DOTE 6756 IS

Business Intelligence Techniques and Applications

Group 1 Project Report

Predicting Income Based on Census Data

Group members：

Name	Student ID
LIN Guanyi	1155218453
LIU Chang	1155220165
WEI Zihan	1155215835
LI Yumiao	1155220698
XIN Yueci	1155220565
LIU Haipo	1155218447
ZHANG Aiwei	1155215891

1. Introduction

1.1 Objective:

The goal of this analysis is to explore the Adult Dataset (also known as the "Census Income" dataset) to predict whether an individual's income exceeds $50,000 per year based on various demographic and socioeconomic attributes. This dataset is widely used in machine learning for classification tasks, particularly in predicting income levels.

1.2 Business Problem or Challenge:

Understanding the factors that influence income levels is crucial for businesses, policymakers, and social scientists. For instance, businesses may use this information to tailor marketing strategies, while policymakers might use it to design targeted social programs. The challenge lies in accurately predicting income categories based on available data, which can help stakeholders make informed decisions.

1.3 Data and context:

The data source is UC Irvine Machine Learning Repository, and the data link is https://archive.ics.uci.edu/dataset/2/adult. The dataset contains information about individuals, including age, education, occupation, work hours, and more. By analyzing this data, we aim to identify patterns and build predictive models that can classify individuals into one of two income categories: ≤ $50K or > $50K. This analysis will provide insights into the key factors that contribute to higher income levels and evaluate the performance of different machine learning models in predicting income.

We had a total of 48,842 data with 14 features and found 52 duplicate values. There are 2795 missing values for Workclass, 2805 missing values for occupation, and 856 missing values for native countries.

2. Research Question and Scope

2.1 Question: Can we accurately predict whether an individual's income exceeds $50,000 per year based on demographic and socioeconomic attributes?

2.2 Scope:

(1) Included: The analysis focuses on the provided dataset, which includes attributes such as age, education, occupation, work hours, marital status, and more. We will perform exploratory data analysis (EDA) to understand the data distribution and relationships between variables. We will also build and evaluate three machine learning models: K-Nearest Neighbors (KNN), Logistic Regression, and Random Forest.

(2) Excluded: The analysis does not include external data sources or additional attributes not present in the dataset. We also do not address causal relationships or policy recommendations directly.

3. Data Analysis and Modeling Result

3.1 EDA

The cleaned dataset consists of 45,175 records, providing statistical insights into variables such as age, years of education, capital gain, capital loss, and hours worked per week. It can be found that our data features are either int type or category type, and there are more of the latter, so we may convert int type and category type to each other to satisfy each model.

3.1.1 Listing of attributes:

Dependent variable: Income: $> \$50K$ or $\leq \$50K$

Independent Variables have numerical variables and category variables.

(1) Numerical variables:

1. age: continuous.
2. fnlwgt: the number of people the census believes the entry represents.
3. education-num: continuous.
4. capital-gain: continuous.
5. capital-loss: continuous.
6. hours-per-week: continuous.

(2) Category variables:

7. workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
8. education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
9. marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
10. occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspect, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
11. relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
12. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
13. sex: Female, Male.

14. native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad & Tobago, Peru, Hong, Holand-Netherlands.

3.1.2 Descriptive statistics for numerical variables

The average age is 38.56 years, with a minimum of 17 and a maximum of 90 years. Education levels vary, with an average of 10.12 years and a maximum of 16 years. Capital gain and capital loss exhibit significant skewness, as most values are zero, but extreme values are present, with the highest capital gain reaching 99,999 and the highest capital loss at 4,356. This suggests that only a small subset of individuals report substantial capital transactions. The distribution of weekly working hours indicates that most individuals work around 40 hours per week, as reflected in the median and quartiles. However, the maximum value of 99 hours suggests the presence of outliers or individuals working exceptionally long hours.

	age	fnlwgt	education-num	capital-gain	capital-loss	hours-per-week
count	45175.000000	4.517500e+04	45175.000000	45175.000000	45175.000000	45175.000000
mean	38.556170	1.897388e+05	10.119314	1102.576270	88.687593	40.942512
std	13.215349	1.056524e+05	2.551740	7510.249876	405.156611	12.007730
min	17.000000	1.349200e+04	1.000000	0.000000	0.000000	1.000000
25%	28.000000	1.173925e+05	9.000000	0.000000	0.000000	40.000000
50%	37.000000	1.783120e+05	10.000000	0.000000	0.000000	40.000000
75%	47.000000	2.379030e+05	13.000000	0.000000	0.000000	45.000000
max	90.000000	1.490400e+06	16.000000	99999.000000	4356.000000	99.000000

Table 3.1.1

3.1.3 Preprocessing

Our dependent variable, income, is a categorical variable with 50K as the dividing line. If it exceeds 50K, it is 1, and if it is lower than or equal to 50K, it is 0. We divide age into four categories: teenager, adults, mid-adults and elderly.

3.1.4 Relationship among categorical variables

(1) Age and Income

The analysis reveals a clear relationship between age and income. Adults in their prime working years (referred to as "Adult") tend to have the highest income levels, followed by middle-aged adults ("Mid Adults"). The elderly and teenagers, who are either retired or not yet fully integrated into the workforce, exhibit lower income levels. This pattern aligns with

expectations, as income typically peaks during one's prime working years and declines post-retirement or during early career stages.

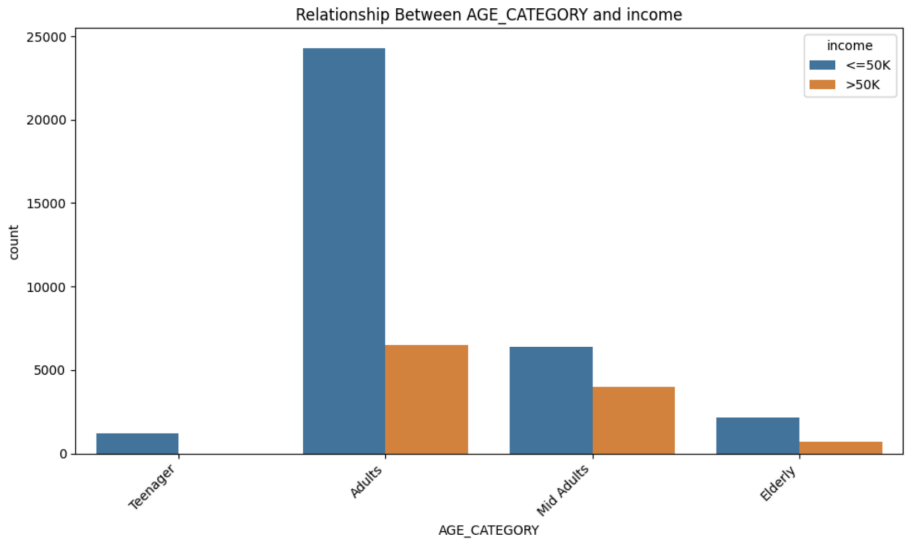

Figure 3.1.1

(2) Workclass and Income

The relationship between workclass and income levels is significant. The private workclass stands out as the category with the highest number of individuals earning below 50K. However, it is also the workclass with the highest number of individuals earning above 50K. This dual trend suggests that while the private sector offers opportunities for high earnings, it also encompasses a large proportion of lower-income workers, possibly due to the diversity of roles within this sector.

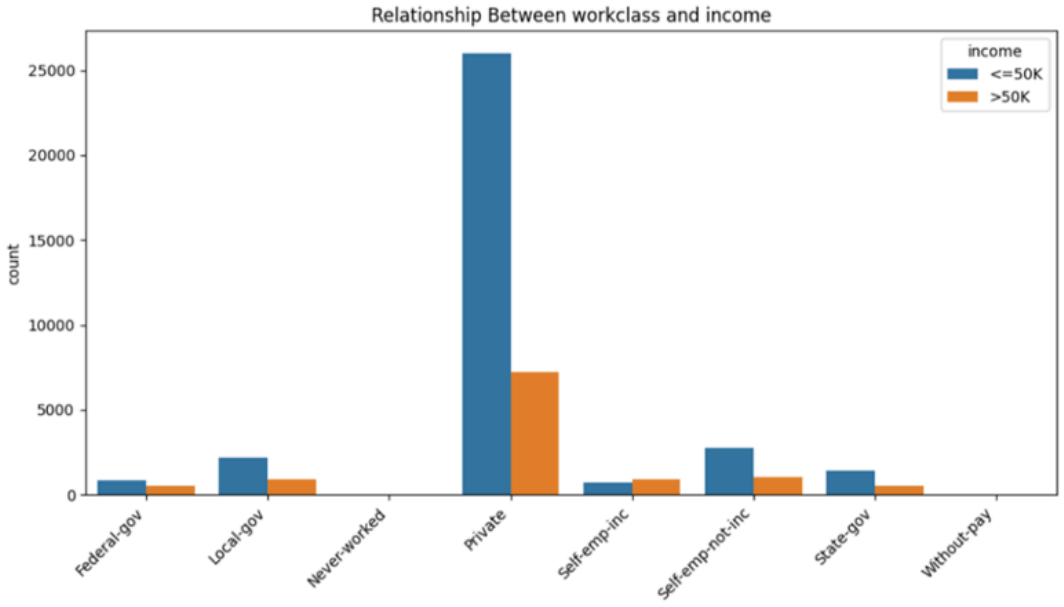

Figure 3.1.2

(3) Marital Status and Income

Marital status appears to be strongly correlated with income levels. Individuals in stable marriages, particularly those who are married, tend to have higher incomes. In contrast, those who are single, divorced, or have an absent spouse are more likely to fall into lower income brackets. This finding suggests that marital stability may contribute to financial stability, possibly due to dual incomes or shared financial responsibilities in married households.

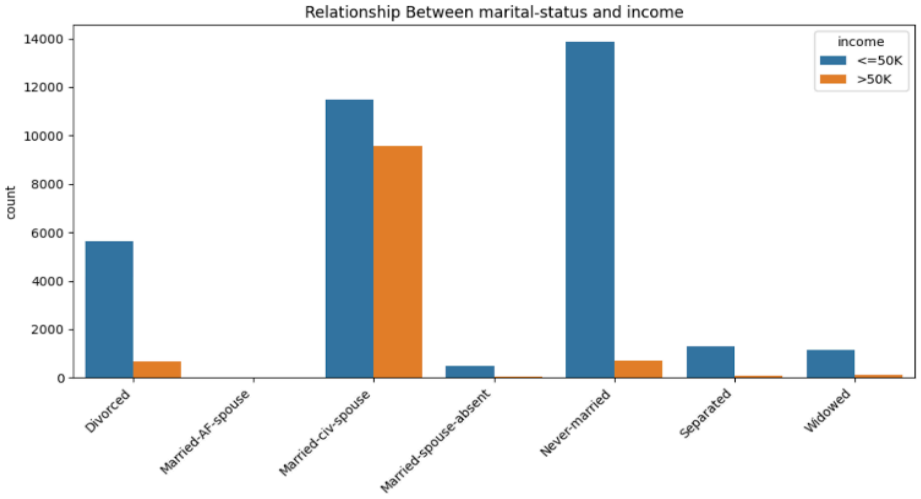

Figure 3.1.3

(4) Family Role (relationship) and Income

The analysis of family roles and income reveals that husbands have the highest proportion of high-income earners. In contrast, wives, children, and other relatives are more likely to be in lower-income categories. Unmarried individuals also tend to have lower earnings. This indicates that the role of being a husband is strongly associated with higher income, while other family roles are more likely linked to lower earnings. This disparity may reflect traditional gender roles and the distribution of labor within households.

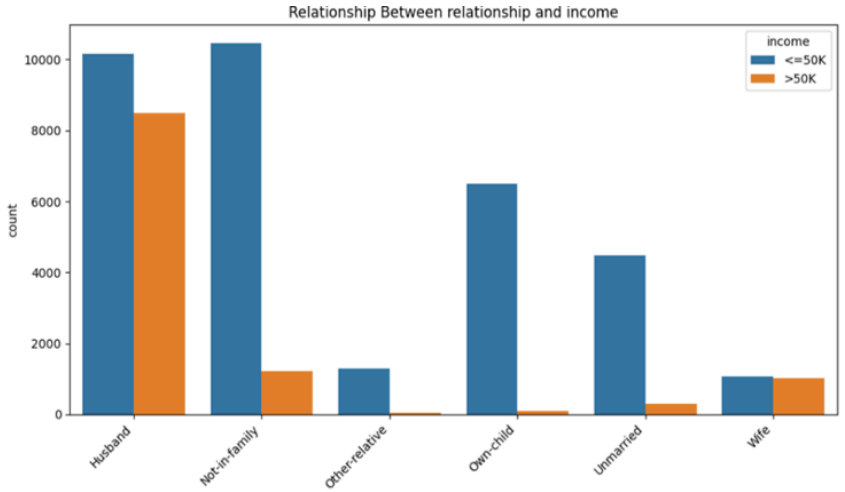

Figure 3.1.4

(5) Occupation and Income

Occupation type has a significant impact on income levels. Certain occupations, such as administrative, managerial, and sales roles, have a higher proportion of individuals earning over 50K. Conversely, occupations in service and craft-related fields are predominantly associated with incomes of 50K or less. This suggests that higher-paying occupations are concentrated in specific sectors, while lower-paying jobs are more prevalent in others.

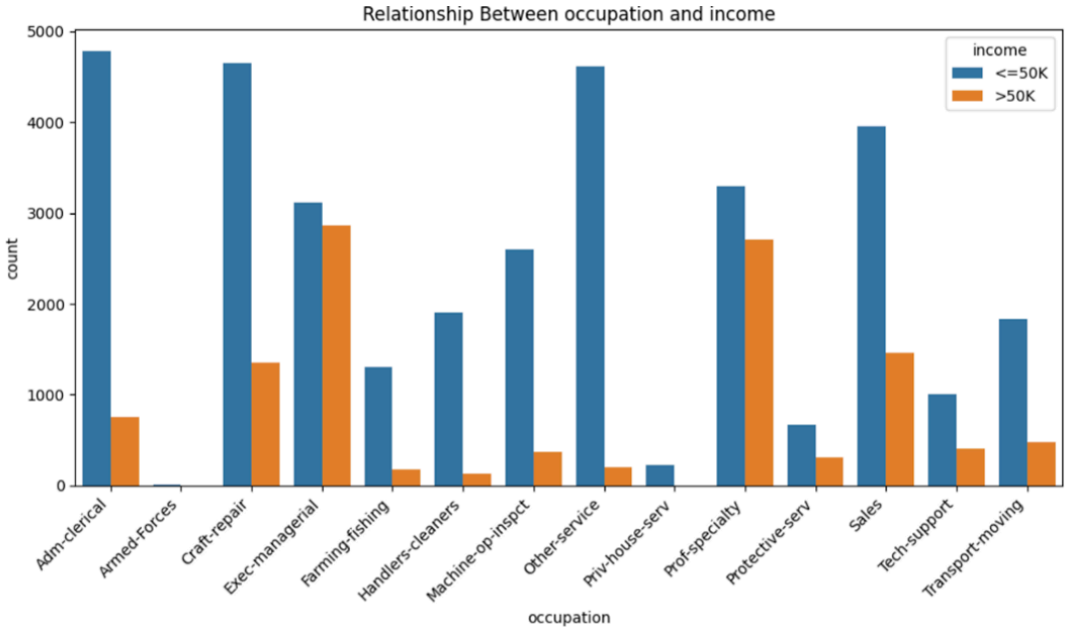

Figure 3.1.5

(6) Race and Income

Income disparities are evident across racial groups. The majority of individuals earning over 50K are White. This highlights significant racial disparities in income distribution, with White individuals more likely to achieve higher earnings compared to other racial groups.

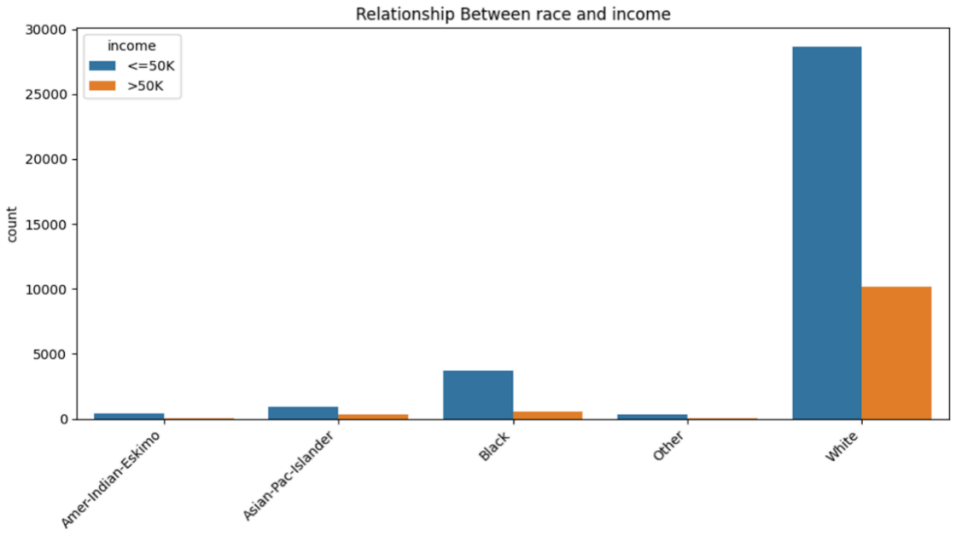

Figure 3.1.6

(7) Sex and Income

The analysis shows a notable income disparity based on sex. A larger number of males earn over 50K compared to females, although the majority of both sexes fall into the <=50K income range. This indicates that males have a higher propensity to earn above the 50K threshold, reflecting ongoing gender-based income inequality in the workforce.

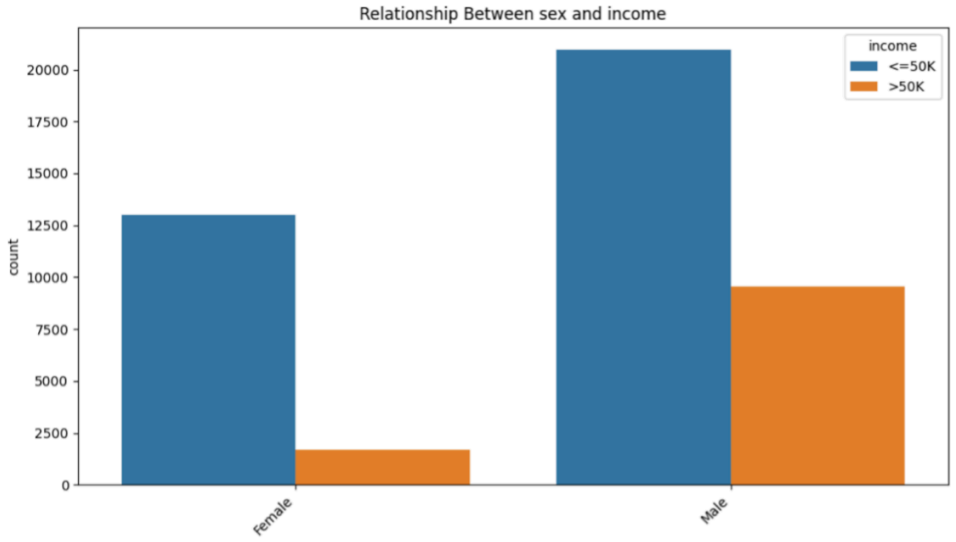

Figure 3.1.7

(8) Native Country and Income

The data visualized in the chart shows the distribution of income levels across different native countries. However, since the study primarily focuses on U.S. citizens, the chart's results may not directly reflect broader trends relating to income across various countries.

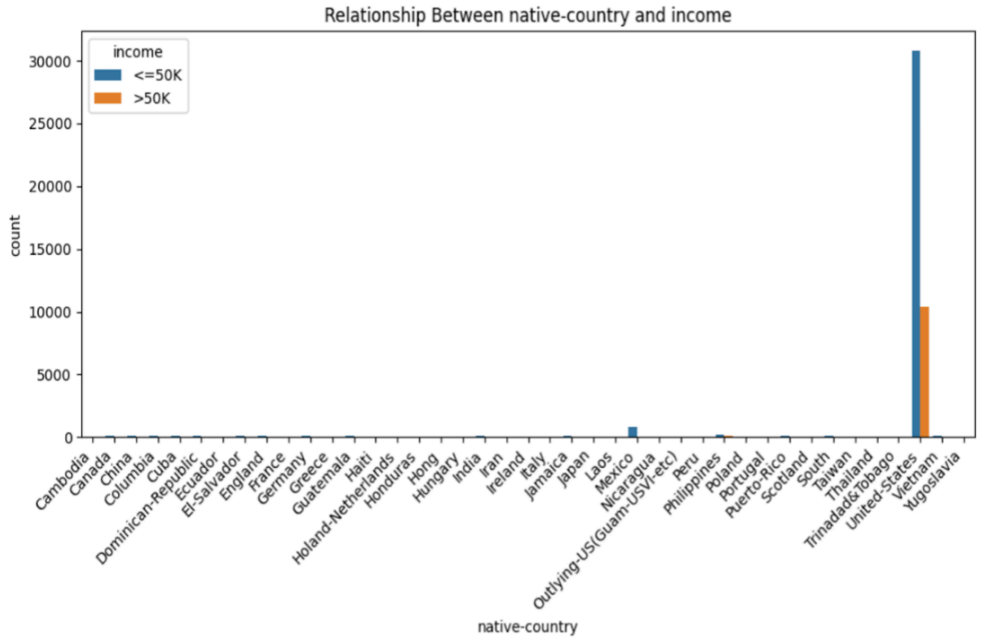

Figure 3.1.8

3.1.5 Relationship among numerical variables

(1) Income and education-num

The median age for those earning over 50K is higher than that for those earning 50K or less, indicating that older individuals tend to have higher incomes. Additionally, the interquartile range (IQR) for the >50K group is narrower, suggesting that higher earners have less variability in age compared to lower earners.

(2) Income and working hours

The median education level for those earning over 50K is higher than that for those earning 50K or less, indicating that individuals with higher education tend to earn more. The interquartile range (IQR) for the >50K group is also wider, suggesting greater variability in education levels among higher earners. The presence of outliers in both income groups highlights a diverse distribution of education levels within each category.

(3) Income and age

It shows that individuals earning over 50K tend to work more hours per week on average compared to those earning 50K or less. The median hours for the >50K group are higher, and the interquartile range (IQR) is also broader, indicating greater variability in hours worked among higher earners.

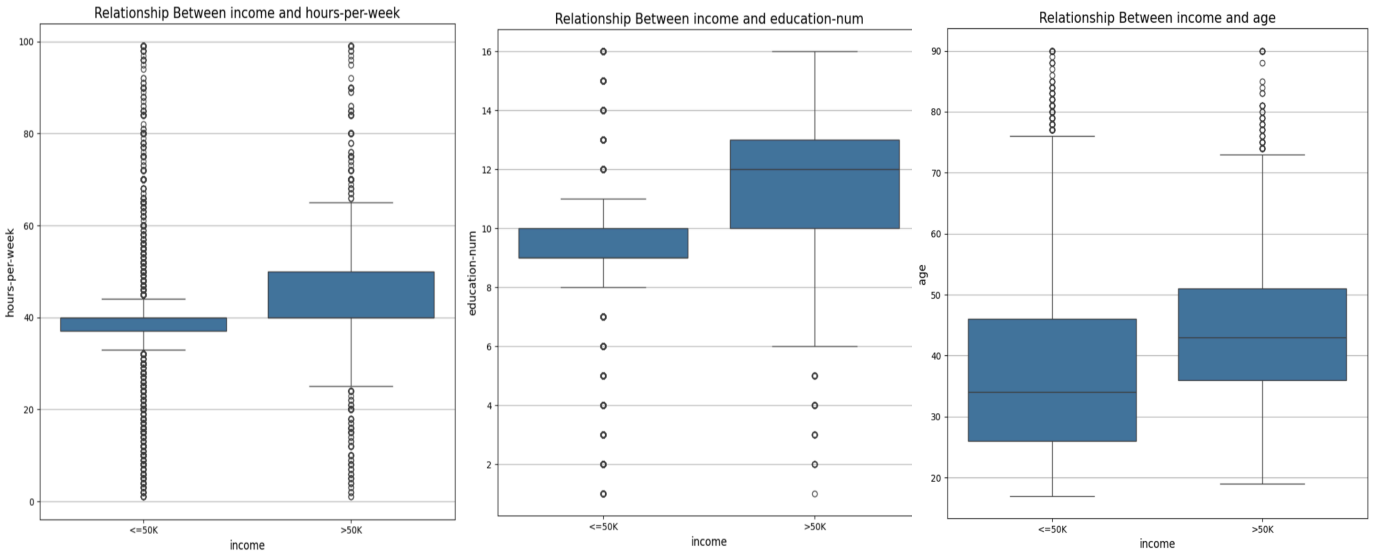

Figure 3.1.9

3.1.6 Pair Plot of numerical variables

To better illustrate, we made a Pair Plot of six numerical variables. The relationship between income and other variables shows that higher income groups tend to have a broader age distribution, with a generally higher median age, suggesting that older individuals often earn

more. Additionally, there is a clear correlation between income and education level, indicating that higher education is associated with higher earnings.

Regarding capital gains, individuals with higher incomes tend to have greater capital gains, highlighting a potential connection between investment success and income level. Overall, these visualizations suggest that age, education, and capital gains are significant factors influencing income levels.

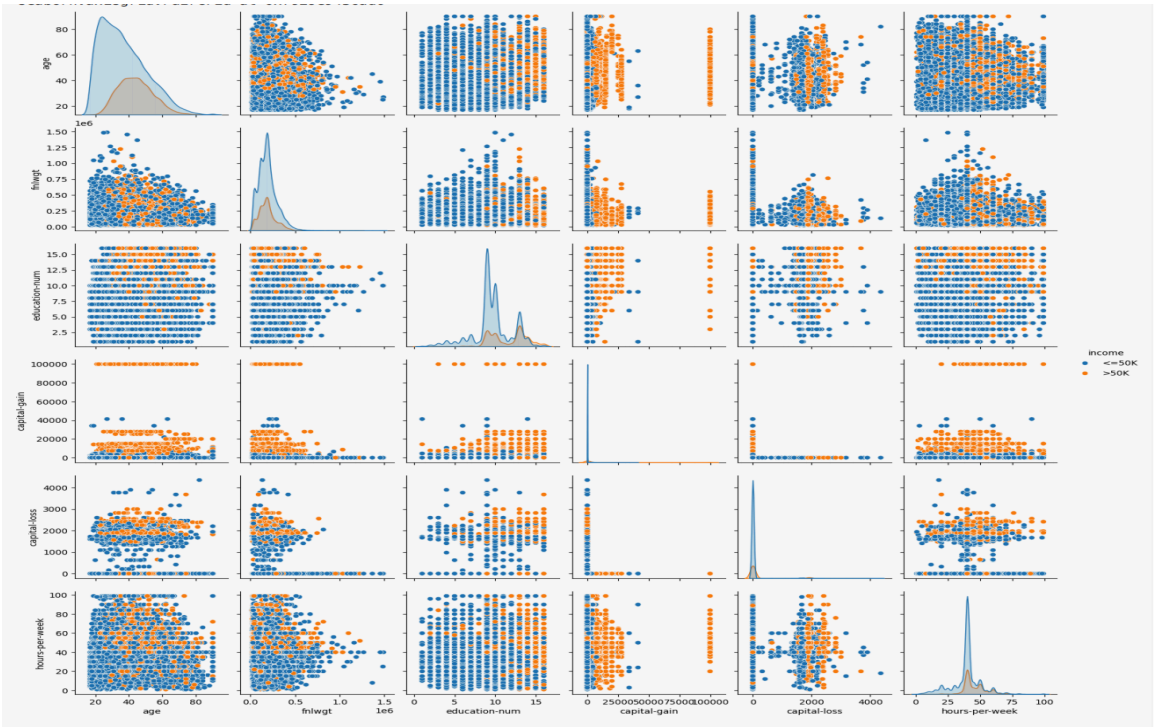

Figure 3.1.10

3.1.7 Education's Impact on Career and Income

In the Sankey diagram, the left side represents education levels, such as High School, Bachelor's, and Graduate degrees. The middle section depicts different occupations, like Manager, Technician, and others. The right side presents income levels, categorized as income larger than 50K and income equal or smaller than 50K. The width of the flows between these sections indicates the number of individuals transitioning from one category to another.

The diagram suggests that individuals with higher education are more likely to occupy higher-paying jobs, indicating a clear pathway where education influences both career choices and income potential.

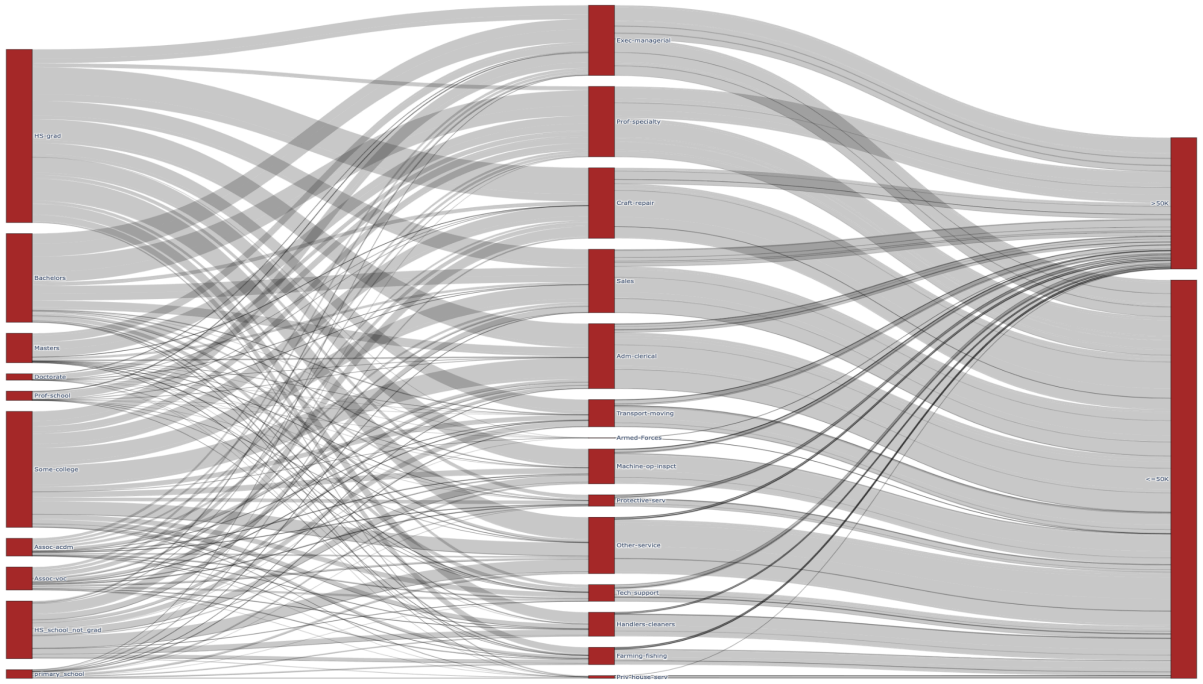

Figure 3.1.11

3.1.8 Skewness and log transformation

The skewness(Figure 3.1.12) is calculated separately for Numerical variables. If the skewness is greater than 1 or less than -1, logarithmic transformation() is carried out to make the data closer to the normal distribution, thereby improving the effect of subsequent analysis and modeling.

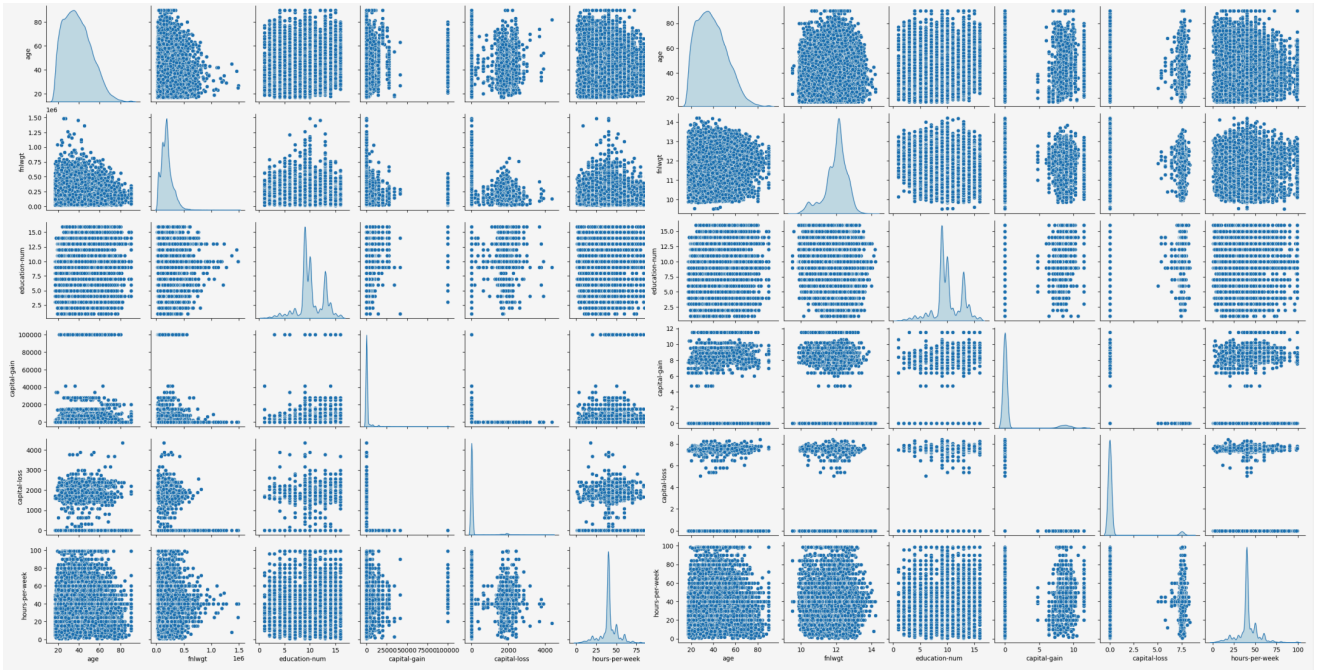

Figure 3.1.12

Figure 3.1.13

3.1.9 Correlation Matrix for Numeric Variables:

Overall, the correlations between these variables are generally low, with no strong relationships observed. Education-num (years of education) has a slight positive correlation with capital-gain(0.13) and hours-per-week(0.15), suggesting that higher education levels may influence capital gains and working hours. Capital-gain and capital-loss have a low correlation (-0.067), indicating that capital gains and losses do not always correspond to each other.

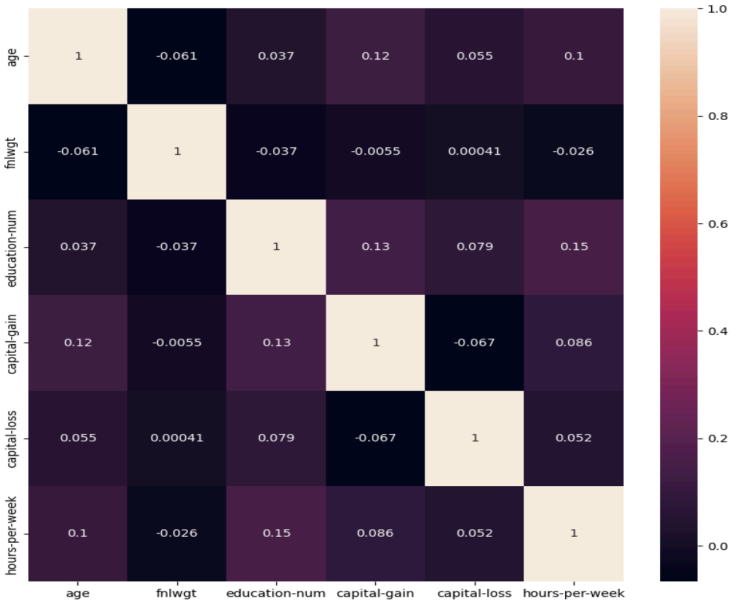

Figure 3.1.14

3.2 KNN

The results of a K-Nearest Neighbors (KNN) classification model are presented. The dataset comprises both categorical and numerical features, with the model evaluated through cross-validation and tested on a holdout dataset. The primary aim of this analysis is to assess the model's performance and provide insights into its predictive capabilities.

The dataset was preprocessed to handle categorical and numerical variables. A preprocessing pipeline was created to standardize the data and encode categorical features. A KNN classifier was integrated into a pipeline that included the preprocessing steps and the classifier. A grid search with 5-fold cross-validation was performed to identify the optimal value of n_neighbors (K) for the KNN model. The tested values for K were: [3, 5, 7, 9, 11, 13, 15, 17]. The model was trained on the training set using the best K value identified during grid search. Its performance was evaluated on a separate test set using accuracy and a confusion matrix.

The best K value identified through grid search is 13, which achieved the highest cross-validated accuracy(0.84). The training set accuracy is 0.86, and the test set accuracy is

0.84. The confusion matrix provides a detailed breakdown of the model's predictions based on testing dataset:

1. **True Negatives (TN)**, where the model correctly identified negative cases (6,180 instances)
2. **False Positives (FP)**, where negative cases were incorrectly classified as positive (662 instances);
3. **False Negatives (FN)**, where positive cases were incorrectly classified as negative (811 instances)
4. **True Positives (TP)**, where the model correctly identified positive cases (1,382 instances).

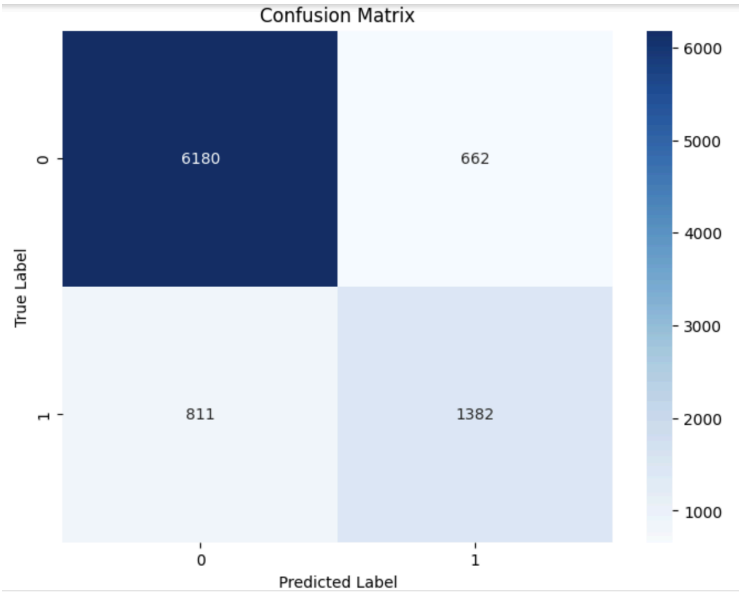

Figure 3.2.1

Based on the confusion matrix we can calculate:

1. The accuracy is calculated as the proportion of correctly classified instances out of the total dataset.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{1382 + 6180}{6180 + 662 + 811 + 1382} \approx 0.84$$

2. Precision measures how many of the instances predicted as positive are actually positive.

$$Precision = \frac{TP}{TP + FP} = \frac{1382}{1382 + 662} \approx 0.64$$

3. Another crucial metric is the recall, which measures the model's ability to correctly identify actual positive cases.

$$Recall = \frac{TP}{TP + FN} = \frac{1382}{1382 + 811} \approx 0.63$$

The accuracy of the model was about 84%, showing that the model performed well on the overall prediction. The precision rate and recall rate is relatively low, in class a prediction shows that the model has certain improvement space.

3.3 Logistic Regression

Logistic Regression is mainly used to predict the case where the dependent variable is binary, by converting the logit function into probability values and thereby predicting the categorical outcome, to visually understand the extent to which each feature affects the predicted outcome.

In this case, we converted the target variable income to a binary label, where 0 means <=$50,000 and 1 means >$50,000. Where if a positive coefficient is shown, it means that an increase in the value of the corresponding feature is associated with an increase in the probability of having an income of more than $50,000, while a negative coefficient means it is irrelevant.

In general, this model has an accuracy of 84% on both of the test set and train set, it means that the model is able to accurately predict whether or not 84% of the individuals have an income of more than $50K. This demonstrates that the model has a high predictive power (which greater than 80%). Figure 3.3.1 shows the confusion matrix for the logistic regression model, it shows the predictions of the model on the test set compared to the actual results:
1. True Positives (TP): the number of instances that the model correctly predicts as positive classes (>50K) (i.e., 1272 in the bottom right corner of the figure);
2. False Positives (FP): the number of instances that the model incorrectly predicts as positive (>50K) (i.e., 545 in the upper right corner of the figure);
3. True Negatives (TN): the number of instances correctly predicted by the model to be in the negative category (<=50K) (i.e., 6297 in the upper left corner of the figure);
4. False Negatives (FN): the number of instances that the model incorrectly predicts to be in the negative category (<=50K) (i.e., 9210 in the lower-left corner of the figure).

Thus, the overall model accuracy is approximately equal to 84%,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{1272 + 6297}{1272 + 6297 + 545 + 9210} = 0.837 \approx 0.84$$

This means that the results of the confusion matrix are consistent with the overall model accuracy shown in the figure being 0.84, which demonstrates that the Logit model has a good predictive power (greater than 80%). Also, the higher number of TP and TN indicates that the model is better at recognizing positive and negative classes.

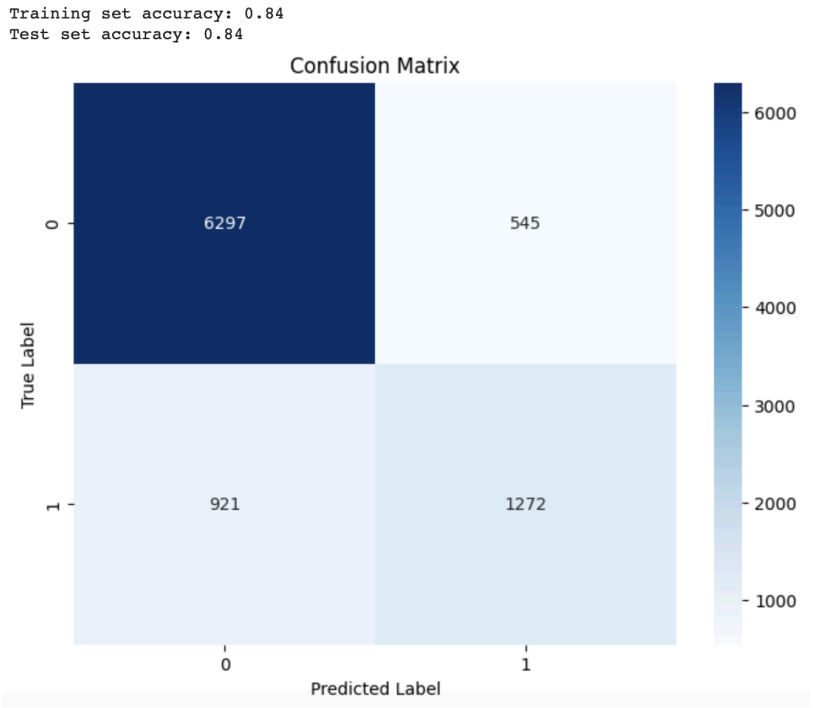

Figure 3.3.1

The bar chart in figure 3.3.2 demonstrates the coefficients of each feature in the logistic regression models, the larger the absolute value of the coefficients of the features, the more it indicates that there is a greater impact of these features on income over $50k. As can be seen from the figure, martial-status_Married-AF-spouse has the largest positive coefficient (close to 1.5), which indicates that it has the largest impact on predicting income levels over $50K among all features. Other characteristics such as martial-status_Married-civ-spouse, native-country_Portugal, relationship_Wife, occupancy_Exce-managerial, native-country_ Canada, native-country_England, native-country_Italy also have high positive coefficients (>0.5), which represents a strong correlation between these characteristics and income level as well. While some features such as native-country_Columbia, martial-status_Never-married have negative coefficients (close to -1.5), which indicates that these features are associated with lower income levels.

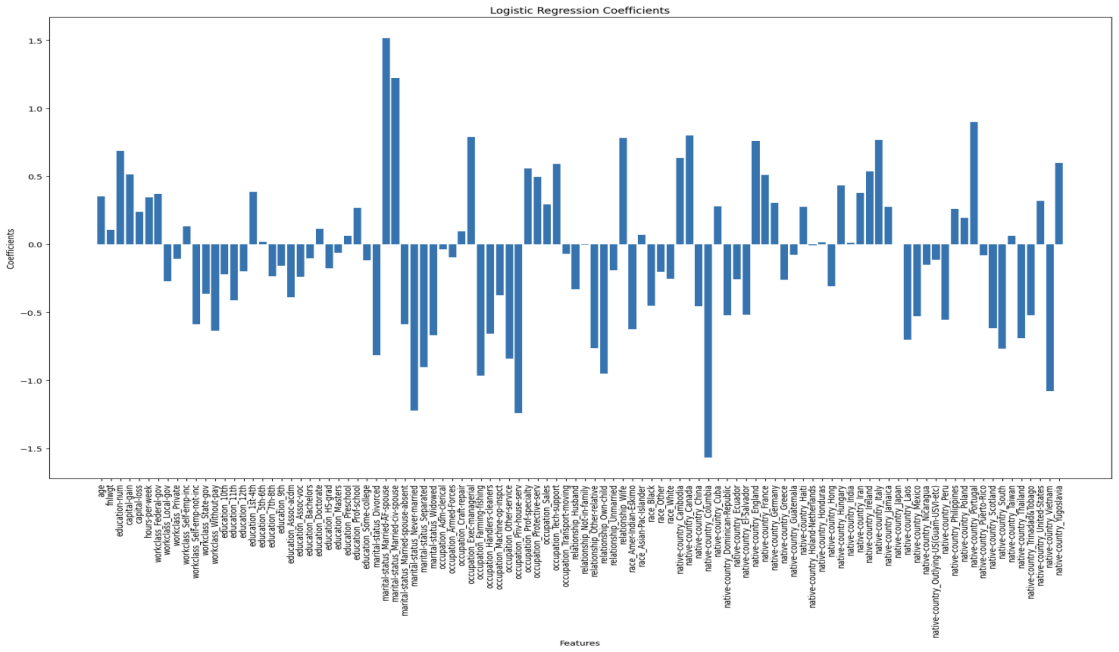

Figure 3.3.2

3.4 LASSO

LASSO regression is a special linear regression that performs feature selection. In this case, we have set the variable to be dichotomous (whether or not the income is more than 50K) and use the penalty='L1' parameter regularization term in Logistic Regression to implement LASSO regression.

In the LASSO regression model, we implemented feature selection by reducing some coefficients to 0. This means that if some features do not contribute to the predictive results of the model, we will remove them, which simplifies the model and provides explainability. The bar chart in Figure 3.4 shows the non-zero coefficients characterizing the main influences and their extent obtained using the LASSO regression analysis.

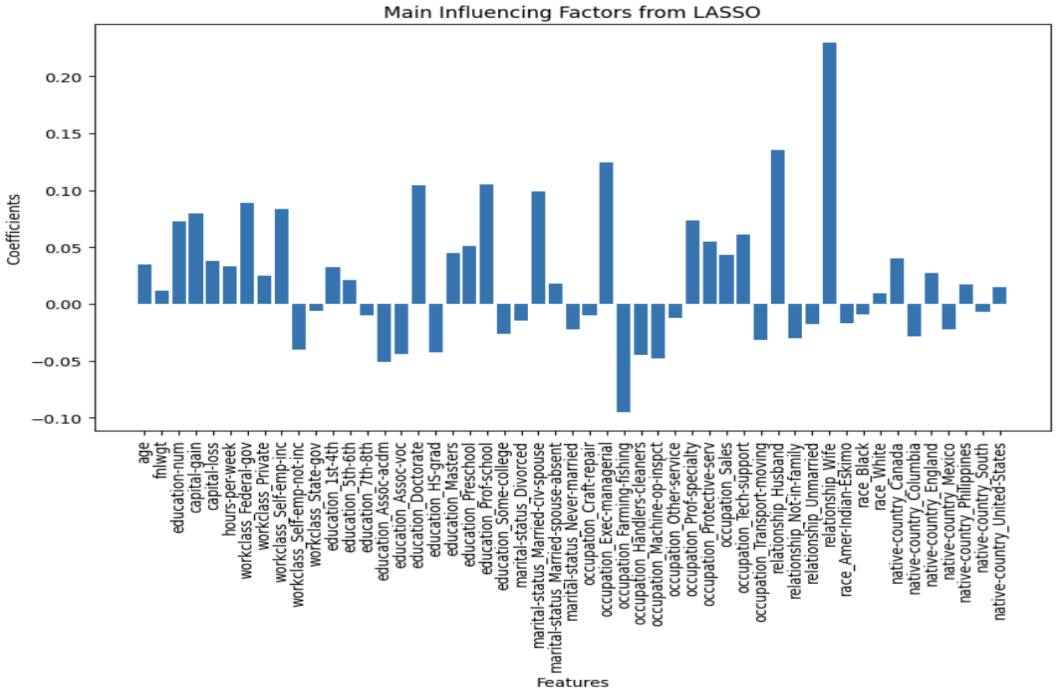

Figure 3.4

In particular, relationships_Wife has the highest positive coefficient (>0.20), and relationships_Husband and occupation_Exec-managerial have positive coefficients of more than 0.10. These features represent strong positive correlations with high income. However, the negative coefficient of occupation_Farming-fishing is close to -0.10, which also represents its relationship with low income. Overall, characteristics with larger absolute values of coefficients have a higher impact on the predictive results of the model. For instance, relationships and occupation are shown to have a significant impact on predicting whether or not the annual income exceeds $50K.

3.5 Decision Tree

In decision tree modeling, we define the feature significance by calculating the information gain of each feature in the tree splits, which in our model includes feature visualization, hyperparameter tuning (pre-pruning and post pruning), and model evaluation. In this case, the training set of the model has an accuracy of 0.99991, which is close to 100%, indicating that the model performs extremely well on the training set and there is overfitting. Meanwhile, the test set accuracy is 0.8152, which also represents a high performance.

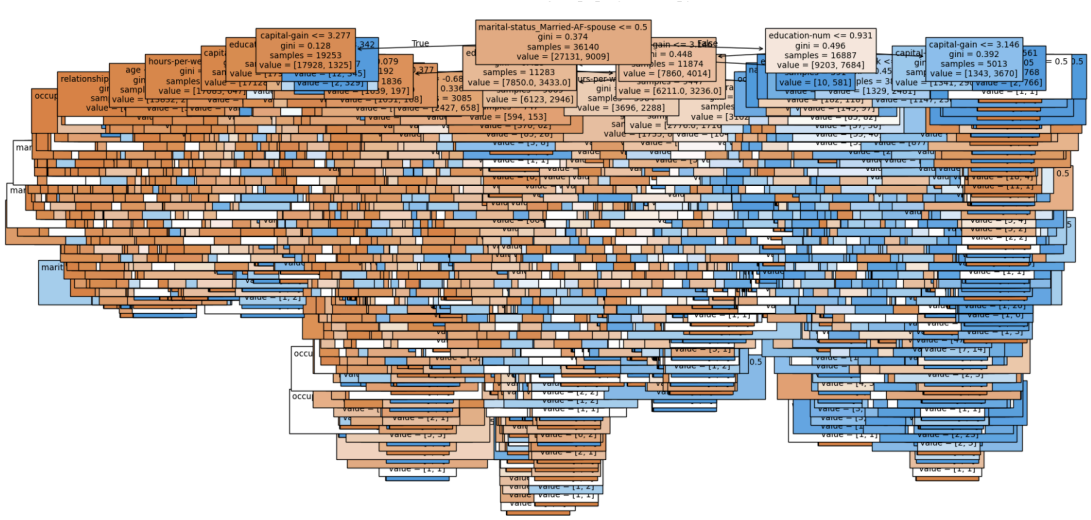

Figure 3.5.1

3.5.1 Feature Visualization

In this part, Figure 3.5.2 shows a bar chart of the top 10 most important features, where it can be seen that the 2 features, martial-status_Married-AF-spouse and fnlwgt, contribute the most to the model's decision impact and predictive power, with 0.200 and 0.195, respectively. It also indicates that marital status has significantly affected the model forecastings.

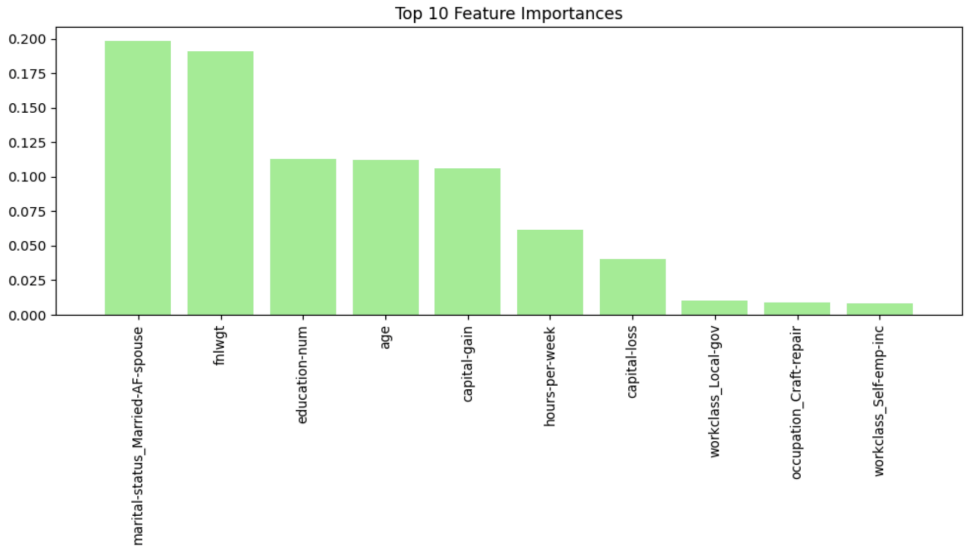

Figure 3.5.2

3.5.2 Hyperparameter Tuning

3.5.2.1 Pre-Pruning

Since the accuracy of the model was overfitted, we used hyperparameter tuning to adjust it. The hyperparameters of the decision tree are scaled by GridSearchCV and a grid search is performed to obtain the best model. The accuracies of the training set and test set are 0.8599

and 0.8548 respectively, which indicates that both the accuracies have reduced and the gap between the training set and the test set is reduced. This represents an improvement in the performance of the model and its ability to generalize and to reduce the overfitting. Figure 3.5.3 shows a pre-pruned decision tree. In particular, features that split up earlier in the decision tree (e.g., marital-status_Married-AF-spouse) have a greater impact on the prediction results of the model.

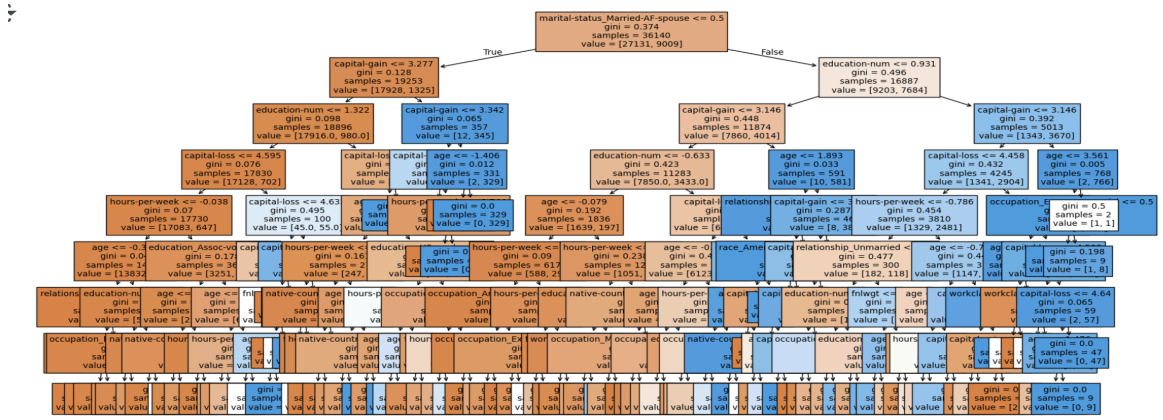

Figure 3.5.3

3.5.2.2 Post-Pruning

Then, we prune on top of prepruning to generate a post pruning model. In the post pruning section, we use ccp_alpha by cost-complexity pruning to further optimize the model and reduce overfitting. By plotting the number of nodes and depth with alpha value (Figure 3.5.4), the optimal alpha value of 0.0002 was chosen to prune the decision tree. From the graph, we can see that the number of nodes increases as the alpha value increases, which means that the model becomes simpler. Meanwhile, as the alpha value increases, the training set accuracy declines, which shows that the model fits over the training data less well.

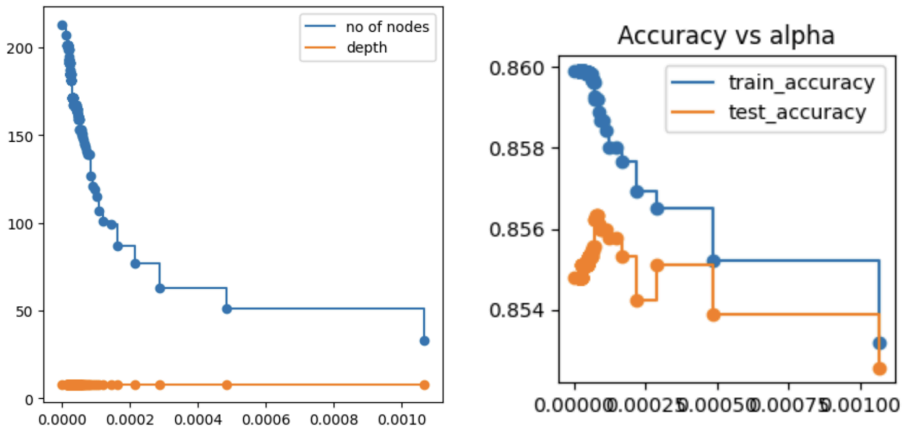

Figure 3.5.4

In the model generated after pruning (shown in Fig. 3.5.5), the training set accuracy is 0.8571 and the test set accuracy is 0.8544. The performance of the post-pruning model on the training and test sets is similar to that of the pre-pruned model, but the accuracy on the test set is higher in the pre-pruning. This means that using the pre-pruned model is more convincing.

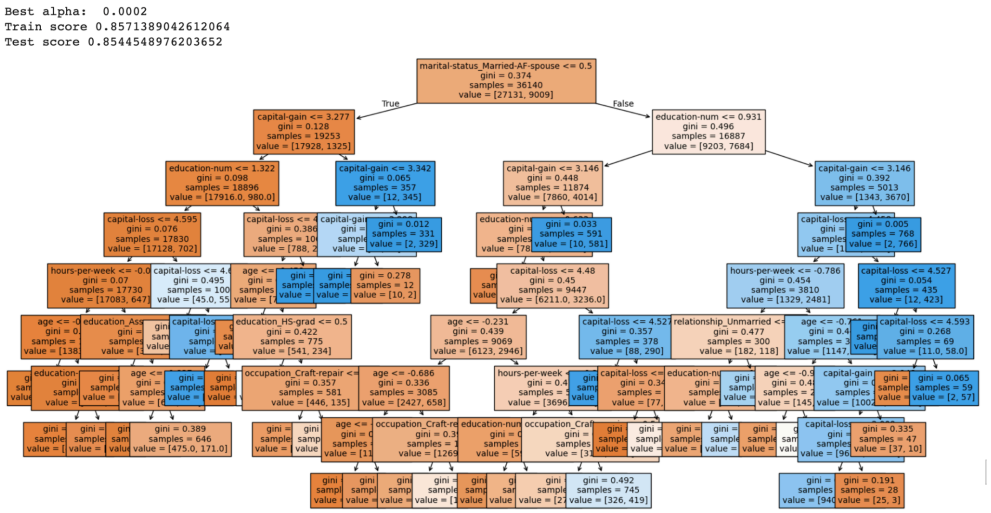

Figure 3.5.5

3.5.3 Model Evaluation

Since the accuracy of the model in pre-pruning is relatively higher, we followed the pre-pruning model in evaluation. After hyperparameter tuning with max depth = 8, min samples split =10, and min samples leaf=2, the train and test accuracy of the final model is 0.8599 and 0.8547 respectively (shown in Fig. 3.5.6), reflecting a better generalization ability, with the overall model predicting over 85% of all samples tested correctly. In the model's classification performance report, category 0 has a precision (the proportion of samples predicted to be positive that are actually positive) of 0.88, a recall (the proportion of samples that are actually positive that are correctly predicted to be positive) of 0.94, and an F1 (the reconciled mean of precision and recall) score of 0.91. Category 1 has a precision of 0.76, a recall of 0.58, and an F1 score of 0.66. In the confusion matrix, the number of correctly categorized samples for category 0 is 6446 and the number of correctly categorized samples for category 1 is 1277. While the number of incorrectly categorized samples for category 0 and category 1 are 396 and 916 respectively. This shows that the model has a prediction ability. Generally, the final adapted decision tree is shown in the figure below (Fig. 3.5.7), which the root node of the decision tree is based on the marital-status_Married-AF-spouse features are split, and the features that contribute the most to the predictive ability of the model are seen.

```
Train score 0.8599059214167127
Test score 0.854786939679026

分类报告:
              precision    recall  f1-score   support

           0       0.88      0.94      0.91      6842
           1       0.76      0.58      0.66      2193

    accuracy                           0.85      9035
   macro avg       0.82      0.76      0.78      9035
weighted avg       0.85      0.85      0.85      9035
```

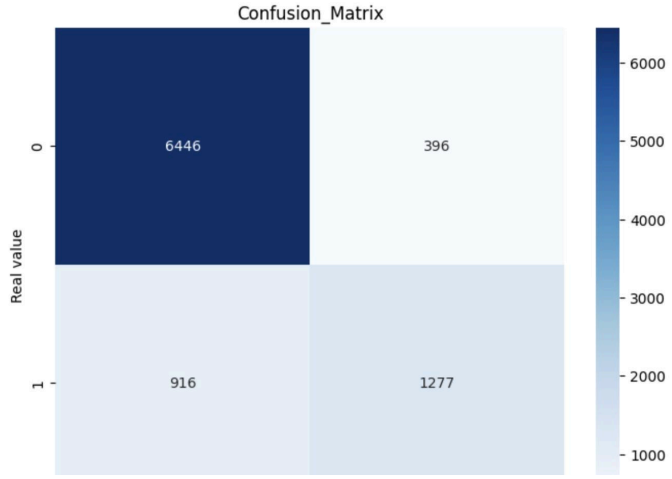

Figure 3.5.6

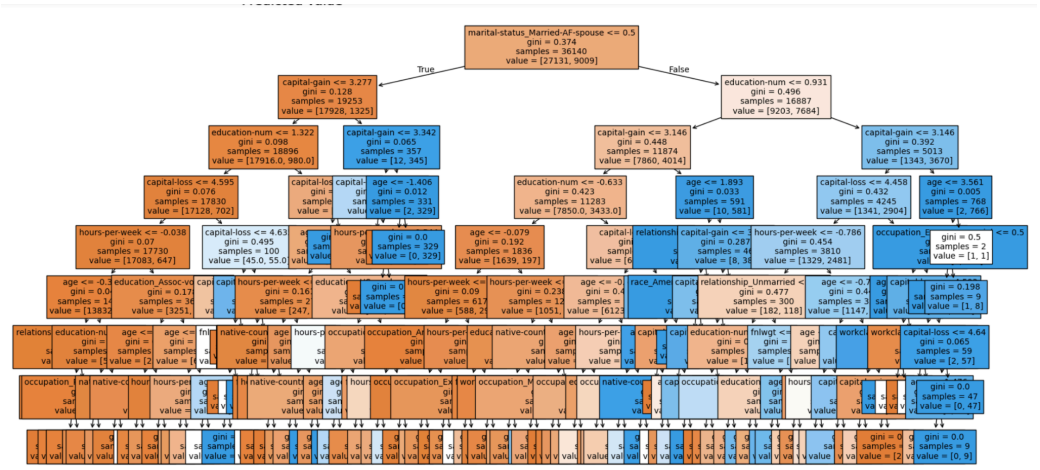

Figure 3.5.7

The following figure shows the 10 important features and their importance scores (Fig. 3.5.8) that the decision tree has been adjusted to re-split, and compared to the previous one it can be seen that the ranking of the different features contributing to the predictive power of the model has changed considerably. Undeniably, marital-status_Married-AF-spouse is still the most important feature (over 0.40), which indicates that marital status has a significant effect on the level of income. Unlike previously, the score for education-num (0.23) moves from third to second place, followed closely by capital-gain (close to 0.20).

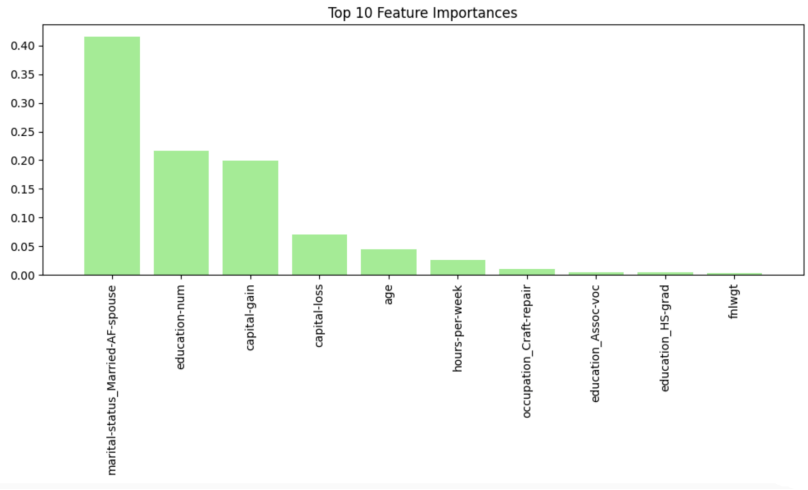

Figure 3.5.8

3.6 Random Forest

In Random Forest, we perform feature visualization analysis, hyperparameter tuning, gradient boosting, and model evaluation. We evaluated the initial accuracy of the model, which was 0.9993 for training set and 0.8465 for testing set, indicating that the model performs very well on the training data, and there is overfitting.

3.6.1 Feature Visualization

We plotted bar graphs (Figure 3.6.1) showing the importance of the top 10 features, rating the average importance of features for each decision tree in the random forest. Among them, the features fnlwgt (0.16) and age (0.14) are shown to contribute the most to the model predictions.

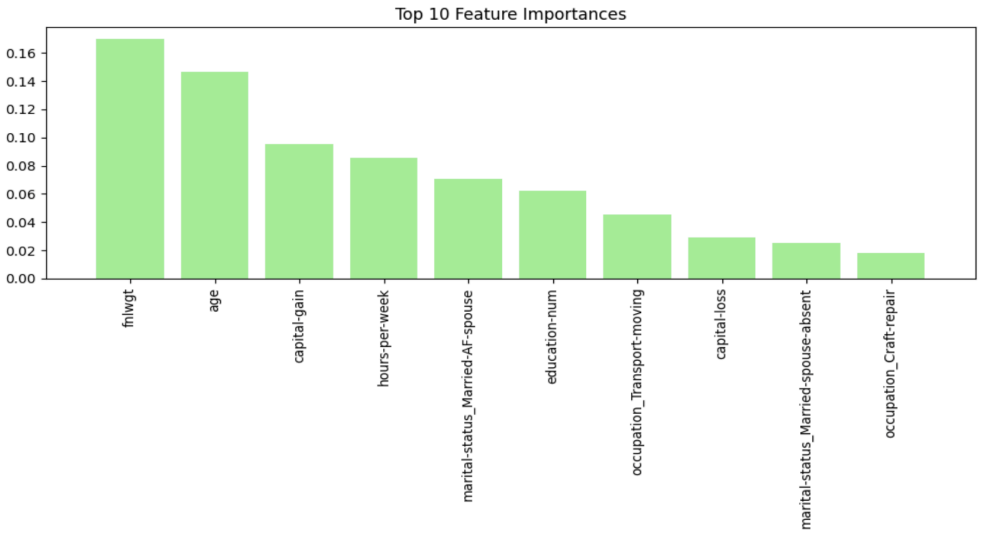

Figure 3.6.1

3.6.2 Hyperparameter Tuning

As there is a tendency of overfitting the model, we performed hyperparameter tuning and used RandomizedSearchCV to find the best hyperparameters. After 50 random samples and 5-fold cross-validation, we got the best hyperparameter: {'max_depth': 9, 'n_estimators': 173}. After hyperparameter tuning, the training set accuracy decreases slightly to 0.8577, which is closer to the test set accuracy and shows that the generalization ability of the model has been improved. Meanwhile, the test set accuracy is 0.8499, which is also slightly improved.

3.6.3 Gradient Boosting

We use Gradient Boosting as an attempt to improve the model performance of the method. We use the GradientBoostingClassifier for training loss and remaining time for model training. However, after generating and comparing, we find that the Gradient Boosting model performs in general (Figure 3.6.2), so we did not use that model in time for the subsequent analysis.

```
gbt = GradientBoostingClassifier(n_estimators=100, learning_rate=1.0, max_depth
```

```
      Iter       Train Loss   Remaining Time
         1           0.8218            6.53s
         2           0.7468            5.10s
         3           0.7083            4.71s
         4           0.6893            4.43s
         5           0.6866            4.37s
         6           0.6816            4.32s
         7           0.6722            4.30s
         8           0.6642            4.17s
         9           0.6603            4.04s
        10           0.6549            3.98s
        20 289885911192084901658443685597172092205496734195915141493837 2021
        30 289885911192084901658443685597172092205496734195915141493837 2021
        40 289885911192084901658443685597172092205496734195915141493837 2021
        50 289885911192084901658443685597172092205496734195915141493837 2021
        60 289885911192084901658443685597172092205496734195915141493837 2021
        70 289885911192084901658443685597172092205496734195915141493837 2021
        80 289885911192084901658443685597172092205496734195915141493837 2021
        90 289885911192084901658443685597172092205496734195915141493837 2021
       100 289885911192084901658443685597172092205496734195915141493837 2021
```

Figure 3.6.2

```
accuracy on training: 0.8653292750415053
accuracy on testing: 0.8573325954620918
```

	0
marital-status_Married-AF-spouse	0.377312
education-num	0.211578
capital-gain	0.148912
marital-status_Divorced	0.135546
capital-loss	0.044586
...	...
race_White	0.000000
education_HS-grad	0.000000
native-country_Columbia	0.000000
native-country_Cuba	0.000000
education_7th-8th	0.000000

Figure 3.6.3

3.6.4 Model Evaluation

We followed the results of hyperparameter tuning in part 3.6.2. We combine the categorical reports and the confusion matrix (Figure 3.6.4) to do the model evaluation. The overall accuracy of the model was 0.8499. 86% of the samples in category 0 were correct and 95% were correctly predicted. 78% of the samples in category 1 were correct and 53% were correctly predicted. Incorporating the precision and recall of macro avg and weighted avg, due to the difference in sample size category 0 has a higher score than category 1. This means

that the model performs better in identifying negative categories, and most of the samples predicted to be negative (TN and FN) are correct.

Eventually, we have rescaled the bars for the importance of the top 10 features in the random forest model (Figure 3.6.5). Among them, capital-gain and marital-status_Married-AF-spouse are the two most important features in the model, and their impact on the prediction results is the largest, with feature importance of about 0.16 and 0.15 respectively. The next important feature is occupation_Transport-moving, which is about 0.12. and education-num has the fourth highest feature importance of about 0.10.

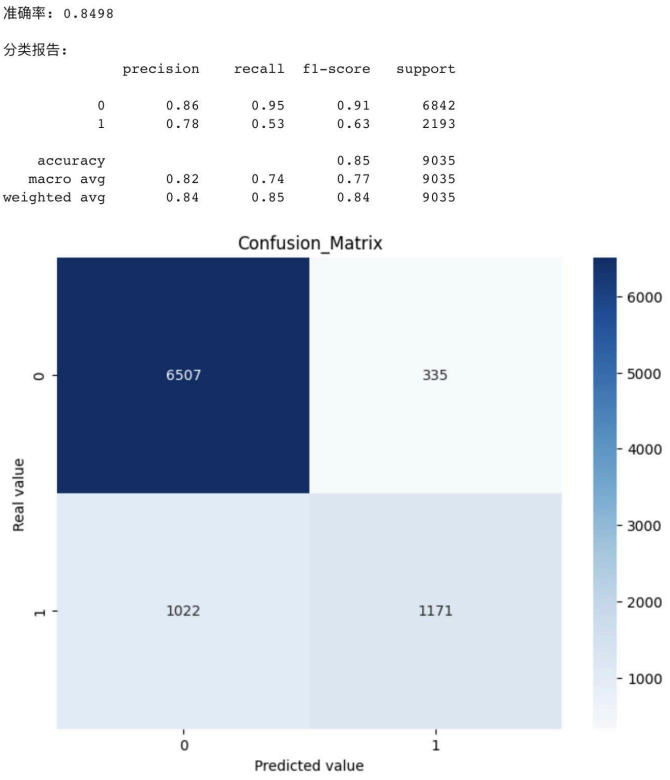

Figure 3.6.4

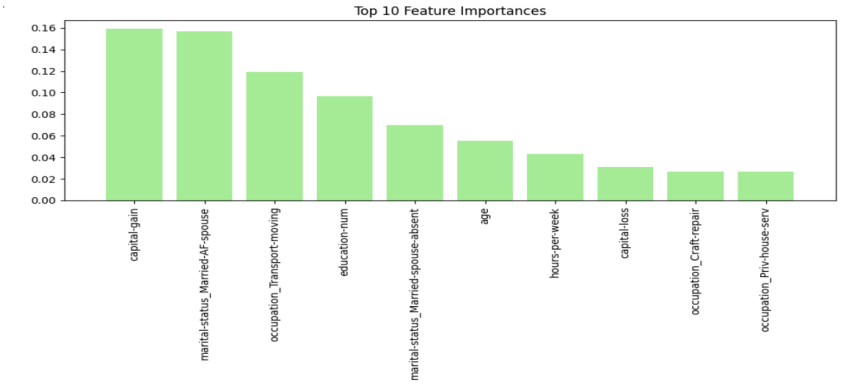

Figure 3.6.5

4. Conclusion

4.1 KNN:

The KNN model demonstrated strong predictive performance, achieving an accuracy of 84% on the testing dataset and 86% on the training dataset. The optimal K value of 13 was identified through grid search, and the model's consistency across cross-validation and testing phases indicates robustness. The confusion matrix provides valuable insights into the model's performance, highlighting its strengths and areas for potential improvement.

4.2 Logistic Regression

In the logistic regression model, the accuracy on the test and train set was 84%, indicating that the model has relatively good predictive ability. In the model, the higher the absolute value of the coefficients, the higher the influence. By visualizing the characteristic coefficients, we can intuitively identify that marital status, occupation, and state have the greatest direct positive and negative impact on an individual's income level.

4.3 LASSO

In the LASSO regression model, by incorporating the L1 regularization term and the non-zero coefficient feature, we improve our ability to predict the model, to identify the features that have a significant impact on the prediction results, and to prevent overfitting while improving the model's explainability. Overall, the three main features: marital status relationship, education's influence and occupation type and status have the greatest impact on the prediction results.

4.4 Decision Tree

In the decision tree, the accuracy of the train set and test set of the model reaches 0.8599 and 0.8548 respectively, which shows a good generalization and predicting ability. The complexity of the decision tree is controlled by adjusting the alpha value to find a balance point and prevent overfitting. However, it can be found that despite the overall accuracy of the model, the model performs better on Category 0 (higher precision and recall) than on Category 1, which suggests that the model is deficient in recognizing category. Combining this with the final bar chart, we can summarize that the features of marital status, years of education, and capital status are key in influencing the overall predicted income level.

4.5 Random Forest

Random Forest reached an accuracy of 0.8499 on the test set and 0.8577 on the train set when dealing with the problem of categorization, and the accuracy of the model on the test set was raised by hyperparameter tuning, indicating that the generalization ability of the model has been improved. But, in the final model, the number of samples of Category 0 is much more

than that of Category 1, which may affect the model's prediction performance for Category 1 to some extent, as the imbalance may affect the model's ability to identify the positive categories. Whereas, in our predicted results, it can be seen that the imbalance between category 0 and 1 does not have much effect on the result. Overall, in combination with the top 10 features at the end of the bar chart, capital, marital status, and occupational characteristics have the greatest impact on income levels.

4.6 Summary

In conclusion, comparing the five models KNN, logistic regression, LASSO, decision tree, and random forest, we found that all of those models generate a similar result with a high accuracy of prediction. Additionally, we combined the accuracy of the test set, recall rate, and precision for KNN, logistic regression, decision tree, and random forest for those four models (refer to the following table). The decision tree presents a higher accuracy on the test set, train set, precision with fewer false positives, and KNN model performs better with fewer false negatives. Meanwhile, we can summarize that the most influential feature on whether or not annual income exceeds $50k is marital status, which is followed by occupational difference, education status and capital status. This would prove that our prediction was successful, as most of the outputs of the models are alike. As a result, combining the similar results of each model we can deem that we can accurately predict whether an individual's income exceeds $50,000 per year based on demographic and socioeconomic attributes.

Table 4.6 Summary of four models

	KNN	Logistic	Decision Tree	Random Forest
Accuracy on test	84%	84%	85.4%	84.9%
Accuracy on train	86%	84%	85.9%	85.7%
Recall	0.63	0.58	0.58	0.53
Precision	0.64	0.70	0.76	0.44

5. Reflection and Recommendations

Based on the findings, it is concluded that our model can accurately predict whether a person's annual income exceeds $50,000 and its main influencing factors. Among these factors, marital status, occupation, education and capital are the key influences on the annual income. Those models help the stakeholders to have a higher ability to adjust salary distribution and identify the income levels of people.

In terms of practical significance, stakeholders (e.g., government, corporations, and other social groups) could analyze the impact of features to perform finer market segmentation and help individuals increase their total income. For married employees in terms of marital status, they have more economic synergy and incentives like family support. On the occupational side, salary variation and job stability between different industries are key factors influencing income. On the capital side, stakeholders can provide financial planning and investment education to help individuals better manage their finances and to subsidize individuals with lower occupational incomes to increase wealth effect. On the education aspect, the government and corporations can conduct no-cost upgrading education programs or other educational resources for individuals who have shorter years of education to improve their education and professional skills. Additionally, stakeholders need to set up regular evaluation and feedback mechanisms to monitor the effectiveness of these measures in the long term. Hence, the overall effect of subsidies, education, and motivation will be to increase the possibility that people may achieve higher incomes.

6. References

Becker, Barry, and Ronny Kohavi. “UCI Machine Learning Repository.” Archive.ics.uci.edu, 30 Apr. 1996, archive.ics.uci.edu/dataset/2/adult. Accessed Feb. 2025.