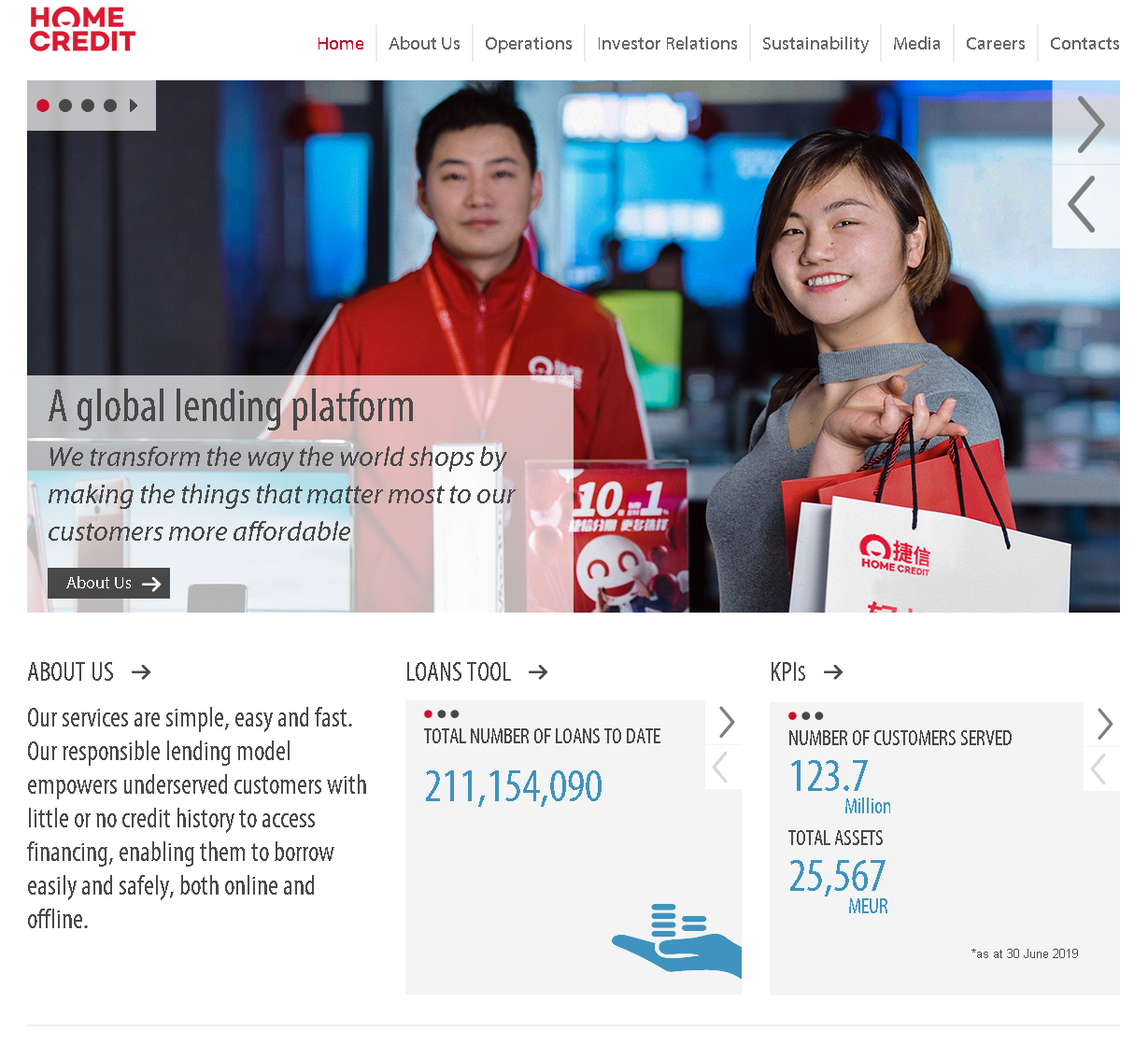
***L’intention de ce projet est de réaliser le modèle de machine learning de bout en bout jusqu’à sa mise en production.***

# **Mise en situation**

Vous êtes *Consultant Data* au sein d'une société financière, nommée "[**home Credit**](http://www.homecredit.net/)", 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](https://www.kaggle.com/c/home-credit-default-risk/data), 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](https://www.kaggle.com/c/home-credit-default-risk/data). 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. (**à vous de vous familiariser avec cette data**!!!).

De plus, 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.



# **Vos missions**

## **Créer la base de données de la Banque**

* MCD, tables, jointures, PK, FK, connexion sécurisée, ..., backup (création dans l’art en respectant les compétences à valider : voir la section **Compétences RNCP à valider**).

## **Créer le modèle de scoring et le mettre en prod**

Construire un modèle de scoring qui donnera une prédiction sur la probabilité de solvabilité d'un client de façon automatique et le déployer via une API sur le Web (Flask, Django par exemple) 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**

Construire un dashboard interactif à destination des gestionnaires de la relation client permettant d’interpréter les prédictions faites par le modèle et d’améliorer la connaissance client des chargés de relation client.

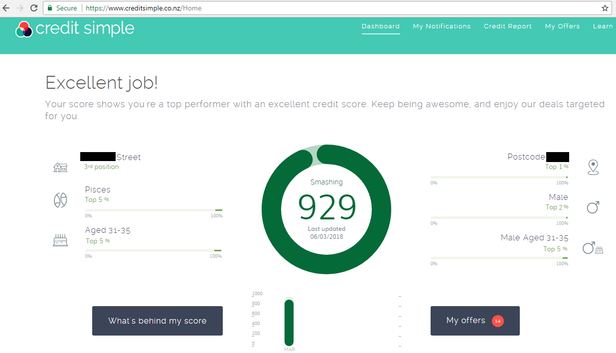
**Cahier des charges rédigé par le manager pour le dashboard :**

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 ?")



# **Livrables attendus**

Un **repo** [**Github**](https://github.com/) 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**](https://trello.com/) et en appliquant la méthode [**Agile**](https://fr.wikipedia.org/wiki/M%C3%A9thode_agile) , [**Scrum**](https://fr.wikipedia.org/wiki/Scrum_(d%C3%A9veloppement)), ...
* 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 modèleest explicitée(1 page max). Entre autres, expliquer que la façon d'interpréter l'importance des variables n'est pas la même pour une régression logistique que pour un random forest) et les limites éventuelles sont précisées
  + 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.

# **Modalités de présentation du travail**

Votre présentation pourra prendre cette forme :

|  |  |
| --- | --- |
| 5 min. | Rappel de la problématique et présentation du jeu de données |
| 10 min | Explication de l’approche de modélisation (à l'aide de la note méthodo-logique) |
| 10 min | Présentation du dashboard |
| 10 min | Séance de questions-réponses |

# **Ressources complémentaires**

* Un [article](https://www.sisense.com/blog/4-design-principles-creating-better-dashboards/) donnant quelques bonnes pratiques pour le design de dashboard.
* Un [document](https://www.psiweb.org/docs/default-source/2018-psi-conference-posters/48-julie-jones.pdf) décrivant les bonnes pratiques pour réaliser des graphiques clairs et pertinents.
* Des informations sur deux librairies permettant de construire des dashboards interactifs en Python : [Dash](https://dash.plot.ly/installation?_ga=2.33133886.1415841438.1542826356-1321721347.1542826356) et [Bokeh](https://bokeh.pydata.org/en/latest/).
* [Ce lien](https://towardsdatascience.com/interpretability-in-machine-learning-70c30694a05f), et [celui-ci](https://christophm.github.io/interpretable-ml-book/) pour aider à construire l’interprétabilité du modèle.

# **Compétences data à valider**

* Déployer un modèle via une API dans le Web
* Réaliser un dashboard pour présenter son travail de modélisation
* Rédiger une note méthodologique afin de communiquer sa démarche de modélisation
* Utiliser un logiciel de version de code pour assurer l’intégration du modèle
* Présenter son travail de modélisation à l'oral

# **Compétences RNCP à valider**

* C2 : *Concevoir et structurer physiquement une BD relationnelle ou non, à partir des besoins, contraintes et données du commanditaire.*
* C3 : *Acquérir des données, les combiner et les structurer en données propres* (*en vue de leur intégration dans la BD*)*.*
* C4 : Intégrer les données propres obtenues dans la BD finale (en utilisant des langages informatiques, logiciels ou outils).
* C5 : Optimiser une BD afin d’en maintenir la fiabilité et la qualité des données.
* C6 : Interroger la BD afin de MAJ les données (brutes ou traitées) stockées, provisoirement ou durablement, en fonction du résultat recherché.
* C7 : Concevoir et réaliser un rendu visuel des données issues du processus d'extraction, à l’aide d’un (des) support(s) adapté(s) répondant aux attentes du commanditaire.
* C8 : Mettre à disposition les rendus visuels et physiques des données, en accès libre ou contrôlé, à l’aide des formats/supports appropriés ou commandés (papier, internet, serveur, mail etc.). Respecter la réglementation associée.
* C9 : Autocontrôler, tout au long du processus de développement, la cohérence des données et la conformité à la demande.
* C10. Suivre, adapter et rendre compte de la réalisation du projet à partir du planning projet validé.
* C11 : Rendre compte de l’avancement du projet.
* C12 : Rechercher des solutions pour la résolution de problèmes techniques rencontrés au moyen des ressources disponibles (documentation, sites Internet, communautés, etc..).