Prevendo o país de destino de um novo usuário

André de Farias Pereira

airbnb

Recruitment Prediction Competition

Airbnb New User Bookings

Where will a new guest book their first travel experience?

Airbnb · 1,458 teams · 6 years ago

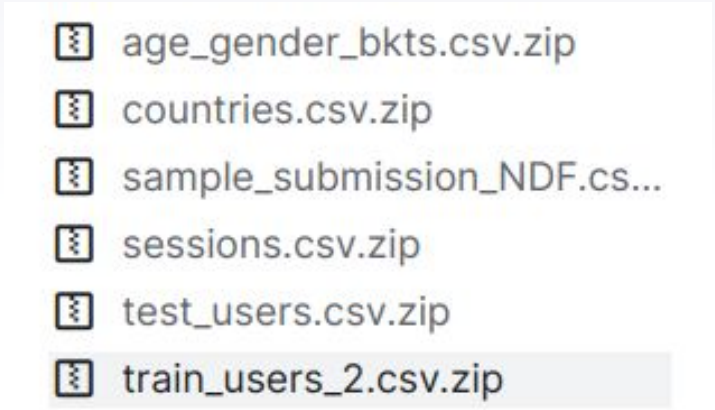
Overview

Instead of waking to overlooked "Do not disturb" signs, Airbnb travelers find themselves rising with the birds in a whimsical treehouse, having their morning coffee on the deck of a houseboat, or cooking a shared regional breakfast with their hosts.

New users on Airbnb can book a place to stay in 34,000+ cities across 190+ countries. By accurately predicting where a new user will book their first travel experience, Airbnb can share more personalized content with their community, decrease the average time to first booking, and better forecast demand.

In this recruiting competition, Airbnb challenges you to predict in which country a new user will make his or her first booking. Kagglers who impress with their answer (and an explanation of how they got there) will be considered for an interview for the opportunity to join Airbnb's Data Science and Analytics team.

Datasets


age_gender_bkts.csv.zip
countries.csv.zip
sample_submission_NDF.cs...
sessions.csv.zip
test_users.csv.zip
train_users_2.csv.zip

Tratamento de dados

train_users_2

	id	date_account_created	timestamp_first_active	date_first_booking	gender	age	signup_method	signup_flow	language	affiliate_channel	affiliate_pr
0	gxn3p5htnn	2010-06-28	20090319043255	NaN	-unknown-	NaN	facebook	0	en	direct	
1	820tgsjxq7	2011-05-25	20090523174809	NaN	MALE	38.0	facebook	0	en	seo	
2	4ft3gnwmtx	2010-09-28	20090609231247	2010-08-02	FEMALE	56.0	basic	3	en	direct	
3	bjjt8pjhuk	2011-12-05	20091031060129	2012-09-08	FEMALE	42.0	facebook	0	en	direct	
4	87mebub9p4	2010-09-14	20091208061105	2010-02-18	-unknown-	41.0	basic	0	en	direct	
...	...	...	...	...	...	...	...	...	...	...	
213446	zxodksqpep	2014-06-30	20140630235636	NaN	MALE	32.0	basic	0	en	sem-brand	
213447	mhewnxesx9	2014-06-30	20140630235719	NaN	-unknown-	NaN	basic	0	en	direct	
213448	6o3arsjbb4	2014-06-30	20140630235754	NaN	-unknown-	32.0	basic	0	en	direct	
213449	jh95kwisub	2014-06-30	20140630235822	NaN	-unknown-	NaN	basic	25	en	other	
213450	nw9fwlyb5f	2014-06-30	20140630235824	NaN	-unknown-	NaN	basic	25	en	direct	

213451 rows × 16 columns

train_users_2

- Unique values

```
id (213451):  ['gxn3p5htnn' '820tgsjxq7' '4ft3gnwmtx'
 'nw9fwlyb5f']


date_account_created (1634):  ['2010-06-28' '2011-05-
 '2014-06-30']


timestamp_first_active (213451):  [20090319043255 200
 20140630235822 20140630235824]


date_first_booking (1977):  [nan '2010-08-02' '2012-0

gender (4):  ['-unknown-' 'MALE' 'FEMALE' 'OTHER']

-unknown-: 95688
MALE: 54440
FEMALE: 63041
OTHER: 282

age (128):  [      nan 3.800e+01 5.600e+01 4.200e+01
 5.000e+01 3.600e+01 3.700e+01 3.300e+01 3.100e+01 2.
 4.000e+01 2.600e+01 3.200e+01 3.500e+01 5.900e+01 4.
 3.400e+01 2.800e+01 1.900e+01 5.300e+01 5.200e+01 3.
```

```
first_browser (52):  ['Chrome'
 'Chrome Mobile' 'RockMelt' 'Ch
 'Palm Pre web browser' 'Mobile
 'Apple Mail' 'Silk' 'Camino'
 'Iron' 'Sogou Explorer' 'IceWe
 'Kindle Browser' 'CoolNovo' '(
 'Crazy Browser' 'Mozilla' 'Omn
 'CometBird' 'Comodo Dragon' 'F
 'Opera Mobile' 'Yandex.Browser
 'Stainless' 'Googlebot' 'Outlo

Chrome: 63845
IE: 21068
Firefox: 33655
Safari: 45169
-unknown-: 27266
Mobile Safari: 19274
Chrome Mobile: 1270
RockMelt: 24
Chromium: 73
Android Browser: 851
AOL Explorer: 245
Palm Pre web browser: 1
Mobile Firefox: 30
```

```
language (25):
 'hu' 'da' 'id'

en: 206314
fr: 1172
de: 732
es: 915
it: 514
pt: 240
zh: 1632
ko: 747
ja: 225
ru: 389
pl: 54
el: 24
sv: 122
nl: 97
hu: 18
da: 58
id: 22
fi: 14
no: 30
tr: 64
th: 24
cs: 32
hr: 2
ca: 5
is: 5
```

```
1928.0: 2
1.0: 2
1936.0: 2
1933.0: 1
1935.0: 1
1925.0: 1
1952.0: 1
150.0: 1
1927.0: 1
132.0: 1
1953.0: 1
1942.0: 1
1995.0: 1
2008.0: 1
1924.0: 2
1929.0: 2
1947.0: 2
1938.0: 1
1926.0: 1
```

train_users_2

- NAN

```
TOTAL: 213451 amostras
id                             0
date_account_created           0
timestamp_first_active         0
date_first_booking        124543
gender                         0
age                       87990
signup_method                  0
signup_flow                    0
language                       0
affiliate_channel              0
affiliate_provider             0
first_affiliate_tracked     6065
signup_app                     0
first_device_type              0
first_browser                  0
country_destination            0
dtype: int64
```

train_users_2

- timestamp e date_account_created

```
date_account_created (1634):  ['2010-06-28' '2011-05-25' '2010-09-28' ... '2014-06-27' '2014-06-29'
 '2014-06-30']


timestamp_first_active (213451):  [20090319043255 20090523174809 20090609231247 ... 20140630235754
 20140630235822 20140630235824]
```

train_users_2

- Estratégia utilizada

1) Número de amostras únicas
2) Verificação sobre a generalidade.
3) Verificação de dados errados
4) Tratamento de NAN

sessions

	user_id	action	action_type	action_detail	device_type	secs_elapsed
0	d1mm9tcy42	lookup	NaN	NaN	Windows Desktop	319.0
1	d1mm9tcy42	search_results	click	view_search_results	Windows Desktop	67753.0
2	d1mm9tcy42	lookup	NaN	NaN	Windows Desktop	301.0
3	d1mm9tcy42	search_results	click	view_search_results	Windows Desktop	22141.0
4	d1mm9tcy42	lookup	NaN	NaN	Windows Desktop	435.0
...	...	...	...	...	...	...
10567732	9uqfg8txu3	dashboard	view	dashboard	Windows Desktop	556.0
10567733	9uqfg8txu3	edit	view	edit_profile	Windows Desktop	6624.0
10567734	9uqfg8txu3	webcam_upload	-unknown-	-unknown-	Windows Desktop	200125.0
10567735	9uqfg8txu3	active	-unknown-	-unknown-	-unknown-	17624.0
10567736	9uqfg8txu3	show_personalize	data	user_profile_content_update	Windows Desktop	1221.0

10567737 rows × 6 columns

sessions

- group_by

	action	action_type	action_detail	device_type	secs_elapsed	id
0	[index, dashboard, header_userpic, dashboard, ...	[view, view, data, view, partner_callback, mes...	[view_search_results, dashboard, header_userpi...	[Mac Desktop, Mac Desktop, Mac Desktop, Mac De...	[20438.0, 787.0, 850.0, 934.0, nan, 129817.0, ...	00023iyk9l
1	[search_results, show, personalize, show, sear...	[click, view, data, nan, click, click, nan, da...	[view_search_results, p3, wishlist_content_upd...	[Mac Desktop, Mac Desktop, Mac Desktop, Mac De...	[1708.0, 21260.0, 1223.0, 26.0, 847.0, 1230.0,...	0010k6l0om
2	[search, search, search, show, social_connecti...	[click, click, click, view, data, -unknown-, v...	[view_search_results, view_search_results, vie...	[Android App Unknown Phone/Tablet, Android App...	[622.0, 1813.0, 1507.0, 6327.0, 927.0, 142.0, ...	001wyh0pz8
3	[show, reviews, show, search, show, search, re...	[view, data, view, click, view, click, data, s...	[user_profile, listing_reviews, p3, view_searc...	[-unknown-, -unknown-, -unknown-, -unknown-, -...	[6162.0, 75.0, 86.0, 13710.0, 25217.0, 10989.0...	0028jgx1x1
4	[social_connections, payment_methods, create, ...	[data, -unknown-, -unknown-, view, data, data,...	[user_social_connections, -unknown-, -unknown-...	[iPhone, iPhone, iPhone, iPhone, iPhone, iPhon...	[17135.0, 711.0, 274.0, 179.0, 483.0, 1.0, 782...	002qnbzfs5
...	...	...	...	...	...	...
135478	[identity, kba, kba_update, kba_update, popula...	[-unknown-, -unknown-, -unknown-, -unknown-, -...	[-unknown-, -unknown-, -unknown-, -unknown-, -...	[Windows Desktop, Windows Desktop, Windows Des...	[1338.0, 10115.0, 23802.0, 16951.0, 49938.0, 7...	zzxox7jnrx
135479	[personalize, header_userpic, create, personal...	[data, data, submit, data, view, click, click,...	[wishlist_content_update, header_userpic, crea...	[Windows Desktop, Windows Desktop, Windows Des...	[501.0, 3671.0, nan, 42612.0, 697.0, 25616.0, ...	zzy7t0y9cm
135480	[hosting_social_proof, create, header_userpic]	[-unknown-, submit, data]	[-unknown-, create_user, header_userpic]	[Windows Desktop, Windows Desktop, Windows Des...	[1533.0, nan, 198.0]	zzysuoqg6x
135481	[header_userpic, personalize, ajax_lwlb_contac...	[data, data, click, click, message_post, submi...	[header_userpic, wishlist_content_update, cont...	[Windows Desktop, Windows Desktop, Windows Des...	[3590.0, 72175.0, 280030.0, 1061.0, 75684.0, 3...	zzywmcn0jv
135482	[similar_listings, show, show, show, similar_l...	[data, view, view, view, data, click, -unknown...	[similar_listings, p3, p3, p3, similar_listing...	[Windows Desktop, Windows Desktop, Windows Des...	[138.0, 0.0, 1036.0, 336.0, 346.0, 2163.0, 23....	zzzlylp57e

sessions_merge_train

- merge com train

first_affiliate_tracked	signup_app	first_device_type	first_browser	country_destination	action	action_type	action_detail	device_type	secs_elapsed
untracked	Web	Mac Desktop	Chrome	NDF	NaN	NaN	NaN	NaN	NaN
untracked	Web	Mac Desktop	Chrome	NDF	NaN	NaN	NaN	NaN	NaN
untracked	Web	Windows Desktop	IE	US	NaN	NaN	NaN	NaN	NaN
untracked	Web	Mac Desktop	Firefox	other	NaN	NaN	NaN	NaN	NaN
untracked	Web	Mac Desktop	Chrome	US	NaN	NaN	NaN	NaN	NaN

sessions_merge_train

- merge com train

Alguns Ids não tinham dados do sessions

first_affiliate_tracked	signup_app	first_device_type	first_browser	country_destination	action	action_type	action_detail	device_type	secs_elapsed
untracked	Web	Mac Desktop	Chrome	NDF	NaN	NaN	NaN	NaN	NaN
untracked	Web	Mac Desktop	Chrome	NDF	NaN	NaN	NaN	NaN	NaN
untracked	Web	Windows Desktop	IE	US	NaN	NaN	NaN	NaN	NaN
untracked	Web	Mac Desktop	Firefox	other	NaN	NaN	NaN	NaN	NaN
untracked	Web	Mac Desktop	Chrome	US	NaN	NaN	NaN	NaN	NaN

sessions_merge_train

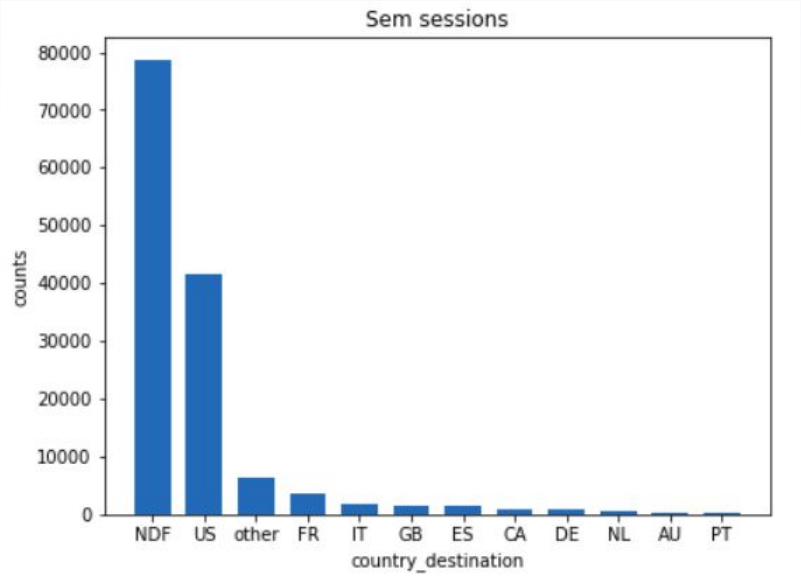
- merge com train

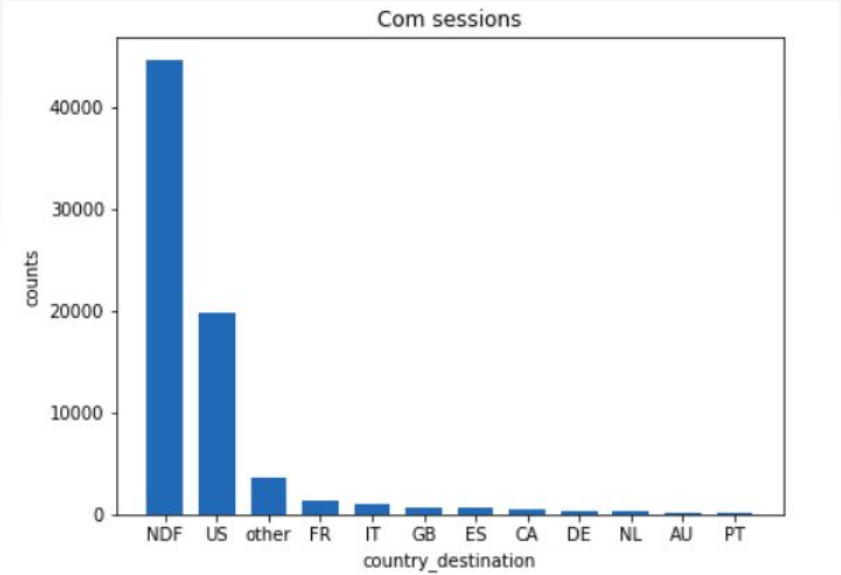
Alguns Ids não tinham dados do sessions

```
id                          0
gender                      0
age                         0
signup_method               0
signup_flow                 0
language                    0
affiliate_channel           0
affiliate_provider          0
first_affiliate_tracked     0
signup_app                  0
first_device_type           0
first_browser               0
country_destination         0
action                 137814
action_type            137814
action_detail          137814
device_type            137814
secs_elapsed           137814
dtype: int64
```

sessions_merge_train

- merge com train





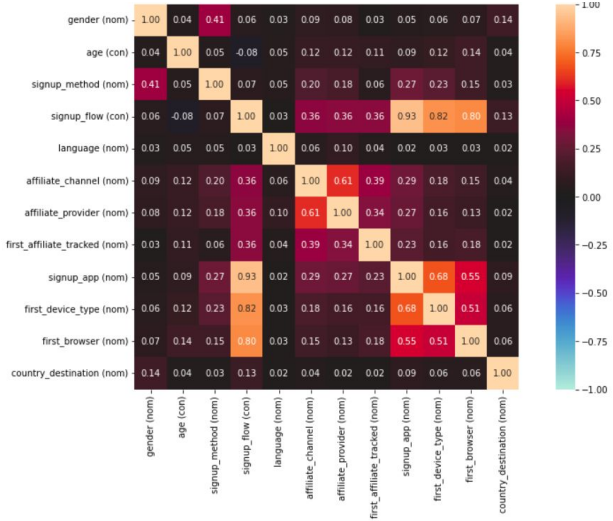
sessions

- Final do sessions

	id	count_actions	total_secs	index	callback	pending	requested	travel_plans_current	show	ajax_refresh_subtotal	...	Android App Unknown Phone/Tablet	Android Phone	Windows Desktop
0	00023iyk9l	40.0	867896.0	1	1	1	1	1	1	1	...	0	0	0
1	0010k6l0om	63.0	586543.0	1	1	0	0	0	1	1	...	0	0	0
2	001wyh0pz8	90.0	282965.0	1	0	0	0	0	1	0	...	1	0	0
3	0028jgx1x1	31.0	297010.0	0	0	0	0	0	1	0	...	0	1	0
4	002qnbzfs5	789.0	6487080.0	1	0	0	0	0	1	0	...	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
135478	zzxox7jnrx	89.0	639436.0	1	1	0	0	1	1	1	...	0	0	1
135479	zzy7t0y9cm	8.0	73771.0	0	0	0	0	0	1	1	...	0	0	1
135480	zzysuoqg6x	3.0	1731.0	0	0	0	0	0	0	0	...	0	0	1
135481	zzywmcn0jv	51.0	2149949.0	1	0	0	0	0	1	1	...	0	0	1
135482	zzzlylp57e	74.0	430959.0	1	0	1	1	0	1	1	...	0	0	1

135483 rows × 339 columns

Correlação


gender (nom) 1.00 0.04 0.41 0.06 0.03 0.09 0.08 0.03 0.05 0.06 0.07 0.14
age (con) 0.04 1.00 0.05 -0.08 0.05 0.12 0.12 0.11 0.09 0.12 0.14 0.04
signup_method (nom) 0.41 0.05 1.00 0.07 0.05 0.20 0.18 0.06 0.27 0.23 0.15 0.03
signup_flow (con) 0.06 -0.08 0.07 1.00 0.03 0.36 0.36 0.36 0.93 0.82 0.80 0.13
language (nom) 0.03 0.05 0.05 0.03 1.00 0.06 0.10 0.04 0.02 0.03 0.03 0.02
affiliate_channel (nom) 0.09 0.12 0.20 0.36 0.06 1.00 0.61 0.39 0.29 0.18 0.15 0.04
affiliate_provider (nom) 0.08 0.12 0.18 0.36 0.10 0.61 1.00 0.34 0.27 0.16 0.13 0.02
first_affiliate_tracked (nom) 0.03 0.11 0.06 0.36 0.04 0.39 0.34 1.00 0.23 0.16 0.18 0.02
signup_app (nom) 0.05 0.09 0.27 0.93 0.02 0.29 0.27 0.23 1.00 0.68 0.55 0.09
first_device_type (nom) 0.06 0.12 0.23 0.82 0.03 0.18 0.16 0.16 0.68 1.00 0.51 0.06
first_browser (nom) 0.07 0.14 0.15 0.80 0.03 0.15 0.13 0.18 0.55 0.51 1.00 0.06
country_destination (nom) 0.14 0.04 0.03 0.13 0.02 0.04 0.02 0.02 0.09 0.06 0.06 1.00
gender (nom)
age (con)
signup_method (nom)
signup_flow (con)
language (nom)
affiliate_channel (nom)
affiliate_provider (nom)
first_affiliate_tracked (nom)
signup_app (nom)
first_device_type (nom)
first_browser (nom)
country_destination (nom)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

sessions

	user_id	action	action_type	action_detail	device_type	secs_elapsed
0	d1mm9tcy42	lookup	NaN	NaN	Windows Desktop	319.0
1	d1mm9tcy42	search_results	click	view_search_results	Windows Desktop	67753.0
2	d1mm9tcy42	lookup	NaN	NaN	Windows Desktop	301.0
3	d1mm9tcy42	search_results	click	view_search_results	Windows Desktop	22141.0
4	d1mm9tcy42	lookup	NaN	NaN	Windows Desktop	435.0
...	...	...	...	...	...	...
10567732	9uqfg8txu3	dashboard	view	dashboard	Windows Desktop	556.0
10567733	9uqfg8txu3	edit	view	edit_profile	Windows Desktop	6624.0
10567734	9uqfg8txu3	webcam_upload	-unknown-	-unknown-	Windows Desktop	200125.0
10567735	9uqfg8txu3	active	-unknown-	-unknown-	-unknown-	17624.0
10567736	9uqfg8txu3	show_personalize	data	user_profile_content_update	Windows Desktop	1221.0

10567737 rows × 6 columns

Pré-processamento

- StandardScaler
- One-Hot-Encoder

	0	1	2	3	4	5	6	7	8	9	10	11
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...
73059	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
73060	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
73061	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
73062	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
73063	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0

73064 rows × 12 columns

	age	signup_flow	count_actions	total_secs	index	callback	pending	requested	travel_plans_current	show	...	38	39	40	41	42	43	44	45	46	47
0	3.173985	-0.521317	0.462929	1.001654	1.0	0.0	0.0	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
1	-0.075396	-0.521317	-0.586666	-0.682087	0.0	0.0	0.0	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
2	-0.075396	-0.521317	-0.524402	-0.196996	0.0	0.0	1.0	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
3	-0.075396	-0.521317	0.685300	1.172957	1.0	0.0	0.0	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
4	-0.075396	2.151240	-0.595561	-0.789442	1.0	0.0	0.0	0.0	0.0	0.0	...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
73059	-0.307495	-0.521317	0.311716	1.897923	1.0	1.0	0.0	0.0	1.0	1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
73060	-0.075396	-0.521317	1.450259	0.714760	1.0	0.0	0.0	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
73061	-0.307495	-0.521317	-0.506612	-0.611415	0.0	0.0	0.0	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
73062	-0.075396	2.151240	0.000395	-0.612133	0.0	0.0	0.0	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
73063	-0.075396	2.151240	-0.302030	0.652155	1.0	0.0	0.0	0.0	0.0	1.0	...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

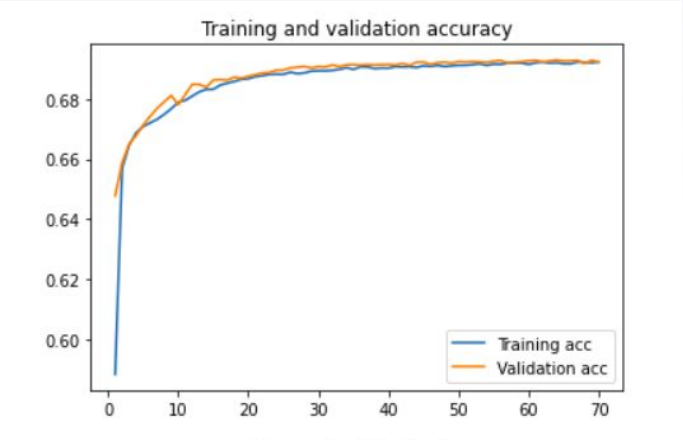
73064 rows × 388 columns

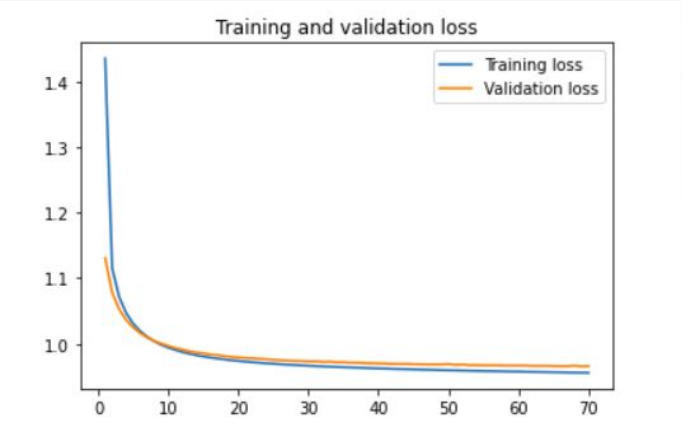
Treinamento

MLP

```
def build_model(table):
    model = Sequential([
        Dense(2*len(table.columns), input_dim=len(table.columns), activation="tanh"),
        Dense(len(table.columns), activation='tanh'),
        Dense(len(table.columns)/3, activation='tanh'),
        Dense(12, activation='softmax'),
    ])
    return model
```

Resultados






```
Epoch 69/70
26/26 [==============================] - 0s 7ms/step - loss: 0.9560 - accuracy: 0.6920 - val_loss: 0.9657 - val_accuracy: 0.6928
Epoch 70/70
26/26 [==============================] - 0s 7ms/step - loss: 0.9558 - accuracy: 0.6922 - val_loss: 0.9660 - val_accuracy: 0.6922
```

Resultados

Resultado no dataset de treino: `0.6929288506507874`

Validação cruzada: `0.6937889635562897`

YOUR RECENT SUBMISSION

sub.csv
Submitted by Andre Pereira · Submitted just now

Score: 0.86926
Private score: 0.87476

MLP vs XGBOOST

	Private Score
0.6929288506507874	0.87476
Acc = 0.6973243002805721	0.87734

Obrigado!