Adult Census Income Prediction

JUNE 17

High level document design
Authored by: Rutvi Gohel

Introduction:

The US Adult Census dataset is a repository of 48,842 entries extracted from the 1994 US Census database.

We explore the data at face value in order to understand the trends and representations of certain demographics in the corpus. We then use this information to form models to predict whether an individual made more or less than $50,000 in 1994. We compare our models as well as that of others in order to find out what features are of significance, what methods are most effective, and gain an understanding of some of the intuition behind the numbers.

Descriptive Analysis

The Dataset

The Dataset is taken from Kaggle.
The Census Income dataset has 48,842 entries. Each entry contains the following information about an individual:

- **age:** the age of an individual
  - Integer greater than 0

- **workclass:** a general term to represent the employment status of an individual
  - Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

- **fnlwgt:** final weight. In other words, this is the number of people the census believes the entry represents.
  - Integer greater than 0

- **education:** the highest level of education achieved by an individual.
  - Bachelors, Somecollege, 11th, HSgrad, Profschool, Assocacdm, Assocvoc, 9th, 7th8th, 12th, Masters, 1st4th, 10th, Doctorate, 5th6th, Preschool.

- **Educational_num:** the highest level of education achieved in numerical form.
  - Integer greater than 0

- **Marital-status:** marital status of an individual. Marriedcivspouse corresponds to a civilian spouse while MarriedAFspouse is a spouse in the Armed Forces.

- Marriedcivspouse, Divorced, Nevermarried, Separated, Widowed, Marriedspouseabsent, MarriedAFspouse.

- **occupation:** the general type of occupation of an individual
  - Techsupport, Craftrepair, Otherservice, Sales, Execmanagerial, Profspecialty, Handlers-cleaners, Machineopinspct, Admclerical, Farmingfishing, Transportmoving, Privhouseserv, Protectiveserv, ArmedForces.

- **relationship:** represents what this individual is relative to others. For example an individual could be a Husband. Each entry only has one relationship attribute and is somewhat redundant with marital status. We might not make use of this attribute at all
  - Wife, Ownchild, Husband, Notinfamily, Otherrelative, Unmarried.

- **race**: Descriptions of an individual's race
  - White, AsianPacIslander, AmerIndianEskimo, Other, Black.

- **gender**: the biological sex of the individual
  - Male, Female

- **capital-gain**: capital gains for an individual
  - Integer greater than or equal to 0

- **capital-loss**: capital loss for an individual
  - Integer greater than or equal to 0

- **hours-per-week:** the hours an individual has reported to work per week
  - continuous.

- **Native-country:** country of origin for an individual
  - UnitedStates, Cambodia, England, PuertoRico, Canada, Germany, OutlyingUS(GuamUSVIetc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, DominicanRepublic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, ElSalvador, Trinadad&Tobago, Peru, Hong, HolandNetherlands.

- **income**: whether or not an individual makes more than $50,000 annually.
  - <=50k, >50k

```python
#features datatype
adult.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48842 entries, 0 to 48841
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   age              48842 non-null  int64 
 1   workclass        48842 non-null  object
 2   fnlwgt           48842 non-null  int64 
 3   education        48842 non-null  object
 4   educational-num  48842 non-null  int64 
 5   marital-status   48842 non-null  object
 6   occupation       48842 non-null  object
 7   relationship     48842 non-null  object
 8   race             48842 non-null  object
 9   gender           48842 non-null  object
 10  capital-gain     48842 non-null  int64 
 11  capital-loss     48842 non-null  int64 
 12  hours-per-week   48842 non-null  int64 
 13  native-country   48842 non-null  object
 14  income           48842 non-null  object
dtypes: int64(6), object(9)
memory usage: 5.6+ MB
```

The dataset doesn't have any null values, but it contains missing values in the form of '?' in feature (workclass, occupation, native-county) which needs to be preprocessed.

```python
# Checking the counts of label categories
income = adult['income'].value_counts(normalize=True)
income
```

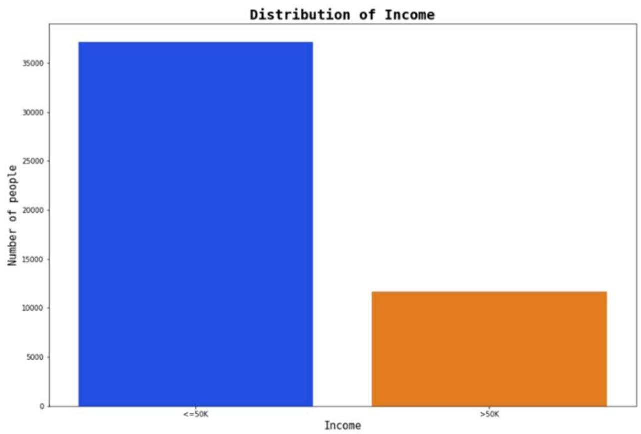
```
<=50K    0.760718
>50K     0.239282
Name: income, dtype: float64
```

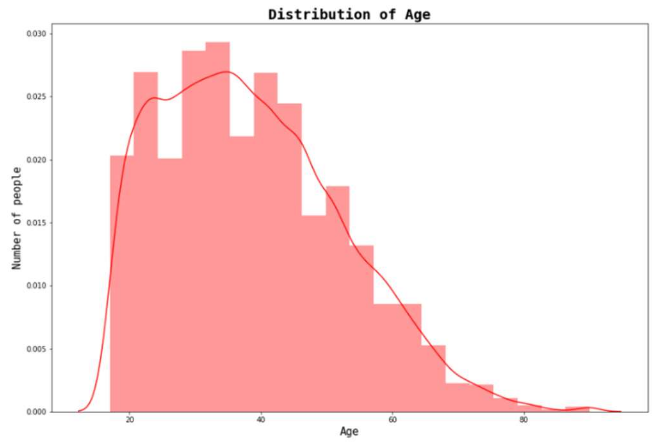
The dataset is unbalanced, as the dependent feature 'income' contains 76.07% values have income less than 50k and 23.92% values have income more than 50k.

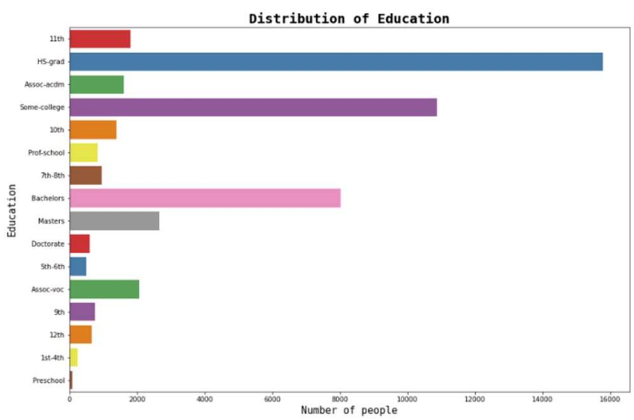
Exploratory Data Analysis

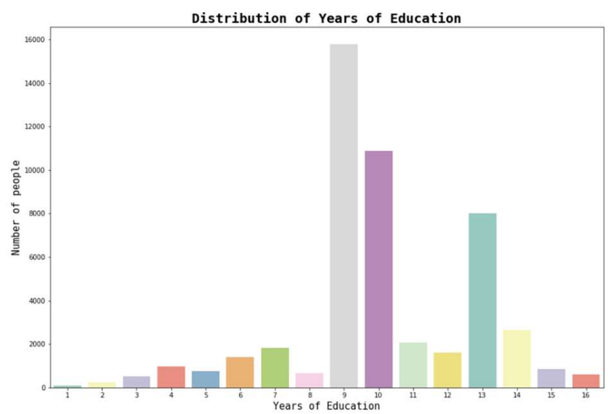
The following Graphs help us to get insight of the data.

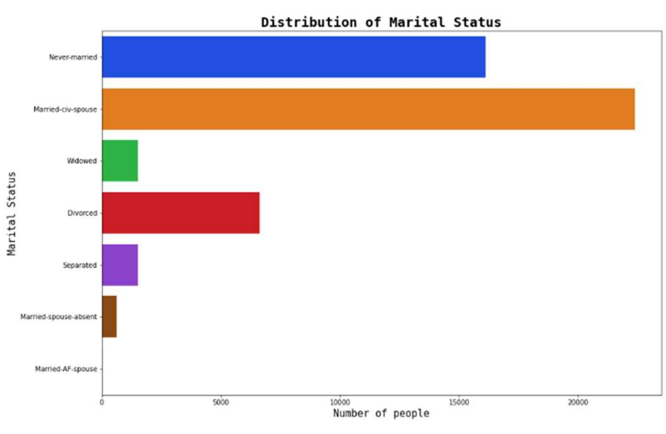
Univariate Analysis

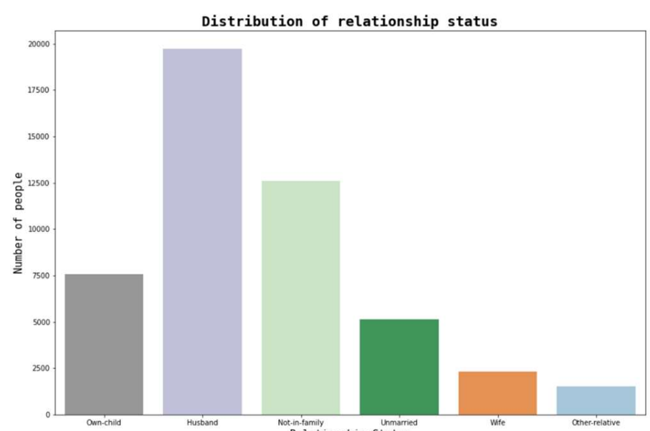


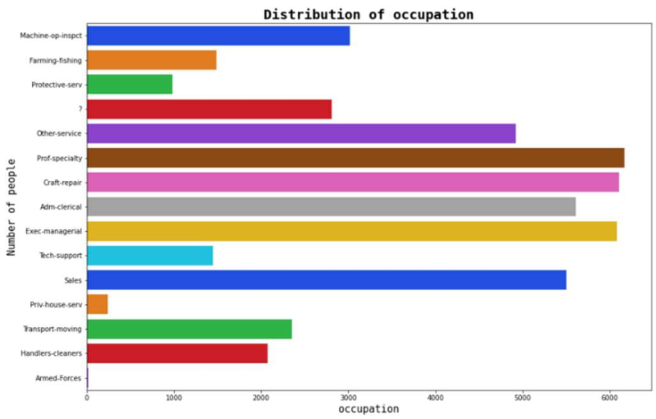


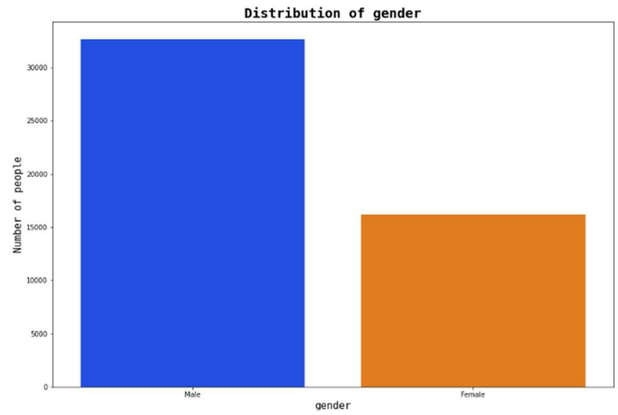


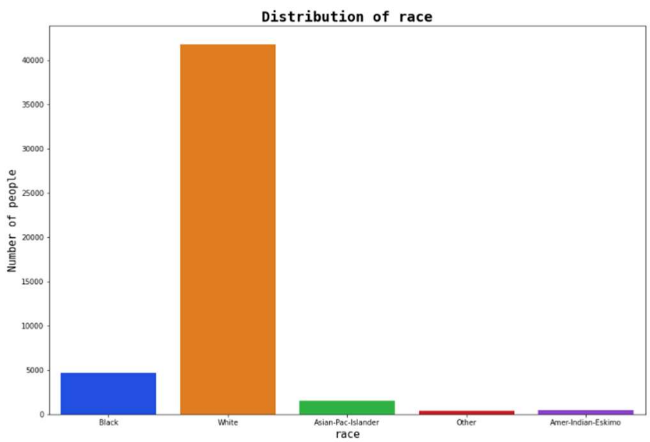


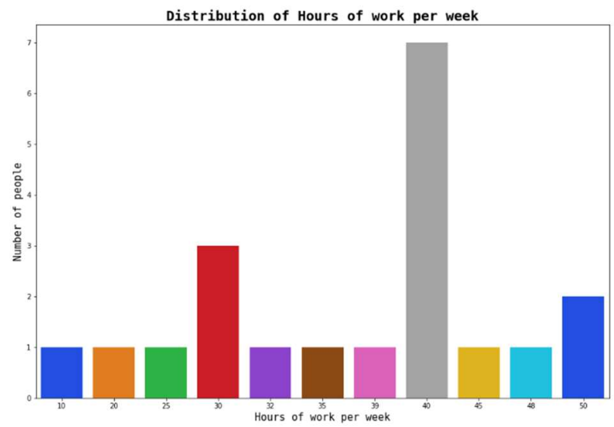




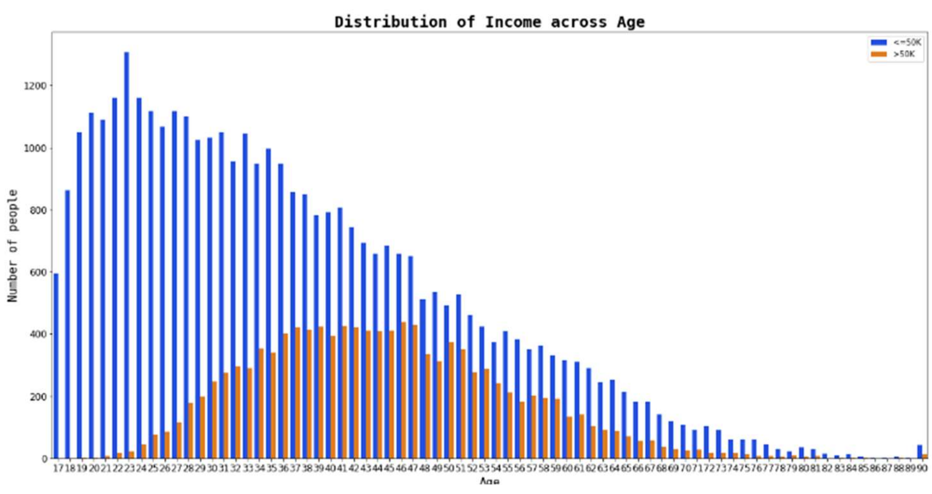


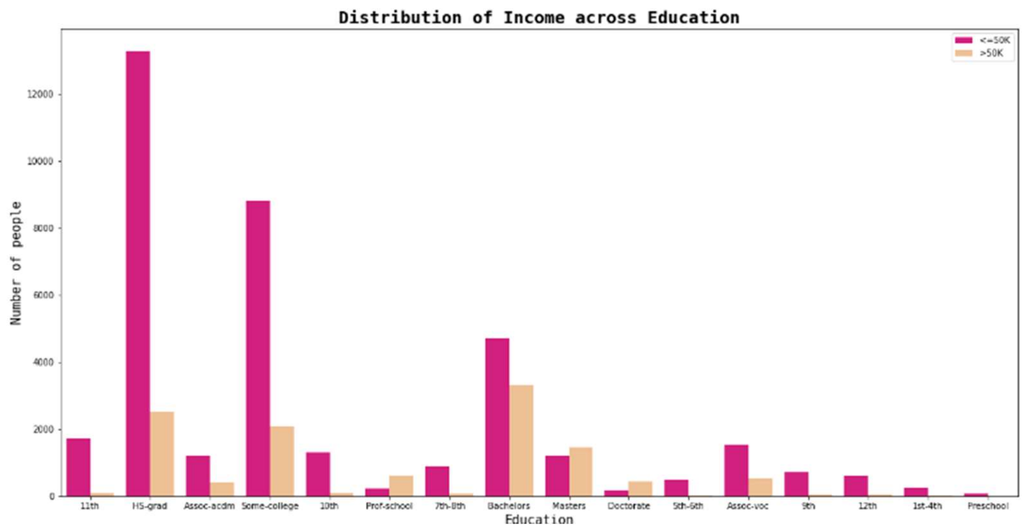


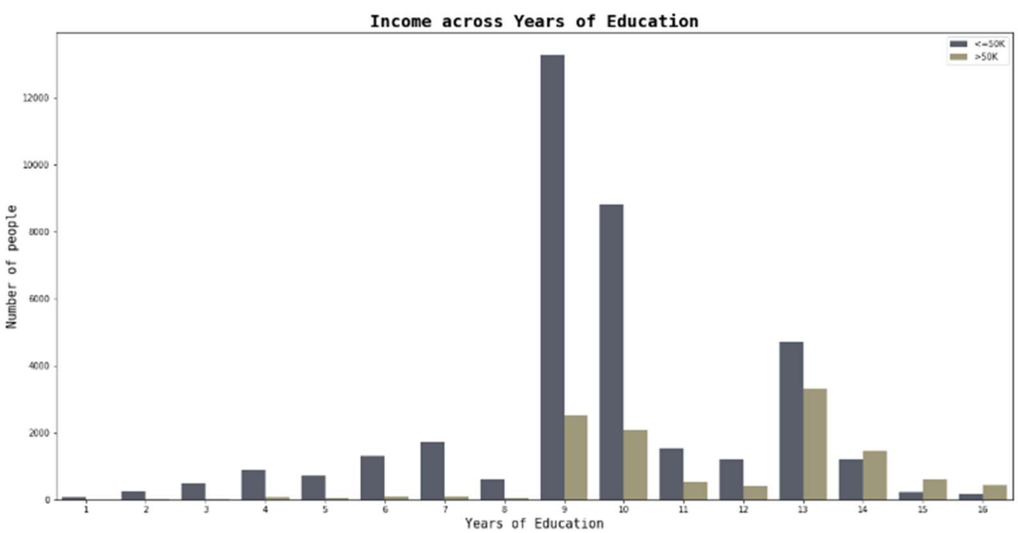


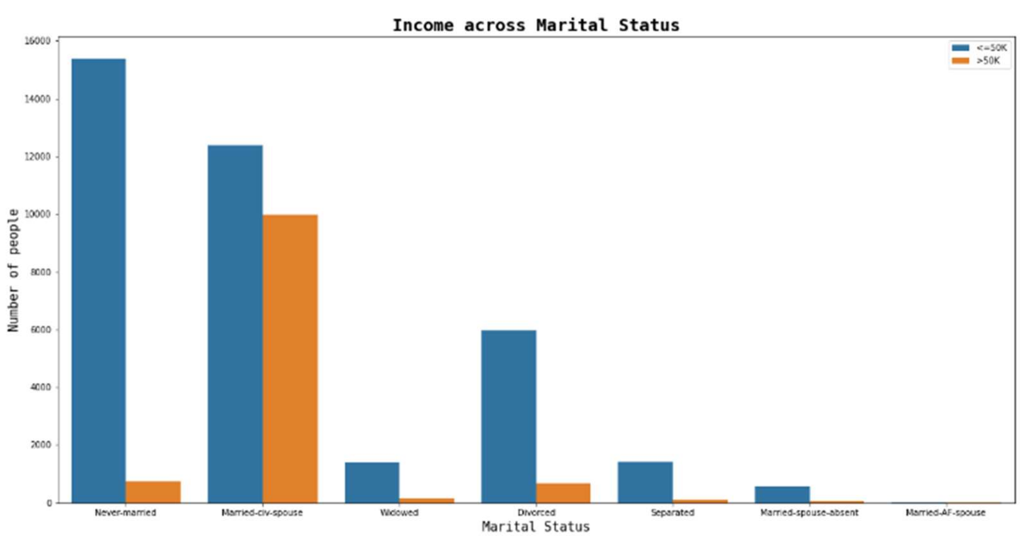


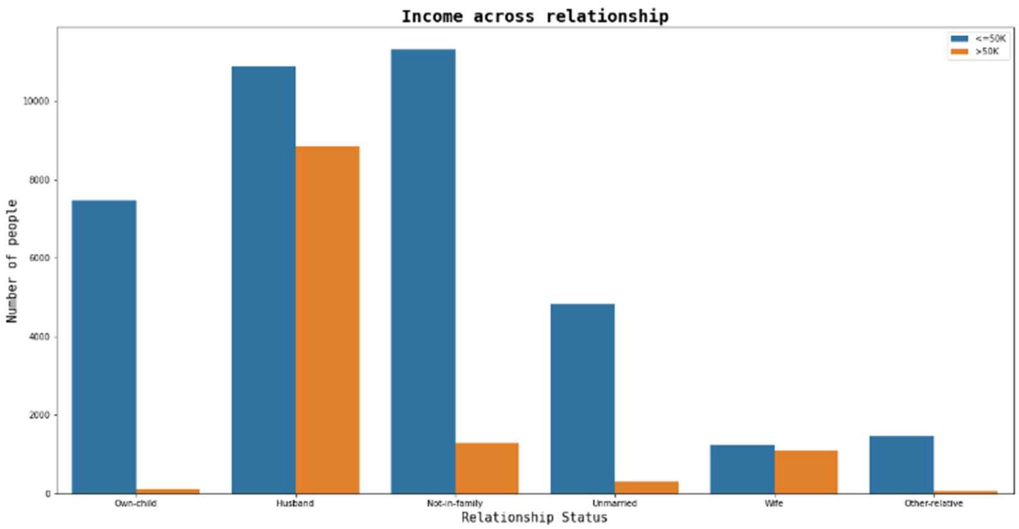
Bivariate Analysis

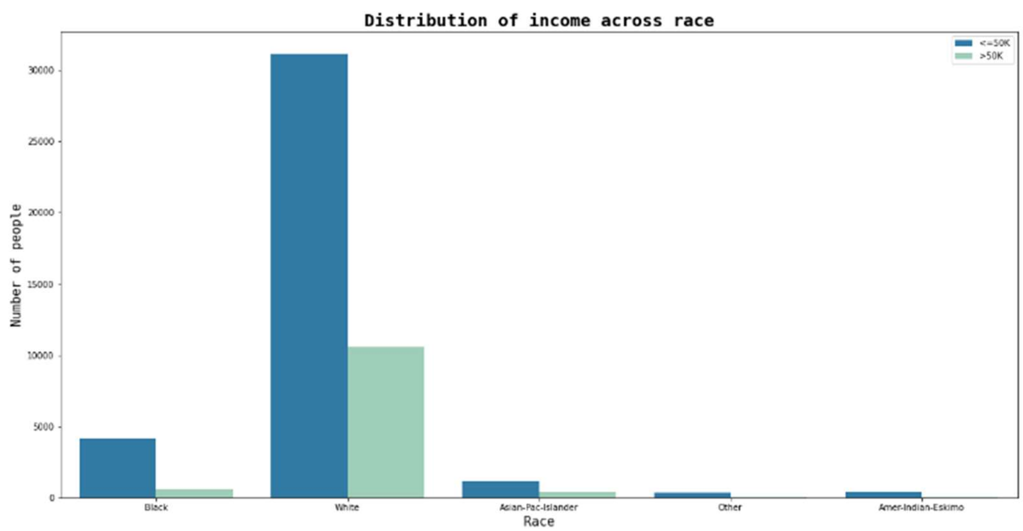


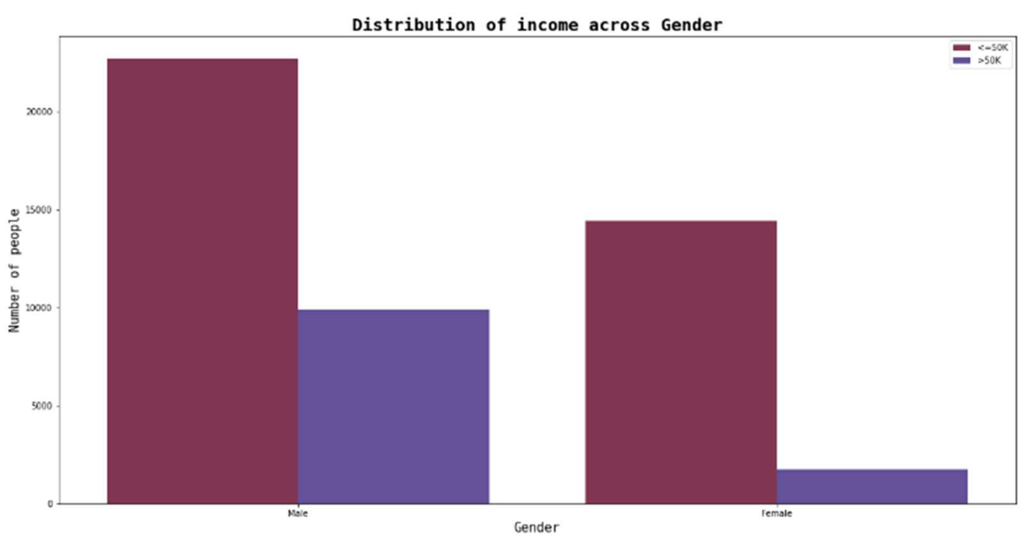

Distribution of Income across Education
<=50K
>50K
Number of people
Education
11th
HS-grad
Assoc-acdm
Some-college
10th
Prof-school
7th-8th
Bachelors
Masters
Doctorate
5th-6th
Assoc-voc
9th
12th
1st-4th
Preschool

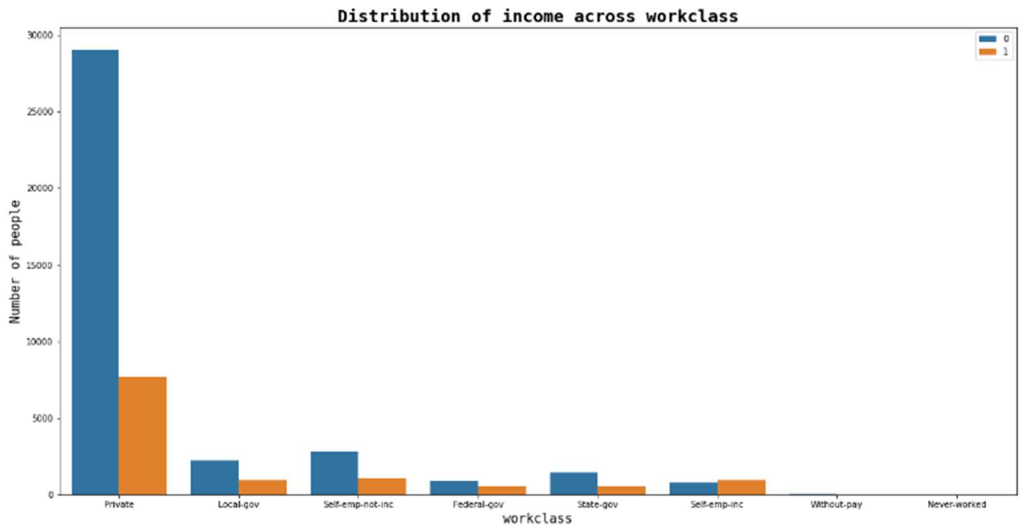

Income across Years of Education
<=50K
>50K
Number of people
Years of Education

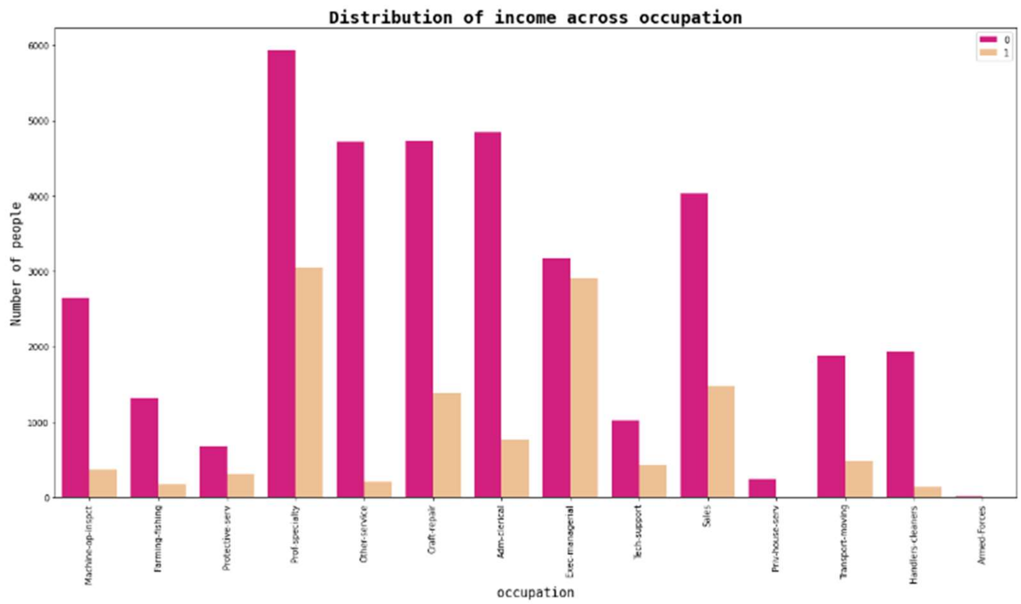

Income across Marital Status
<=50K
>50K
Number of people
Marital Status
Never-married
Married-civ-spouse
Widowed
Divorced
Separated
Married-spouse-absent
Married-AF-spouse


Income across relationship
<=50K
>50K
Number of people
Own-child
Husband
Not-in-family
Unmarried
Wife
Other-relative
Relationship Status


Distribution of income across race
<=50K
>50K
Number of people
Black
White
Asian-Pac-Islander
Other
Amer-Indian-Eskimo
Race


Distribution of income across Gender
<=50K
>50K
Number of people
Male
Female
Gender





We do this in9 hopes to identify features that provide little information in order to simplify our model's complexity and runtime.

The age feature describes the age of the individual. The ages range from 17 to 90 years old with the majority of entries between the ages of 25 and 50 years.

The education feature describes the highest level of education of each individual in the dataset. The Other group represents Preschool through 12th grade. Most of the individuals in the dataset have at most a high school education while only a small portion have a doctorate. We think this is a fair representation. For the most part, a higher level of education is correlated to a higher percentage of individuals with the income >50k. One interesting statistic to note is the ratio of individuals labeled >50k to <=50k is almost the same between those that have a doctorate and those that went to a professional school (Profschool).

The majority of the individuals work in the private sector. The probabilities of making above $50,000 are similar among the work classes except for self-emp-inc and federal government. Federal government is seen as the most elite in the public sector, which most likely explains the higher chance of earning more than $50,000.

There is a somewhat uniform distribution of occupations in the dataset, disregarding the absence of Armed Forces. However, Occupation vs Income, exec-managerial and prof-specialty stand out as having very high percentages of individuals making over $50,000. In addition, the percentages for Farming-fishing, Other-service and Handlers-cleaners are significantly lower than the rest of the distribution.

Looking at the distribution of hours per week, the vast majority of individuals are working 40 hour weeks which is expected as the societal norm.

Looking at the Distribution of income across race, it seems like the feature could be useful in our prediction model, as Whites and Asians have a larger percentage of entries greater than $50,000 than the rest of the races. However, the sample size of Whites in the dataset is disproportionately large in comparison to all other races. The second most represented group is Blacks with about 4000 entries. The lack of equal distribution caused us to consider not utilizing this attribute in our prediction model.

In Distribution of income across Gender, we can see that there is almost double the sample size of males in comparison to females in the dataset. While this may not affect our predictions too much, the distribution of income can. The percentage of males who make greater than $50,000 is much greater than the percentage of females that make the same amount. This will certainly be a significant factor, and should be a feature considered in our prediction model.

Data Preprocessing

Fixing '?' values in the dataset

As our dataset have some '?' values in feature names: workclass, occupation and native-county; we need to fix them. As these features are classes, we need to fill the '?' value with most occurrence class. That is nothing but mode. After replace '?' value with mode; now, we have no '?' values.

Feature Selecting

We also opted to not use the features: 'fnlwgt', 'capital-loss'. These features were not useful for our analysis.

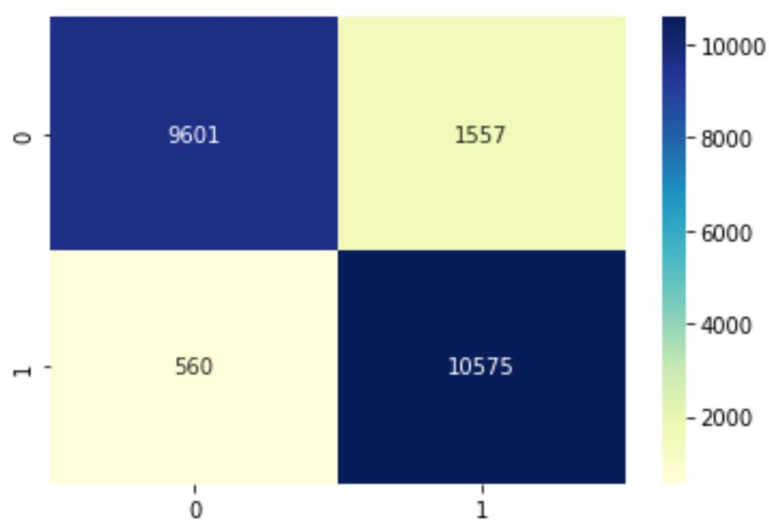
Fixing Imbalance dataset

As mentioned before, we saw a distribution of roughly twenty-four percent entries labeled with >50k and seventy-six percent labeled with <=50k. In order to establish baseline data for our classifiers, we predicted the majority label <=50k for each item. So, we fixed this by using oversampling technique.

Data Modeling

Model	Accuracy Score	F1- Score
Logistic Regression	82.18 %	82.59 %
KNN- Classifier	82.69 %	83.47 %
Support Vector Classifier	81.97 %	82.77 %
Decision Tree Classifier	88.98 %	89.43 %
Random Forest Classifier	90.50 %	90.90 %
XGB Boost	85.55 %	85.99 %
Random Forest Classifier with Hyperparameter tuning	90.33 %	90.80 %

Report



	precision	recall	f1-score	support
0	0.94	0.86	0.90	11158
1	0.87	0.95	0.91	11135
accuracy			0.91	22293
macro avg	0.91	0.91	0.90	22293
weighted avg	0.91	0.91	0.90	22293

In this project, I build various models like logistic regression, KNN classifier, support vector classifier, decision tree classifier, random forest classifier and XGboost classifier.

A Random Forest Classifier (without hyper parameter tunned) gives the highest accuracy score of 90.50% and f1 score of 90.90%.

Future Work

We have a large enough dataset, so we can use neural networks such as an artificial neural network to build a model which can result in better performance.

Thank You,
Rutvi Gohel