UQAC

8INF867 FONDEMENTAUX DE L'APPRENTISSAGE AUTOMATIQUE

Projet de session

December 14, 2023

Ali Ghammaz GHAA27070100
Atoumane Gaye GAYA19039305
Eric Moussinga

Contents

List of Figures

1 Introduction

ProsperPal, se positionne comme un guide essentiel pour accompagner les Canadiens en difficulté financière vers l'atteinte de la richesse. Notre équipe de conseillers financiers experts est dédiée à utiliser nos modèles d'apprentissage automatique pour offrir une assistance proactive dans divers aspects financiers.

Nos services comprennent une compréhension approfondie de la relation de nos clients avec l'argent et la richesse, des analyses approfondies de leur état financier actuel, l'établissement de plans de remboursement de dettes personnalisés, et des propositions avisées pour des opportunités d'investissement.

Ce rapport expose en détail les différentes composantes de ProsperPal, décrivant comment nous appliquons les modèles d'apprentissage automatique pour comprendre les attitudes financières individuelles, réaliser des analyses approfondies des situations financières personnelles, regrouper les clients en segments distincts en fonction de leurs caractéristiques financières, et enfin, comment nous aidons nos clients à prendre des décisions d'investissement éclairées grâce à nos modèles de prédiction des marchés boursiers.

2 Etude du marché

2.1 Clients cible

Notre marché cible s'adresse aux individus souhaitant améliorer leur compréhension financière et maximiser leurs investissements. Nous mettons l'accent sur les jeunes professionnels, les entrepreneurs et toute personne confrontée à des défis de gestion budgétaire, cherchant à bénéficier d'un conseil financier personnalisé.

2.2 Analyse de la demande

Les problèmes financiers au Canada se concentrent principalement sur trois aspects : la difficulté financière, l'endettement et l'anxiété financière.

Difficulté Financière :

- Une part significative des Canadiens, vivant dans des ménages déclarant des difficultés à répondre à leurs besoins financiers de base, a atteint un pic en mai 2020, représentant 22,2% de la population.

Endettement :

- Environ un Canadien sur six (17%) affirme que ses dépenses dépassent ses revenus, et le quart (27%) indique emprunter pour couvrir des dépenses courantes, voire acheter des vivres.

Anxiété Financière :

- Plus de la moitié des Canadiens reconnaissent être à 200 dollars ou moins de l'incapacité de payer toutes leurs factures à la fin du mois. Par ailleurs, près de la moitié des Québécois (48%) ressentent de l'anxiété à l'idée de leurs finances, avec une tendance à l'évitement plus marquée chez les jeunes adultes de 18 à 34 ans, les femmes et les chefs de famille (33%).

Cette analyse souligne l'ampleur des défis financiers rencontrés par de nombreux Canadiens, renforçant ainsi la nécessité d'une solution proactive pour les aider à mieux gérer leurs finances.

2.3 Analyse de la Pratique Budgétaire au Canada

Selon le gouvernement canadien, l'établissement d'un budget demeure un élément indispensable pour de nombreux Canadiens afin de gérer efficacement leurs finances courantes, les paiements de factures et le remboursement des dettes. Près de la moitié des participants (49%), selon l'Enquête canadienne sur les capacités financières de 2019, déclarent avoir un budget, en hausse par rapport à 2014. Les outils numériques, tels que les applications mobiles, sont largement utilisés (20%), tandis que des méthodes traditionnelles comme le budget manuscrit persistent (14%). Cependant, l'enquête révèle qu'un Canadien sur six (17%) pourrait bénéficier de l'établissement d'un budget. Les raisons évoquées pour l'absence de budget incluent le manque de temps ou d'intérêt (9%) et le sentiment d'être dépassé par la gestion financière (6%).

Comparativement, ceux qui ont un budget sont moins susceptibles de manquer à leurs obligations financières (8% contre 16%). Ces individus

gèrent plus efficacement leurs finances mensuelles, évitant de vivre au-dessus de leurs moyens (18% contre 29%) ou d'emprunter pour couvrir des dépenses courantes (31% contre 42%). Notamment, ceux qui utilisent des outils numériques sont parmi les plus aptes à payer leurs factures à temps et à maintenir le contrôle de leurs finances. De plus, par rapport à ceux qui ne font pas de budget en raison du manque de temps ou du sentiment d'être dépassés, ceux qui établissent un budget sont plus enclins à accélérer le remboursement de leur prêt hypothécaire (35% contre 24%) et de leurs autres dettes (57% contre 47%), avec une différence d'environ 10 points de pourcentage dans chaque cas. Ces résultats soulignent l'importance de l'établissement d'un budget dans la gestion financière, particulièrement lorsqu'il est réalisé de manière proactive à l'aide d'outils numériques.

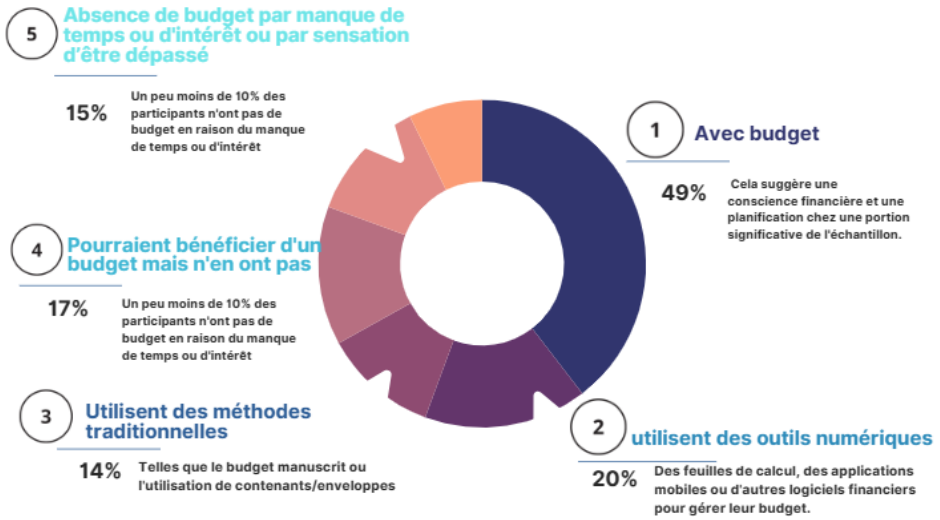


Figure 1: Gestion financière des canadiens

2.4 Identification des concurrents

Dans le paysage concurrentiel, ProsperPal se distingue au milieu de divers concurrents : Les concurrents directs incluent les firmes de conseil financier traditionnelles, telles que les banques et les sociétés de gestion de patrimoine, qui offrent des services personnalisés, mais souvent à des tarifs élevés. Les plateformes de robo-advisors, comme Wealthsimple et Questwealth, sont également des concurrents directs en fournissant des conseils d'investissement automatisés en ligne. Les applications de ges-

tion budgétaire, telles que Mint et YNAB, bien qu'elles ne proposent pas un conseil financier complet, sont des concurrents directs en matière de gestion quotidienne des finances.

D'autre part, les concurrents indirects englobent les applications de suivi de crédit, comme Credit Karma, qui offrent des fonctionnalités similaires, mais se concentrent davantage sur la gestion du crédit. Les plateformes éducatives en ligne, telles qu'Investopedia, fournissent des conseils financiers généraux, mais ne proposent pas nécessairement des conseils personnalisés.

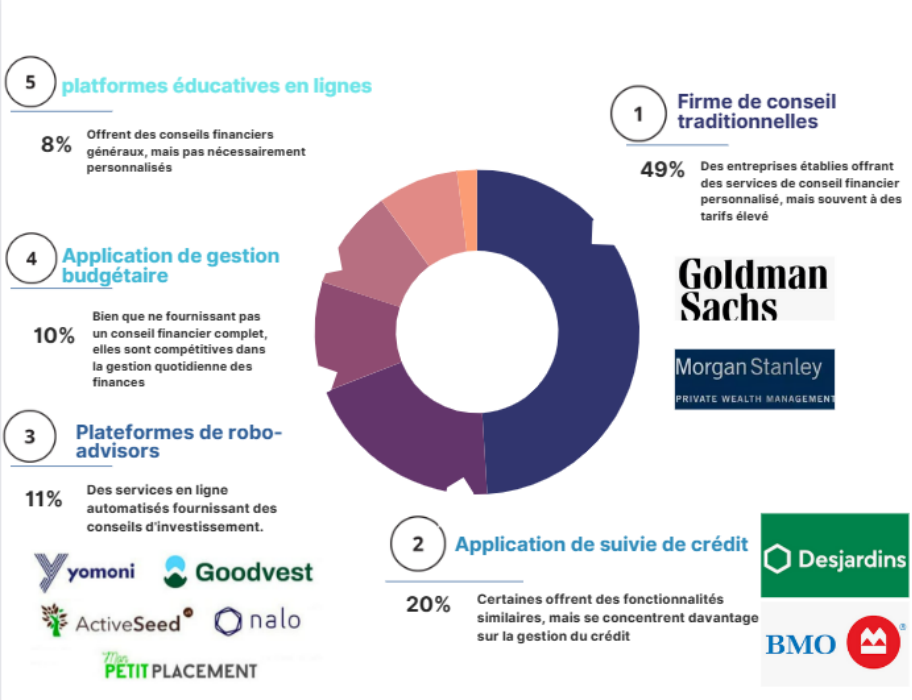


Figure 2: Répartition des concurrents

3 Description du travail effectué

Notre solution repose sur trois modèles clés, visant à aider les clients à gérer proactivement leurs finances. Ces modèles sont conçus pour offrir une vue approfondie de la situation financière des clients, les aider à surmonter les difficultés financières, et les guider dans des décisions d'investissement judicieuses.

Notre Modèle de Classification du Crédit Score, le premier volet de notre approche chez ProsperPal, se concentre sur l'évaluation de la solvabilité des clients. En utilisant des techniques de régression logistique et SVM, le modèle attribue des scores de crédit aux clients, facilitant ainsi la gestion proactive des finances. Ces scores sont cruciaux pour comprendre la capacité de remboursement des clients et orientent nos conseillers financiers dans la recommandation de produits financiers adaptés.

Le deuxième volet, notre Modèle de Clustering des Clients, basé sur l'algorithme K-Means, offre une segmentation des clients en groupes homogènes. Cette approche permet d'identifier des similitudes financières au sein de chaque groupe, offrant ainsi la possibilité de personnaliser les conseils financiers. En regroupant les clients ayant des caractéristiques financières similaires, nos conseillers sont mieux équipés pour proposer des solutions spécifiques à chaque groupe, traitant ainsi les aspects financiers particuliers de manière plus précise.

Enfin, le troisième volet consiste en notre Modèle de Prédiction de Bourses. En utilisant un modèle LSTM, il offre des recommandations d'investissement personnalisées. Cette prédiction repose sur une analyse approfondie des tendances du marché financier, guidant ainsi nos clients vers les meilleurs choix d'investissement. Ces recommandations sont intégrées dans la plateforme globale, permettant aux conseillers financiers d'offrir des conseils éclairés et personnalisés en matière d'investissement.

En somme, cette trilogie de modèles permet à Prosper Pal de fournir des conseils financiers complets et individualisés, en considérant tant la solvabilité des clients que leurs profils financiers spécifiques, pour les aider à surmonter les défis financiers et à prospérer.

4 Classification du Crédit Score

4.1 Jeu de données

Le jeu de données provient de Kaggle et comprend des informations sur plus de 100 personnes à travers le monde. Les données incluent des variables telles que l'âge, le genre, le revenu, le niveau d'éducation, le statut matrimonial, le nombre d'enfants, la propriété du logement et le score de crédit. Aucune valeur manquante n'est présente, avec 3 variables numériques et 5 variables catégoriques.

4.2 Analyse Univariable : Distribution et Tendance Centrale

L'analyse univariable offre une vision approfondie des caractéristiques clés de nos clients. Par exemple, l'âge, variant de 25 à 53 ans avec une moyenne d'environ 38 ans, est essentiel pour établir des profils financiers. Cette diversité d'âges indique une représentation significative des différentes générations parmi nos clients. De plus, les revenus, oscillant entre 25 000 et 162 500, illustrent une distribution équilibrée, permettant une analyse précise des capacités financières des clients. La plupart de nos clients n'ont pas d'enfants, mais une petite proportion ayant un nombre plus élevé suggère la nécessité d'approches financières spécifiques pour cette catégorie.

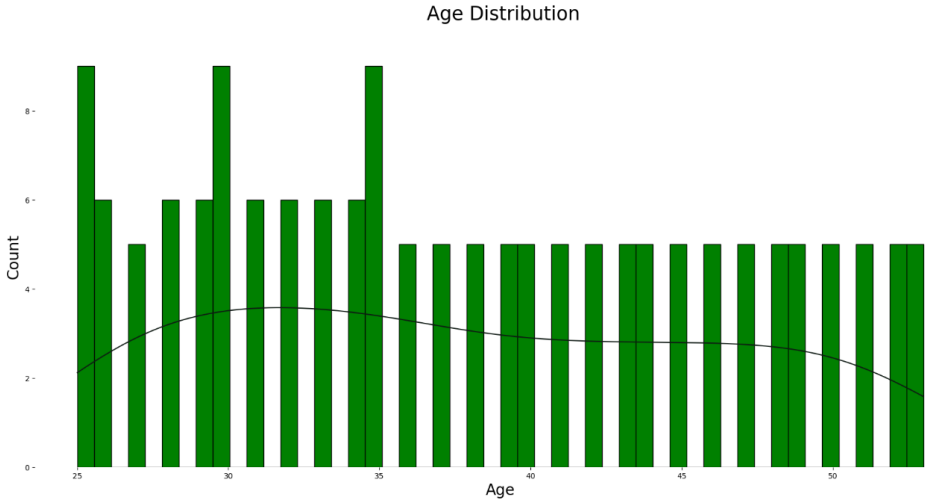


Figure 3: Distribution d'ages

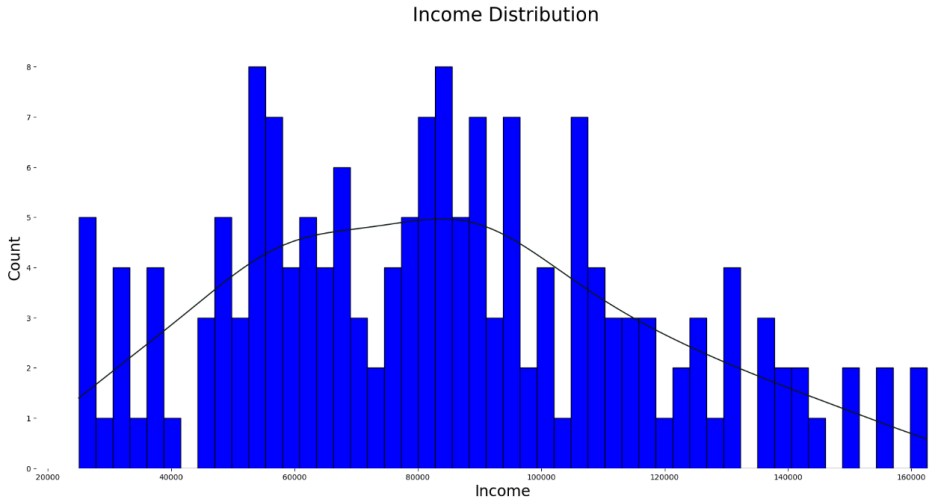


Figure 4: Distribution des salaires

4.3 Analyse Bivariable : Corrélation et Relations Cruciales

L'identification d'une corrélation positive entre l'âge et le revenu est un élément clé de notre analyse bivariable. Cette corrélation suggère une tendance significative à l'augmentation du revenu avec l'âge, ce qui informe nos conseillers financiers sur l'évolution potentielle des ressources financières de nos clients au fil du temps. De plus, l'analyse révèle des relations complexes entre le niveau d'éducation, le revenu et le score de crédit. Ces informations sont cruciales pour comprendre les fondements éducatifs qui peuvent influencer les choix financiers et le score de crédit.

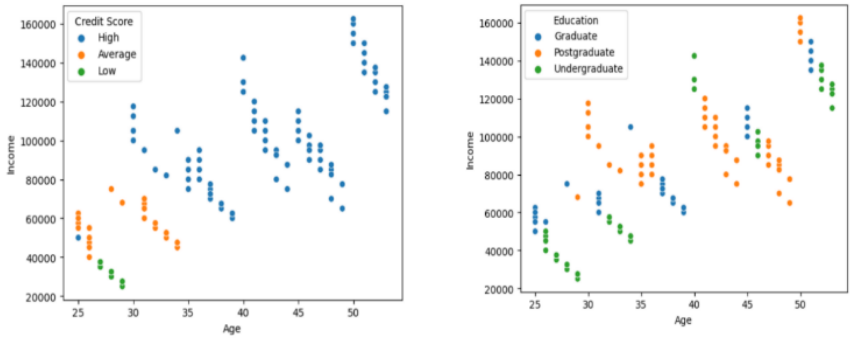


Figure 5: Analyse Bivariable

4.4 Implémentation du pipeline :

Le pipeline de traitement des données a pour objectif de préparer les caractéristiques de notre ensemble de données avant de les utiliser pour l'entraînement des modèles de classification. Les caractéristiques sont séparées en deux catégories : numériques et catégoriques. Les caractéristiques numériques, de type entier (int64), sont transformées à l'aide d'un scaler, le StandardScaler, pour standardiser les valeurs numériques. Les caractéristiques catégoriques, de type objet, sont transformées avec un encodeur onehot, le OneHotEncoder, pour les convertir en une représentation numérique. Un transformateur de colonne (ColumnTransformer) combine ces transformateurs numériques et catégoriques, appliquant les transformations spécifiques à chaque type de caractéristique. Étant donné que notre ensemble de données ne contient pas de valeurs manquantes, l'ajout d'un imputeur dans le pipeline n'est pas nécessaire. Enfin, un pipeline global est créé en associant le transformateur de colonne (preprocessor) à un classificateur donné, garantissant une préparation cohérente des données avant de les alimenter aux modèles de classification.

4.5 Entraînement et Évaluation des Modèles:

Nous avons évalué les performances de nos modèles de classification en utilisant une technique de validation croisée à 5 plis avec la fonction cross validate de scikit-learn. Cette approche nous permet d'estimer la capacité de notre modèle à généraliser sur des données non vues. Nous avons choisi la métrique de précision (accuracy) pour mesurer la justesse de nos prédictions. Les données d'entraînement ont été divisées en 5 ensembles, et le modèle a été entraîné et évalué cinq fois, chaque fois sur un ensemble de test différent. Cette méthodologie offre une évaluation robuste de la performance de notre modèle sur divers sous-ensembles de données.

Ensuite, nous avons comparé les performances de plusieurs algorithmes de classification sur notre jeu de données de scores de crédit:

- Régression Logistique: La régression logistique a démontré une précision moyenne de 96.97 avec une classification très précise des catégories de crédit.
- SVM (Support Vector Machine): Le SVM a montré une précision moyenne de 90.91, indiquant une performance solide bien que légèrement inférieure à la régression logistique.

- K-Nearest Neighbors (KNN): Le modèle KNN a fourni une précision moyenne de 90.91, montrant des résultats similaires au SVM.
- Random Forest: Le Random Forest a présenté une précision moyenne de 90.91, alignée avec les résultats du SVM et du KNN.
- Decision Tree: L'arbre de décision a démontré une performance exceptionnelle avec une précision de 100, suggérant une adaptation parfaite aux données d'entraînement.
- Gradient Boosting: Le modèle Gradient Boosting a également atteint une précision de 100, montrant une adaptation et une capacité de généralisation exceptionnelles.

	fit time mean	test score mean	F1 score	Precision	Recall	Accuracy
Logistic Regression	0.0126	0.969	0.9696	0.9696	0.9696	0.9696
SVM	0.00846	0.954	0.9090	0.9090	0.9090	0.9090
KNN 3	0.0171	0.977	0.9090	0.9090	0.9090	0.9090
Random F	0.1144	0.977	0.9090	0.9090	0.9090	0.9090
Gradient B	0.2194	0.969	1.0	1.0	1.0	1.0
Decision T	0.0115	0.969	1.0	1.0	1.0	1.0

Figure 6: Résultats

Dans l'ensemble, ces résultats mettent en évidence l'efficacité de divers algorithmes de classification pour résoudre le problème de classification des scores de crédit. La régression logistique, le SVM, le KNN, le Random Forest, l'arbre de décision et le Gradient Boosting ont tous montré des performances robustes, bien que l'arbre de décision et le Gradient Boosting se soient démarqués avec une précision parfaite. Ces résultats soulignent l'importance de choisir le modèle approprié en fonction des caractéristiques spécifiques du problème et des données disponibles.

Nos conseillers financiers chez ProsperPal peuvent tirer parti des résultats du modèle de classification du score de crédit pour offrir un service de conseil personnalisé et efficace. En utilisant ce modèle, ils évaluent la capacité de remboursement des clients en se basant sur des données telles que l'âge, le revenu, le niveau d'éducation, le statut matrimonial, le nombre d'enfants, et la propriété du logement. Cette évaluation fournit des

informations cruciales sur le profil financier de chaque client, permettant ainsi de recommander des produits financiers adaptés à leur situation spécifique.

Par exemple, si le modèle prédit un score de crédit élevé, nos conseillers peuvent suggérer des options de prêt avantageuses ou des produits d'investissement avec des conditions favorables. En revanche, si le score de crédit est plus faible, nos conseillers peuvent recommander des stratégies de remboursement de dettes ou des solutions financières plus adaptées à la situation du client.

5 Clustering des clients

L'objectif central de ce modèle de clustering est de segmenter précisément les clients en fonction de leur comportement d'utilisation des cartes de crédit. Cette segmentation offre une meilleure compréhension des habitudes de dépenses, des préférences d'achat, et des comportements financiers spécifiques à chaque groupe de clients. Ces informations sont cruciales pour nos conseillers financiers, car elles leur permettent de personnaliser leurs recommandations et de proposer des solutions adaptées aux besoins financiers spécifiques de chaque segment.

5.1 Jeu de données

Le jeu de données sélectionné, provenant de Kaggle, résume le comportement d'environ 9000 détenteurs de cartes de crédit actifs au cours des 6 derniers mois. Il comprend 18 variables comportementales, telles que le solde, la fréquence du solde, les achats, les avances de fonds, etc. Les données ont été traitées en gérant les valeurs manquantes et en appliquant une transformation logarithmique pour améliorer la distribution et réduire l'impact des valeurs aberrantes. Le dictionnaire des données pour le jeu de données sur les cartes de crédit est le suivant : *CUST ID*, *BALANCE*, *BALANCE FREQUENCY*, *PURCHASES*, *ONEOFF PURCHASES*, *INSTALLMENTS PURCHASES*, *CASH ADVANCE*, *PURCHASES FREQUENCY*, *ONEOFF PURCHASES FREQUENCY*, *PURCHASES INSTALLMENTS FREQUENCY*, *CASH ADVANCE FREQUENCY*, *CASH ADVANCE TRX*, *PURCHASES TRX*, *CREDIT LIMIT*, *PAYMENTS*, *MINIMUM PAYMENTS*, *PRC FULL PAYMENT*, *TENURE*. Pour des descriptions plus détaillées de ces variables, veuillez vous référer à l'annexe.

# Age	Gender	# Income	Education	Marital Sta...	# Number of...
25	Female	50000	Bachelor's Degree	Single	0
30	Male	100000	Master's Degree	Married	2
35	Female	75000	Doctorate	Married	1
40	Male	125000	High School Diploma	Single	0
45	Female	100000	Bachelor's Degree	Married	3
50	Male	150000	Master's Degree	Married	0

Figure 7: Aperçu de la *dataframe*

5.2 Gestion des valeurs manquantes

Initialement, une vérification a révélé que l'ensemble de données comportait au total 313 valeurs manquantes, dont 312 étaient présentes dans la colonne 'MINIMUM PAYMENTS' et une seule dans la colonne *'CREDIT LIMIT'*. Pour remédier à cela, la stratégie suivante a été mise en œuvre : la ligne contenant la valeur manquante dans la colonne *'CREDIT LIMIT'* a été supprimée, car il s'agissait d'une seule observation. En ce qui concerne la colonne *'MINIMUM PAYMENTS'*, les valeurs manquantes ont été remplacées par la médiane de cette colonne. Cette approche a permis de conserver l'intégrité des données tout en traitant efficacement les valeurs manquantes, fournissant ainsi un ensemble de données prêt pour une analyse plus approfondie.

5.3 Gestion des variables catégoriques

La seule variable catégorique qui se présente dans notre jeu de donnée est la *CUST ID* .Nous avons décidé de supprimer cette colonne vu qu'elle ne porte pas d'informations pertinentes pour le processus de clustering. elle est utilisée pour identifier de manière unique chaque Client, mais ils ne contribuent pas à la similarité ou à la distance entre les observations.

5.4 Distribution et valeurs aberrantes

Afin d'améliorer la distribution des données et de réduire l'impact des valeurs aberrantes, nous avons appliqué une transformation logarithmique sur l'ensemble de jeu de données. Cette transformation facilite à notre modèle l'interprétation des résultats du *clustering* en rendant les distributions plus symétriques.

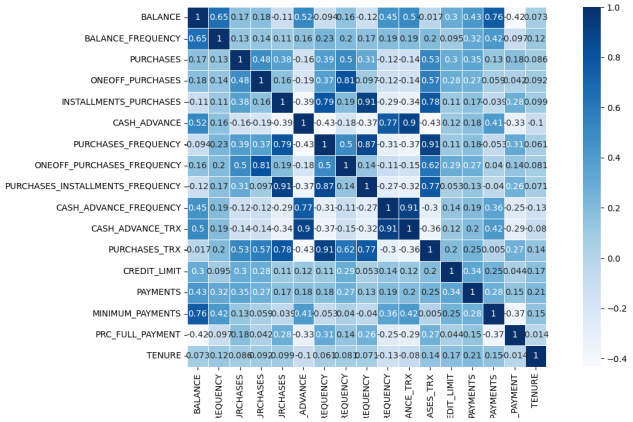


Figure 8: Matrice de corrélation

5.5 Analyse bivariable

L'analyse de la matrice de corrélation révèle des relations intéressantes entre les variables du jeu de données.

La corrélation la plus forte, de 0.92, entre *PURCHASES* et *ONEOFF PURCHASES*, suggère une relation presque parfaite entre le montant total des achats effectués et le montant maximum d'achat réalisé en une seule fois. Cela indique que les clients qui effectuent des achats plus importants ont tendance à réaliser des achats significatifs en une seule fois.

D'autre part, la corrélation négative faible entre *BALANCE* et *PRC FULL PAYMENT* indique une tendance où les clients ayant un solde plus élevé sont moins enclins à effectuer des paiements intégraux. Cette relation pourrait être attribuée au fait que les clients qui maintiennent un solde plus élevé préfèrent peut-être payer partiellement plutôt que de régler intégralement leur solde.

La corrélation de -0.16 entre *BALANCE FREQUENCY* et *PRC FULL PAYMENT* indique une faible relation négative entre la fréquence de mise à jour du solde et le pourcentage de paiement intégral. Cela suggère que les clients qui mettent fréquemment à jour leur solde ont légèrement moins tendance à effectuer des paiements intégraux.

L'analyse de ces relations met en lumière des tendances générales, mais il est important de souligner que ces corrélations ne capturent pas nécessairement toute la complexité des comportements financiers des clients. Les relations entre ces variables sont multifactorielles, et d'autres facteurs non pris en compte dans l'analyse de corrélation peuvent influencer les comportements financiers individuels.

C'est précisément pour cette raison que l'utilisation d'algorithmes de

clustering devient pertinente. Les algorithmes de clustering permettent d'identifier des caractéristiques similaires au sein du jeu de données, regroupant les clients ayant des comportements financiers plus similaires. En utilisant ces techniques de regroupement, nous pourrions obtenir des segments de clients plus homogènes en termes de comportements financiers, offrant ainsi une compréhension plus approfondie et nuancée des différents profils de clients.

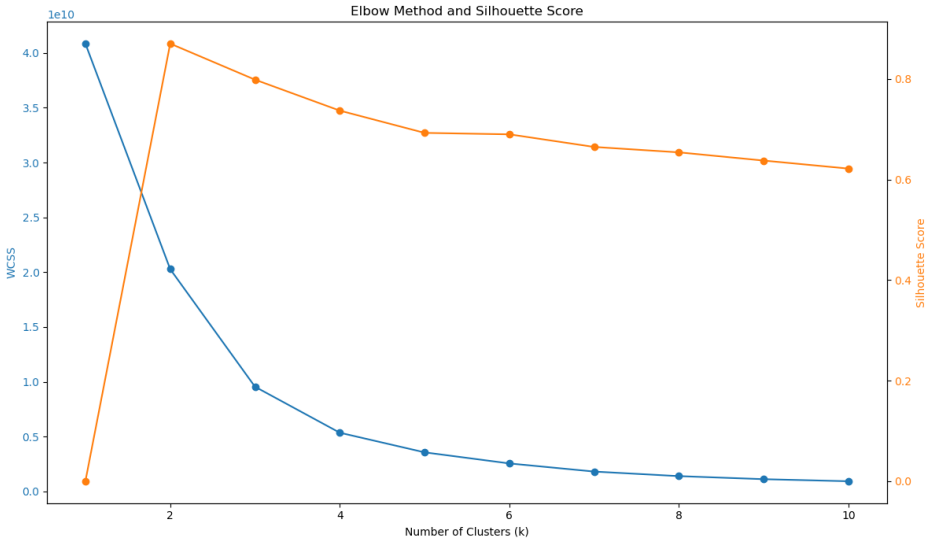
6 Implémentation de l'algorithmes de *Clustering*

6.1 Réduction de dimensionnalité PCA

La première étape de cette démarche consiste à appliquer une réduction de dimensionnalité sur la dataset à l'aide de l'analyse en composantes principales (PCA). Cela est particulièrement bénéfique dans le contexte du clustering, où la recherche de patterns significatifs est plus efficace dans un espace de dimension réduite.

6.2 Nombre optimal de clusters (k)

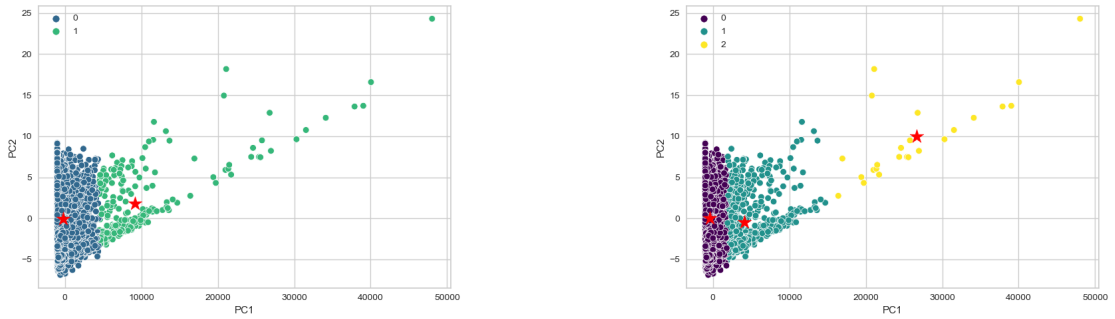
Ensuite, la démarche se poursuit avec la détermination du nombre optimal de *clusters* (k), en utilisant la méthode du coude (*Elbow Method*) et l'analyse du *silhouette score*. Ces deux métriques sont des outils cruciaux pour évaluer la cohérence et la compacité des *clusters*. La méthode du coude vise à identifier le point où l'ajout de clusters supplémentaires n'apporte plus une réduction significative de la variance intra-cluster (*WCSS*), tandis que le silhouette score mesure à quel point chaque point de données est similaire à son propre cluster par rapport aux clusters voisins.



les valeurs de k=2 et k=3 sont deux options qui présentent des scores de silhouette élevés et des *WCSS* relativement bas, offrant une indication que la structure des clusters pourrait être bien capturée avec l'un ou l'autre de ces deux choix de k. Nous allons donc établir une analyse plus approfondie afin de déterminer la meilleure option en fonction de notre contexte spécifique .

6.3 Comparaison des résultats avec différents valeur de k

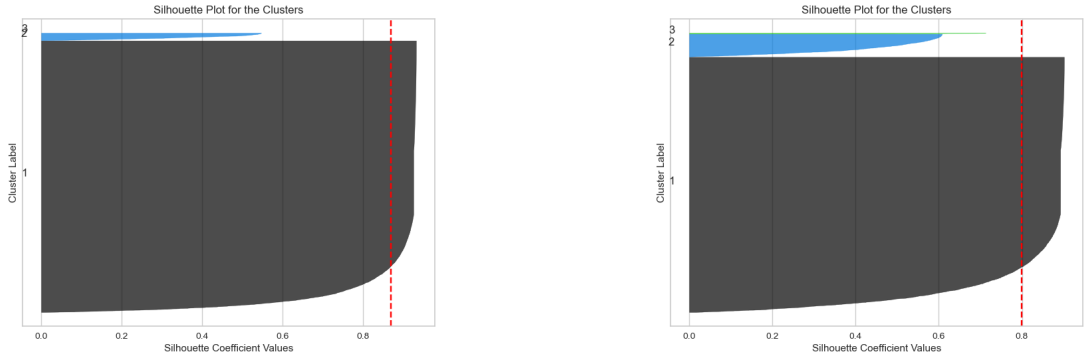
2D clustering:



(a) k=2 (b) k=3

Figure 9: Comparison of 2D clustering

Les graphiques 2D pour les clusters k=2 et k=3 permettent de visualiser la répartition des points de données dans l'espace réduit par PCA. Visuellement k=3 permet une meilleur séparation Silouhette Score:



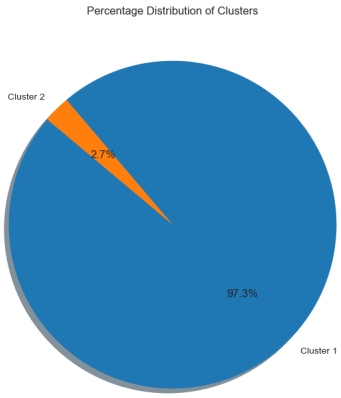
(a) k=2 (b) k=3

Figure 10: Comparison of Silouhette Score

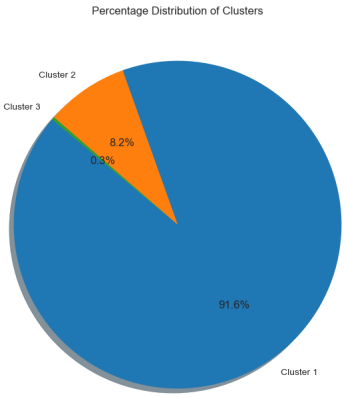
Les scores de silhouette mesurent la cohésion et la séparation des *clusters*. Un score de silhouette élevé indique que les points d'un cluster sont

similaires entre eux et différents des points des autres clusters.pour k=2 le score de silhouette est légèrement supérieur à celui obtenu pour k=3.

Clustering percentage:



(a) k=2



(b) k=3

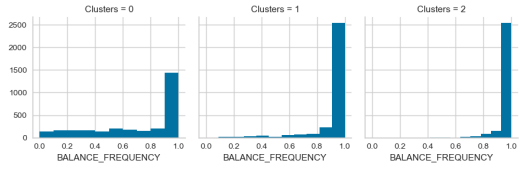
Figure 11: Comparison of Clustering percentage

En examinant la répartition des clients dans les clusters, pour k=2, le premier cluster représente 97.3 des clients, tandis que le deuxième cluster en représente seulement 2.7. Pour k=3, le premier cluster englobe la majorité des clients avec 91.6, suivi du deuxième cluster avec 8.2, et enfin, le troisième cluster représente une proportion marginale de 0.3.

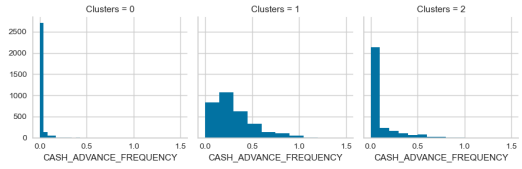
Cette observation suggère que le modèle avec k=2 offre une segmentation plus concentrée, avec un cluster dominant et un autre moins fréquent, tandis que le modèle avec k=3 introduit une complexité supplémentaire avec un troisième cluster.

En fonction de ces résultats, le modèle avec k=3 semble mieux représenter la structure sous-jacente des données, offrant une segmentation plus nette et concise de la clientèle.

6.4 Caractéristiques des Clusters

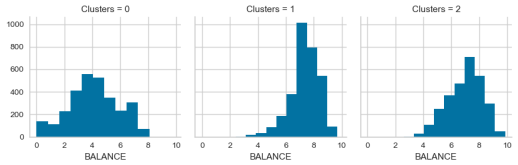


(a) BALANCE FREQUENCY



(b) CASH ADVANCE FREQUENCY

(a) MINIMUM PAYMENTS

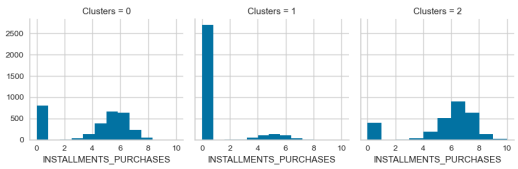


(b) BALANCE

(a) CASH ADVANCE

(b) CASH ADVANCE TRX

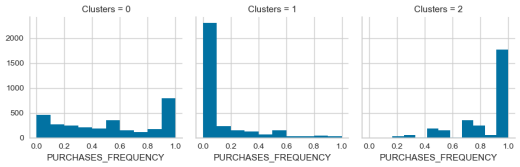
(a) CREDIT LIMIT



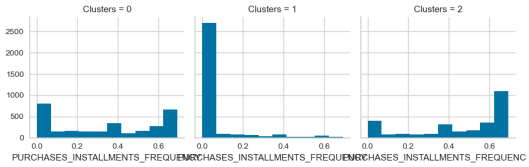
(b) INSTALLMENTS PURCHASES

(a) ONEOFF PURCHASES FREQUENCY

(b) PAYMENTS



(a) PF



(b) PIF

6.5 Cluster 1:

Cash Advance Frequency : Les clients de ce cluster ont une fréquence d'utilisation faible des avances de trésorerie, indiquant une utilisation peu fréquente de ce service.

Balance: Les clients de ce cluster ont tendance à avoir un solde faible sur leur compte. **Cash Advance** : Le montant des avances de trésorerie pris par les clients de ce cluster est relativement élevé.

Installments Purchases : Les clients de ce cluster effectuent des achats en versements avec une fréquence faible.

Purchase Frequency : La fréquence globale d'achats pour ce cluster est élevée.

Cash Advance TRX : Les clients de ce cluster ont un nombre élevé de transactions d'avance de trésorerie.

6.6 Cluster 2:

Cash Advance Frequency : Les clients de ce cluster utilisent fréquemment les services d'avance de trésorerie.

Balance : Les clients de ce cluster maintiennent un solde élevé sur leur compte.

Cash Advance : Le montant des avances de trésorerie pris par les clients de ce cluster est faible.

Installments Purchases : Les clients de ce cluster effectuent des achats en versements avec une fréquence élevée.

Purchase Frequency : La fréquence globale d'achats pour ce cluster est élevée.

Cash Advance TRX : Les clients de ce cluster ont un nombre faible de transactions d'avance de trésorerie.

6.7 Cluster 3:

Cash Advance Frequency : Les clients de ce cluster ont une fréquence d'utilisation faible des avances de trésorerie.

Balance : Les clients de ce cluster ont tendance à avoir un solde faible sur leur compte.

Cash Advance : Le montant des avances de trésorerie pris par les clients de ce cluster est faible.

Installments Purchases : Les clients de ce cluster effectuent des achats en versements avec une fréquence faible.

Purchase Frequency : La fréquence globale d'achats pour ce cluster est non élevée.

Cash Advance TRX : Les clients de ce cluster ont un nombre élevé de transactions d'avance de trésorerie.

6.8 Résultats:

- Cluster 1 : Ce cluster représente des clients qui utilisent peu fréquemment les avances de trésorerie, ont un solde faible, mais effectuent des achats relativement fréquents. Recommandations : Les conseillers pourraient conseiller ces clients sur la gestion prudente de leurs avances de trésorerie, en les encourageant à maintenir un solde plus élevé

pour réduire les frais liés aux avances. Ils pourraient également suggérer des plans d'épargne ou des produits financiers adaptés à leurs fréquentes habitudes d'achat.

- Cluster 2 : Ce cluster regroupe des clients qui utilisent fréquemment les avances de trésorerie, ont un solde élevé, et effectuent des achats réguliers, souvent en versements. Recommandations : Pour ce cluster, les conseillers pourraient discuter des opportunités d'investissement en utilisant le solde élevé, proposer des solutions de crédit plus avantageuses et suggérer des options d'achat en versements pour optimiser leur utilisation fréquente.
- Cluster 3 : Ce cluster concerne des clients avec une faible utilisation des avances de trésorerie, un solde faible, et des achats peu fréquents. Recommandations : Les conseillers peuvent orienter ces clients vers des options d'économies, discuter des avantages d'un solde plus élevé, et proposer des plans de crédit adaptés à leurs habitudes d'achat moins fréquentes.

Bien que KMeans ait permis d'obtenir les résultats de clustering souhaités, nous avons également exploré d'autres algorithmes tels que DBSCAN et OPTICS pour évaluer leur performance. Cependant, ces méthodes n'ont pas produit des résultats aussi satisfaisants que KMeans dans le contexte spécifique de notre ensemble de données sur les cartes de crédit.

DBSCAN et OPTICS sont des algorithmes de clustering basés sur la densité qui peuvent être plus adaptés à des distributions de données spécifiques. Dans notre cas, les caractéristiques de l'ensemble de données et la structure des clusters semblaient mieux correspondre au modèle de KMeans, qui divise les données en clusters de forme sphérique.

7 Modèle LSTM pour la Prédiction des Stocks:

Le modèle LSTM (Long Short-Term Memory) représente une approche avancée en matière de prévision des valeurs boursières. Son objectif principal est de prédire avec précision les fluctuations futures des cours des actions. Cette capacité prédictive offre une opportunité majeure pour nos conseillers financiers d'apporter une valeur ajoutée significative à nos clients.

7.1 Jeu de donnée:

Nous avons utilisé l'API Yahoo Finance via la bibliothèque pandas pour récupérer les données boursières. Ces données comprennent les prix de

clôture des actions pour un symbole boursier spécifique sur une période donnée. Elles offrent une vision détaillée des fluctuations quotidiennes des cours boursiers, fournissant ainsi un aperçu précieux pour évaluer les tendances et les mouvements du marché. En utilisant ces informations, nos conseillers financiers peuvent formuler des recommandations éclairées, permettant aux clients de prendre des décisions d'investissement informées et stratégiques.

7.2 Prétraitement des données :

Après avoir récupéré les données, une étape cruciale de prétraitement a été entreprise pour garantir la qualité et la pertinence des informations utilisées dans notre modèle. Dans cette phase, nous avons extrait spécifiquement les prix de clôture ('Close') des actions. Ce choix est fondamental, car les prix de clôture fournissent une indication directe de la valeur finale d'un actif à la fin d'une séance de négociation.

Par la suite, nous avons normalisé ces données en utilisant le MinMaxScaler. Cette étape est essentielle pour mettre toutes les valeurs à l'échelle entre 0 et 1, assurant ainsi une comparaison et une interprétation cohérentes, indépendamment de l'ampleur initiale des prix. Cette normalisation facilite également la convergence du modèle pendant l'entraînement, contribuant à des prédictions plus précises.

Enfin, nous avons segmenté nos données en ensembles distincts d'entraînement et de test. Cette division permet d'évaluer la performance du modèle sur des données non vues, garantissant sa capacité à généraliser au-delà des exemples sur lesquels il a été initialement formé.

7.3 Architecture du Modèle LSTM:

Le choix d'utiliser un modèle LSTM (Long Short-Term Memory) pour résoudre le problème de prédiction des prix des actions est justifié par la nature séquentielle et temporelle des données financières. Les LSTM sont des réseaux de neurones récurrents (RNN) dotés d'une capacité inhérente à traiter des séquences de données, ce qui les rend adaptés à la modélisation de séries temporelles, comme les variations des prix des actions au fil du temps.

L'architecture du modèle LSTM que nous avons mise en place se compose de plusieurs couches, chaque couche jouant un rôle spécifique dans la capture des motifs complexes des données temporelles:

- Couche LSTM (128 unités) - Return Sequences=True : Cette première couche LSTM, avec 128 unités, est configurée pour retourner les séquences complètes. Elle permet au modèle de conserver des in-

formations sur chaque pas de temps, capturant ainsi les relations à long terme dans les données séquentielles.

- Couche LSTM (64 unités) - Return Sequences=False : La deuxième couche LSTM, avec 64 unités, est configurée pour ne pas retourner les séquences complètes, mais plutôt fournir une représentation consolidée des informations apprises par la première couche. Cette couche contribue à réduire la complexité tout en préservant les caractéristiques importantes.
- Couche Dense (25 unités) : Une couche dense de 25 unités suit les couches LSTM. Cette couche introduit une non-linéarité dans le modèle et extrait des caractéristiques haut niveau des données.
- Couche Dense (1 unité) : La dernière couche dense, avec une unité, produit la sortie finale du modèle, représentant la prédiction du prix de clôture.

7.4 Résultats:

Le RMSE de 4.98 dollars suggère que, en moyenne, les prédictions du modèle LSTM diffèrent d'environ 4.98 dollars par rapport aux valeurs réelles. En effet, dans le monde financier, une volatilité quotidienne de quelques dollars peut être considérée comme normale, en particulier pour des actions dont les prix sont relativement élevés. Surtout si l'objectif est de capturer les tendances générales du marché. Si le modèle parvient à reproduire les tendances de manière satisfaisante, un RMSE de cette ampleur peut être toléré.

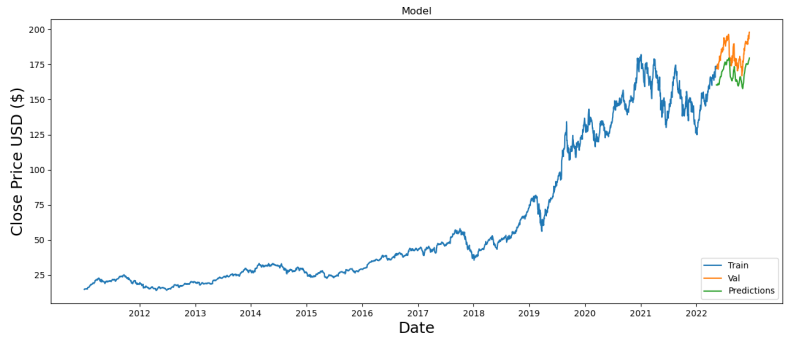


Figure 18: Résultats

La courbe de prédiction affiche graphiquement les résultats du modèle par rapport aux données réelles. La conformité de la courbe de prédiction à la courbe réelle indique à quel point les prédictions du modèle se rapprochent des véritables valeurs. L'ajustement serré entre la courbe de prédiction et la courbe réelle suggère une bonne performance du modèle.

8 Synthèse des résultats:

8.1 Modèle de Classification du Crédit Score :

Dans l'ensemble, les résultats démontrent l'efficacité du modèle de classification du crédit score. Les algorithmes tels que la régression logistique, le SVM, le KNN, le Random Forest, l'arbre de décision et le Gradient Boosting ont tous montré des performances solides. Notamment, l'arbre de décision et le Gradient Boosting ont atteint une précision parfaite, soulignant l'importance de choisir le modèle en fonction des caractéristiques spécifiques du problème. Ces résultats permettent d'atteindre notre objectif en fournissant une évaluation précise de la solvabilité des clients, orientant nos conseillers financiers dans la recommandation de produits adaptés.

8.2 Modèle de Clustering des Clients :

Le modèle de clustering a permis de segmenter les clients en groupes homogènes, facilitant une personnalisation plus précise des conseils financiers. Chaque cluster offre des recommandations spécifiques en fonction des comportements financiers observés. Par exemple, pour les clients du Cluster 1, des conseils sur la gestion prudente des avances de trésorerie et des suggestions d'épargne peuvent être avancés. Ce modèle contribue à atteindre notre objectif en offrant des solutions spécifiques à chaque groupe, traitant ainsi les aspects financiers particuliers de manière plus précise.

8.3 Modèle de Prédiction de Bourses :

Le modèle LSTM pour la prédiction des bourses a montré une performance satisfaisante avec un RMSE de 4.98 dollars. Bien que des différences subsistent entre les prédictions et les valeurs réelles, ces résultats sont tolérables compte tenu de la volatilité normale du marché financier. La courbe de prédiction illustre l'ajustement serré entre les prédictions du modèle et les données réelles, indiquant une bonne performance globale. Ces résultats permettent d'atteindre notre objectif en fournissant

des recommandations d'investissement personnalisées, guidant nos clients vers des choix éclairés.

En somme, la trilogie de modèles de ProsperPal permet d'offrir des conseils financiers complets et individualisés, contribuant à aider nos clients à surmonter les défis financiers et à prospérer. L'utilisation de ces modèles dans notre approche globale renforce notre capacité à fournir un service de conseil de haute qualité, aligné sur les besoins spécifiques de chaque client.

9 Identification des risques, inconvénients et avantages du projet :

9.1 Risques potentiels:

- Risques technologiques : Les modèles d'apprentissage automatique peuvent être sensibles aux données sur lesquelles ils sont formés. Des erreurs dans les algorithmes pourraient entraîner des recommandations inappropriées.
- Sécurité des données : Les informations financières des clients sont sensibles. Tout incident de sécurité pourrait entraîner une perte de confiance et des conséquences légales.
- Réglementation financière : Les changements dans les réglementations financières au Québec pourraient avoir un impact sur la manière dont nos services sont fournis et réglementés.
- Concurrence accrue : L'entrée de nouveaux concurrents sur le marché du conseil financier peut augmenter la pression sur les prix et la nécessité d'innover constamment.

9.2 Inconvénients potentiels :

- Dépendance technologique : En cas de défaillance technique, les clients pourraient être temporairement privés d'accès à leurs conseils financiers.
- Coût initial de développement : La création des modèles d'apprentissage automatique et des outils nécessite des investissements significatifs avant de générer des revenus..
- Besoin de données précises : Les modèles d'apprentissage automatique dépendent de données précises. Des erreurs dans les données d'entrée peuvent affecter la qualité des conseils fournis.

9.3 Avantages potentiels :

- Personnalisation : La capacité à fournir des conseils financiers personnalisés basés sur des modèles d'apprentissage automatique peut attirer un large éventail de clients.
- Évolutivité : Les solutions basées sur l'IA peuvent évoluer et s'adapter rapidement aux changements du marché et aux besoins des clients.
- Efficacité opérationnelle : L'automatisation des processus peut améliorer l'efficacité opérationnelle, permettant de servir un plus grand nombre de clients avec une équipe plus petite.
- Coûts compétitifs : En automatisant une partie du processus de conseil financier, l'entreprise peut offrir des services à des coûts compétitifs par rapport aux conseillers traditionnels.

9.4 Stratégies d'atténuation des risques :

- Sécurité des données : Mettre en place des mesures de sécurité robustes, y compris le cryptage des données et les meilleures pratiques de gestion des accès.
- Suivi réglementaire : Maintenir une veille constante sur les changements réglementaires et ajuster les pratiques commerciales en conséquence.
- Tests rigoureux des modèles : Effectuer des tests approfondis des modèles d'apprentissage automatique pour s'assurer de leur précision et de leur fiabilité.
- Communication proactive : Éduquer les clients sur les avantages de l'automatisation tout en soulignant la présence de conseillers humains pour renforcer la confiance..

10 Conclusion

En conclusion, ProsperPal se distingue en tant que guide essentiel pour accompagner les Canadiens en difficulté financière vers l'atteinte de la richesse. Grâce à notre équipe dévouée de conseillers financiers experts et à l'utilisation de modèles d'apprentissage automatique, nous offrons une assistance proactive dans divers aspects financiers.

Nos services englobent une compréhension approfondie de la relation de nos clients avec l'argent, des analyses approfondies de leur état financier actuel, des plans de remboursement de dettes personnalisés, et des propositions avisées pour des opportunités d'investissement.

Ce rapport détaille les différentes composantes de ProsperPal, mettant en lumière comment nous appliquons les modèles d'apprentissage automatique pour comprendre les attitudes financières individuelles, réaliser des analyses approfondies des situations financières personnelles, et regrouper les clients en segments distincts en fonction de leurs caractéristiques financières. Enfin, nous montrons comment nous aidons nos clients à prendre des décisions d'investissement éclairées grâce à nos modèles de prédiction des marchés boursiers.

Bien que le projet présente des risques et des inconvénients, les avantages potentiels en termes de personnalisation, d'efficacité opérationnelle et de coûts compétitifs peuvent positionner l'entreprise comme une force innovante sur le marché du conseil financier au Canada. La gestion proactive des risques et l'adaptation aux évolutions du marché seront essentielles pour assurer le succès continu du projet.

11 Annexe

- CUST ID : Identification du titulaire de la carte de crédit (Catégorique)
- BALANCE : Montant du solde restant sur le compte pour effectuer des achats
- BALANCE FREQUENCY : Fréquence de mise à jour du solde, score entre 0 et 1
- PURCHASES : Montant des achats effectués depuis le compte
- ONEOFF PURCHASES : Montant maximum d'achat effectué en une seule fois
- INSTALLMENTS PURCHASES : Montant des achats effectués en versements
- CASH ADVANCE : Avance de trésorerie accordée par l'utilisateur
- PURCHASES FREQUENCY : Fréquence des achats, score entre 0 et 1
- ONEOFF PURCHASES FREQUENCY : Fréquence des achats
- PURCHASES INSTALLMENTS FREQUENCY : Fréquence des achats en versements
- CASH ADVANCE FREQUENCY : Fréquence du paiement de l'avance de trésorerie
- CASH ADVANCE TRX : Nombre de transactions effectuées avec "Cash in Advanced"
- PURCHASES TRX : Nombre de transactions d'achat effectuées
- CREDIT LIMIT : Limite de la carte de crédit pour l'utilisateur
- PAYMENTS : Montant du paiement effectué par l'utilisateur
- MINIMUM PAYMENTS : Montant minimum des paiements effectués par l'utilisateur
- PRC FULL PAYMENT : Pourcentage du paiement intégral effectué par l'utilisateur
- TENURE : Durée du service de la carte de crédit pour l'utilisateur