NOTE METHODOLOGIQUE

Implémentez un modèle de scoring

Eric Blanvillain

17/02/2022

Diagram, shape, circle

Description automatically generated

TABLE DES MATIERES

1. PRESENTATION DES DONNEES ................................................................................................ 1

Arborescence des tables de données...........................................................................................1

Modèle de classification et cible (TARGET)..................................................................................1

2. PREPARATION DES DONNEES .................................................................................................. 2

Pipeline de pré-traitement appliqué aux tables ......................................................................... 2

Assemblage des tables et sélection de variables..........................................................................3

3. MODELISATION ........................................................................................................................ 4

Méthodologie de rééquilibrage de la variable cible ....................................................................4

Séparation des données................................................................................................................5

Choix de l’algorithme....................................................................................................................5 Problématique « métier » : métrique « bancaire » et seuil de solvabilité……………………………….6

Méthode d’optimisation des hyperparamètres de LightGBM.....................................................6

Modèle finalisé..............................................................................................................................7

Fixation du seuil de solvabilité optimum......................................................................................9

4. INTERPRETABILITE ................................................................................................................. 10

Importance relative des variables pour le modèle ................................................................... 10

1. CONCLUSION – LIMITES ET AMELIORATIONS .............................................................. 10
2. PRESENTATION DU DASHBOARD .................................................................................. 11
3. LIENS............................................................................................................................... 12

Déploiement du dashboard ...................................................................................................... 12

SYNOPSIS

**Introduction**

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

L’entreprise souhaite mettre en œuvre un outil de “scoring crédit” pour calculer la probabilité qu’un client rembourse son crédit, puis classifie la demande en crédit accordé ou refusé. Elle souhaite donc développer un algorithme de classification en s’appuyant sur des sources de données variées (données comportementales, données provenant d'autres institutions financières, etc.).

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.

Prêt à dépenser décide donc de développer un dashboard interactif pour que les chargés de relation client puissent à la fois expliquer de façon la plus transparente possible les décisions d’octroi de crédit, mais également permettre à leurs clients de disposer de leurs informations personnelles et de les explorer facilement.

**Les données**

Voici les données dont vous aurez besoin pour réaliser le dashboard. Pour plus de simplicité, vous pouvez les télécharger à cette adresse.

Vous aurez sûrement besoin de joindre les différentes tables entre elles.

**Votre mission**

Construire un modèle de scoring qui donnera une prédiction sur la probabilité de faillite d'un client de façon automatique.

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.

Michaël, votre manager, vous incite à sélectionner un kernel Kaggle pour vous faciliter la préparation des données nécessaires à l’élaboration du modèle de scoring. Vous analyserez ce kernel et l’adapterez pour vous assurer qu’il répond aux besoins de votre mission.

Vous pourrez ainsi vous focaliser sur l’élaboration du modèle, son optimisation et sa compréhension.

**Spécifications du dashboard**

Michaël vous a fourni des spécifications pour le dashboard interactif. Celui-ci devra contenir au minimum les fonctionnalités suivantes :

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 visualiser des informations descriptives relatives à un client (via un système de filtre). Permettre de comparer les informations descriptives relatives à un client à l’ensemble des clients ou à un groupe de clients similaires.

**Livrables**

Le dashboard interactif répondant aux spécifications ci-dessus et l’API de prédiction du score, déployées chacunes sur le cloud.

Un dossier sur un outil de versioning de code contenant : Le code de la 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 méthodologie d'entraînement du modèle (2 pages maximum) La fonction coût métier, l'algorithme d'optimisation et la métrique d'évaluation (1 page maximum) L’interprétabilité globale et locale du modèle (1 page maximum) Les limites et les améliorations possibles (1 page maximum) Un support de présentation pour la soutenance, détaillant le travail réalisé.

Pour faciliter votre passage au jury, déposez sur la plateforme, dans un dossier nommé “P7\_nom\_prenom”, tous les livrables du projet. Chaque livrable doit être nommé avec le numéro du projet et selon l'ordre dans lequel il apparaît, par exemple “P7\_01\_dashboard”, “P7\_02\_dossier”, et ainsi de suite.

**Soutenance**

Pendant la soutenance, l’évaluateur jouera le rôle de Michaël, à qui vous présentez votre travail.

Présentation (20 minutes) Rappel de la problématique et présentation du jeu de données (5 minutes) Explication de l’approche de modélisation (10 minutes) Présentation du dashboard (5 minutes) Discussion (5 minutes) L’évaluateur, jouant le rôle de Michaël, vous challengera sur vos choix. Débriefing (5 minutes)

À la fin de la soutenance, l'évaluateur arrêtera de jouer le rôle de Michaël pour vous permettre de débriefer ensemble.

**Skills**

Présenter son travail de modélisation à l'oral 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 Déployer un modèle via une API dans le Web

PRESENTATION DES DONNEES

1. **Arborescence des tables de données**

Sept tables de données sont mises à disposition pour mener à bien notre démarche. Elles sont constituées de données anonymisées d’informations personnelles et bancaires des clients.

Timeline

Description automatically generated

FIGURE 1 – Arborescence des 7 tables de données fournies par Home Crédit

1. **Modèle de classification et cible**

La valeur à prédire est contenue dans la variable « TARGET ». Elle peut prendre 2 valeurs :

* 1 qui indique que le client n’a pas remboursé son crédit. (Défaillant, 8%)
* 0 qui indique que le client l’a remboursé. (Non défaillant, 92%)

Ce déséquilibre devra être pris en compte lors de la construction du modèle, car certains algorithmes sont sensibles au déséquilibre.

Chart, pie chart

Description automatically generated

FIGURE 2– Répartition de la variable TARGET

PREPARATION DES DONNEES

1. **Pipeline de pré-traitement appliqué aux tables**

Chaque table est traitée indépendamment. Des anomalies détectées pendant l’analyse exploratoire des données sont corrigées. Cette analyse a permis de constater @qu’un tiers des variables totales ont 50% de valeurs manquantes. Une double stratégie est appliquée : suppression de colonnes (NAN > 60%) et imputation selon le type de données.

Diagram

Description automatically generated

Deux processus de création manuelle de nouvelles variables (feature enginneering) sont appliqués :

Création systématique de variables statistiques créées à partir des variables existantes : (moyenne, minimum, maximum, écart-type, compte)

Création de variables « métier » basées sur les connaissances du domaine bancaire :

* **Ratios** : montant total du crédit/revenu du client, montant crédit/ intérêt du prêt, ancienneté au travail/âge, nombre d’enfants/revenu, revenu par tête
* **Sommes** : des 3 sources extérieures, des docs (flags) à fournir, des contacts client (adresse, téléphones, mail)
* **Différences** : âge - ancienneté emploi
* **Booléen** : crédit>somme demandée ? (Credit>GoodPrice)
* **Variable liée à la cible** : moyenne de la cible pour les 500 voisins les plus proches de chaque client en fonction de 4 variables : les 3 sources extérieures (EXT\_SOURCE 1, 2 et 3) et un ratio créé par nos soins (ratio CREDIT/ANNUITY)

Les données sont ensuite encodées selon leur type et leur signification :

Variables qualitatives : transformation en variables quantitatives pour les rendre exploitables

* **Label encoder** : encodage binaire (0 ou 1) (modalités < 3)
* **One Hot Encoder** : pour les autres variables qualitatives

Variables quantitatives : mise à l’échelle : ensemble des valeurs ramenées de 0 à 1

* **MinMaxScaler** : pour pouvoir tester nos données sur des algorithmes de machine learning sensibles aux variations d’échelle des variables lors de la phase de sélection du meilleur modèle.