Implémentez un modèle de scoring

Projet 7 Parcours Data Scientist

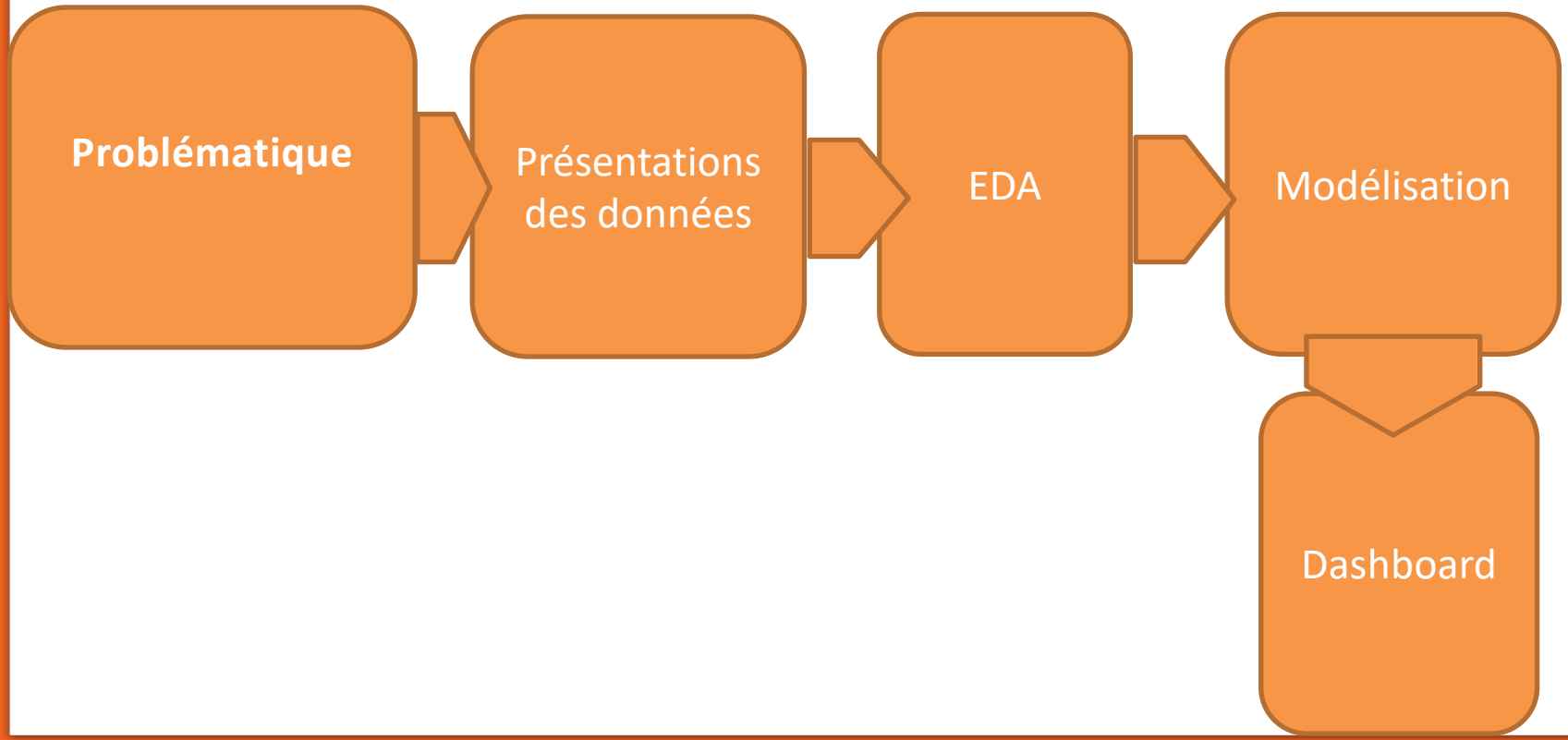
Python Jupyter Notebook – VSCode

Github > Streamlit > Heroku

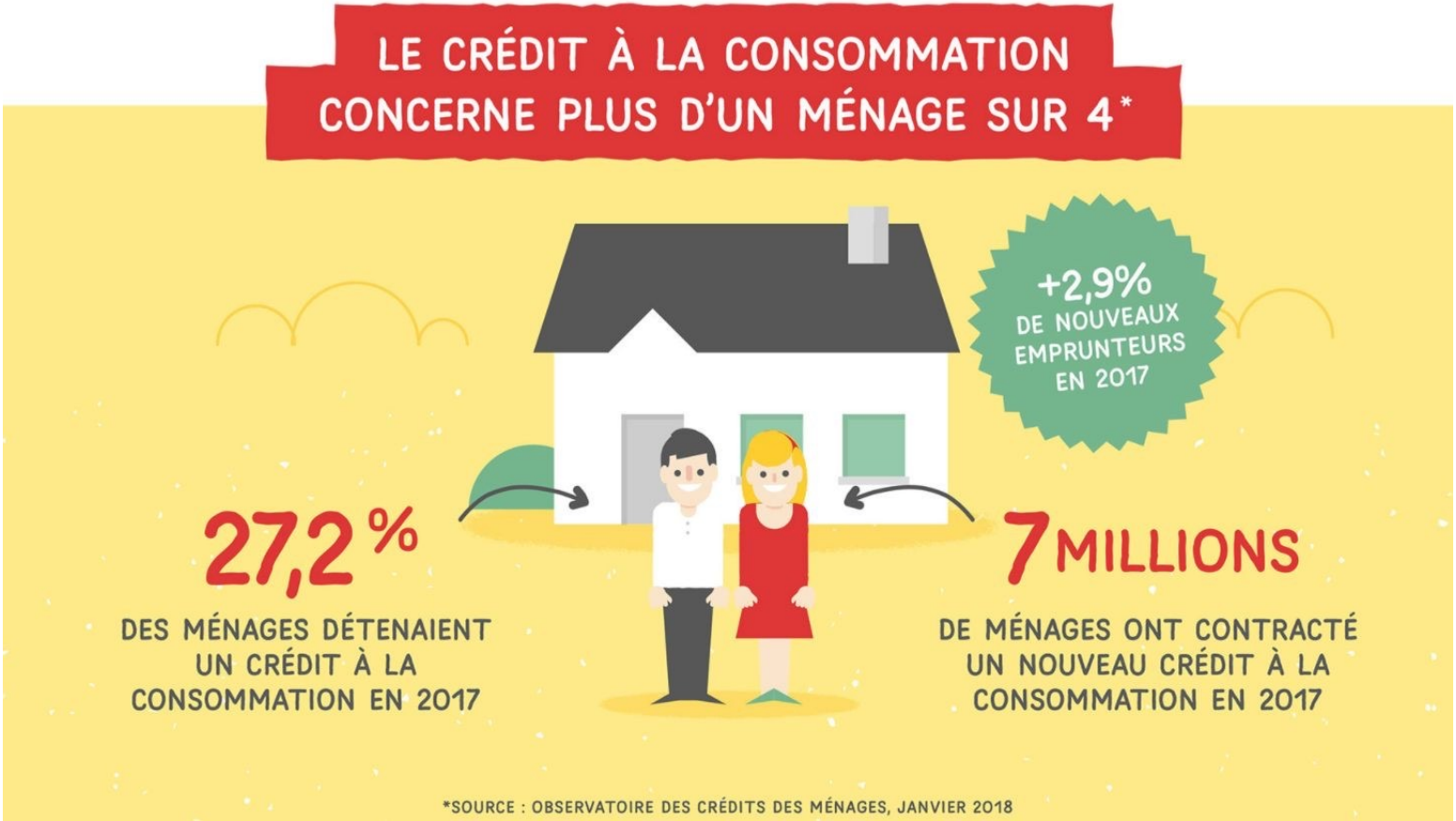
Yanes Khereddine

OpenClassrooms - Centrale Supélec

Sommaire


Problématique
Présentations des données
EDA
Modélisation
Dashboard

Contexte:
Impact marcher

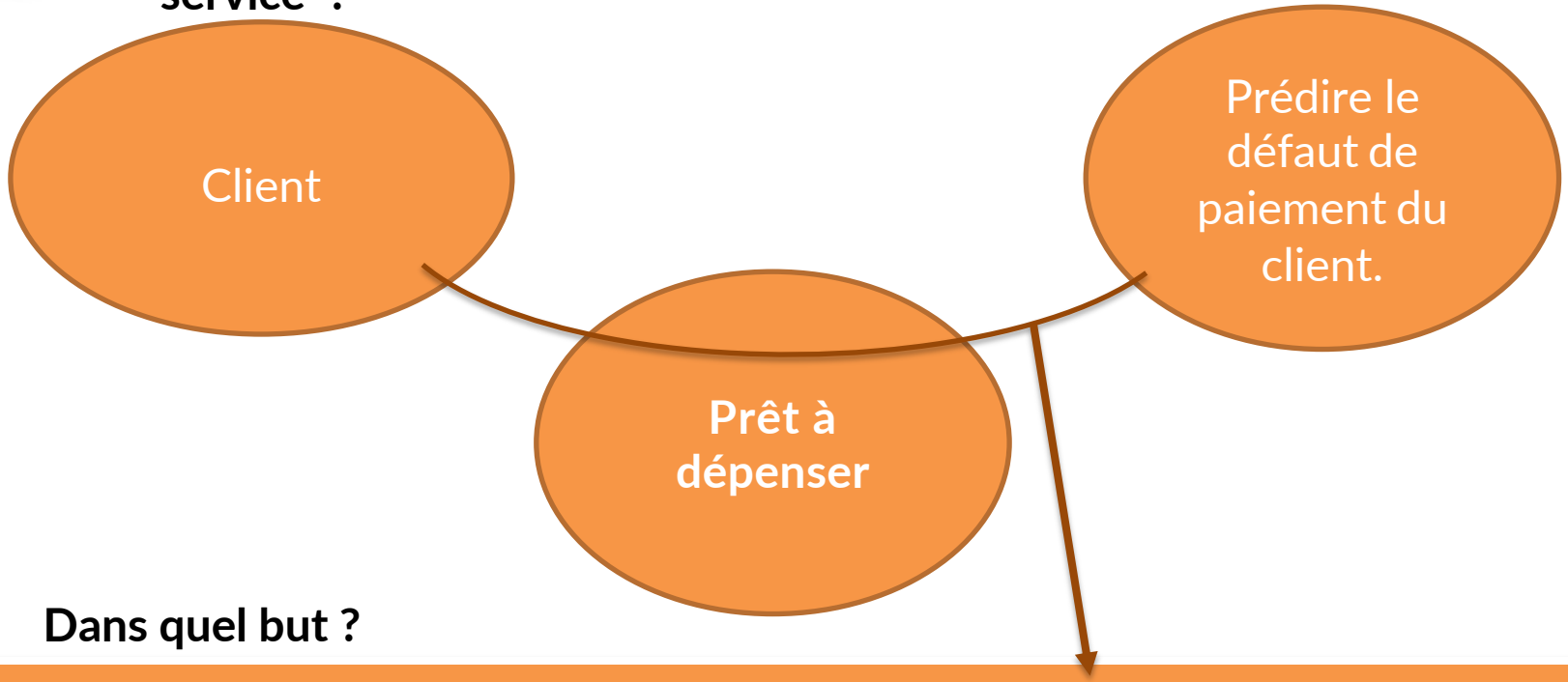


Problématique :

Problématique

A qui rend-il service ?

Sur quoi agit-il ?



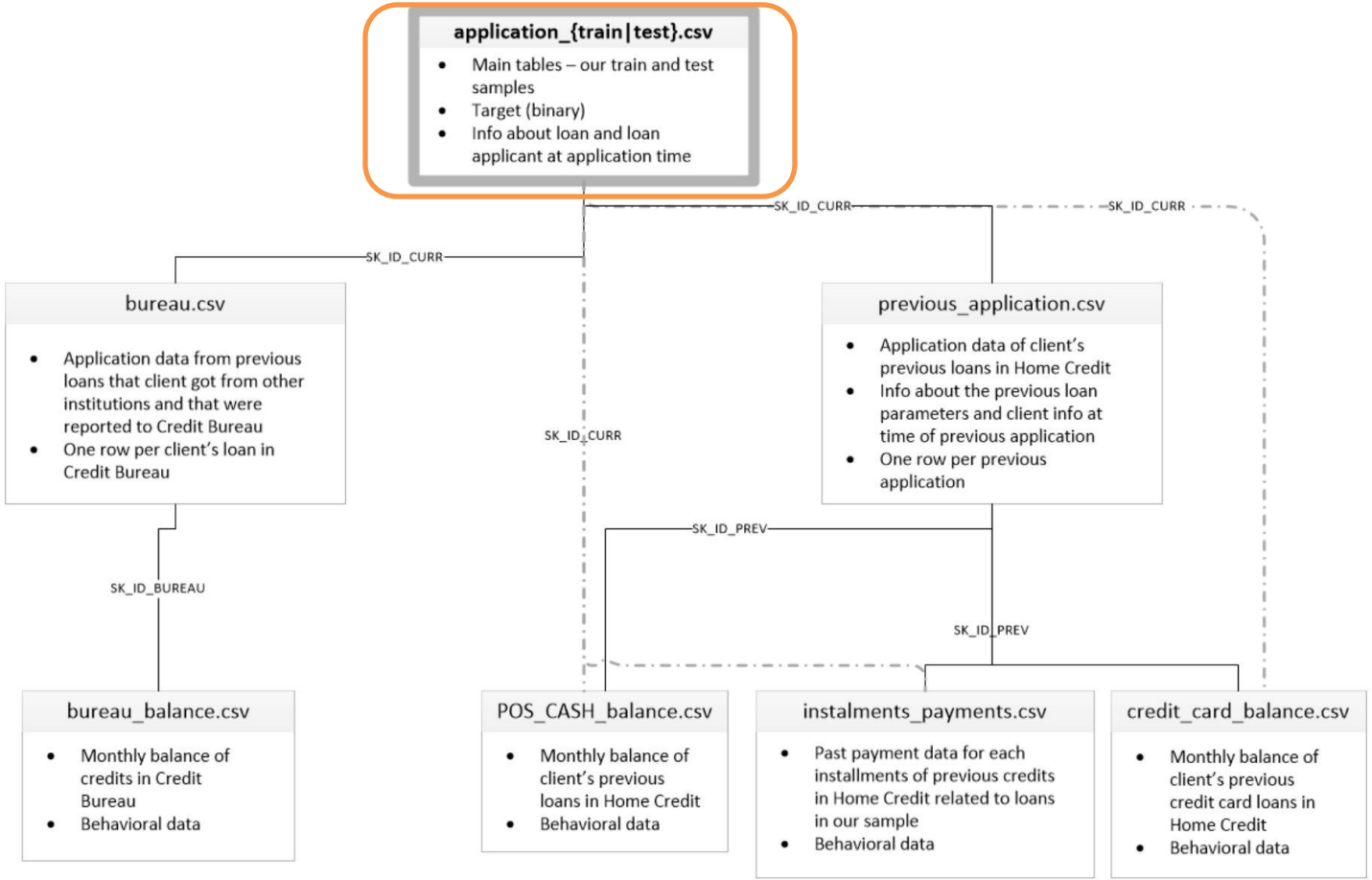
Dans quel but ?

Dashboard interactif à destination des chargés de relation client, afin d'avoir plus de transparence sur la façons dont on octroie un crédit.

– Présentation des données

Kaggle "Home Credit Default Risk" :
https://www.kaggle.com/c/home-credit-default-risk/data

Schéma de la BDD *source Kaggle*



Description rapide des données

	file.csv	n_rows	n_cols	null_amount	qty_null_columns	null_columns
0	application_test	48744	121	1404419	64	AMT_ANNUITY, NAME_TYPE_SUITE, OWN_CAR_AGE, OCCUPATION_TYPE, EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3, APARTMENTS_AVG, BASEMENTAREA_AVG, YEARS_BEGINEXPLUATATION_AVG, YEARS_BUILD_AVG, COMMONAREA_AVG, ELEVATORS_AVG, ENTRANCES_AVG, FLOORSMAX_AVG, FLOORSMIN_AVG, LANDAREA_AVG, LIVINGAPARTMENTS_AVG, LIVINGAREA_AVG, NONLIVINGAPARTMENTS_AVG, NONLIVINGAREA_AVG, APARTMENTS_MODE, BASEMENTAREA_MODE, YEARS_BEGINEXPLUATATION_MODE, YEARS_BUILD_MODE, COMMONAREA_MODE, ELEVATORS_MODE, ENTRANCES_MODE, FLOORSMAX_MODE, FLOORSMIN_MODE, LANDAREA_MODE, LIVINGAPARTMENTS_MODE, LIVINGAREA_MODE, NONLIVINGAPARTMENTS_MODE, NONLIVINGAREA_MODE, APARTMENTS_MEDI, BASEMENTAREA_MEDI, YEARS_BEGINEXPLUATATION_MEDI, YEARS_BUILD_MEDI, COMMONAREA_MEDI, ELEVATORS_MEDI, ENTRANCES_MEDI, FLOORSMAX_MEDI, FLOORSMIN_MEDI, LANDAREA_MEDI, LIVINGAPARTMENTS_MEDI, LIVINGAREA_MEDI, NONLIVINGAPARTMENTS_MEDI, NONLIVINGAREA_MEDI, FONDKAPREMONT_MODE, HOUSETYPE_MODE, TOTALAREA_MODE, WALLSMATERIAL_MODE, EMERGENCYSTATE_MODE, OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE, AMT_REQ_CREDIT_BUREAU_HOUR, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_QRT, AMT_REQ_CREDIT_BUREAU_YEAR
1	pos_cash_balance	10001358	8	52158	2	CNT_INSTALMENT, CNT_INSTALMENT_FUTURE
2	credit_card_balance	3840312	23	5877356	9	AMT_DRAWINGS_ATM_CURRENT, AMT_DRAWINGS_OTHER_CURRENT, AMT_DRAWINGS_POS_CURRENT, AMT_INST_MIN_REGULARITY, AMT_PAYMENT_CURRENT, CNT_DRAWINGS_ATM_CURRENT, CNT_DRAWINGS_OTHER_CURRENT, CNT_DRAWINGS_POS_CURRENT, CNT_INSTALMENT_MATURE_CUM
3	installments_payments	13605401	8	5810	2	DAYS_ENTRY_PAYMENT, AMT_PAYMENT
4	application_train	307511	122	9152465	67	AMT_ANNUITY, AMT_GOODS_PRICE, NAME_TYPE_SUITE, OWN_CAR_AGE, OCCUPATION_TYPE, CNT_FAM_MEMBERS, EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3, APARTMENTS_AVG, BASEMENTAREA_AVG, YEARS_BEGINEXPLUATATION_AVG, YEARS_BUILD_AVG, COMMONAREA_AVG, ELEVATORS_AVG, ENTRANCES_AVG, FLOORSMAX_AVG, FLOORSMIN_AVG, LANDAREA_AVG, LIVINGAPARTMENTS_AVG, LIVINGAREA_AVG, NONLIVINGAPARTMENTS_AVG, NONLIVINGAREA_AVG, APARTMENTS_MODE, BASEMENTAREA_MODE, YEARS_BEGINEXPLUATATION_MODE, YEARS_BUILD_MODE, COMMONAREA_MODE, ELEVATORS_MODE, ENTRANCES_MODE, FLOORSMAX_MODE, FLOORSMIN_MODE, LANDAREA_MODE, LIVINGAPARTMENTS_MODE, LIVINGAREA_MODE, NONLIVINGAPARTMENTS_MODE, NONLIVINGAREA_MODE, APARTMENTS_MEDI, BASEMENTAREA_MEDI, YEARS_BEGINEXPLUATATION_MEDI, YEARS_BUILD_MEDI, COMMONAREA_MEDI, ELEVATORS_MEDI, ENTRANCES_MEDI, FLOORSMAX_MEDI, FLOORSMIN_MEDI, LANDAREA_MEDI, LIVINGAPARTMENTS_MEDI, LIVINGAREA_MEDI, NONLIVINGAPARTMENTS_MEDI, NONLIVINGAREA_MEDI, FONDKAPREMONT_MODE, HOUSETYPE_MODE, TOTALAREA_MODE, WALLSMATERIAL_MODE, EMERGENCYSTATE_MODE, OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE, DAYS_LAST_PHONE_CHANGE, AMT_REQ_CREDIT_BUREAU_HOUR, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_QRT, AMT_REQ_CREDIT_BUREAU_YEAR
5	bureau	1716428	17	3939947	7	DAYS_CREDIT_ENDDATE, DAYS_ENDDATE_FACT, AMT_CREDIT_MAX_OVERDUE, AMT_CREDIT_SUM, AMT_CREDIT_SUM_DEBT, AMT_CREDIT_SUM_LIMIT, AMT_ANNUITY
6	previous_application	1670214	37	11109336	16	AMT_ANNUITY, AMT_CREDIT, AMT_DOWN_PAYMENT, AMT_GOODS_PRICE, RATE_DOWN_PAYMENT, RATE_INTEREST_PRIMARY, RATE_INTEREST_PRIVILEGED, NAME_TYPE_SUITE, CNT_PAYMENT, PRODUCT_COMBINATION, DAYS_FIRST_DRAWING, DAYS_FIRST_DUE, DAYS_LAST_DUE_1ST_VERSION, DAYS_LAST_DUE, DAYS_TERMINATION, NFLAG_INSURED_ON_APPROVAL
7	bureau_balance	27299925	3	0	0	
8	sample_submission	48744	2	0	0	

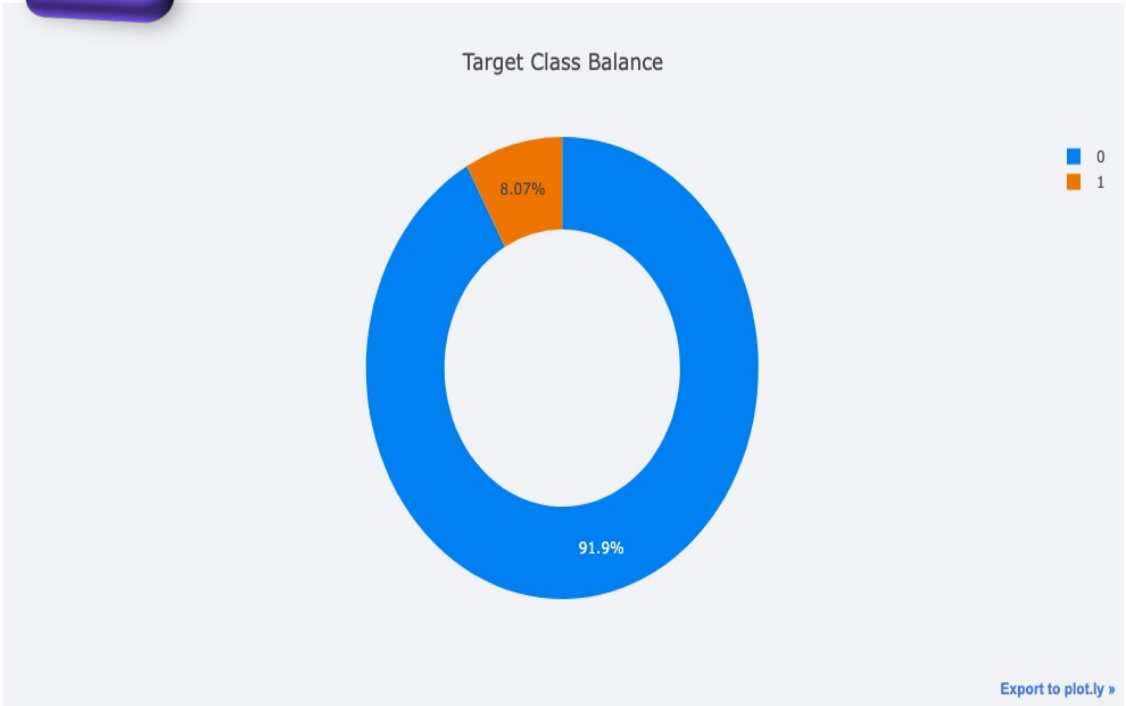
Jeu de données principal

Analyse exploratoire des données

Inspiré par le Kernel :
https://www.kaggle.com/thiagopanini/predicting-credit-risk-eda-viz-pipeline
https://www.kaggle.com/willkoehrsen/start-here-a-gentle-introduction

Inspiré par le Github :
https://github.com/nalron/project_credit_scoring_model

Distribution de la cible



ICI LA VARIABLE TARGET EST CE QU'ON NOUS DEMANDE DE PRÉDIRE :

- 0: POUR LE PRÊT A ÉTÉ REMBOURSÉ À TEMPS.
- 1: INDIQUANT QUE LE CLIENT A EU DES DIFFICULTÉS DE PAIEMENT.

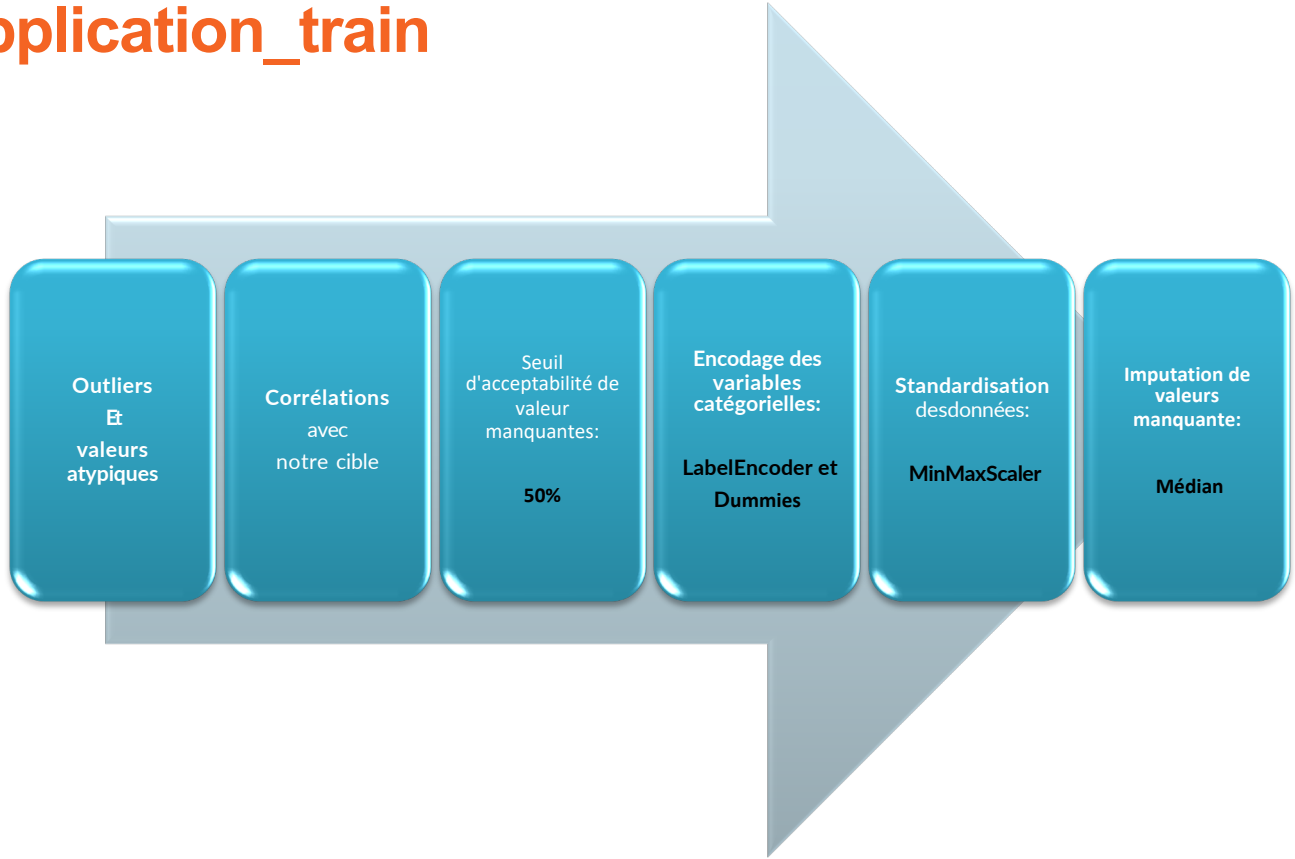
SMOTE

ET

Classe Weight >>> Hyperparamètres des Model

Technique utilisée pour traiter des ensembles de **données déséquilibrées**.

Preprocessing : Application_train



Opération de Merging

Enrichissement de l'échantillon de travail :

Combinaison des **7 jeux de données.**

Avant 122 features - Après **165 features**

Dont 3 features de moyenne et de comptage :

PREVIOUS_LOANS_COUNT : nombre des précédents crédits pris par le client

MONTHS_BALANCE_MEAN : solde mensuel moyen des précédents crédits

PREVIOUS_APPLICATION_COUNT : nombre de demandes antérieures au crédit immobilier

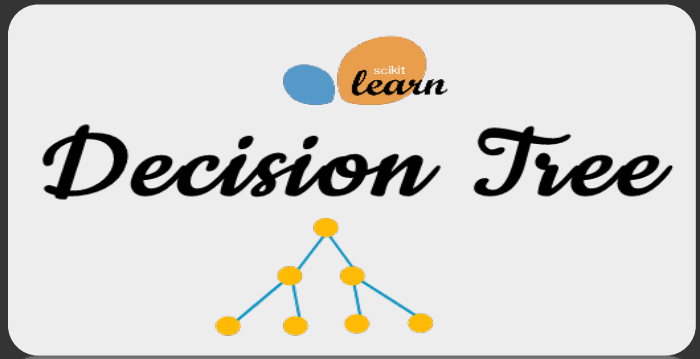
Feature engineering

Enrichissement de l'échantillon par **4 ratios explicatifs** :

CREDIT_INCOME_PERCENT : % montant du crédit par rapport au revenu d'un client
ANNUITY_INCOME_PERCENT : % rente de prêt par rapport au revenu d'un client
DAYS_EMPLOYED_PERCENT : % jours employés par rapport à l'âge du client
CREDIT_TERM : durée du paiement en mois

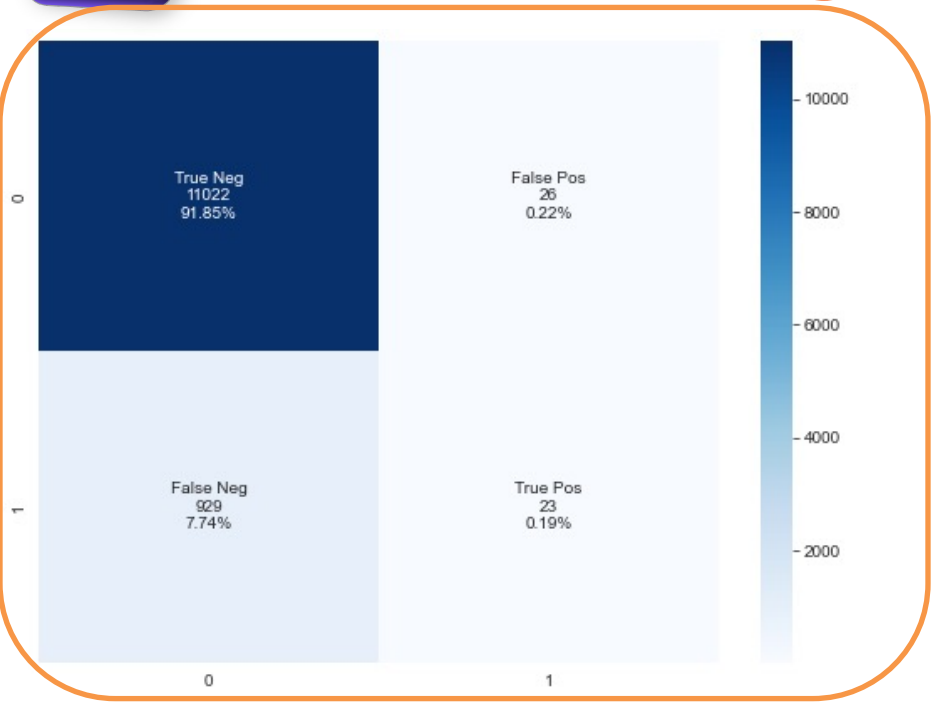
Echantillon de travail obtenu : 356255 x 169

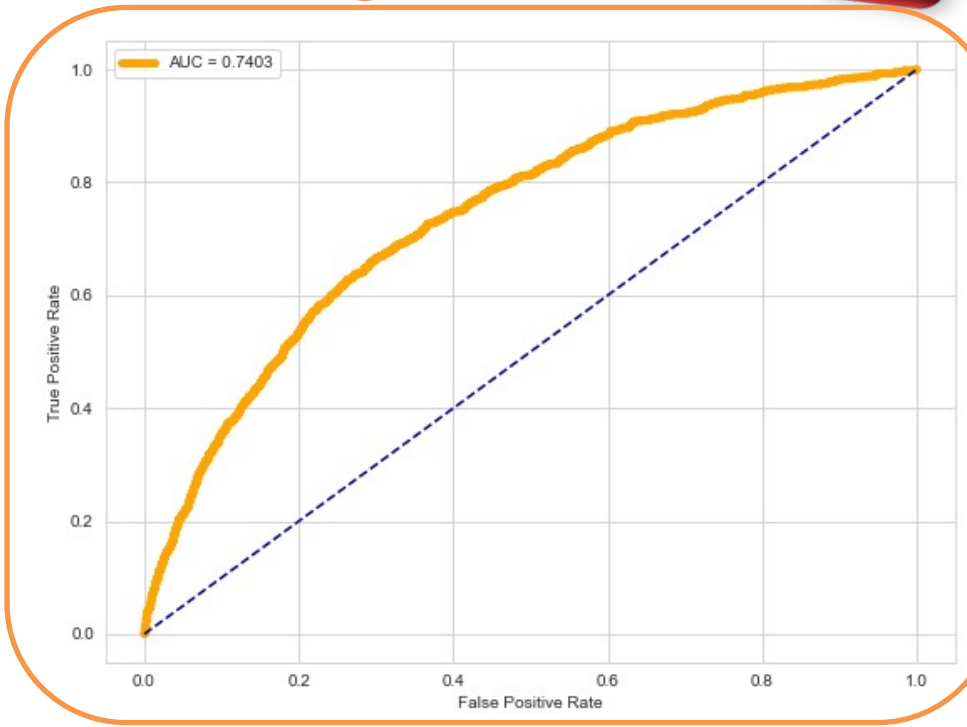
Modélisation :
Baseline fixée par régression logistique



Elaboration d'un modèle, optimisation et compréhension.

Baseline Régression logistique





Performances de la "baseline" avec toutes les features.

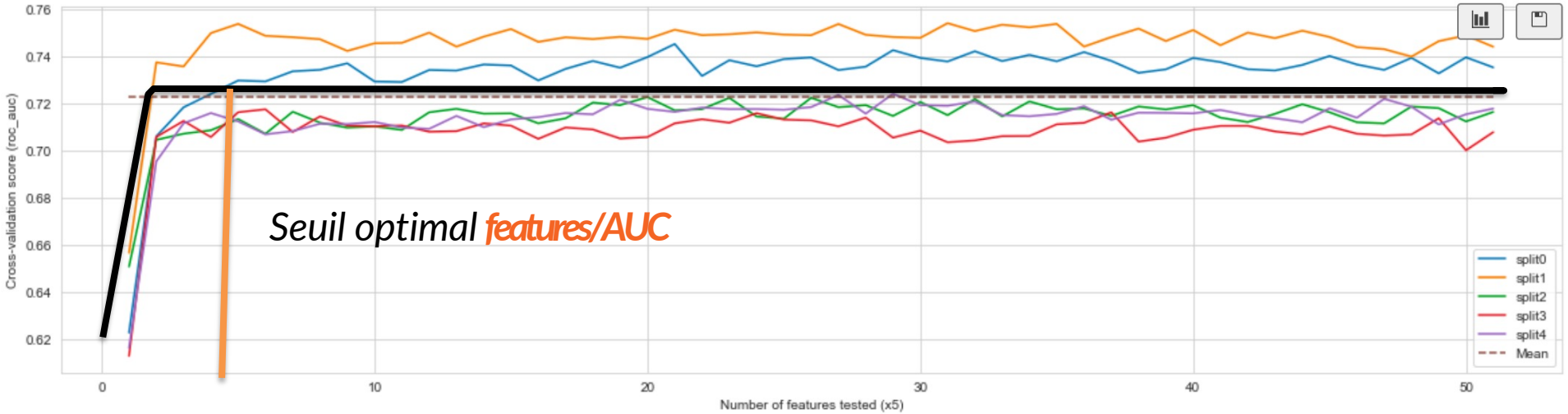
Synthèse des modèle

	Model	AUC	Accuracy	Precision	Recall	F1	Time
0	LogisticRegression	0.739456	0.920167	0.428571	0.018908	0.036217	5.940477
1	RandomForestClassifier	0.736040	0.920417	0.000000	0.000000	0.000000	52.483439
2	DecisionTreeClassifier	0.693081	0.918667	0.306452	0.019958	0.037475	0.713839

Feature selection

Recursive Feature Elimination : RFECV

Identification des best features par validation croisée en optimisant la métrique AUC.



Recursive elimination ⟶ *126 features*

Fonction coût

Limiter les risques de perte financière :

Pénaliser les *Faux Positifs* et les *Faux Négatifs*.

Quantification de l'importance relative entre *Recall* et *Precision*.

Estimation du coût moyen d'un défaut de paiement.

Estimation du coût d'opportunité d'un client refusé par erreur.

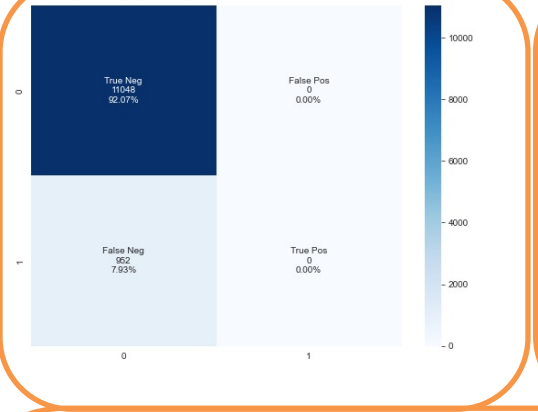
Connaissance métier nécessaire ou hypothèses à fixer.

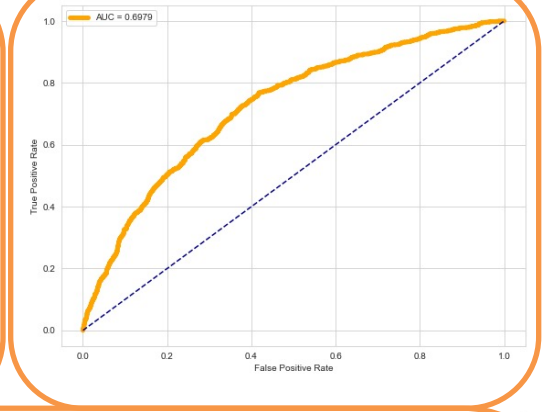
Métrique d'évaluation

Amélioration de la métrique et **pénalisation** des erreurs FP et FN.

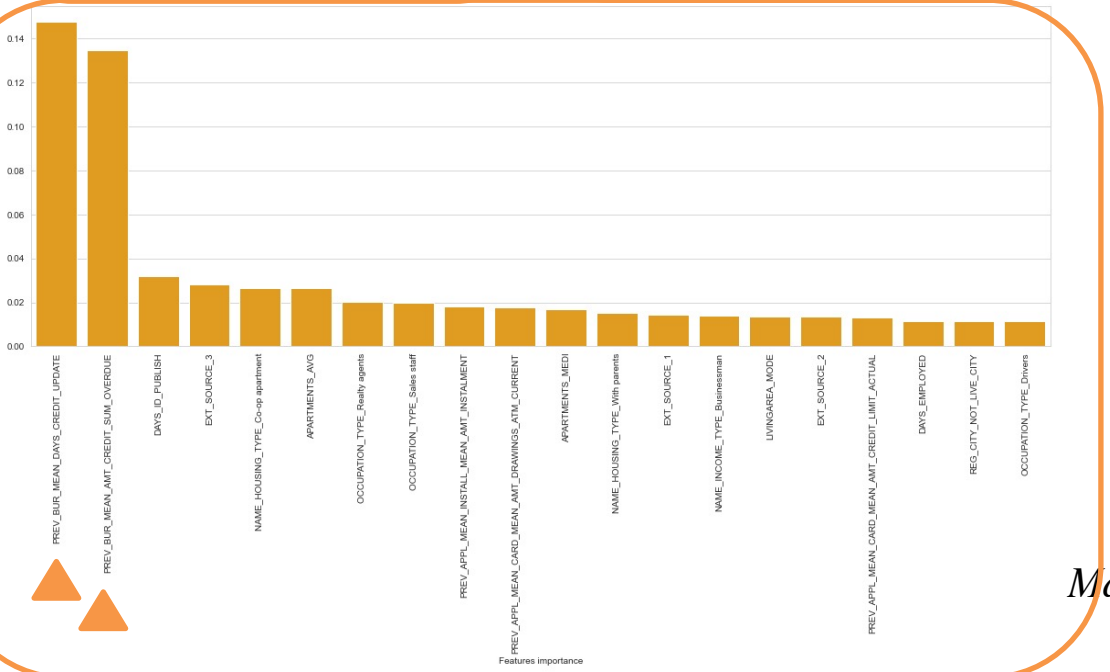
Meilleures performances

RandomForest.





- Features Engineering



Matrice de confusion + Courbe ROC / AUC score + Features Selection.

Présentation du dashboard

Versioning Github : https://github.com/DeepScienceData/Projet-OpenClassRoms
Heroku : https://model-scoring-openclassrom.herokuapp.com/
Streamlit : https://share.streamlit.io/deepsciencedata/projet-openclassroms/app/app.py

Streamlit

Streamlit : framework open-source Python spécialisé ML.

Application web : modèles de ML et visualisation des données.

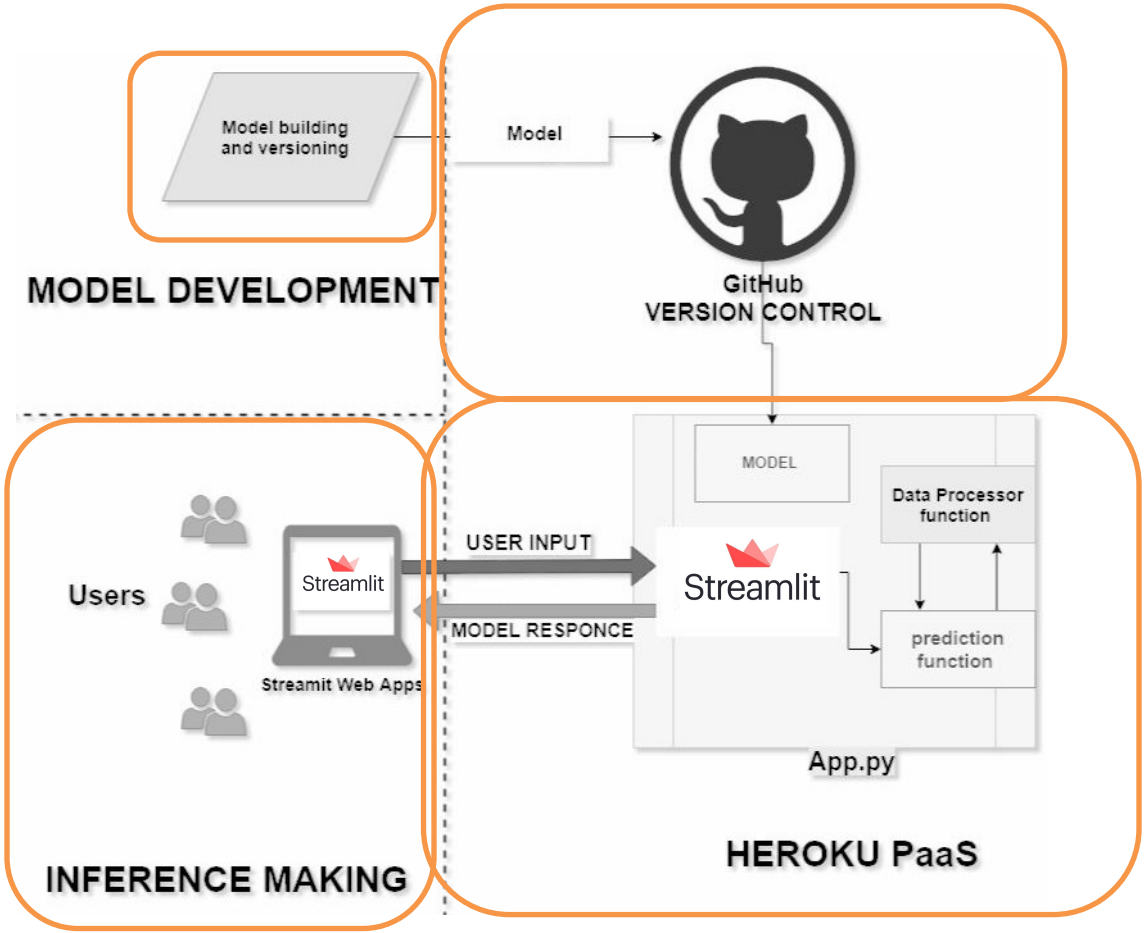
Application performante : mise en cache via une annotation.

Création facile : sans nécessité d'implémenter du HTML.

	Maturity	Popularity	Simplicity	Adaptability	Focus	Language support
Streamlit	C	A	A	C	Dashboard	Python
Dash	B	A	B	B	Dashboard	Python, R, Julia
Panel	C	B	B	B	Dashboard	Python
Shiny	A	B	B	B	Dashboard	R
Voila	C	C	A	C	Dashboard	Python, R, Julia
Jupyter	A	A	B	B	Notebook	Python, R, Julia
Flask	A	A	B	A	Web framework	Python

source : www.datarevenue.com

Schéma fonctionnel de l'application



Conclusion

Utilisation et modification d'un Kernel Kaggle.

Entraînement d'un modèle de scoring.

Fonction coût, optimisation et évaluation.

Interprétabilité du modèle RandomForest.

Dashboard interactif.

STREAMLIT : DASHBOARD

HEROKU: DASHBOARD