Création d’un modèle Machine Learning de prédiction des fermetures d’entreprises

Christophe Brun - Le Wagon - 4 juillet 2024

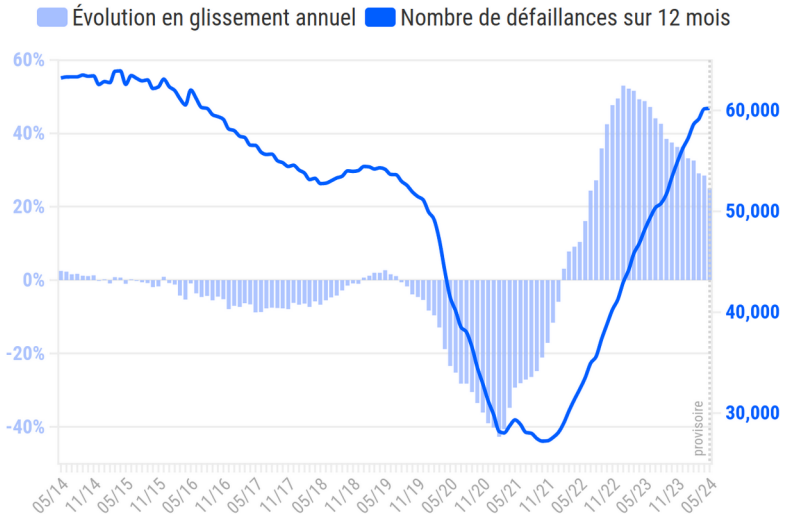
Sommaire

- Problématique et postula
- Data cleaning
- Création du modèle
- Analyse du modèle
- Prédiction
- Conclusion

Problématique et postula

Les finances des entreprises sont la principale cause de l’augmentation des fermetures que l’on observe actuellement.

Plus 27,7 % sur l’année selon la baromètre des défaillances Société.com.



Problématique et postula

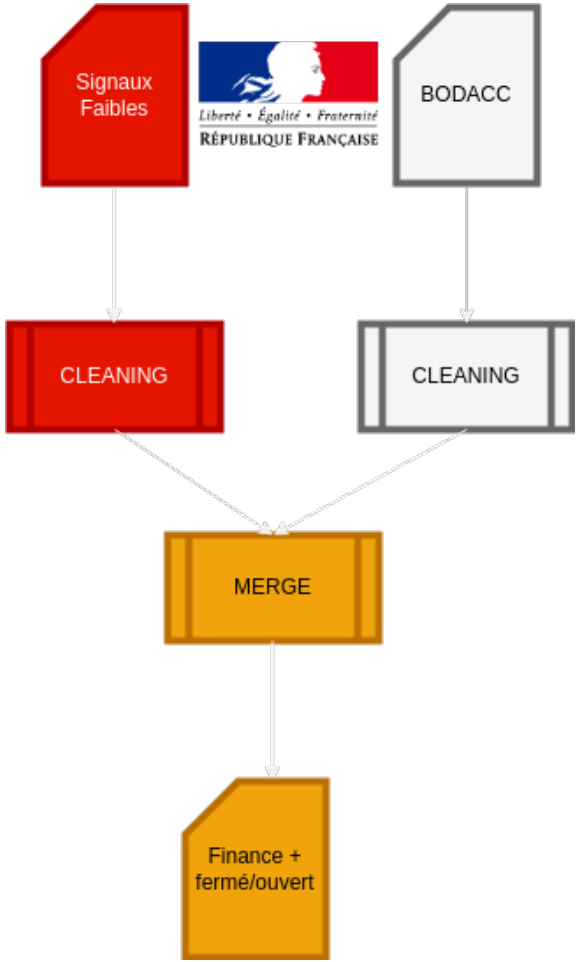
L’enjeu est d’anticiper les fermetures d’entreprises pour informer les acteurs économiques, les investisseurs, banquiers, assureurs, décisionnaires au sens large, *etc*.

Si les fermetures d’entreprises sont motivées par l’aspect financier principalement, avec les données des tribunaux de commerce qui publient les comptes et le BODACC qui publie les fermetures. On doit pouvoir créer un modèle prédictif.

Data cleaning

Consiste à extraire des jeux de données du BODACC et des tribunaux de commerce/INPI/Signaux Faibles seulement les données utiles et les mettre en forme.

- BODACC : *Le bulletin officiel des annonces civiles et commerciales.*
- Signaux Faibles : *Valoriser la richesse des données administratives pour produire un outil d'analyse prédictive des difficultés des entreprises.*



Data cleaning

- 2,2 Go de data dans un fichier parquet. Découpé avec Pyarrow en chunk de 100 000 lignes car ne rentre pas en RAM.
- Sur chaque chunk, filtration des données sur les 3 ans de l'entraînement.
- Sur chaque chunk, la date devient un entier, l'année de la liasse.
- Sur chaque chunk, les liasses fiscales, i.e., les *features* sont mise en forme et sélectionnées ou non.
- Sur chaque chunk, suppression des colonnes.
- Pivot de tous les chunks concaténés et nommage des colonnes avec le pattern liasse_<code liasse>_<année target – année liasse>.

Data cleaning

- Dataset de 31 Mo dans un parquet.
- 15 *features* explicatives, 5 liasses sur 3 ans.
- L’identifiant de Société, le SIREN, en index.

siren	liasse_fy_3	liasse_fy_2	liasse_fy_1	liasse_fz_3	liasse_fz_2	liasse_fz_1	liasse_fj_3
5420120	656957	490913	449215	279954	272761	224681	342381
5520176	1820749	1686707	1792024	765906	719568	868025	6082217
5520242	792695	0	847142	209864	0	253325	4874985
5580113	31493	0	0	15669	0	0	3321087
5580501	0	0	0	0	0	0	0

Data cleaning

- 654 Mo de data dans un fichier parquet. Découpé avec Pyarrow en chunk de 100 000 lignes car ne rentre pas en RAM
- Sur chaque chunk, filtration des données sur la seule année suivant les 3 ans des features.
- Sur chaque chunk, la date devient un entier, l'année de la fermeture.
- Sur chaque chunk, on ne garde que les numéros SIREN des entreprises fermées.
- Sur chaque chunk, suppression des colonnes.

Data cleaning

- L’identifiant de Société fermée, le SIREN, en index.
- Merge avec le dataset des données financières.
- Cast du booléan en UInt8 sans impact car PyCaret fait le *label encoding*.

registre
845122308
809696826
513412387
850572967
404955080

liasse_fx_3	liasse_fx_2	liasse_fx_1	closed
123770	104216	103060	0
210981	151987	137289	0
188193	0	114651	0
5687	0	0	0
0	0	0	0

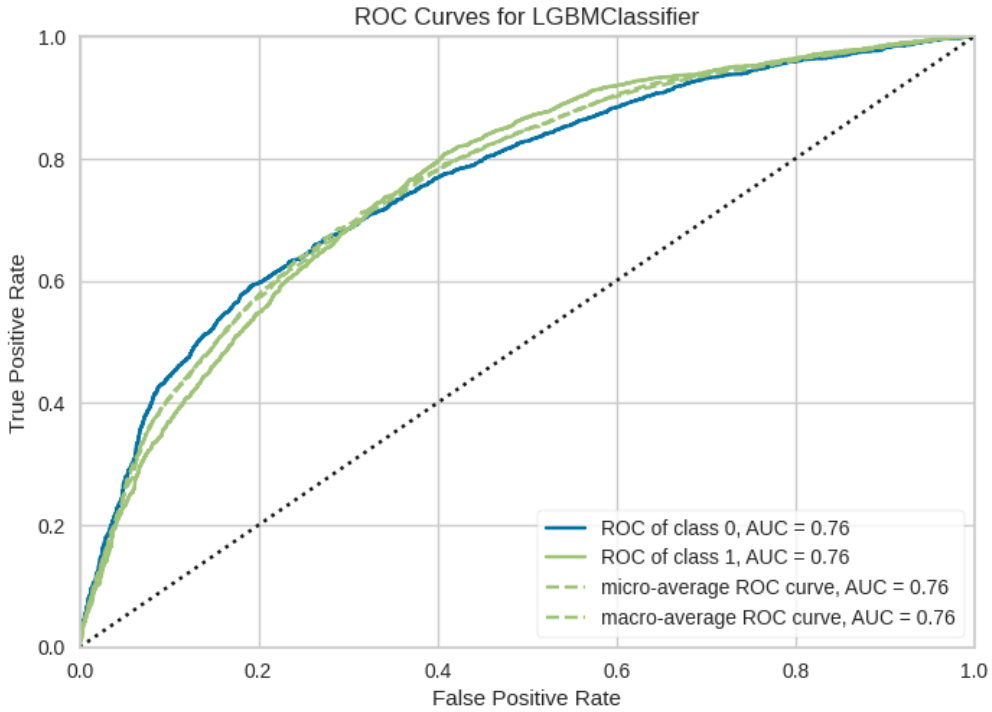
Création du modèle

- Le dataset a 663 108 entreprises ouvertes et ~2 % (15 295) d'entreprises fermées en 2024, sans équilibrage un modèle fiable à 98 % ne verrait rien.
- Dataset équilibré de 15 295 entreprises ouvertes et autant de fermées.
- *Shuffle* avec Pandas.
- *Split* en 80/20 pour training et testing avec SciKit Learn.
- Génération d'un rapport avec YData Profiling pour valider que toutes les transformations donnent bien le dataset attendu.
- La normalisation avec Standard Scaler de SciKit Learn « casse » le modèle sans raison évidente et à donc été laissée à PyCaret.

Création du modèle

Avec le framework de Machine Learning low code PyCaret on compare les modèles et on *tune* les hyper-paramètres du meilleur modèle.

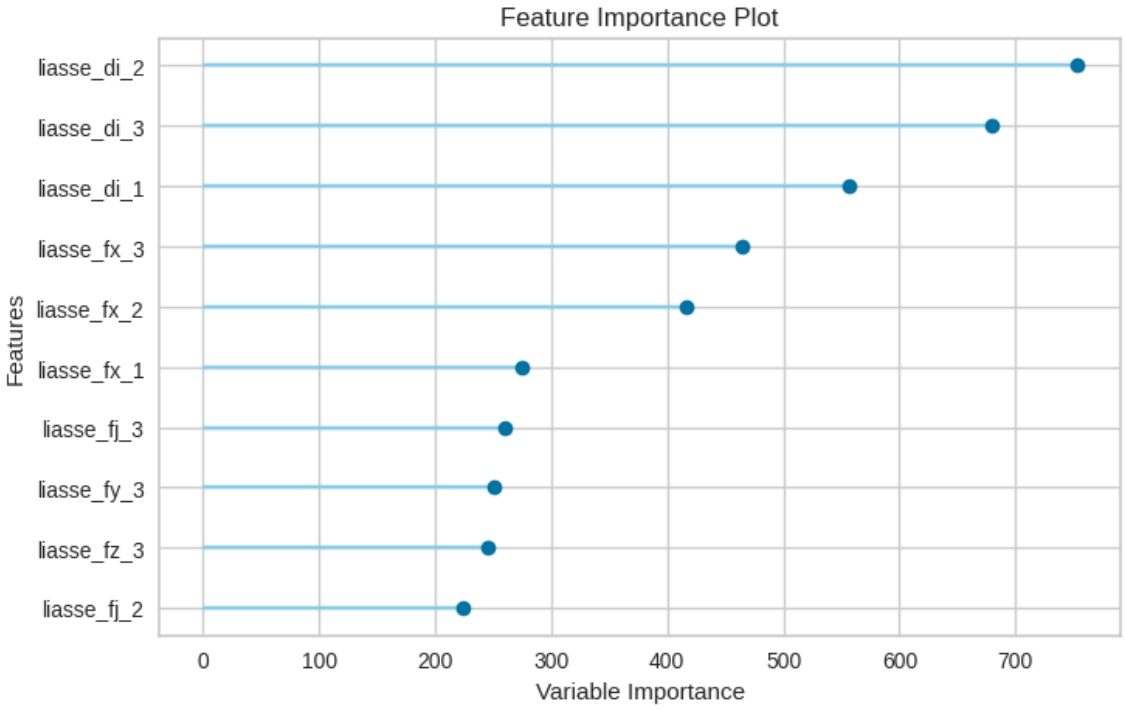
Le meilleur modèle trouvé est un Gradient Boosted Decision Tree qui montre une *accuracy* de 0,76 seulement.



Analyse du modèle

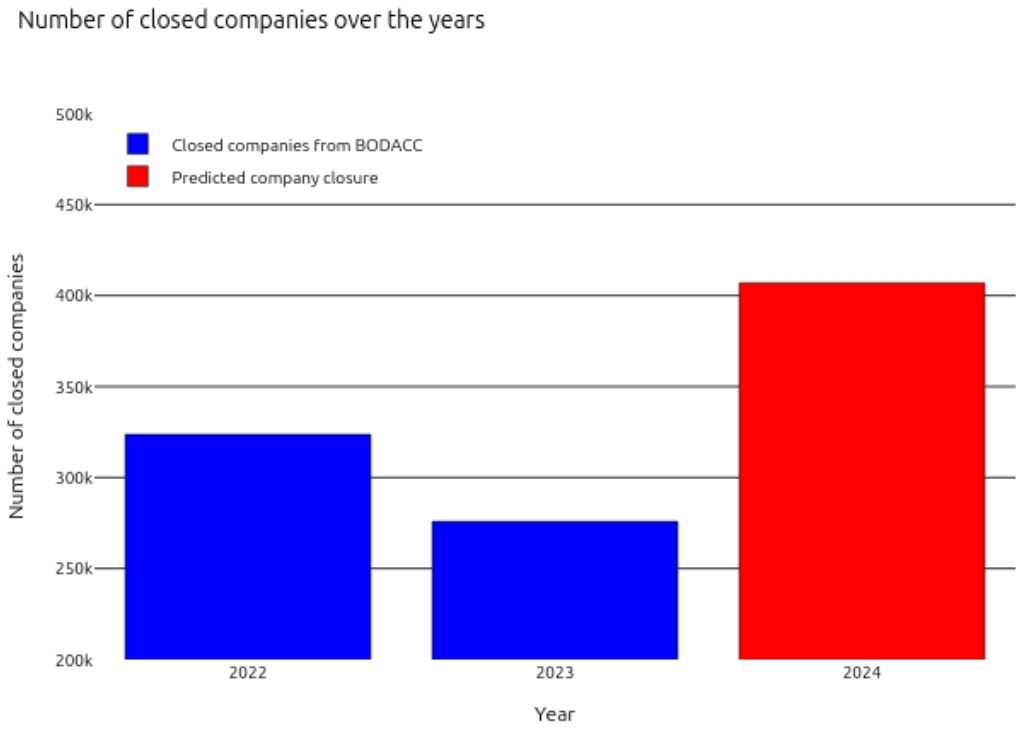
Une *accuracy* de 0,76 est faible, mais c'est le maximum trouvé quelque soit les liasses fiscales en *feature*. Les finances ne sont pas la seule cause de fermeture.

Les features qui ont le plus d'importance sont quand même le résultat (code liasse di) et les impôts (code liasse fx).



Prédiction

- On applique les mêmes transformations aux données de 2021, 2022, 2023 pour prédire les fermetures 2024.
- On passe ces données au modèle précédent.
- Création du graphique avec Plotly.



407 k fermetures !

Conclusion

- L’IA est parfois considérée comme une boîte noire mais le low code de PyCaret (problème de normalisation avec StandardScaler)
- Les données financières à elles seules n’expliquent pas les fermetures. Ou trouver/penser à d’autres source de données.
- Travailler le modèle avec un autre framework moins boîte noire comme SciKit Learn par exemple.
- Les prédictions semblent très cohérentes (contexte économique mauvais) en terme de quantité de fermeture et d’importance des *features*.

Merci à Le Wagon et mes camarades de promotion