DataSuits

# Formation Data Analyst
# Projet Final :
## Conception de modèles prédictifs pour l'approbation d'un crédit bancaire.

Du 18 aout au 24 octobre 2025

Roger Mamaty

DataSuits

# Table des Matières

# 1- Présentation

Dans le cadre du rapport final de la formation de « Data Analyst » auprès de **DataSuits**, j'ai choisi de travailler sur des modèles prédicatifs pour obtention du crédit : oui / non. Le jeu de données retenue est Loan.csv (https://www.kaggle.com/datasets/lorenzozoppelletto/financial-risk-for-loan-approval/data).

L'objectif est de développer des modèles prédictifs afin de 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 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). |

DataSuits

### 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 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 charges et endettement

| Variable | Description |
|---|---|
| **Paiements_mensuels_dettes** | Paiements 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 de risque

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

### 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_Defauts_de_paiement** | **Historique de défauts de paiement sur prêts précédents.** |

# 2- Analyse exploratoire des données sous Python

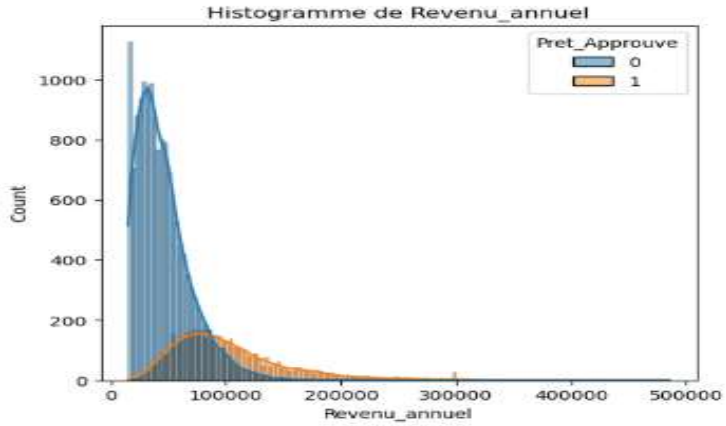
## 2-1 Intégration et conversion des données

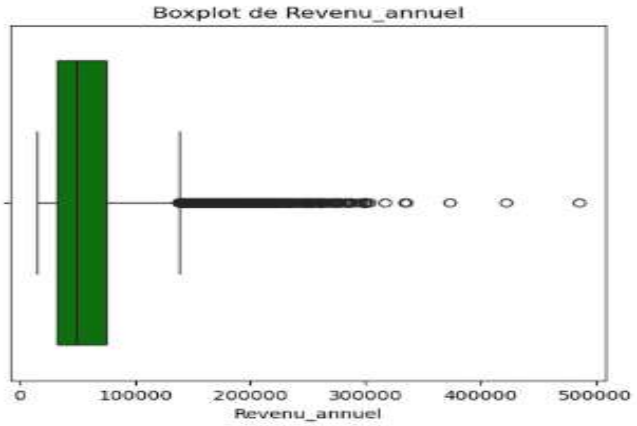
Le fichier Loan.CSV est intégré dans Python. Apres intégration des données, nous procédons au formatage des données :

- La colonne date est transformé en type Date
- 3 colonnes sont transformées en type Catégorie.
- Pas de doublons détectés dans les données du datagramme.
- De nouvelles création de colonnes seront crées 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 un regroupement sur des visuels de type PowerPoint 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 ci-dessous, on constate de possibles valeurs Outliers et ou aberrantes. Après réduction de certaines valeurs Outliers, nos avons un jeu de données de 19830 lignes et 36 colonnes.



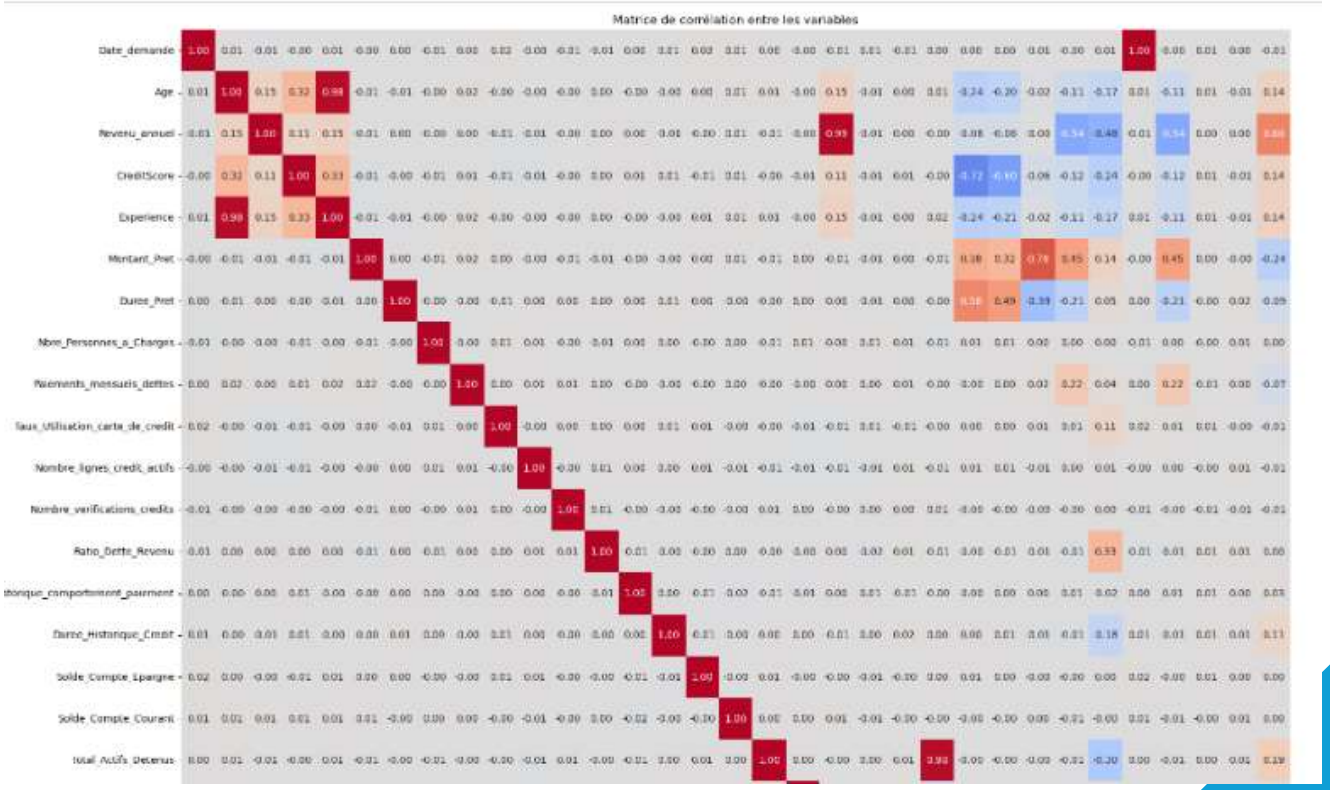



## 2-3 Matrice de Corrélation

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).

Extrait de la matrice de corrélation



# 3- Régression Logistique

Notre premier modèle a évaluer est la « Régression Logistique ». Ce modèle va nous permettre de détecter les « mosts features » lié a ce modèle,  et de réaliser un modèle de prédiction l'approbation ou non du crédit bancaire.

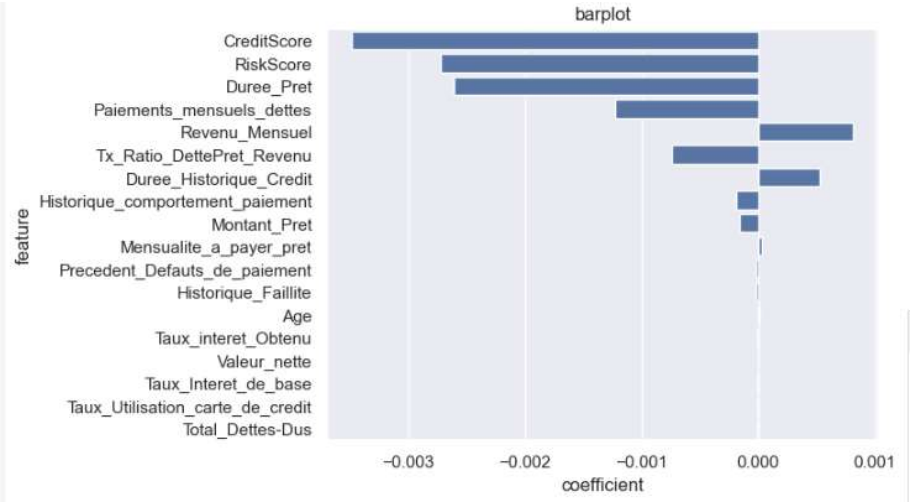
## 3-1 Évaluation du modèle

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

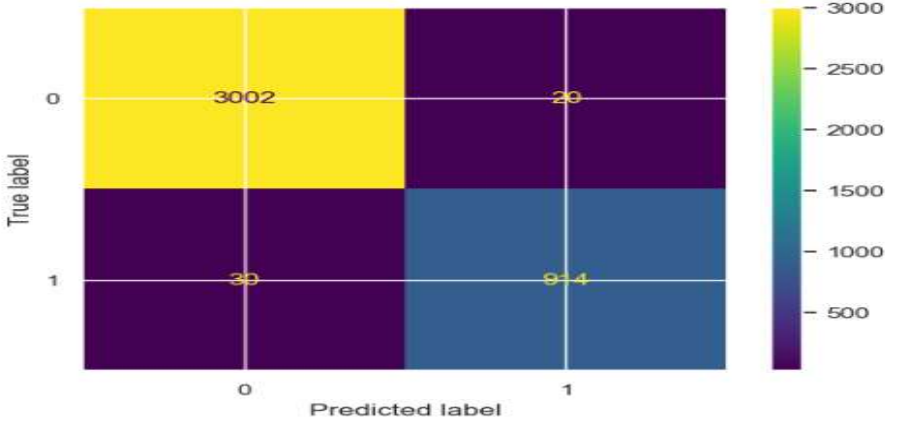
- **Précision du modèle en %** **: 98.74%**
- Nombre de prédictions positifs bien effectuées : 97.86%
- Nombre de positifs (approuvé) bien prédit : 96.82%

## 3-2 matrice de confusion

- Vrais négatifs (TN) = 3002.
- Faux positifs (FP) = 20.
- Faux négatifs (FN) = 30.
- Vrais positifs (TP) = 914.




```
Matrice de confusion :
[[3002   20]
 [  30  914]]
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x24a1
```



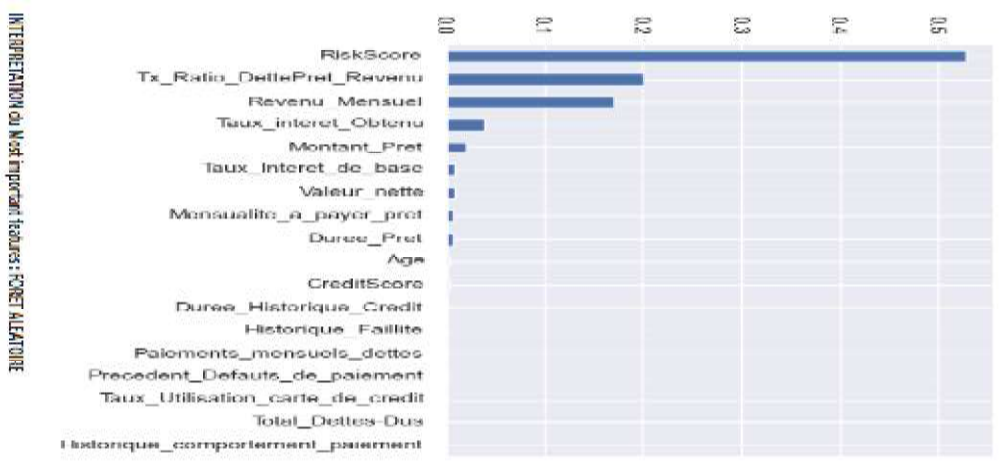
# 4- Forets Aléatoire

Nous allons utiliser un autre modèle, « la Foret Aléatoire » et évaluer sa performance a générer un modele de prédiction d'acceptation du crédit

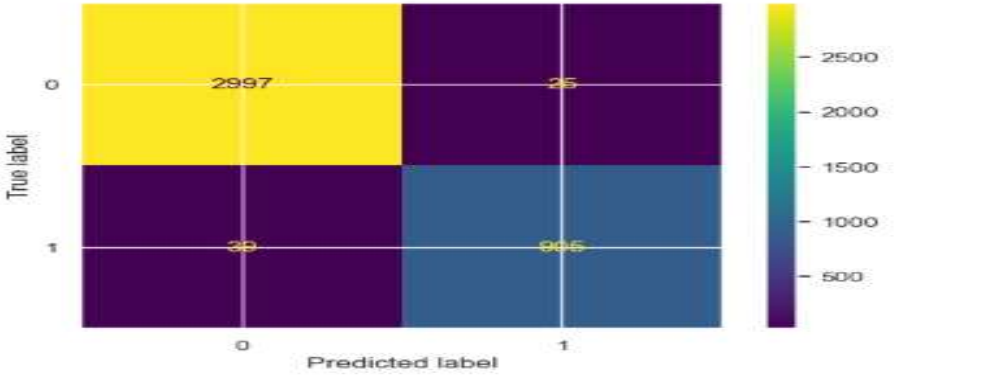
**4-1 Évaluation du modèle**

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%




```
Matrice de confusion Foret Aleatoire:
[[2997   25]
 [  39  905]]
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x1bd05292350>
```

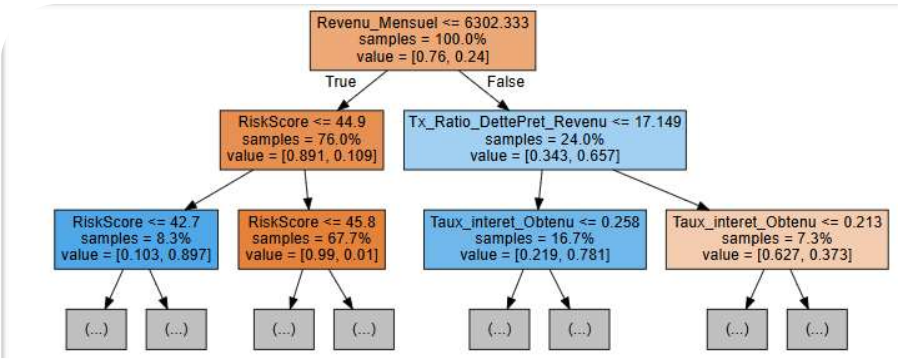


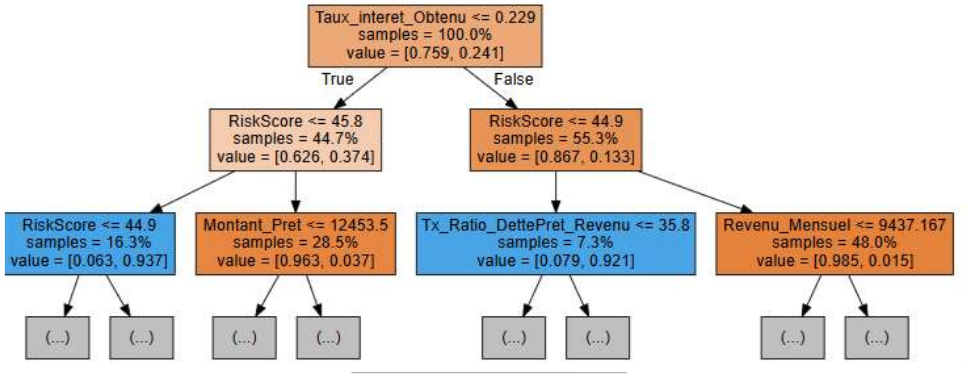

**4-2 matrice de confusion**

- Vrais négatifs (TN) = 2997.
- Faux positifs (FP) = 25.
- Faux négatifs (FN) = 39.
- Vrais positifs (TP) = 905.

# 4- Forets Aléatoire

**4-3 visualisation de l'arbre a 2 niveaux**



# 5- Comparaison Régression Logistique / Forets Aléatoire

| Most feature Regréssion 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).

# 6- Conclusion

DataSuits

À la suite de l'analyse Python, les deux modèles prédictifs présentent des critères impactant différents.

Concernant la « **Régression Logistique** », ses 3 premiers critères impactant sont ***CreditScore***, ***RiskScore*** et la ***durée_prêt***. Tandis que la « **Foret Aléatoire** », ses 3 premiers critères impactant sont ***RiskScore, Tx_Ratio_DettePret_Revenu*** et ***Revenu_Mensuel***.

La « **Régression Logistique** » privilégie les features qui sont des indicateurs calculés par la banque, alors que la « Foret Aléatoire » privilégie les informations du demandeur (revenu et dettes).

La précision générale (*accuracy*), est élevé pour les deux modèles, plus de 98%, « la **Régression Logistique** » a le taux le plus élevé de réussite globale (98,74% contre 98,39%).

Concernant l'efficacité à trouver les vrais positifs (*precision*), %, « la **Régression Logistique** » (97,86%) a un taux supérieur à celui de la « **Forets Aléatoire** » (97,31%).

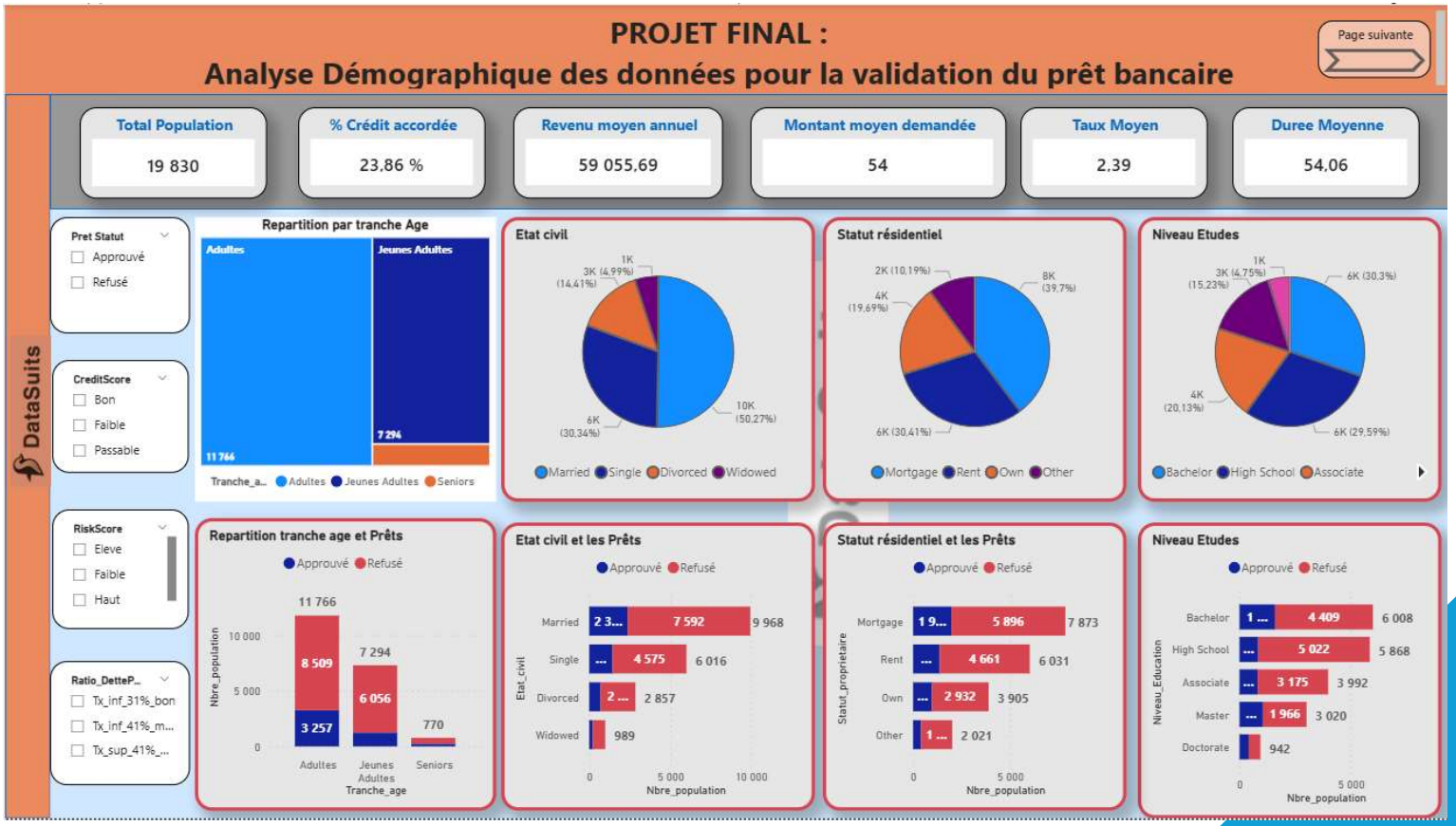
Finalement, la « **Régression Logistique** » est plus adaptée pour prédire au mieux un profil client peut voir sa demande de crédit acceptée ou non. Elle se passe en priorité sur les indicateurs établis par la banque a partir des données du demandeur.

La « **Foret Aléatoire** » utilise au mieux les informations usuels du client pour la prédiction, malgré son taux de réussite a plus de 98%, elle est moins efficace.

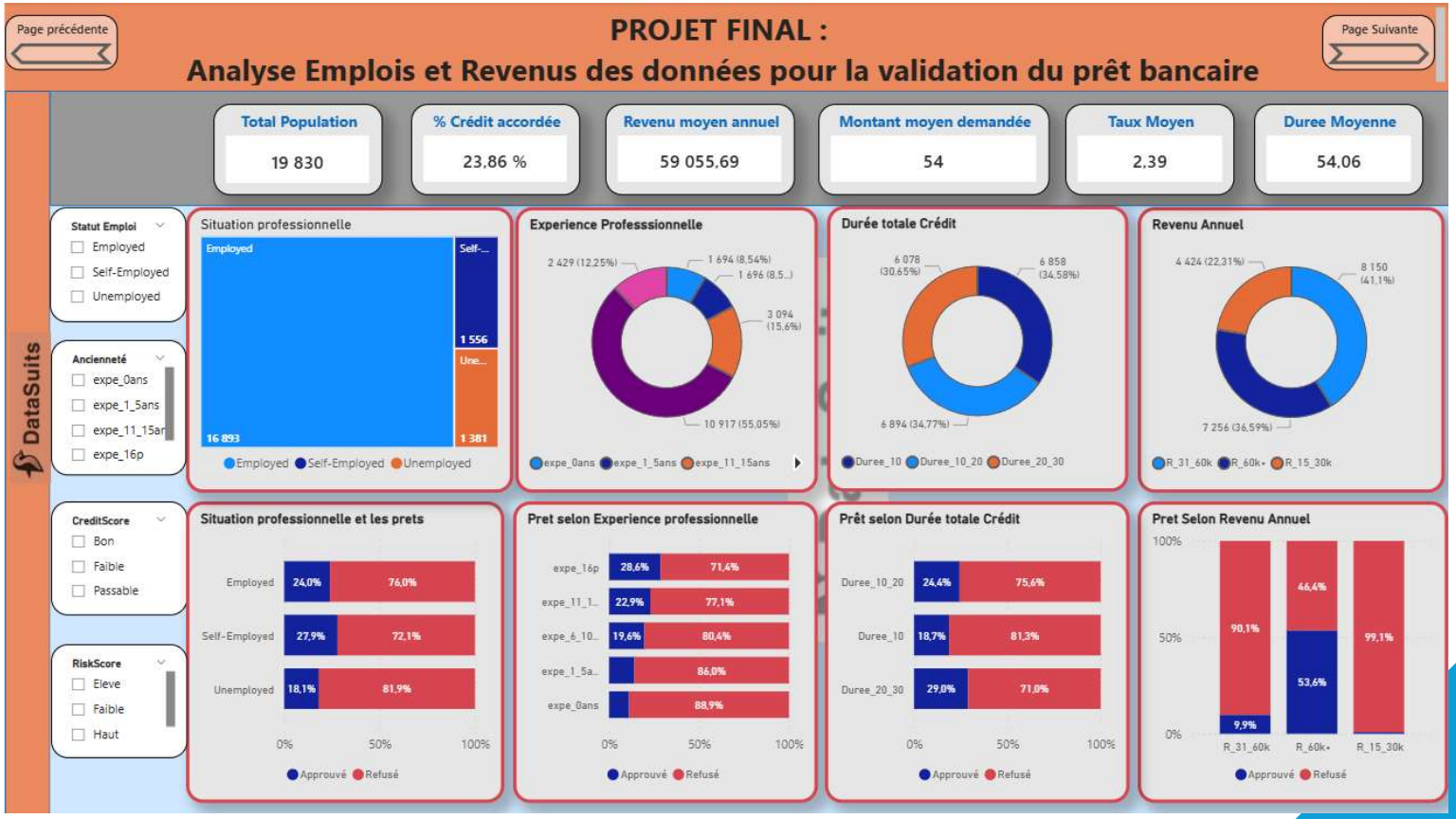
**Nb**: Les documents Notebook Python et PowerBy sont consultables via ce lien : GitHub - Rogermamaty/projet-final: Projet final - Formation Data Analyst

# 7- Annexe : Présentation des données sous PowerPoint

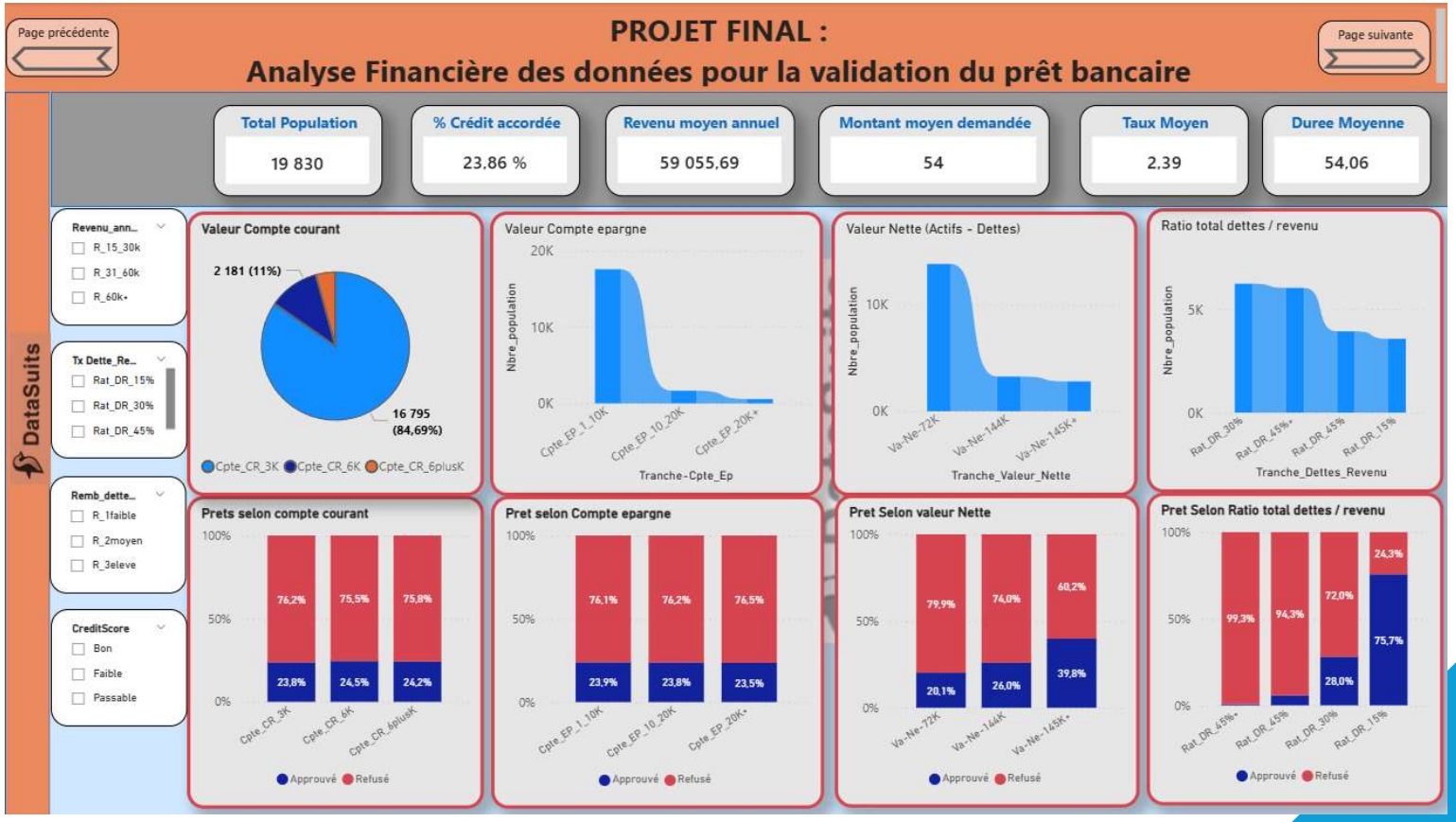
## 7-1 Analyse Démographique



## 7-2 Analyse Emplois etRevenus



## 7-3 Analyse Financière



## 7-4 Analyse Financière