

Census Income

capstone project

Rishabh Goriya

Table of Contents

1. Problem Statement

2. Project Objective

3. Data Description

4. Data Pre-processing Steps and Inspiration

5. Choosing the Algorithm for the Project

6. Motivation and Reasons for Choosing the Algorithm

7. Assumptions

8. Model Evaluation and Techniques

9. Inferences from the project

10. Future Possibilities of the Project

11. References

# Problem Statement

The main aim of this project is to develop a predictive model which is capable of accurately classifying individuals into two class –

* Individuals earning more than $50,000 annually
* Individuals earning less than $50,000 annually

The task is to develop the model with the help of various demographic and socio-economic features.

The developed predictive model should be capable of assisting in identifying factors that can contribute to get the individual in the higher income group and to understand the underlying patterns within the data.

# Project Objective

### Data Exploration

* Perform a thorough exploration of the “Census Income” dataset to understand the distribution of features and the nature of the data.
* Identifying the types of features available in the dataset.
* Calculating the basic statistics of the data.

### Data Manipulation

* Performing task to handle the missing values, outliers and perform the important preprocessing steps.
* Encode categorical variables to make it easier for the algorithm to understand.
* Transform the data to ensure the data is suitable for training and machine learning models.

### Model Development

* Experiment with various machine learning algorithms such as logistic regression, decision tree, random forests and others for classification of “Income”.
* All models will be built on the “Census Income” dataset.
* The dataset will be partitioned into two subsets, X and Y, each further divided into training and testing sets.

### Model Tuning and Model Evaluation

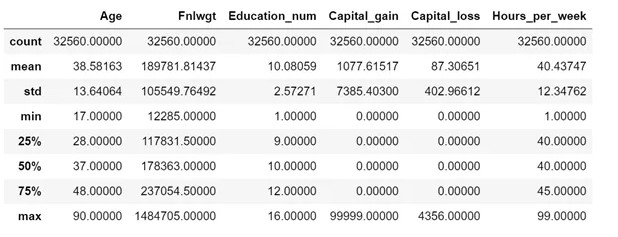
* Fine-tune the hyperparameters to optimize the performance of the chosen model.
* Evaluate the model performance using “confusion matrix” and “accuracy score” to choose the better model for the objective.

# Data Description

* “Census Income” dataset is also known as “Adult” dataset. It was extracted by Barry Becker from the 1994 census dataset.
* Census Income is a multivariate dataset with 48,842 instances and 14 features.
* The dataset is focused on the subject of social science.
* It has features of categorical and integer type.
* Since the data predicts 2 values (>50K or <=50K), this clearly is a classification problem.

# Insights from the data

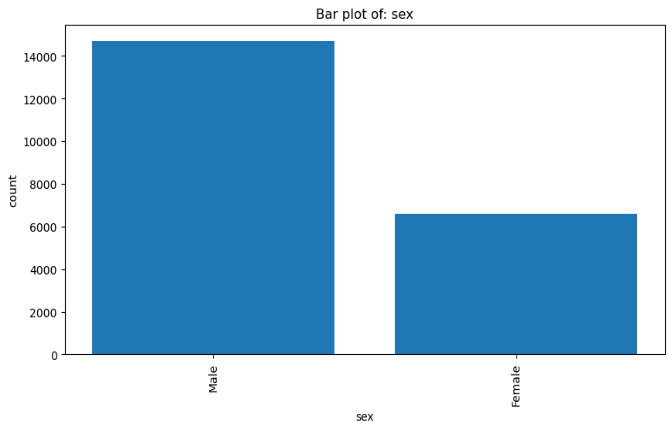
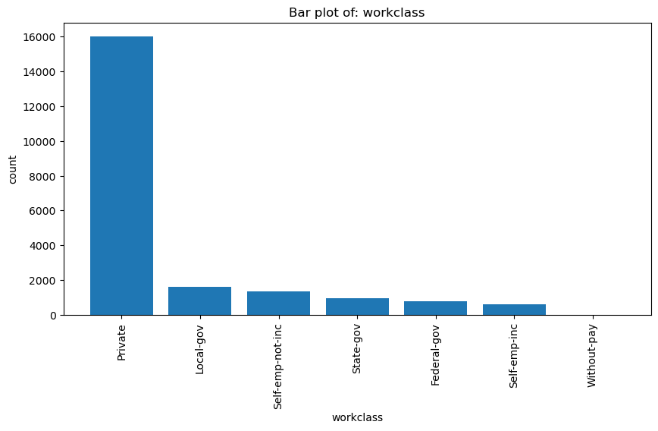
Insights gathered from the Census Income dataset are:



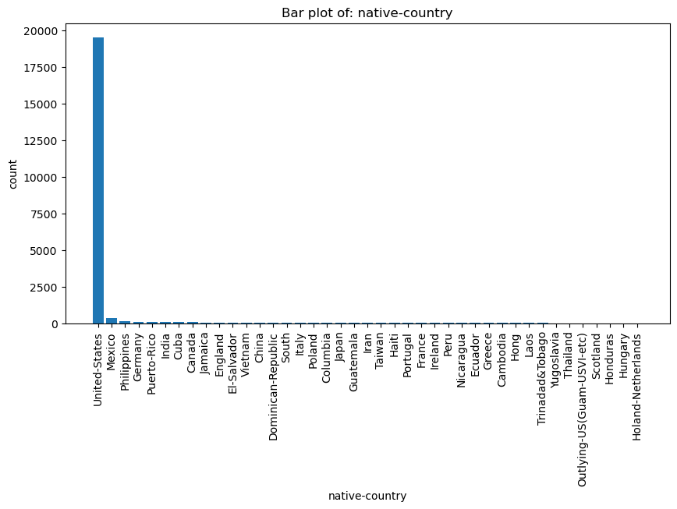
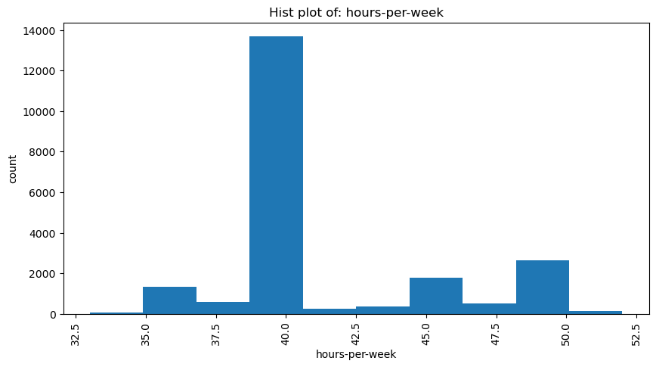
Insights gathered from- “.describe()” function from integer columns

* Age - The age of an individual ranges from 17 to 90.
* The education number has a range of 1 to 16
* There are outliers expected in Capital gain column as the values till 75% are 0. Same is the case with capital loss as well.
* Hours per week range between 1–99.

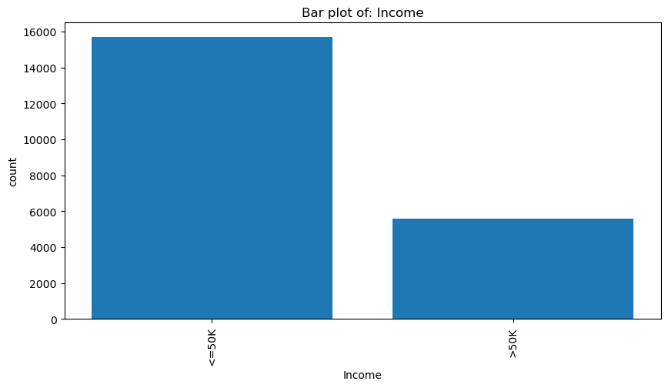
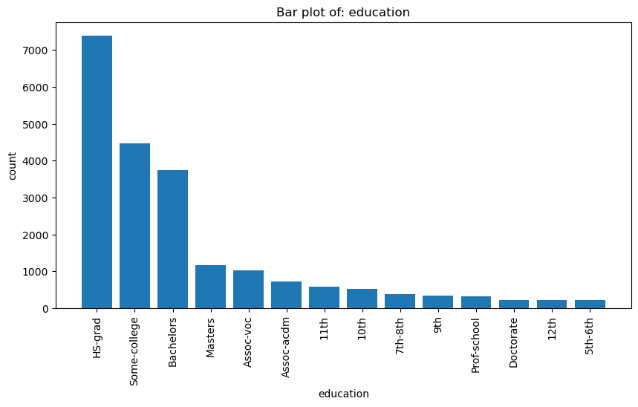
Some insights gathered from Bar plot and Histograms -



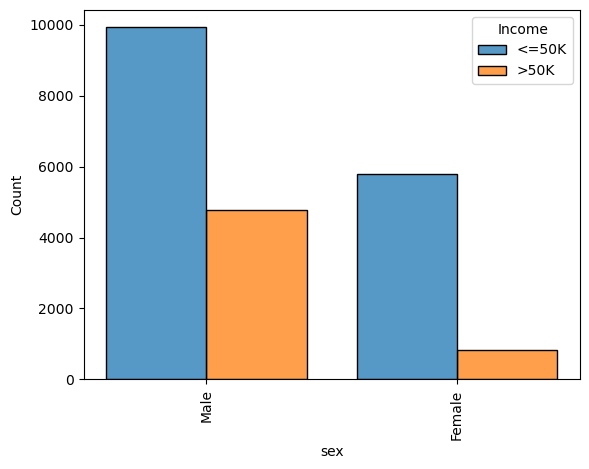
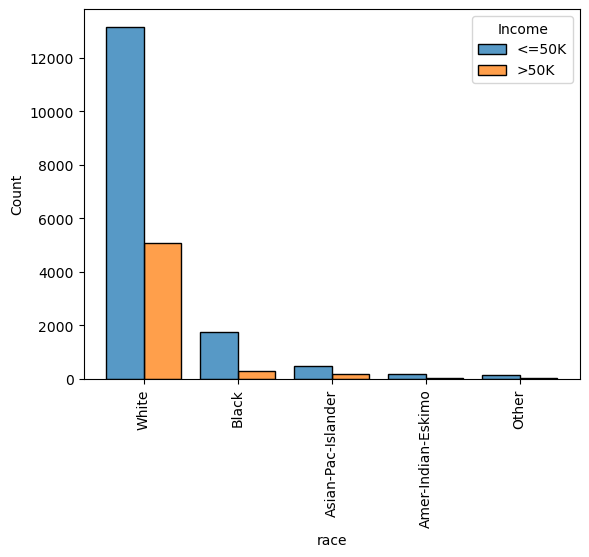
* The data have majority of people working in private work class.
* Census Income data have twice the number of males in comparison to females.



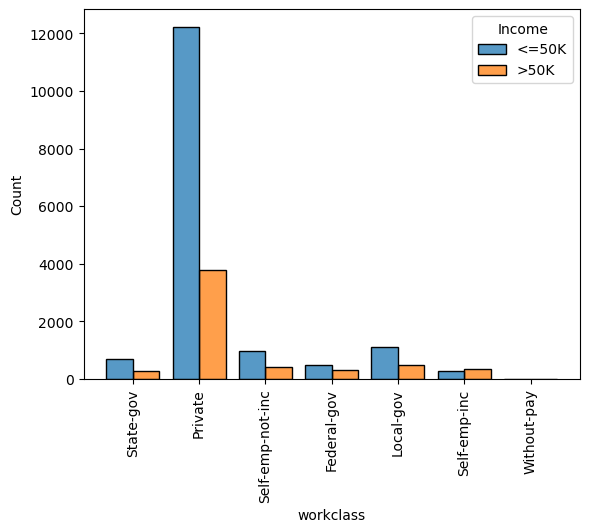
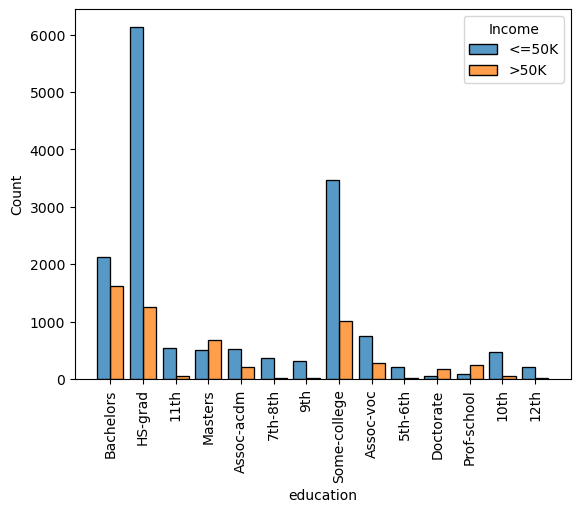
* According to the data the maximum number of people work 40 hours per week.
* The data consist of information of people from 41 countries but maximum number of data is of United States.



* People earning less than $50k are almost thrice in number in comparison to people earning more than $50k.
* Highest level of education for most of the people is “HS-grad”
* According to the data the maximum amount of capital gain has been done by the people working in “Self-emp-inc (Self-employed incorporated)” work class. The top 3 capital gain mean of work class are as follows-
* Self-emp-inc 3994.00
* Self-emp-not-inc 2215.05
* Without-pay 1104.00
* According to the dataset only 5.86% of people who are just High-School graduate are earning more than $50k annually.
* According to the dataset 56.43% of the people of United States are earning less than $50k annually.



* The data says more people of white race are earning more than $50k annually.
* Males in different countries have a higher chance of earning more than $50k annually than females



* Individuals with education as “Masters”, “Doctorate”, “Prof-school” have a higher chance of earning more than $50k annually.
* Individuals working in “Self-emp-inc” have higher ratio of earning more than $50k annually.

# Data Preprocessing Steps and Inspiration

Data preprocessing is a crucial step in preparing a dataset for machine learning models.

1. Handling missing values:
   * Inspiration: Real-world datasets often contain missing values, which can significantly affect the performance of machine learning models. Handling missing values is essential to ensure that your model is trained on complete and reliable data.
   * Actions taken:
     1. The dataset contained some initial space in the dataset to handle that I used “skipinitialspace= True”
     2. Dataset had symbol “?” in some places I replaced that with “nan”
     3. There were some null and duplicated values and they were not a lot so I decided to drop them.
     4. I observed the target column in the dataset is empty so I renamed it to “Income”
2. Removing Outliers:
   * Inspiration: Outliers can have a significant impact on the performance of certain models. Removing extreme values can improve the model's robustness and generalization.
   * Actions Taken: With the help of boxplot from matplotlib I plotted box plot of integer columns to check the presence of outliers in the dataset. The outliers were removed with the help Inter Quartile Method. I had observed box plot diagram of two columns (‘capital-gain’, ‘capital-loss’) are not visible suggesting most of the data in these columns are outliers so I haven’t used the IQR method on these columns.
3. Encoding Categorical Variables:
   * Inspiration: Machine learning models typically require numerical input, so categorical variables need to be encoded. This is necessary to represent the categorical information in a format that the algorithm can understand.
   * Actions Taken: I encoded the categorical variables using “label encoding” technique from “sklearn.preprocessing”.
4. Data Splitting:
   * Inspiration: Splitting the dataset into training and testing sets is crucial to evaluate the model's performance on unseen data.
   * Actions Taken: I split the dataset into training and testing sets using ratio of 80% training and 20% testing set.
5. Feature Scaling:
   * Inspiration: Features with different scales can impact the performance of machine learning algorithms. Scaling features helps ensure that each feature contributes equally to the model.
   * Action: I applied standardization using “StandardScaler” from “sklearn.preprocessing” to normalize the range of feature values.

# Choosing the Algorithm for the Project

I had applied multiple algorithms on the Census Income dataset to predict the Income class of the individuals. After that based on the performance of the models, I have chosen the best model for predicting the Income class.

1. Logistic Regression:

Reason:

* + Logistic Regression is a simple yet effective algorithm for binary classification problems, and it's well-suited for problems where the relationship between features and the target variable is approximately linear.
  + It provides probabilities of class membership, making it interpretable and allowing for easy understanding of the impact of individual features on the predicted outcome.
  + Logistic Regression is computationally efficient and less prone to overfitting, making it a good baseline model.

1. Decision Tree:

Reason:

* + Decision Trees are versatile and can handle both binary and multiclass classification problems.
  + They are capable of capturing complex relationships in the data, including non-linear patterns.
  + Decision Trees are interpretable, as they provide a clear decision-making process based on feature splits.

1. Random Forest:

Reason:

* + Random Forest is an ensemble method that builds multiple decision trees and combines their predictions, providing improved generalization and robustness compared to a single Decision Tree.
  + It handles non-linearity and complex relationships well, making it suitable for datasets with diverse features.
  + Random Forest is less prone to overfitting than a single Decision Tree, as it aggregates predictions from multiple trees, reducing the impact of noise in the data.

# Assumptions

1. The following assumptions were made while building Logistic Regression model.
   * The response variable is binary.
   * The observations in the dataset are independent of each other.
   * There is no severe multicollinearity among the explanatory variables.
   * There are no extreme outliers or influential observations in the dataset.
   * Linear relationship between the dependent and independent variables are not required.
   * The error terms (residuals) do not need to be normally distributed; Homoscedasticity is not required.
   * The sample size is sufficiently large.
2. Decision Tree:
   * Decision trees does not follow any assumptions.
3. Following assumptions were made while building Random Forest model:
   * The data is independent and identically distributed (i.i.d.). This means that each data point is sampled independently from the same distribution.
   * The target variable is categorical for classification tasks and continuous for regression tasks.
   * Random Forests do not make any assumptions about the distribution of the data or the relationship between the features and the target variable.
   * Random Forests assume that the sampling is representative.

# Model Evaluation and Technique

The following techniques were involved in the evaluation of the Logistic Regression, Decision Tree and Random Forest model.

1. Accuracy score:

The accuracy score is a metric used to evaluate the overall performance of a classification model. It is calculated as the ratio of correctly predicted instances to the total instances.

1. Confusion matrix:

A confusion matrix is a table used to evaluate the performance of a classification model. It summarizes the results of a classification task, showing the counts of true positive, true negative, false positive, and false negative predictions.

# Inferences from the Project

With the help of training and test dataset performance of the models are calculated. Which are as follows:

|  |  |
| --- | --- |
| Logistic Regression: | 81.66% |
| Decision Tree: | 83.11% |
| Random Forest: | 84.61% |

I had observed Random Forest performed best among the other models, in terms of accuracy.

With the help of Confusion Matrix, I had confirmed the accuracy.

# Future Possibilities

* Although the data is 30 years old still it’s capable of giving an outline on what metrics are capable of affecting income of an individual.
* Newer data can be incorporated into this dataset to assess whether the parameters continue to exert the same influence on income as they did 30 years ago.
* New columns/parameters can be added to this data as per the need of this era.
* This dataset primarily comprises data from individuals in the United States of America. To enhance its global significance, data from around the world can be added in equal measure.

# References

1. <https://archive.ics.uci.edu/dataset/20/census+income>
2. <https://medium.com/@sagar.rawale3/understanding-decision-tree-algorithm-drawbacks-and-advantages-4486efa6b8c3>
3. https://www.statology.org/assumptions-of-logistic-regression/
4. <https://www.analyticsvidhya.com/blog/2023/01/step-by-step-working-of-decision-treealgorithm/#:~:text=In%20those%20algorithms%2C%20the%20major%20disadvantage%20is%20that,any%20assumption.%20And%20it%20also%20handles%20non-linear%20data>.
5. <https://medium.com/@cristianleo120/building-random-forest-from-scratch-in-python16d004982788#:~:text=Random%20Forests%20make%20the%20following%20assumptions%3A%201%20The,for%20classification%20tasks%20and%20continuous%20for%20regression%20tasks>.
6. https://towardsmachinelearning.org/random-forest/