Lab Oriented Project

On

US Census income

Abhishek, Anuj, Rishabh Verma

2011981005,2011981014,2011981099

Supervised By

Mr. Shivam Singh

Department of Computer Science and Engineering,

Chitkara University, Himachal Pradesh

Problem Statement

CHITKARA UNIVERSITY

The aim of this project is to predict if an individual earns over $50,000 annually or not. With features like age, education, occupation, and more, the goal is to create a user-friendly model if an individual falls into the ">50K" or "<=50K" income category with high accuracy. Challenges include handling diverse data types, dealing with imbalanced income distributions, and choosing the right features for effective predictions.

About Dataset

CHITKARA UNIVERSITY

The US Census Income dataset, also known as the "Census Income" or "Adult Income" dataset, is widely used in machine learning and data analysis. It comprises diverse features describing individuals, with the primary task being the prediction of whether an individual's income surpasses or falls below $50,000 annually. Key features include age, workclass, education, marital status, occupation, and more. The dataset poses challenges like handling categorical and numerical data, addressing potential class imbalances, and selecting relevant features. The ultimate aim is to build a simple yet effective predictive model for binary classification, distinguishing individuals with incomes above or below the $50,000 threshold.

Exploratory Data Analysis

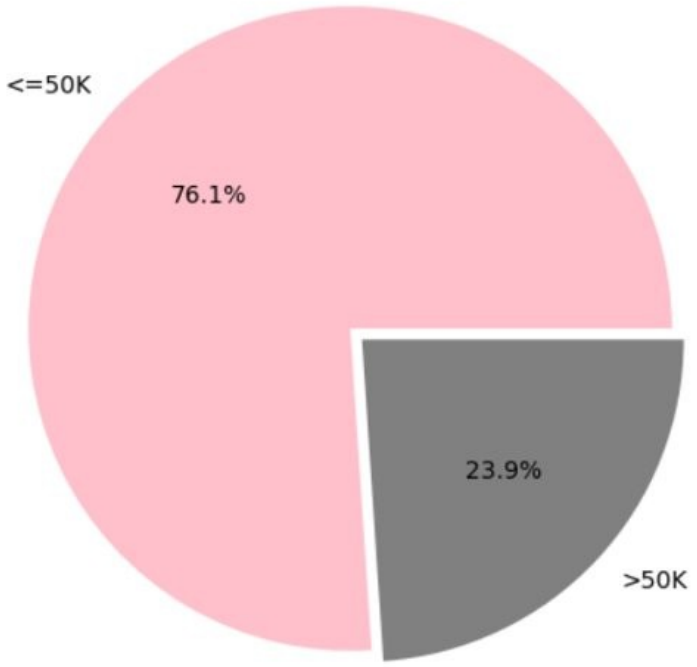
CHITKARA UNIVERSITY

In exploring the US Census Income dataset, we start by understanding basic statistics, such as age and income distribution. Dive into categorical features like workclass and education, observing demographic patterns. Analyze numerical features like capital gain and loss to uncover financial dynamics. Check for class imbalances in income categories and address them. Investigate correlations between features to grasp relationships. Visualize insights through histograms and scatter plots, aiming to enhance predictive model understanding. This concise EDA paves the way for effective feature selection and model building in the quest to predict income levels.

Analysis of Target Variable

CHITKARA UNIVERSITY

- From this pie chart, we can observe that around 76.1% of the individuals are earning <=50k annually and 23.9% of the individuals are earning >50K annually

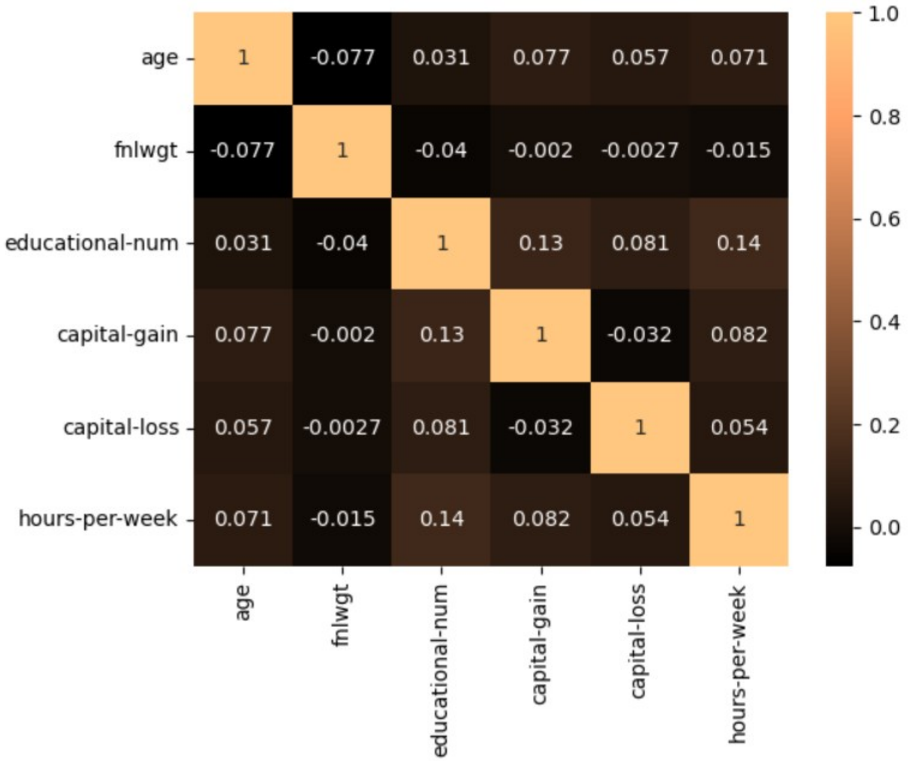


Analysis of Numerical Data

CHITKARA UNIVERSITY

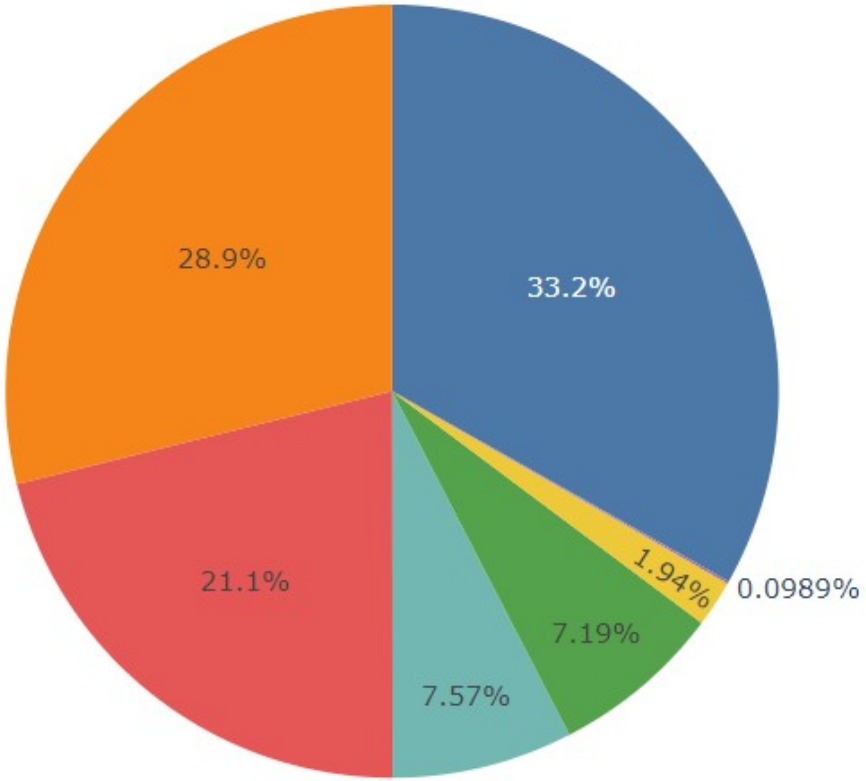
- From the above plot, we are getting some interesting insights about the dataset and those insights are following.

1. hours-per-week & educational-num is having strong correlation between them and correlation value is 0.14
2. Educational-num & capital gain is having average relation between them since the correlation value is 0.13
3. And the final observation from this Heatmap is that hours-per-week and fnlwgt are having negative relation between them.



Analysis of Numerical Data

From this pie chart,
we can observe that around
majority(33.2%)
of the individuals have
taken higher education
than high school .



Analysis of categorical Data

CHITKARA UNIVERSITY

- In categorical data, we can observe that each data contains so many distinct categories; there are only few variables where we are having just 2, 3, 4 or 5 categories. Although we have analyzed all these variables with the count plot and got some interesting insights. Some of those insights are as following:

1. The majority of individuals have taken higher education than high school .
2. We can observe that number of males are more than females who have annual income <=50k and >50k as well.
3. We can observe that ratio of individuals who earn <=50k and >50k is approximately same who has occupation of Exec-managerial.

Feature Engineering

CHITKARA UNIVERSITY

- Feature engineering is like the creative process of crafting valuable aspects from the data we already have, aligning them with what we want our machine learning model to learn. It's the art of reshaping the data to make it more in tune with what we're trying to predict. In our dataset, we've applied various techniques to refine and mold the features, ensuring they provide a clearer connection to the target we aim to predict.

1. Handling Categorical Data: We have handled the categorical data using Label Encoding technique.
2. Handling Outliers: The dataset contains an outlier and to handle that we have used IQR (Interquartile Range).
3. Feature Selection: For selecting the best feature in a dataset, we have used Heat Map and drop those features which were making negative impact on target variable.

Balancing the Data

CHITKARA UNIVERSITY

- In dealing with imbalanced data, the SMOTE (Synthetic Minority Oversampling Technique) method is utilized. Diverging from conventional approaches, SMOTE generates synthetic instances for the minority class through a K-Nearest Neighbor (KNN) mechanism, taking into account neighboring data points. This technique introduces novel and varied observations, bolstering the model's capacity to comprehend and forecast instances from the minority class. The KNN-based SMOTE method is preferred for its optimized and stable characteristics, steering clear of mere duplication of existing data.

Proposed Machine Learning Algorithms

CHITKARA UNIVERSITY

- Logistic Regression Classifier : Logistic Regression is a statistical method used for binary classification, predicting the probability of an observation belonging to a specific class. Despite its name, it's employed for classification rather than regression.

- Decision Tree classifier : A decision tree is a predictive modeling algorithm that recursively partitions data into subsets based on the values of input features, creating a tree-like structure of decisions to ultimately make predictions about the target variable.

Proposed Machine Learning Algorithms

CHITKARA UNIVERSITY

- Random Forest : Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the average prediction for regression or the mode for classification, providing a robust and accurate predictive model.

- XG Boost Classifier: XG Boost is again a very popular and effective ensemble learning technique but unlike Random Forest, XG Boost is a boosting technique. Boosting builds models from individual weak learners and unlike bagging it's a sequential learning process.

Proposed Machine Learning Algorithms

CHITKARA UNIVERSITY

- Adaptive Boosting: Adaptive Boosting is an ensemble learning method in machine learning that combines the strengths of multiple weak learners to create a robust and accurate predictive model. The algorithm operates iteratively, assigning varying weights to instances in the dataset based on their classification accuracy.

Hyperparameter Optimization

CHITKARA UNIVERSITY

- We have used Random Search Cross-Validation. It is a technique used in machine learning for hyperparameter tuning. Instead of searching through all possible combinations of hyperparameter values (as in grid search), it randomly samples a fixed number of hyperparameter combinations from specified distributions. This approach can be more efficient in terms of time and computational resources.

Performance Metrics

CHITKARA UNIVERSITY

1. **Accuracy Score:** It is calculated as total number of correctly classified points divided by total number of points is test set.
2. **Precision Score:** Precision is calculated as the ratio of correctly classified positive points (True Positives) to the total number of correctly classified positive points (True Positives + False Positives).
3. **Recall Score:** Recall tells us what proportion of the positive class got correctly classified. It is calculated as the total number of true positive points divided by true positive and false negative.
4. **F1-score:** In Precision, we are focusing more on false positive values where as in case of recall we focused on false negative value. Now, if we want to minimize false positive as well as false negative both then we can use f1-score. It is calculated as harmonic mean of precision and recall.

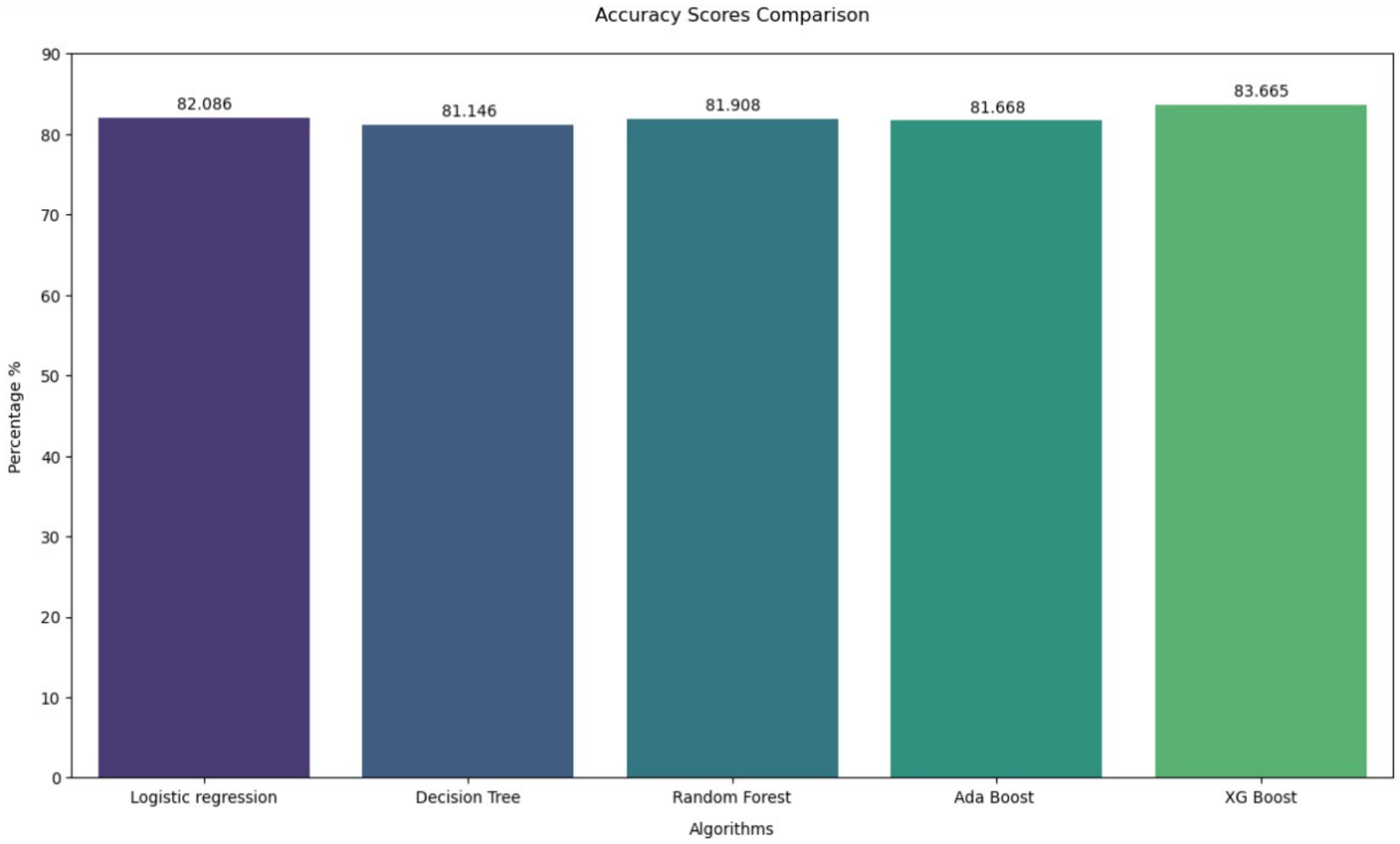
Confusion Matrix

CHITKARA UNIVERSITY

- Confusion Matrix can give us the visual representation of total sum of correctly classified points and sum of misclassified points. In this heatmap, we can observe that 12,190 points are classified correctly where 2380 points are mis-classified.

Performance comparison

CHITKARA UNIVERSITY



CHITKARA UNIVERSITY

Thank You