Income Classification using Adult Census Data

DS 861 - Data Mining and Advanced Statistical Methods for Business Analysts
Group Project - Spring 2023 - Professor Minh Pham

By:

Saksham Motwani

Agenda

1. Objective
2. Dataset Description
3. Data Preprocessing
    a. Cleaning
    b. Feature Engineering & Feature Selection
4. Exploratory Data Analysis
5. Data Preprocessing - Encoding Categorical Variables
6. Models & Techniques
7. Results & Conclusions
8. Final Thoughts
9. Q&A

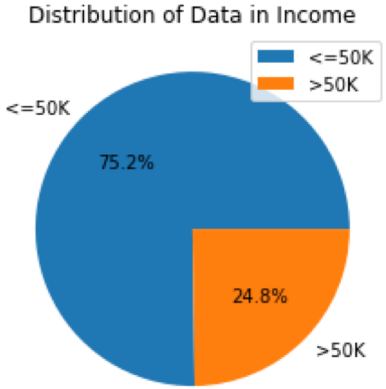
Objective

For this project, we are trying to predict whether an individual's income will be greater than $50,000 per year based on several attributes from the census data. A binary classification problem.

There is imbalance in the response variable.

First we explore the data at face value, then clean and transform the data, and finally train the data on 4 different models - Logistic Regression (5-Fold CV with threshold tuning), Decision Trees, Random Forest Classifier and Gradient Boosting Classifier (all with 5-Fold CV & hyperparameter tuning).

We also compare models based on their F1 Score.



Adult Census Income - Dataset Description

- Fetched from UCI ML repository.
- 48,842 entries extracted from 1994 US Census database. Each entry contains the following information about an individual

Categorical (8+1)

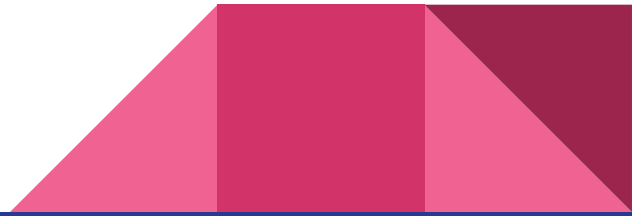
- **workclass**: represents employment status
State-gov, Self-emp-not-inc, Private, Federal-gov, Local-gov, Self-emp-inc, Without-pay, Never-worked
- **education**: highest level of education
Bachelors, HS-grad, 11th, Masters, 9th, Some-college, Assoc-acdm, Assoc-voc, 7th-8th, Doctorate, Prof-school, 5th-6th, 10th, 1st-4th, Preschool, 12th
- **marital-status**
Never-married, Married-civ-spouse, Divorced, Married-spouse-absent, Separated, Married-AF-spouse, Widowed
- **occupation**: general type of occupation
Adm-clerical, Exec-managerial, Handlers-cleaners, Prof-specialty, Other-service, Sales, Craft-repair, Transport-moving, Farming-fishing, Machine-op-inspct, Tech-support, Protective-serv, Armed-Forces, Priv-house-serv
- **relationship**: represents what this individual is relative to others
Not-in-family, Husband, Wife, Own-child, Unmarried, Other-relative
- **race**
White, Black, Asian-Pac-Islander, Amer-Indian-Eskimo, Other
- **sex**
Male, Female
- **native-country**: country of origin
41 countries, however majority category is United-States
- **income**: whether or not an individual makes more than $50,000 annually
<=50K, >50K

Numeric (6)

- **age**: integer greater than 0
- **fnlwgt**: final weight. integer greater than 0
weight assigned by the census bureau. allocate similar weights to people with similar demographic characteristics
- **education-num**: highest level of education
Label encoding of education (redundant)
- **capital-gain**: capital gains for a person (integer>0)
- **capital-loss**: capital loss for a person (integer>0)
- **hours-per-week**: hours an individual reported to work

Data Preprocessing - Data cleaning

- So to begin with, our dataset did not contain NaN values, but after we observed more closely, we found out that there were lot of ‘?’ values in the workclass (2799 instances), occupation (2809 instances) and native-country (857 instances) columns.
- We replaced these ‘?’ values with NaN, treated them as missing values, and decided to drop them.
- After dropping them, our dataset had 45222 observations and 15 features.
- We also checked for duplicates and dropped them. Finally, our dataset had 45175 observations and 15 features



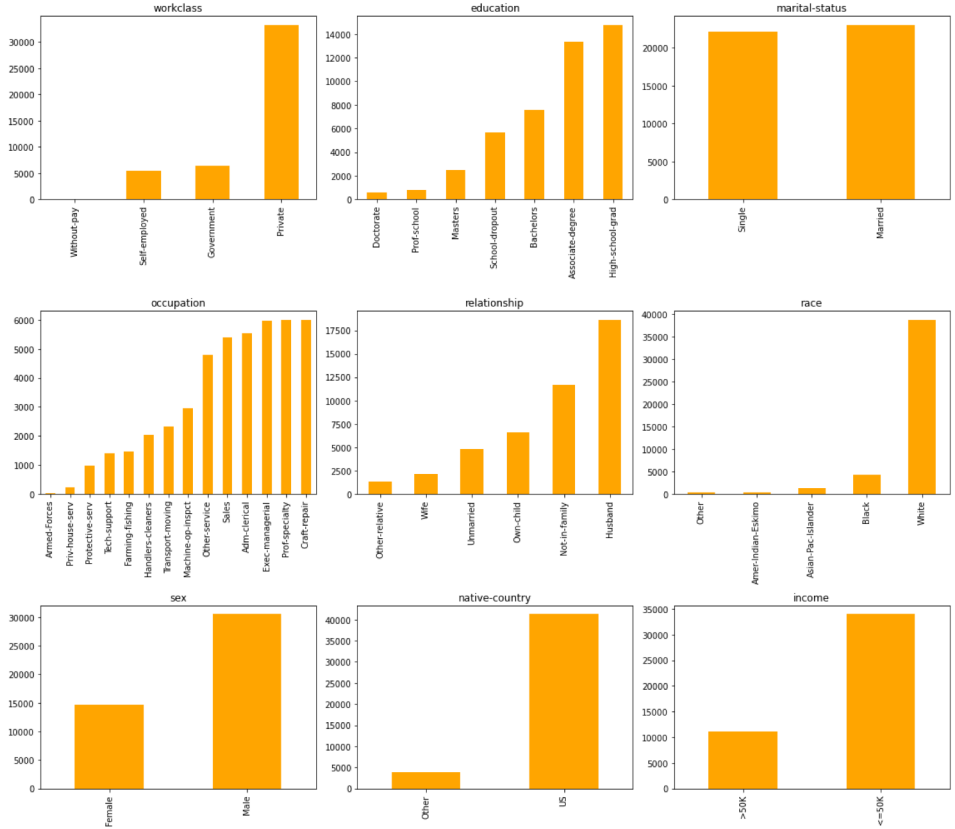
Data Preprocessing - Feature Engineering & Feature Selection

We decided to re-categorize some features in categories in a way with a minimal loss of information and dropped the ‘education-num’ column, since that was just redundant (label-encoded form of education) and proceeded for EDA

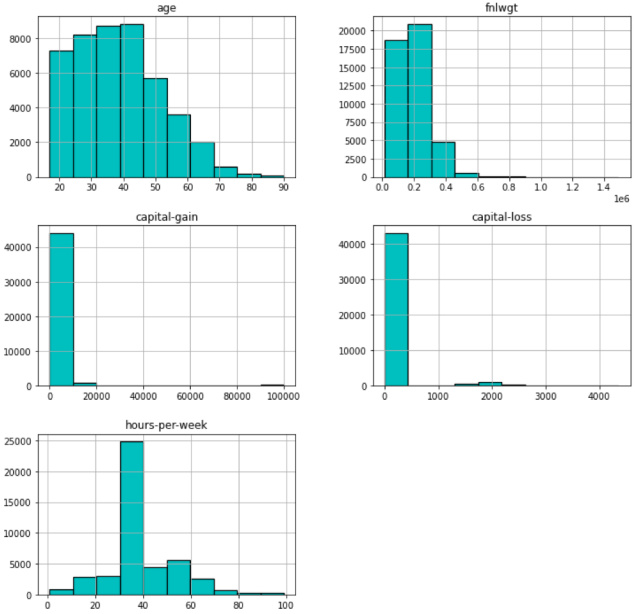
Attribute name	Old values	Recategorized value
Native country	‘United states’	United States
	'Philippines', 'Germany', 'Puerto-Rico', 'Canada', 'El-Salvador', 'India', 'Cuba', 'England', 'China'.....	Others
Education	'Preschool', '1st-4th', '5th-6th', '7th-8th', '9th','10th', '11th', '12th'	'School-dropout'
	'Assoc-voc', 'Assoc-acdm', 'Some-college'	‘Associate degree’
	‘HS-grad’	'High-school-grad'
Work class	'Local-gov','State-gov','Federal-gov'	Government
	'Self-emp-not-inc','Self-emp-inc'	Self-Employed
Marital status	'Married-civ-spouse', 'Married-AF-spouse','Married-spouse-absent', 'Separated'	Married
	'Never-married', 'Divorced', 'Widowed'	Single

Exploratory Data Analysis

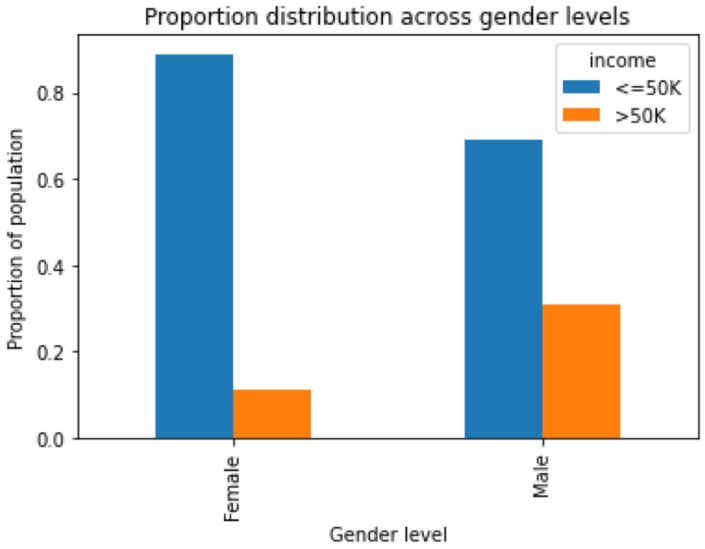
Distribution of categorical features



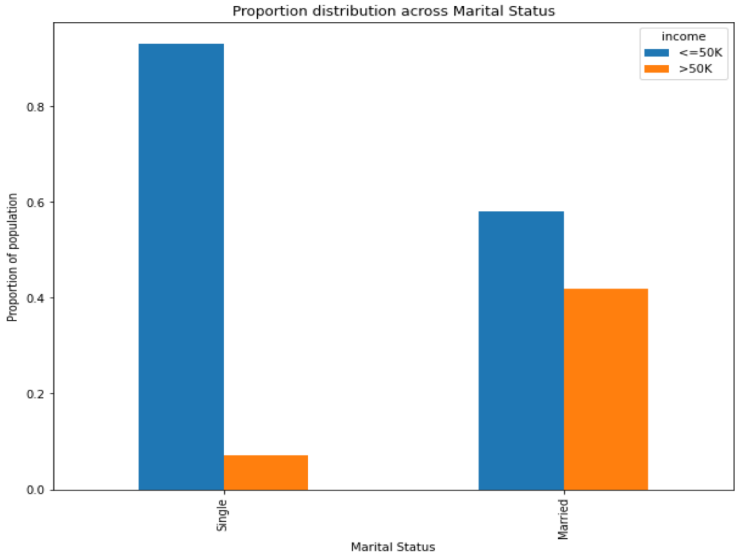
Distribution of numeric features



Exploratory Data Analysis

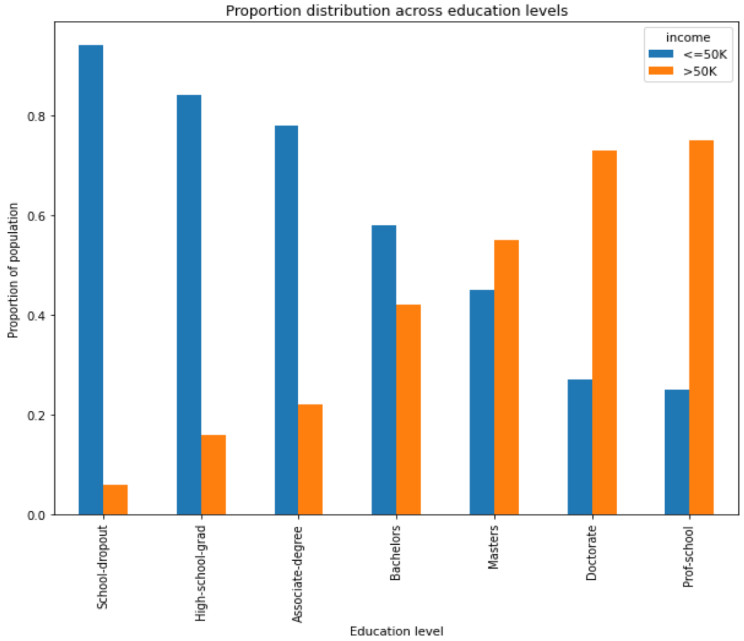


It can be observed that the proportion of men who earn over $50,000 annually is more than double of their female counterparts.



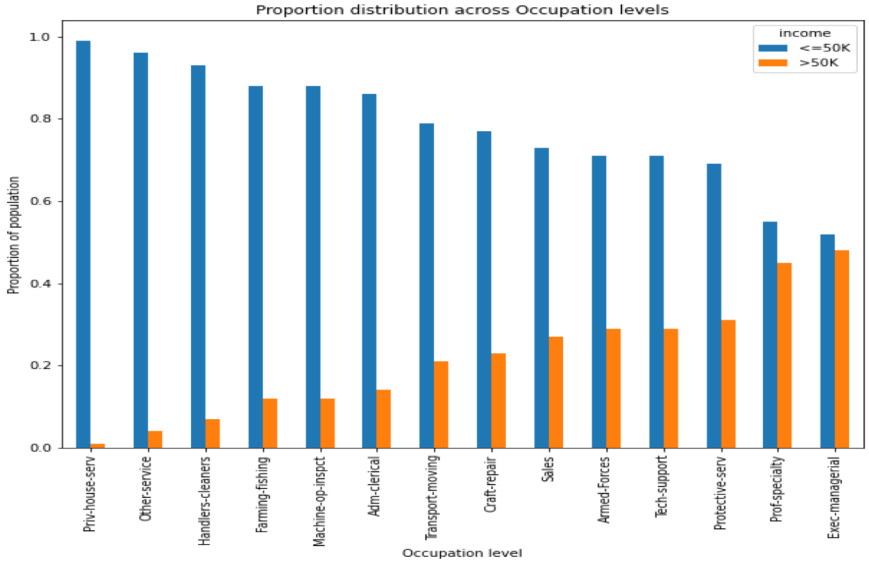
Proportion of married people who earn over $50,000 annually is much higher than the proportion of single people who earn more than $50K annually.

Exploratory Data Analysis

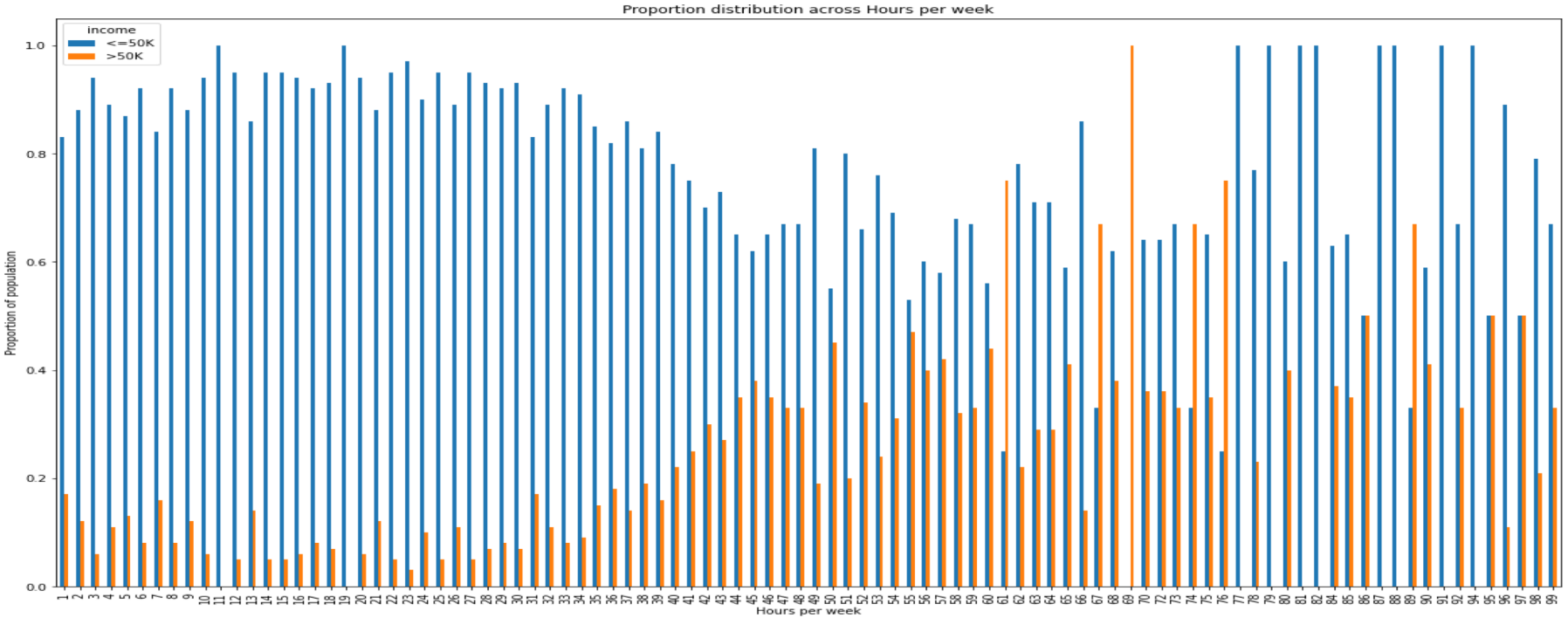


This graph shows the proportion of income classes across education levels. As expected, as the education level increases, the proportion of people who earn more than 50k a year also increases. But what is interesting is that only after a masters degree the proportion of people earning more than 50k a year is a majority.

We can see that for occupations such as 'Prof-specialty' and 'Exec-managerial', the proportion of people whose salary is higher than 50k per year is more than other occupations.



Exploratory Data Analysis



It was anticipated that there would be a correlation between the number of hours worked per week and the percentage of the population earning over 50k annually, but this was not consistently observed in the graph. Specifically, there were instances (such as when individuals worked 77, 79, 81, 82, 87, or 88 hours per week) where nobody earned more than 50k per year

Data Preprocessing - Encoding Categorical Variables

- Since our dataset contains categorical variables, it was important to handle them before we train our models.

- We created dummy variables for each categorical variable. (education, workclass, sex, race, native-country, occupation, marital-status, relationship, income)

- The reference values we chose (to allow easy interpretability) for each variable are
  workclass - Government
  education - School-dropout
  marital-status - Single
  occupation - Craft-repair
  relationship - Unmarried
  race - Other
  sex - Female
  Native-country - Other
  income - <=50K

- We drop reference columns and our dataset (features) dimensions are (45175,39)
- We split the data into 80% training, 20% test, perform scaling on the numerical features to fit in the statsmodels logistic regression function.

Logistic Regression Models - Stats models & train-test

Generalized Linear Model Regression Results

Dep. Variable:	income_>50K	No. Observations:	36140
Model:	GLM	Df Residuals:	36100
Model Family:	Binomial	Df Model:	39
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-11808.
Date:	Sun, 14 May 2023	Deviance:	23616.
Time:	12:15:59	Pearson chi2:	1.81e+06
No. Iterations:	8	Pseudo R-squ. (CS):	0.3713
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.8359	0.258	-18.763	0.000	-5.341	-4.331
age	0.3513	0.020	17.840	0.000	0.313	0.390
fnlwgt	0.0683	0.017	4.084	0.000	0.036	0.101
capital-gain	2.3914	0.072	33.113	0.000	2.250	2.533
capital-loss	0.2741	0.014	19.420	0.000	0.246	0.302
hours-per-week	0.3716	0.018	20.233	0.000	0.336	0.408
workclass_Private	0.0275	0.049	0.559	0.576	-0.069	0.124
workclass_Self-employed	-0.2595	0.064	-4.050	0.000	-0.385	-0.134
workclass_Without-pay	-0.7227	0.796	-0.908	0.364	-2.282	0.837
education_Associate-degree	1.3508	0.079	17.143	0.000	1.196	1.505
education_Bachelors	2.0712	0.084	24.628	0.000	1.906	2.236
education_Doctorate	2.8457	0.158	17.956	0.000	2.535	3.156
education_High-school-grad	0.9531	0.077	12.366	0.000	0.802	1.104
education_Masters	2.3182	0.099	23.383	0.000	2.124	2.513
education_Prof-school	2.9609	0.148	20.070	0.000	2.672	3.250
marital-status_Married	0.4988	0.108	4.633	0.000	0.288	0.710
occupation_Adm-clerical	-0.0138	0.072	-0.191	0.849	-0.156	0.128
occupation_Armed-Forces	0.9752	0.962	1.014	0.311	-0.911	2.861
occupation_Exec-managerial	0.7373	0.059	12.531	0.000	0.622	0.853
occupation_Farming-fishing	-1.1683	0.118	-9.927	0.000	-1.399	-0.938
occupation_Handlers-cleaners	-0.8401	0.124	-6.779	0.000	-1.083	-0.597
occupation_Machine-op-inspct	-0.4165	0.083	-5.030	0.000	-0.579	-0.254
occupation_Other-service	-1.0089	0.103	-9.839	0.000	-1.210	-0.808
occupation_Priv-house-serv	-2.2595	0.993	-2.275	0.023	-4.206	-0.313
occupation_Prof-specialty	0.4235	0.067	6.330	0.000	0.292	0.555
occupation_Protective-serv	0.2836	0.107	2.641	0.008	0.073	0.494
occupation_Sales	0.2058	0.063	3.269	0.001	0.082	0.329
occupation_Tech-support	0.5200	0.094	5.556	0.000	0.337	0.703
occupation_Transport-moving	-0.1537	0.078	-1.975	0.048	-0.306	-0.001
relationship_Husband	1.5999	0.126	12.685	0.000	1.353	1.847
relationship_Not-in-family	0.0376	0.090	0.418	0.676	-0.139	0.214
relationship_Other-relative	-0.1054	0.199	-0.530	0.596	-0.495	0.285
relationship_Own-child	-0.9313	0.148	-6.290	0.000	-1.221	-0.641
relationship_Wife	2.6807	0.134	19.996	0.000	2.418	2.944
race_Amer-Indian-Eskimo	-0.4977	0.305	-1.631	0.103	-1.096	0.100
race_Asian-Pac-Islander	0.2741	0.246	1.113	0.266	-0.209	0.757
race_Black	-0.0909	0.238	-0.382	0.702	-0.557	0.376
race_White	0.1237	0.229	0.541	0.588	-0.324	0.572
sex_Male	0.6400	0.072	8.929	0.000	0.500	0.780
native-country_US	0.2843	0.073	3.920	0.000	0.142	0.426

Insignificant features: workclass_private, workclass_Without-pay, occupation_Adm-clerical, occupation_Armed-Forces, relationship_Not-in-family, relationship_Other-relative, race_Amer-Indian-Eskimo, race_Asian-Pac-Islander, race_Black, race_White

Some interpretations of significant features:

- Odds of income being >50K incr. by 42% for each 1 std. dev. incr. in age
- For men, odds of income being >50K incr. by 90% (w.r.t women)
- Odds of income being >50K incr. by 159%, 286%, 693%, 915%, 1831%, 1621% for HS grad, Assoc. degree, Bachelors, Masters, Prof-school, and Doctorate respectively (compared to school dropout)
- For married, odds of income being >50K incr. by 65% (w.r.t single)

Metrics (using Sklearn), train-test split, w/o threshold tuning

```
Confusion Matrix for testing set
 [[1327  955]
 [ 483 6270]]
False Positive Rate = 0.07152376721457131
False Negative Rate = 0.41849255039439087
True Positive Rate/Recall = 0.5815074496056091
True Negative Rate = 0.9284762327854287
Accuracy = 0.840841173215274
Precision = 0.7331491712707182
F1 score of Testing Set = 0.6485826001955034
```

- FN (955) > FP (483): Higher tendency to miss positive samples than incorrectly classify negative samples.
- Low FPR
- High FNR - difficulty identifying positives

- Recall (proportion of actual positives that are correctly identified as positive by the model) - **0.58 (not a high recall rate)**
- Accuracy - **84%** - may not be a good metric since class dist. is imbalanced
- Precision (proportion of positive predictions that are correctly classified by the model) - **0.733 (decent)**

- F1 score - **0.648 (can be improved)**

Logistic Regression Model - 5-Fold CV & Threshold tuning

- Since the response variable is imbalanced, we tune the threshold value to remove the bias towards the majority class. We tune for the threshold that maximizes the F1 score.
- 50 candidate values for threshold taken from 0 to 1
- Best threshold value that maximizes validation set F1 score = 0.28

Metrics on Testing Set

```
Confusion Matrix for testing set
 [[1832  450]
 [1239 5514]]
False Positive Rate = 0.18347401155042203
False Negative Rate = 0.1971954425942156
True Positive Rate/Recall = 0.8028045574057844
True Negative Rate = 0.8165259884495779
Accuracy = 0.8130603209739901
Precision = 0.5965483555845001
F1 score of Testing Set = 0.6844759947692881
```

- Performed better than the previous model
- Recall is higher (**0.803**) -> better at correctly identifying higher proportion of positive cases
- Precision is slightly lower (**0.597**) -> expected because threshold value is tuned to optimize F1 score
- F1 score is higher (**0.684**) -> overall a better performing model

Decision Tree Model - 5-Fold CV & Hyperparameter tuning

- Tuned 3 parameters: max_depth, max_leaf_nodes, criterion
- max_depth: no. of levels. limits number of splits and prevents overfitting. Candidate values - np.arange(2,50,2)
- max_leaf_nodes: sets max. leaf nodes in the tree. limits tree complexity and prevents overfitting.
Candidate values: [2, 4, 8, 16, 32, 64, 128, 256, 512, 1024]
- criterion: measures quality of split. Candidate values: gini, entropy

Metrics on Testing Set

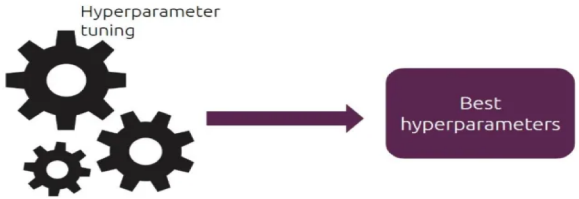
```
Decision Tree – Best Hyperparameters: {'criterion': 'gini', 'max_depth': 24, 'max_leaf_nodes': 128}
Decision Tree – F1 Score on Testing Set: 0.6710050614605929
Decision Tree – Confusion Matrix for Testing Set:
 [[1392  890]
 [ 475 6278]]
Decision Tree – Accuracy of Testing Set: 0.8489208633093526
Decision Tree – Features ranked according to their importance are as below:

                          feature  importance
14       marital-status_Married    0.351797
2                  capital-gain    0.232972
3                  capital-loss    0.101893
9            education_Bachelors    0.049664
0                           age    0.039830
12             education_Masters    0.039698
4                hours-per-week    0.034175
17   occupation_Exec-managerial    0.028621
23     occupation_Prof-specialty    0.026032
8      education_Associate-degree    0.015725
32           relationship_Wife    0.012107
11     education_High-school-grad    0.011634
13          education_Prof-school    0.009389
29    relationship_Not-in-family    0.008566
1                         fnlwgt    0.006428
28          relationship_Husband    0.006092
10          education_Doctorate    0.005852
21     occupation_Other-service    0.004759
18    occupation_Farming-fishing    0.003371
26       occupation_Tech-support    0.001937
6         workclass_Self-employed    0.001770
19  occupation_Handlers-cleaners    0.001531
5           workclass_Private    0.001192
36                 race_White    0.001171
15       occupation_Adm-clerical    0.001124
20  occupation_Machine-op-inspct    0.000802
38          native-country_US    0.000662
37                   sex_Male    0.000651
27   occupation_Transport-moving    0.000555
```

- Performed worse than Logistic Regression
- Best hyperparameters: 'criterion': 'gini', 'max_depth': 24, 'max_leaf_nodes': 128
- F1 score is slightly higher (**0.671**) -> overall a lesser performing model
- Imp features: marital-status_Married, capital-gain, capital-loss

Hyperparameter tuning on Random Forest

1. GridSearchCV : .GridSearchCV performs an exhaustive search over a specified grid of hyperparameters.
2. RandomizedSearchCV : .RandomSearchCV, on the other hand, randomly samples from a distribution of hyperparameters for a fixed number of iterations. With a 100 iterations, we got the below parameters

	n_estimators	min_samples_leaf	max_features
RandomizedSearchCV	600	2	9
GridSearchCV	450	3	6



Random Forest Model Evaluation

Following is model evaluation of Random Forest model on Census data using GridSearchCV using 5-Fold CV.

- The **F1 score of 0.673**, which represents the harmonic mean of precision and recall, indicates the balance between the two metrics. It suggests that the classifier achieves a reasonable trade-off between correctly identifying positive instances (precision) and capturing all positive instances (recall).
- An **Accuracy of 0.8567** indicates that the classifier is able to correctly classify approximately 85.67% of the instances in the testing set.
- In summary, the results suggest that the Random Forest classifier with the specified parameters and 5-fold cross-validation demonstrates a reasonable ability to handle the binary classification problem.
- Best parameters : **max_features: 6, min_samples_leaf: 3 n_estimators: 450**.
- Performed better than DT but still not as good as Log regression.
- Imp features : **capital _gain , marital-status _Married, age**

```
Random Forest - Best Hyperparameters: {'max_features': 6, 'min_samples_leaf': 3, 'n_estimators': 450}
Random Forest - F1 Score on Testing Set: 0.6736214605067065
Random Forest - Confusion Matrix for Testing Set:
 [[1356  926]
 [ 388 6365]]
Random Forest - Accuracy of Testing Set: 0.8545655783065855
```

Random Forest - Features ranked according to their importance are as below:

	feature	importance
2	capital-gain	0.167997
14	marital-status_Married	0.112531
0	age	0.110246
28	relationship_Husband	0.098998
1	fnlwgt	0.070412
4	hours-per-week	0.067252
3	capital-loss	0.047863
17	occupation_Exec-managerial	0.035258
9	education_Bachelors	0.034523
23	occupation_Prof-specialty	0.032433
32	relationship_Wife	0.025161
12	education_Masters	0.021008
29	relationship_Not-in-family	0.019707
37	sex_Male	0.016669
11	education_High-school-grad	0.016064
31	relationship_Own-child	0.015121
13	education_Prof-school	0.012910
8	education_Associate-degree	0.012852
21	occupation_Other-service	0.011225
5	workclass_Private	0.008879
10	education_Doctorate	0.007342
6	workclass_Self-employed	0.006519
25	occupation_Sales	0.006384
18	occupation_Farming-fishing	0.005955
38	native-country_US	0.004901
26	occupation_Tech-support	0.004736
15	occupation_Adm-clerical	0.004123
20	occupation_Machine-op-inspct	0.004067
36	race_White	0.004056
27	occupation_Transport-moving	0.003384
19	occupation_Handlers-cleaners	0.003335

Gradient Boosting Model Evaluation

Following is model evaluation of Gradient Boosting model on Census data using GridSearchCV using 5-Fold CV.

- The **F1 score of 0.705**, which represents the harmonic mean of precision and recall, indicates the balance between the two metrics.
- An **accuracy of 0.86596** indicates that the classifier is able to correctly classify approximately 86.59% of the instances in the testing set.
- In summary, the results suggest that the Gradient Boosting classifier with the specified parameters and 5-fold cross-validation demonstrates a reasonable good ability to handle the binary classification problem.
- Best parameters : **Learning rate: 0.1,Maximum depth of each tree (max_depth): 4,Number of estimators (n_estimators): 300**
- Performed **the best** out of all the other models.
- Imp features: **capital _gain, relationship-Husband, marital-status _Married**

```
Gradient Boosting - Best Hyperparameters: {'learning_rate': 0.1, 'max_depth': 4, 'n_estimators':
300}
Gradient Boosting - F1 Score on Testing Set: 0.7048501096758469
Gradient Boosting - Confusion Matrix for Testing Set:
 [[1446  836]
 [ 375 6378]]
Gradient Boosting - Accuracy of Testing Set: 0.8659656889872718
```

```
Boosting - Features ranked according to their importance are as below:

                          feature  importance
2                   capital-gain    0.231591
28          relationship_Husband    0.196180
14        marital-status_Married    0.122277
3                   capital-loss    0.077494
0                            age    0.063394
4                hours-per-week    0.047945
32             relationship_Wife    0.044703
23     occupation_Prof-specialty    0.036462
9            education_Bachelors    0.035509
17    occupation_Exec-managerial    0.031691
1                         fnlwgt    0.020606
12             education_Masters    0.019430
13          education_Prof-school    0.011141
21      occupation_Other-service    0.009992
10           education_Doctorate    0.008080
11    education_High-school-grad    0.007285
18    occupation_Farming-fishing    0.006115
8     education_Associate-degree    0.005528
26       occupation_Tech-support    0.003626
25             occupation_Sales    0.003429
5              workclass_Private    0.002320
20  occupation_Machine-op-inspct    0.002188
37                    sex_Male    0.002117
6       workclass_Self-employed    0.001946
```

Results and Conclusions

Based on these results, we can draw some conclusions:

- Gradient Boosting outperforms the other three models in terms of both accuracy (0.866) and F1 score (0.705). Therefore, it seems to be the most effective model among the four.
- Logistic Regression: Tuning the threshold really helped improving F1 score.
- Gradient Boosting exhibits a better balance between precision and recall, as indicated by its higher F1 score. This suggests that it can effectively handle both false positives and false negatives, making it a more reliable model.

Model	Accuracy	F1 Score
Logistic Regression	0.813	0.6844
Decision Tree	0.8489	0.671
Random Forest	0.8546	0.6736
Gradient Boosting	0.8659	0.7048

Final Thoughts

Challenges

1. Handling categorical variables: We were confused about which technique to use (Label Encoding/One hot encoding) as some categorical variables had a high number of unique values and one hot encoding was leading to 107 features, which would have been tough to interpret. So we decided to recategorize (bucket) some categories based on logic, with minimal loss of information/meaning.
2. Running GBT with > 10 candidate values for each of three hyperparameters due to computational and time limitations. Despite the model running an entire night (>12 hours), did not get a result. So we had to restrict candidate values to 2 or 3 and over a smaller range of values.

Learnings

1. Data preprocessing: Took 70-80% of time we spent on the project.
2. Apply learnings from the course by training 4 models and tuning hyperparameters for each.
3. Working on assignments where direction was given v/s taking your own direction (completely different).

Things we would have done if we had more time

1. Definitely PCA to reduce dimensionality - Analysing PCA variance ratios and dropping features step-wise with the least ratios.
2. Bagging and KNN.
3. Regularization methods to our log regression model.