U.S. Adult Income Analysis & Prediction

A demo on building a machine learning model for a classification problem and related interpretation

- Python codes on Jupyter notebook -

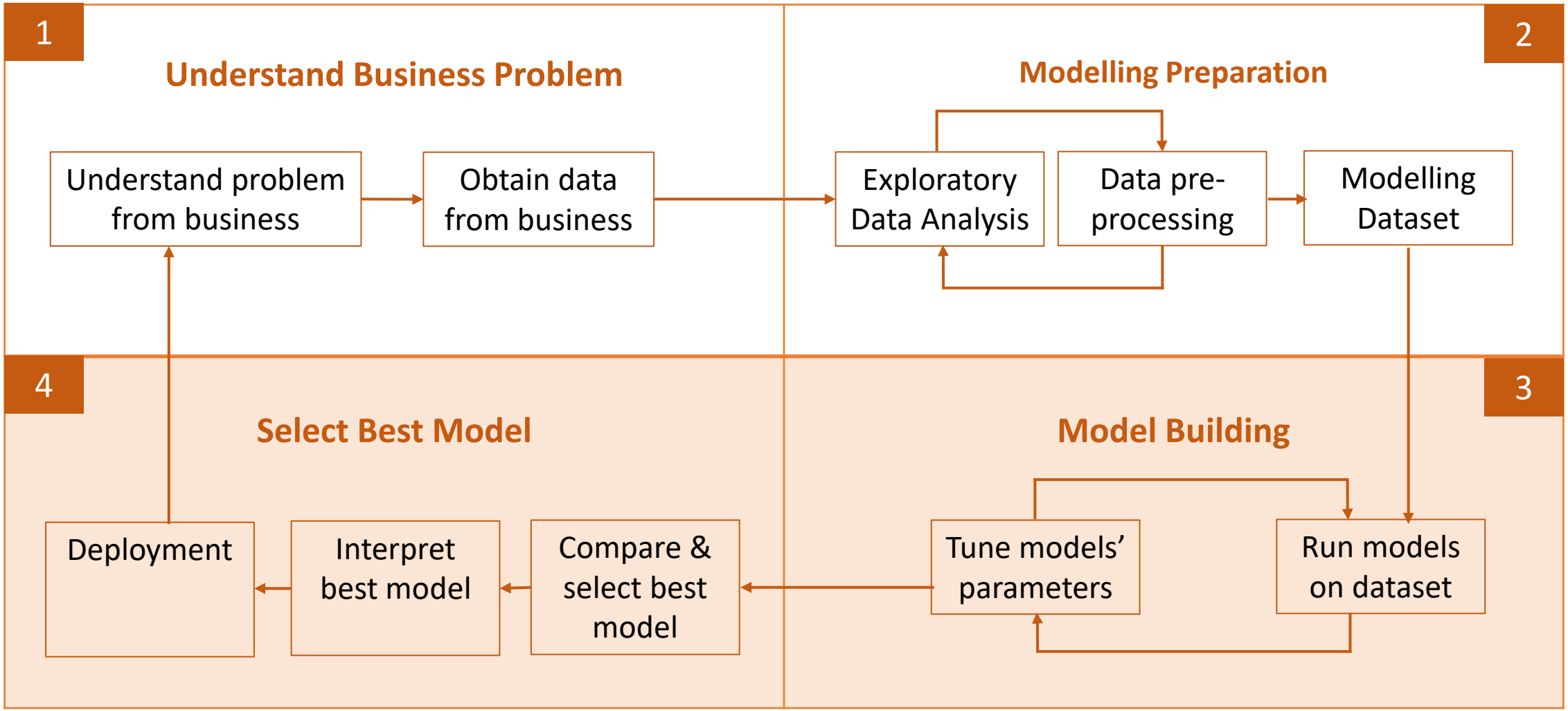
Objective

Using a small dataset from U.S. Census Database in 1994, I will demonstrate how a machine learning pipeline is implemented, in building a model for predicting adult income based on individual characteristics. This slide deck and accompanied codes will showcase the following:

1. Exploratory data analysis for understanding the data and features selection
2. Data pre-processing techniques
3. Various machine learning algorithms and hyperparameters tuning
4. Model performance measures for model selection
5. Explainable AI methods that helps to understand how and why the ML model makes certain predictions

⌨ Python codes are shown in Jupyter notebook.

Machine Learning Pipeline



The Model in this Demo

Data source

- UCI Machine Learning Repository (also available in Kaggle)
- Extraction was done by Barry Becker from the 1994 Census database

A Classification Model

- Prediction: Binary classification - a person having income <= 50K or >50K a year
- Features: 10+ individual characteristics such as age, education, work hours per week

Exploratory Data Analysis

The Dataset

Number of records
32,561

Target variable
Income <= 50K a year (‘0’)
Income > 50K a year (‘1’)

6 numerical attributes
Age
Education level
Working hours per week
Capital gain
Capital loss
Final weight of record

8 categorical attributes
Sex
Education
Occupation
Work class
Marital status
Relationship
Nationality
Race

Exploratory Data Analysis – Analysis Method

Numerical attributes

- Univariate analysis – histogram
- Bivariate analysis with income – box plots

Categorical attributes

- Univariate analysis – bar charts
- Bivariate analysis with income – normalised stacked bar charts

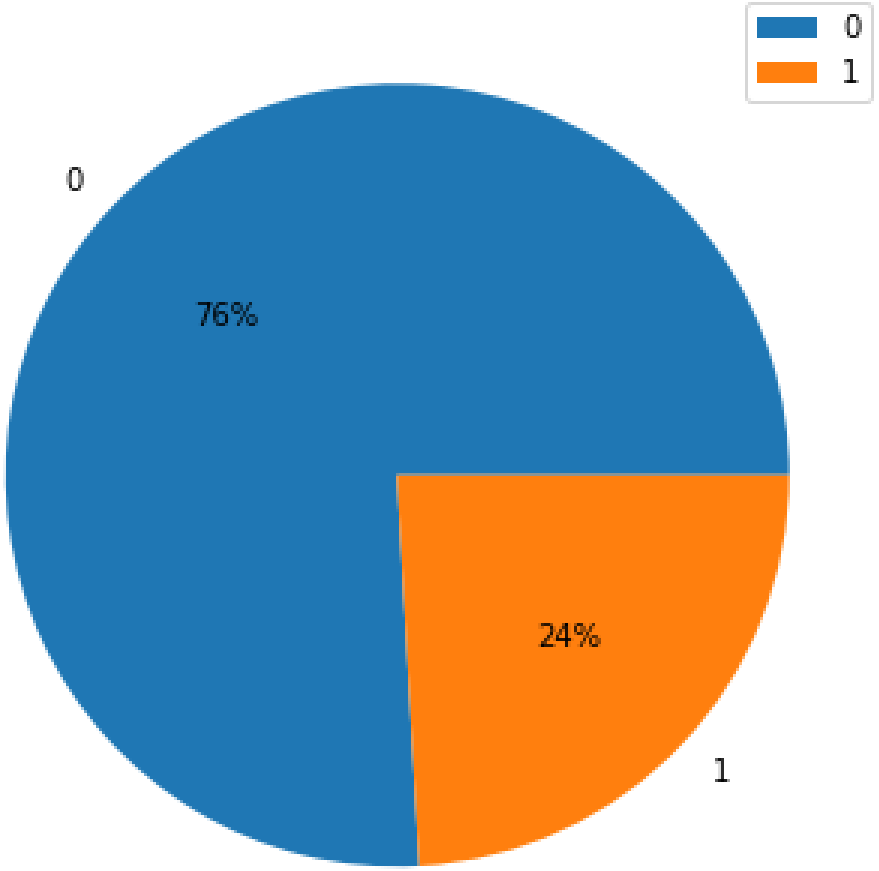
All attributes

- Bivariate analysis between selected variables – box plots
- Multivariate analysis – scattered plots and correlation matrix

Exploratory Data Analysis – Income

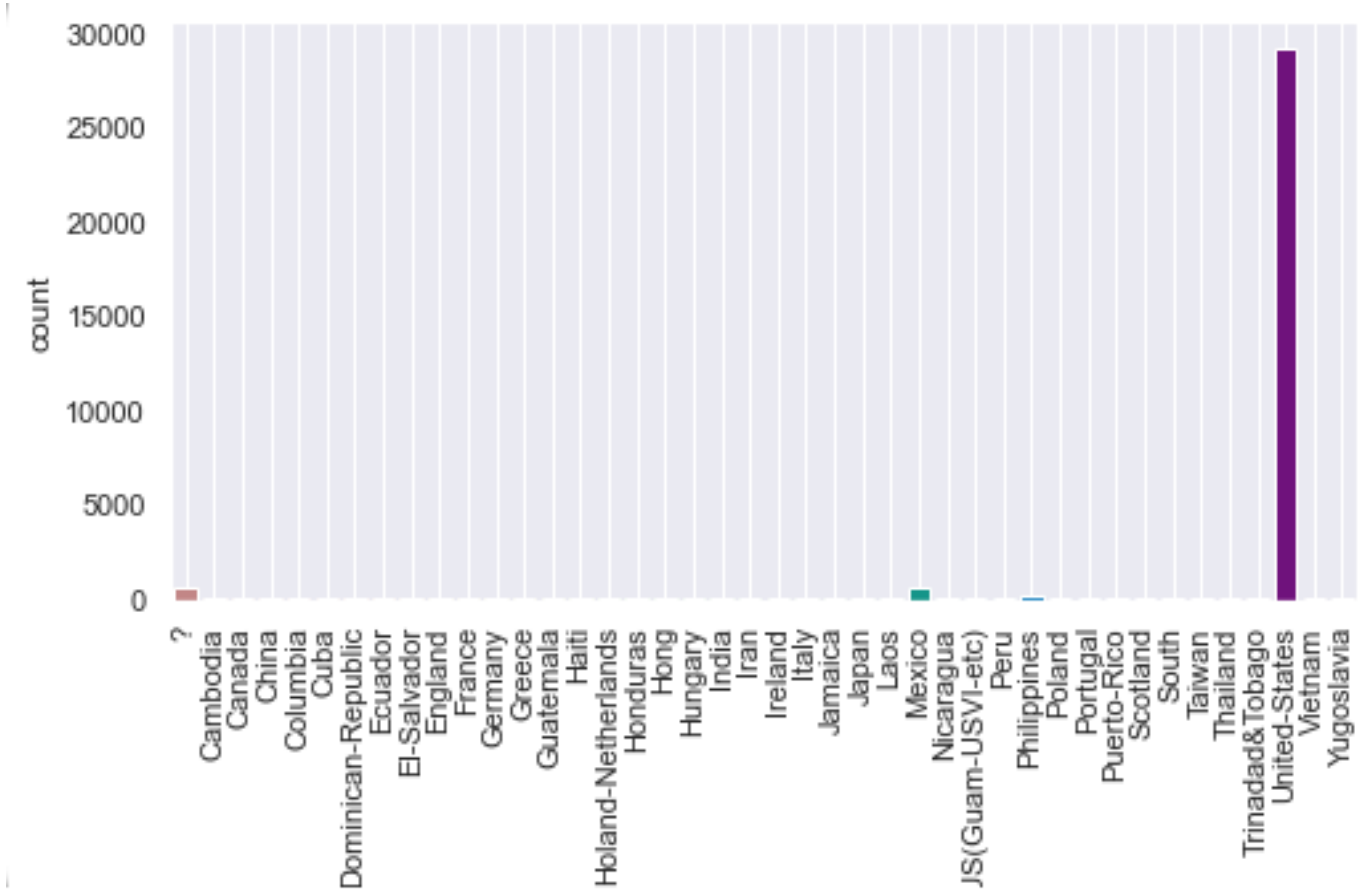
- Imbalanced target variable
  - **‘0’**: Income <= 50K - 76%
  - **‘1’**: Income > 50K - 24%

→ We will explore various treatment for imbalanced dataset in Data Pre-processing section.



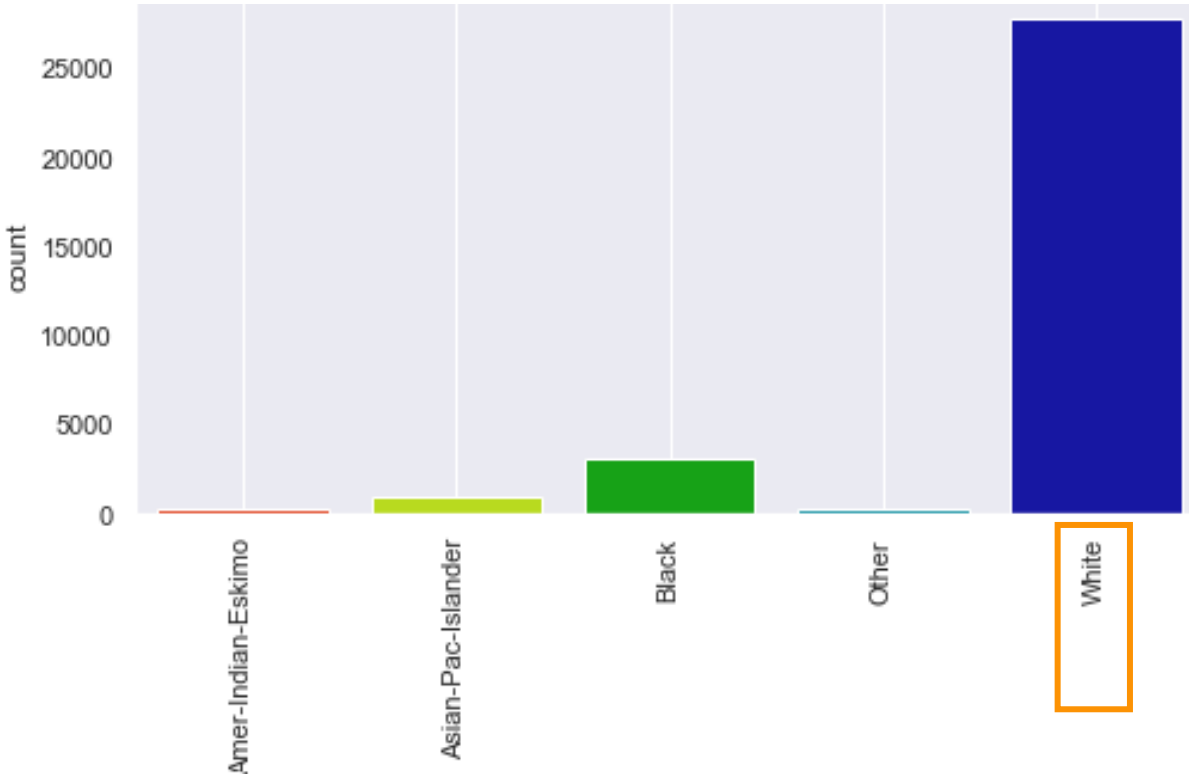
Exploratory Data Analysis – Native Country

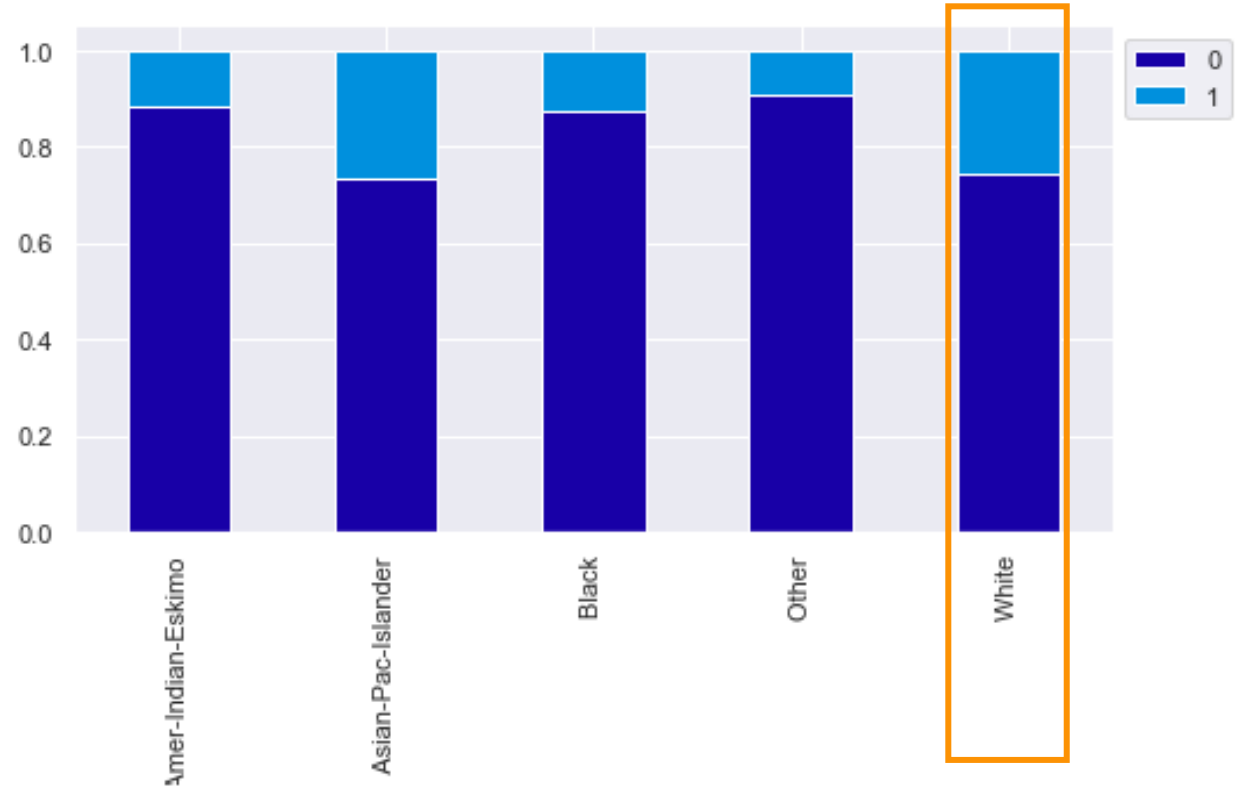
- This data set is from US Census Bureau database, so unsurprisingly majority of the people were born in the US.



Exploratory Data Analysis – Race

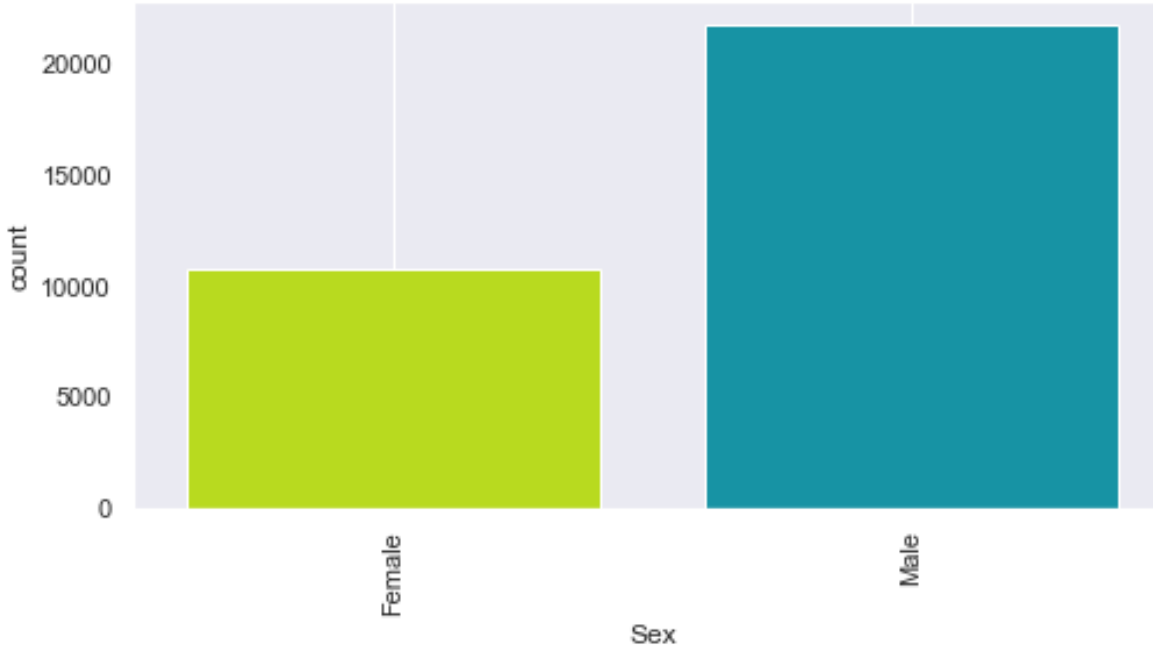
- Majority of the people in the data are White.
- Cannot conclude on the differences between races as sample size of other groups is relatively small.

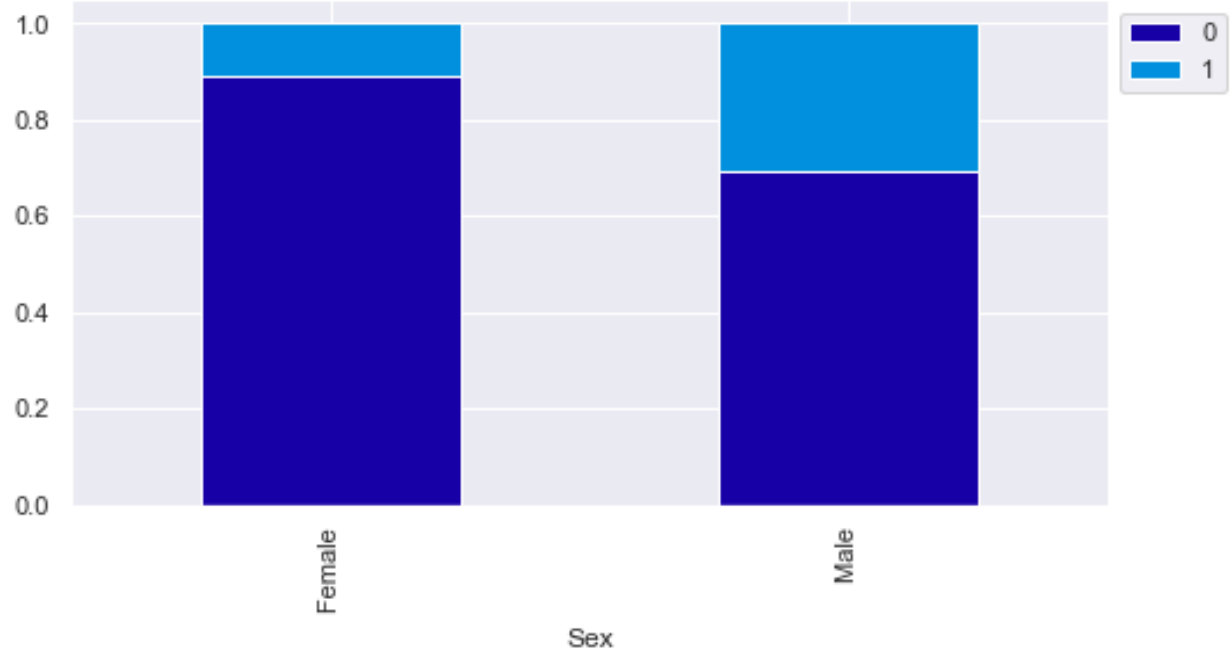




Exploratory Data Analysis – Sex

- There are more males in the dataset.
- Male appears to have higher percentage of income >50K than female.

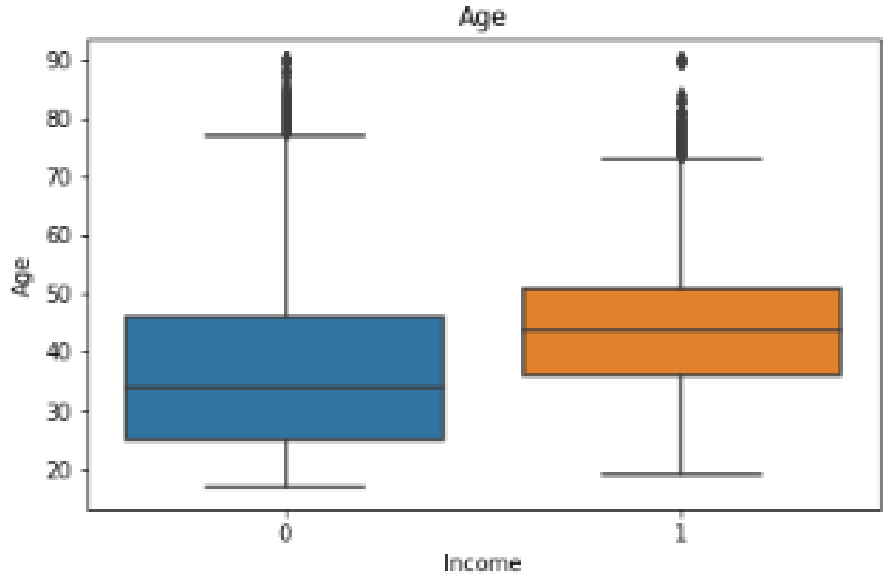


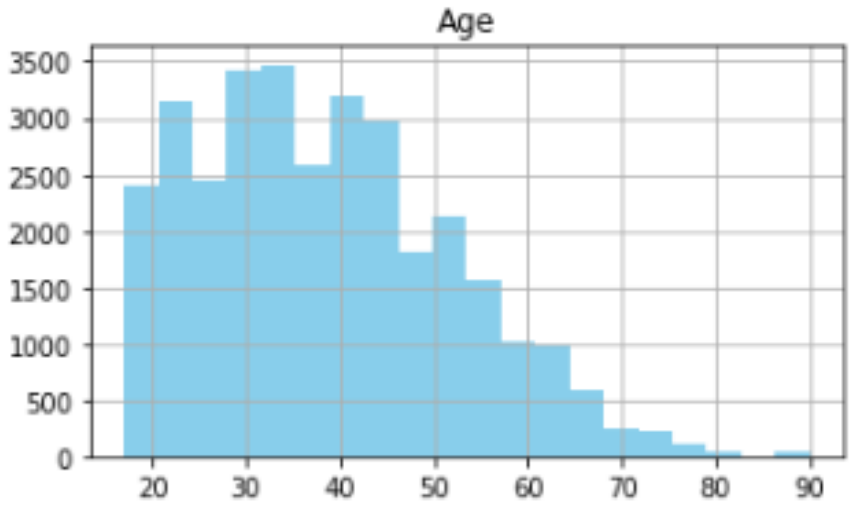


Exploratory Data Analysis – Age

- Majority are in 20 to 55 age range.
- People in >50K income group are generally older.

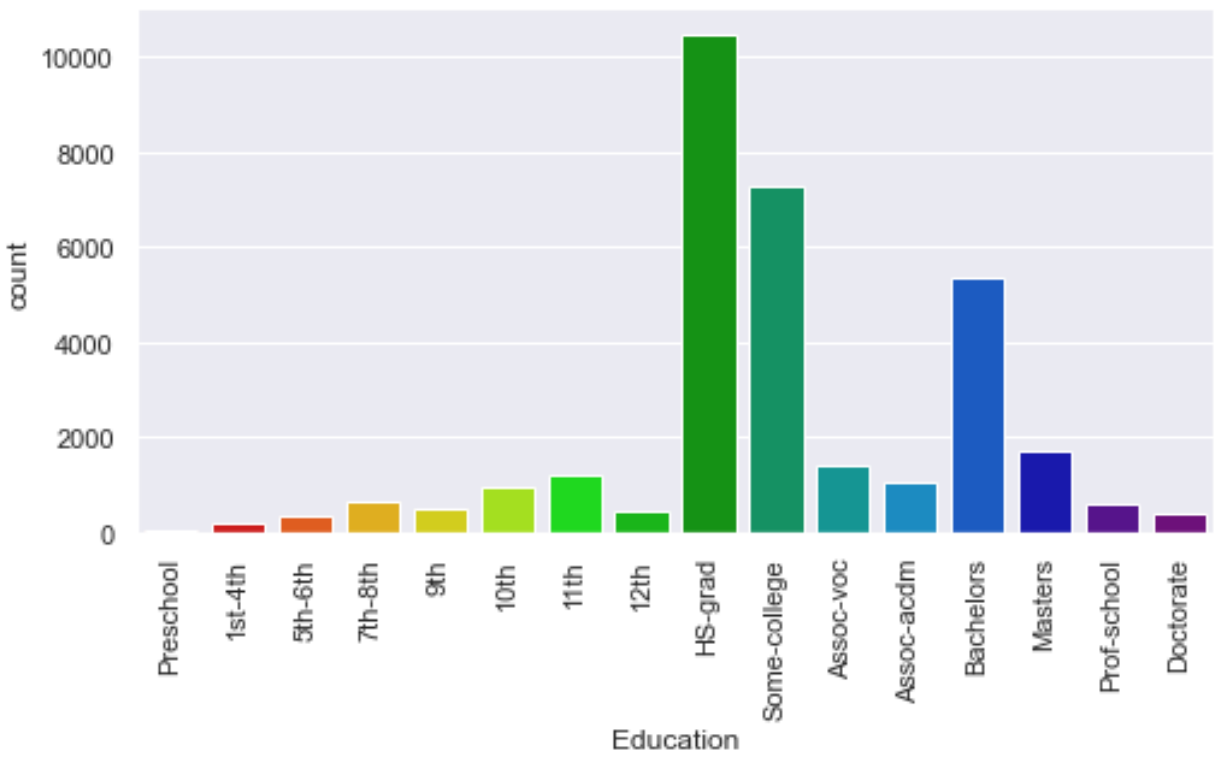
Age	Income <=50K	Income >50K
25th	25	36
Median	34	44
75th	46	51

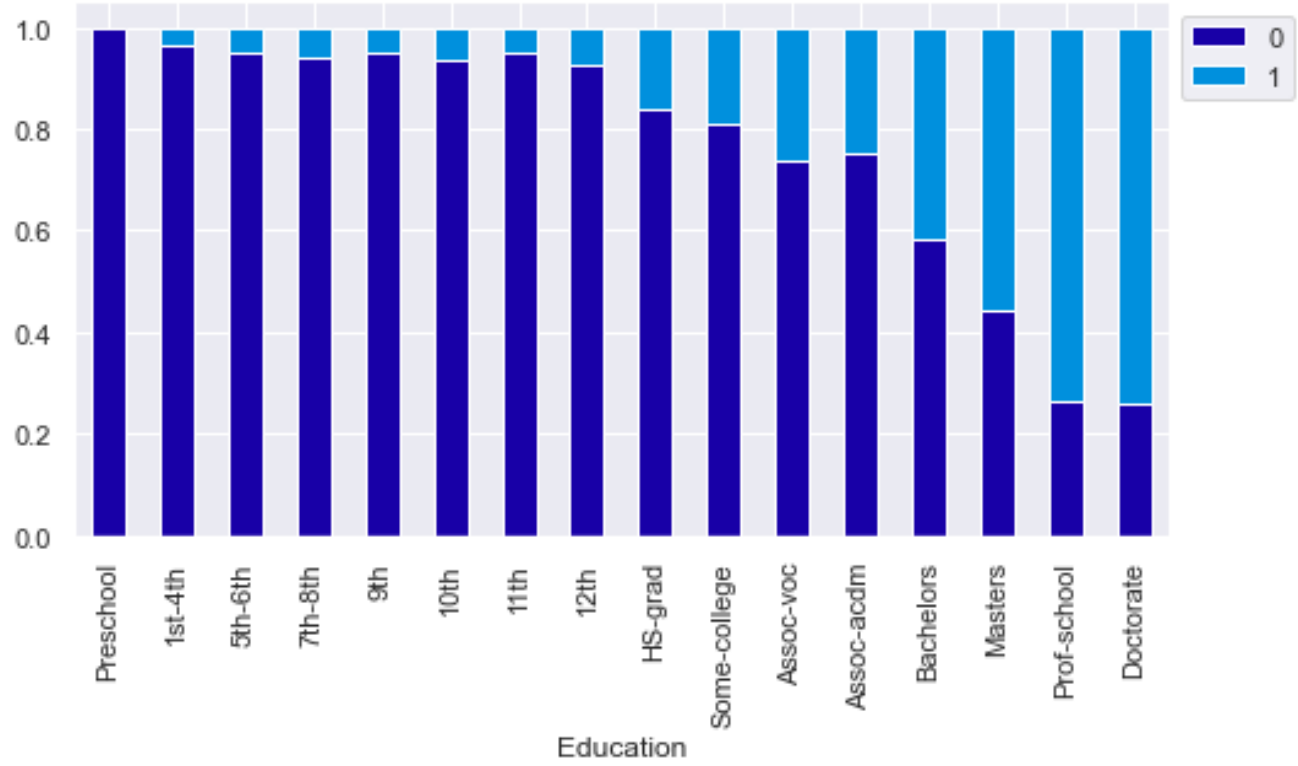




Exploratory Data Analysis – Education

- The higher the education level, the higher proportion of people having income >50K.
  - Bachelors: 40%+
  - Masters, Prof-school, Doctorate 50%+ of people income >50K.

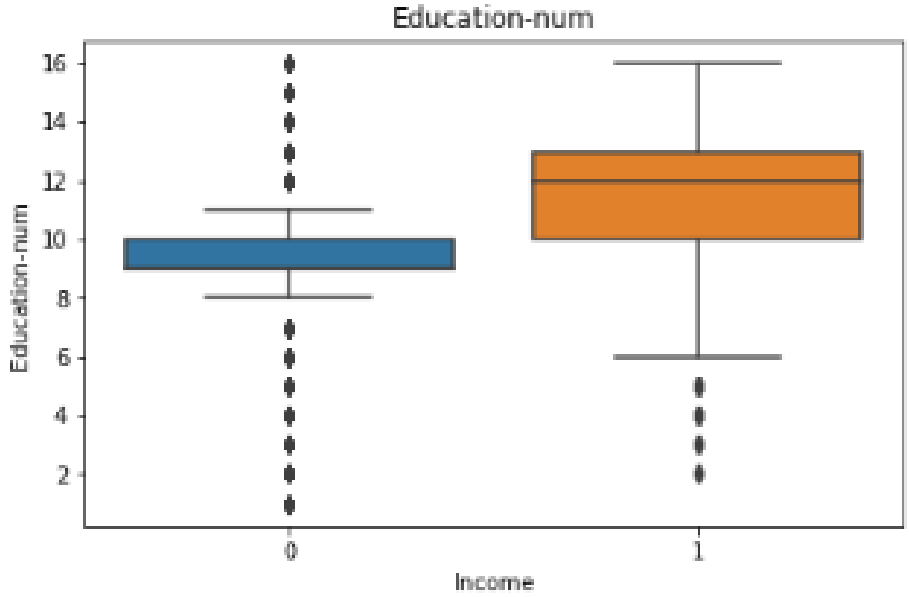


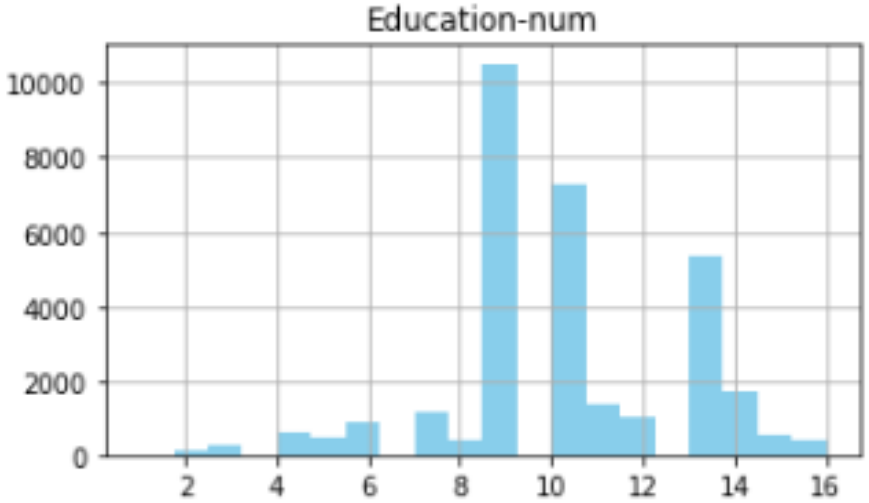


Exploratory Data Analysis – Education Years

- Similar to previous slide, people in >50K income group generally have higher number of years of education:

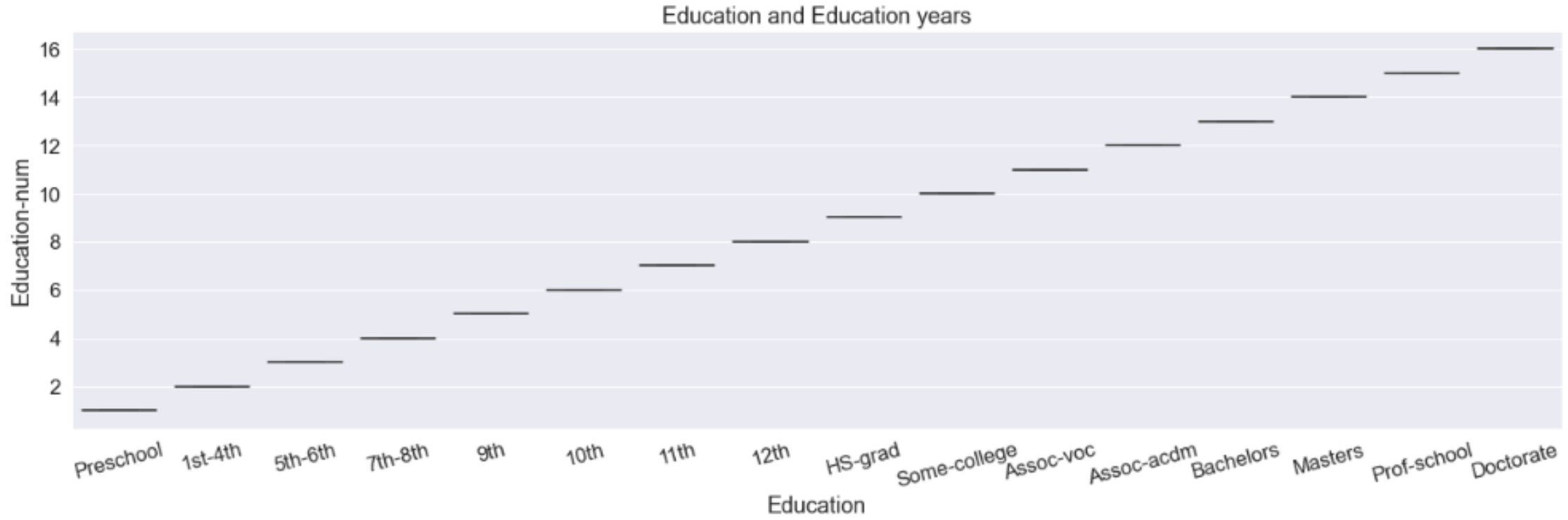
Edu years	Income <=50K	Income >50K
25th	9	10
Median	9	12
75th	10	13





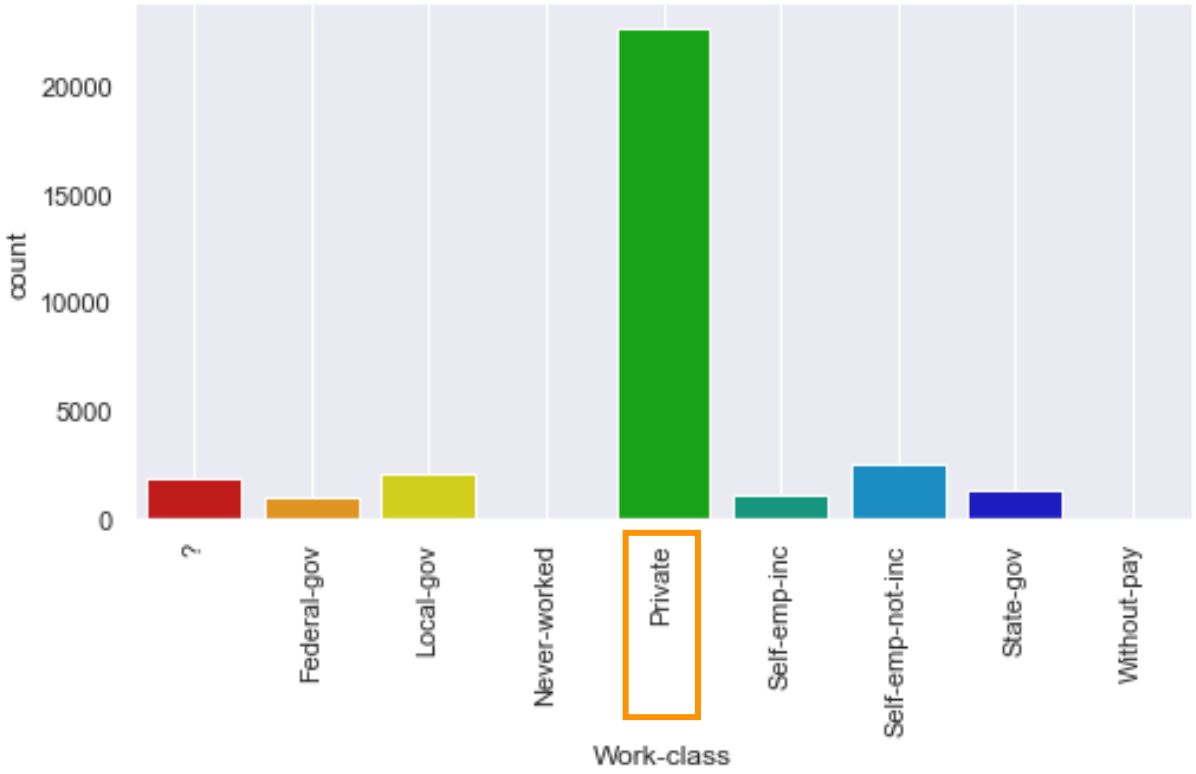
Exploratory Data Analysis – Education and Education Years

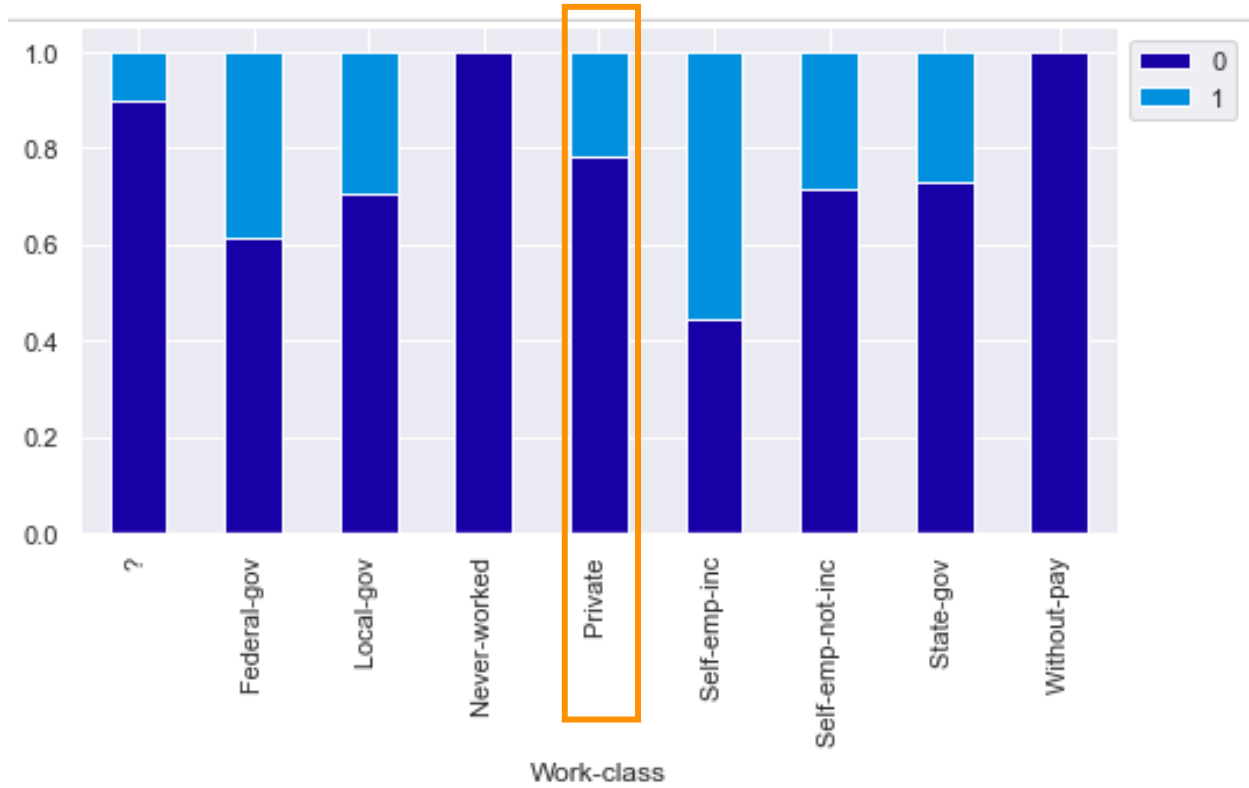
- In fact, Education and Education Years feature exactly correlates – there is a specific number of years for each education level.
- We will drop one of the features later.



Exploratory Data Analysis – Work class

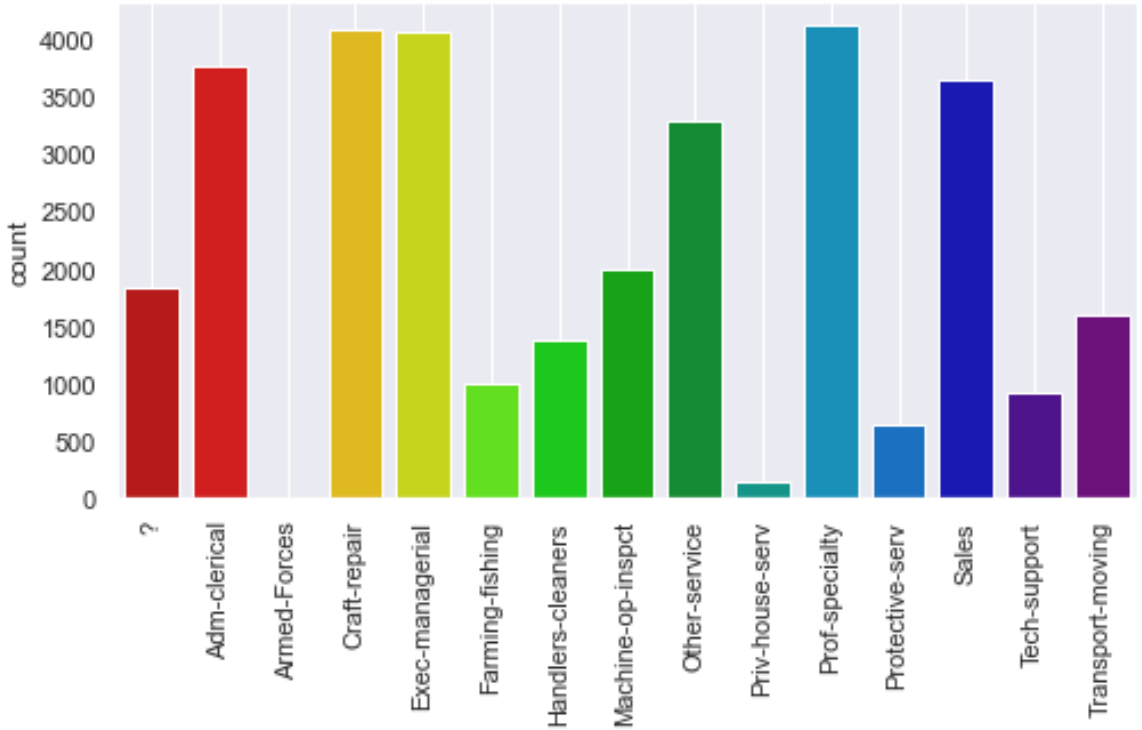
- Majority are ‘private’.
- The sample size of most groups other than ‘private’ may be too small to conclude about the difference in proportion of income >50K.

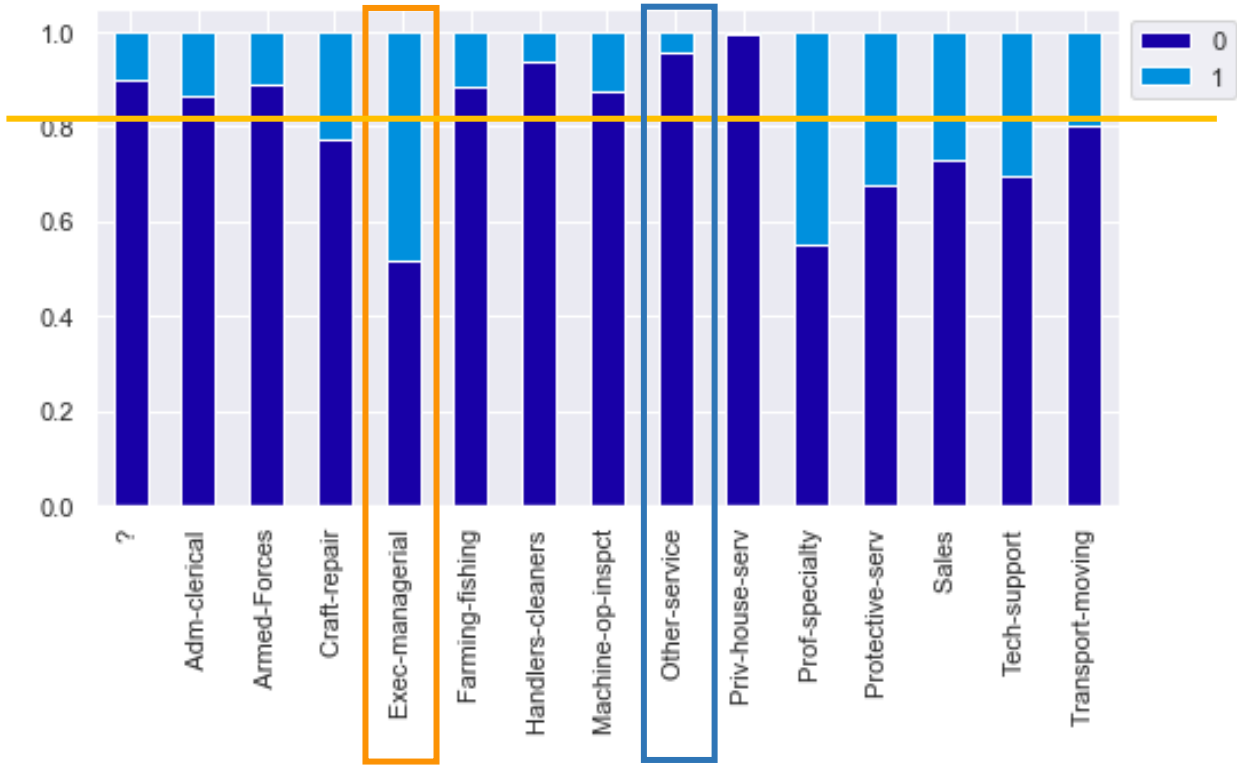




Exploratory Data Analysis – Occupation

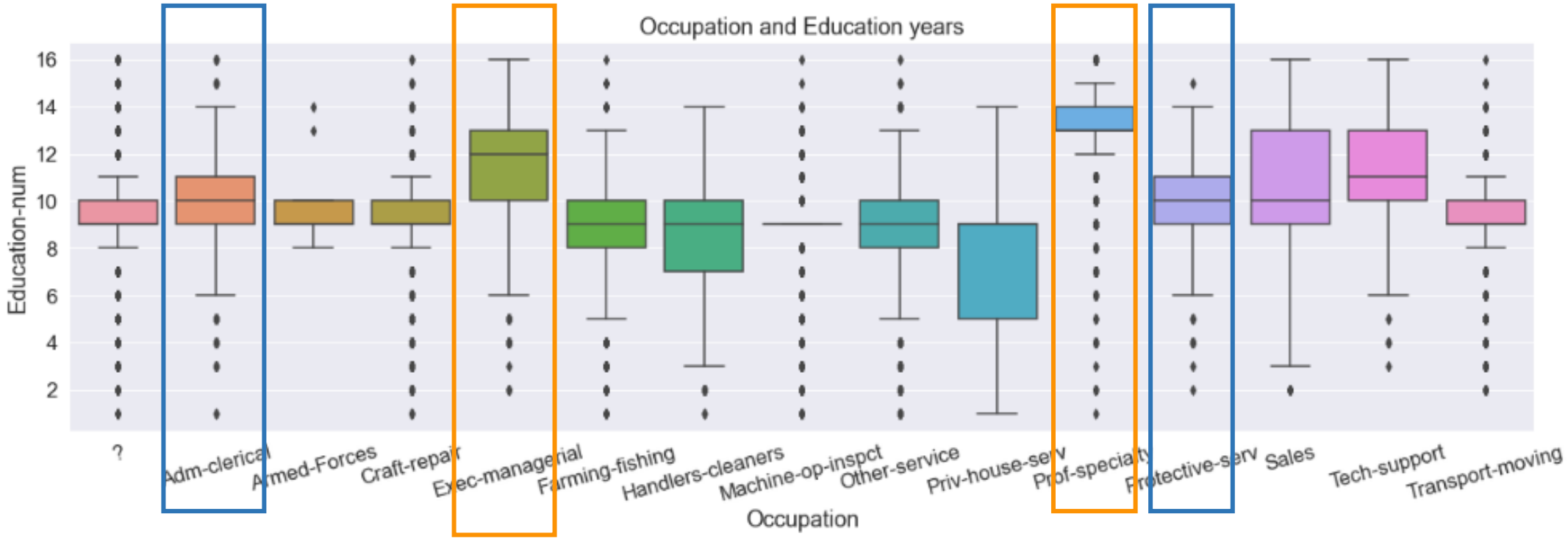
- Top 3 job in % of income >50K: Exec-management, Prof-specialty, Protective-service
- Bottom 3 job in % income >50K: Other-service, Handlers-cleaners, Admin-clerical





Exploratory Data Analysis – Occupation and Education Years

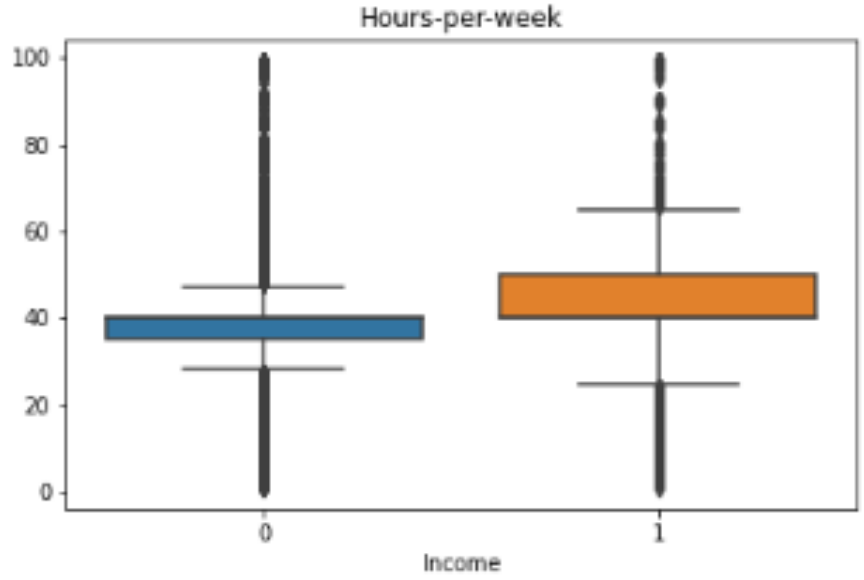
- High-paid jobs – executive managerial and professional specialty both have higher education years distribution.
- But education years is not especially high for another high-paid job – protective service.
- And ‘low-paid’ job – admin-clerical have similar education distribution with protective service.

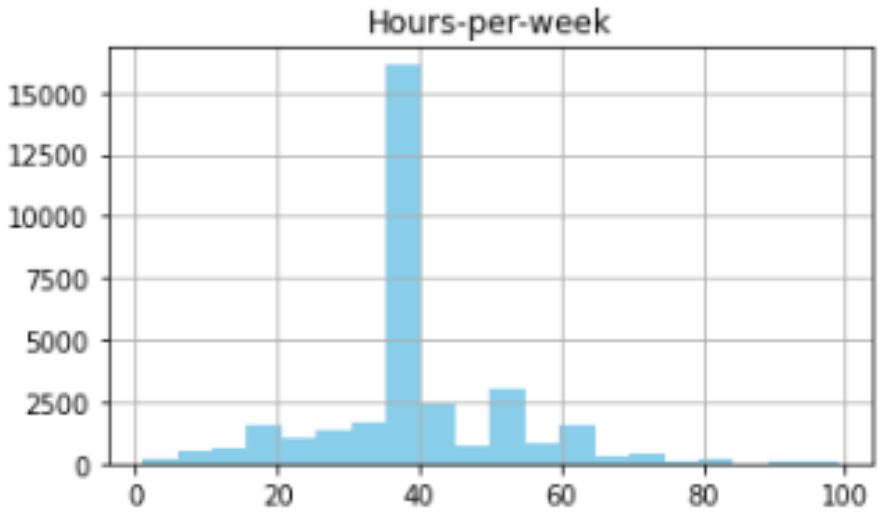


Exploratory Data Analysis – Work hours per week

- Most people work approx. 40 hours per week.
- People in >50K income group work slightly more hours a week:

Working hours	Income <=50K	Income >50K
25th	35	40
Median	40	40
75th	40	50



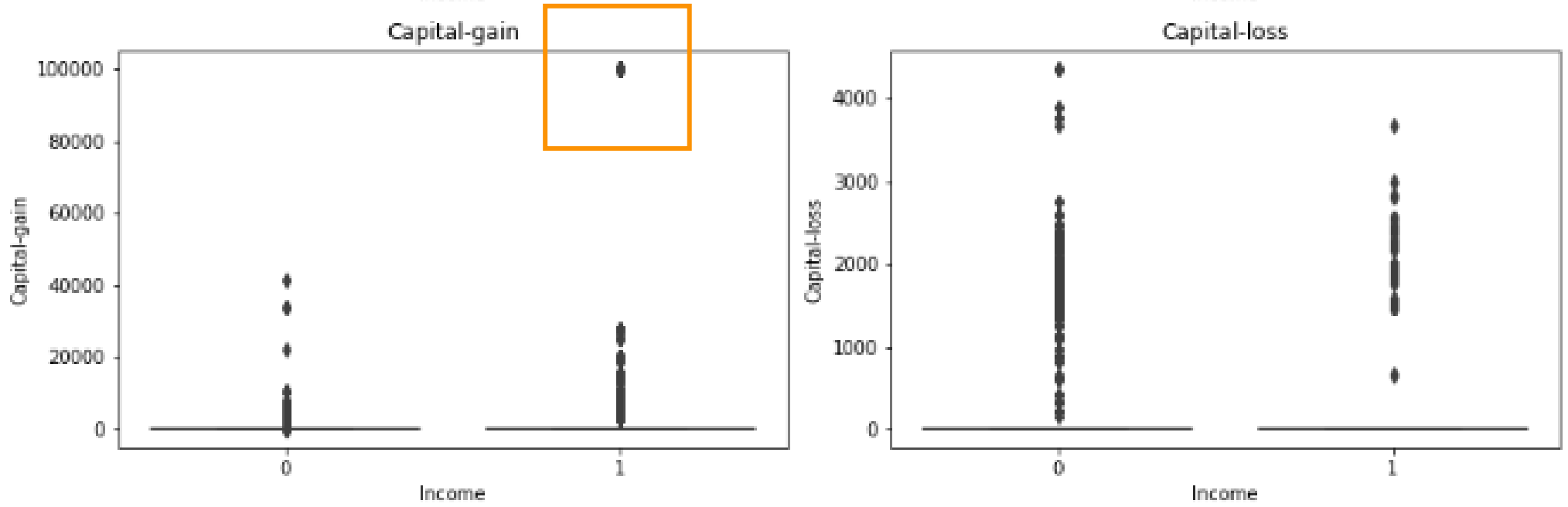


Exploratory Data Analysis – Occupation and Work hours per week

- Top 3 occupation in % of income >50K: work 40 hours or ***above*** per week in general
- Bottom 3 occupation in % income >50K: work 40 hours or ***below*** per week in general

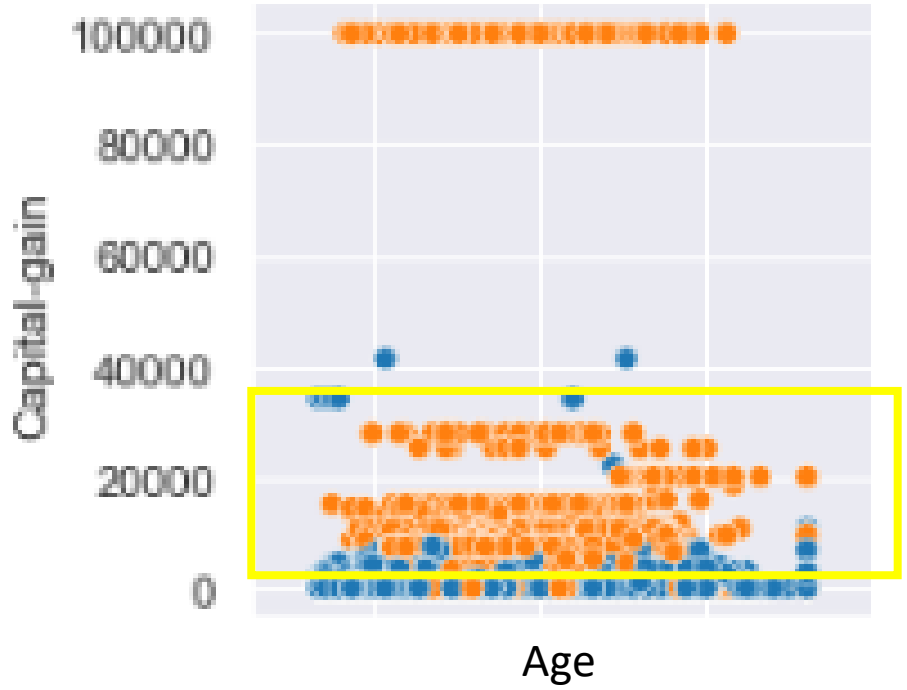
Exploratory Data Analysis – Capital gain and loss in the year

- Most people have zero capital gain and loss from the sale of capital assets in the year.
- There are some people having capital gain of 99999. The data appears weird and need to be treated.



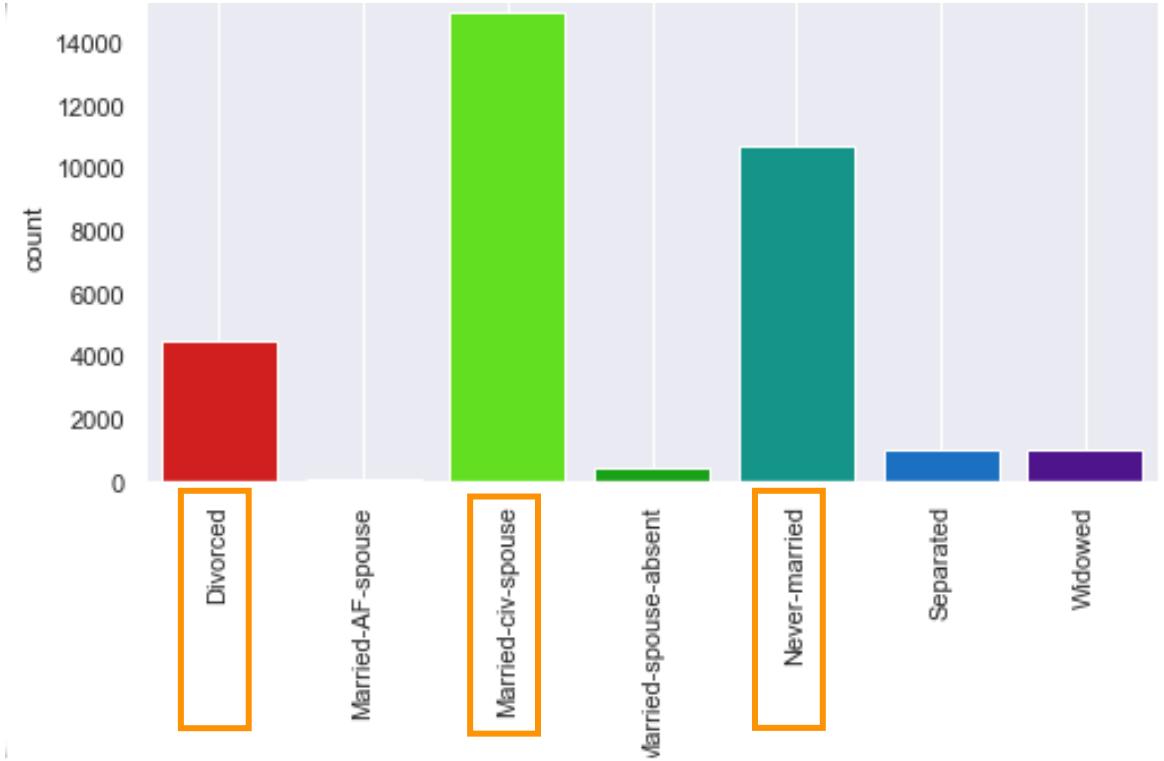
Exploratory Data Analysis – Capital gain and loss in the year

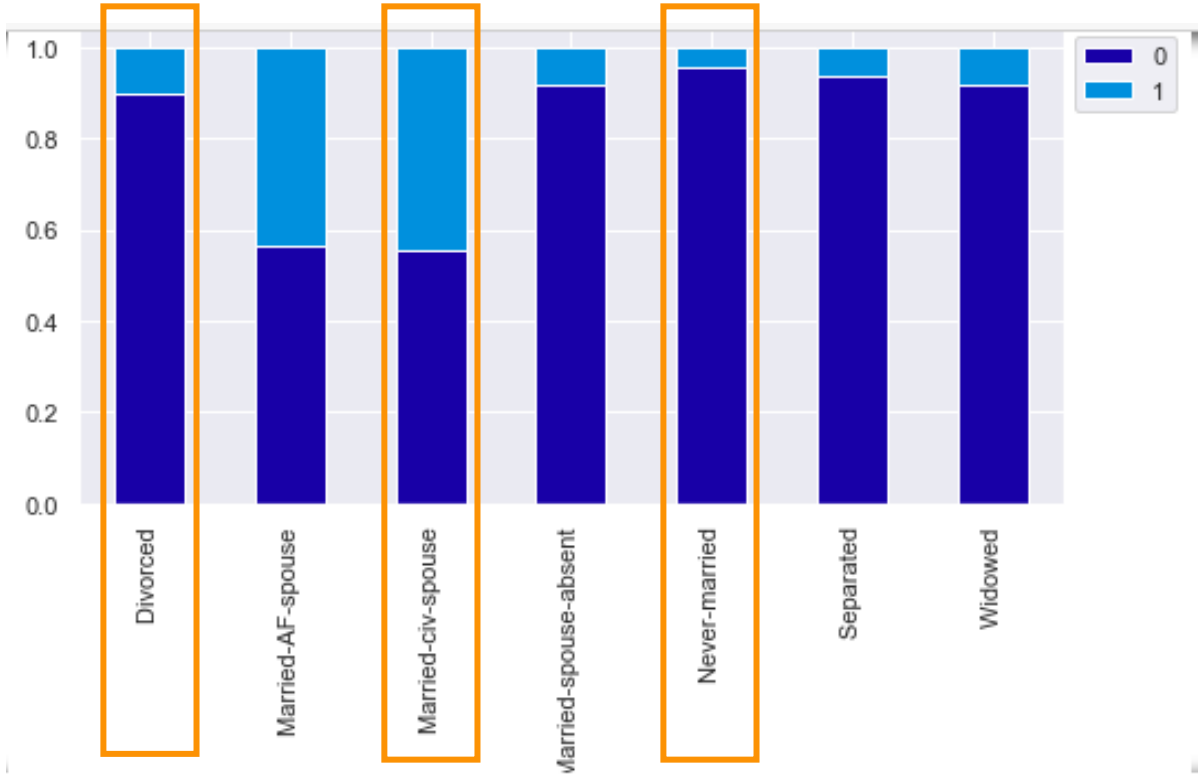
- If we look at the positive outliers only,
  - The positive capital gain mostly belongs to income >50K group.
  - Difference in positive capital loss appears not significant between two groups by scatter plots.



Exploratory Data Analysis – Marital Status

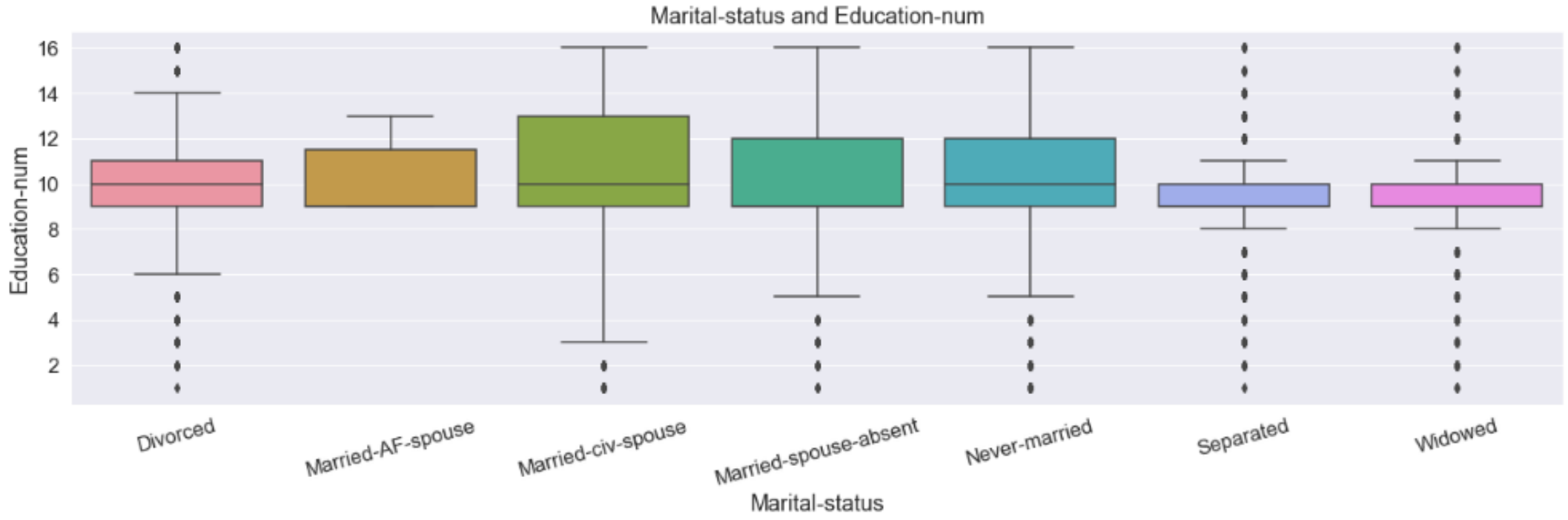
- Majority are never-married, married-civ-spouse or divorced.
- Married group shows highest percentage of income >50K.





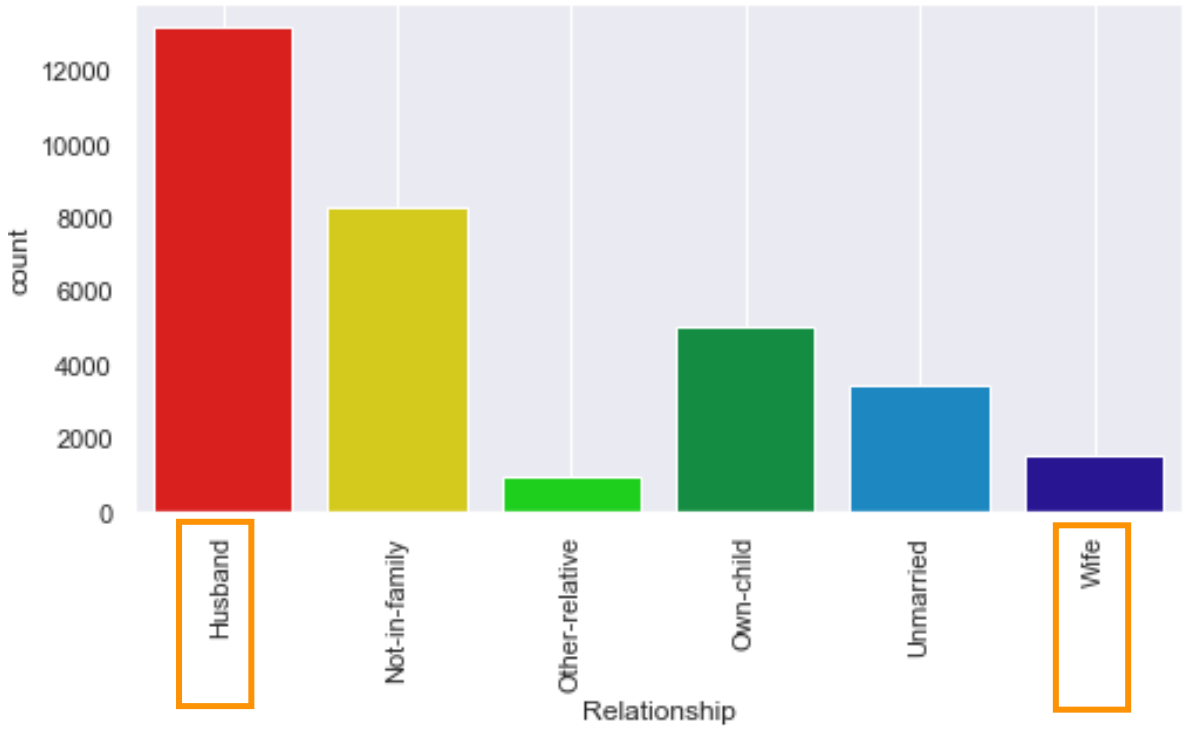
Exploratory Data Analysis – Marital Status and Education Years

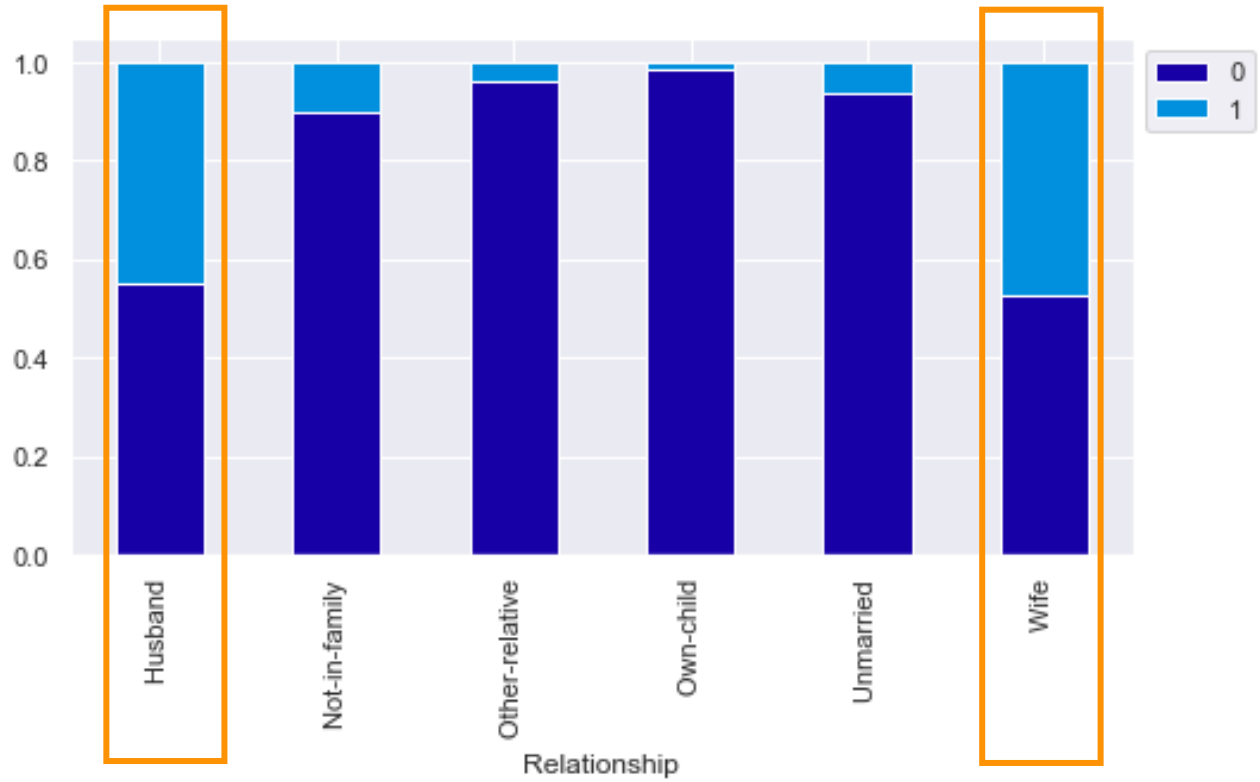
- Married and never-married group have similar distribution in education years.
- Divorced group has slightly narrower distribution but median is also similar.



Exploratory Data Analysis – Relationship

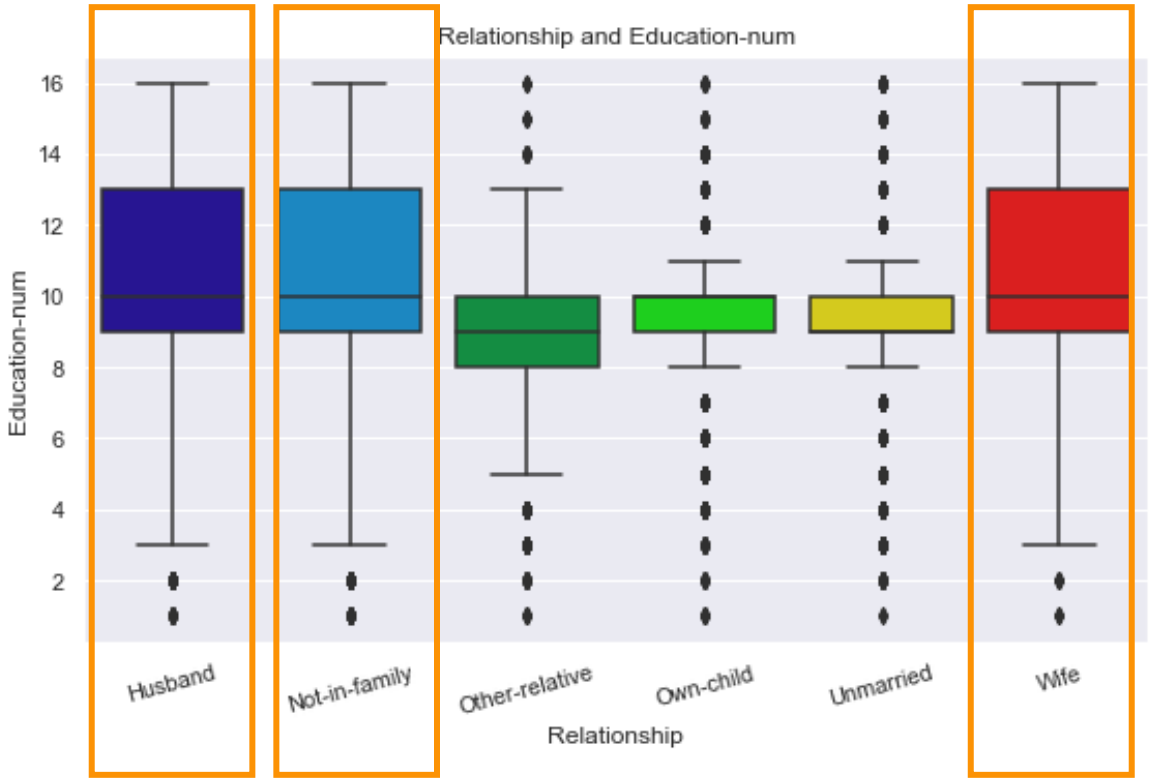
- Being ‘Husband’ or ‘Wife’ has highest and comparable percentage of income >50K.
- This is consistent with the pattern by marital status, where married group has highest rate of income >50K.

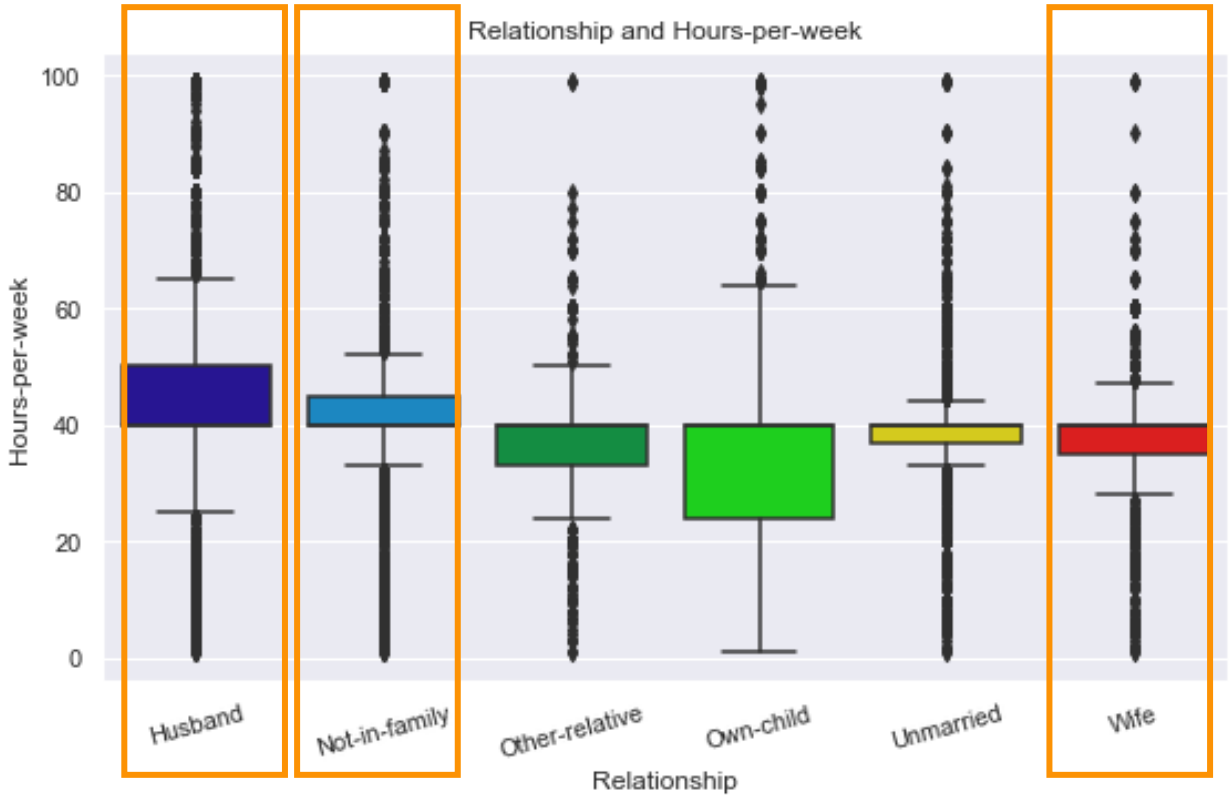




Exploratory Data Analysis – Relationship, Education Years, Hours-per-week

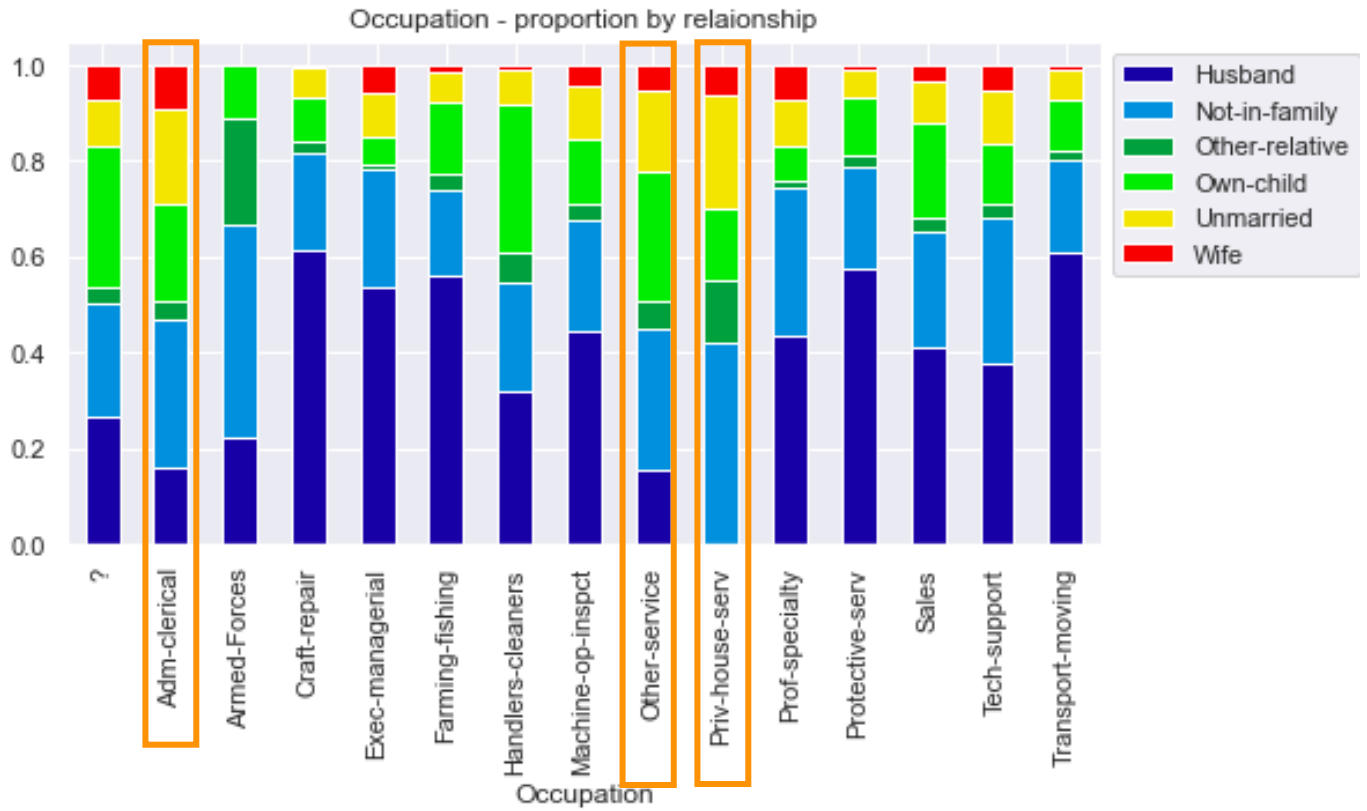
- Husband, Wife and Not-in-family have similar distribution in education years.
- But Husband work slightly longer hours in general than the other two groups.



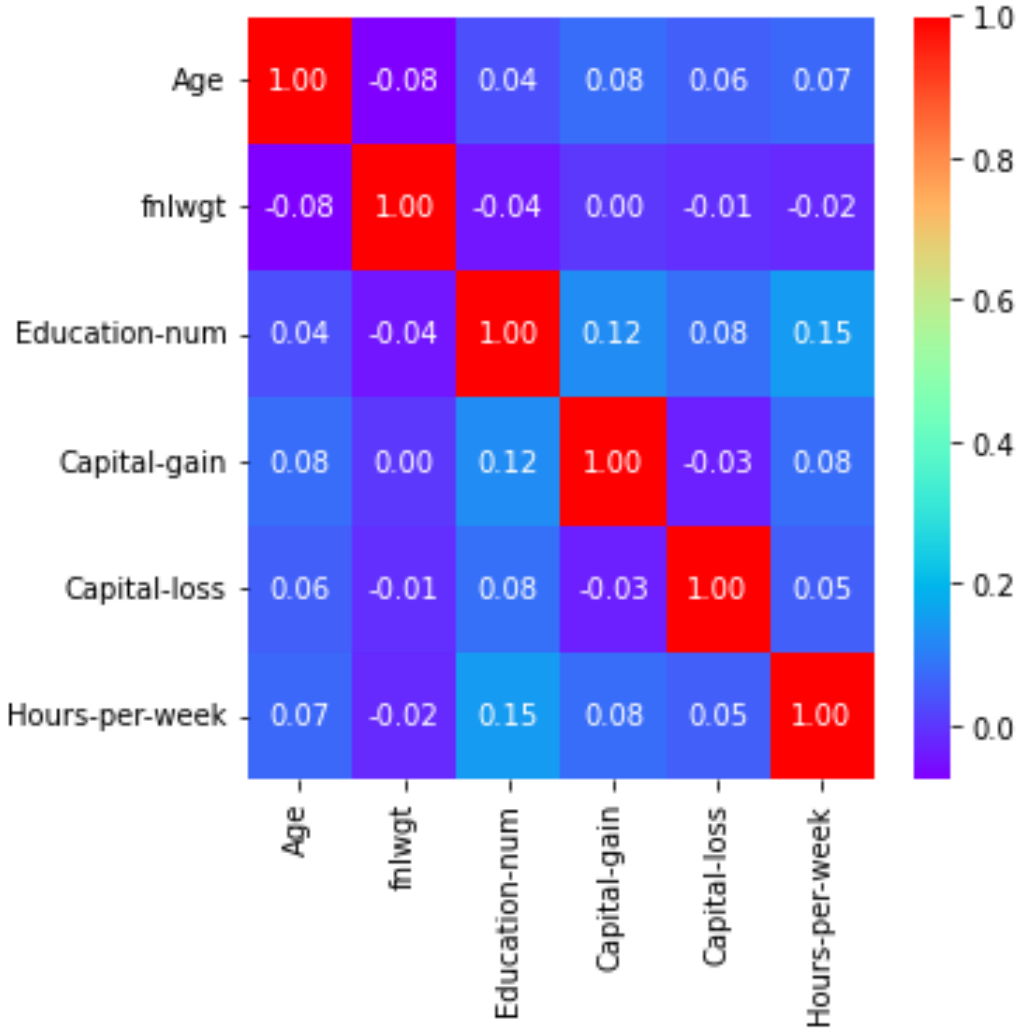


Exploratory Data Analysis – Relationship and Occupation

- Higher proportion of Wife and Not-in-family work in jobs which have lower working hours (as identified earlier), such as admin-clerical, private-house-servant and other-service.



Exploratory Data Analysis – Correlations between numerical features



- No significant correlations observed between numerical features though.

Exploratory Data Analysis – summary

Sex	• Higher proportion of males earning high income than females.
Age	• People in income >50K group are generally older.
Education	• People in income >50K group generally have higher education level attained. • Each education level has a specific number of education years.
Occupation	• People in occupation with higher rate of income >50K tend to have higher education level. This is not a must true though, e.g. protective-service vs admin-clerical.
Working hours per week	• People in income >50K group work slightly more hours. • High-paid occupation also tend to associate with longer working hours, and vice versa.

Exploratory Data Analysis – summary (cont)

Capital gain and loss	• Most people have zero capital gain and loss from the sale of capital assets in the year. • Most positive capital gain outliers belongs to >50K income group.
Marital status	• More people in married group earned high income than in never-married and divorced group.
Relationship	• Husband or Wife have highest and comparable rate of income >50K, consistent with marital status results. • Husband, Wife and Not-in-family have similar education distribution, but Wife and Not-in-family have shorter work hours / work in jobs that have shorter work hours.

Exploratory Data Analysis – summary (cont)

- In short, people in >50K income group may have these characteristics:
  - More likely to be male
  - Older
  - Have higher education level
  - Work longer hours
  - Married
  - Have some capital gain on asset during the year

Data Pre-processing

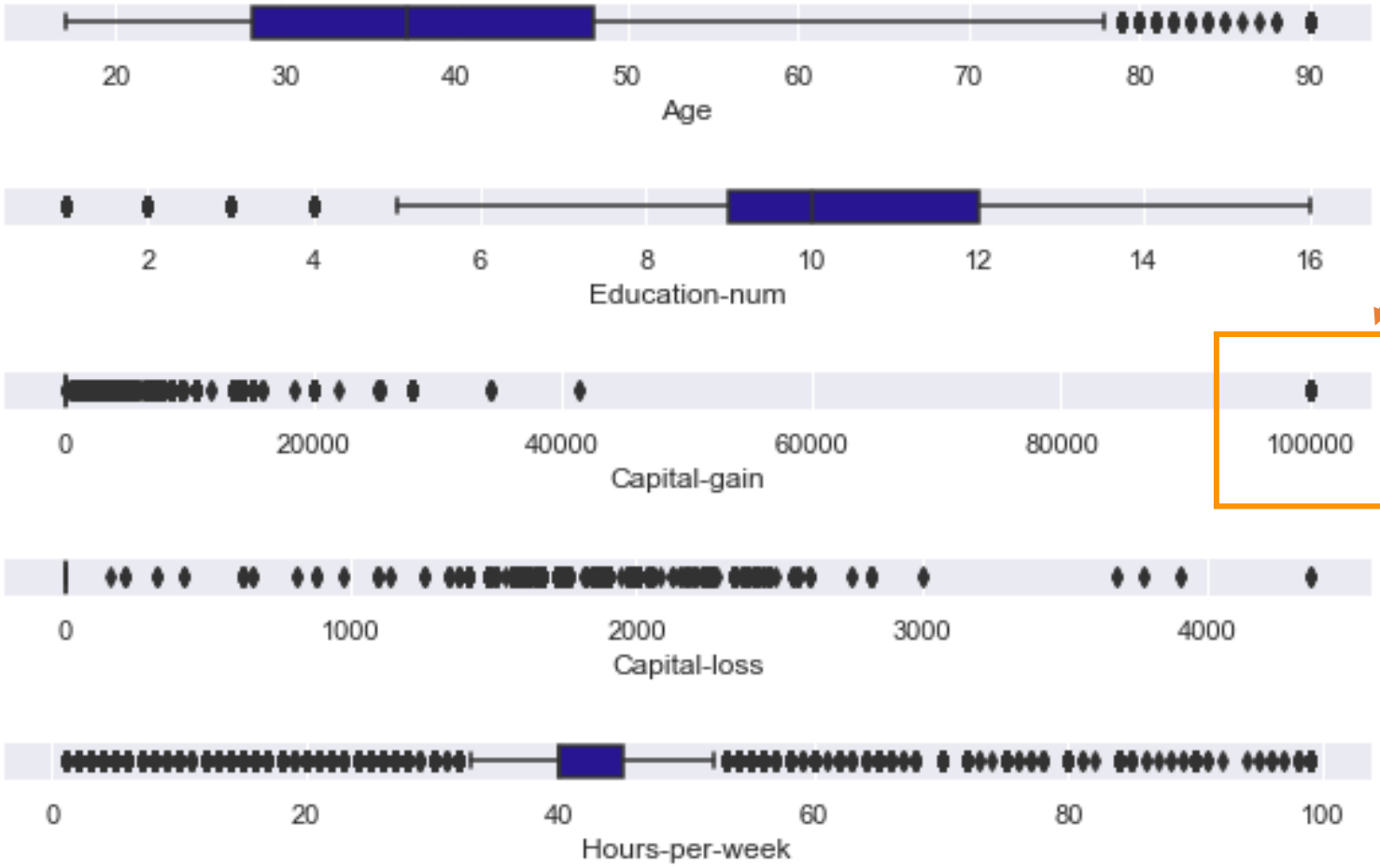
Data Pre-processing

- Below table shows the aspects considered and treatment methods:

Issues	EDA results	Treatment
Missing values	No missing values	Not required
Outliers	Certain unreasonable outliers exist	Replace with summary (mode)
Imbalanced dataset	Income <= 50K: 76% Income > 50K: 24%	Adjusting class weight parameter
Categorical attributes	Certain attributes having numerous low frequency categories	• Grouping very low frequency categories as ‘others’ • One-hot encoding
Different unit or scale of measurement	Except for capital-gain and capital-loss, the difference in absolute values across different numerical features is not very significant.	Only capital-gain and capital-loss were standardized by Z-Score standardization. Values of other features kept for explainability.

Dealing with outliers

Before treatment



- Multiple samples having same capital gain of 99,999 considered as abnormal.
- Other outliers considered reasonable.

Dealing with outliers

After treatment – capital gain

- Replaced the 99,999 with zero, which is the mode of the samples.

Dealing with imbalanced dataset

- Below table shows an example which Logistic Regression model results were compared across different methods.
- Test result:

Treatment method	AUC	Precision	Recall	F1-score
No treatment	0.90	0.73	0.60	0.66
Adjusted Class Weight	0.90	0.56	0.84	0.67
Random Under Sampling	0.90	0.56	0.85	0.68
Random Over Sampling	0.90	0.56	0.84	0.67
SMOTE	0.90	0.66	0.68	0.67

Note: Precision, Recall, F1-score indicate minority class result (Income>50K)

- Model performance is not significantly different across different imbalanced data treatment methods.
- We will pick adjusted class weight in the models training.

Feature Selection

Dropped features

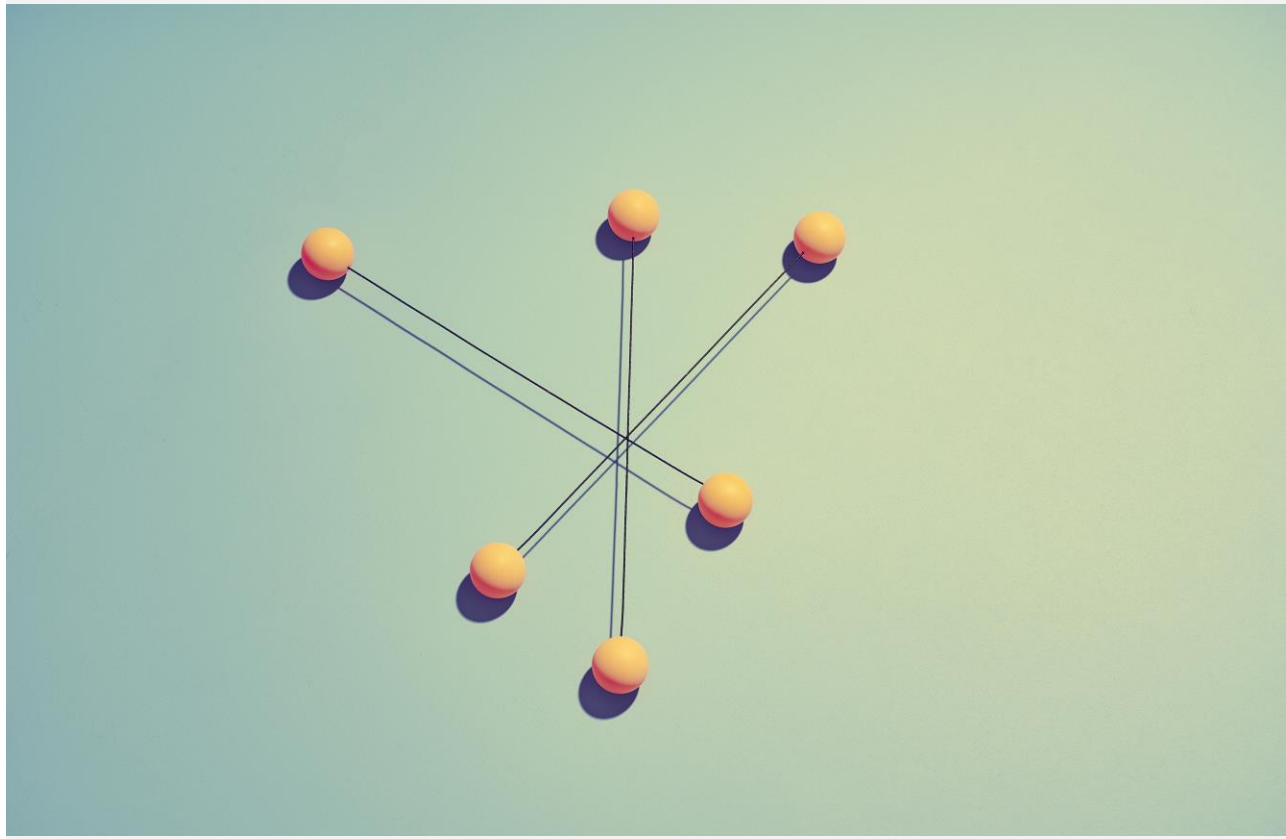
- Education
  - Completely correlate with education years
- Final weight of record
  - It is ‘the number of people that the U.S. Census estimates that the specific record represents’. Considered not useful for our prediction purpose as we only want to focus on the characteristics of high and low income group regardless of how popular they are in the population.

Based on exploratory data analysis, the remaining features are considered relevant.

Finalised list of features for model building

0	Age	32537 non-null	int64
1	Education-num	32537 non-null	int64
2	Capital-gain	32537 non-null	float64
3	Capital-loss	32537 non-null	float64
4	Hours-per-week	32537 non-null	int64
5	Income	32537 non-null	int64
6	Work-class_Federal-gov	32537 non-null	uint8
7	Work-class_Local or State gov	32537 non-null	uint8
8	Work-class_Others	32537 non-null	uint8
9	Work-class_Private	32537 non-null	uint8
10	Work-class_Self-emp-inc	32537 non-null	uint8
11	Work-class_Self-emp-not-inc	32537 non-null	uint8
12	Marital-status_Divorced	32537 non-null	uint8
13	Marital-status_Married-civ-spouse	32537 non-null	uint8
14	Marital-status_Never-married	32537 non-null	uint8
15	Marital-status_Others	32537 non-null	uint8
16	Marital-status_Separated	32537 non-null	uint8
17	Occupation_Adm-clerical	32537 non-null	uint8
18	Occupation_Craft-repair	32537 non-null	uint8
19	Occupation_Exec-managerial	32537 non-null	uint8
20	Occupation_Farming-fishing	32537 non-null	uint8
21	Occupation_Handlers-cleaners	32537 non-null	uint8
22	Occupation_Machine-op-inspct	32537 non-null	uint8
23	Occupation_Other-service	32537 non-null	uint8
24	Occupation_Others	32537 non-null	uint8
25	Occupation_Prof-specialty	32537 non-null	uint8
26	Occupation_Protective-serv	32537 non-null	uint8
27	Occupation_Sales	32537 non-null	uint8
28	Occupation_Tech-support	32537 non-null	uint8
29	Occupation_Transport-moving	32537 non-null	uint8
30	Relationship_Husband	32537 non-null	uint8
31	Relationship_Not-in-family	32537 non-null	uint8
32	Relationship_Other-relative	32537 non-null	uint8
33	Relationship_Own-child	32537 non-null	uint8
34	Relationship_Unmarried	32537 non-null	uint8
35	Relationship_Wife	32537 non-null	uint8
36	Race_Black	32537 non-null	uint8
37	Race_Others	32537 non-null	uint8
38	Race_White	32537 non-null	uint8
39	Sex_Female	32537 non-null	uint8
40	Sex_Male	32537 non-null	uint8
41	Native-country_Others	32537 non-null	uint8
42	Native-country_United-States	32537 non-null	uint8

Model Building



Model Building - Models

5 models are explored:

- Linear Classifier
  - Logistic Regression
- Decision Tree
- Ensemble Models
  - Random Forest
  - XGBoost
  - LightGBM

Model Building - Model Parameters

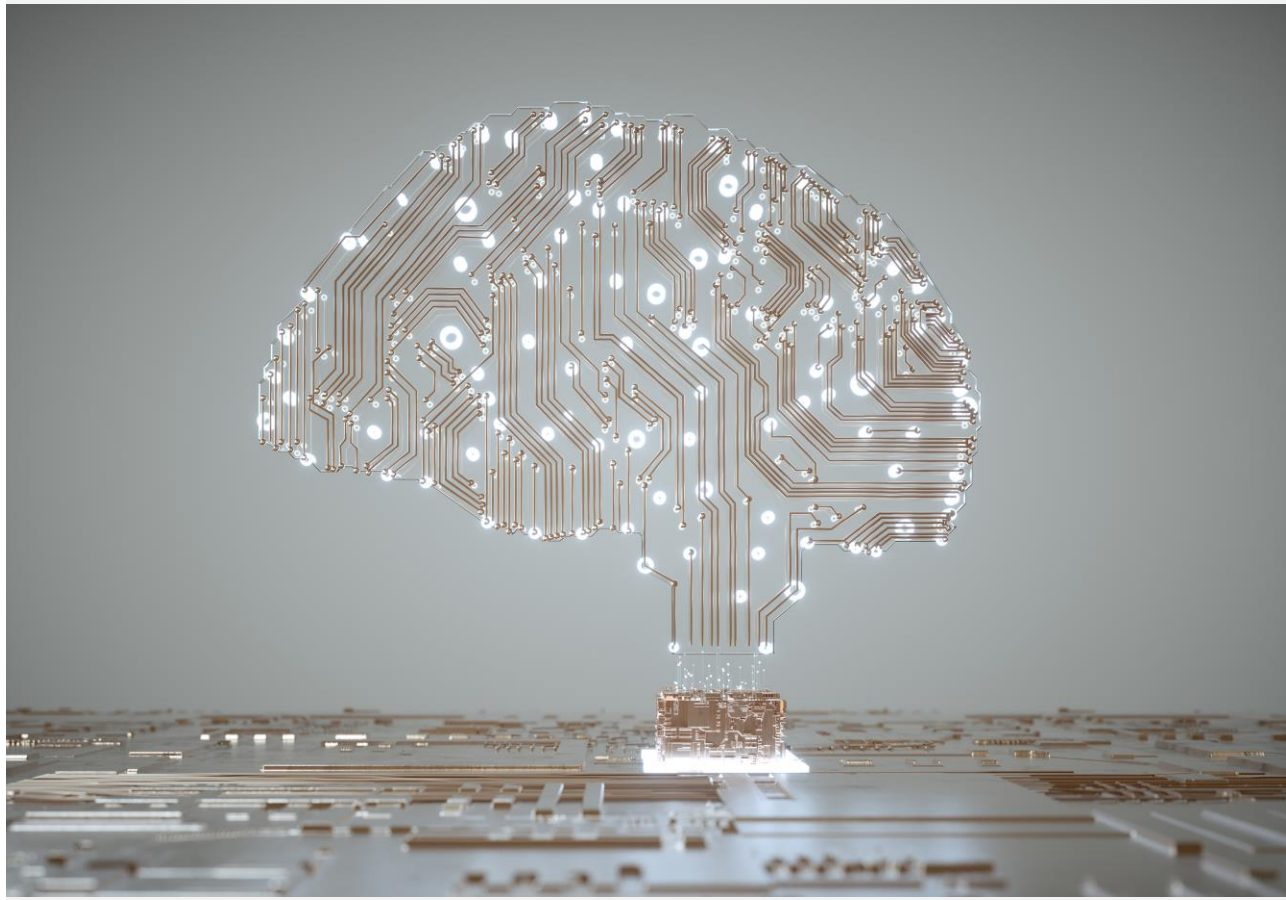
Model	Major pre-set parameters	Hyperparameters tuning
Logistic Regression	penalty = L2, class_weight = balanced	• C: 0.01, 0.1, 0.5, 1, 10, 100
Decision Tree	min_samples_split = 2, min_samples_leaf = 1, max_leaf_nodes = None, class_weight = balanced	• Max_depth: 4, 6, 8, 10, 15 • Criterion: entropy, gini
Random Forest	n_estimators=100, max_samples = None, min_samples_split = 2, min_samples_leaf = 1, max_leaf_nodes = None, class_weight = balanced	• Max_features: None, sqrt, 0.33, 0.5 • Max_depth: 6, 8, 10, 15 • Criterion: entropy, gini
XGBoost	min_child_weight=1, gamma=0, reg_alpha=0, reg_lambda=1, objective='binary:logistic'	• n_estimators: 100,200 • max_depth: 6, 10 • subsample: 0.5, 0.8, 1 • colsample_bytree: 0.5, 0.8, 1 • learning_rate: 0.1, 0.3
LightGBM	num_leaves=600, objective='binary', max_bin=255, lambda_l1=0, lambda_l2=0	• n_estimators: 50, 100, 200 • max_depth: 6, 8, 10 • feature_fraction: 0.5, 0.8, 1 • min_data_in_leaf: 100, 500, 1000 • learning_rate: 0.1, 0.2, 0.3

Model Building - Considerations

- GridSearchCV was utilized in hyperparameters tuning
  - Cross-validation - 5-fold, scoring = roc_auc
- Hyperparameters selected for tuning with these considerations in mind:
  - Balancing bias-variance trade-off
  - Model training speed and processing capacity

    -> Searching for all possible set of hyperparameters is not possible with regular laptop
- Train-test-split: 0.7 : 0.3, stratified by target variable

Model Performance



Model Best Parameters

Model	Major pre-set parameters	Best parameters
Logistic Regression	penalty = L2, class_weight = balanced	• C: 0.5
Decision Tree	min_samples_split = 2, min_samples_leaf = 1, max_leaf_nodes = None, class_weight = balanced	• Max_depth: 8 • Criterion: entropy
Random Forest	n_estimators=100, max_samples = None, min_samples_split = 2, min_samples_leaf = 1, max_leaf_nodes = None, class_weight = balanced	• Max_features: 0.33 • Max_depth: 15 • Criterion: entropy
XGBoost	min_child_weight=1, gamma=0, reg_alpha=0, reg_lambda=1, objective='binary:logistic'	• n_estimators: 200 • max_depth: 6 • subsample: 1 • colsample_bytree: 0.5 • learning_rate: 0.1
LightGBM	num_leaves=600, objective='binary', max_bin=255, lambda_l1=0, lambda_l2=0	• n_estimators: 200 • max_depth: 6 • feature_fraction: 0.5 • min_data_in_leaf: 100 • learning_rate: 0.1

Evaluation Metrics

- Since the dataset is imbalanced, Accuracy is not a suitable evaluation metrics as it will be heavily skewed towards majority class.
- Instead, we will focus more on below metrics:
  - AUC
  - Precision (what percent of positive predictions were correct)
  - Recall (how many positive cases did the model catch)
  - F1 score
- For this case study, we don't consider which class is more important than the other. The cost of making incorrect predictions in terms of false positives or false negatives is also comparable. We only need a more accurate result overall for both class and in both scenario (precision and recall).
- Therefore we would use F1 score as the major metric.

$$F\text{ - measure} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall} + \text{Precision}}$$

Model Results

Below table summarizes the performance for each model:

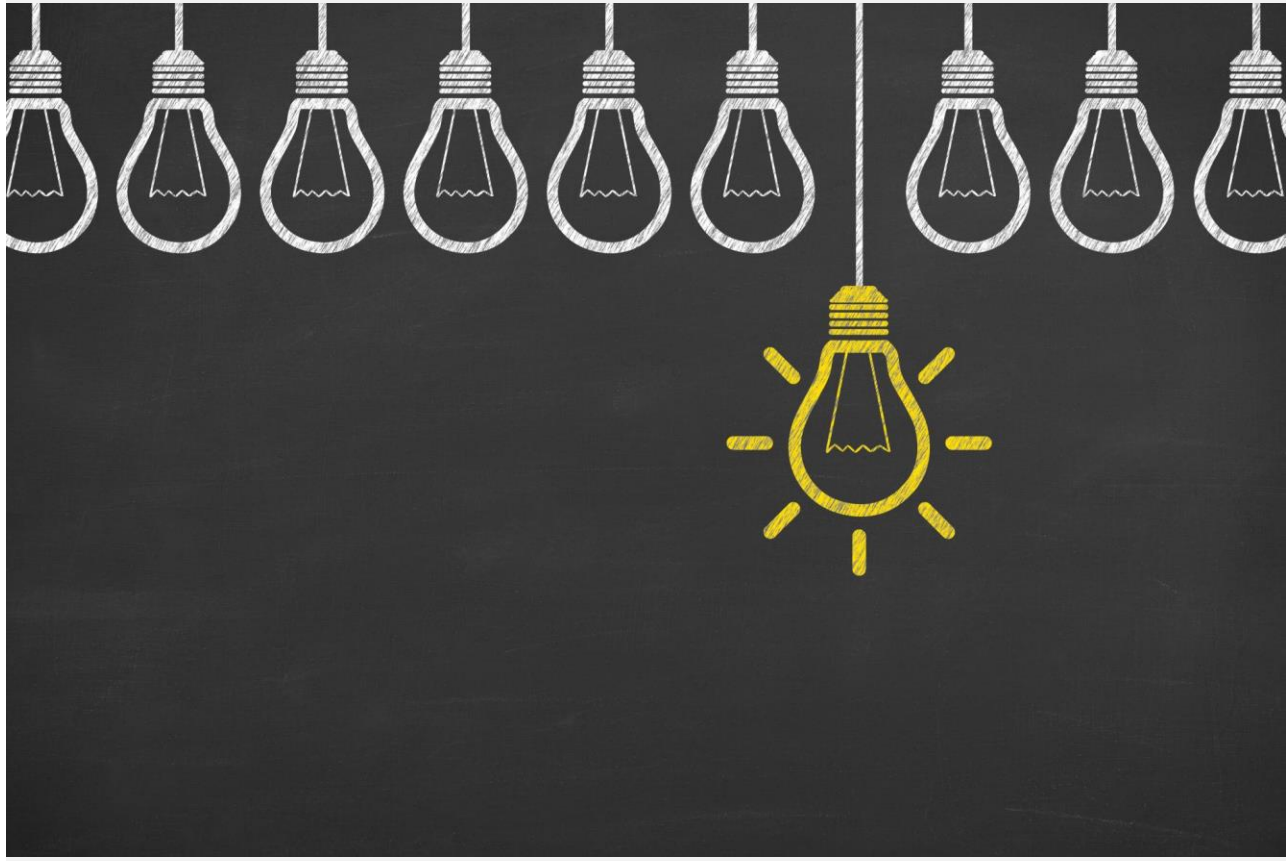
		Majority class (Income <=50K)			*Minority class (Income >50K)*		
Model	**AUC**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Logistic Regression	0.90	0.94	0.79	0.86	0.56	0.84	0.67
Decision Tree	0.90	0.95	0.77	0.85	0.54	0.86	0.67
Random Forest	0.91	0.94	0.83	0.88	0.60	0.82	0.69
XGBoost	0.93	0.89	0.94	0.92	0.78	0.65	0.71
LightGBM	0.93	0.89	0.94	0.92	0.78	0.65	0.71

- AUC is high across all models (>0.90), indicating all are good ranking classifiers.
- XGBoost and LightGBM model achieve highest F1 scores for both majority and minority class.

Interpret Best Model

Ensemble methods like LightGBM show improvement in performance due to greater complexity but there is loss of interpretability, as compared to a simpler decision tree.

In this section we will utilise several explainable AI methods to interpret the model results.



Model Interpretability

Specifically, we will address these questions:

Global Feature Importance

- What features are most influential to the model's output?
- What are the relative sizes of feature values influencing the model?
- Does a feature influence the output to be a lower ('0': income <=50K) or higher number ('1': income >50K)?
- How strong are the influences?

Global Feature Behavior

- What is the marginal effect of a feature on the predicted outcome?
- What is the relationship between feature and the predicted outcome?

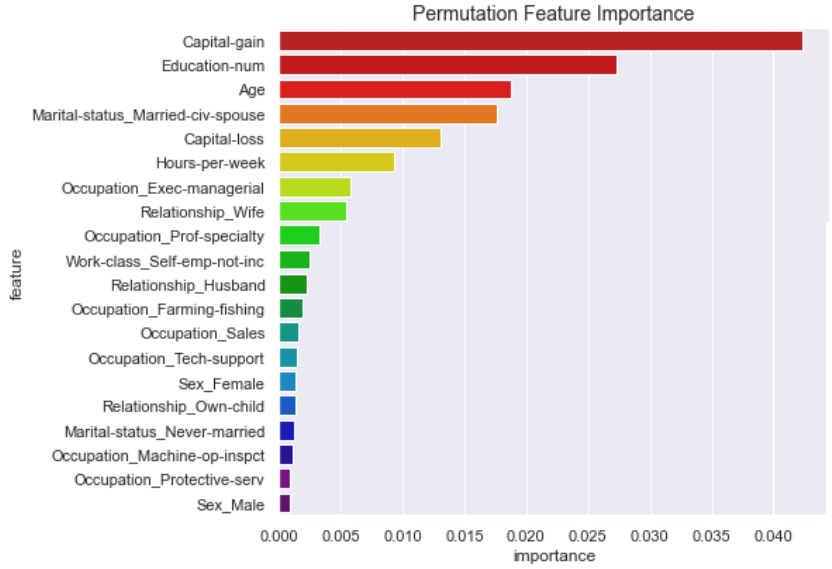
Local Feature Importance

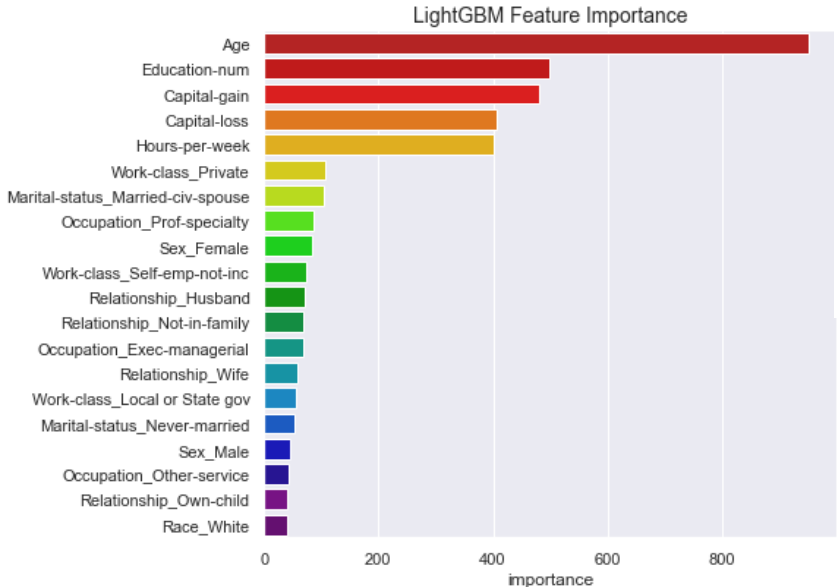
- Same questions as global feature importance but applying to a specific data record.

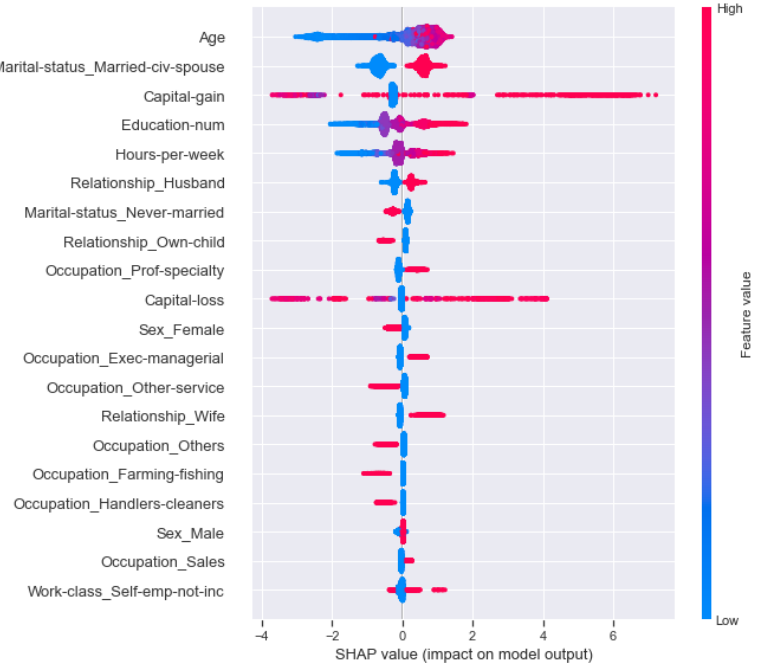
Top 5 influential features to LightGBM model output

Three methods to compute feature importance are explored:

- Permutation feature importance
- LightGBM feature importance
- Shapley Additive exPlanations feature importance







Top 5 influential features to LightGBM model output

- Top 5 ranking slightly different.
- Permutation vs LightGBM :
  - The main difference is Married-yes/no feature identified as top 5 in Permutation and Work hours per week identified as top 5 in LightGBM instead.
  - Impurity-based feature importance of LightGBM has more tendency to biased towards high cardinality features. This may explain why work hours per week is picked over the Married-yes/no feature.

	Permutation feature importance	LightGBM feature importance	SHAP feature importance
1	Capital gain	Age	Age
2	Education years	Education years	Married-yes/no
3	Age	Capital gain	Capital gain
4	Married-yes/no	Capital loss	Education years
5	Capital loss	Work hours per week	Work hours per week

Top 5 influential features to LightGBM model output

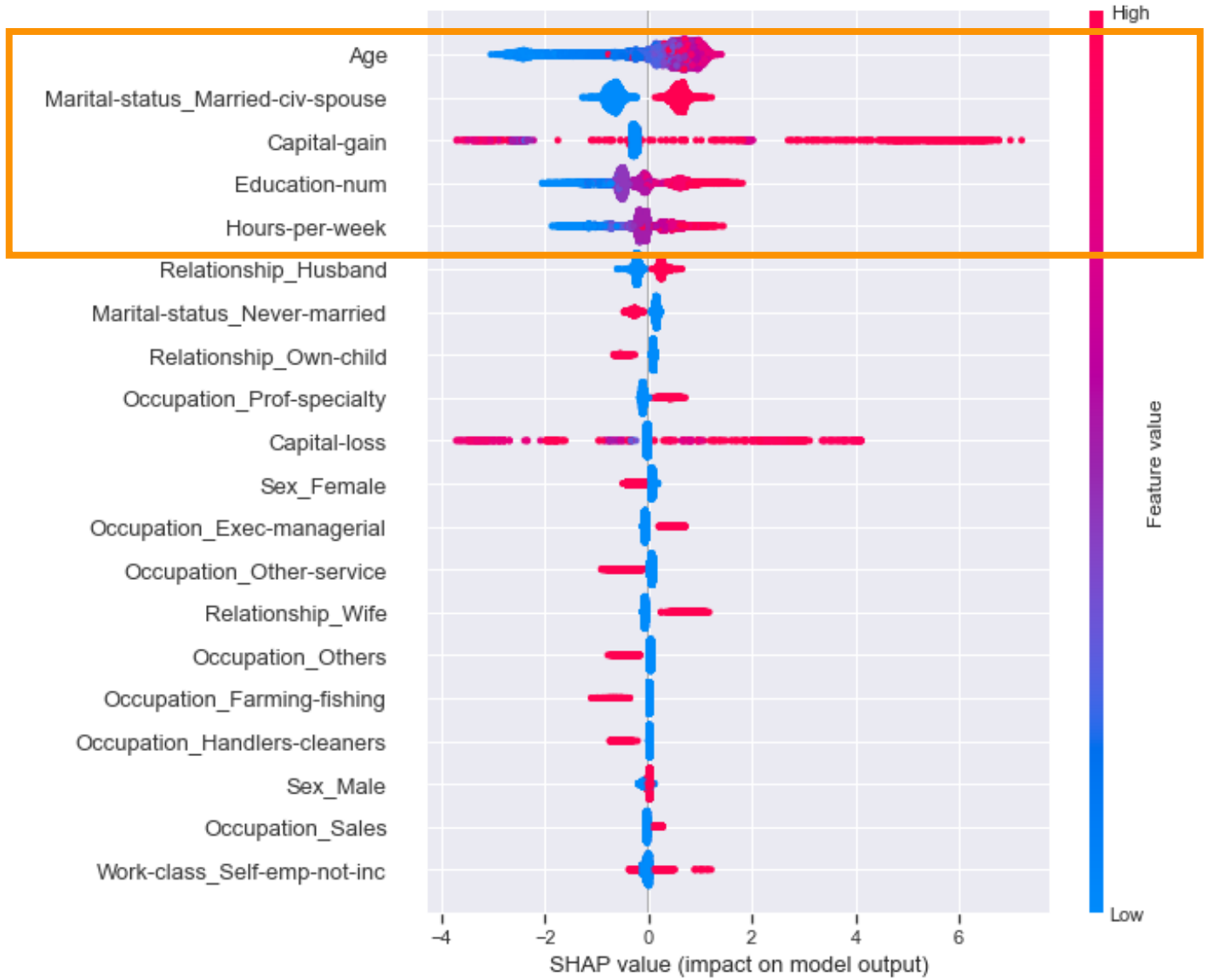
- SHAP identified both Married-yes/no and work hours per week as top 5 features, but didn't identify capital loss.
- SHAP also have solid math theory behind. In addition, it can be applied to both global and local interpretation.
- Next we will use SHAP plots to look into the impact of top features to target variable.

	Permutation feature importance	LightGBM feature importance	SHAP feature importance
1	Capital gain	Age	Age
2	Education years	Education years	Married-yes/no
3	Age	Capital gain	Capital gain
4	Married-yes/no	Capital loss	Education years
5	Capital loss	Work hours per week	Work hours per week

Top 5 influential features to LightGBM model output

Based on SHAP summary plot:

1. Age
2. Married-yes/no
3. Capital gain
4. Education years
5. Work hours per week

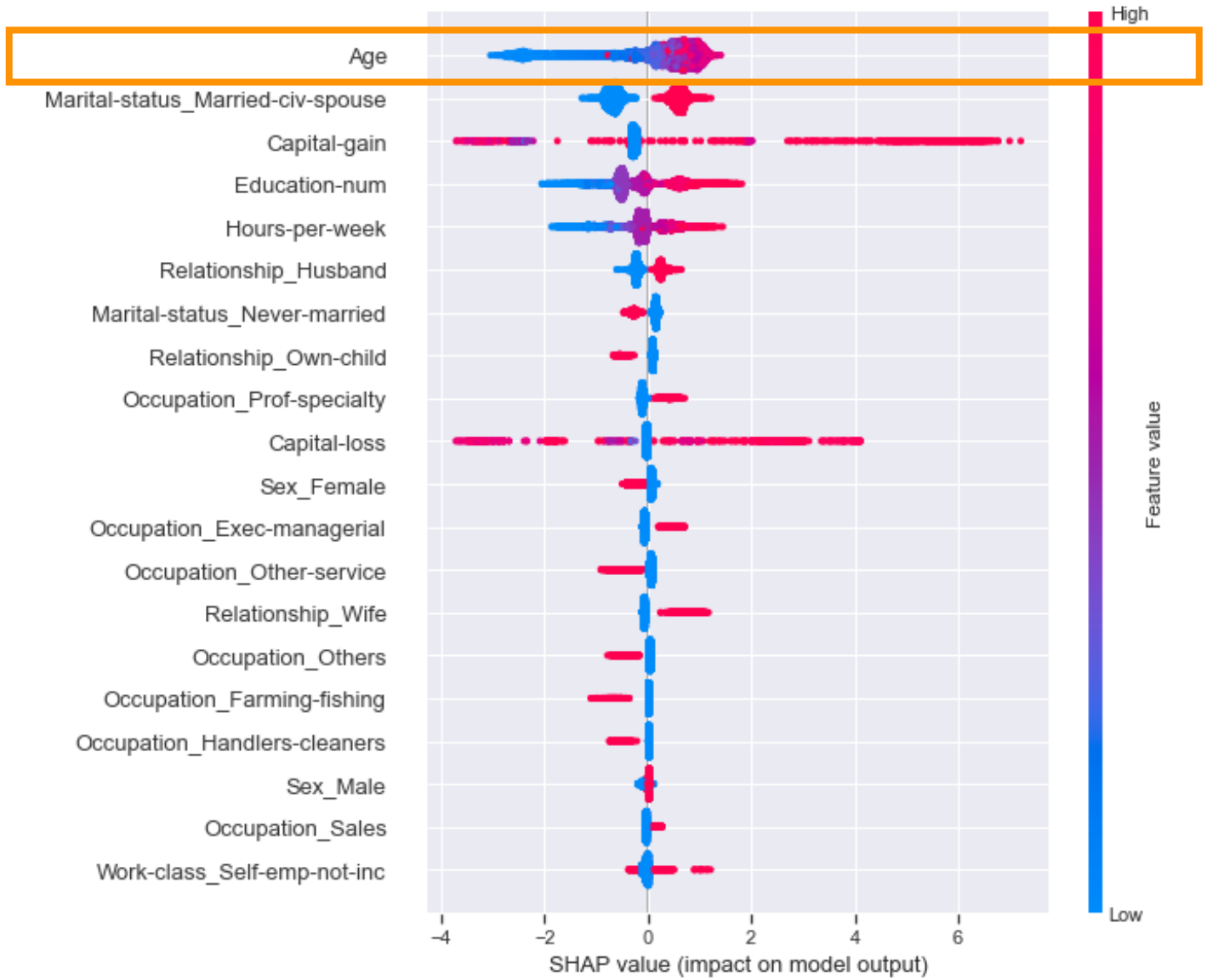


Top 5 influential features to LightGBM model output

Age

- Low age value impact the model output negatively (i.e. resulting ‘0’: income <=50K)

- High age value impact the output positively (i.e. resulting ‘1’: income >50K) but impact is not as strong, as reflected by the lower positive SHAP value.

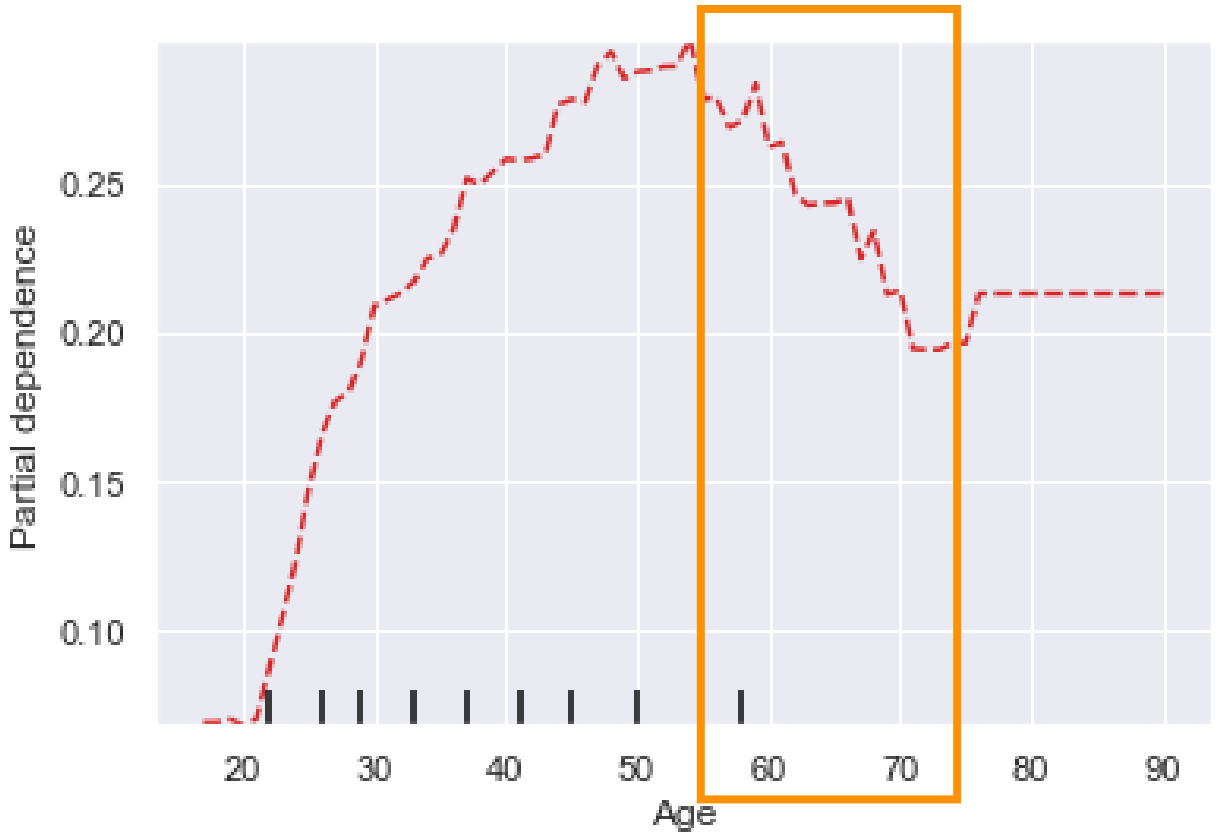


Top 5 influential features to LightGBM model output

Age on Partial Dependence Plot

- Upward slope of PD Plot curve confirmed again the positive correlation of age and target variable (income>50K = 1).

- However, beyond age 55, higher age would reduce the chance of having income >50K.

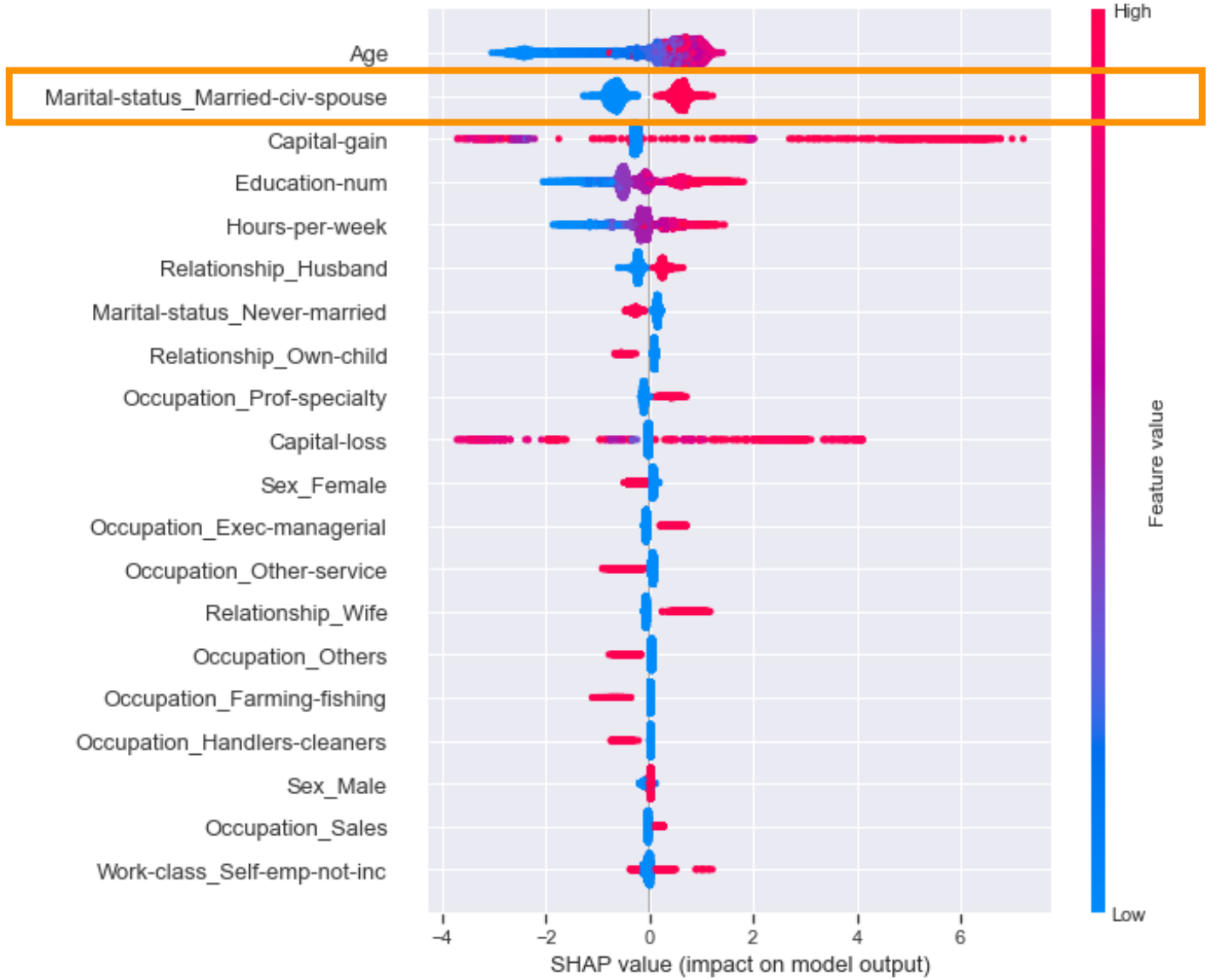


Top 5 influential features to LightGBM model output

Marital status

- High feature value (i.e. Married yes/no = yes) impact model output positively, low feature value impact model output negatively, in similar extent.

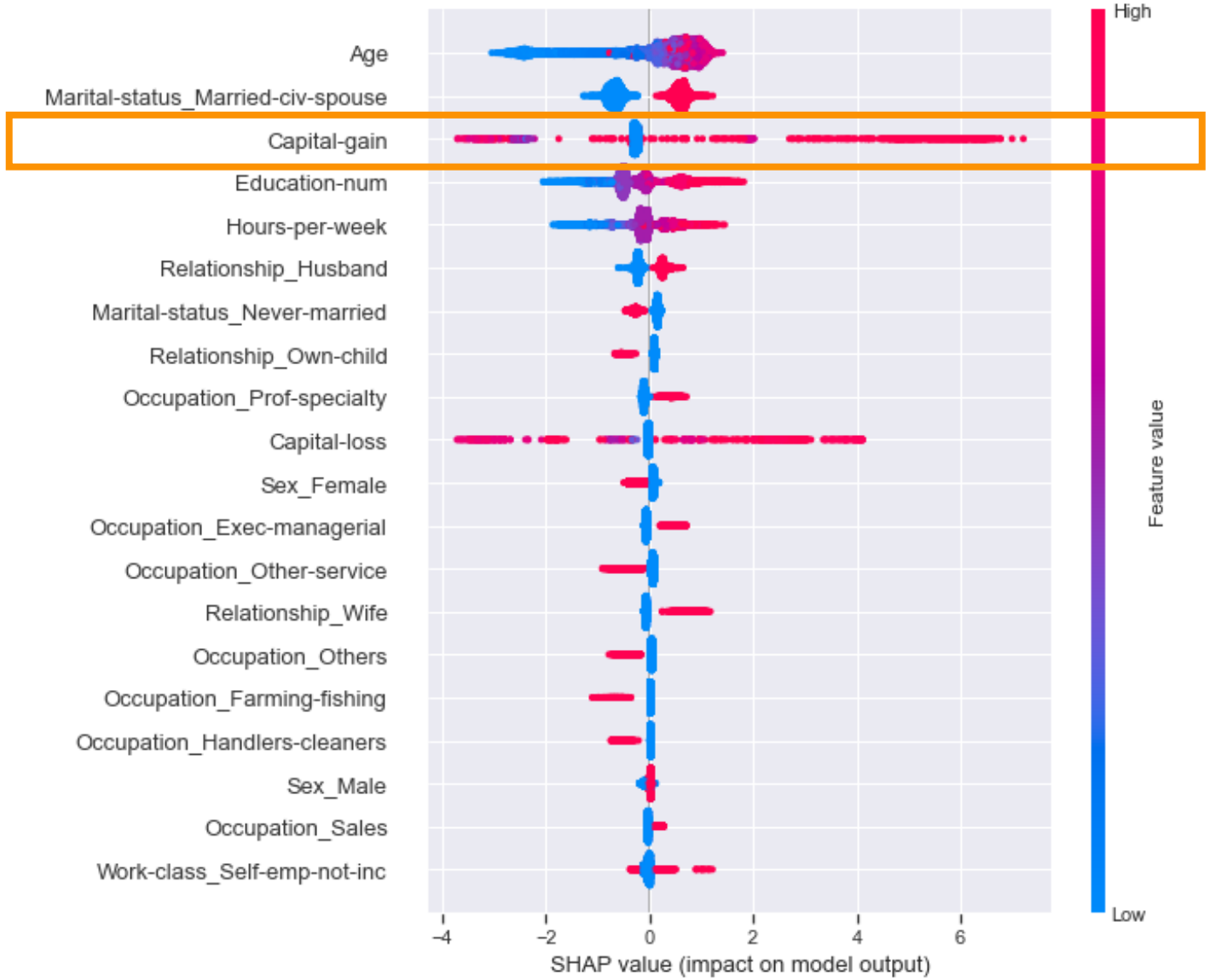
- Married person correlated with income >50K and unmarried correlated income <=50K



Top 5 influential features to LightGBM model output

Capital gain in a year

- Low capital gain value has no impact to model output.
- High capital gain value may have positive or negative impact to model output, as reflected by red colour appearing on both sides.
- This is partly consistent with previous EDA results, where positive capital gain outliers are mainly associated with >50K income group.

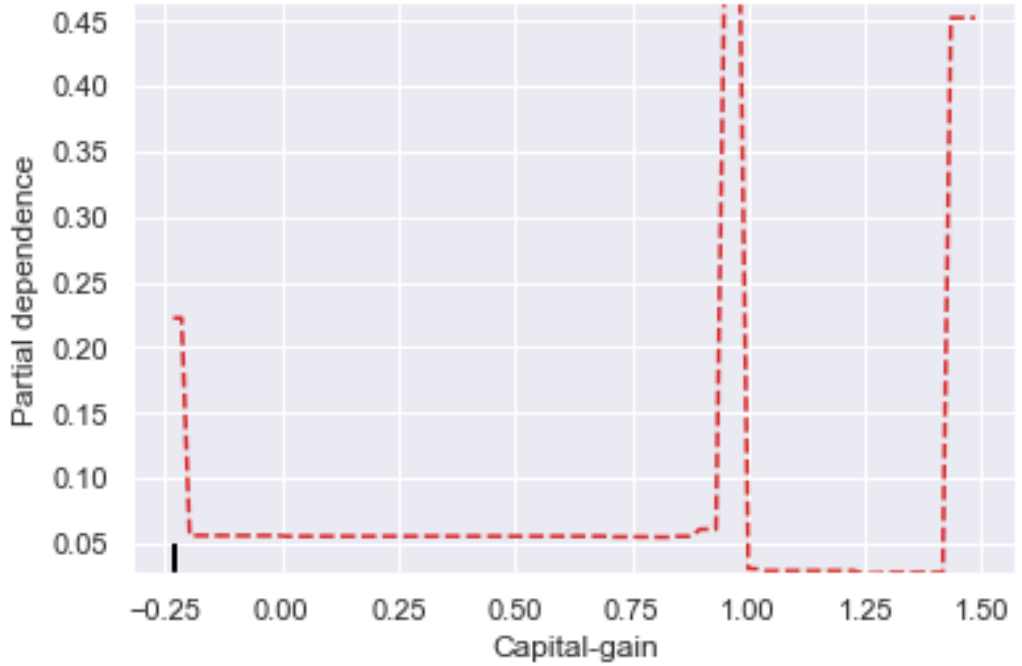


Top 5 influential features to LightGBM model output

Capital gain on Partial Dependence Plot

- In general there is no effect to income outcome when capital gain is at low range.

- Some significant effect is seen when capital gain of a person is 1 standard deviation away from the mean. This may be consistent with previous EDA results where positive capital gain outliers are mainly associated with >50K income group.



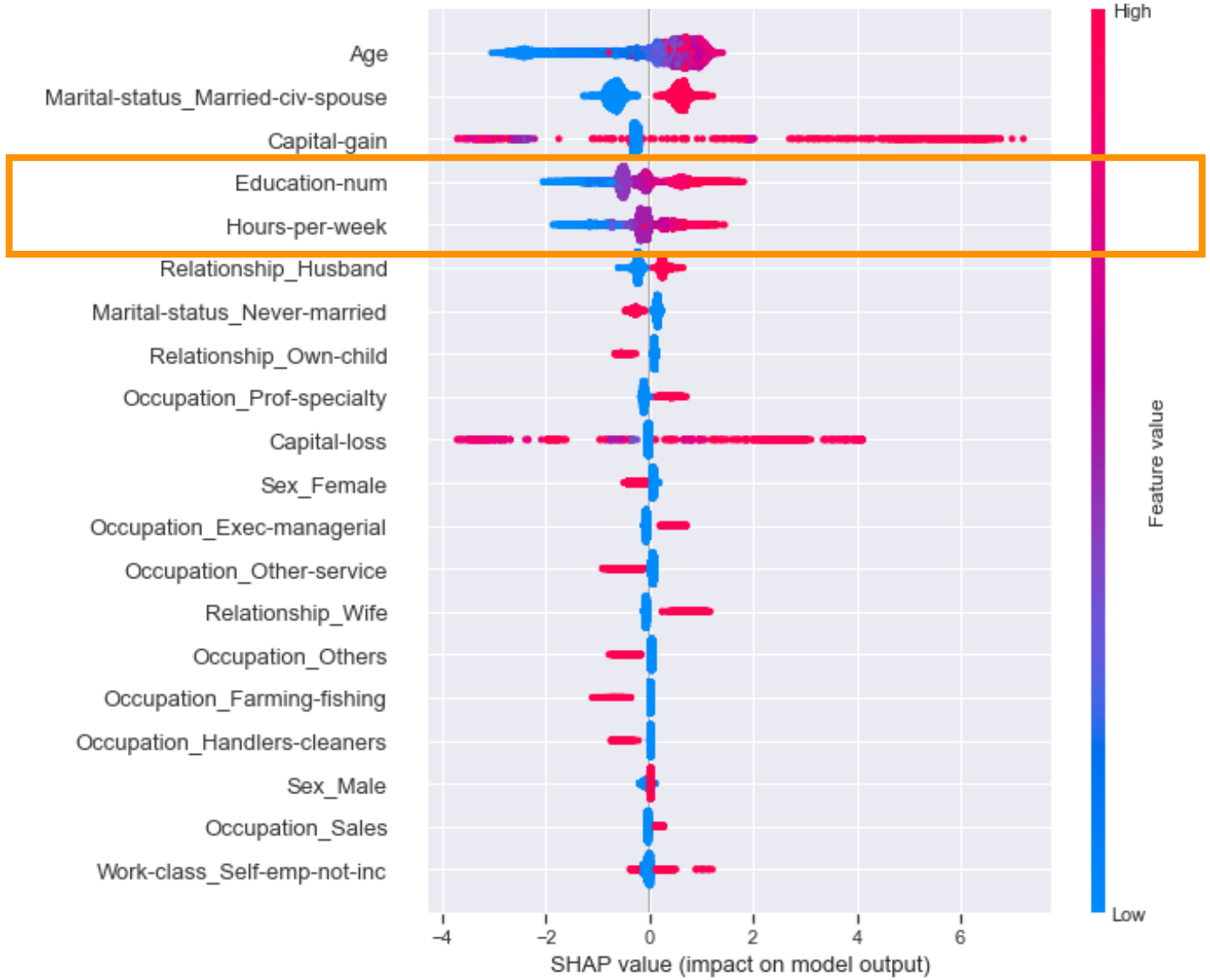
Top 5 influential features to LightGBM model output

Education years

- High education years impact model output positively, low education years impact model output negatively, in similar extent.

Work hours per week

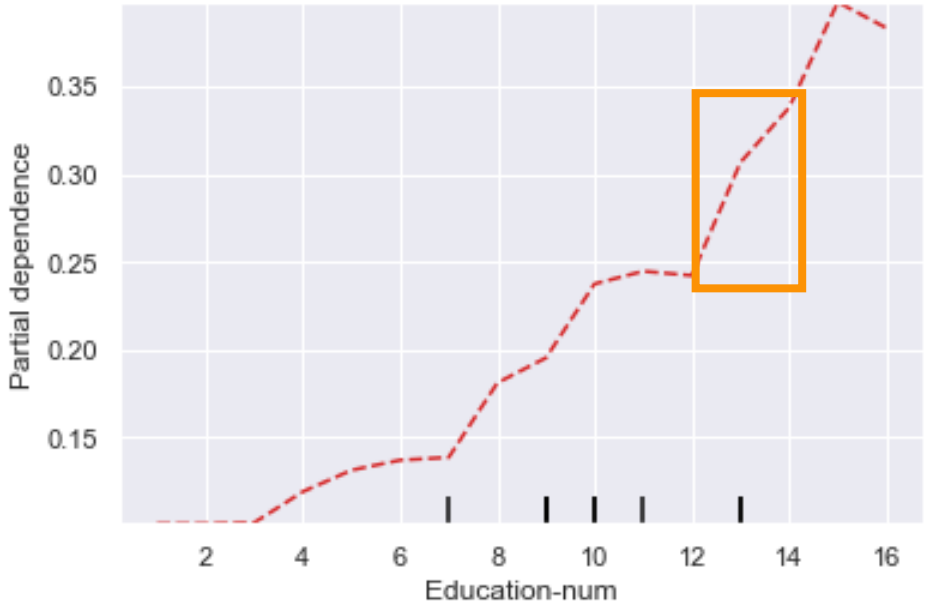
- Same case as education years.

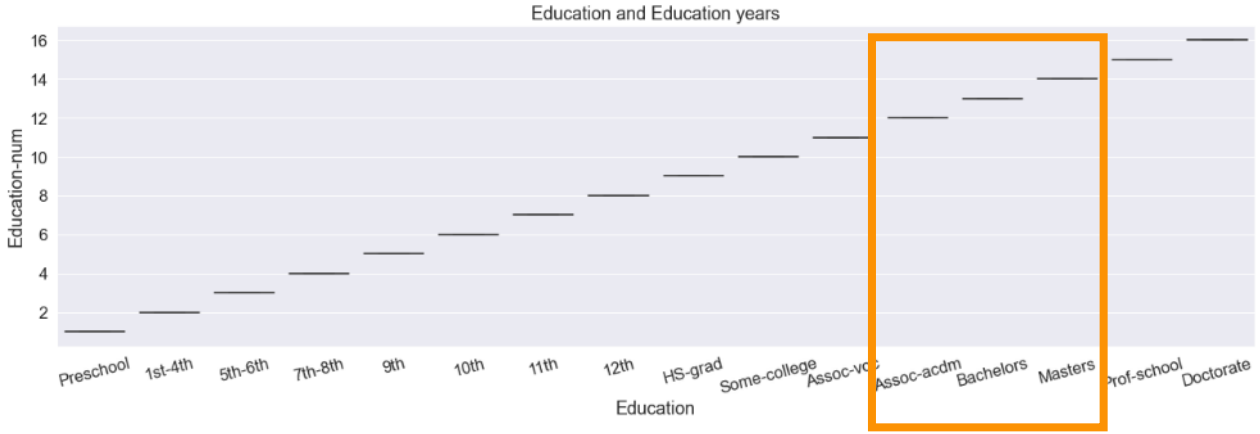


Top 5 influential features to LightGBM model output

Education years on Partial Dependence Plot

- Upward sloping curve confirmed again the positive correlation between education and income >50K.
- The increase is not even throughout. For instance, increase edu. years from 10 to 12 has lesser impact than an increase from 12 to 14.
- This can be explained by the kind of degree one is getting. 12 to 14 means upgrading from associate to master degree, while 10 to 12 means some-college to associate degree only.



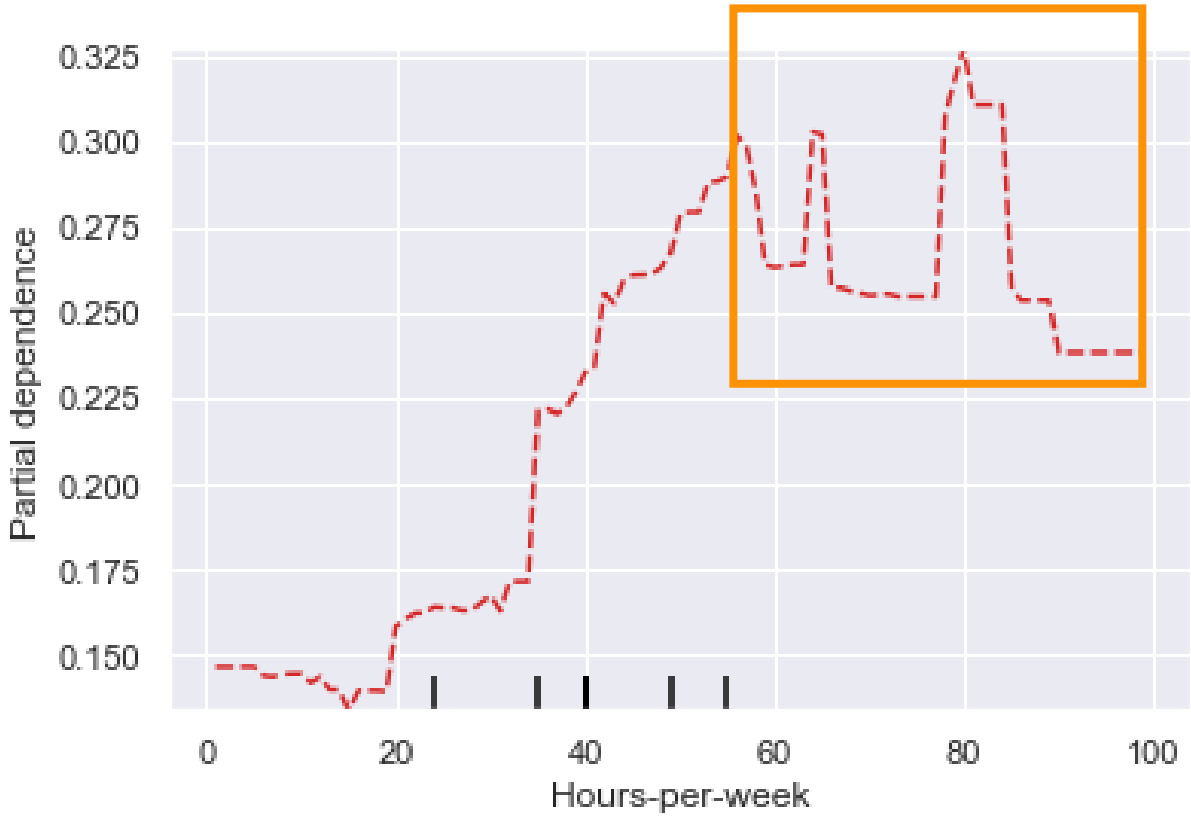


Top 5 influential features to LightGBM model output

Work hours per week on Partial Dependence Plot

- Upward sloping curve confirmed again the positive correlation between work hours and income >50K.

- However, beyond 55 hours, increasing working hours further may not always have positive impact to achieving income >50K.

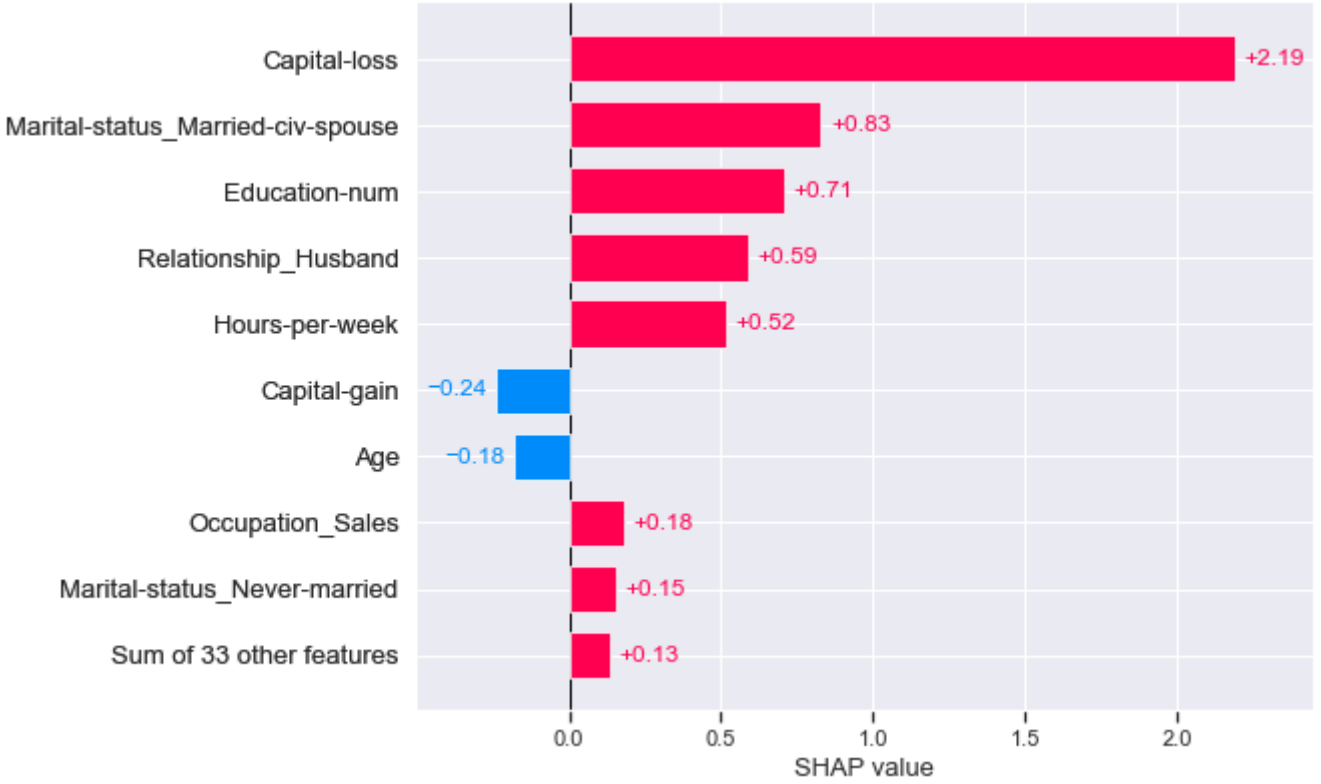


Impact of different features on an individual instance

Here we pick one person in the test sample and demonstrate how the features contribute to its prediction outcome in lightGBM model.

Sample ID 31139	Income <= 50K	Income > 50K
Actual class		✓
Predicted class	5.1% chance	94.9% chance

Feature	Values
Age	27
Sex	Male
Education years	13
Occupation	Sales
Work hours per week	55
Marital status	Married
Relationship	Husband
Capital gain	-0.23
Capital loss	4.37



Impact of different features on an individual instance

- Noticed that capital loss is the most important feature positively impacted the model predicted result (income >50K) for this individual, although capital loss is not in the top 5 important features by SHAP globally.
- Despite being relatively young at 27, the individual has numerous characteristics that contribute to having income >50K.

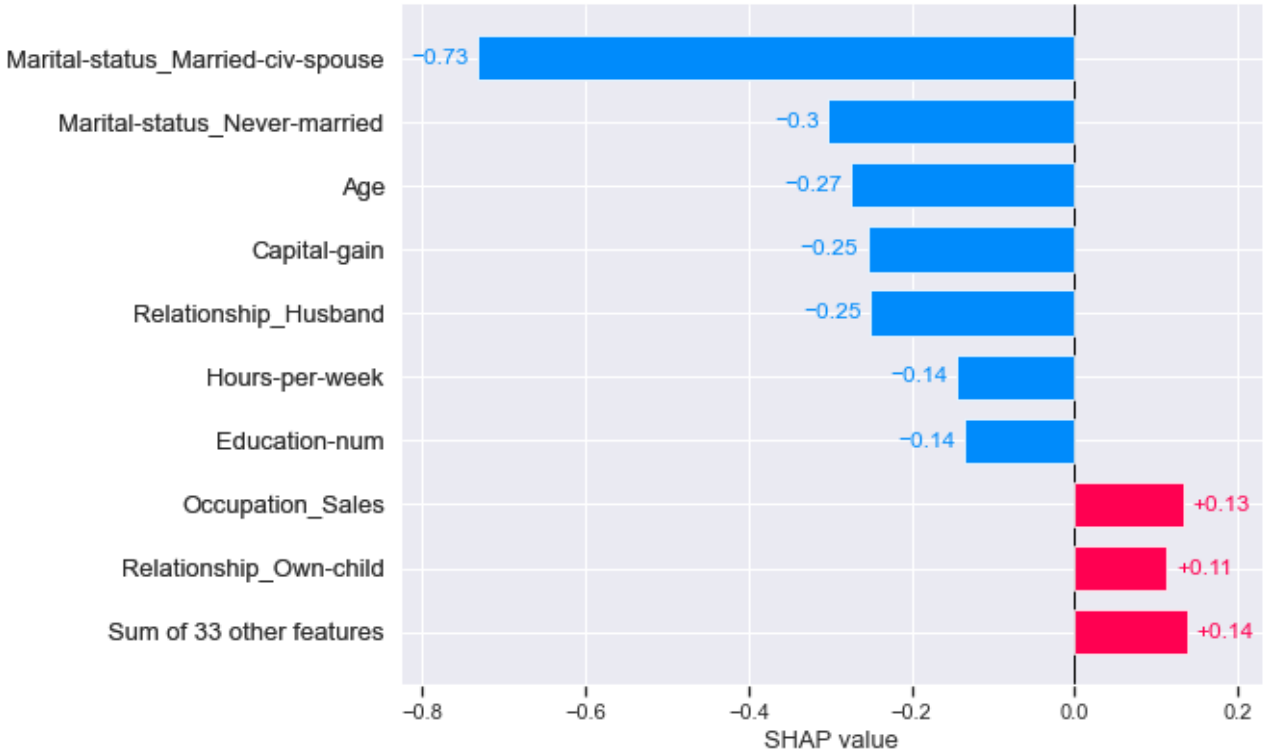
Feature	Values
Age	27
Sex	Male
Education years	13
Occupation	Sales
Work hours per week	55
Marital status	Married
Relationship	Husband
Capital gain	-0.23
Capital loss	4.37

Impact of different features on an individual instance

We look at one more test sample with income <=50K instead.

Sample ID 20866	Income <= 50K	Income > 50K
Actual class	✓	
Predicted class	97.5% chance	2.5% chance

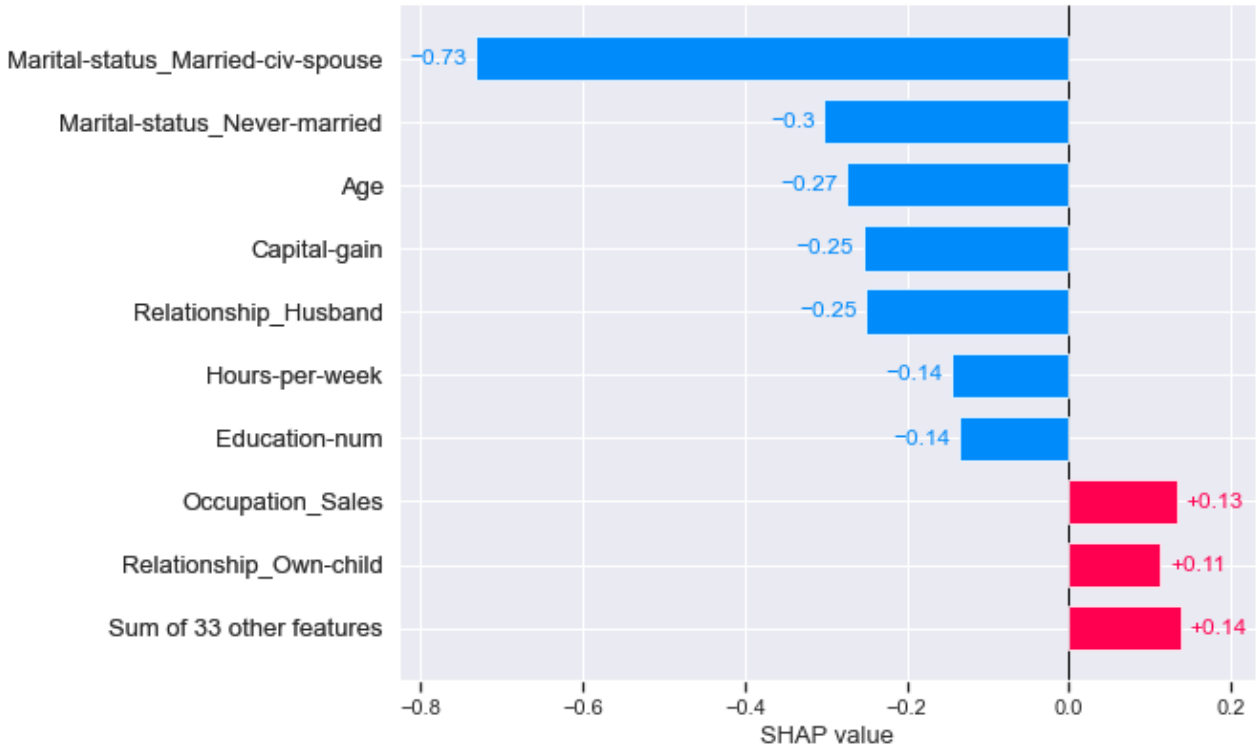
Feature	Values
Age	27
Sex	Male
Education years	10
Occupation	Sales
Work hours per week	40
Marital status	Never married
Relationship	Not in family
Capital gain	-0.23
Capital loss	-0.22



Impact of different features on an individual instance

- Although having the same age of 27 as previous sample, the fact that he is not married, have low capital gain, work lower hours and have lower education years, have contributed the most to the predicted outcome of income <= 50K.

Feature	Values
Age	27
Sex	Male
Education years	10
Occupation	Sales
Work hours per week	40
Marital status	Never married
Relationship	Not in family
Capital gain	-0.23
Capital loss	-0.22

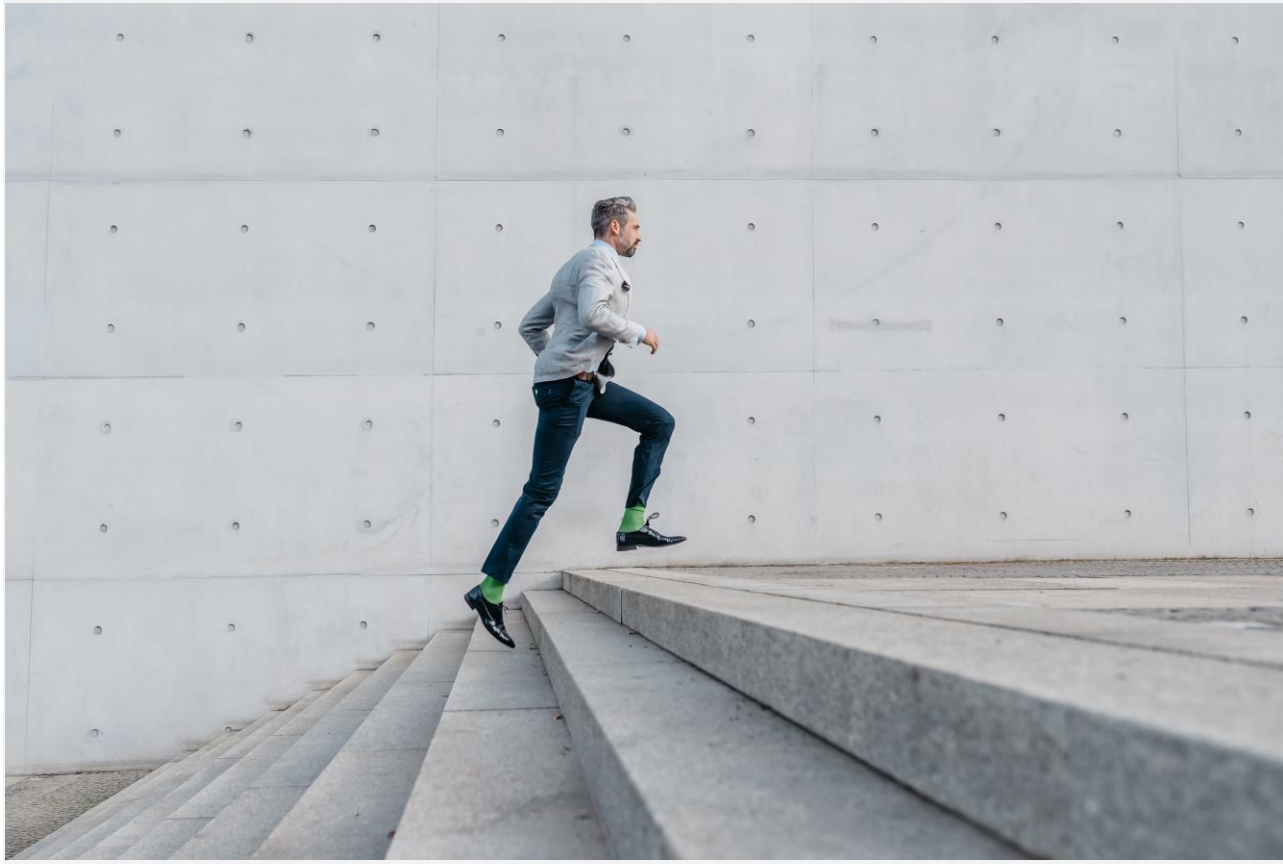


Model Interpretability - summary

Model interpretability methods have provided some additional insights apart from those in EDA:

	From exploratory data analysis	Additional insights from model interpretability methods
Age	Older - more likely for income >50K.	Confirming such trend. However, beyond age 55, higher age would reduce the chance of having income >50K.
Married-yes/no	Married - more likely for income >50K.	Showed same result.
Capital gain	Most positive capital gain outliers belongs to >50K income group.	Similar trend observed.
Education	Higher education level attained - more likely for income >50K, and vice versa.	Confirming such trend. But the increase is not linear. For example, the increase in probability of having income >50K is not the same for increase in education years from 10 to 12 vs increase from 12 to 14.
Work hours per week	Longer work hours - more likely for income >50K.	Confirming such trend. However, beyond 55 hours, increasing working hours further may not always have positive impact to achieving income >50K.

Final Thoughts



Final Thoughts

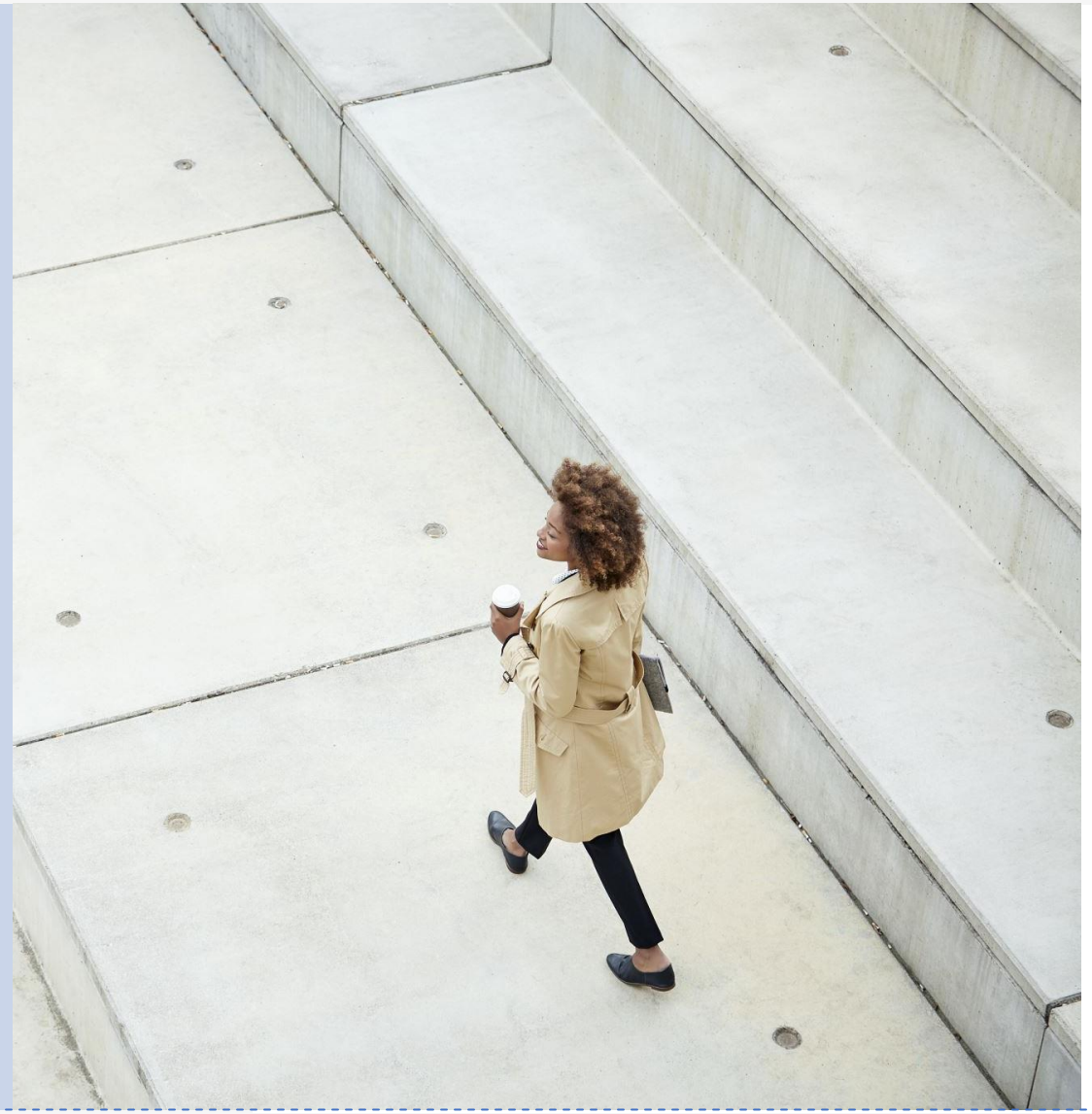
- In these slides, we have looked at how a machine learning pipeline is implemented using a small dataset from U.S. Census in 1994.
- Of course this is not a perfect business case, because:
  - Some of the insights are conventional, for example, higher education level by logic means higher chance of getting high income).
  - We did not specify the use case in this income prediction. For example, are we targeting the low or high income group in some welfare or marketing campaign?
- However, similar machine learning techniques are useful in many other business cases, for example:
    - Predicting purchase/non-purchase likelihood
    - Predicting customer churn
    - Predicting credit default
    - Predicting getting illness
    - Predicting machine failure
    - Many many more.

Final Thoughts

- Always aware of some limitations -
  - This study is not causation analysis. For example, inferring that investing more for having positive capital gain or loss will *lead to* income >50K, is not correct. The more logical interpretation is that the higher income group tend to invest more, so we see more positive capital gains being associated with them.
  - Actual environment is dynamic. For instance this dataset is from 1994 and so it would not represent the situation today. In actual implementation, models should be continuously updated and monitored.
  - The size of dataset and number of features is low here. In reality, there are tons of data every day. Selecting appropriate features, data pre-processing, modelling them and generating a timely prediction results would be even more complicated and require additional big data techniques. But the concepts discussed here surely applies.

References

- Permutation Importance vs Random Forest Feature Importance (MDI) — scikit-learn 1.0.2 documentation
- bar plot — SHAP latest documentation



The End

Feel free to view other projects on my GitHub page:
https://github.com/wktracy

June 2022