Predicting >50K and <=50K Income Class

Springboard Data Science - Capstone Project 1 - Final Project Report

Project Goals

This capstone project is based on a dataset called ‘Adult Data Set’ from UCI’s Machine Learning Repository.

This project aims to solve following problems:

1. Predict whether a person’s income exceed $50K/yr based on the census data
2. Identify segments which may exist in the population based on the data collected

Target Applications of The Exercise

Census data is almost like a Gold Mine for Data Scientists, Marketers, Advertisers

- True population data not just based on a sample of a population
- Income predictor can be used for targeting specific customer segments
- Features identified as having impact on income can be used for segmenting
- Who can use this data - marketing (traditional, search, social), advertising

Adult Data Set

- The dataset is based on the 1994 US Census Database
- Prediction task is to determine whether a person makes over 50K a year
- Attributes: Income, Age, Workclass, Education, Marital Status, Occupation, Relationship, Race, Sex, Capital Gain, Capital Loss, Hours-per-Week, Native-Country
- Total Records: 32,561 Entries, 15 Columns

Key Sections Ahead

- Data Wrangling
- Data Story
- In-Depth Analysis using Statistics
- In-Depth Analysis using Machine Learning
- Key Findings
- Closing Thoughts, Conclusion

Data Wrangling

Data Wrangling

- Initial data conversion: .data → .txt → .csv
- Data Type: 6 Features - int64, 9 Features - Object

```
adult.head()
```

	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

Data Wrangling

Missing Values?

- Native-Country, Workclass and Occupation has ? as one of the labels
- ? later replaced with ‘unknown-’ label.
- High Counts, hence kept them as separate labels and not add to any other labels.

Outliers

- Capital Gain feature has a record with value of $100,000, however this is not treated as an outlier as it is a possible value

Data Story

Data Story

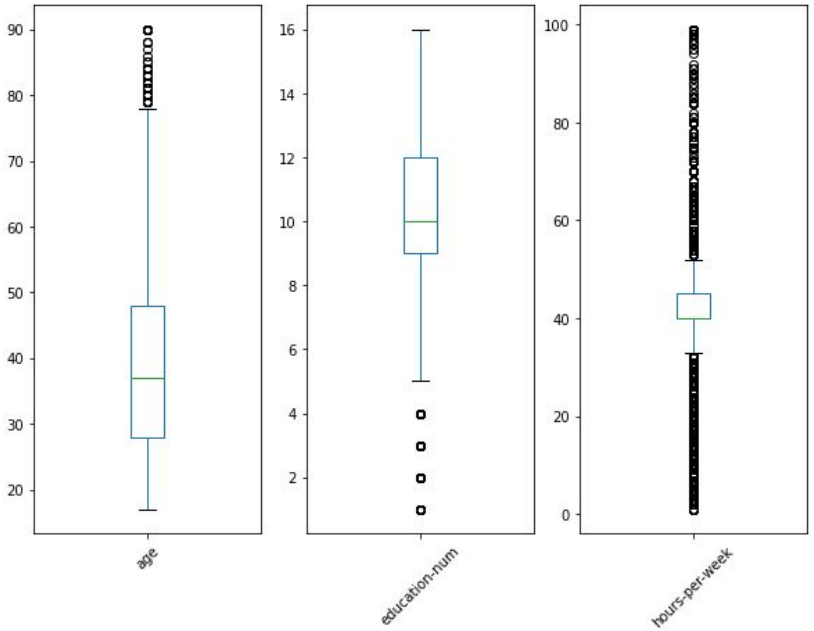
```
adult.describe()
```

	age	fnlwgt	education-num	capital-gain	capital-loss	hours-per-week
count	32561.000000	3.256100e+04	32561.000000	32561.000000	32561.000000	32561.000000
mean	38.581647	1.897784e+05	10.080679	1077.648844	87.303830	40.437456
std	13.640433	1.055500e+05	2.572720	7385.292085	402.960219	12.347429
min	17.000000	1.228500e+04	1.000000	0.000000	0.000000	1.000000
25%	28.000000	1.178270e+05	9.000000	0.000000	0.000000	40.000000
50%	37.000000	1.783560e+05	10.000000	0.000000	0.000000	40.000000
75%	48.000000	2.370510e+05	12.000000	0.000000	0.000000	45.000000
max	90.000000	1.484705e+06	16.000000	99999.000000	4356.000000	99.000000

Data Story

```
adult3.plot(kind='box', figsize=(10,7), rot=45, subplots=True)
```

```
age                AxesSubplot(0.125,0.125;0.227941x0.755)
education-num      AxesSubplot(0.398529,0.125;0.227941x0.755)
hours-per-week     AxesSubplot(0.672059,0.125;0.227941x0.755)
dtype: object
```



Data Story

Further Data Transformations

- Non-Integer Attributes from Object → Category
- Income Class: <= 50K, >50K → 0, 1
- Dropped Columns - Fnlwgt, Capital-Gain, Capital-Loss

```
replace_map = {'income':{'<=50K' : 0, '>50K' : 1}}
adult4.replace(replace_map, inplace=True, regex=True)
adult4.head()
```

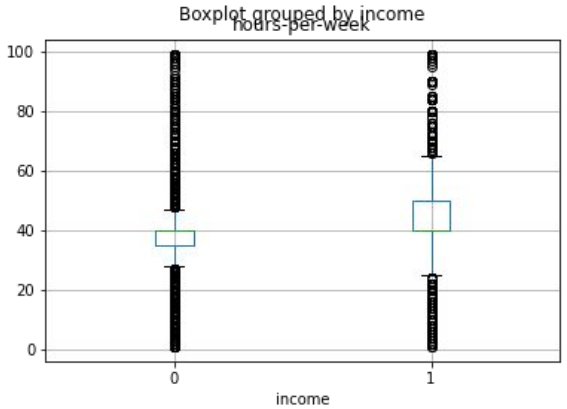
	age	workclass	education	education-num	marital-status	occupation	relationship	race	sex	hours-per-week	native-country	income
0	39	State-gov	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	40	United-States	0
1	50	Self-emp-not-inc	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White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

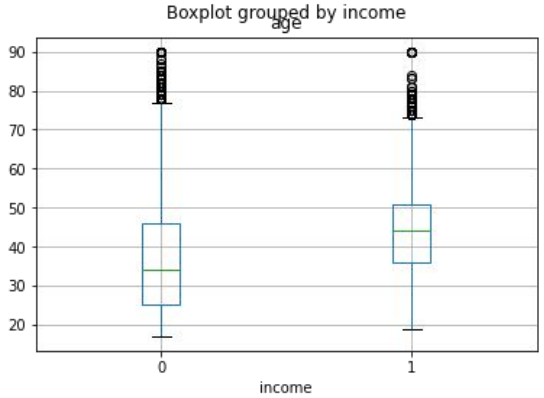
Data Story

- Total sample = 32,561
- Majority, 24,720 people (76%) falls in the low income, <=50K group
- Only a small population - 7841 people (24%) falls in high income, >50K group

Data Story

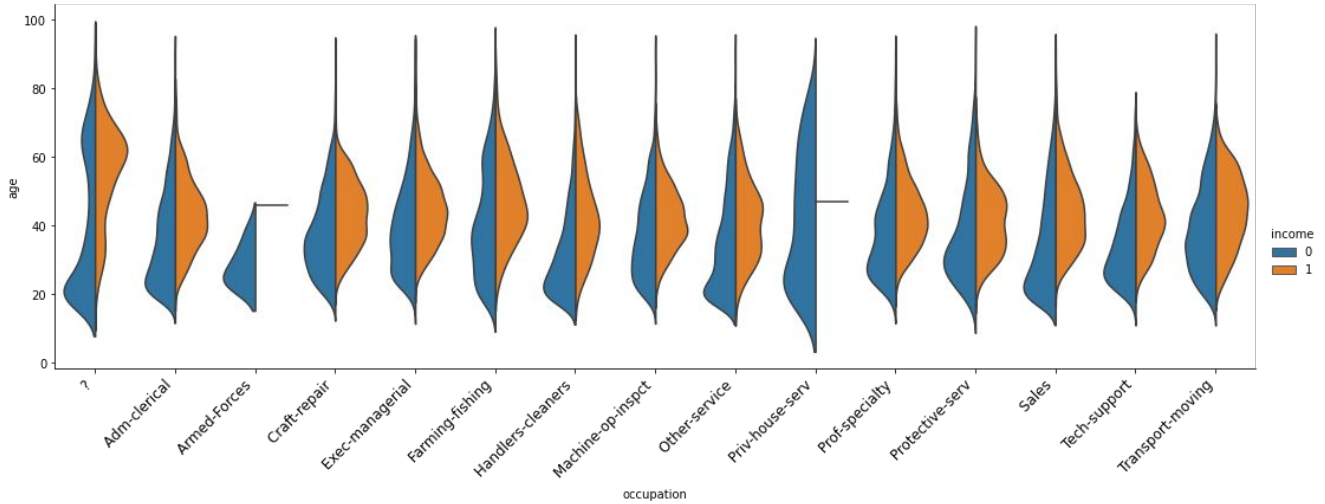
Low Income (<=50K) group and High Income (>50K) group clearly differ from each other





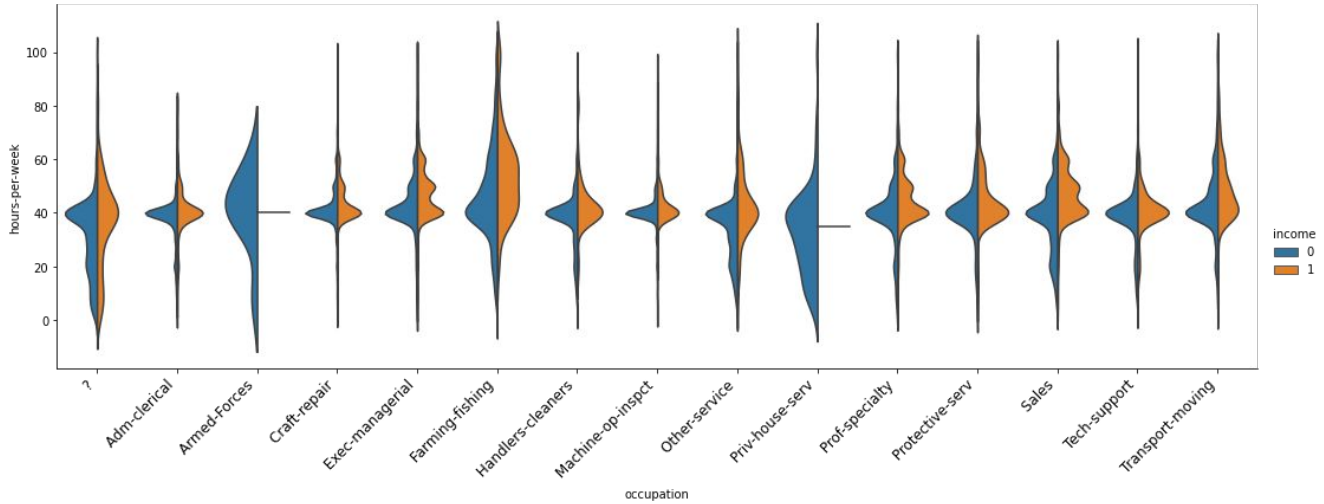
Data Story

The overall trend of age difference continues across various occupation categories as well



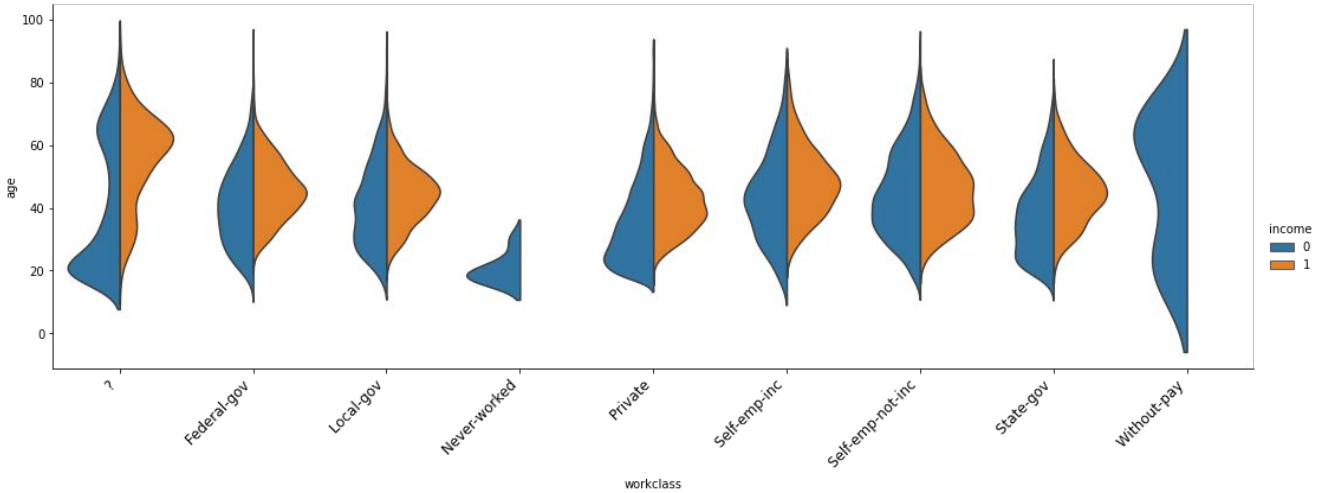
Data Story

The overall trend of 'hours worked per week' difference continues across various occupation categories as well



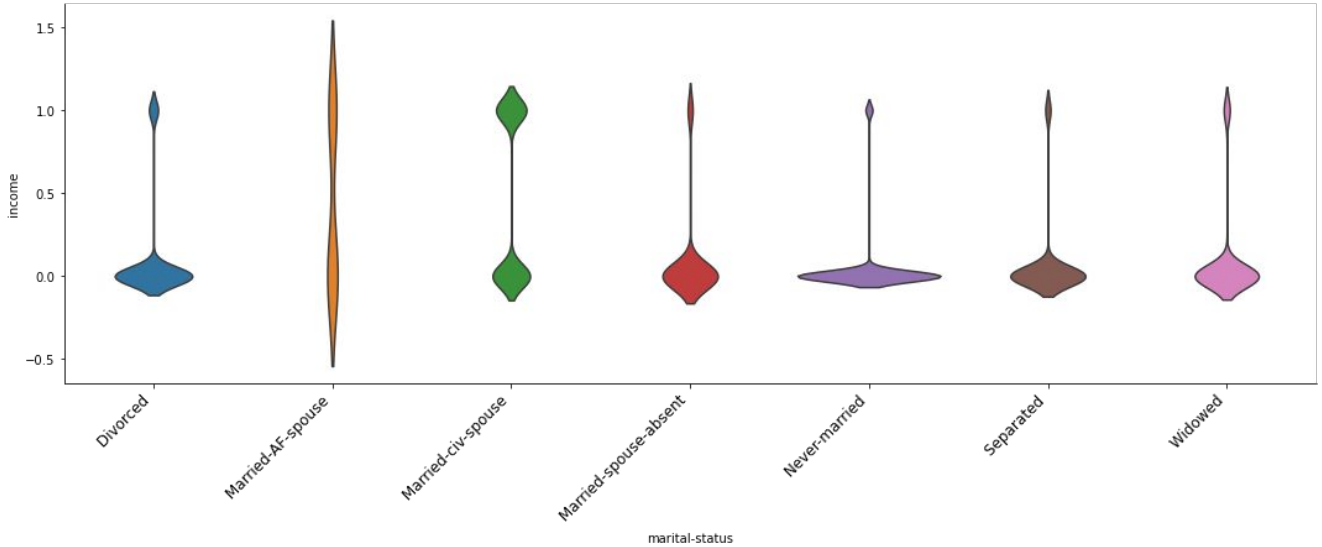
Data Story

Lower income group is more concentrated between 20 to 40 years age block vs. higher income group which is more concentrated between 40 to 60 age block



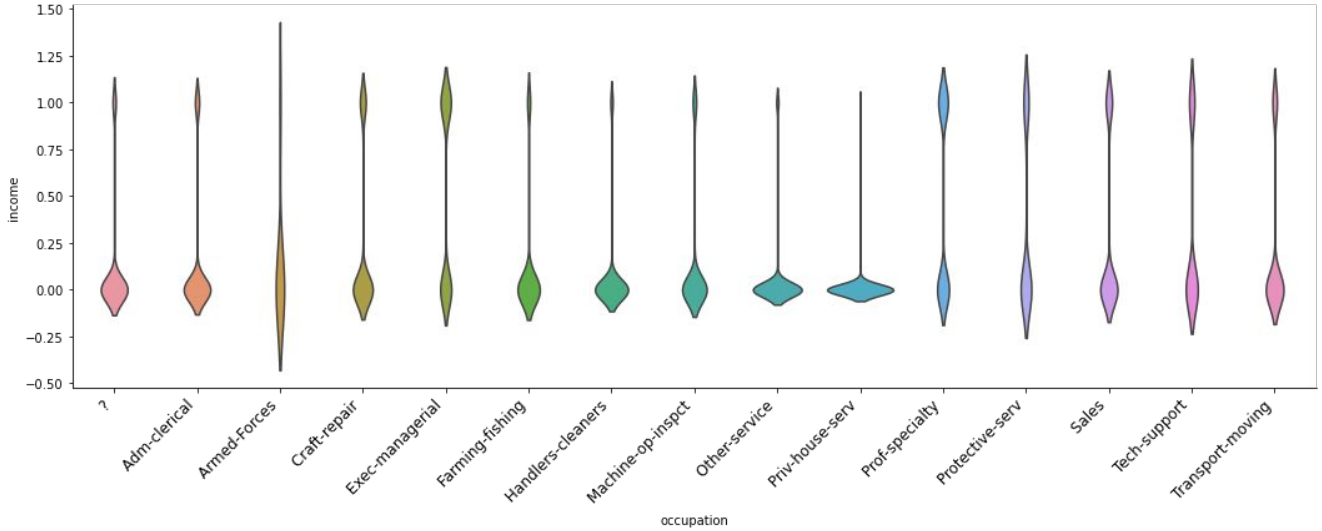
Data Story

'Married with Civilian Spouse' group which clearly shows a higher proportion of distribution in the high income group



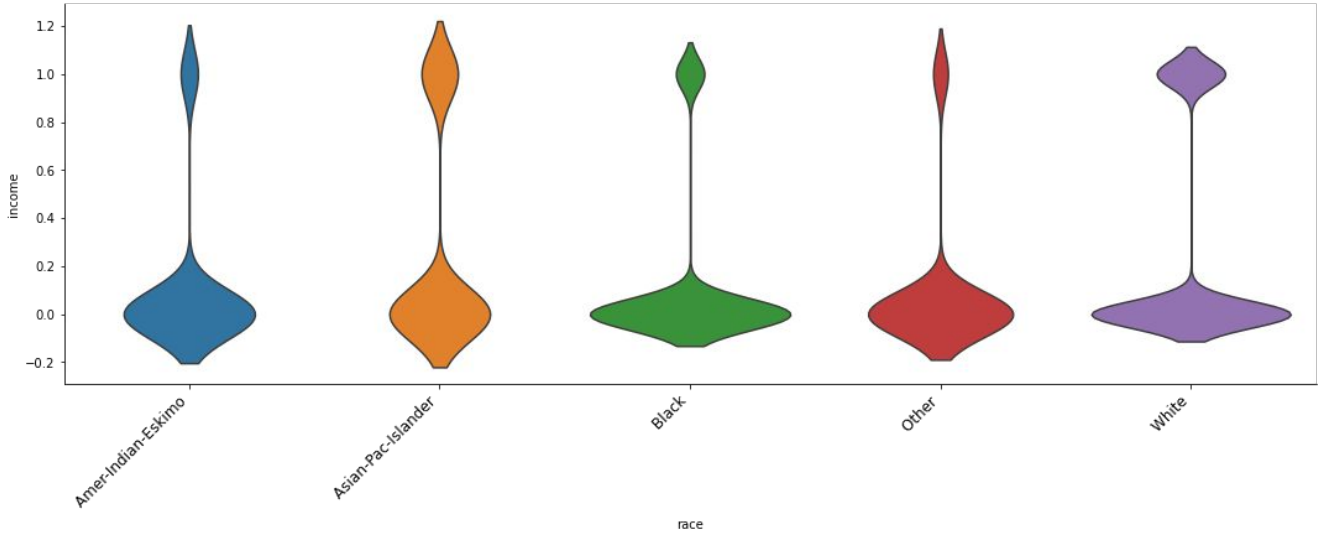
Data Story

prof-speciality, protective services, Sales, Tech Support and Transport-moving show more proportion of people falling in high income group



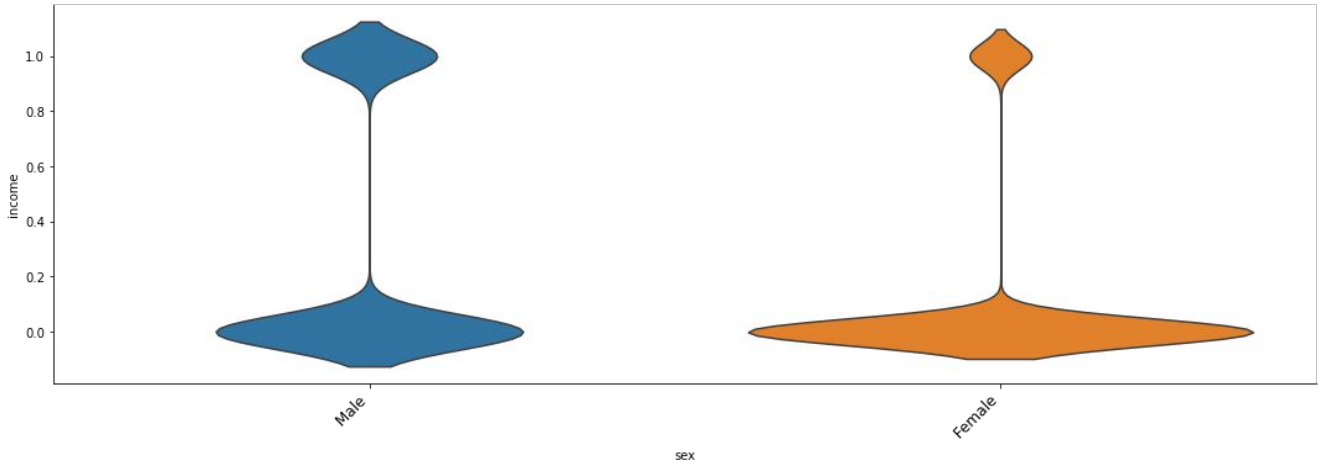
Data Story

More representation of White and Asian Pacific Islander among high income groups



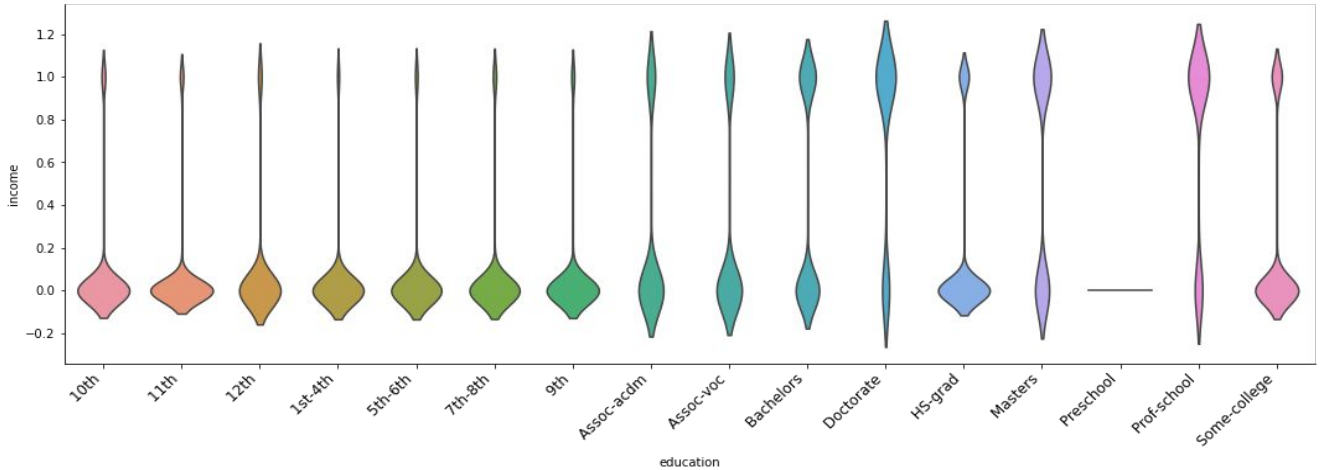
Data Story

Females have a significantly lower representation in higher income group compared to Males



Data Story

Education level clearly seems to influence income groups as lower education categories have lower income and vice versa



Data Story

- Some attributes had significant impact on affecting the income level of a person (e.g. age, hours-worked, profession, sex, education, etc.)
- While some attributes did not seem to have much impact on the income earned (e.g. workclass)
- It remains to be seen however; whether this impact is statistically significant or not (will be addressed in Statistical Analysis section)

In-Depth Analysis Using Statistics

In-Depth Analysis Using Statistics

Sample is first split between High Income (>50K) and Low Income (<=50K) groups

```
highinc = adult5[adult5.income==1]
print(highinc.shape)
print(highinc.describe())
highinc.head()
```

```
(7841, 13)
                age  education-num  hours-per-week  income
count  7841.000000    7841.000000     7841.000000  7841.0
mean     44.249841      11.611657       45.473026     1.0
std      10.519028       2.385129       11.012971     0.0
min      19.000000       2.000000        1.000000     1.0
25%      36.000000      10.000000       40.000000     1.0
50%      44.000000      12.000000       40.000000     1.0
75%      51.000000      13.000000       50.000000     1.0
max      90.000000      16.000000       99.000000     1.0
```

```
lowinc = adult5[adult5.income==0]
print(lowinc.shape)
print(lowinc.describe())
lowinc.head()
```

```
(24720, 13)
                age  education-num  hours-per-week  income
count  24720.000000   24720.000000    24720.000000  24720.0
mean      36.783738       9.595065       38.840210      0.0
std       14.020088       2.436147       12.318995      0.0
min       17.000000       1.000000        1.000000      0.0
25%       25.000000       9.000000       35.000000      0.0
50%       34.000000       9.000000       40.000000      0.0
75%       46.000000      10.000000       40.000000      0.0
max       90.000000      16.000000       99.000000      0.0
```

In-Depth Analysis Using Statistics

Age

Let's state the null and alternative hypothesis here. And use the t-test for the difference between means.

Null Hypothesis: There is no significant difference between the age distribution between people with high income vs. people with low income

Alternate Hypothesis: There is a significant difference between the age distribution between people with high income vs. people with low income

In-Depth Analysis Using Statistics

Age

- Pooled Standard Deviation of the two groups (Sp) = 13.261453837048414

Method	T-test Statistic	P Value
Manual Calculation	43.43740936207961	0.0
Scipy.Stats, equal_var = True	43.43740936207961	0.0
Scipy.Stats, equal_var = False	50.26662666736956	0.0

- T-value is significantly higher compared to mean value of age for 2 groups
- P-value (probability of t-value due to random sampling error) is 0.00
- Considering the results above, we **REJECT** the Null Hypothesis
- And **ACCEPT** the Alternate Hypothesis that there is indeed a significant difference between the Age distribution of High Income and Low Income groups.

In-Depth Analysis Using Statistics

Education

- Pooled Standard Deviation of the two groups (Sp) = 13.261453837048414

Method	T-test Statistic	P Value
Manual Calculation	64.18992220536272	0.0
Scipy.Stats, equal_var = True	64.18992220536272	0.0
Scipy.Stats, equal_var = False	64.89945481161963	0.0

- T-value is significantly higher compared to mean value of education for 2 groups
- P-value (probability of t-value due to random sampling error) is 0.00
- Considering the results above, we **REJECT** the Null Hypothesis
- And **ACCEPT** the Alternate Hypothesis that there is indeed a significant difference between the Education distribution of High Income and Low Income groups.

In-Depth Analysis Using Statistics

Hours Per Week

- Pooled Standard Deviation of the two groups (Sp) = 13.261453837048414

Method	T-test Statistic	P Value
Manual Calculation	42.58510982311484	0.0
Scipy.Stats, equal_var = True	42.58510982311485	0.0
Scipy.Stats, equal_var = False	45.12541444076267	0.0

- T-value is significantly higher compared to mean value of hours per week for 2 groups
- P-value (probability of t-value due to random sampling error) is 0.00
- Considering the results above, we **REJECT** the Null Hypothesis
- And **ACCEPT** the Alternate Hypothesis that there is indeed a significant difference between the Hours per Week distribution of High Income and Low Income groups.

In-Depth Analysis Using Machine Learning

In-Depth Analysis Using Machine Learning

Categorical features changed to continuous integers so that ML models can be used

```
X1 = adult7.drop(['income'], axis=1)
```

```
X2 = pd.get_dummies(X1)
X2.head()
```

	age	education-num	hours-per-week	workclass_?	workclass_Federal-gov	workclass_Local-gov	workclass_Never-worked	workclass_Private	workclass_Self-emp-inc	workclass_Self-emp-not-inc	...	relationship_Wife	race_Amer-Indian-Eskimo	race_Asian-Pac-Islander	race_Black	ra O
0	39	13	40	0	0	0	0	0	0	0	...	0	0	0	0	0
1	50	13	13	0	0	0	0	0	0	1	...	0	0	0	0	0
2	38	9	40	0	0	0	0	1	0	0	...	0	0	0	0	0
3	53	7	40	0	0	0	0	1	0	0	...	0	0	0	1	0
4	28	13	40	0	0	0	0	1	0	0	...	1	0	0	1	0

5 rows × 49 columns

- The original data frame with 12 features transformed into a dataframe with 49 features

In-Depth Analysis Using Machine Learning

Apply ‘Scaling’ so that features are scaled proportionaly

```
In [38]: X2.values

Out[38]: array([[39, 13, 40, ...,  1,  0,  1],
                [50, 13, 13, ...,  1,  0,  1],
                [38,  9, 40, ...,  1,  0,  1],
                ...,
                [58,  9, 40, ...,  0,  0,  1],
                [22,  9, 20, ...,  1,  0,  1],
                [52,  9, 40, ...,  0,  0,  1]])

In [39]: X3

Out[39]: array([[ 0.03067056,  1.13473876, -0.03542945, ...,  0.70307135,
                 -0.34095391,  0.34095391],
                [ 0.83710898,  1.13473876, -2.22215312, ...,  0.70307135,
                 -0.34095391,  0.34095391],
                [-0.04264203, -0.42005962, -0.03542945, ...,  0.70307135,
                 -0.34095391,  0.34095391],
                ...,
                [ 1.42360965, -0.42005962, -0.03542945, ..., -1.42233076,
                 -0.34095391,  0.34095391],
                [-1.21564337, -0.42005962, -1.65522476, ...,  0.70307135,
                 -0.34095391,  0.34095391],
                [ 0.98373415, -0.42005962, -0.03542945, ..., -1.42233076,
                 -0.34095391,  0.34095391]])
```

In-Depth Analysis Using Machine Learning

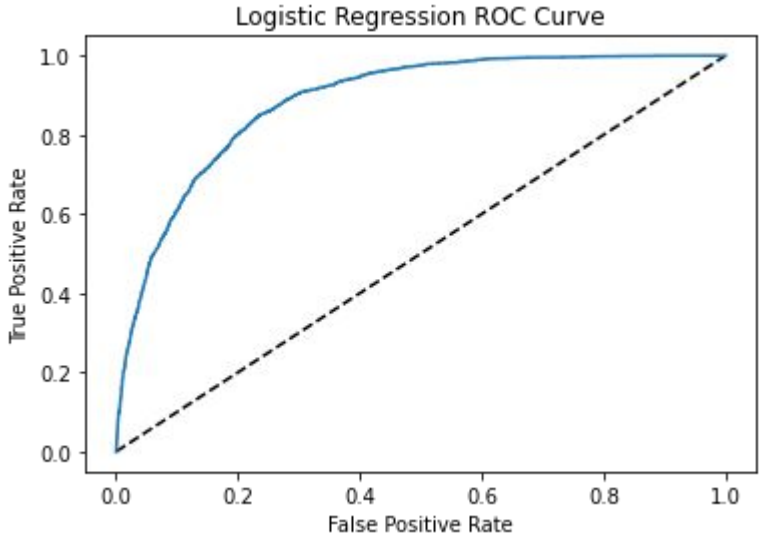
Now let's apply Logistics Regression Model (preferred for Classification problem)

```
print(confusion_matrix(yte, ypred))
```

```
[[6871  584]
 [1049 1265]]
```

```
print(classification_report(yte, ypred))
```

	precision	recall	f1-score	support
0	0.87	0.92	0.89	7455
1	0.68	0.55	0.61	2314
accuracy			0.83	9769
macro avg	0.78	0.73	0.75	9769
weighted avg	0.82	0.83	0.83	9769



- Results for low income group (0) are good, but not so great for high income group (1)
- Large area under the ROC curve. ML model used is a good for predicting income
- Issue of under representation of high income group in sample

In-Depth Analysis Using Machine Learning

Oversampling using SMOT (Synthetic Minority Oversampling Technique)

```
import imblearn
print(imblearn.__version__)
```

```
0.7.0
```

```
from imblearn import under_sampling, over_sampling
from imblearn.over_sampling import SMOTE
smote = SMOTE()
```

```
X4, y4 = smote.fit_resample(X3 , y)
```

```
X4.shape
```

```
(49440, 49)
```

```
y4.shape
```

```
(49440,)
```

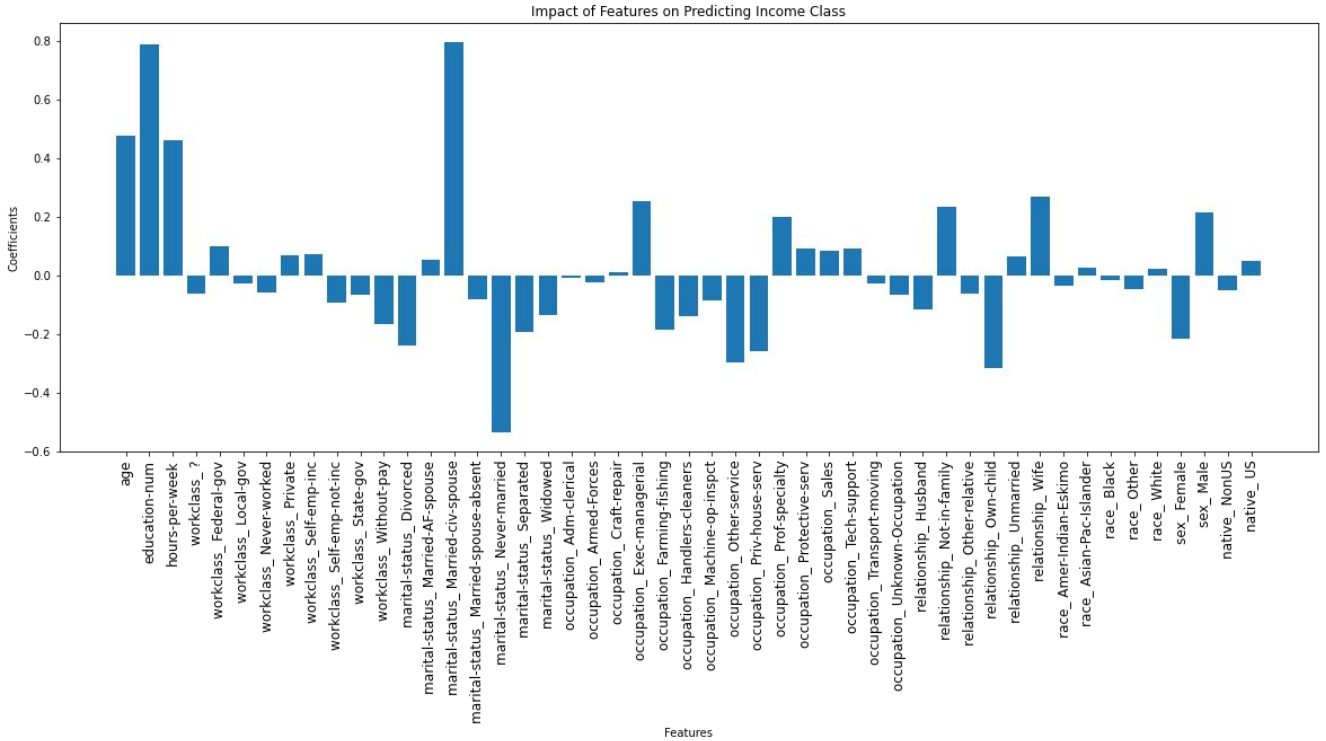
```
print(classification_report(yte2, ypred2))
```

	precision	recall	f1-score	support
0	0.84	0.77	0.80	7391
1	0.78	0.85	0.82	7441
accuracy			0.81	14832
macro avg	0.81	0.81	0.81	14832
weighted avg	0.81	0.81	0.81	14832

- Precision, Recall and F1 score for class 1 (>50K income class) have gone up significantly
- Better and balanced representation of both the <=50K and >50K income groups

In-Depth Analysis Using Machine Learning

Which Features Have High/Significant Impact on Determining/Predicting Income?



In-Depth Analysis Using Machine Learning

Key findings based on analysis of coefficient values for features

- *Age, Education, Hours-per-Week, Married W Civil Souse, Executive-Managerial Occupation and Male* categories have high and positive impact on predicting income
- *Marital Status of Never Married, Relationship of Own Child* have significant but negative correlation with income
- *Male* class is positively correlated with income while *Female* class is negatively correlated with income (this could be because of lower employment status among female class
- *Race* does not seem to have any impact on predicting the income class
- Many low skilled income categories (e.g. *Farming, Fishing, Handlers, Cleaners*) have a negative correlation with income

Closing Thoughts, Conclusions

Closing Thoughts, Conclusions

- Two income groups, High Income (>50K) and Low Income (<=50K) have distinct characteristics as defined by the features of the two groups
- To use Statistical and Machine Learning techniques, one would need to do quite a bit of
  - Data cleanup (changing data types from Objects, Categories, Integers, etc,),
  - Data transformations (e.g. scaling the data, dropping columns, splitting samples, etc.)

Closing Thoughts, Conclusions

- Machine Learning models can be applied more effectively by first transforming the data to reduce impact of inherent sample structure imbalances
    - E.g. small number of samples for high income group
    - Data scaling, converting categorical features into integer types via using get_dummies
    - Oversampling using SMOT
- ***Age, Education, Hours-per-Week, Married w Civil Souse, Executive-Managerial Occupation, and Male categories*** have high and positive impact on determining/predicting income of individuals

References

References

Adult Data Set: http://archive.ics.uci.edu/ml/datasets/Adult

Github Repository: https://github.com/harshrk/Springboard-Data-Science-Capstone-Project-1

Thank You