



Formation Data Analyst

Projet Final :

**Conception de modèles prédictifs pour
l'évaluation des risques sur l'approbation
d'un crédit bancaire.**

Du 18 aout au 24 octobre 2025

Roger Mamaty

Table des Matières



1- [Présentation](#).....3

1-1 Jeu de données.....3

1-2 Catégorie des données.....3



2- [Analyse exploratoire des données sous Python](#).....5

2-1 Intégration et conversion des données.....5

2-2 Traitement des valeurs aberrantes.....5

2-3 Matrice de Corrélation.....6



3- [Régression Logistique](#).....7

3-1 Évaluation du modèle.....7

3-2 matrice de confusion.....7



4- [Forets Aléatoire](#).....8

4-1 Évaluation du modèle.....8

4-2 matrice de confusion.....8

4-3 visualisation de l'arbre a 2 niveaux9



5- [Comparaison Régression Logistique / Forets Aléatoire](#).....10



6- [Présentation des données sous Power Bi](#).....11

6-1 Analyse Démographique.....11

6-2 Analyse Emplois et Revenus.....12

6-3 Analyse Financière.....13

6-4 Analyse Financière.....14

1- Présentation

Dans le cadre du rapport final de la formation de « Data Analyst » auprès de **DataSuits**, j'ai choisi d'utiliser un jeu de données publique sur les données personnelles financières permettant de développer des modèles prédictifs pour l'évaluation des risques bancaires. Je vais l'utiliser pour :

- Régression du score de risque : Prédire un score de risque continu associé à la probabilité de défaut de paiement ou d'instabilité financière de chaque individu.
- Classification binaire : Déterminer le résultat binaire de l'approbation d'un prêt, indiquant si un demandeur est susceptible d'être approuvé ou refusé pour un prêt.

1-1 Jeu de données

Le jeu de données retenue est Loan.csv (<https://www.kaggle.com/datasets/lorenzozoppelletto/financial-risk-for-loan-approval/data>). Le fichier **Loan.csv** contient 36 colonnes et 20 000 enregistrements. Une phase de traduction est réalisée sur les colonnes origines afin de simplifier l'usage de ces colonnes dans les différents processus d'analyse.

Les données comprennent de nombreuses informations, nous allons les présenter en les regroupant par type d'information : *Temporelles, Démographiques, Emplois et Revenus, logements, Financières, Charges et endettement, Crédit, Prêt, Risque.*

1-2 Catégorie des données

Variables temporelles

Variable	Description
Date_demande	Date de soumission de la demande de prêt.

Variables liées au logement

Variable	Description
Statut_proprietaire	Statut résidentiel (locataire, propriétaire...).

Variables démographiques

Variable	Description
Age	Âge du demandeur.
Etat_civil	État matrimonial (célibataire, marié, divorcé...).
Nbre_Personnes_a_Charges	Nombre de personnes à charge.
Niveau_Education	Plus haut niveau d'éducation atteint.

Variables liées à l'emploi et aux revenus

Variable	Description
Statut_Emploi	Situation professionnelle (CDI, CDD, indépendant...).
Experience	Années d'expérience professionnelle.
Duree_Emploi_Actuel	Durée dans l'emploi actuel.
Revenu_annuel	Revenu annuel.
Revenu_Mensuel	Revenu mensuel (souvent dérivé de AnnualIncome).

Variables financières générales

Variable	Description
Solde_Compte_Courant	Solde du compte courant.
Solde_Compte_Epargne	Solde du compte épargne.
Total_Actifs_Detenus	Valeur totale des actifs détenus.
Total_Dettes-Dus	Montant total des dettes.
Valeur_nette	Valeur nette (TotalAssets - TotalLiabilities).

Variables de charges et endettement

Variable	Description
Paielements_mensuels_dettes	Paielements mensuels de dettes existantes.
Ratio_Dette_Revenu	Ratio dettes / revenu.
Ratio_DettePret_Revenu	Ratio total dettes / revenu en tenant compte du prêt demandé.
Historique_paiement_charges	Historique de paiement des charges/abonnements (eau, électricité...).

Variables liées au crédit

Variable	Description
CreditScore	Score de crédit mesurant la solvabilité.
Duree_Historique_Credit	Durée totale de l'historique de crédit.
Historique_comportement_paiement	Historique des paiements passés (ponctualité, retards...).
Taux_Utilisation_carte_de_credit	Pourcentage d'utilisation des cartes de crédit.
Nombre_lignes_credit_actifs	Nombre de lignes de crédit actives.
Nombre_verifications_credits	Nombre de vérifications de crédit récent.
Historique_Faillite	Historique de faillites.
Precedent_Defaults_de_paiement	Historique de défauts de paiement sur prêts précédents.

Variables liées au prêt

Variable	Description
Montant_Pret	Montant du prêt demandé.
Duree_Pret	Durée de remboursement du prêt.
Motif_Pret	Objet du prêt (voiture, travaux...).
Taux_Interet_de_base	Taux d'intérêt de base du marché.
Taux_interet_Obtenu	Taux d'intérêt appliqué au demandeur.
Mensualite_a_payer_pret	Mensualité à payer pour le prêt.
Pret_Approuve	Statut d'approbation du prêt (oui/non).

Variables de risque

Variable	Description
RiskScore	Score global de risque calculé par l'institution.

2- Analyse exploratoire des données sous Python

2-1 Intégration et conversion des données

Le fichier Loan.CSV est intégré dans Python. Après intégration des données, nous procédons au formatage des données :

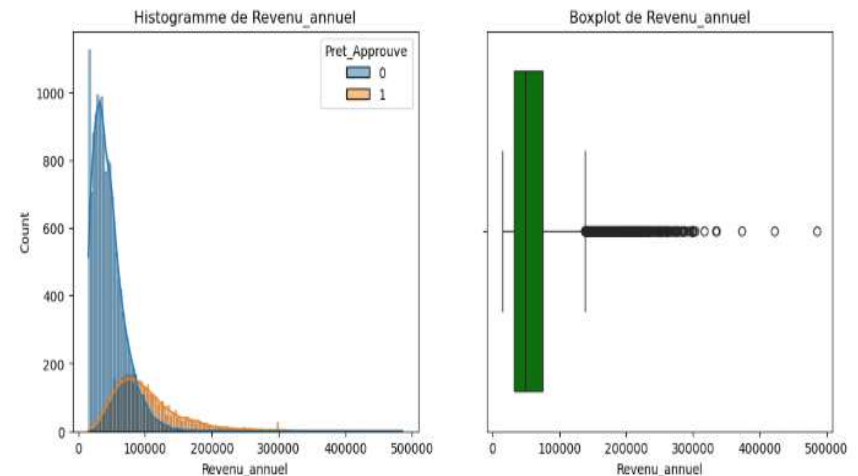
- La colonne date est transformé en type Date
- Les colonnes « *Historique_Faillite* », « *Precedent_Defaults_de_paiement* » et « *Pret_Approuve* » sont transformés en type Catégorie pour un meilleur traitement.
- Pas de doublons détectés dans les données du datagramme.
- De nouvelles création de colonnes seront créés afin de regrouper les colonnes à valeurs numériques en libellé du type *faible*, *moyen*, *élevé*. Qui vont regrouper des intervalles de valeur. Ces colonnes vont nous permettre une meilleure analyse sur des visuels de type PowerBy ou Excel.

2-2 Traitement des valeurs aberrantes

Nous allons utiliser l'affichage des Boxplot des colonnes pour visualiser d'éventuelles valeurs aberrantes. Dans l'exemple à droite, on constate de possibles valeurs Outliers et ou aberrantes.

Au vu des résultats, en tenant compte des différentes catégories de données et en faisant un test avec la méthode IQR pour supprimer automatiquement les valeurs aberrantes, je prends le choix de faire une suppression manuelle en me basant des Boxplot, tout en essayant de garder un jeu de données conséquent pour analyse ultérieure.

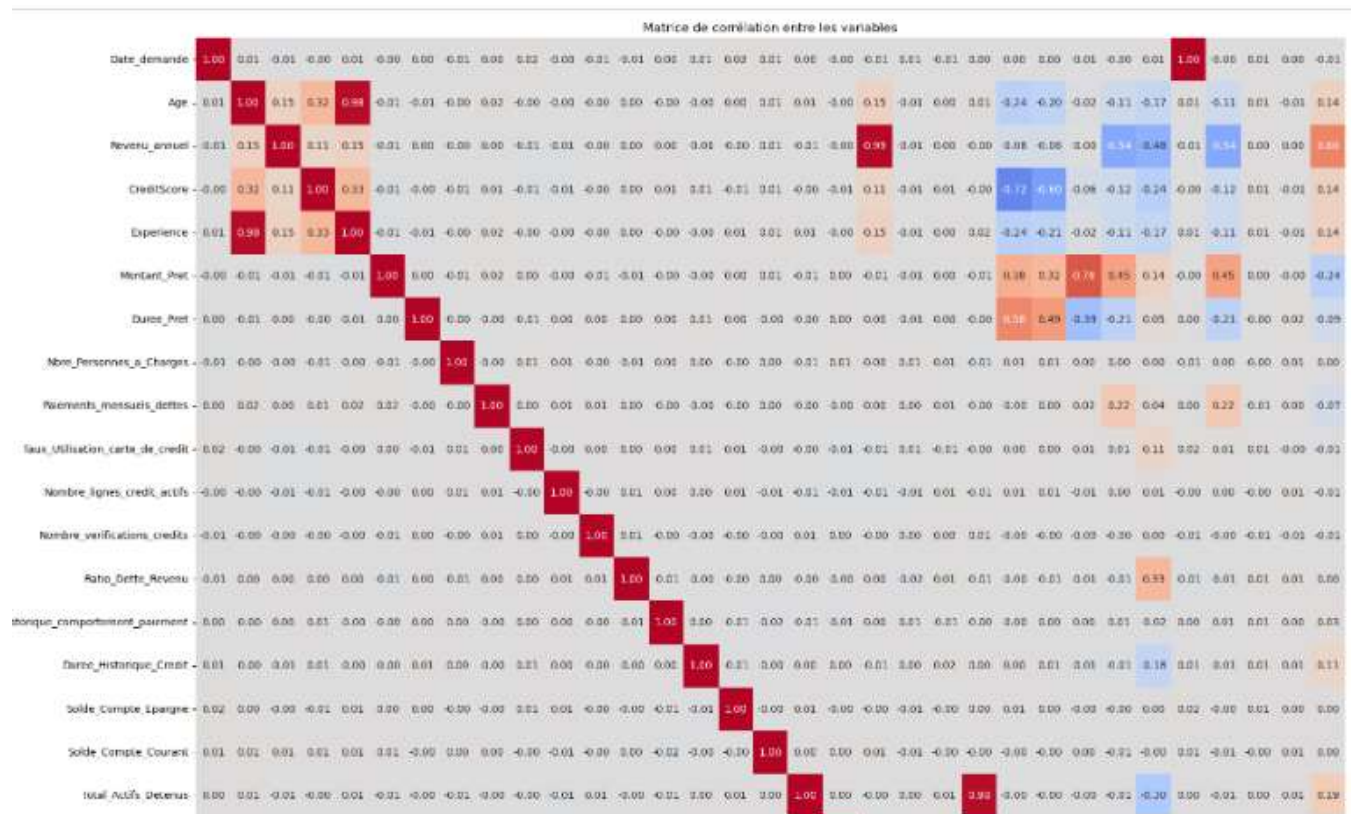
Pour finir, après réduction de certaines valeurs extrêmes, nous avons un jeu de données de 19830 lignes et 36 colonnes.



2-3 Matrice de Corrélation

La matrice de corrélation va nous permettre de voir :

- Les features qui ont le plus d'impact (positif ou négatif) sur l'obtention du crédit (Pret_Approuve).
- Exclure les features ayant une forte multi colinéarité.
- Exclure les features ayant un impact nul sur l'obtention du crédit (Pret_Approuve).



3- Régression Logistique

Après sélection des features, nous exécutons une régression logistique pour déterminer :

- Le taux de précision du modèle a prédire correctement les profils clients (positifs ou négatifs) par rapport aux données
- Les features impactant dans le modèle de régression pour générer un modèle prédiction de décision pour l'obtention ou non du crédit.

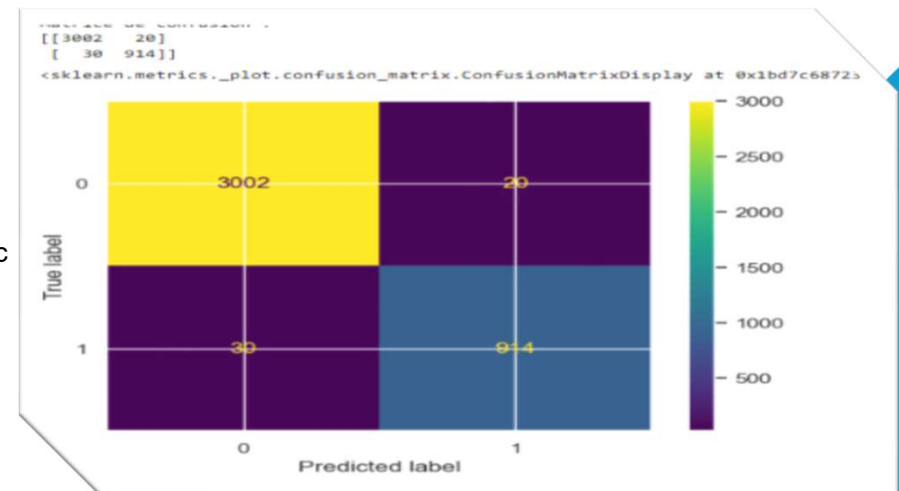
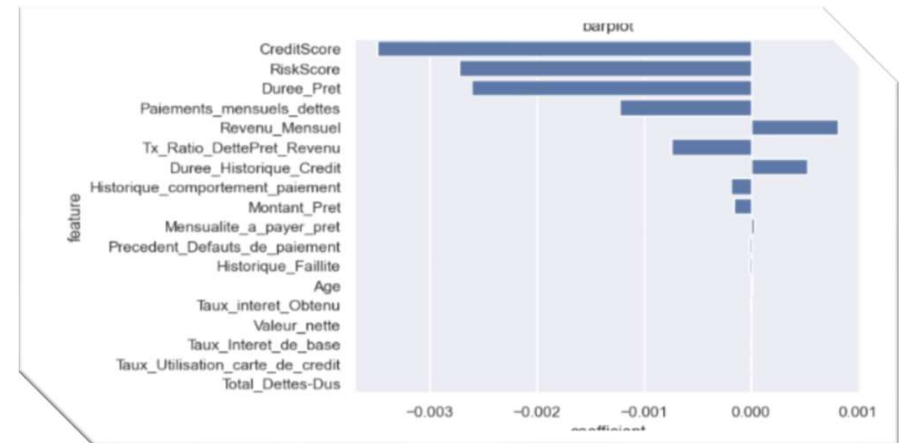
3-1 Évaluation du modèle

Le résultat du modèle de régression logistique nous donne les informations suivantes :

- **Précision du modèle en %** : **98.74%**
- Nombre de prédictions positifs bien effectuées : 97.86%
- Nombre de positifs (approuvé) bien prédit : 96.82%

3-2 matrice de confusion

- Le modèle est bon pour identifier si le dossier client du crédit sera accepté ou non. Avec 914 vrais dossiers acceptés et un taux général de précision de 98,74%, le modèle est capable de détecter les dossiers qui seront acceptés avec un taux de 97.86%.
- Le modèle a une marge d'amélioration pour réduire les erreurs : Les 20 faux positifs et 30 faux négatifs montrent que le modèle fait des erreurs.



4- Forêts Aléatoire

Nous allons utiliser un autre modèle qui est la Forêt Aléatoire et comparer le taux de précision de chaque modèle. Une sélection des most features qui seront intégrés dans le modèle

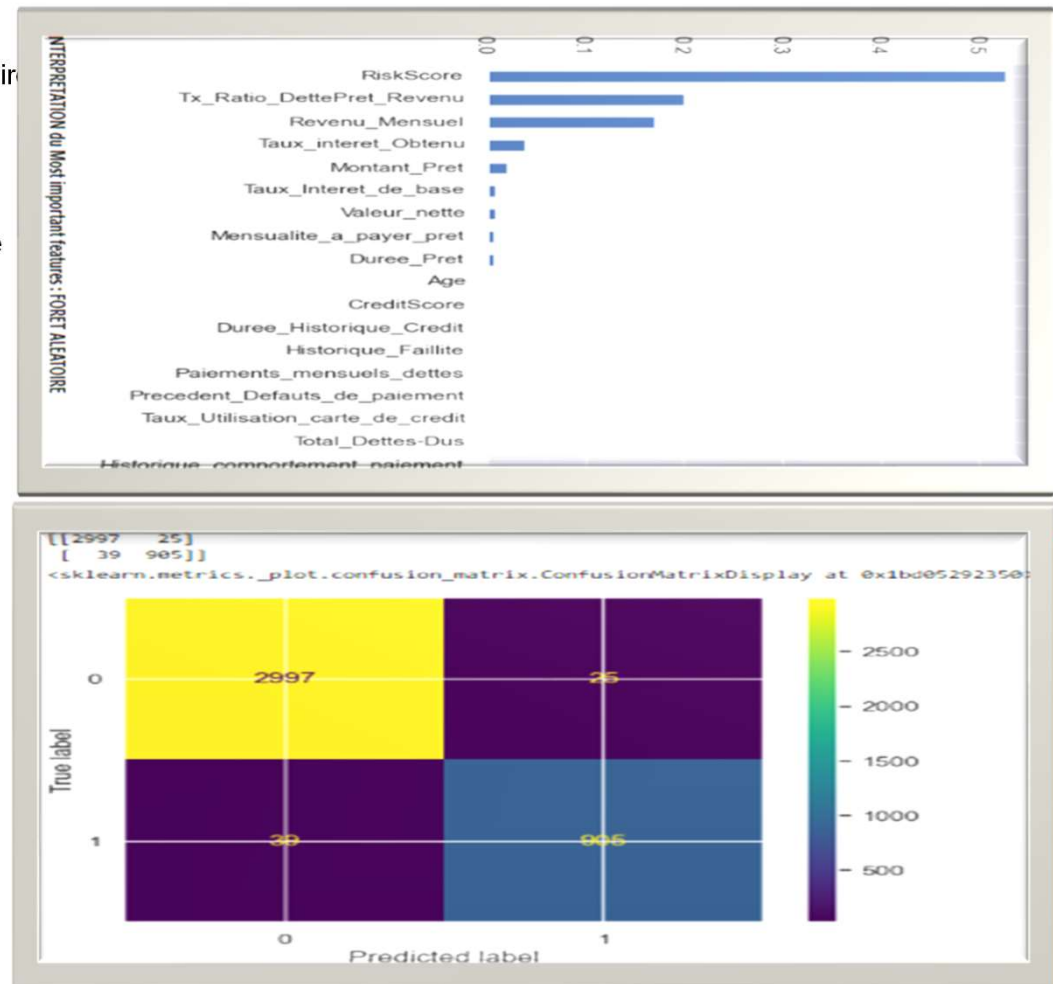
4-1 Évaluation du modèle

Le résultat du modèle de régression logistique nous donne les informations suivantes :

- **Précision du modèle en %** : **98.39%**
- Nombre de prédictions positifs bien effectuées : 97.31%
- Nombre de positifs (approuvé) bien prédit : 95.87%

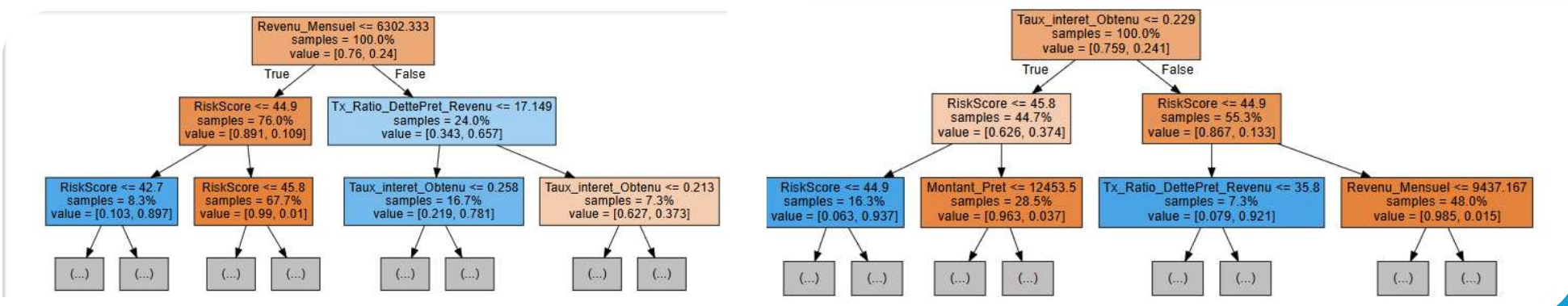
4-2 matrice de confusion

- Le modèle est bon pour identifier si le dossier client du crédit sera accepté ou non. Avec 905 vrais dossiers acceptés et un taux général de précision de 98,39%, le modèle est capable de détecter les dossiers qui seront acceptés avec un taux de 97.31%.
- Le modèle a une marge d'amélioration pour réduire les erreurs : Les 25 faux positifs et 39 faux négatifs montrent que le modèle peut progresser sur les erreurs.



4- Forets Aléatoire

4-3 visualisation de l'arbre a 2 niveaux



5- Comparaison Régression Logistique / Forets Aléatoire

Most feature Régression Logistique

1-CreditScore
2-RiskScore
3-Duree_Pret
4-Paiements_mensuels_dettes
5-Revenu_Mensuel
6-Tx_Ratio_DettePret_Revenu
7- Duree_Historique_Credit
8-Historique_comportement_paiement
9-Montant_Pret
10- Mensualite_a_payer_pret

Comparaison Régression Logistique / Foret Aléatoire

Libellé	Régression Logistique	Forets Aleatoire
Accuracy	98,74%	98,39%
Precision	97,86%	97,31%
Recall	96,82%	95,87%

Most Feature Forets Aléatoire

1-RiskScore
2-Tx_Ratio_DettePret_Revenu
3-Revenu_Mensuel
4-Taux_interet_Obtenu
5-Montant_Pret
6-Taux_interet_de_base
7-Valeur_nette
8- Mensualite_a_payer_pret
9-Duree_Pret

Interprétation :

- **accuracy** calcule la précision du modèle, c'est-à-dire le pourcentage de prédictions correctes par rapport à l'ensemble de test.
- **precision** est utilisée pour calculer la précision du modèle concernant la proportion de vrais positifs (client ayant le pret approuvé) parmi les predictions des clients positifs. c'est le nombre de positifs bien prédit (Vrai Positif) divisé par l'ensemble des positifs prédit (Vrai Positif + Faux Positif).
- **recall** est utilisée pour calculer le rappel du classificateur, c'est le nombre de positifs bien prédit (Vrai Positif) divisé par l'ensemble des positifs (Vrai Positif + Faux Négatif).

Concernant la précision générale (accuracy) des modèles, la Régression Logistique (98,74%) a un taux supérieur a celui de la Forets Aléatoire (98,39%). Concernant l'efficacité a trouver les vrais positifs (precision), la Régression Logistique (97,86%) a un taux supérieur a celui de la Forets Aléatoire (97,31%).

Conclusion :

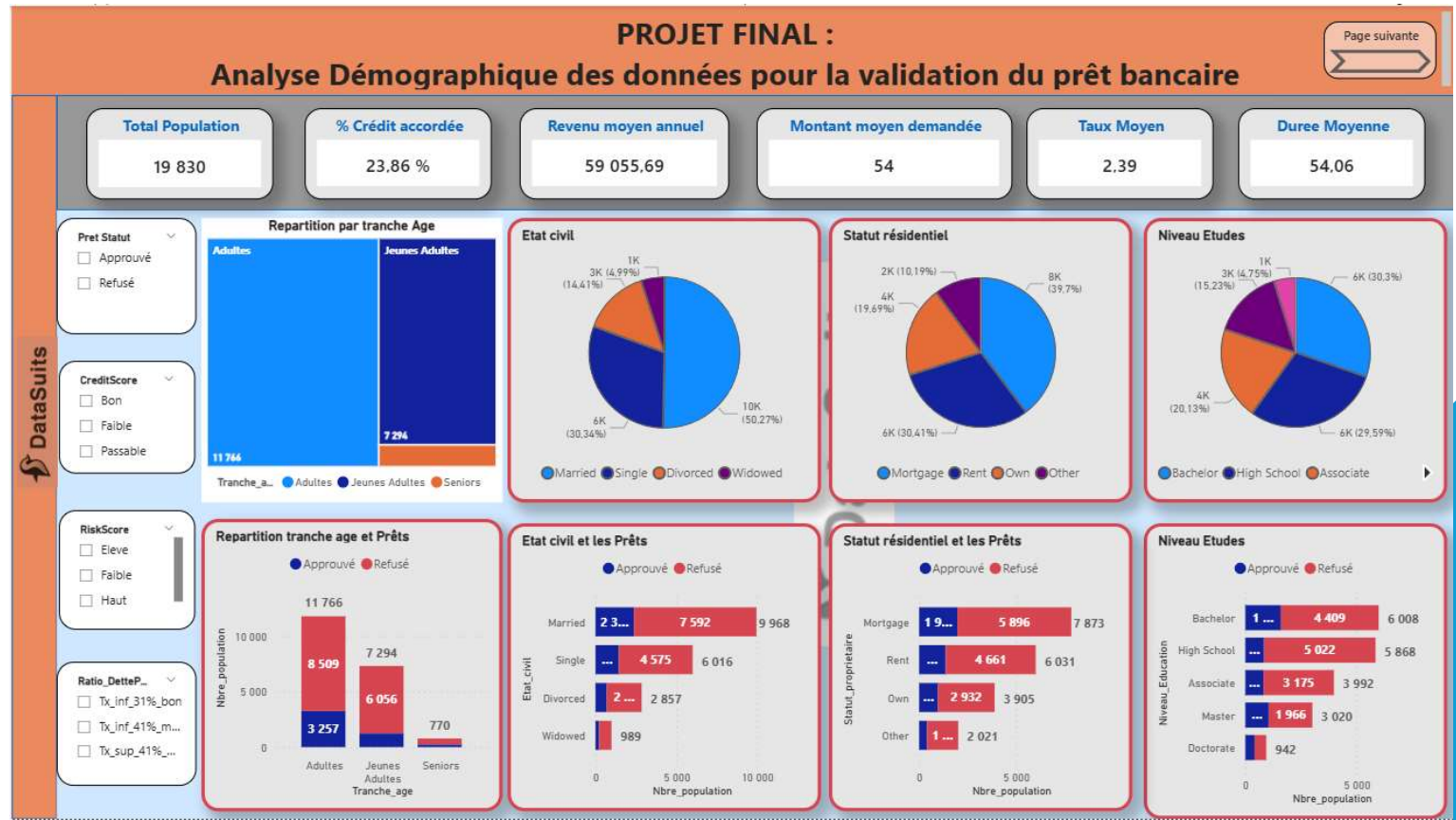
De façon générale, sur cet exemple, la Régression Logistique a un taux supérieur de réussite supérieur au modèle de la Forets Aléatoire. On peut dire que le modèle de Régression Logistique est plus adapté pour la prédiction des crédits (approuvé ou pas) dans la banque.

6- Présentation des données sous Power Bi

6-1 Analyse Démographique

On a 23,88% des demandes qui sont acceptés sur le volume total (19 830).

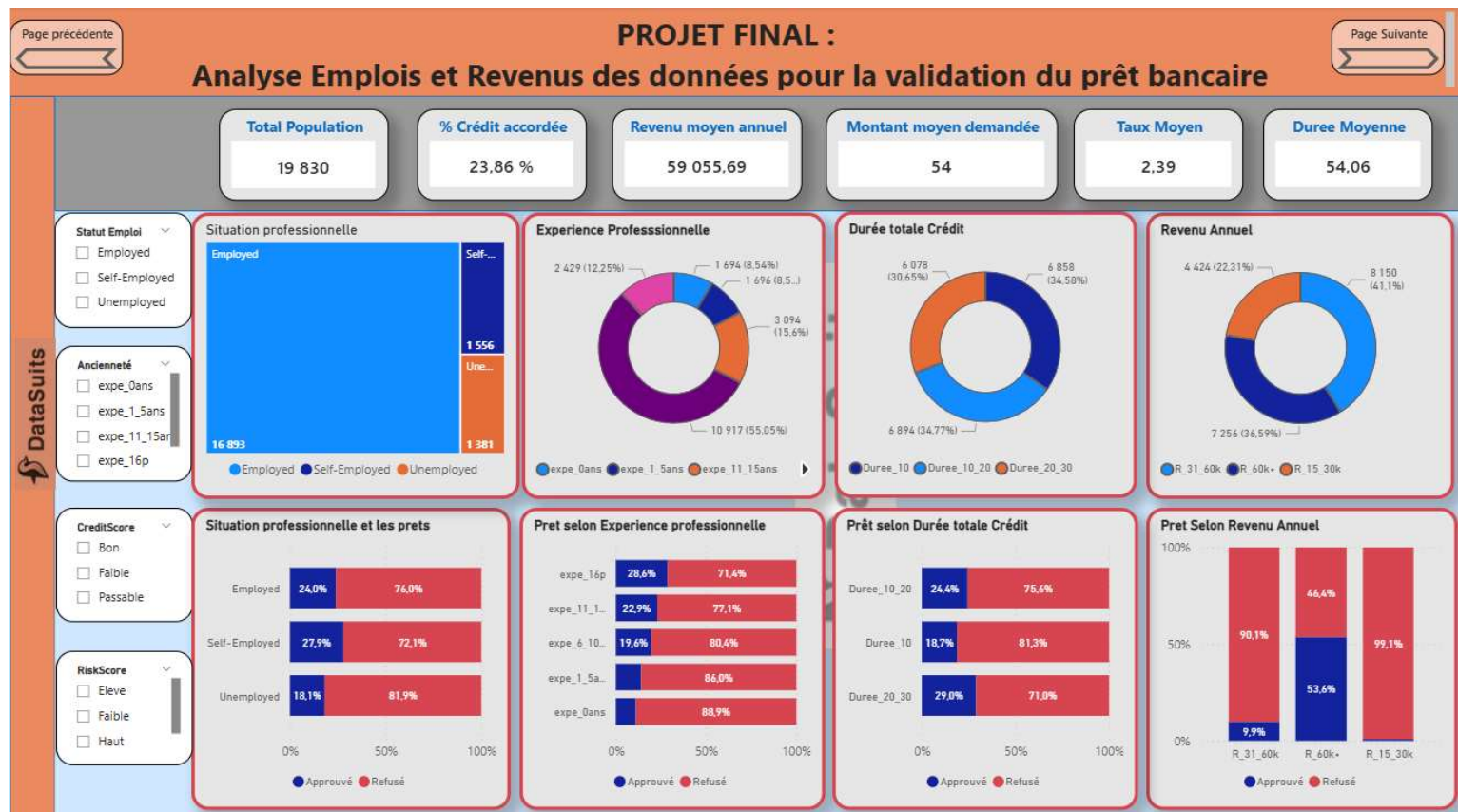
- Sur la tranche d'âges, c'est le profil « Adultes » qui obtient le plus de crédit (3257 sur 4732 soit 68,83%).
- Au niveau de l'état civil, c'est le profil « Marié » qui a obtenu le plus de crédit (2376 sur 4732 soit 50,21%).
- Au niveau du statut résidentiel, c'est la tranche « Mortgage (Hypothèque) » qui obtient le plus de crédit (1977 sur 4732 soit 41,78%).
- Au niveau des études, c'est le profil « Bachelor (Bac +X) » qui obtient le plus de prêt (1599 sur 4732 soit 33,78%).



6- Présentation des données sous Power Bi

6-2 Analyse Emplois et Revenus

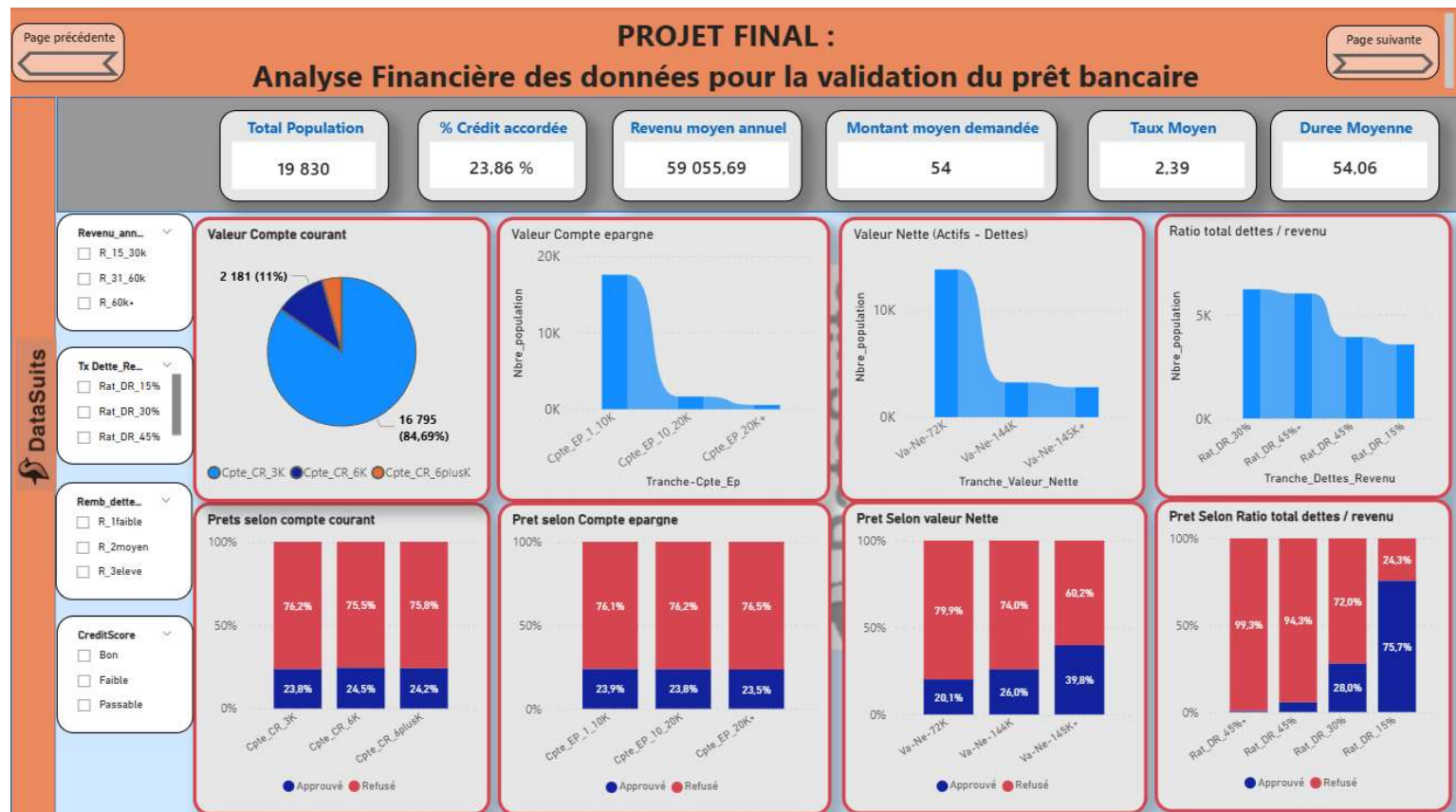
- Au de la situation professionnelle, c'est le profil employé qui « Entrepreneur/ auto-entrepreneur » qui obtient le plus de crédit soit 27,9%.
- Au niveau de l'expérience professionnelle, c'est le profil « plus de 16 ans » qui obtient le plus de crédit, soit 28,6%.
- Au niveau de la durée du prêt, c'est la durée 20-30 mois qui obtient le plus de crédit, soit 29%
- Au niveau des Revenus, c'est la tranche la plus haute plus de 60K qui obtient le plus de crédit soit 53,6%.



6- Présentation des données sous Power Bi

6-3 Analyse Financière

- Au niveau de la valeur du compte courant et du compte épargne, nous avons une distribution en entre 23% et 24% pour chaque profil.
- Concernant la valeur Nette, c'est le profil ayant le plus d'économie qui obtient le plus grand nombre de crédit soit 39,8%.
- Concernant le aux d'endettement, c'est le profil ayant un taux inférieur ou égal a 15% de ses revenus mensuels qui obtient le plus de crédit soit 75,7%. C'est le critère ayant le taux de succès le plus élevé en dehors du 'RiskScore' qui un élément de calcul de la banque



6- Présentation des données sous Power Bi

6-4 Analyse Financière

- Le « CreditScore » avec le profil bon obtient le meilleur taux d'acceptation, soit 18,4%.
- Le « RiskScore » avec le profil « Faible » obtient le meilleur taux d'acceptation, soit 98,2%. Cela signifie que ce critère est déterminant dans l'obtention du crédit.
- La durée des crédits obtenus dépassant les 20 mois a le taux d'acceptation le plus important soit 11,5%.
- Le critère « Motif Prêt » avec le profil « Auto » et « Education » ont le taux d'acceptation les plus importants. Soit 9,5% et 9,9%.

