Le CAP

(Cumulative Accuracy Profile)

Comment utiliser la courbe CAP (Cumulative Accuracy Profile)

Pour mieux évaluer votre modèle et formuler des intuitions

Author	Version	Date
Carlos Carvalho	Version 1.0	23.03.2019

INDEX

1 Introduction

Dans la suite de mon cours de Machine Learning et Deep Learning j'ai été amené à effectuer plusieurs développements, évaluer et tester différentes méthodes utilisées pour construire et évaluer de modèles statistiques.

2 Objectifs

L'objectif de ce document c'est comparer le score obtenu avec les outils de machine Learning et le Réseaux de Neurones Artificiel (RNA) pour les mêmes observations d'un dataset en utilisant le modèle de Régression Logistique.

Pour ce fait, nous devons extraire une liste avec les probabilités prédites des clients plus susceptibles de quitter « une banque », et comparer le score CAP (Cumulative Accuracy Profile) obetnu pour les deux modèles.

Une liste contenant le taux de départs des clients d'une banque « Churn Model » pendant un intervalle de temps a servi comme jeu de données.

3 Environnement

Les tests de comparaison ont été effectués sur 2 environnements distincts sur la même machine, un environnement virtuel pour les teste avec les réseaux de neurones artificiel, et l'environnement local.

Les tests ont été effectués à intervalles de temps différents, c'est-à-dire, es tests n'ont pas été exécutés en simultané dans la même machine.

Cependant GretL a été utilisé comme outils pour construire e modèle de référence, pour valider les méthodes et évaluer les modèles entre le train et le test set.
Par la suite les mêmes procédures ont été développé et implémentés dans l'environnement Python.

3.1 *Environnement technique*

- OS X
- GretL
- Conda
- Python 3.7
- Environment virtuel
    - Python 3.6
    - TensorFlow
    - Keras

3.2 *Train set et Test Set*

La validation du model a été effectué à partir des résulta obtenues entre le Train set et le test set avec un ratio de précision de AR test proche du ratio de précision de AR.train
L'objectif c'est se servir comme modèle de référence pour le test de comparaison entre les méthodes de Régression logistique et cette même méthode via les Réseaux de Neurones

4 Le jeu de données « Dataset »

4.1 *Description*

Le dataset « Churn_model ou Taux de départs » c'est extrait d'une liste de clients d'une hypothétique banque.

Les informations contenues dans cette liste sont le résultat d'une enquête effectué par la banque auprès de ses clients suite à un taux de départ très signifiant pendant un intervalle de temps.

Cette liste contient le noms et prénoms de ses clients, leur âge, les creditscore (c'est un indicateur interne à la banque qui informe sur la capacité du client à rembourser ses crédit), le solde du compte, si le client possédé ou pas une carde de crédit, le nombre de produits achetés par le client, si le client est un membre actif ou pas de leur site et une dernière information avec si le client à quitte ou non la banque.

4.2 *Échantillon extrait du jeu de données « Churn Modeling »*

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMem	timatedSalary	Exited
2	1	15634602	Hargrave	619	France	Female	42	2	0	1	1	1	101348.88	1
3	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
4	3	15619304	Onio	502	France	Female	42	8	159660.8	3	1	0	113931.57	1
5	4	15701354	Boni	699	France	Female	39	1	0	2	0	0	93826.63	0
6	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.1	0
7	6	15574012	Chu	645	Spain	Male	44	8	113755.78	2	1	0	149756.71	1
8	7	15592531	Bartlett	822	France	Male	50	7	0	2	1	1	10062.8	0
9	8	15656148	Obinna	376	Germany	Female	29	4	115046.74	4	1	0	119346.88	1
10	9	15792365	He	501	France	Male	44	4	142051.07	2	0	1	74940.5	0
11	10	15592389	H?	684	France	Male	27	2	134603.88	1	1	1	71725.73	0
12	11	15767821	Bearce	528	France	Male	31	6	102016.72	2	0	0	80181.12	0
13	12	15737173	Andrews	497	Spain	Male	24	3	0	2	1	0	76390.01	0
14	13	15632264	Kay	476	France	Female	34	10	0	2	1	0	26260.98	0
15	14	15691483	Chin	549	France	Female	25	5	0	2	0	0	190857.79	0
16	15	15600882	Scott	635	Spain	Female	35	7	0	2	1	1	65951.65	0
17	16	15643966	Goforth	616	Germany	Male	45	3	143129.41	2	0	1	64327.26	0
18	17	15737452	Romeo	653	Germany	Male	58	1	132602.88	1	1	0	5097.67	1
19	18	15788218	Henderson	549	Spain	Female	24	9	0	2	1	1	14406.41	0
20	19	15661507	Muldrow	587	Spain	Male	45	6	0	1	0	0	158684.81	0
21	20	15568982	Hao	726	France	Female	24	6	0	2	1	1	54724.03	0
22	21	15577657	McDonald	732	France	Male	41	8	0	2	1	1	170886.17	0
23	22	15597945	Dellucci	636	Spain	Female	32	8	0	2	1	0	138555.46	0

5 Construction du Modèle

5.1 *Modèle*

Le modèle utilisé : Régression Logistique Multivarié Logit Binaire avec la variable dépendante connue.

5.2 *Méthode*

La méthode BackWard Élimination.a été utilisé pour la construction du modèle.

5.3 *Choix des Variables*

La méthode Backward Elimination a permis de sélectionner les variables avec les plus de pouvoir de prédiction, en d'autres mots, avec les plus de signification statistique.

- CreditScore
- Age
- Tenure,
- NumOfProducts
- ActiveMember
- Germany
- Female
- Log_Balance

5.4 *Transformation de variables*

Afin de garder la consistance de la variable « Balance » celle-ci a été transformé en utilisant le Log10 afin de garder le même effet sur l'ensemble de la population de clients et ne pas restreindre à une segmentation de clients.

5.5 *Évaluation du modèle*

5.5.1 Statistiques du Modèle

Les variables indépendantes retenues avec le plus de signification statistique, c'est-à-dire, avec le p-value inférieur au seuil de signification de 0.5, exception faite pour la variable « tenure » avec un p-value faiblement supérieur à alpha, mais retenu dans le modelé avec un faible pouvoir de prédiction.

Logit Regression Results

Dep. Variable:	Exited	No. Observations:	10000
Model:	Logit	Df Residuals:	9991
Method:	MLE	Df Model:	8
Date:	Fri, 22 Mar 2019	Pseudo R-squ.:	0.1528
Time:	11:19:56	Log-Likelihood:	-4282.6
converged:	True	LL-Null:	-5054.9
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.9126	0.237	-16.497	0.000	-4.377	-3.448
CreditScore	-0.0007	0.000	-2.408	0.016	-0.001	-0.000
Age	0.0727	0.003	28.221	0.000	0.068	0.078
Tenure	-0.0159	0.009	-1.699	0.089	-0.034	0.002
NumOfProducts	-0.0950	0.048	-1.999	0.046	-0.188	-0.002
IsActiveMember	-1.0758	0.058	-18.662	0.000	-1.189	-0.963
Geography_Germany	0.7476	0.065	11.492	0.000	0.620	0.875
Gender_Female	0.5267	0.054	9.672	0.000	0.420	0.633
Log_Balance	0.0690	0.014	4.945	0.000	0.042	0.096

1.1.1 Analyse de la Multi-colinéarité

Des valeurs supérieures à 10 peuvent indiquer un problème de colinéarité.
Le tableau ci-dessous affiche aucune colinéarité entre nous variables indépendantes.

5.5.1.1 Tableaux VIF – Variance Inflation Factor

Features	**VIF Factor**
CreditScore	1.00
Age	1.01
Tenure	1.00
NumOfProducts	1.15
IsActiveMember	1.01
Geography_Germany	1.27
Gender_Female	1.00
Log_Balance	1.42

5.6 *Matrice de corrélation*

Les matrices de corrélation extraites des 2 systèmes (GretL et Python) affichent exactement les mêmes résultats.

La valeur 0 indique aucune corrélation et la valeur 1 les variables sont très fortement corrélé.
On constate dans le tableau ci-dessous de corrélation asymétrique entre certaines variables (e.g.) (log_balance et NumOfProducts et aussi entre LogBalance et Tenure …)

5.6.1 Matrice de Corrélation (Tableau)

	CreditScore	Age	Tenure	NumOfProduct	ActiveMember	Geo_Germany	Female	Log_Balance
CreditScore	1.000000	-0.003965	0.000842	0.012238	0.025651	0.005538	0.002857	0.008159
Age	-0.003965	1.000000	-0.009997	-0.030680	0.085472	0.046897	0.027544	0.034530
Tenure	0.000842	-0.009997	1.000000	0.013444	-0.028362	-0.000567	-0.014733	-0.014941
NumOfProducts	0.012238	-0.030680	0.013444	1.000000	0.009612	-0.010419	0.021859	-0.329162
ActiveMember	0.025651	0.085472	-0.028362	0.009612	1.000000	-0.020486	-0.022544	-0.004769
Geo_Germany	0.005538	0.046897	-0.000567	-0.010419	-0.020486	1.000000	0.024628	0.435979
Female	0.002857	0.027544	-0.014733	0.021859	-0.022544	0.024628	1.000000	-0.005406
Log_Balance	0.008159	0.034530	-0.014941	-0.329162	-0.004769	0.435979	-0.005406	1.000000

6 Analyse CAP

Avant de démarrer la construction du tableau CAP nous devons calculer le taux de départs.
Nous avons le Nombre total de Clients, et le nombre de clients ayant quitté la banque.
Count = 10.000
Exit =2037
Exit Ratio (Churn ratio) = 0.2037 = 20%

Dans la liste de 10.000 clients, 2037 ont quitté la banque, ce que correspond à un taux de 20% de clients qui ont quitté la banque.
Ce taux sera utilisé dans les calculs de la courbe CAP par la suite.

6.1 *Construire le tableaux CAP*

Le tableaux cap est composé de 9 colonnes. Les 3 premières colonnes sont obtenues à partir du dataset et des probabilités prédites de notre model.

Les colonnes (**RowNumber et Exited**) contiennent les valeurs extraites à partir du dataset initial.
A cette grille, nous allons rajouter la colonne « **Predicted** » qui correspond aux résultats des probabilités de nos prédictions pour le modèle et trier le tout par l'ordre décroissant des prédictions.
Nous allons ainsi avoir en haut de notre tableau les probabilités prédites les plus élevés.

Les deux colonnes **(Total Select et % Total Selected)** correspondent aux nombres de sélections de notre dataset et au pourcenatge.de sélections correspondant pour chaque observation (e.g.) dans le premier tableau CAP le pourcentage de 0.001 dans la colonne **% Total Selected** correspond aux 10 premiers clients

Le deux colonnes suivantes **(Random Select et % Random Select)** correspondent aux nombres de clients ayant quitté la banque pour chaque client sélectionné au hasard
En d'autres mots**, « Random Select »** correspond a la fréquence cumulative absolue par rapport au nombre total des départs

La première ligne du tableau ou le premier client équivaut à 0,2037 sur le nombre total de départ (2037). Pour le 10 ème client, c'est l'équivalent à 2,037 ou 2% du total de départs.

La colonne **« % Random Select »** c'est tout simplement la fréquence cumulative relative des valeurs de la colonne « **Random Select »** par rapport au nombre total des clients (10.000)**,** (e.g.) le 10 ème client correspond ainsi à 0.1% sur le nombre total de départs.

Les deux dernières colonnes **(Model Select et % Model Select)** correspondent aux clients ayant quitté la banque calculés et classées selon un rang.

La colonne **(Model Select)** nous dit combien de clients quitte la banque à chaque fois qu'on sélectionne un client dans notre colonne de total de clients sélectionnés de notre classement de probabilités prédites de notre modèle

Comment calculer ? c'est le solde ou (running total) par rapport à la colonne « Exited » de notre grille.

Nous allons retrouver pour cette variable à la fin du tableau un total de 2037 soit disons les 2037 client qui ont quitté la banque

La colonne **(% Model Select)** correspond à la fréquence cumulative relative des valeurs de la colonne précédente par rapport aux nombres de clients qui ont quitté la banque. Pour la 10 ème ligne de notre tableau nous allons obtenir la valeur en effectuant l'opération 6/2037= 0.3%

6.2 *La règle d'évaluation du modèle en quantifiant le CAP*

La règle utilisée pour évaluer le modèle en quantifiant le CAP est affiché dans le tableau ci-dessous. Dans notre cas, pour la valeur de X=80% nous pouvons affirmer que c'est un très bon modèle

90% < X < 100%	Trop bon
80% < X < 90%	Très bon
70% < X < 80%	Bon
60% < X < 70%	Faible
X < 60%	Poubelle

6.3 *Le tableau CAP pour la régression logistique*

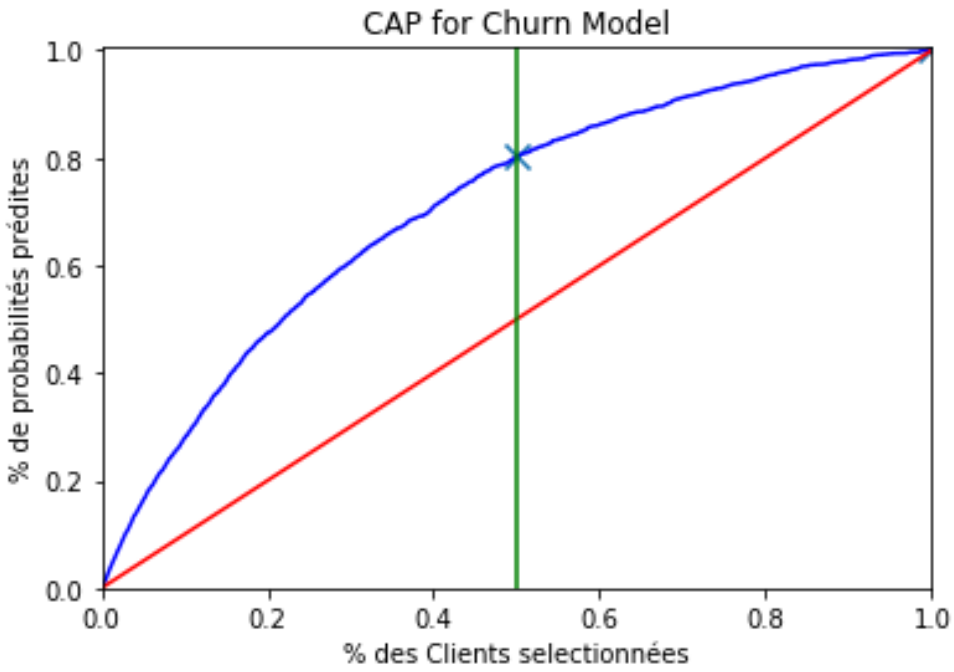
Les résultats dans le tableau ci-dessous sont extrait directement du dataframe. On constante que pour 50% des observations sélectionnées le model retrouve 80% des probabilités prédites.

Reste à représenter graphiquement (6.1.1) avec l'ensemble des valeurs de notre tableau CAP (page suivante)

RowNumber	Predicted	RowNumber	Exited	Total_selected	%_Total_Selected	Random_select	%-Random_select	Model_select	%-Model_select
3298	0.156294	3299	0	5000	0.5	1018.5	0.5	1633	0.801669

Index	predicted	RowNumber	Exited	Total_selected	%_Total_Selected	Random_select	%-Random_select	Model_select	%-Model_select
3531	0.931089	3532	1	1	0.0001	0.2037	0.0001	1	0.000490918
4815	0.921648	4816	0	2	0.0002	0.4074	0.0002	1	0.000490918
9587	0.914199	9588	0	3	0.0003	0.6111	0.0003	1	0.000490918
7499	0.901453	7500	1	4	0.0004	0.8148	0.0004	2	0.000981836
9555	0.893385	9556	1	5	0.0005	1.0185	0.0005	3	0.00147275
8488	0.886364	8489	1	6	0.0006	1.2222	0.0006	4	0.00196367
7629	0.873938	7630	1	7	0.0007	1.4259	0.0007	5	0.00245459
7692	0.871082	7693	0	8	0.0008	1.6296	0.0008	5	0.00245459
9747	0.864885	9748	1	9	0.0009	1.8333	0.0009	6	0.00294551
8156	0.862709	8157	0	10	0.001	2.037	0.001	6	0.00294551
7008	0.855857	7009	1	11	0.0011	2.2407	0.0011	7	0.00343643
4463	0.852046	4464	1	12	0.0012	2.4444	0.0012	8	0.00392734
4435	0.849614	4436	1	13	0.0013	2.6481	0.0013	9	0.00441826
416	0.839625	417	1	14	0.0014	2.8518	0.0014	10	0.00490918
4559	0.838719	4560	1	15	0.0015	3.0555	0.0015	11	0.0054001
4501	0.83692	4502	0	16	0.0016	3.2592	0.0016	11	0.0054001
7813	0.834129	7814	1	17	0.0017	3.4629	0.0017	12	0.00589102

6.3.1 Représentation graphique



6.3.2 Intuitions

La droite linaire rouge
Donne le nombre de clients qui ont quitté la banque pour chaque client sélectionné au hasard

La courbe bleue
Combien de client ont quitté la banque parmi le client sectionnées
(E.G) Pour les 10 Clients notre model donne 6 clients qui ont l'intention de quitter la banque

6.3.3 Matrice de confusion

La matrice de Confusion pour l'implémentation du modèle de Régression logistique affiche un taux de 81% pour le nombre de cas correctement prédits.

Nous pouvons ainsi calculer l'Accuracy Rate et Error Rate pour chacun des modèles implémentés.

Matrice de confusion pour les probabilités prédites du 1er modèle

		Prédite	
		0	1
Actuel	0	7698	265
	1	1599	438

Accuracy Rate = (7698+438)/10000 = 0.8136 = 81%

Error Rate = (1599+265)/100 = 0.19 = 19%

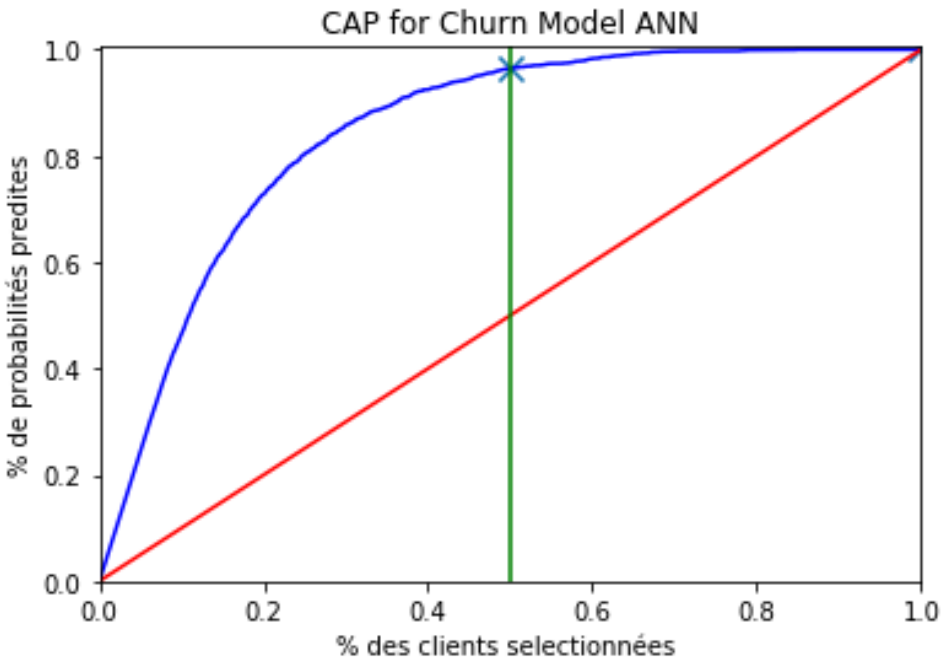
6.4 *Tableaux CAP pour le RNA*

La première observation dans le tableau ci-dessous, on constate que pour 96% des probabilités prédites de notre model des clients susceptibles de quitter la banque on les retrouve dans 50% de notre sélection.

RowNumber	predicted	RowNumber	Exited	Total_selected	%_Total_Selected	Random_select	%-Random_select	Model_select	%-Model_select
4057	0.07305	4058	0	5000	0.5	1018.5	0.5	1965	0.964654

Index	predicted	RowNumber	Exited	Total_selected	%_Total_Selected	Random_select	%-Random_select	Model_select	%-Model_select
2542	1.0	2543	1	1	0.0001	0.2037	0.0001	1	0.000490918
4822	1.0	4823	1	2	0.0002	0.4074	0.0002	2	0.000981836
4889	1.0	4890	1	3	0.0003	0.6111	0.0003	3	0.00147275
9438	1.0	9439	1	4	0.0004	0.8148	0.0004	4	0.00196367
8590	1.0	8591	1	5	0.0005	1.0185	0.0005	5	0.00245459
4516	1.0	4517	1	6	0.0006	1.2222	0.0006	6	0.00294551
4166	1.0	4167	1	7	0.0007	1.4259	0.0007	7	0.00343643
4520	1.0	4521	1	8	0.0008	1.6296	0.0008	8	0.00392734
893	1.0	894	1	9	0.0009	1.8333	0.0009	9	0.00441826
2720	1.0	2721	1	10	0.001	2.037	0.001	10	0.00490918
1469	1.0	1470	1	11	0.0011	2.2407	0.0011	11	0.0054001
4869	1.0	4870	1	12	0.0012	2.4444	0.0012	12	0.00589102
3615	1.0	3616	1	13	0.0013	2.6481	0.0013	13	0.00638193
8071	1.0	8072	1	14	0.0014	2.8518	0.0014	14	0.00687285
376	1.0	377	1	15	0.0015	3.0555	0.0015	15	0.00736377
2351	1.0	2352	1	16	0.0016	3.2592	0.0016	16	0.00785469
7567	1.0	7568	1	17	0.0017	3.4629	0.0017	17	0.00834561

6.4.1 Représentation Graphique – Courbe CAP pour le RNA



6.4.2 Intuitions

Dans l'axe des abscisses nous allons retrouver le pourcentage de clients sélectionnées et dans l'axe des ordonnées le pourcentage de probabilité prédites.

Une première intuition qualitative, plus notre droite bleue s'éloigne de la droite linaire rouge, plus le modèle est performant et à l'inverse, plus la courbe s'approche de la droite linaire rouge, moins performant est le modèle.

Une deuxième intuition quantitative, c'est considérer une droite verticale de 50% dans le graphique. On regarde en suite où la droite verticale coupe la courbe bleue, on projette l'abscisse 50% sur la courbe bleue et on projette à nouveau ce point de la courbe bleu sur l'axes des ordonnées

La droite linaire rouge
Donne le nombre de clients qui quitte la banque pour chaque client sélectionné au hasard.

La courbe bleue
Combien de client ont quitté la banque parmi le client sectionnées dans notre model.
(e.g) Pour les 10 premiers Clients de notre liste, le model donne exactement 10 clients qui quittent la banque.

*Ainsi, pour les valeurs correspondant à la courbe bleu représenté dans ce graphique on retrouve que pour **50% de des clients** sélectionnés dans notre modèle (la ligne verte) nous **pouvons cibler plus de 90%** des clients avec l'intention de quitter la banque (dans le cas de ce modèle la valeur exacte est de 96%).*

6.4.3 Matrice de confusion

La matrice de Confusion pour l'implémentation du modèle de Régression logistique démontre un **taux de 90%** pour le nombre de cas correctement prédits.
Autre constant flagrant c'est la valeur de false négative(685) est très nettement inférieur à la valeur de falses négatives affiché dans notre matrice de confusion pour le modèle précédente(1559)

Matrice de confusion pour les probabilités prédites du 2ème modèle

		Prédite	
		0	1
Actuel	0	7652	311
	1	685	1352

Accuracy Rate = (7652+1352)/10000 = 0.9004= 90%

Error Rate = (685+311)/100 = 0.9996 = 10%

7 Prévisions

Lors de la construction de notre tableaux CAP nous avons les probabilités estimées triés par l'ordre décroissante ceux qui nous permet classer les clients qui sont le plus susceptibles de quitter la banque.

C'est valide pour le cas de cette étude, mais dans le cas général ce classement de probabilités nous donne les observations pour laquelle les probabilités que le résultat soit égal à 1 est la plus élevé.

Ces probabilités prédites vont permettre de formuler de prévision puisque vont nous permettre de cibler nos recherches et dans notre cas cibler les clients le plus susceptibles de partir ce que va permettre à la banque de prendre les mesures appropriées pour effectuer les changements nécessaires comme proposer et ou adapter des offre le mieux correspondant à leurs attentes avec l'objectif de garder ces clients.

Avec notre courbe CAP nous pouvons aussi tirer d'autres insights.

Un premier insight révélé pour le premier modèle démontre que pour 50% des observations nous pouvons cibler 80% à 96% des clients susceptibles de quitter la banque.

Avec le premier modèle nous pouvions cibler 81% des clients qui serait susceptibles de quitter la banque avec seulement 50% des nous observations. Et c'est un très bon score,

Avec le deuxième modèle nous pouvons cibler 96% de des clients plus susceptibles de quitter la banque avec exactement 50% des mêmes observations.

Ainsi, en comparant avec la sélection de clients au hasard nous aurions obtenu un score de 50% pour 50% des clients sélectionnées au hasard.

Maintenant si on compare les score pour 20% des clients ou des observations entre la sélection au hasard et le model, nous pouvons constate que pour 20% des clients nous pouvons cibler 20% des clients susceptibles de quitter la banque, tandis que dans le 2ème modèle, avec uniquement 20% des clients sélectionnées par le model nous retrouvons plus de 80% de clients susceptibles de quitter la banque. C'est 3 fois plus.

8 Insights de la courbe CAP

En résume nous pouvons identifier 2 insights que nous pouvons extraire de de la courbe CAP

1) Le premier Insight de la courbe CAP on l'obtiens lors de la construction du tableau et va permettre le classement des probabilités (de la plus fort à la plus faible probabilité que le résultat soit 1).
2) Le deuxiéme insight, c'est l'optimisation de la rentabilité qui nous offre la possibilité de sélectionner le pourcentage optimal de clients (observations) qui répondent aux questions soulèves par le modèle

9 Annexes

Vous retrouverez dans le lien ci-dessous le documents permettant construire votre courbe CAP ainsi que le model de données.
https://github.com/Ccarl19/dataset

- Dataset «Churn Modeling.xlsl »
- Tableau CAP
- Tableau CAP pour le RNA