REPUBLIQUE DE COTE D'IVOIRE
Union – Discipline – Travail

Institut National Polytechnique Félix Houphouët-Boigny

REPUBLIC OF IVORY COAST
Union – Discipline – Work

National Polytechnic Institute Félix Houphouet-Boigny

Master Data Science – Big Data

2024-2025

Projet SPARK

Olaf Kouamo

Les compteurs intelligents d'eau fournissent des données de consommation quotidiennes. En cas de blocage du compteur, on constate l'absence d'évolution de l'index donc une consommation nulle.

Cependant, les consommations nulles peuvent aussi être le fait d'une absence du client, plus ou moins longue. Dans ce cas on dira que le compteur est « arrêté » et non "bloqué". Pour une résidence principale, les périodes sans consommation sont généralement courtes (moins d'un mois), pour une résidence secondaire, ces périodes peuvent être longues.

Aussi, la caractérisation de la consommation avant la dernière consommation nulle est traduite en fonction de plusieurs variables qui sont autant d'entrées de l'algorithme. Outre les données de consommation, d'autres variables peuvent également être utilisées si elles sont disponibles.

Les données métrologiques : Il apparaît que certains modèles et années de fabrication des compteurs sont plus sujet à des blocages que d'autres. On peut donc introduire une variable supplémentaire qui est la probabilité pour un (modèle, année de fabrication) d'être bloqué.

Données sur les clients : La base de données clients peut comporter des variables qui indiquent le mode d'occupation du client. En particulier, on peut utiliser le champ « **logement vacant** », qui permet par exemple de justifier l'absence de consommation dans ce logement.

Plusieurs variables différentes ont déjà été testées. Les variables présentées ci-dessous (nommées v1, v2, ... v9), sont celles qui ont un intérêt pour la différenciation entre compteur « arrêté » et compteur « bloqué ».

Variable v1 : Temps écoulé sans consommation.

L'analyse est nécessairement effectuée à une date donnée. Lorsque le compteur est bloqué ou arrêté, un certain temps s'est écoulé. La durée sans consommation est un facteur très important pour détecter les compteurs bloqués.

En effet, plus la durée est importante, plus le blocage est probable. La variable v1 est donc le nombre de jours depuis la dernière consommation. Parfois, le blocage n'est pas simple, et nous avons des consommations très faibles (quelques litres) avant un blocage définitif. Nous proposons donc la règle suivante pour positionner le dernier jour avec la consommation :

v1 = Différence entre le jour de l'analyse et le dernier jour de consommation significative (plus de 20 litres pendant 2 jours consécutifs).

Définition : Nous notons To le dernier jour de consommation selon la définition ci-dessus et T le jour courant. On a alors v1 = T - To

Cas d'exclusion : La valeur de To peut être indéterminée si l'index est resté fixe depuis la pose de l'émetteur. On peut prendre une convention pour fixer To à une valeur maximale (2 ans par exemple).

Dans ce cas, certaines variables ci-dessous ne peuvent pas être calculées (v2, v3, v4, v5, v6, v8, v11, v12), et l'algorithme ne peut pas être appliqué.

Variable v2 : Taux de consommation non nul sur les 90 jours précédant le To. Cette variable décrit le pourcentage de jours avec une consommation sur les 90 jours avant To. Si certaines sont manquantes, le taux est calculé sur les consommations disponibles. Il se peut qu'il n'y ait pas de données avant To.

Variable v3 : Taux de consommation non nul sur 6 mois (183 jours) avant le T_o - 90 jours. Cette variable décrit le pourcentage de jours avec une consommation sur les 183 jours avant To - 90 jours. Si certaines consommations sont manquantes, on calcule le taux sur les consommations disponibles. Cette variable permet de caractériser le présentéisme sur une année complète, et donc de distinguer assez facilement les résidences secondaires. L'analyse est effectuée en neutralisant la période précédente de 90 jours qui peut correspondre à un fonctionnement dégradé du compteur. Il apparaît clairement que les données concernent la période comprise entre To - 273 jours et To – 90 jours. Si certaines consommations sont manquantes, nous calculons le taux sur les valeurs disponibles. Si nous n'avons pas de valeur, nous prenons le même taux que la variable précédente (v2)

Variable v4 : durée moyenne des périodes de consommation nulle.

Sur une période de 1 ans avant le To . Chaque série de données est identifiée lorsque la consommation est nulle. Une série est définie par une consommation nulle et sa longueur est la différence de dates entre le dernier indice strictement inférieur et le premier indice strictement supérieur.

En cas d'indices manquants, on considère qu'il n'y a pas de consommation nulle entre deux indices strictement croissants. Inversement, s'il y a des indices manquants mais que les indices sont identiques, les consommations sont considérées comme nulles entre les 2 bornes.

Les consommations négatives sont considérées comme non nulles. S'il n'y a pas de période de consommation nulle, la longueur retournée sera 0.

Variable v5 : Durée maximale des périodes de consommation nulles. Parmi les périodes de consommation zéro définies pour la variable v4, il s'agit d'identifier la plus longue durée de consommation zéro. La plus longue durée de consommation nulle. S'il n'existe pas de période de consommation nulle, on prend la valeur 0.

Variable v6 : longueur maximale de consommation non nulle. Contrairement à la variable précédente, nous calculons la longueur maximale de la série de consommation non nulle. En cas de manque de données, on suppose que la consommation est non nulle si l'indice est strictement croissant.

Variable v7 : Dernier indice de consommation. Cette variable est l'index mécanique du compteur sur le To. L'intérêt commercial de cette variable est que les compteurs qui ont peu travaillé sont moins souvent bloqués que les compteurs qui ont subi beaucoup de passage.

Variable v8 : Nombre de périodes de consommation. Avec la même définition que la variable v6, les périodes de consommation sont comptabilisées sur une période de 1 an avant To.

Variable v9 : Dans le dataset training, le champ Millesime désigne les 5 premiers caractères du sensor. La variable v9 est définie comme suit.

- Si le millesine appartient à ('C10FA','C10LA','C10SA', 'C11FA','C11LA','C11SA') et string(v7)>3 alors v9=0
- Si millesime (‘D16BU’, ‘Z12ER’, ‘C07AA’) alors v9=1 sinon v9=2

Questions : Vous avez recu deux datasets le dataset training_data_csv qui represente le dataset pour la modélisation avec les valriables v1 à v9 déjà calculées et le dataset dataset_sample_spark qui représente les consommations et index des compteurs.

1. Considérons le ***Dataset training_data*.csv**.
    a. Calculer la variable v9
    b. On considére que dans le dataset training, la variable PDC représente une pseudonimisation de la triplette (sensor, servicePoint, transmitter). On souhaite prédire si un compteur est bloqué ou stoppé en fonction des différentes variables v1 à v9.
    c. Quelle est la proportion de compteur bloqué/stoppé par typologie de localite
    d. Modéliser à l’aide d’un modèle de machine learning la prediction de compteur bloqué.
        i. On testera plusieurs modèles de classification et on gardera celui pour lequel les performances seront meilleurs sur l’ensemble test.
        ii. Préciser le modèle choisi et les critères d’évaluation
2. Considérons à présent le ***dataset sample_spark*.csv**. L’objectif est de calculer les différentes features v1 à v9 afin de pouvoir appliquer le modèle sur ce nouveau dataset et calculer ainsi la probabilité pour un compteur d’être bloqué ou stoppé. La date T correspond à la date maximale pour laquelle on a des observations dans le dataset. Par

exemple pour la triplette [C04AE134021, 984588253128, C01E00D9] la date T est le 08-08-2024.

a. Rassurez-vous que la triplette (sensor, servicePoint, transmitter) qui définit un compteur est unique pour une date donnée. Si ce n'était pas le cas, supprimez les doublons.
b. Quelle est la consommation moyenne par diametre de compteur ? En déduire quels sont les gros compteurs (compteurs qui consomment le plus).
c. Calculer les variables v1 à v9 comme décrites ci-dessous
d. Appliquer ainsi le modèle obtenu précédemment et calculez la probabilité pour chaque compteur d'être bloqué