Sean McLean

ALY 6020

Module 1 Project

Understanding Income Inequality

Introduction

Companies are increasingly focused on ensuring equitable pay across various demographics, requiring a thorough understanding and awareness of the factors that contribute to income disparities. This assignment seeks to assess and create a predictive model to classify individuals into low-income and high-income categories using variables such as occupation, education, and race from the provided census dataset. By utilizing a k-nearest neighbors (KNN) algorithm, the objective is to accurately distinguish between these income classes and identify predictors of affluence. These insights can guide policymaking and support initiatives that will over the long term promote financial equality. This assignment summarizes the data cleansing techniques used to improve data quality and the development and interpretation of the KNN model, offering insights for stakeholders that can be used in the future.

Analysis

The initial phase of the analysis focused on thorough data cleansing to ensure high-quality data for modeling. The provided census dataset, which includes attributes such as occupation, education, gender, and race, underwent various data cleaning techniques. The first step in the process is loading and inspecting the data by loading it into a Pandas DataFrame. Basic information, including data types and the first few records, was displayed to understand the dataset's structure. The next step is to check for any missing value, using the SimpleImputer function with a strategy of replacing missing values with the most frequent value was used. This step ensured that the dataset had no missing values, which is essential for building a reliable predictive model.

The data types are then converted with numeric columns such as 'age', 'fnlwgt', 'education-num', 'capital-gain', 'capital-loss', and 'hours-per-week'. They are converted from the 'object' data type to appropriate numeric types, enabling accurate computations and modeling. The target variable 'salary' was separated from the features and then label encoding was applied to convert the categorical target variable into numeric form, making it suitable for machine learning algorithms. Categorical columns with a manageable number of unique values were then one-hot encoded, creating binary columns for each category and thus avoiding any ordinal assumptions. The last step is standardizing the features by using the StandardScaler function to ensure all features have a mean of zero and a standard deviation of one. This standardization is crucial for algorithms like KNN, which can be sensitive to the scale of the data (Anaconda Assistant, 2024).

After completing the data cleansing phase, the subsequent step involved constructing and evaluating a k-nearest neighbors (KNN) model. Initially, the cleaned and standardized dataset was divided into two subsets: a training set for model training and a test set for evaluating its performance. This partitioning ensured that the model's effectiveness could be assessed on unseen data, validating its generalizability.

To optimize the KNN model, a GridSearchCV approach was employed to determine the optimal number of neighbors. This process involved systematically testing K values ranging from one to 30 while using five-fold cross-validation. By going through these configurations, the model selection process was robust and aimed at identifying the K value that yielded the best performance metrics. Subsequently, leveraging the optimal K value identified through GridSearchCV, a KNN model was trained using the designated training dataset. This model was then deployed to make predictions on the test set, providing a practical assessment of its predictive capabilities in a real-world scenario (Anaconda Assistant, 2024).

Data visualizations are then created using matplotlib and seaborn from the census dataset for analysis. It includes a pair plot that illustrates relationships between selected numerical features like age and education-num, categorized by salary class (low-income vs. high-income). Additionally, a boxplot highlights age distributions across salary classes, revealing potential differences between income groups. Multiple count plots further depict the distribution of categorical features such as workclass and education across salary classes. These visualizations provide valuable insights into how different variables relate to income levels, facilitating the interpretation and understanding of the dataset's characteristics that are essential for further modeling and analysis.

Findings

The analysis of the census dataset, comprising approximately 48,000 rows, successfully classified individuals into low-income and high-income categories based on their salary using a k-nearest neighbors (KNN) model. The model achieved an accuracy of 83.28%, demonstrating a high rate of correct salary class predictions. For low-income predictions, the precision was 0.86 and the recall was 0.93, indicating the model's strong performance in identifying low-income individuals. For high-income predictions, the precision was 0.69 and the recall was 0.52, highlighting areas for improvement while still providing valuable insights into factors affecting higher salaries (See Appendix A).

The optimal value of K was determined to be 16 through a grid search with cross-validation, meaning the model bases its salary predictions on the 16 nearest neighbors to each data point. This value of K helps balance overfitting and underfitting, ensuring the model is neither overly sensitive to noise nor is too generalized. The model with that K value considers a

broad yet focused set of neighbors, contributing to its robust performance. These findings suggest that the KNN model is effective for the given dataset, and its high accuracy and balanced metrics provide a solid foundation for understanding salary disparities and informing policy decisions. Further validation and refinement could enhance its predictive capabilities, particularly for high-income predictions.

The pair plot indicated that while the two income groups are similar, some observations highlight disparities in income. The plot showed that higher-income individuals generally have higher capital gains despite working similar hours. Capital gains also correlated with the level of education, with higher-income individuals showing a linear increase in gains with more education, while lower-income individuals' gains remained consistently lower (See Appendix B).

The boxplots illustrated the distribution of age by salary category, revealing several differences. Individuals earning less than $50,000 annually had a larger range in age distribution but a lower median age and fewer outliers. Conversely, individuals earning more than $50,000 had a more compact age range, more outliers, and a higher median age, suggesting that higher-income individuals tend to seek more education and potentially receive more promotions (See Appendix C).

Count plots showed educational differences impacting both income groups, with most individuals working in the private sector. Lower-income males were more likely to be single, divorced, or not in a family, while most higher-income individuals were married and identified as a husband. Few individuals from either income group were identified as wives, suggesting either fewer women in the dataset overall or a higher proportion of unmarried women. These findings suggest that higher education, stable marriage, and career focus are prevalent in the

higher income group, whereas the lower income group is characterized by younger working ages, less education, and higher rates of divorce or singleness (See Appendix D).

Recommendations

Based on these findings, it is recommended to continue using the KNN model for classifying salary categories, as it shows high accuracy and balanced performance, especially in identifying low-income individuals. However, to enhance the precision and recall for high-income predictions, further refinement of the model is suggested. This could involve exploring additional features, adjusting preprocessing techniques, or considering alternative machine learning algorithms. Additionally, a more detailed analysis of misclassified instances might reveal patterns that can be addressed to improve overall model performance. These steps will aid in better understanding the factors influencing salary disparities and making more informed policy decisions.

From the findings in the plots, it is recommended to implement policies that promote educational attainment and career advancement opportunities, as these factors are strongly linked to higher income levels. Programs supporting continuing education and skill development can help lower-income individuals achieve better financial outcomes. Additionally, initiatives to support stable family structures, such as marriage counseling and family support services, may also contribute to higher income levels, given the correlation between marital status and income. Furthermore, efforts to address gender disparities observed in the dataset, such as encouraging female participation in higher-paying sectors and providing support for working mothers, could help balance income distribution. These targeted interventions can help bridge the income gap and promote financial equality across different demographics.

Conclusion

Looking at the analysis and KNN model implementation, recommendations include refining the model by exploring additional features and alternative algorithms to enhance its predictive accuracy. This stems from its strong performance with an 83.28% accuracy rate in distinguishing between low-income and high-income individuals. Policies should focus on promoting educational attainment and career advancement opportunities, which strongly correlate with higher incomes. Supporting continuing education and stable family structures through initiatives like marriage counseling can also positively impact income levels. Addressing gender disparities in higher-paying sectors and supporting working mothers are crucial for achieving income balance and financial equality across demographics.

References

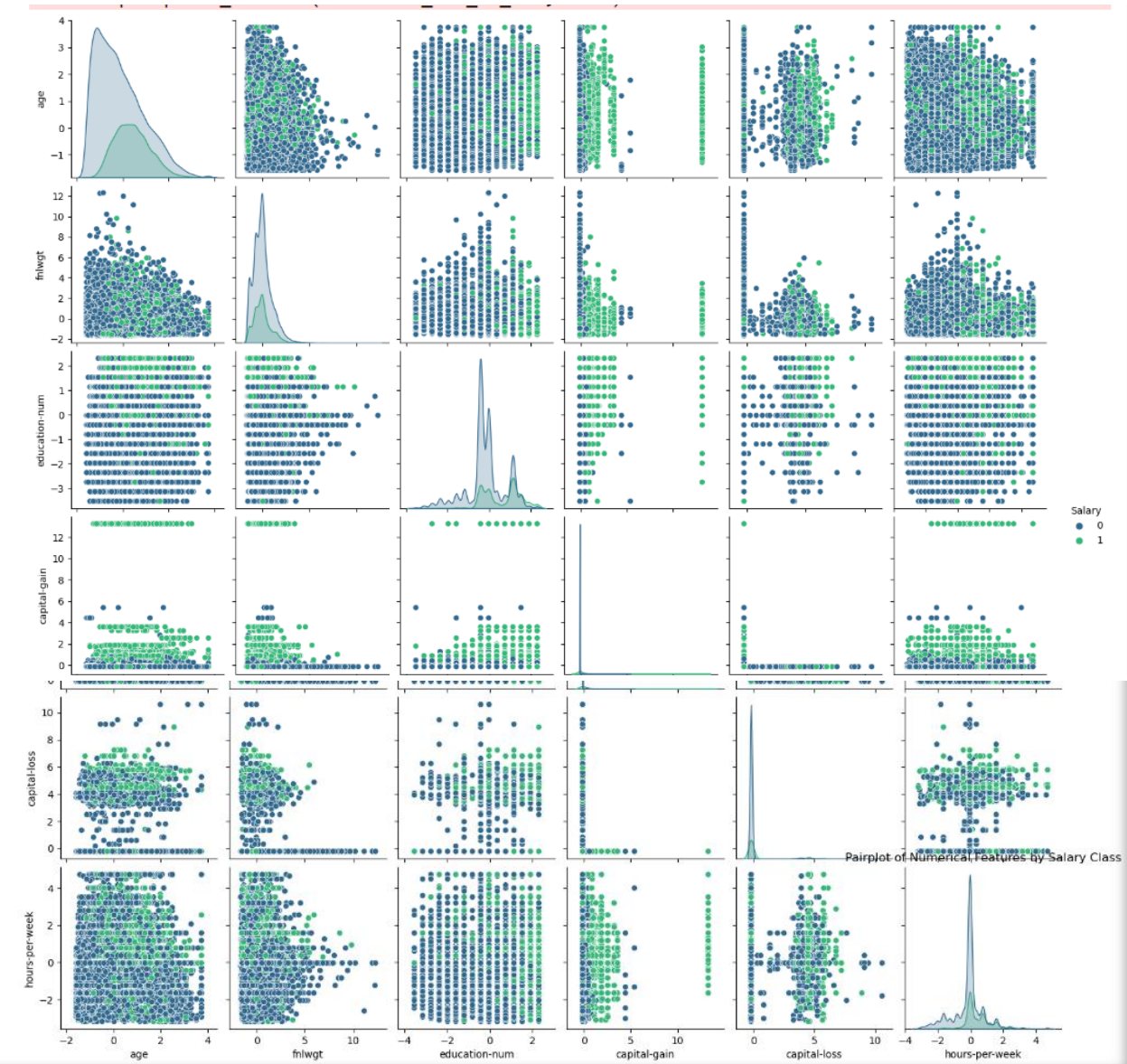
Anaconda Distribution. (2024). Anaconda (Version 4.12.3). Retrieved July 15, 2024, from https://www.anaconda.com

Appendix

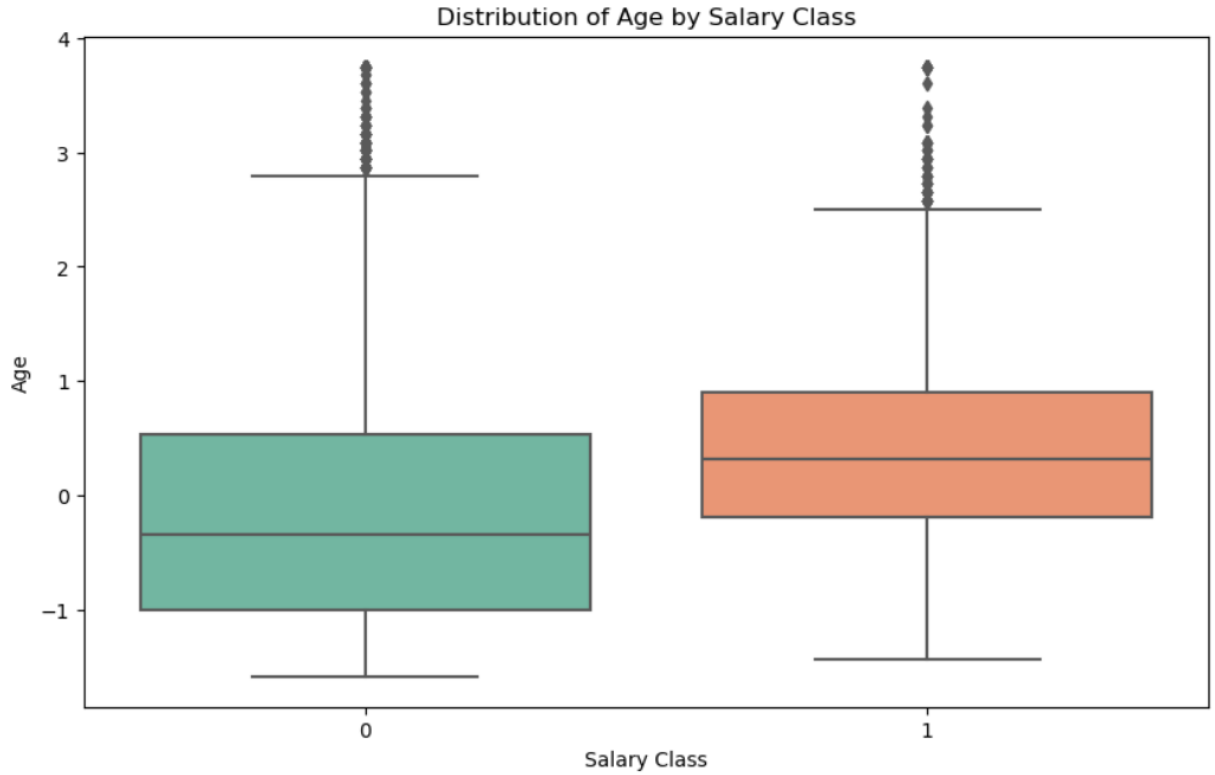
Appendix A:

```
Accuracy: 0.8328385709898659
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.93      0.89      7464
           1       0.69      0.52      0.59      2305

    accuracy                           0.83      9769
   macro avg       0.78      0.72      0.74      9769
weighted avg       0.82      0.83      0.82      9769

Confusion Matrix:
[[6937  527]
 [1106 1199]]
```

Appendix B:

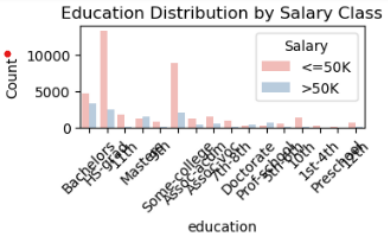

age
fnlwgt
education-num
capital-gain
capital-loss
hours-per-week
Salary
0
1
Pairplot of Numerical Features by Salary Class

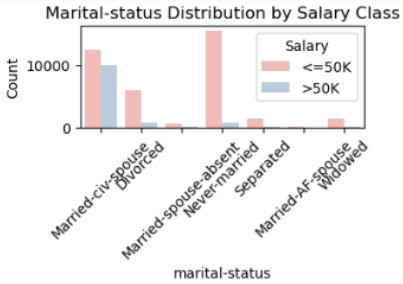
Appendix C:

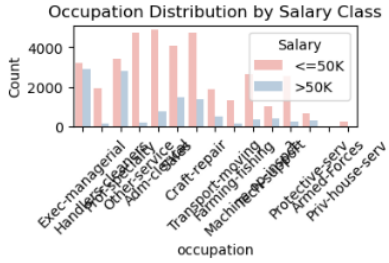

Distribution of Age by Salary Class
Age
Salary Class
4
3
2
1
0
−1
0
1

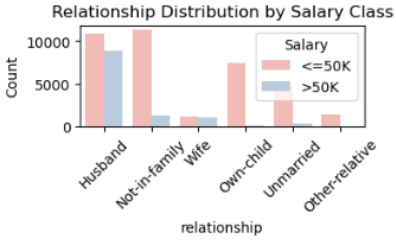
Appendix D:

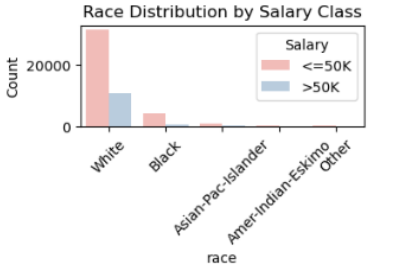

Workclass Distribution by Salary Class
Salary
<=50K
>50K
Count
20000
10000
0
Private
State-gov
Federal-gov
Local-gov
?
Self-emp-inc
Without-pay
Never-worked
workclass

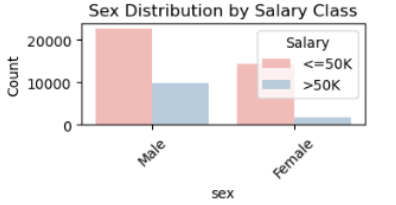

Education Distribution by Salary Class
Salary
<=50K
>50K
Count
10000
5000
0
Bachelors
Masters
Some-college
Doctorate
1st-4th
Preschool
education


Marital-status Distribution by Salary Class
Salary
<=50K
>50K
Count
10000
0
Married-civ-spouse
Divorced
Married-spouse-absent
Never-married
Separated
Married-AF-spouse
Widowed
marital-status


Occupation Distribution by Salary Class
Salary
<=50K
>50K
Count
4000
2000
0
Exec-managerial
Craft-repair
Transport-moving
Protective-serv
Armed-Forces
Priv-house-serv
occupation


Relationship Distribution by Salary Class
Salary
<=50K
>50K
Count
10000
5000
0
Husband
Not-in-family
Wife
Own-child
Unmarried
Other-relative
relationship


Race Distribution by Salary Class
Salary
<=50K
>50K
Count
20000
0
White
Black
Asian-Pac-Islander
Amer-Indian-Eskimo
Other
race


Sex Distribution by Salary Class
Salary
<=50K
>50K
Count
20000
10000
0
Male
Female
sex