**Rapport d'Analyse et de Modélisation pour la Prédiction d'Approbation de Prêt**

**Par Chayma Miledi – M1 ECOFIQ**

**1. Introduction**

Ce rapport présente une analyse approfondie et une modélisation prédictive pour évaluer et prédire les approbations de prêt. L'étude utilise un ensemble de données comprenant diverses caractéristiques financières et personnelles des demandeurs de prêt. L'objectif principal est de développer des modèles de machine learning capables de prédire avec précision si une demande de prêt sera approuvée ou rejetée.

Les données utilisées dans ce projet ont été extraites du lien suivant : [Loan Approval Prediction Dataset](https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset/data), mis à disposition par Archit Sharma sur Kaggle. Ce jeu de données, dénommé "loan approval dataset", constitue une compilation de dossiers financiers et d'informations associées utilisées pour déterminer l'éligibilité des individus ou des organisations à l'obtention de prêts auprès d'une institution financière. Il englobe divers facteurs tels que le score CIBIL, le revenu, le statut professionnel, la durée du prêt, le montant du prêt, la valeur des actifs et le statut du prêt. Couramment utilisé dans le domaine de l'apprentissage automatique et de l'analyse de données, ce jeu de données permet le développement de modèles et d'algorithmes visant à prédire la probabilité d'approbation d'un prêt en fonction des caractéristiques fournies.

**2. Exploration des Données (EDA)**

**2.1 Chargement et Inspection des Données**

Nous avons commencé par charger l'ensemble de données, inspecter les premières lignes et vérifier la qualité des données.

Ce projet utilise plusieurs bibliothèques Python essentielles :

1. **Pandas :** Manipulation et analyse des données.
2. **Matplotlib et Seaborn :** Visualisation des données.
3. **Scikit-learn :** Apprentissage automatique, modèles de classification et régression.
4. **XGBoost :** Gradient boosting pour des performances élevées.
5. **Scikit-plot :** Création de courbes ROC.
6. **Statsmodels :** Estimation de modèles statistiques.
7. **GridSearchCV :** Recherche des meilleurs hyperparamètres.

Ces outils offrent une gamme complète pour l'analyse de données, la création de modèles prédictifs et l'évaluation de leurs performances.

**2.2 Prétraitement des Données**

Nous avons effectué plusieurs étapes de prétraitement, notamment la suppression des espaces dans les noms de colonnes, la gestion des valeurs négatives dans la colonne 'residential\_assets\_value', et l'encodage des variables catégorielles.

**2.3 Analyse Exploratoire des Données (EDA)**

Nous avons exploré la matrice de corrélation entre les différentes caractéristiques et visualisé la distribution de la variable cible 'loan\_status'.

Une image contenant texte, capture d’écran, diagramme, Parallèle

Description générée automatiquementLa caractéristique la plus corrélée avec le Loan\_status est le cibil\_score : **0,77**.

Une image contenant texte, capture d’écran, Rectangle, diagramme

Description générée automatiquement

Les données contiennent **2 656** prêts approuvés et **1 613** prêts rejetés.

**3. Modèles de Machine Learning**

**3.1 Préparation des Données**

Nous avons séparé l'ensemble de données en fonction des caractéristiques (X) et de la variable cible (y). Ensuite, nous avons divisé les données en ensembles d'entraînement et de test.

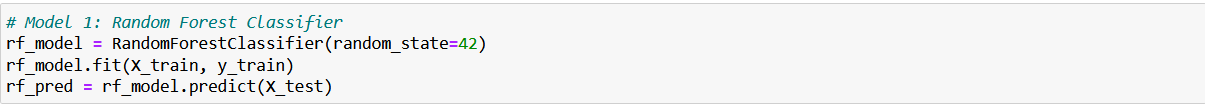
Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

**3.2 Modèles de Machine Learning**

**3.2.1 Random Forest Classifier**

Nous avons utilisé un modèle Random Forest pour sa capacité à capturer des relations complexes dans les ensembles de données mixtes.



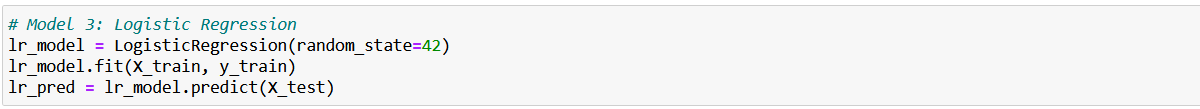
**3.2.2 XGBoost**

Le modèle XGBoost, réputé pour ses performances élevées, a été utilisé pour traiter des motifs complexes dans de grands ensembles de données.



**3.2.3 Logistic Regression**

La régression logistique a été incluse pour améliorer l'interprétabilité et éclairer l'impact individuel des caractéristiques.



**4. Résultats**

Nous avons évalué les performances des modèles à l'aide de matrices de confusion et de métriques telles que la précision, le rappel et l'aire sous la courbe ROC (AUC).

**4.1 Résultats de Random Forest Classifier**

La matrice de confusion montre que le modèle Random Forest Classifier a une capacité élevée à prédire les prêts approuvés et rejetés.

Une image contenant texte, capture d’écran, diagramme, Rectangle

Description générée automatiquement

* Le modèle a correctement prédit 529 cas de prêts approuvés (Vrais Positifs).
* Il a prédit correctement 308 cas de prêts rejetés (True Negatives).
* Il y a eu 10 cas où le modèle a prédit de manière incorrecte l’approbation de prêts rejetés (faux positifs).
* Il y a eu 7 cas où le modèle a prédit de manière incorrecte le rejet des prêts approuvés (faux négatifs).

**4.2 Résultats de XGBoost**

Le modèle XGBoost présente des performances similaires à Random Forest, avec une légère amélioration de la précision.

Une image contenant texte, capture d’écran, diagramme, Rectangle

Description générée automatiquement

* Le modèle XGBoost a correctement prédit 530 cas de prêts approuvés (True Positives).
* Il a prédit correctement 309 cas de prêts rejetés (True Negatives).
* Il y a eu 9 cas où le modèle a prédit de manière incorrecte l’approbation des prêts rejetés (faux positifs).
* Il y a eu 6 cas où le modèle a prédit de manière incorrecte le rejet des prêts approuvés (faux négatifs).

Haut du formulaire

Bas du formulaire

**4.3 Résultats de Logistic Regression**

Le modèle de régression logistique présente des performances inférieures en termes d'exactitude, de précision et d'AUC.

Une image contenant texte, capture d’écran, diagramme, conception

Description générée automatiquement

* Le modèle de régression logistique a prédit correctement 535 cas de prêts approuvés (vrais positifs).
* Il a prédit correctement 2 cas de prêts rejetés (True Negatives).
* Il y a eu 316 cas où le modèle a prédit de manière incorrecte l’approbation de prêts rejetés (faux positifs).
* Il y a eu 1 cas où le modèle a prédit de manière incorrecte le rejet d’un prêt approuvé (faux négatif).

**4.4 Évaluation des Performances**

Une image contenant texte, capture d’écran, ligne, Tracé

Description générée automatiquementUne image contenant texte, capture d’écran, ligne, Tracé

Description générée automatiquementNous avons évalué les performances de chaque modèle en termes d'exactitude, de précision, de rappel et d'AUC.

Une image contenant texte, capture d’écran, ligne, Tracé

Description générée automatiquement

Une image contenant texte, Police, capture d’écran, ligne

Description générée automatiquement

**5. Analyse des Performances des Modèles**

**5.1 Random Forest et XGBoost**

Les modèles Random Forest et XGBoost démontrent des performances globales élevées avec une exactitude, une précision et un rappel équilibrés. Ils parviennent à identifier correctement une partie significative des prêts approuvés et rejetés.

**5.2 Régression Logistique**

La régression logistique, bien que moins complexe computationnellement, présente une exactitude, une précision et un AUC nettement inférieurs. Elle peut ne pas être le meilleur choix pour cette tâche particulière en raison de sa tendance à prédire de manière excessive l'approbation de prêts, conduisant à un grand nombre de faux positifs.

**6. Conclusion**

Après une évaluation approfondie, le modèle XGBoost émerge comme le choix optimal pour la prédiction d'approbation de prêt. Son équilibre entre précision et rappel, ainsi que ses performances globales élevées, en font le modèle le plus adapté à cette tâche.

**7. Optimisation du Modèle avec Grid Search**

Nous avons effectué une optimisation du modèle XGBoost en utilisant une recherche par grille pour trouver les hyperparamètres optimaux.

Une image contenant texte, Police, capture d’écran, ligne

Description générée automatiquement

le modèle XGBoost avec une profondeur maximale de 10 et 50 estimateurs a montré des performances exceptionnelles, avec une précision de prédiction de 98.01% sur l'ensemble de test. Ces résultats suggèrent que le modèle est robuste et généralise bien aux données non vues.

**8. Analyse des Importances des Caractéristiques**

En analysant les importances des caractéristiques du modèle XGBoost, nous identifions les facteurs critiques, modérés et moins influents.

Une image contenant texte, capture d’écran, nombre, ligne

Description générée automatiquement

**8.1 Facteurs Critiques**

* Montant du Prêt (195.0) : Le plus influent.
* Score CIBIL (126.0) : Très important.
* Revenu Annuel (118.0) : Influence significative.

**8.2 Facteurs Modérés**

* Valeur des Biens Résidentiels (113.0)
* Valeur des Biens Commerciaux (74.0)
* Valeur des Biens de Luxe (67.0)
* Nombre de Personnes à Charge (47.0)
* Durée du Prêt (47.0)
* Valeur des Biens de la Banque (38.0)

**8.3 Moins Influents**

* Travailleur Indépendant (14.0)
* Éducation (9.0)

**9. Conclusion Finale**

En conclusion, ce projet a démontré l'efficacité des modèles Random Forest et XGBoost dans la prédiction d'approbation de prêt. Le modèle XGBoost, en particulier après une optimisation minutieuse, se démarque comme le choix préféré en raison de ses performances globales supérieures.

L'analyse des importances des caractéristiques fournit des informations précieuses sur les facteurs qui influent le plus sur la décision d'approbation de prêt. Ces informations peuvent être utilisées pour prendre des décisions plus éclairées dans le processus d'approbation de prêt.

## 10. Recommandations et Perspectives

### 10.1 Recommandations

* **Utilisation du Modèle XGBoost :** Nous recommandons l'implémentation du modèle XGBoost dans le processus d'approbation de prêt en raison de ses performances supérieures.
* **Surveillance Continue :** La surveillance continue des performances du modèle est cruciale. Les modèles doivent être réévalués périodiquement pour garantir leur adaptation aux évolutions des données.

### 10.2 Perspectives

* **Intégration de Données Supplémentaires :** L'ajout de données supplémentaires, telles que des informations économiques externes, pourrait améliorer la précision du modèle.
* **Optimisation Continue :** L'exploration d'algorithmes d'optimisation continue, en particulier pour XGBoost, pourrait conduire à des améliorations continues des performances.
* **Interprétabilité :** Bien que XGBoost soit performant, son manque d'interprétabilité peut être un défi. L'exploration de techniques d'interprétabilité, telles que SHAP (SHapley Additive exPlanations), pourrait rendre le modèle plus compréhensible.
* **Mise en Production :** Si ce n'est pas déjà fait, la mise en production du modèle XGBoost pour une utilisation opérationnelle serait une étape logique.