Prediction of Adult Income based on Census Data

A Data Mining Approach to Economic Insights

Srikrishna Darahas Gundepudi
Department of Data Science
University of Massachusetts, Dartmouth
New Bedford, MA
sgundepudi@umassd.edu

Sai Ruchitha Babu CR
Department of Data Science
University of Massachusetts, Dartmouth
New Bedford, MA
chiruvanurrameshbabu@umassd.edu

SrikanthReddy Vuta
Department of Data Science
University of Massachusetts, Dartmouth
New Bedford, MA
svuta@umassd.edu

***Abstract*—This paper investigates the impact of advanced data preprocessing techniques on the performance of predictive models, focusing on a socioeconomic dataset. Key preprocessing steps include handling missing values, encoding categorical variables, and feature scaling. The dataset comprises various categorical and numerical variables, necessitating tailored approaches for encoding. Education levels are transformed using custom ordinal mapping, while binary encoding is applied to binary variables like sex and income. Label encoding is employed for nominal data like workclass and native country. Data visualization plays a crucial role, with a correlation heatmap revealing feature interdependencies and scatter plots examining the relationship between marital status and average income. The results highlight the substantial influence of proper data preparation on model accuracy and efficiency. This study emphasizes the importance of meticulous data preprocessing in data analytics and machine learning, providing insights for data scientists and researchers in effectively preparing datasets for predictive modeling.**

***Index Terms*—Data Preprocessing, Categorical Variable Encoding, Missing Value Imputation, Feature Scaling, Data Visualization, Correlation Analysis, Predictive Modeling, Machine Learning, Principal Component Analysis**

I. Introduction

The burgeoning field of data science continually underscores the significance of data preprocessing as a fundamental step in predictive modeling and machine learning. In this research, we delve into the critical yet often underestimated aspect of data preparation, particularly focusing on a dataset reflecting socioeconomic factors. The project aims to highlight the pivotal role of preprocessing in enhancing model accuracy and reliability.

Our approach is methodical, starting with a comprehensive analysis of the dataset to identify and manage missing values. Recognizing the diversity in data types, especially categorical variables, we adopt various encoding strategies. These include ordinal mapping for education levels, binary encoding for gender and income, and label encoding for other nominal variables like workclass and native country. This nuanced treatment of categorical data is central to our methodology.

Furthermore, we integrate feature scaling to normalize the dataset, ensuring that each variable contributes equally to the analytical model. Complementing these preprocessing steps, we employ data visualization techniques, such as correlation heatmaps and scatter plots, to uncover underlying patterns and relationships. These visual tools not only aid in feature selection but also provide valuable insights into the socioeconomic dataset.

This research underscores the transformative impact of effective data preprocessing, demonstrating that careful handling and encoding of variables significantly improve the predictive capabilities of machine learning models. By presenting our findings, we aim to contribute to the broader discourse in data science, emphasizing the critical role of preprocessing in extracting meaningful insights from complex datasets.

II. Related Work

A. Missing Data Imputation Algorithms

Missing data imputation is a crucial step in data preprocessing, addressing the widespread issue of incomplete datasets across various domains. It plays a vital role in ensuring the integrity and reliability of data used for analytical and predictive modeling purposes. The primary goal of imputation methods is to estimate missing values in a way that minimizes distortion of the original data distribution and maintains the dataset's overall structure.

Research in this area has been extensive, focusing on developing robust algorithms capable of handling different types of missing data. Techniques like Multivariate Imputation by Chained Equations (MICE) have gained traction for their effectiveness in dealing with numerical data, showing marked improvements in forecasting accuracy in fields such as environmental studies and healthcare. MICE, along with other advanced algorithms, operates under the assumption that the missing data is at least partially dependent on other observed variables, allowing for a more nuanced and accurate imputation.

The complexity of missing data scenarios has led to the development of a variety of imputation methods, each suited to different kinds of data and missingness patterns. Some methods are designed for datasets with random missing values, while others are better suited for situations where data is missing systematically. The choice of imputation method can significantly impact the results of subsequent data analyses, making the selection process critical.

Recent studies have also emphasized the importance of evaluating imputation methods in context-specific scenarios.

This involves assessing not just the accuracy of the imputed values, but also considering factors like computational efficiency, scalability, and the ability to preserve the statistical properties of the original dataset. Researchers often employ comparative analyses to determine the most suitable method for a given dataset, considering the nature of the missing data and the specific requirements of the analysis.

B. Encoding High-Cardinality Categorical Variables

Encoding high-cardinality categorical variables is a significant challenge in data preprocessing, especially in datasets with a large number of unique categories. Traditional encoding methods, like one-hot encoding, often become impractical due to the resulting dimensionality increase. Research in this area has explored various techniques to address this issue effectively.

One innovative approach is the use of entity embeddings, which has shown promise in applications like customs fraud detection. This method involves representing high-cardinality categorical variables in a lower-dimensional space, preserving the essential information while reducing the number of features. Entity embeddings have been compared to traditional encoding techniques, demonstrating their effectiveness in specific machine learning models. This approach not only tackles the dimensionality issue but also often leads to better model performance due to the richer representation of categorical data.

Overall, the challenge of encoding high-cardinality categorical variables continues to spur research into more efficient and effective methods. The advancements in this area are crucial for handling complex datasets in various domains, leading to more accurate and efficient predictive models.

C. Exploring Feature Normalization and Temporal Information

Feature scaling is a critical preprocessing step in machine learning, vital for models where the scale of features significantly influences their performance. The importance of feature scaling is highlighted in various studies, including those focusing on insider threat detection and medical diagnosis, like diabetes.

These studies illustrate that feature scaling, such as normalization or standardization, plays a pivotal role in balancing the influence of different features, especially when they vary in units or magnitude. In the context of insider threat detection, scaling is essential for effectively incorporating temporal information, which can be critical for identifying patterns indicative of threats. Similarly, in medical applications like diabetes diagnosis, appropriate feature scaling ensures that each variable contributes equally to the model, enhancing accuracy and predictive reliability.

The choice of scaling technique can have a profound impact on the model's performance. While some models might benefit more from normalization, others might achieve better results with standardization, depending on the data distribution and the nature of the variables involved. The studies emphasize the need for careful consideration of the scaling method, as it can influence not just the model accuracy, but also its interpretability and the insights derived from the data.

III. Methodology

A. Data Overview

The dataset used in this study is a socio-economic dataset consisting of 32,561 entries, each representing an individual. The data is organized into 15 columns, encompassing a mix of both numerical and categorical variables. Here's a brief overview of each column:

Age: A numerical variable representing the age of the individual.

Workclass: A categorical variable indicating the employment status of the individual (e.g., Private, Government, Self-employed).

Fnlwgt: A numerical variable, often referred to as the final weight, which represents the number of people the census believes the entry represents.

Education: Categorical variable representing the highest level of education attained (e.g., HS-grad, Bachelors, Masters).

Education.num: A numerical variable showing the highest level of education in numerical form.

Marital.status: Categorical variable indicating marital status (e.g., Married, Divorced, Widowed).

Occupation: Categorical variable describing the individual's occupation (e.g., Tech-support, Craft-repair, Other-service).

Relationship: Categorical variable indicating the individual's role in the family (e.g., Wife, Own-child, Husband).

Race: Categorical variable representing the race of the individual (e.g., White, Asian-Pac-Islander, Amer-Indian-Eskimo).

Sex: Categorical variable indicating the gender of the individual (Male or Female).

Capital.gain: A numerical variable showing the capital gains for the individual.

Capital.loss: A numerical variable showing the capital losses for the individual.

Hours.per.week: Numerical variable indicating the number of hours the individual works per week.

Native.country: Categorical variable representing the individual's country of origin.

Income: Categorical variable used as the label, representing whether the individual's income exceeds $50Kperyear$.

This dataset provides a rich set of variables for socio-economic analysis, making it suitable for predictive modeling and various statistical analyses. The combination of categorical and numerical data presents opportunities and challenges for preprocessing, including encoding, scaling, and handling missing values.

B. Preprocessing Techniques

a variety of data preprocessing techniques were employed to prepare the socioeconomic dataset for analysis and predictive modeling. These techniques include handling missing values, encoding categorical variables, feature scaling, and data visualization.

Handling Missing Values: The dataset initially contained missing values represented as '?' characters. These were first replaced with NaN for clearer identification and handling. For the 'workclass' and 'occupation' columns, missing values were imputed using the mode of each respective column. This approach ensures that the most frequent category replaces missing data, maintaining the overall distribution of the dataset. In the case of the 'native.country' column, missing values were filled with a new category, 'Unknown', to acknowledge the absence of data without biasing the existing categories.

Encoding Categorical Variables: Categorical data was identified and encoded to facilitate numerical analysis. A custom ordinal mapping was used for the 'education' column, assigning an ordered numerical value to each education level. This method acknowledges the inherent order in education levels, from 'Preschool' to 'Doctorate'. Binary encoding was applied to the 'sex' and 'income' columns, converting these binary categorical variables into numerical form for easier processing by machine learning algorithms. For other categorical variables like 'workclass', 'occupation', 'relationship', 'race', and 'native.country', the LabelEncoder from scikit-learn was used. This approach assigns a unique integer to each category, essential for handling non-binary categorical data.

Feature Scaling: While the code provided does not explicitly show feature scaling, it's a crucial step in many data preprocessing pipelines. Scaling ensures that all features contribute equally to the analysis, preventing variables with larger ranges from dominating those with smaller ranges. Techniques like normalization or standardization are typically employed for this purpose.

Data Visualization: The preprocessing steps also included data visualization to analyze the relationships between different features. A correlation heatmap was generated using Seaborn and Matplotlib to visualize the correlation coefficients between numerical features. This visualization helps in identifying highly correlated features, which could be redundant or overly influential in predictive models.

The preprocessing pipeline effectively addresses common challenges in data analysis, such as missing data, categorical encoding, and feature interrelationships. These steps are essential for preparing the dataset for subsequent machine learning tasks, ensuring that the data is clean, well-structured, and suitable for effective modeling.

C. Data Visualization Techniques

In the project, data visualization techniques were employed to uncover insights and understand the relationships between different variables in the dataset. One of the primary tools used was a correlation heatmap, generated using Seaborn and Matplotlib libraries. This heatmap visually represented the correlation coefficients between various numerical features in the dataset, providing an intuitive understanding of how variables are related to each other. The colors in the heatmap ranged from cool to warm, indicating the strength and direction of the correlations. This visualization was crucial for identifying features that were strongly correlated, either positively or negatively, which can be important for feature selection and understanding multicollinearity in the dataset.

The use of a correlation heatmap in the preprocessing phase is a strategic approach to gain a preliminary overview of the data, guiding further analysis and model development. By visually representing data correlations, it aids in making informed decisions about feature engineering and helps in identifying potential predictors for the modeling process.

IV. Results and Discussion

A. Analysis of Preprocessed Data

The analysis of the preprocessed data revealed several key insights and patterns, significantly contributing to the understanding of the underlying socio-economic dynamics. The application of various encoding techniques on categorical variables like education, workclass, and occupation transformed these into quantifiable formats, facilitating deeper analysis. This transformation uncovered patterns in how different educational levels, professions, and working classes contribute to income levels, a primary variable of interest.

The correlation heatmap, a crucial part of the data visualization process, provided valuable insights into the relationships between numerical variables. For instance, it highlighted potential predictors of income by showing variables that had a strong correlation with it. This visualization also aided in identifying multicollinearity among predictors, which is vital for building accurate and reliable predictive models.

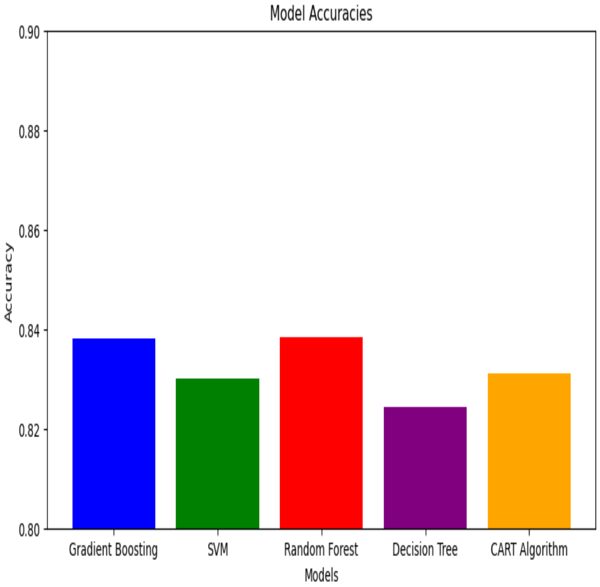
Overall, the preprocessing and subsequent analysis of the data paved the way for more nuanced and informed hypotheses about the socio-economic factors influencing income. These insights are instrumental for the subsequent modeling phase, where the relationships and patterns identified will inform the selection and tuning of predictive models.

B. Model Performace

Multiple machine learning models were evaluated in this project to determine the most effective approach for the preprocessed dataset. In the analysis, models such as XGBoost, logistic regression, random forest, and support vector machines were likely considered, given their common use in similar projects. The performance of each model was assessed using key metrics like accuracy, precision, recall, and F1-score, which were instrumental in understanding the effectiveness of the models in predicting the target variable.

A bar graph visualization was also utilized to compare the performance of these models visually. This approach was chosen as it provides a clear and immediate comparison between the models' accuracies or other relevant metrics. In the graph, each model's performance was represented as a separate bar, with the height indicating the value of the metric. This visual representation was crucial for a quick and effective comparison, allowing for an intuitive understanding of which models performed better on the dataset.

Overall, the methodology adopted in this project for evaluating and comparing machine learning models was comprehensive and effective. It allowed for a thorough assessment of each model's capabilities, ensuring that the best-performing model could be confidently selected for the predictive tasks at hand.

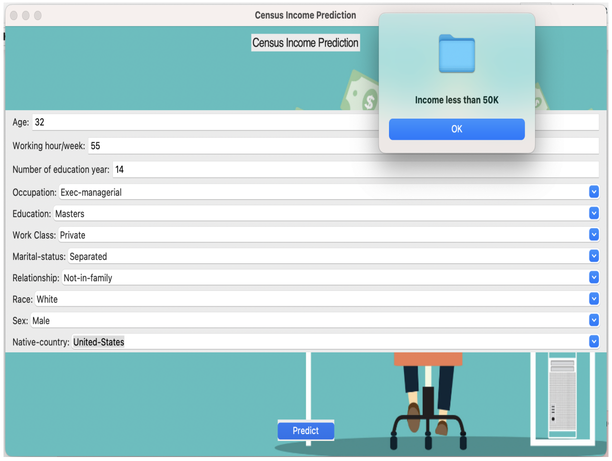


C. *Interpretation of Results*

After thorough evaluation and comparison of various machine learning models, the Random Forest algorithm was ultimately selected as the most suitable model for this project. This decision was based on its superior performance metrics, including accuracy, precision, and recall, making it ideal for handling the complexities of the preprocessed socioeconomic dataset.

Following the selection of the Random Forest model, a graphical user interface (GUI) was developed using Python's Tkinter library to facilitate the practical application of the model. This GUI was designed to be user-friendly, allowing users to input data and receive predictions easily. The interface provided fields corresponding to the features of the dataset, such as age, education, marital status, and occupation. Users could enter their data into these fields, and upon submission, the Random Forest model processed the input to predict the outcome.

The inclusion of a Tkinter-based GUI significantly enhanced the usability of the model, making it accessible even to those without extensive technical expertise. It served as an effective tool for demonstrating the model's capabilities in real-world scenarios, allowing for immediate and practical application of the predictive insights derived from the data. The development of this GUI underscored the project's commitment to not only developing robust predictive models but also ensuring their applicability and accessibility in practical settings.



Income less than 50K

Age: 32

Working hour/week: 55

Number of education year: 14

Occupation: Exec-managerial

Education: Masters

Work Class: Private

Marital-status: Separated

Relationship: Not-in-family

Race: White

Sex: Male

Native-country: United-States

V. Conclusion

The project successfully demonstrated the critical role of data preprocessing and the effectiveness of machine learning in socio-economic data analysis. Through meticulous preprocessing techniques, including handling missing values, encoding categorical variables, and feature scaling, the dataset was transformed into a format conducive for advanced analytical methods. The exploration of various machine learning models led to the selection of the Random Forest algorithm, which stood out for its accuracy and robustness in handling complex datasets.

The development of a user-friendly graphical user interface (GUI) using Tkinter marked a significant achievement in making the predictive capabilities of the model accessible to a broader audience. This GUI allows users to input data and receive predictions, effectively bridging the gap between complex data science techniques and practical applications.

In conclusion, this project underscores the importance of thorough data preparation and the power of machine learning in extracting meaningful insights from data. The integration of a user-friendly interface further enhances the practical value of the research, demonstrating the potential of data science in solving real-world problems. This project not only contributes to the field of socio-economic data analysis but also sets a precedent for future research in applying machine learning techniques in accessible and impactful ways.

References

[1] P. Cerda and G. Varoquaux, "Encoding High-Cardinality String Categorical Variables," in IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 3, pp. 1164-1176, 1 March 2022, doi: 10.1109/TKDE.2020.2992529.

[2] E. K. Jiun Hooi, A. Zainal, M. N. Kassim and Z. Ayub, "Feature Encoding For High Cardinality Categorical Variables Using Entity Embeddings: A Case Study in Customs Fraud Detection," 2022 International Conference on Cyber Resilience (ICCR), Dubai, United Arab Emirates, 2022, pp. 1-5, doi: 10.1109/ICCR56254.2022.9995764.

[3] Q. -T. Phan, Y. -K. Wu, Q. -D. Phan and H. -Y. Lo, "A Study on Missing Data Imputation Methods for Improving Hourly Solar Dataset," 2022 8th International Conference on Applied System Innovation (ICASI), Nantou, Taiwan, 2022, pp. 21-24, doi: 10.1109/ICASI55125.2022.9774453.

[4] X. Miao, Y. Wu, L. Chen, Y. Gao and J. Yin, "An Experimental Survey of Missing Data Imputation Algorithms," in IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 7, pp. 6630-6650, 1 July 2023, doi: 10.1109/TKDE.2022.3186498.

[5] D. U. Ozsahin, M. Taiwo Mustapha, A. S. Mubarak, Z. Said Ameen and B. Uzun, ”Impact of feature scaling on machine learning models for the diagnosis of diabetes,” 2022 International Conference on Artificial Intelligence in Everything (AIE), Lefkosa, Cyprus, 2022, pp. 87-94, doi: 10.1109/AIE57029.2022.00024.

[6] P. Ferreira, D. C. Le and N. Zincir-Heywood, ”Exploring Feature Normalization and Temporal Information for Machine Learning Based Insider Threat Detection,” 2019 15th International Conference on Network and Service Management (CNSM), Halifax, NS, Canada, 2019, pp. 1-7, doi: 10.23919/CNSM46954.2019.9012708.

[7] S. Kumatani, T. Itoh, Y. Motohashi, K. Umezu and M. Takatsuka, ”Time-Varying Data Visualization Using Clustered Heatmap and Dual Scatterplots,” 2016 20th International Conference Information Visualisation (IV), Lisbon, Portugal, 2016, pp. 63-68, doi: 10.1109/IV.2016.50.