



Group 9

Predicting Family Income for Marketing

Training and comparing Naive Bayes, Logistic regression and CART models to predict if family income is above 50K/Y

Jeriel
Yin
Kieran

Executive Summary

1

Random Forest classification is the best method to predict household income greater than \$50K since it had the highest accuracy rate predicting family income **(83.4%)**

2

\$25 Million Direct Marketing Campaign Target Group:

- Married couples - particularly the husband
- Bachelors & Masters level education



Agenda

- **Understanding raw data** - Variables, data types
- **Cleaning the data** - Sanitize Missing values, outliers
- **Preprocessing** - Encode variables
- **Model Exploration** - Build and evaluate models
- **Model Comparison** - Find the best predicting model

Understanding Raw Data

Explore, Check and Convert Data Types

age	int64
workclass	object
fnlwgt	int64
education	object
education_num	int64
marital_status	object
occupation	object
relationship	object
race	object
sex	object
capital_gain	int64
capital_loss	int64
hours_per_week	int64
native_country	object
income	object



age	int64
workclass	category
fnlwgt	int64
education	category
education_num	category
marital_status	category
occupation	category
relationship	category
race	category
sex	category
capital_gain	int64
capital_loss	int64
hours_per_week	int64
native_country	category
income	category
dtype:	object

- There are **32561 observations** and **15 variables** in the raw data
- **Employee_num** should be treated as 'category'

Cleaning the Data

```
age          143
workclass    0
fnlwgt       992
education    0
education-num 0
marital-status 0
occupation  0
relationship 0
race         0
sex          0
capital-gain 2712
capital-loss 1519
hours-per-week 9008
native-country 0
class        0
dtype: int64
```

	age	fnlwgt	capital-gain	capital-loss	hours-per-week
mean	38.581647	1.897784e+05	1077.648844	87.303830	40.437456

- Before imputing corrections, we needed to **identify outliers** in the numeric variables
- Replaced all numeric columns null values to their respective **mean** values
- liminated unusual unknown values such as ' ?' and replacing the values with the **mode**

Preprocessing: Improve Model Efficiency

Dummy Variables

	workclass_ Local-gov	workclass_ Never- worked	workclass_ Private
0	0	0	0
1	0	0	0
2	0	0	1
3	0	0	1
4	0	0	1
...
32556	0	0	1
32557	0	0	1
32558	0	0	1
32559	0	0	1
32560	0	0	0

32561 rows × 51 columns

- Using binary dummy - allowing the computer to calculate them with the numeric variables

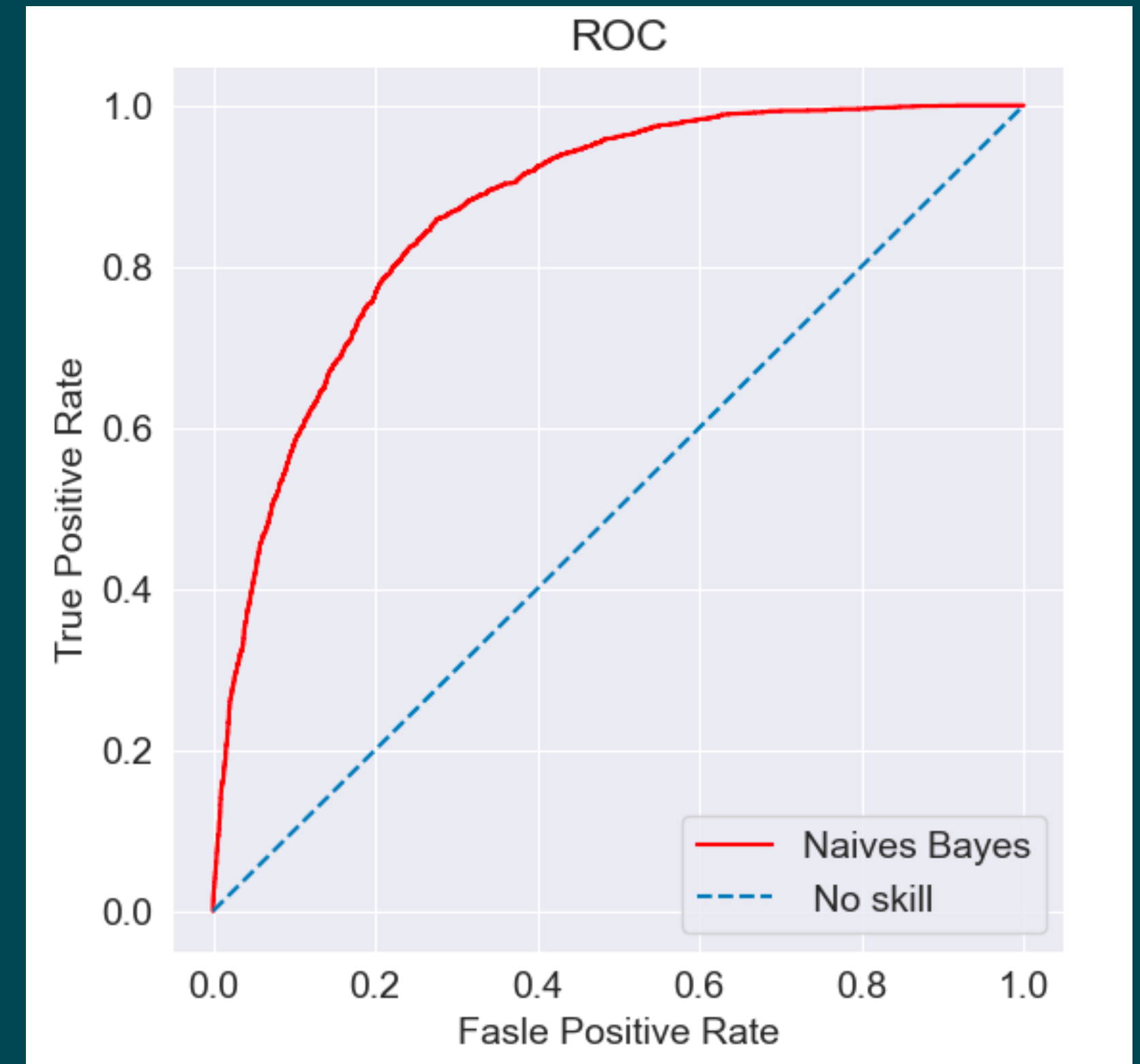
	age	fnlwgt	capital-gain	capital-loss	hours-per-week
count	32561.000000	32561.000000	32561.0	32561.0	32561.000000
mean	38.381609	179631.339130	0.0	0.0	41.565533
std	13.300577	86028.457849	0.0	0.0	3.415436
min	17.000000	12285.000000	0.0	0.0	33.000000
25%	28.000000	117827.000000	0.0	0.0	40.000000
50%	37.000000	178356.000000	0.0	0.0	40.000000
75%	47.000000	226196.000000	0.0	0.0	41.565533
max	78.000000	415847.000000	0.0	0.0	52.000000

- Normalization the numbers (scale the variables to similar sizes) makes the graphing of the values more efficient for the ML methods

Naive Bayes

Accuracy	Misclassification	True Positive	False Positive
0.816	0.171	0.926	0.458
Specificity	Precision	Prevalence	ROC
0.541	0.61	0.83	0.868

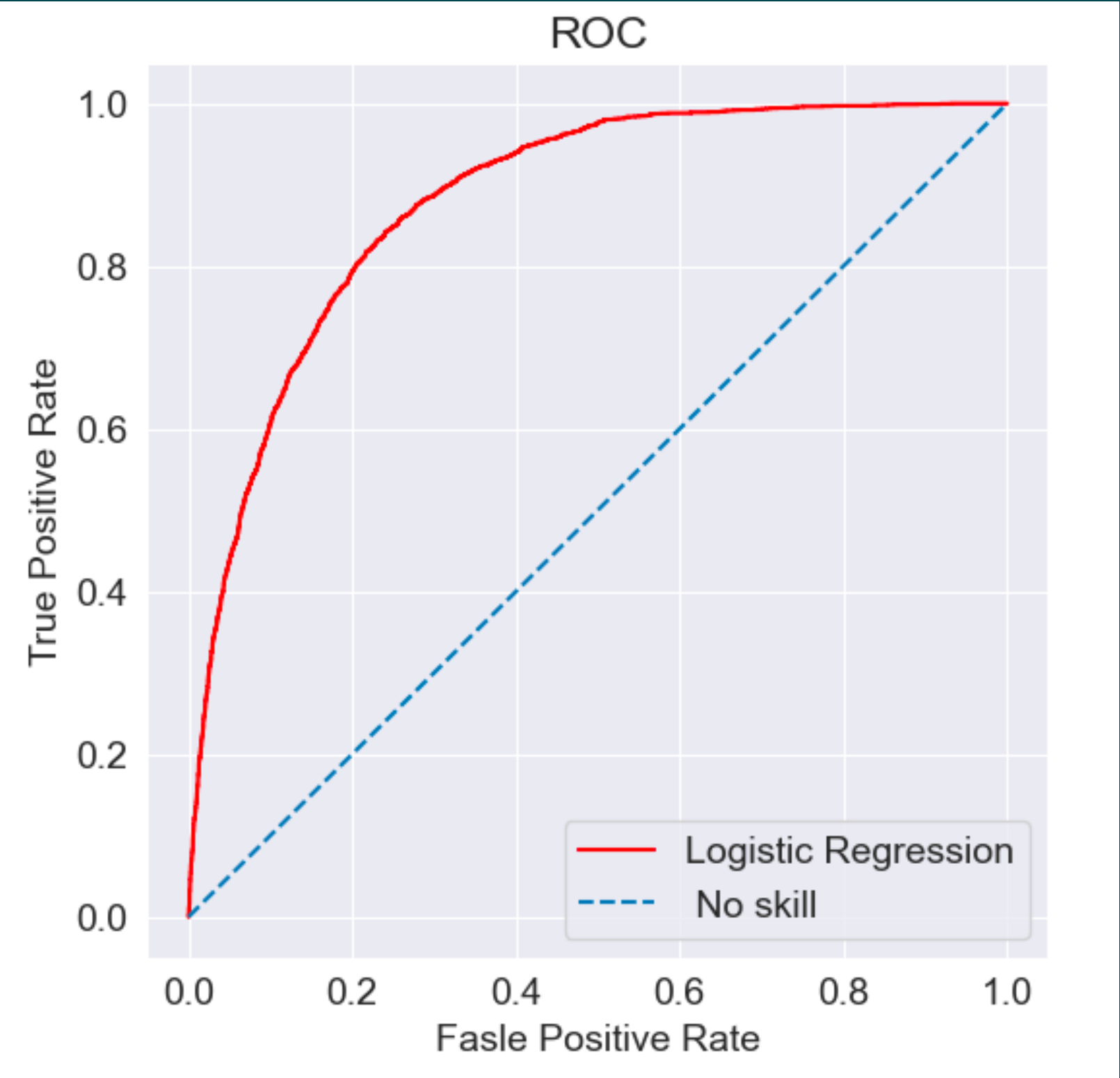
True Class	Predicted	
	True	Negative
True	15018	2279
Negative	1999	3497



Logistic Regression

Accuracy	Misclassification	True Positive	False Positive
0.829	0.171	0.826	0.459
Specificity	Precision	Prevalence	ROC
0.541	0.69	0.834	0.88

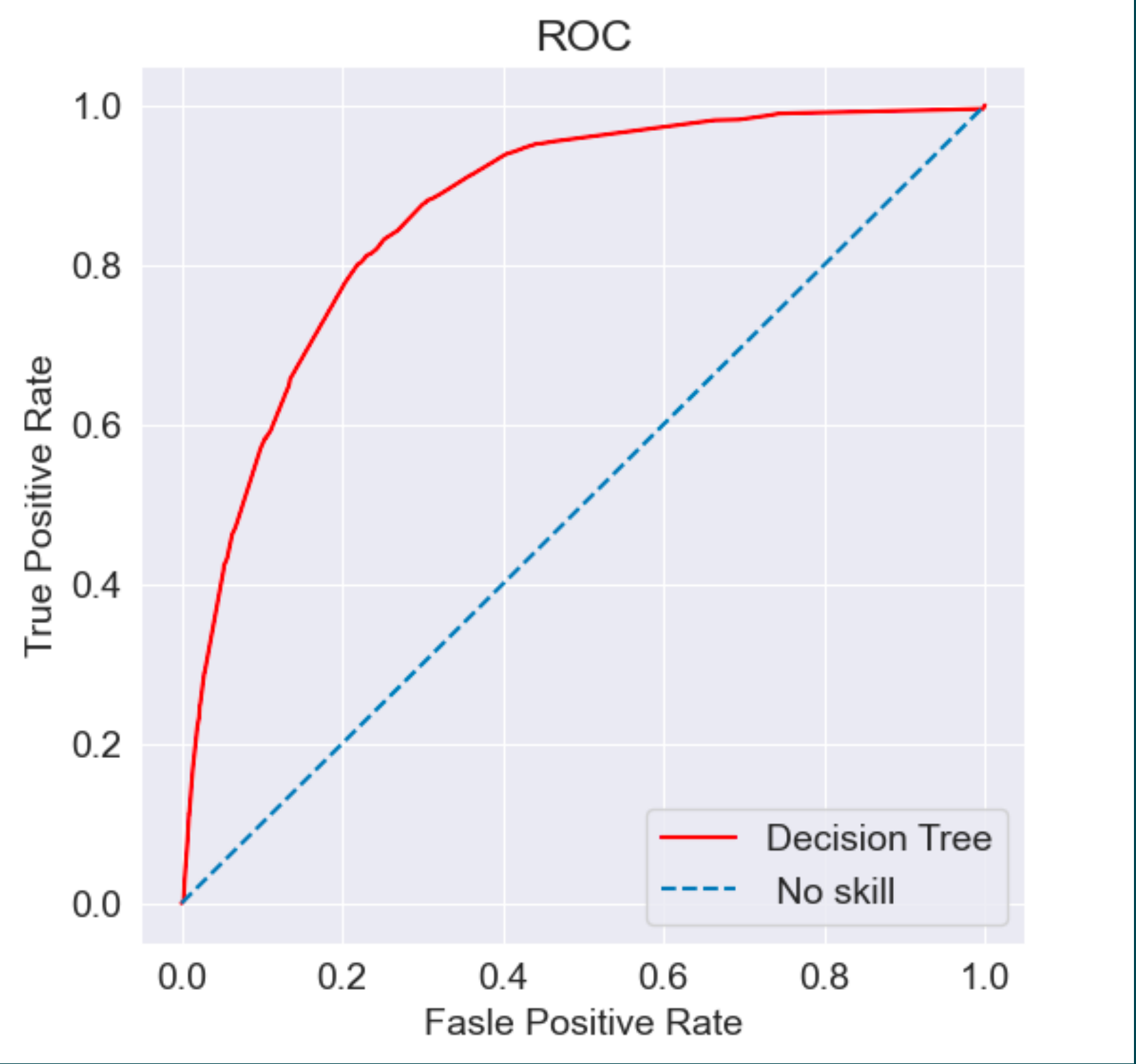
		Predicted	
		True	Negative
True Class	True	16057	1278
	Negative	2504	2953



Decision Tree

Accuracy	Misclassification	True Positive	False Positive
0.82	0.171	0.915	0.402
Specificity	Precision	Prevalence	ROC
0.598	0.65	0.839	0.865

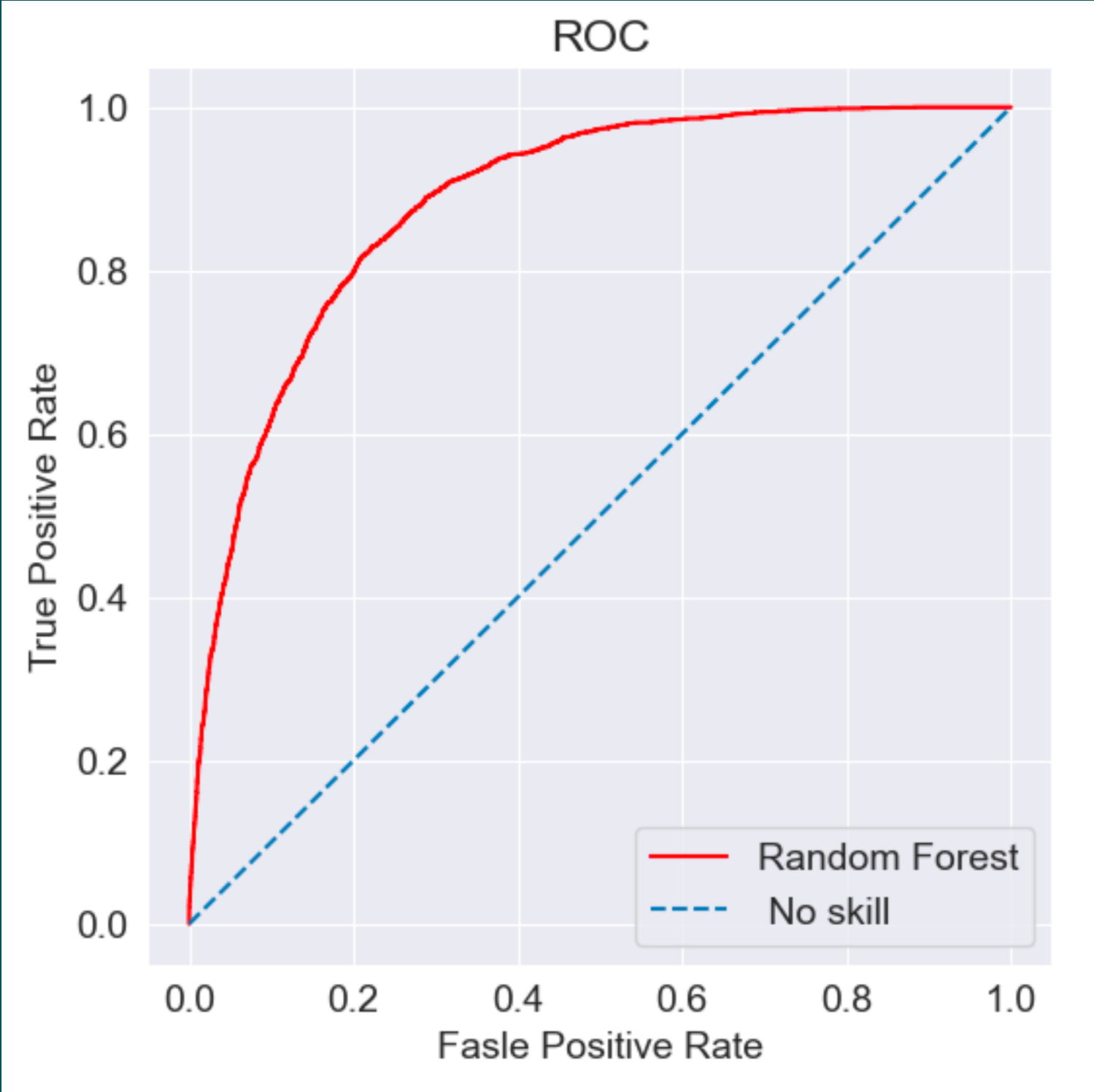
		Predicted	
		True	Negative
True Class	True	15853	1482
	Negative	2192	3265



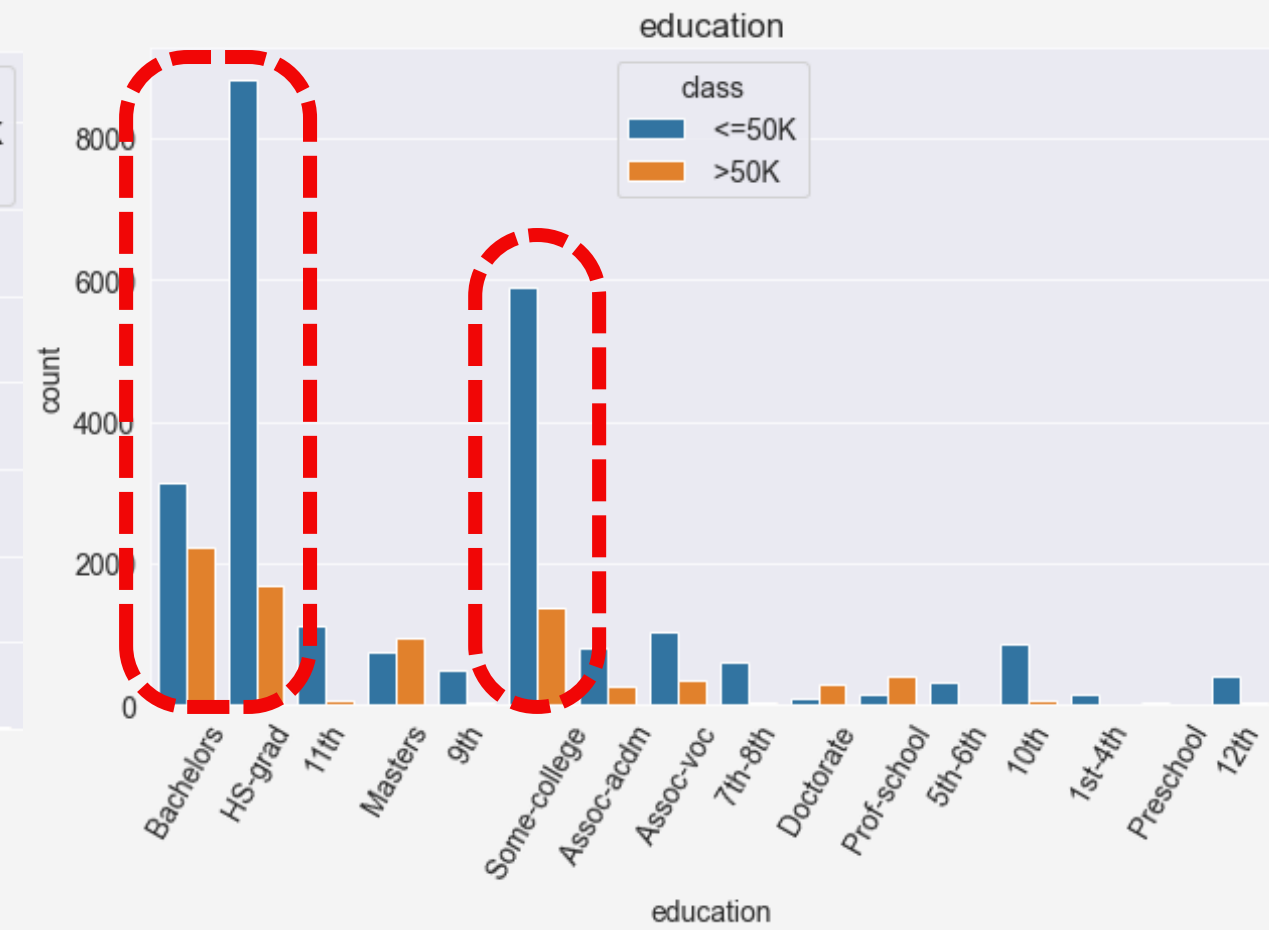
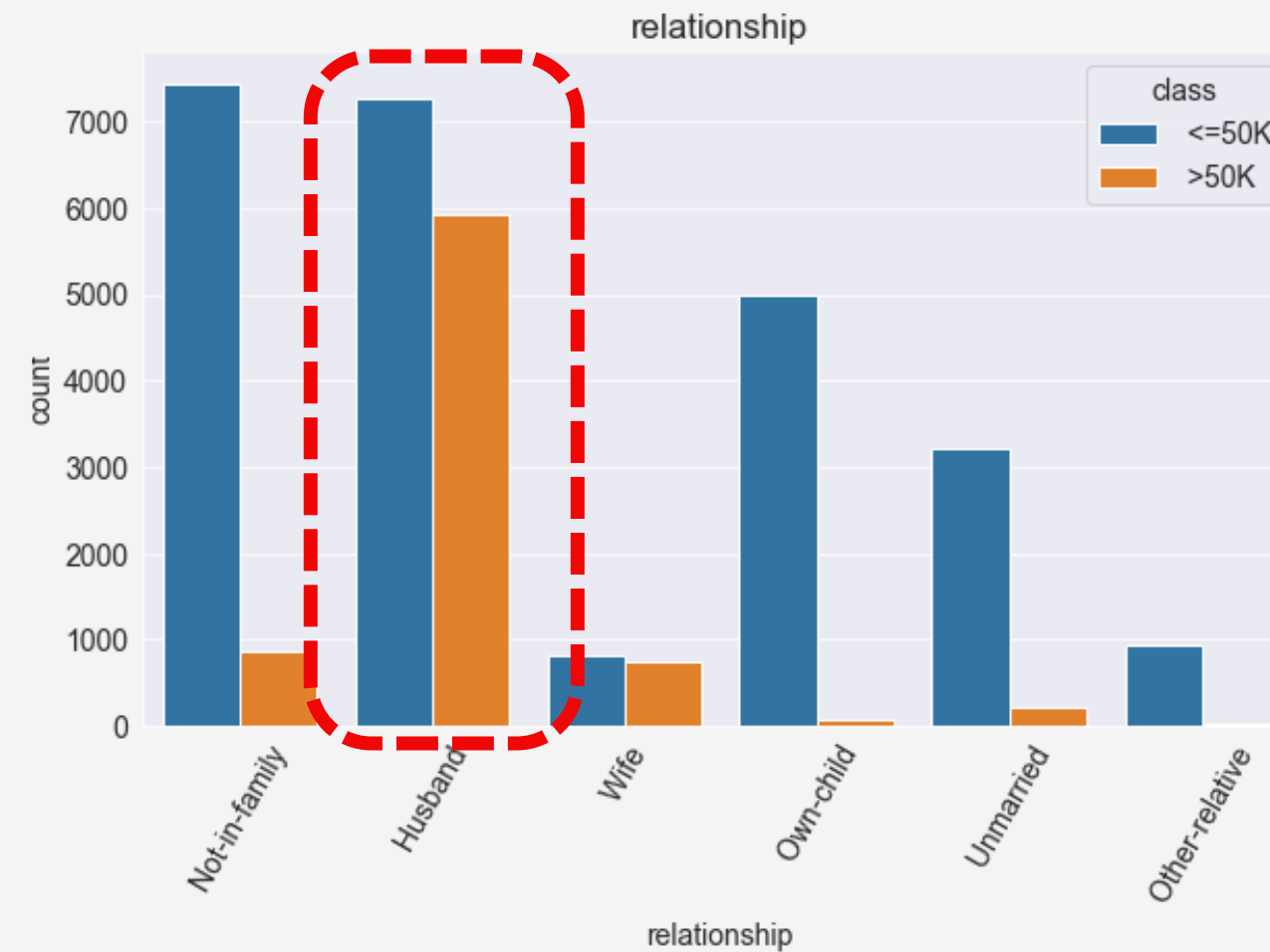
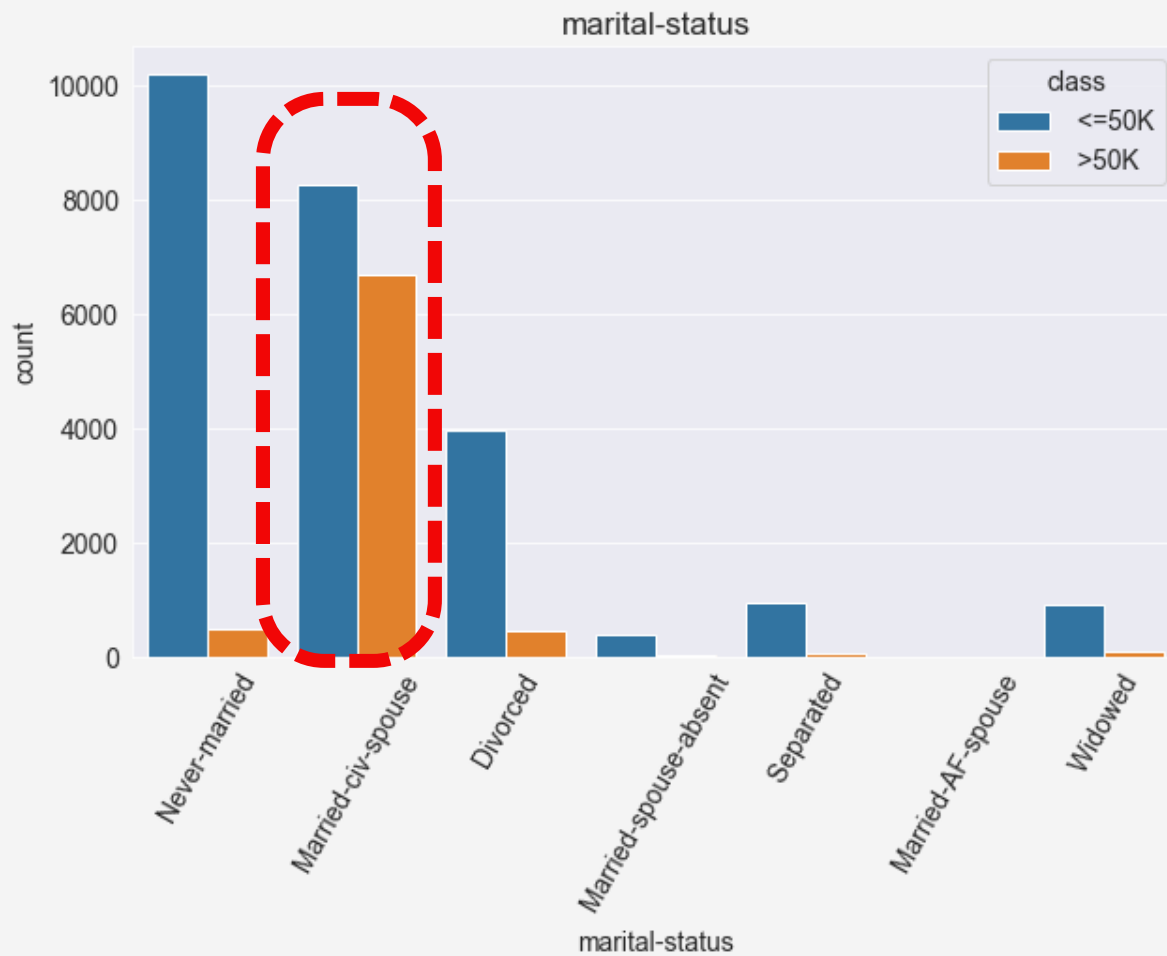
Random Forest: Most Accurate Model

Accuracy	Misclassification	True Positive	False Positive
0.834	0.171	0.967	0.306
Specificity	Precision	Prevalence	ROC
0.694	0.72	0.901	0.883

		Predicted	
		True	Negative
True Class	True	16755	580
	Negative	1669	3788



Recommendations



- Most accurate model
 - Random Forest Model 83.4%

- Target
 - Married couples - particularly the husband
 - Higher educated individuals - particularly Bachelors and Master Graduates.

Appendix (1)

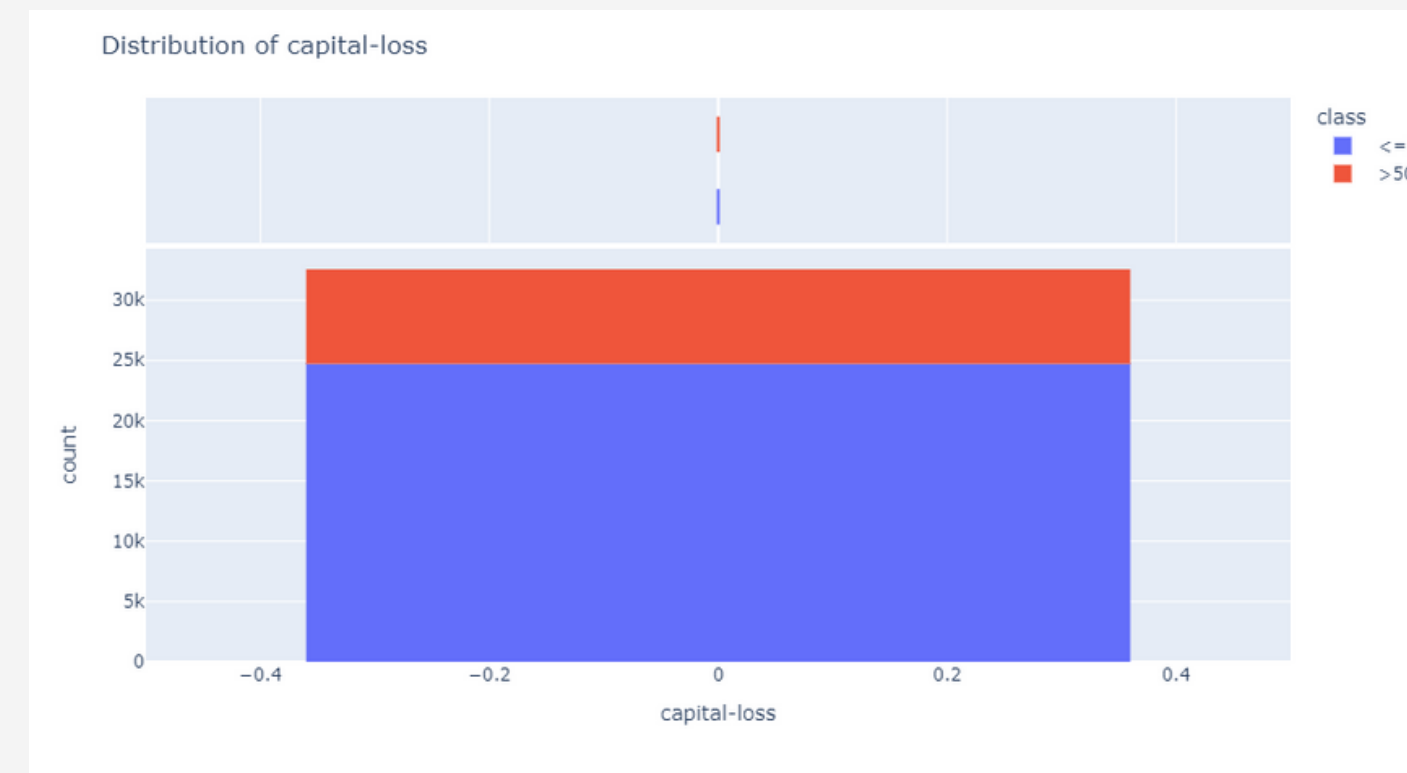
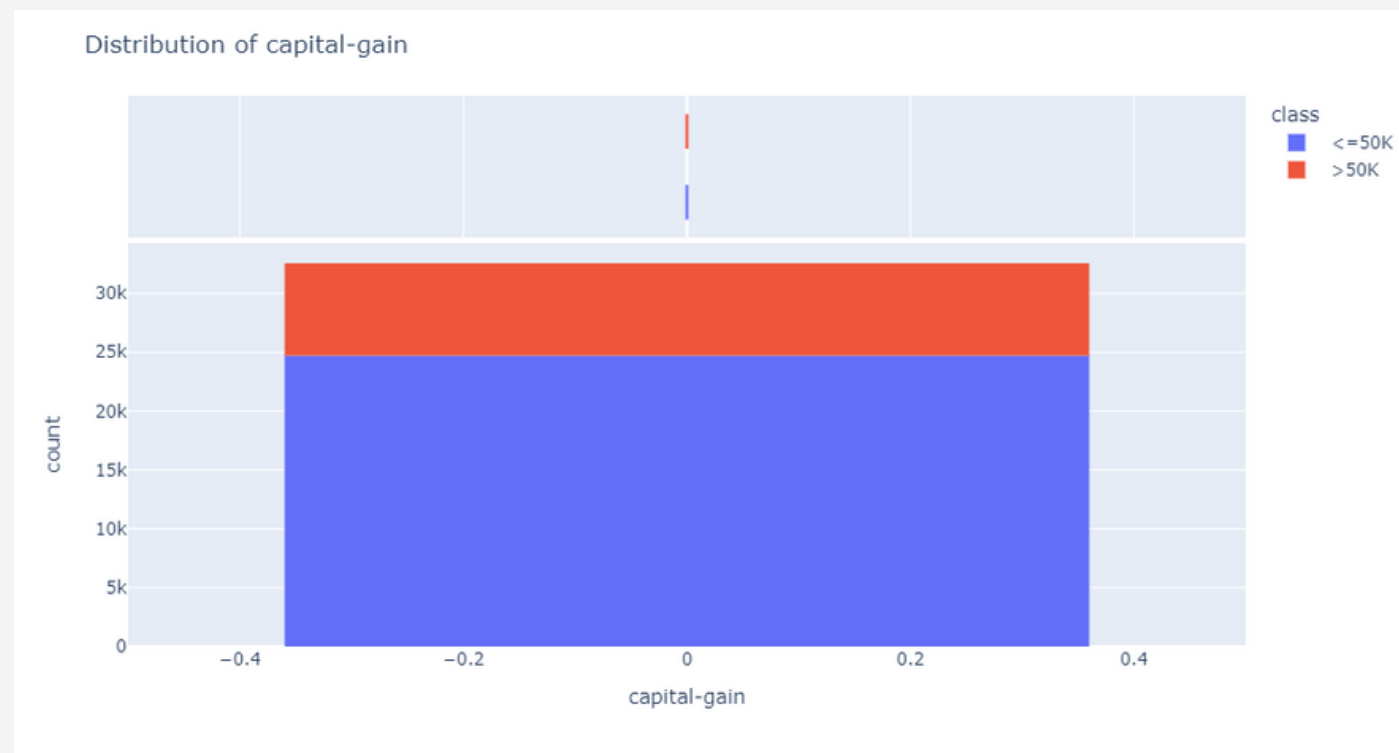
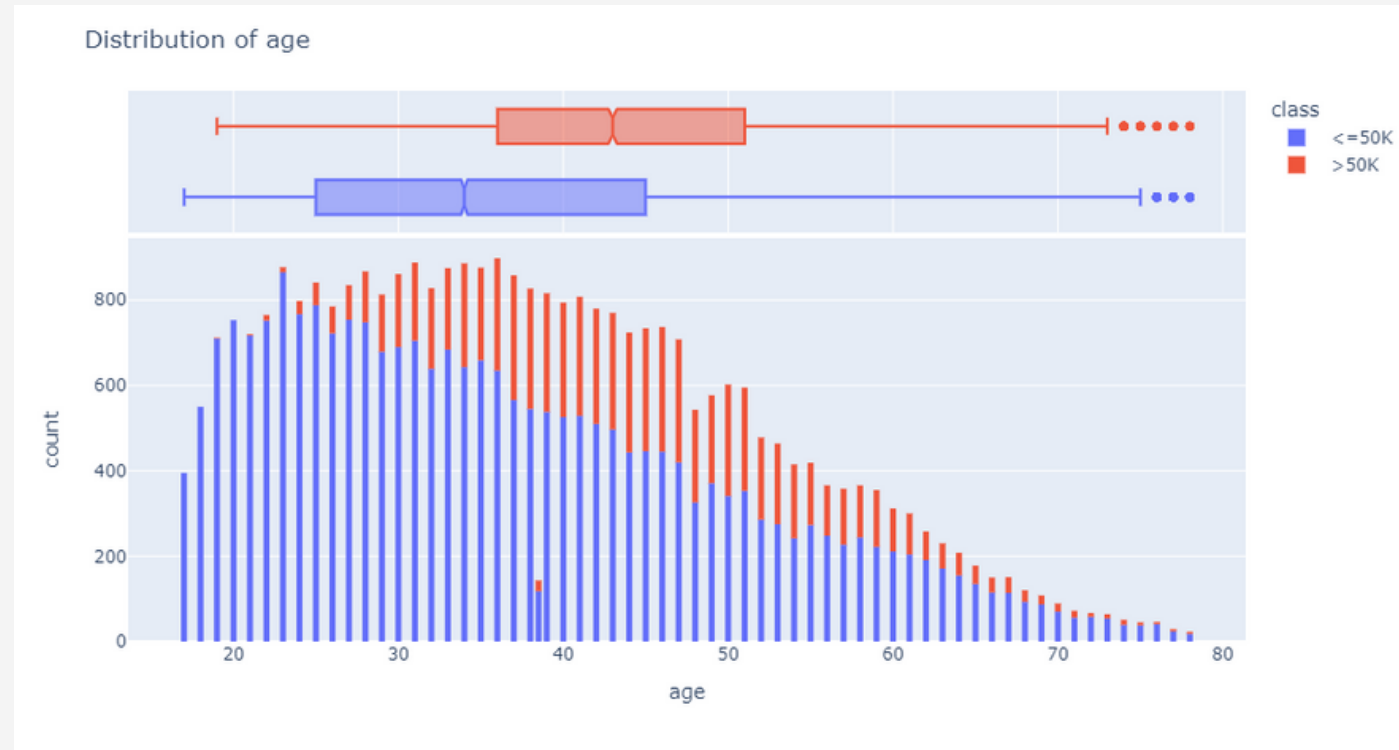
age	int64
workclass	object
education	object
education_num	int64
marital_status	object
occupation	object
relationship	object
race	object
sex	object
capital_gain	int64
capital_loss	int64
hours_per_week	int64
native_country	object
income	object
dtype:	object

Raw Data Variable Types (Q2)

	age	capital_gain	capital_loss	hours_per_week
min	17.000000	0.000000	0.000000	1.000000
max	90.000000	99999.000000	4356.000000	99.000000
median	37.000000	0.000000	0.000000	40.000000
mean	38.581647	1077.648844	87.303830	40.437456
std	13.640433	7385.292085	402.960219	12.347429
skew	0.558743	11.953848	4.594629	0.227643
kurt	-0.166127	154.799438	20.376802	2.916687

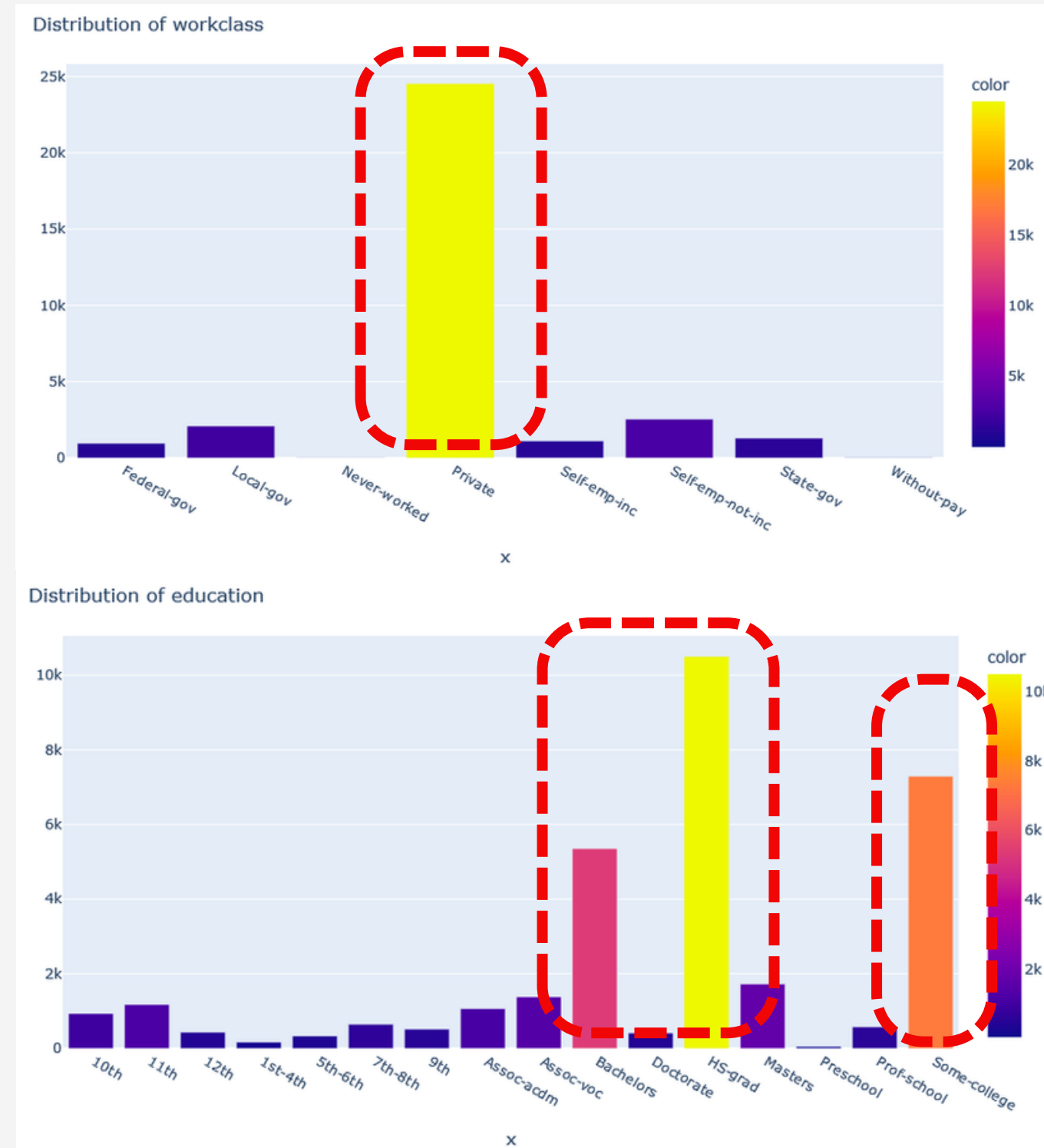
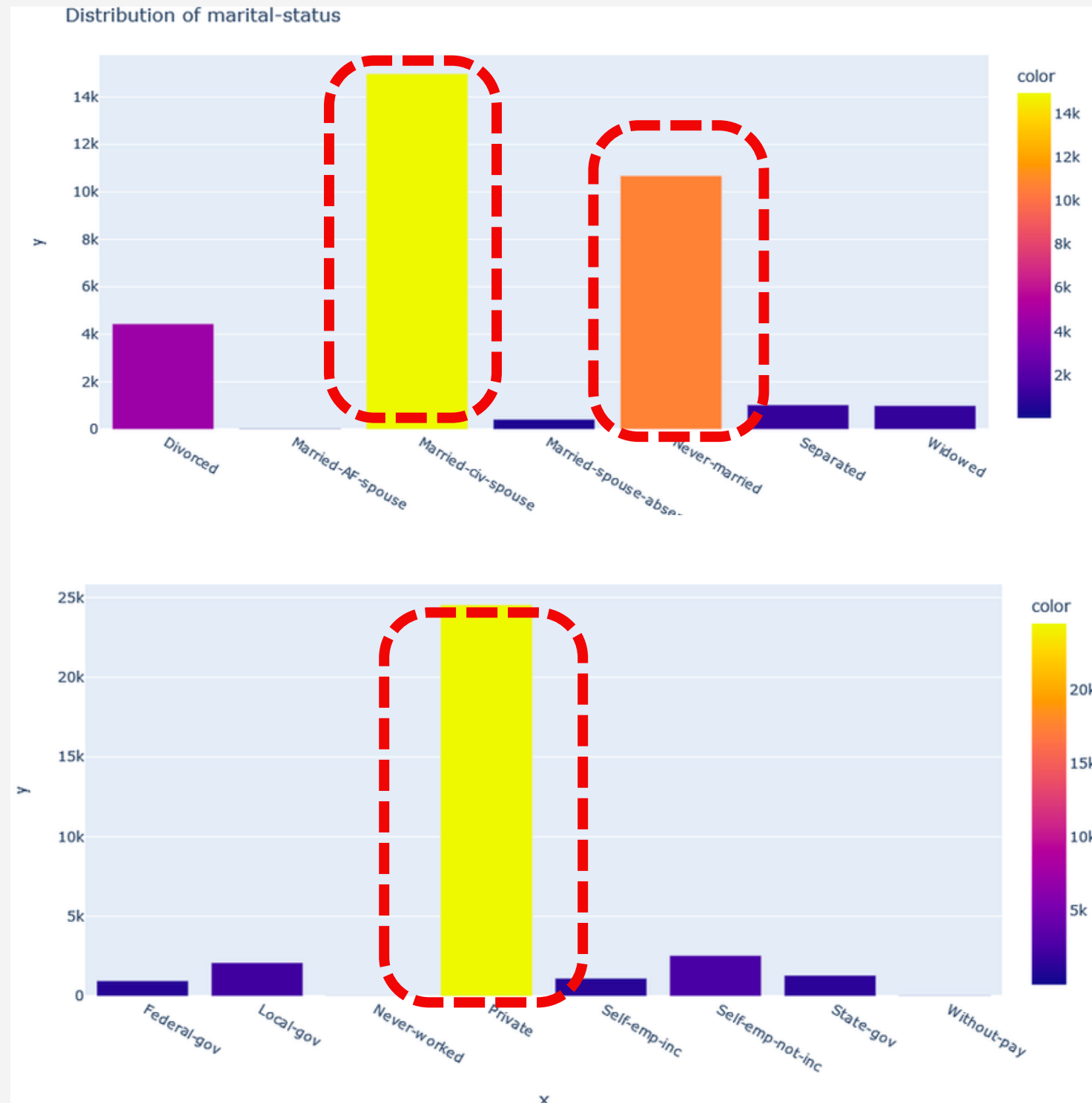
Statistics of Numeric Values (Q4)

Appendix (2)



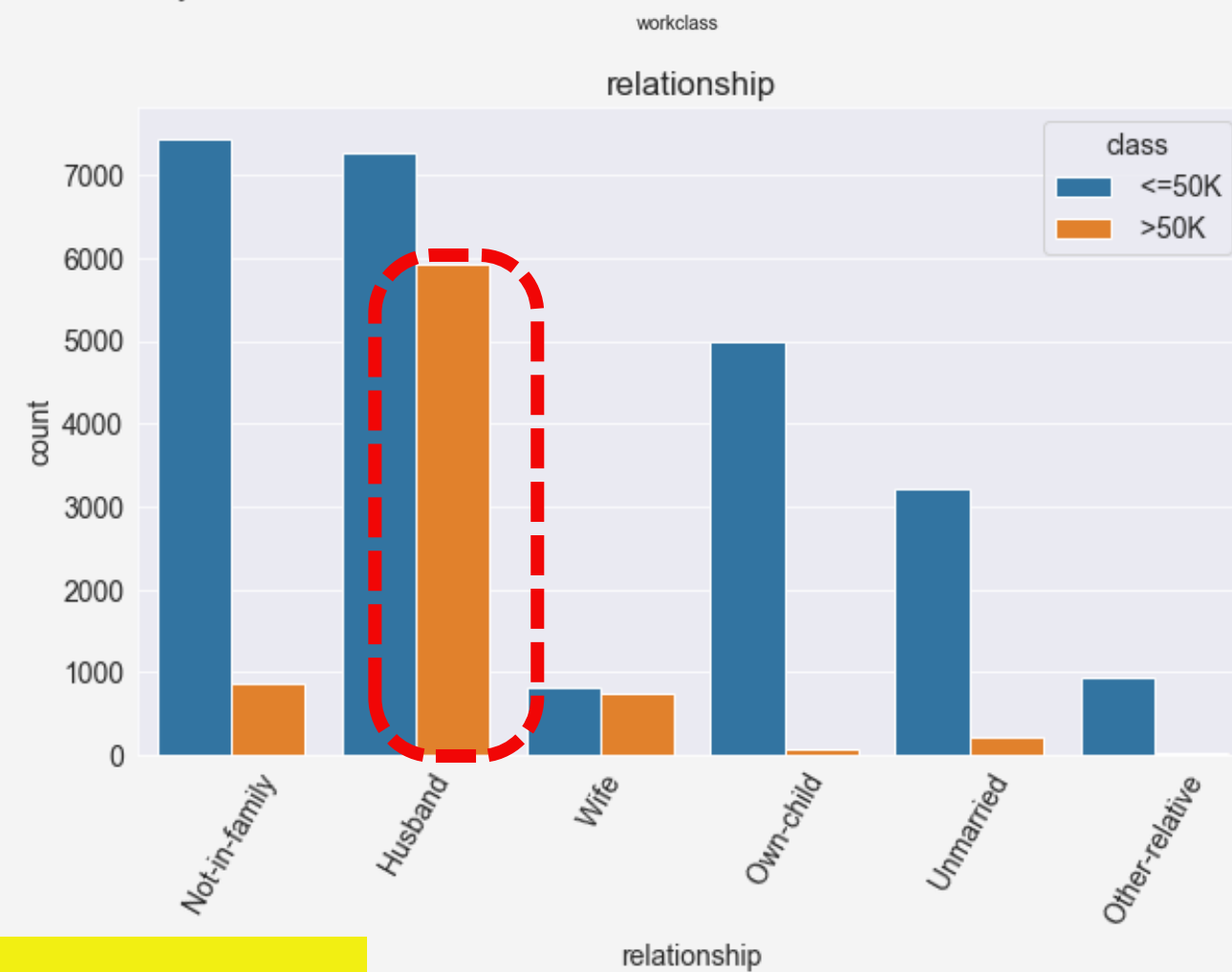
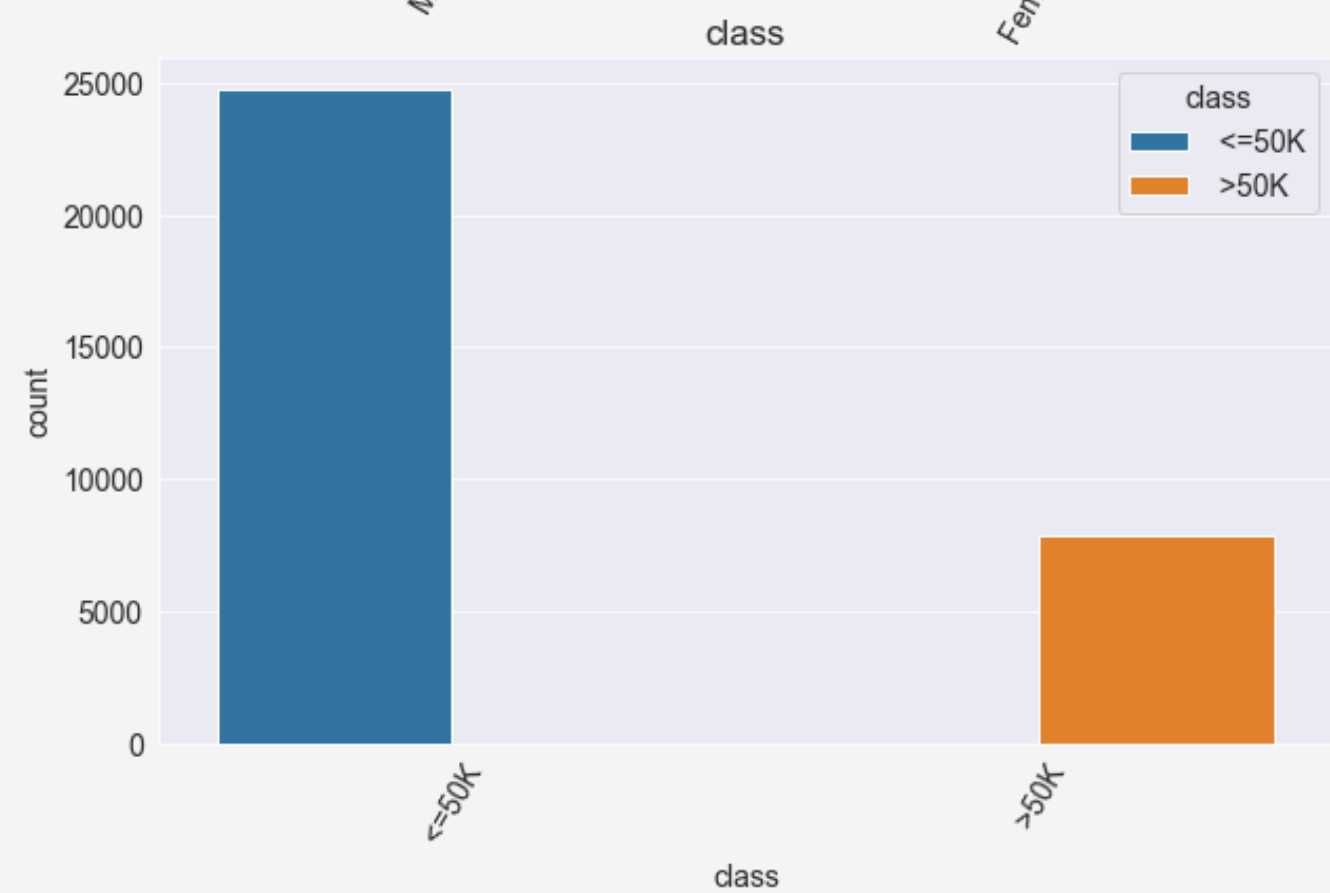
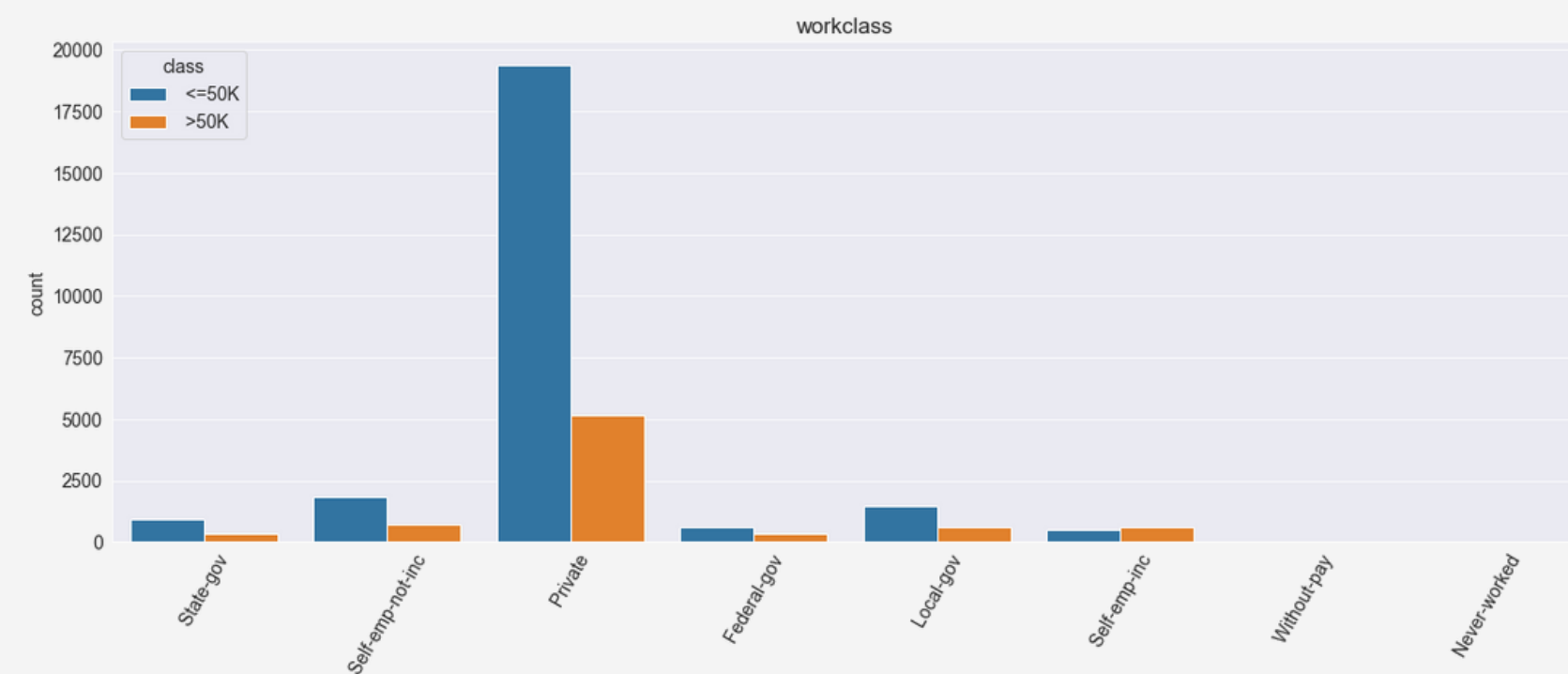
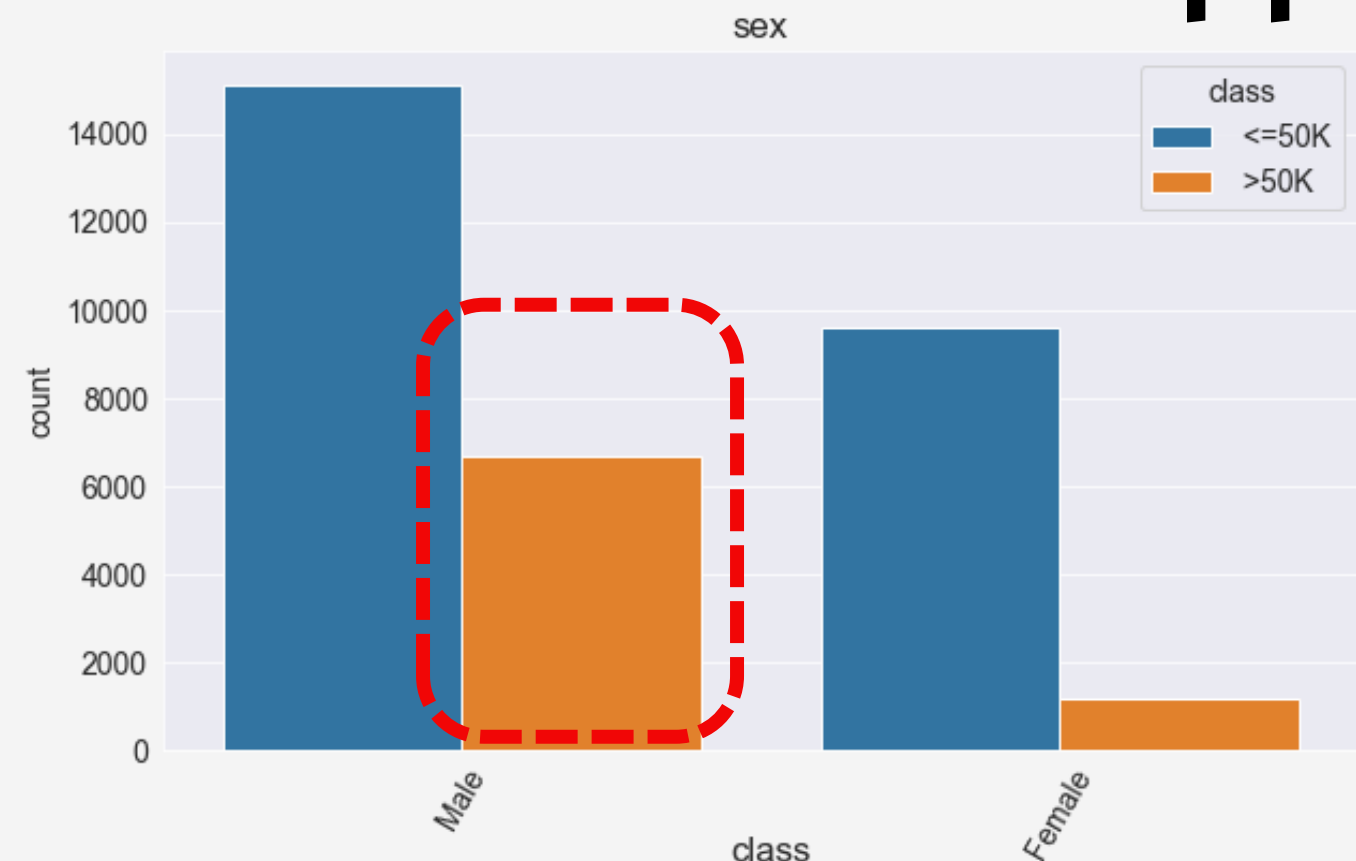
Bar Chart and Histogram of Numeric Values (Q8)

Appendix (3)



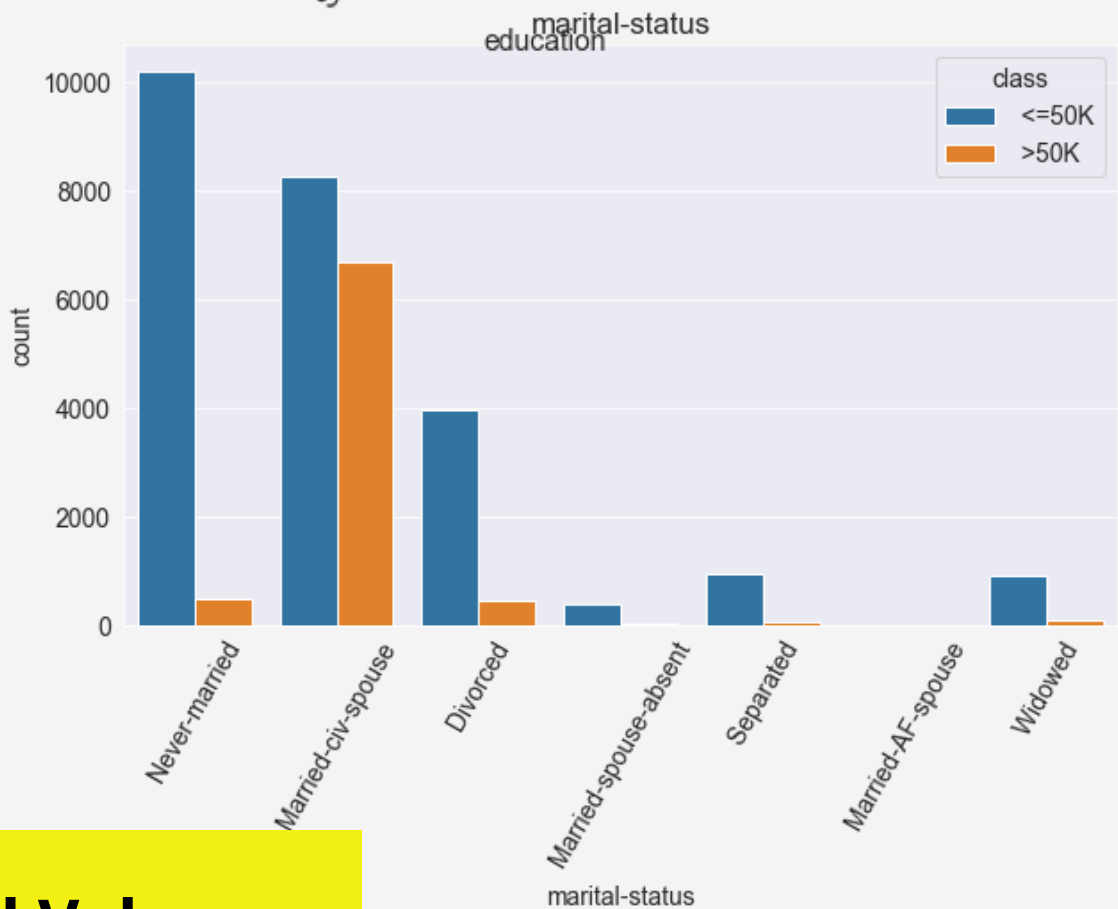
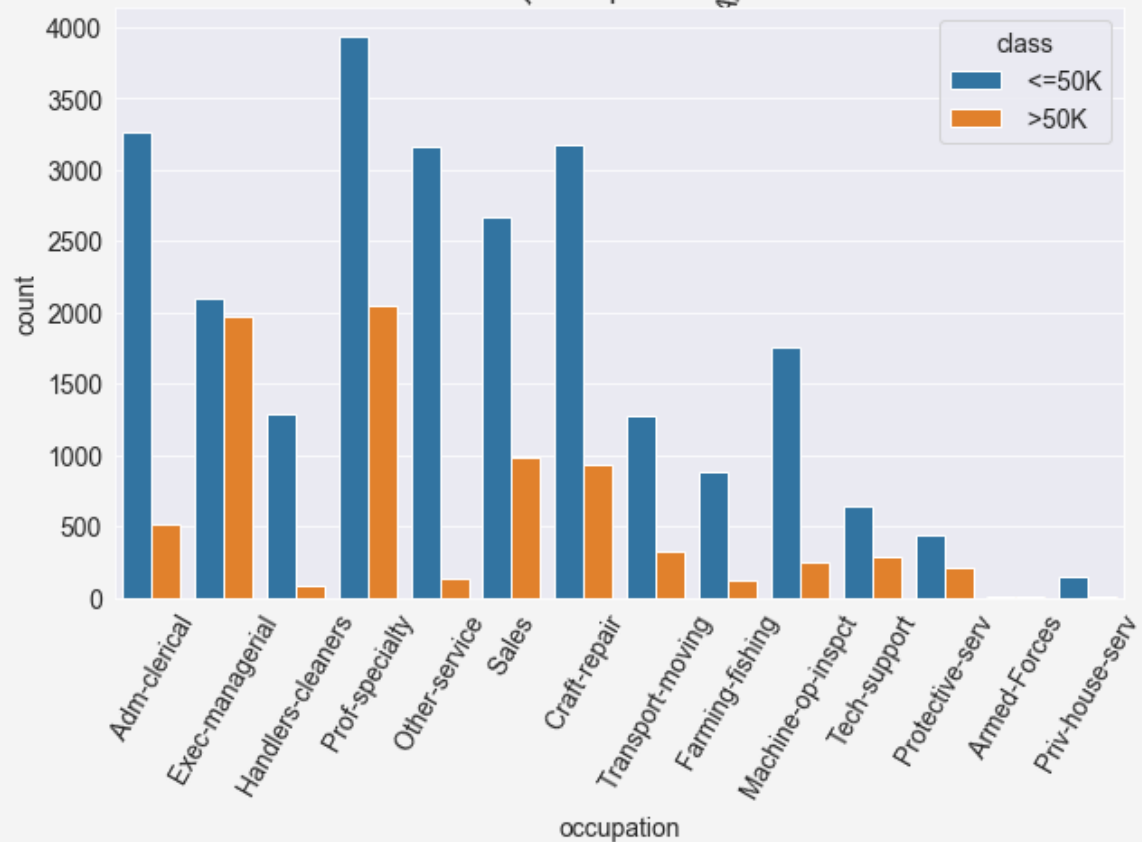
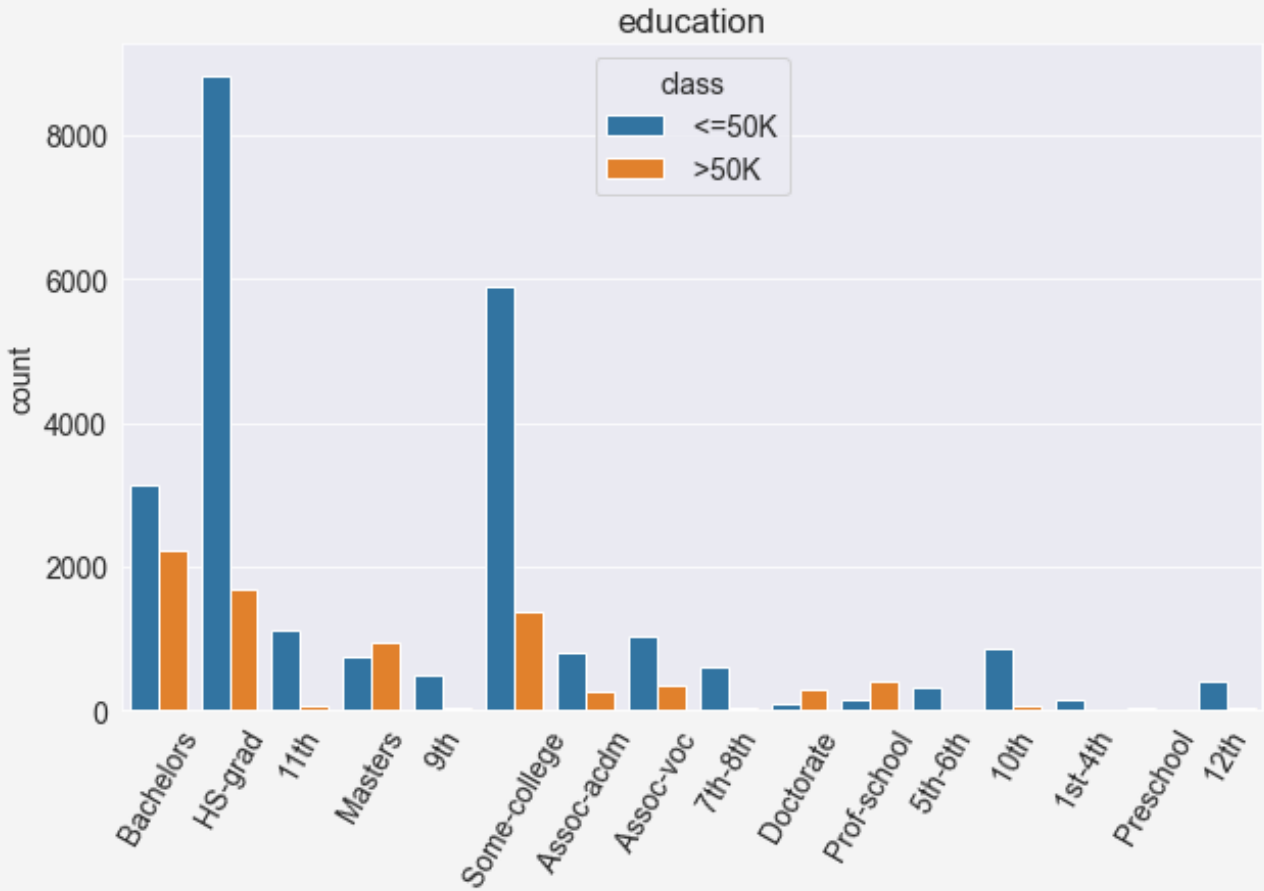
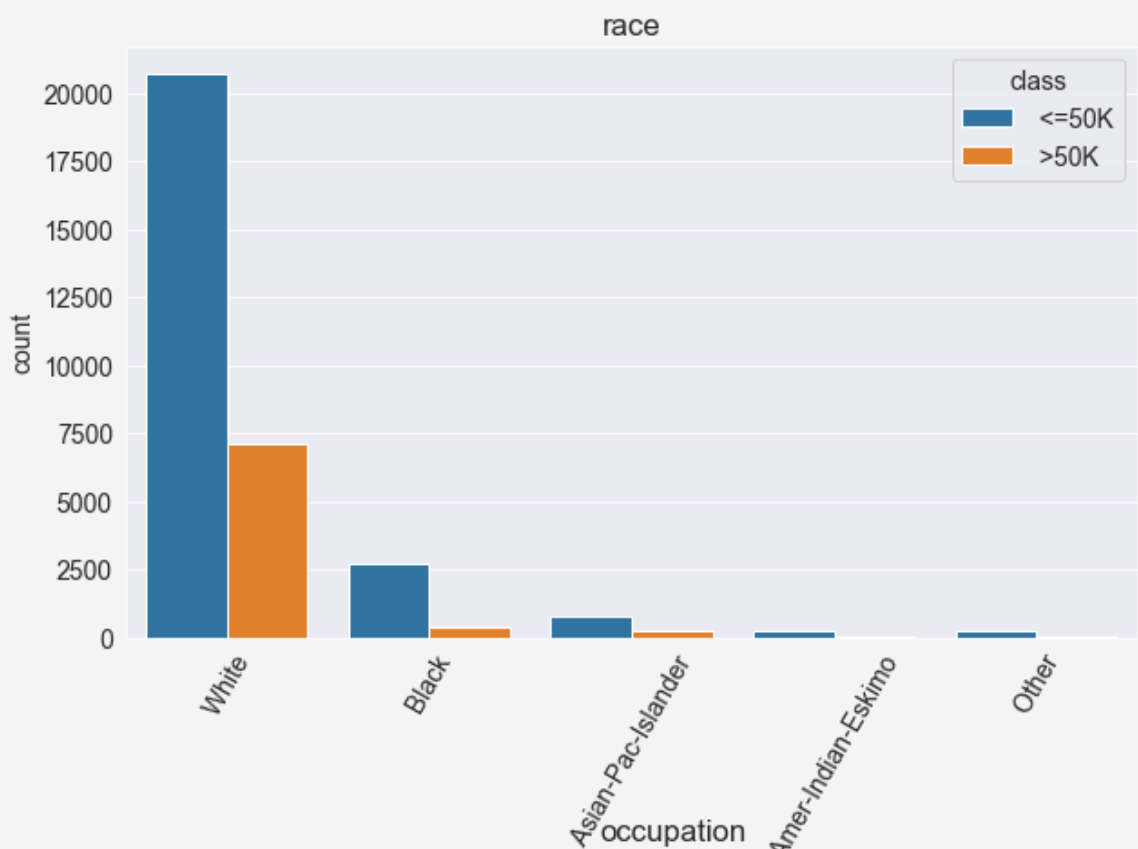
Bar Chart of Categorical Values (Q9)

Appendix (4)



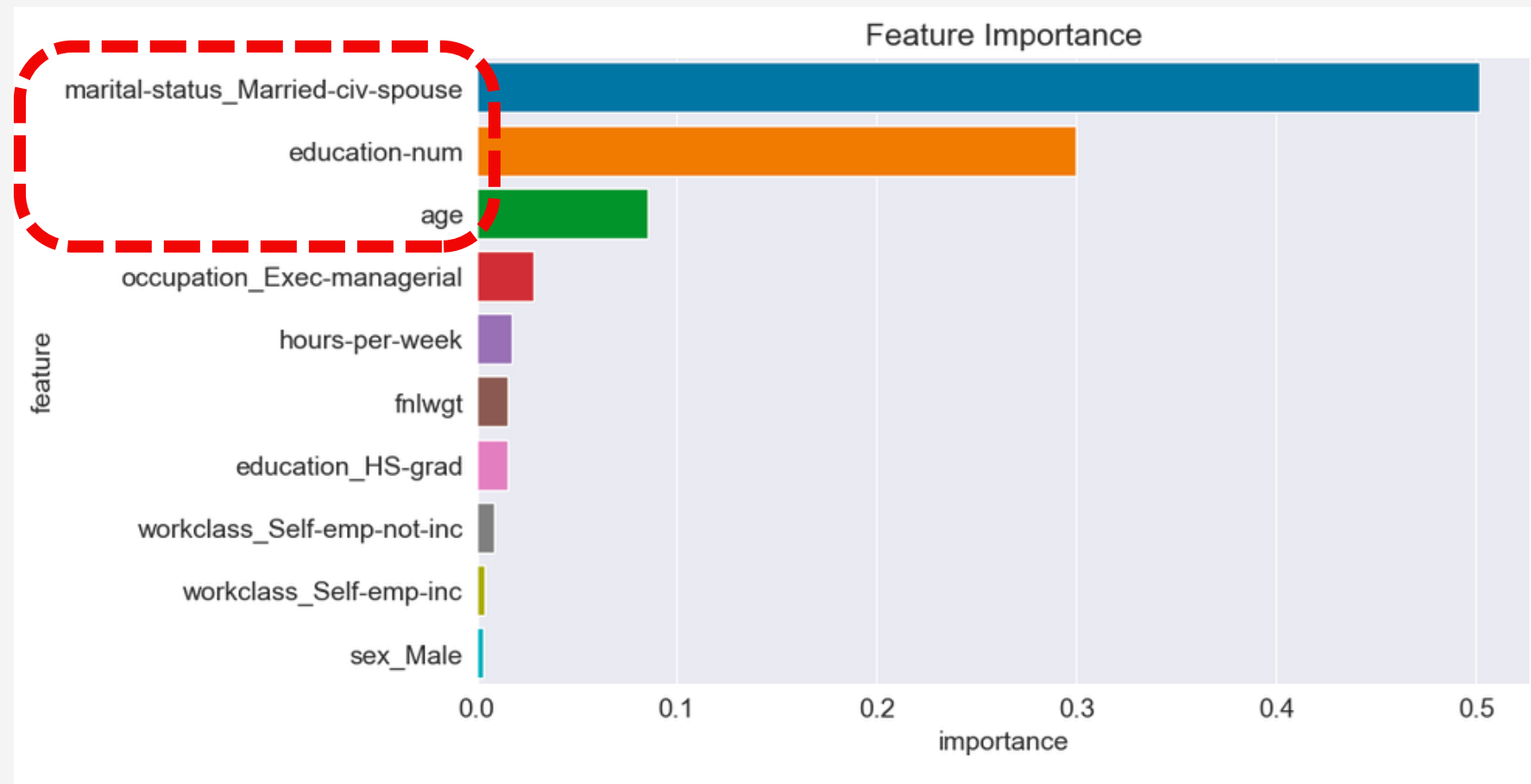
**Bar Chart of Categorical Values
Based on Income Class (Q9)**

Appendix (5)

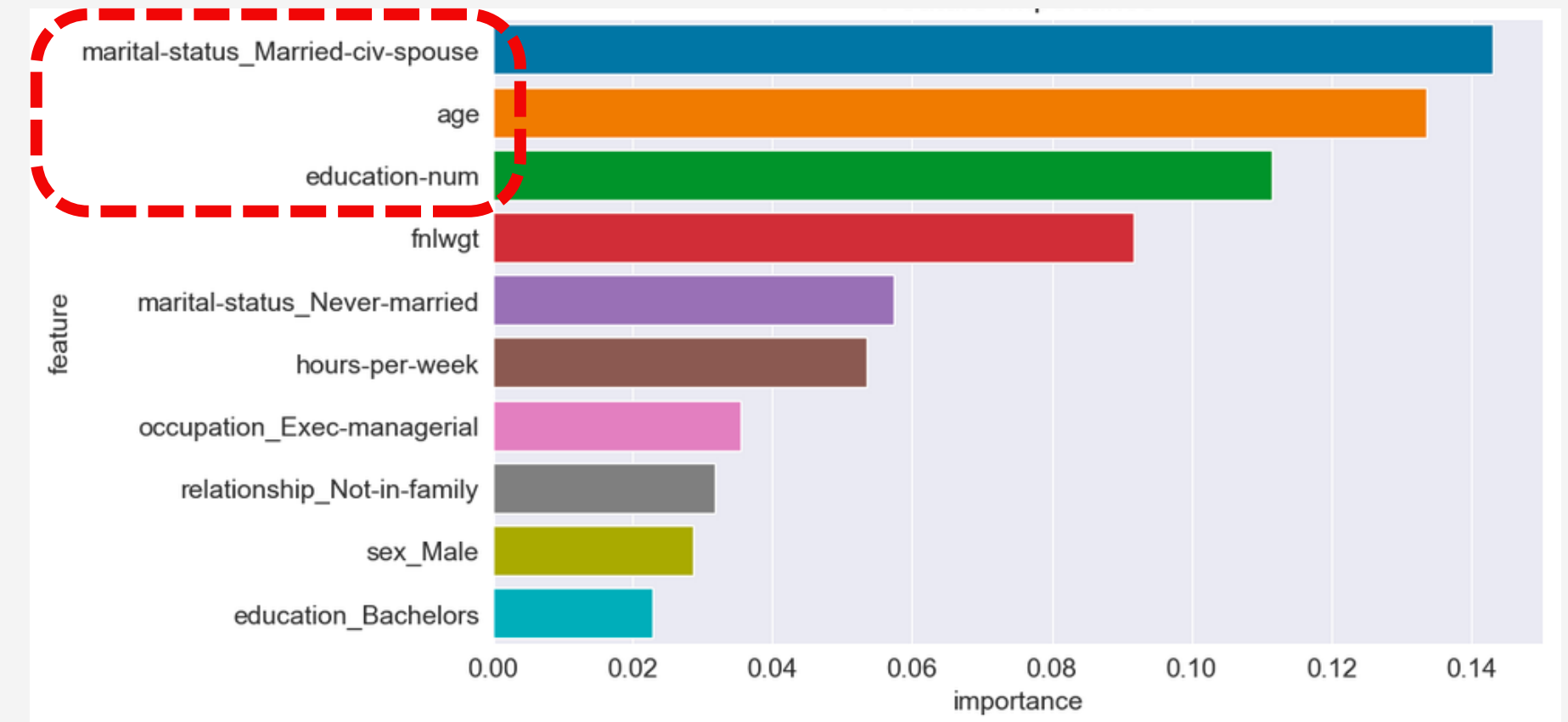


**Bar Chart of Categorical Values
Based on Income Class (Q9)**

Appendix (6)

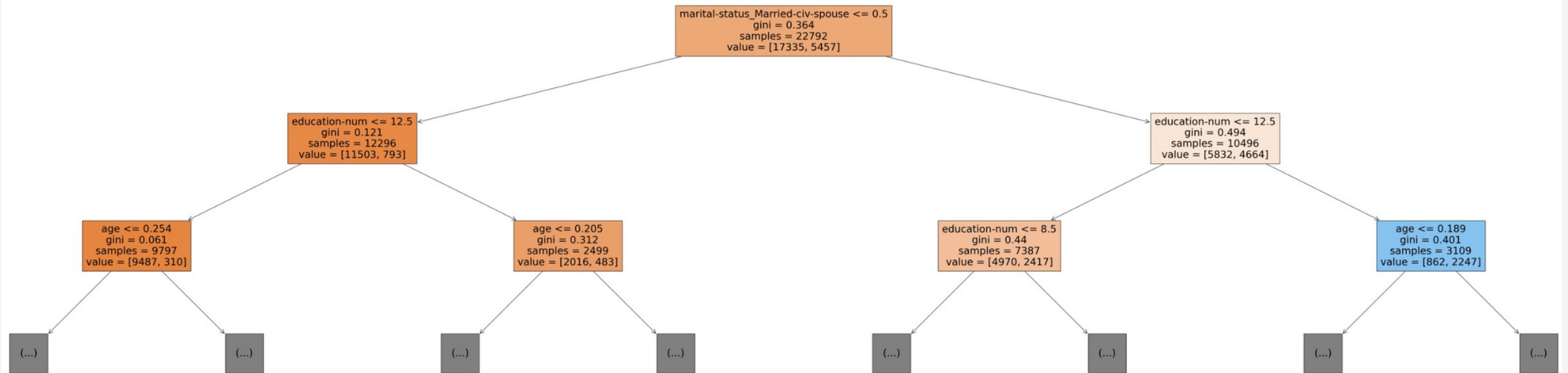


Decison Tree
Most Important Features (Q12.2)



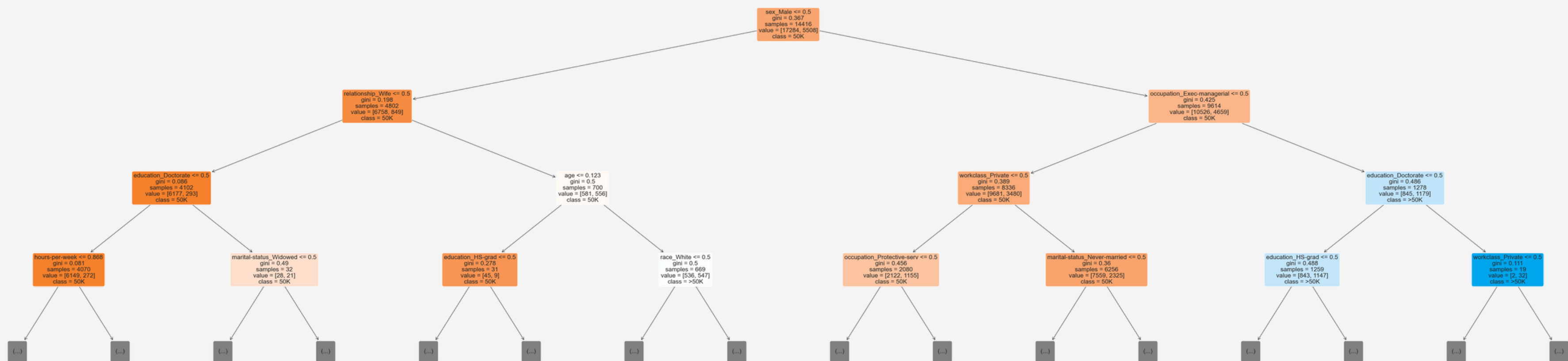
Random Forest
Most Important Features (Q12.2)

Appendix (7)



Decison Tree Structure Plot (Q12.3)

Appendix (8)



Random Forest Structure Plot (Q12.3)

Appendix (9)

	models_name	acc_train	acc_test	misclassification_rate	True Positive rate	False Positive rate	Specificity	Precision	Prevalence	ROC
0	Naive Bayes	81.580000	81.380000	17.100000	92.600000	45.800000	54.100000	61	83.000000	86.000000
1	Logistic Regression	83.400000	82.870000		82.600000	45.900000	54.100000	69	83.400000	88.800000
2	Decision Tree	83.880000	82.050000		91.500000	40.200000	59.800000	65	83.900000	86.500000
3	Rnadam Forest	90.130000	83.400000		96.700000	30.600000	69.400000	72	86.500000	88.300000

Model Metrics Comparisom (Q13)

Appendix (10)

```

=====
                        OLS Regression Results
=====
=
Dep. Variable:          class    R-squared:
0.344
Model:                OLS    Adj. R-squared:
0.341
Method:             Least Squares    F-statistic:
126.5
Date:                Thu, 10 Nov 2022    Prob (F-statistic):
0.00
Time:                15:54:38    Log-Likelihood:
8129.7
No. Observations:    22792    AIC:
1.645e+04
Df Residuals:        22697    BIC:
1.721e+04
Df Model:            94
Covariance Type:    nonrobust
=====
=====
                                coef    std err          t
P>|t|    [0.025    0.975]
-----
const                -0.1296    0.101    -1.286
0.198    -0.327    0.068
age                  0.1655    0.014    11.967
0.000    0.138    0.193
fnlwgt               0.0490    0.011    4.460
0.000    0.027    0.070
capital-gain         5.437e-15    3.99e-15    1.361
0.173    -2.39e-15    1.33e-14
capital-loss        -7.774e-16    4.15e-16    -1.874
0.061    -1.59e-15    3.59e-17

```

```

education-num        0.0338    0.002    17.281
0.000    0.030    0.038
workclass_Local-gov  -0.0963    0.016    -5.880
0.000    -0.128    -0.064
workclass_Never-worked -0.0240    0.174    -0.138
0.890    -0.365    0.317
workclass_Private     -0.0712    0.014    -5.140
0.000    -0.098    -0.044
workclass_Self-emp-inc  0.0287    0.019    1.543
0.123    -0.008    0.065
workclass_Self-emp-not-inc -0.1245    0.016    -7.694
0.000    -0.156    -0.093
workclass_State-gov   -0.1166    0.018    -6.559
0.000    -0.151    -0.082
workclass_Without-pay -0.2704    0.106    -2.561
0.010    -0.477    -0.063
education_11th       -0.0166    0.018    -0.932
0.351    -0.051    0.018
education_12th       -0.0360    0.023    -1.546
0.122    -0.082    0.010
education_1st-4th     0.0992    0.038    2.615
0.009    0.025    0.174
education_5th-6th     0.0485    0.029    1.654
0.098    -0.009    0.106
education_7th-8th    -0.0059    0.023    -0.258
0.796    -0.051    0.039
education_9th        -0.0139    0.024    -0.591
0.555    -0.060    0.032

```

t-test and p-value for logistic regression

Appendix (11)

education_Assoc-acdm	-0.0877	0.017	-5.064	occupation_Protective-serv	0.0738	0.018	4.050
0.000 -0.122 -0.054				0.000 0.038 0.109			
education_Assoc-voc	-0.0542	0.016	-3.369	occupation_Sales	0.0386	0.010	3.862
0.001 -0.086 -0.023				0.000 0.019 0.058			
education_Bachelors	-0.0123	0.014	-0.853	occupation_Tech-support	0.0684	0.015	4.512
0.394 -0.040 0.016				0.000 0.039 0.098			
education_Doctorate	0.1546	0.023	6.663	occupation_Transport-moving	-0.0245	0.013	-1.877
0.000 0.109 0.200				0.061 -0.050 0.001			
education_HS-grad	-0.0642	0.013	-5.134	relationship_Not-in-family	-0.1512	0.028	-5.326
0.000 -0.089 -0.040				0.000 -0.207 -0.096			
education_Masters	0.0540	0.017	3.202	relationship_Other-relative	-0.1263	0.028	-4.488
0.001 0.021 0.087				0.000 -0.181 -0.071			
education_Preschool	0.1649	0.061	2.700	relationship_Own-child	-0.1439	0.028	-5.080
0.007 0.045 0.285				0.000 -0.199 -0.088			
education_Prof-school	0.1727	0.021	8.137	relationship_Unmarried	-0.1364	0.029	-4.633
0.000 0.131 0.214				0.000 -0.194 -0.079			
education_Some-college	-0.0542	0.013	-4.236	relationship_Wife	0.1166	0.013	9.017
0.000 -0.079 -0.029				0.000 0.091 0.142			
marital-status_Married-AF-spouse	0.2818	0.091	3.113	race_Asian-Pac-Islander	0.0219	0.030	0.724
0.002 0.104 0.459				0.469 -0.037 0.081			
marital-status_Married-civ-spouse	0.1326	0.029	4.641	race_Black	0.0443	0.025	1.772
0.000 0.077 0.189				0.076 -0.005 0.093			
marital-status_Married-spouse-absent	0.0285	0.021	1.340	race_Other	0.0087	0.035	0.251
0.180 -0.013 0.070				0.802 -0.059 0.076			
marital-status_Never-married	-0.0040	0.009	-0.457	race_White	0.0557	0.024	2.333
0.648 -0.021 0.013				0.020 0.009 0.103			
marital-status_Separated	0.0147	0.014	1.020	sex_Male	0.0613	0.007	8.891
0.308 -0.014 0.043				0.000 0.048 0.075			
marital-status_Widowed	0.0097	0.015	0.638	native-country_Canada	-0.1027	0.105	-0.980
0.524 -0.020 0.039				0.327 -0.308 0.103			
occupation_Armed-Forces	-0.1780	0.132	-1.349	native-country_China	-0.1418	0.108	-1.313
0.177 -0.437 0.081				0.189 -0.353 0.070			
occupation_Craft-repair	-0.0074	0.010	-0.727	native-country_Columbia	-0.2565	0.115	-2.231
0.467 -0.027 0.013				0.026 -0.482 -0.031			
occupation_Exec-managerial	0.1445	0.010	14.478	native-country_Cuba	-0.1814	0.107	-1.698
0.000 0.125 0.164				0.090 -0.391 0.028			

t-test and p-value for logistic regression

Appendix (12)

sex_Male			0.0613	0.007	8.891	native-country_Honduras	-0.1529	0.143	-1.067
0.000	0.048	0.075				0.286	-0.434	0.128	
native-country_Canada			-0.1027	0.105	-0.980	native-country_Hong	-0.1906	0.131	-1.449
0.327	-0.308	0.103				0.147	-0.448	0.067	
native-country_China			-0.1418	0.108	-1.313	native-country_Hungary	-0.0349	0.157	-0.223
0.189	-0.353	0.070				0.824	-0.342	0.272	
native-country_Columbia			-0.2565	0.115	-2.231	native-country_India	-0.1851	0.105	-1.768
0.026	-0.482	-0.031				0.077	-0.390	0.020	
native-country_Cuba			-0.1814	0.107	-1.698	native-country_Iran	-0.0489	0.115	-0.425
0.090	-0.391	0.028				0.671	-0.274	0.177	
native-country_Dominican-Republic			-0.1611	0.110	-1.460	native-country_Ireland	-0.0149	0.127	-0.117
0.144	-0.377	0.055				0.907	-0.264	0.235	
native-country_Ecuador			-0.1088	0.123	-0.885	native-country_Italy	-0.0450	0.108	-0.415
0.376	-0.350	0.132				0.678	-0.258	0.168	
native-country_El-Salvador			-0.1188	0.106	-1.116	native-country_Jamaica	-0.1514	0.109	-1.391
0.264	-0.327	0.090				0.164	-0.365	0.062	
native-country_England			-0.0827	0.107	-0.775	native-country_Japan	-0.0193	0.109	-0.177
0.439	-0.292	0.127				0.860	-0.234	0.195	
native-country_France			-0.0679	0.123	-0.554	native-country_Laos	-0.1963	0.130	-1.515
0.580	-0.308	0.172				0.130	-0.450	0.058	
native-country_Germany			-0.0972	0.103	-0.939	native-country_Mexico	-0.1620	0.099	-1.630
0.348	-0.300	0.106				0.103	-0.357	0.033	
native-country_Greece			-0.1648	0.122	-1.356	native-country_Nicaragua	-0.1929	0.120	-1.609
0.175	-0.403	0.073				0.108	-0.428	0.042	
native-country_Guatemala			-0.0656	0.111	-0.591	native-country_Outlying-US (Guam-USVI-etc)	-0.2829	0.151	-1.869
0.554	-0.283	0.152				0.062	-0.580	0.014	
native-country_Haiti			-0.1668	0.116	-1.443	native-country_Peru	-0.1756	0.126	-1.393
0.149	-0.393	0.060				0.164	-0.423	0.071	
native-country_Holand-Netherlands			-0.0704	0.360	-0.195	native-country_Philippines	-0.0691	0.101	-0.687
0.845	-0.777	0.636				0.492	-0.266	0.128	

t-test and p-value for logistic regression

Appendix (13)

native-country_Peru	-0.1756	0.126	-1.393
0.164	-0.423	0.071	
native-country_Philippines	-0.0691	0.101	-0.687
0.492	-0.266	0.128	
native-country_Poland	-0.1724	0.111	-1.552
0.121	-0.390	0.045	
native-country_Portugal	-0.1442	0.124	-1.167
0.243	-0.387	0.098	
native-country_Puerto-Rico	-0.1472	0.105	-1.404
0.160	-0.353	0.058	
native-country_Scotland	-0.1217	0.157	-0.776
0.438	-0.429	0.186	
native-country_South	-0.1094	0.107	-1.023
0.306	-0.319	0.100	
native-country_Taiwan	-0.1542	0.113	-1.367
0.172	-0.375	0.067	
native-country_Thailand	-0.1692	0.142	-1.191
0.234	-0.448	0.109	
native-country_Trinidad&Tobago	-0.1172	0.133	-0.883
0.377	-0.377	0.143	
native-country_United-States	-0.1107	0.098	-1.132
0.258	-0.302	0.081	
native-country_Vietnam	-0.1276	0.108	-1.183
0.237	-0.339	0.084	
native-country_Yugoslavia	-0.0940	0.147	-0.640
0.522	-0.382	0.194	

=====

Omnibus: 991.186 Durbin-Watson: 2.005

Prob(Omnibus):	0.000	Jarque-Bera (JB):	
1124.379			
Skew:	0.543	Prob(JB):	6.98e-
245			
Kurtosis:	2.922	Cond. No.	
1.03e+16			
=====			
=			

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The smallest eigenvalue is 2.4e-26. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.

t-test and p-value for logistic regression