Classifying Customers by Earnings

Luis Escárcega

Introduction

- We have a dataset containing information about several customers.
- Numerical:
  - Age of client, final weight, number of years of education, capital gain, capital loss and hours of work per week.
- Categorical:
  - Type of work, education level, marital status, ocupation, type of relationship, type of race, sex and native country.

- For this problem we will assume that we are interested in those individuals who earn more than $50k (e.g., to offer them loans). We will refer to these as the positive class. The remainder we will refer to as the negative class.
- We will try to find a model that gives us scores that separate as much as possible the positive class from the negative class.
- Throughout the problem we will assume that the records come from a random sample, i.e., they are independent and identically distributed.

Train and test split

- Before starting to explore the database, it is necessary to split the data into training and test sets.
- In this problem we will not use a validation (also known as development) set, which is necessary to obtain unbiased estimates of the performance of various models. Instead, we will use cross-validation.
    - Number of individuals in the training set: 29304
    - Number of individuals in the test set: 3257

Exploratory Data Analysis

- In this section we show some summaries regarding the variables we took from the customers and their relationship with the target variable.
- We present:
  - The number of categories
  - The Herfindahl-Hirschman index, which measures how "diverse" the categorical variable is. Values close to zero indicate that the distribution of the training sample in each of the categories is approximately uniform, while values close to one indicate that a single category encompasses a large part of the sample.

- Also:
    - The Gini index of the variables, as well as 99% bootstrap confidence intervals. This metric takes values in [-1, 1]. Absolute values close to 1 indicate that the variable has a lot of predictive power, while values around zero indicate null predictive power.

Summary of type of work:

- Number of categories: 9
- Most common category: Private
- Highest proportion of positives: Self-emp-inc
- Lowest proportion of positives: Never-worked
- Herfindahl–Hirschman index: 44.07%
- Gini index: 16.62%
- Gini index confidence interval: (14.97%, 18.33%)

Summary of type of work:

type_of_work	mean_target	count
Never-worked	0.000000	7
Without-pay	0.000000	13
?	0.102454	1630
Private	0.218311	20425
State-gov	0.270178	1177
Self-emp-not-inc	0.280122	2299
Local-gov	0.296158	1874
Federal-gov	0.384259	864
Self-emp-inc	0.556650	1015

Summary of education level:

- Number of categories: 16
- Most common category: HS-grad
- Highest proportion of positives: Doctorate
- Lowest proportion of positives: Preschool
- Herfindahl–Hirschman index: 13.68%
- Gini index: 43.33%
- Gini index confidence interval: (41.63%, 45.14%)

Summary of education level:

education_level	mean_target	count
Preschool	0.000000	45
1st-4th	0.038217	157
5th-6th	0.046053	304
11th	0.050095	1058
9th	0.057906	449
7th-8th	0.065719	563
10th	0.067296	847
12th	0.078534	382
HS-grad	0.157778	9469
Some-college	0.191073	6542
Assoc-acdm	0.247156	967
Assoc-voc	0.261847	1245
Bachelors	0.413380	4843
Masters	0.554906	1539
Prof-school	0.724806	516
Doctorate	0.735450	378

Summary of marital status:

- Number of categories: 7
- Most common category: Married-civ-spouse
- Highest proportion of positives: Married-civ-spouse
- Lowest proportion of positives: Never-married
- Herfindahl–Hirschman index: 23.03%
- Gini index: 53.91%
- Gini index confidence interval: (52.56%, 55.28%)

Summary of marital status:

marital_status	mean_target	count
Never-married	0.045384	9629
Separated	0.063373	931
Married-spouse-absent	0.077540	374
Widowed	0.084842	884
Divorced	0.106083	3978
Married-AF-spouse	0.434783	23
Married-civ-spouse	0.445532	13485

Summary of ocupation:

- Number of categories: 15
- Most common category: Prof-specialty
- Highest proportion of positives: Exec-managerial
- Lowest proportion of positives: Priv-house-serv
- Herfindahl–Hirschman index: 3.25%
- Gini index: 46.09%
- Gini index confidence interval: (44.43%, 47.74%)

Summary of ocupation:

ocupation	mean_target	count
Priv-house-serv	0.007042	142
Other-service	0.042396	2972
Handlers-cleaners	0.064019	1234
?	0.102016	1637
Farming-fishing	0.112335	908
Armed-Forces	0.125000	8
Machine-op-inspct	0.125207	1813
Adm-clerical	0.134977	3408
Transport-moving	0.200000	1430
Craft-repair	0.227273	3696
Sales	0.271895	3277
Tech-support	0.305325	845
Protective-serv	0.326425	579
Prof-specialty	0.443577	3713
Exec-managerial	0.484898	3642

Summary of type of relationship:

- Number of categories: 6
- Most common category: Husband
- Highest proportion of positives: Wife
- Lowest proportion of positives: Own-child
- Herfindahl–Hirschman index: 12.13%
- Gini index: 55.92%
- Gini index confidence interval: (54.58%, 57.24%)

Summary of type of relationship:

type_of_relationship	mean_target	count
Own-child	0.013345	4571
Other-relative	0.035915	891
Unmarried	0.063933	3097
Not-in-family	0.102764	7454
Husband	0.446250	11879
Wife	0.483003	1412

Summary of type of race:

- Number of categories: 5
- Most common category: White
- Highest proportion of positives: Asian-Pac-Islander
- Lowest proportion of positives: Other
- Herfindahl–Hirschman index: 67.47%
- Gini index: 7.44%
- Gini index confidence interval: (6.42%, 8.60%)

Summary of type of race:

type_of_race	mean_target	count
Other	0.087866	239
Amer-Indian-Eskimo	0.113553	273
Black	0.125000	2808
White	0.255324	25027
Asian-Pac-Islander	0.258098	957

Summary of sex:

- Number of categories: 3
- Most common category: Male
- Highest proportion of positives: ?
- Lowest proportion of positives: Female
- Herfindahl–Hirschman index: 33.38%
- Gini index: 23.65%
- Gini index confidence interval: (22.40%, 24.88%)

Summary of sex:

sex	mean_target	count
Female	0.110174	9721
Male	0.304763	19566
?	0.352941	17

Summary of native country:

- Number of categories: 42
- Most common category: United-States
- Highest proportion of positives: France
- Lowest proportion of positives: Outlying-US(Guam-USVI-etc)
- Herfindahl–Hirschman index: 79.78%
- Gini index: 6.03%
- Gini index confidence interval: (5.05%, 7.12%)

Summary of native country:

native_country	mean_target	count
Outlying-US(Guam-USVI-etc)	0.000000	14
Holand-Netherlands	0.000000	1
Dominican-Republic	0.030303	66
Columbia	0.036364	55
Mexico	0.054329	589
Guatemala	0.054545	55
Nicaragua	0.060606	33
Peru	0.066667	30
Vietnam	0.079365	63
Honduras	0.083333	12
Puerto-Rico	0.089109	101
El-Salvador	0.089888	89
Haiti	0.097561	41
Trinadad&Tobago	0.111111	18
Portugal	0.117647	34
Laos	0.117647	17
Ecuador	0.125000	24

Jamaica	0.135135	74
Thailand	0.176471	17
Hungary	0.181818	11
South	0.184211	76
Poland	0.200000	50
?	0.245136	514
United-States	0.245208	26239
China	0.246377	69
Ireland	0.250000	20
Cuba	0.253012	83
Greece	0.285714	28
Philippines	0.306011	183
Hong	0.312500	16
Canada	0.317757	107
Scotland	0.333333	9
Germany	0.333333	126
Italy	0.343284	67
England	0.352941	85

Cambodia	0.352941	17
Yugoslavia	0.375000	16
Taiwan	0.382979	47
Iran	0.400000	35
India	0.406593	91
Japan	0.418182	55
France	0.444444	27

Remark:

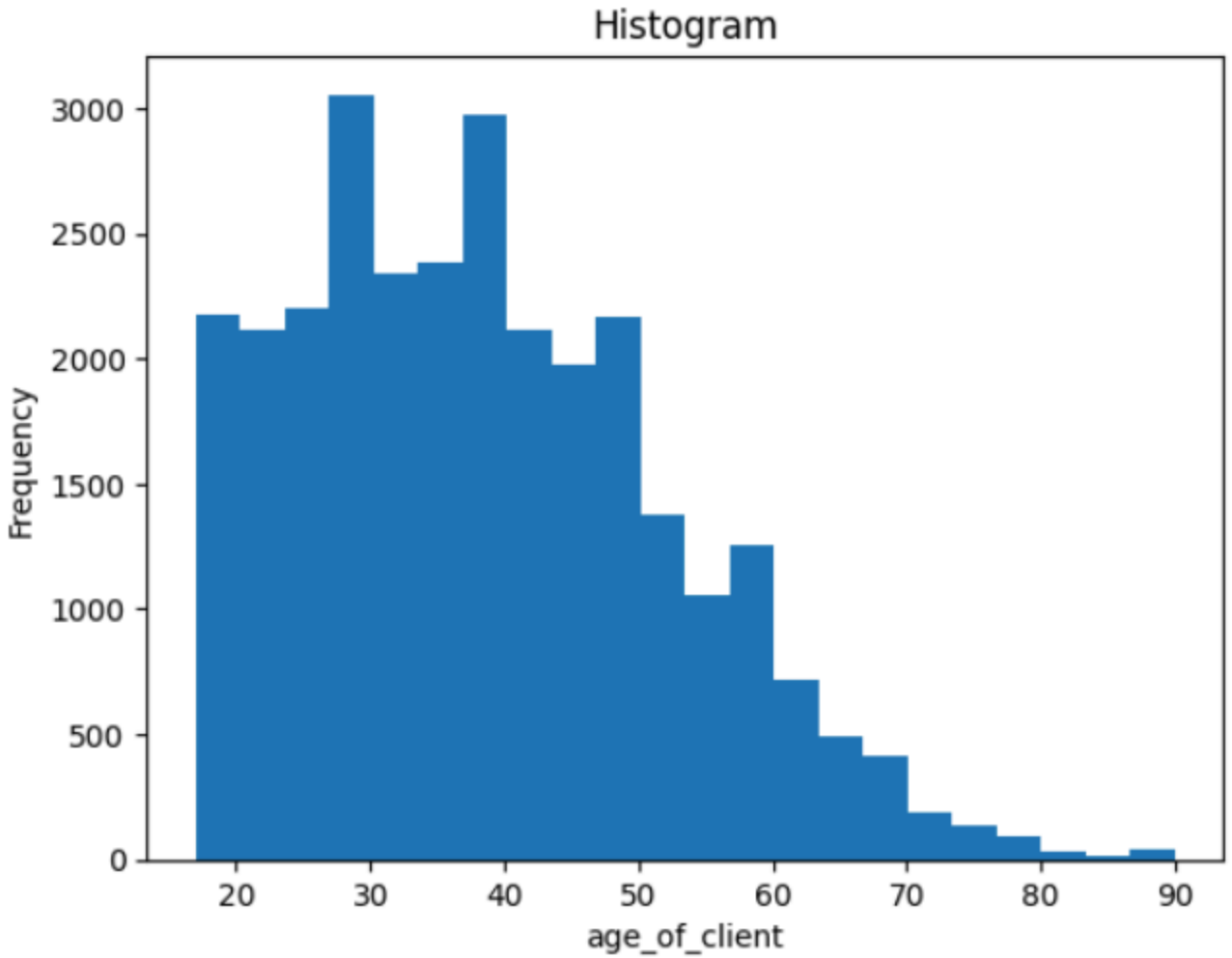
- We can see that the most diverse categorical variable is, according to the HH index, “occupation” followed by “type_of_relationship”. The least diverse is “native_country” followed by “type_of_race”.
- Note that the most predominant native nationality and race are respectively United States and white. Therefore, it is not to be expected that any model fitted on this training set would have good predictions in populations where the majority native nationality is not the United States or where the predominant race is not white.

- From the confidence intervals we can see that all categorical variables have predictive power (since they do not contain zero), although not all are equally powerful. The most powerful is “type_of_relationship”, while the least powerful is “native_country”.
- Note that all levels corresponding to an elementary school level of education have about the same odds of individuals having an income above $50k, of about 5%. Therefore, categories such as “1st-4th”, “5th-6th”,..., “12th” can be joined. This will be done below in an automatic way.

Summary of age of client:

- Mean: 38.55
- Median: 37.00
- Standard Deviation: 13.65
- Range: 73.00
- Skewness: 0.56
- Kurtosis: -0.17
- Gini index: 36.94%
- Gini index confidence interval: (35.24%, 38.52%)

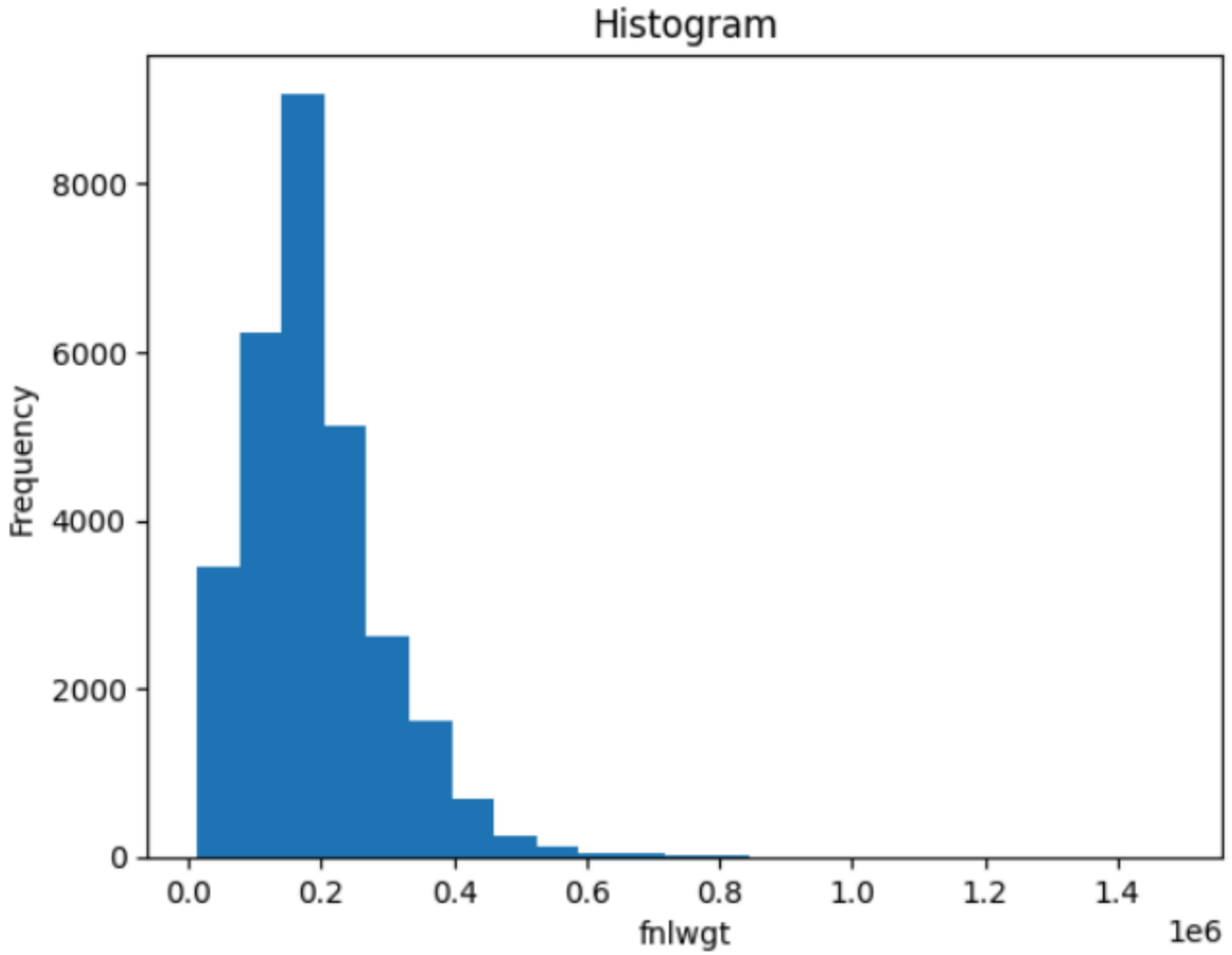
Summary of age of client:



Summary of final weight:

- Mean: 189909.67
- Median: 178469.50
- Standard Deviation: 105655.82
- Range: 1472420.00
- Skewness: 1.47
- Kurtosis: 6.50
- Gini index: -1.07%
- Gini index confidence interval: (-3.06%, 1.09%)

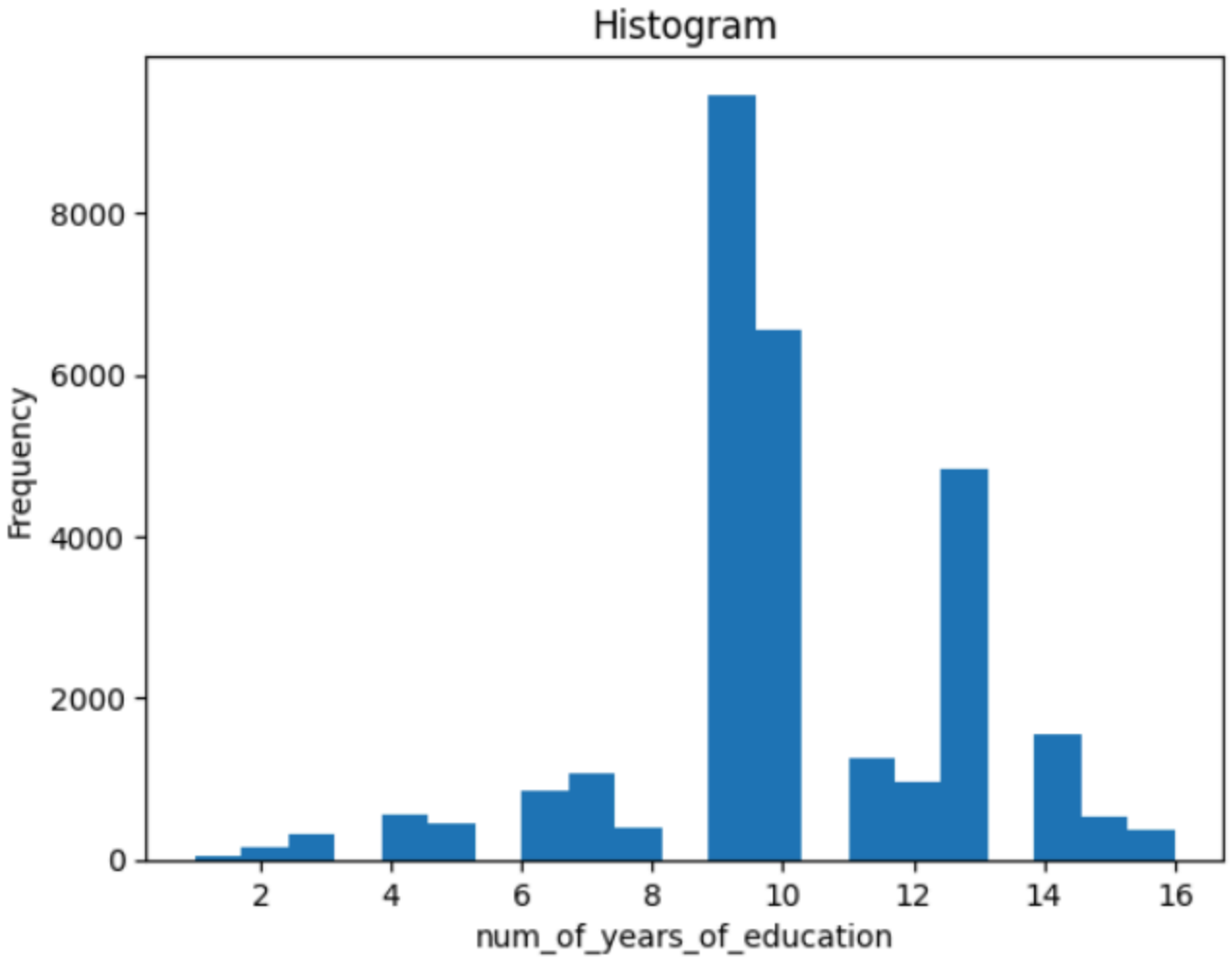
Summary of final weight:



Summary of number of years of education:

- Mean: 10.09
- Median: 10.00
- Standard Deviation: 2.57
- Range: 15.00
- Skewness: -0.31
- Kurtosis: 0.63
- Gini index: 43.27%
- Gini index confidence interval: (41.55%, 45.09%)

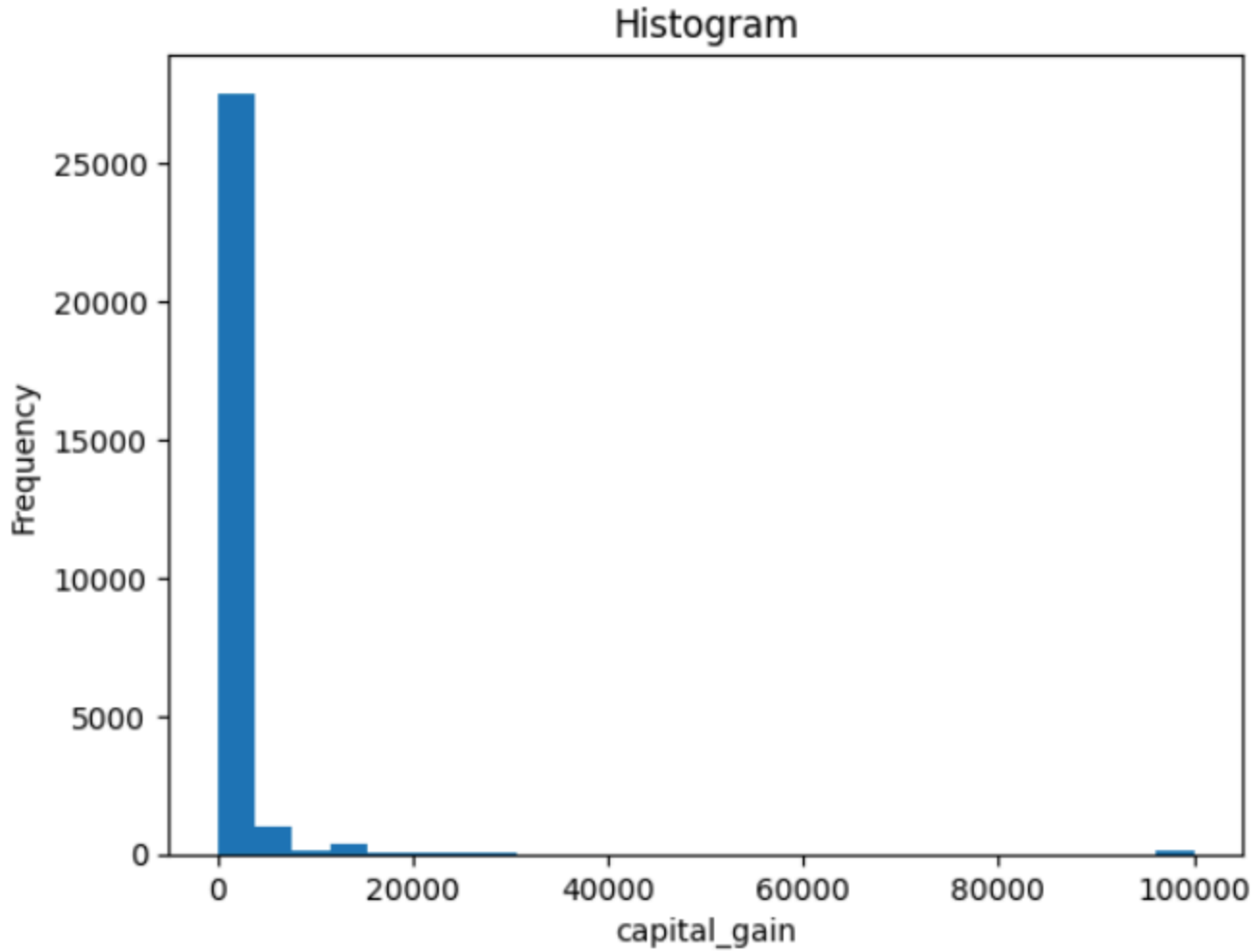
Summary of number of years of education:



Summary of capital gain:

- Mean: 1071.23
- Median: 0.00
- Standard Deviation: 7363.07
- Range: 99999.00
- Skewness: 11.98
- Kurtosis: 155.52
- Gini index: 17.90%
- Gini index confidence interval: (16.56%, 19.13%)

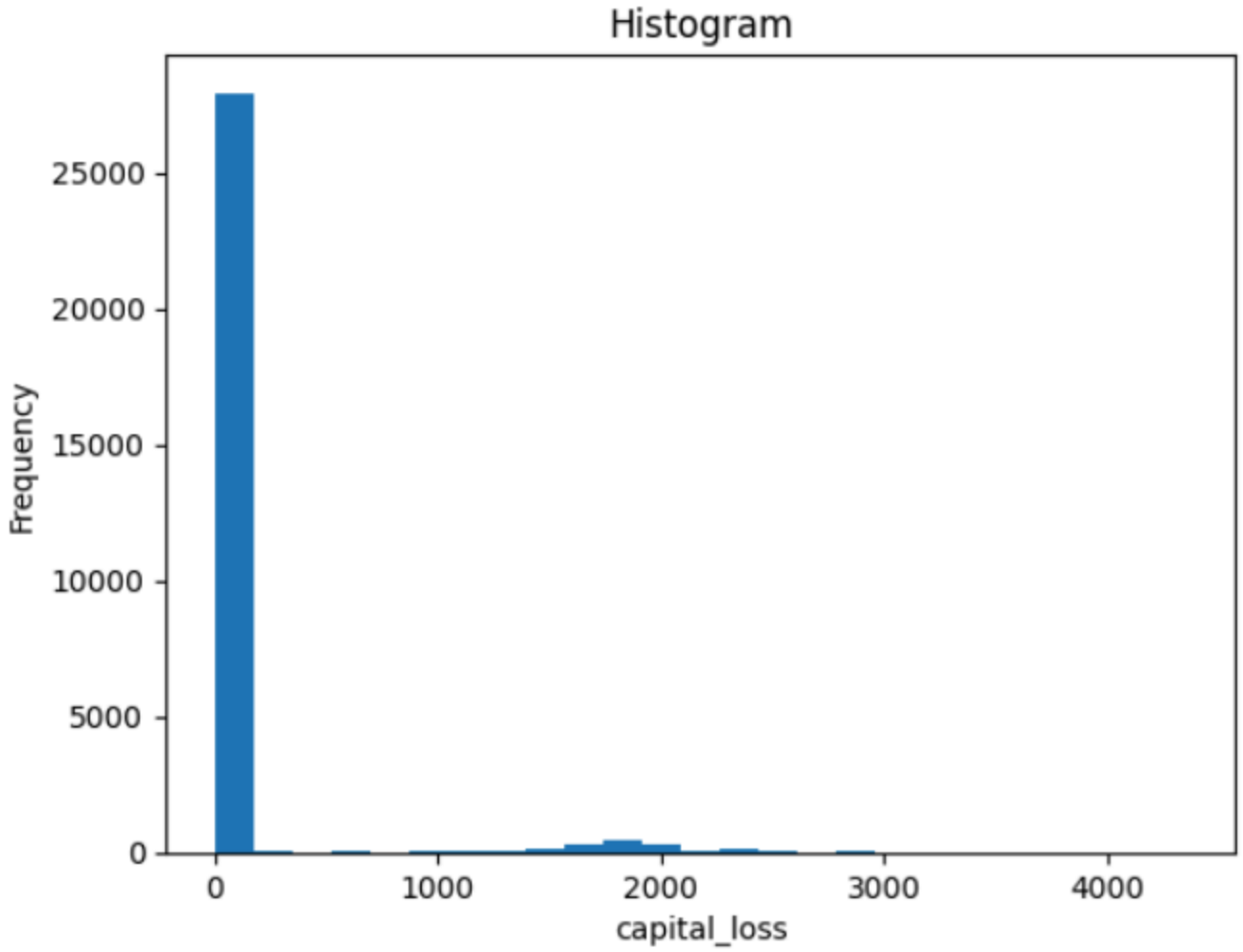
Summary of capital gain:



Summary of capital loss:

- Mean: 86.73
- Median: 0.00
- Standard Deviation: 401.25
- Range: 4356.00
- Skewness: 4.62
- Kurtosis: 20.70
- Gini index: 6.80%
- Gini index confidence interval: (5.88%, 7.71%)

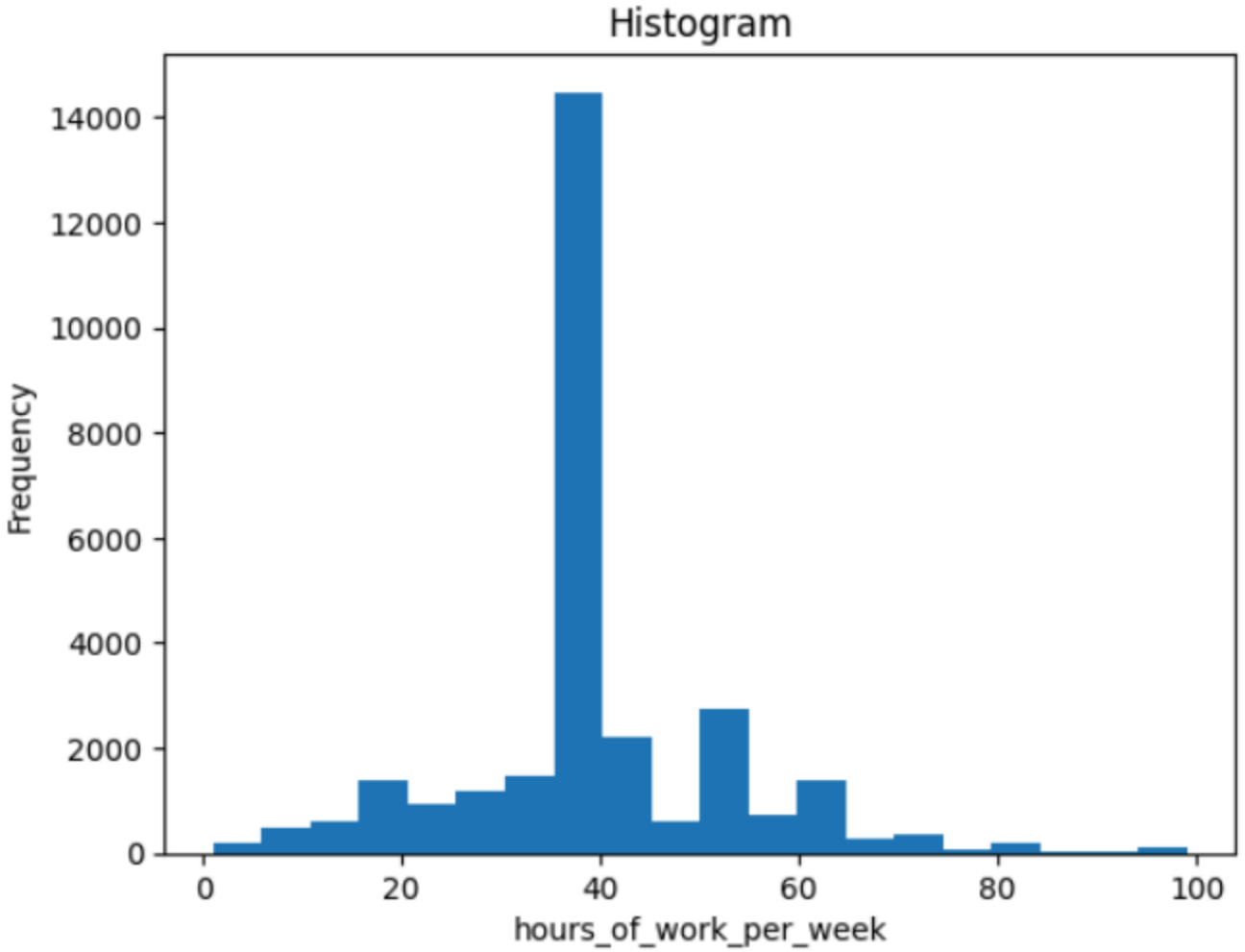
Summary of capital loss:



Summary of hours of work per week:

- Mean: 40.46
- Median: 40.00
- Standard Deviation: 12.39
- Range: 98.00
- Skewness: 0.24
- Kurtosis: 2.92
- Gini index: 34.35%
- Gini index confidence interval: (32.51%, 35.84%)

Summary of hours of work per week:



Remarks:

- We see that all numerical variables have predictive power, because their gini is statistically different from zero, except for “final weight” whose confidence interval contains zero.
- Some distributions are highly skewed to the right, for example, capital gain and capital loss.

Preprocessing

In this part, we are going to make some treatments to the training base. Specifically:

- We will reduce the number of levels per categorical variable.
    - As described above, in some variables we can see categories that have approximately the same proportion of positives. There are some statistical tests that can be used to determine whether there is evidence that the proportions at two different levels are equal. We will be using tests based on Agresti-Coull 99% confidence intervals. If it cannot be rejected that the proportions are different, the categories are merged.
    - This is done iteratively until there are no levels left that can be joined, i.e., all confidence intervals are disjoint.

- It is evident from the histograms that some numerical variables show outliers. We will eliminate the observations below the 0.05% percentile and above the 99.95% percentile. There are other techniques to do this, but we prefer this one for its simplicity.
- Since the variable final weight has a Gini index statistically equal to zero, we will eliminate this variable.
- We use one hot encoding to transform the categorical variables, including the target.
- Finally, the numerical predictors are standardized to have zero mean and unit standard deviation.

Merged levels for type of work:

	negative	positive	proportion
(?, Never-worked, Without-pay)	1483	167	0.101212
(Private,)	15966	4459	0.218311
(Local-gov, Self-emp-not-inc, State-gov)	3833	1517	0.283551
(Federal-gov,)	532	332	0.384259
(Self-emp-inc,)	450	565	0.556650

Merged levels for education level:

	negative	positive	proportion
(10th, 11th, 12th, 1st-4th, 5th-6th, 7th-8th, 9th, Preschool)	3582	223	0.058607
(HS-grad,)	7975	1494	0.157778
(Some-college,)	5292	1250	0.191073
(Assoc-acdm, Assoc-voc)	1647	565	0.255425
(Bachelors,)	2841	2002	0.413380
(Masters,)	685	854	0.554906
(Doctorate, Prof-school)	242	652	0.729306

Merged levels for marital status:

	negative	positive	proportion
(Never-married, Separated)	10064	496	0.046970
(Divorced, Married-spouse-absent, Widowed)	4710	526	0.100458
(Married-AF-spouse, Married-civ-spouse)	7490	6018	0.445514

Merged levels for ocupation:

	negative	positive	proportion
(Handlers-cleaners, Other-service, Priv-house-serv)	4142	206	0.047378
(?, Adm-clerical, Armed-Forces, Farming-fishing, Machine-op-inspct)	6817	957	0.123103
(Craft-repair, Transport-moving)	4000	1126	0.219664
(Protective-serv, Sales, Tech-support)	3363	1338	0.284620
(Exec-managerial, Prof-specialty)	3942	3413	0.464038

Merged levels for type of relationship:

	negative	positive	proportion
(Own-child,)	4510	61	0.013345
(Other-relative, Unmarried)	3758	230	0.057673
(Not-in-family,)	6688	766	0.102764
(Husband, Wife)	7308	5983	0.450154

Merged levels for type of race:

	negative	positive	proportion
(Amer-Indian-Eskimo, Black, Other)	2917	403	0.121386
(Asian-Pac-Islander, White)	19347	6637	0.255426

Merged levels for sex:

	negative	positive	proportion
(Female,)	8650	1071	0.110174
(?, Male)	13614	5969	0.304805

Merged levels for native country:

	negative	positive	proportion
(Columbia, Dominican-Republic, El-Salvador, Guatemala, Mexico, Puerto-Rico, Vietnam)	957	61	0.059921
(?, Cambodia, Canada, China, Cuba, Ecuador, England, France, Germany, Greece, Haiti, Holand-Netherlands, Honduras, Hong, Hungary, India, Iran, Ireland, Italy, Jamaica, Japan, Laos, Nicaragua, Outlying-US(Guam-USVI-etc), Peru, Philippines, Poland, Portugal, Scotland, South, Taiwan, Thailand, Trinadad&Tobago, United-States, Yugoslavia)	21307	6979	0.246730

Possible models

- For this type of customer classification problem there are a large number of model options, some more complex than others. For example:

1. Logistic Regression:
    - This is a classic statistical model. It is based on the idea that the probability of belonging to the positive class depends only on a linear combination of predictors.
    - Pros: Simple to understand and study
    - Cons: Tends to be suboptimal.

2. Neural Networks
    - It is a generalization of logistic regression and is based on the idea of iteratively constructing new predictors through linear combinations and activation functions.
    - Pros: It is a flexible model that adapts to any type of problem.
    - Cons: It tends to be over-parameterized, causing over-fitting. Can be complicated to put into production.

3. Decision Trees
    - It is based on the idea of building optimal customer segments. The prediction of a customer can be obtained iteratively, being at each step the verification or not of a condition.
    - Pros: They are easy to understand, adjust and implement.
    - Cons: Tends to tend to have a lot of variance. Also, decision boundaries are complex.

4. Gradient Boosting Decision Trees
    - It is a generalization of decision trees. It is based on iteratively adjusting several decision trees, where each one learns from the mistakes of the previous one.
    - Pros: It is perhaps the best model for tabular data (as in our case). It has proven to be superior to other more complex neural network models. It is easy to fit.
    - Cons: It is complex and can be difficult to put into production.

5. Support vector machine
    - This model, like logistic regression, is linear in the sense that predictions are based only on a linear combination of predictors.
    - Pros: It is a simple model, since it has a linear decision boundary.
    - Cons: It is not a probabilistic model. It only gives binary predictions.

- For this problem we will choose the Logistic Regression model because it is one of the simplest, most studied and easy to implement models. Moreover, being a probabilistic model, it can not only provide binary predictions, but also the likelihood of falling into a certain class.

Model hypotheses

For the development of this Logistic Regression model we assume several things:

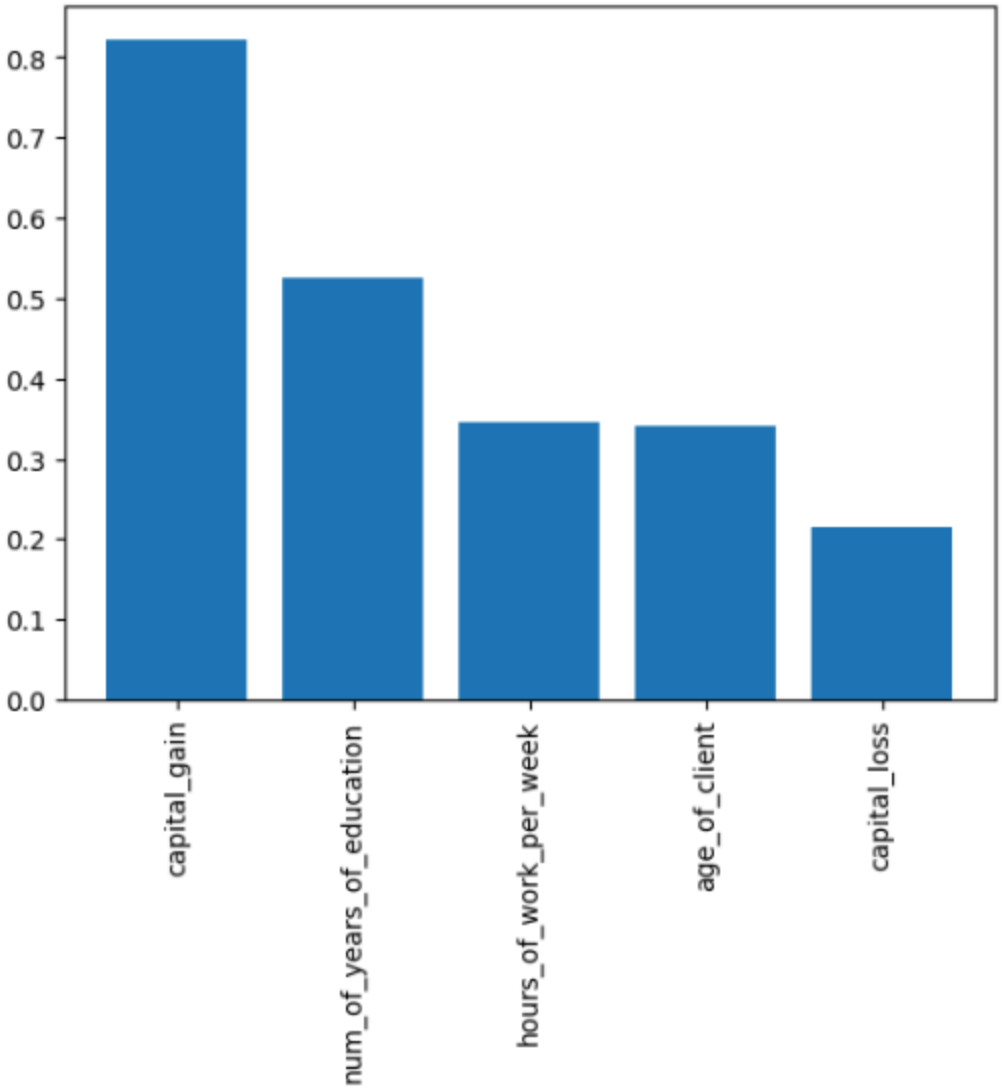
- There are no more outliers in the training base. Being a linear model, it is sensitive to these.
- The categories observed in the training base are all the possible ones.
- In the bases where this model is used to make predictions, there are no null values (except for the categorical classes, where nulls are reported with “?”).
- Positive and negative clients are almost linearly separable (in particular, they do not form complex structures).

Model results

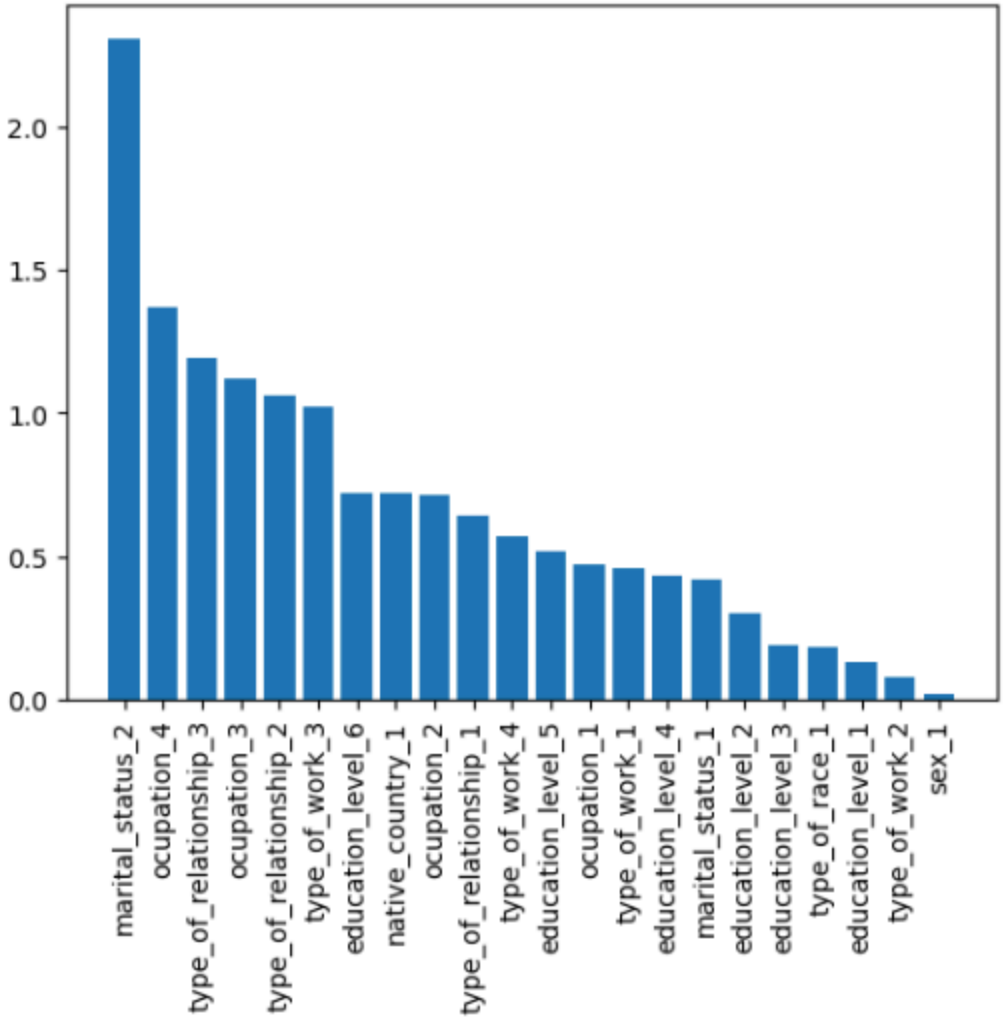
- We fit several logistic regression models with L1 penalty, with C = 1.6237 being the best regularization hyperparameter. It was obtained with cross-validation.
- This value is the one that maximizes the Gini index among all the options studied.
- One of the most important variables is one related to marital status, followed by one related to the customer's occupation.

feature	coefficient	data_type
marital_status_2	2.306760	categorical
ocupation_4	1.369528	categorical
type_of_relationship_3	1.194124	categorical
ocupation_3	1.120242	categorical
type_of_relationship_2	1.059258	categorical
type_of_work_3	1.021000	categorical
capital_gain	0.822077	numerical
education_level_6	0.723864	categorical
native_country_1	0.721989	categorical
ocupation_2	0.713112	categorical
type_of_relationship_1	0.639748	categorical
type_of_work_4	0.569881	categorical
num_of_years_of_education	0.525212	numerical
education_level_5	0.518614	categorical
ocupation_1	0.473275	categorical
type_of_work_1	0.456266	categorical

Next we can see the weights of the numerical variables.

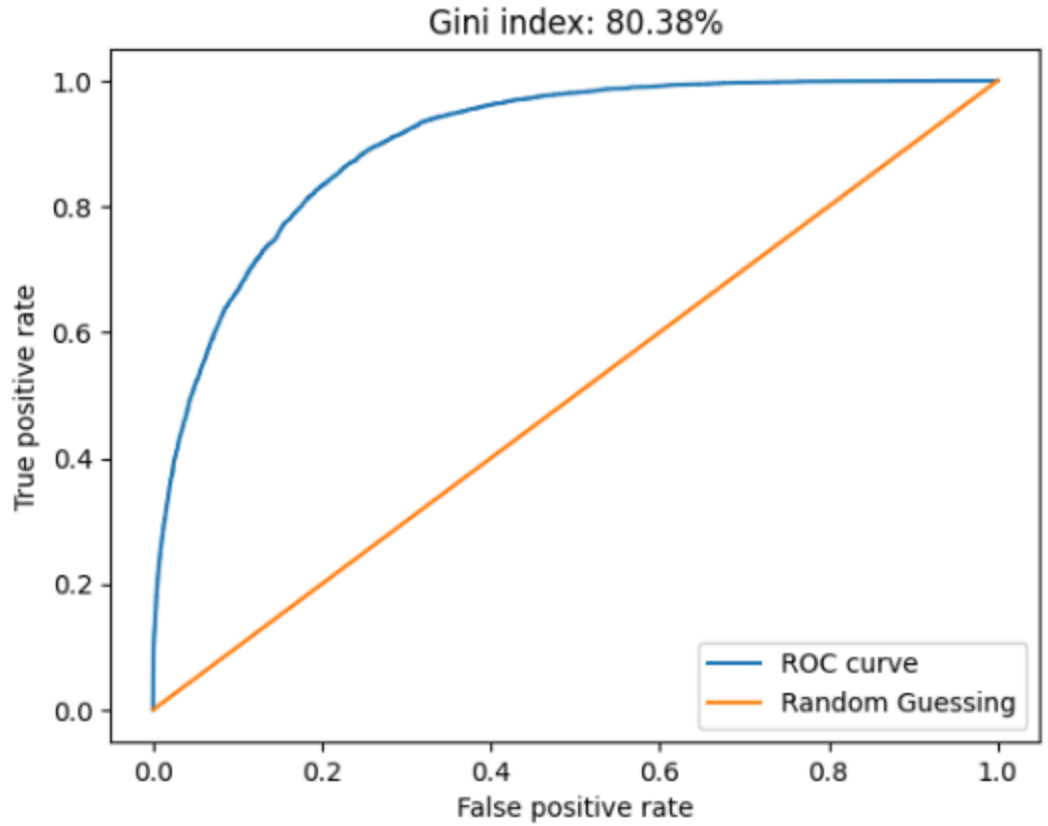


The weights of the categorical dummy variables are shown below.

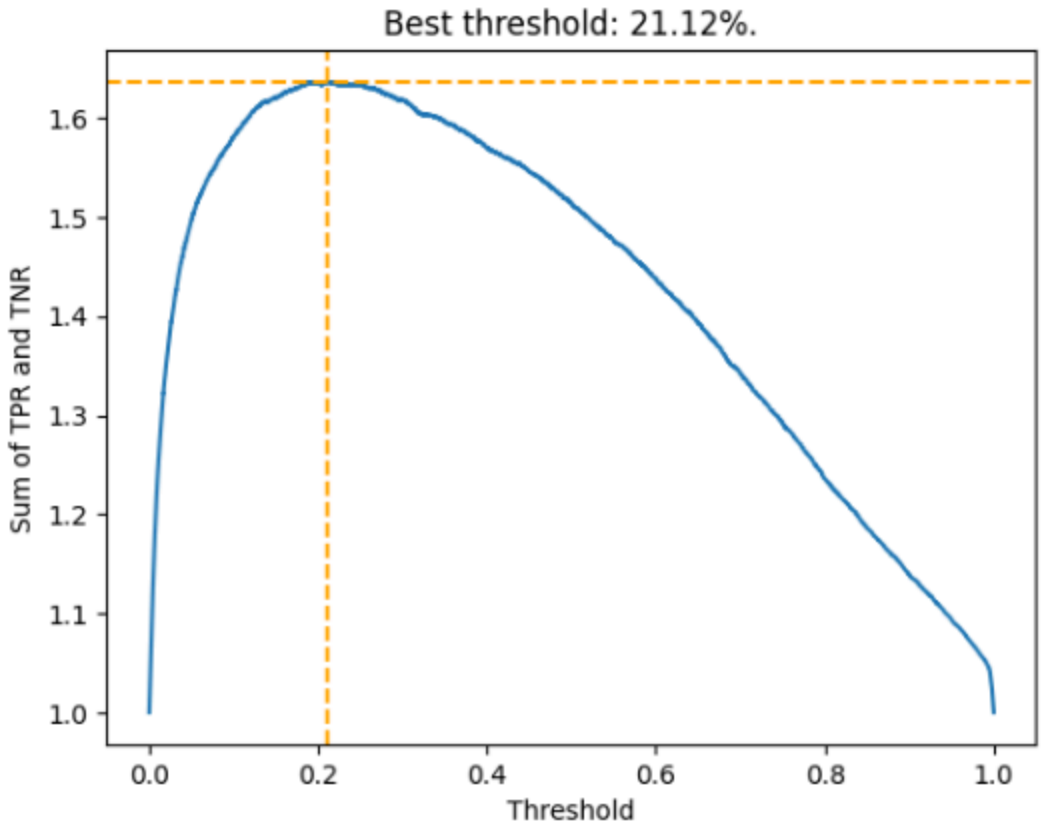


Results in the training set

We obtain a Gini index of 80%, which is excellent (in the industry, we usually only obtain around 60%).



A commonly taken cutoff point is the one that maximizes the sum of the true positive and true negative rates.

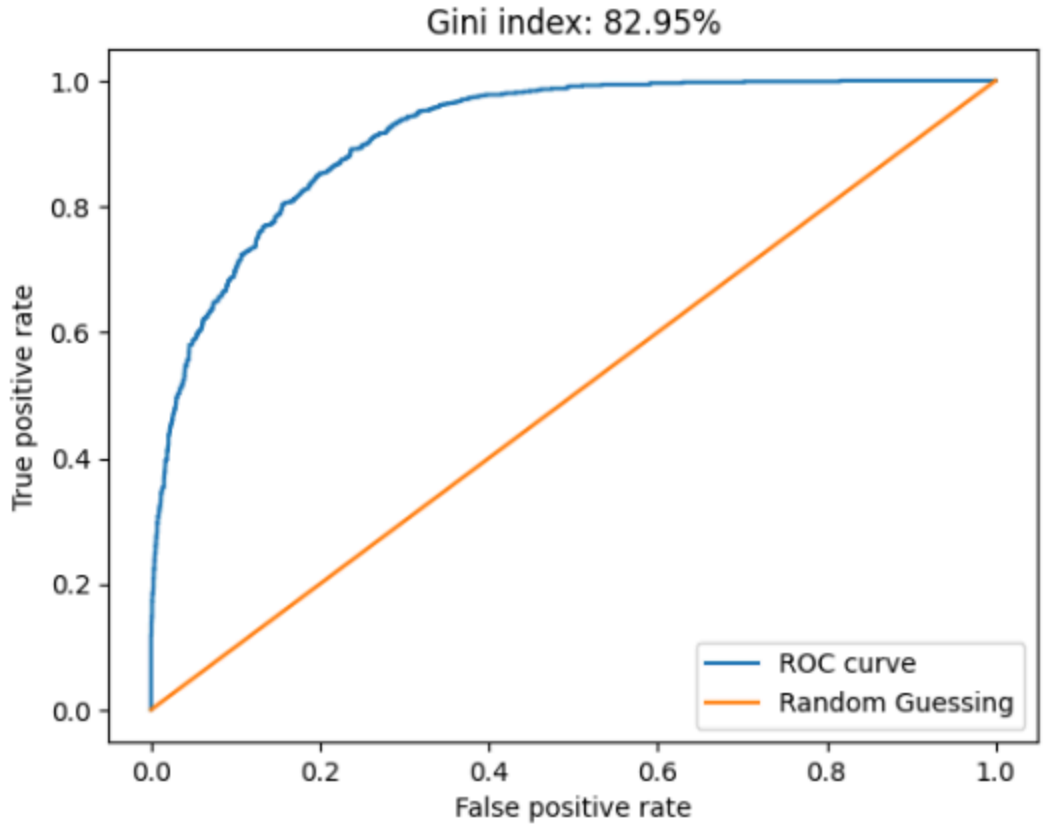


We can observe good adjustments because we have a recall of 87%.

	precision	recall	f1-score	support
0	0.95	0.77	0.85	21847
1	0.54	0.87	0.66	6731
accuracy			0.79	28578
macro avg	0.74	0.82	0.76	28578
weighted avg	0.85	0.79	0.81	28578

Results in the test set

We obtained a Gini index of about 83%, quite close to that obtained in the training set.



In the following table we can see the metrics obtained in the test set. We can observe a recall of 87%, which coincides with the training set, suggesting that the model has no overfitting.

	precision	recall	f1-score	support
0	0.95	0.78	0.85	2456
1	0.56	0.87	0.68	801
accuracy			0.80	3257
macro avg	0.75	0.82	0.77	3257
weighted avg	0.85	0.80	0.81	3257