

Hackthon 6

Hugo Castilho

2018-01-27

Abstract

During a 10 day period, a deployed estimator received new data and outcomes.

The received data had significant differences from the data used in training, but the models was still able to produce a reasonable estimation of the outcomes. The AUC ROC score for the new outcomes was 0.8752.

Using the new data and outcomes to retrain the data does not result in very impressive improvement, but should progressively increase as more outcomes arrive.

1 Intro

In a previous report we analysed a dataset with characteristics of an unknown population and if they became unemployed in the next 12 months. We will assume knowledge of this report, if you are not familiar please read it.

Several predicted models were tested and one was selected. This model was deployed online to receive further data during a 10 day period. In this time we received both more population samples and outcomes.

In this report we analyse the behaviour of our deployed model in light of the new information.

2 Exploratory data analysis

During this 10 day period we received 9943 new samples and 498 new outcomes. In this section we will have a look at the new data and check for any anomalies relative to the original data used to train the model.

Having a look at the data (appendix A on page 3) it's easy to spot that there are significant differences from the data used to train our model. We now have samples with *earned dividends* and different genders. The proportions of the categories for *domestic relationship type* and *domestic status* do not match. There are also some changes in *job types* and *profession*. This is a considerable amount of change for our model to deal with.

3 Model Analysis

Our model achieved an AUC ROC score of:

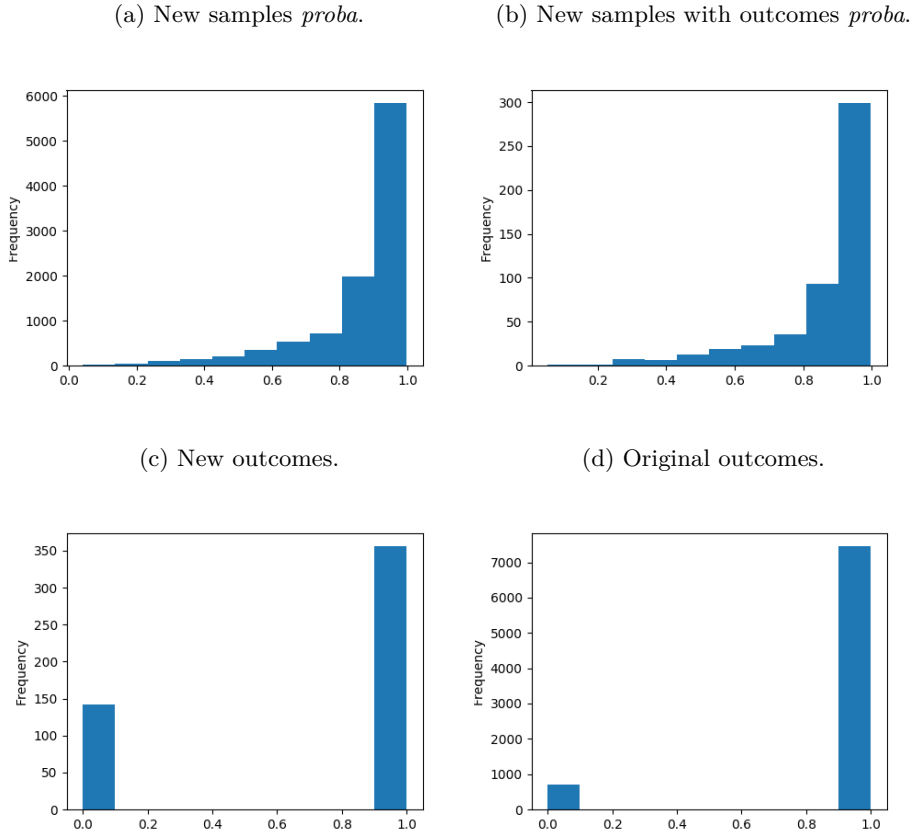
0.8752

Taking in consideration that the new population profile does not match the one for which our model was trained for it is still a high score.

We can also have a look at the output of the model for the samples independently of our knowledge of the outcome. Ideally our models should give a lot more outcomes near 0 and 1 than in the middle. So let's look at the histogram of our *proba* output in fig. 1a.

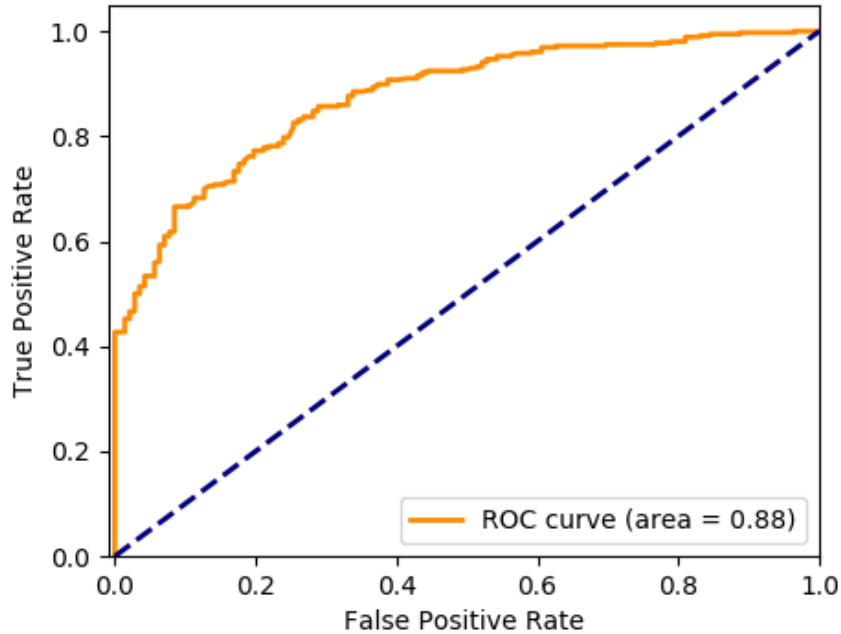
While most of our outputs are near the edge, we see that this is only happening towards 1. This could mean that our new samples are a skewed view of the population or that the population itself is skewed toward 1 (becoming unemployed during the 12 month period after the sample is collected). So let's compare with the histogram for the samples for which we do know the outcome (fig. 1b) and the true outcomes (fig. 1c).

Figure 1: Outcome and *proba* histograms.



We can see that we are not in an ideal situation, our population outcomes are skewed, we have to be extra careful. But the AUC ROC score and the RUC curve (fig. 2 on the following page) tells us that we should have a good predictive capability by carefully selecting a threshold.

Figure 2: ROC curve.



4 Retraining

We joined the new data (with outcomes) with the old and split into training and test sets. Afterward we trained the same model (`GradientBoostingClassifier`) with the new data to compare the score with our deployed estimator, see table 1.

Table 1: Country of origin top 10.

(a) Original dataset.

Deployed estimator	0.9182
New estimator	0.9232

As we can see there is not a big difference, but there is so little new data that this just means that no new "insights" came from the new data.

A Tables & Figures

Table 2: Country of origin top 10.

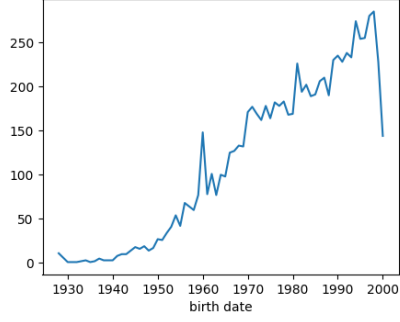
(a) Original dataset.		(b) New dataset.	
u.s.	7330	u.s.	8935
unknown	126	mexico	210
mexico	111	unknown	185
philippines	60	philippines	53
de	50	de	39
puerto rico	39	india	37
jamaica	34	canada	32
cuba	34	gb	32
el-salvador	30	puerto rico	30
canada	28	el-salvador	29
(c) New dataset samples with target.			
u.s.	452		
mexico	9		
unknown	8		
de	4		
india	3		
hong	2		
puerto rico	2		
poland	2		
china	2		
ireland	2		

Table 3: Domestic relationship type value counts.

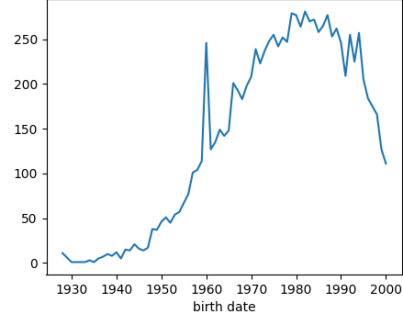
(a) Original dataset.		(b) New dataset.	
not living with family	2919	has wife	5398
never married	2063	not living with family	2227
living with child	1750	living with child	1353
has husband	1106	never married	532
living with extende family	325	living with extende family	245
has wife	1	has husband	188
(c) New dataset samples with target.			
has wife	266		
not living with family	119		
living with child	66		
never married	23		
living with extende family	14		
has husband	10		

Figure 3: Birth date histograms.

(a) Original samples.



(b) New samples.



(c) New samples with outcomes.

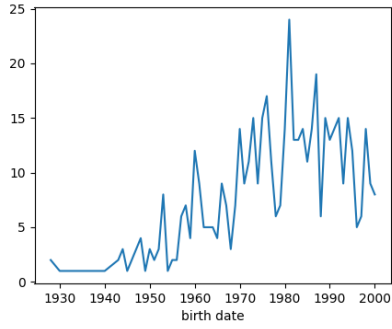


Table 4: Domestic status value counts.

(a) Original dataset.

single	3662
d	2073
married 2	1170
spouse passed	599
divorce pending	486
married not together	163
married 1	11

(b) New dataset.

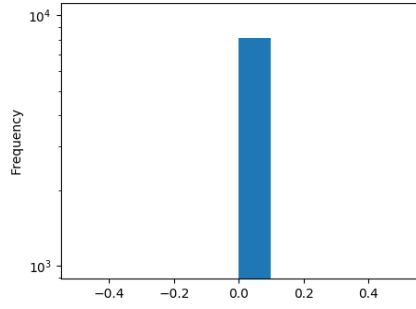
married 2	5645
single	2853
d	983
divorce pending	211
spouse passed	154
married not together	93
married 1	4

(c) New dataset samples with target.

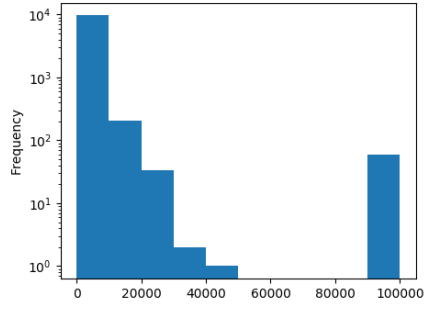
married 2	281
single	145
d	52
spouse passed	8
married not together	6
divorce pending	6

Figure 4: Earned dividends histograms.

(a) Original samples.



(b) New samples.



(c) New samples with outcomes.

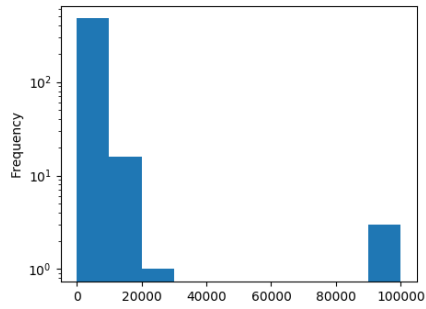


Table 5: Ethnicity value counts.

(a) Original dataset.

white and privileged	6523
afro american	1210
asian	262
american indian	88
other	81

(b) New dataset.

white and privileged	8679
afro american	778
asian	315
american indian	92
other	79

(c) New dataset samples with target.

white and privileged	431
afro american	46
asian	17
american indian	2
other	2

Table 6: Gender value counts.

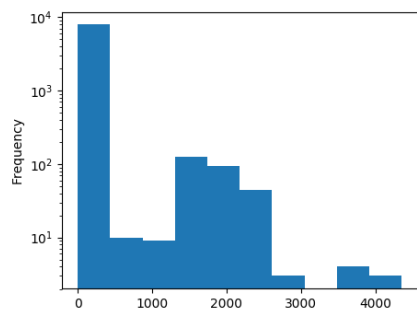
(a) Original dataset.		(b) New dataset.	
Female	8164	Male	8904
		Female	1039
(c) New dataset samples with target.			
Male	453		
Female	45		

Table 7: Job type value counts.

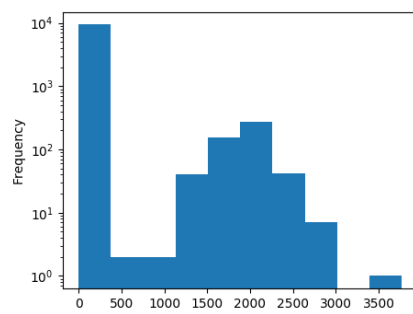
(a) Original dataset.		(b) New dataset.	
private	5919	private	6824
unknown	620	self-emp-not-inc	905
local-gov	618	local-gov	573
state-gov	368	unknown	489
self-emp-not-inc	303	self-emp-inc	426
federal-gov	236	state-gov	415
self-emp-inc	94	federal-gov	304
without-pay	4	without-pay	4
never-worked	2	never-worked	3
(c) New dataset samples with target.			
private	341		
self-emp-not-inc	44		
unknown	28		
local-gov	28		
self-emp-inc	24		
state-gov	20		
federal-gov	13		

Figure 5: Interest earned histograms.

(a) Original samples.



(b) New samples.



(c) New samples with outcomes.

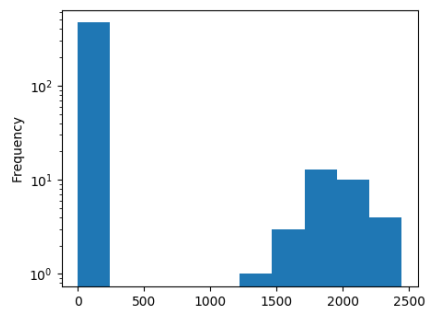
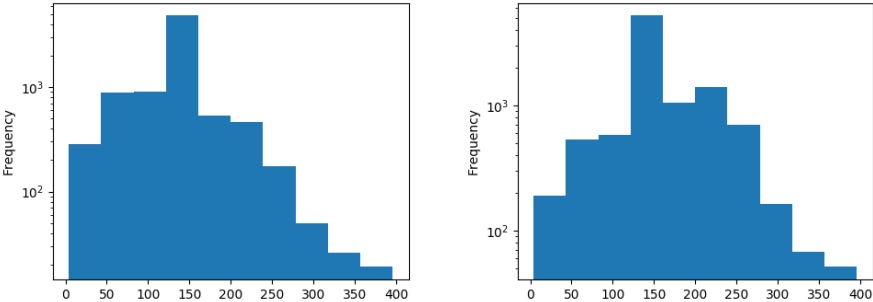


Figure 6: Monthly work histograms.

(a) Original samples. (b) New samples.



(c) New samples with outcomes.

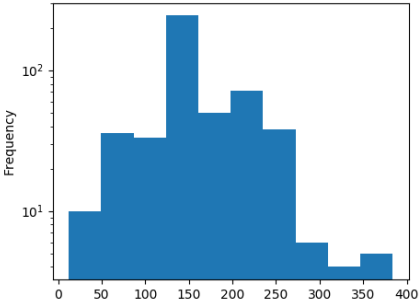


Table 8: Profession value counts.

(a) Original dataset.		(b) New dataset.	
secretarial	1949	vocational	1561
other	1423	C-level	1316
specialist technician	1096	specialist technician	1213
sales	978	sales	1112
C-level	842	other	780
unknown	622	secretarial	736
mechanic	420	trucking	666
technology support	247	mechanic	637
vocational	184	household labor	501
household labor	131	unknown	492
estate employee	108	agriculture	388
defense contractor	58	technology support	279
trucking	58	defense contractor	244
agriculture	48	estate employee	16
		army	2
(c) New dataset samples with target.			
C-level	68		
vocational	68		
sales	64		
specialist technician	49		
other	42		
trucking	38		
household labor	33		
mechanic	30		
secretarial	29		
unknown	28		
defense contractor	18		
agriculture	17		
technology support	13		
army	1		

Table 9: School level value counts.

(a) Original dataset.		(b) New dataset.	
secondary	2594	secondary	3243
entry level college	2165	entry level college	2087
college graduate	1188	college graduate	1691
basic vocational	373	some post graduate	577
some post graduate	355	basic vocational	402
secondary 11	341	secondary 11	317
advanced vocational	326	advanced vocational	303
10th	248	10th	277
secondary-7 through 8	123	secondary-7 through 8	208
secondary 12	106	primary school	202
secondary-9	104	secondary-9	172
secondary-5 through 6	72	advanced post graduate	142
advanced post graduate	61	secondary 12	140
primary school	58	secondary-5 through 6	117
primary 1 through 4	37	primary 1 through 4	49
kindergarten	13	kindergarten	16
(c) New dataset samples with target.			
secondary	179		
entry level college	110		
college graduate	76		
some post graduate	26		
secondary 11	20		
primary school	13		
10th	12		
advanced vocational	11		
secondary 12	10		
secondary-9	9		
basic vocational	9		
advanced post graduate	8		
secondary-5 through 6	5		
secondary-7 through 8	5		
kindergarten	3		
primary 1 through 4	2		