Article on Census Income (Evaluation Project – Flip Robo)

**Project Overview: Census Income Prediction**

The Census Income Prediction Project aims to predict whether a person earns more than $50,000 a year based on information about them, like their age, education, job and work hours. The data comes from the 1994 U.S. Census which was extracted by Ronny Kohavi and Barry Becker Which includes details about many different people.

The main goal that we have here is to build a model which could predict whether a person has income above or below $50,000 using the pattern found in the data based on various properties. This will help us understand what factors are linked to higher earnings and why some pe ople earn more than others.

**Understanding the Dataset**

The Dataset consists of 32,560 records and 15 features including Target Variable (Income). Each Record contains Several Details of an Individual like Age, Sex, Occupation, Relationship status etc.., along with their Demographic details of which the Target Variable is Income which is Binary representing whether person earns more than $50,000 or less than $50,000.

Key Features are:-

a. Age – Age of an Individual

b. Education – Highest Level of Education Attained.

c. Work class – Type of Employment

d. Hours per week – No. of Hours worked per week

e. Marital status, Occupation, Relationship, Race and Sex are also among the features.

So we observe that Dataset provide rich source of variables which can be helpful in exploring Relationship between Personal Characteristics and Income Levels.

**Exploratory Data Analysis (EDA)**

Then we proceeded with Exploratory Data Analysis which is an important step in Machine Learning Project as it helps us to understand the structure of the Data, identify patterns and spot potential Issue such as missing data or outliers.

In EDA we analyzed the Data Set properly and noted the types of data each column contain, shape of the data, checked the missing value. We observed following details from the dataset:

* 6 are int dtypes and 9 are object dtypes.
* Few columns in Dataset contains have missing values those rows have ‘?’ mark.
* 76 % earn less than $50,000 and 24% earn more than $50,000

Thus EDA helps in understanding which features might have the most impact on income prediction. For example, education and age seem to have a strong correlation with income.

**Data Visualization part :**

In Data Visualization part we displayed relationship between income and various features which gave an insight as how and which data is contributing and which is having adverse affect on Target column.

* Univariate Analysis - we saw the income distribution in the Data set 76 % earn less than $50,000 and 24% earn more than $50,000
* Bivariate Analysis - we saw relation of target column with various other features column, we found many private sector individual have earners above $50k and also has very huge no. of individual who earn less than $50k. On the other hand, job types like local-gov,self-emo-not-inc, Self-emp-inc and Federal-gov have a good number of individuals making over $ 50K.

Individual with high degree have good no. of earners who earn more than $50k.

Age 30 to 50 have good no. of earners above $ 50k.

Mostly Married-civ-spouse have good no. of earners above $ 50k.

Mostly Husbands are earning more than $50k.

Mostly white people are earning more than $50k.

Most Individual are from USA earning more than $50k.

* Multivariate Analysis - In Multivariate analysis we analysed pairwise relation of Numeric features.

**Data Processing**

To prepare the dataset for training a machine learning model, several preprocessing steps were necessary:

* Columns like Work class, Occupation and Native country have most of the rows missing its value having ‘?’ mark instead which is handled by Replace() fn.
* We used Replace() fn for substitution.
* Later on we encoded following column 'Workclass', 'Education', 'Marital\_status', 'Occupation', 'Relationship', 'Race', 'Sex', 'Native\_country', 'Income' which were in categorical form.

**Outliers in Dataset:**

Using Boxplot we got to know as which numeric columns have outliers and we handled it using z-score function

* **Capital Gain and Capital Loss** columns had significant outliers. These financial variables showed extreme values that could skew the results.
* **Age** also displayed some outliers, particularly with individuals aged 90, which could affect predictions as the majority of people fall into a younger age range.
* **Hours per Week** had a few extreme outliers, where individuals reported working almost 99 hours per week, which is very unusual and may represent errors or anomalies.

**Skewness in dataset:**

We observed

**Highly Positively Skewed**:

**Capital Gain (11.95)** and **Capital Loss (4.59)**: Most values are low with a few extreme high values.

**Negative Skewness**:

**Race (-2.43)** and **Native Country (-4.21)**: Majority of values are in lower categories, with fewer extreme values

**Removing Skewness:**

Using Power Transform (Ye Jonhshon) we dealt with it. Which left skewness with

Capital\_gain 3.016951

Capital\_loss 4.299511

Furthur we dropped both the columns as both were unnecessary and were not contributing , would have created problem in model training.

**Feature Scaling:**

Certain machine learning models, like Logistic Regression, require the features to be on the same scale. We used StandardScaler to standardize numerical features such as age and hours worked per week.

### Splitting the Dataset:

We divided the dataset into training and testing sets to evaluate the performance of our model:

* **Training Set**: 80% of the data, which is used to train the model.
* **Testing Set**: 20% of the data, which is used to check how well the model predicts outcomes on unseen data.

Applied SMOTE:

we also applied SMOTE (Synthetic Minority Over-sampling Technique) to handle the imbalance in our target variable, Income. Since a large percentage of the dataset (about 76%) consisted of individuals earning less than $50,000, the model could have been biased toward predicting this majority class, leading to poor performance in identifying people earning more than $50,000.

**Multicollinearity Check:**

For Multicollinearity check we did VIF Calculation on smote data and found no strong multicollinearity.

### Observations from the Model Performance:

After evaluating five different models (Logistic Regression, Decision Tree, Random Forest, Support Vector Classifier, and K-Nearest Neighbors), the **Random Forest Classifier** stands out as the best-performing model based on several key metrics:

1. **Accuracy:** Random Forest achieved the highest accuracy of 0.8305.
2. **Precision and Recall:** It maintains a good balance between precision (0.6455) and recall (0.6562).
3. **F1-Score:** The F1-Score of 0.6509 suggests that Random Forest generalizes well, with a decent trade-off between precision and recall.
4. **Overall Robustness:** Random Forest outperforms the other models consistently across multiple metrics.

## Hyper parameter Tuning for Random Forest Classifier:

After evaluating several models, we identified the **Random Forest Classifier** as the best-performing model for our dataset, achieving an accuracy of **83.05%.** To further improve this model's performance, we decided to conduct hyper parameter tuning using **Grid Search.** This technique helped us to find the optimal combination of hyper parameters to maximize the model's accuracy.

Upon executing the grid search, we found the following optimal hyperparameters for our Random Forest model:

* **Best Hyper parameters:**
  + max\_depth: 10
  + min\_samples\_leaf: 2
  + min\_samples\_split: 2
  + n\_estimators: 150
* **Best Accuracy Score:** 0.8384 (83.84%)

These results indicated that our model was likely to perform better with the tuned hyper parameters. We then proceeded to retrain our Random Forest model using these settings and evaluated its performance on the test dataset.

**Key Insight:**

Through our analysis, several important factors stood out as being strongly correlated with income level:

1. **Education**: Individuals with higher education levels, especially those with bachelor's degrees or higher, are significantly more likely to earn above $50K.
2. **Hours Worked**: Working more than 40 hours per week is positively correlated with higher income.
3. **Occupation**: Certain occupations, such as managerial roles or technical jobs, have a higher likelihood of falling in the >$50K category.
4. **Age**: Older individuals, particularly those in the 30-50 age range, tend to have higher incomes compared to younger workers.

**Conclusion:**

In conclusion, the Census Income Prediction Project successfully showcased the influence of various personal characteristics on income levels. Through our analysis, we found that factors like education, age, and job type play a significant role in determining whether an individual earns over $50,000 a year. The Random Forest Classifier emerged as the most effective model, achieving an impressive accuracy of 83.84% after fine-tuning its parameters.

This project not only deepened our understanding of income disparities but also highlighted the importance of data-driven decision-making. By recognizing the attributes linked to higher earnings, we can better inform policies and programs aimed at improving educational and job opportunities for individuals.