A Data-Driven Approach to Predict whether income exceeds $50K/yr based on census data

Curated By:
Pyimoe Than

San Francisco State University

Table of Contents

1. Executive Summary

This document outlines our approach to help government organizations overcome its challenge in effectively allocating its low income housing. Through the use of advanced machine learning techniques, we are confident in our ability to predict which individuals have an earning potential of less than 50 K. Our project involves identifying the most effective ML model for predicting responses and ensuring that our low income housing programs are personalized and targeted toward a person in need. To ensure the success of the project and provide reliable services to those in need, we aim for an accuracy score of 90 % or higher. This high threshold is essential to guarantee the effectiveness and reliability of our offerings.

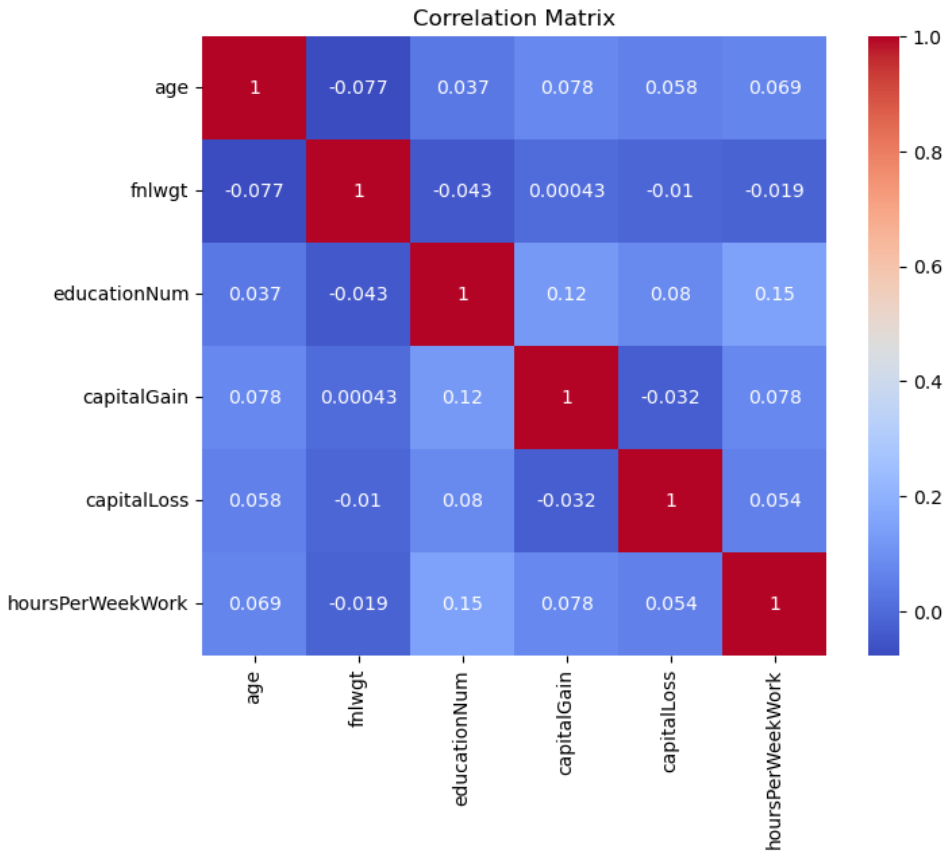
2. Introduction

The dataset, acquired from the University of California, Irvine Repository, from this project contains information about individuals income. It removed sensitive information. It was conducted by the U.S. Census. The ML techniques aim to enhance the government organizations to understand the predicting factors that potentially a classified “<=50K” or “>50K” response to a person's income.

3. Data Description

The data utilized in this project is sourced from the UCI Repository, specifically from the Adult Data Set. This dataset comprises 48,842 observations including 14 columns.

3.1 Correlation Matrix Graph Between Non-numeric Predictor Variables



Correlation coefficients between many pairs of predictor variables in the data appear to be less than 0.5. This suggests a weak linear relationship between the variables.

Variable overview:

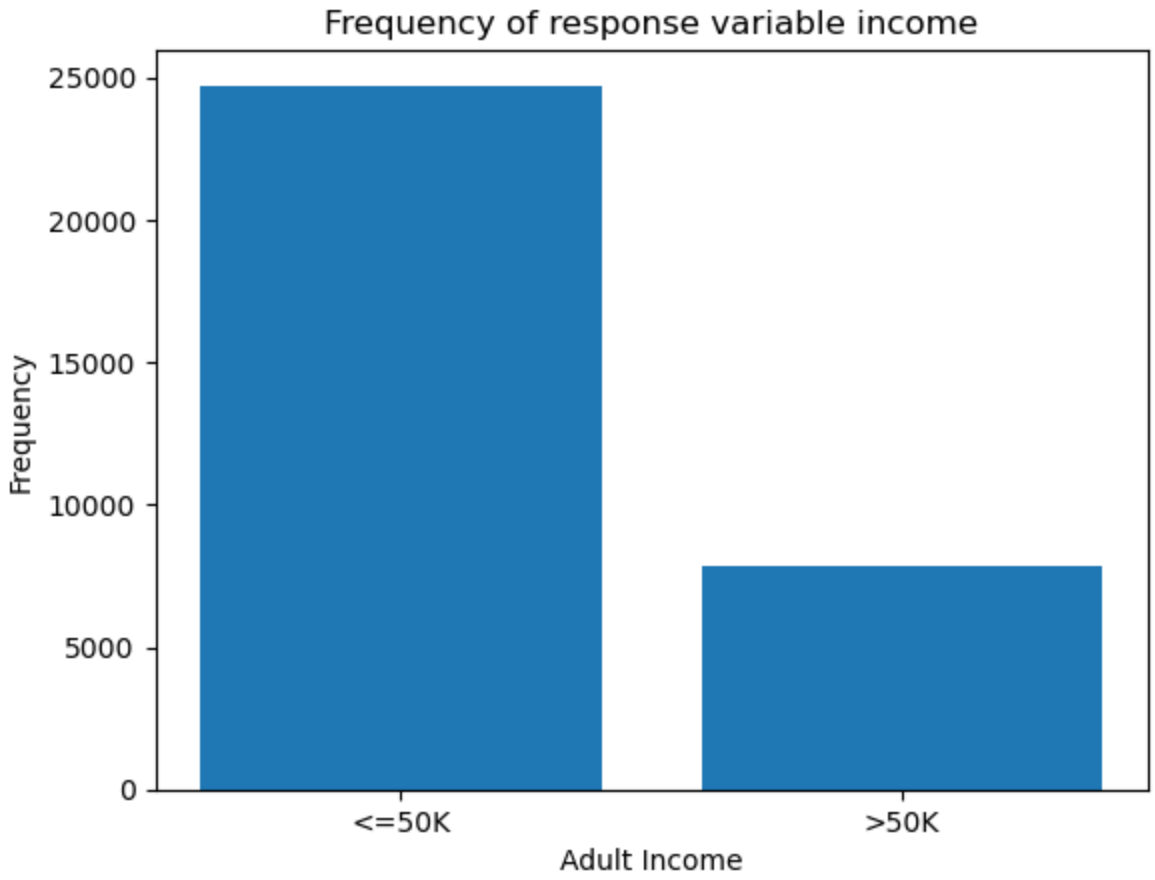
1. ***age***: (numeric)
2. ***workClass***: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
3. ***Fnlwgt***: (numeric)
4. ***Education***: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
5. ***marital-status***: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
6. ***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.
7. ***relationship***: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
8. ***race***: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
9. ***sex***: Female, Male.
10. ***capital-gain***: continuous.

11. ***capital-loss***: continuous.
12. ***hours-per-week***: continuous.
13. ***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, Trinidad Tobago, Peru, Hong, Holland-Netherlands.
14. Output variable (desired target)
    income: a person income which has two classes: <=50K and >50K.

3.2 Data Cleaning

During the data cleaning process, we conducted a series of integrity checks to ensure the dataset accuracy, completeness, and consistency. We were pleased to discover that there were no missing values or duplicated data in the dataset. However, “occupation” and “workClass” columns have ? value. We rename the “?” into the unknown. We also noticed an unbalanced response variable, with a higher count of “<=50K” responses than “>50K” responses. This class imbalance has the potential to introduce bias in our prediction model, which is something we must address.

Figure 1: Frequency of Responses




```
Percentage of adult who earn less than or equal to 50K: 75.92%
Percentage of adult who earn greater than 50K :24.08%
```

The figure above depicts the total number of adults who earn > 50K responses and total number of adults who earn <=50K in the dataset. From the graph above we can see there are far more adults who earn less than 50K than adults who earn more than 50K.

An oversampling method was applied to address this issue to balance the data. The result of this technique exhibits a balanced distribution in our dataset and is now ready for further analysis using machine learning techniques. Below, in Figure 2, you can visualize the frequency of responses after including the oversampling method.

Figure 2: Frequency of Responses (After Oversampling Method)

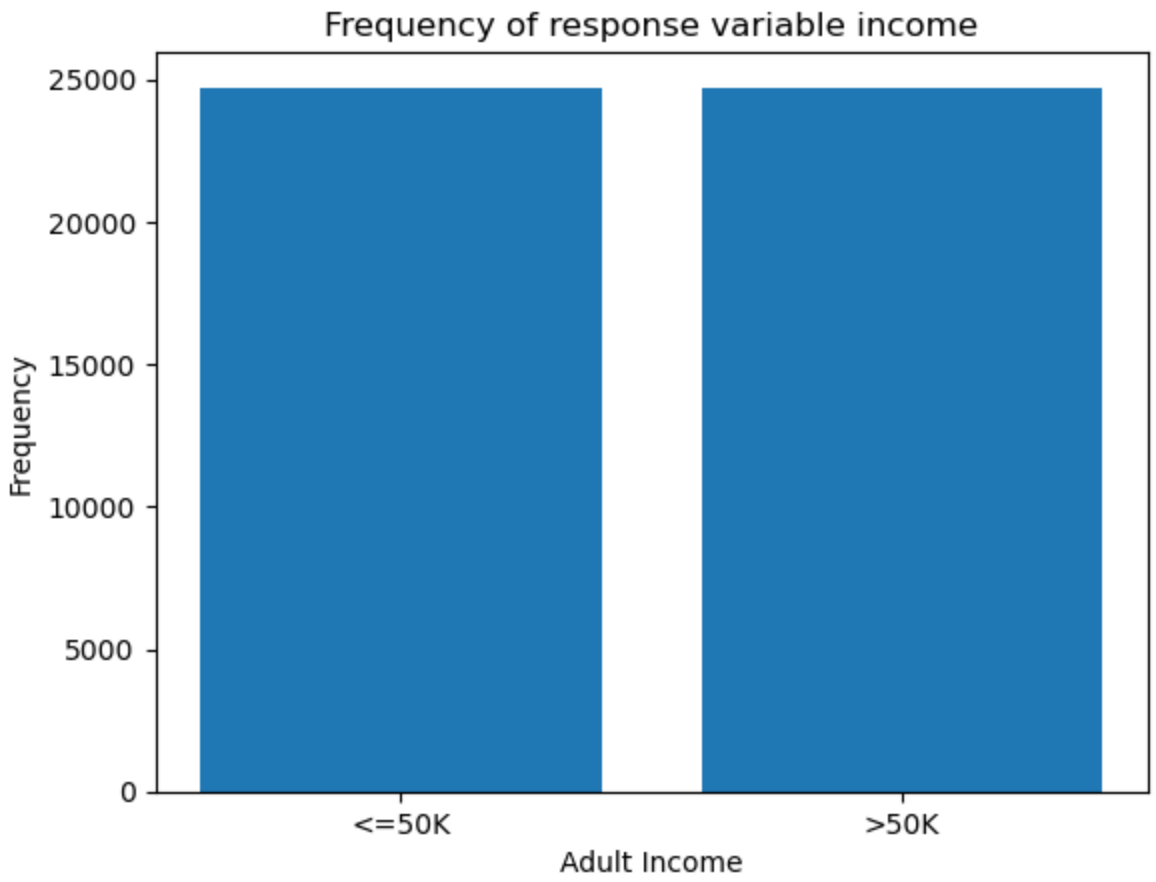


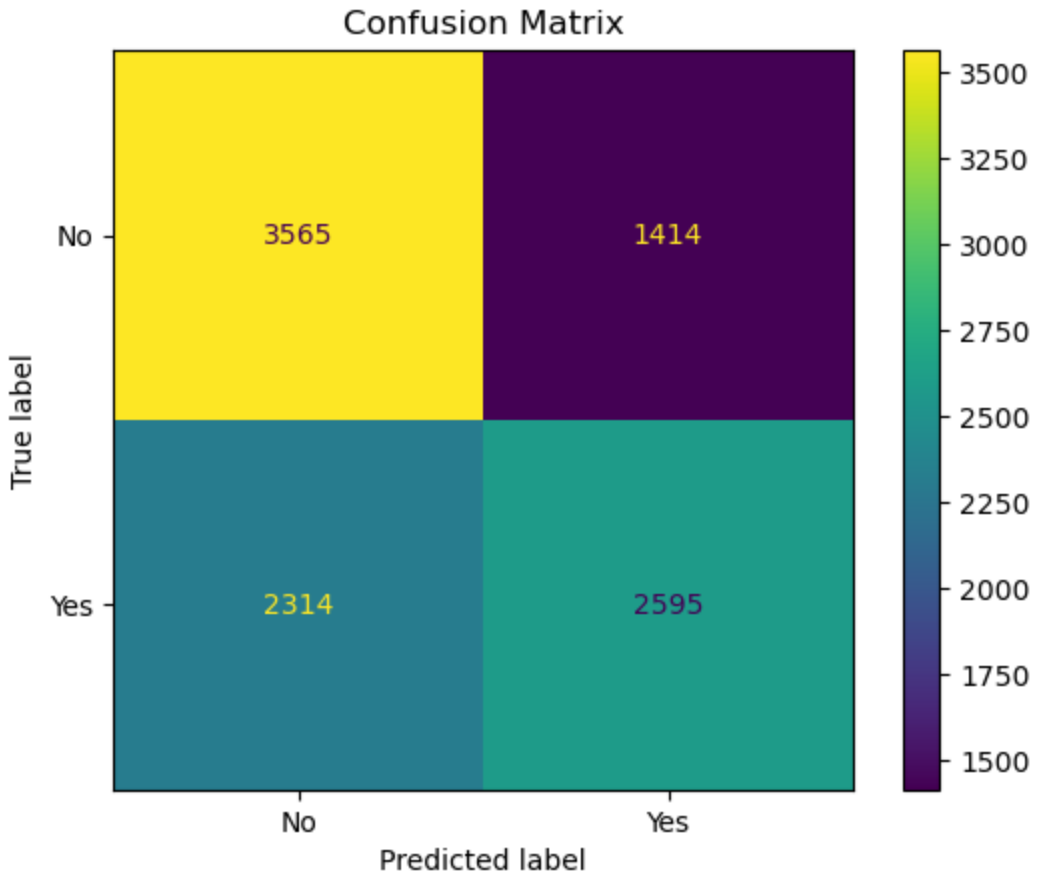
Figure 2 above displays a balanced number of adults who earn less than or equal to 50K and adults who earn more than 50K, after incorporating the oversampling method.

4. **Model Selection**

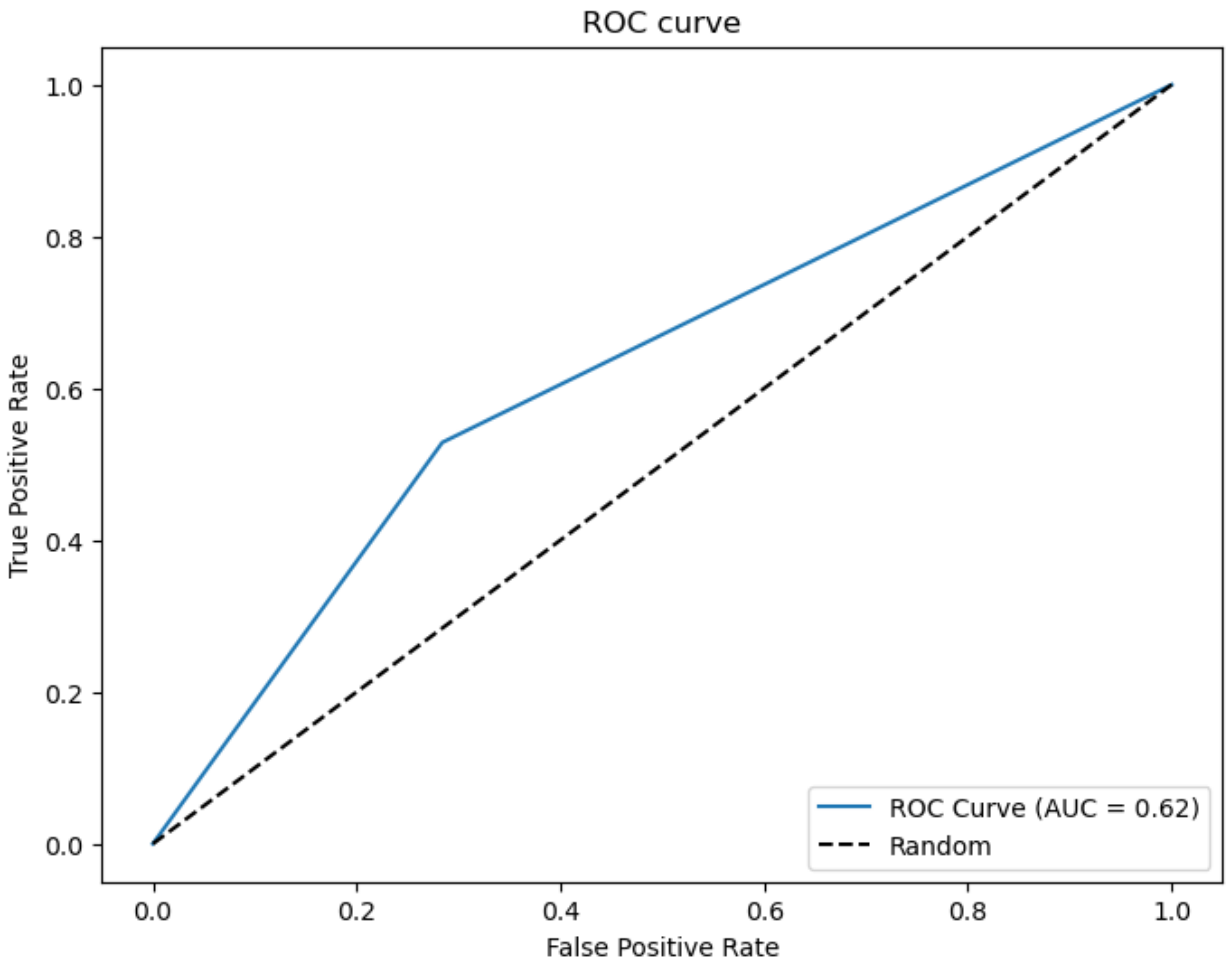
4.1 Logistic Regression

Following the data cleaning process, we divided the data into 80% training and 20% testing data prior to fitting. The Logistic Regression model was trained using both k-fold cross-validation on the training data. This comprehensive validation approach provides robust estimates of the model's performance. We generated a Confusion Matrix which shows the accuracy, AUC, sensitivity, and specificity.

Figure 3: Logistic Regression Confusion Matrix



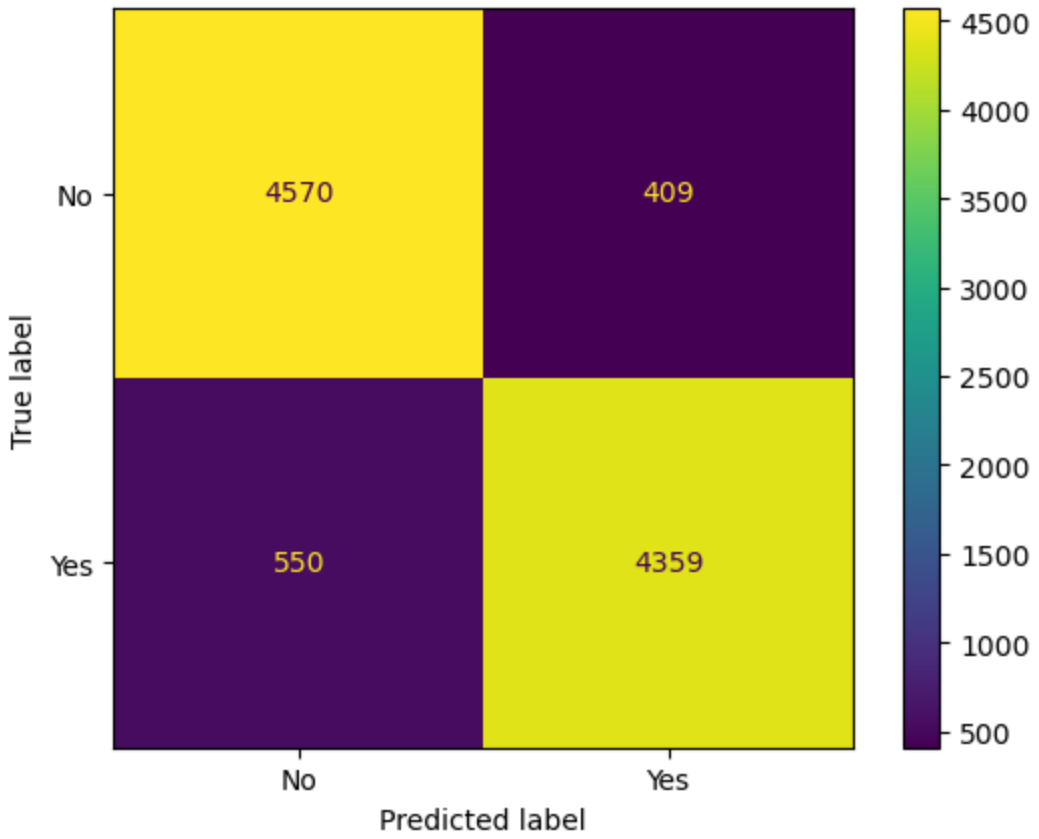
Accuracy	Sensitivity	Specificity	AUC	# False Positive	# False Negative	misClassification Rate
0.62	0.53	0.72	0.62	1414	2314	0.377



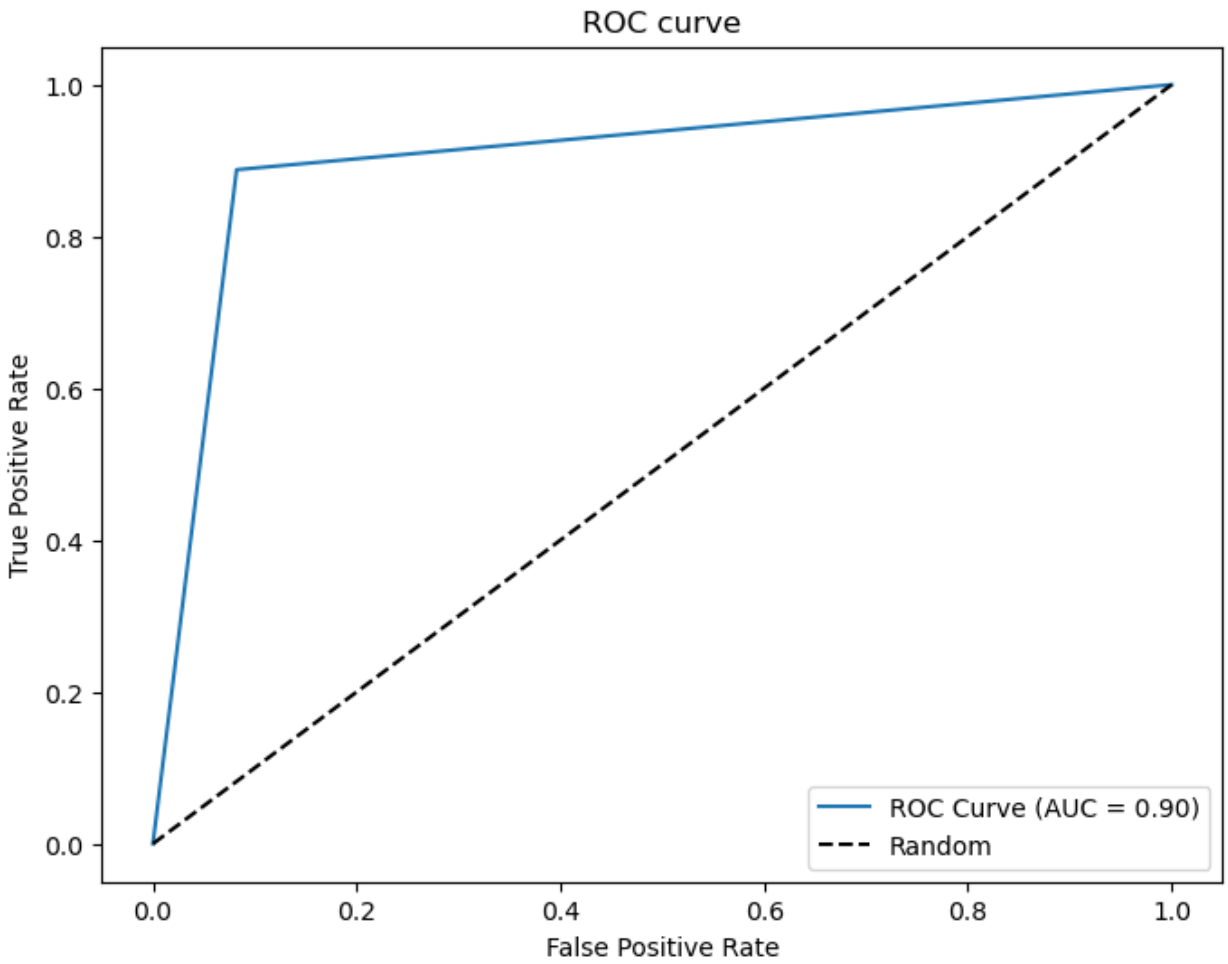
- Accuracy score of 0.62 means the model correctly predicted the outcome for **62.0 %** of adults' incomes.
- AUC score of 0.62 means the model performs better than the random guessing since it is greater than 0.5
- Sensitivity means the percentage of positive outcomes the model is able to detect
- Specificity means the percentage of negative outcomes the model is able to detect.
- misClassificationRate of 0.377 means the model incorrectly predicted the outcome for **37.7%** of the adults' incomes.

4.2 Random Forest

After fitting the logistic regression, I then fit the model using the random Forest machine learning model on the training data. We generated a Confusion Matrix which shows the accuracy, AUC, sensitivity, and specificity. We further create a ROC curve. It gives us the AUC score



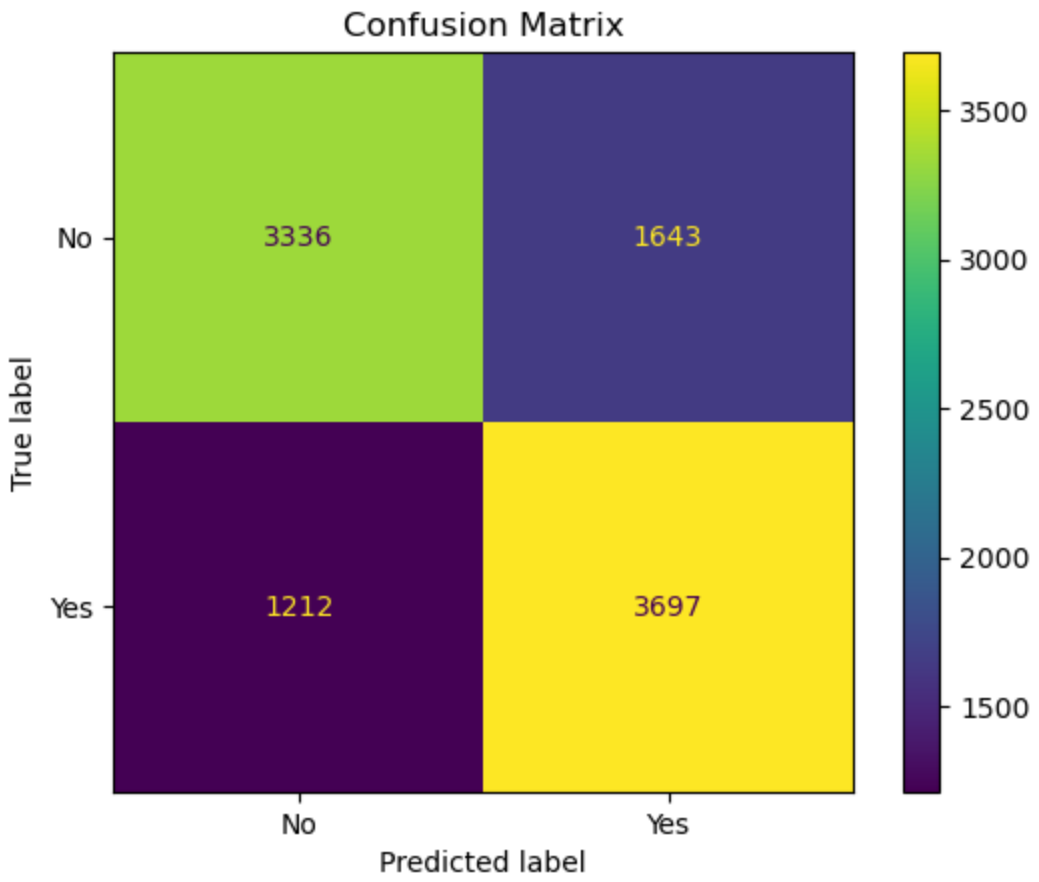
Accuracy	Sensitivity	Specificity	AUC	# False Positive	# False Negative	misClassification Rate
0.90	0.89	0.91	0.90	409	550	0.097



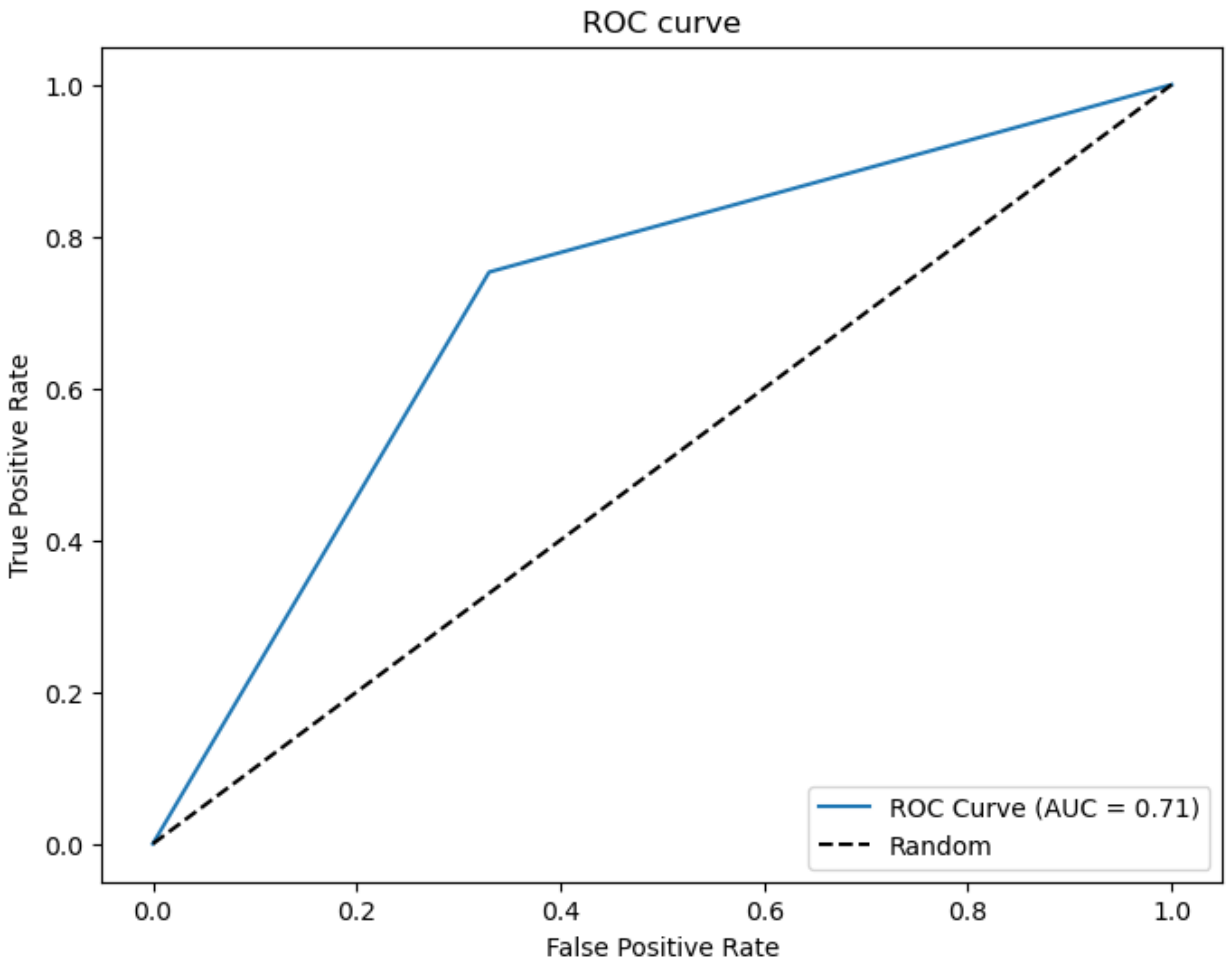
- Accuracy score of 0.90 means the model correctly predicted the outcome for **90.0 %** of adults' incomes.
- AUC score of 0.90 means the model performs better than the random guessing since it is greater than 0.5
- Sensitivity means the percentage of positive outcomes the model is able to detect
- Specificity means the percentage of negative outcomes the model is able to detect.
- misClassificationRate of 0.097 means the model incorrectly predicted the outcome for **9.7%** of the adults' incomes.

4.3 K-Nearest Neighbor

Since Random Forest is memory intensive, I will try to fit the model using a fast, efficient, and simple method which is KNN.



Accuracy	Sensitivity	Specificity	AUC	# False Positive	# False Negative	misClassification Rate
0.71	0.75	0.67	0.71	1643	1212	0.288



- Accuracy score of 0.71 means the model correctly predicted the outcome for **71.0 %** of adults' incomes.
- AUC score of 0.71 means the model performs better than the random guessing since it is greater than 0.5
- Sensitivity means the percentage of positive outcomes the model is able to detect
- Specificity means the percentage of negative outcomes the model is able to detect.
- misClassificationRate of 0.288 means the model incorrectly predicted the outcome for **28.8%** of the adults' incomes.

4.4 Proposed Solution

I highly recommend utilizing the Random Forest algorithm as the chosen machine learning model, as it has proven to significantly enhance accuracy in various scenarios. However, considering the company's focus on cost efficiency, I suggest implementing the K-nearest neighbor(KNN) algorithm. This method offers a favorable balance between accuracy and cost -effectiveness. With an accuracy rate of 71% Knn outperforms the logistic regression model and surpasses the results achieved by random guessing. By adopting KNN, the company can achieve reliable results while optimizing resource allocation.

Appendix

5.1 Data Source - UCI Repository: http://archive.ics.uci.edu/dataset/2/adult.

5.2 Code - GitHub: Project Code