BBVA DATA CHALLENGE

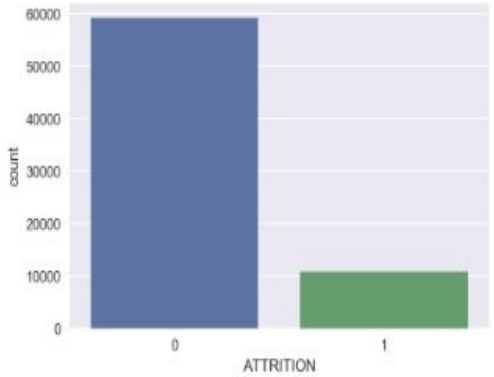
El reto que desafía tus conocimientos

Robert Alonso Aduviri Choque

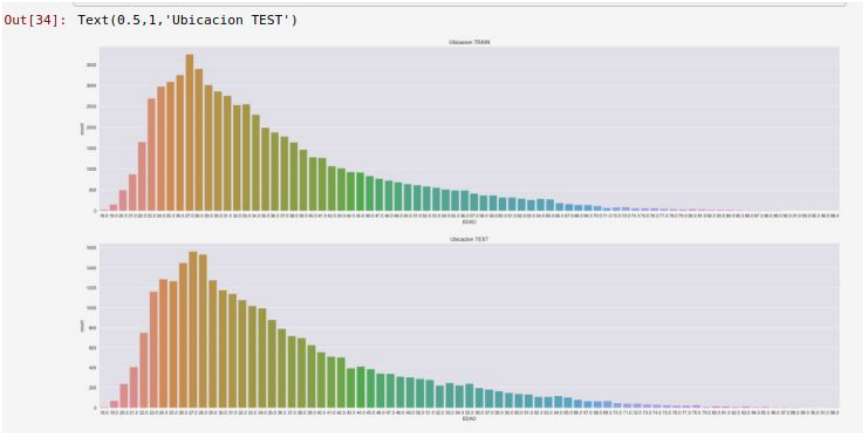
Exploratory Data Analysis (EDA)

```
In [33]: # Hay desbalance de 6:1 en la data
         sns.countplot(train_clientes['ATTRITION'])
```

```
Out[33]: <matplotlib.axes._subplots.AxesSubplot at 0x255023ec5c0>
```




```
Out[34]: Text(0.5,1,'Ubicacion TEST')
```



Data Preprocessing

```
# Solo hay nulos en:
# RANG_INGRESO        (60583 / 70000: 86.54%)
# FLAG_LIMA_PROVINCIA (66614 / 70000: 95.16%)
# EDAD                (64674 / 70000: 92.39%)
# ANTIGUEDAD          (68238 / 70000: 97.48%)
# train_clientes.info()
```

```
# Hay solo un nulo en:
# DICTAMEN (1 / 51417)
# train_requerimientos.info()
```

```
# Solo hay nulos en:
# RANG_INGRESO        (25862 / 30000: 86.21%)
# FLAG_LIMA_PROVINCIA (28487 / 30000: 94:96%)
# EDAD                (27649 / 30000: 92.16%)
# ANTIGUEDAD          (29254 / 30000: 97.51%)
# test_clientes.info()
```

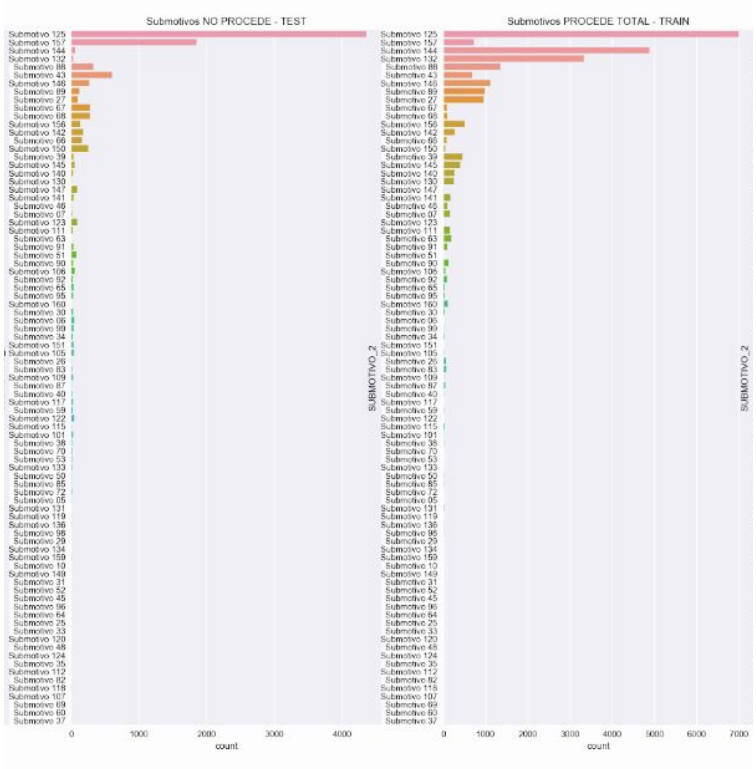
```
# No hay nulos
# test_requerimientos.info()
```

	ID_CORRELATIVO	TIPO_REQUERIMIENTO2	DICTAMEN	CODMES	PRODUCTO_SERVICIO_2	SUBMOTIVO_2
0	64216	Reclamo	NO PROCEDE	201206	Producto 20	Submotivo 43
1	64216	Solicitud	NO PROCEDE	201205	Producto 20	Submotivo 157

	ID_CORRELATIVO	CODMES	FLG_BANCARIZADO	RANG_INGRESO	FLAG_LIMA_PROVINCIA	EDAD	ANTIGUEDAD
0	35653	201208	1	Rang_ingreso_06	Lima	25.0	6.0
1	66575	201208	1	Rang_ingreso_03	Provincia	27.0	0.0
2	56800	201208	1	Rang_ingreso_01	Provincia	34.0	4.0
3	8410	201208	1	Rang_ingreso_04	Provincia	63.0	5.0
4	6853	201208	1	NaN	Lima	25.0	0.0

Data Preprocessing

- Merge clientes & requerimientos (column-category)
- Label encoding
- One-hot encoding
- Null values: (mean + median) / 2, 0
- Outlier clipping: [1%, 99%]
- Normalization: (min-max)
- Is-null feature (Feature Engineering)



Modeling - Baseline

- Cross validation, 5 folds, fixed random seed
- Merged datasets: (70,000, 73) (30,000, 72)
- Top results per model:
  - CatBoost (0.3024)
  - RandomForest (0.3114)
  - XGBoost (0.3162)
  - ExtraTrees (0.3420)
  - LogisticRegression (0.3489)
  - SGD (0.3489)

	dataset	model	params	aucroc	aucroc-std	logloss	logloss-std	time
16	clipped	CatBoostClassifier	default	0.856796	0.001835	0.302478	0.001686	558997
24	norm	CatBoostClassifier	default	0.856575	0.001992	0.302597	0.001660	571138
2	preprocessed	CatBoostClassifier	default	0.856396	0.001619	0.303038	0.001421	393646
5	merge	CatBoostClassifier	default	0.855920	0.001691	0.303143	0.001529	607605
8	full	CatBoostClassifier	default	0.855452	0.001837	0.303650	0.001699	756572
45	norm	RandomForestClassifier	n_est:200 max_d:15 min_samp_leaf:5	0.850701	0.002111	0.311489	0.001323	46719
42	clipped	RandomForestClassifier	n_est:200 max_d:15 min_samp_leaf:5	0.850276	0.002412	0.311869	0.001705	46448
39	full	RandomForestClassifier	n_est:200 max_d:15 min_samp_leaf:5	0.846945	0.002880	0.315545	0.001936	45321
3	merge	XGBClassifier	default	0.843034	0.002004	0.316268	0.001360	100629
14	clipped	XGBClassifier	default	0.843009	0.002103	0.316304	0.001336	94696
22	norm	XGBClassifier	default	0.843009	0.002103	0.316304	0.001336	105525
0	preprocessed	XGBClassifier	default	0.843053	0.002046	0.316309	0.001531	51635
6	full	XGBClassifier	default	0.842251	0.002612	0.316945	0.001697	210007
33	clipped	RandomForestClassifier	n_est:200	0.844048	0.002991	0.330048	0.005200	66193
36	norm	RandomForestClassifier	n_est:200	0.843689	0.003163	0.330961	0.005063	66263
30	full	RandomForestClassifier	n_est:200	0.843167	0.003365	0.331042	0.007104	61264
46	norm	ExtraTreesClassifier	n_est:200 max_d:15 min_samp_leaf:5	0.814461	0.002607	0.342048	0.002137	52661
43	clipped	ExtraTreesClassifier	n_est:200 max_d:15 min_samp_leaf:5	0.813167	0.002087	0.342761	0.001748	50135
28	norm	LogisticRegression	default	0.800786	0.002054	0.348578	0.002038	14378
47	norm	SGDClassifier	max_iter:1000	0.799823	0.001892	0.348906	0.002008	141958
40	full	ExtraTreesClassifier	n_est:200 max_d:15 min_samp_leaf:5	0.804185	0.002372	0.350379	0.001822	46284
38	norm	SGDClassifier	tol:1e-4	0.791592	0.010746	0.360025	0.012450	3539

Modeling - Baseline

- Cross validation, 5 folds (stratified), fixed random seed
- Merged datasets: (70,000, 73) (30,000, 72)
- Top results per model:
  - CatBoost (0.3024)
  - RandomForest (0.3114)
  - XGBoost (0.3162)
  - ExtraTrees (0.3420)
  - LogisticRegression (0.3489)
  - SGD (0.3489)

	dataset	model	params	aucroc	aucroc-std	logloss	logloss-std	time
16	clipped	CatBoostClassifier	default	0.856796	0.001835	0.302478	0.001686	558997
24	norm	CatBoostClassifier	default	0.856575	0.001992	0.302597	0.001660	571138
2	preprocessed	CatBoostClassifier	default	0.856396	0.001619	0.303038	0.001421	393646
5	merge	CatBoostClassifier	default	0.855920	0.001691	0.303143	0.001529	607605
8	full	CatBoostClassifier	default	0.855452	0.001837	0.303650	0.001699	756572
45	norm	RandomForestClassifier	n_est:200 max_d:15 min_samp_leaf:5	0.850701	0.002111	0.311489	0.001323	46719
42	clipped	RandomForestClassifier	n_est:200 max_d:15 min_samp_leaf:5	0.850276	0.002412	0.311869	0.001705	46448
39	full	RandomForestClassifier	n_est:200 max_d:15 min_samp_leaf:5	0.846945	0.002880	0.315545	0.001936	45321
3	merge	XGBClassifier	default	0.843034	0.002004	0.316268	0.001360	100629
14	clipped	XGBClassifier	default	0.843009	0.002103	0.316304	0.001336	94696
22	norm	XGBClassifier	default	0.843009	0.002103	0.316304	0.001336	105525
0	preprocessed	XGBClassifier	default	0.843053	0.002046	0.316309	0.001531	51635
6	full	XGBClassifier	default	0.842251	0.002612	0.316945	0.001697	210007
33	clipped	RandomForestClassifier	n_est:200	0.844048	0.002991	0.330048	0.005200	66193
36	norm	RandomForestClassifier	n_est:200	0.843689	0.003163	0.330961	0.005063	66263
30	full	RandomForestClassifier	n_est:200	0.843167	0.003365	0.331042	0.007104	61264
46	norm	ExtraTreesClassifier	n_est:200 max_d:15 min_samp_leaf:5	0.814461	0.002607	0.342048	0.002137	52661
43	clipped	ExtraTreesClassifier	n_est:200 max_d:15 min_samp_leaf:5	0.813167	0.002087	0.342761	0.001748	50135
28	norm	LogisticRegression	default	0.800786	0.002054	0.348578	0.002038	14378
47	norm	SGDClassifier	max_iter:1000	0.799823	0.001892	0.348906	0.002008	141958
40	full	ExtraTreesClassifier	n_est:200 max_d:15 min_samp_leaf:5	0.804185	0.002372	0.350379	0.001822	46284
38	norm	SGDClassifier	tol:1e-4	0.791592	0.010746	0.360025	0.012450	3539

Hyperparameter Optimization

Random Search for Hyper-Parameter Optimization

James Bergstra JAMES.BERGSTRA@UMONTREAL.CA
Yoshua Bengio YOSHUA.BENGIO@UMONTREAL.CA
Département d'Informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC, H3C 3J7, Canada

Model	**LL (before)**	**LL (after)**	**Main parameters**
LightGBM	0.396168	0.298089	num_leaves, reg_alpha, reg_lambda
XGBoost	0.316268	0.298051	max_depth, n_estimators, min_child_weight
CatBoost	0.302478	0.299388	depth, od_type, od_pval, l2_leaf_reg
RandomForest	0.311489	0.302411	n_estimators, max_depth, min_samples_leaf
ExtraTrees	0.342048	0.300252	n_estimators, max_depth, max_features
LogisticRegression	0.348906	0.348610	C, class_weight, max_iter
MLP	0.338674	0.318510	hidden_layer_sizes, alpha, early_stopping
KNN	1.646710	0.356579	n_neighbors, alpha, max_iter

Ensembling (Stacking)

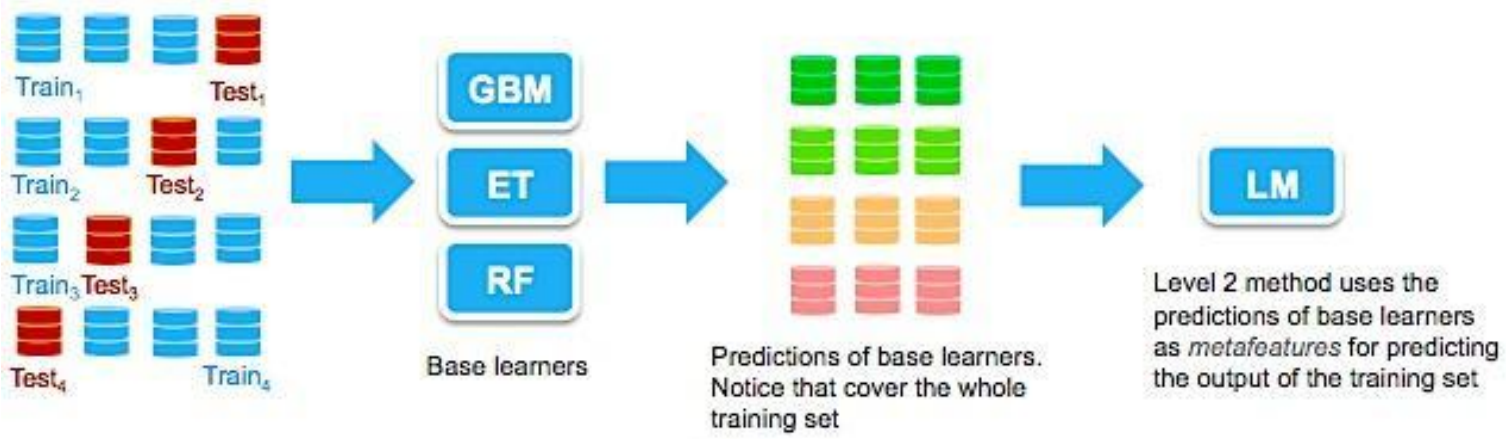
Best base model:

LightGBM (0.299388)

Metamodel:

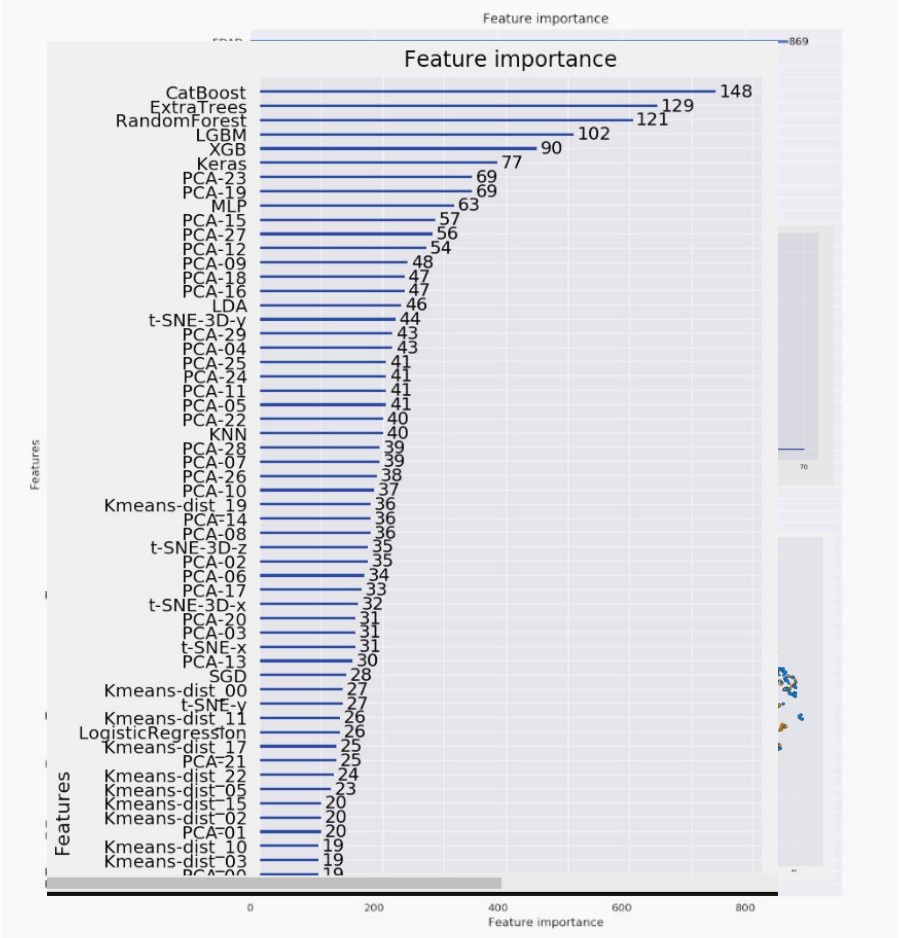
CatBoost (0.294979)

submission11_Ensemble.csv 14 days ago by Robert Aduviri Ensemble (9 models)	0.29425	0.29203



Feature Engineering

- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- t-distributed Stochastic Neighbor Embedding (t-SNE) (2D, 3D)
- SVM as features
- Distance to centroids of KMeans clustering
- Deep Neural Network with Dropout and Batch Normalization layers
- Alternative metamodel: StackNet
- Public / Private LB: 0.29125 / 0.29425



Conclusiones

- Las GBM se reafirman como mejor tipo de modelo para data tabular
- La efectividad de un ensamble de modelos depende de la variedad de los modelos base (GBMs, tree-based, linear, KNN), y es importante que sean optimizados previamente
- La transformación de la data debe alineada con el tipo de modelos que se utilicen (normalization, outlier clipping)
- Es útil contar con representaciones que resuman la información de la data original en el ensamble (PCA, LDA, t-SNE). Al mismo tiempo, permiten filtrar ruido
- En general es bueno considerar la complejidad del modelo versus la ganancia en poder predictivo (interpretability vs accuracy tradeoff)