ÉTANDEX

Et vos ouvrages durent plus longtemps

Junior CentraleSupélec

AUTOMATANTS

MISSION JCS X ETANDEX

Estimation des chances de succès d'une opportunité commerciale par un algorithme de Machine Learning

29 AOUT 2023

Clement Wang

clementwang2001@gmail.com

Table des matières

Mise en contexte

Objectifs de la mission

L'objectif de la mission menée conjointement par la Junior CentraleSupélec et l'entreprise Etandex est d'accroître la compétence de l'entreprise à évaluer ses probabilités de réussite de conversion d'une opportunité en un contrat conclu.

Les principaux objectifs à long terme de cette mission incluent :

- Une amélioration de la gestion des efforts commerciaux : En identifiant les opportunités présentant les meilleures perspectives de réussite, l'entreprise peut optimiser l'allocation de ses ressources commerciales en se focalisant sur les clients les plus enclins à conclure une transaction.
- Une meilleure estimation du chiffre d'affaires potentiel : Grâce à une évaluation plus précise des probabilités de succès, Etandex peut anticiper de manière plus réaliste les revenus générés par les opportunités en cours, ce qui revêt une importance cruciale pour la planification financière et la gestion des ressources.

Méthodologie

Pour accomplir ces objectifs, l'équipe de la Junior CentraleSupélec se base sur les données historiques des opportunités de l'entreprise collectées à partir de Salesforce, en se focalisant sur des variables telles que l'activité, la localisation géographique (ville/département), ainsi que les spécificités du client.

Au moyen d'outils statistiques et d'analyses de données, l'équipe examine les données recueillies pour déceler des liens entre ces caractéristiques et le succès des opportunités. En se fondant sur les conclusions de cette analyse, l'équipe élabore un modèle de prédiction qui évalue les probabilités de succès d'une opportunité en fonction de ses attributs. Par ailleurs, un aspect crucial de la mission consiste à décrypter les raisons sous-tendant les prédictions formulées par l'algorithme développé.

Nettoyage et mise en forme des données

Le projet repose sur l'utilisation de deux tables fournies par Etandex, à savoir les tables "*Opportunités*" et "*Staff*". De plus, des opérations de jointure sont effectuées en intégrant des données provenant de l'INSEE.

Opportunités

La table "*Opportunités*" renferme toutes les informations fondamentales concernant les contrats potentiels. Dans le cadre du traitement des données, plusieurs étapes sont entreprises :

- Une première étape consiste à filtrer les lignes présentant le même numéro d'opportunité, en ne conservant que la première occurrence afin d'éviter les doublons.
- Les lignes dépourvues d'informations concernant le montant ou celles ayant une valeur négative sont supprimées, et les valeurs sont converties en format numérique (float).
- Le codage binaire "one hot encoding" est appliquée aux variables Opération, Domaine et Position.
- La méthode de codage "one hot encoding" est également employée pour les variables Origine de l'opportunité, Type de compte, Activité et Rôle du contact.

Jointure avec la table Staff

La table "Staff" est associée au projet pour fournir des informations sur les employés. Les étapes suivantes sont effectuées dans le cadre de cette jointure :

- L'identifiant du membre du personnel "LIL" est corrigé pour devenir "MIL", car il est probablement le résultat d'une simple erreur d'inattention.
- Deux nouvelles colonnes, "Âge" et "Ancienneté", sont créées pour contenir les informations correspondantes sur les employés.
- La méthode de codage "one hot encoding" est appliquée aux variables DiplomeFinal et FicheDePoste.
- Une opération de jointure est effectuée entre la table "Opportunités" (en tant que responsable) et la table "Staff" en utilisant l'identifiant comme clé de jointure.

Jointure avec la table INSEE communes

La table "INSEE communes" est fusionnée avec d'autres données pour enrichir l'ensemble. Elle contient diverses informations sur les communes, telles que le code postal, le nom, le numéro de département, la superficie, la population, etc. La colonne "Ville" de la table "Opportunités" contient généralement un code postal suivi du nom de la commune. Les étapes de cette fusion sont les suivantes :

- Remplacement de la valeur "75019" par "75019 PARIS-19E__ARRONDISSEMENT" dans le champ "Ville" de la table "Opportunités".
- Création de deux nouvelles colonnes, "nom_ville" et "code_postal", en séparant le champ "Ville" de la table "Opportunités". Cette séparation n'est effectuée que si le champ commence par cinq chiffres. Dans le cas contraire, le champ est laissé vide (NaN). Cette opération permet notamment de filtrer les communes qui ne sont pas en France.
- Certaines communes correspondent à plusieurs codes postaux, répertoriés dans le champ "Code Postal" de la table "INSEE communes" et séparés par des slashes ("/"). Une première étape consiste à créer une ligne distincte pour chaque code postal, préparant ainsi les données pour la jointure.
- Pour traiter les codes postaux incorrects dans la table "Opportunités", une recherche manuelle des noms de villes manquants dans la table "Opportunités" est effectuée dans la table "INSEE communes". Si une correspondance est trouvée, le code postal incorrect

est ajouté à la table "INSEE communes". Des lignes manuelles sont également ajoutées pour gérer des cas spéciaux, tels que Paris, Lyon et Marseille, qui fonctionnent avec des arrondissements.
- La jointure finale est effectuée en utilisant le champ "Code Postal" pour associer les données de la table "Opportunités" à celles de la table "INSEE communes".
- Les codes départementaux "2A" et "2B" sont remplacés par "20" pour simplifier le traitement en les convertissant en entiers (int).
- Seule la colonne "département" est conservée dans cette table fusionnée.

Labels

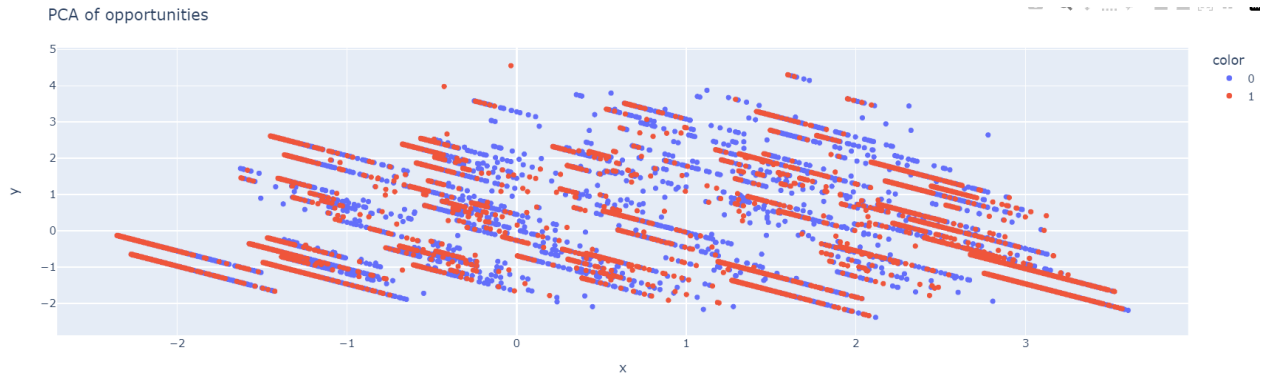
La valeur effective d'une opportunité est évaluée en se basant sur la colonne "Étape". Cette colonne peut avoir l'une des valeurs suivantes :

- 0- Etude amont / Budget
- 1- Priorisation
- 2- Dépôt de candidature
- 3- Réponse en préparation
- 4- Réponse envoyée
- 5- Négociation prix
- 6- Gagnée
- 7- Perdue
- 8- Perdue archivée
- 9- Gagnée archivée

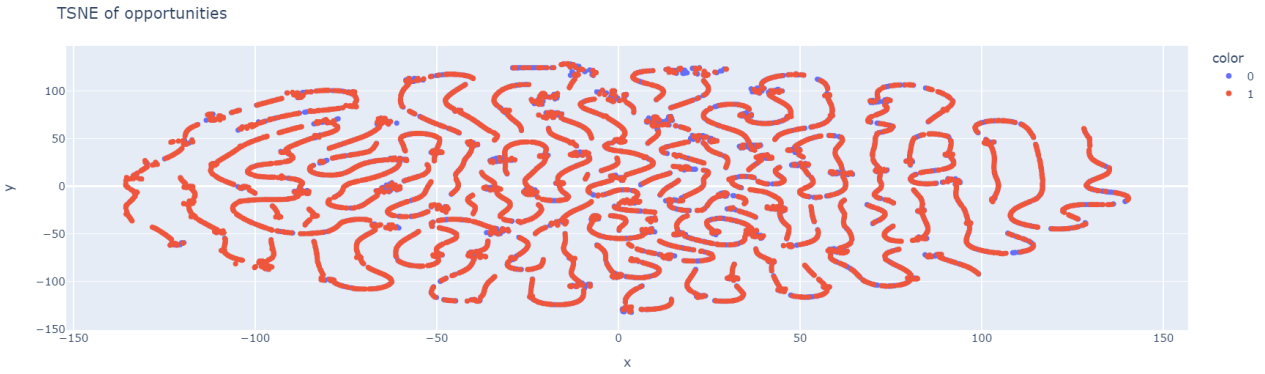
Si la colonne "Étape" est définie comme "Gagnée" ou "Gagnée archivée", l'opportunité est classée comme une victoire. En revanche, si elle est marquée comme "Perdue" ou "Perdue archivée", elle est répertoriée comme une défaite. Pour les autres cas, si la "date de dernière modification" remonte à plus de deux ans, l'opportunité est également considérée comme perdue, et toutes les autres entrées sont retirées de la table.

Visualisation des données

Lorsqu'on aborde un problème de machine learning, il s'avère toujours bénéfique de examiner visuellement les données afin d'éviter d'appliquer un algorithme excessivement sophistiqué dans une situation qui est en réalité assez simple.



PCA de la table Opportunités uniquement



TNSE de la table Opportunités uniquement

La première visualisation consiste en une réduction de la dimension en projetant les données dans un espace 2D à l'aide d'une projection linéaire, tandis que la seconde repose sur une projection non linéaire qui préserve au maximum les distances. Ces deux visualisations mettent en évidence une difficulté significative du problème, car il est impossible de discerner des tendances claires entre les classes dans ces représentations visuelles.

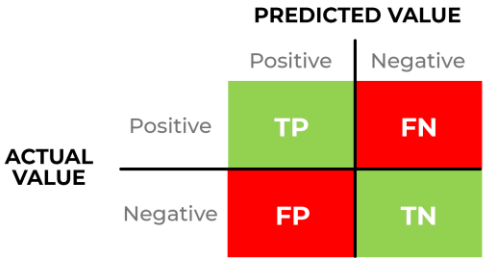
Entraînement des algorithmes

Dataset d'entraînement et dataset de test

Après avoir nettoyé et structuré les données, la table finale est partitionnée en deux sous-tables : 80% de la table est réservée à l'entraînement de l'algorithme, tandis que les 20% restants sont utilisés pour évaluer les performances des algorithmes. Cette approche permet de quantifier l'overfitting, qui se manifeste lorsque le modèle apprend spécifiquement à partir des données d'entraînement et de leurs étiquettes plutôt que de capturer des relations générales entre les données et leurs étiquettes. En outre, la division est effectuée de manière à maintenir les mêmes proportions entre l'ensemble de données d'entraînement et l'ensemble de données de test.

Métrique d'évaluation

La métrique d'évaluation choisie est l'exactitude (accuracy). Cette métrique mesure la capacité du modèle à classer correctement les données, en fournissant le pourcentage de prédictions correctes par rapport à l'ensemble total des données. Une exactitude élevée indique que le modèle est performant dans sa capacité à prédire avec précision les étiquettes des données, tandis qu'une exactitude faible peut signaler des problèmes de classification.



$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Lorsqu'il s'agit de traiter des données tabulaires, les algorithmes d'arbres de décision sont très répandus et appréciés, aussi bien dans le domaine de la recherche que dans des applications pratiques. Dans le cadre de ce projet, quatre algorithmes ont été évalués :

- **Régression logistique (Logistic Regression)** : La régression logistique est un algorithme classique utilisé pour la classification. Il modélise la probabilité qu'une observation appartienne à une classe donnée en utilisant une fonction logistique. C'est un modèle linéaire qui peut être adapté à des données tabulaires et est souvent utilisé comme point de départ pour la classification.
- **Random Forest** : Les forêts aléatoires sont une technique d'ensemble basée sur les arbres de décision. Elles construisent plusieurs arbres de décision de manière aléatoire et combinent leurs prédictions pour améliorer la précision. C'est une méthode puissante pour la classification et la régression, souvent utilisée pour traiter des données tabulaires complexes.
- **XGBoost** : XGBoost est un algorithme de gradient boosting qui est devenu très populaire dans les compétitions de science des données. Il améliore progressivement les prédictions en construisant des arbres de décision en série. XGBoost est apprécié pour sa robustesse et sa performance, en particulier sur des ensembles de données tabulaires.
- **LightGBM** : LightGBM est un autre algorithme de gradient boosting qui se distingue par sa rapidité d'exécution. Il utilise une technique appelée "Gradient Boosting with Histogram-based Learning" pour accélérer le processus d'apprentissage. LightGBM est particulièrement efficace sur des ensembles de données tabulaires de grande taille.

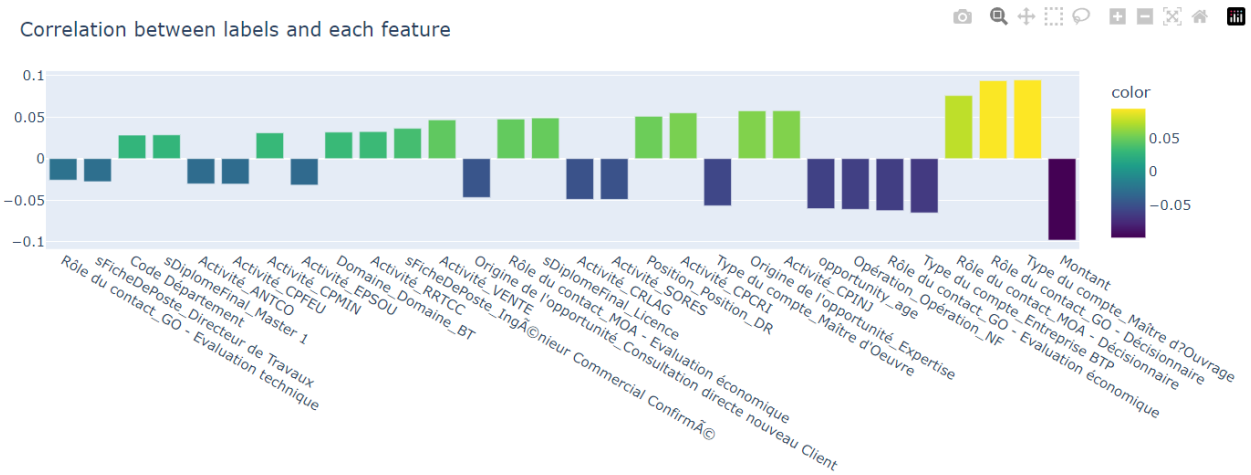
Lors de la phase d'entraînement, une pondération accrue est attribuée aux données récentes. Toutes les opportunités datant de moins de 3 ans se voient attribuer un poids de 1 pendant l'entraînement. Ensuite, ce poids diminue de manière linéaire jusqu'à atteindre une valeur de 0,2 pour l'opportunité la plus ancienne.

Nom	Accuracy sur train set (%)	Accuracy sur test set (%)
Régression logistique	64.4	64.4
Random forest	75.4	68.9
XGBoost	77.0	69.0
LightGBM	72.9	69.0

Explicabilité

Graphe des corrélations

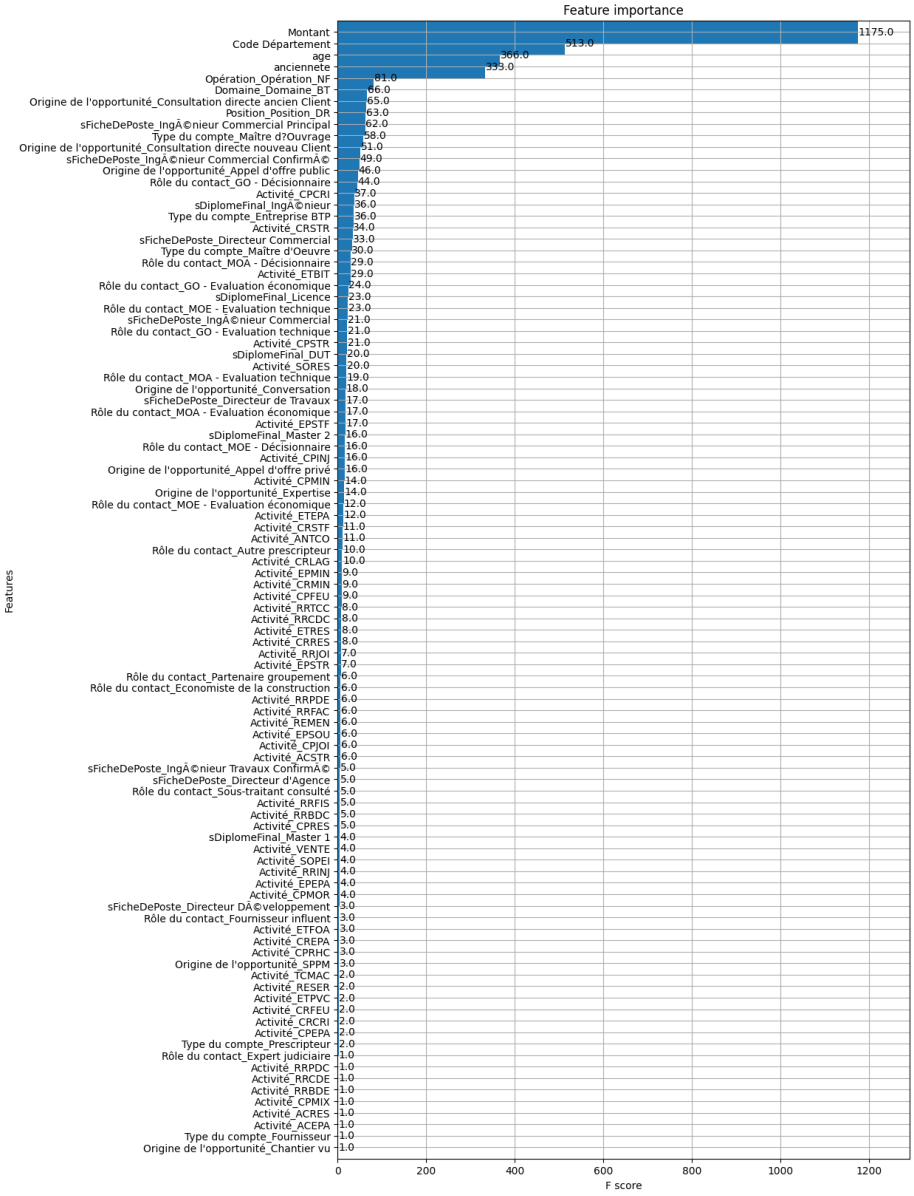
La corrélation constitue une première étape vers l'explication. Elle offre une vue initiale de la relation entre les variables d'entrée et l'étiquette attendue.



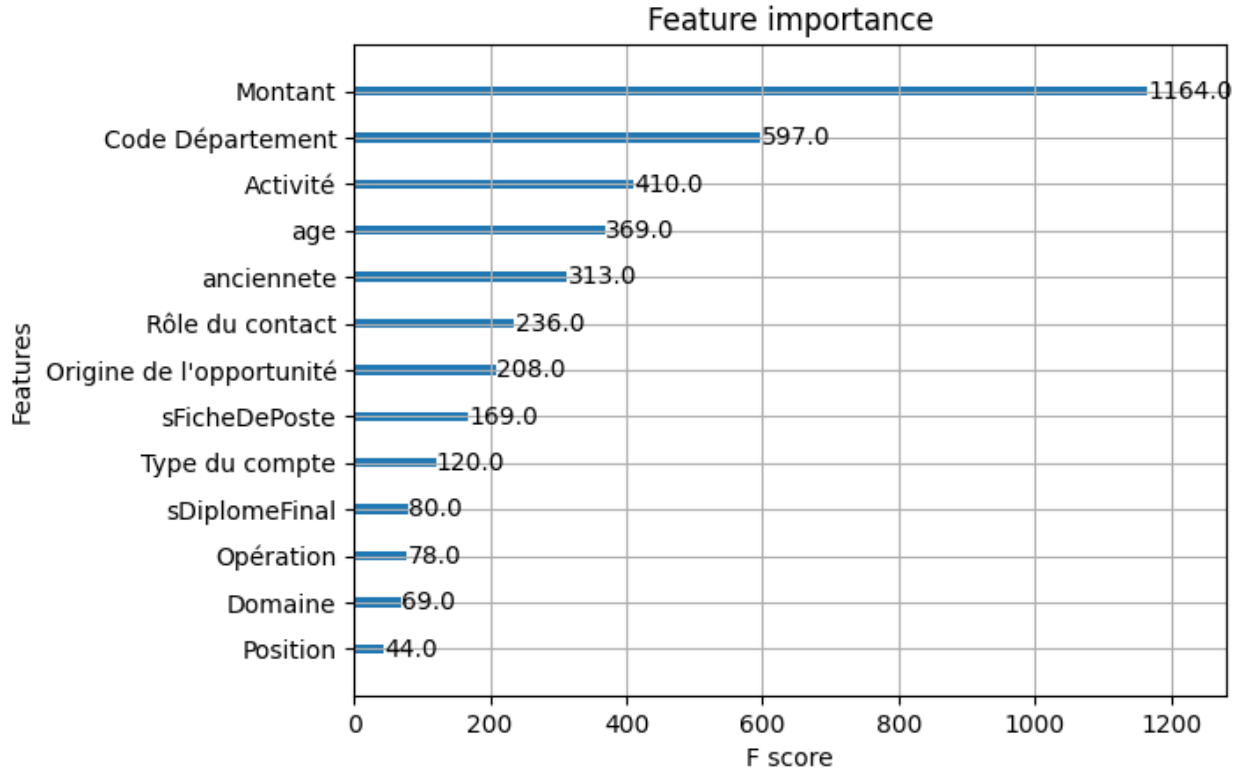
À titre d'exemple, le paramètre relatif au "montant" présente une corrélation négative, ce qui signifie qu'en règle générale, à mesure que le montant augmente, le taux de réussite d'une opportunité a tendance à diminuer.

Explicabilité globale de XGBoost par Feature importance

XGBoost et LightGBM sont des algorithmes de Gradient boosting. Ces méthodes ajoutent progressivement des arbres de décision en sélectionnant de manière itérative une caractéristique à chaque nœud. L'importance des caractéristiques (feature importance) est évaluée en fonction du nombre de fois où chaque caractéristique a été utilisée lors de la création des divisions (splits) dans les arbres.



Feature importance issue de XGBoost

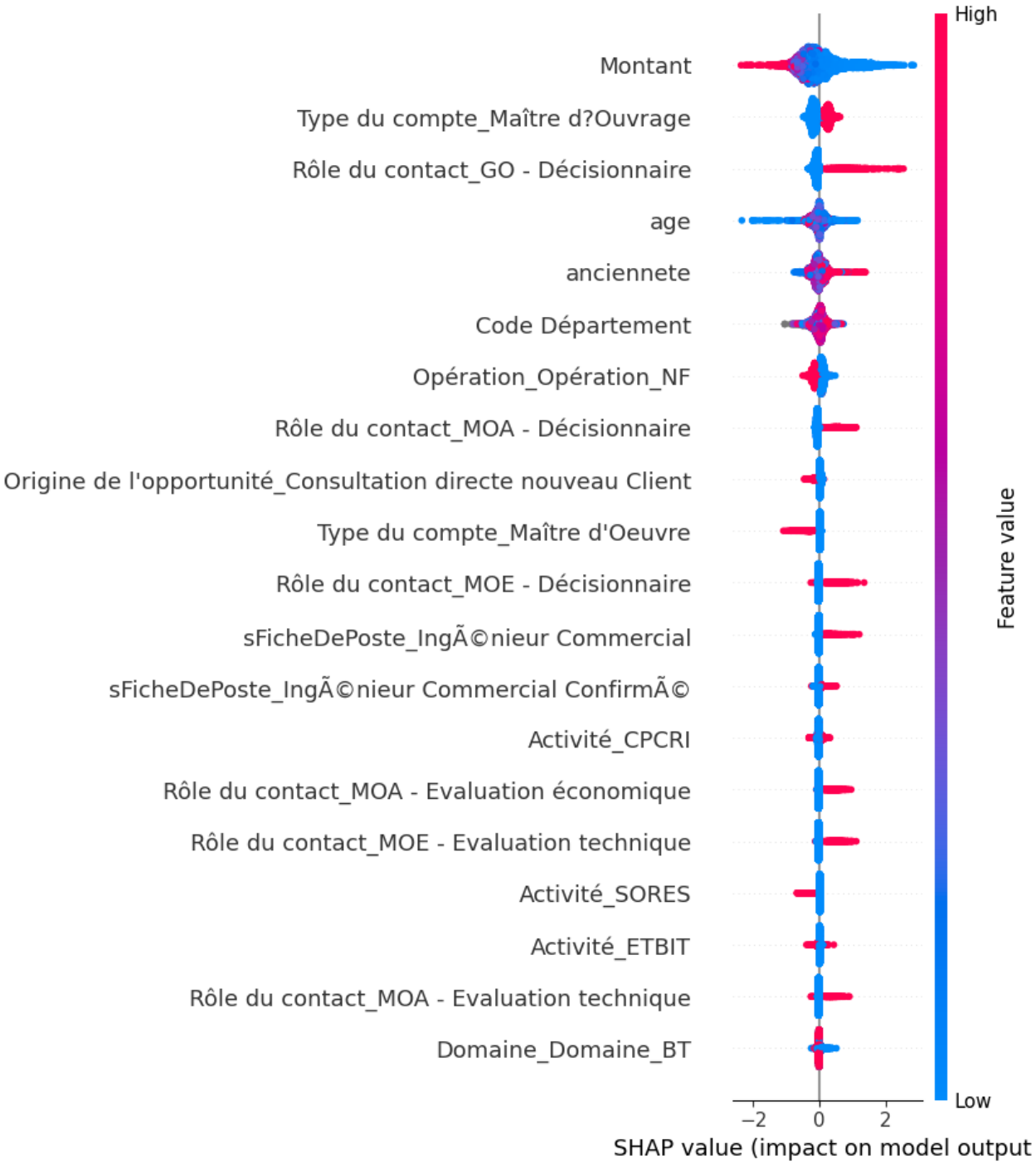


Feature importance issue de XGBoost avec réduction des catégories

Il convient de noter que l'algorithme ne conduit à aucune conclusion de causalité ; tout repose sur des corrélations. Par exemple, l'âge du commercial responsable de l'opportunité peut être attribué à une importance élevée dans la décision de l'algorithme. Cependant, cette importance pourrait découler de la présence d'un employé exceptionnel dans une tranche d'âge spécifique, plutôt que d'une relation de cause à effet directe entre l'âge et le succès de l'opportunité.

Explicabilité globale par SHAP value

La SHAP value offre une manière plus détaillée d'analyser l'influence d'une caractéristique sur la prédiction de probabilité générée par un algorithme. Elle se base sur la théorie de Shapley, un concept de la théorie des jeux, pour attribuer une contribution à chaque caractéristique dans la prédiction finale. En d'autres termes, elle mesure comment chaque caractéristique a contribué à la différence entre la prédiction du modèle pour une observation spécifique et la valeur de prédiction moyenne pour l'ensemble des observations.



Graphe agrégé des SHAP values

Prenons un exemple : le paramètre "Rôle du contact - décisionnaire" conserve constamment une influence positive sur la décision, tandis que l'ancienneté peut avoir soit une influence positive, soit une influence négative, en fonction des cas.

Explicabilité locale par SHAP value

La valeur SHAP fournit une explication plus approfondie sur la manière dont un paramètre spécifique influence une prédiction particulière.

Dans cet exemple, le montant apporte une contribution positive, et cette contribution est de magnitude supérieure à la contribution négative attribuée au type de compte, qui est celui d'un maître d'œuvre.

Explicabilité locale par Anchor

Le principe d'Anchor est une technique d'explicabilité utilisée pour expliquer les prédictions de modèles de machine learning, y compris ceux basés sur XGBoost. L'idée principale de l'Anchor est de fournir une règle simple et interprétable qui définit une condition sous laquelle la prédiction d'un modèle est valable. Concrètement, Anchor identifie un sous-ensemble de caractéristiques et une plage de valeurs pour ces caractéristiques, de telle sorte que si ces conditions sont remplies, la prédiction du modèle reste cohérente avec la prédiction d'origine.

En d'autres termes, Anchor cherche à établir une "règle ancrée" qui explique pourquoi le modèle a fait une certaine prédiction. Cela permet aux utilisateurs de comprendre plus facilement comment le modèle prend des décisions, en fournissant une explication simple et compréhensible sous forme de règles conditionnelles.

Améliorations

Plusieurs pistes d'amélioration peuvent être envisagées pour optimiser ce projet. Tout d'abord, il serait judicieux de procéder au fine-tuning de l'algorithme utilisé afin de minimiser l'overfitting et d'améliorer ses performances globales. Cela pourrait impliquer une recherche plus approfondie des hyperparamètres de l'algorithme. En outre, une méthodologie plus rigoureuse de sélection des caractéristiques pourrait être mise en place pour réduire l'overfitting en éliminant les variables moins pertinentes ou redondantes, simplifiant ainsi le modèle tout en préservant sa capacité prédictive.

Annexe : Table finale

La table finale englobe un total de 39 488 lignes et comprend 107 colonnes distinctes. Ci-dessous, vous trouverez les libellés de chacune de ces colonnes :

```
['Montant',
 'opportunity_age',
 'Opération_Opération_NF',
 'Domaine_Domaine_BT',
 'Position_Position_DR',
 "Origine de l'opportunité_Appel d'offre privé",
 "Origine de l'opportunité_Appel d'offre public",
 "Origine de l'opportunité_Chantier vu",
 "Origine de l'opportunité_Consultation directe ancien Client",
 "Origine de l'opportunité_Consultation directe nouveau Client",
 "Origine de l'opportunité_Conversation",
 "Origine de l'opportunité_Expertise",
 "Origine de l'opportunité_SPPM",
 'Type du compte_Entreprise BTP',
 'Type du compte_Fournisseur',
 "Type du compte_Maître d'Oeuvre",
 'Type du compte_Maître d?Ouvrage',
 'Type du compte_Prescripteur',
 'Activité_ACEPA',
 'Activité_ACRES',
 'Activité_ACSTR',
 'Activité_AMNFR',
 'Activité_ANTCO',
 'Activité_CPCRI',
 'Activité_CPEPA',
 'Activité_CPFEU',
 'Activité_CPINJ',
 'Activité_CPJOI',
 'Activité_CPMIN',
 'Activité_CPMIX',
 'Activité_CPMOR',
 'Activité_CPRES',
 'Activité_CPRHC',
 'Activité_CPSTR',
 'Activité_CRCRI',
 'Activité_CREPA',
 'Activité_CRFEU',
 'Activité_CRLAG',
 'Activité_CRMIN',
 'Activité_CRRES',
 'Activité_CRSTF',
 'Activité_CRSTR',
 'Activité_EPEPA',
 'Activité_EPEQP',
 'Activité_EPMIN',
 'Activité_EPSOU',
```

```
'Activité_EPSTF',
'Activité_EPSTR',
'Activité_ETBIT',
'Activité_ETEPA',
'Activité_ETFOA',
'Activité_ETPVC',
'Activité_ETRES',
'Activité_REMEN',
'Activité_RESER',
'Activité_RRADC',
'Activité_RRADE',
'Activité_RRBDC',
'Activité_RRBDE',
'Activité_RRCDC',
'Activité_RRCDE',
'Activité_RRDEM',
'Activité_RRFAC',
'Activité_RRFIS',
'Activité_RRINJ',
'Activité_RRJOI',
'Activité_RRPDC',
'Activité_RRPDE',
'Activité_RRTCC',
'Activité_SOPEI',
'Activité_SORES',
'Activité_TCBET',
'Activité_TCMAC',
'Activité_VENTE',
'Rôle du contact_Autre prescripteur',
'Rôle du contact_Concurrent',
'Rôle du contact_Economiste de la construction',
'Rôle du contact_Expert judiciaire',
'Rôle du contact_Fournisseur influent',
'Rôle du contact_GO - Décisionnaire',
'Rôle du contact_GO - Evaluation technique',
'Rôle du contact_GO - Evaluation économique',
'Rôle du contact_MOA - Décisionnaire',
'Rôle du contact_MOA - Evaluation technique',
'Rôle du contact_MOA - Evaluation économique',
'Rôle du contact_MOE - Décisionnaire',
'Rôle du contact_MOE - Evaluation technique',
'Rôle du contact_MOE - Evaluation économique',
'Rôle du contact_Partenaire groupement',
'Rôle du contact_Programmiste de la construction',
'Rôle du contact_Sous-traitant consulté',
'Label',
'sDiplomeFinal_DUT',
'sDiplomeFinal_IngÃ©nieur',
'sDiplomeFinal_Licence',
'sDiplomeFinal_Master 1',
'sDiplomeFinal_Master 2',
```

```
'sFicheDePoste_Directeur Commercial',
'sFicheDePoste_Directeur Commercial France',
'sFicheDePoste_Directeur DÃ©veloppement',
"sFicheDePoste_Directeur d'Agence",
'sFicheDePoste_Directeur de Travaux',
'sFicheDePoste_IngÃ©nieur Commercial',
'sFicheDePoste_IngÃ©nieur Commercial ConfirmÃ©',
'sFicheDePoste_IngÃ©nieur Commercial Principal',
'sFicheDePoste_IngÃ©nieur Travaux ConfirmÃ©',
'age',
'anciennete',
'Code Département']
```