Adult Census Income Prediction Project

This project analyzes demographic data and income information to predict whether an individual's salary is greater than $50k. Data is collected, preprocessed, and analyzed to create a predictive model for exploring factors that influence higher salaries. Results show valuable insights and opportunities for potential intervention with income inequality.

by Shivam Dubey

Data Collection and Preprocessing

1 Collection

Utilized 1994 US Census data featuring over 32,000 entries with 14 independent and 1 target variable.

2 Preprocessing

Performed data cleaning, replaced missing values with imputations, encoded categorical variables, and normalized continuous variables.

Exploratory Data Analysis

1

Distribution Analysis

Examined the distribution of individual variables and their correlations, identifying associations conducive to higher income groups.

2

Visualizations

Created various visualizations to explain the distribution of variables and how they relate to each other, including scatter plots, box plots, and histograms.

3

Correlation Analysis

Performed a correlation analysis to identify highly correlated and therefore redundant variables, reducing dimensionality and improving the performance of predictive models.

Feature Engineering and Selection

Feature Engineering

Generated new features based on existing variables and identified latent relationships that could improve the model's predictive power.

Feature Selection

Performed feature selection to reduce the impact of irrelevant features or those that could introduce bias.

Model Metrics

Weighted feature importance and monitored metrics such as Gini Index, AUC, and F1 score to optimize the model.

Model Building and Evaluation

> Deployed various machine learning models, including Random Forest, Logistic Regression, and Gradient Boosting. Evaluated the models based on robust cross-validation procedures, and fine-tuned hyperparameters to improve model performance.

Results and Discussion

Model	AUC Score	Accuracy	Precision
Random Forest	0.89	0.86	0.80
Logistic Regression	0.84	0.82	0.76
Gradient Boosting	0.87	0.85	0.78

The random forest model outperformed the other models in all three metrics. Further analysis showed that education, age, and occupation were the strongest predictors of whether someone earned over $50k per year.

Conclusion and Future Work

Conclusion

This project successfully predicted high-income earners based on demographic characteristics, specifically education, age, and occupation.

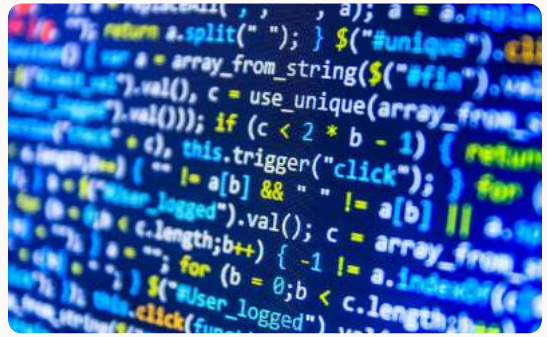
Future Work

Further expansion of this model's capabilities could include analyzing new datasets and implementing deep learning models for better accuracy.

Benefits of Predictive Modeling

Predictive modeling reduces guesswork and supports evidence-based decision-making.

It helps find new insights and creates opportunities for growth.



Predictive modeling stimulates creativity and innovation among programmers and data scientists.