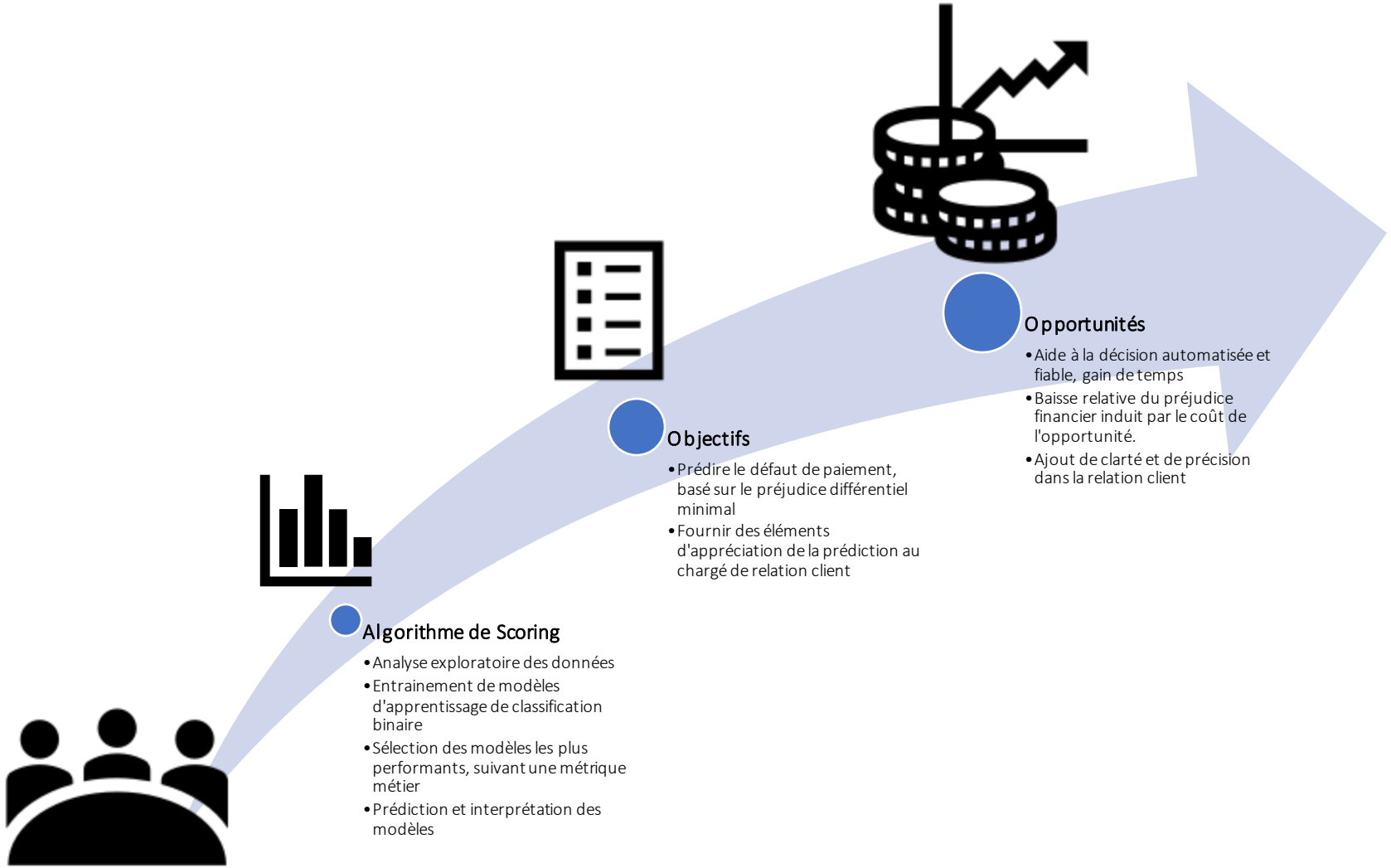
Solution de Scoring sur dossiers d'emprunt

Société "Prêt à dépenser"

Bailly DIOUNOU, Data Scientist junior – 02/12/2020

Contexte & périmètre du projet

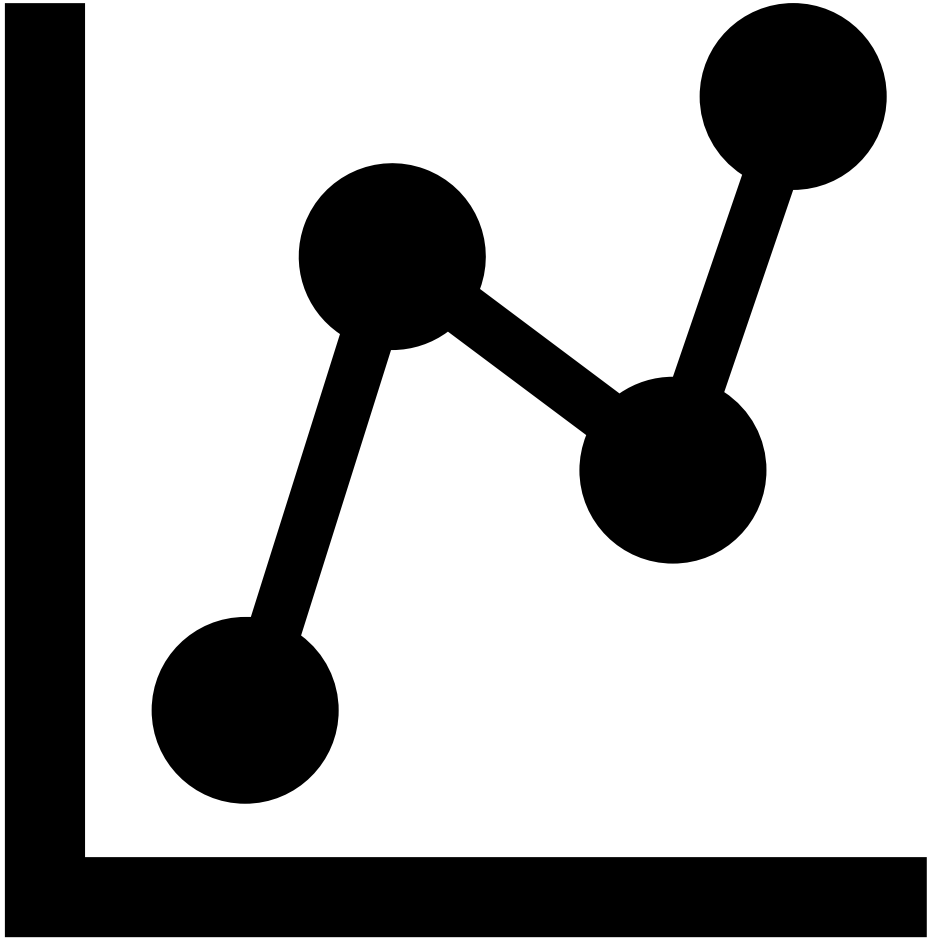


Société "Prêt à dépenser"

Présentation générale du jeu de données

Données statistiques de base - Données brutes

- Problème de classification binaire
- Ensemble de données financières de challenge Kaggle
  - Historique de prêts
  - Historique d'informations financières
  - Données sur le comportement antérieur de l'emprunteurs
- *Application_*.csv*: données sur le dossier de la candidature courante (informations financières et personnelles)
- Nombre de variables avant (encoding): 121 + 1, dont 16 catégorielles. Variable cible: *Target*
- Taille des échantillons:
  - Application_train: 307k
  - Application_test: 48.7k



Analyse exploratoire des données

Importée telle quelle du site de compétition Kaggle

Analyse Exploratoire des données

Faits marquants

Label & One Hot Encodings + Ajustement train-test

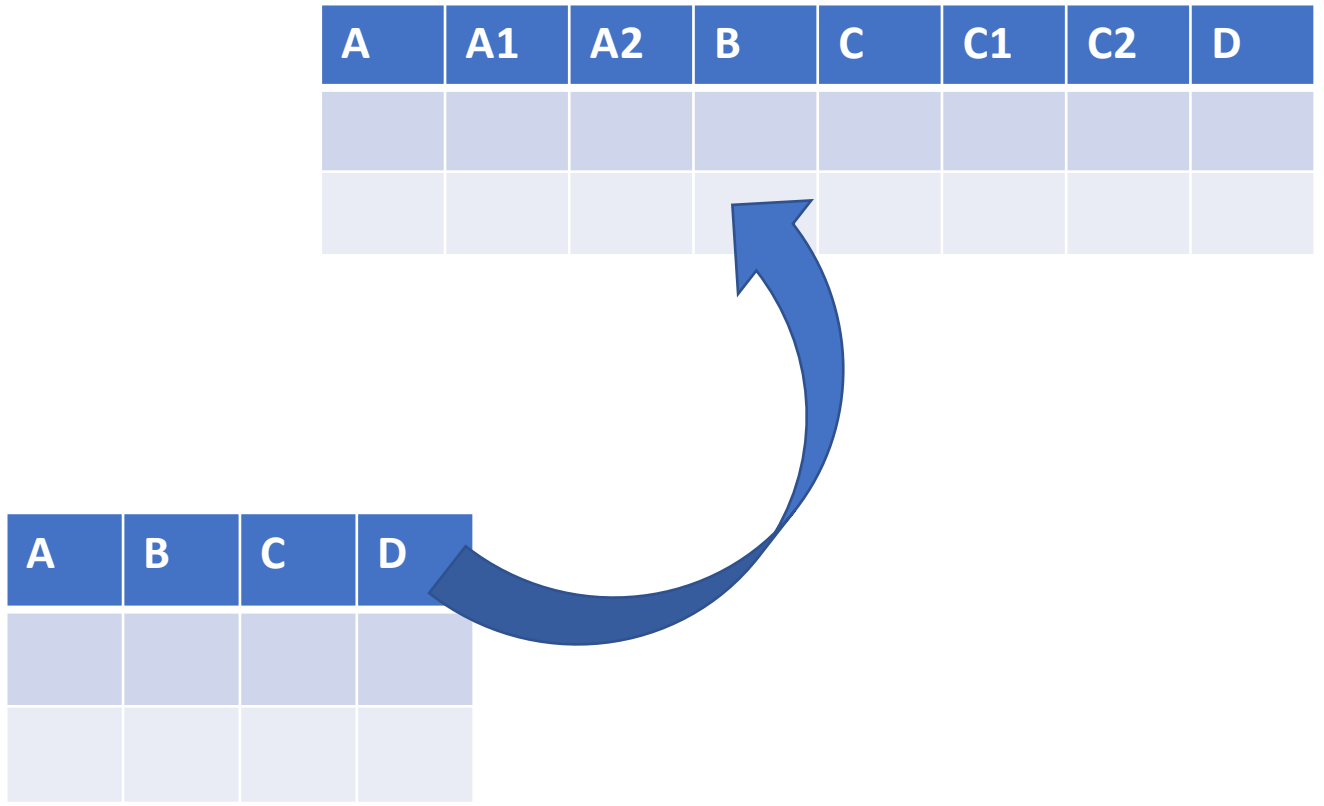
- App_train #var.: 122 -> 243 -> 240
- App_test #var. : 121 -> 239 -> 239

Corrélations à la variable cible

- *Ext_Source_** (benchmark score) : corrélation négatives plus élevées avec *Target*

Effet de l'âge

- Plus de défauts de paiement chez l'individu jeune



Analyse Exploratoire des données

Faits marquants

Label & One Hot Encodings + Ajustement train-test

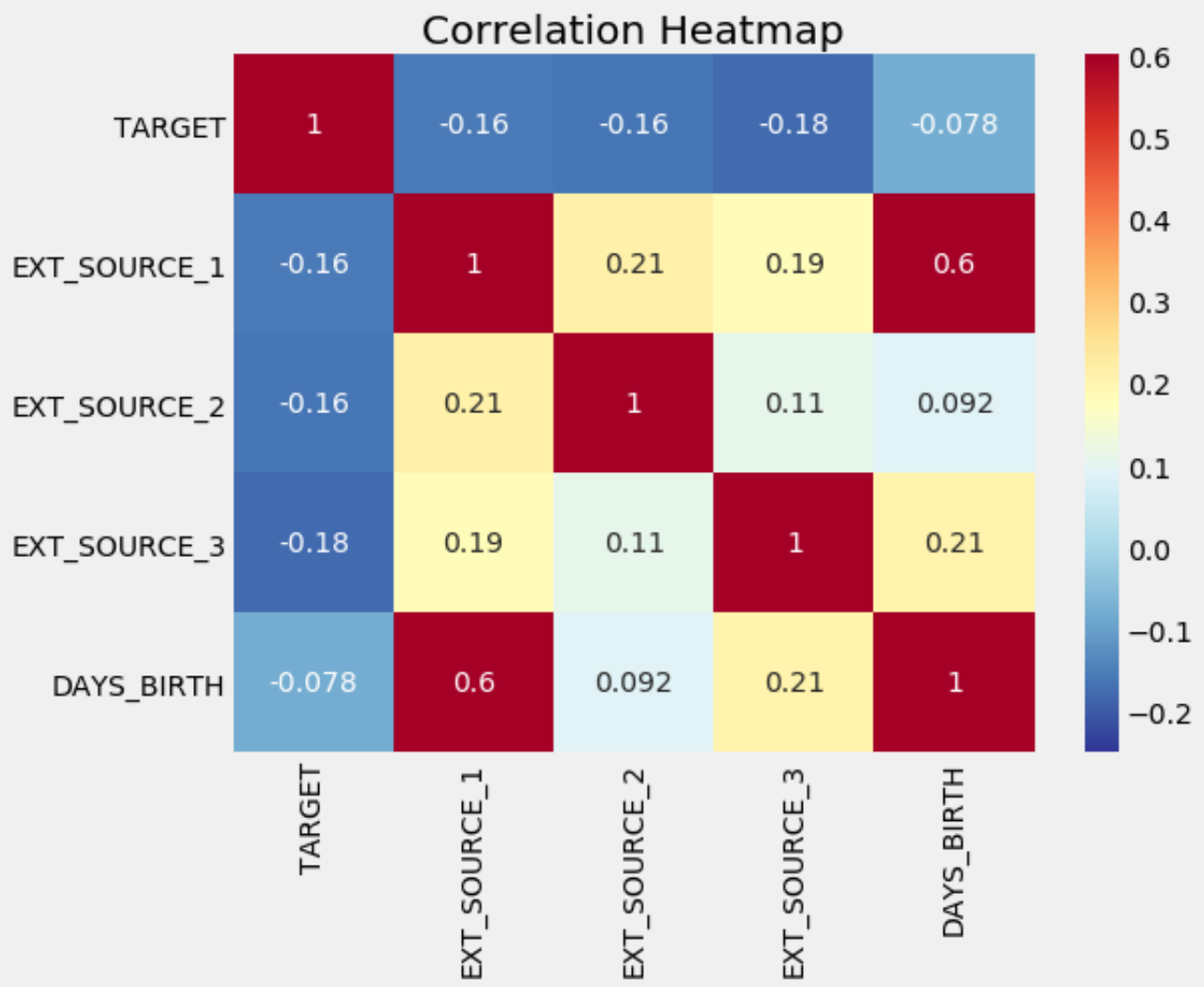
- App_train #var.: 122 -> 243 -> 240
- App_test #var. : 121 -> 239 -> 239

Corrélation à la variable cible

- *Ext_Source_** (benchmark score) : corrélation négatives plus élevées avec *Target*

Effet de l'âge

- Plus de défauts de paiement chez l'individu jeune



Analyse Exploratoire des données

Faits marquants

Label & One Hot Encodings + Ajustement train-test

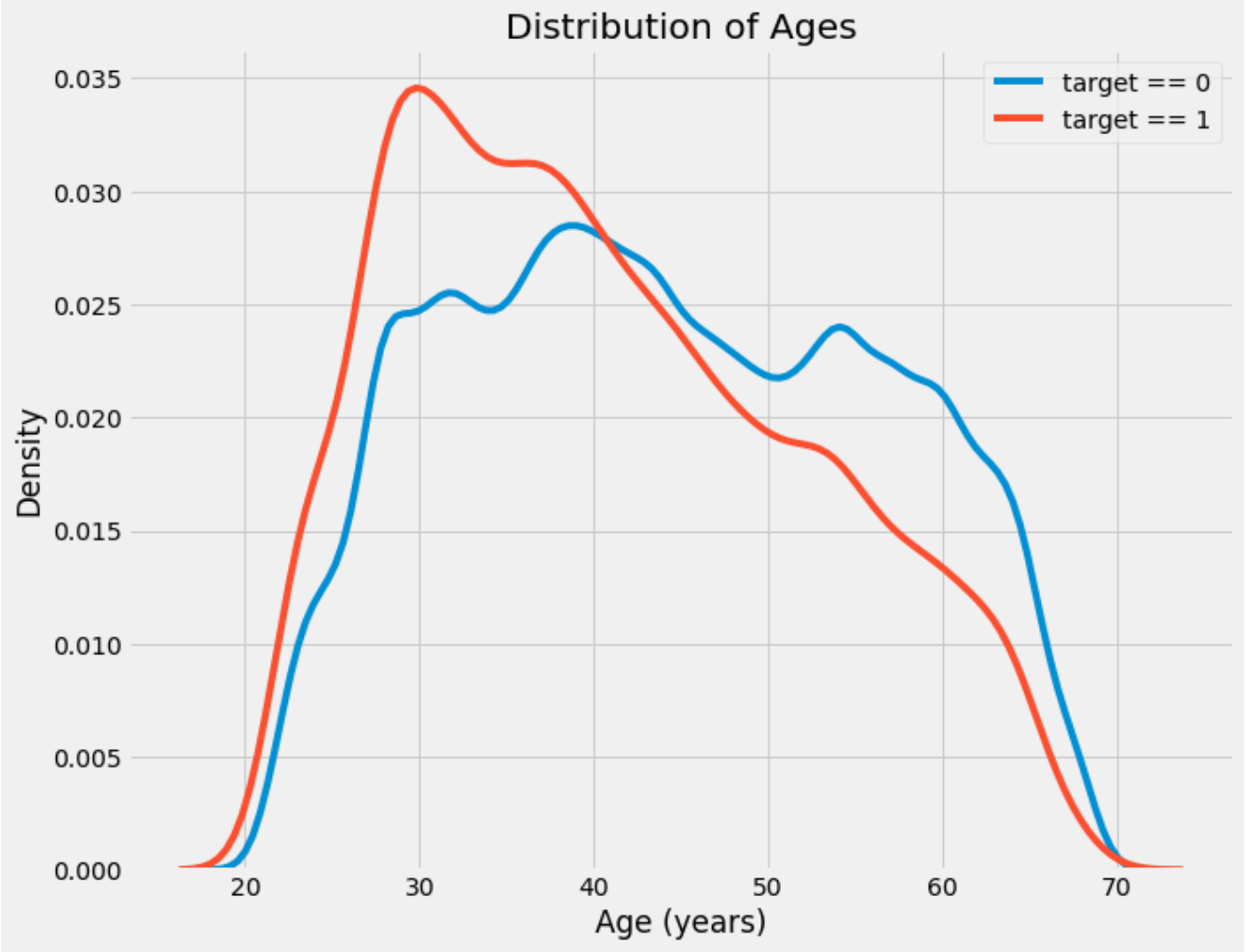
- App_train: #var.: 122 -> 243 -> 240
- App_test: #var. : 121 -> 239 -> 239

Corrélation à la variable cible

- *Ext_Source_** (benchmark score) : corrélation négatives plus élevées avec *Target*

Effet de l'âge

- Plus de défauts de paiement chez l'individu jeune



Apprentissage et Evaluation de modèles

Classification binaire

Prétraitement des données

- Imputation des valeurs manquantes: *impute.SimpleImuter* ("median")

- Séparation de App_train en X_train et X_test avec *model_selection.train_test_split (test_size=* 0.3) - App_test ne contient pas la variable cible "Target".

- Préprocessing avec *preprocessing.StandardScaler*.

Présentation des modèles

- **CV= 5 (n_folds), Score= précision**

GridSearchCV

Dummy Classifier (modèle de référence)

Arbre de décision

Extra Trees (XTree)

Forêts aléatoires (RF)

Régression logistique (LR)

X Gradient Boosting (XGB)

Présentation des modèles

GridSearchCV

Dummy Classifier (modèle de référence)

Arbre de décision

Extra Trees (XTree)

Forêts aléatoires (RF)

Régression logistique (LR)

X Gradient Boosting (XGB)

- **CV= 5 (n_folds), Score= précision**
- **Dummy**
  - Strategy = "Stratified" (default)

Présentation des modèles

GridSearchCV

Dummy Classifier - *modèle de référence*

Arbre de décision

Extra Trees (XTree)

Forêts aléatoires (RF)

Régression logistique (LR)

X Gradient Boosting (XGB)

- **CV= 5 (n_folds), Score= précision**
- Dummy
  - strategy = "Stratified" (default)
- **Arbre de décision**
  - max_depth = 2:1:10

Présentation des modèles

GridSearchCV

Dummy Classifier - *modèle de référence*

Arbre de décision

Extra Trees (XTree)

Forêts aléatoires (RF)

Régression logistique (LR)

X Gradient Boosting (XGB)

- **CV= 5 (n_folds), Score= précision**
- Dummy
  - strategy = "Stratified" (default)
- Arbre de décision
  - max_depth = 2:1:10
- **XTree**
  - n_estimators=100,
  - max_depth = 2:1:10

Présentation des modèles

GridSearchCV

Dummy Classifier - *modèle de référence*

Arbre de décision

Extra Trees (XTree)

Forêts aléatoires (RF)

Régression logistique (LR)

X Gradient Boosting (XGB)

- **CV= 5 (n_folds), Score= précision**
- **RF**
  - n_estimators=100,
  - max_depth = 2:1:10

Présentation des modèles

GridSearchCV

Dummy Classifier - *modèle de référence*

Arbre de décision

Extra Trees (XTree)

Forêts aléatoires (RF)

Régression logistique (LR)

X Gradient Boosting (XGB)

- **CV= 5 (n_folds), Score= précision**
- RF
  - n_estimators=100,
  - max_depth = 2:1:10
- **LR**
  - C = Log_10(-3:3:5)

Présentation des modèles

GridSearchCV

Dummy Classifier - *modèle de référence*

Arbre de décision

Extra Trees (XTree)

Forêts aléatoires (RF)

Régression logistique (LR)

X Gradient Boosting (XGB)

- **CV= 5 (n_folds), Score= précision**
- RF
  - n_estimators = 100,
  - max_depth = 2:1:10
- LR
  - C = Log_10(-3:3:5)
- **XGB**
  - n_estimators = 100,
  - max_depth = 2:1:10
  - learn_rate = Log_10(-3:3:7)

Démarche générale d'apprentissage, d'évaluation et d'interprétation des modèles

1. Apprentissage sur set large de modèles (large spectre de complexité) et évaluation par un score classique
   -> Sélection des modèles performants
2. Apprentissage des modèles sélectionnés avec introduction de métrique métier pour l'évaluation
3. Apprentissage et évaluation avec introduction de variables métier additionnelles – feature engineering
4. Apprentissage et évaluation avec introduction de l'équilibrage des classes
5. Interprétation des modèles

Apprentissage & Evaluation Perfo Modèles

score GridSCV= 'accuracy'

- **Set large de modèles**

Score GridSCV= accuracy	roc_auc	f1_score	max_depth_opt	C_opt	learning_rate_opt
XGBoosting	0,749	0,02202	3		0,1
Regression Logistique	0,740	0,00415		0,001	
Random Forests	0,707	0,00000	2		
Extra Trees	0,694	0,00000	2		
Arbre de décision	0,644	0,00000	2		
Dummy Classifier	0,500	0,08214			

Apprentissage & Evaluation Perfo Modèles

score GridSCV= 'accuracy'

- Set large de modèles
- **Métrique d'évaluation perfo = Roc_AUC**

Score GridSCV= accuracy	roc_auc	f1_score	max_depth_opt	C_opt	learning_rate_opt
XGBoosting	**0,749**	0,02202	3		0,1
Regression Logistique	**0,740**	0,00415		0,001	
Random Forests	**0,707**	0,00000	2		
Extra Trees	**0,694**	0,00000	2		
Arbre de décision	**0,644**	0,00000	2		
Dummy Classifier	**0,500**	0,08214			

Apprentissage & Evaluation Perfo Modèles

score GridSCV= 'accuracy'

- Set large de modèles
- Métrique d'évaluation perfo = Roc_AUC
- **Modèles retenus**

Score GridSCV= accuracy	roc_auc	f1_score	max_depth_opt	C_opt	learning_rate_opt
XGBoosting	**0,749**	0,02202	3		0,1
Regression Logistique	**0,740**	0,00415		0,001	
Random Forests	**0,707**	0,00000	2		
Extra Trees	**0,694**	0,00000	2		
Arbre de décision	**0,644**	0,00000	2		
Dummy Classifier	**0,500**	0,08214			

Apprentissage & Evaluation Perfo Modèles

Introduction de la métrique métier

- **Définition de la métrique métier**
  - **Sans coût de l'opportunité**

	Fait défaut	**Rembourse**
Défaut prédit {1}	VP: pas d'investissement (pai)	FP: pas d'investissement (pai)
Remb. prédit {0}	FN: perte d'investissement (pri)	VN: retour sur investissement (rsi)

Apprentissage & Evaluation Perfo Modèles

Introduction de la métrique métier

- **Définition de la métrique métier**
  - **Sans coût de l'opportunité**
    - **Min FN, Max VN**
      -> Max F_beta (beta>1)

	Fait défaut	**Rembourse**
Défaut prédit {1}	VP: pas d'investissement (pai)	FP: pas d'investissement (pai)
Remb. prédit {0}	FN: perte d'investissement (pri)	VN: retour sur investissement (rsi)

Apprentissage & Evaluation Perfo Modèles

Introduction de la métrique métier

- **Définition de la métrique métier**
  - Sans coût de l'opportunité
    - Min FN, Max VN
      -> Max F_beta (beta>1)
  - **Avec coût de l'opportunité**

	Fait défaut	**Rembourse**
Défaut prédit {1}	VP: rsi "bas"+ rsi partiel "haut" + pri partiel	FP: rsi "bas"+ - rsi (total) "haut"
Remb. prédit {0}	FN: rsi partiel "haut" + pri partiel - rsi "bas"	VN: rsi (total) "haut" - rsi "bas"

Apprentissage & Evaluation Perfo Modèles

Introduction de la métrique métier

- **Définition de la métrique métier**
  - Sans coût de l'opportunité
    - Min FN, Max VN
      -> Max F_beta (beta>1)
  - **Avec coût de l'opportunité**
    - **Max combinaison linéaire {VP, FP, FN, VN}** (
      rsi "bas" = 1.5%, rsi "haut" = 5%,
      p: partie du principal remboursée avec intérêt = 50%)

	Fait défaut	**Rembourse**
Défaut prédit {1}	VP: rsi "bas"+ rsi partiel "haut" + pri partiel	FP: rsi "bas"+ - rsi (total) "haut"
Remb. prédit {0}	FN: rsi partiel "haut" + pri partiel - rsi "bas"	VN: rsi (total) "haut" - rsi "bas"

Apprentissage & Evaluation Perfo Modèles

Introduction de la métrique métier

- **Définition de la métrique métier**
  - Sans coût de l'opportunité
    - Min FN, Max VN
      -> Max F_beta (beta>1)
  - **Avec coût de l'opportunité**
    - **Max combinaison linéaire {VP, FP, FN, VN}** (
      rsi "bas" = 1.5%, rsi "haut" = 5%,
      p: partie du principal remboursée avec intérêt = 50%)
    - **Métrique non-corrélée avec #FN !!**

	Fait défaut	**Rembourse**
Défaut prédit {1}	VP: rsi "bas"+ rsi partiel "haut" + pri partiel	FP: rsi "bas"+ - rsi (total) "haut"
Remb. prédit {0}	FN: rsi partiel "haut" + pri partiel - rsi "bas"	VN: rsi (total) "haut" - rsi "bas"

Apprentissage & Evaluation Perfo Modèles

Introduction de la métrique métier

- **Définition de la métrique métier**
  - Sans coût de l'opportunité
    - Min FN, Max VN
      -> Max F_beta (beta>1)
  - Avec coût de l'opportunité
    - Max combinaison linéaire {VP, FP, FN, VN} (
      rsi "bas" = 1.5%, rsi "haut" = 5%,
      p: partie du principal remboursée avec intérêt = 50%)
    - Métrique non-corrélée avec #FN
- **Métrique conservée: F_beta (beta=2)**

	Fait défaut	**Rembourse**
Défaut prédit {1}	VP: pas d'investissement (pai)	FP: pas d'investissement (pai)
Remb. prédit {0}	FN: perte d'investissement (pri)	VN: retour sur investissement (rsi)

Apprentissage & Evaluation Perfo Modèles

Introduction de la métrique métier

- **Sur modèles sélectionnés précédemment**

Score GridSCV= accuracy	business_metrics	roc_auc	f1_score	max_depth_opt	C_opt	learning_rate_opt
XGBoosting	0,0768	0,716	0,107	3		1
Regression Logistique	0,0181	0,740	0,028		31,623	
Random Forests	0,0000	0,701	0,000	2		
Extra Trees	0,0000	0,695	0,000	2		

Apprentissage & Evaluation Perfo Modèles

Introduction de la métrique métier

- Sur modèles sélectionnés précédemment
- **Métrique d'évaluation perfo = business_metrics**

Score GridSCV= accuracy	business_metrics	roc_auc	f1_score	max_depth_opt	C_opt	learning_rate_opt
XGBoosting	**0,0768**	0,716	0,107	3		1
Regression Logistique	**0,0181**	0,740	0,028		31,623	
Random Forests	**0,0000**	0,701	0,000	2		
Extra Trees	**0,0000**	0,695	0,000	2		

Apprentissage & Evaluation Perfo Modèles

Introduction de 4 variables métier - *feature engineering*

- **Les 4 variables métier**

- ['CREDIT_ANNUITY_RATIO'] = ['AMT_CREDIT'] / ['AMT_ANNUITY']
- ["credit_income_ratio"] = ["AMT_CREDIT"] / ["AMT_INCOME_TOTAL"]
- ["annuity_income_ratio"] = ["AMT_ANNUITY"] / ["AMT_INCOME_TOTAL"]
- ["credit_good_ratio"] = ["AMT_CREDIT"] / ["AMT_GOODS_PRICE"]

Apprentissage & Evaluation Perfo Modèles

Introduction de 4 variables métier - *feature engineering*

- **Les 4 variables métier**
- Amélioration des performances sur la métrique métier (et sur le **roc_auc).**

Score GridSCV = accuracy	business_metrics	roc_auc	f1_score	max_depth_opt	C_opt	learning_rate_opt
XGBoosting	0,0819	**0,724**	0,112	3		1
Regression Logistique	0,0196	**0,742**	0,031		1	
Random Forests	0,0000	**0,708**	0,000	2		
Extra Trees	0,0000	**0,687**	0,000	2		

Apprentissage et évaluation avec

Introduction de l'équilibrage des classes

- Constat d'un déséquilibre des classes de l'ordre de:
  #(y=1) / #(y=0) ~ 11

Apprentissage et évaluation avec

Introduction de l'équilibrage des classes

- Constat d'un déséquilibre des classes de l'ordre de:
  #(y=1) / #(y=0) ~ 11
- **Introduction de l'hyperparamètre *class_weight* dans les modèles**:
  - class_weight = 'balanced' ou 11 (pour XGB)

Score GridSCV = accuracy	business_metrics	roc_auc	**f1_score**	**max_depth_opt**	**C_opt**	**learning_rate_opt**
XGBoosting	0,426	0,758	0,274	3		0,178
Regression Logistique	0,411	0,744	0,257		0,001	
Random Forests	0,395	0,727	0,246	5		
Extra Trees	0,372	0,707	0,232	9		

Apprentissage et évaluation avec

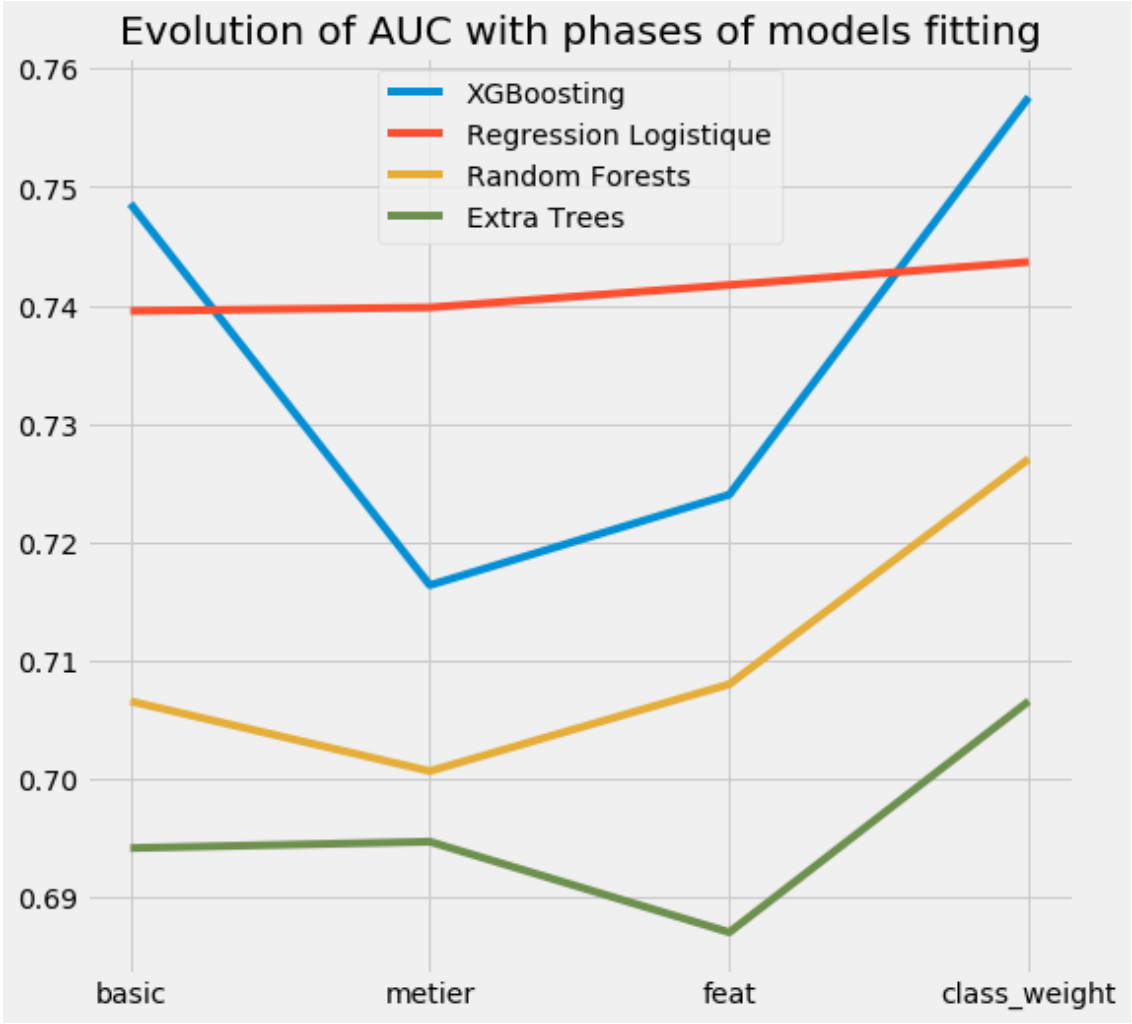
Introduction de l'équilibrage des classes

- Constat d'un déséquilibre des classes de l'ordre de:
  #(y=1) / #(y=0) ~ 11
- Introduction de l'hyperparamètre *class_weight* dans les modèles:
  - class_weight = 'balanced' ou 11 (pour XGB)
- Amélioration des performances sur la métrique métier (et sur le **roc_auc).**

Score GridSCV = accuracy	business_metrics	roc_auc	f1_score	max_depth_opt	C_opt	learning_rate_opt
XGBoosting	0,426	**0,758**	0,274	3		0,178
Regression Logistique	0,411	**0,744**	0,257		0,001	
Random Forests	0,395	**0,727**	0,246	5		
Extra Trees	0,372	**0,707**	0,232	9		

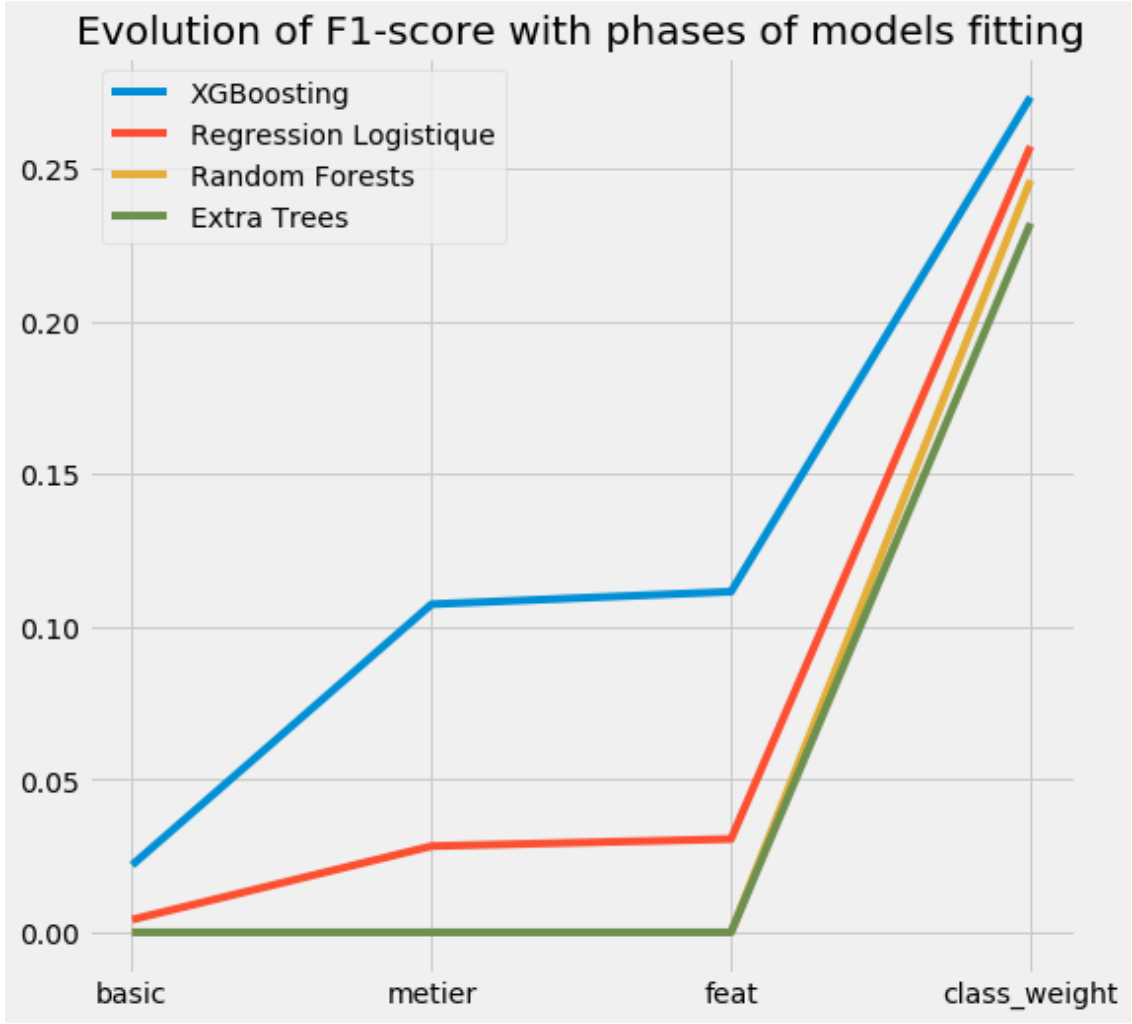
Evolution des métriques principales

- **Evolution de la ROC AUC**
- Evolution du F1 Score
- Evolution de la métrique métier = F1_beta



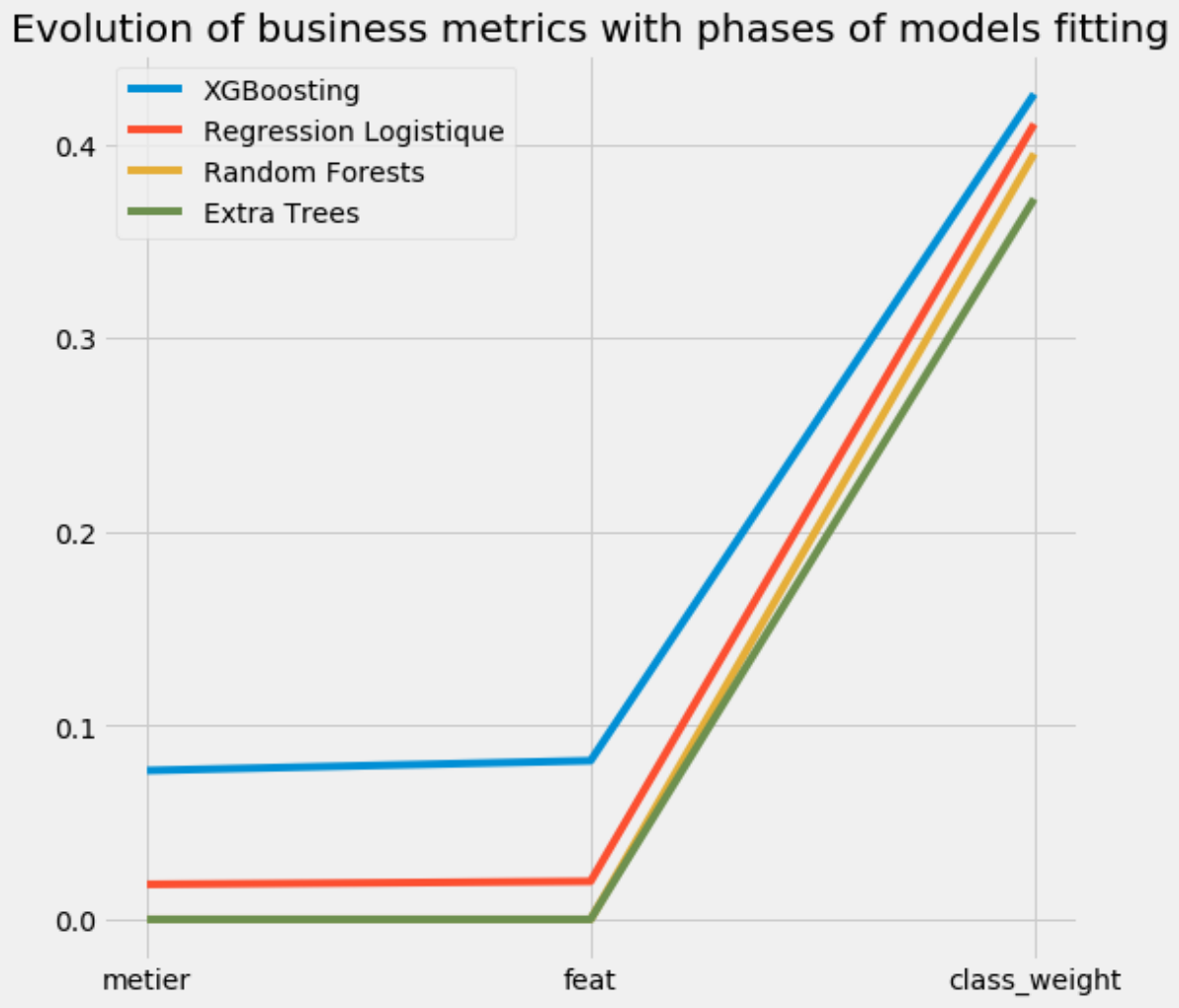
Evolution des métriques principales

- Evolution de la ROC AUC
- **Evolution du F1 Score**
- Evolution de la métrique métier = F1_beta



Evolution des métriques principales

- Evolution de la ROC AUC
- Evolution du F1 Score
- **Evolution de la métrique métier = F1_beta**



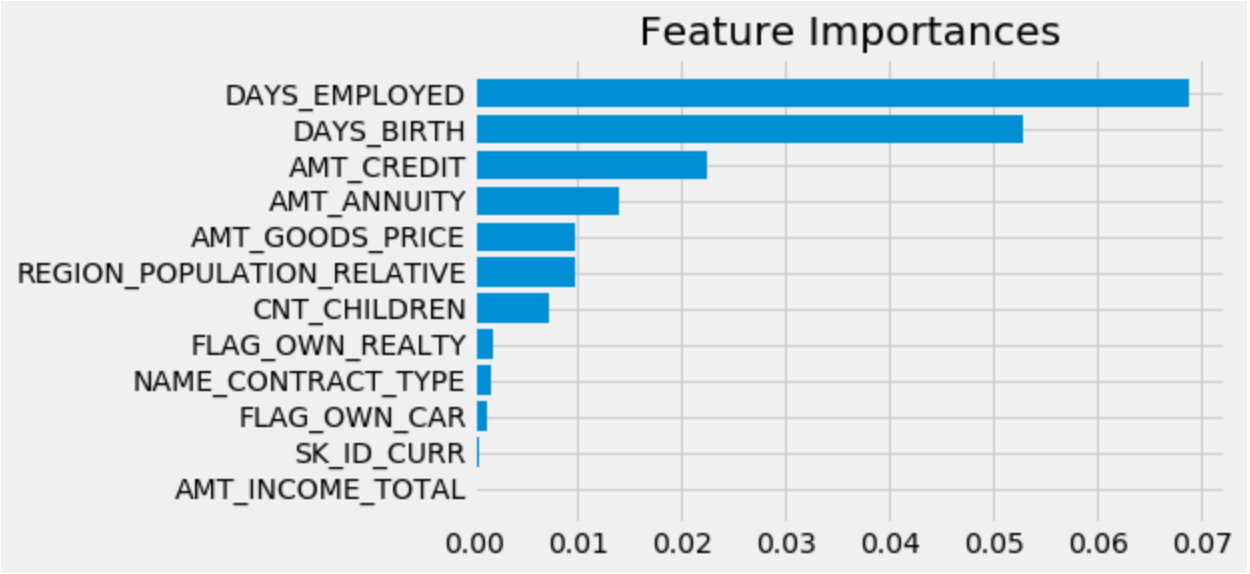
Interprétation des modèles

Classification binaire

Feature importance

Forêts aléatoire

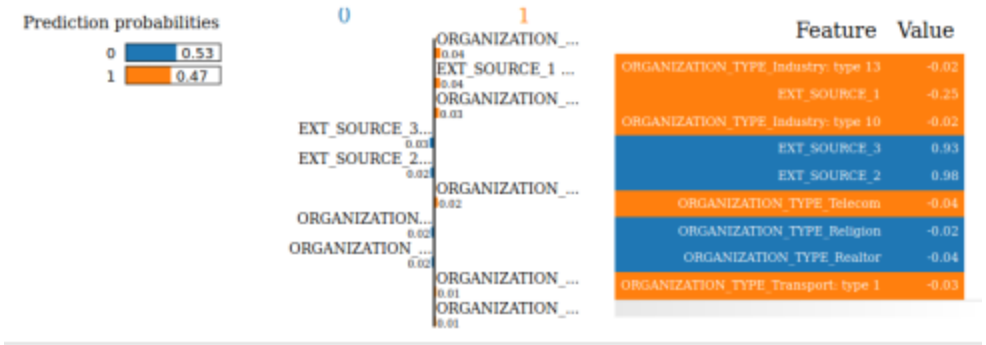
Principe: La forêt aléatoire est constituée de plusieurs arbres. Chaque arbre grandit jusqu'à une taille espérée qui correspond au maximum du *gain d'information*. A chaque nœud, une valeur particulière d'une variable parmi toutes, va minimiser l'*impureté de gini* (du nœud), fixant ainsi les deux prochaines branches. C'est la moyenne sur tous les arbres de la décroissance d'impureté de gini par une variable qui est la mesure de son importance.
C'est donc une mesure globale sur tous les individus.



Lime

Principe: Lime permet l'interprétation de la prédiction d'un modèle sur chaque individu de l'échantillon de données. Ceci est rendu possible par l'apprentissage d'un modèle linéaire localement, autour de l'individu - dans l'espace contenant le nuage des points-individus. Cet apprentissage est fait sur un dataset reconstitué après une attribution de poids chaque autre individu, dont la valeur est inversede sa distance à l'individu cible.

Dans Lime, l'importance d'une variable dans la prédiction du modèle est quantifiée (avec, au cas échéant une précision de la plage de valeurs de la variable particulièrement concernée).



Merci pour votre attention

Disponible pour des questions/réponses