Income Classification Report

Executive Summary

This project aims to predict an individual's income category ('>50K' or '<=50K') based on demographic attributes. The project utilized diverse machine learning algorithms and advanced data preprocessing techniques to achieve high prediction accuracy. Insights from the analysis can support socio-economic research and HR analytics.

Introduction

Income prediction is crucial for socio-economic studies and business insights, such as understanding labor market dynamics and tailoring financial products. Using the Adult Income Dataset from the UCI Machine Learning Repository, this project applies machine learning classifiers to predict income categories accurately.

Methodologies

1. Data Collection and Cleaning

- **Dataset Source**: UCI Machine Learning Repository.
- Missing values were addressed using techniques like replacing with most frequent or "unknown" categories.
- Replaced '?' with NaN for clean analysis.

2. Exploratory Data Analysis (EDA)

- Analyzed the distribution of demographic attributes such as age, education, occupation, and income levels.
- Visualized the correlation between features and income.
- Noted significant trends, such as higher education levels correlating with higher income.

3. Data Preprocessing

- Categorical features were one-hot encoded.
- Numerical features were scaled to a [0, 1] range.
- Addressed class imbalance using:
  - **Over-sampling**: SMOTE, ADASYN
  - **Under-sampling**: NearMiss, Tomek Links
  - **Algorithmic Techniques**: Class weight adjustments in models.

4. Feature Selection and Clustering

- Utilized **Random Forest Classifier** for feature importance ranking.
- Performed **K-Means Clustering** to identify patterns within data.

5. Modeling Approaches

Implemented multiple classifiers: - Logistic Regression - Decision Tree - Random Forest - Support Vector Classification (SVC) - K-Nearest Neighbors (KNN) - Gaussian Naive Bayes - Quadratic Discriminant Analysis (QDA) - AdaBoost - Gradient Boosting - Multi-layer Perceptron (MLP)

6. Evaluation Metrics

- **Accuracy**: Correct predictions ratio.
- **Precision & Recall**: Evaluated the positive prediction quality.
- **F1-Score**: Harmonic mean of precision and recall.
- **ROC-AUC Score**: Assessed classification model performance.

Results & Insights

- **Random Forest** and **Gradient Boosting** emerged as top-performing models.
- Over-sampling with **SMOTE** improved model recall.

- Key features influencing income prediction included:

  - **Education Level**
  - **Hours Worked Per Week**
  - **Age**
  - **Capital Gain**

- Visual analyses showed a strong correlation between higher education levels and income categories.

- Resampling methods significantly mitigated the class imbalance challenge, leading to better prediction accuracy.

Conclusion

- The project successfully developed an accurate income prediction model using advanced machine learning algorithms and resampling techniques.
- Addressing data imbalance significantly enhanced model performance.
- Future work can focus on:
  - Deploying the model into an interactive dashboard.
  - Collecting more real-time income data for model retraining.
  - Exploring deep learning techniques for further optimization.

References

- UCI Machine Learning Repository - Adult Income Dataset
- Scikit-learn Documentation for Machine Learning Models
- Kaggle for Data Analysis Insights

This report aims to provide a comprehensive overview of the project from data preprocessing to model deployment insights.