REPUBLIQUE DU SENEGAL

UN PEUPLE-UN BUT-UNE FOI

Une image contenant texte, clipart

Description générée automatiquement

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Institut Supérieur de Management « ISM »

Logo

Description automatically generated with low confidence

Rapport

**THÈME : Datamining**

**Prédictions**

**Pour une**

**Banque Américaine**

Présenté par : Sous la direction de

Axel DIMA GANMO Mr Diouf Mbachir

ANNEE ACADEMIQUE : 2021-2022

Sommaire

# 1. Traitement d’apurement préliminaire ou Pré traitement de la base de données

# 2. Analyse descriptive de certaines variables clés

# 3. Choix du modèle et modélisation du phénomène

# 4. Recommandations

### 1. Pré traitement des données

**Contexte**

Les dépôts à terme sont une source majeure de revenus pour une banque. Un dépôt à terme est un placement en espèces détenu auprès d'une institution financière. Votre argent est investi à un taux d'intérêt convenu sur une durée ou une durée déterminée. La banque a divers plans de sensibilisation pour vendre des dépôts à terme à ses clients.

Il est crucial d'identifier au préalable les clients les plus susceptibles de convertir afin de pouvoir les cibler spécifiquement pour probablement leur accorder des prêts

L’objectif de ce projet est de prédire pour une banque américaine les bons et les mauvais Clients... (variable y).

**Objectif :**

- Comprendre du mieux possible nos données (un petit pas en avant vaut mieux qu'un grand pas en arrière)

- Développer une première stratégie de modélisation

- Rendre nos données propices à l’élaboration d’un modèle de Machine Learning

### Checklist de base

**Analyse de Forme** :

- **variable Target** : y

- **lignes et colonnes** :

- **types de variables** : qualitatives, quantitatives

- **Analyse des valeurs manquantes** : Aucune valeur d’attributs manquantes

**Analyse de Fond :**

- **Visualisation de la Target** :

- 10% de positifs (558 / 5000)

- **Signification des variables** :

1 - âge (numérique)

2 - emploi : type d'emploi (catégoriel : "admin.","inconnu","chômeur","gestionnaire","femme de ménage","entrepreneur","étudiant", "col bleu","indépendant","retraité","technicien","services")

3 - matrimonial : état civil (catégoriel : "marié","divorcé","célibataire", "célibataire" ;

4 - éducation (catégoriel : "inconnu","secondaire","primaire","tertiaire","illettré",)

5 - défaut : à crédit en défaut ? (Binaire : "oui", "non")

6 - solde : solde annuel moyen,

7 - logement : a prêt logement ? (Binaire : "oui", "non")

8 - prêt : a prêt personnel ? (Binaire : "oui", "non")

9 - contact : (catégoriel : "inconnu", "téléphone", "cellulaire")  
10 - jour : dernier jour de contact du mois (numérique)  
11 - mois : mois du dernier contact de l'année (catégoriel : "jan", "feb", "mar", …, "nov", "dec")  
12 - duration : durée du dernier contact, en secondes (numérique)

# autres attributs :  
13 - campagne : nombre de contacts effectués durant cette campagne et pour cette client (numérique, inclut le dernier contact)

14 - pdays : nombre de jours qui se sont écoulés depuis que le client a été contacté pour la dernière fois lors d'une campagne précédente (numérique, -1 signifie que le client n'a pas été contacté auparavant)

15 - précédent : nombre de contacts effectués avant cette campagne et pour ce client (numérique)

16 - résultat final : résultat de la campagne marketing précédente (catégoriel : "inconnu", "autre", "échec", "succès")

**Variable de sortie (cible souhaitée) :**  
17 - y - le client est-il un bon client ou pas ? (Binaire : "oui", "non")

### Sélectionner les features appropriées.

Selon l’analyse précédente, les caractéristiques suivantes sont sélectionnées pour mettre en œuvre un modèle de régression logistique :

**Categorical features: ["job","marital","education","default","housing", "loan", "contact", "poutcome", "y"]**

**Numerical features: ["age", "balance", "duration", "campaign"]**

3.1. Pre-processing

3.2. Removing noise

3.3. Transforming – Encoding

### 3.1. Preprocessing / Pré-traitement :

Cela consiste donc à combiner le train et les ensembles de test pour maintenir la cohérence entre eux. En outre, parce que je dois encoder des variables catégoriques.

### 3.2. Suppression du bruit :

Les caractéristiques suivantes seront supprimées parce qu’elles ne représentent pas un impact sur l’ensemble de données (vous trouverez ci-dessus l’explication respective de chacun des éléments énumérés ci-dessous) :

**\* a. Remplacer l’attribut "unknown".**

**\* b. Drop features: ["month", "previous", "day", "pdays"].**

### 3.3 Transformation - Encodage :

**Codage des caractéristiques catégorielles : coder les variables catégorielles en 1 et 0 (oui et non deviennent 1 et 0) etc.**

### 2. Analyse des variables clés

### a. features job

### La banque a beaucoup plus contacté les personnes avec des profils professionnels. La plupart des preneurs ont une qualification élevée par rapport à d’autres.

### b. features mariage

### La banque s’intéresse davantage aux personnes mariées et célibataires qu’aux divorcés. Les trois variables sont présentées par ordre décroissant. Relation directe des échantillons avec la colonne cible. Les échantillons les plus "mariés" signifiaient meilleurs clients.

### c. features éducation

### Plus de personnes ayant un diplôme d’études supérieures sont considérés comme bons clients. Relation proportionnelle.

### d. Fonction par défaut

### Une forte proportion de non défaillants correspond aux bons clients. Il semble logique que les gens qui ont du crédit ne peuvent pas bénéficier d’un prêt bancaire (mauvais clients).

### e. Target Y

### Un grand nombre de personnes contactées sont considérés comme mauvais clients. Seulement les 13,2% sont bons. Cette différence énorme montre un problème de classe déséquilibré. Cette question sera traitée en utilisant le suréchantillonnage.

### f. Features âge

### Cette fonction montre la relation entre "âge" et la variable cible catégorique en utilisant une représentation visuelle de la boîte. En outre, l’histogramme affiche une image en forme de cloche, présente ci-dessous, qui a une distribution normale décalée à gauche. Le groupe de population se situe entre 20 et 60 ans. La case indique un groupe d’âge ciblé précis entre 30 et 50 ans. Probablement, parce que c’est quand les gens sont plus productifs et stables. Cette tendance devient plus forte si l’on tient compte de la caractéristique de l’emploi. L’analyse bivariée est excellente parce qu’elle montre les clients qu’il doit cibler un segment précis de clients.

### d. duration

### Cette caractéristique a une influence sur le résultat "y". On peut voir que lorsque la durée est comprise entre 0 et les 2 premières minutes la plupart des personnes sont de mauvais clients. Les échantillons restants prennent une décision dans une gamme de plus de 2 min jusqu’à 12 min.

### 3. Modélisation du phénomène

**Choix du modèle** : Un modèle de régression logistique doit être mis en œuvre afin d’effectuer la classification binaire de chaque instance, à savoir si le client est bon ou pas.

**Evaluation du modèle :** Le classificateur qui en résulte doit être évalué, avec « exactitude » comme mesure de notation.

Comme nous pouvons le voir ci-dessus, 92 % des précisions étaient exactes, mais le recall ne présente que 30 % des cas positifs correctement identifiés et 42 % des prédictions positives. Dans cet esprit, nous confirmons une fois de plus le problème de déséquilibre lié ci-dessus.

### 4. Recommandations

**Recommandations**: Ainsi, la technique de suréchantillonnage est mise en œuvre ci-dessous pour résoudre ce problème, aussi essayer de normaliser les données avec les transformers comme RobustScaler ou StandartScaler pour des résultats plus optimaux