Projet 7

Implémentez un modèle de

scoring Note méthodologique

Kouaouci Asma

Formation Data Scientist

I.Présentation de la mission

1- Objet :

L'entreprise "Prêt à dépenser", qui propose des crédits à la consommation pour des personnes ayant peu ou pas du tout d'historique de prêt, souhaite développer un modèle de scoring de la probabilité de défaut de paiement du client pour étayer la décision d'accorder ou non un prêt à un client potentiel en s'appuyant sur des sources de données variées (données comportementales, données provenant d'autres institutions financières, etc.).

Elle 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 objectifs sont donc :

- 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

La description du jeu de données, le code de la modélisation et génération du Dashboard sont présentés dans des documents annexes.

Le présent document est une note méthodologique décrivant :

- La méthodologie d'entraînement du modèle,
- La fonction coût, l'algorithme d'optimisation et la métrique d'évaluation,
- L'interprétabilité du modèle,
- Les limites et les améliorations possibles.

2- Données à disposition :

Pour mener à bien cette mission, nous disposons d'un jeu de données fourni Ce jeu de données est composé :

- D'un jeu de données d'entraînement (*train.csv*), qui sera utilisé pour entraîner le modèle et tester la qualité
- D'un jeu de données de test (*test.csv*, qui ne contient pas le résultat de la variable à prédire – TARGET- et qui sera utilisé dans le Dashboard interactif

Pour la préparation du jeu de données, nous avons utilisé un Kernel Kaggle

Pour rappel, nous sommes ici dans un problème de classification binaire (0 et 1) où la TARGET que l'on cherche à prédire prend deux formes : soit un 0, indiquant que le prêt a été remboursé à temps, soit un 1 indiquant que le client a eu des difficultés de paiement.

2- Modèles testés et recherche des hyper-paramètres

Pour cette mission, j'ai testé les différents modèles :

.Dummy classifier

"Dummy Classifier donne la prédiction selon la règle prédéfinie cette classification Dans notre cas, prédit la classe la plus fréquente dans la target en l'occurrence la classe 0

• Logistic Regression

Il s'agit d'un modèle linéaire généralisé avec une fonction logistique,il est utilisé pour estimer une valeur discrète
la variable cible n'a que deux résultat possible accordé prêt ou pas
l'algorithme de régression logistique est un algorithme de classification plutôt que de régression il se base sur un ensemble de données indépendante

• Decision Tree

est un algorithme capable d'effectuer une classification binaire et multiclasse l'objectif est de Créer Un modèle qui prédit la valeur d'une variable cible en apprenant des règles de décision simples déduites des caractéristique des données

il est simple à interpréter et à comprendre ,les arbres peuvent être visualisé
nécessite peu de préparation de données car d'autres technique nécessite une normalisation des données et des variables fictives doivent être crée et des valeurs vides doivent être supprimées .. ect
Par contre les apprenants des arbres de décision peuvent Créer des arbres trop complexe ça ce qu'on appel le sur apprentissage du modèle

• Random Forest

Comme son nom l' indique, cette méthode agrège un ensemble d'arbres de décision (on parle de la méthode ensembliste). Les arbres de décision qui constituent le forêt sont très similaires entre eux mais à chaque fois légèrement différents lors de l'apprentissage on cherche un meilleur découpage de fait uniquement sur un sous ensembles de caractéristique d'origines pris au hasard elles sont différentes pour chaque noeud divisé Les résultats de tous les arbres de décision sont alors combinés pour donner une réponse finale. Chaque arbre ""vote"" (oui ou non) et la réponse finale est celle qui a eu la majorité de vote.

Important

avant de lancer l'apprentissage des modèles il faut résoudre problème on a la répartition des valeurs sur la target clients remboursé le prêt ou pas trop déséquilibré avantage de 0 que de 1 les modèle plus haut ne sont pas fait pour classification très déséquilibré on obtient des résultats médiocre

Donc la solution est de créer un data artificielle

1 cas est le oversampling : on multiplie les exemples de la Classe minoritaire de façon lui donner plus de poids

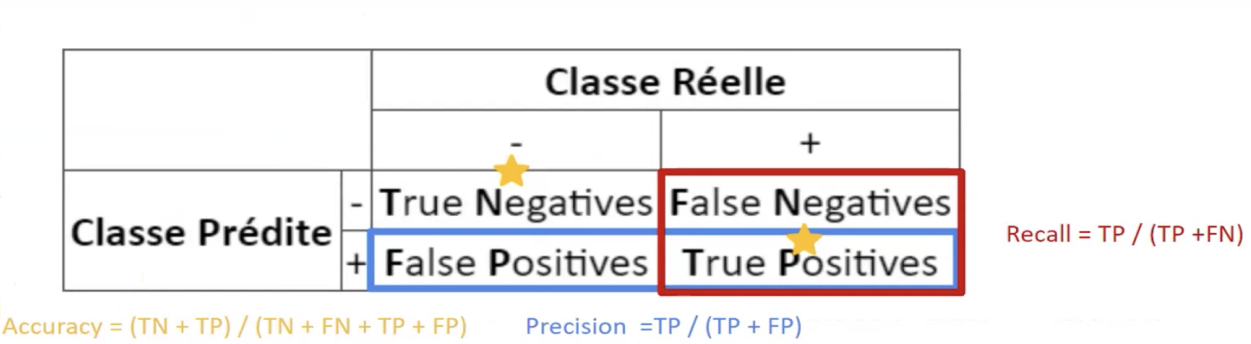
2 cas est under sampling : on réduit nombre d'exemple de la Classe majoritaire sans altérer la capacité du modèle à trouver une bonne solution le Clients a remboursé à temps cela consiste à enlever des exemples loin de la frontière de classification on lui donne un poids important

on dit au modèle si tu fais un erreur sur la classe minoritaire tu vas prendre des pénalité plus importante sur la Classe majoritaire

Donc le modèle vas éviter de faire des erreurs et comme ça on augmente la performance du modèle pour notre cas on a attribuée du poids pénalisant des erreur commis lors de l'apprentissage

3. Algorithme d'optimisation et métrique d'évaluation

Dans le cadre de cette mission, l'objectif est de prédire la probabilité d'un client de rembourser son prêt.



1 première métrique d'évaluation c'est **la précision**

la réponse à la question quelle est la proportion d'identification positive était effectivement correcte ?
Parmis la classe prédit Positives mais quelle sont les vraie positives

2 métrique la notion recall Rappel

la réponse à la question quelle est la proportion des résultas réel qui ont été identifiés correctement classe réel positives combien on l'a identifié avec prédiction

3• Accuracy : mesure la proportion des clients correctement classés parmi l'ensemble des clients

4 • F-Mesure : est la moyenne harmonique de Précision et Recall en donnant à chacun la même pondération
F-Mesures = { 2* Précision * Recall) / (Precision + Recall)

5• Fbeta-Mesure: La Fbeta-score est une généralisation de la F-mesure qui ajoute un paramètre de configuration appelé bêta
Fbeta-Mesures = ((1+beta au 2)* Precision * Recall) / (beta au 2* Precision + Recall)

4 Interprétabilité du modèle final

Le modèle présentant les meilleurs résultats F beta score le plus important Sur la base des différents résultats, le modèle utilisé est **Régression logistique**

L'objectif est de vérifier si la qualité de la prédiction est sensiblement identique (voire améliorée) ou si le modèle entraîné conduisait à de l'overfitting ou underfitting.

Scénarios	Situation	Gain / Perte
Des clients non payeurs classés comme des clients non payeurs (FP)	moyen	refus de crédit → l'intérêt généré par le prêt vaut 0
Des clients ayant remboursé classés comme des clients ayant remboursé (TN)	bien	accord de crédit → l'intérêt généré par le prêt
Des clients non payeurs classés comme des clients ayant remboursé (FN)	mauvais	accord de crédit → non remboursement & perte d'intérêt généré par le prêt
Des clients non payeurs classés comme des clients non payeurs (TP)	moyen	refus de crédit → l'intérêt généré par le prêt vaut 0

Parmi les 4 scénarios listés celui qui est à éviter en priorité est le scénario 3 (FN). Un volume important de« FN »génère un coût important pour l'entreprise car il représente l'ensemble des clients à qui un crédit a été accordé et qui finira par ne pas être remboursé."

"C'est pourquoi j'ai privilégié la métrique Recall par rapport à la Précision, car le Recall permet de plus tenir compte du volume de FN."

"Afin d'accorder plus d'importance au Recall dans mon Beta-mesure, on attribue Beta une valeur 2."

"Le modèle de régression logistique est le modèle qui obtient les meilleurs résultats selon la Beta-mesure, c'est pourquoi je décide de retenir ce modèle

5 Limites et améliorations possibles

Dans ce travail, compte tenu des temps de calcul pouvant être relativement importants, les modèles ont été entraînés sur des jeux de données restreints échantillons. En ayant accès à une puissance de calcul plus importante (exemple AWS), certaines caractéristiques spécifiques de certains clients pourraient plus facilement être prises en compte dans le modèle lors de la phase d'entraînement et conduire à des erreurs moins importantes de prédictions.

Aussi, un travail plus conséquent de FE pourrait permettre d'aller plus loin dans les résultats et de réduire les erreurs. Il est toutefois important de préciser que certaines variables explicatives qui ont une importance dans la probabilité de rembourser un prêt sont peu claires voire floues (ex:Ext Source), ce qui rend difficile la réalisation d'opérations de FE avec ces variables, dans le cadre d'une problématique métier précise.