Les compétitions Kaggle gagnent l'industrie financière

Kaggle[1], créé en 2010, est un site très prisé par la communauté data science qui héberge depuis son origine des « compétitions » et un forum à destination de celles-ci. Par la suite, il a renforcé son volet enseignement via l'ajout de tutoriaux et mini-cours, de questions-réponses plus techniques et de « datasets publiques » [2] permettant de s'entraîner ou de tester de nouveaux algorithmes.

Bien que Kaggle ne soit pas le seul site à proposer des compétitions[3], il demeure, et de loin, le plus connu et le plus prisé de la communauté data science. Les sociétés qui souhaitent résoudre ou améliorer leurs modèles exposent à la communauté de data scientists leur problématique avec un jeu de données. Les trois meilleures solutions reçoivent un prix et plus celui-ci est élevé, plus il y a de compétiteurs et meilleures sont les solutions proposées.

Selon les compétitions, les solutions sont accessibles aux autres compétiteurs qui peuvent alors jauger leur travail et apprendre de nouvelles approches, techniques ou outils. Le site offre aussi des compétitions d'entraînement, dont les solutions publiques contiennent codes et explications qui permettent aux débutants d'apprendre sur des cas concrets, très proches de cas réels.

L'intérêt des compétitions Kaggle pour l'industrie financière et les sociétés proposant une compétition.

Les sociétés de l'industrie financière ont un avantage certain à se positionner sur ce créneau en tant qu'organisatrices comme en tant que participantes :

- S'ouvrir sur de nouvelles techniques et promouvoir son image en tant que société innovante et active en data sciences dans un secteur traditionnellement discret et conservateur dont les modèles sont jalousement gardés secrets.
- Obtenir une ébauche de modèle prédictif, voire un modèle quasi final en laissant la communauté chercher des solutions selon différentes approches en parallèle. Ainsi, la société peut espérer obtenir plus rapidement des modèles pertinents tout en différant le recrutement ou l'affectation de ses ressources internes avec un gain de temps et d'argent.
- Améliorer son image employeur : en sponsorisant son équipe de data scientists, elle met avant de réelles capacités en data sciences et se présente comme une société innovante utilisant des canaux différents de communication et de recrutement pouvant ainsi attirer des talents et de nouvelles compétences à moindre coût publicitaire.

1. www.kaggle.com
2. Ensemble de données 'propres' librement accessibles
3. https://www.quora.com/What-are-some-alternatives-to-Kaggle

L'Intérêt des compétitions Kaggle pour les data scientists.

En dehors de l'aspect financier, l'intérêt est double :

- Se former en cherchant à résoudre un problème en conditions réelles, tout en bénéficiant également des recherches faites par ses pairs et ainsi améliorer ses compétences.
- Se jauger via des compétitions qui leur permettent de se comparer aux meilleurs de la place mais aussi de se promouvoir auprès de la communauté et se faire remarquer par des recruteurs potentiels.

En effet, participer à des compétitions sur d'autres domaines que le sien permet au data scientist d'apprendre de nouvelles techniques pas ou peu utilisées dans son secteur qu'il pourra alors adapter à son activité. Ces compétitions présentent ainsi une dimension gagnant-gagnant car participant à la promotion à la fois du data scientist et de la banque.

Les institutions financières, historiquement grandes consommatrices de statistical learning, se sont intéressées à l'IA et aux data sciences plus tardivement que d'autres secteurs tels que le marketing ou le digital. La présence de data scientists y est encore assez faible proportionnellement et ils y sont rarement regroupés dans des unités communes. Il alors difficile au data scientist d'être challengé et de progresser par émulation.

Ainsi, finir bien classé à des compétitions Kaggle est un moyen de prouver sa compétence de manière très factuelle, assurant promotion interne et évolutions vers de nouvelles fonctions, voire une mobilité externe…

Opportunités et Limites.

Ces compétitions apportent un réel plus tant au niveau des pratiquants que des entreprises en mettant en exergue factuellement les compétences ou centres d'intérêts des participants.

Mais se pose également la question de la mise à disposition des données utilisées, en particulier si elles s'avèrent critiques comme des données de tiers, même anonymisées. Des recherches[3] ont prouvé qu'il était possible d'identifier des individus à partir de leurs données anonymisées. Exposer ainsi des données critiques de type contrepartie, c'est livrer à ses concurrents une partie de son patrimoine. C'est aussi prendre le risque d'enfreindre la réglementation RGPD[4] en divulguant des données personnelles insuffisamment anonymisées.

Conclusion

Pour pallier à cette limite, la meilleure approche en l'état actuel choses est peut-être pour les banques de recourir à une adaptation interne des compétitions Kaggle : la compétition n'est ouverte qu'au personnel de la société, sur des données qu'il est interdit de sortir du réseau. La faiblesse est que les résultats obtenus seront forcément moins exhaustifs que si toute la communauté de Kaggle avait pu répondre. Cette solution assure néanmoins le succès des deux autres composantes : la promotion interne des participants et la promotion de la société envers ses collaborateurs en tant que société innovante.

À PROPOS DE CAPTEO

CAPTEO est un cabinet de conseil en Stratégie, en Organisation et Management, dédié à l'industrie financière et aux marchés financiers. Cabinet de référence dans le secteur financier, nous accompagnons nos clients depuis plus de 12 ans dans leurs réflexions stratégiques, dans la mise en œuvre de leurs projets de transformation et l'amélioration de leurs performances. www.capteo.com

CONTACTS OFFRE/PUBLICATION

Éric JULLIEN

Manager
En charge de l'Offre Data
ejullien@capteo.com

3. https://arxiv.org/pdf/cs/0610105.pdf
4. RGPD: https://www.cnil.fr/fr/rgpd-de-quoi-parle-t-on