CASE STUDY MACHINE LEARNING ENGINEER

I. Case study principal

Temps conseillé : 3-5h

Objectif : à partir des données à disposition, **prédire la conversion de nos leads en clients**, puis proposer une façon de **servir les prédictions du modèle**.

Nous te proposons de structurer le rendu de ton travail comme suit :

- un livrable contenant tes développements qui soit rapidement intégrable en production dans un environnement Cloud
- une note d'une page résumant ton travail et expliquant tes choix techniques

L'esprit de ce case study :

- Nous n'attendons pas que tu développes une compréhension profonde des données dans le temps imparti.
- L'objectif de ce test est de capturer la vision end-to-end d'un projet de Machine Learning. Tu n'as pas besoin d'aller au bout de chaque partie dans le test mais nous souhaitons que tu puisses présenter oralement les éléments additionnels que tu aurais mis en place.
- Une attention particulière sera portée sur la manière de modéliser l'objectif business et les contraintes qui peuvent l'entourer.
- Tu n'auras pas non plus le temps de tester différentes méthodes, nous te conseillons de t'appuyer sur ton expérience pour choisir directement une modélisation qui te paraît optimale d'après l'objectif et les données.
- Le résultat du modèle en lui-même n'est pas l'essentiel, nous évaluerons plutôt la démarche générale, la qualité du travail mené ainsi que la méthode de mise à disposition des prédictions.
- L'entretien sera l'occasion d'un échange afin de présenter ta solution, tes choix, puis d'extrapoler sur les améliorations possibles.

De préférence, envoie-nous ton case study la veille de l'entretien afin que nous ayons le temps de le lire, cela rendra notre échange plus constructif 🙂

Bonus → Pas de code nécessaire ici ! L'objectif est de discuter des éléments ci-dessous **si tu as des éléments de réponse**

Supposons qu'il faille mettre ce modèle en production dans une architecture Cloud, avec les contraintes suivantes :

- Les données sources à disposition proviennent d'un Data warehouse
- Le modèle doit pouvoir être mis à jour régulièrement et facilement
- L'équipe Tech - client principal de ton modèle - a besoin d'avoir une prédiction sous 1 seconde

1. Est-ce que ta solution serait scalable ? Quelles limites y vois-tu ?
2. Peux-tu proposer une idée d'architecture répondant à ces besoins ?

CASE STUDY MACHINE LEARNING ENGINEER

3. Quels points seraient à garder à l'esprit dans une approche FinOps ?
4. Quel(s) élément(s) devrais-tu changer si le temps de réponse devait être de quelques millisecondes ?

Données

Note : les données fournies ont été modifiées dans le cadre de ce case study

Tu trouveras en pièce jointe un CSV contenant les données disponibles :

long_quotes.csv : Table de production qui contient des informations relatives aux devis d'assurance auto entrants

- long_quote_id : ID du devis
- lead_id : ID du lead
- last_utm_source : source du trafic arrivant sur notre site
- has_been_proposed_formulas : Si Ornikar a proposé d'assurer le lead
- has_chosen_formula : Si le lead a cliqué sur une formule proposée
- has_subscribed_online : Si le lead a souscrit un contrat en ligne
- submitted_at : Timestamp de complétion du devis
- effective_start_date : Date de début souhaitée du contrat
- rbs_result : Résultat du test psychologique qui est proposé aux jeunes conducteurs
- provider : Fournisseur de contrat d'assurance
- product_third_party : Produit au tiers proposé
- product_intermediate : Produit intermédiaire proposé
- product_all_risks : Produit tous risques proposé
- annual_price_third_party : niveau de prix au tiers proposé
- annual_price_intermediate : niveau de prix intermédiaire proposé
- annual_price_all_risks : niveau de prix tous risques proposé
- chosen_formula : Formule sur laquelle le lead a cliqué
- chosen_product : Produit correspondant à la formule sur laquelle le lead a cliqué
- policy_subscribed_at : Timestamp de souscription du contrat en ligne
- payment_frequency : Fréquence de paiement choisie
- main_driver_age : Catégorie d'âge du lead
- main_driver_gender : Genre du lead
- main_driver_licence_age : Catégorie d'ancienneté du permis du lead
- main_driver_bonus : Catégorie de bonus/malus du lead
- vehicle_age : Catégorie d'ancienneté du véhicule du lead
- vehicle_class : Catégorie de classe du véhicule du lead
- vehicle_group : Catégorie de groupe du véhicule du lead
- vehicle_region : Région de parking du véhicule du lead
- has_secondary_driver : Si le devis inclut un conducteur secondaire
- has_subscribed : Si le lead a souscrit un contrat (en ligne ou via le service clients)

CASE STUDY MACHINE LEARNING ENGINEER

CASE STUDY MACHINE LEARNING ENGINEER

II. Deux petits casse-têtes

Temps conseillé : 20 min
Pas besoin de support, on pourra simplement en discuter !

1. Présence de graphes au sein d'un dataset

On prédit la résiliation de l'assurance de nos clients, avec un certain nombre d'informations les concernant (données de souscription, variables socio-démographiques de leur zone géographique...).

Une ligne = 1 client

Certains clients sont de la même famille (on dispose de cette information, avec l'adresse postale par ex). Ils risquent donc de s'influencer les uns les autres. Par exemple, si l'un des parents résilie son assurance, le risque que l'autre résilie augmente fortement.

Dans ce contexte, comment faire pour avoir une performance mesurée sur Test fiable ? (plusieurs réponses possibles)

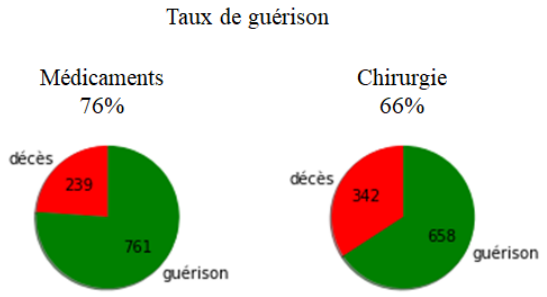
- On ne garde qu'un seul client par famille dans la base (on retire les autres clients de sa famille)
- On ne garde qu'un seul client de chaque famille dans train, et un dans test
- On split train-test sur les clients (un split train-test normal)
- On split train-test sur les familles (chaque famille est soit intégralement dans train, soit intégralement dans test)
- On doit régulariser fortement le modèle
- On n'est pas obligé de régulariser fortement le modèle

2. Etudes contradictoires

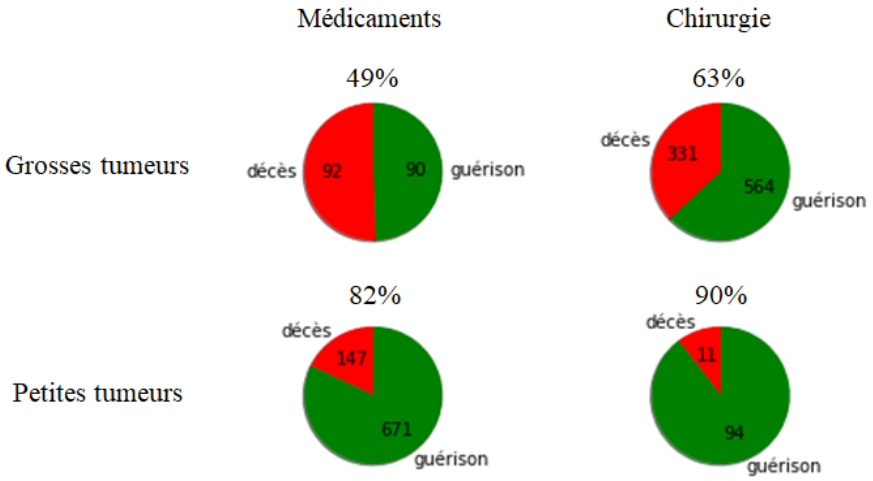
Deux études ont été menées dans deux cliniques différentes pour déterminer la meilleure méthode pour soigner le cancer du petit doigt :

- La clinique A trouve que les patients soignés par médicaments ont un meilleur taux de guérison :

CASE STUDY MACHINE LEARNING ENGINEER



- La clinique B trouve que la chirurgie est la meilleure méthode, que le patient ait une petite ou une grosse tumeur :



Qu'en penses-tu ? (Il n'est pas nécessaire de faire des calculs).

Bon courage !

La team Data Ornikar

En cas de questionnement, n'hésite pas à contacter Charles, charles.tremblay@ornikar.com