XYZ Corporation

Measure -> Interpret -> Act

Exploratory Data Analysis + Income Prediction for UVW College

June 28, 2020

Data Science Team:

Shashi Kiran Chilukuri

James Graves

Yinghai Zhao

Kenneth Han

Problem Statement

- To develop marketing profiles using data supplied by the US Census Bureau with a focus on $50K as a key number for salary.
- To identify the factors that determine whether an individual's income is above or below $50K.
- To predict the whether the income of an individual is above or below $50K.

Dataset Information

Data Source:

- United States Census Bureau

Features:

- 14 features (8 categorical and 6 continuous variables)

Label:

- Income with 2 categories (“>50K”, “<=50K”)

Initial Length of Dataset:

- 32,561 records

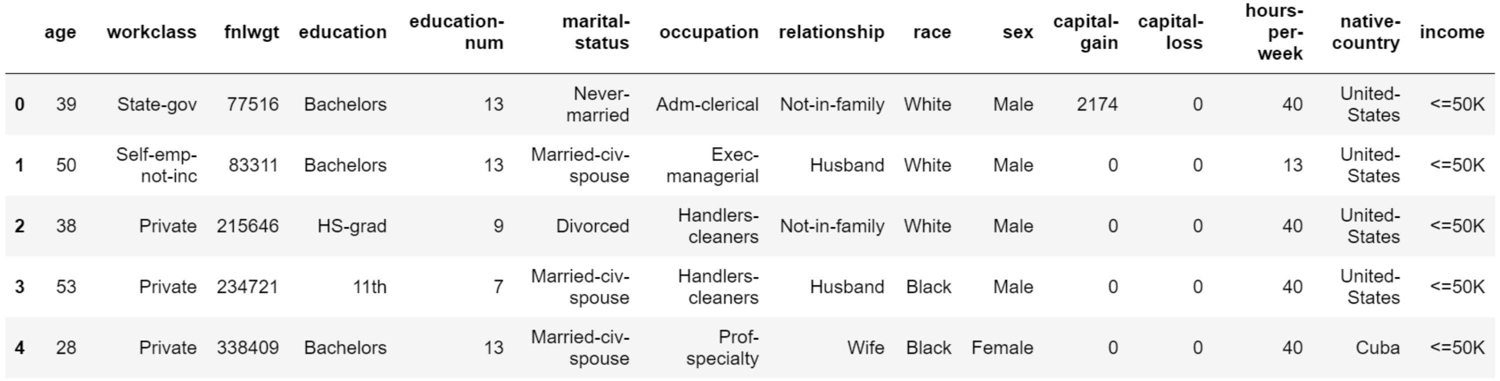
Data Cleaning:

- No Null records
- Removed records with unknow (“?”) data
- Stripped white space in categorical features

Final Length of Dataset:

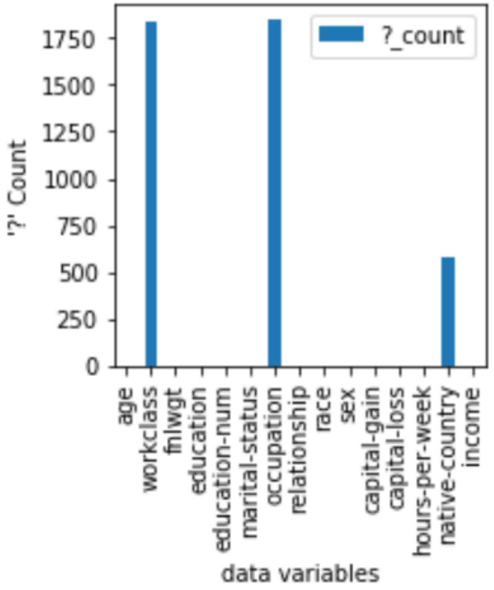
- 30,162 records
- Class distribution
  - <=50K: 22,654
  - >50K: 7,508

Sample Final Dataset after data cleaning:

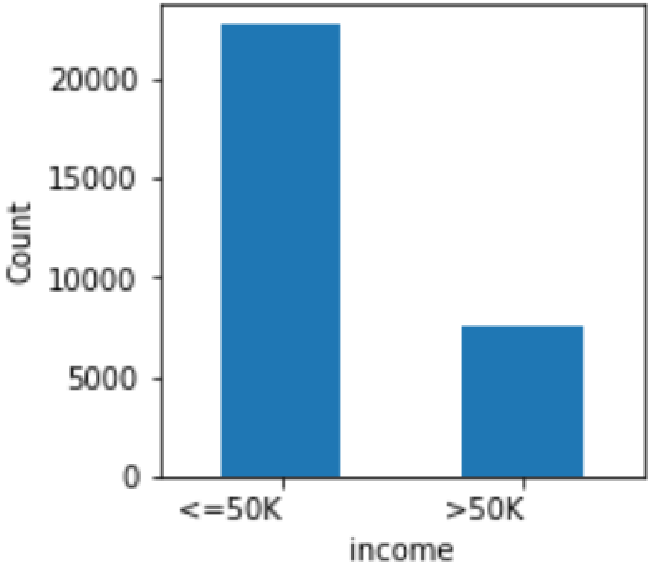


	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

Unknown Data Distribution:



Final Class (‘Income’) Distribution



Exploratory Data Analysis: Data Correlation

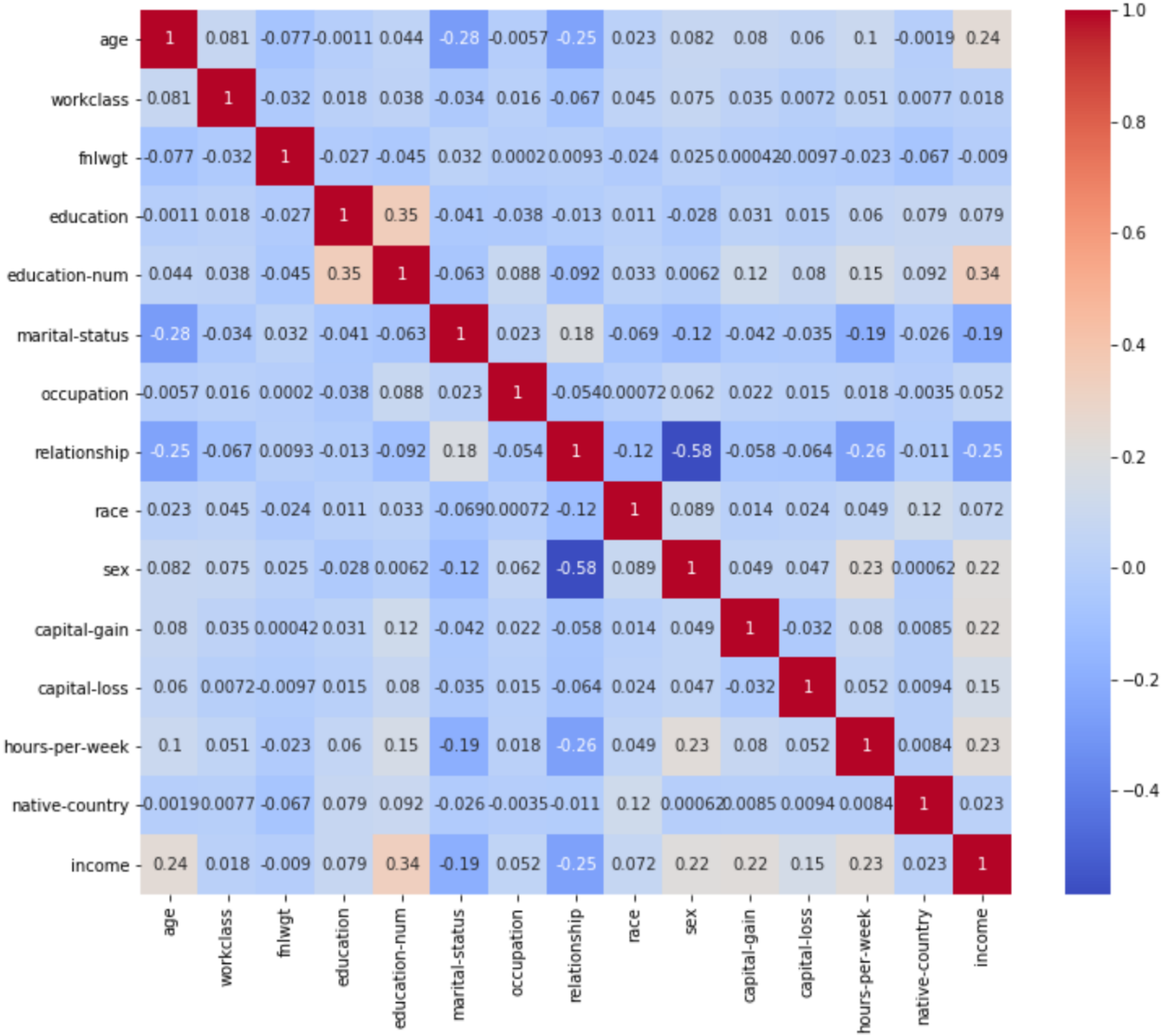
To find correlation between the variables:

- Converted categorical columns into numbers.
- Plotted correlation heatmap between them.

Based on correlation heat map:

- Identified important features that are correlated with class label “Income”.
- Applied correlation >0.20 as a threshold to identify important features that are positively correlated with label “Income”. Here is the list:
  - Education-num
  - Age
  - Sex
  - Capital-gain
  - Hours-per-week
- Applied correlation < -0.18 as a threshold to identify important features that are negatively correlated with label “Income”. Here is the list:
  - Relationship
  - Marital-status

Data Correlation Heat Map:



	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
age	1	0.081	-0.077	-0.0011	0.044	-0.28	-0.0057	-0.25	0.023	0.082	0.08	0.06	0.1	-0.0019	0.24
workclass	0.081	1	-0.032	0.018	0.038	-0.034	0.016	-0.067	0.045	0.075	0.035	0.0072	0.051	0.0077	0.018
fnlwgt	-0.077	-0.032	1	-0.027	-0.045	0.032	0.0002	0.0093	-0.024	0.025	0.00042	-0.0097	-0.023	-0.067	-0.009
education	-0.0011	0.018	-0.027	1	0.35	-0.041	-0.038	-0.013	0.011	-0.028	0.031	0.015	0.06	0.079	0.079
education-num	0.044	0.038	-0.045	0.35	1	-0.063	0.088	-0.092	0.033	0.0062	0.12	0.08	0.15	0.092	0.34
marital-status	-0.28	-0.034	0.032	-0.041	-0.063	1	0.023	0.18	-0.069	-0.12	-0.042	-0.035	-0.19	-0.026	-0.19
occupation	-0.0057	0.016	0.0002	-0.038	0.088	0.023	1	-0.054	0.00072	0.062	0.022	0.015	0.018	-0.0035	0.052
relationship	-0.25	-0.067	0.0093	-0.013	-0.092	0.18	-0.054	1	-0.12	-0.58	-0.058	-0.064	-0.26	-0.011	-0.25
race	0.023	0.045	-0.024	0.011	0.033	-0.069	0.00072	-0.12	1	0.089	0.014	0.024	0.049	0.12	0.072
sex	0.082	0.075	0.025	-0.028	0.0062	-0.12	0.062	-0.58	0.089	1	0.049	0.047	0.23	0.00062	0.22
capital-gain	0.08	0.035	0.00042	0.031	0.12	-0.042	0.022	-0.058	0.014	0.049	1	-0.032	0.08	0.0085	0.22
capital-loss	0.06	0.0072	-0.0097	0.015	0.08	-0.035	0.015	-0.064	0.024	0.047	-0.032	1	0.052	0.0094	0.15
hours-per-week	0.1	0.051	-0.023	0.06	0.15	-0.19	0.018	-0.26	0.049	0.23	0.08	0.052	1	0.0084	0.23
native-country	-0.0019	0.0077	-0.067	0.079	0.092	-0.026	-0.0035	-0.011	0.12	0.00062	0.0085	0.0094	0.0084	1	0.023
income	0.24	0.018	-0.009	0.079	0.34	-0.19	0.052	-0.25	0.072	0.22	0.22	0.15	0.23	0.023	1

Exploratory Data Analysis on Identified Features

Education-num vs Income

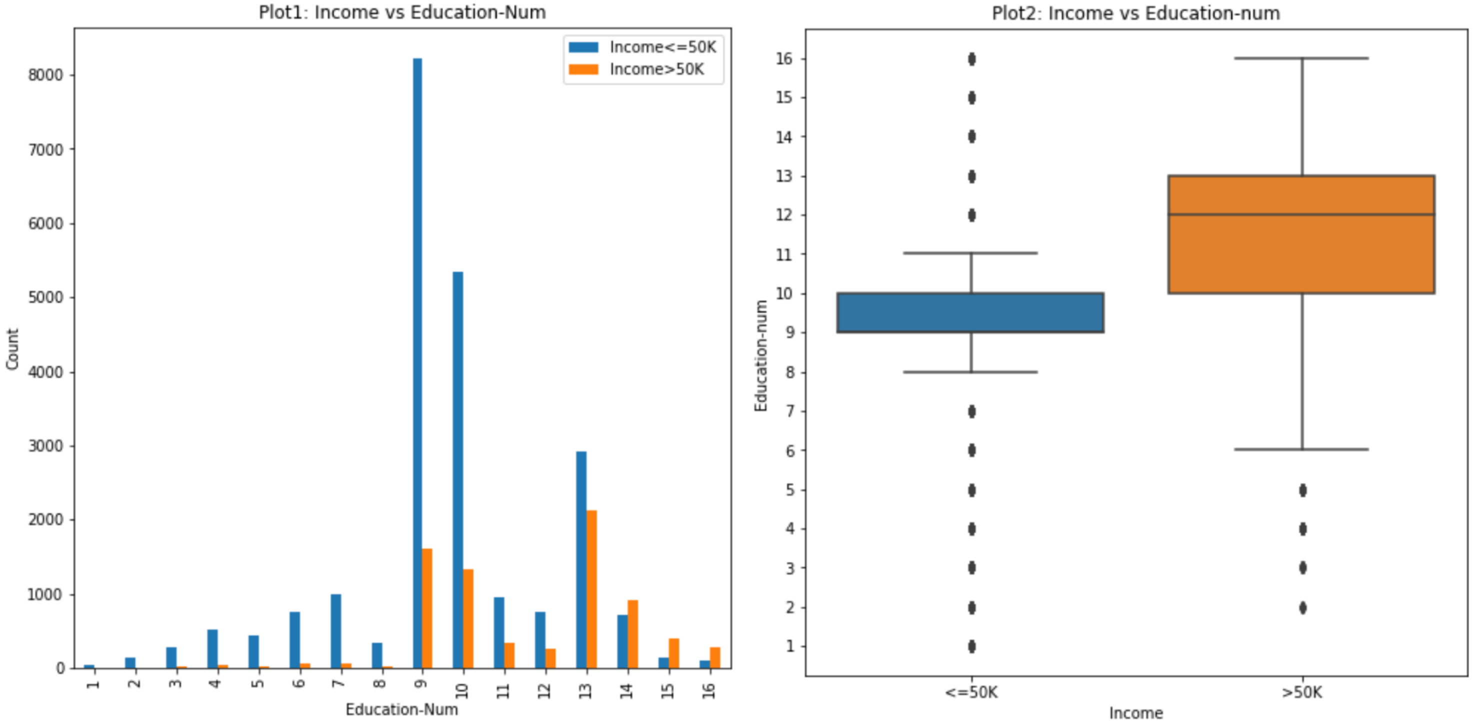
Plot:

- Plot1: Bar plot between "Education-num" and "income"
- Feature "Education-num" on X-axis and Label "income" on Y-axis
- Plt2: Box plot between "income" and "Education-num"
- Label "income" on X-axis and Feature "Education-num" on Y-axis

Inference:

- From Plot1: Individuals with education more than 13 years have higher chances making more than $50k.
- From Plot2: 75% of the group making =< 50K have less than 10th grade education as opposed to only 25% of > 50K income group.
- From Plot2: For the =<50K income group, 12 years of education is considered as an outlier.
- From Plot2: For the >50K income group, roughly 50% have less than 12 years of education.

Plots between Education-num vs Income



Exploratory Data Analysis on Identified Features Cont.

Age vs Income

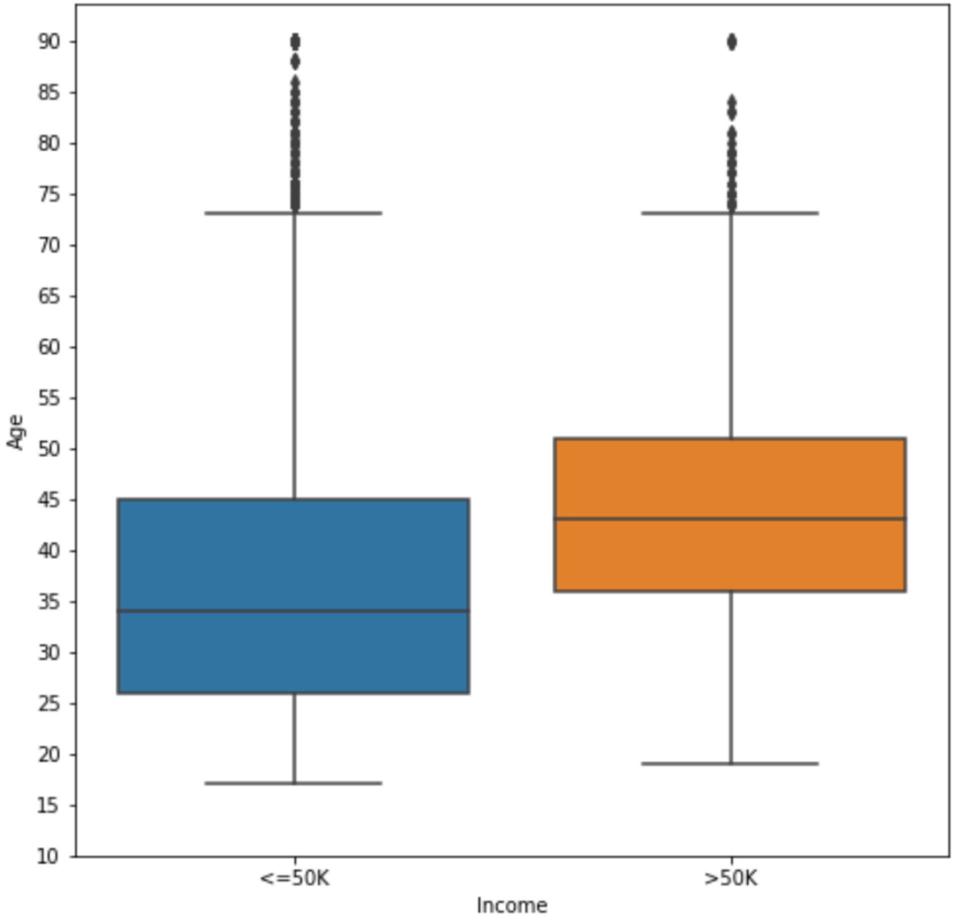
Plot:

- Box plot between "age" and "income".
- Label "income" on X-axis and feature "age" on Y-axis.

Inference:

- Individuals who are making income more than 50K are older than those making less than or equal to 50K at the Quartile1, Quartile2 (median) and Quartile3.
- Those making less than or equal to 50K have a greater Inter Quartile Range (IQR) which means greater group diversity with respect to age.
- Box plot establishes ages exceeding 74 are outliers.

Plot between Age vs Income



Exploratory Data Analysis on Identified Features Cont.

Age vs Capital-gain vs Income

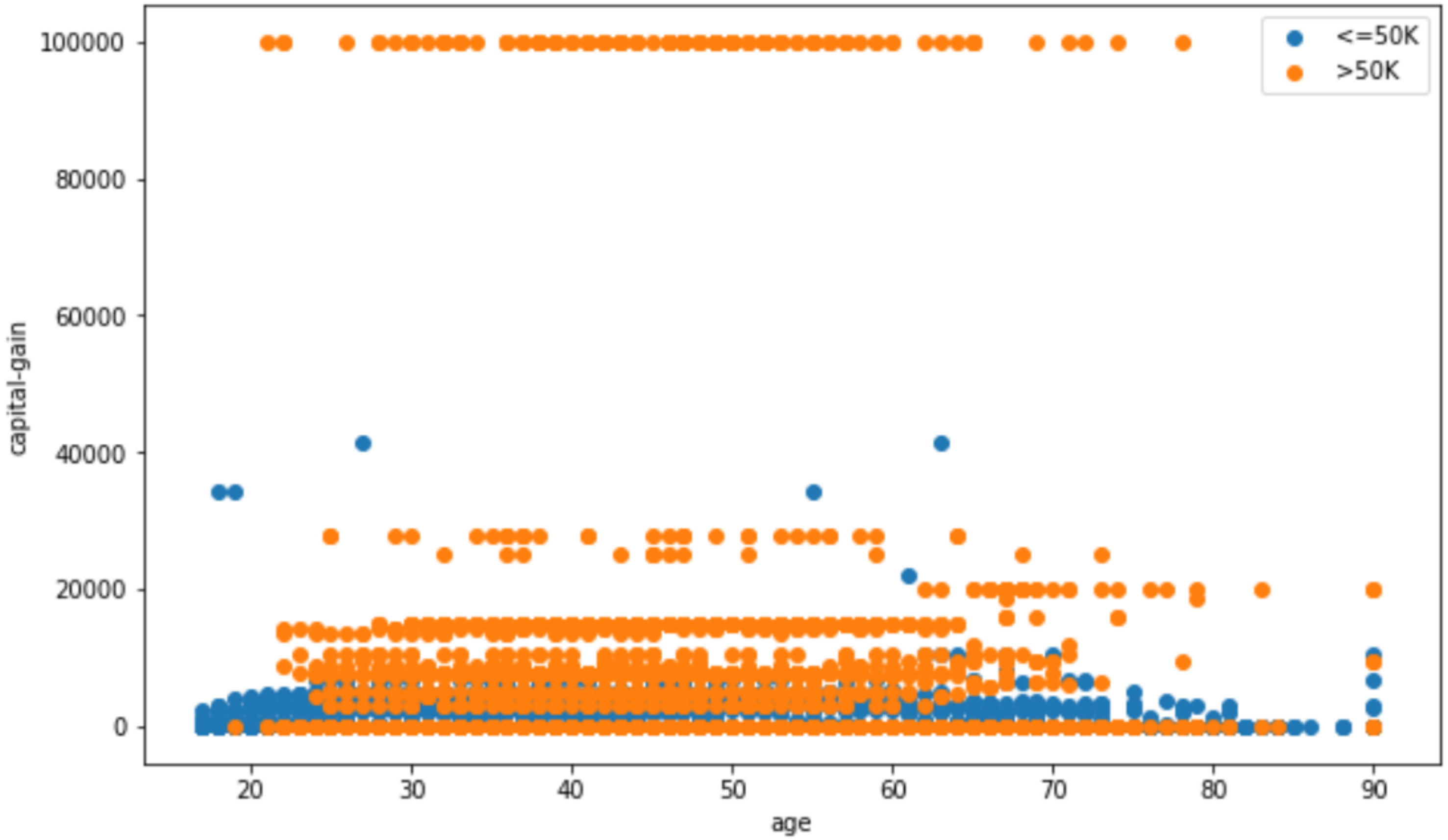
Plot:

- Scatter plot between age, capital-gain and income.
- Features ‘age’ and ‘capital-gain’ on x and y axis respectively.
- Color to categorize label ‘Income’. Blue for income <= 50K and Orange for income >50K.

Inference:

- Individuals with higher capital-gain, have higher probability of having income greater than 50K.
- With increase in Capital-gain, the separation between the two class categories becomes more clear.
- Individuals under age 23 likelier to have lower capital gains and income less than 50K.

Plot between age vs capital-gain vs Income



Exploratory Data Analysis on Identified Features Cont.

Sex vs Income

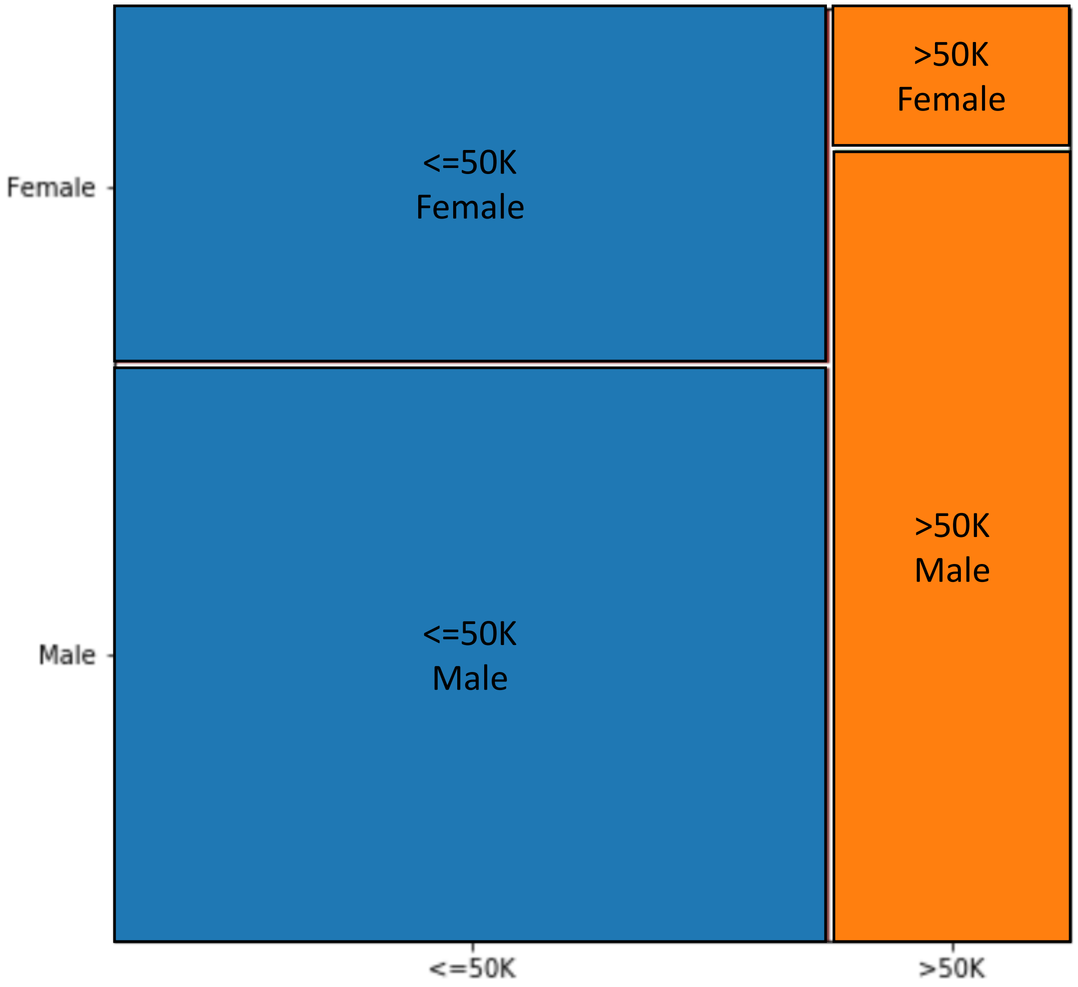
Plot:

- Mosaic plot between sex and income.
- Label 'Income' on X-axis and Feature 'Sex' on Y-axis.

Inference:

- There is a high probability that individual making more than 50K is a Male.

Plot between Sex vs Income



Exploratory Data Analysis on Identified Features Cont.

Hours-per-week vs Income

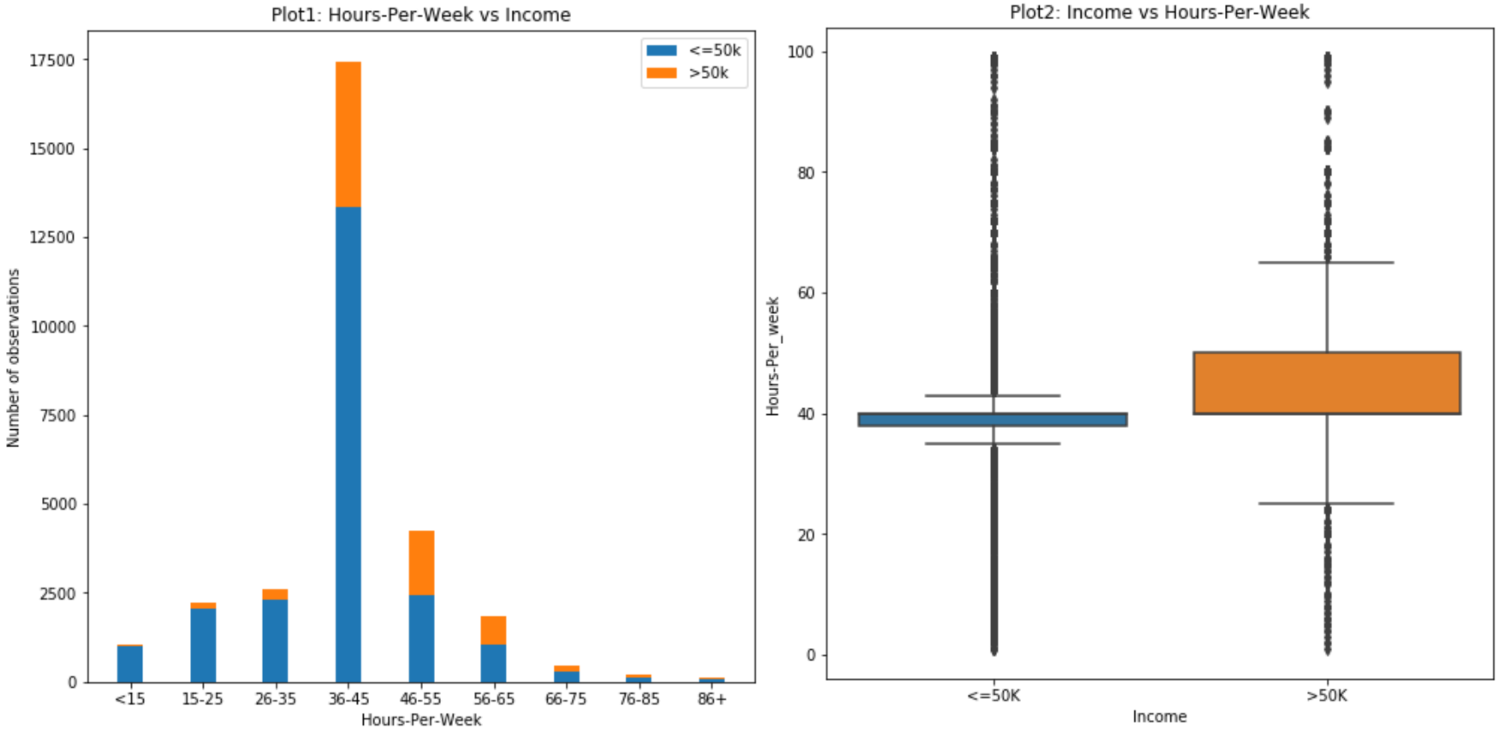
Plot:

- Plot1: Stacked bar plot between "Hours-Per-Week" and "income"
- Feature "Hours-Per-Week" on X-axis and Label "income" on Y-axis
- Plt2: Box plot between "income" and "Hours-Per-Week"
- Label "income" on X-axis and Feature "Hours-Per-Week" on Y-axis

Inference:

- From Plot1: Individuals making more than 50K are likely to be working more than 35 hours per week.
- From Plot2: Inter Quartile range is small for the <=50K group which indicates a high concentration of hours worked near 40 hour work week. In case of over 50K income group, IQR is larger and, Q1, Q2 (median) and Q3 are all higher than the respective values in the <=50K income group.

Plot between Hours-per-week vs Income



Exploratory Data Analysis on Identified Features Cont.

Relationship vs Income

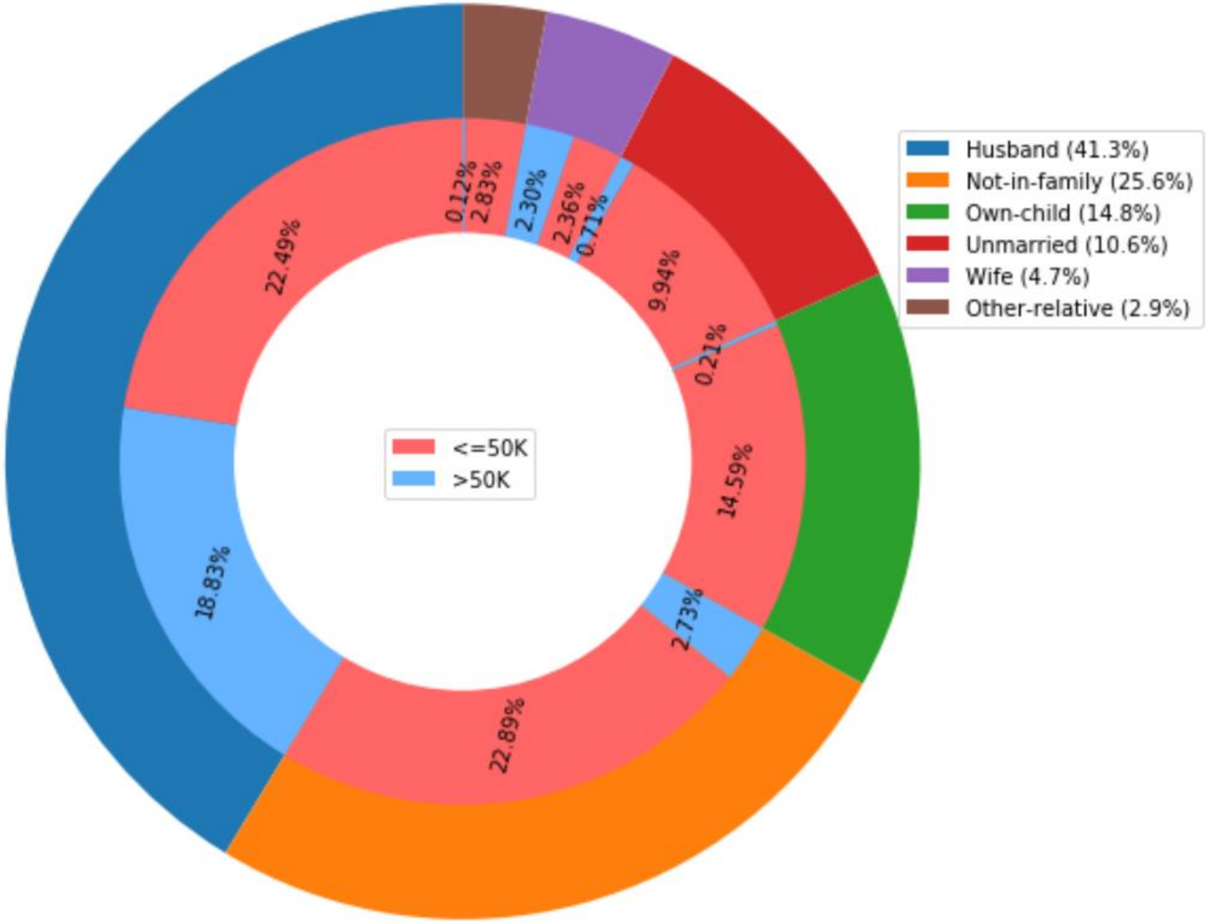
Plot:

- Donut plot of Relationship and also Income (Inner donut plot)

Inference:

- Individual with Relationship category “Husband” or “Wife” has a higher probability of making more than 50K as compared to categories “Not-in-family”, “Own-child”, “Unmarried” and “Other-relative”.

Plot between Relationship vs Income



Exploratory Data Analysis on Identified Features Cont.

Marital-status vs Income

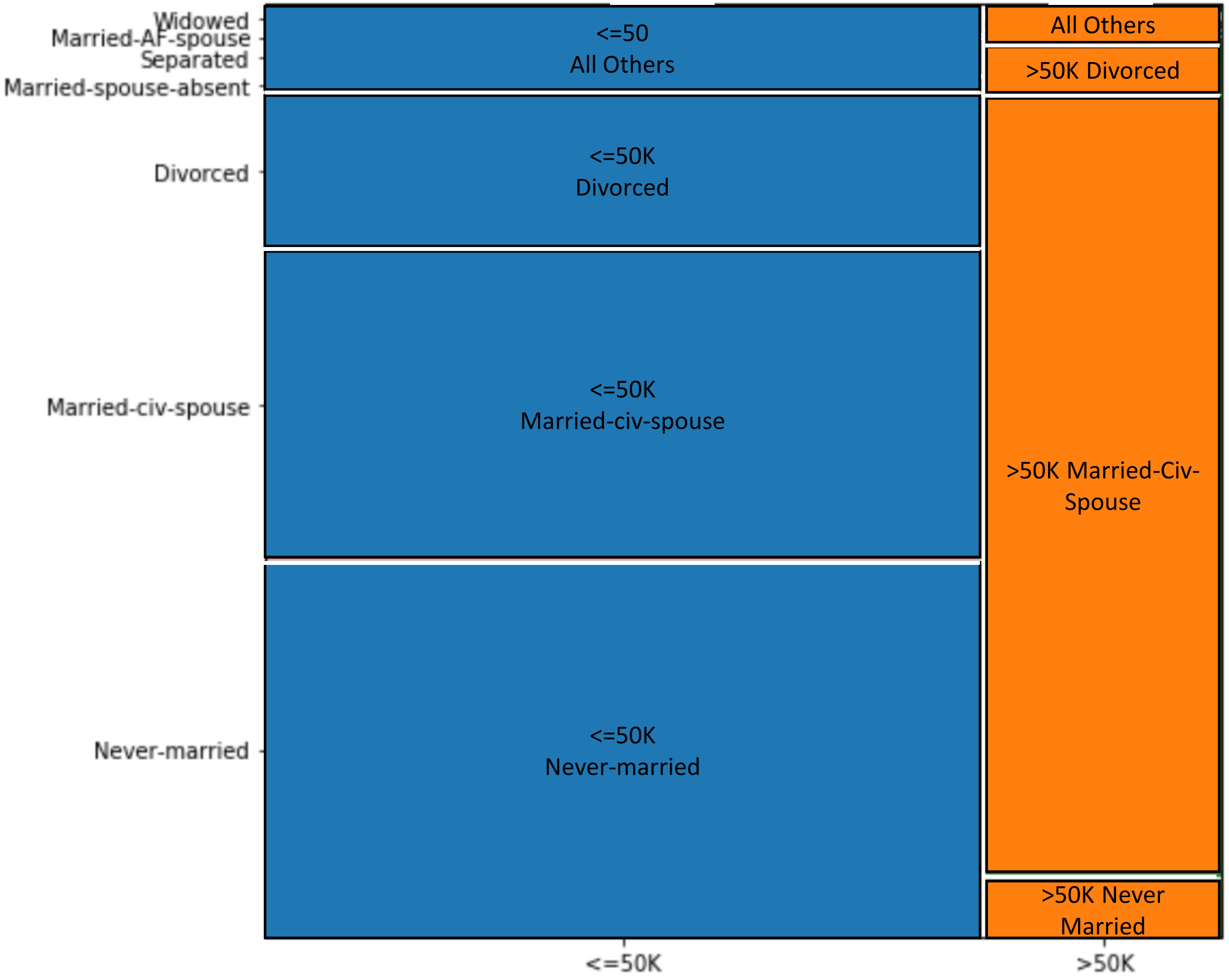
Plot:

- Mosaic plot between marital-status and income.
- Label 'Income' on X-axis and Feature 'marital-status' on Y-axis.

Inference:

- Individual making more than 50K very likely to have marital-status "Married-civ-spouse".

Plot between Marital-status vs Income



Interesting Inferences from Analysis

- For the <=50K income group, 12 years of education is considered as an outlier.
- The median ages for the <=50K income group is 34 whereas for the >50K income group it is 44.
- There is a high probability that individual making more than 50K is a Male.
- Individuals making more than 50K are likely to be working more than 35 hours per week.
- Individual with Relationship category “Husband” or “Wife” has a higher probability of making more than 50K as compared to categories “Not-in-family”, “Own-child”, “Unmarried” and “Other-relative”.
- Individual making more than 50K very likely to have marital-status “Married-civ-spouse”.

Machine Learning: Data Pre-processing

Separating Dataset:

- Separating Dataset into Features (with 14 variables) and Label “Income”.

Converting Categorical variables:

- Identified 8 categorical features and converted them into 1’s and 0’s resulting in a total of 96 different features.

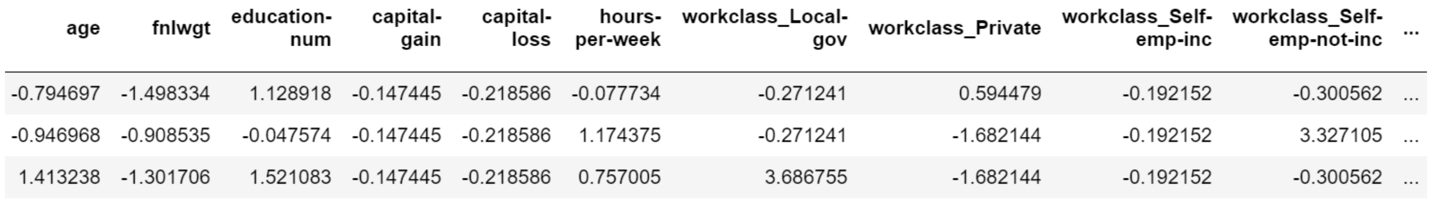
Feature Scaling:

- For the model to give equal importance to all the features, normalized the data.

Splitting Data into Train and Test:

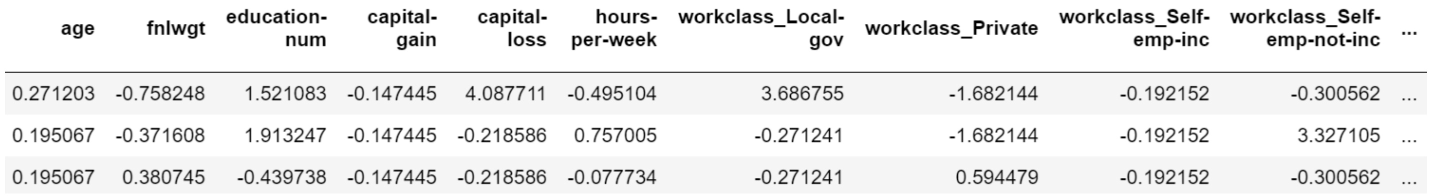
- Split the data into 70/30 ratio for model training and testing. As a result, the following train and test datasets were created:
  - Features Train set
  - Features Test set
  - Label Train set
  - Label Test set

Features Train set: 21113 rows x 96 columns



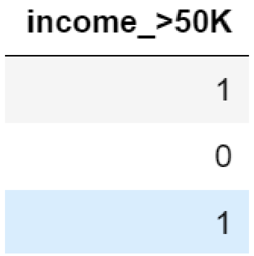
age	fnlwgt	education-num	capital-gain	capital-loss	hours-per-week	workclass_Local-gov	workclass_Private	workclass_Self-emp-inc	workclass_Self-emp-not-inc	...
-0.794697	-1.498334	1.128918	-0.147445	-0.218586	-0.077734	-0.271241	0.594479	-0.192152	-0.300562	...
-0.946968	-0.908535	-0.047574	-0.147445	-0.218586	1.174375	-0.271241	-1.682144	-0.192152	3.327105	...
1.413238	-1.301706	1.521083	-0.147445	-0.218586	0.757005	3.686755	-1.682144	-0.192152	-0.300562	...

Features Test set: 9039 rows x 96 columns



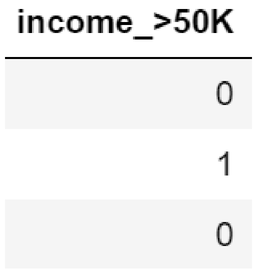
age	fnlwgt	education-num	capital-gain	capital-loss	hours-per-week	workclass_Local-gov	workclass_Private	workclass_Self-emp-inc	workclass_Self-emp-not-inc	...
0.271203	-0.758248	1.521083	-0.147445	4.087711	-0.495104	3.686755	-1.682144	-0.192152	-0.300562	...
0.195067	-0.371608	1.913247	-0.147445	-0.218586	0.757005	-0.271241	-1.682144	-0.192152	3.327105	...
0.195067	0.380745	-0.439738	-0.147445	-0.218586	-0.077734	-0.271241	0.594479	-0.192152	-0.300562	...

Label Train set:
21113 rows x 1 column



income_>50K
1
0
1

Label Test set:
21113 rows x 1 column



income_>50K
0
1
0

Machine Learning: Logistic Regression Implementation

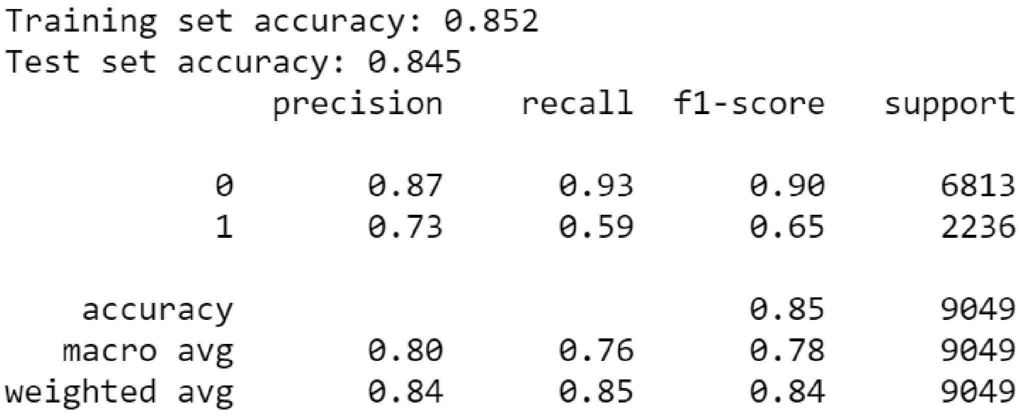
Model:

- Model created: Logistic Regression
- This model will classify given features into two class categories “Income more than 50K” and “Income less than or equal to 50K”

Model Results:

- Logistic Regression Model classified the test data with 84.5 % accuracy.
- Important features suggested by this model are
  - Capital-gain
  - Martial-status with “Married-civ-spouse”
  - Education-num
  - Sex “Male”
  - Age

Logistic Regression Model Results

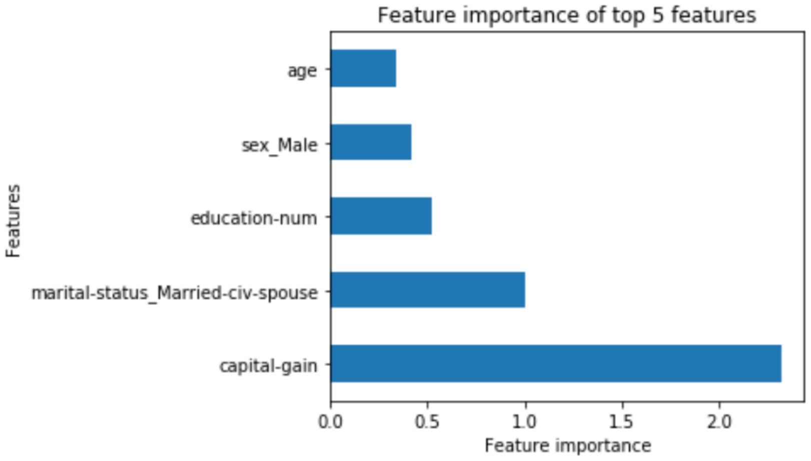


```
Training set accuracy: 0.852
Test set accuracy: 0.845
              precision    recall  f1-score   support

           0       0.87      0.93      0.90      6813
           1       0.73      0.59      0.65      2236

    accuracy                           0.85      9049
   macro avg       0.80      0.76      0.78      9049
weighted avg       0.84      0.85      0.84      9049
```

Logistic Regression Model’s Top 5 Features



Machine Learning: Decision Tree Implementation

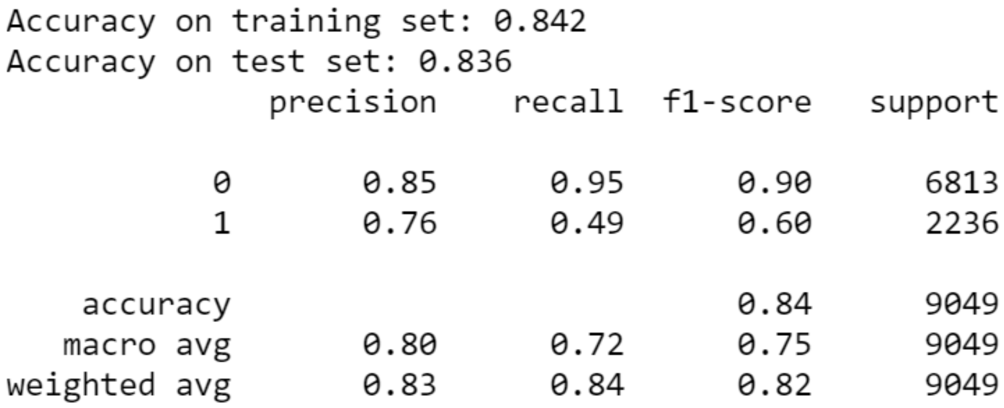
Model:

- Model created: Decision Tree
- This model will classify given features into two class categories “Income more than 50K” and “Income less than or equal to 50K”

Model Results:

- Decision Tree Model classified the test data with 83.6 % accuracy.
- Important features suggested by this model are
  - Martial-status with “Married-civ-spouse”
  - Education-num
  - Capital-gain
  - Age
  - Capital-loss

Decision Tree Model Results

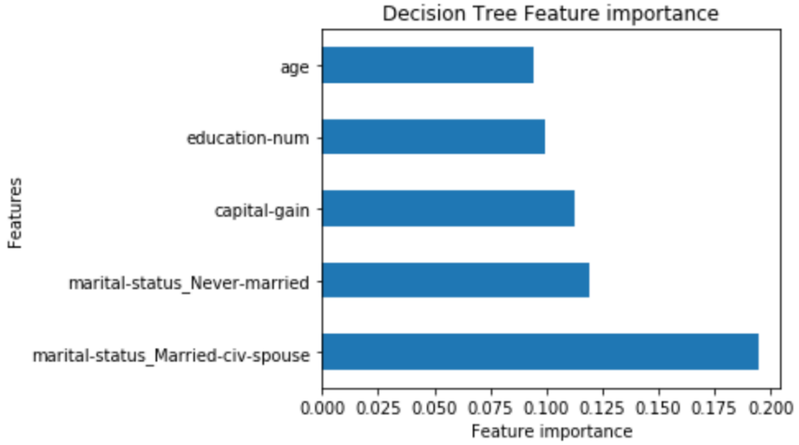


```
Accuracy on training set: 0.842
Accuracy on test set: 0.836
              precision    recall  f1-score   support

           0       0.85      0.95      0.90      6813
           1       0.76      0.49      0.60      2236

    accuracy                           0.84      9049
   macro avg       0.80      0.72      0.75      9049
weighted avg       0.83      0.84      0.82      9049
```

Decision Tree Model’s Top 5 Features



Machine Learning: Random Forest Implementation

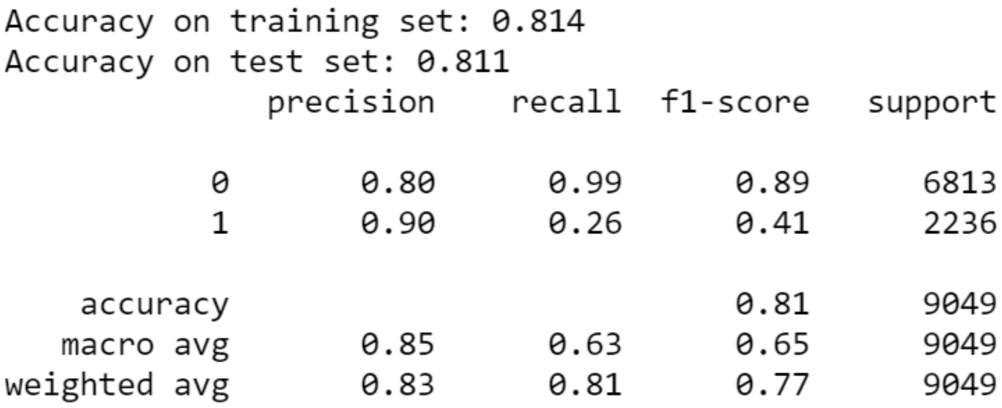
Model:

- Model created: Random Forest
- This model will classify given features into two class categories “Income more than 50K” and “Income less than or equal to 50K”

Model Results:

- Random Forest Model classified the test data with 81.1 % accuracy.
- Important features suggested by this model are
  - Martial-status with “Married-civ-spouse”
  - Capital-gain
  - Martial-status with “Never-Married”
  - Education-num
  - Age

Random Forest Model Results

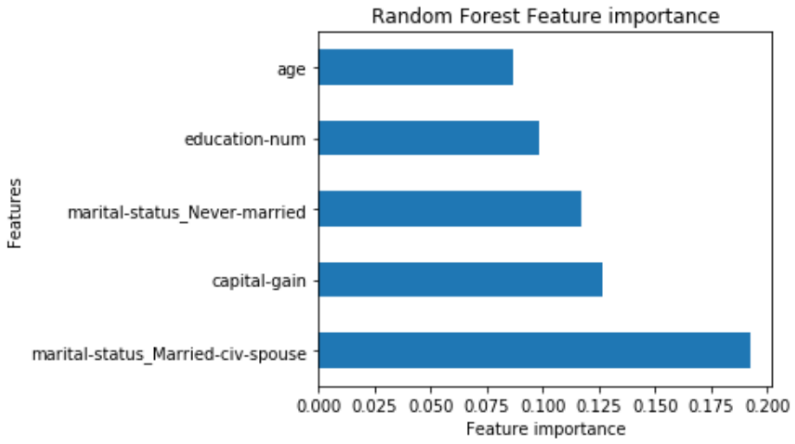


```
Accuracy on training set: 0.814
Accuracy on test set: 0.811
              precision    recall  f1-score   support

           0       0.80      0.99      0.89      6813
           1       0.90      0.26      0.41      2236

    accuracy                           0.81      9049
   macro avg       0.85      0.63      0.65      9049
weighted avg       0.83      0.81      0.77      9049
```

Random Forest Model’s Top 5 Features



Machine Learning: SVM Implementation

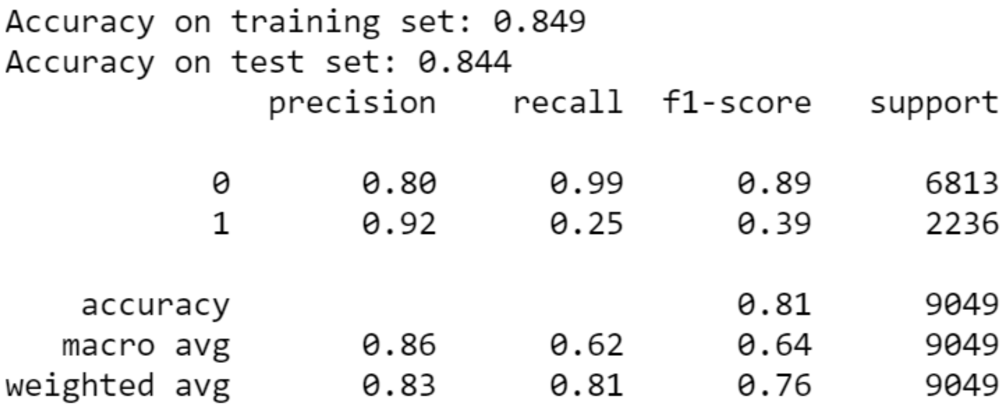
Model:

- Model created: Support Vector Machine with Linear kernel
- This model will classify given features into two class categories “Income more than 50K” and “Income less than or equal to 50K”

Model Results:

- SVM Model classified the test data with 84.4 % accuracy.
- Important features suggested by this model are
  - Capital-gain
  - Martial-status with “Married-civ-spouse”
  - Education-num
  - Occupation with “Exec-Managerial”
  - Education with “Bachelors”

SVM model Results

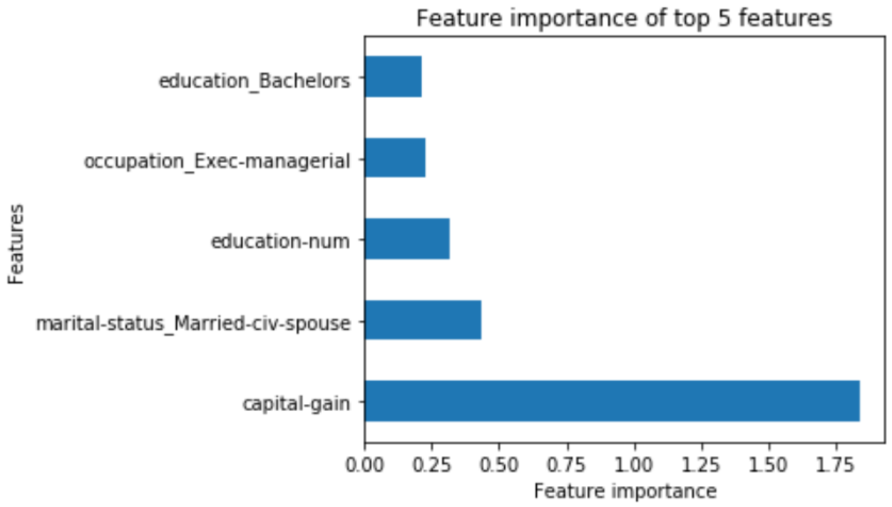


```
Accuracy on training set: 0.849
Accuracy on test set: 0.844
              precision    recall  f1-score   support

           0       0.80      0.99      0.89      6813
           1       0.92      0.25      0.39      2236

    accuracy                           0.81      9049
   macro avg       0.86      0.62      0.64      9049
weighted avg       0.83      0.81      0.76      9049
```

SVM Model’s Top 5 Features



Machine Learning: Models Evaluation

Models Created:

- Logistic Regression, Decision Tree, Random Forest, Support Vector Machine with Linear kernel.

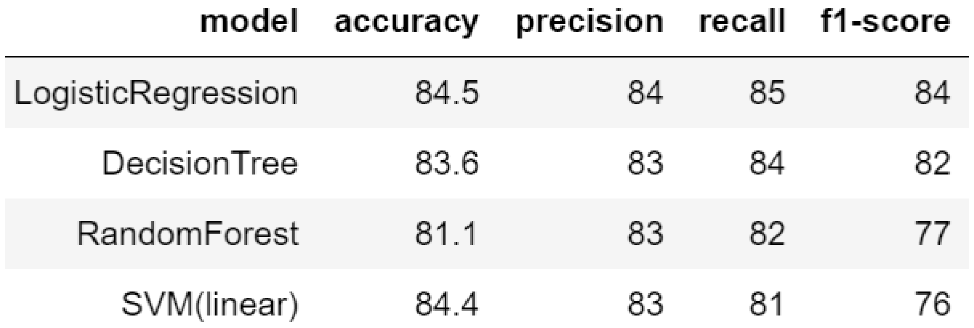
Model Results:

- All the models showed similar accuracy with Logistic Regression having a slight edge in predicting the results.

Feature Importance:

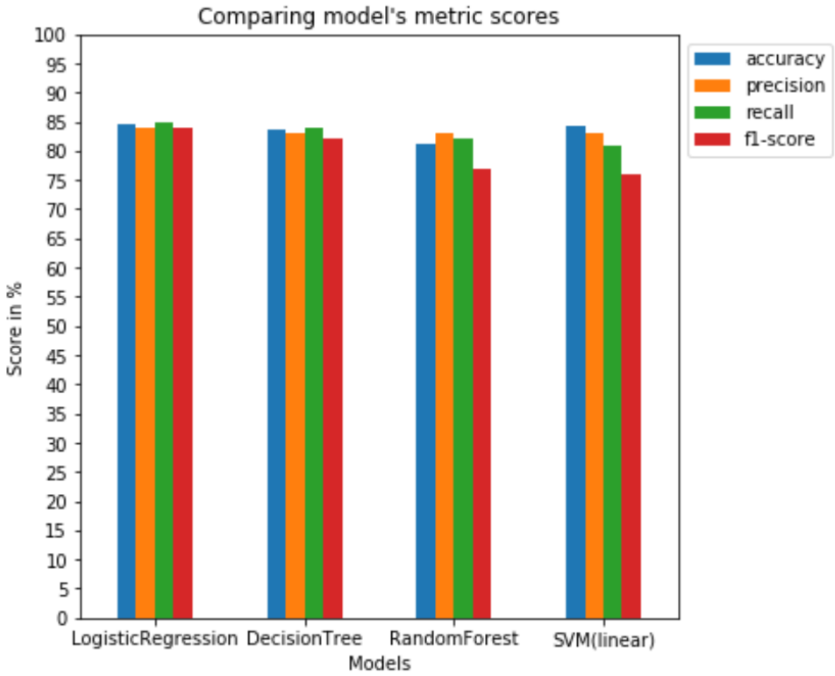
- Top features suggested by the 4 models are similar with a couple of exceptions. Here is the list of features common to all models:
  - Martial-status with “Married-civ-spouse”
  - Education-num
  - Capital-gain
- Top features suggested by the 4 models are quite similar to those obtained through Exploratory Data analysis.

Models Evaluation Matrix



model	accuracy	precision	recall	f1-score
LogisticRegression	84.5	84	85	84
DecisionTree	83.6	83	84	82
RandomForest	81.1	83	82	77
SVM(linear)	84.4	83	81	76

Plot Comparing Machine Learning Model Used



Machine Learning Model Recommendations

- Based on our study of 4 popular machine learning models, we recommend the logistic regression model for predicting whether the income is above or below 50K. Here is the ranked list of the important features suggested by the model:
  1. Capital-gain
  2. Martial-status “Married-civ-spouse”
  3. Education-num
  4. Sex “Male”
  5. Age
- Based on our analysis, the following features were found to be significant across all the machine learning models and in our exploratory data analysis:
  - Martial-status “Married-civ-spouse”
  - Education-num
  - Capital-gain

Questions ?