Thèse Professionnelle
Mastère Spécialisé Big Data

Télécom ParisTech

Contribution au développement d'outils d'apprentissage automatique et d'aide à la décision dans le processus de gestion et d'optimisation des levées de capitaux à destination des sociétés privées

---

Kaëlig CASTOR

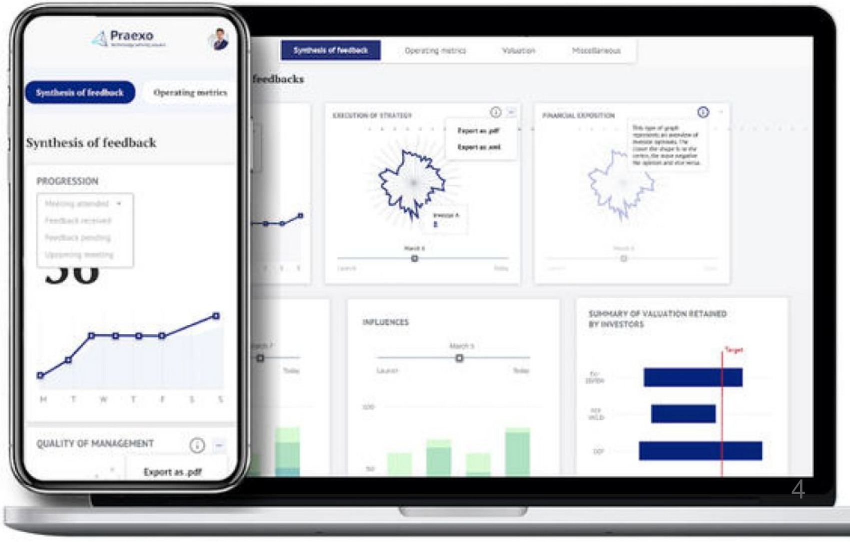
27 Septembre 2021

Plan de l’exposé

1. Introduction

La société Praexo : digitalisation de la levée de capitaux

1. **Création en 2019** par Guillaume Moinet, banquier d'affaires chez Rothschild & Co
2. 1ère levée de 1.6M€ en mai 2020
3. Exploiter et analyser les données pour optimiser la relation entreprise-investisseur
    - Accompagner les sociétés en préparant et améliorant l’exécution de leurs levées de capitaux (3 ou 4 ans avant leur possible IPO)
    - **Mieux cibler les bons investisseurs, anticiper et comprendre leurs attentes** (*equity story*)
    - Dynamiser l’intérêt de l’investisseur - aide à la décision
4. **~10 personnes** :
    - Business Team
      (3 banquiers senior + 2 stagiaires)
    - Data Science Team (1 + 2 stagiaires) :
      data lake, ML
    - Tech Team
      (1 CTO + 2 front-end developers) :
      data viz, back-end, infra



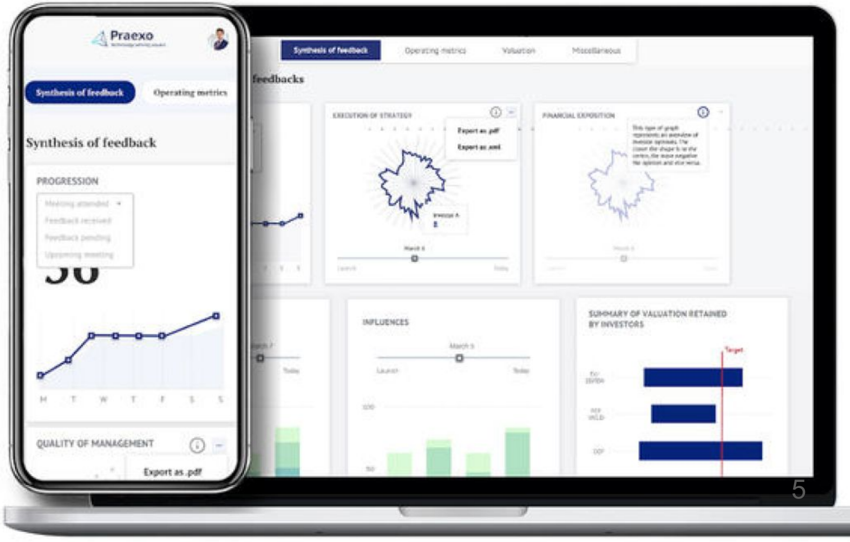
La plateforme Praexo : une communication directe

1. Tableau de bord (web + app) - état d’avancement de la transaction
   - Liste investisseurs impliqués + retours investisseurs synthétisés
   - Suivi en temps réel des métriques de valorisation et des KPI

2. Module d’interface
   - Système de dialogue / commentaires - questions/réponses entre entreprise et investisseurs
   - Système de réservation de commandes

3. Modèle d’analyse financière
   - Outils d’estimations - comparaison d’hypothèses
   - Mesures d’impact sur les notations

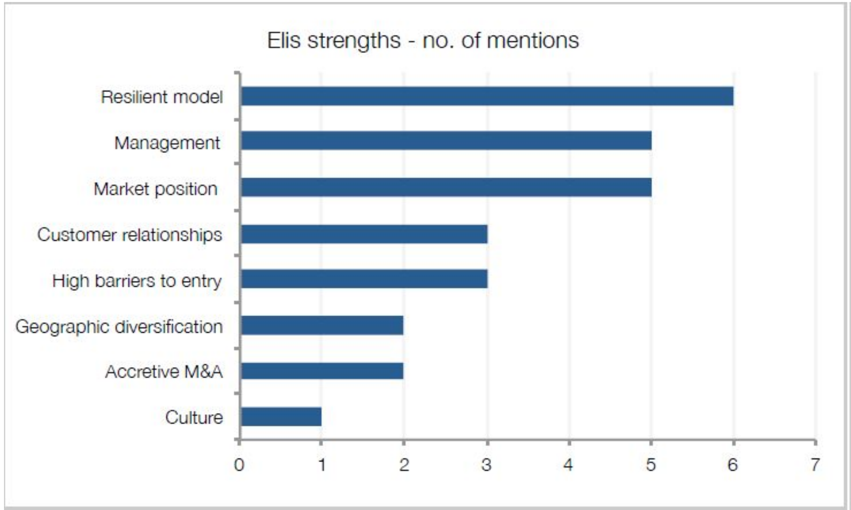
4. Accompagnement des clients avec l’expertise métier des banquiers d’affaires

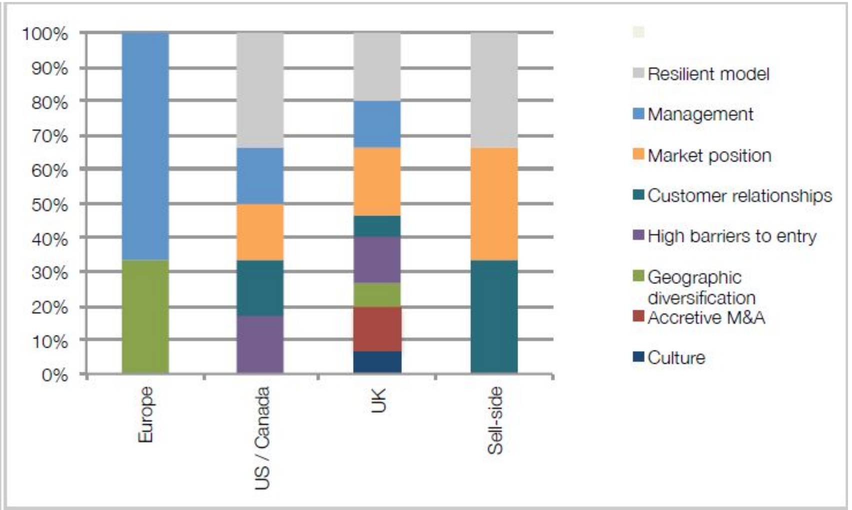


Contexte des Projets

1. Système de recommandation d'investisseurs
    - User (transaction) - Item (investisseur)
    - Caractéristiques des investisseurs et des investissements passés (critères géographiques, secteurs d'activité, style d'investissement, montant des transactions, du ticket moyen des investisseurs, etc)

2. Classification de commentaires d'investisseurs
    - Synthèse et catégorisation automatique des retours des analystes
    - Forces, faiblesses, et perception générale de la compagnie



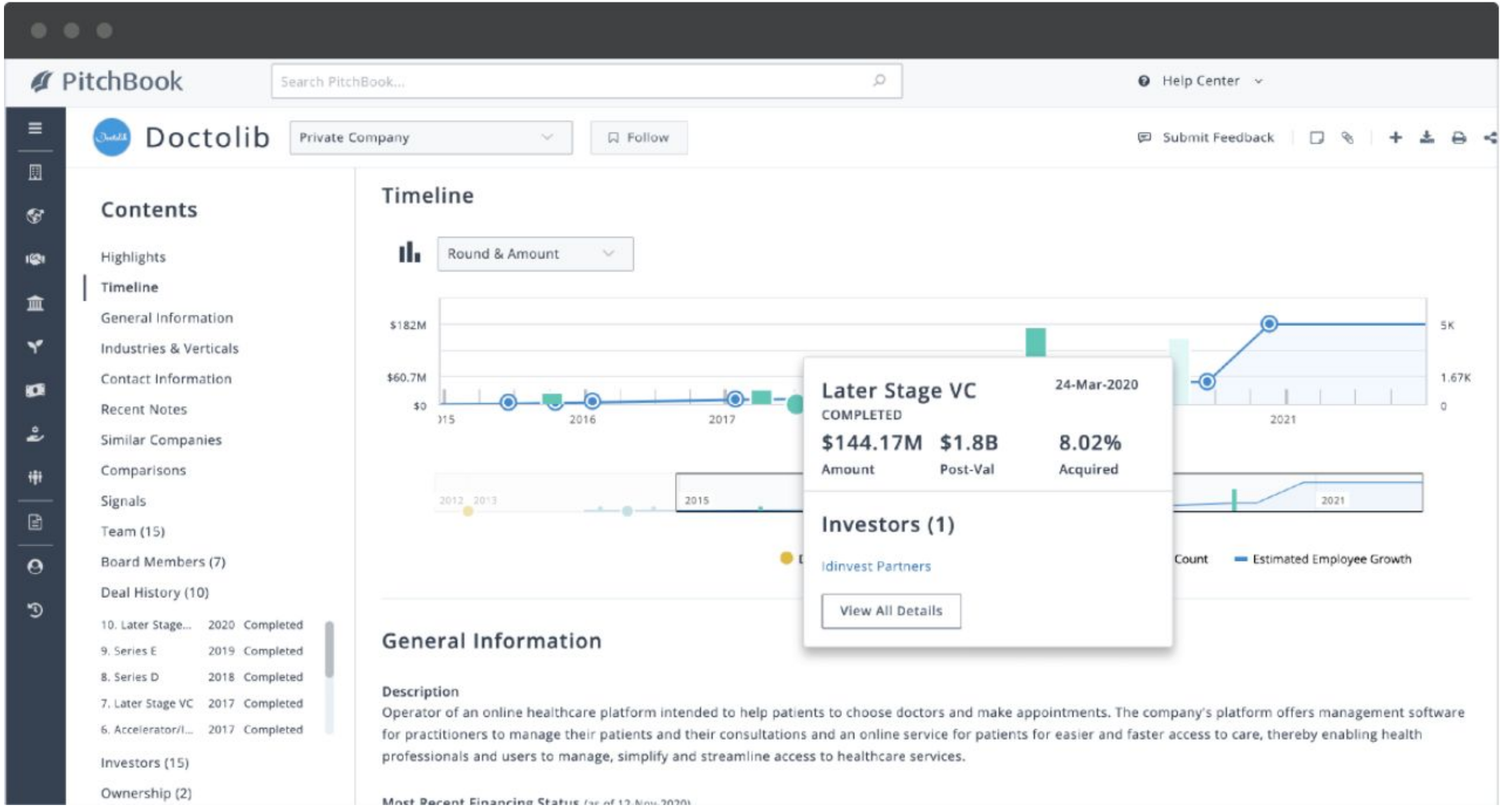


2. Système de recommandation d’investisseurs

Création de la Base de Données : scraping de pitchbook

PitchBook

Scraping de 5 fichiers raw :
- ○ **deals.json** (15687 deals)
- ○ **investors.json** (10770 investors)
- ○ (companies.json)
- ○ (advisors.json)
- ○ (funds.json)



3,234,754 Private companies	97,906 Publicly traded companies
94,716 Venture capital-backed companies	96,003 Private equity-backed companies
54,221 Startups and stealth startups	816 Pre-IPO companies

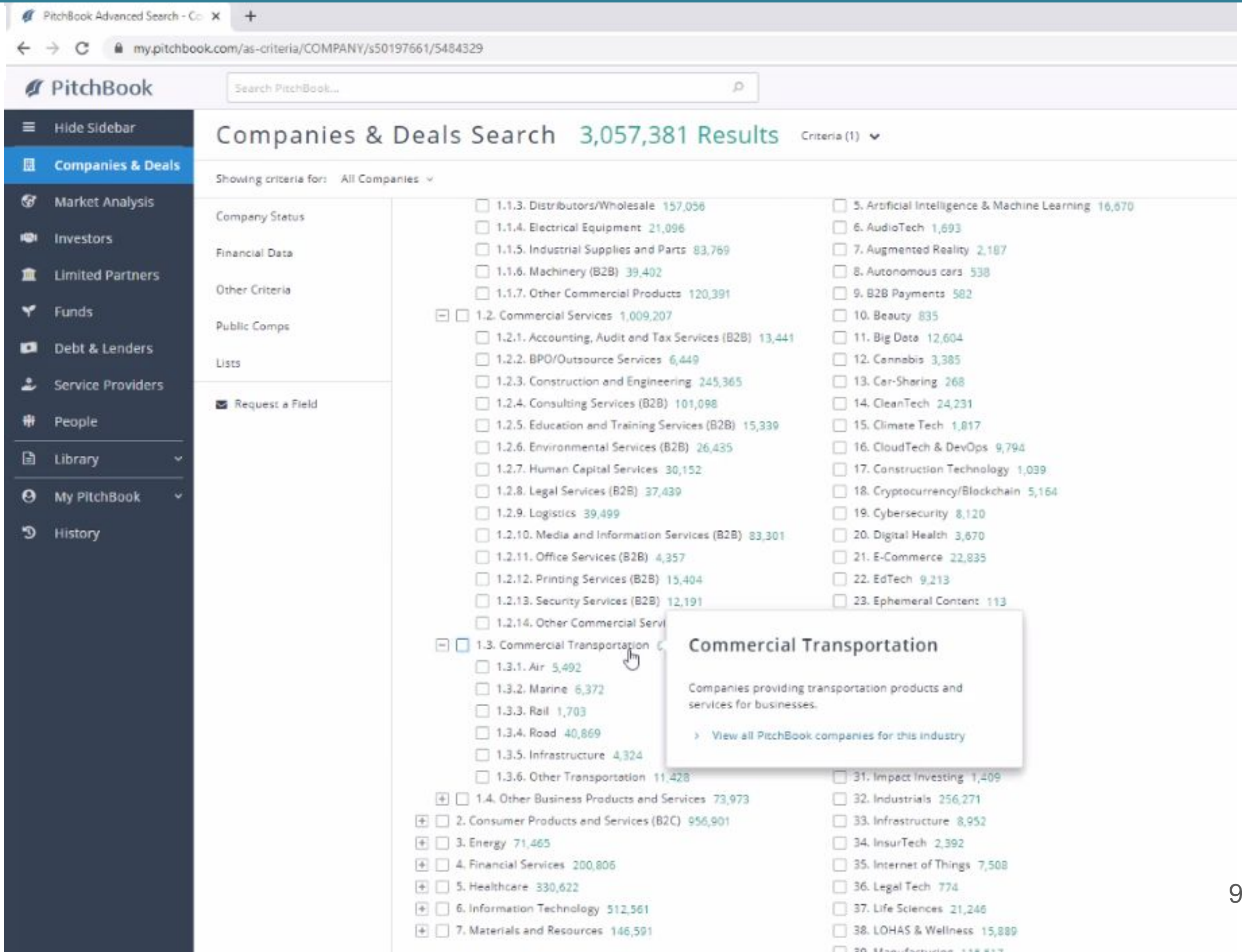
3,293,727 global companies and 1,038,440 European companies,

Création de la Base de Données : scraping de pitchbook

PitchBook

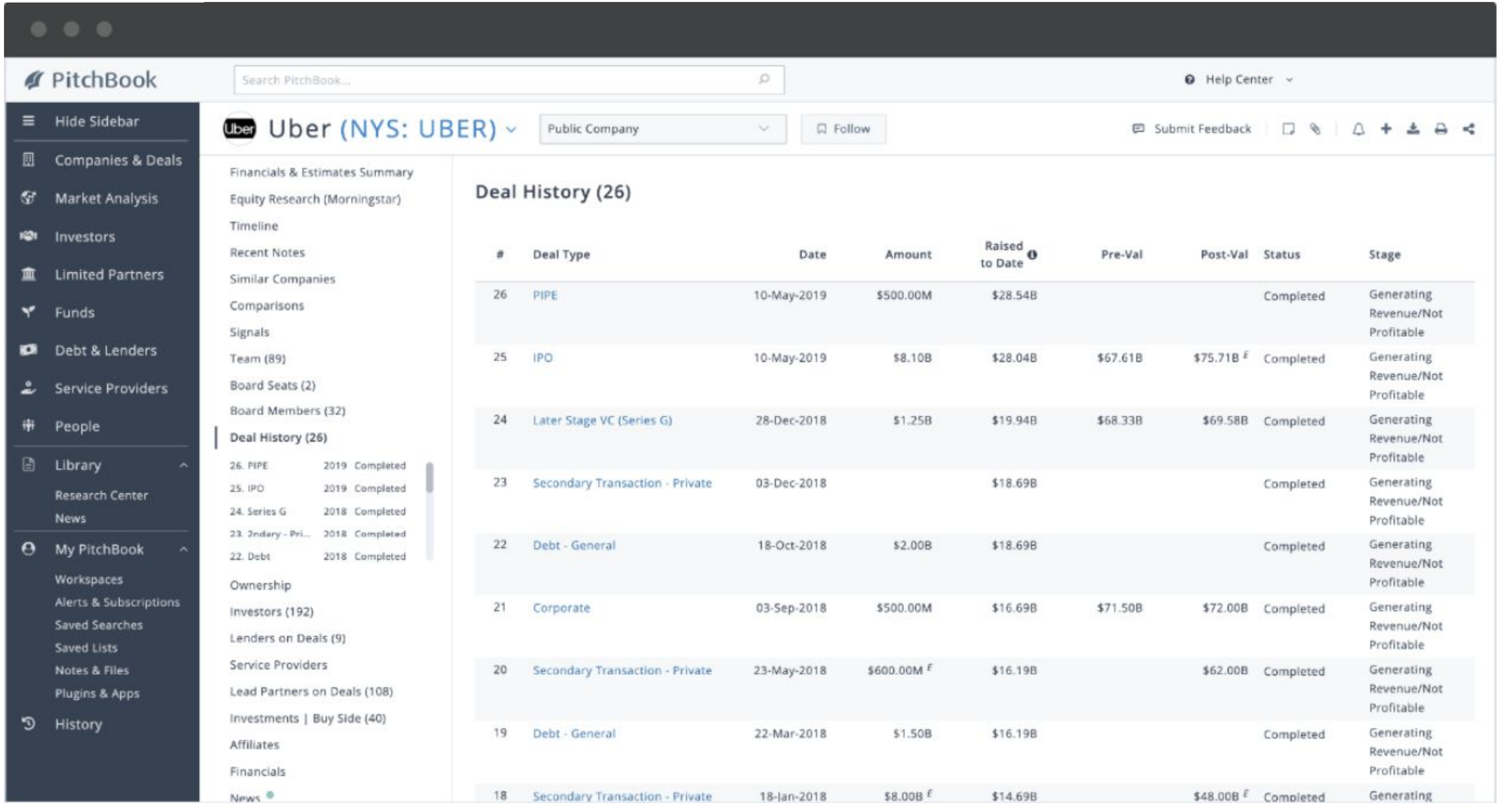
Scraping de 5 fichiers raw :

- **deals.json** (15687 deals)
- **investors.json** (10770 investors)
- (companies.json)
- (advisors.json)
- (funds.json)



Création de la Base de Données : scraping de pitchbook

Scraping de 5 fichiers raw :
- **deals.json** (15687 deals)
- **investors.json** (10770 investors)
- (companies.json)
- (advisors.json)
- (funds.json)



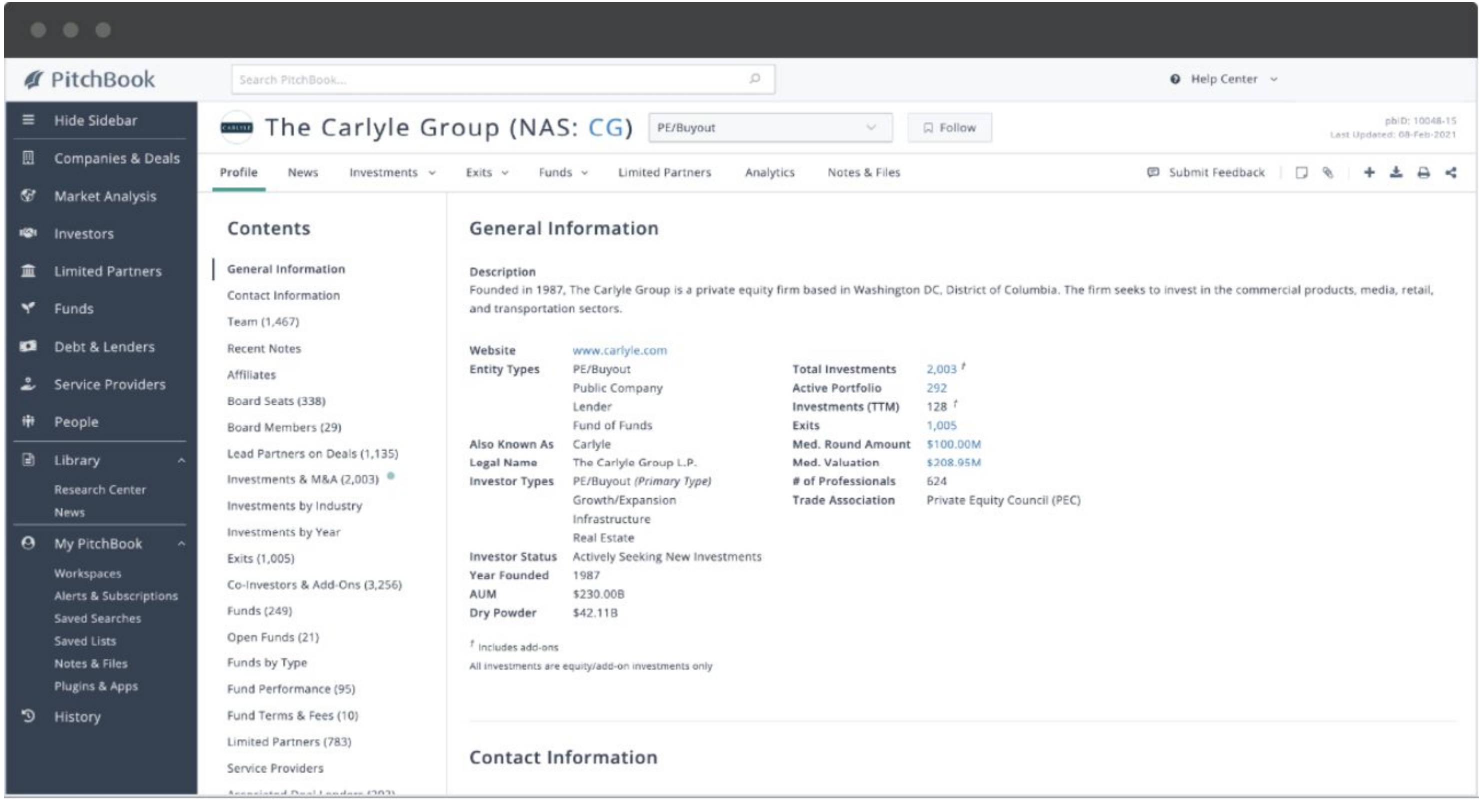
1,614,699 global deals and 465,638 European deals,

Création de la Base de Données : scraping de pitchbook

PitchBook

Scraping de 5 fichiers raw :

- **deals.json** (15687 deals)
- **investors.json** (10770 investors)
- (companies.json)
- (advisors.json)
- (funds.json)



What you can research

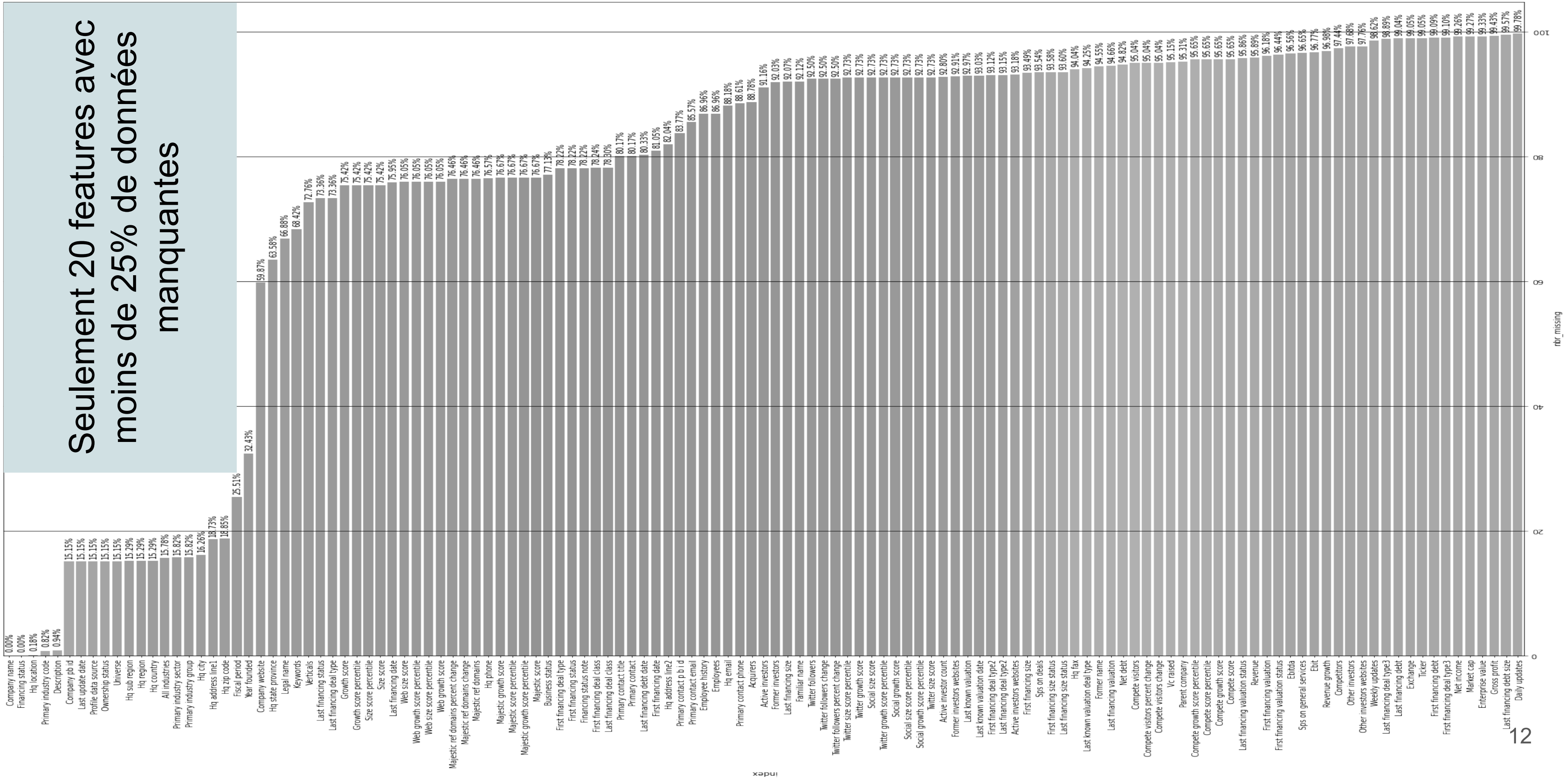
Counts are updated daily

376,425 Investors	39,168 Venture capital firms	19,988 Private equity firms
61,404 Angel investors and angel groups	8,891 Incubators and accelerators	232,346 Strategic acquirers

376,425 global investors and 67,983 European investors,

Création de la Base de Données : données manquantes

Seulement 20 features avec moins de 25% de données manquantes



Création de la Base de Données : exemple de deal


```
'2473": {
    "name": "D2iQ",
    "primaryIndustryGroup": "IT Services",
    "primaryIndustrySector": "Information Technology",
    "dealNo": 5,
    "dealDate": "24-Mar-2016",
    "dealType": "Later Stage VC",
    "Series": "Series C",
    "VCRound": "5th Round",
    "dealSize": 73.57,
    "postValuation": 467.94,
    "impliedEV": "\u00a0",
    "valuationRevenue": -1,
    "ValuationEbitda": -1,
    "impliedEVCashFlow": -1,
    "revenue": -1,
    "ebitda": -1,
    "PercentAcquired": 15.72,
    "investors": "A Capital, a_capital, Andreessen Horowitz (Peter Levine), FUEL Capital, Hewlett Packard Pathfinder, Intel (NAS: INTC),
    "investorCount": 10,
    "FollowOnInvestorCount": 3,
    "FollowOnInvestors": "Andreessen Horowitz, FUEL Capital, Khosla Ventures",
    "firstTimeInvestorCount": 7,
    "FirstTimeInvestors": "A Capital, a_capital, Hewlett Packard Pathfinder, Intel, Microsoft, Triangle Peak Partners, Two Sigma Ventures
    "dealSynopsis": "The company raised $73.5 million of Series C venture funding in a deal led by Hewlett Packard Pathfinder and Microso
    "dealStatus": "Completed",
    "dealFinancingStatus": "Venture Capital-Backed",
    "primaryIndustryCode": "Systems and Information Management",
    "verticals": "Big Data, CloudTech & DevOps, SaaS",
    "description": "Developer of a data center operating system intended to simplify operational efforts for maximum impact. The company
    "hqLocation": "San Francisco, CA",
    "companyWebsite": "www.d2iq.com",
    "hqCountryCode": "us",
    "hqCountryLat": 36.7014631,
    "hqCountryLng": -118.7559974,
    "emission": "Series C",
```

Création de la Base de Données : exemple d’investisseur


```json
{
    "investor_id": "10074-61",
    "investor_name": "Warburg Pincus",
    "description": "Founded in 1966, Warburg Pincus is a private equity firm headquartered in New York, New York. The firm invests in companies based in the Americas, Europe, Middle East, and Asia. The firm seeks to invest in the like energy, financial services, healthcare and consumer, industrial and business services, technology, real estate, media, and telecommunication sectors.",
    "primary_investor_type": "PE/Buyout",
    "aum": 56000.0,
    "last_updated_date": "2020-12-06",
    "year_founded": 1966,
    "hq_location": "New York, NY",
    "investments": 131,
    "active_portfolio": 92,
    "last_investment": {
        "company": {
            "company_id": "51146-20",
            "company_name": "Infoblox"
        },
        "date": "2020-12-01",
        "size": null,
        "valuation": null
    },
    "preferred": {
        "industry": "Buildings and Property, Capital Markets/Institutions, Commercial Banks, Communications and Networking, Consumer Durables, Energy Equipment, Energy Services, Exploration, Production and Refining, Healthcare Devices and Supplies, Healthcare Services, Healthcare Technology Systems, Industrial Supplies and Parts, Insurance, Media, Media and Information Services (B2B), Other Business Products and Services, Other Commercial Products, Other Commercial Services, Other Consumer Products and Services, Other Energy, Other Financial Services, Other Healthcare, Other Information Technology, Other Services (B2C Non-Financial), Pharmaceuticals and Biotechnology, Real Estate Services (B2C), Software, Transportation, Utilities",
        "investment_types": "Buyout/LBO, Early Stage VC, Later Stage VC, PE Growth/Expansion, Seed Round",
        "company_valuation": "100.00 - 15,000.00",
        "deal_size": "50.00 - 500.00",
        "investment_amount": "20.00 - 1,000.00",
        "other_stated_preferences": "Long-Term Investor, Prefers minority stake, Seeks ESG investments, Will syndicate",
        "geography": "Brazil, Canada, China, Europe, Hong Kong, India, Middle East, Singapore, Taiwan, United States, Vietnam",
        "verticals": "AdTech, Advanced Manufacturing, B2B Payments, CleanTech, Construction Technology, Digital Health, FinTech, Gaming, HealthTech, Industrials, Infrastructure, Life Sciences, LOHAS & Wellness, Manufacturing, Real Estate Technology, SaaS, TMT"
    },
    "investments_in_the_last_12_months": 27,
    "keywords": "business model, growth capital, growth investing, hedging fund, venture capital",
    "verticals": "",
    "primary_industry_group": "Capital Markets/Institutions",
    "primary_industry_sector": "Financial Services",
    "website": "www.warburgpincus.com",
    "primary_contact": {
        "name": "Mark M. Colodny",
        "title": "Managing Director, Executive Management, Healthcare, Technology, Media and Telecommunications",
        "phone": "+1 (212) 878-0600",
        "email": "mcolodny@warburgpincus.com"
    }
},
```

Prétraitement des données

1. Extraction et nettoyage des données
   - Transactions de moins de 2 investisseurs
   - Transactions antérieures à 2018
   - Transactions de certains types : PIPE, Secondary Transaction_ Private, Seed Round, Grant and Equity for Service

2. Création de nouvelles features
   - Standardisation des données géographiques en une seule variable : `Hq_Country_Code`
   - Fusion de deux features en élargissant la tolérance suivant une nomenclature `(deal_Type, deal_Serie) => deal_Emission`
   - Génération de l'historique de certaines features (taille de deals, géographie et secteur des investissements, etc)
   - Comptage d'occurrence de certaines features

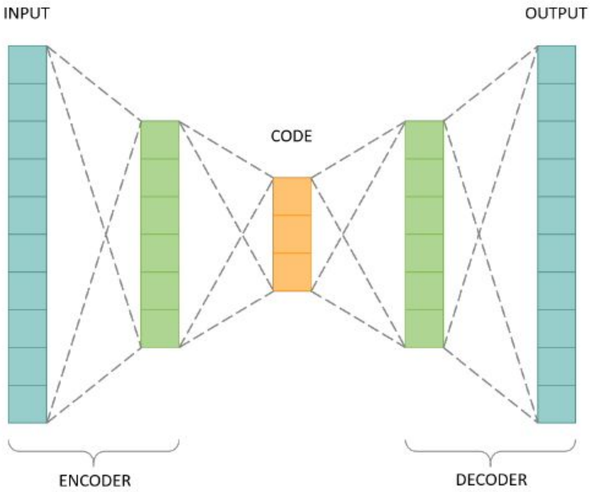
3. Encodage des variables catégorielles

4. Vectorisation des données (Embedding)

Premier Modèle : Auto-Encodeur

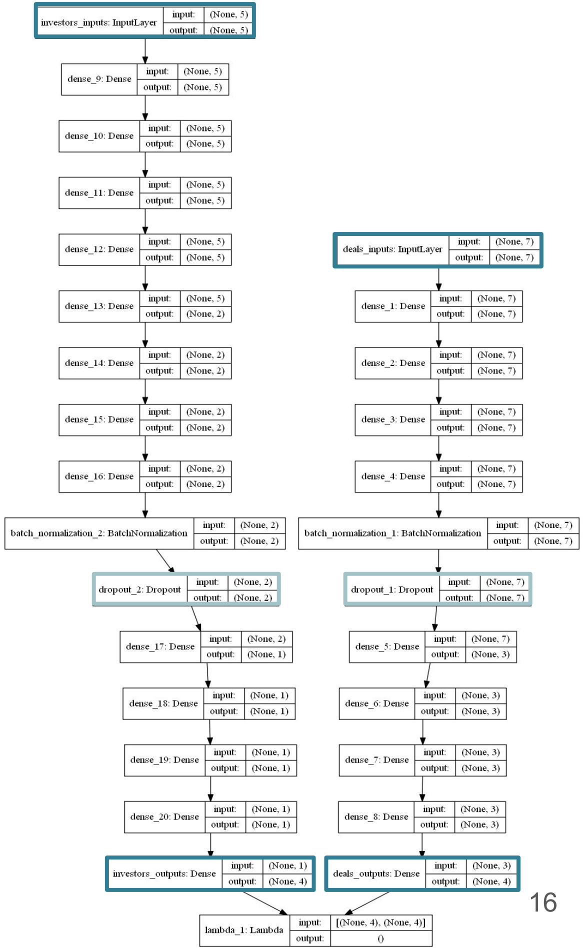
- ❖ Réseau de neurones avec TensorFlow et Keras
- ❖ `Label Encoding` pour features catégorielles puis `MinMaxScaler` pour normaliser les données
- ❖ Split des données : 80% entrainement, 20% validation
- ❖ 8 couches cachées
- ❖ Modèle compilé avec l'optimisation ADAM, la fonction de coût MSE, et la métrique MAE

➔ En sortie de prédiction du modèle, les sorties ne sont pas une liste d'investisseurs mais des poids de **<u>caractéristiques estimées</u>**. Une mesure de distance détermine la similarité pour sortir une liste 100 meilleurs investisseurs suggérés. On calcule la précision du modèle en comparant ces investisseurs avec les investisseurs réels qui ont investi dans la transaction.



7 caractéristiques d'entrée pour les transactions :
- deals_primaryIndustryGroup,
- deals_primaryIndustrySector,
- deals_primaryIndustryCode,
- deals_emission,
- deals_dealSize,
- deals_hqCountryCode,
- deals_FollowOnInvestors,

et 5 caractéristiques d'entrée pour les investisseurs :
- investors_primaryInvestorType,
- investors_histIndustryGroup,
- investors_histHqCountryCode,
- investors_histPrimaryIndustryCode,
- investors_hqCountryCode.



Deuxième Modèle : PCA

- **<u>Seules les données des transactions sont utilisées</u>**

- <u>6 Inputs</u> : 5 caractéristiques catégorielles `('deals_primaryIndustryGroup', 'deals_primaryIndustrySector', 'deals_primaryIndustryCode', 'deals_emission', 'deals_hqCountryCode')` + 1 caractéristique numérique `('deals_dealSize')`
  <u>1 Output</u> : **`'deals_investors'`**

- <u>Prétraitement</u> : Duplication des lignes de variables catégorielles à champs multiples de one-to-many à one-to-one,
  Dropna (35619 -> 17855 lignes),
  Label Encodage des variables catégorielles => **X[17855, 6] <u>user matrix</u>**

- Construction de la matrice de correspondance OneHot = [17855, 13842] (=[len(deals_investors), len(index_to_investor))
- Factorisation et compression - Embedding avec SVD/PCA : **OneHot ~ Y[17855, 200]** x S x E[200, 13842]
- Reconstruction de chaque ligne (transaction) => 100 coefficients les plus élevés = liste de prédiction

```
Total recall:  71.4 %
Total accuracy:  4.5 %
Nb of non-null predictions:  13219 over 17855 (74.0%)
```

- On ne garde que les deals dont la prédiction est non-nulle lors la reconstruction (X -> X_sliced, Y -> Ysliced)
- Train-test split (80-20) : **X_sliced[13219, 6]** -> X_train[10575, 6] X_test[2644, 6]
  Y_sliced[13219, 200] -> Y_train[10575, 200] Y_test[2644, 200])

- Y_prediction = Regression_Model(X, Y) => reconstruction OneHot_prediction => 100 coefficients les plus élevés = liste de prédiction

Deuxième Modèle : PCA

- 2 différents modèles :
  - Arbre de régression à gradient boosté (XGBoost)
  - Réseau de neurones à propagation avant (FFNN)

- 4 tests différents :

	Résultats
FFNN (sortie : liste de 100 prédictions)	4525 intersections non-nulles sur 10575 (42.8%) => 25.8 % rappel total
FFNN (sortie : liste de 1000 prédictions)	8173 intersections non-nulles sur 10575 (77.3%) => 44.7 % rappel total
XGBoost (sortie : liste de 100 prédictions)	4585 intersections non-nulles sur 10575 (43.4%) => 25.8 % rappel total
XGBoost (sortie : liste de 1000 prédictions)	8265 intersections non-nulles sur 10575 (78.2%) => 45.2 % rappel total

Troisième Modèle : LGBM

- Méthode d'ensemble de type régression d'arbre de décision optimisé (Implémentation LightGBM du gradient boosting)
- Création de **nouvelles features** (`Featurer`) :
    - **Temporalité** : comptage - des occurrences de différentes features sur la dernière année (`DealCounter`)
      - du nombre de jours depuis la dernière transaction pour des features spécifiques (`DaysSinceLastDealCounter`)
- `aikit NumericalEncoder` pour features catégorielles puis MinMaxScaler pour normaliser les données
- <u>Training</u> :
    - **L'ensemble des investisseurs est labellisé, suivant deux classes, ayant investi ou non**
    - **On sélectionne aléatoirement 2% des données pour l'entraînement du modèle (limite les ressources temps de calcul, rééquilibre la classe minoritaire)**

➔ AUC : 90.9%
Average Precision : 3.8%
Mean AUC : 89.7%
Mean Average Precision : 8.7%
Mean Recall : **35.9%**

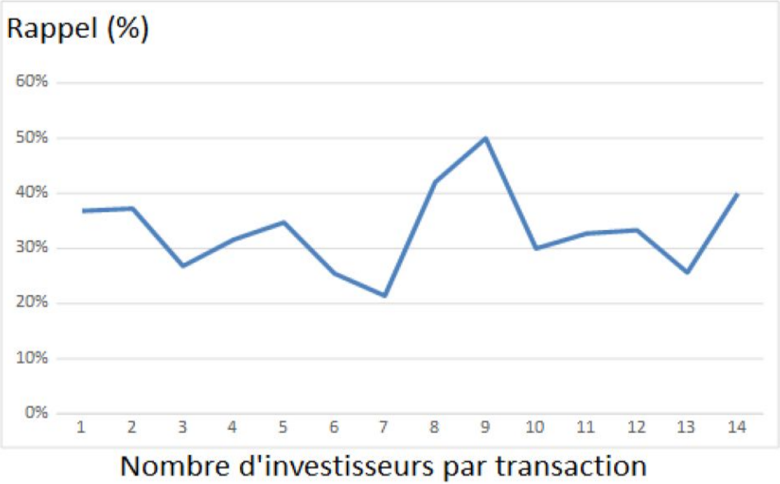
Comparaison des trois approches

lgbm est favorite (création de nouvelles features)

	Précision Moyenne (%)	Rappel Moyen (%)
Méthode 1 (Auto-Encodeur)	10	17
Méthode 2 (SVD)	2	28
Méthode 3 (Lgbm)	**10**	**38**

Troisième Modèle : Variations du Rappel

Case 1	Case 2	Case 3	Case 4
Original Database + drop incomplete deals	Enlarged Database (including small deals) + drop incomplete deals	Enlarged Database (including small deals)	Enlarged Database (including small deals) + historical new features
Keeping 10487/18828 deals	Keeping 23048/61919 deals	Keeping 53412/61919 deals	Keeping 53412/61919 deals
25.5% mean Recall	27.8% mean Recall	33% mean recall	**37.5% mean recall**



Troisième Modèle : Explicabilité

- Comprendre le modèle : pourquoi une recommandation a été faite et l'importance des différents facteurs contribuant à ce choix

- Librairie **SHAP** d'**explication des caractéristiques** (*i.e. SHAP, SHapley Additive exPlanations*) quantifie la contribution que chaque feature apporte à la prédiction effectuée par le modèle.

- Exemple : entreprise Linkcy pour laquelle on a vérifié un problème de **surreprésentation d'investisseurs locaux** dans les recommandations

- Trouver ajustement de l'importance des features dans les prédictions en accord avec nos besoins business

FEATURE	OCCURRENCE
Investor located in company state	96
Number of deals last year	90
Company state	76
Investor located in company region	58
Number of deals on country last year	34
Number of deals of deal type last year	28
Days since last deal on industry sector last year	24
Days since last deal on deal type last year	23
Days since last deal on region last year	21
Investor located in company city	19
Days since last deal last year	14
Investor AUM	6
Number of deals on verticals last year	6
Number of deals on region last year	2
Deal size (m$)	1
Investor age (in years)	1
Number of deals on industry sector last year	1

Troisième Modèle : Analyse des prédictions nulles

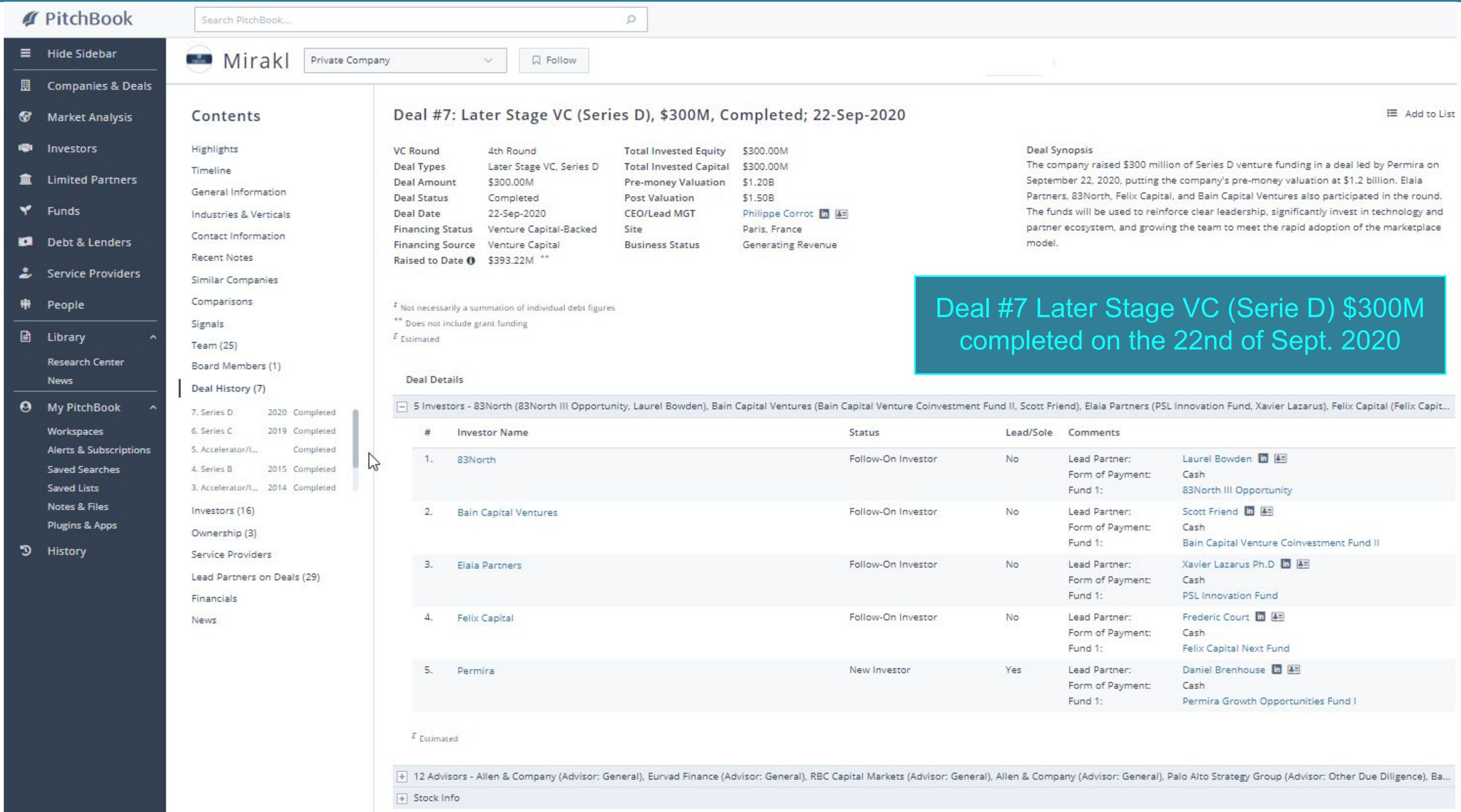
Analyse séries

	recall = 0	recall = 0 (%)	Total	Total (%)	recall = 0/Total
Series A	63	52 %	123	42 %	51 %
Series B	33	27 %	78	27 %	42 %
Series C	15	12 %	56	19 %	27 %
Series D	5	4 %	21	7 %	24 %
Series E	5	4 %	13	4 %	38 %
Series F	0	0 %	2	1 %	0 %
Series H	0	0 %	1	0 %	0 %
	121		294		

Le nombre de prédictions nulles (prédictions ne contenant aucun bon investisseur) décroît au fur et à mesure du nombre de tours de financement (de la série A à H).

On peut donc comprendre que les caractéristiques intégrant l'historique des transactions pour une même société joue un rôle important dans la qualité des prédictions des investisseurs.

Troisième Modèle : Cas d’usage - entreprise Mirakl



Deal #7 Later Stage VC (Serie D) $300M completed on the 22nd of Sept. 2020

Troisième Modèle : Cas d’usage - entreprise Mirakl

Input Parameters

```
possible_deal_type:  ['Later Stage VC', 'Restart - Later VC']
possible_primary_industry_group:  ['Media', 'Software', 'Other Information Technology', 'Commercial Services', 'Other Business Products and Services']
possible_verticals:  ['E-Commerce', 'SaaS', 'TMT']
{
    "auc": 0.9070784667353022,
    "ap": 0.04482640137759274,
    "mean_auc": 0.896487725303293,
    "mean_ap": 0.09180516231395112,
    "mean_recall": 0.35786888633188446
}
```

Résultats d'entraînement du modèle

```
deal = {
'deal_type': possible_deal_type_choice,
'series': 'D',
'vc_round': '4th',
'deal_size': 300,
'percent_acquired': 0.3,
'company_id': 'Unknown',
'primary_industry_sector': 'Business Products and Services (B2B)',
'primary_industry_group': possible_primary_industry_group_choice,
'company_hq_location': 'Paris, France',
'verticals': possible_verticals_choice,
}
```

=> 2 x 5 x 3 = **30 listes de 100 investisseurs prédits**

Model pipeline

```
self.pipeline = make_pipeline(
    Featurer(),
    DealCounter(),
    DealCounter(['deal_type']),
    DealCounter(['series']),
    DealCounter(['vc_round']),
    DealCounter(['company_id']),
    DealCounter(['primary_industry_sector']),
    DealCounter(['primary_industry_group']),
    DealCounter(['company_hq_country']),
    DealCounter(['company_hq_region']),
    DealCounter(['verticals']),
    DaysSinceLastDealCounter(),
    DaysSinceLastDealCounter(['deal_type']),
    DaysSinceLastDealCounter(['primary_industry_sector']),
    DaysSinceLastDealCounter(['primary_industry_group']), # added
    DaysSinceLastDealCounter(['company_hq_region']),
    SelectiveNumericalEncoder(self.categories),
    ColumnsSelector(columns_to_drop=[
        'deal_id', 'deal_date', 'invest',
        'company_id', 'company_name',
        'investor_id', 'investor_name', 'investor_hq_location'
    ])
)
```

Troisième Modèle : Cas d'usage - entreprise Mirakl

5 Investisseurs réels (ground truth)

last_updated_date	investor_name	aum	investments	investments_in_the_last_12_months	active_portfolio	aum_per_investments	aum_per_active_portfolio	hq_location
2020-12-03	83North	1000.00	59	19.0	35.0	16.95	28.57	London, United Kingdom
2020-12-04	Bain Capital Ventures	5200.00	130	32.0	77.0	40.00	67.53	Boston, MA
2020-12-02	Elaia Partners	499.82	7	3.0	5.0	71.40	99.96	Paris, France
2020-12-07	Felix Capital	600.00	29	9.0	21.0	20.69	28.57	London, United Kingdom
2020-12-06	Permira	56407.71	10	4.0	8.0	5640.77	7050.96	London, United Kingdom

2 Investisseurs réels trouvés dans les suggestions

	investor_id	investor_name	primary_investor_type	aum	investor_hq_location	investor_hq_country	investor_hq_region	probability	occurrence
25	11170-63	Elaia Partners	Venture Capital	499.82	Paris, France	France	Europe	0.020360	18
118	11287-81	Permira	PE/Buyout	56407.71	London, United Kingdom	UK	Europe	0.009983	1

Parmi les 30 listes de 100 investisseurs, on a **190 investisseurs uniques suggérés** et triés par probabilité.
2 investisseurs (Elaia Partners and Permira**) sont trouvés parmi les 5 investisseurs réels**
(comme attendu, le score de prédiction est rappel=40%).

Système de recommandation d’investisseurs :

Conclusions et perspectives

Système de Recommandation : Conclusion et Perspectives

Conclusion

- Scraping des données Pitchbook de transactions pour le financement des entreprises à l’échelle globale
- 3 approches RecSys totalement différentes ont été développées, elles sont basées sur :
  - un auto-encodeur
  - une PCA
  - **Régression d’arbre de décision à gradient boosté**
- Importance de la **création de nouvelles features** avec statistiques (rapport de features, comptages, temporalité)
- Présence de nombreuses sorties non-pertinentes. Minimiser l’occurrence de faux-positifs dans les résultats
- Améliorer le modèle d'extraction : Filtrage des investisseurs non pertinents pour la transaction afin d’optimiser ensuite l'étape de ré-équilibrage des classes.
- Cas d’usage -> bon accueil par une communauté de professionnels tels que BNP Paribas (*early adopter*)

Perspectives

- Automatisation et actualisation fréquente de la base de données afin de ré-entraîner régulièrement les modèles
- Revue régulière de la littérature concernant les méthodes (Embedding / Categorical Encoding Methods)
  Modèles de Word Embedding (Word2Vec, glove). Bonne représentation vectorielle des mots du champs lexical spécifique
- L’influence contextuelle relative aux critères géographiques doit être mieux maîtrisée. On doit pouvoir demander au système une flexibilité dans le paramétrage afin d’obtenir des recommandations plus ou moins localisées dans certaines régions, en fonction du choix de l’utilisateur.
- Considérations business pour augmenter les performances de la solution actuelle

3. Classification de commentaires d’investisseurs

Classification de commentaires d’investisseurs : roadmap

- **Création d’un corpus de texte non-labellisé** :
  - Extraction et prétraitement du texte RAW à partir de fichiers .pdf (`tika parser`)
  - Extraction des phrases à partir du texte (`nltk tokenizer`)
  - Nettoyage final du corpus (lowercase, borner le nombre de mots par phrase, etc)

- Utiliser **des catégories déterminées** par l’équipe business (50 classes qualitatives)

- **Classification Zero-Shot** de BERT (*Bidirectional Encoder Representations from Transformers*) (`transformers "zero-shot-classification" pipeline`)
  - L’apprentissage Zero-Shot : On force le réseau de neurones à effectuer la classification dans des classes pour lesquelles il n’a jamais été entraîné. Modèle pré-entraîné sur un corpus de texte de taille massif (*i.e.* `BartModel`, `BertForClassification`).
  - Possibilité immédiate de faire de la classification Force / Faiblesse (`transformers "sentiment-analysis" Text Classification pipeline`)

Classification de commentaires d’investisseurs : cas d’usage

- Une analyse de perception permet d’extraire un corpus de 3802 phrases
- Le pipeline suppose par défaut qu'une seule des étiquettes candidates est vraie, renvoyant une liste de scores pour chaque étiquette qui totalisent 1 (multi_class=False). <u>On utilise l’option Multi-classe</u> (multi_class=True) pour laquelle les scores sont indépendants, mais chacun est compris entre 0 et 1.

```
00:02:07   ---   301 / 3802
====================================   301   ================================
they execute well on their strategic m&a acquisitions, how they manage the business and how they integrate what they acquire etc - bdl the lacklustre organic growth is because of both the markets they are in
and also management.
```

{'sequence': 'they execute well on their strategic m&a acquisitions, how they manage the business and how they integrate what they acquire etc - bdl the lacklustre organic growth is because of both the
markets they are in and also management.', 'labels': ['Mergers and Acquisitions', 'Market exposure', 'Management', 'Strategy execution', 'Industry dynamic', 'Deal structure', 'Business model', 'Track record',
'Size ou Scale', 'Share liquidity', 'Leadership', 'Backlog', 'Industry specificities', 'Social', 'Return to shareholders', 'Geographical specificities', 'Industry capacity', 'Financial Position',
'resiliency', 'Overhang', 'Geographical diversification', 'Governance', 'Patents', 'Leverage ', 'Pricing power', 'Shareholders', 'Environment', 'Technology', 'Currency', 'Capital intensity', 'Revenues KPI',
'Margin evolution', 'Regulation', 'Macro economics', 'Dilution', 'Revenues growth', 'Profitability', 'Innovation', 'Digitalisation', 'Barriers to entry', 'Product diversification', 'Competition', 'Cyclicality
', 'Cash-flow', 'Commodity prices', 'Research and Development', 'Restructuring', 'Geopolitics', 'Interest rates', 'Cost inflation'], 'scores': [0.992043137550354, 0.9834339022636414, 0.9633303284645081,
0.9348139762878418, 0.9016832709312439, 0.7484057545661926, 0.6934672594070435, 0.6770192384719849, 0.5795993804931641, 0.5737428069114685, 0.5387290120124817, 0.5325745940208435, 0.45081984996795654,
0.4351787567138672, 0.4317941963672638, 0.3716159760951996, 0.35180842876434326, 0.3291803002357483, 0.31250181794166565, 0.3036833703517914, 0.2940145432949066, 0.2746160328388214, 0.27413833141326904,
0.27213940024375916, 0.26324185729026794, 0.25546085834503174, 0.2510124742984772, 0.23918011784553528, 0.23427194356918335, 0.23026570677757263, 0.21836815774440765, 0.20879895985126495, 0.17868317663669586,
0.15872441232204437, 0.1574268490076065, 0.1427176594734192, 0.12703293561935425, 0.1246345117688179, 0.11663249880075455, 0.10842921584844589, 0.09768812358379364, 0.09149441123008728, 0.07520044595003128,
0.0654749646782875, 0.05744219943881035, 0.04376499727368355, 0.028134217485785484, 0.022899789735674858, 0.014329720288515091, 0.004501082003116608]}

Classification de commentaires d'investisseurs : cas d'usage

Elis Perception study 01.02.19 - Class occurrence for best classification score

resiliency	656
Market exposure	541
Industry dynamic	524
Return to shareholders	379
Social	301
Mergers and Acquisitions	198
Track record	196
Management	189
Deal structure	188
Leverage	171
Cash-flow	152
Size ou Scale	149
Capital intensity	148
Geographical specificities	147
Business model	109
Shareholders	108
Pricing power	100
Strategy execution	78
Profitability	73
Share liquidity	69
Financial Position	64
Geographical diversification	62
Interest rates	61
Industry specificities	47
Barriers to entry	40

Margin evolution	38
Overhang	37
Leadership	35
Cost inflation	32
Competition	30
Dilution	24
Governance	20
Revenues growth	16
Cyclicality	14
Restructuring	14
Macro economics	14
Currency	11
Environment	11
Innovation	9
Industry capacity	6
Backlog	6
Technology	6
Commodity prices	4
Product diversification	4
Geopolitics	2
Revenues KPI	2
Regulation	1

No 'best_score' occurrence for {'Digitalisation', 'Patents', 'Research and Development'}

Classification de commentaires d’investisseurs : cas d’usage

```
output.best_class.value_counts()

Industry dynamic              621
resiliency                    467
Market exposure               323
Management                    289
Social                        269
Mergers and Acquisitions      213
Capital intensity             106
Cash-flow                     104
Pricing power                  90
Geographical specificities     89
Geographical diversification   79
Leverage                       79
Shareholders                   73
Business model                 73
Track record                   71
Size ou Scale                  65
Financial Position             64
Leadership                     59
Overhang                       59
Return to shareholders         56
Environment                    51
Profitability                  50
Interest rates                 49
Share liquidity                37
Barriers to entry              36
Strategy execution             34
Deal structure                 33
Industry specificities         31
Technology                     29
Margin evolution               29
Cost inflation                 25
Governance                     22
Competition                    21
Macro economics                15
Dilution                       14
Revenues growth                12
Product diversification        11
Cyclicality                    10
Innovation                      8
Restructuring                   7
Currency                        6
Backlog                         6
Geopolitics                     5
Industry capacity               5
Commodity prices                3
Revenues KPI                    3
Regulation                      1
Name: best_class, dtype: int64
```

best_class	sentence	max_score
Cash-flow	these type of things, we can see it, but our first focus is in terms of cash flow.	84.0
Cost inflation	then there is the cost inflation, so they have to deal with some market topic.	84.0
resiliency	the second thing is probably the resilience, except probably the hospitality part of the business.	82.3
Governance	we rank governance quite high at 7.5, so 7 or 8 is fine for me.	81.0
Social	i am talking about the social background, which will impact the economy.	78.8
Leverage	leverage is one of these things where people think you should have more and more leverage and then they suddenly decide you should not have leverage and those two things can be a week apart.	78.0
Shareholders	in this area, we are shareholders, as you may know.	75.7
Capital intensity	i suppose a weakness you could point to on the strategy is that it is capital-intensive, although that is a barrier, so it can have a reasonable level of debt.	75.0
Financial Position	perception study - january 2019 financial position (cont.)	70.9
Revenues growth	not so good at organic revenue growth - dk partners i like them very much.	67.8
Commodity prices	the last negative is just in terms of raw materials; it does impact them, any movement in oil price or cotton price and stuff - franklin templeton we worry in general that elis is very focused on maximising its cost base.	66.2
Management	perception study - january 2019 management (cont.)	66.0
Mergers and Acquisitions	in order to get the top line going, you need to invest, and it is stuff that for the first few years just does not generate a lot of profitable margin and so he is just more focused on m&a.	64.8
Restructuring	it is just execution on the restructuring and not dropping the ball on the organic growth.	64.4
Strategy execution	on a scale of 1 to 10 how do you rate the execution of elis' strategy?	61.3
Business model	the first strength is probably the business model, but business model is quite a large word.	59.2
Interest rates	the point is more market perception and with interest rates rising and maybe market slowdown, there are many fears around high indebtedness.	58.9
Leadership	as i said at the beginning, it is a leader.	55.5
Market exposure	not on the management, but on the market, on this balance sheet structure, on the uk exposure, which is part of the business - lfde on the negatives, it is hard to know, berendsen might turn out to be brilliant or it might turn out to be a step too far.	54.2
Profitability	in terms of profitability, they have been doing quite well and everything is fine, according to me.	52.3
Geographical specificities	at this point, i think there is no particular mis-executions on their different geographies, in terms of integration of acquisitions, for example.	52.1
Track record	basically, if i summarise it, good execution in france, a good track record there of running the operations.	51.3
Pricing power	if they are unable to pass it through, there is a question around their pricing power and there is a question around the competitive dynamics, so that is another point.	51.3
Return to shareholders	forget dividends, i would do a buyback.	46.7
Industry dynamic	it is an industry in consolidation, then consolidation brings higher margins, higher returns.	46.4
Dilution	in terms of their strategy so far, the issue here is that they said we are going to consolidate markets and see margin expansion and i feel like that is being diluted a little bit.	46.1

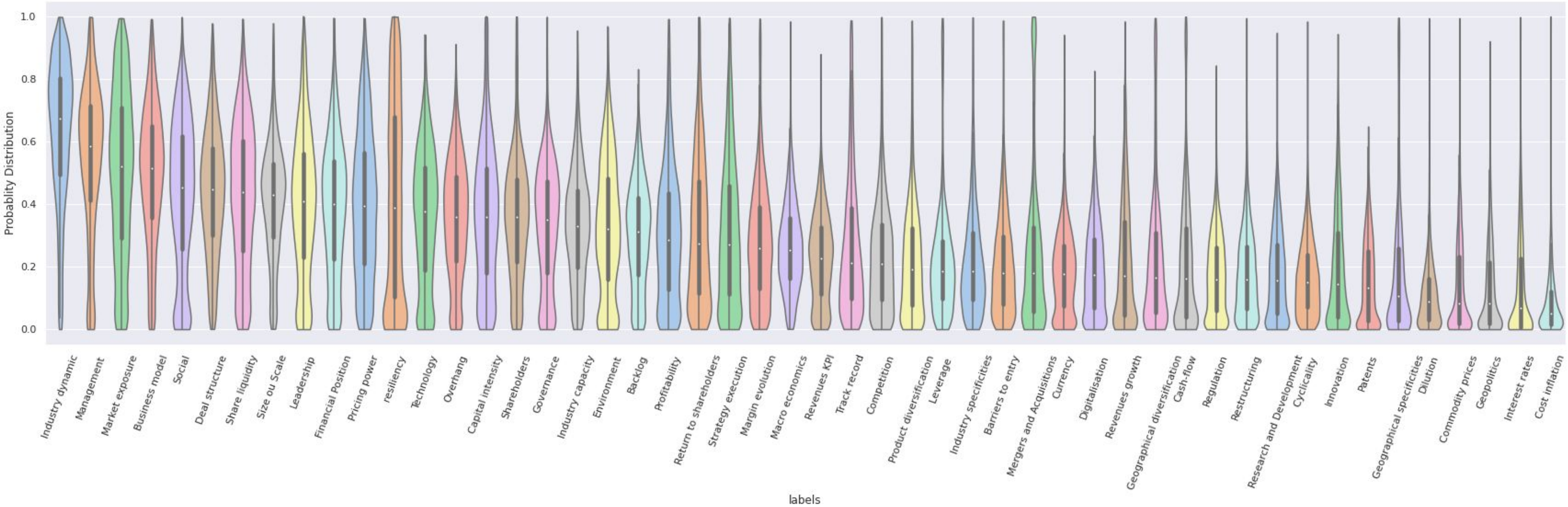
Classification de commentaires d’investisseurs : cas d’usage

```
output.best_class.value_counts()
```

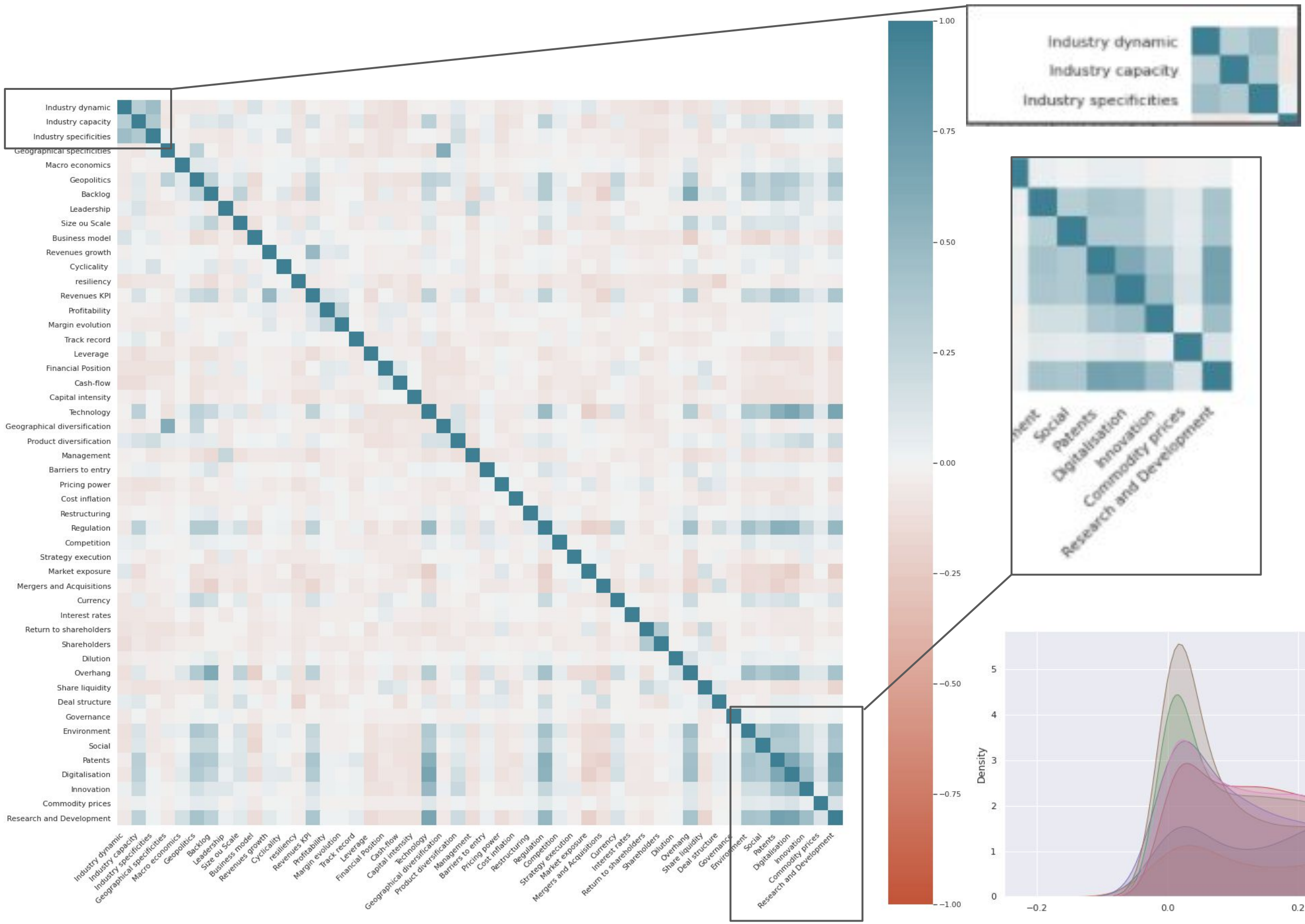
```
Industry dynamic               621
resiliency                     467
Market exposure                323
Management                     289
Social                         269
Mergers and Acquisitions       213
Capital intensity              106
Cash-flow                      104
Pricing power                   90
Geographical specificities      89
Geographical diversification    79
Leverage                        79
Shareholders                    73
Business model                  73
Track record                    71
Size ou Scale                   65
Financial Position              64
Leadership                      59
Overhang                        59
Return to shareholders          56
Environment                     51
Profitability                   50
Interest rates                  49
Share liquidity                 37
Barriers to entry               36
Strategy execution              34
Deal structure                  33
Industry specificities          31
Technology                      29
Margin evolution                29
Cost inflation                  25
Governance                      22
Competition                     21
Macro economics                 15
Dilution                        14
Revenues growth                 12
Product diversification         11
Cyclicality                     10
Innovation                       8
Restructuring                    7
Currency                         6
Backlog                          6
Geopolitics                      5
Industry capacity                5
Commodity prices                 3
Revenues KPI                     3
Regulation                       1
Name: best_class, dtype: int64
```

Barriers to entry	you can see that with mainly distributors, like bunzl, but the barrier to entry is even higher, which is even more positive.	45.3
Competition	there is competition, but the berendsen brand is a strong brand, so there is no reason why it should not continue.	42.2
Industry specificities	there are no synergies, that is exactly what we like in this industry.	41.7
Cyclicality	they are going to face some cost pressures in the short-term, but we see that more cyclical.	41.1
Geographical diversification	regarding the geography, the business is more diversified now after the different acquisitions they made this past three years.	40.8
Macro economics	we will see how they are able to continue to grow in this different, maybe more complicated macroeconomic context.	39.7
Margin evolution	it is a good mix between the organic sales growth and the margin evolution.	36.2
Size ou Scale	that goes then in pair with doing bolt-on acquisitions or a larger scale deal, but i do not think larger deals are on the horizon right now.	30.4
Share liquidity	that is the struggle the shares have at the moment.	28.4
Product diversification	regarding the business, they are quite balanced between workwear, linen, etc.	26.3
Currency	after that, it is also about the free cash flow generation and the roce.	24.1
Deal structure	keep doing small deals, rather than a very big structuring deal, leading to higher indebtedness and maybe capital increase.	23.3
Industry capacity	i think it is going to change now that, for example, they have hired a new ir.	21.1
Technology	seven, for their technical capacity.	18.3
Innovation	for example, for me, indusal was a material acquisition.	17.7
Revenues KPI	there is only a revenue number, so i do not really have much to say about it.	15.6
Environment	you are in a different environment.	13.6
Overhang	uk, that is a question mark.	11.9
Geopolitics	the prospects of a more challenging macro economic / geopolitical environment does not help.	10.9
Backlog	in uk, i hope they could restore the growth base in the future.	10.1
Regulation	i cannot express these kind of views for compliance reasons.	6.7

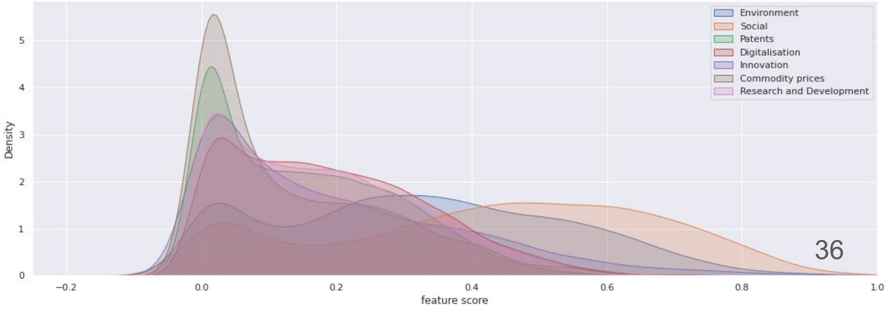
Classification de commentaires d’investisseurs : cas d’usage



Classification de commentaires d’investisseurs : cas d’usage



labels	labels	
Revenues KPI	Revenues growth	0.783029
Geographical diversification	Geographical specificities	0.774720
Governance	Regulation	0.767967
Management	Leadership	0.755918
Technology	Digitalisation	0.750049
Regulation	Patents	0.732459
Digitalisation	Research and Development	0.727452
Social	Environment	0.727086
Overhang	Backlog	0.691443
Technology	Research and Development	0.681035
Profitability	Margin evolution	0.680931
Digitalisation	Regulation	0.679336
	Innovation	0.677048
Technology	Innovation	0.675495
Business model	Deal structure	0.674131
Digitalisation	Patents	0.668108
Social	Governance	0.664948
Regulation	Research and Development	0.662363
Product diversification	Geographical diversification	0.661608
Environment	Regulation	0.656419
Governance	Environment	0.655406
	Patents	0.647457
Product diversification	Innovation	0.644897
Market exposure	Share liquidity	0.641109
Backlog	Patents	0.634053
Social	Technology	0.632980
Research and Development	Patents	0.632525
Currency	Patents	0.630769
Size ou Scale	Backlog	0.629644
Share liquidity	Pricing power	0.624484
Management	Governance	0.624161
Leadership	Governance	0.623814



Classification de commentaires d’investisseurs :

Conclusions et perspectives

Classification de commentaires d'investisseurs

Conclusion

- Création corpus propre, non-labellisé
- Classification multi-classes (multi_class=True) pour prendre en compte des sous-catégories
- Peut être combinée avec l'analyse de sentiments pour une classification quantitative (positif/négatif) force/faiblesse
- Summarizer (`gensim, lexrank, lsa, textrank, luhn`)
- Une liste exhaustive de catégories à quatre niveaux hiérarchiques a été établie et servira de base pour les travaux futurs

Perspectives

- Prospection de différentes méthodes de classification multi-classes hiérarchisées
- Création en cours d'un corpus labellisé, de grande taille, et spécialisé au problème à l'aide de la plateforme https://prodi.gy/ développée par les créateurs de `SpaCy`
    - => Affinage du modèle pré-entrainé
    - => Mieux comprendre les problèmes et interpréter les résultats.