Programming Assignment 2 Final Report

Bingchang Song bingchang@wustl.edu
Ruoyao Wen ruoyao@wustl.edu
Yonghao Chen c.yonghao@wustl.edu

April 27, 2024

1 Introduction

1.1 Motivation for the Study

This project delves into the "Adult" dataset, which stems from the 1994 U.S. Census database, initially prepared by Barry Becker. The principal aim of analyzing this dataset is to predict whether an individual's annual income exceeds $50,000, based on various demographic and employment-related features. Such predictions are highly beneficial for a range of stakeholders:

- **Governments and Policymakers:** They can use the insights from the analysis to identify demographic groups or regions that might benefit from specific economic policies or educational programs aimed at elevating income levels. For instance, if a significant correlation between educational attainment and higher income brackets is identified, policies could focus on improving educational access and quality.
- **Financial Institutions:** Banks and other lending agencies could utilize the model to better assess the risk profiles of loan applicants, thus enhancing their decision-making processes and potentially reducing the risk of defaults. Insights from the model could help in adjusting credit scoring systems and lending terms accordingly.

Understanding these dynamics aids in crafting targeted interventions that can lead to more effective policy formulations and business strategies.

1.2 Overview of the Dataset

The dataset includes a range of variables that are both categorical and continuous, capturing various aspects of an adult's life. These include:

- **Demographic Information:** Age, sex, race, and native country.
- **Employment Details:** Workclass (e.g., private, government), occupation, and hours worked per week.
- **Education Information:** Highest level of education achieved and a numerical representation of education years.
- **Family and Relationship Status:** Marital status and relationship.
- **Financial Details:** Capital gain, capital loss, and ultimately, the income classification indicating whether the individual earns more than $50,000 annually.

The final variable, income, serves as the binary target for our predictive models, with the classes labeled as "less than $50K" or "higher than or equal to $50K". This classification not only allows us to assess the financial status of individuals but also helps in understanding the impact of various factors on income levels. The exploration and modeling of these variables provide crucial insights that contribute to the objectives of our study.

2 Methods

2.1 Data Preprocessing

The preprocessing of the dataset involved several steps to prepare it for effective model training and validation:

- **Handling Missing Values:** Entries with missing data were identified and removed to ensure the quality and integrity of the models' training process.
- **Data Encoding:** Categorical variables were encoded using label encoding to transform them into a format suitable for model input, facilitating the analysis process without altering the number of features.
- **Data Segmentation:** Based on the hypothesis that education significantly influences income, the dataset was divided into three subsets based on educational attainment using the Education Number variable. This created three binary classification problems, each predicting the likelihood of an income exceeding $50,000 in different educational brackets.

2.2 Model Selection and Description

Various classification models were selected to address the binary classification problems:

- **k-Nearest Neighbors (kNN):** A simple yet effective model that classifies new data points based on the majority vote of its nearest neighbors.
- **Decision Tree:** This model uses a tree-like structure of decisions, where each node represents a feature in the dataset, and the branches to children represent the possible values that lead to decision outcomes.
- **Random Forest:** An ensemble method that builds multiple decision trees and merges them together to get a more accurate and stable prediction.
- **Artificial Neural Networks (ANN):** Inspired by biological neural networks, these models are capable of capturing complex patterns through layers of neurons, suitable for non-linear data structures.
- **Support Vector Machines (SVM):** This model constructs a hyperplane in a high-dimensional space to separate different classes with as wide a margin as possible.
- **Naive Bayes:** A probabilistic model that applies Bayes' Theorem, with the assumption of independence between predictors.

2.3 Hyperparameter Tuning

To optimize model performance, hyperparameters for each model were carefully selected based on the results of cross-validation:

- **Grid Search Technique:** Employed to systematically work through multiple combinations of parameter tunes, cross-validating as it goes to determine which tune gives the best performance.
- **Validation Approach:** The data was split into training and validation sets to ensure that the tuning process did not overfit to the data, maintaining the model's ability to generalize to new data.

2.4 Dimension Reduction Technique

- **Principal Component Analysis (PCA):** Given the high dimensionality of the data post-categorical encoding, PCA was used to reduce the number of features by half while retaining most of the variance in the data. This reduction was crucial for enhancing model training efficiency and avoiding overfitting.

These methods formed the foundation of our approach to tackling the binary classification problems presented by the dataset, enabling a thorough analysis of the factors influencing income levels.

3 Results

3.1 Model Performance Evaluation Without Dimension Reduction

Initially, models were evaluated without applying dimension reduction techniques to understand their baseline performance. The performance metrics focused on were accuracy, precision, recall, and F1-score. Here are the summarized results for each model before dimension reduction:

- **Random Forest:** Achieved the highest overall accuracy and F1-score, particularly excelling in the Higher Education subset with an accuracy of 97.23% and F1-score of 0.97.

- **Decision Tree:** Showed strong performance with an accuracy of 81.61% and an F1-score of 0.82 in the same subset.

- **Naive Bayes:** While it was the fastest, it showed lower accuracy at 67.61% and an F1-score of 0.65.

- **k-Nearest Neighbors:** Produced an accuracy of 76.88% and an F1-score of 0.77.

- **Support Vector Machine (SVM) and Artificial Neural Network (ANN):** SVM and ANN exhibited competitive results with accuracies of 80.63% and 80.21%, respectively, and similar F1-scores around 0.80.

3.2 Model Performance Evaluation With Dimension Reduction

Following the application of PCA, which reduced the dimensionality of the dataset by half, the models were re-evaluated to assess the impact on performance:

- **Random Forest:** Maintained a high accuracy of 97.23% and F1-score of 0.97, showing minimal impact from the reduction.

- **Decision Tree:** Experienced a slight decrease in performance, dropping to an accuracy of 73.80% and F1-score of 0.74.

- **Naive Bayes:** Showed a significant decrease to an accuracy of 64.75% and F1-score of 0.59.

- **k-Nearest Neighbors:** Reduced to an accuracy of 66.81% and F1-score of 0.67.

- **Support Vector Machine (SVM) and Artificial Neural Network (ANN):** Both models showed a notable decline, with SVM dropping to an accuracy of 68.91% and ANN to 71.54%.

3.3 Impact of Dimension Reduction

Dimension reduction was implemented using PCA, which was particularly effective in reducing training time and improving model manageability without significantly compromising accuracy:

- **Pre vs. Post PCA Comparison:** Models trained on data reduced by PCA generally showed a slight decrease in accuracy but benefitted from faster training times and reduced overfitting.

- **Model Specific Impact:** Random Forest and Decision Tree models were less affected by the reduction in feature space, maintaining high levels of accuracy, while models like SVM and kNN experienced more notable declines in performance.

3.4 Visualizations and Findings

The figures illustrate the cross-validation results and the impact of hyperparameter tuning across different models and subsets are available in our GitHub repository: cse514pa2.

These results clearly demonstrate the varied impact of dimension reduction on different models, highlighting the importance of choosing the right model and preprocessing techniques based on specific dataset characteristics and desired outcomes.

4 Discussion

4.1 Analysis of Model Performance

The analysis revealed significant variances in performance among the different models:

- **Random Forest and Decision Tree:** These models consistently outperformed others across various metrics and subsets, suggesting their robustness in handling the dataset's complexity and feature interdependencies. The ensemble nature of Random Forest particularly helped in mitigating overfitting, a common issue observed with Decision Trees.

- **Naive Bayes:** Despite its speed, the Naive Bayes model struggled with the dataset's feature dependencies, underperforming in terms of accuracy and precision. This outcome highlights the model's limitations in handling real-world data complexities where features are often interrelated.

- **SVM and kNN:** These models showed sensitivity to the dimensionality and distribution of the data, with performance dropping notably post-PCA. This suggests that while these models are powerful in high-dimensional spaces, they require careful consideration of feature selection and dimensionality reduction techniques.

4.2 Recommendations

Based on the findings, the following recommendations can be made to stakeholders:

- **Policymakers and Governments:** Should consider the robustness of Random Forest models when analyzing demographic data for policy formulation. Additionally, targeted educational programs could be developed based on the strong performance linkages observed between education levels and income.

- **Financial Institutions:** Could integrate Decision Tree or Random Forest models into their risk assessment processes to enhance credit scoring accuracy, given these models' ability to handle diverse and complex data sets effectively.

4.3 Limitations and Future Work

While the study provided valuable insights, it had several limitations:

- **Data Quality and Completeness:** The removal of entries with missing data might have introduced bias or affected the generalizability of the findings.
- **Assumption of Independence:** The reliance on this assumption by Naive Bayes could be revisited by exploring models that can handle feature interdependencies more effectively.
- **Dimension Reduction:** The impact of PCA on model performance suggests further exploration into other dimension reduction techniques that might preserve more relevant information for certain models.

Future research could explore the integration of more sophisticated machine learning techniques, such as deep learning for classification tasks, or the application of feature engineering to enhance model performance further. Additionally, gathering more comprehensive and diverse datasets could help in developing more generalized models capable of wider applicability.

5 Conclusion

This study provided a comprehensive analysis of the ”Adult” dataset to predict whether individuals' incomes exceed $50,000 per year, utilizing a range of machine learning models and techniques. The findings highlighted the varying effectiveness of different models, with Random Forest and Decision Tree models demonstrating superior performance across multiple metrics and data subsets. These models not only offered high accuracy but also proved robust against overfitting, making them particularly valuable for practical applications in economic policy formulation and risk assessment in financial sectors.

The application of Principal Component Analysis for dimension reduction revealed the trade-offs between model complexity and training efficiency. While PCA helped in reducing the training time and model complexity, it also led to a slight decrease in performance for certain models, underscoring the need for careful consideration of dimension reduction techniques in predictive modeling.

Furthermore, this project emphasized the importance of hyperparameter tuning in achieving optimal model performance, showcasing how different settings can significantly impact the results. The extensive use of cross-validation ensured that the models were not just tailored to the peculiarities of the training data but also generalized well to unseen data.

Key Recommendations:

- Stakeholders in policy-making and financial sectors should consider integrating robust models like Random Forest in their analytical arsenals to better understand and predict economic behaviors and risks.
- Future studies should aim to explore more advanced dimension reduction techniques and incorporate more granular data that could provide deeper insights into the income prediction task.

In conclusion, this project underscores the critical role of machine learning in socioeconomic data analysis, offering insights that can drive more informed decisions and strategies. The methodologies and findings from this study pave the way for further research that could extend beyond the scope of income prediction to other areas of economic and social policy analysis.

For further details, the complete code and datasets used in this study are available in our GitHub repository: cse514pa2.