Introduction à la prévision des performances des campagnes à l'aide de l'apprentissage automatique

Dans le paysage marketing concurrentiel actuel, comprendre le comportement des clients et prédire les résultats des campagnes sont essentiels pour optimiser les stratégies marketing et maximiser le retour sur investissement (ROI). En exploitant les données historiques des campagnes, les entreprises peuvent identifier les facteurs clés influençant l'acceptation des clients et améliorer leurs efforts de ciblage. Ce projet vise à utiliser des techniques d'apprentissage automatique, en particulier la régression logistique et les arbres de décision, pour prédire la probabilité d'acceptation d'une campagne et évaluer les implications financières de ces efforts de marketing.

Préparation des données et ingénierie des fonctionnalités

Pour commencer, nous organiserons nos données historiques de campagne en fonctionnalités et étiquettes structurées. Les fonctionnalités engloberont divers aspects des campagnes marketing, tels que :

- Dépenses publicitaires : le budget total alloué à chaque campagne.
- Segmentation du public : informations sur les facteurs démographiques, notamment le revenu, l'âge et la structure familiale (par exemple, la présence d'enfants).
- Mesures d'engagement : taux de clics, fréquence d'achat sur différents canaux (par exemple, Web, magasin, catalogue) et récence du dernier achat.

L'étiquette que nous prédisons indique si une campagne a été « acceptée » ou « non acceptée ». En divisant l'ensemble de données en ensembles de formation et de test, nous pouvons garantir que notre modèle est formé sur des données historiques tout en étant évalué sur des données invisibles pour une évaluation impartiale des performances.

Implémentation de modèles d'apprentissage automatique

Nous appliquerons la régression logistique, une méthode populaire de classification binaire, et un modèle d'arbre de décision pour analyser les données d'acceptation de la campagne. Ces modèles nous aideront à comprendre comment différentes fonctionnalités contribuent à la probabilité d'acceptation de la campagne.

Évaluation du modèle avec matrice de confusion

Après avoir entraîné nos modèles, nous ferons des prédictions à l'aide de l'ensemble de test et évaluerons les performances à l'aide de matrices de confusion. Ces matrices fourniront des informations sur les vrais positifs (TP), les faux positifs (FP), les vrais négatifs (TN) et les faux négatifs (FN), nous permettant de quantifier l'efficacité de nos campagnes.

Calcul des bénéfices

En utilisant les résultats des matrices de confusion, nous calculerons les bénéfices en fonction des performances de nos campagnes. Plus précisément, le profit proviendra des campagnes acceptées (TP) moins les coûts associés aux faux positifs (FP). Nous envisagerons également un calcul de profit révisé qui inclut les opportunités manquées potentielles, offrant ainsi une vue complète des résultats financiers de nos efforts de marketing.

Segmentation client et valeur à vie

De plus, nous explorerons la segmentation des clients en fonction des données démographiques et des mesures d'engagement, ce qui peut conduire à des stratégies marketing plus ciblées. L'analyse des modèles de dépenses et de la valeur à vie du client améliorera encore notre compréhension de la façon dont les différents segments réagissent aux campagnes.

Conclusion

Ce projet vise non seulement à prédire l'acceptation des campagnes à l'aide de l'apprentissage automatique, mais cherche également à fournir

des informations exploitables sur le comportement des clients et l'efficacité des campagnes. En comprenant la dynamique des performances des campagnes, les entreprises peuvent affiner leurs stratégies, générant ainsi des revenus plus élevés et réduisant les coûts associés à des efforts marketing inefficaces. Grâce à cette approche systématique, nous espérons établir un cadre pour une prise de décision basée sur les données dans les campagnes marketing.

Préparation des données:

Organisez les données de vos campagnes passées en fonctionnalités et en étiquettes. Par exemple, les fonctionnalités peuvent inclure des facteurs tels que les dépenses publicitaires, la segmentation de l'audience, les taux de clics, les données démographiques et les mesures d'engagement. L'étiquette indiquerait si la campagne a été « acceptée » ou « non acceptée ».

Divisez l'ensemble de données en ensembles de formation et de test.

Modèle d'apprentissage automatique simple:

Utilisez un classificateur de régression logistique ou un arbre de décision pour la classification binaire. Il s'agit de méthodes simples mais efficaces pour générer des prédictions dans un scénario d'acceptation de campagne.

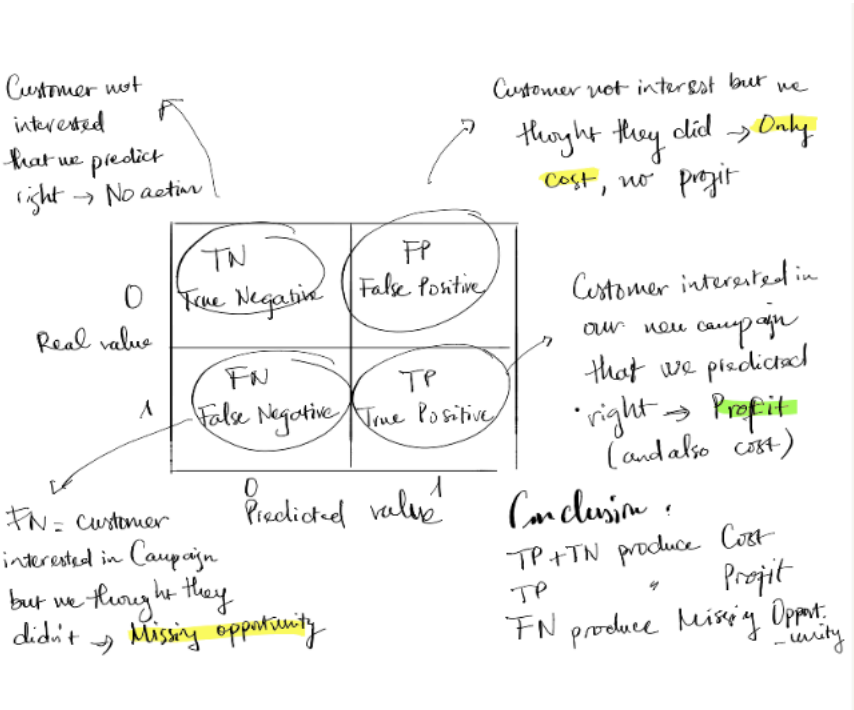
Matrice de prédiction et de confusion:

Entraînez le modèle sur l'ensemble d'entraînement et utilisez l'ensemble de test pour faire des prédictions.

Générez une matrice de confusion pour évaluer les performances du modèle en termes de vrais positifs (TP), de faux positifs (FP), de vrais négatifs (TN) et de faux négatifs (FN).

Calcul des bénéfices:

Utilisez les valeurs TP et FP de la matrice de confusion pour calculer le profit. Supposons que les vrais positifs génèrent des bénéfices et que les faux positifs entraînent des coûts (par exemple, pour l'envoi de campagnes qui ne conduisent pas à des conversions). Le profit peut être calculé sur la base du nombre de campagnes correctement acceptées (TP) moins le coût des faux positifs (FP).



Calculate Profit

The profit is calculated as:

$$\text{Profit} = \text{Campaign Revenue} - \text{Campaign Cost}$$

or a revised Profit which reflect both potential revenue and missed opportunity:

$$\text{Revised Profit} = \text{Campaign Revenue} - \text{Campaign Cost} - \text{Missed Opportunity}$$

Where:

- $\text{Campaign Revenue} = \text{Revenue per accepted customer} \times TP$
- $\text{Campaign Cost} = \text{Cost per customer} \times (TP + FP)$
- $\text{Missed Opportunity} = (\text{Revenue per accepted customer} - \text{Cost per customer}) \times FN$

Fonctionnalités à mettre en œuvre :

Segmentation client:

Utiliser **Revenu**, **Âge**, et **Structure familiale (Kidhome, Teenhome)** pour regrouper les clients en segments significatifs pour des stratégies marketing ciblées.

Dépense totale:

Total Min est déjà présent, mais vous pourriez envisager d'analyser les modèles de dépenses dans différentes catégories de produits telles que **M. Vins**, **MntMeatProduits**, etc., pour comprendre quels produits génèrent le plus de valeur.

Engagement client:

Récence mesure le nombre de jours depuis le dernier achat du client. Cela peut être un élément clé pour déterminer l'engagement des clients ou la probabilité de répondre aux campagnes.

NumWebAchats, **NumStoreAchats**, et **NumCatalogueAchats** indiquer les canaux d'achat des clients. Vous pouvez les regrouper pour analyser leur méthode d'achat préférée.

Analyse des réponses à la campagne:

L'acceptation de différentes campagnes (**AcceptéCmp1-5**) et l'acceptation globale de la campagne (**AcceptéCmpOverall**) peuvent être analysés pour évaluer l'efficacité de la campagne.

Vous pouvez créer une fonctionnalité supplémentaire pour le **nombre total de campagnes acceptées** en résumant les colonnes binaires.

Prédiction de la valeur à vie:

Client_Days représente le nombre de jours pendant lesquels le client est dans l'entreprise. Ceci peut être utilisé pour prédire la valeur à vie du client lorsqu'il est combiné avec **Total Min** et les réponses à la campagne.

État civil et éducation:

État civil (**marital_* domaines) et le niveau d'éducation (**education_*) peuvent être utilisés comme variables catégorielles pour comprendre leur impact sur les dépenses et la réponse à la campagne.

Focus sur les données:

Le code se concentre désormais sur les fonctionnalités liées à la campagne (AcceptéCmp1, AcceptéCmp2, etc.) et prédit si une campagne a été acceptée (CampagneAcceptée).

Matrice de confusion:

Deux matrices de confusion sont imprimées (une pour la régression logistique, une pour XGBoost), et elles sont également visualisées à l'aide de seaborn.heatmap.

Faux positifs et faux négatifs:

Le code extrait les faux positifs (FP) et les faux négatifs (FN) des matrices de confusion pour les deux modèles.

Sorties :

Matrice de confusion:

Les matrices de confusion pour les deux modèles sont imprimées et tracées.

FP et FN:

Vous verrez le nombre de faux positifs et de faux négatifs pour la régression logistique et XGBoost.

Rapport de classement:

Un rapport de classification est généré pour XGBoost, qui inclut la précision, le rappel, le score f1 et l'exactitude.

Hypothèses pour la régression logistique

La variable dépendante est binaire :
La variable dépendante doit être classée en deux catégories. Cela signifie que la régression logistique prédira la probabilité d'un événement selon deux scénarios : l'événement se produit, 1, ou l'événement ne se produit pas, 0.

Distribution gaussienne:
La régression logistique suppose que la relation entre les variables (entrée et sortie) est linéaire.

Les variables indépendantes ne doivent pas avoir de multi-colinéarité :
Cela signifie qu'il ne devrait y avoir aucune ou très peu de corrélation entre la variable indépendante/prédictive.

Taille d'échantillon plus grande :
L'analyse de régression logistique nécessite un échantillon de grande taille. Un échantillon de grande taille génère des résultats d'analyse fiables.

Quels sont les types de régression logistique ?