight, capital gain, capital loss, hours per week are removed. They have used Weka, which is a collection of machine learning algorithms for data mining tasks. After running all the classification algorithms, they conclude that Naïve Bayesian is the best with an accuracy of 84.31%. Though Naïve Bayesian is followed by Zero R then Random Forest and K Star, these three algorithms remain in

the same range.[3]

The second study uses Principal Component Analysis, PCA and Support Vector Machine, SVM for the classification. Scaling of the features in the range of [-1,1] is done because large values can cause computational problems and may also lead to wrong results. As SVM deals very well with real numbers, it is inferred that scaling is a much better approach than discretization. Accuracy of 84.92% is obtained.[4]

The third study used random forest classifier to predict income levels of an individual. Random forest is preferred to decision tree since using results from many decision trees will avoid the overfitting problem associated with using a decision tree classifier. Results from the fitted classifier model show that marital status, capital gain, education, age and work hours determine much of the difference between low and high income levels. Though the model accuracy is 85%, the model is weak in predicting high income individuals.[1]

The fourth study has used various data pre-processing techniques like label encoding, one-hot encoding, and shuffling. Based on the scores of the Extra Tree Classifier for different attributes the most relevant features have been selected. The learning algorithm, used to build the predictive model is an Ensemble Learning and Boosting Algorithm known as Gradient Boosting Classifier. The model was deployed which clocked the highest accuracy of 88.16%, eventually breaking the benchmark accuracy of existing works.[2]

III. Our approach

A. Problem Statement

To analyze the dataset to find the attributes that cause higher income in some individuals and to classify income level as $\leq 50k$ or $> 50k$ from the dataset using the relevant attributes.

B. Dataset

Our dataset is taken from Kaggle.com which was provided by the U.S. Census Bureau. It contains 15 attributes and 32561 observations. The target variable is income which is binary, is either $\leq 50k$ or $> 50k$. "Fig. 1" shows a snapshot of the dataset being used.

class	fnlwgt	education	education.num	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.country	income
?	77053	HS-grad	9	Widowed	?	Not-in-family	White	Female	0	4356	40	United-States	<=50K
ivate	132870	HS-grad	9	Widowed	Exec-managerial	Not-in-family	White	Female	0	4356	18	United-States	<=50K
?	186061	Some-college	10	Widowed	?	Unmarried	Black	Female	0	4356	40	United-States	<=50K
ivate	140359	7th-8th	4	Divorced	Machine-op-inspct	Unmarried	White	Female	0	3900	40	United-States	<=50K
ivate	264663	Some-college	10	Separated	Prof-specialty	Own-child	White	Female	0	3900	40	United-States	<=50K

Fig. 1. Snapshot of the dataset

C. Approach

We will use many different approaches for cleaning and preprocessing data like outlier detection and removal. Different classification algorithms will be tried out. Through this study we aim to get an accuracy of more than 88%. We also aim to throw light on the attributes that are most responsible for the income inequality.

References

[1] Sisay Menji Bekena. "Using decision tree classifier to predict income levels". In: (2017).

[2] Navoneel Chakrabarty and Sanket Biswas. "A statistical approach to adult census income level prediction". In: *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. IEEE. 2018, pp. 207–212.

[3] S Deepajothi and S Selvarajan. "A comparative study of classification techniques on adult data set". In: *International Journal of Engineering Research & Technology (IJERT)* 1 (2012).

[4] Alina Lazar. "Income prediction via support vector machine." In: *ICMLA*. Citeseer. 2004, pp. 143–149.