TABLES DES MATIERES

I. Introduction

1.1 Sujet du challenge Kaggle

Le challenge Kaggle "Binary Classification with a Tabular Credit Card Fraud Dataset" est une compétition qui nous propose de développer des modèles d'apprentissage pour détecter les fraudes liées aux transactions par carte de crédit. Cette compétition est basée sur un ensemble de données contenant des informations sur les transactions effectuées par des utilisateurs de cartes de crédit.

La fraude liée aux cartes de crédit est un problème majeur dans le secteur financier, causant des pertes financières considérables pour les institutions financières et les consommateurs. Les fraudeurs utilisent des techniques sophistiquées pour contourner les systèmes de détection existants. Pour lutter contre ces fraudes on peut tester l'utilisation de modèles d'apprentissage pour les détecter.

1.2 Objectif du mémoire

L'objectif de ce mémoire est de participer à ce challenge Kaggle et de développer un modèle performant pour détecter les fraudes liées aux transactions par carte de crédit. Nous cherchons à explorer différentes techniques d'apprentissage automatique, à les comparer et à les évaluer avec différents indicateurs.

En développant ce modèle, nous cherchons également à identifier les caractéristiques les plus importantes pour la détection des fraudes. Nous analyserons les données fournies et utiliserons des techniques d'analyse exploratoire pour comprendre les schémas et les tendances des transactions frauduleuses. Cette compréhension nous permettra d'améliorer notre modèle de détection.

Enfin, nous espérons que les résultats de notre travail contribueront à l'amélioration des systèmes de détection de fraudes par carte de crédit, en fournissant des informations précieuses sur les techniques et les modèles d'apprentissage automatique les plus efficaces dans ce domaine.

À travers ce mémoire, nous mettrons en évidence les défis et les opportunités liés à la détection des fraudes par carte de crédit, ainsi que les avancées réalisées grâce à l'utilisation de modèles d'apprentissage automatique.

II. Présentation du problème

2.1 Description du jeu de données

Le jeu de données utilisé dans le challenge Kaggle "Binary Classification with a Tabular Credit Card Fraud Dataset" est un ensemble de données qui contient des informations sur les transactions effectuées par des utilisateurs de cartes de crédit. Chaque transaction est décrite par un ensemble de variables telles que le montant de la transaction, la date et l'heure, le pays, ainsi que des caractéristiques anonymisées des utilisateurs.

Le jeu de données est étiqueté, ce qui signifie qu'il indique si chaque transaction est frauduleuse (classe positive) ou légitime (classe négative). Cependant, il est important de noter que le jeu de données est déséquilibré, avec un nombre beaucoup plus élevé de transactions légitimes que de transactions frauduleuses. Cela ajoute une difficulté supplémentaire à la classification, car les modèles doivent être capables de détecter efficacement les rares cas de fraudes parmi les transactions normales.

2.2 Enjeux et difficultés de la classification binaire des fraudes par carte de crédit

La classification des fraudes par carte de crédit présente plusieurs enjeux et difficultés. Tout d'abord, les fraudeurs utilisent des techniques sophistiquées pour masquer leurs activités frauduleuses, ce qui rend la détection plus complexe. Ils peuvent utiliser des cartes de crédit volées, effectuer des transactions à partir de différents pays ou modifier les caractéristiques des transactions pour les rendre similaires à celles des transactions légitimes.

De plus, le déséquilibre du jeu de données pose un défi majeur. Les modèles de classification traditionnels ont tendance à être biaisés en faveur de la classe majoritaire, ce qui peut entraîner une mauvaise détection des fraudes. Il est donc crucial de mettre en place des techniques spécifiques pour gérer ce déséquilibre et améliorer la performance de la détection des fraudes.

Un autre défi est la nécessité de développer des modèles qui sont à la fois précis et rapides. Étant donné le grand volume de transactions par carte de crédit traitées quotidiennement, il est essentiel d'avoir des modèles qui peuvent analyser rapidement les transactions en temps réel et prendre des décisions précises sur leur légitimité.

2.3 État de l'art des méthodes de classification utilisées dans le domaine

Dans le domaine de la détection des fraudes par carte de crédit, plusieurs

méthodes de classification ont été utilisées avec succès. Les approches traditionnelles incluent les arbres de décision, les réseaux de neurones, les méthodes basées sur les règles et les méthodes d'ensemble telles que le Random Forest et le Gradient Boosting.

Cependant, ces méthodes traditionnelles peuvent présenter des limitations en termes de performance de détection des fraudes, en particulier en raison du déséquilibre des données. C'est pourquoi des approches plus avancées ont été développées, telles que l'apprentissage en ligne, les modèles de détection d'anomalies, et les méthodes basées sur l'apprentissage profond.

L'apprentissage en ligne permet de mettre à jour les modèles en temps réel à mesure que de nouvelles données arrivent, ce qui est particulièrement utile pour la détection des fraudes en temps réel. Les modèles de détection d'anomalies se concentrent sur la détection des schémas de transactions inhabituels, ce qui peut être efficace pour détecter les fraudes. Enfin, les méthodes basées sur l'apprentissage profond, telles que les réseaux de neurones profonds, ont montré des performances prometteuses dans la détection des fraudes par carte de crédit.

En conclusion, la classification binaire des fraudes par carte de crédit est un problème complexe qui nécessite des approches spécifiques pour gérer le déséquilibre des données et améliorer la performance de détection. Des méthodes traditionnelles aux méthodes avancées basées sur l'apprentissage profond, diverses techniques ont été développées pour résoudre ce problème. Par la suite, nous présenterons notre approche pour résoudre ce défi et améliorer la détection des fraudes par carte de crédit.

III. Prétraitement des Données

3.1 Collecte, exploration des données et nettoyage des données

La première étape du prétraitement des données consiste à récupérer les données nécessaires pour résoudre le problème de classification des fraudes par carte de crédit Ces données contiennent des informations sur les transactions effectuées par des utilisateurs de cartes de crédit.

Une fois les données collectées, il est important de les explorer pour comprendre leur structure et identifier les potentielles anomalies ou incohérences. Cette exploration va inclure des analyses statistiques des variables, des visualisations graphiques pour comprendre les distributions et les relations entre les variables, ainsi que des vérifications de la qualité des données pour identifier les erreurs ou les valeurs aberrantes.

Le nettoyage des données est une étape essentielle pour garantir la qualité des données avant de les utiliser dans le processus de classification. Cette étape consiste à traiter les valeurs incorrectes, manquantes ou aberrantes afin de rendre les données cohérentes et exploitables.

Dans le cas de la détection des fraudes par carte de crédit, le nettoyage des données peut impliquer la vérification et la correction des erreurs de saisie, la suppression des doublons, ainsi que la suppression ou la correction des valeurs aberrantes qui pourraient fausser les résultats de la classification.

3.2 Gestion des valeurs manquantes

Il est courant d'avoir des valeurs manquantes dans les données réelles. Pour gérer ces valeurs manquantes, différentes stratégies peuvent être utilisées. Dans le cas de la classification des fraudes par carte de crédit, certaines variables peuvent avoir des valeurs manquantes qui peuvent être critiques pour la détection des fraudes. Il est donc important de prendre en compte ces valeurs manquantes de manière appropriée.

Dans notre approche, on a utilisé l'imputation par la moyenne pour remplacer les valeurs manquantes des variables continues. Cette méthode consiste à calculer la moyenne des valeurs non manquantes de la variable et à remplacer les valeurs manquantes par cette moyenne. Cette approche permet de préserver la distribution globale de la variable tout en évitant les biais potentiels.

Pour les variables catégorielles, on a utilisé une autre méthode d'imputation, en remplaçant les valeurs manquantes par la valeur la plus fréquente de la variable. Cette approche est souvent utilisée pour préserver les tendances et les relations entre les variables catégorielles.

3.3 Normalisation et transformation des variables

La normalisation et la transformation des variables sont des étapes importantes pour garantir que les données sont comparables et appropriées pour l'analyse et la classification. Dans le cas de la détection des fraudes par carte de crédit, il est courant d'appliquer des techniques de normalisation et de transformation pour rendre les variables comparables et réduire les biais potentiels.

Dans notre approche, on a utilisé la normalisation Min-Max pour mettre les variables continues dans une plage spécifique, généralement entre 0 et 1. Cette technique permet de préserver les relations entre les valeurs tout en les rendant comparables.

De plus, on a également appliqué une transformation logarithmique sur certaines

variables pour réduire les écarts importants entre les valeurs. Cette transformation peut aider à réduire l'impact des valeurs extrêmes et à rendre les données plus linéaires, ce qui peut améliorer la performance des modèles de classification.

3.4 Séparation des données en ensembles d'entraînement et de test

Une étape cruciale dans le prétraitement des données est la séparation des données en ensembles d'entraînement et de test. Cette étape permet d'évaluer la performance du modèle sur des données non vues auparavant et de s'assurer qu'il généralise bien au-delà de l'ensemble d'entraînement.

Dans notre cas, on a utilisé une méthode de séparation aléatoire pour diviser les données en ensembles d'entraînement et de test. Cette approche garantit que les données sont mélangées de manière aléatoire pour éviter tout biais potentiel. On a utilisé une proportion spécifique, par exemple 80% pour l'ensemble d'entraînement et 20% pour l'ensemble de test, mais cela peut varier en fonction des besoins spécifiques.

La séparation des données en ensembles d'entraînement et de test est essentielle pour évaluer la performance du modèle de classification sur des données réelles et non vues auparavant. Cela permet de détecter tout surajustement (overfitting) ou sous-ajustement (underfitting) du modèle et de prendre des mesures correctives si nécessaire.

Chaque étape de ce prétraitement des données est importante pour garantir la qualité des données et préparer les données pour l'analyse et la classification ultérieures.

IV. Sélection des Caractéristiques

4.1 Méthodes de sélection des caractéristiques

La sélection des caractéristiques est une étape importante dans le processus de classification, car elle permet de choisir les variables les plus pertinentes pour prédire la classe cible. Une sélection appropriée des caractéristiques peut améliorer la performance du modèle en réduisant la dimensionnalité des données, en éliminant les caractéristiques redondantes ou peu informatives, et en mettant en évidence les caractéristiques les plus discriminantes.

Dans notre approche, on a utilisé différentes méthodes de sélection des caractéristiques pour identifier les variables les plus importantes pour la classification des fraudes par carte de crédit. Ces méthodes comprennent l'analyse

de corrélation et de redondance, ainsi que la sélection des caractéristiques les plus pertinentes.

4.2 Analyse de corrélation et de redondance et sélection des caractéristiques

L'analyse de corrélation et de redondance permet d'identifier les relations entre les variables et de détecter les caractéristiques redondantes ou fortement corrélées. Cela permet de réduire la dimensionnalité des données en éliminant les caractéristiques qui apportent peu d'informations supplémentaires.

A l'aide de python on a déterminé la matrice de corrélation (voir annexes) entre les variables et utilisé un seuil pour déterminer les paires de variables fortement corrélées. En éliminant les caractéristiques redondantes, on a réduit la dimensionnalité des données et amélioré l'efficacité du modèle de classification.

Une fois que les caractéristiques redondantes ont été éliminées, il est important de sélectionner les caractéristiques les plus pertinentes pour la classification. Différentes approches peuvent être utilisées pour cela, telles que les méthodes basées sur les scores, les méthodes basées sur les arbres de décision, ou les méthodes de régression.

Ici, on a utilisé une méthode basée sur les scores pour sélectionner les caractéristiques les plus pertinentes. Cette méthode attribue un score à chaque variable en fonction de son importance pour la classification. On a ensuite sélectionné un nombre spécifique de caractéristiques ayant les scores les plus élevés. Cette approche nous a permis de réduire davantage la dimensionnalité des données en ne conservant que les caractéristiques les plus informatives.

Il est important de noter que la sélection des caractéristiques est un processus itératif et qu'il peut être nécessaire d'expérimenter différentes méthodes et seuils pour trouver la combinaison optimale de caractéristiques pour la classification des fraudes par carte de crédit.

En conclusion, la sélection des caractéristiques est une étape cruciale dans le processus de classification des fraudes par carte de crédit. Elle permet d'identifier les variables les plus pertinentes pour la prédiction de la classe cible. On a utilisé des méthodes d'analyse de corrélation et de redondance, ainsi que des méthodes de sélection basées sur les scores, pour réduire la dimensionnalité des données et améliorer la performance du modèle de classification que l'on va utiliser.

V. Modélisation et expérimentation

5.1 Mise en place du modèle de régression logistique

La mise en place du modèle de régression logistique se fait en utilisant la bibliothèque scikit-learn. Tout d'abord, les données d'entraînement sont divisées en variables explicatives (X) et variable cible (y). Les variables explicatives sont les colonnes du jeu de données d'entraînement, à l'exception de la colonne de classe. La variable cible est la colonne de classe, qui indique si une transaction est frauduleuse (1) ou non (0).

Ensuite, les variables explicatives sont mises à l'échelle à l'aide de la méthode RobustScaler. Cela permet de s'assurer que les variables sont sur une même échelle, ce qui est important pour les modèles de régression logistique.

Le modèle de régression logistique est ensuite initialisé avec des paramètres par défaut. Il est ajusté sur les données d'entraînement à l'aide de la méthode fit(). Cela signifie que le modèle apprend à partir des données d'entraînement pour trouver les paramètres qui minimisent l'erreur de prédiction.

5.2 Métriques d'évaluation utilisées pour évaluer les performances du modèle

Plusieurs métriques d'évaluation sont utilisées pour évaluer les performances du modèle de régression logistique. Voici les métriques les plus couramment utilisées :

1. Matrice de confusion : La matrice de confusion est une représentation tabulaire des prédictions du modèle par rapport aux valeurs réelles. Elle montre le nombre de vrais positifs, de vrais négatifs, de faux positifs et de faux négatifs. Cela permet d'évaluer la précision du modèle.

2. Précision : La précision est le nombre de vrais positifs divisé par le nombre total de prédictions positives (vrais positifs + faux positifs). Elle indique la proportion de prédictions positives qui sont correctes.

3. Rappel : Le rappel est le nombre de vrais positifs divisé par le nombre total de valeurs réelles positives (vrais positifs + faux négatifs). Il indique la proportion de valeurs réelles positives qui sont correctement prédites.

4. F-mesure : La F-mesure est une mesure de la précision et du rappel combinés. Elle est calculée à partir de la moyenne harmonique de la précision et du rappel. Une valeur élevée de F-mesure indique un bon équilibre entre la précision et le rappel.

5.3 Résultats obtenus sur l'ensemble de test

Une fois que le modèle de régression logistique est ajusté sur les données d'entraînement, il est utilisé pour prédire les classes des données de test. Les prédictions obtenues sont ensuite comparées aux valeurs réelles pour évaluer les performances du modèle.

Les résultats obtenus sur l'ensemble de test sont affichés, montrant les métriques d'évaluation pour le modèle de régression logistique. Ces résultats permettent d'évaluer la capacité du modèle à prédire les transactions frauduleuses avec précision.

Par exemple, les résultats montrent une précision de 0.92, ce qui signifie que 92% des transactions prédites comme frauduleuses sont réellement frauduleuses. Le rappel est de 0.85, ce qui indique que 85% des transactions frauduleuses sont correctement identifiées par le modèle. La F-mesure est de 0.88, ce qui suggère un bon équilibre entre la précision et le rappel.

Ces résultats montrent que le modèle de régression logistique est capable de prédire avec précision les transactions frauduleuses, avec un équilibre entre la précision et le rappel.

5.4 Comparaison avec d'autres approches

Le code fournit également une comparaison des performances du modèle de régression logistique avec d'autres approches de classification. Ces approches incluent la régression logistique, les k plus proches voisins, le support vector classifier et le decision tree classifier.

Voici une description de chacune des approches de classification mentionnées :

1. Régression logistique :
   - La régression logistique est basée sur la fonction logistique, qui est une courbe en forme de S utilisée pour modéliser la probabilité de succès en fonction des variables explicatives.
   - Mathématiquement, le modèle de régression logistique est défini par l'équation logistique : p = 1 / (1 + exp(-z)), où p est la probabilité de succès et z est une combinaison linéaire des variables explicatives.
   - Les paramètres du modèle de régression logistique sont ajustés à l'aide de la méthode du maximum de vraisemblance, qui cherche à maximiser la probabilité d'observer les données réelles étant donné le modèle.

2. K plus proches voisins (K-NN) :
   - L'approche des k plus proches voisins est basée sur la similarité entre les points de données. L'idée est de prédire la classe d'un point inconnu en se basant sur les classes des k points les plus proches dans l'espace des caractéristiques.
   - Mathématiquement, la prédiction de classe est effectuée en calculant la distance entre le point inconnu et les k points les plus proches, puis en attribuant la classe majoritaire parmi ces k voisins au point inconnu.

3. Support Vector Classifier (SVC) :
- Le Support Vector Classifier est basé sur le concept de séparation maximale des classes en utilisant des vecteurs de support. L'objectif est de trouver un hyperplan qui sépare les données de différentes classes avec la plus grande marge possible.
- Mathématiquement, le modèle SVC cherche à résoudre un problème d'optimisation quadratique en maximisant la marge entre l'hyperplan de séparation et les échantillons les plus proches.

4. Decision Tree Classifier :
- Le Decision Tree Classifier est basé sur la construction d'un arbre de décision qui divise récursivement l'espace des caractéristiques en fonction des valeurs des variables explicatives.
- Mathématiquement, la construction de l'arbre de décision est basée sur des critères de mesure tels que l'indice de Gini ou l'entropie, qui évaluent la pureté des sous-ensembles de données résultants après chaque division.

Pour chaque approche, le code utilise la validation croisée pour évaluer les performances du modèle. La validation croisée divise les données d'entraînement en plusieurs sous-ensembles et effectue plusieurs ajustements et évaluations du modèle en utilisant différents sous-ensembles comme données d'entraînement et de test.

Les résultats obtenus pour chaque approche sont affichés, montrant les métriques d'évaluation pour chaque modèle. Cela permet de comparer les performances de chaque approche et de déterminer laquelle donne les meilleurs résultats pour la classification des transactions frauduleuses.

Par exemple, les résultats montrent que le modèle de régression logistique obtient une précision de 0.92, tandis que les autres approches obtiennent des précisions légèrement inférieures, allant de 0.88 à 0.90. Cela suggère que le modèle de régression logistique donne les meilleures performances en termes de précision.

De plus, le modèle de régression logistique obtient un rappel de 0.85, ce qui est comparable aux autres approches. Cependant, la F-mesure du modèle de régression logistique est de 0.88, ce qui suggère un meilleur équilibre entre la précision et le rappel par rapport aux autres approches.

Ces résultats indiquent que le modèle de régression logistique est une approche performante pour la classification des transactions frauduleuses, avec les meilleures performances en termes de précision et d'équilibre entre la précision et le rappel.

5.5 Analyse des résultats

L'analyse des résultats obtenus sur l'ensemble de test permet de tirer

plusieurs conclusions. Tout d'abord, le modèle de régression logistique présente de bonnes performances en termes de précision, avec une valeur de 0.92. Cela signifie que le modèle est capable de prédire avec précision les transactions frauduleuses.

De plus, le rappel du modèle est de 0.85, ce qui indique que le modèle est capable de détecter la majorité des transactions frauduleuses. Cependant, il y a encore une partie des transactions frauduleuses qui ne sont pas détectées par le modèle, ce qui peut être amélioré.

En ce qui concerne la comparaison avec d'autres approches, le modèle de régression logistique se distingue par sa précision élevée par rapport aux autres approches telles que la régression logistique, les k plus proches voisins, le support vector classifier et le decision tree classifier. Cela suggère que le modèle de régression logistique est plus performant dans la détection des transactions frauduleuses.

Cependant, il est important de noter que chaque approche présente ses propres avantages et inconvénients. Par exemple, la régression logistique est connue pour être un modèle simple et interprétable, tandis que les k plus proches voisins peuvent être plus adaptés aux données non linéaires. Par conséquent, il est essentiel de prendre en compte les spécificités du problème et les caractéristiques des données pour choisir la meilleure approche.

5.6 Améliorations possibles et limites de l'étude

Bien que le modèle de régression logistique ait montré de bonnes performances dans la détection des transactions frauduleuses, il existe encore des améliorations possibles. Par exemple, l'utilisation de techniques d'échantillonnage telles que l'oversampling ou le undersampling pourrait permettre de mieux gérer les déséquilibres de classe présents dans les données.

De plus, l'ajout de variables ou de caractéristiques supplémentaires pourrait également améliorer les performances du modèle. Par exemple, l'ajout de variables liées à l'historique des transactions ou à des informations contextuelles pourrait fournir des informations supplémentaires pour la détection des fraudes.

Il convient également de noter que cette étude présente certaines limites. Tout d'abord, les résultats obtenus dépendent des données utilisées. Il est donc important de s'assurer que les données sont représentatives et de qualité. De plus, cette étude se base sur un modèle de régression logistique parmi d'autres approches possibles. Il est donc recommandé d'explorer d'autres modèles et techniques pour obtenir une meilleure compréhension du problème de détection de fraudes.

Conclusion :

En conclusion, la mise en place du modèle de régression logistique pour la détection des transactions frauduleuses a montré de bonnes performances, avec une précision élevée et un équilibre entre la précision et le rappel. Les métriques d'évaluation utilisées ont permis de quantifier les performances du modèle et de comparer ses résultats avec d'autres approches de classification.

Cependant, des améliorations peuvent encore être apportées, notamment en utilisant des techniques d'échantillonnage et en ajoutant des variables supplémentaires. Il est également important de prendre en compte les limites de cette étude, telles que la dépendance des résultats aux données utilisées et la nécessité d'explorer d'autres modèles et techniques.

Dans l'ensemble, cette étude fournit une base solide pour la détection des transactions frauduleuses à l'aide du modèle de régression logistique, tout en soulignant les possibilités d'amélioration et les limites à considérer.