Midterm Kaggle Competition: Predicting Income Level

Selena Cho

CS6530

University of Utah

Introduction

The dataset used in this project was curated by Barry Becker, who extracted it from the extensive 1994 Census database. He selected a set of records that met specific criteria. The primary objective of this project is to address the challenging task of predicting whether an individual's annual income exceeds $50,000 based on the information available in the census. This prediction task holds significant importance in various applications, including social and economic analysis.

Referred to as the "Adult" dataset or the "Census Income" dataset, this dataset is widely recognized in the fields of machine learning and data mining, primarily used for binary classification tasks. The core goal is to determine whether an individual's income surpasses $50,000 per year by leveraging a range of demographic and socioeconomic attributes. The dataset encompasses 14 attributes, featuring a combination of continuous, categorical, and integer data types. Notably, some of these attributes may contain missing values, denoted by question marks. This dataset serves as a valuable resource for practicing and evaluating machine learning algorithms, particularly those related to classification models.

This report offers an overview of the project's progress, which revolves around predicting an individual's income status, with a particular focus on whether their income exceeds $50,000. The project encompasses various phases, including data preprocessing, model selection, and the initial implementation of a predictive model.

Methods

Dataset Description

Training Dataset: *age, workclass, fnlwgt, education, education.num, marital.status, occupation, relationship, race, sex, capital.gain, capital.loss, hours.per.week, native.country, income>50K*

Testing Dataset: *ID, age, workclass, fnlwgt, education, education.num, marital.status, occupation, relationship, race, sex, capital.gain, capital.loss, hours.per.week, native.country*

Feature Encoding

Categorical variables often require special treatment to be used effectively in machine learning models. In our dataset, several columns were categorical, including *'workclass,' 'education,' 'marital.status,' 'occupation,' 'relationship,' 'race,' 'sex,' and 'native.country.'* We applied one-hot encoding to these categorical columns, effectively converting them into numerical form, which is essential for many machine learning algorithms.

Feature Scaling

Standardization of numerical features is a crucial step to ensure that the scale of different attributes does not adversely affect the performance of machine learning algorithms. We standardized the numeric features, which included *'age,' 'fnlwgt,' 'education.num,' 'capital.gain,' 'capital.loss,' and 'hours.per.week,'* using the *StandardScaler* via *sklearn* library. This procedure ensured that all numerical features had a mean of zero and a standard deviation of one, thus promoting optimal model performance.

Model Selection

For our predictive model, we chose the Random Forest Classifier, a versatile ensemble method renowned for its effectiveness in handling diverse datasets. The Random Forest Classifier is a prominent ensemble learning technique that combines the predictive power of multiple decision trees. Ensemble methods have emerged as a cornerstone of modern machine learning due to their capacity to aggregate diverse and complementary information, thereby enhancing the overall predictive accuracy. The Random Forest Classifier has garnered widespread acclaim for its robustness, versatility, and ability to mitigate common challenges encountered in predictive modeling. This model was selected due to its ability to handle both categorical and numerical data, reducing overfitting, and providing accurate predictions.

One of the critical reasons behind our choice of the Random Forest Classifier is its inherent capability to seamlessly manage both categorical and numerical data. In our dataset, we encountered a heterogeneous mix of attributes, spanning from demographic and socioeconomic features, represented as categorical variables, to numerical attributes capturing continuous variables such as age and working hours per week. The Random Forest Classifier adeptly accommodates these different data types without necessitating extensive data preprocessing, which often consumes valuable resources and time.

Model Pipeline and Training

The data preprocessing component of our pipeline is instrumental in preparing the dataset for predictive modeling. It involves a series of operations, including data imputation for missing values, encoding categorical variables, and scaling numerical features. By encapsulating these operations within the pipeline, we ensure consistency and reproducibility in the treatment of both the training and test datasets. This uniformity in data preprocessing is pivotal to the reliability and generalizability of the predictive model.

The training data consists of two key components: the feature variables (X) and the target variable (y). The feature variables (X) encompass the demographic and socioeconomic attributes gathered from the dataset, which are employed to make predictions regarding income status. The target variable (y) serves as the ground truth, indicating whether an individual's annual income exceeds $50,000. It is this variable that the Random Forest Classifier seeks to predict accurately.

Results

The effectiveness of our predictive model is assessed through the utilization of a robust evaluation metric known as the Area Under ROC (AUC) curve. The AUC is a quantitative measure that provides valuable insights into the model's ability to discriminate between positive and negative instances, which, in the context of our research, corresponds to predicting income levels exceeding $50,000 or not. The AUC score ranges from 0 to 1, with higher values indicative of superior predictive performance.

From the following evaluation, we have a prediction score of 0.90139. From the time of the result, it was the third best performance.

Future Plans

One of our immediate priorities is the investigation and mitigation of missing data within our dataset. Missing data can introduce biases and inaccuracies, potentially impeding the model's predictive capabilities. Our approach to this challenge involves a thorough examination of the dataset to identify instances of missing values. Subsequently, we will adopt appropriate strategies for handling these gaps in the data.

Imputation: Missing data will be addressed through imputation techniques, where we will infer and assign values to the missing entries based on the information available in the dataset. Imputation ensures data completeness and integrity, permitting the model to make informed predictions on a comprehensive dataset.

To further enhance the predictive power of our Random Forest Classifier, we are planning an extensive exploration of hyperparameter tuning. The hyperparameters of a machine learning model play a pivotal role in shaping its behavior and performance. The Random Forest Classifier, being a versatile ensemble method, offers various hyperparameters to fine-tune the model.

As our class and learning journey continues, we recognize the invaluable insights that can be gained by exploring and implementing a diverse array of machine learning models. While the Random Forest Classifier is a robust choice for its versatility and predictive power, there exists a rich landscape of other models that offer unique strengths and capabilities. For instance, the Nonlinear classifier and kernel tricks can be particularly advantageous when dealing with complex, non-linear relationships within the data, allowing for the representation of intricate patterns that may be challenging for decision tree-based ensembles. Furthermore, Probabilistic learning, Naive Bayes classifier, Logistic regression, and Risk Minimization strategies offer distinct advantages, such as the ability to capture probabilistic relationships and minimize the risk of model overfitting. As the class progresses, we eagerly anticipate delving into these alternative models, carefully assessing their suitability for our specific predictive task and the unique insights they can provide for our socioeconomic analysis.

Conclusion

This report serves as a milestone in our project, toward predicting income status with precision and insight. Our research has thus far encompassed meticulous data preprocessing, model selection in the form of the Random Forest Classifier, and the generation of predictive outcomes on the test data. These achievements have not only provided valuable insights but have also revealed avenues for further enhancement and exploration.

Data preprocessing addressed missing values and standardized the dataset. Robust data processing pipeline has paved the way for consistent and reproducible treatment of both training and test data, ensuring the integrity and quality of our predictive model.

Model Selection: The selection of the Random Forest Classifier, as discussed in the earlier sections, was a judicious choice, leveraging its strengths in handling diverse data types and mitigating overfitting. This ensemble method stands as a formidable tool for predicting income status, but our commitment to model improvement and exploration extends to embracing alternative models as the project progresses. The multifaceted nature of machine learning will allow us to explore nonlinear classifiers, kernel-based models, artificial neural networks, probabilistic learning, and others, each with its unique attributes and potential advantages. These models are primed for consideration as we progress through the final steps of the competition.

Reference

GitHub Repository

https://github.com/SelenaChoUtah/CS6350/tree/main/Kaggle