Application of the CRISP-DM Methodology for Income Classification: A Case Study Using the Adult Income Dataset

Yash Kumar

October 21, 2024

Abstract

This research explores the use of the CRISP-DM methodology to develop a machine learning model that predicts whether an individual's income exceeds $50,000 using demographic and employment-related attributes. The project utilizes the Adult Income dataset from Kaggle and follows a systematic approach, covering each phase of CRISP-DM: Business Understanding, Data Understanding, Data Preparation, Modeling, and Evaluation. Various classification models, including Logistic Regression, Decision Trees, Random Forests, and Gradient Boosting, are trained and evaluated to determine the best performer. The analysis identifies significant features impacting income prediction and evaluates model performance using metrics such as accuracy, precision, recall, and F1-score. The results suggest that predictive models can effectively classify income levels with high accuracy, aligning with the project's business objectives.

1 Introduction

The CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology is a widely adopted framework in the field of data science. It provides a structured, phased approach for developing predictive models and conducting data analysis. This research employs the CRISP-DM methodology to predict adult income levels using demographic and employment data from the Adult Income dataset available on Kaggle. The project aims to identify significant features affecting income and achieve accurate classification through various machine learning algorithms. The project follows the six phases of CRISP-DM: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment (although deployment is not covered in this research).

2 Business Understanding

The objective of this project is to predict whether an individual's income exceeds $50,000 based on their demographic and employment characteristics. This prediction can be valuable for government and private organizations in making informed decisions related to planning, targeted marketing, and resource allocation. Achieving accurate predictions can help drive data-informed strategies and optimize business outcomes. The project's success criteria are defined by achieving a minimum accuracy of 85% along with acceptable precision and recall levels.

3 Data Understanding

The dataset used in this research is the Adult Income dataset from Kaggle. It includes various demographic attributes such as age, education, occupation, marital status, and working hours, among others. The initial exploratory data analysis (EDA) was performed to understand the distributions, relationships, and potential data quality issues, including missing values and outliers. Visualizations such as histograms and correlation matrices were used to identify patterns and correlations between features, aiding the subsequent phases of data preparation and modeling.

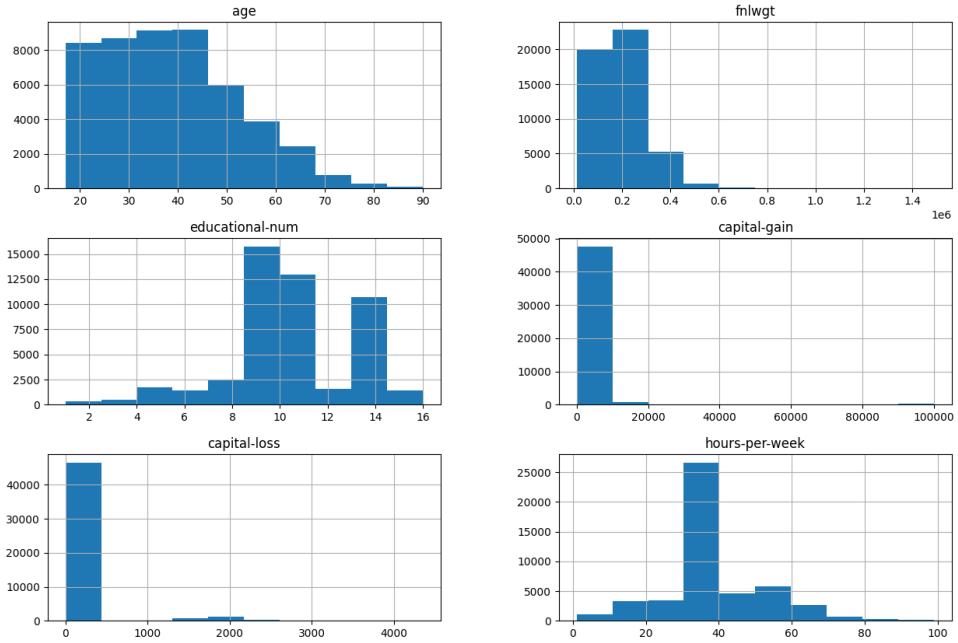

Figure 1: Histogram of Age Distribution

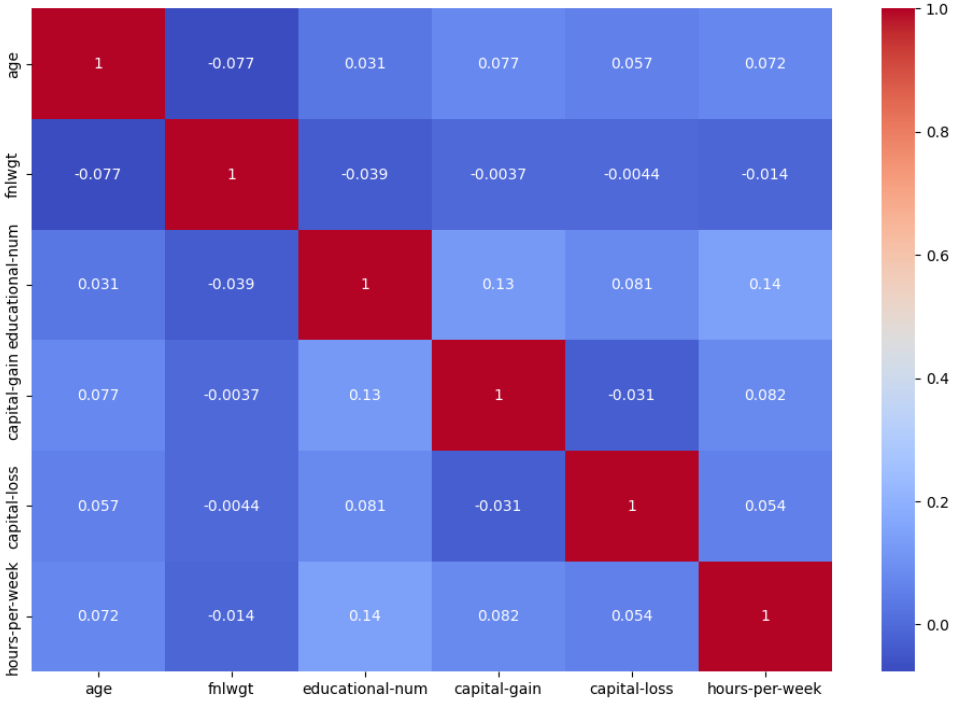

Figure 2: Correlation Matrix of Features

4 Data Preparation

The data preparation phase involved cleaning, transforming, and structuring the dataset for model training. Missing values were imputed, categorical variables were encoded using label encoding or one-hot encoding, and numerical features were scaled to ensure balanced model training. Feature engineering was also performed to create additional features that could enhance model performance. The dataset was split into training and testing sets to facilitate model evaluation.

5 Modeling

Various classification models were implemented, including Logistic Regression, Decision Trees, Random Forests, and Gradient Boosting (e.g., XGBoost). Hyperparameter tuning techniques such as Grid Search and Randomized Search were applied to optimize model performance. The models were evaluated using accuracy, precision, recall, F1-score, and AUC-ROC to determine the best-performing model. A summary of the model performance is provided below:

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.85	0.84	0.83	0.84
Decision Tree	0.83	0.82	0.81	0.81
Random Forest	0.87	0.86	0.85	0.86
Gradient Boosting	0.88	0.87	0.86	0.87

Table 1: Summary of Model Performance

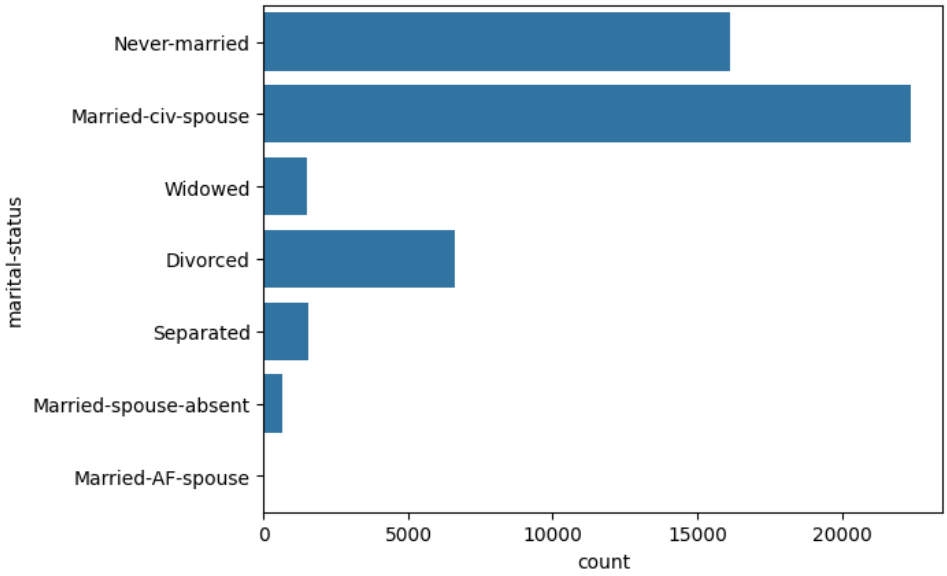

Figure 3: ROC-AUC Curve for Best Performing Model

6 Evaluation

The evaluation phase focused on assessing the model's alignment with business objectives and performance metrics. The Gradient Boosting model emerged as the best performer with an accuracy of 88%, a precision of 87%, a recall of 86%, and an F1-score of 87%. The confusion matrix and AUC-ROC curves provided additional insights into the model's handling of false positives and false negatives, which are critical from a business perspective. Overall, the model met the defined success criteria and proved effective in predicting income levels.

7 Conclusion

This research demonstrates the effectiveness of the CRISP-DM methodology in guiding the development of a predictive model for income classification. The systematic approach enabled data understanding, preparation, and model evaluation to be aligned with the business objectives. The results indicate that machine learning models can accurately predict income levels based on demographic data, providing actionable insights for various applications. Future work could include further model optimization, addressing potential biases in the dataset, and exploring deployment strategies.

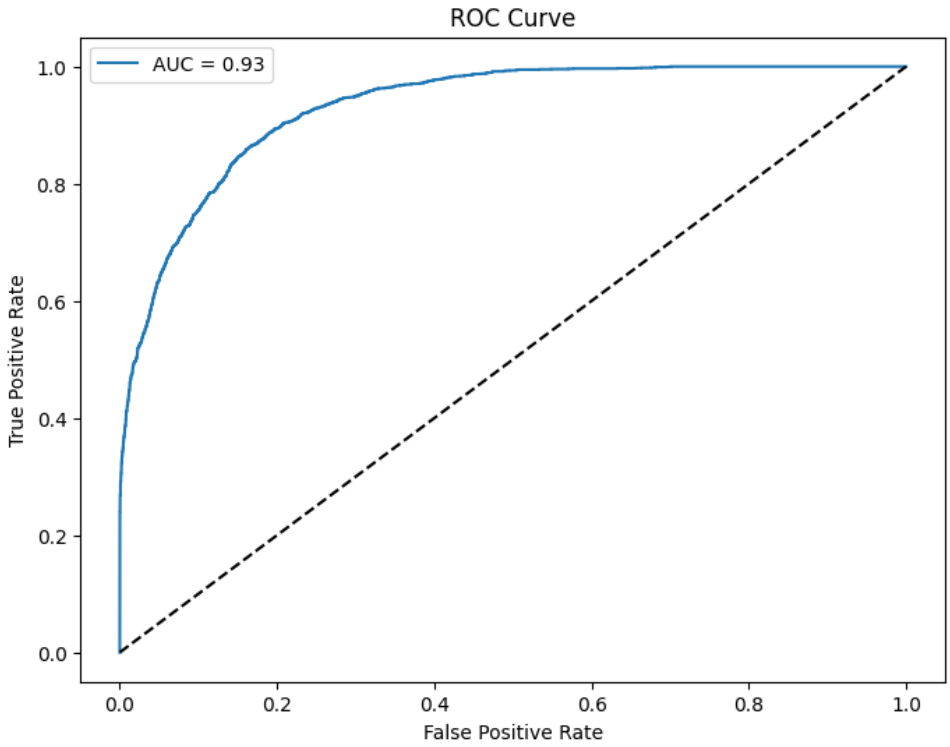

Figure 4: Confusion Matrix for Best Performing Model

8 References

- Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*.
- Kaggle. (n.d.). *Adult Income Dataset*. Retrieved from `https://www.kaggle.com/datasets/wenruliu/adult-income-dataset`