

Interview 2 - Ilan

Interviewer1

OK donc en gros, les objectifs de l'entrevue, c'est de faire ... en fait pourquoi on fait ça ? C'est parce qu'on veut faire un catalogue de Quality issues on the machine Learning software system. Là tu peux demander c'est quoi les Quality issues ? C'est n'importe quel problème qui n'affecte pas la fonctionnalité d'un système, mais seulement la qualité des résultats, puis un exemple de ça ce serait un système de recommandations. On disait sur Netflix, peu importe. Dans lequel les prédictions sont peut-être de bonne qualité. C'est de bonnes recommandations, mais on n'arrive pas à les expliquer. Et ça, ça peut être un Quality issue. C'est pas explicable, ok ? Les Quality aspect en ML qu'on a robustesse, scalability, explainability ... La raison pourquoi on fait cette étude, c'est pour améliorer ... pour donner des pistes de solutions à des problèmes qui existe évidemment dans le but d'améliorer la qualité des systèmes en machine Learning. Donc on a demandé ta permission, pour l'enregistrement t'as dit que, c'est correct, mais aussi pour être transparent, quand on va publier notre rapport, Pour être transparent avec le reviewer, on va partager les transcripts de l'entrevue. Donc en gros, c'est la version écrite de notre conversation. Est ce qu'on entend le consentement pour ça ?

Interviewee1

Oui, vous avez mon consentement.

Interviewer1

et aussi à toutes est anonymes et il n'y aura pas de problème. Et puis sinon, est ce que tu peux commencer à parler un peu de ... Je pense qu'on a parlé, ça fait sept mois que je travaille à Compagnie 1, en fait ton expérience avec le ML est en général ce que tu travailler ?

Interviewee1

J'ai commencé, on va dire ma carrière dans la data à travers des stages. Lorsque j'étais en Pays 1, j'ai travaillé du coup dans le domaine bancaire et en Software Engineering aussi e pour les villes connectées. Donc j'ai fait ça à Ville 1, en Pays 1 et à Ville 2. J'ai surtout commencé. On va dire ma carrière en tant que data analyste dans une entreprise qui était au carrefour du coût de la finance et de l'immobilier. J'ai commencé à travailler avec un outil qui s'appelle Dataiku. C'est une plateforme du coup qui permet en fait a

plusieurs postes dans une entreprise, de comprendre en fait ce qui se passe au niveau des données. On a en fait une visualisation d'un ETL de sa collecte jusqu'à son extraction et Dataiku est donc une solution qui permet aussi en fait de proposer donc des modèles ML qui fonctionne avec du AutoML C'est quelque chose que j'ai fait. On va dire pendant un mois, donc je ne pourrais pas parler précisément et dans le détail en fait de cette partie-là ensuite au sein de compagnie 1 aujourd'hui. Moi je m'occupe de tout ce qui va être en fait pipeline données donc ça va être pareil. Collecte, automatisation, extractions, CICD pour autant en fait tout ce qui va se rattacher. Donc en thématique de data science, je ne pense pas être la personne la plus qualifiée pour vous en parler, mais je vais essayer de répondre le mieux possible à vos questions.

Interviewer1

Non mais nous, on est intéressé à toutes les étapes dans le pipeline pour dans l'objectif de produire un modèle. Donc ton expérience, c'est tout à fait intéressante. OK, merci et bon, je connais encore une fois je connais, je l'ai dit, je connais Compagnie 1, je sais, je sais que vous étiez consultant là, mais on a une question, c'est pour quelle raison d'utiliser l'intelligence artificielle?

Interviewee1

L'objectif, en fait, c'est de promouvoir du coup les entreprises québécoises grâce justement à des solutions d'intelligence artificielle pour optimiser. Différentes parties du coup dans leur entreprise. Alors ça peut être du côté business, ça peut être dans le côté RH Voilà, ça peut être purement dans le côté financier, les principaux bénéfices, on va dire de notre domaine, c'est qu'on peut l'appliquer à différents univers, ce qui ne nous restreint absolument pas.

Interviewer1

Et puis là, en ce moment j'imagine tu travailles sur certains projets. C'est quoi le genre de données que vous utilisez ?

Interviewee1

alors en ce moment je travaille donc pour Compagnie 2. Donc un manufacture de moteurs d'avion on va collecter des données donc relatifs à des moteurs. Donc ça va être le temps de vol d'un moteur. Quel opérateur a fait voler un avion dans lequel il y avait un tel moteur ? Et ça va être des données détaillées. On

va dire dans ces deux domaines là et avec des données que l'on récolte du coup sur plus de trente quarante ans. On a mis en place donc des modèles d'intelligence artificielle pour émettre des hypothèses. sur la prochaine panne du coup qu'un moteur pourrait avoir. On est donc dans un sujet de maintenance prédictive.

Interviewer1

ok, c'est intéressant et le genre de données. Vous avez j'imagine ces données tabulaires ?

Interviewee1

Oui, exactement. On travaille avec des fichiers CSV, Excel, Parquet, mais c'est uniquement voilà des données tabulaires. On n'est pas avec des images vidéo où du son.

Interviewer1

et vos informations vous les obtenez de votre client?

Interviewee1

notre Tech aujourd'hui, notre TechStack elles se basent effectivement sur Azure Devops. Pour tout ce qui est CI/CD .tout ce qui va être, on va dire exploration de données, traitement. Dans un premier temps, on va se faire avec Databricks à travers des NoteBook et ensuite on a eu l'opportunité de travailler avec Compagnie 3. Qui est donc une grande firme de consulting et qui a aussi une entreprise tierce qui s'appelle McKinsey et qui a développé du côté du génial.... Ils ont inventé donc une solution open source qui s'appelle Kendro qui est utile en fait, pour faire de l'extraction, du traitement et du chargement de données, c'est un ETL, donc accessible car open source et en même temps qu'il propose un système de data visualisation pour savoir en fait où est notre input ou va notre output et qui fonctionne sous un système de

Interviewer1

OK, et je veux savoir pour vos modèle est ce que vous utilisez des plateformes comme AWS, Google Cloud ?

Interviewee1

notre Tech aujourd'hui, notre TechStack elles se basent effectivement sur Azure Devops. Pour tout ce qui est CI/CD .tout ce qui va être, on va dire exploration de données, traitement. Dans un premier temps, on va se faire avec Databricks à travers des NoteBook et ensuite on a eu l'opportunité de travailler avec M. Qui est

donc une grande firme de consulting et qui a aussi une entreprise tierce qui s'appelle Q et qui a développé du coût du génial.... Ils ont inventé donc une solution open source qui s'appelle K qui est utile en fait, pour faire de l'extraction, du traitement et du chargement de données, c'est un ETL, donc accessible car open source et en même temps qu'il propose un système de data visualisation pour savoir en fait où est notre input ou va notre output et qui fonctionne sous un système de

Interviewer1

Donc le on s'est présentée. On va vraiment sauté dans le corps de l'entrevue. Et puis on va passer sur chacun des étapes de workflow en ML Donc on va commencer par des questions qui sont plus en lien avec la data collection. On va terminer avec des questions en lien avec les modèles et là, je sais que t'es plus data engineering. Évidemment, aura plus d'expérience dans les premières phases et pour les autres, on ne peut pas ... On ne s'attend pas à répondre à toutes les questions ont plus seulement abordé les premières. Je vais commencé avec une question en général, c'est quoi les plus grands problèmes, Les principaux problèmes de qualité que tu as rencontrées avec les données ou le modèle de ton de système. jusqu'à présent ?

Interviewee1

le plus gros problème, ça va être des trous dans la requête. Quand on va essayer de les connecter des données historiques, on va se retrouver par exemple pour ... Je vais prendre le cas de Compagnie 2 d'un moteur d'avion qui va avoir des informations, par exemple des années quatre-vingt-dix jusqu'à deux mille. Ensuite de deux mille jusqu'à deux mille dix, c'est quelque chose qui va réapparaître du coup en deux mille dix. Pour ça, on va utiliser donc de l'inférence pour essayer donc de d'imaginer en fait retracer la donnée manquante. Un autre point aussi. Ça va être toutes les valeurs null que l'on va récolter. On a une différence en fait, entre des valeurs nulles et des valeurs None. On va dire ça comme ça par exemple, comme si je prends un contrat pour n'importe quel service. Je vais avoir de temps en temps des contrats qui sont nuls, donc des contrats par exemple qui e n'existe pas en tant que tel pour un client donné. Mais je peux avoir aussi des contrats None, ça va signifier en fait deux choses soit l'information rentrer par un être humain a volontairement été laissés vides car il n'y a pas de contrat associés. et donc là c'est une information à prendre en compte, soit ça peut être un oubli de sa part et donc il est toujours en fait important de savoir faire la différence entre une valeur volontairement laissé nul et une valeur qu'il est réellement. Et donc pour ça sa demande des allers-retours justement avec des PO en interne enfin avec les gestionnaires de projets.

Interviewer1

je comprends et c'est les quelles qui sont volontairement laissé les None et les nulles ?

Interviewee1

justement. En fait, il faut discuter avec la partie business pour savoir quelles informations en fait, est la vraie et juste.

Interviewer1

Je comprends, ça donne un processus qui est long et compliqué.

Interviewee1

c'est exactement ça, c'est toujours un ... C'est beaucoup d'allers retours en fait pour comprendre exactement ce que veut le client, quelles sont ces données ? Car il n'est pas toujours sûr de lui-même en fait de l'authenticité de celle-ci et leur véracité.

Donc c'est quelque chose qu'il faut aller vérifier avec lui et ensuite lorsque vous va vouloir aller construire un modèle, on va se rendre compte qu'il va peut-être manqué certaines données qui n'ont pas été assez bien référencé en fait en amont du projet du convoi de temps en temps revenir en arrière pour récolter ces données là et les faire rentrer dans un modèle que l'on a déjà construit.

Donc de temps en temps ça peut être assez difficile.

Interviewer1

OK je vois en gros d'une certaine manière des fois ta des données de bonne qualité et en plus ils demandent certaines informations en plus de tout ça. Donc je vais sauter avec dans section de data collection tu m'a déjà parlé que c'est P qui vous donnent des données puis et si vous avez des problèmes donnez-vous communiquer avec les clients qui vous êtes. il manque, c'est pour ce temps-là il manque certaines données ou bien cette information-là cette information est None, puis Est ce que c'est volontaire ou c'est une erreur ? Mais en outre les données qui sont données par le client est ce que vous avez d'autres moyens d'aller chercher des données comme par exemple je sais que vous avez un client avec Compagnie 1 et vous avez

un API un pour connaître la température et puis vous pouvez prédire c'est quoi les la vendent aux Hotdogs pour le prochains mois ? Quelque chose comme ça ? Est-ce que vous avez d'autres sources de données dans votre projet ?

Interviewee1

Euh oui, on va utiliser des données du coup de Compagnie 3 qui est une Marketplace qui va recensés. En fait toutes les informations lorsqu'un avion est vendu et donc ça, c'est des données avec lesquelles Compagnie 2 travaille mais nous ont aussi donné accès. J'en parle parce que ce n'est pas de la même entité que Compagnie 2, mais je travaille avec

Interviewer1

OK, c'est quoi les données, des données de ?

Interviewee1

Ça s'appelle Compagnie 3 Marketplace et donc ce sont des données relatives à des avions lorsqu'ils sont vendus. Donc on va retrouver en fait le propriétaire, le contact, le numéro de série en fait du moteur et c'est avec justement un numéro de série qu'on arrive à retracer les données dans la base de Compagnie 2

Interviewer1

et la raison que vous utilisez ces données externes, c'est parce que ça vous aide à prédire si les moteurs vont faire des fonds ?

Interviewee1

Euh non, on va s'en servir plutôt pour faire du contact Matching. C'est à dire que avec les données de Compagnie 2, on arrive a donc euh, faire de la maintenance prédictive mais Le problème, c'est que si on sait pour que le moteur A qu'il va tomber en panne, en fait, l'année prochaine, il faut qu'on donne l'information à Compagnie 2 de la personne qui possède ce moteur pour l'appeler.

Interviewer1

ok, je comprends. Est-ce que vous avez des fois des problèmes de qualité avec ces données ?

Interviewee1

Énormément justement parce que on ne va pas toujours retrouvé. En fait un contact associé à un moteur. Le contact n'est peut-être plus bon où l'avion est tellement. Enfin, le moteur associé à un avion est tellement récent qu'il n'est pas en fait pris en compte dans la base de Jetnet et du coup, le contact n'existe pas.

Interviewer1

Et comment Vous gérer ce problème quand les données sont pas bonnes ?

Interviewee1

On fait comme on peut, c'est à dire que l'on va en fait essayer de comprendre s'il y a plus de source de données. En fait, dans ce que nous propose compagnie 2, on va essayer aussi d'aller dans d'autres systèmes informatiques. Donc ça va être par exemple un CRM de chez compagnie 2 qui est n'est pas directement relié en fait avec leurs autres services et c'est là qu'on va essayer de récolter leurs données

Interviewer1

Puis de manière générale, est ce que tu as rencontré d'autres problèmes ? J'imagine que oui. Mais est-ce que tu as des problèmes de qualité des données ? Que tu aimerais partager avec nous ? Et ça pourrait ne pas être dans compagnie 1 ça peut être dans d'autre expériences ?

Interviewee1

Oui, c'est une information en fait identique qui rentrait dans deux systèmes d'information que l'on va essayer de relier, sauf qu'elles ne sont pas rentrés de la même manière. Et donc là, il y a un travail en fait de normalisation de la donnée qui a effectué par exemple, je vais rentrer un nom de famille qui a deux noms par exemple, Je ne sais pas Jaque Paul, ça c'est c'est un nom. Certains vont écrire en fait en minuscules, d'autres vont écrire uniquement en majuscules. Certains vont ajouter un tiret entre Jacques et Paul et donc à ce travail de normalisation effectuer.

Interviewer1

Ouais, donc ça fait le tour pour tout ce qui est par rapport à la collection de données, on commence à parler un peu de la préparation des données. Mais là, je me demandais. C'est quoi les outils que t'aide a la préparation de donner des trucs qui automatise certaines tâches.

Interviewee1

Alors avec les outils sur lesquels je travaille actuellement non en revanche sur Dataiku, il y avait une fonction automatique justement, qui permettait de faire de la normalisation de données très rapidement. Donc qui essayé de reconnaître des patterne dans certaines valeurs pour les catégoriser ensemble

Interviewer1

Est-ce que tu peux me dire comment on épelle Dataiku?

Interviewee1

Data et I K U

Interviewer1

Merci il y a un ... c'est quelque chose général. Mais bon, ça s'applique un peu moins à toi. On dit qu'on passe beaucoup plus de temps à préparer les donnes plutôt qu'à préparer le modelé ?

Interviewee1

D'accord à cent pour cent d'accord. J'ai enfin, j'ai travaillé un petit peu sur des modèles d'intelligence artificielle. Quand j'étais à l'université en Pays 1, j'ai fait mes études à Ville 2. Effectivement, fait, on passe le plus clair de son temps à améliorer en fait nos données plutôt que le modèle en soi.

Interviewer1

Ouais, c'est ça. Et puis de ton expérience. Pourquoi?

Interviewee1

Parce que la qualité en fait de la donnée que l'on reçoit est loin d'être parfaite. Et on essaie justement d'atteindre un niveau acceptable pour notre modèle afin qu'il ne soit pas biaisé.

Interviewer1

Donc les problèmes que tu rencontrent de manière répétitives quand tu prepare tes donnees pour la ML ?

Interviewee1

c'est la compréhension en fait de la donnée, le client ne fournit pas obligatoirement en fait un catalogue. Donc qui va recenser la description de la donnée et quand il le fait parfois cela peut être faux quand on fouille en fait en profondeur. Du coup, c'est plutôt en fait un travail de compréhension qui est long, plutôt que le travail technique en soi.

Interviewer1

Ouais, je n'en sais, c'est intéressant parce que ça parallèle avec quand on parle de ça, je pense que je connais des outils qui justement cherchent à répondre aux problèmes que on se pose. Je donc là, on a fait le tour de tout ce qui a rapport aux données.

On peut passer rapidement sur la question des modèles. Ce que tu sais ? Comment évaluer la qualité des modèles dans les campagnes de travail en général ?

Interviewee1

Dans mes expériences passées, on n'a pas eu l'occasion en fait de regarder précisément les métriques de nos modèles car ils n'ont pas été mis en fait en production n'était vraiment sur des états aujourd'hui par contre on va utiliser MLFlow pour versionner et comparer du coup à nos expériences et toutes les métriques fait son associé à l'intérieur. donc enfin la courbe F1, AUC mais après ce ne sont pas en fait mes domaines de prédilection, du coup je ne pouvais pas en parler dans le détail.

Interviewer1

pour le modèle que vous déployez dans des expériences passées. Est-ce que vous avez des user acceptance measures donc par exemple, vous avez un modèle pour vérifier si ça répond vraiment besoin de ce que vous avez, les moyens de vérifier que ça a vraiment besoin des clients ?

Interviewee1

C'est une très bonne question. On essaie en fait en amont du projet, de définir donc des cas payés pour savoir les seuils minimum en fait à atteindre pour nos résultats. Pour autant, Nos modèles, en fait, ne suffisent pas toujours parce qu'il faut une réelle expérience humaine en fait d'utilisation de notre modèle. Donc, il y a une partie en fait de pilotes qui s'effectue environ sur deux semaines, pendant laquelle, en fait, on

va aller proposer notre solution à nos utilisateurs. Sinon, et c'est pendant cette période-là qu'il faut s'assurer de récolter un maximum de Feedback pour pouvoir améliorer du coup notre modèle. Pendant une phase d'optimisations qui apparaît on va dire dans la dernière phase en fait du processus de consultation.

Interviewer1

Ok, je vois et est-ce que tes modèle sont exposées dans des scénarios qui dans lequel il y a plusieurs groupes de personnes donc ton modèle pourrait bien performé sur un certain groupe et mal performé sur d'autre est ce que ça arrive parfois ?

Interviewee1

euh oui, complètement et du coup, il faut faire en sorte de modifier son modèle suivant ce que l'on souhaite prédire. Donc là, le cas sur lequel je travaille, ce sont des moteurs d'avion. Donc on a fait un modèle pour un type de moteur d'avion. On en a d'autres en fait, qui sont du coup pris en considération. Et du coup les modèles doivent être modifiées,

Interviewer1

OK

Interviewee1

pour rebondir en fait là-dessus c'est la force en fait de Kendro donc le logiciel développé par McKinsey qui permet en fait à travers donc son ETL de pouvoir concevoir un data model. Ddata modèle en fait, il va reprendre les domaines spécifiques de l'industrie donc euh en l'occurrence chez Compagnie 2 on va avoir des tables qui vont faire référence par exemple à des moteurs, à des événements moteur, à l'utilisation des moteurs, à une table contact à une table de société qui gère un avion qui liait le même moteur. Et ça, du coup, ça va être un premier modèle de données. Ce modèle-là, il peut être ensuite utilisé pour un modèle futur qui va servir en fait à alimenter derrière nos algorithmes ML On a du coup un modèle axé sur le business est un modèle en fait, qui va s'axer sur les features que l'on peut extraire du modèle précédent. Et c'est ce feature modèle qui va être très important car c'est celui-ci en fait sur lequel va se baser d'autres développement justement de modèles de prédiction de ML ce qu'on appelle un un feature layer un feature store.

Interviewer1

OK, Je comprends donc en fait t'as l'outil de Q qui mémorise. C'est quoi les schémas des données?

Interviewee1

C'est à nous de concevoir. Pour le coup, ce modèle, c'est nous qui allons extraire une Big table dont nous avons besoin pour un modèle primaire. C'est un du coup, ERD en anglais, un modèle d'entité relations. et de ce modèle-là. Donc qui va prendre les différents domaines que j'ai évoqués. Donc moteur, événement, utilisation etc. On va créer un feature layer où on va retrouver en fait un store de feature qui se base justement sur le primary layer donc juste en amont.

Interviewer1

OK et ça, ça permet d'accélérer le processus. Pour reconnaître les features important c'est ça ?.

Interviewee1

Du coup, on va se retrouver avec un features store et de ce feature store. On va pouvoir créer notre modèle de ML à partir justement de plusieurs features, qu'elles sont dans un magasin et de la lorsqu'on va vouloir on va faire la prédiction sur des moteurs d'hélicoptères.

On va directement aller dans ce feature store et ça nous permet en fait, c'est ça l'avantage de ne pas reconstruire tout un ETL car on va avoir en fait un magasin. Toutes les features vont être disponibles.

Interviewer1

Et quand tu dit reconstruit l'ETL, c'est en gros davantage. L'avantage c'est que ça permet de savoir quel feature c'est important mais aussi tu peux le réutilisé pour faire les transformation de donees

Interviewee1 : c'est exactement ça.

Interviewer1

C'est quoi le nom de l'outil? Déjà?

Interviewee1

Euh ça s'appelle K. Je vais envoyer ça dans le chat. Toutes les références. ça et la ETL, dont je te parlais tout à l'heure, c'était ça. voilà

Interviewer1

merci, C'est tout pour l'évaluation des modèles on va passer au déploiement du modèle une question par rapport à ça. Donc bon, ton modèle est ce que vous le déployer automatiquement ou manuellement.

Interviewee1

toujours manuellement dans un premier temps, forcément. Et ensuite on va passer donc à des processus de CI/CD donc c'est la première fois en fait que je travaille sur le déploiement d'un modèle, comment ça se passe? Euh, on va en fait récupérer donc un artefact de notre ETL et notre modèle, donc une Pythonwheel que l'on va en fait aller exécuter avec Azure DevOps évoque donc notre plateforme de CI/CD qui va ensuite aller exécuter tout cela dans Databricks plate-forme ETL aussi qui va être reliées du coup à AWS. Donc c'est un endroit de stockage choisie par Compagnie 2 on n'a pas été ... en clair on va dire de choisir leur stack. Elle était déjà faite par d'autres personnes. Quand on est arrivé sur le projet

Interviewer1

est t'aurais faire autrement. Vous n'avez pas utilisé Azure AWS?

Interviewee1

Euh je sais pas parce que c'est quand même des solutions euh bah utiliser, on va dire en grande partie par la communauté. Et lorsqu'on va avoir des interrogations, l'information va se retrouver très vite sur les internautes.

Interviewer1

Ok, tu m'a dit qu'au début vous commencez par deployer le modèle manuellement? Et après ça, vous déployez vos modèles automatiquement. Est ce que c'est quelque chose qui est compliqué? E le déploiement automatique du modèle, ou C'est relativement simple ?

Interviewee1

Je ne pouvais pas parler pour le moment, parce que justement, on n'est pas encore arrivée à cette phase.

Interviewer1

OK, c'est bon. Est ce que est ce que ça fait déjà que ton équipe n'est pas un modèle qui performant bien localement mes moins Une fois déployé, il était surpris d'avoir des mauvais résultats. Peu importe,

Interviewee1

ça ne m'est pas arrivé parce que je n'ai pas assez d'expérience pour parfait. C'est bon je pense. Modèle maintenance.

Interviewer1

Là, il reste des questions, c'est sur les mesures de qualité des modèles ML, mais en fait c'est comme je vais les poser et voire. Est ce que vous avez déjà Ok, voila est ce que la robustnesse est une significant quality issue donc c'est un problème de qualité importante pour vous Quand vous construisez des modèles, ça c'est un exemple des autos de Tesla qui conduit automatiquement. mais parfois il ne fonctionne pas dans des situations que visiblement pour un être humain, c'est très clair. On ne comprend pas pourquoi ça ? Parce que par exemple, il y a un stop et elle est passée tout droit est ce que la robustesse c'est quelque chose que vous considérez important quand vous faites des modèles ML

Interviewee1

je suppose que oui, j'aille avoir du mal aussi a détaillé ce point-là.

Interviewer1

Ouais, ouais bon, c'est vrai. Ouais, c'est un peu ouais, c'est un peu en dehors de, on est très loin des données, donc c'est tout. Est-ce que tu as des problèmes, Des projets AI qui ont duré plus long que ce que tu pensais ?

Interviewee1

oui, ça arrive quotidiennement parce que le client a peut-être mal formulé son besoin en fait en amont ou se rend compte que la solution qu'il avait voulu n'est pas la bonne. Du coup, il y a des allers-retours qui s'opère des deux côtés commercial pour justement essayer de prolonger le projet pour faire en sorte d'avoir une solution acceptable et durable pour le client.

Interviewer1

j'ai compris que tu as dit que des fois le client donne des mauvais requirements et ça fait que le projet est prolongé.

Interviewee1

C'est ça, c'est ça, Parce qu'ils, parce qu'il a donné des mauvais requirements comme tu le dis ou parce que la solution qu'il avait envisagé n'est pas celle qui lui plaît au final.

Interviewer1

Et pourquoi tu penses que souvent il est difficile pour les clients de spécifier Qu'est ce qu'il veut vraiment ?

Interviewee1

excuse-moi Je n'ai pas entendu la question formulée

Interviewer1

désolé pourquoi? Pourquoi tu crois qu'il est difficile pour les clients parfois des spécialistes? C'est quoi? Qu'est ce qu'il veut?

Interviewee1

Le client fait appel du coup en fait à une une firme de consulting justement parce qu'il n'a pas les moyens en interne de réaliser justement son besoin lorsqu'il va essayer de le formuler. C'est à notre charge en fait de bien comprendre ce dont il souhaite faire de son projet quelles données est ce qu'il a et où il veut aller. Donc pour répondre à ta question

Interviewer1

en gros les clients il n'y a pas d'expertise en ML suffisante pour bien exprimer ce qu'il veut et c'est un challenge pour vous de interpréter C'est quoi ce besoin s'est faite là-dessus?

Interviewee1

Ouais ouais

Interviewer1

clairement ok Est la dernière question, c'est est ce que tu as des projets AI qui ont été cette fois-ci cancelled à cause que tu n'avais pas de bons résultats avec les Datasets qui s'étaient donnés

Interviewee1

ce n'est pas arrivé. Par contre, je sais que certains produits justement en AI en fait ne voient pas le jour, non pas à cause des données en tant que tel, mais peut-être parce que la solution créée n'arrive pas à être vendue du fait qu'il y a déjà la concurrence du fait que le produit à une mauvaise interface utilisateur peut avoir plusieurs facteurs.

Interviewer1

Et ça c'est la dernière question à ton avis c'est quoi le quality issue Ou que les recherches sur lesquels les recherches travailler. Qui est le plus important dans ton expérience? C'est quoi le plus gros problème que tu rencontres de manière précise que j'aimerais voir si on veut

Interviewee1

à mon sens. En fait, ce ne sont pas des problèmes techniques, mais des problèmes de communication qui se créent en fait entre justement les personnes qui vont aller développer le modèle et ceux qui en ont besoin. c'est surtout Là-dessus qu'il faut travailler et donc faire en sorte que une équipe dans un projet d'intelligence artificielle soit complète. Donc ça va du projet à avoir un Business un Data analyste un data Scientist et un Engineer. C'est comme ça qu'on fait de belles équipes.

Interviewer1

Ouais, c'est tout et j'aimerais te remercier pour le temps que t'as passé

Et puis ça va être des bonnes informations. Nous avec grand plaisir,

Interviewee1

interviewer 3. T'as encore une question.

Interviewer2

Ouais, j'ai une petite question. En fait, c'est rapide. En fait, je veux savoir comment évaluer la qualité de votre data Preprocessing

Interviewee1

sur notre projet, on va se servir de la librairie Great Expectations, qui permet d'aller définir les sources de données. Du coup, à vérifier donc ça peut être le nom des colonnes, le nombre de colonnes, le data type, les statistiques numéraire attendu justement de ces colonnes.

Donc il faut les définir en amont. Et puis après les codés et c'est en faisant tourner justement notre pipeline que l'on va avoir les métriques qui vont nous permettre d'évaluer notre qualité de données. E on a aussi la possibilité d'aller mettre en place un système d'alerte. On peut imaginer l'envoi d'un e-mail par exemple dès qu'on va dépasser un certain seuil où faire en sorte de bloquer notre processus si jamais un seuil est atteint, donc la va s'arrêter.

Interviewer2 2

d'accord Merci

Interviewer1

merci beaucoup merci beaucoup alors bon a ouvert

Interviewee1

Merci beaucoup à vous trois puis passez une bonne journée et bon courage pour cette recherche. J'espère vous avoir été utile.

Created with the Delve Qualitative Analysis Tool (<http://www.delvetool.com>)