WiDS Datathon

Keven Ronald Fernández Carrillo
Analista de Data Intelligence en E-Business y Marketing Digital - BCP
kevenfernandezc@gmail.com

Meetup Data Science Lima, 06 Abril 2018

Institute for Computational & Mathematical Engineering

WOMEN IN DATA SCIENCE
MARCH 5, 2018
@ STANFORD UNIVERSITY

Conferencia Global de Mujeres en Ciencia de Datos (WiDS)

Ubicaciones globales de WiDS
80 eventos, 73 ciudades, 30 países, 1 misión

ICME
Stanford Institute for Computational & Mathematical Engineering

kaggle

Equipo: KKLL

Keven Fernández
Data Intelligence
BCP

Michael Larico
Big Data Engineer
BCP

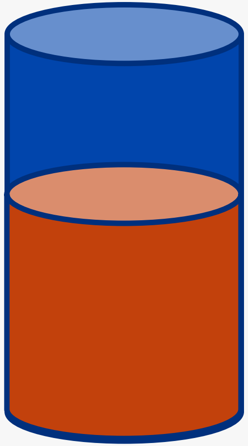
Linda Anicama
Customer Intelligence Analyst
Claro

Katerine Clavo
Big Data Specialist
BCP

Private Leaderboard

#	Team Name	Score
1	Minions	0.97501
2	icf-cdr - Peru - Trust you	0.97469
3	KKLL - Peru	0.97460
4	WomenInKaggle	0.97448
5	Women Who Code Kyiv	0.97432
6	Leetle	0.97427
7	Dracarys	0.97424

Revisión Previa de la Data



Train

18,255 registros (40%)

Test

27,285 registros (60%)

45,540 registros

Target: **is_female**
- 1: es femenino
- 0: es masculino

Data Train:
1: 9,805 (53.7%)
0: 8,450 (46.3%)

1,234 features
- 238 vars integer
- 900 vars float
- 96 vars string

Entre los tipos de variables predictoras se contaba con información:

- Sociodemográfica
- Familiar
- Tipos de Documentos
- Financiera (Personal, Hogar)
- Bancaria
- Telefonía móvil
- Programas del Estado
- Uso de Servicios de Dinero Móvil
- Seguros
- Otros

Enfoque de la Solución

1. **Eliminar variables con porcentaje de nulos mayor a 97%. Variables pre-seleccionadas: 575**
2. **[** *Reemplazar Nulos* **]**

```
#data_ini.fillna(-9009, inplace = True)
```

3. **Transformar variables:**

- Vars Categóricas Ordinales → Numérica
- Vars Categóricas Cardinales → Dummies (0,1)
- Agrupar Clases Minoritarias

2. Var: DG6

DG6

	Atributo	Cantidad	%Total
0	1	6725	36.839222
1	2	6569	35.984662
2	3	3437	18.827718
3	7	632	3.462065
4	4	430	2.355519
5	5	205	1.122980
6	99	190	1.040811
7	6	53	0.290331
8	9	14	0.076691

```
# DG6
feature = 'DG6'
#data_ini['FLG_'+feature] = data_ini[feature]
data_ini.loc[(data_ini[feature] != 1) &
             (data_ini[feature] != 2) &
             (data_ini[feature] != 3), feature ] = 10001
```

Enfoque de la Solución

4. Creación de nuevas Variables:

- Generación según edad
- Agrupar niveles de educación
- Agrupar clases entre diferentes variables (pero del mismo concepto):

```
In [63]: # Having_bank_account
         list_vars = ['FF7_1',
         'FF7_2',
         'FF7_3',
         'FF7_4',
         'FF7_5',
         'FF7_6',
         'FF7_7',
         'FF7_96']
         data_ini[list_vars].head()
```

```
data_ini['flg_bank_account'] = 0
data_ini.loc[(data_ini['FF7_1']==1)|
         (data_ini['FF7_2']==1)|
         (data_ini['FF7_3']==1)|
         (data_ini['FF7_6']==1)|
         (data_ini['FF7_7']==1),'flg_bank_account'] = 1
```

```
cross_target(data_ini,"flg_bank_account")
```

```
--------------------------------------------------------------
---- Var:  flg_bank_account
   Atributo  Cantidad     %Total
0         0     16867  92.396604
1         1      1388   7.603396
   flg_bank_account  % ratio_conv
0               0.0     54.722239
1               1.0     41.426513
```

- Entre otros (Flag_Viuda, Actividad Anual, Flag_Housewife, Relación con la cabeza del hogar, etc)

Enfoque de la Solución

5. Selección de Variables:

- Creación de un modelo Inicial: Ejm **Random Forest**
- Proceso Iterativo de “n mejores variables”:

```
list_n_features = list()
list_metrica = list()

vars_importances_ordered = ['var_best_1',
                            'var_best_2',
                            'var_best_3',
                            'var_best_4']

for i in range(0,len(vars_importances_ordered)):
    n_features = len(vars_importances_ordered) - i # nro de variables a usar
    tmp_vars_importances_ordered = vars_importances_ordered[0:n_features]
    print(tmp_vars_importances_ordered)

    model_tmp = fit(data_train[tmp_vars_importances_ordered]) # Entrenar Modelo (puede ser diferente al modelo inicial)
    metrica_tmp = calcular_metrica(model_tmp) # Calcular la métrica según la que te evalúan en la competencia

    list_n_features.append(n_features)
    list_metrica.append(metrica_tmp)
```

```
['var_best_1', 'var_best_2', 'var_best_3', 'var_best_4']
['var_best_1', 'var_best_2', 'var_best_3']
['var_best_1', 'var_best_2']
['var_best_1']
```

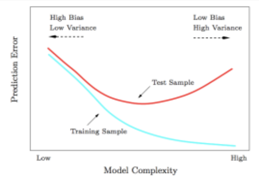
Enfoque de la Solución

6. Entrenamiento y Tunning del Modelo

LightGBM

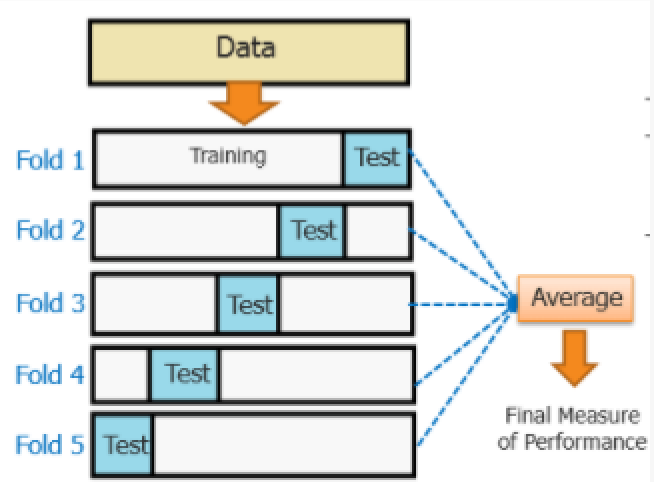
```
num_boost_round = 10000

lgb_params = {}

lgb_params['learning_rate'] = 0.009
lgb_params['colsample_bytree'] = 0.2
lgb_params['min_child_samples'] = 40
lgb_params['seed'] = 123
lgb_params['objective'] = 'binary'
lgb_params['max_depth'] = 10

lgb_params['num_leaves'] = 80

lgb_params['metric'] = 'auc'
lgb_params['training_metric'] = True
```

```
model = lgb.train(lgb_params,
                  lgb_train,
                  num_boost_round=num_boost_round,
                  valid_sets = lgb_eval,
                  early_stopping_rounds = 80,
                  verbose_eval = 0)
```



Enfoque de la Solución

6. Validación del modelo



Resultados Cross Validation

K folds = 10

Ok - 0 : 0.999996 0.972194
Ok - 1 : 0.998921 0.972479
Ok - 2 : 0.997454 0.976474
Ok - 3 : 0.999759 0.976989
Ok - 4 : 0.999974 0.972798
Ok - 5 : 0.999649 0.981537
Ok - 6 : 0.998333 0.973183
Ok - 7 : 0.999844 0.972989
Ok - 8 : 0.999641 0.972255
Ok - 9 : 0.999073 0.966340

auc - train: 0.999264463884
auc - test : 0.973760712252

```
cross_validation.KFold( len(X_train_cv),      n_folds=10, shuffle = True, random_state = 16)
vs
cross_validation.StratifiedKFold( y_train_cv, n_folds=10, shuffle = True, random_state = 16)
```

Train: 52.5 Test: 54.3

Train: 53.7 Test: 53.7

Enfoque de la Solución

6. Ensamblado – Promedio Ponderado

LightGBM_10

CV: 0.972421

PUB: 0.97447

LB PRIV: 0.97384

x 5/8

XGB_19

CV: 0.972767

PUB: 0.97440

LB PRIV: 0.97414

x 1/8

LightGBM_26

CV: 0.973760

PUB: 0.97437

LB PRIV: 0.97444

x 2/8

CV: 0.973982

PB: 0.97488

PRIV: 0.97460

	test_id	is_female_x	is_female_y	is_female_z	is_female
0	0	0.996615	0.994650	0.994433	0.995824
1	1	0.061414	0.097544	0.057922	0.065057
2	2	0.986808	0.984562	0.978972	0.984568
3	3	0.986533	0.978258	0.986918	0.985595
4	4	0.467452	0.589511	0.637860	0.525312

	a	b	c	metric
0	0	0	1	0.972421
10	0	1	0	0.972767
120	1	0	0	0.973760
132	1	1	1	0.973839

	a	b	c	metric
0	5	1	2	0.973982

Keven Ronald Fernández Carrillo

Analista de Data Intelligence en E-Business y Marketing Digital - BCP

kevenfernandezc@gmail.com

Github: https://github.com/KevenRFC

LinkedIn: https://www.linkedin.com/in/keven-fern%C3%A1ndez-carrillo-50b07aa2/