Hackthon 6

Hugo Castilho

2018-01-07

Abstract

The objective is to to predict the employment outcome of an individual in the next 12 months. To this effect number of machine learning models and techniques were applied.

The model with the best predictive capabilities for this problem was a gradient boost classifier with an AUC ROC score of 0.9369 with the validation set.

1 Exploratory Data Analysis

1.1 Data description

The provided dataset contains personal information of several people (8164 samples), such as date of birth and current employment and maps it to the observation if the person became unemployed in the next 12 months.

The dataset was made available without any information about it's contents except for the property names. Each one was analysed for insights and anomalies.

The following list details what could be obtained exploring the data (column name, value types and values).

id Integer, identifier for each subject, all values distinct no problems detected.

target Integer, did the subject become unemployed in the following 12 months.

birth date Date (YYYY-MM-DD), date of birth of the subject. The youngest being 2016-02-10 and the oldest 1928-01-09. Frequency plot (see fig. 3 on page 10) shows nothing surprising.

country of origin Categorical, country names come in a variety of inconsistent formats.

domestic relationship type Categorical, (see table 6 on page 8) who the subject lives with. Categories are unclear and ill defined. Moreover it is inconsistent with the *domestic status*, there are a considerable number of entries classified as *domestic relation type–never married* and *domestic status–d* (presumably divorced) (see table 5 on page 8).

domestic status Categorical, marital status or if has married several times (see table 7 on page 8). By elimination category *d* is supposedly divorced.

earned dividends Numerical, monetary amount (currency not specified). Return from distribution of corporate earnings, it's 0 for all samples.

ethnicity Categorical, categories have funny names (see table 8 on page 9).

gender Categorical, all female dataset (see table 9 on page 9).

job type Categorical, current job type of the subject (see table 10 on page 9), government, self employed, item etc...

interest earned Numerical, monetary amount (currency not specified). Returns from loaning money (see fig. 4 on page 10).

monthly work Numerical, number of hours of work per month (see fig. 5 on page 10)

profession Categorical, type of profession (see table 11 on page 9).

school level Categorical, subject level of schooling (see table 12 on page 9).

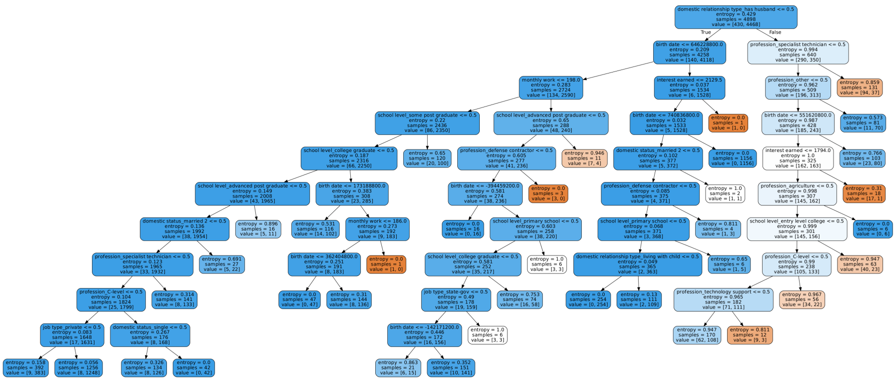
1.2 Data exploration

To get some further insights from the data the correlation matrix was inspected. It was split in multiple plots to ease this process fig. 6 on page 10. Some preprocessing was necessary to produce the correlation matrix see section 2 on the following page for a description.

Some insights were expected, for example, the categorical dummy variables have a strong negative correlation between themselves. These dummy variables are not independent which will present a challenge to linear models (consider applying PCA if using linear models). There is a strong (negative or positive) correlation between the *target* and the *domestic status–married 2*, *domestic relationship type–has husband*, which could be interesting to explore. Also as expected the *birth date* is connected with the *domestic status* classes, the connection with *school level* is not evident. Finally there are severall connections between specific *countries of origin* and *school level* and *ethnicity*.

To help understand the dataset we trained a small decision tree see fig. 1 on the next page. Although the decision tree model does not give good results (see section 4 on page 4) it is still interesting to observe what features split the data. In fact, this model with a `max_depth` of 4 we outperformed the same model using the standard parameters obtaining a 0.8927 AUC ROC score. But this is the maximum obtained using simple decision trees.

Figure 1: Date of birth frequency.



2 Data pre-processing

This section details any and all data pre-processing before modelling. The first section section 2.1 explains what was done to convert the dataset to usable, unambiguous types. Afterwards section section 2.2 details the what was done to select and improve the features feed to the model. Finally section section 2.3 breaks down the train/test split.

2.1 Data cleaning

On data import the *birth dates* were converted to naive dates as no timezone information was provided. It's doubtful the timezone would provide any useful information.

The *countries of origin* were converted to the corresponding *ISO 3166–1 alpha-2* representation. Some inputs required special rules. Especially ambiguous was *dr* which represents no country code this was converted to Dominican Republic (DO) even if the race for these inputs suggests it's not (mostly white). Also in the list is *schotland* which is not part of the *ISO* country list.

All categorical data columns was kept as is, there was not enough information to reach any conclusion.

2.2 Feature engineering

birth dates were converted to timestamps. In the dataset earned dividends and gender do not change, these properties were dropped since they convey no useful information. If new samples include this value this decision will be reconsidered. All categorical data was turned into dummy class variables.

2.3 Train/test splitting

The train/test split was done holding out .4 of the data for final validation. Furthermore training was done using a shuffle split with .3 for testing.

3 Modelling

4 Initial tests

To get a feelling of the baseline performance of the models available in the *scikit* package severall were tried with the default parameters (except were not possible), see table 1.

Table 1

Model	AUC ROC Score
GradientBoostingClassifier	0.9212
AdaBoostClassifier	0.9008
BaggingClassifier	0.8428
VotingClassifier	0.8233
RandomForestClassifier	0.8160
ExtraTreesClassifier	0.8154
QuadraticDiscriminantAnalysis	0.8053
KNeighborsClassifier	0.7267
GaussianProcessClassifier	0.7022
DecisionTreeClassifier	0.6729
SVC	0.6698
SGDClassifier	0.6377
GaussianNB	0.6718

With the default parameters there is, as expected, a clear dominance of ensemble models. The top 5 were selected for further parameter tuning.

5 Model tuning & selection

Each of the models was optimized by randomly searching a small part of the parameter space. The portion to explore was determined empirically by careful study of each of the parameter. The results are detailed in table 2.

Table 2

Model	AUC ROC Score
GradientBoostingClassifier	0.9369
AdaBoostClassifier	0.9329
RandomForestClassifier	0.9308
VotingClassifier	0.9299
BaggingClassifier	0.9143

After tuning all models were able to achieve AUR ROC scores in the .9 range, but *GradientBoostingClassifier* outperformed the others. As there are no other constraints model selection is based solely on the score.

6 Feature Elimination

Running recursive feature elimination on the model identified in the previous section the optimal number of features was determined to be 31 and are the following:

- birth date
- interest earned
- monthly work
- job type–federal-gov
- job type–self-emp-not-inc
- school level–10th
- school level–advanced post graduate
- school level–college graduate
- school level–primary school
- school level–secondary
- school level–some post graduate
- domestic status–married 1
- domestic status–married 2
- domestic status–spouse passed
- profession–C-level
- profession–defense contractor
- profession–mechanic
- profession–other
- profession–secretarial
- profession–specialist technician
- profession–trucking
- profession–vocational
- domestic relationship type–has husband
- domestic relationship type–not living with family
- ethnicity–afro american
- country of origin–GR
- country of origin–HU

- country of origin–IE
- country of origin–JP
- country of origin–PH
- country of origin–US

Figure 2: Score change.

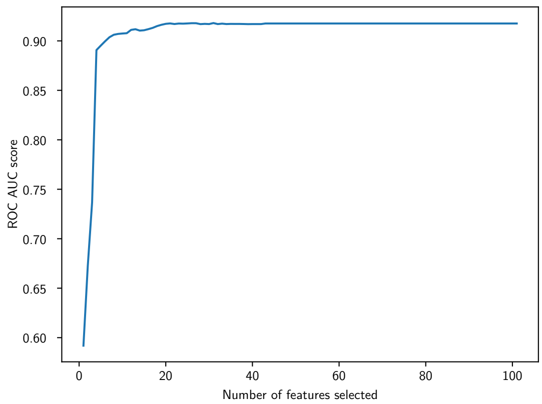


Figure 2 displays how the AUC ROC score changes with the number of features. Since there are no computational limitations and the prediction does not deteriorate with the number of features this exercise is a mere curiosity.

Is is interesting however to compare the selected features with their correlation (see table 3 on the next page) with the target, remember that a positive correlation means it more likely to end up unemployed in 12 months. The selected features that correlate positively with being unemployed soon are presented in the following list. Note that this list says nothing of the decision surface of our model, having any combination of these features does not mean that you are more likely to become unemployed.

- birth date
- school level–10th
- school level–secondary
- domestic status–spouse passed
- profession–mechanic
- profession–other
- profession–secretarial
- profession–trucking
- profession–vocational

- domestic relationship type–not living with family
- ethnicity–afro american
- country of origin–US

Table 3: Selected features and respective correlation.

birth date	0.114102
interest earned	-0.148656
monthly work	-0.110180
job type–federal-gov	-0.042745
job type–self-emp-not-inc	-0.065876
school level–10th	0.047029
school level–advanced post graduate	-0.139795
school level–college graduate	-0.129046
school level–primary school	-0.077406
school level–secondary	0.083645
school level–some post graduate	-0.145206
domestic status–married 1	-0.047914
domestic status–married 2	-0.469246
domestic status–spouse passed	0.050169
profession–C-level	-0.131167
profession–defense contractor	-0.010123
profession–mechanic	0.046294
profession–other	0.105133
profession–secretarial	0.036197
profession–specialist technician	-0.171728
profession–trucking	0.005404
profession–vocational	0.011720
domestic relationship type–has husband	-0.481157
domestic relationship type–not living with family	0.094195
ethnicity–afro american	0.068792
country of origin–GR	-0.016763
country of origin–HU	-0.016763
country of origin–IE	-0.009930
country of origin–JP	-0.027360
country of origin–PH	-0.024337
country of origin–US	0.004980

7 Evaluation

Applying the selected model (a gradient boost classifier) to the validation data we obtain a final score of:

$$0.9369$$

We can therefore with a high degree of certainty predict the employment outcome in the next 12 months.

A EDA Tables

Table 4: Country of origin value counts (trimmed).

u.s.	7330
unknown	126
mexico	111
philippines	60
de	50
puerto rico	39
jamaica	34
cuba	34
el-salvador	30
canada	28
dr	27
gb	22
south	20
italy	18
columbia	17
haiti	17
china	17
vietnam	17
guatemala	16
japan	15
poland	14
peru	11
taiwan	11
thailand	11

Table 5: Domestic relationship type grouped by domestic status counts.

domestic status	domestic relationship type	count
d	living with child	116
	living with extende family	47
	never married	1006
	not living with family	904
divorce pending	living with child	40
	living with extende family	25
	never married	293
	not living with family	128
married 1	has husband	10
	living with child	1
married 2	has husband	1096
	has wife	1
	living with child	28
	living with extende family	42
	not living with family	3
married not together	living with child	25
	living with extende family	7
	never married	73
	not living with family	58
single	living with child	1530
	living with extende family	174
	never married	449
	not living with family	1509
spouse passed	living with child	10
	living with extende family	30
	never married	242
	not living with family	317

Table 6: Domestic relationship type value counts.

not living with family	2919
never married	2063
living with child	1750
has husband	1106
living with extende family	325
has wife	1

Table 7: Domestic status value counts.

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

Table 8: Ethnicity value counts.

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

Table 9: Gender value counts.

Female	8164

Table 10: Job type value counts.

private	5919
unknown	620
local-gov	618
state-gov	368
self-emp-not-inc	303
federal-gov	236
self-emp-inc	94
without-pay	4
never-worked	2

Table 11: Profession value counts.

secretarial	1949
other	1423
specialist technician	1096
sales	978
C-level	842
unknown	622
mechanic	420
technology support	247
vocational	184
household labor	131
estate employee	108
defense contractor	58
trucking	58
agriculture	48

Table 12: School level value counts.

secondary	2594
entry level college	2165
college graduate	1188
basic vocational	373
some post graduate	355
secondary 11	341
advanced vocational	326
10th	248
secondary-7 through 8	123
secondary 12	106
secondary-9	104
secondary-5 through 6	72
advanced post graduate	61
primary school	58
primary 1 through 4	37
kindergarten	13

B EDA Figures

Figure 3: Date of birth frequency.

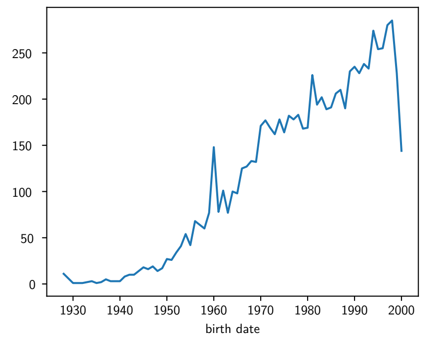


Figure 4: Interest earned frequency (logarithmic scale).

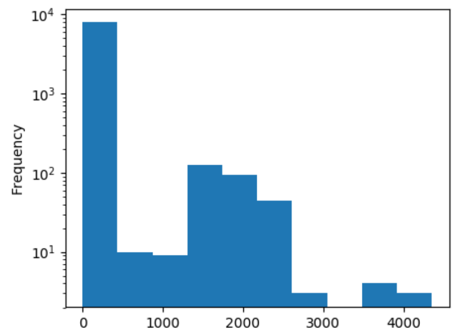


Figure 5: Monthly work frequency.

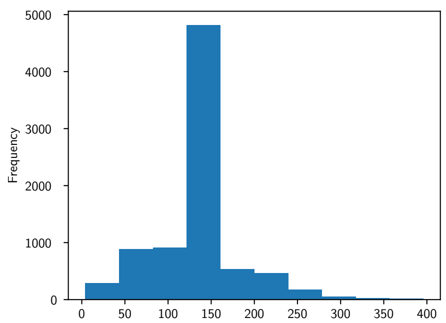


Figure 6: Sections of the cross-correlation matrix.