Mise en situation

Vous êtes Consultant Data au sein d'une société financière, nommée "home Credit", qui propose des crédits à la consommation pour des personnes ayant peu ou pas du tout d'historique de prêt.

Fraîchement embauché depuis une semaine avec ce salaire annuel, vous avez fait connaissance avec vos collègues et votre nouveau bureau. Mais revenons à vos missions : il est temps de mettre les mains dans le cambouis ! Le DSI vous a donné l'accès à la base de données. L'entreprise souhaite développer un modèle de scoring de la probabilité de défaut de paiement du client et de le mettre en production. Les données à disposition sont variées : données comportementales, données provenant d'autres institutions financières, etc.

Mission

Créer le modèle de scoring et le mettre en prod

Construire un modèle de scoring qui donnera une prédiction de la probabilité de solvabilité d'un client et le déployer via une API sur le Web (Flask, Django par exemple, help1 et help2) via un outil gratuit et disponible plusieurs mois (en vue de vos entretiens techniques, jury,…). Cette API renvoie bien la prédiction correspondante à un client en réponse à un identifiant client.

Conseil : Pour déployer le modèle via une API dans le Web, penser à utiliser par exemple un fichier de format pickle contenant le modèle de machine learning sérialisé (et tester que son chargement fonctionne).

Le focus pour le modèle de machine learning sera mis sur :

- La conception du modèle, son évaluation et son interprétation compréhensible pour les métiers

- La systématisation de la création de features, via les fonctions min, max, mean ou via la combinaison de features (rapport de 2 features, notamment montants, ...)

    - Vous pouvez chercher un kernel sur kaagle qui propose le codage, la préparation et la transformation des données (jointures, groupby, labelEncoder, oneHotEncoder,…). Surtout adapter ce kernel avec votre façon de coder.

- Dans le cadre de l'optimisation du modèle, penser à utiliser SMOTE (génération de lignes pour ré-équilibrer le nombre de valeur cible à 1 par rapport à 0) et Hyperopt (optimisation des hyperparamètres).

- N'oublier pas de mettre en œuvre une matrice de coût adaptée au contexte de crédit afin de proposer une optimisation orientée métier et non pas technique :

    - Par exemple : le coût d'un faux positif (bon crédit considéré comme mauvais constitue un manque à gagner modéré pour la banque, une perte de marge) est différent d'un faux négatif (mauvais crédit considéré comme bon = constitue une perte importante pour la banque, un défaut de paiement et/ou une perte de capital non remboursé). Idéalement, vous montrez que l'optimum « métier » est différent de l'optimum du fscore ou autres mesures purement « techniques ».

Créer un dashbord interactif

Les chargés de relation client ont fait remonter le fait que les clients sont de plus en plus demandeurs de transparence vis-à-vis des décisions d'octroi de crédit. Cette demande de transparence des clients va tout à fait dans le sens des valeurs que l'entreprise veut incarner. Votre manager décide donc de développer un dashboard interactif pour que les chargés de relation client puissent expliquer de façon la plus transparente possible les décisions d'octroi de crédit.

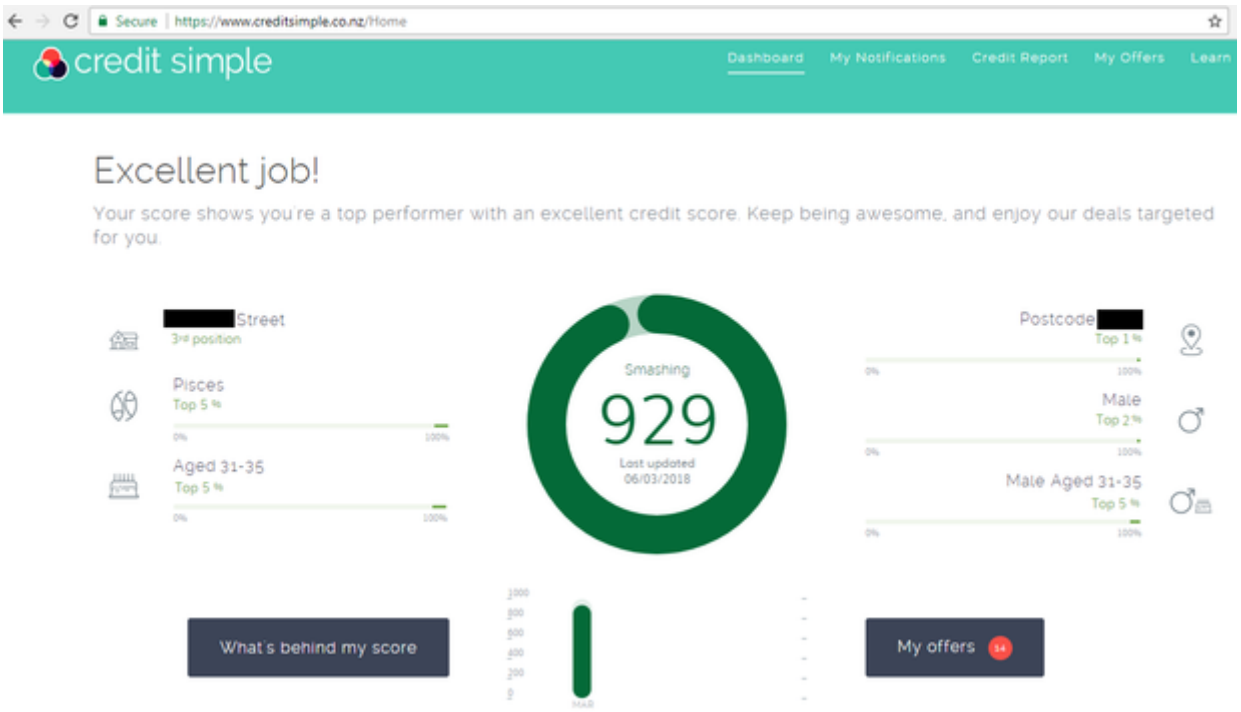
Cahier des charges rédigé par le manager pour le dashboard :

Les spécifications du Dashboard devront a minima contenir les fonctionnalités suivantes :

- Permettre de visualiser des informations descriptives relatives à un client (via un système de filtre).
- Permettre de visualiser le score et l'interprétation de ce score pour chaque client de façon intelligible pour une personne non experte en data science.
- Permettre de comparer les informations descriptives relatives à un client à l'ensemble des clients ou à un groupe de clients similaires.

Le focus sera mis :

- Le dashboard est accessible pour d'autres utilisateurs sur leurs postes de travail (déploiement dans le web)
- Les graphiques réalisés sont pertinents (ils permettent de répondre à la problématique métier) avec aussi au moins deux graphiques interactifs permettant aux utilisateurs d'explorer les données clients ont été développés (l'objectif est de répondre à des questions comme "quel est le client avec le plus de transactions ?")


Secure | https://www.creditsimple.co.nz/Home
credit simple
Dashboard
My Notifications
Credit Report
My Offers
Learn
Excellent job!
Your score shows you're a top performer with an excellent credit score. Keep being awesome, and enjoy our deals targeted for you.
[redacted] Street
3rd position
Pisces
Top 5 %
Aged 31-35
Top 5 %
Smashing
929
Last updated
06/03/2018
Postcode [redacted]
Top 1 %
Male
Top 2 %
Male Aged 31-35
Top 5 %
What's behind my score
My offers
14


AcmeBank
CreditView Dashboard
Menu
MyCredit Dashboard / Home
TransUnion Credit Score
755
850 / A
B
300 / F
Where You Stand
74%-100%
A
781-850
52%-73%
B
720-780
755
35%-51%
C
658-719
22%-34%
D
601-657
1%-21%
F
300-600
My Score History
755
745
735
725
715

Livrables attendus

Un repo Github contenant :

- Un court fichier README, contenant les explications pour lancer vos scripts.
- La liste des problèmes rencontrés sur le jeu de données et les éventuelles stratégies pour les résoudre suivant un trello et en appliquant la méthode Agile , Scrum, ...
- Le script destiné à nettoyer le jeu de données.
- Le script de modélisation (du prétraitement à la prédiction).
- Le code générant le dashboard.
- Le code permettant de déployer le modèle sous forme d'API

- Une note méthodologique décrivant :
    - La démarche de modélisation et la méthodologie d'entraînement du modèle est présentée de manière synthétique (2 pages max)
    - La fonction coût, l'algorithme d'optimisation et la métrique d'évaluation (1 page max)
    - L'interprétabilité du modèle est explicitée (1 page max). N'oubliez pas : la façon d'interpréter l'importance des variables n'est pas la même pour une régression logistique que pour un random forest (par exemple). Préciser les limites éventuelles ?
    - Les limites et les améliorations possibles pour gagner en performance et en interprétabilité (1 page max)
- L'application interactive répondant au cahier des charges précisé ci-dessus.