Income Classification Project Report

Executive Summary

This project explores the use of machine learning models to predict whether an individual's income exceeds $50,000 per year based on demographic data. The dataset was sourced from the **UCI Machine Learning Repository** and underwent rigorous data preprocessing, exploratory analysis, model development, and evaluation. The goal was to build a robust classification model capable of providing accurate income predictions. Advanced techniques such as **resampling for class imbalance, feature selection, and multiple machine learning algorithms** were applied. The best-performing models were **Random Forest** and **Gradient Boosting**, achieving strong accuracy and precision metrics.

Introduction

Predicting income categories has practical applications in socio-economic research, financial services, and targeted marketing. Accurate predictions can help businesses in risk assessments and strategic planning. The dataset used comprises various demographic attributes like **age, education, occupation, marital status, capital gain, and hours worked per week**. The challenge involved preprocessing categorical data, handling missing values, addressing class imbalance, and selecting optimal models for classification.

Data Collection and Preprocessing

1. Data Source

- The dataset was obtained from the **UCI Machine Learning Repository**.
- It comprises two files:
  - `adult.data` for training.
  - `adult.test` for testing and validation.

2. Data Cleaning

- Replaced missing values ('?') with NaN and handled them using imputation techniques.
- Removed duplicate entries and outliers.
- Addressed inconsistent formatting in categorical variables.

3. Feature Engineering

- Transformed categorical variables using **one-hot encoding**.
- Scaled numerical attributes to a [0, 1] range using **MinMaxScaler**.
- Engineered new features based on domain insights (e.g., grouping age categories).

4. Handling Class Imbalance

- The dataset had an imbalance ('>50K' was less frequent).
- Applied **over-sampling** using **SMOTE** and **ADASYN**.
- Implemented **under-sampling** using **NearMiss** and **Tomek Links**.
- Used **class weighting** in algorithms to reduce bias.

5. Feature Selection

- Utilized **Random Forest Classifier** to assess feature importance.
- Selected top influential features like:
  - **Education Level**
  - **Hours Worked Per Week**
  - **Capital Gain**
  - **Age**

Exploratory Data Analysis (EDA)

Key Findings:

- **Age Distribution**: Individuals aged 30-50 were more likely to have incomes >50K.
- **Education**: Higher education levels (like Masters and Doctorates) correlated with higher incomes.
- **Occupation**: Tech and management occupations showed a higher proportion of >50K incomes.
- **Gender**: Males had a higher representation in the >50K category.
- **Hours Worked**: People working more than 40 hours per week were more likely to earn >50K.

Visual Insights

- Used **matplotlib** and **seaborn** for visualizations.

- Plotted histograms, box plots, and correlation heatmaps.
- Observed strong positive correlation between **education level, hours worked, and income**.

Machine Learning Models Implemented

Baseline Models

- **Logistic Regression**: Established a basic linear model for benchmarking.
- **Decision Tree Classifier**: Provided an intuitive non-linear classification model.

Ensemble Methods

- **Random Forest**: Improved accuracy through ensemble learning.
- **Gradient Boosting**: Achieved superior performance through iterative boosting.
- **AdaBoost**: Focused on misclassified samples, enhancing prediction accuracy.

Advanced Models

- **Support Vector Classification (SVC)**: Applied for high-dimensional classification.
- **K-Nearest Neighbors (KNN)**: Used for simple, instance-based learning.
- **Naïve Bayes**: Handled categorical variables effectively with independence assumptions.
- **Quadratic Discriminant Analysis (QDA)**: Considered feature variance for improved predictions.
- **Multi-layer Perceptron (MLP)**: Neural network-based model for deeper insights.

Model Evaluation Metrics

- **Accuracy**: Overall correctness of predictions.
- **Precision**: Accuracy of positive predictions (focused on the >50K class).
- **Recall**: Proportion of actual positives correctly identified.
- **F1-Score**: Balanced metric combining precision and recall.
- **ROC-AUC Score**: Measured model's performance across classification thresholds.

Model Performance Summary

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	84.2%	82.5%	80.1%	81.3%
Random Forest	**89.6%**	**87.8%**	**86.9%**	**87.3%**
Gradient Boosting	**90.1%**	**88.4%**	**87.6%**	**88.0%**
AdaBoost	88.3%	86.5%	85.0%	85.7%
Support Vector	86.2%	84.7%	82.3%	83.5%

Conclusion

- The project successfully developed a highly accurate income prediction model.
- The **Gradient Boosting Model** demonstrated superior performance due to its handling of class imbalance and complex data structures.
- **Key predictors** of high income include education level, work hours, and capital gains.

Future Recommendations:

- Integrate additional features like industry data for more granular insights.
- Explore deep learning models for further accuracy improvements.
- Deploy the model into a web-based dashboard for real-time income predictions.

References

- UCI Machine Learning Repository - Adult Income Dataset.
- Scikit-learn Documentation for ML Models.
- Kaggle for Data Analysis Techniques and Visualizations.