Présentation du projet Data Sciences

Karine Tribouley

Master 2 ISIFAR

Promotion 2023- 2024

Contexte

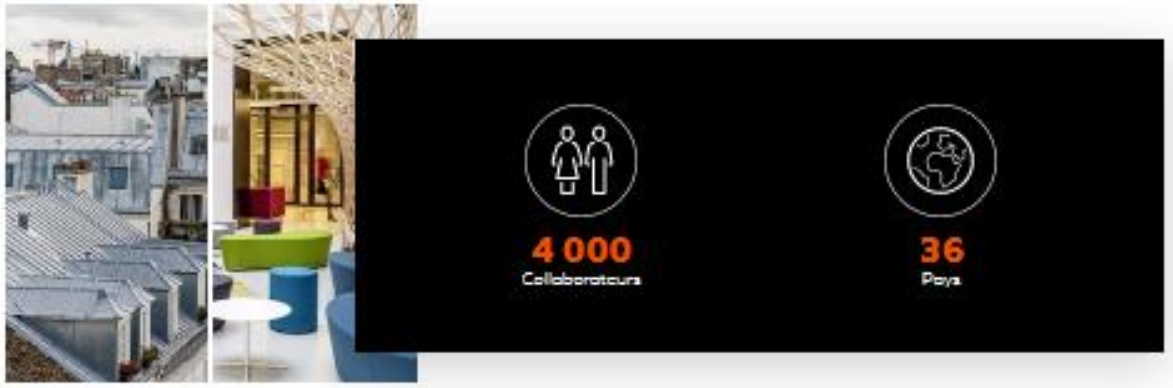
Partenariat RCI – Master ISIFAR 2017 – 2018

Marques de l'Alliance

Renault Crédit International est

- une filiale du groupe français Renault (100%) crée en 1974
- spécialisée dans le financement, les services automobiles, l'assurance et les activités liées aux marques du groupe Renault et Nissan
- présente en Europe, en Russie et en Amérique du Sud
- affichant un CA = 1,9 Milliards Euros en 2020
- faisant travailler 3 700 collaborateurs en 2019

https://www.mobilize-fs.com/fr/notre-groupe/nos-chiffres-cles



Mission

Objectif Data science : construire un score d'octroi de crédit pour les clients du groupe automobile

Périmètre :

➔ Irlande
➔ Historique : 01/01/2014 à 01/11/2015
➔ Niveau contrat – un client peut avoir plusieurs contrats - **N = 8 457** contrats.
➔ Cible : **98** contrats sont tombés en défaut dans l'année qui a suivi la date de début de contrat
➔ Variables disponibles : **p = 17**

Fichier : « Octroi.csv »

Objectif Master ISIFAR :

- Travailler sur une mission à but opérationnel
  - Restitution de qualité professionnelle
  - Résultats combinant sens métier/performance
- Avec des données réelles
- En mode « projet »
  - En équipe
  - Encadré par un CP

Challenge :

- Taux de cible très petit : solutions ?
- Création de features appropriées

Brief

Contexte : un vendeur de véhicules adossé à une banque-crédit

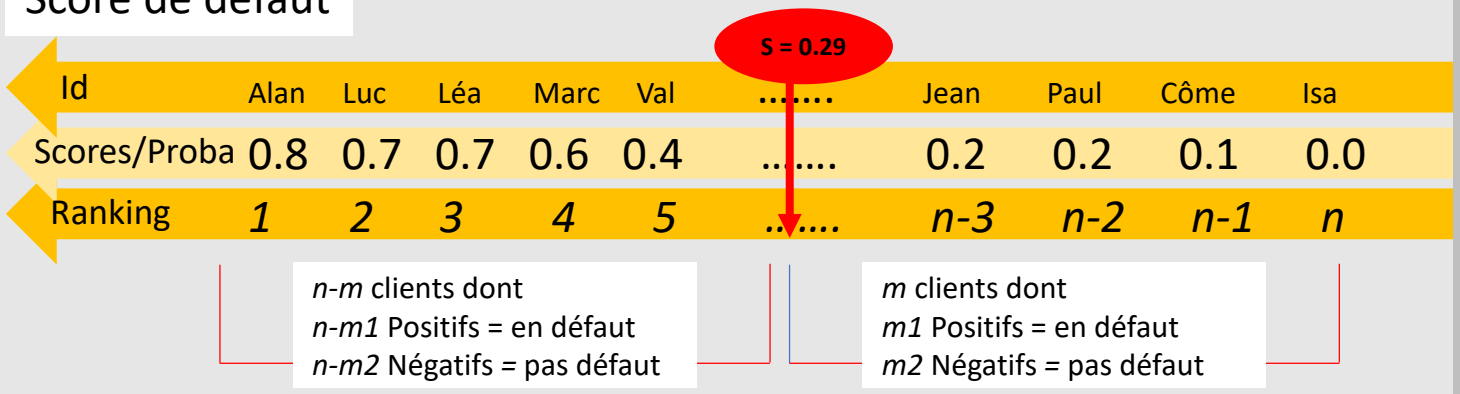
Objectif Métier :

- Ne pas proposer de crédit aux personnes qui vont faire défaut
- Proposer un crédit aux personnes qui ne vont pas faire défaut

Objectif Data Science : Construire un score de défaut

On connait la variable de défaut
1 = Défaut
0 = Non défaut

Score de défaut



Si le business est sponsor – risque mini

1. le business fixe l'assiette *m* = nombre de clients à qui on propose un crédit
2. Le data scientist propose
   - les clients à qui on attribue un crédit sont les *m* plus bas scorés
   - la performance est évaluée en calculant le taux de
     - Risk ➔ FN = Faux Négatifs parmi les *m* plus bas scorés
     - VP = Vrais Positifs parmi les *n-m* plus hauts scorés

Si le risk est sponsor – assiette maxi

1. le risk fixe le nombre de FN
2. le data scientist propose l'assiette
   - le nombre *m* de clients à qui on offre un crédit ou, de manière équivalente,
   - le seuil *s* à partir duquel le client n'obtient pas de crédit

Cadrage

Décisions à prendre :

➔ Score d'Octroi ou Score de Défaut ?
- ✓ La cible est-elle DEFAUT ou 1 - DEFAUT ?

➔ **Feature engineering**

Prétraiter les données
- ✓ Que faire des NA ?
- ✓ Les variables sont elles toutes d'intérêt ?

Créer de nouvelles features – indicateurs …
- ✓ Impact sur la population cible ?

➔ Créer un échantillon TRAIN et un échantillon TEST
- ✓ Quid du taux de cible ?
- ✓ Re-échantillonner ? Pondérer ?

Taux de cible loin de 50%

- Sur échantillonnage des « CIBLE = 1 » si peu de données ➔ est ce prédictif ?
- Sous échantillonnage des « CIBLE = 0 » si beaucoup de données ➔ perd-on de l'information ?

data algo

➔ Déterminer l'algorithme pour modéliser ?

Comment sélectionner les inputs ?
- ✓ Paramètres de tuning
- ✓ Features

➔ Mesurer les performances
- ✓ Courbes : Lift (Marketing) ? ROC (Risk) ?
- ✓ Indicateurs : Gini ? AUC ? Alpha-Lift ?

VP et FN (classification) ?

Attention : Les modèles ont été appris sur la base d'apprentissage. Les qualités doivent être évaluées sur la base de test

Principe du « on ne peut être juge et partie »

- AUC toujours bon
- VP et FN toujours bons

Planification

Mission en binôme

Réception des données

Restitution 1

Point d'étape Restitution 2

Point d'étape Restitution 3

Fin de la production

Evaluation étudiants

GO / NOGO

Kick off **21/12**

10/01 | **24/01** | **14/02** | **06/03**

Livraison **29/03**

Recette

Clôture/PPR **24/05**

Constitution Teams **10/01**

- ✓ Audit et Pré-traitement des données
- ✓ Création des features

- ✓ **Re-échantillonage**
- ✓ Modélisation Logistique
- ✓ Performance

- ✓ Modèle Alternatif
- ✓ Performance

Rédaction des livrables :
- ➢Codes documentés
- ➢Document final

Janvier / Février / Mars

19/01 | 26/01 | 02/02 | 09/02 | 16/02 | 23/02 | 01/03 | 08/03 | 15/03 | 22/03

Avril / Mai

CP méthodo Paris Diderot : TRIBOULEY
Retours Flash Report mail ***tous les vendredi***
Format **nom1_nom2_FR1.pdf**
Restitutions 1/2/3/ finale : livraison par mail
(la veille de la date convenue avant minuit)

Cours magistral ou atelier : mercredi à 13h
Point d'étape 1 : 24 janvier
Point d'étape 2 : 14 février
Point d'étape 3 : 6 mars

Livrables

Objectif : Construction d'un score d'octroi

Contrainte méthodologique : 2 méthodes doivent être proposées dont la régression logistique

Résultats indispensables:

- La performance de la méthode utilisée pour palier au déséquilibre de la cible est quantifiée
- la qualité de chaque score est évaluée
- les 2 scores sont comparés notamment
  - Calcul des indices de Gini
  - Calcul de l'indice 10/X : pourcentage de défauts sur les 10%

    plus hauts scorés –score octroi

    plus bas scorés – score défaut

Description des livrables :

Pour chaque restitution, un document sous forme ppt comprenant obligatoirement les éléments suivants :

- Introduction : objectifs du jalon et démarche proposée
- Présentation de la méthodologie et de sa mise en oeuvre
- Résultats
- Performances
- Conclusion : REX et recommandation éventuelle
- Aperçu sur le jalon suivant

Pour la livraison finale :

- Un document ppt retraçant la mission complète
- Les codes documentés et fichiers des INPUT + OUTPUT
  - ✓ Le langage de codage est au choix de chaque équipe : SAS / R / Python

Evaluation

Etapes d'évaluation :

- Point d'étape 1 : pas d'évaluation
  - ➔ **Go ou No Go**
- Point d'étape 2 : évaluation E2
- Point d'étape 3 : évaluation E3
- Livraison finale: évaluation EF

Note finale : Si **Go au point d'étape 1**

Note = (E2+E3)/4 + EF/2 **- 1 par jour de retard**

Critères :

33% lié au comportement

- Respect du planning
- Retours et relations avec CP
- Travail en équipe, prise du leadership

33% lié à la qualité de la présentation :

- Respect des éléments demandés
- Travail de synthèse et de restitution des résultats
- Forme de la présentation

33% lié à l'aspect scientifique

- Bonne application des méthodo vues en cours
- Pertinence de l'analyse statistique
- Travail de recherche sur une solution pour le problème du petit taux de cible

	Variable	Libellé	Description
	mois_gestion	Mois d'entrée en gestion	
Cible	def12_31	Indicateur de défaut	0= Non 1 = Oui
Variables candidates à la modélisation	ANC_EMPLOI	Ancienneté à l'emploi	
	PRIX_VEH	Prix du véhicule	
	MT_APPORT	Montant de l'apport	
	MT_FINANCE	Montant financé	
	MT_MENS	Montant de la mensualité	
	VR_BALLON	Montant ballon	Fait référence à un crédit ballon. Pendant X mois le client rembourse des mensualités (intérêts) et à la fin de cette période, le client peut acheter le véhicule ou bien le restituer. La valeur de rachat est le ballon.
	DUREE_CONTRAT	Durée du contrat	
	MT_PREST	Montant des prestations	
	MT_ASSUR	Montant des assurances	
	age_cli	Age du client	
	ancienete_rci	Ancienneté relation RCI	
	pc_appo	Pourcentage d'apport	
	AGE_VEH	Age du véhicule	
	STITUATION_FAM	Situation familiale	1 = Marié 2 = Célibataire 3 = Divorcé 4 = Veuf 5 = Separé 11 = Collocation
	MODE_LOGT	Mode logement	1 = Locataire 2 = Propriétaire 3 = Autre 4 = Chez les parents
	MARQUE	Marque	
	VNVO	Type véhicule	VN = véhicule neuf VO= véhicule occasion

Exemple de rationalisation pour un CR

Tout point de discussion, coproj, copil, doit pouvoir être tracé.

C'est le responsable de la réunion qui s'en charge

- Ex Carrouf : Cap Gémini
- Ex RCI : Equipe Etudiant

Cela permet

- agilité
- partage
- évaluation des risques

Cela évite

- les alertes non remontées
- les contestations
- les oublis
- les incompréhensions

FLASH REPORT

Émetteur : Karine TRIBOULEY

Participants : Tribouley, Equipe étudiants 1

Période : 5 janvier 2019 – 15 janvier 2019

Dest. : **Zied DRIDI** Internal Credit Risk Modeler
DEPARTMENT OF ANALYTICS

1. Principales actions menées

- Chargement des data
- Audit des data

2. Risques identifiés

- Congés du 15 janvier au 20 janvier
- Démission d'un membre de l'équipe étudiants
- SAS impossible à installer

3. Décisions prises

- Ne pas utiliser la variable TRUC car trop de NA
- Filtrer sur les individus majeurs
- Ok congés
- Remplacement SAS par Python

4. Réunions prévues et échéances

- Point d'échange le 21 janvier
- Point d'étape 1 : le 30 janvier ➔ restitution partie « Data » à RCI

5. Problèmes à régler – Décisions à prendre

- Proposer à un autre étudiant de rentrer dans l'équipe ➔ cf étudiants qui rendent compte à Tribouley

6. Actions en cours / à venir

- Finir la partie Data en exploratoire
- Rédiger les slides Partie Data

format du fichier pdf à m'envoyer tous les vendredi soir avant minuit
MAUFFRET_PRUVOT_FlashReport1

Utiliser le email

Adresse email

Utiliser son mail pro ou étudiant.

Alternative : adresse de messagerie privée avec un « vrai » identifiant ➔ roudoudou@gmail.com à bannir.

Objet du mail

Un email sans **objet** n'est pas ouvert

➔ concis, doit permettre le désarchivage

Utilisation du cc

Toute personne en copie d'un email est considérée comme ayant été informée

DOIVENT être en copie de tout email pro

- Le supérieur hiérarchique ➔ doit pouvoir suivre toutes les démarches
- Les participants au projet ➔ pas de rétention d'information
- L'administratif concerné

Exemples :

- Mail concernant le projet DataScience adressé à tribouley + membres de l'équipe en cc

Attention, toujours utiliser le « répondre à tous »

Pas de réponse si tout le monde n'est pas en cc

Corps du mail

PJ en format pdf

Un email DOIT

- être court. Utiliser PJ pour faire des rapports ou comptes rendus longs
- d'un niveau de langage professionnel, sans «décoration»
- ne pas être écrit en majuscules, avec des abréviations hors contexte, des émoticons
- ne pas comporter de fautes d'orthographe
- commencer par **Bonjour,**
- finir par **Cordialement** ➔ même position hiérarchique ou inconnu,
  A disposition si besoin ➔ supérieurs hiérarchiques

Checker AVANT d'envoyer

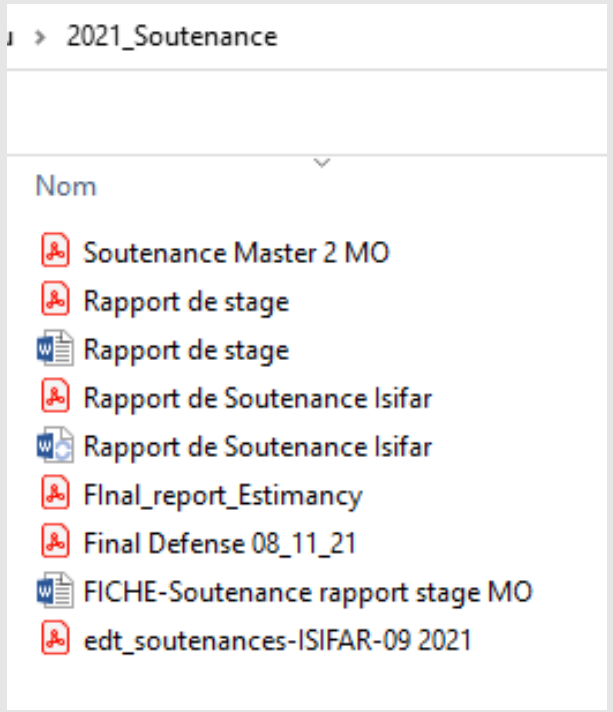
- **Objet** est-il renseigné ?
- **Reply all** est-il utilisé ? Le **n+1** y est-il ?
- En cas de **transfert**, le nettoyage est-il ok ? – si commentaires perso
- La **pièce jointe** est-elle jointe ?
- **Fautes d'orthographe** sont-elles absentes ?
- **Cordialement, A disposition si besoin** figurent-ils ? si même niveau hiérarchique ou inconnu.
- **Signature** ? Nom ET prénom.

TOUJOURS GARDER LES EMAILS ENVOYES

Bonnes pratiques

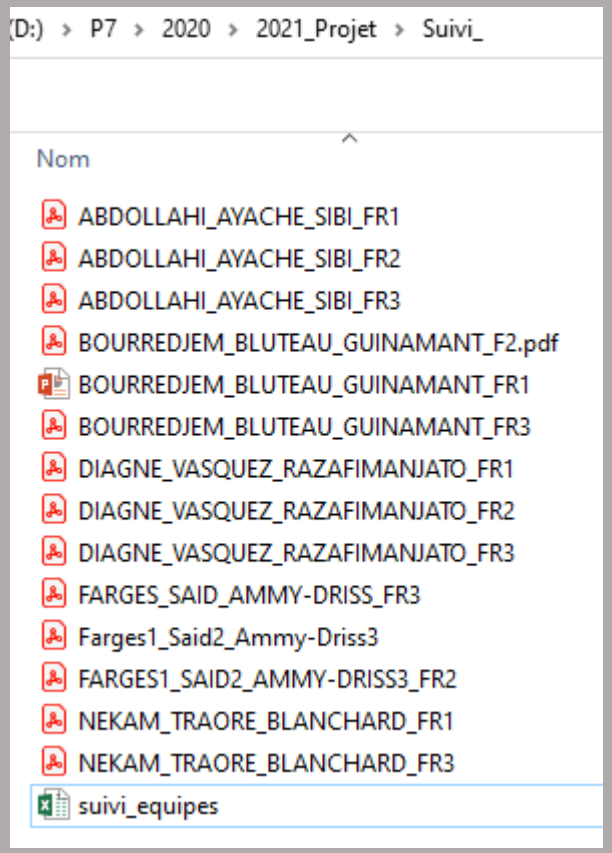
Noms fichiers

Exemple de fichiers reçus par mail en octobre

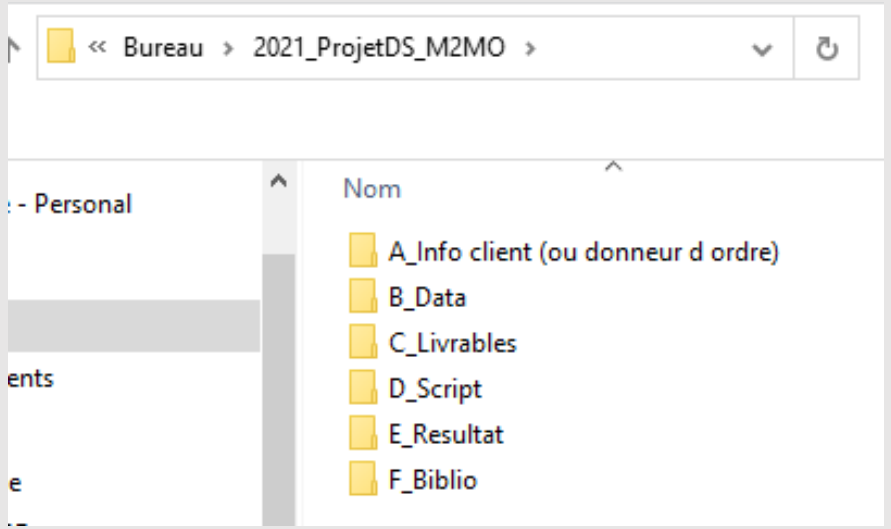


Se mettre à la place du destinataire

Exemple : suivi des projets



Répertoire projet dédié



Propre à chaque entreprise

En général, les fichiers

- ont un nom avec la date YYYYMMDD en préfixe
- ont un nom précis et non générique

Ex : **TRIBOULEY_Rapport_stage.pdf** et non **rapport.pdf**

- sont des fichiers pdf

QUESTIONS ?