Université paris 8

PARIS 8 UNIVERSITE VINCENNES - SAINT-DENIS

Transaction connect

MÉMOIRE

pour obtenir le grade de Master délivré par

Université Paris 8 Vincennes à Saint-Denis

Mention *Informatique*

Parcours MIASHS Big data et fouille de données

présenté et soutenu publiquement par

Komlan Jean-Marie DANTODJI

le 22 septembre 2021

Classification et affectation des transactions aux magasins des détaillants

Encadrant universitaire : Rakia JAZIRI

Tuteur de stage : Thomas MOULIN

Stage effectué à : Transaction Connect
18 Rue du Faubourg du Temple, 75011 Paris

Université Paris 8
Laboratoire d'Informatique Avancée de Saint Denis
EA n° 4383 Saint Denis, France

MASTER

Sommaire

Remerciements

Pendant mon parcours de stage, j'ai reçu le soutien moral et technique venant de plusieurs personnes. Ce qui m'a permis d'atteindre les objectifs qui m'ont été assignés.

J'aimerais remercier en premier lieu, mes très chers parents, qui se sont surpassés tout au long de leur vie pour nous offrir, une éducation exemplaire, un modèle de labeur et de persévérance.

Je tiens à exprimer également mon remerciement à mon tuteur Monsieur Thomas MOULIN, qui m'a accordé sa confiance afin de travailler à l'amélioration du scoring. Il m'a assisté tout le temps avec ses pistes qu'il me donne afin d'avancer.

Par ailleurs, j'adresse mes remerciements à mon encadrant Madame Rakia JAZIRI, Maître de Conférences à l'Université Paris 8 pour le temps qu'elle m'a accordé dans le suivi de mon stage.

Mes remerciements vont aussi à Aurissa TIV, Data Scientist dans la même équipe qui m'a assisté dans mes tâches quotidiennes. Je remercie aussi la Data Analyst, Lisa AÏT-MOULOUD qui m'a accompagnée depuis le début de mon stage chez Transaction Connect à ma montée en compétence sur les différents sujets.

Enfin, je remercie également toute l'équipe de Transaction Connect, et spécialement l'Equipe Data pour leur accueil, leur esprit d'équipe, ainsi que les conditions favorables dans lesquelles j'ai évolué au quotidien.

Résumé

Dans ce document de mémoire, j'ai effectué mon stage de validation du master 1 au sein de l'équipe Data de Transaction Connect. Cette dernière est une start-up spécialisée dans la transformation de tout moyen de paiement en cartes de fidélité en se basant sur les données d'achats des clients. Pendant mon parcours, j'ai participé à l'amélioration des moyens d'affectation des transactions aux magasins. L'un des sujets principaux sur lequel j'ai travaillé est la mise en place d'une nouvelle version de l'algorithme du scoring. La donnée intervenant dans l'amélioration est calculée à partir des données de transaction en base.

Dans le but de répondre efficacement à ce besoin j'ai d'un côté migré toutes les requêtes de calcul des features de la dataset d'entraînement de postgres vers Amazon Redshift. Ensuite, j'ai amélioré la détection de transaction des magasins en dehors de la France. Pour les centres commerciaux qui ont suffisamment de transaction, j'ai mis en place un procédé automatique pouvant créer un modèle propre à ces centres. Dans le but d'avoir des modèles aussi efficaces sur de nouvelles transactions je suis amené à automatiser le calcul de nouvelle dataset à chaque 30 jours.

Introduction Génarale

Dans le cadre de validation de mon master 1 mention Informatique parcours Big Data et Fouille de Données massives, j'ai intégré en tant que stagiaire la start-up Transaction Connect au poste de Data scientist et Data Analyst.
L'objectif principal de Transaction connect est d'un côté enrichir les transactions des clients en utilisant des techniques d'algorithme classique telle que l'application des patterns sur les transactions. D' un autre côté grâce au machine learning, en mettant en place des modèles de classification pouvant prédire les transactions qui sont faites dans le programme du client ou à l'extérieur.
Le plan de ce mémoire s'articulera autour de plusieurs points. Dans un premier temps, on présentera l'entreprise d'accueil, son activité principale et ses clients. Ensuite poser le problème que va traiter l'ensemble de ce document. Pour répondre au mieux à la problématique, on passera en revue l'état de l'art des méthodes qui existent afin d'envisager proposer la meilleure solution qui répondra au besoin. Dans le système réalisé on verra premièrement le calcul des features et la présentation du jeu de données autours duquel les modèles seront construits. Enfin on évaluera différents modèles appliqué à la dataset permettant de valider le modèle de classification optimal.

Première partie

Présentation de l'entreprise

Sommaire

Chapitre 1

Présentation de l'entreprise

Sommaire

1.1 L'histoire de Transaction Connect

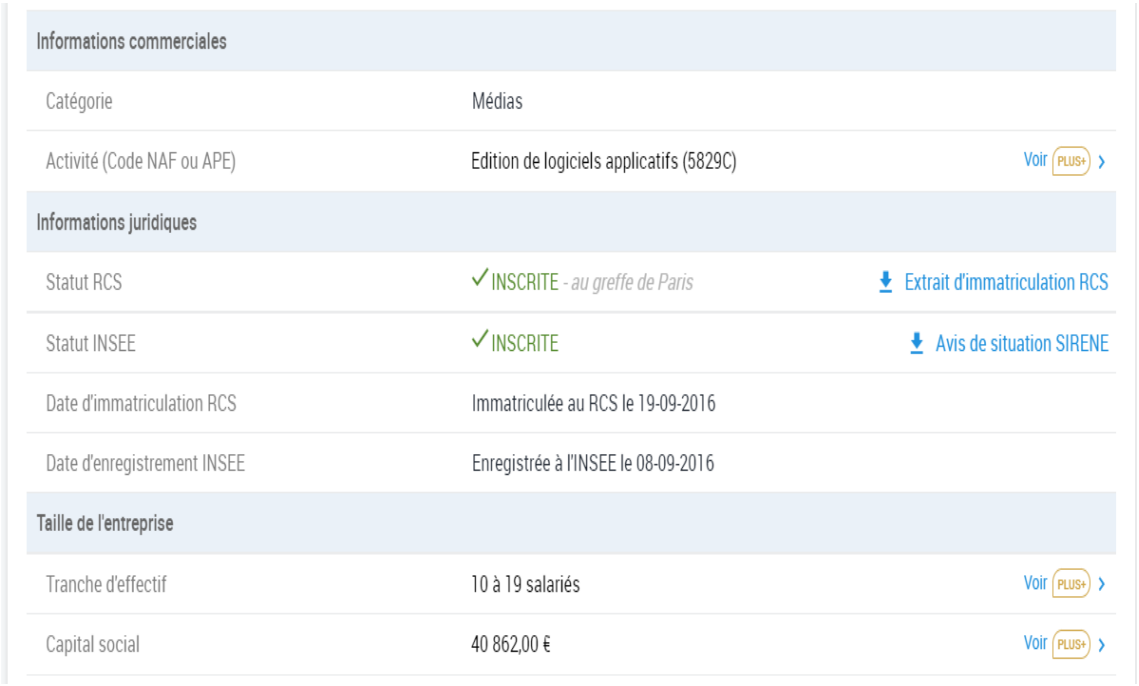
Transaction Connect est une start-up française fondée en 2016 par Didier GASTÉ après ses huit années passées chez le leader de l'immobilier commercial européen Unibail-Rodamco dans le secteur du digital. Il s'associe à son cofondateur Maxime Dellerie pour inventer une technologie permettant de transformer tout moyen de paiement électronique en une carte de fidélité. Avec un capital de 30.315€, Transaction Connect commence ses activités le 08 septembre 2016 de manière officielle et fixe son siège social au 86 rue du Faubourg Saint Denis 75010 Paris. En moins d'une année son capital passe à 31.575€.

Dans ses activités la start-up estime une perte de moitié des capitaux propres et est sous la menace de dissolution. Au terme de l'Assemblée Générale ordinaire du 28 juin 2019, il a été décidé de ne pas dissoudre la société bien que ses capitaux propres soient devenus inférieurs à la moitié du capital social. Aux termes du PV des Décisions du 25 juillet 2019, Le Président a constaté la réalisation d'une augmentation du capital social d'un montant de 9.287€, ce qui permet au capital de revenir à 40.862€ et qui prend effet au 25 juillet 2019. En décembre 2020, par la décision du président , Transaction Connect a vu son siège social transféré au 210 Quai de Jemmapes, Bureaux 2e étage 75010 et prend effet le 31 décembre 2020.

En quelques mois d'activités dans cette nouvelle adresse, suite à la décision du président fondateur, la start-up retourne à son ancienne adresse 86 rue du Faubourg Saint Denis 75010 Paris

1.2 Chiffres de Transaction Connect

Fig. 1.1: Chiffres sur actuels sur l'évolution de Transaction Connect



Informations commerciales		
Catégorie	Médias	
Activité (Code NAF ou APE)	Edition de logiciels applicatifs (5829C)	Voir PLUS+ ›
Informations juridiques		
Statut RCS	✓ INSCRITE - *au greffe de Paris*	Extrait d'immatriculation RCS
Statut INSEE	✓ INSCRITE	Avis de situation SIRENE
Date d'immatriculation RCS	Immatriculée au RCS le 19-09-2016	
Date d'enregistrement INSEE	Enregistrée à l'INSEE le 08-09-2016	
Taille de l'entreprise		
Tranche d'effectif	10 à 19 salariés	Voir PLUS+ ›
Capital social	40 862,00 €	Voir PLUS+ ›

FIG. 1.2: Informations sur la start-up Transaction Connect

1.3 Les Clients de Transaction Connect

1. Clients foncières B2B ; En 2017, Transaction Connect a signé son premier contrat avec le géant européen de l'immobilier commercial Unibail Rodamco en testant sa solution de la mise en relation d'un moyen de paiement à un programme de fidélité qui permet l'accès à des données de comportement d'achat. Par la suite, la start-up ne cesse de signer d'autres contrats et compte aujourd'hui une vingtaine de clients foncières en Europe notamment SOCRI, Nepi, Rockcastle, Hyper U, Eurocommercial, Accessite, CHEESE, ...

   Ces foncières bénéficient donc de deux modes d'intégration sur contrat :

   - Intégration en White Label
     Les clients Business inscrits en White Label bénéficient des services comme la page web de connexion, la synchronisation des comptes bancaires de leurs clients, l'enrichissement des données récupérées, la fourniture des dashboard analytics, la notification et envoie des récompenses.
   - Intégration en External Front

Pour les clients foncières qui optent à ce mode d'intégration Transaction connect fournit tous les services précédents sauf la page web de connexion et la notification de cashback. Dans le premier cas le foncier fait appelle à Transaction Connect à lui mettre en place la solution avec tous les services alors dans ce cas le client dispose de certains services en interne.

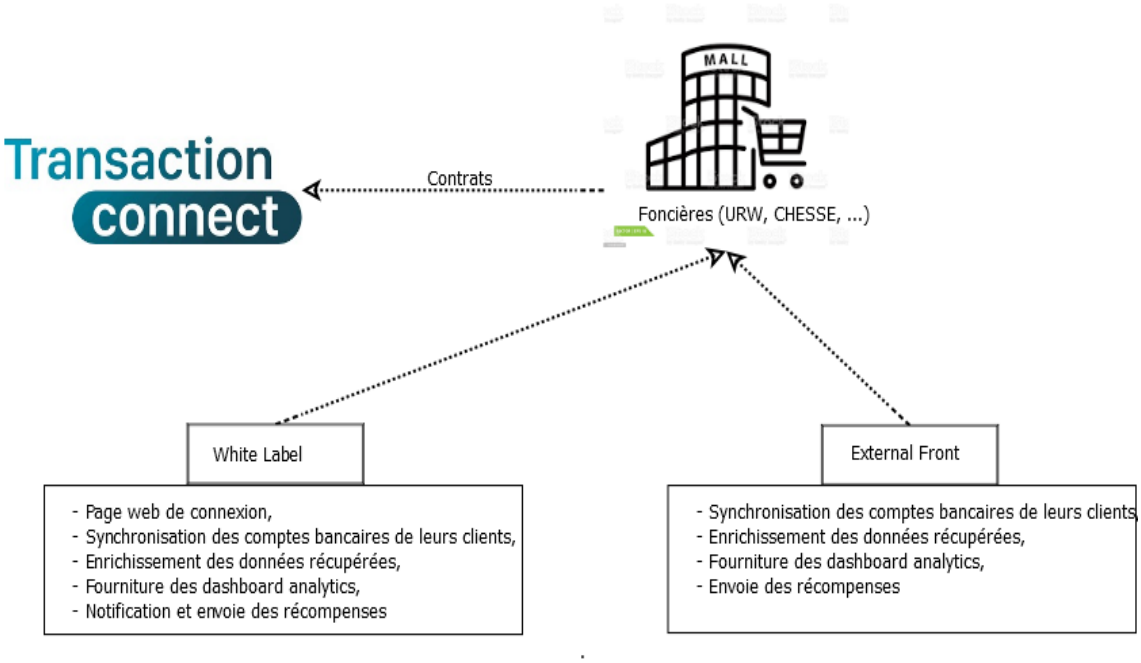


Fig. 1.3: Modes d'intégration des foncières

2. Clients acheteur B2C
   Les clients acheteurs sont la source génératrice de la matière première de transaction Connect. Ils s'inscrivent à un programme de fidélité qu' offre un Centre Commercial en donnant leur consentement sur le mode de connexion de leur transaction. Après avoir enrichi ces transactions en appliquant les solutions mise en place par Transaction Connect, les clients foncières s'en servent pour prendre des décisions et des recommandations. Les clients bénéficient au choix de trois modes de connexion des données d'achats.

   - Account Linking
     Les clients sous cette formule de connexion acceptent que Transaction Connect accède à l'historique de toutes ses transactions à travers leur Banque. En effet, Transaction ne reçois pas des informations personnelles du client, elle a accès à certaines informations intéressante notamment : un label de description du produit acheté Le nom du retail chez qui il a effectué l'achat, la date de transaction, le montant, … Exemple : "CB CARREFOUR 045 02/03/2021 10€"
   - Card Linking
     Dans cette formule de connexion, le client désire transmettre seulement les transactions effectuées dans les Centres dans lesquels ils sont inscrits.
   - Receipt Linking
     Pour les clients qui hésitent à donner accès à leur compte bancaire, ils ont la possibilité de scanner les tickets de caisse et de les envoyer. Transaction se charge d'extraire les informations pertinentes afin de le considérer comme une transaction.

Le client en contrepartie reçoit des récompenses après chaque montant dépensé dans le Centre dans lequel il est fidélisé. En Pologne par exemple, le client cumule des points pour les achats supérieurs à 30 PLN, reçoit 30 points pour chaque achat de ce type. Il se voit récompensé par virement sur son compte 30 PLN à chaque 300 points atteints.

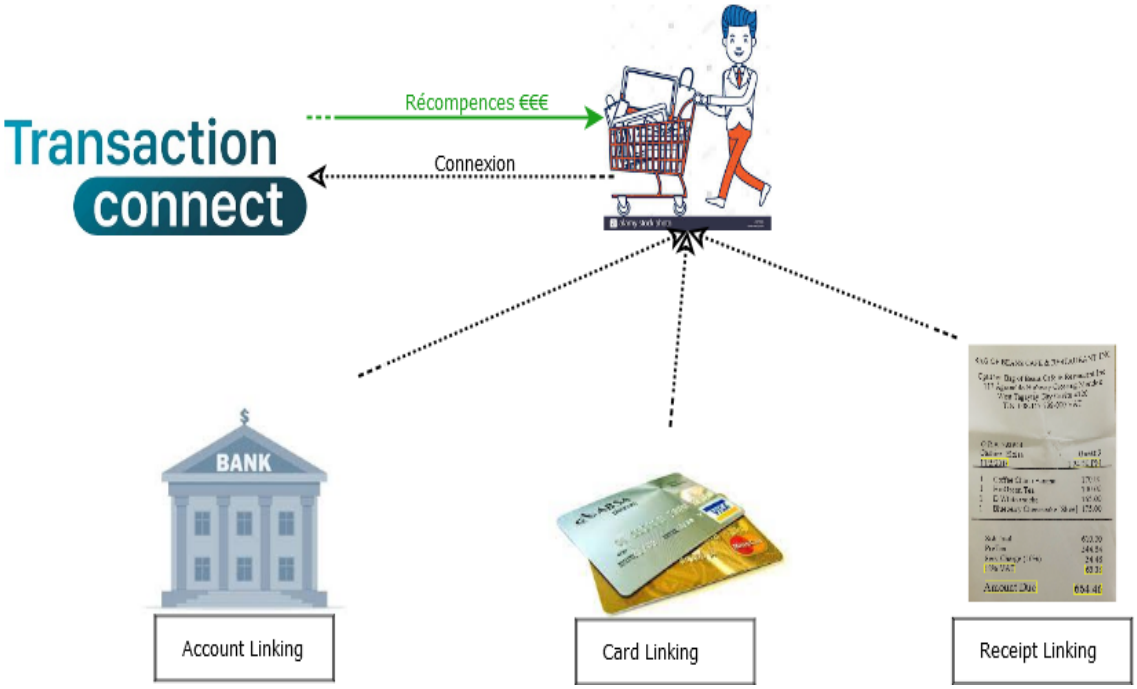


FIG. 1.4: Modes de connexion des Clients aux programmes de fidélités

1.4 Les services clés

Transaction Connect propose une solution révolutionnaire aux centres commerciaux qui remplacent les cartes de fidélité en se basant sur les données d'achats des clients. En rattachant un moyen de paiement à une entité, Transaction Connect permet aux acteurs du commerce physique, d'améliorer leur connaissance client et de proposer des expériences innovantes, fluides et personnalisées à leurs clients. Les services qu'offre Transaction Connect :

- Page web de connexion
  Pour les clients foncières en white label, Transaction Connect leur fournit une page web que vont utiliser leurs clients acheteurs afin de s'inscrire aux fidélités et de charger des tickets de caisse.

- Synchronisation des comptes bancaires de leurs clients :
  Transaction Connect intervient aussi dans la récupération des données de transactions auprès des fournisseurs (Fidel, Fintecture, Manual, Salt Edge Dsp2, Swedbank, Budget Insight) qui sont en contact avec les Banques.

- Enrichissement des données récupérées :
  Grâce aux différentes solutions de Transactions Connects les données sont enrichies afin de définir la source de la transaction notamment le Centre commercial hôte, le retailer qu'appartient le magasin d'achat et le magasin lui-même..

- Fourniture des dashboard analytics :
  Après que les données sont enrichies, on construit des indicateurs de performances et d'évolution afin de comprendre le comportement d'achat des clients.

- Notification et envoie des récompenses :
  Lorsque un client acheteur est éligible à percevoir une récompense, Transaction Connect envoie un message d'information du montant débloqué et envoie un virement du montant correspondant au client.

1.5 Transaction Connect en interne

1. L' équipe d'accueil :
   Tout d'abord, Transaction connect compte deux équipes techniques notamment l'équipe de Plateforme et l'équipe de Data. Tout mon parcours chez Transaction Connect s'est passé au sein de l'équipe data de composé de six membres. Au sein de l'équipe, on retrouve deux catégories de profils : les data scientistes et les data analystes. Les tâches analytiques sont subdivisées en deux : les tâches à destination des clients externes notamment les foncières qui s'en servent pour suivre la performance d'accueil des clients. La deuxième analyse gérée par un autre membre de l'équipe vise à monitorer les qualités d'affectation des transactions en interne. Les Data scientist interviennent dans la mise en place des algorithmes de machine learning permettant d'enrichir des transactions.

2. Organisation de travail : le scrum Agile
   Scrum est une méthode agile pour la gestion de projet informatique et a pour objectif d'améliorer la productivité d'une équipe. C'est un cadre de travail au sein duquel les acteurs peuvent aborder des problèmes complexes et adaptatifs, en livrant de manière efficace et créative des produits tout en créant de la valeur ajoutée. Le sprint dure deux semaines et sur cette période chacun travail sur une ou plusieurs tâches dans le but de fournir un résultat en fin de sprint.

3. Outils techniques :
   Dans le but de réaliser mes missions au sein de Transaction Connect, cette dernière a mis à ma disposition un ordinateur portable. Comme outils technique ou de communication on dispose entre autres de DataGrip, Pycharm, Slack, Dashlane, VPN,...

Deuxième partie

Problématique

Sommaire

Chapitre 2

Le contexte de résolution du problème

Sommaire

2.1 Introduction

La matière première de Transaction Connect est la donnée de transaction que génèrent les clients en faisant les achats dans les magasins. Ces données brutes sont soumises à des traitements afin de ressortir d'autres informations importantes notamment le centre commercial visité, et le magasin d'achat.


```
{
   "id": "cb89d0c5b843a2659fc013a85f7cb96f78c3c0e9d1f10577ccd66b70903f458d",
   "type": "transactions",
   "attributes": {
       "transaction_id": "17d65f5e-1e3c-4611-9c3e-7a7ba0748d80",
       "booking_date": "23/04/2021",
       "value_date": "23/04/2021",
       "amount": "-150",
       "currency": "EUR",
       "communication": "CARTE\n0533625\nCB  CARREFOUR  22/04/21\nCBLM PASCAL
FUHRY\n\n\n",
       "status": "BOOKED",
       "customer_id": "2974f09a-fefd-4db1-b18e-1037491f5dec",
       "credit_debit": "DEBIT"
   }
}
```


date	détaillant	montant	customer_id	magasin
22/04/21	CARREFOUR	150 €	2974f09a-fefd-4db1-b18e-1037491f5dec	??

FIG. 2.1: Modes de connexion des Clients aux programmes de fidélités

Après avoir extrait les informations nécessaires, l'objectif maintenant est d'identifier l'origine de la transaction, le magasin de provenance de cette transaction. La réponse à cette question nous permet d'identifier si ce magasin fait partie du centre commercial où le client est fidélisé. Dans le cas affirmatif, le client gagne des points correspondant au programme de fidélité de ce Centre commercial.

2.2 Le problème à résoudre

En interne, il existe plusieurs méthodes d'affectation des transactions tant en machine learning que algorithmique. Les principaux algorithmes d'affectation des transactions :

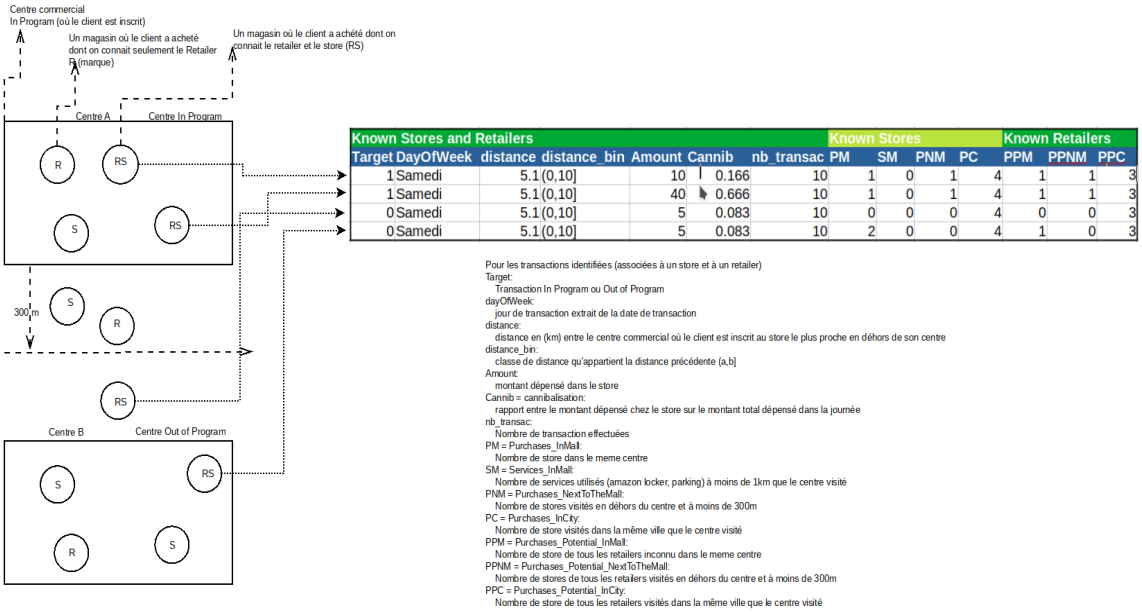
- Le store locator
- Alpha / Alpha City
- Patterns regex
- Le Scoring

Le problème à résoudre durant mon parcours en tant que stagiaire sera autour de la classification et l'affectation aux magasins des transactions en utilisant l'algorithme du scoring. Le scoring est le quatrième algorithme de Transaction Connect permettant l'affectation des transactions remontant avec moins d'informations dans les labels qui peuvent permettre l'identification de l'origine de transaction. C'est une méthode de machine learning basée sur la probabilité visant à analyser le comportement d'achat des clients dans une journée afin de retracer une transaction inconnue à partir d'autres transactions identifiables dans la même journée. L'objectif final sera dans un premier temps d'améliorer la qualité d'affectation des transactions grâce au Scoring. Dans le but d'améliorer l'ancienne version du scoring, je serai amené à :

- Procéder au feature engineering en faisant la migration des requêtes de Postgresql vers Redshift dans le but de diminuer la durée des calculs.
- Mettre en place des modèles de scoring spécifique d'affectation des transactions de certains grands centres commerciaux
- Automatiser le choix du modèle idéal correspondant en se basant sur les métriques de test.
- Automatiser la construction de nouveaux modèles en prenant en compte des données récentes sur une période mensuelle.

2.3 Features engineering (Scoring des transactions)

Dans le schéma ci-dessous on peut remarquer deux centres commerciaux (Centre A et Centre B) à gauche. Le centre A est le centre commercial où le client est inscrit,



Known Stores and Retailers						Known Stores				Known Retailers		
Target	DayOfWeek	distance	distance_bin	Amount	Cannib	nb_transac	PM	SM	PNM	PC	PPM	PPNM	PPC
1	Samedi	5.1	(0,10]	10	0.166	10	1	0	1	4	1	1	3
1	Samedi	5.1	(0,10]	40	0.666	10	1	0	1	4	1	1	3
0	Samedi	5.1	(0,10]	5	0.083	10	0	0	0	4	0	0	3
0	Samedi	5.1	(0,10]	5	0.083	10	2	0	0	4	1	0	3

Pour les transactions identifiées (associées à un store et à un retailer)
Target:
Transaction In Program ou Out of Program
dayOfWeek:
jour de transaction extrait de la date de transaction
distance:
distance en (km) entre le centre commercial où le client est inscrit au store le plus proche en déhors de son centre
distance_bin:
classe de distance qu'appartient la distance précédente (a,b]
Amount:
montant dépensé dans le store
Cannib = cannibalisation:
rapport entre le montant dépensé chez le store sur le montant total dépensé dans la journée
nb_transac:
Nombre de transaction effectuées
PM = Purchases_InMall:
Nombre de store dans le meme centre
SM = Services_InMall:
Nombre de services utilisés (amazon locker, parking) à moins de 1km que le centre visité
PNM = Purchases_NextToTheMall:
Nombre de stores visités en déhors du centre et à moins de 300m
PC = Purchases_InCity:
Nombre de store visités dans la même ville que le centre visité
PPM = Purchases_Potential_InMall:
Nombre de store de tous les retailers inconnu dans le meme centre
PPNM = Purchases_Potential_NextToTheMall:
Nombre de stores de tous les retailers visités en déhors du centre et à moins de 300m
PPC = Purchases_Potential_InCity:
Nombre de store de tous les retailers visités dans la même ville que le centre visité

FIG. 2.2: Systeme de calcule des variables

en d'autre terme le centre A est "In program". Les petits cercles sont les magasins visités par un client dans la même journée : quatre magasins dans le centre In program dont deux sont identifiables (R : détaillant, S : magasin) et deux autres dont on connaît seulement respectivement le détaillant (R) et le magasin (S). Tous les autres magasins dans le centre B et à l'extérieur sont considérés comme "out of program".

La dataset entrant dans l'apprentissage de notre modèle est construite en appliquant le scoring sur les transactions effectuées dans la même journée dont le détaillant et le magasin sont identifiables.

On charge seulement les transactions des clients qui sont inscrits dans un programme (centre commercial) donné ex : Les 4 Temps, Okabe,...

- Target :
  Une transaction est in-programme si la transaction est effectuée dans un magasin du centre où le client est inscrit.

- distance :
  Calculer la distance entre le centre où le client est inscrit aux autres magasins qu'il a aussi visité et considérer celui qui est proche du centre.

- distance-bin :
  Classer la distance du magasin le plus proche dans une classe de distance €(a,b]
- DayOfWeek :
  Jour de la transaction extrait de la date de la transaction.
- Amount :
  Montant dépensé dans le magasin
- Cannibalisation :
  Poids du détaillant dans le centre : Rapport entre le montant dépensé par les clients chez le détaillant sur le montant total dépensé dans le centre sur une période.
- Nb-transaction :
  Quantité totale d'articles achetés par le client dans la même journée.
- ServicesInMall :
  Nombre de services utilisés par le client dans la journée (amazon locker, parking) à moins de un kilomètre du centre commercial
- Purchases-InMall :
  Nombre de magasin visité dans le centre dans la journée
- Purchases-NextToTheMall :
  Nombre de magasins visités en dehors du centre et à moins de 300 mètres.
- Purchases-InCity :
  Nombre de magasins visités dans la même ville que le centre visité.
- Purchases-Potential-InMall :
  Nombre de magasins de tous les détaillants non identifiés que le client a visité dans le centre.
- Purchases-Potential-NextToTheMall :
  Nombre de magasins de tous les détaillants que le client a visité en dehors du centre et à moins de 300 mètres.
- Purchases-Potential-InCity :
  Nombre de magasins de tous les détaillants non identifiés visités dans la même ville que le centre.

Le calcul de ces variables étant coûteux en temps d'exécution, dans mon travail il a fallu migrer toutes les requêtes de calcul Postgres et certaines opérations dataframes Pandas vers Amazon Redshift.

Classification et affectation des transactions aux magasins des détaillants

2.4 Présentation des données

2.4.1 Dataset :

Après calcul sur l'ensemble des transactions en base le jeu de données comporte 483725 lignes sur 14 variables.

	target	DayOfWeek	amount	distance	distance_bin	nb_transac	Purchases_InMall	Purchases_NextToTheMall	Services_InMall	Purchases_InCity	Purchase
0	1	Monday	39.90	294.316896	(200.0, 500.0]	4	1	1	0	0	
1	1	Monday	64.94	5.605244	(5.0, 10.0]	4	1	1	0	1	
2	1	Monday	35.98	294.114634	(200.0, 500.0]	2	1	1	0	1	
3	1	Monday	75.90	5.458404	(5.0, 10.0]	2	1	1	0	1	
4	1	Saturday	118.96	5.627760	(5.0, 10.0]	6	4	4	0	4	
5	1	Saturday	22.90	5.433361	(5.0, 10.0]	6	4	5	0	4	
6	1	Saturday	10.00	294.203458	(200.0, 500.0]	6	4	4	0	3	
7	1	Saturday	29.00	5.600936	(5.0, 10.0]	6	4	4	0	3	
8	1	Saturday	16.90	5.458404	(5.0, 10.0]	6	4	4	0	3	
9	1	Saturday	30.00	294.206739	(200.0, 500.0]	5	3	2	0	2	
10	1	Saturday	13.98	473.669856	(200.0, 500.0]	5	3	2	0	3	
11	1	Saturday	13.98	294.252744	(200.0, 500.0]	5	3	2	0	2	
12	1	Saturday	18.00	294.247216	(200.0, 500.0]	5	3	3	0	2	
13	0	Saturday	19.00	5.620421	(5.0, 10.0]	5	4	3	0	3	
14	1	Thursday	39.98	294.152306	(200.0, 500.0]	3	2	2	0	2	

FIG. 2.3: Jeu de donnée

	Purchases_Potential_InMall	Purchases_Potential_InCity	Purchases_Potential_NextToTheMall	cannibalisation
	0.0	0.0	0.0	0.593874
	0.0	0.0	0.0	0.418474
	0.0	0.0	0.0	0.484127
	0.0	0.0	0.0	0.376787
	0.0	0.0	0.0	0.070770
	0.0	0.0	0.0	0.355689
	0.0	0.0	0.0	0.569227
	0.0	0.0	0.0	0.222634
	0.0	0.0	0.0	0.375748
	0.0	0.0	0.0	0.379135
	0.0	0.0	0.0	0.661259
	0.0	0.0	0.0	0.252171
	0.0	0.0	0.0	0.934046
	0.0	0.0	0.0	0.210076
	0.0	0.0	0.0	0.528370

FIG. 2.4: Jeu de donnée

Troisième partie

État de l'art

Sommaire

Chapitre 3

État de l'art des techniques

Sommaire

3.1 État de l'art des techniques classiques

Pour pouvoir affecter une transaction à un magasin ou identifier son origine, il existe plusieurs procédés permettant de résoudre ce problème. Pour cela nous ferons le parcours des meilleurs algorithmes utilisés comme Store locator, Alpha / Alpha City et Patterns regex.

3.1.1 Store Locator :

- Contexte :
  Store Locator vise à enrichir les transactions qui respectent les critères suivantes :
  - Présence du nom du magasin dans lequel la transaction a été effectuée (identifiant du magasin). L'idée derrière cet algorithme est que si nous savons qu'il n'y a qu'un seul magasin d'un détaillant dans une zone géographique et si nous sommes capables de détecter cette zone géographique et que le nom du détaillant se trouve dans le label de transaction ("CB CARREFOUR 045 02/03/2021 10€"), nous pouvons l'attribuer à ce magasin spécifique.
  - Présence de la ville dans laquelle la transaction a été effectuée (identifiant de la ville).
  - S'il n'a pas été possible d'associer la transaction à un magasin, déterminer s'il s'agit d'une transaction hors centre commercial, c'est-à-dire hors du centre commercial dans lequel le client est enregistré.

- Notion de zone géographique :
  Tout d'abord on crée toutes les zones géographiques de manière hiérarchique sous forme d'un arbre n-aire avec une relation de parent à fils et à la racine le monde. Le centre commercial est placé comme étant fils de la ville (ex : Les 4 Temps dans la ville de Puteaux). Cette configuration permet de faire un parcours de manière hiérarchique en base à la recherche de la zone géographique correspondante à une transaction.

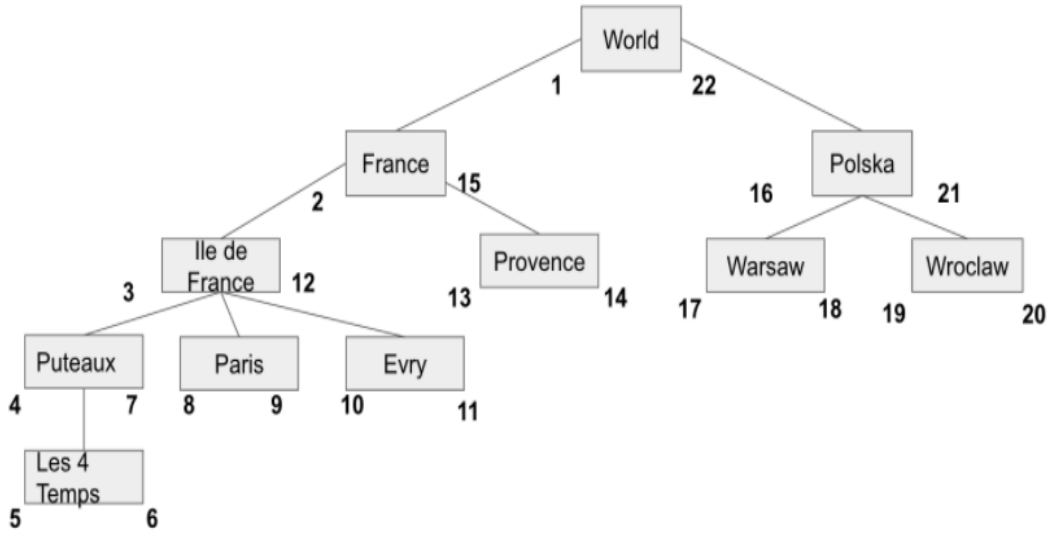


FIG. 3.1: Arbre des zones géographiques

- Affectation ou correspondance :
  La première étape consiste à trouver si les transactions ont un pattern correspondant.

- Correspondance des zones géographiques :
  Pour attribuer une transaction à un magasin ou à une ville, nous devons d'abord connaître sa zone géographique. Pour chaque pattern de zone géographique en base, on recherche les transactions qui correspondent à ce pattern.
  - Affectation de la ville et du magasin :
  Une fois la correspondance effectuée, nous pouvons procéder aux affectations. Lorsqu'on n'est pas en mesure d'associer la transaction à un magasin, on veut au moins savoir si elle a été effectuée en dehors du centre commercial. On rappelle qu' une transaction peut potentiellement être considérée comme hors centre commercial en fonction des valeurs du détaillant et de la zone géographique dans la matrice ci-dessous.

Zone geo / Retailer	Dans le Centre	Hors du Centre	Null
Dans le Centre	Potentiellement Dans le Centre	Hors du Centre	Potentiellement Dans le Centre
Hors du Centre	Hors du Centre	Hors du Centre	Hors du Centre
Null	Hors du Centre	Hors du Centre	Hors du Centre

FIG. 3.2: Arbre des zones géographiques

Les transactions passent par différents processus en fonction de leur zone géographique et de la valeur des détaillants :

– Transactions en bleu :
Pour les transactions dont leur zone géographique est non nul on garde celle qui matche avec des patterns de zone géographique - On regroupe les transactions suivant leur patterns de zone géographique afin d'avoir pour chaque ensemble la zone géographique liés à l'ensemble des transactions qu'ils ont appariées.
- Pour chaque ensemble de zones géographiques, nous voulons d'abord réduire l'ensemble des zones géographiques potentielles aux plus discriminantes si l'ensemble des zones géographiques potentielles est composé de plusieurs zones. Les zones géographiques sont organisées sous forme de hiérarchie. Ainsi, si une zone géographique contient une autre zone géographique, nous conservons la zone géographique au niveau le plus bas. À ce stade, il peut encore y avoir plusieurs zones géographiques potentielles.
- Après cette première présélection, nous choisissons les magasins qui se trouvent dans l'ensemble réduit de zones géographiques.
- Une fois que nous avons la liste des magasins potentiels, nous voulons savoir si nous pouvons attribuer la transaction à un magasin unique.

Pour chaque zone géographique de la liste des magasins potentiels, nous regardons s'il n'y a qu'un seul magasin dans la zone géographique. Si à la fin il n'y a qu'un seul magasin, nous pouvons attribuer nos transactions à ce magasin unique et à la ville correspondant à ce magasin.

Si nous ne sommes pas en mesure d'attribuer nos transactions à un magasin, nous essayons de l'attribuer à une ville. Le processus est le même que pour les magasins mais nous regardons dans la liste des zones géographiques potentielles.

Pour les transactions que nous n'avons pas réussi à attribuer à un magasin, nous voulons au moins préciser si elles étaient en dehors du centre commercial.

– Transaction en jaune :
Ces transactions n'ont pas de zone géographique connue mais ont un détaillant dans le centre commercial connu. Elles peuvent être potentiellement dans le centre.

– Transactions en rouge :
Ces transactions n'ont pas de zone géographique connue et le détaillant n'est non plus connu ou sont hors du centre où le client est enregistré. On peut conclure que ces transactions sont effectuées en dehors du centre du client.

3.1.2 Alpha :

Alpha est un algorithme qui prend en entrée le label de la transaction, effectue une requête à l'API des lieux de Google, et si l'API renvoie un seul résultat et si ce résultat est de type détaillant, il attribue le magasin à la transaction. Afin d'éviter de créer des doublons de magasins, nous utilisons l'identifiant de google appelé google-place-id (ex : ChIJo82pu26q2EcRRvXT8RUamHo) retourné.
De plus, afin d'éviter de faire la même requête à google à chaque fois, lorsque la requête a été postée, elle est stockée dans une table en base. Pour chaque requête, on vérifie d'abord dans cette table pour éviter tout coût supplémentaire.

3.1.3 Alpha City :

Alpha city est une évolution d'alpha où au lieu d'interroger le label de transaction à l'API google de lieux, la requête est de la forme nom-détaillant + ville (donc on doit avoir détecté une ville et un détaillant dans le label pour pouvoir faire une requête). A part cela, le mécanisme est exactement le même que celui d'Alpha.

Détection de la ville : Afin de détecter la ville dans le label, une façon de procéder est d'utiliser des mots-clés. Cette méthode a l'avantage d'être très précise si les mots-clés sont bien conçus mais l'inconvénient est de ne pas être évolutive du tout : si on veut détecter n'importe quelle ville dans le monde, on doit déclarer des mots-clés pour chaque ville. C'est pourquoi les mots-clés sont utilisés pour la localisation des magasins (lorsque le but est d'attribuer une transaction à l'intérieur du centre et que toute erreur peut conduire à une récompense cliente inappropriée).

Pour la détection de masse des villes, la méthode consiste à calculer la distance de Levenshtein entre le label des transactions et les villes. De manière informelle, la distance de Levenshtein entre deux mots est le nombre minimum de modifications d'un seul caractère (insertions, suppressions ou substitutions) nécessaires pour transformer un mot en l'autre. Si la distance entre le label et la ville est suffisamment faible, l'algorithme attribue la ville à la transaction.

La distance de Levenshtein entre deux cordes a, b un B (de longueur $|a|$ et $|b|$ respectivement) est donnée par $lev(a, b)$ où

$$\mathrm{lev}(a,b) = \begin{cases} |a| & \text{if } |b| = 0, \\ |b| & \text{if } |a| = 0, \\ \mathrm{lev}(\mathrm{tail}(a), \mathrm{tail}(b)) & \text{if } a[0] = b[0] \\ 1 + \min \begin{cases} \mathrm{lev}(\mathrm{tail}(a), b) \\ \mathrm{lev}(a, \mathrm{tail}(b)) \\ \mathrm{lev}(\mathrm{tail}(a), \mathrm{tail}(b)) \end{cases} & \text{otherwise.} \end{cases}$$

3.2 Etat de l'art des algorithmes de machine learning

3.2.1 Méthode de Support Vector Machine SVM :

D'après l'article Proposé par Boser, Guyon, et Vapnik en 1992, SVM est une technique de classification linéaire et non linéaire. SVM est un modèle d'apprentissage supervisé basé sur la détermination d'un hyperplan qui sépare les données d'une classe des autres classes.

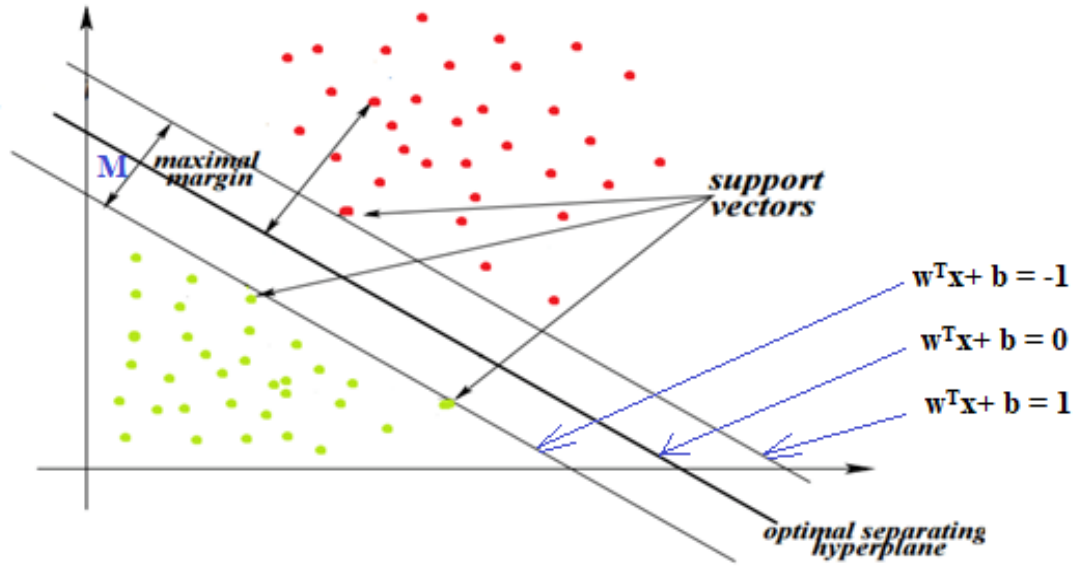


Fig. 3.3: Hyperplan de séparaton des points de données

Soit x_0 et x_1 deux vecteurs supports aux deux extrémités et

$$Soit\ l'hyperplant\ (P): w^T x + b = 0, \quad w \in R^p, \quad b \in R$$

$$Soit\ f(x) = w^T x + b$$

$$y_i \in \{-1,1\}, \quad i \in \{0,n\}$$

$$\begin{cases} if\ y_i = 1 => w^T x + b \geq 1\ (en\ haut) \\ if\ y_i = -1 => w^T x + b \leq -1\ (en\ dessous) \end{cases} => y_i(w^T x + b) \geq 1$$

$$M = d(x_0, P) + d(x_1, P) = \frac{|w^T x_0 + b|}{\|w\|} + \frac{|w^T x_1 + b|}{\|w\|} = \frac{|1|}{\|w\|} + \frac{|-1|}{\|w\|} = \frac{2}{\|w\|}$$

$$\begin{cases} Max\ (M) = Max(\frac{2}{\|w\|}) \\ on\ y_i(w^T x + b) \geq 1 \end{cases} => \begin{cases} Min\ (\|w\|) \\ on\ y_i f(x) \geq 1 \end{cases}$$

Coût minimum de la fonction

$$J = \gamma\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} Max(0, 1 - y_i(w^T x + b)) = \gamma\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} Max(0, 1 - y_i f(x))$$

$$Max(0, 1 - y_i f(x)) = \begin{cases} 0 & if\ y_i f(x) \geq 1 \\ 1 - y_i f(x) & else \end{cases}$$

Application de la descente de Gradient :

if $y_i f(x) \geq 1$:

$$\frac{dJ_i}{dw_k} = 2\gamma w_k\ ; \qquad \frac{dJ_i}{db} = 0$$

Sinon :

$$\frac{dJ_i}{dw_k} = 2\gamma w_k - y_i . x_i; \qquad \frac{dJ_i}{db} = y_i$$

Mise à jour des paramètres :

$$\begin{cases} w = w - \gamma \frac{dJ_i}{dw_k} \\ b = b - \gamma \frac{dJ_i}{db} \end{cases}$$

3.2.2 Méthode du Decision tree :

L'apprentissage par arbre de décision ou induction d'arbres de décision est l'une des approches de modélisation prédictive utilisées en statistique, en exploration de données et en apprentissage automatique. Elle utilise un arbre de décision pour passer des observations sur un élément aux conclusions sur la valeur cible de l'élément. Considérons l'exemple dans lequel on souhaite construire l'arbre de décision sur une donnée pouvant prévoir le moment idéal de sortir en fonction de la météo. Pour chaque nœud , la règle consiste à trouver la meilleure colonne (ex : Time) et le séparateur idéal sur cette colonne (ex : Time > 10 ?)

On choisit le meilleur feature et le meilleur threshold en se basant sur le calcul du gain d'information à partir de l'entropie ou le gini de l'information. On construit donc les fils du nœud en faisant la séparation à partir de la meilleure coupure (Best

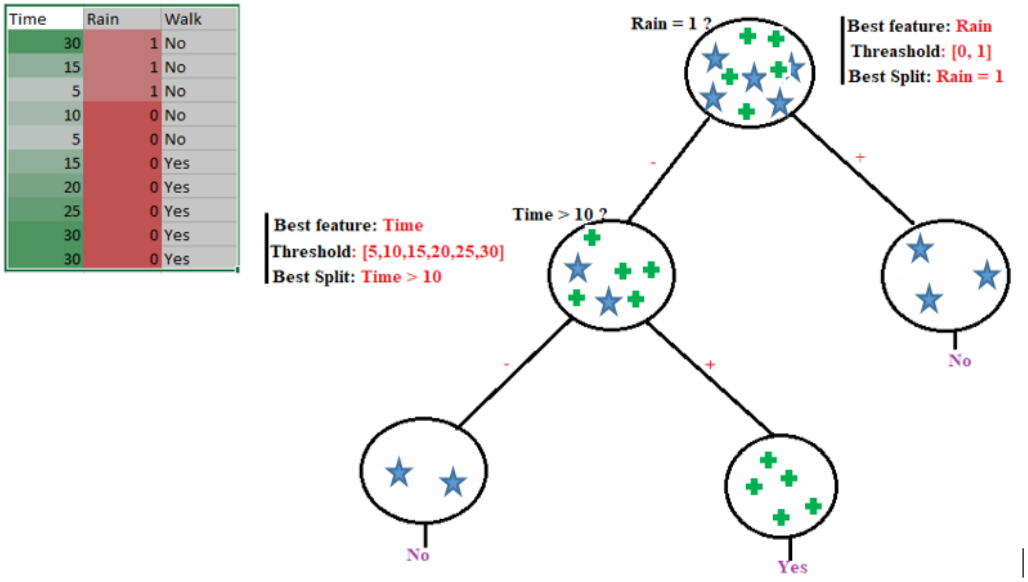


FIG. 3.4: Arbre de décision

Split). Au bout de la profondeur de l'arbre fixée, la décision dans chaque feuille est la classe avec une majorité de données.

Application de l'Entropie:

$$X: \text{label} \in [\text{"Yes","No"}]$$

$$E = -\sum_{i=0}^{nb\_labels} P(X) * log_2(P(X)) \text{ and}$$

$$P(X) = \frac{Label\_i\_population\_in\_node}{Total\_population}$$

Dans cet exemple : $E = -\frac{5}{10} log_2\left(\frac{5}{10}\right) - \frac{5}{10} log_2\left(\frac{5}{10}\right) = 1$

Calcul du gain d'Information :

$$IG = E(Parent) - \sum_{i=0}^{nb\_childs} \frac{Total\_population\_in\_node}{Total\_population} E(childs)$$

Dans cet exemple : $IG = E(Parent) - [\frac{7}{10} * E(child1) + \frac{3}{10} * E(child2)]$

Application du Gini:

$$X: \text{label} \in [\text{"Yes","No"}]$$

$$G = 1 - \sum_{i=0}^{nb\_labels} P(X)^2 \text{ and}$$

$$P(X) = \frac{Label\_i\_population\_in\_node}{Total\_population}$$

Calcul du gain d'Information :

$$IG = G(Parent) - \sum_{i=0}^{nb\_childs} \frac{Total\_population\_in\_node}{Total\_population} G(childs)$$

3.2.3 Méthode du Random Forest :

Le random forest un algorithme de classification supervisé basé sur la combinaison de plusieurs arbres N de décision appelés Décision Tree. Comme définie dans [3] chaque arbre décisionnel est réalisé en sélectionnant aléatoirement les données parmi les données disponibles. Par exemple, une forêt aléatoire pour chaque arbre décisionnel peut être construite par l'échantillonnage aléatoire d'un sous-ensemble de caractéristiques, et/ou par l'échantillonnage aléatoire d'un sous-ensemble de données d'entraînement pour chaque arbre décisionnel. Pour avoir la décision

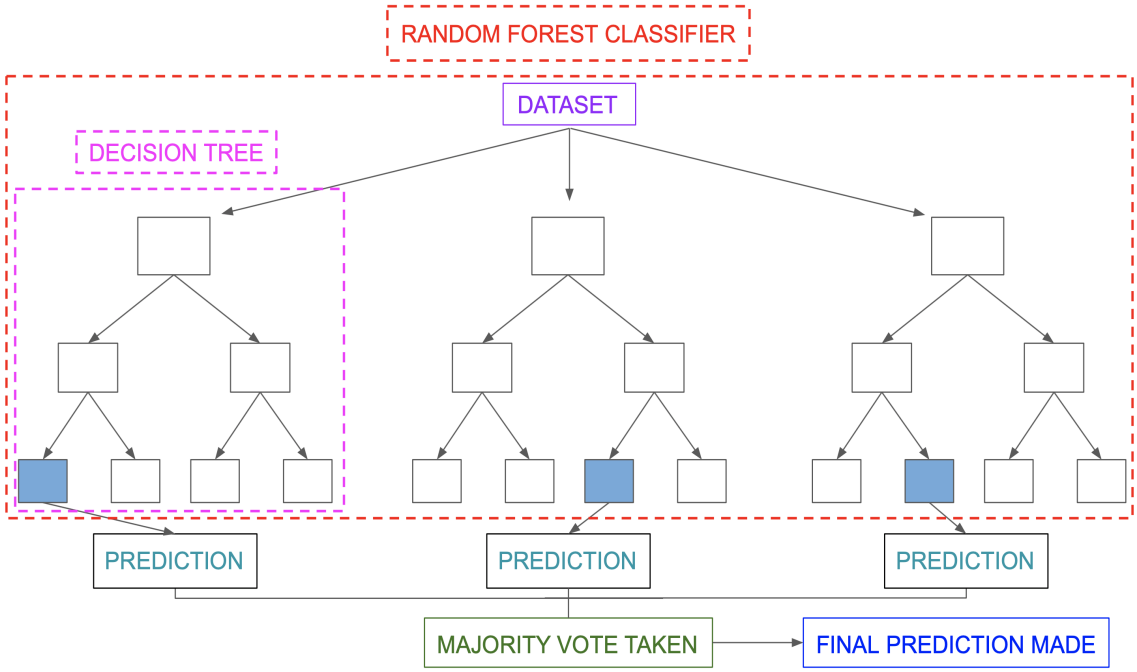


Fig. 3.5: Foret aléatoire

finale, on calcule la moyenne des décisions prises par l'ensemble des arbres de décision de la forêt.

3.2.4 K- Nearest Neighbor

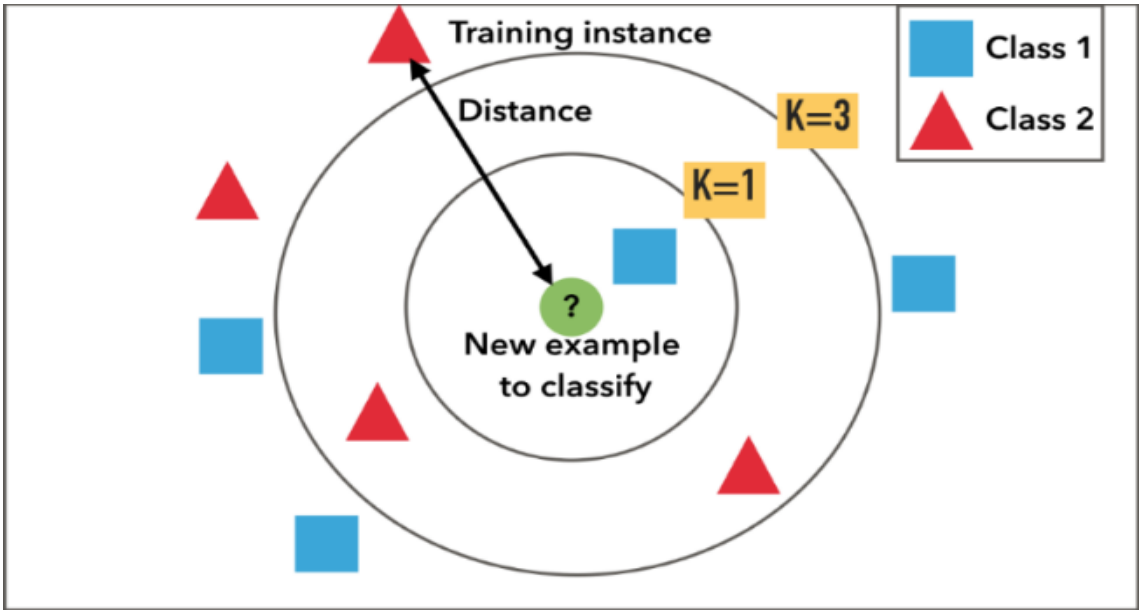


FIG. 3.6: K- Nearest Neighbor

1. Pricipe du K- Nearest Neighbor :
   Le K-Nearest Neighbor est un algorithme de machine learning d'apprentissage supervisé qui a pour principe de déterminer les K premiers points de données les plus proches en termes de distance par rapport à un point de donnée.

2. Les types de distances :
   On distingue différentes sortes de distances applicable au KNN :

   - Distance euclidienne

$$d(A, X) = \sqrt{\sum_{i=1}^{n}(a_i - x_i)^2}$$

   - Distance de Manhattan

$$d(A, X) = \sum_{i=1}^{n} \mid a_i - x_i \mid$$

- Distance de Minkowski

$$d(A, X) = \sqrt[p]{\sum_{i=1}^{n} \mid a_i - x_i \mid^p}$$

3. Choix du paramètre K :

   - Utilisation de K Le choix du paramètre K peut être effectué en prenant la partie entière de la racine carré du nombre de jeu de donnée

$$K = \sqrt{nombre - de - donnees}$$

   - Choisir K suivant celui qui donne une meilleure prédiction :
     La meilleure valeur de K est aussi déterminée en faisant des tests sur différentes valeurs de K.

Quatrième partie

Système réalisé

Sommaire

Chapitre 4

Implémentation du système

Sommaire

4.1 Introduction

Les transactions effectuées par les clients ont des informations importantes contenues dans les “label de transaction”. Par exemple, une transaction peut ressortir avec un label : “PAIEMENT PAR CARTE CARREFOUR DAC VL AMIENS 05/01” qui contient les informations comme le nom du détaillant Carrefour et la ville AMIENS. Par conséquent, si on sait qu'il n'y a qu'un seul carrefour à Amiens, grâce à l'algorithme de store locator, détaillé plus haut, cette transaction est identifiée. Maintenant dans le cas où le label de la transaction ne contient aucune information pertinente (ex : ”59091383100 §§”), on aimerait quand même savoir l'origine de la transaction. L'algorithme du Scoring nous permet donc d'affecter les transactions en cas de manque d'informations dans le label de transaction. L'algorithme analyse le comportement d'achat des clients au cours de la même journée pour décider si les transactions étaient dans le programme ou non. Enfin, à partir des algorithmes de machine learning on pourra appliquer des modèles capables d'affecter ces types de transaction au magasin correspondant.

4.2 Analyse de données

4.2.1 Informations sur la taille et les variables du jeu de donnée :

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 483725 entries, 0 to 483724
Data columns (total 14 columns):
target                              483725 non-null int64
DayOfWeek                           483725 non-null object
amount                              483725 non-null float64
distance                            483725 non-null float64
distance_bin                        483725 non-null object
nb_transac                          483725 non-null int64
Purchases_InMall                    483725 non-null int64
Purchases_NextToTheMall             483725 non-null int64
Services_InMall                     483725 non-null int64
Purchases_InCity                    483725 non-null int64
Purchases_Potential_InMall          483725 non-null float64
Purchases_Potential_InCity          483725 non-null float64
Purchases_Potential_NextToTheMall   483725 non-null float64
cannibalisation                     483725 non-null float64
dtypes: float64(6), int64(6), object(2)
memory usage: 51.7+ MB
```

Fig. 4.1: Informations sur la taille du jeu de donnée
On distingue 3 types de variable dont 6 variables entiers, 6 float et 2 variables catégorielles.

4.2.2 Statistiques sur la dataset :

	target	amount	distance	nb_transac	Purchases_InMall	Purchases_NextToTheMall	Services_InMall	Purchases_InCity	Purchases_
count	483725.000000	483725.000000	483725.000000	483725.000000	483725.000000	483725.000000	483725.000000	483725.000000	
mean	0.760469	33.524185	1213.432949	6.976536	1.538781	1.179689	0.000420	1.141192	
std	0.426798	74.058961	3130.164561	8.403424	1.882426	1.675003	0.020481	1.587473	
min	0.000000	0.002181	0.082108	2.000000	0.000000	0.000000	0.000000	0.000000	
25%	1.000000	9.950000	9.120183	4.000000	0.000000	0.000000	0.000000	0.000000	
50%	1.000000	19.990000	13.363057	5.000000	1.000000	1.000000	0.000000	1.000000	
75%	1.000000	39.580000	294.114634	8.000000	2.000000	2.000000	0.000000	2.000000	
max	1.000000	18449.092344	10000.000000	301.000000	35.000000	32.000000	1.000000	27.000000	

FIG. 4.2: Statistiques sur le jeu de donnée

Purchases_Potential_InMall	Purchases_Potential_InCity	Purchases_Potential_NextToTheMall	cannibalisation
483725.000000	483725.000000	483725.000000	483725.000000
0.515351	1.393953	0.499461	0.389588
1.587876	7.013884	1.907540	0.396687
0.000000	0.000000	0.000000	-73.081067
0.000000	0.000000	0.000000	0.200696
0.000000	0.000000	0.000000	0.382113
1.000000	1.000000	0.000000	0.543593
83.000000	558.000000	111.000000	103.161466

FIG. 4.3: Statistiques sur le jeu de donnée

4.2.3 Corrélation entre les variables numériques :

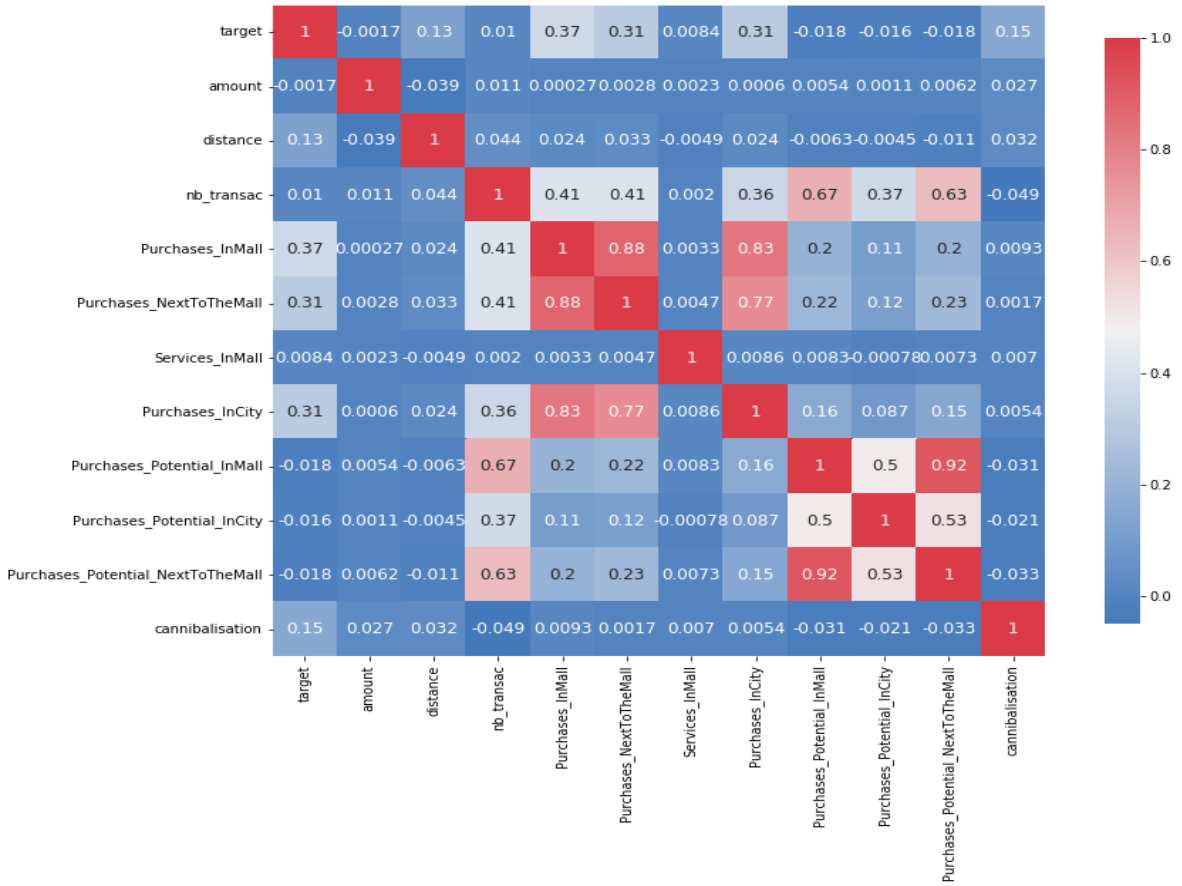


Fig. 4.4: Matrice de corrélation

On peut remarquer par exemple à partir de cette matrice de corrélation une forte liaison à 0.92 entre les variables Purchases-Potential-NextToTheMall et Purchases-Potential-InMall. Il est donc possible d'éliminer une de ces deux variables lors de la construction du modèle.

4.2.4 Relation entre les variables catégoriels :

- Variable DayOfWeek :
On remarque que les clients ont naturellement tendance à effectuer des tran-

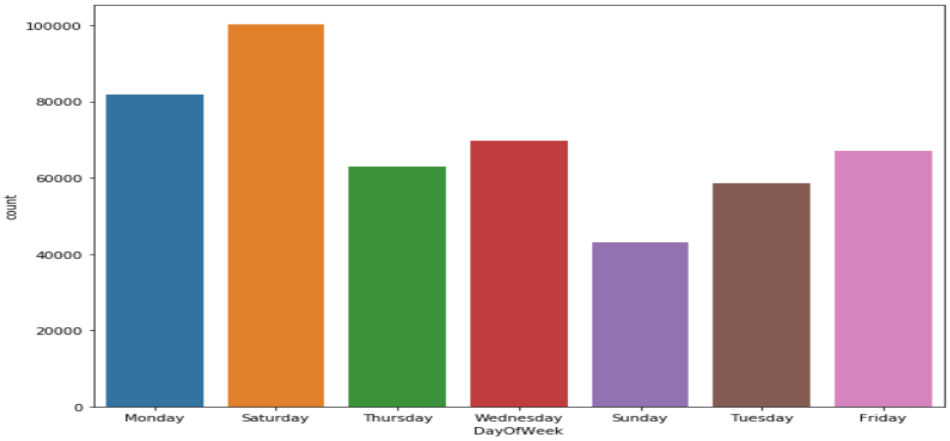


FIG. 4.5: Proportion des clients en fonction des jours

sactions le samedi puis le lundi que les autres jours. Ce qui rend cette variable importante dans la décision.

- Variable Distance-bin :
La distance Bin ici constitue le magasin le plus proche du centre visité par le client.

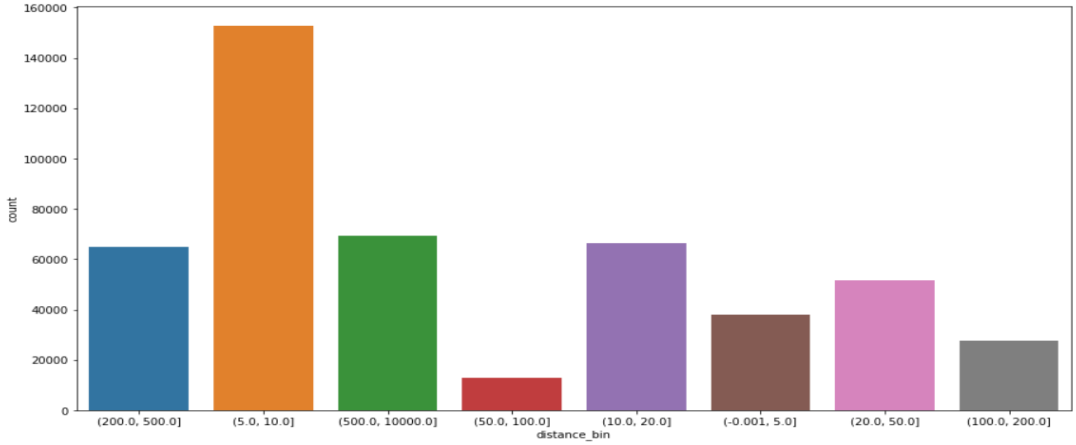


FIG. 4.6: Proportion des clients en fonction des jours

On remarque que la majorité des clients visitent d'autres magasins en dehors du centre entre 5 à 10 mètres.

4.3 Application des différents modèles au jeu de données

4.3.1 Résultats de l'application des modèles sur le jeu de donnée

Modèle	Accuracy	Couverture	AUC	Temps execution
Decision Tree	0.86	-------	-------	0.23 s
Random Forest	0.85	-------	-------	6.21 s
X GBoost	0.89	0.91	0.87	13.30 s
K-NN	0.81	-------	-------	9.33 s
Regression Linéaire	0.83	-------	-------	2.94 s
SVM	----	-------	-------	Trop de temps

FIG. 4.7: Statistiques d'apprentissage

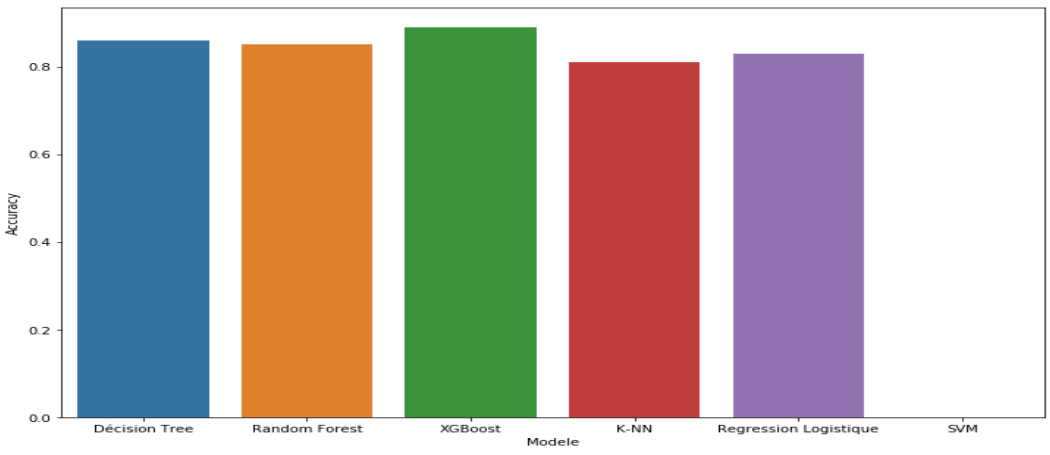


FIG. 4.8: Score des modèles

A travers ces résultats on remarque que le modèle XGBoost est plus efficace en score que les autres modèles sur le jeu de données.

- Cas particulier du X GBoost :

L'avantage que présente XG Boost est qu'il donne en plus de la prédiction, la probabilité d'appartenance à chaque classe. Ce qui nous permet d'évaluer le taux de succès en fonction de la couverture de la prédiction sur le jeu de données.

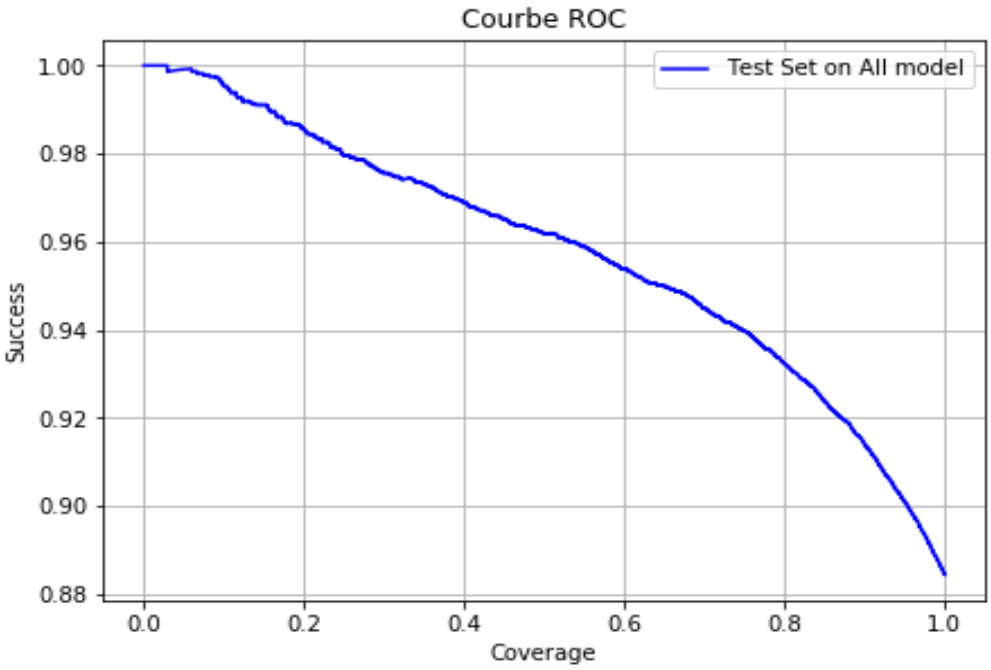


Fig. 4.9: Courbe de AUC

4.3.2 Tests et validation : Matrice de Confusion

Afin d'évaluer le taux d'erreur, on a essayé d'évaluer la matrice de confusion de chaque modèle. Sur la figure suivante seul le modèle XGBoost présente un faible taux de Faux Positif et de Vrai Négatif à 0.11. Ce qui rend ce modèle plus précis que les autres.

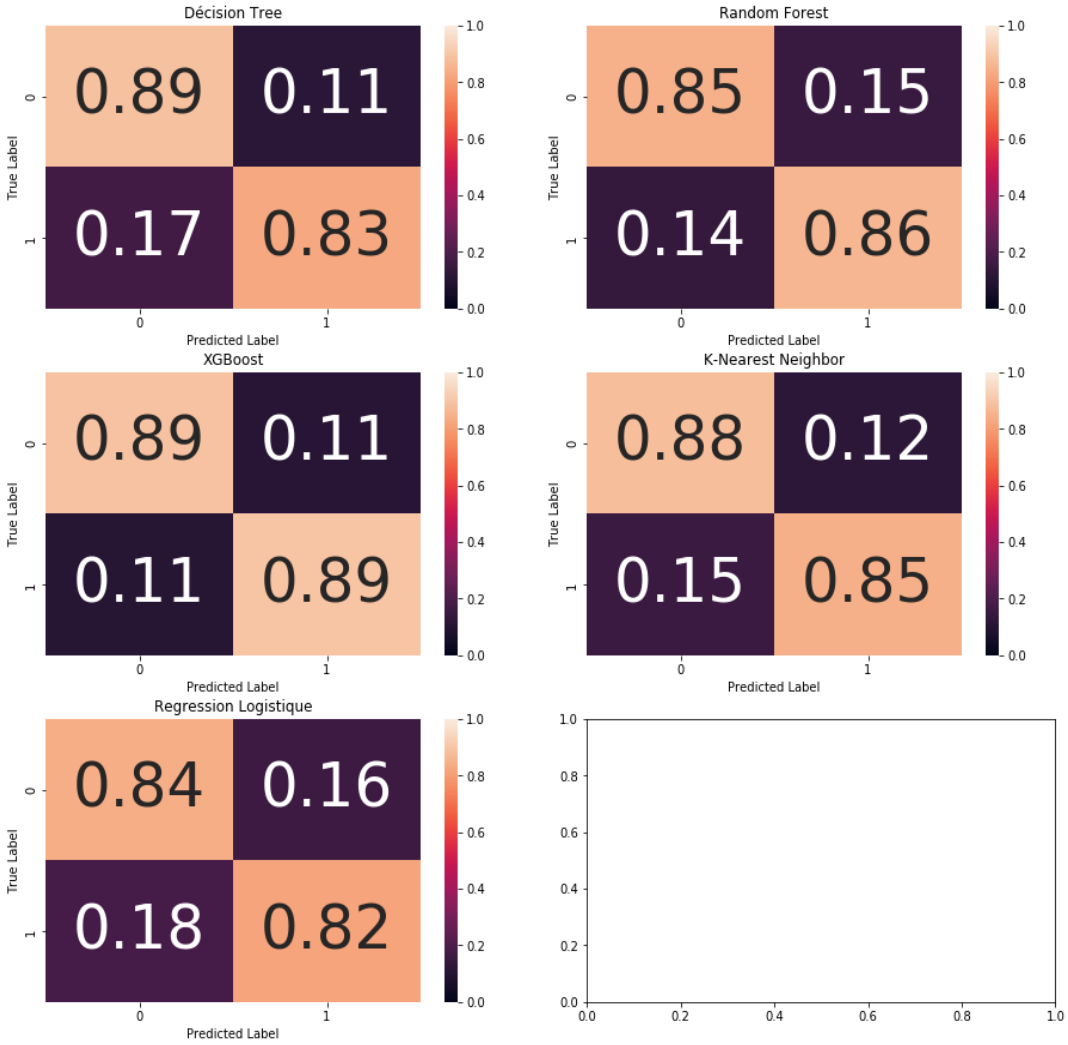


FIG. 4.10: Matrice de Confusion

4.4 Conclusion

Au vu de tous ces résultats qui montrent la performance de XGBoost, on a décidé de choisir cette dernière comme modèle classification qui sera déployé en production. Des améliorations pourront éventuellement intervenir prochainement notamment au niveau du temps de calcul des variables. Dans quelques mois les données seront multipliées par deux et il sera nécessaire de passer de redshift vers peut être Spark afin de distribuer le calcul des variables.

Cinquième partie

Conclusion

Sommaire

Conclusion

Ce stage de 6 mois effectué chez Transaction Connect m'a permis de monter en compétence d'une part de manière professionnelle et d'autre part techniquement. J'ai en effet eu la chance d'utiliser des outils de data et de développement qui m'ont permis d'être opérationnelle. Il arrive quelquefois d'être bloqué sur certains sujets mais à travers l'aide de l'équipe en place j'ai toujours eu la solution à mes inquiétudes. Grâce à l'effort fourni, j'ai donc pu terminer mon sujet principal qui est l'automatisation et l'amélioration de l'algorithme du scoring. Durant la réalisation de ma mission j'ai développé la compétence en validation de modèles de machine learning et la mise en production. Les points de progrès à mettre en place seront donc sur la restitution de l'information à des personnes non techniques.

Bibliographie

[1] Yingjie Tian, Yong Shi, Xiaohui Liu. RECENT ADVANCES ON SUPPORT VECTOR MACHINES RESEARCH. in TECHNOLOGICAL AND ECONOMIC DEVELOPMENT OF ECONOM, 2012 Volume 18(1) : 5–33

[2] Jehad Ali, Rehanullah Khan, Nasir Ahmad, Imran Maqsood. Random Forest and Decision Tree. In IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012 ISSN (Online) : 1694-0814

[3] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. KNN Model-Based Approach in Classification. In School of Computing and Mathematics, University of Ulster Newtownabbey, BT37 0QB, Northern Ireland, UK

[4] Ramraj S, Nishant Uzir, Sunil R and Shatadeep Banerjee. Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets. In International Journal of Control Theory and Applications ISSN : 0974–5572 International Science Press Volume 9 • Number 40 , 2016

[5] C. Mitchell Dayton. LOGISTIC REGRESSION ANALYSIS. Department of Measurement, Statistics and Evaluation. In Room 1230D Benjamin Building University of Maryland September 1992

4.5 Webographie

Chiffres sur l'évolution de Transaction Connect
https://incubateur-dauphine-prod.alwaysdata.net/startup-sortie/transaction-conn

Identité de Transaction Connect :
https://www.societe.com/societe/transaction-connect-822619185.html

Historique de l'évolution de Transaction Connect :
https://entreprises.lefigaro.fr/transaction-connect-75/entreprise-822619185

Premier pilote déployé par Transaction Connect :
https://business.lesechos.fr/entrepreneurs/aides-reseaux/transaction-connect-et
php

Distance de Levenshtein :
https://en.wikipedia.org/wiki/Levenshtein_distance

Random Forest :
https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

Table des figures

Liste des tableaux

Table des matières