Income Prediction - US Census

Index

1. Business understanding

The given dataset is from the United States Census Bureau - a principal agency of the U.S. Federal Statistical System, responsible for producing data about the American people and economy.

Goal:
Predict whether an individual income is <=50K/yr or >50K/yr based on census data.

Features:

workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never Worked.

fnlwgt: continuous

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.

education-num: continuous.

marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

sex: Female, Male.

capital-gain: continuous.
capital-loss: continuous.

hours-per-week: continuous.

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.

Target Variable

Income: >50K, <=50K

2. Data Understanding & Preparation

a) Two files are provided **'data.csv'** to train the model and **'final.csv'** to test the model.
b) The training dataset contains **6 numerical** variables and **8 categorical** variables.
c) There are total **40935 records** in the training dataset.
d) All the columns in the dataset seems to have some missing value. (Except for column Income)

Below table shows the percentage of missing values for each columns

Column Name	Percentage of missing values (%)
Age	6.70
WorkClass	6.81
fnlwgt	6.81
Education	6.70
EducationNum	6.85
MaritalStatus	7.14
Occupation	6.74
Relationship	6.88
Gender	6.66
CapitalGain	6.85
CapitalLoss	6.64
HoursPerWeek	6.55
NativeCountry	6.67

e) The target variable 'Income' has duplicate records and represents the same value twice. <=50K and <=50K. are same, hence both were merged. Same process was done for >50K and >50K.

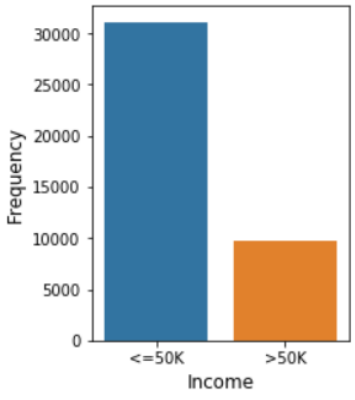
Before Merging

<=50K 21121
<=50K. 10002
>50K 6719
>50K. 3093

After Merging

<=50K 31123
>50K 9812

f) Plot showing the frequency of people having Income <=50K and >50K.



g) The data seems to be highly imbalanced and might Require use of some sampling techniques to Balance the minority class,

h) Workclass, Occupation & NativeCountry columns contained a Special character ('?') which was replaced with null value.

i) In order to make the analysis if Education simpler - 'HS-grad','11th','10th','9th','12th' were merged into one set of category as 'High school graduates' and '1st-4th','5th-6th','7th-8th' into

another set of category as 'elementary'. Similarly, 'Married-spouse-absent','Married-civ-spouse','Married-AF-spouse' were merged into one set of category as 'Married' and 'Separated', 'Divorced' into another set of category as 'seprated' for MaritialStatus column.

j) Majority of the people belongs to United states with additional remaining people belonging to 40 different country. Hence in order to simplify the analysis, countries were divided into 6 geographical region.

Region/Country	Count
United_States	34265
Central_America	1482
Europe_West	506
East_Asia	400
Central_Asia	173
South_America	128
Europe_East	105

k) The missing value for Age, HoursPerWeek, CapitalGain, CapitalLoss were imputed using mean.

l) In the dataset there were some records which represented people with age >=90 working for more than 90 hours, which seems unrealistic and hence 2 records were removed.

m) Capital gain of $9999 doesn't followed any pattern for any age group and was clearly an outlier. Hence, 891 records were removed from the dataset.

n) EducationNum and Education were giving similar information. Similary, Relationship and MaritalStatus imply similar information. Hence only one is kept (Education & MaritialStatus)

o) Not enough information was provided for 'fnlwgt' hence it was removed.

p) NativeCountry was also removed, since we have already created new feature name Native Region

3. Model Selection

Since we need to predict weather a person ears an Income less than or greater than a particular threshold, we can consider this problem belonging to classification.

1) For the choosing our prediction model Logistic regression was used as a baseline model and Random forest for feature selection and feature importance.

2) One Hot Encoding was used to convert all the categorical variables into dummy variables.
3) The value of Target variable was set to binary (0 for <=50 K and 1 for >50K), which will easily help in classification.

4. Model Creation

Without Oversampling

a) **Logistic Regression:** The data was split in training and test set and was fitted with Logistic regression model. Model accuracy was **83.6%**

K-Fold Cross validation was used to check If our model is getting overfitted or not.

The average score for K-Fold Cross validation was **83.4%** which doesn't show much deviation from our baselined predicted score. Hence, we can neglect the assumption that the was overfitted.

Apart from accuracy we want to know how well the model can specifically classify Income. In statistics, this is called recall, and it's the number of correctly predicted "positives" divided by the total number of "positives".

Here we are predicting Income with **81.3%** accuracy.

b) **Random Forest:** The Random Forest Classifier was used as a second predictive model. Model accuracy was **81.7%**

Conclusion: Among Logistic Regression and Random Forest, **Logistic Regression** seems to be performing well and giving better result.

With Oversampling: Using **SMOTE** algorithm (Synthetic Minority Oversampling Technique)

SMOTE creates synthetic observations of the minority class by:

- Finding the k-nearest-neighbors for minority class observations (finding similar observations)
- Randomly choosing one of the k-nearest-neighbors and using it to create a similar, but randomly tweaked, new observation.

a) **Logistic Regression:** After performing oversampling using SMOTE, the model accuracy came to be **81.5%**. In other words, model accuracy got decreased from **83.6% to 81.5%** after performing oversampling.

b) **Random Forest:** Random Forest performed little better with oversampling in compare to without and accuracy of the model of increased from **81.7% to 82.2%**.

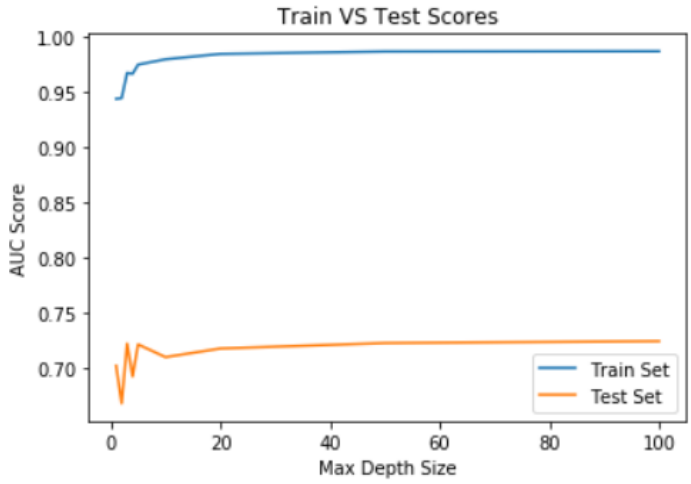
Recall: 55.2%
Precision: 61.32%

5. Model Tuning

After tuning the Random Forest model (SMOTE) following values of the hyperparameter were chosen:

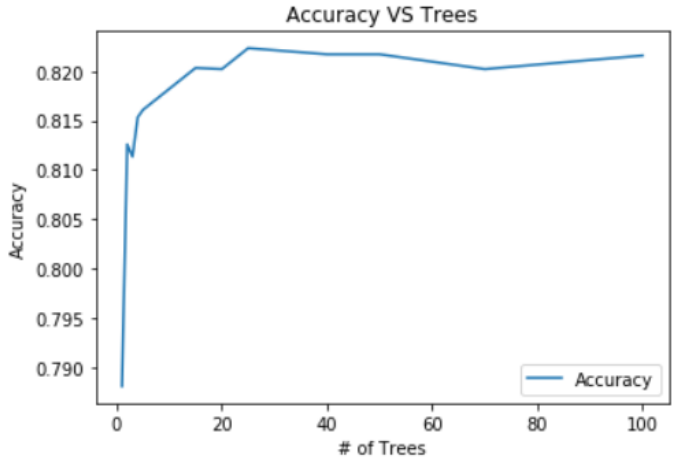
a) Tree depth:

Tree with depth 50 gives best value for AUC Score, after depth of 50 there seems to be no significant change in score.



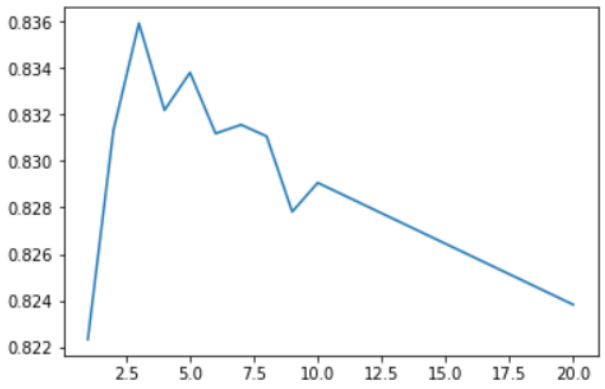
b) Number of Trees:

Highest accuracy is achieved when number of trees were 25.



c) Min Sample Leaf:

Highest accuracy is achieved when value for min_sample_leaf is 3.

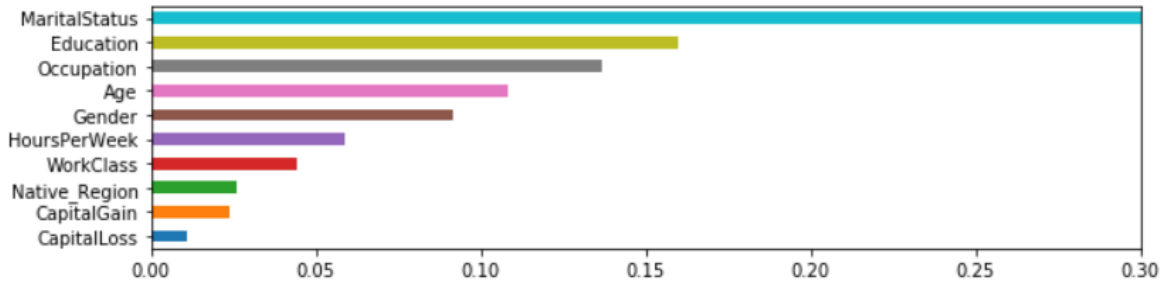


After hyperparameter tuning the Random Forest model (SMOTE) was again ran.
The Accuracy of the model got increased from **82.2% to 83.6%** along with significant increase in Recall from **55.2% to 63.1%**.

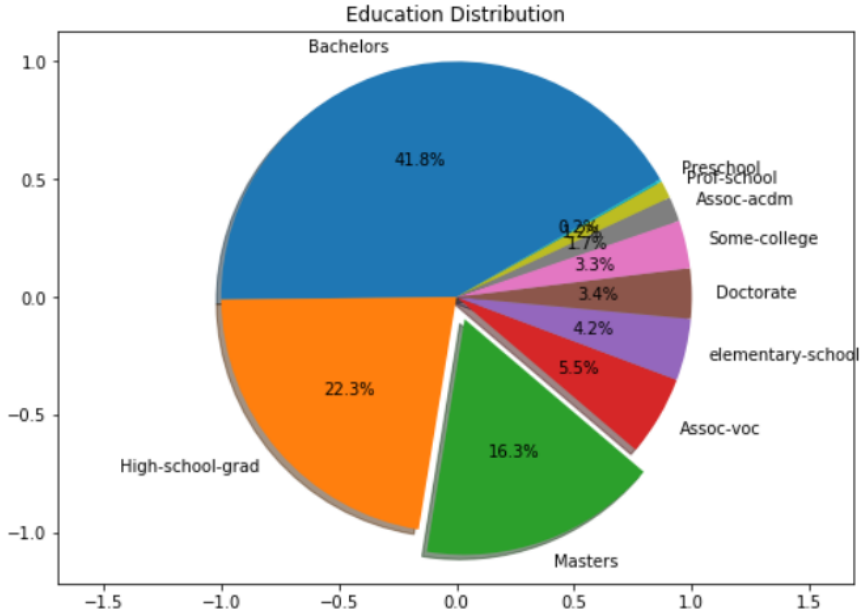
Conclusion : Our Final choice of model will be Random Forest with Oversampling.

Key Insights & Take Always

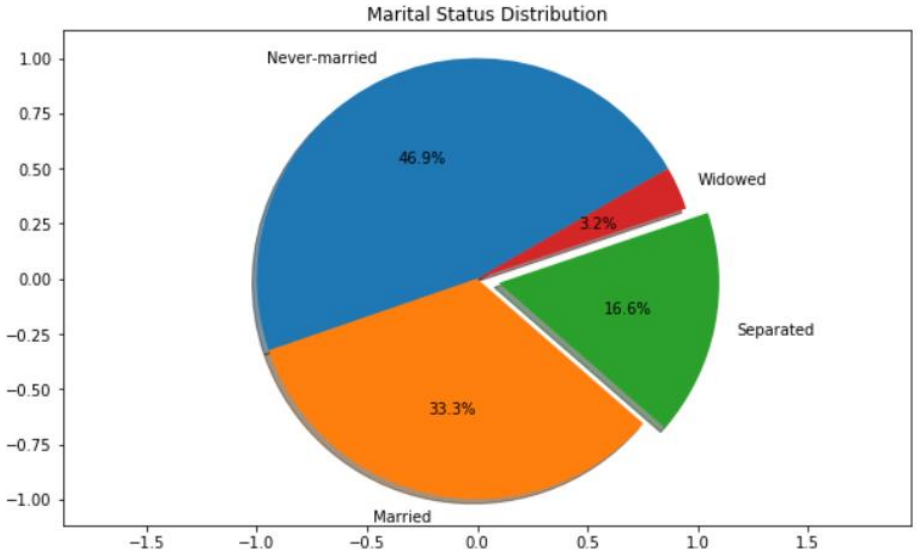
1) The average age of people is approx. **38 years** with minimum being 17 years and maximum being 90 years.
2) On average, people work for **40 hours** per week
3) Majority of the people had **High school Education** followed by some college degree.
4) CapitalGain represents profit from the sale of property or an investment. The average CapitalGain of the population from dataset is around 1700.
5) Based on feature importance we can see that MaritialStatus seems to have higher significance followed by Education & Occupation



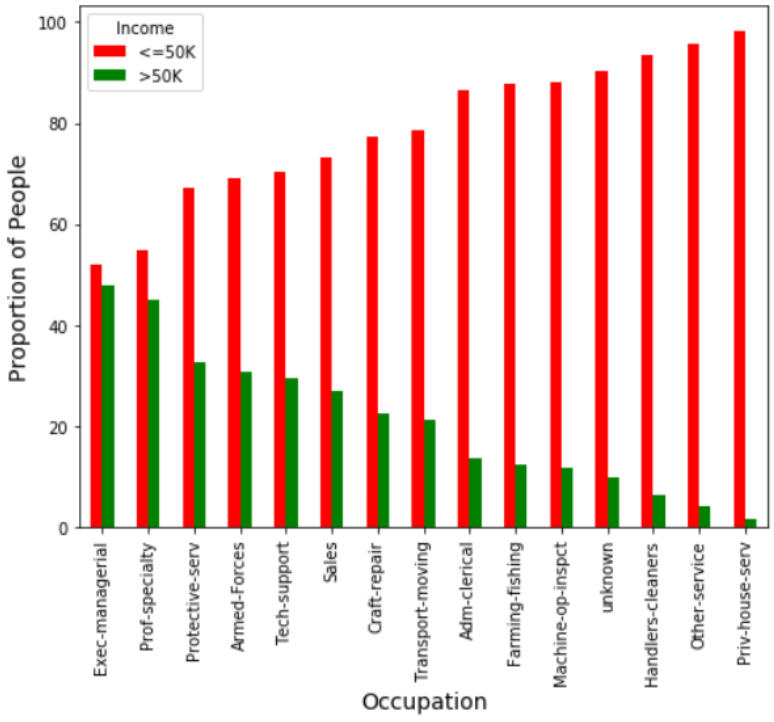
6) **41.8%** of the people had **bachelor's degree** followed by **High school education** and **Masters**



7) **46.9%** of the people have never married

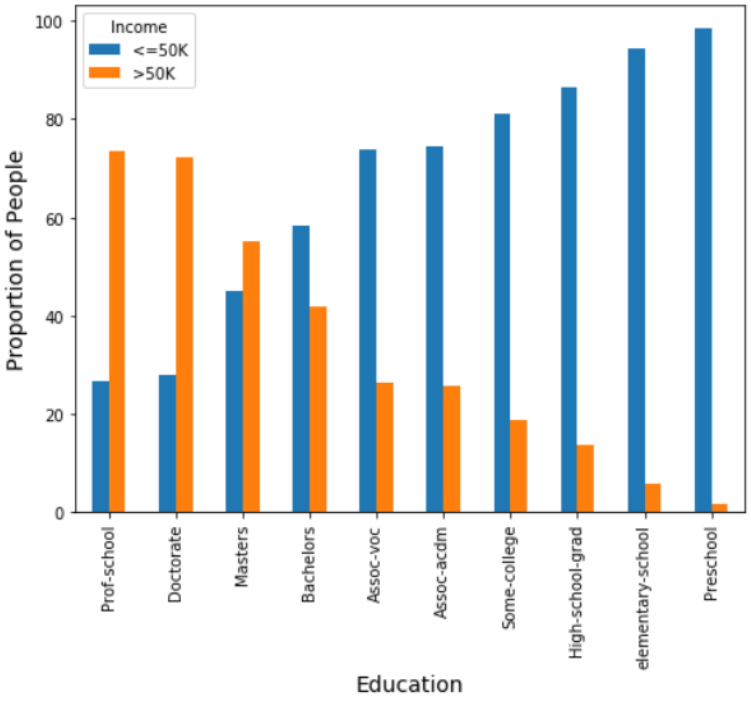


8) People who belongs to Exec-managerial category and prof-specialty earns more than 50K per year in compare to people belonging to other categories.

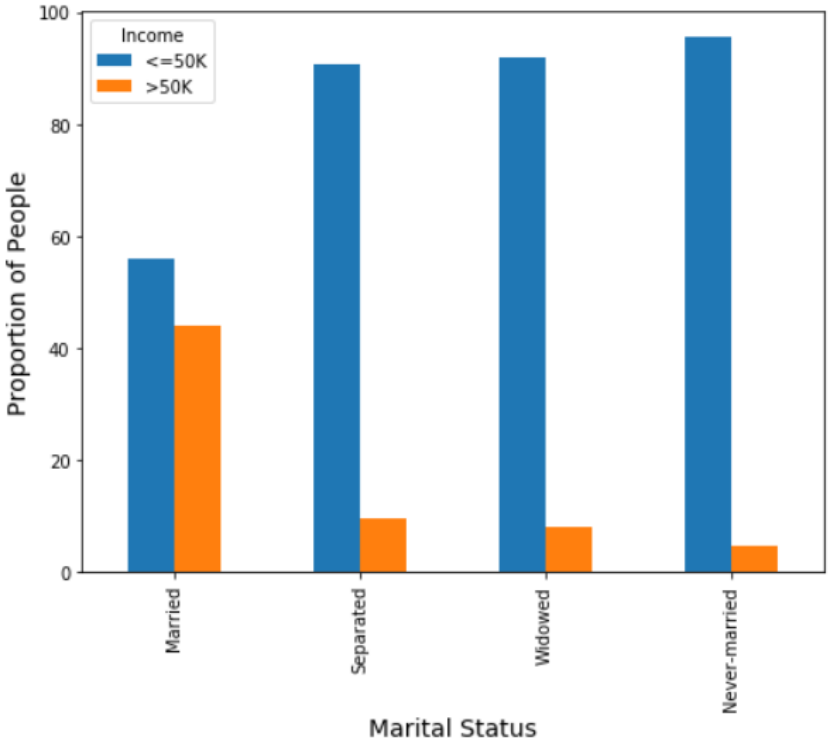


.

9) People who have **Doctorate, master's degree and attended a prof school** earns more than 50K per year, whereas people who only attended pre-school earns less than 50K per year.



10) People who are **Married** appears to be earning more than 50k per year.



11) The top 5 countries which represents people earning more than 50k per year were Yugoslavia, India, Iran, Greece and France.