TeleCom+ - Prédiction de Churn Client (Scikit-learn vs Spark Mllib)

Synthèse exécutive

Objectif: prédire le churn (départ client) afin de cibler des actions de rétention.

Le dataset contient 7043 clients et un taux de churn de 14.39% (classe déséquilibrée).

Nous comparons trois modèles supervisés sous Scikit-learn et proposons une transposition Spark MLlib pour le passage à l'échelle.

Données et préparation

Variables: 5 numériques (SeniorCitizen, tenure, InternetCharges, MonthlyCharges, TotalCharges) et 13 catégorielles.

Pré-traitements: One-hot encoding (catégorielles) + StandardScaler (numériques), via un pipeline (ColumnTransformer) pour éviter les fuites de données.

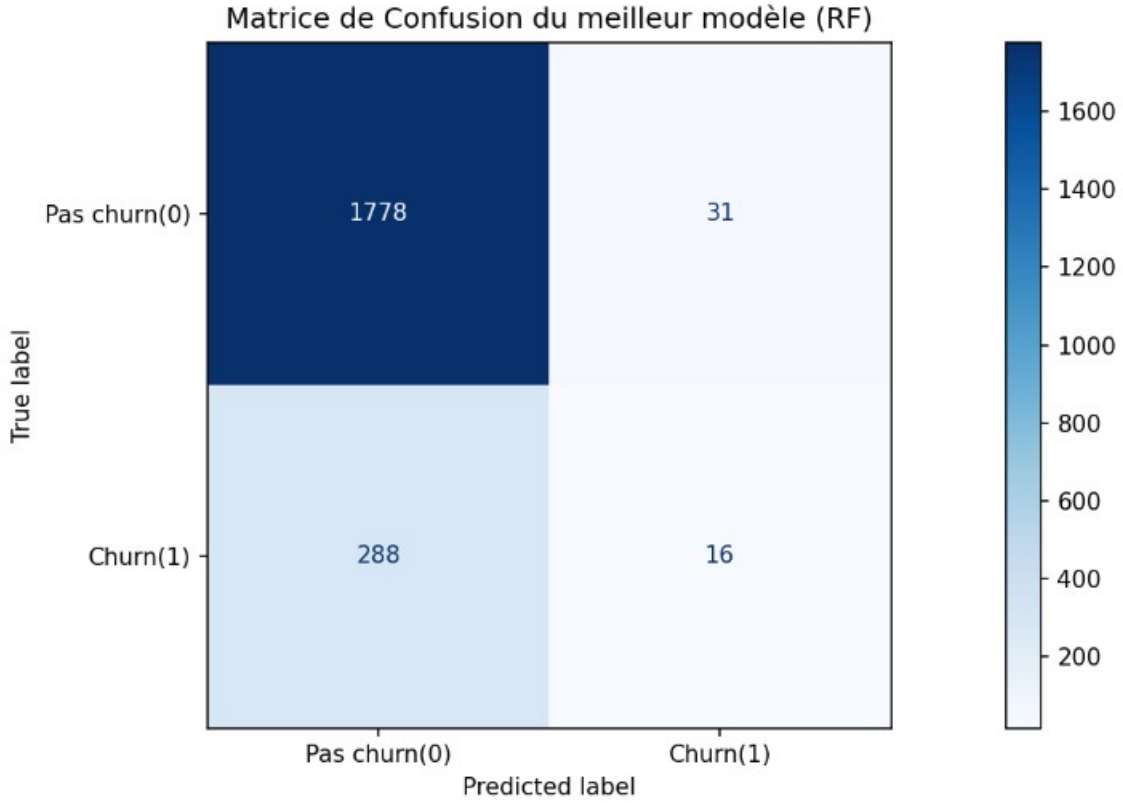
Résultats - comparaison des modèles (test 30%)

Modèle	Time (sec)	Accuracy	Precision	Recall	F1-score
LogisticRegression	0.088	0.857	0.667	0.013	0.026
RandomForest	0.444	0.849	0.340	**0.053**	**0.091**
GradientBoosting	0.733	0.855	0.440	0.036	0.067

/!\ : Le temps peut changer d'une exécution à une autre.

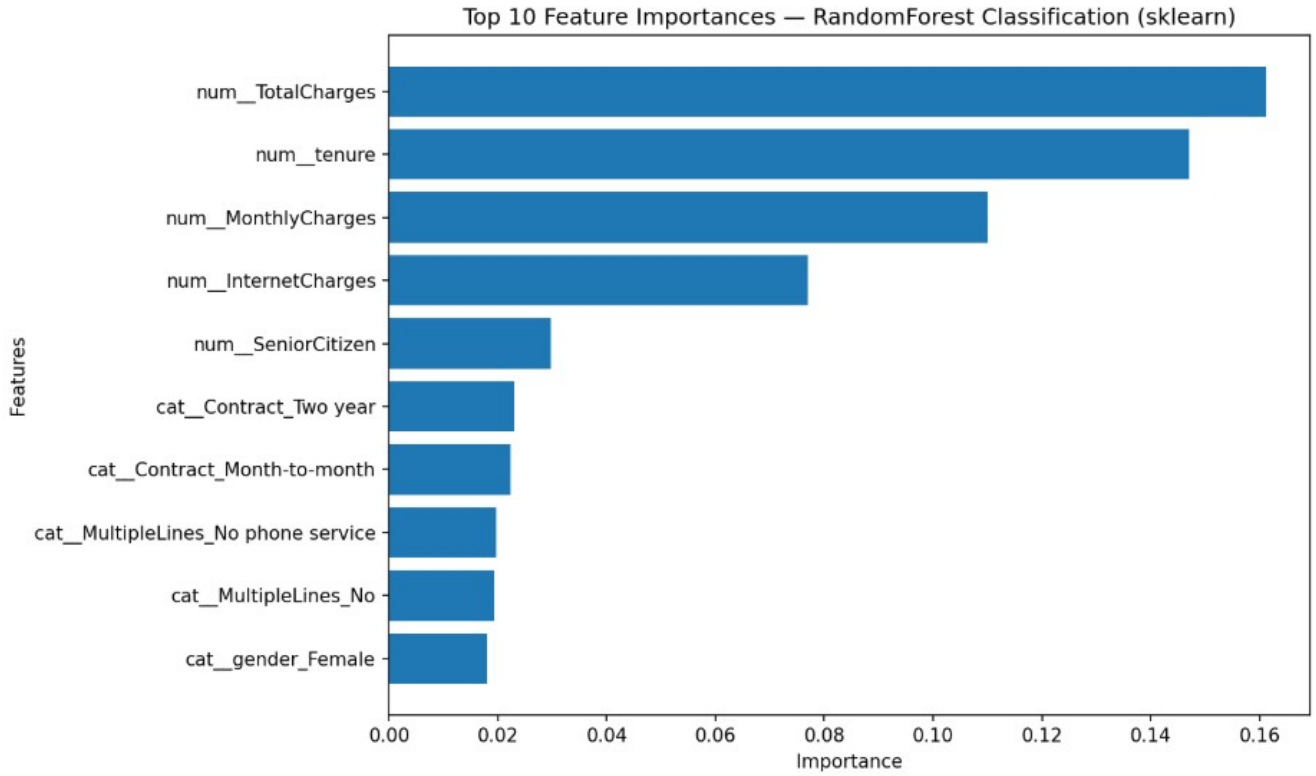
Modèle retenu: **Random Forest**. Dans un contexte churn (classe minoritaire), le rappel (Recall) et le F1-score sont prioritaires : ils maximisent la détection des clients à risque.

Matrice de confusion (modèle retenu)



=> Avant optimisation, même si que le Random Forest est le meilleur des 3 modèles, il détecte peu de clients à risque.

Variables les plus explicatives (Top 10)



=> Les **charges à payer** et **l'ancienneté** sont les **facteurs dominants** (qui influence le plus sur la prédiction)

=> Le **type de contrat** ainsi que le fait d'être **senior** ou pas influence également la prédiction d'une manière **secondaire**

=> Le **reste** des facteurs est **marginale** ou influence très peu la prédiction

Validation et optimisation (pour le meilleur modèle R.F)

- **Performances de base** : Recall : 0.053, accuracy : 0.849, precision : 0.340, f1-score : 0.091

- **Cross-validation** (5-fold, moyenne) : Recall=0.5120, F1=0.4930 , Accuracy=0.8494, Precision=0.5785.

- **GridSearch** rapide (2 candidats) confirme des performances proches du modèle de base.

	Recall	F1-score	Precision	Accuracy	Remarques
Performances de base (1)	0.05	0.09	0.34	0.85	Le modèle de base est mauvais.
Cross-validation (5-fold, moyenne)	0.5120	0.4930	0.5785	0.8494	La cross-validation montre un bon potentiel
Après optimisation (GridSearch) (1)	**0.441**	0.330	0.263	0.742	Le modèle optimisé est exploitable

Ciblage business – top clients à haut risque (p(churn) >= 80%)

Le fichier top_risk_churn_client.csv liste les clients les plus à risque selon le modèle optimisé.
Leur profil est souvent associé à :
-> Une surreprésentation des contrats **Month-to-month**
-> Une surreprésentation de la fibre optique (**Fiber optic**) comme service internet
-> Une **tenure** majoritairement faible (nouveaux clients)
-> **MonthlyCharges** élevés (prix perçu trop cher)

3 actions de rétention recommandées

- Pour les clients qui ont la fibre optique : après recueil de leurs avis (via des appels proactifs), proposer des services supplémentaires gratuitement
- Migration vers un contrat 1 ou 2 ans (en les encouragent avec une remise et/ou 1 ou quelques mois offerts)
- Pour les nouveaux clients (tenure faible) : appel proactif + résolution des éventuels problèmes de facturation, débit ...etc

Estimation ROI (hypothèses simples)

D'après les données qu'on a :

- Client perdu = 500€ de perte mensuelle
- Nombre de clients ciblés (qui risquent de churn : poba. > 80%)

Suppositions :

- On va invertir 100 € en actions pour essayer de convertir chaque client (actions comme mails, SMS, appels ...)

- On va gagner : 10% des clients pendant 3 mois de plus

Avec ces données on peut réaliser un ROI de 50.00 % ce qui est un très bon retour sur investissement.

Scikit-learn vs Spark MLlib - recommandation production

Scikit-learn: idéal jusqu'à quelques millions de lignes si la mémoire permet, itérations rapides, grande richesse d'algorithmes et de tuning.

Spark MLlib: à privilégier si les données ne tiennent plus en mémoire, si les features/joins proviennent d'un data lake, ou si l'entraînement doit être distribué.