Classifying Income from 1994 Census Data

Prasanti Das

Supervised Learning Capstone

May 2018

About the Data

Approx. 48,000 working people over the age of 16, who made over $100 that year

Attributes:

age, work class,
education, marital-status,
occupation, relationship,
race, sex, capital-gain,
capital-loss, hours-per-week, income

Research Question

What:

- Determine whether a person In USA earns over $50K a year.
- Which features are the best determinant for a person's income to be over $50K a year?

Why?

Income is a primary concern that dictates the standard of living of an individual. So knowing the factors that help to earn more and live a good life would be interesting.

Basic Statistics

* Displaying only major categories

Attribute	Values
Education Level	High School (32%), Some college (22%), Bachelors (16%), Masters (5%) etc.
Work Class	Private (69%), Local-gov (6%), State-gov(4%) etc.
Race	White (85%), Black (10%) etc.
Marital Status	Married-civ-spouse (46%), Never-married (33%), Divorced (14%), Separated (3%) etc.
Gender	Male (67%), Female (33%)
Salary [Label]	<=$50K (75%), >$50K (25%)

Data Cleaning

Missing values:

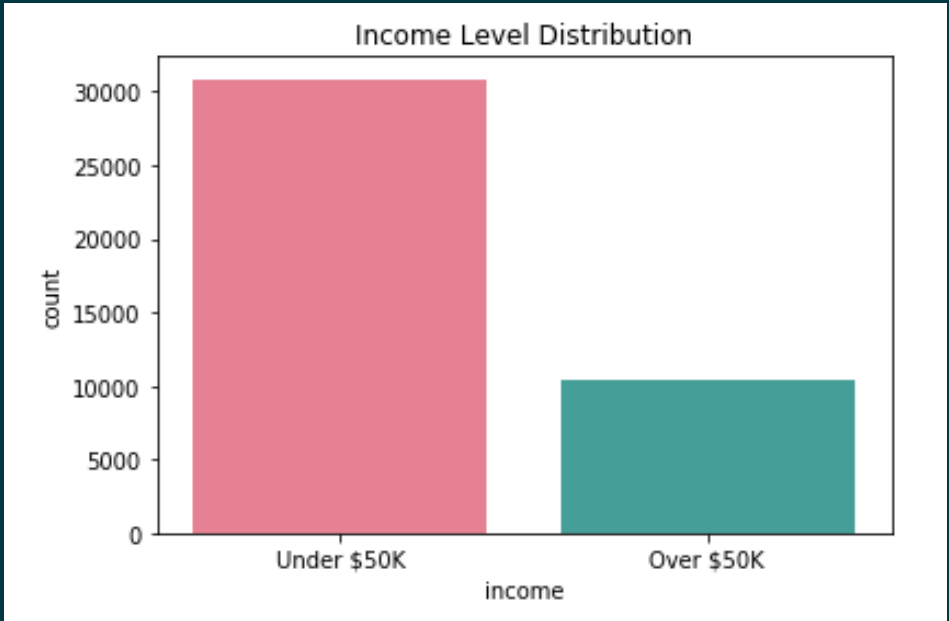
Removed observations that had a missing value for attribute : Occupation, work class.

Removal of Features:

‘fnlwgt’, ‘relationships’, education number. These features were not useful for our analysis.
Also dropped the attribute native country, as I am limiting my research only to USA.

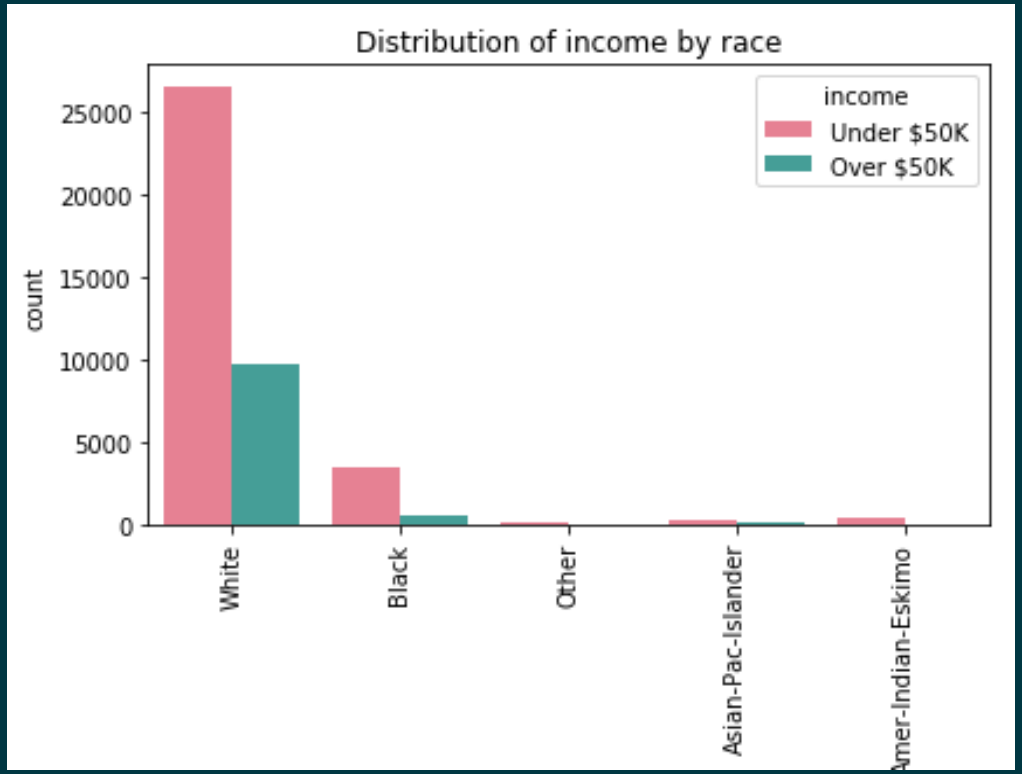
Exploratory Data Analysis

Distribution of Income



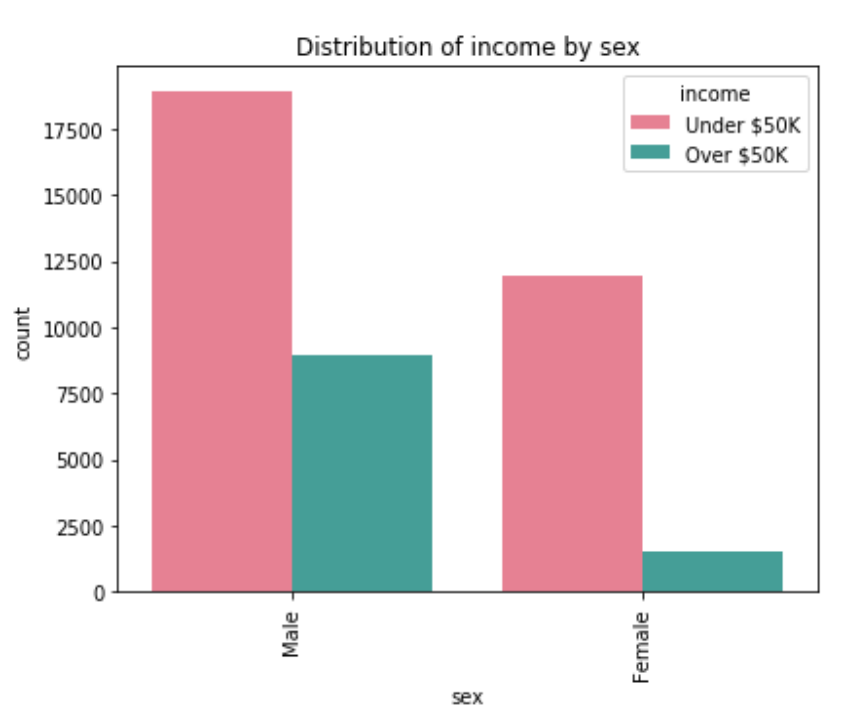
The dataset contains
a distribution of 25% entries labeled with over $50k and 75%
entries labeled with under $50k.

Distribution of Income By Race



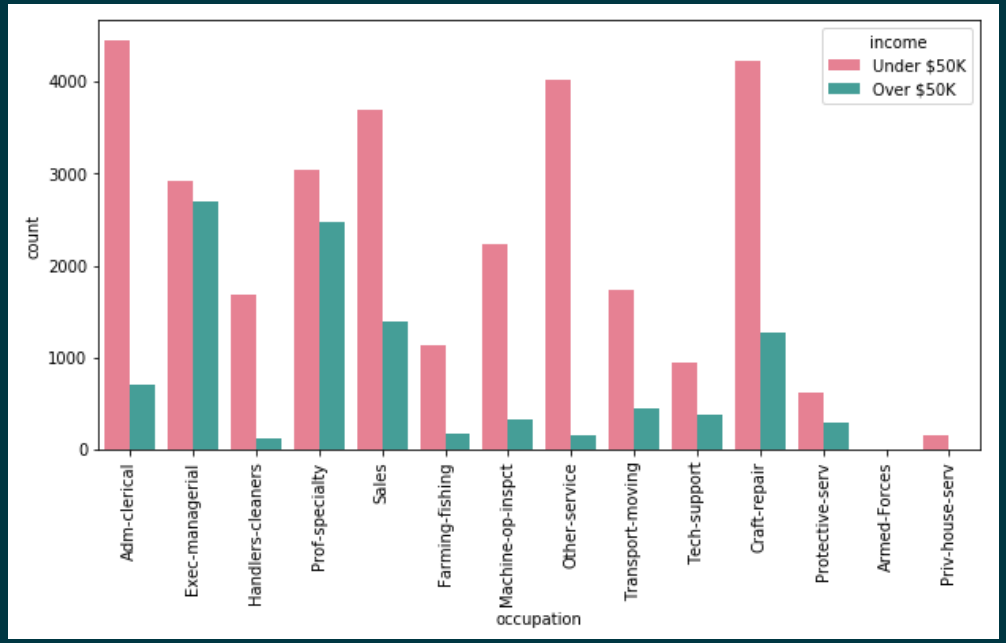
Whites have a larger percentage of entries for over $50,000 than the rest of the races.

Distribution of Income By Sex

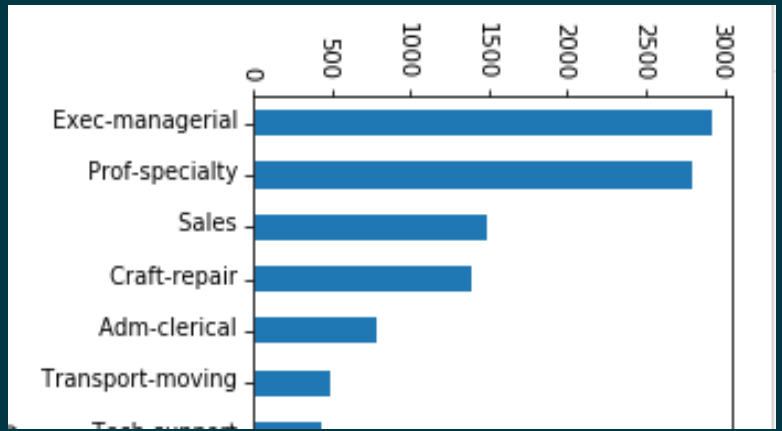


The number of males who make over $50K is much greater than the number of females that make the same amount.

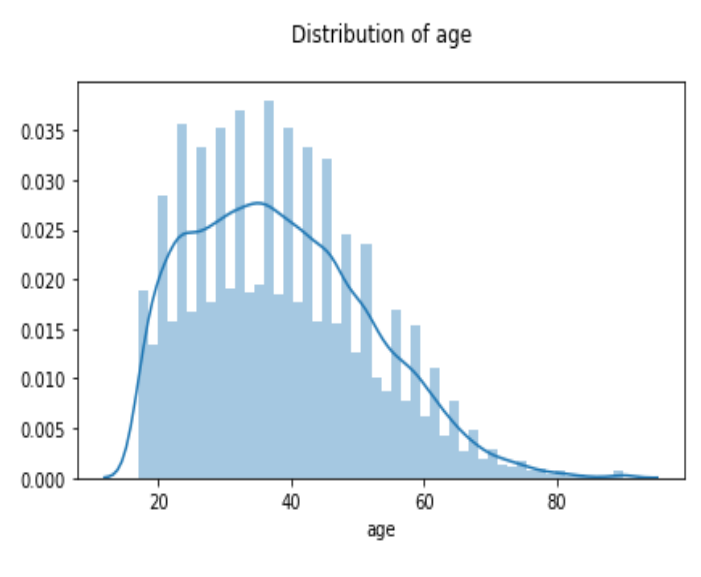
Distribution of Income By Occupation

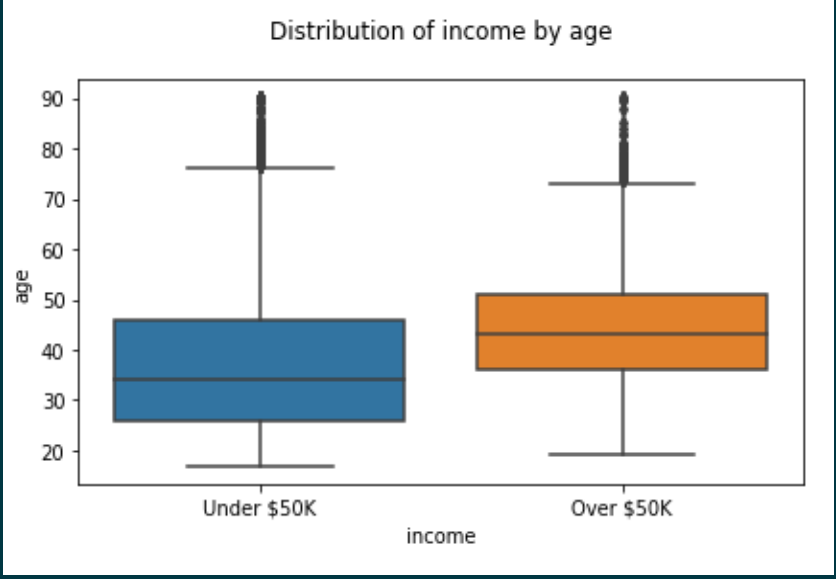


Exec managerial and prof specialty stand out as having very high number of individuals making over $50,000.



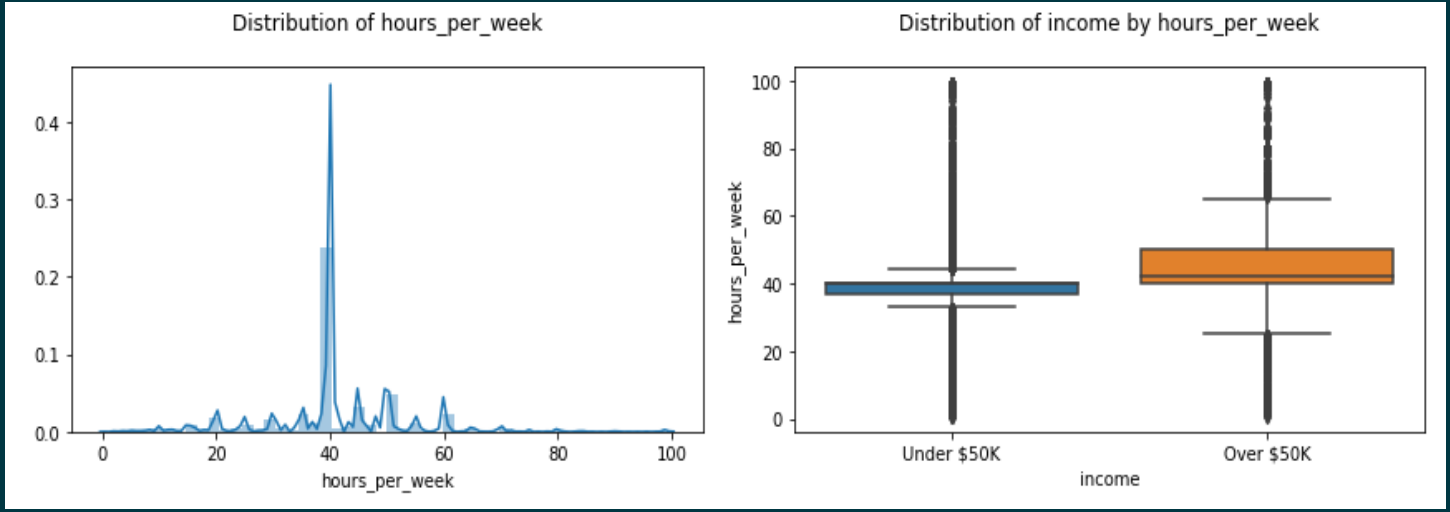
Distribution of Income By Age





The ages range from 17 to 90 years old. Older people are more likely to earn more than $50K.

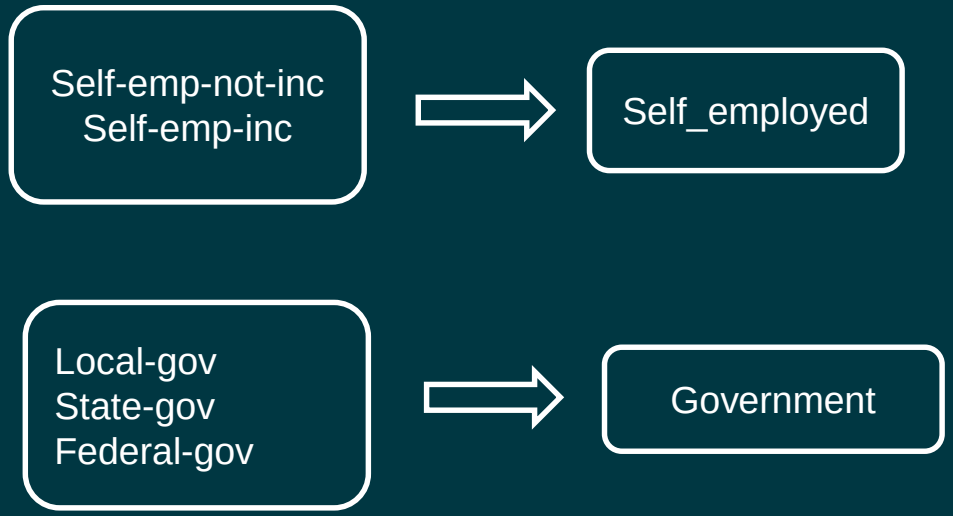
Distribution of Income By hours per week



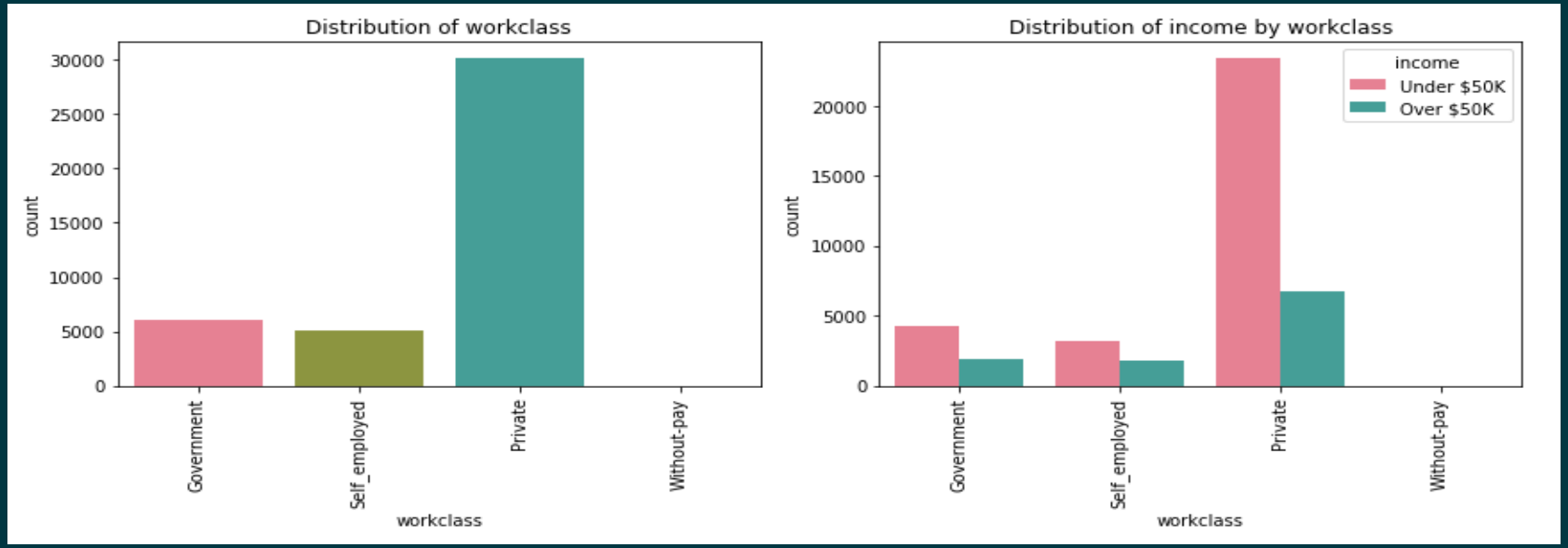
The majority of people are working 40 hour per weeks. Those who invest more time at work tend to be earning more.

Feature Engineering

Working Class

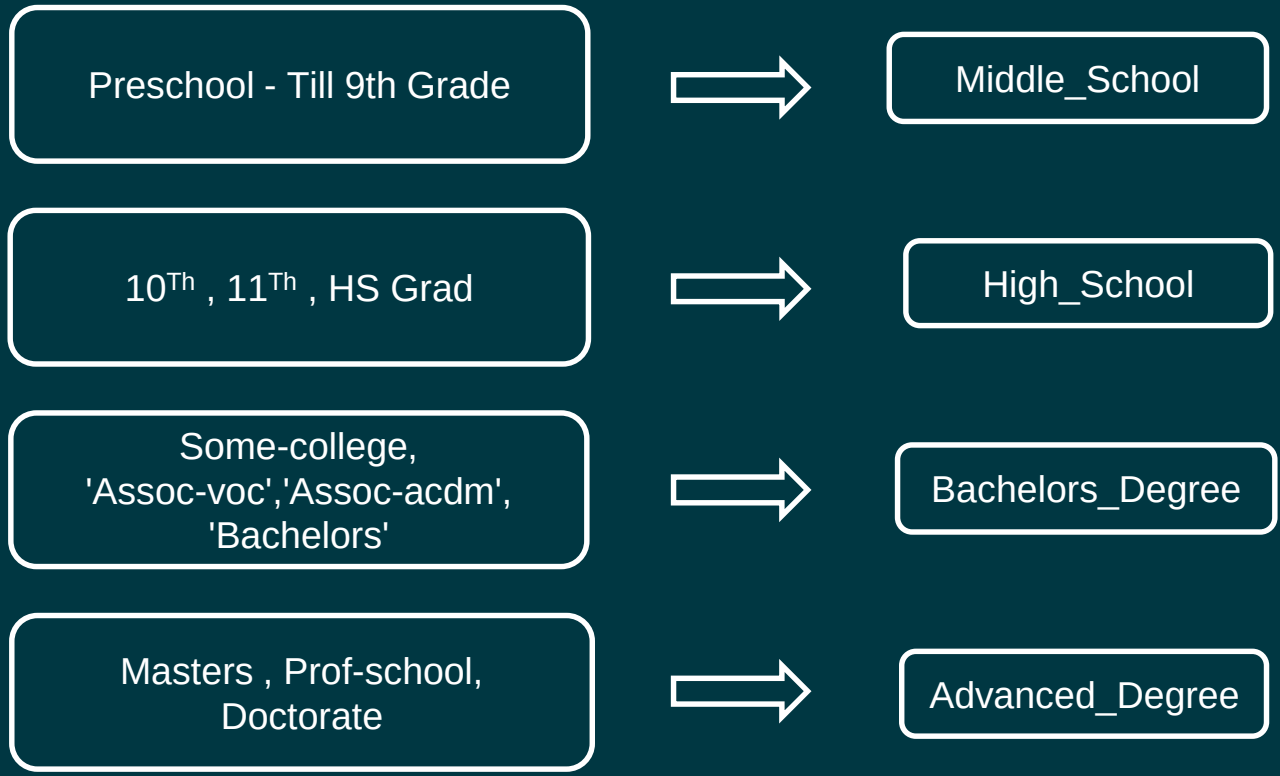


Distribution of workclass and its effect on the likelihood of earning more than $50K per year

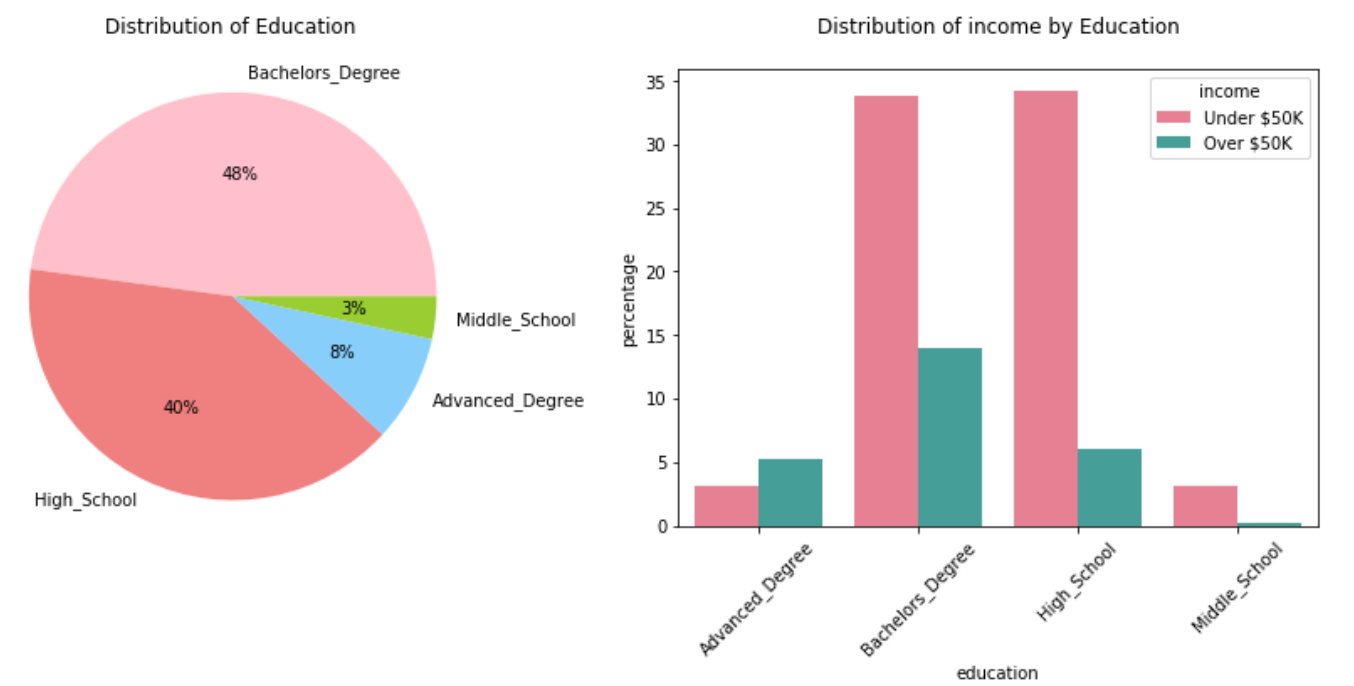


Majority of the individuals work in the private sector. A majority of private sector employees are earning under $50K.

Education

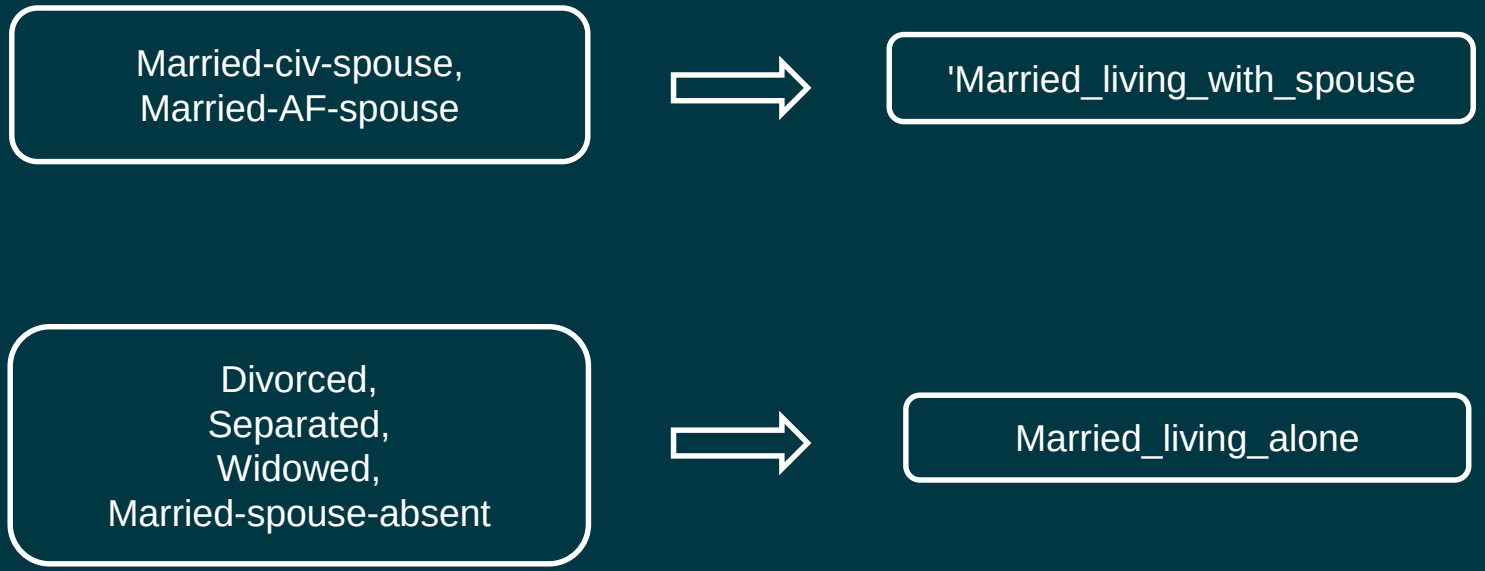

Preschool - Till 9th Grade
Middle_School
10Th , 11Th , HS Grad
High_School
Some-college, 'Assoc-voc','Assoc-acdm', 'Bachelors'
Bachelors_Degree
Masters , Prof-school, Doctorate
Advanced_Degree

Distribution of Education and its effect on the likelihood of earning more than $50K per year

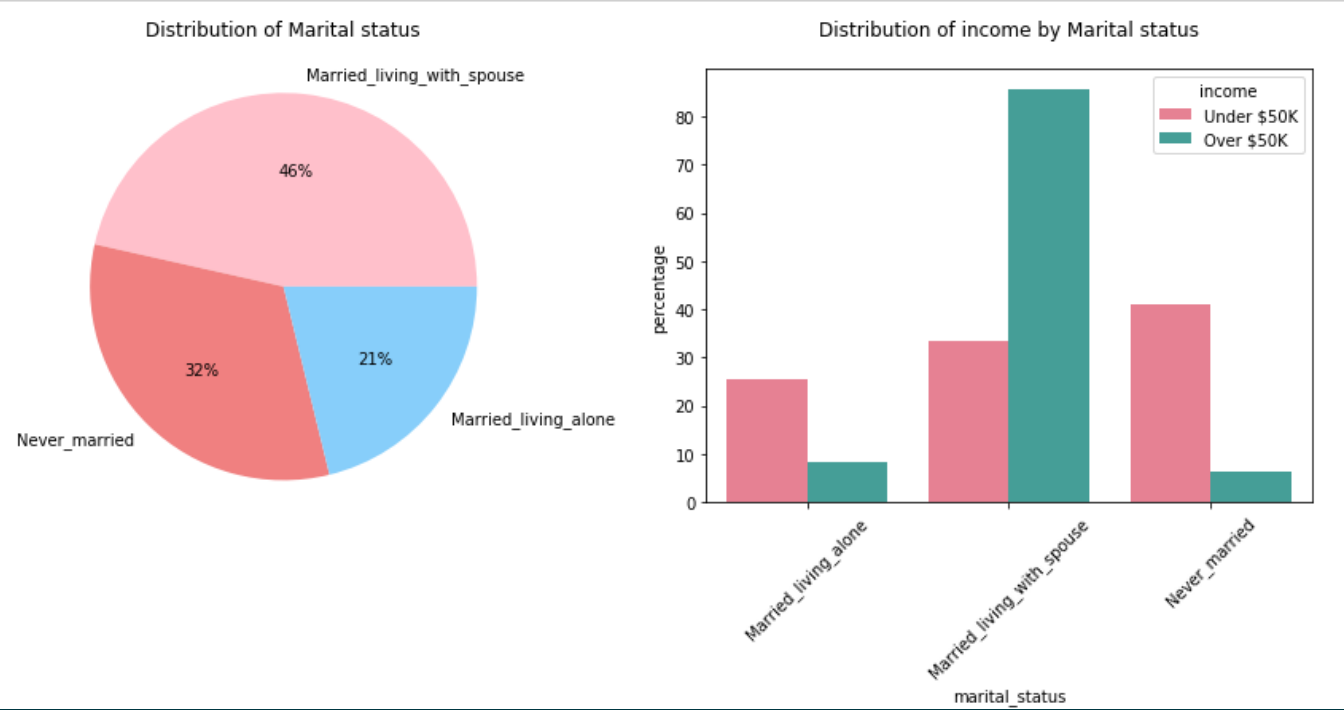


Having an Advance degree increases the chance of earning over $50K.

Marital Status



Distribution of Marital Status and its effect on the likelihood of earning more than $50K per year



Married and living with spouse increases the chance of earning over $50K.

Capital Gain/loss

Modelling

- Bernoulli Naive Bayes
- Logistic Regression
- K Nearest Neighbors Classifier
- Random Forest
- Gradient Boosting Classifier

** under-sampling on majority class
test data size 25%, training data size 75%
5 fold cross validation

Bernoulli Naive Bayes

Accuracy: 78 %

ROC Score: 0.865 (+/-0.02)

Logistic Regression

Default Settings:

Accuracy: 80.5 %

ROC Score: 0.892 (+/-0.01)

Optimization:

Accuracy: 80.7 %

ROC Score: 0.893 (+/-0.02)

Parameters Optimized :

regularization parameter C [0.01, 0.1, 1, 10, 100]
solver algorithm - liblinear
L1 or L2 penalties

Optimal Values:
'C': 10, 'penalty': 'l2', 'solver': 'liblinear'

K Nearest Neighbors Classifier

Default Settings:

Accuracy: 77 %

ROC Score: 0.840 (+/-0.02)

Optimization:

Accuracy: 77 %

ROC Score: 0.861 (+/-0.02)

Parameters Optimized :

Number of Neighbors k = 2 to 20

Optimal Values:
19

Random Forest

Default Settings:

Accuracy: 77 %

ROC Score: 0.851 (+/-0.02)

Optimization:

Accuracy: 80 %

ROC Score: 0.890 (+/-0.01)

Parameters Optimized :

Number of estimators [100,200,500]
Minimum samples split [2, 8]
Maximum depth [4, 6,7, 8, None]
Maximum features ['auto', 'sqrt', 'log2’]

Optimal Values:
'max_depth': 8, 'max_features': 'auto',
'min_samples_split': 8, 'n_estimators': 100

Gradient Boosting Classifier

Default Settings:

Accuracy: 81.7 %

ROC Score: 0.898 (+/-0.01)

Optimization:

Accuracy: 82.1 %

ROC Score: 0.899 (+/-0.01)

Parameters Optimized :

Learning Rate [0.1,0.01,0.005,0.025]
number of estimators [110,440,1000,2000]
maximum depth [5,7]
minimum samples split [800]
minimum samples per leaf [50,60]
number of features considered [11,15]
fraction of observations used to subsample [0.9,0.85]

Optimal Values:
0.005 ,2000, 5, 800 ,
60 , 15, 0.85

Model Comparison

* 5 fold cross validation

	Model	Training Data Accuracy	Test Data Accuracy	ROC AUC Score
4	Gradient_Boost	0.825103	0.821019	0.899372
1	Logistic_Regression	0.816425	0.807237	0.893154
3	Random_Forest	0.811256	0.803025	0.890410
0	Bernoulli NB	0.782351	0.780057	0.864660
2	KNN	0.794346	0.773546	0.860583

The gradient boosting classifier performed better than others in predicting the income class with 82% accuracy and a ROC AUC score of 0.899. The Logistic Regression did farely well with accuracy of 80% and ROC AUC Score of 0.893

Error Analysis - Gradient Boosting

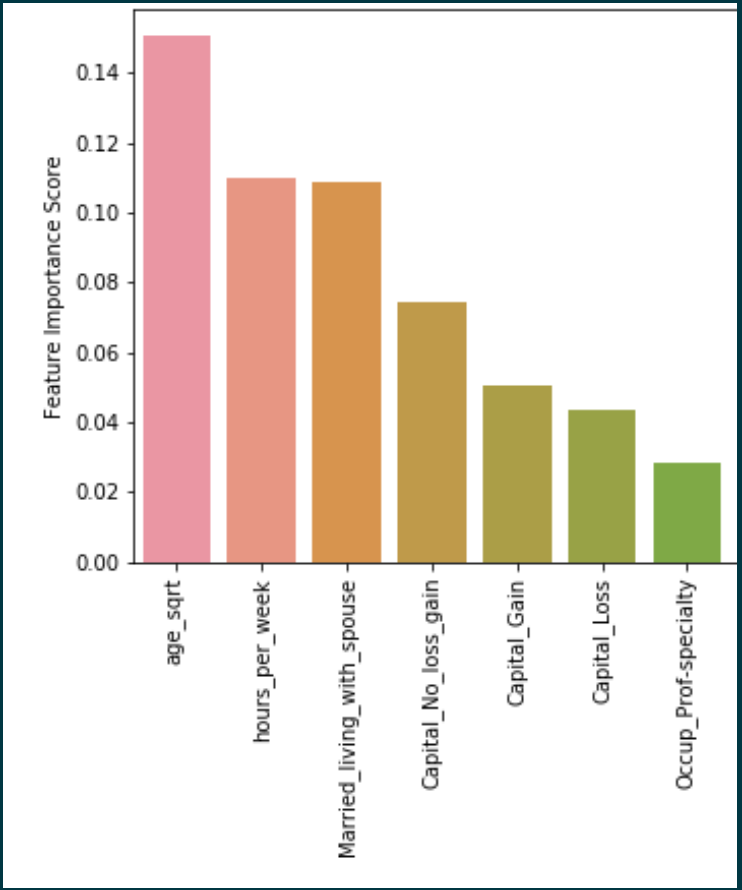
	precision	recall	f1-score	support
Under $50K	0.85	0.79	0.82	2656
Over $50K	0.80	0.86	0.82	2568
avg / total	0.82	0.82	0.82	5224

Precision : precision score was higher for predicting incomes under $50,000

Recall: recall score was higher for predicting incomes over $50,000.

F1-score : Overall, the model was able to identify both classes with 82% accuracy.

Feature Importance



According to the gradient boosting classifier model, the "age of the person", "hours per week", and "Married and living with spouse" were the three most important demographic features to predict if a person's income is over $50,000.

Conclusion

For this capstone project, I wanted to predict if a person's income is over or under $50,000 in a year based on their demographic information. I used the features like age, the number of hours worked per week, their net capital gains, level of education, marital status, occupation, sex, working class, race, and native country.

I applied models like Bernoulli Naive Bayes, Logistic Regression, KNN, Decision Trees, Random Forest Classifier and Gradient Boosting classifier.

The Gradient Boosting classifier performed best on this dataset followed by Logistic Regression. These two models had ROC AUC scores of 0.899372 and 0.893154 respectfully which are nearly equal. As per the Gradient Boosting classifier model the 3 most important features are age, number of hours worked per week, married and living with Spouse.

Opportunities for further exploration

- How do these indicators compare to the income levels of other developed countries?
- How have these indicators changed since 1994?

THANK YOU