Income Level Insights: A Data-Driven Approach to Economic Census Analysis

Jason Ie

DS3000

Abstract:

Economic Census data is a comprehensive source of information about the economic activity of millions of businesses and industries. Conducted every five years by the United States Census Bureau, this rich dataset provides detailed insights into the nation's economic structure, performance, and operations [3]. The economic census offers valuable specifics on industry revenues, business expenditures, employment, and payroll, among others. This data is vital for gaining a comprehensive view of the economic landscape, informing policymakers, assisting city planners, and aiding businesses in making data-driven decisions.

Problem Definition & Objectives

This project dives into economic census data to categorise American individuals into income brackets of <=50K and >50K. Our aim is to illuminate economic disparities and support informed decision-making for businesses and policymakers [3].

We chose this dataset for its detailed representation of the nation's economic landscape, reflecting the financial realities faced by its citizens. Our analysis began with a detailed examination of the data, identifying key variables crucial in understanding income variations. These included age, education, occupation, and capital gains, which our findings revealed as significantly impactful in determining income levels.

Our methodology combined meticulous data inspection, comprehensive cleaning to rectify over 6,000 unclear data points, and exploratory data analysis (EDA). This process allowed us to make initial assessments and set the stage for more advanced analysis using machine learning techniques, particularly the Random Forest model [3].

The results of our analysis revealed that the Random Forest model, known for its accuracy and reliability, demonstrated an impressive ability to predict income brackets, achieving an accuracy of about 85.88%. This high level of accuracy underscored the significance of the variables like 'fnlwgt', 'age', and 'education-num', which emerged as the most influential factors in income classification. 'fnlwgt' or 'final weight' represents the number of people in the U.S. population that each respondent in the census data stands for, calculated based on demographic characteristics to ensure representativeness.

Through this project, we hope to contribute valuable insights into the factors driving income disparities. Our ultimate goal is to provide stakeholders with a clearer understanding of income distribution, aiding in the development of strategies that promote equitable resource allocation and thoughtful policymaking for community enhancement.

Related Work

In socio-economic analysis, accurately classifying income levels is crucial for understanding societal structures. Two studies were found to be closely aligned to the research and outcome of our project.

The Pew Research Center's study categorizes nearly half of American adults as middle-class, offering a key reference point for income distribution. This research is instrumental for our project as it provides a comparative baseline and underscores the importance of income stratification in socio-economic studies [1].

Additionally, a CNBC article from January 2023 explores what constitutes a middle-class income in major U.S. cities. It highlights the variation in income levels needed to maintain a middle-class lifestyle across different regions, underscoring the impact of cost of living on income classification. This variability is crucial for our project, emphasizing the need to consider regional economic conditions when analyzing income data [4].

Methodology

Our analysis commenced with the 1994 UCI economic census data, a comprehensive dataset encapsulating socio-economic aspects of the U.S. population, such as demographics, education, employment, and income.

In the initial stages of our exploration, we utilized techniques like `.head()`, `.describe()`, `.info()`, and `.shape()` to gain a fundamental understanding of our dataset. These steps were instrumental in grasping the data structure and planning our analysis [3].

Addressing data cleanliness, we replaced ambiguous '?' entries with NaN values, ensuring consistency and accuracy. To prepare our data for machine learning, we segregated the dataset into categorical and numerical variables [3]. Utilizing the `SimpleImputer`, we imputed missing values in numerical data using the 'mean' strategy and in categorical data with the 'most frequent' value, thereby preserving the integrity and distribution of our data.

Feature engineering played a pivotal role in our methodology. We created new features like `edu_occ`, which combined education and occupation, and `hours_per_week_cat` to categorize working hours. These engineered features were designed to provide deeper insights and enhance model performance.

We performed one-hot encoding on categorical variables to convert them into a machine-readable format. This process involved creating dummy variables for each category within a feature, which is essential for models that require numerical input.

For the KNeighbors Classifier, we implemented feature scaling using `StandardScaler`. Scaling was necessary to normalize the feature space, as KNN is sensitive to the magnitude of data and requires features to be on a similar scale for accurate distance computation.

Our data was divided into a 70/30 split for training and testing. This division allowed for a robust learning process and an effective validation of the model's predictive capabilities.

Our approach included evaluating various algorithms, namely Random Forest, KNeighbors Classifier, and Decision Tree. The selection criteria were based on a combination of accuracy, precision, recall, and interpretability. The Random Forest algorithm emerged as the optimal choice due to its robustness in handling diverse data types and its capability to provide meaningful insights through feature importance.

We further refined our Random Forest model using `GridSearchCV` to determine the optimal number of trees (`n_estimators`). This step was crucial in balancing the model's bias and variance, leading to improved accuracy and reliability. The Mean Squared Error metric was also calculated to assess the average squared difference between the predicted and actual values, providing another layer of model evaluation.

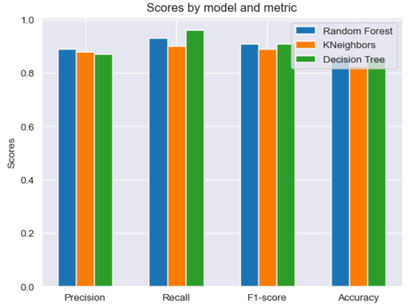
To comprehensively evaluate our chosen model, we utilized a confusion matrix, which offered a detailed breakdown of the model's predictive accuracy in terms of true positives, false negatives, and more. This analysis was vital in understanding the model's strengths and areas for improvement.

Finally, we conducted a feature importance analysis using our Random Forest model. This analysis was key in identifying which variables most significantly influenced income levels, aligning with our project's objective of gaining deeper insights into the socio-economic factors affecting income distribution.

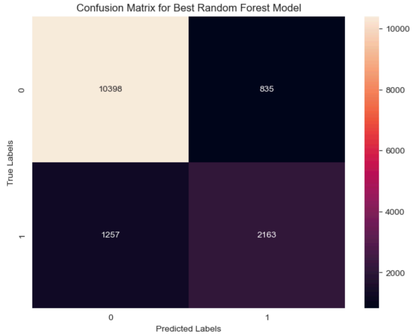
Results & Evaluation

In our evaluation of machine learning models, the Random Forest Classifier emerged as the superior model, achieving an accuracy of 85.65%. When compared to the KNeighborsClassifier and Decision Tree Classifier, which attained accuracies of 82.64% and 85.80% respectively, the Random Forest model demonstrated a more favorable balance between precision and recall. Specifically, for the <=50K income bracket, the Random Forest model achieved high precision (89%) and recall (93%), outperforming the KNeighborsClassifier's precision (88%) and recall (90%), and was on par with the Decision Tree's precision (87%) but had a significantly better recall (96%).

For the >50K income bracket, the Random Forest model's precision (72%) and recall (63%) were considerably higher than the KNeighborsClassifier's precision (64%) and recall (58%), and although the Decision Tree model had a slightly better precision (79%), its recall (54%) was lower. This indicates that the Random Forest model is better at identifying individuals who actually have an income above 50K.

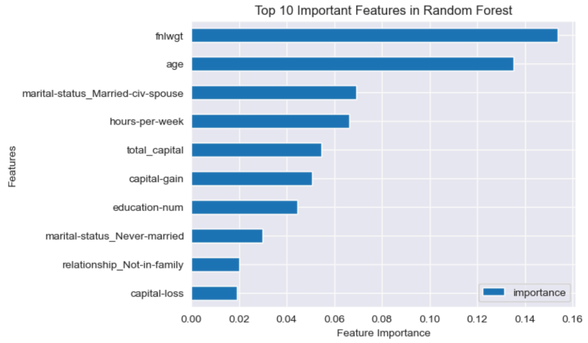


From the confusion matrix, the Random Forest model correctly predicted 10,398 true negatives and 2,163 true positives. In contrast, the KNeighborsClassifier and Decision Tree models exhibited a lower number of true positives, which is critical for accurately identifying individuals within the higher income bracket.



The feature importance analysis from the Random Forest model revealed that 'fnlwgt' and 'age' were the most significant predictors of income level, with 'fnlwgt' holding the highest importance score. This contrasted with the simpler Decision Tree model, where the importance might be concentrated on fewer features, potentially leading to less robust insights.

In summary, the Random Forest Classifier was not only accurate but also offered a rich interpretation of feature relevance. Its performance was consistently strong across various metrics, outshining the other models in both the comprehensive understanding of influential factors and the correct classification of individuals across different income levels. This made it particularly suitable for our objective of gaining deep insights into income distributions for informed socio-economic policy and business strategy formulation.



Conclusion & Impact

In the conclusion of our data science project, we've generated insights from the economic census data to highlight the key differentiators in income levels. Our analysis, leveraging the Random Forest Classifier, has provided a nuanced understanding of the intricate factors influencing income disparities.

Summary of Findings:
Our model underscored the prominence of features like 'fnlwgt', age, and hours worked per week as significant predictors of income. The 'fnlwgt' feature, indicative of the population demographics and labor force characteristics, emerged as the most influential, revealing insights on the socioeconomic fabric that shapes income distribution. With an accuracy rate of around 85.72%, our model has demonstrated a notable capacity to segment the population into income brackets, serving as a valuable tool for both business and policy applications.

Model Application:
Beyond the theoretical factors, our model's practical applications span various domains. For businesses, it can fine-tune marketing strategies, optimise location planning, and tailor product offerings to different income segments. In the public sector, it can inform welfare policies, educational grants, and infrastructure development by targeting the needs of specific income groups. The model's ability to identify the higher-income bracket, despite being an area for improvement, still provides a springboard for targeted economic initiatives.

Impact and Future Work:
Our project's implications extend beyond the immediate results. The insights derived have the potential to guide strategic decisions that foster community prosperity.

The model's limitations must be acknowledged. The dataset, rooted in the 1990s, may not perfectly mirror today's economic landscape, and the model's predictive power for higher-income brackets could benefit from further refinement.

To refine our model further, we're aiming on a range of enhancements that align with the evolving landscape of data-driven socio-economic analysis.

Firstly, we plan to incorporate our dataset with up-to-date information, reflecting the latest economic conditions. This step is crucial to ensure that our model mirrors the contemporary economic environment, thereby enhancing its relevancy and accuracy.

In our pursuit of greater precision, particularly in predicting higher-income brackets, we're opting to delve into more sophisticated machine learning techniques. Advanced algorithms such as Gradient Boosting and Neural Networks are on our exploration goals. These methods promise to refine our model's accuracy, offering a more nuanced understanding of income dynamics.

To enrich our analysis, we aim to integrate additional variables into our dataset, with a particular focus on local economic indicators. This inclusion will provide a more layered and comprehensive perspective, enabling our model to capture the diverse economic realities across different regions.

Lastly, recognising the dynamic nature of economic data, we're committed to implementing a regular cycle of model updates and validation. This ongoing process will ensure that our model adapts to changing economic trends and maintains its accuracy and relevance over time.

These planned improvements represent a significant leap forward in our investigation to develop a model that's not only robust in its predictive capabilities but also agile in adapting to the ever-evolving economic landscape.

Final Remarks:
To conclude, our effort has yielded a predictive model that serves as an approach to the dynamics of income distribution. Its potential for societal impact is significant, offering a beacon for data-driven decision-making in both the commercial and public sectors. The journey thus far has been enlightening, yet we are keen to explore further, enhance our methodologies, and amplify our model's impact on communities.

References

[1] Bennett, J. (2020, July 23). *Are you in the american middle class? find out with our income calculator*. Pew Research Center. https://www.pewresearch.org/short-reads/2020/07/23/are-you-in-the-american-middle-class/#:~:text=About%20half%20of%20U.S.%20adults,were%20in%20upper%2Dincome%20households.

[2] Bureau, U. C. (2021, November 22). *Purposes and uses of Economic Census Data*. Census.gov. https://www.census.gov/programs-surveys/economic-census/guidance/data-uses.html

[3] *Census income*. UCI Machine Learning Repository. (n.d.). https://archive.ics.uci.edu/ml/datasets/Census+Income

[4] McNair, K. (2023, January 2). *Here's how much money it takes to be considered middle class in 20 major U.S. cities*. CNBC. https://www.cnbc.com/2023/01/02/middle-class-income-in-major-us-cities.html