SALARY PREDICTION USING UNITED STATES CENSUS BUREAU DATA

ABHAY JOSHI
ADITYA CHAYAPATHY
JAGADEESH BASAVARAJU

PROBLEM STATEMENT

- To develop marketing profiles of individuals with a focus on $50,000 as a key number for salary.
- To identify the factors that determine the individual's income.
- To develop an application to predict the income of an individual.

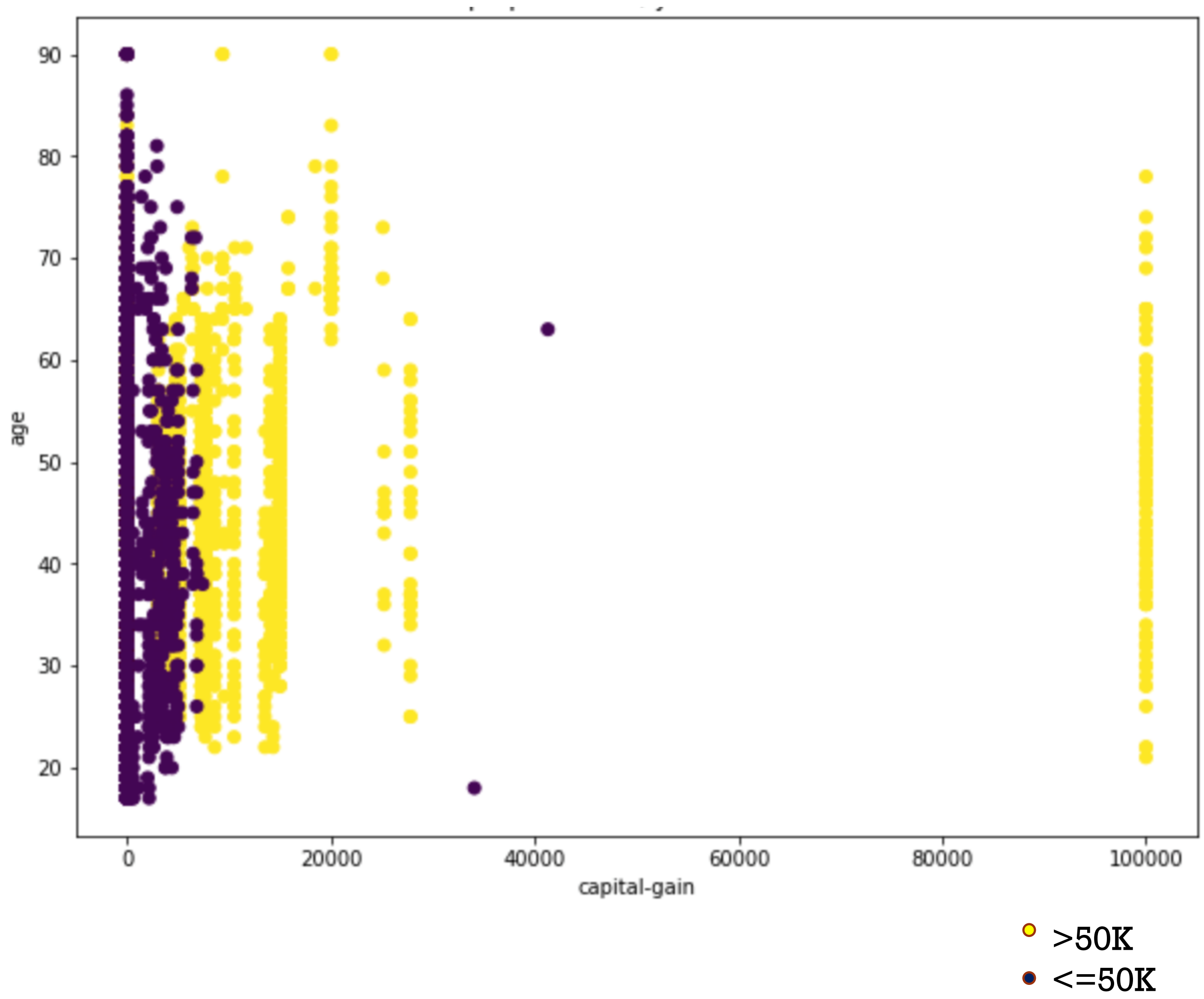
DATA SET

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

- Source :
  - United States Census Bureau
- Data cleaning:
  - Removed records having incomplete ("?") data present in them.
- Classes:
  - Above 50K (">50K")
  - Below 50K ("<=50K")
- Features:
  - 14 features with 8 features having categorial data.
- Skewed dataset (train data + test data):
  - 34014 records belonging to "<=50K" class
  - 11208 records belonging to ">50K" class
- Data used for analysis:
  - <=50K – 11208 (randomly sampled from 34014 records)
  - >50K - 11208

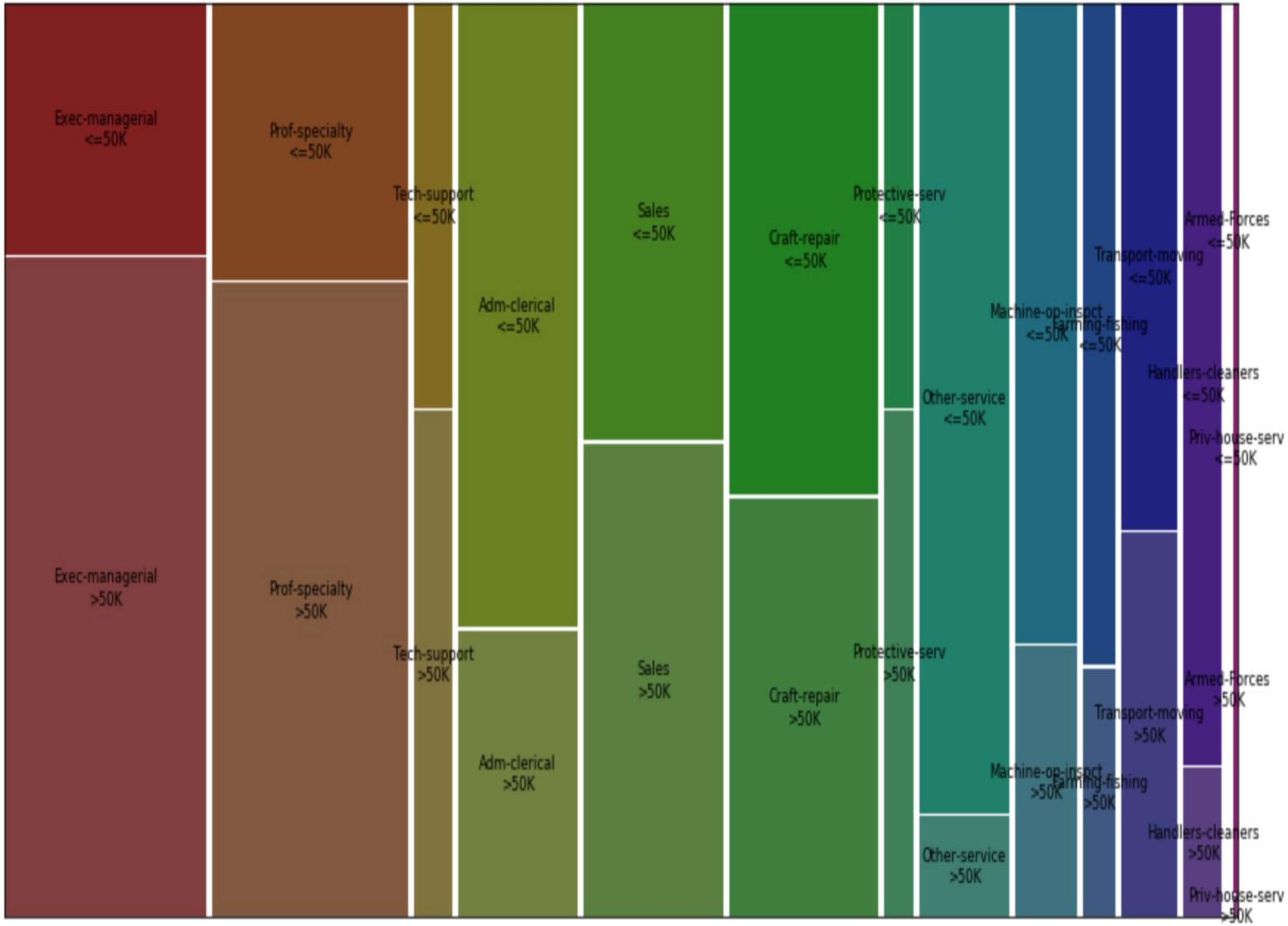
INITIAL ANALYSIS

- Top 5 important features based on initial analysis through data exploration:
  - Capital-gain
  - Age
  - Occupation
  - Education-num
  - Marital-status
- Redundant features based on initial analysis through data exploration:
  - Capital-loss
  - Fnlwgt
  - education



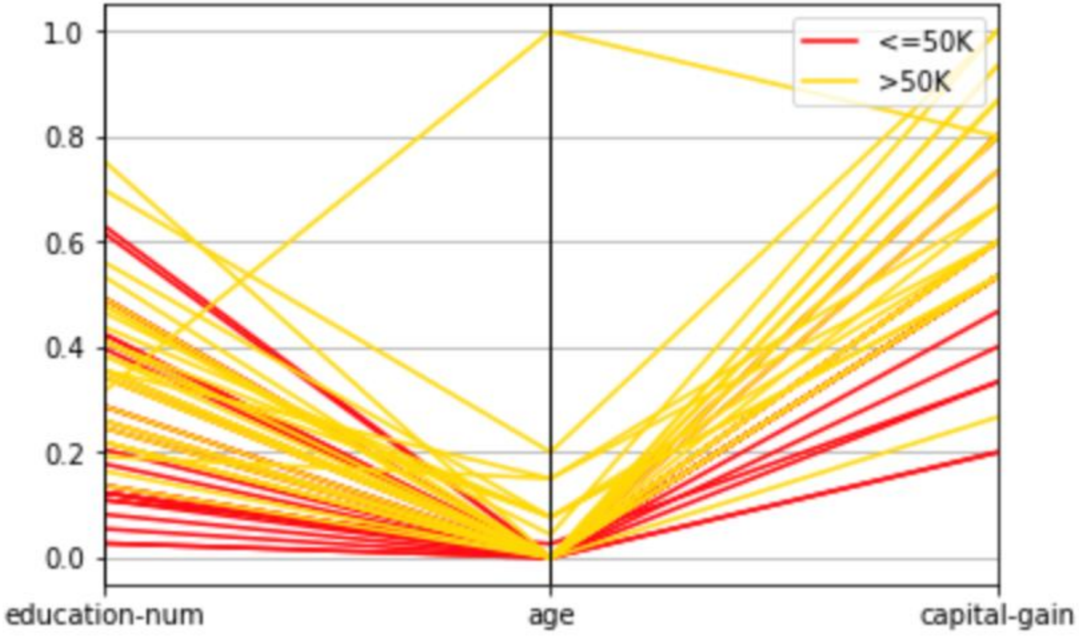
IMPORTANT FEATURES

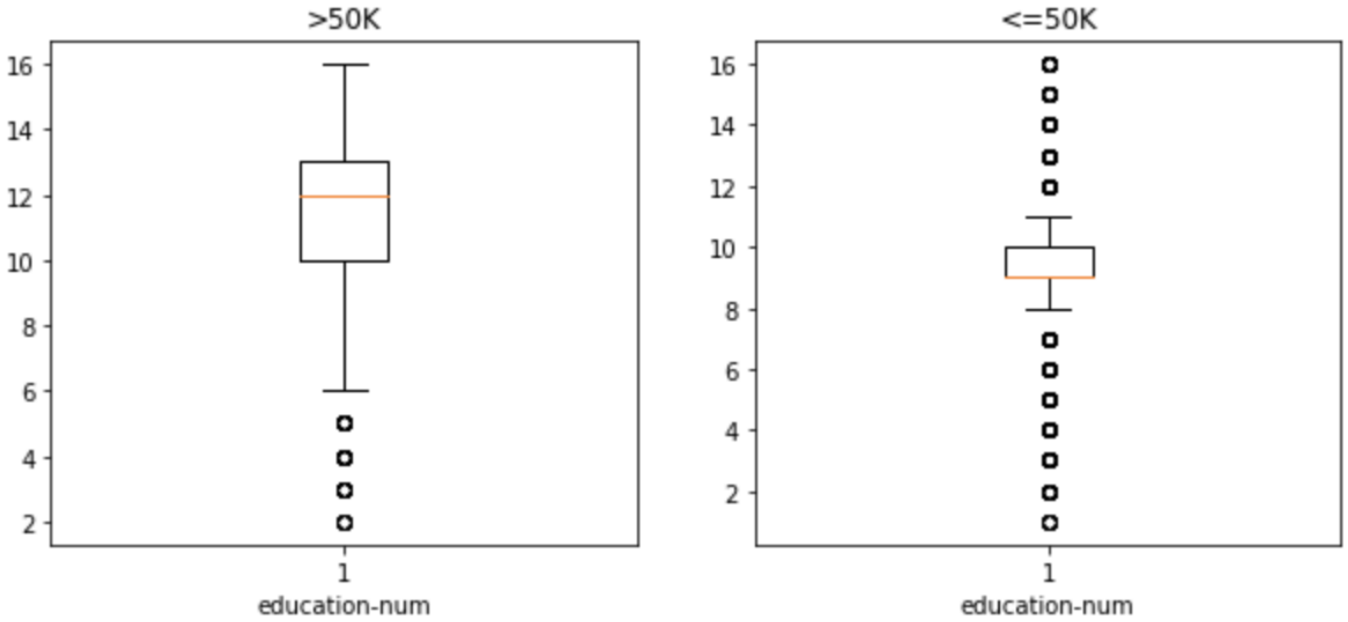
- Scatter plot:
  - Features Covered: age, capital-gain
  - X axis – Capital-gain
  - Y axis – Age
- Inferences:
  - There seems to be a separation between the two classes of data with the exception of a few outliers.
  - Individuals with high capital gain are more likely to earn more than 50K income.



IMPORTANT FEATURES CONT.

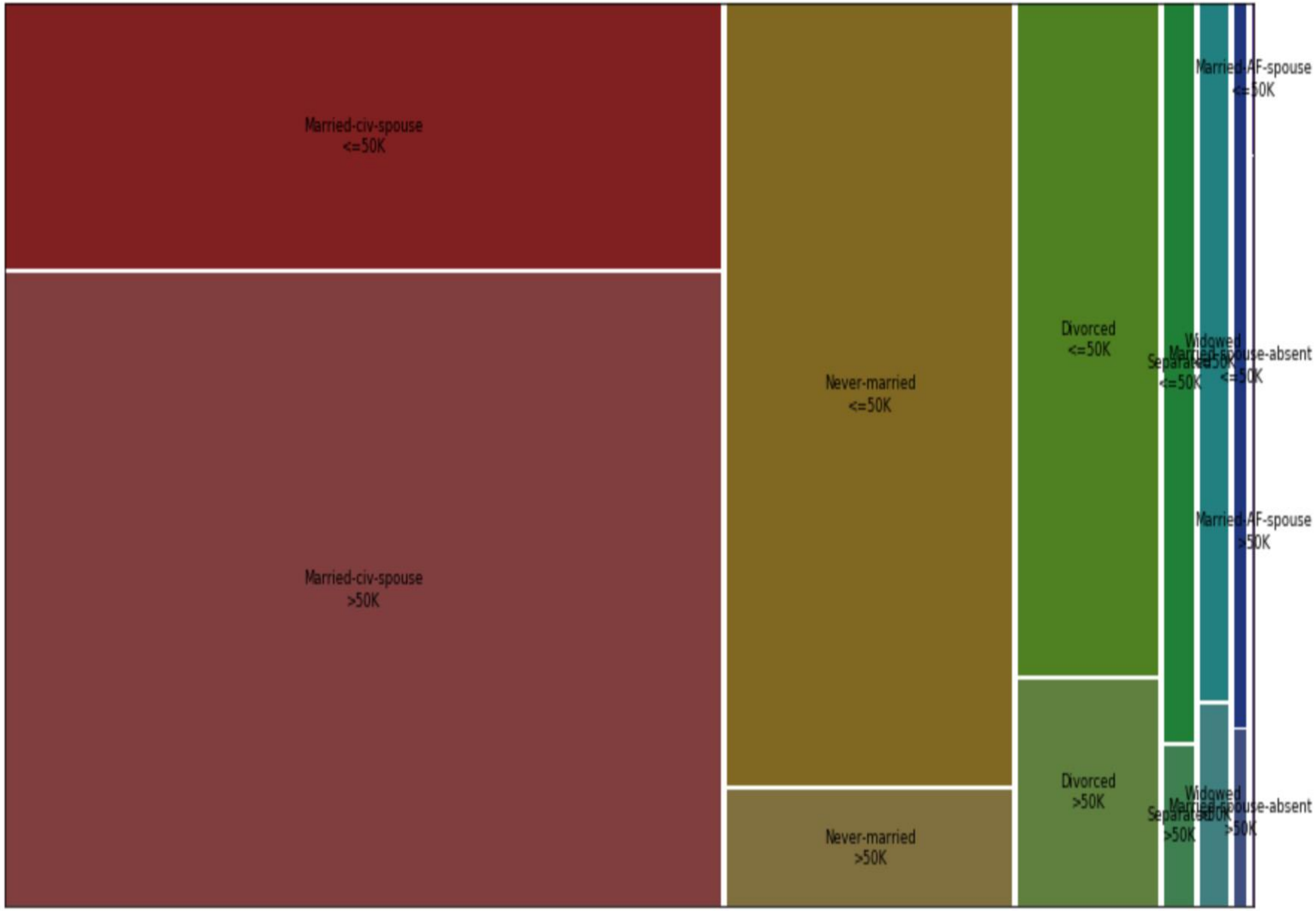
- Mosaic Plot:
  - Features covered: occupation
  - Categories in the order as they appear:
    - Adm-clerical
    - Exec-managerial
    - Handlers-cleaners
    - Prof-specialty
    - Other-service
    - Sales
    - Transport-moving
    - Farming-fishing
    - Machine-op-inspct
    - Tech-support
    - Craft-repair
    - Protective-serv
    - Armed-Forces
    - Priv-house-serv
- Inferences:
  - For most categorial data, the distribution of the two classes are highly skewed hinting that this feature can be used to distinguish among the two classes.
  - Individuals with occupations such as "Exec-managerial", "Prof-speciality" are more likely to earn 50K income.





IMPORTANT FEATURES CONT.

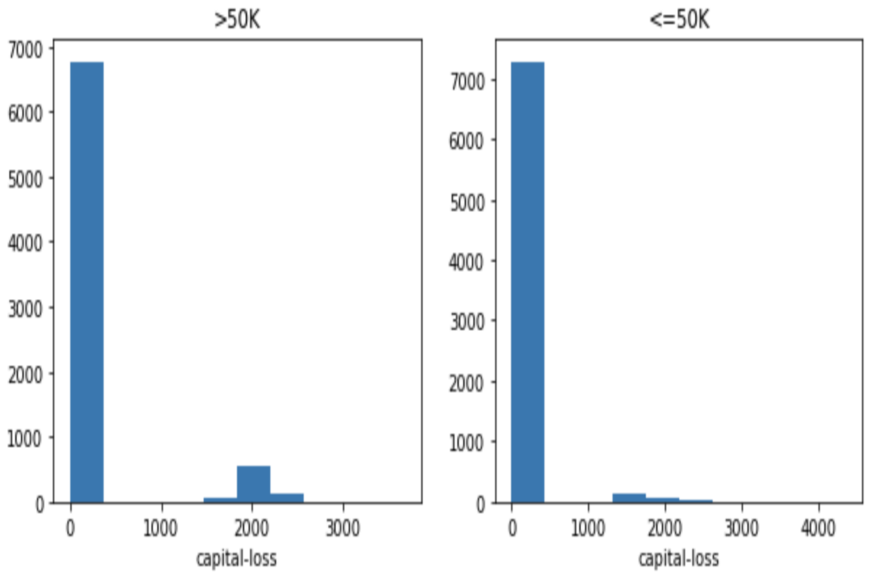
- Parallel coordinate plot:
  - Features Covered: education-num, age, capital-gain
  - Each of the features are scaled to value between 0 and 1.
- Box plot:
  - Features Covered : education-num
- Inferences:
  - From the parallel coordinate plot, we can see that the yellow lines and the red lines can be distinguished using the combination of these three features.
  - From the box plot, we can see that the distribution of the education among the two classes of data vary drastically.
  - Individuals with high education number are more likely to earn greater than 50K income.
  - Older individuals are likely to earn more than younger individuals.



IMPORTANT FEATURES CONT.

- Mosaic Plot:
  - Features covered: marital-status
  - Categories in the order as they appear:
    - Never-married
    - Married-civ-spouse
    - Divorced
    - Married-spouse-absent
    - Separated
    - Married-AF-spouse
    - Widowed
- Inferences:
  - For most categorial data, the distribution of the two classes are highly skewed hinting that this feature can be used to distinguish among the two classes.
  - Individuals with marital-status of "married-civ-spouse" are more likely to earn more than 50K income.
  - Individuals with marital-status of "never-married" are more likely to earn less than 50K income.

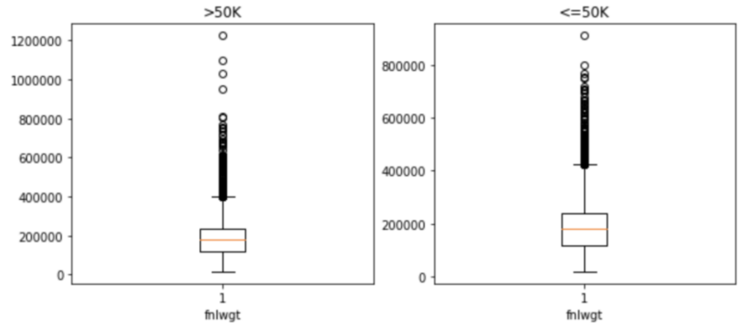
REDUNDANT FEATURES




```
Mean
Above 50K = 188149.96217368142
Below 50K = 189325.58364411295

Median
Above 50K = 176185.0
Below 50K = 178615.0

Standard Deviation
Above 50K = 102814.88940721683
Below 50K = 103992.09230884453
```

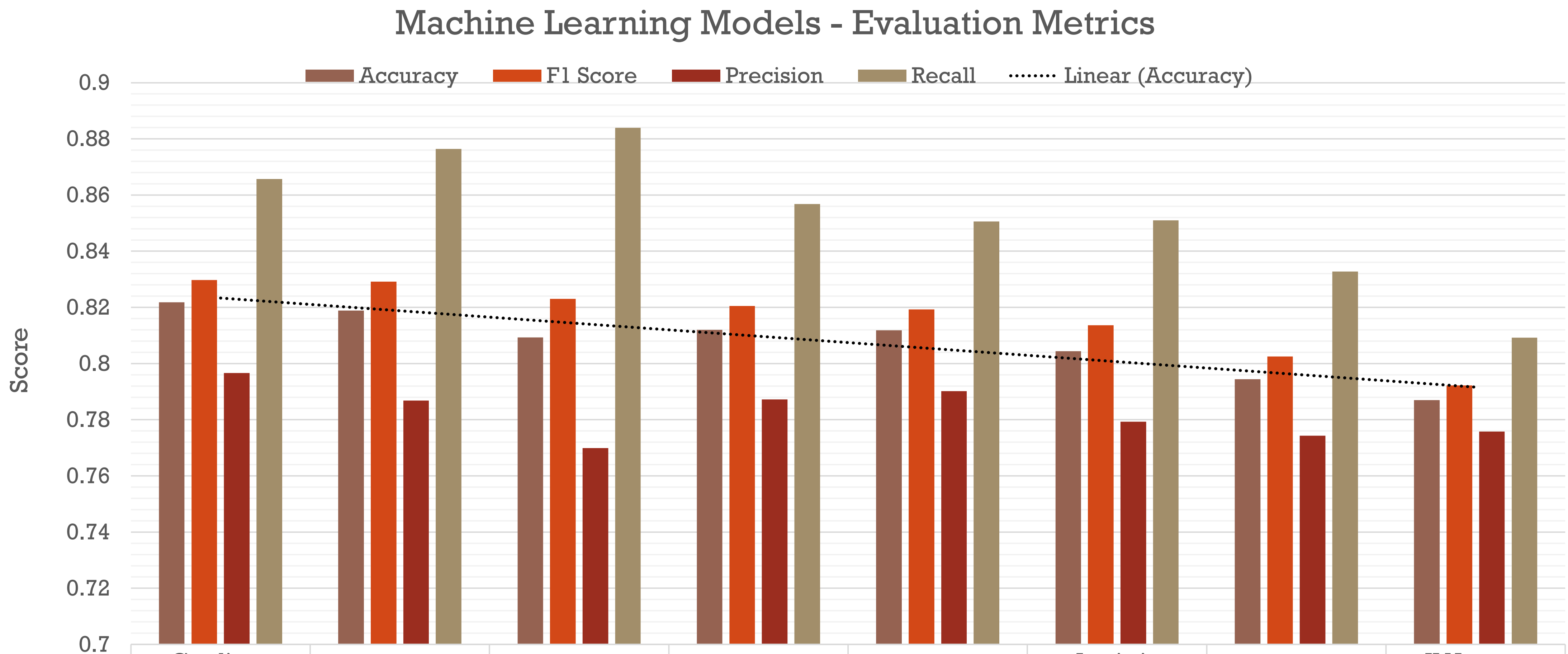


1. Education:
   - Similar information is encoded in ”education-num” feature. Hence, this feature can be ignored.
2. Capital-loss:
   - As seen in the figure, for both classes of data, they show similar distribution indicating that this feature may not help in distinguishing between the two classes of data.
3. Fnlwgt:
   - As seen in the figure, for both classes of data, fnlwgt has similar statistical properties. Also, their distribution is similar as it is evident from the box-and-whisker plot. Hence, this feature may not help in distinguishing between the two classes of data.

MACHINE LEARNING ANALYSIS

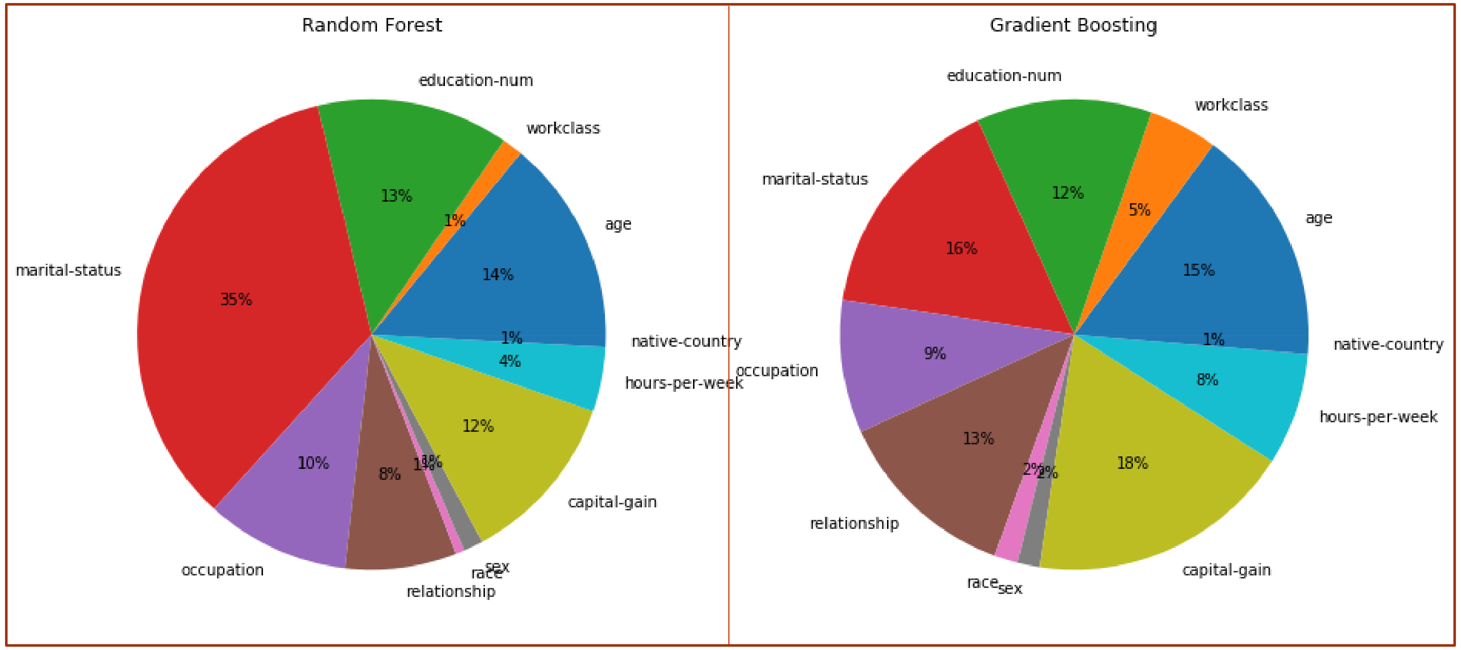
age	workclass	education-num	marital-status	occupation	relationship	race	sex	capital-gain	hours-per-week	native-country	class
38	3	9	2	3	1	2	1	0	35	2	1
54	3	9	2	1	2	2	2	0	40	2	0
19	6	10	7	3	6	2	1	0	30	2	0
49	6	13	2	1	2	2	2	0	43	2	1
25	6	13	7	1	3	2	2	0	50	2	0

1. Features excluded:
   - Fnlwgt
   - Education-num
   - Captial-loss
2. Feature Engineering:
   - All the numerical features are left as is.
   - For each categorial data, a numerical number is assigned based on the distinguishing factor of that category from our initial data exploration analysis.
3. Data Normalization:
   - Each feature is scaled to a value between 0 and 1. This is done to ensure that the ML algorithms give equal importance to each feature.
4. Data Division:
   - Data is split in the ratio of 80:20 where 80 percent of the data is used for training and 20 percent of the data is used for testing.



	Gradient Boosting	Random Forest	SVM	Neural Network	Bagging	Logistic Regression	Decision Tree	K Nearest Neighbors
Accuracy	0.8218	0.8189	0.8093	0.812	0.8118	0.8044	0.7944	0.787
F1 Score	0.8297	0.8292	0.823	0.8205	0.8193	0.8136	0.8025	0.7922
Precision	0.7966	0.7868	0.7699	0.7872	0.7902	0.7793	0.7743	0.7758
Recall	0.8657	0.8764	0.8839	0.8568	0.8506	0.851	0.8328	0.8092

FEATURE IMPORTANCE



- Pie Chart:
  - Shows the importance of the features based on the top 2 accurate ML models.
  - ML models covered:
    - Random Forest
    - Gradient Boosting
- Inferences:
  - Both the trained models more or less infer the same level of importance to each of the features.
  - As per our initial analysis, the algorithms too provide high importance to the same set of features.

QUESTIONS