Income Prediction

Muhammad Daniyal Qureshi, Hafiz Muhammad Ahmad, Muhammad Sufyan

November 27, 2024

Abstract

This project aims to predict income levels using demographic and work-related attributes. Data preprocessing, feature selection, and predictive modeling were conducted on the "Census Income" dataset. The models—Logistic Regression, Decision Tree, and Random Forest were evaluated using metrics like accuracy, confusion matrix, precision, recall, F1 score, and AUC to further improved model performance. Results highlight Random Forest's effectiveness in classification, with an AUC score of 0.93 and slight difference in recall for both classes.

1 Introduction

Income classification is a crucial task with applications in socio-economic research and policy-making. Using the "Census Income" dataset, we predict whether individuals earn more than $50K annually based on attributes like age, workclass, educational-num, marital-status, occupation, relationship, gender and hours-per-week. The study utilizes machine learning techniques to preprocess data, select important features, and build predictive models. These insights provide a foundation for developing scalable income prediction systems.

2 Methodology

2.1 Data Preprocessing

The "Census Income" dataset contains 48,842 rows and 15 columns. After preprocessing, 10 features were retained.

- **Cleaning**: Removed redundant columns like `capital-gain`, `capital-loss`,`education` and `fnlwgt`.

- **Encoding**: Applied `LabelEncoder` for categorical variables such as `workclass` and `marital-status` etc.
- **Feature Selection**: Recursive Feature Elimination (RFE) identified that all features are important in Random Forest.
- **Splitting**: Split the data into training(70 percent) and testing(30 percent).

2.2 Machine Learning Models

Three models were used:

- **Logistic Regression**: A linear model capturing relationships between features.
- **Decision Tree**: A non-linear model that creates decision boundaries.
- **Random Forest**: An ensemble model that averages predictions from multiple trees to reduce overfitting.

2.3 Evaluation Metrics

The models were evaluated using:

- **Accuracy**
- **Training and Testing Errors**
- **Confusion Matrix**
- **Precision, Recall, F1 Score**
- **AUC Score**

3 Results & Discussion

The evaluation results for the three models are presented in the table below:

Metric	Logistic Regression	Decision Tree	Random Forest
Training Accuracy	74.34%	91.28%	97.27%
Testing Accuracy	74.41%	83.23%	85.55%
Training Error (MSE)	25.7%	8.72%	2.7%
Testing Error (MSE)	25.59%	16.8%	14.4%
AUC Score	0.82	0.88	0.93
Precision (<=50K)	0.75	0.82	0.86
Precision (>50K)	0.74	0.85	0.85
Recall (<=50K)	0.73	0.85	0.84
Recall (>50K)	0.76	0.81	0.87
F1-Score (<=50K)	0.74	0.83	0.85
F1-Score (>50K)	0.75	0.83	0.86

Table 1: Evaluation Metrics for Logistic Regression, Decision Tree, and Random Forest

Among all the models tested, Random Forest emerged as the best choice due to its highest AUC score of 0.93, indicating its superior ability to distinguish between income classes. Despite the Decision Tree having a lower AUC score (0.88), Random Forest showed more balanced performance and reduced overfitting, making it the most reliable model for this task.

4 Conclusion & Future Work

Random Forest emerged as the most balanced model with a high AUC score of 0.93, indicating its ability to distinguish income classes. The project highlights the importance of feature selection in enhancing model performance.

4.1 Future Work

Future improvements can include:

- Experimenting with advanced models like XGBoost or Gradient Boosting.
- Increasing dataset diversity to improve generalization.
- Optimizing computational cost by further dimensionality reduction.