**Note EDA**

**Analyse Exploratoire des données**

**Objectif :**

* Comprendre du mieux possible nos données
* Développer une première stratégie de modélisation

#### **Analyse de la forme**

* **variable target** : Attrition\_Flag
* **lignes et colonnes** : 10127, 21
* **Pas de colonnes dupliquées ni de lignes dupliquées**
* **types de variables**: catégoriques : 6, numériques : 15
* **Valeurs manquantes**: pas de valeurs manquantes.

**Analyse de fond**

* On a deux classes : **'Existing Customer', 'Attrited Customer'**
* 84% (8500) clients maintiennent le service de carte bancaire
* 16% (1627) clients se sont désabonnés du service de carte bancaire
* Dataset déséquilibré
* **Signification des variables**
* Variables continues suivant une distribution normale pour certaines (symétriques) et d’autres asymétriques (right-skewed);
* Certaines variables numériques contiennent des valeurs abérrantes (outliers): Customer\_Age, Months\_on\_book, Months\_Inactive\_12\_mon, Contacts\_count\_12\_mon, Credit\_Limit, Avg\_open\_To\_Buy, Total\_Amt\_Chng\_Q1\_Q4, Total\_Trans\_Amt, Total\_Trans\_Ct, Total\_Ct\_Chng\_Q1\_Q4,
* Variables qualitatives multi-catégoriques et binaires. Les variables catégoriques n’ont pas une grande cardinalité (pas beaucoup de valeurs uniques)
* Il y a des labels qui sont très rares dans certaines variables (moins de 5% souvent)
* **Relation Variables / Target :**
* Target/variables numériques: les variables Total\_Revolving\_Bal, Total\_Trans\_Amt, Total\_Trans\_Ct, Total\_Ct\_Q4\_Q5 semblent être liées aux raisons de désabonnement.

**Analyse un peu avancée**

* Il y a une forte correlation entre certaines variables numériques don’t les plus fortes valeurs sont 0.5 et 1.
* Il n’existe pas de relations linéaires entre toutes les variables numériques.
* Il n'existe à priori pas de relation entre l'age des clients et leurs catégories de salaire, leur statut matrimonial et leur niveau éducatif

**Model Building**

**Modèles entraînés avec la Cross Validation**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | RForest | SVM | KNN | AdaBoost | GrdtBost | NB |
| Precision (Précision) | .97/.91  .94/.96 | .96/.83  .89/.94 | 94/80  .87/.92 | .97/.86  .91/.95 | .97/.92  .94/.96 | .93/.61  .77/.88 |
| Recall (Sensibilité) | .98/.84  .91/.96 | .97/.75  .86/.94 | .97/.68  .82/.93 | .98/.82  90/.95 | .99/.84  .91.96 | .93/.62  .77/.88 |
| Score F1 (Compromis  Recall-Precision | .98/.87  .92.96 | 96./.79  .88/.92 | .96/73  .85/.92 | .97/.84  .90/.95 | .98/.88  .93/.96 | .93/.62  .77/.88 |
| Overfitted | oui | Non | Non | Non | Non | Bizarroîde |

**Modèles Optimisés**

|  |  |  |
| --- | --- | --- |
|  | RForest (RandomiezdSeaech) | SVM  (threshold =0.5) |
| Precision (Précision) | .97/.89  .93/.96 | .72 |
| Recall (Sensibilité) | .98/.85  .91/.96 | .84 |
| Score F1 (Compromis  Recall-Precision | .98/87  .92/96 | .78 |



