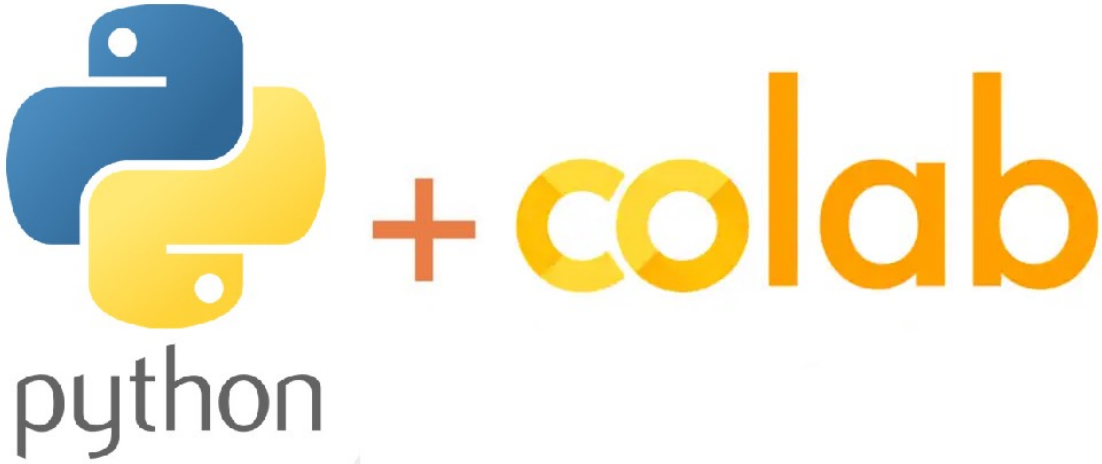
Soutenance projet Spark

Erwan MEIGNEN

Sommaire

- Présentation du projet, objectifs…
- Chaîne de traitements mise en place
- Valeur ajoutée de Spark dans cette chaîne
- Résultats/Démonstration



Présentation du projet

Choix du thème : Télécommunications

Base de données : *Telco Customer Churn (IBM)*

Objectifs : 1- Prédire quels clients vont se désabonner

Objectifs : 2- Identifier les causes de désabonnement

```
customerID,gender,SeniorCitizen,Partner,Dependents,tenure,PhoneService,MultipleLines,InternetService,OnlineSecurity,OnlineBackup,D
eviceProtection,TechSupport,StreamingTV,StreamingMovies,Contract,PaperlessBilling,PaymentMethod,MonthlyCharges,TotalCharges,Churn
7590-VHVEG,Female,0,Yes,No,1,No,No phone service,DSL,No,Yes,No,No,No,No,Month-to-month,Yes,Electronic check,29.85,29.85,No
5575-GNVDE,Male,0,No,No,34,Yes,No,DSL,Yes,No,Yes,No,No,No,One year,No,Mailed check,56.95,1889.5,No
3668-QPYBK,Male,0,No,No,2,Yes,No,DSL,Yes,Yes,No,No,No,No,Month-to-month,Yes,Mailed check,53.85,108.15,Yes
```

Aperçu du début du fichier .csv

Présentation des données

Catégorie	Variable	Description	Type attendu
Identification	customerID	Identifiant unique du client	String
Démographie	gender	Genre du client (Male/Female)	Catégoriel
Démographie	SeniorCitizen	Est-ce une personne âgée ? (1/0)	Binaire (Num)
Démographie	Partner	A-t-il un partenaire ? (Yes/No)	Binaire
Démographie	Dependents	A-t-il des personnes à charge ?	Binaire
Services	PhoneService	A-t-il un service téléphonique ?	Binaire
Services	MultipleLines	A-t-il plusieurs lignes ?	Catégoriel
Services	InternetService	Type de fournisseur internet	Catégoriel
Services	OnlineSecurity	Option sécurité en ligne	Catégoriel
Services	OnlineBackup	Option sauvegarde en ligne	Catégoriel
Services	DeviceProtection	Option protection de l'appareil	Catégoriel
Services	TechSupport	Option support technique	Catégoriel
Services	StreamingTV	Option streaming TV	Catégoriel
Services	StreamingMovies	Option streaming films	Catégoriel
Compte	tenure	Nombre de mois (ancienneté)	Numérique (Int)
Compte	Contract	Type de contrat	Catégoriel
Compte	PaperlessBilling	Facturation sans papier ?	Binaire
Compte	PaymentMethod	Moyen de paiement	Catégoriel
Compte	MonthlyCharges	Montant facturé mensuellement	Numérique (Float)
Compte	TotalCharges	Montant total facturé depuis le début	Numérique (Float)
Cible	Churn	Le client a-t-il résilié ?	Binaire (Target)

7000 lignes, peu de données mais grande qualité (21 colonnes)

Chaîne de traitements

Ingestion : Chargement CSV avec inférence de schéma

Nettoyage (Anomalie → Typage fort)

Préparation (Transformation des variables, Assemblage)

Modélisation (Comparaison approche linéaire/non linéaire)

Interprétation des résultats

Valeur ajoutée de Spark

- Scalabilité native (Scale-Invariant)
- Rigueur méthodologique (Pipelines)
- Architecture Distribuée
- Robustesse de production

Résultats/Démonstration

- Logistic Regression :

Accuracy : 80.92%

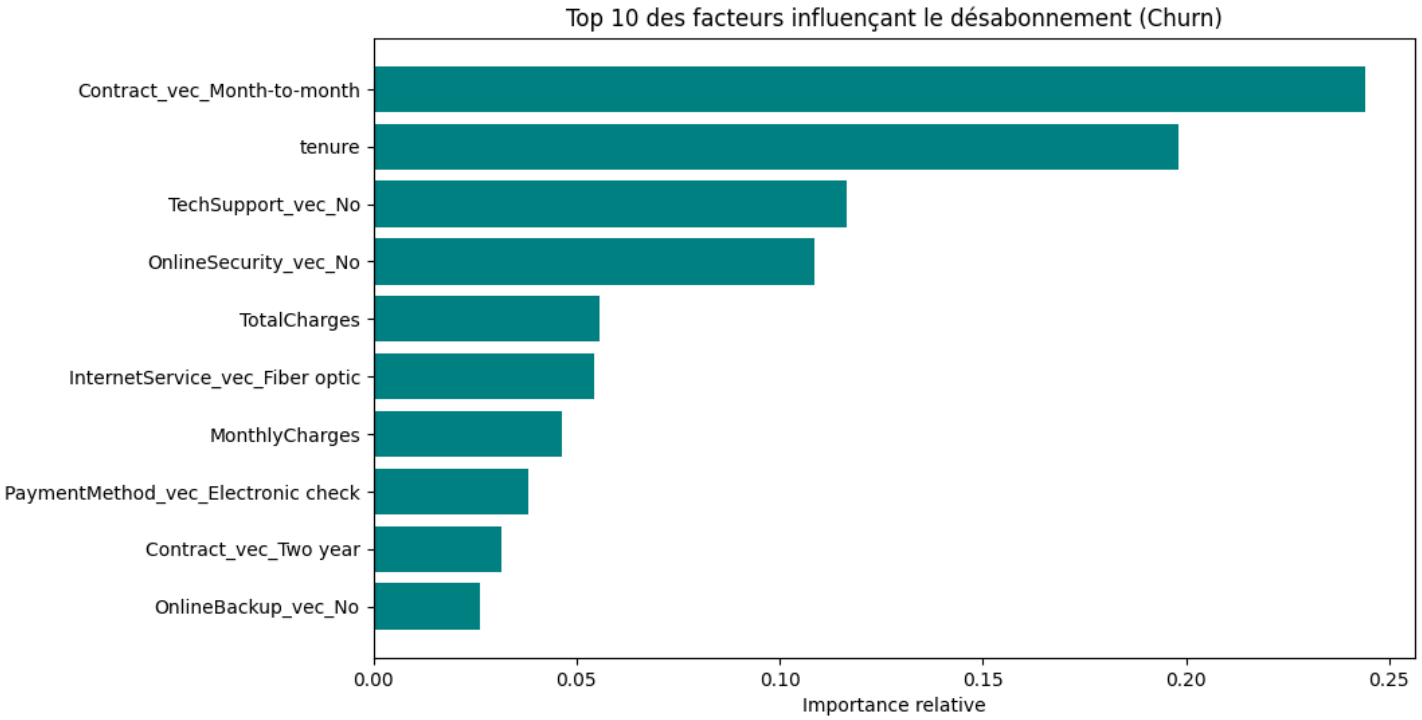
AUC (Area Under ROC) : 0.8556

- Random Forest :

Accuracy : 79.58% , AUC : 0.8497

Gain vs Logistic Regression : -0.0059

Résultats/Démonstration



Conclusion