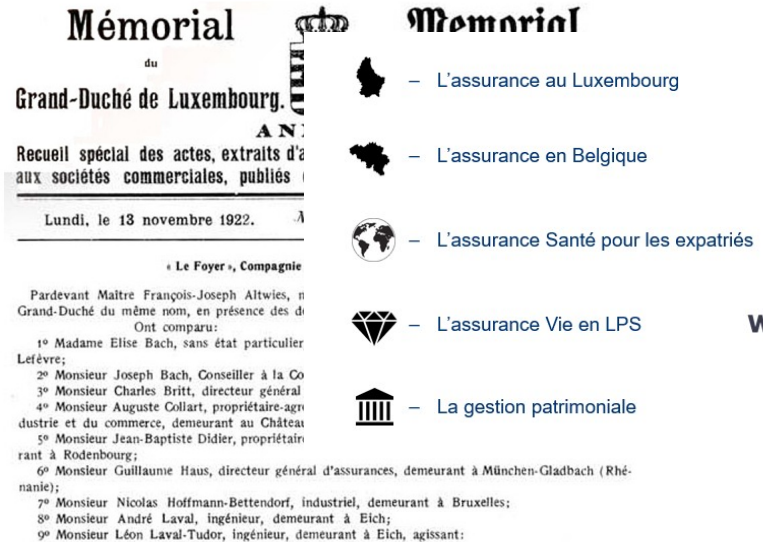
Challenge Data Foyer

30/08/2021

Le Groupe Foyer

1922



Mémorial

du

Grand-Duché de Luxembourg.

AN

Recueil spécial des actes, extraits d'a
aux sociétés commerciales, publiés

Lundi, le 13 novembre 1922.

« Le Foyer », Compagnie

Pardevant Maître François-Joseph Altwies, n
Grand-Duché du même nom, en présence des d
Ont comparu:

1° Madame Elise Bach, sans état particulier
Lefèvre;
2° Monsieur Joseph Bach, Conseiller à la Co
3° Monsieur Charles Britt, directeur général
4° Monsieur Auguste Collart, propriétaire-agr
dustrie et du commerce, demeurant au Châtea
5° Monsieur Jean-Baptiste Didier, propriétair
rant à Rodenbourg;
6° Monsieur Guillaume Haus, directeur général d'assurances, demeurant à München-Gladbach (Rhé-
nanie);
7° Monsieur Nicolas Hoffmann-Bettendorf, industriel, demeurant à Bruxelles;
8° Monsieur André Laval, ingénieur, demeurant à Eich;
9° Monsieur Léon Laval-Tudor, ingénieur, demeurant à Eich, agissant:

Memorial

- L'assurance au Luxembourg
- L'assurance en Belgique
- L'assurance Santé pour les expatriés
- L'assurance Vie en LPS
- La gestion patrimoniale

2021

Foyer

storique de
e au Luxembourg

global health
Foyer Group

aborateurs et agents

WEALINS WEALTH INSURER
FOYER GROUP

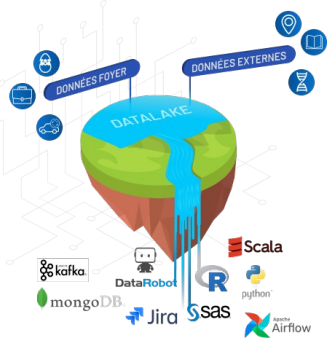
GBLife Luxembourg
Foyer Group

12 pays européens

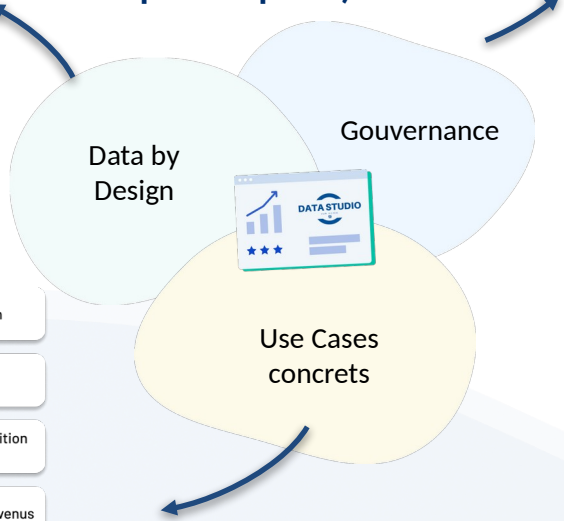
capital at work
Wealth Management
Foyer Group

276 000 clients au Luxembourg
et en Europe

La stratégie Data



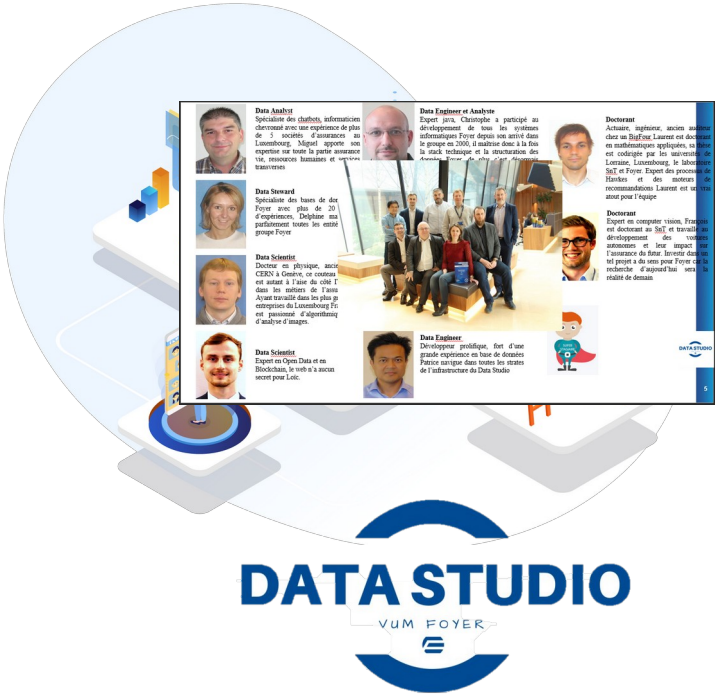
Mise en place d'un Datalake ouvert à des données externes (briques techniques en place)



- Mise en place d'une politique des données
- Le prédictif au quotidien
- Cartographie des données
- Un meilleur pricing pour nos produits
- Mise en place de contrôles et d'audits
- Génération et/ou acquisition de nouvelles données
- Acculturation de Foyer à l'utilisation de la donnée
- Nouvelles sources de revenus
- Centraliser le reporting
- Nouveaux produits et business models
- Un tag management interne et externe efficace
- Des personas à l'image de notre portefeuille

- **Mise en place d'une stratégie Data avec une Gouvernance quant à l'utilisation des données.**
- **Mise en place d'un Data Studio qui a 4 missions fondamentales (OKR) :**
  - Garantir la qualité et l'intégrité des données stockées et traitées
  - Garantir une Business Intelligence efficace avec la mise en place de Dashboard pertinents.
  - Organiser l'Open Data et l'intégration efficace de nouvelles données dans nos systèmes .
  - Elaborer des Use Cases pour le *machine learning* et l'IA.

Data Studio

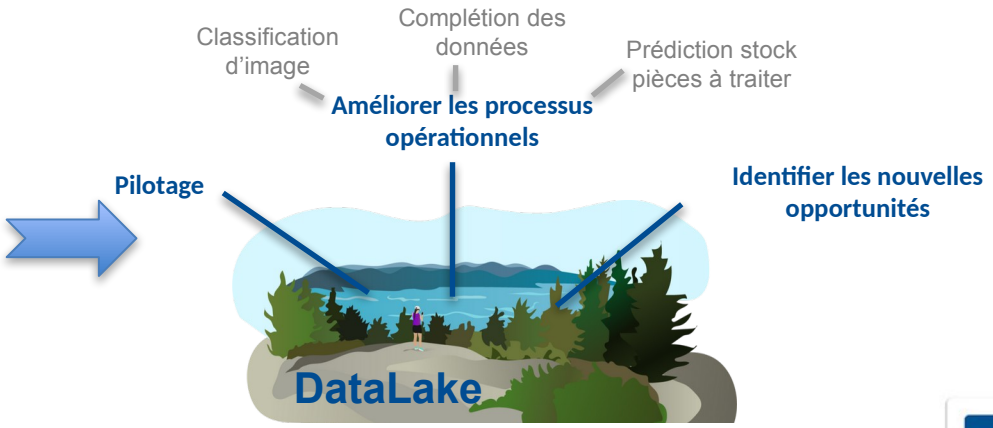


- Data {scientist, engineer, analyst, *} bref tout ce qui commence par data
- Actuellement : Une vingtaine de personnes
  - 4 devs
  - 2 data scientist (physicien)
  - 1 spécialiste open data
  - 5 business analyst
- + 2 doctorants du SNT (UNI Luxembourg)
- + 1 postdoc du SNT
- + 4 stagiaires

La qualité des données
C’est la clé ☺

QUALITE DES DONNEES

- Structuration (Dataset)
- Outils (SAS, Datarobot)
- Rationalisation (reporting)
- Contrôles clés



Etat des lieux

- Règles de contrôles déterministes
- Tentative d’introduire des indicateurs statistiques et des méthodes de Machine Learning sans résultats convaincants

C’est pour ça que l’on a besoin de vous ! ☺

Sujet

Comment évaluer de manière automatique la qualité d'une table de données ?

Plateforme de contact - Discord

- Participants ont reçu une invitation sur leur mail
- Contacts (écrits ou vocaux) sur la plateforme Discord
- Chaque équipe aura un salon privé pour s'organiser
- Permanences organisées sur Discord 3 fois par jour (9h-14h-17h)
- N'hésitez pas à m'appeler ou m'envoyer un mail en cas de problème avec Discord

Base de travail

Index	SalesID	SalePrice	MachineID	ModelID	datasource	auctioneerID	YearMade	ineHoursCurrentI	UsageBand	saledate
0	1139246	66000	999089	3157	121	3	2004	68	Low	11/16/2006 0:00
1	1139248	57000	117657	77	121	3	1996	4640	Low	3/26/2004 0:00
2	1139249	10000	434808	7009	121	3	2001	2838	High	2/26/2004 0:00
3	1139251	38500	1026470	332	121	3	2001	3486	High	5/19/2011 0:00
4	1139253	11000	1057373	17311	121	3	2007	722	Medium	7/23/2009 0:00
5	1139255	26500	1001274	4605	121	3	2004	508	Low	12/18/2008 0:00
6	1139256	21000	772701	1937	121	3	1993	11540	High	8/26/2004 0:00
7	1139261	27000	902002	3539	121	3	2001	4883	High	11/17/2005 0:00
8	1139272	21500	1036251	36003	121	3	2008	302	Low	8/27/2009 0:00
9	1139275	65000	1016474	3883	121	3	1000	20700	Medium	8/9/2007 0:00
10	1139278	24000	1024998	4605	121	3	2004	1414	Medium	8/21/2008 0:00
11	1139282	22500	319906	5255	121	3	1998	2764	Low	8/24/2006 0:00
12	1139283	36000	1052214	2232	121	3	1998	0	nan	10/20/2005 0:00
13	1139284	30500	1068082	3542	121	3	2001	1921	Medium	1/26/2006 0:00
14	1139290	28000	1058450	5162	121	3	2004	320	Low	1/3/2006 0:00
15	1139291	19000	1004810	4604	121	3	1999	2450	Medium	11/16/2006 0:00
16	1139292	13500	1026973	9510	121	3	1999	1972	Low	6/14/2007 0:00
17	1139299	9500	1002713	21442	121	3	2003	0	nan	1/28/2010 0:00
18	1139301	12500	125790	7040	121	3	2001	994	Low	3/9/2006 0:00
19	1139304	11500	1011914	3177	121	3	1991	8005	Medium	11/17/2005 0:00
20	1139311	41000	1014135	8867	121	3	2000	3259	Medium	5/18/2006 0:00

Base de travail

- CSV classique
- 100 000 lignes pour 53 colonnes
- Enrichie d’erreurs que nous stockons (1% des lignes environ)
- Ces erreurs vous permettront d’estimer la performance de votre méthode
- Un dictionnaire de données est aussi disponible

Base de travail – Evaluation des performances

- Déposer votre fichier CSV d'erreurs sur Discord dans votre salon d'équipe
- Une ligne par erreur détectée
- Structure Fichier : SalesID – ColonneError – TypeError (peut être flexible selon les types d'erreurs)
- Calcul de votre score de performance
- Maximum 1 fois par jour
- Possibilité de tester votre méthode sur nos bases durant le challenge

Base de Travail – Point d’attention

- Support pour tester vos algorithmes
- La performance de votre méthode doit à minima être prouvée sur cette base de tests et être généralisable à d’autres tables
- La finalité sera d’utiliser cette méthode sur nos bases, potentiellement différente de celle-ci

Rendu

A fournir à aog@foyer.lu,nig@foyer.lu,nif@foyer.lu avant la fin du challenge :

1. Proposition d'une définition de la qualité d'une donnée
2. Algorithme (descriptif + code source en Python ou R) qui permet d'évaluer la qualité d'une donnée et de l'ensemble des données d'une table
3. Bilan des erreurs de la table fournie en entrée comportant pour chaque erreur son type ainsi que son ID + colonne impactée

Jury

- Une présentation publique en visio-conférence sera organisée le vendredi 10/09 à 9h.
- Foyer + Uni délibère le verdict
- Nous vous proposons une session d’entrainement deux jours avant le jury

Ce à quoi nous avons pensé

- Détection d’anomalies
- Indicateurs statistiques (kurtosis,skewness…)
- Méthodes de drift
- Distance de Levenshtein
- Génération de règles logiques
- Tests classiques (null…)

Ressources

- https://www.win.tue.nl/~mpechen/publications/pubs/Gama_ACMCS_AdaptationCD_accepted.pdf
- https://github.com/online-ml/river
- https://paperswithcode.com/paper/pyod-a-python-toolbox-for-scalable-outlier
- https://mobidev.biz/blog/unsupervised-machine-learning-improve-data-quality
- https://www.vldb.org/pvldb/vol11/p1781-schelter.pdf
- https://arxiv.org/ftp/arxiv/papers/1810/1810.07132.pdf
- https://arxiv.org/ftp/arxiv/papers/2009/2009.06672.pdf

Contacts

- Alexandre HOTTON (AOG) - aog@foyer.lu
- François NIEDERCORN (NIF) – nif@foyer.lu
- Geoffrey NICHIL (NIG) – nig@foyer.lu

Rendez vous sur le serveur Discord 11H15 !

Questions ?