INVESTIGATING THE INHERENT REASONS FOR INCOME DISPARITIES IN THE U.S.

ABSTRACT

Income inequality refers to the unequal distribution of income within a society, where a small percentage of individuals or households have a significantly larger share of wealth than others. In recent decades, due to the rise of income inequality in the United States, there has also been an increase in poverty and a decrease in social mobility and economic growth. Our model aims to identify and rationalize the causes of income equality, exploring factors that contribute to the problem including age, sex, race, and education level. Understanding the root causes of income inequality can provide insight into creating a more equitable society as well as developing effective policy solutions to mitigate this problem.

INTRODUCTION

BACKGROUND

Income inequality is widespread throughout the United States and every country worldwide. However, it's highest in the US out of all of the G7 nations. In addition, in the US, the wealth divide between upper-income families and middle- and lower-income families is steep and rising (Schaeffer, 2020). There are a variety of factors that lead to these gaps including levels of education, gender, age, and more. Discovering which factors affect these gaps the most can help show where we need to do better work as a country to close the gaps.

PROBLEM

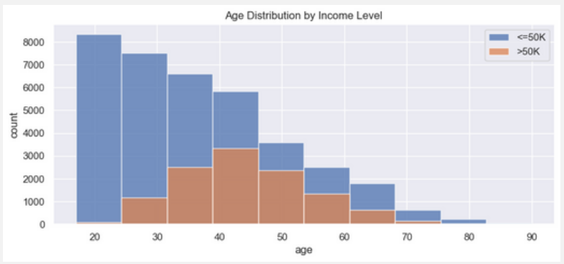
We are looking to understand the inherent reasons for income disparities from demographic data. In the United States, the average Black or Latino household make about 50% of an average white household and own 15-20% as much net worth (Aladangady & Forde, 2023). In terms of gender, in 2020, women earned around 84% of what men made (Barroso & Brown, 2022). Historically, income inequality has put money in the hands of the rich, resulting in low social and economic mobility for large portions of the population. This can have a ripple effect in relation to economic instability, financial crises, and inflation.

MOTIVATION

We want to examine the underlying drivers to income to see how those factors affect the income inequality gap that Americans face. Income inequality can affect the economy, inhibit economic growth, and increase poverty. In our goal to predict income based on key factors, we will be able to gauge how important targeting societal inequities like the wage gap and education disparities are. Data from the Federal Reserve Bank of Boston has also recently shown that there is a direct connection between inequality of outcomes and inequality of opportunity (Bradbury & Triest, 2014).

GOALS AND OBJECTIVES

The Census Income Dataset extracted from Kaggle provides information on over 48,000 US citizens on 13 predictive factors such as education level, age, sex, race, hours worked per week, and more. For each person, there is a binary income label indicating whether or not they make over $50K in a year. The goal of this project is to be able to predict whether or not a person makes over $50K a year given the 13 predictive factors.



RELATED WORK

From Kaggle, there have been five projects that people have worked on but three stand out. The Adult Income Case Study project graphically investigated the relation between income and race, gender, and education level. They reached conclusions on the gender pay gap only using simple aggregated graphs. The K-Means Clustering project focused on trying to cluster all of the income groups, but as their end goal was to simply produce the clustering model, they did not come to any direct conclusion. The last project we looked at, the ML-Donation Target project, utilized ML models (Logistic Regression on SMOTE, SVClassifier on SMOTE, and Random Forest Classifier on SMOTE) to understand the American donation behavior. They reached a few strong observations based off of the data, such as women tending to donate more compared to men and people being more likely to donate their time to the community the older they got.

METHODOLOGY

ALGORITHMS

With the binary nature of our target variable, we explored three different classifier algorithms: Random Forest Classifier, Decision Tree Classifier, and K-NN Classifier. Because our goal is to predict a binary target, we're choosing classification modes over regression models. A K-NN Classifier is reliable and accurate but not as efficient as other algorithms with larger datasets. It classifies new data points based on their proximity to other points in the training set. Since we have a large dataset, we're also looking at Decision Tree Classifiers and Random Forest Classifiers. Using both a Decision Tree and a Random Forest allows us to visualize possible outcomes of a series of choices by plotting a single Decision Tree, and then we can reduce overfitting by analyzing a Random Forest model.

MODEL TRAINING AND EVALUATION

Since we are using a classification model we split our data set using stratified k-folds to cross-validate the data to evenly train the model and balance the model's class weights to minimize biases in the data. Following that, we determined the feature importance using mean decrease in impurity for the attributes. To evaluate our model we utilized confusion matrices and classification reports to tune our model.

HYPERPARAMETER TUNING

For the Decision Tree Classifier, we adjusted max_depth, max_features, min_samples_leaf, and min_samples_split, with max_depth limiting the tree's complexity to prevent overfitting, and min_samples_split and min_samples_leaf ensuring adequate samples at each node and leaf. For the Random Forest Classifier, we employed max_depth, max_features, min_samples_leaf, min_samples_split, n_estimators, and class_weight, where increasing n_estimators enhances model accuracy, and constraining max_depth and selecting the optimal max_features prevents overfitting. Lastly, for the K-NN classifier, we applied n_neighbors and weights, which consider the influence of nearby points while making the prediction.

MEASURES OF ACCURACY

We used a classification report to evaluate the decision tree classifier and random forest classifier models, which measures precision, recall, and F1 scores for each class, as well as the overall accuracy of the model. The K-NN classifier model's results are assessed by a confusion matrix, which can be used to calculate other evaluation metrics, such as precision, recall, and F1 scores. These metrics take into account different aspects of the model's performance and can help to identify areas of improvement. We took the accuracy, sensitivity, and specificity scores of each accuracy model to identify the performance of the models as a whole.

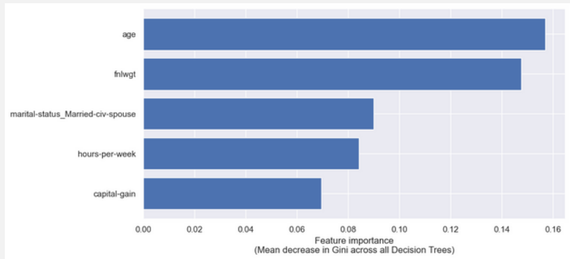
RESULTS & EVALUATION

In the context of our dataset, our target variable is income (0 = less than or equal to $50K, 1 = more than $50K), and our x features are factors such as age, workclass, education, relationship, race, and sex. After building the three models (Random Forest Classifier, Decision Tree Classifier, and K-NN Classifier), we compared a multitude of accuracy metrics to determine the best model for results evaluation. Out of the three models, the Random Forest Classifier performed best, outscoring the other models with its weighted average for precision, recall (aka sensitivity), and accuracy being .84, .85, and .85 respectively. Precision measures how many positive predictions (>50K) were true out of all actual y labels (>50K and <=50K). Recall measures a model's ability to predict true positives (<=50K) out of all predictions made.

	accuracy	sensitivity	specificity
Random Forest Classifier	0.847	0.601	0.929
Decision Tree Classifier	0.840	0.556	0.934
K-NN Classifier	0.838	0.560	0.929

The Random Forest classifier consistently shows upticks in performance compared to the other models, however, it does lag behind in specificity. Specificity measures a model's ability to predict true negatives (>50K) out of all predictions made. This usually comes at a trade-off for sensitivity, and considering the Random Forest's sensitivity is significantly larger than the other models, a 1-2% decrease in specificity is manageable.

With confidence that the Random Forest Classifier is strong enough to perform a feature importance on, we plotted the mean decrease in Gini across all trees in the forest and identified the top 5 features which influence a person's income.



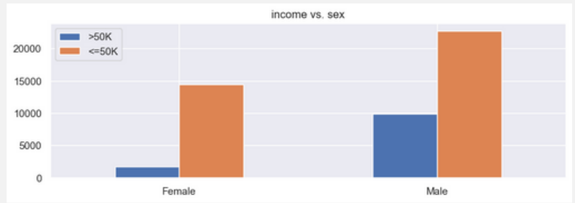
These five influential features are age, final weight, married (civilian spouse), hours worked per week, and capital gains. Final weight is an estimate of how many people in the total US population are represented by the features in each row. In this dataset, the census only includes people 16+ in age. It seems that age and final weight are the most influential factors in terms of determining income, with hours worked per week, marriage status, and capital gains trailing behind them. In order to understand the direction of this influence, we calculated the correlation of each feature to our income variable (>=50K).

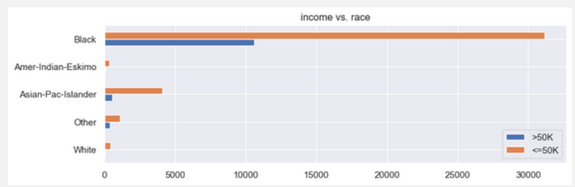
capital-gain	0.221034
marital-status_Married-civ-spouse	0.446192
hours-per-week	0.227199
age	0.237040
fnlwgt	-0.007264

Based on these results, we conclude that **a person who is married to a civilian, is older in age, works more hours per week, and has higher capital gains is more likely to earn an income of over $50,000.** The correlation of final weight to income is almost 0 so we will disregard it as an influential factor. Also, data in this column may represent overlapping populations and we do not want to over-count features.

IMPACTS

Luckily, our model did not output race or sex as a top feature, but it is important to note how these types of results can impact the real world. A biased Random Forest Classifier can have significant impacts in the real world if it produces discriminatory predictions. In the context of predicting income, if the model is biased against certain demographic groups such as race and sex, it can perpetuate existing social and economic inequalities by unfairly predicting lower incomes for those groups. This can result in discriminatory hiring practices, lending decisions, and other important societal outcomes that affect people's opportunities and quality of life. For example, if the model systematically predicts lower incomes for women and people of color, employers may use the model to screen job candidates, leading to discriminatory hiring practices that perpetuate the gender and racial wage gaps. Similarly, banks and other financial institutions may use the model to make lending decisions, leading to unfair denial of loans and perpetuation of systemic economic disparities. To mitigate these risks, it is important to train and evaluate the model on a diverse and representative dataset, while also being transparent about the model's limitations and potential biases. Additionally, it is essential to have safeguards in place to prevent the model from being used in discriminatory ways and to allow for regular monitoring and updating of the model's performance.





CONCLUSION

There are many factors which may influence income level of a resident in the United States. From our analysis of three different supervised machine learning models, we have concluded that age is a leading factor in income determination; however, further analysis should be done to inspect the relationship between income and race and sex because of the real-world impacts it can have on hiring practices, lending decisions, and quality of life.

REFERENCES

Aladangady, A., & Forde, A. (n.d.). Wealth inequality and the Racial Wealth Gap. The Fed - Wealth Inequality and the Racial Wealth Gap. Retrieved February 17, 2023.

Barroso, A., & Brown, A. (2022, June 8). Gender pay gap in U.S. held steady in 2020. Pew Research Center. Retrieved February 17, 2023, from https://www.pewresearch.org/fact-tank/2021/05/25/gender-pay-gap-facts/

Bradbury, K., & Triest, R. (2014). Inequality of Opportunity and Aggregate Economic Performance. http://www.bostonfed.org/inequality2014/papers/bradbury-triest.pdf

Schaeffer, K. (2020, February 7). 6 facts about economic inequality in the U.S. Pew Research Center. Retrieved February 18, 2023, from https://www.pewresearch.org/fact-tank/2020/02/07/6-facts-about-economic-inequality-in-the-u-s/

Vivamoto, M. (2021). US Adult Income [Data set]. Kaggle. https://www.kaggle.com/datasets/vivamoto/us-adult-income-update