PARTICIPEZ À UNE COMPÉTITION KAGGLE

BOURBON Vicente

PROBLÉMATIQUE

- Kaggle: plateforme qui organise des compétitions de Data Science
- Mission: participer à une compétition réelle et en cours
- Partager ses résultats avec la communauté

SOMMAIRE

- Présentation de la compétition
- Importation et exploration des données
- Modèles de prédiction
- Conclusion

I. PRÉSENTATION DE LA COMPÉTITION

VUE D'ENSEMBLE DE LA COMPÉTITION

Mise en contexte

Commodité des cartes bancaires dans la vie moderne, mais problématiques liées aux remboursements des factures.

Situation actuelle

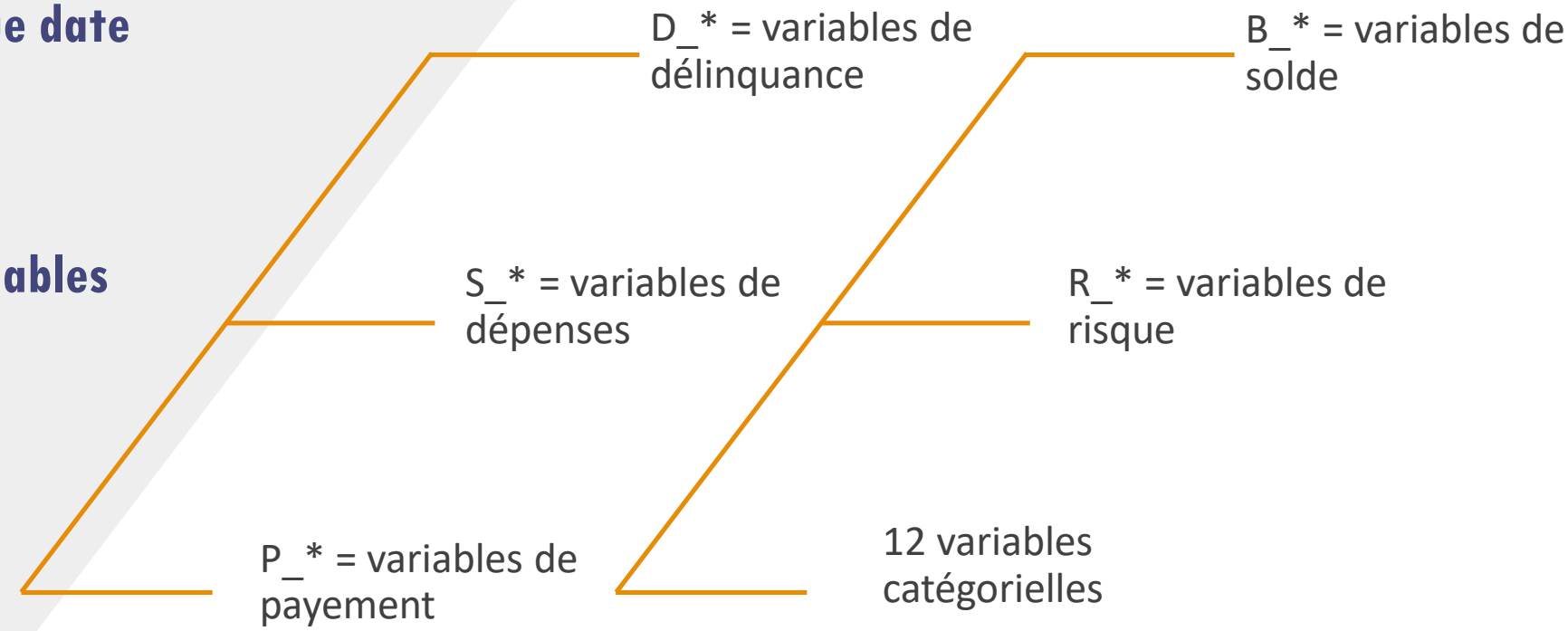
Importance de la prévision de défaut de payement pour la gestion des risques. Existence de modèles mais travail de recherche et d'amélioration continu.

Données

Données comportementales et chronologiques anonymisées de clients d'American express pour essayer de concurrencer les modèles en production.

APERÇUE DES DONNÉES

- **Caractéristiques de profil agrégées pour chaque client à chaque date de relevé**
- **Anonymes et normalisées**
- **Regroupées en catégories**
- **Pas de description des variables**



Objectif

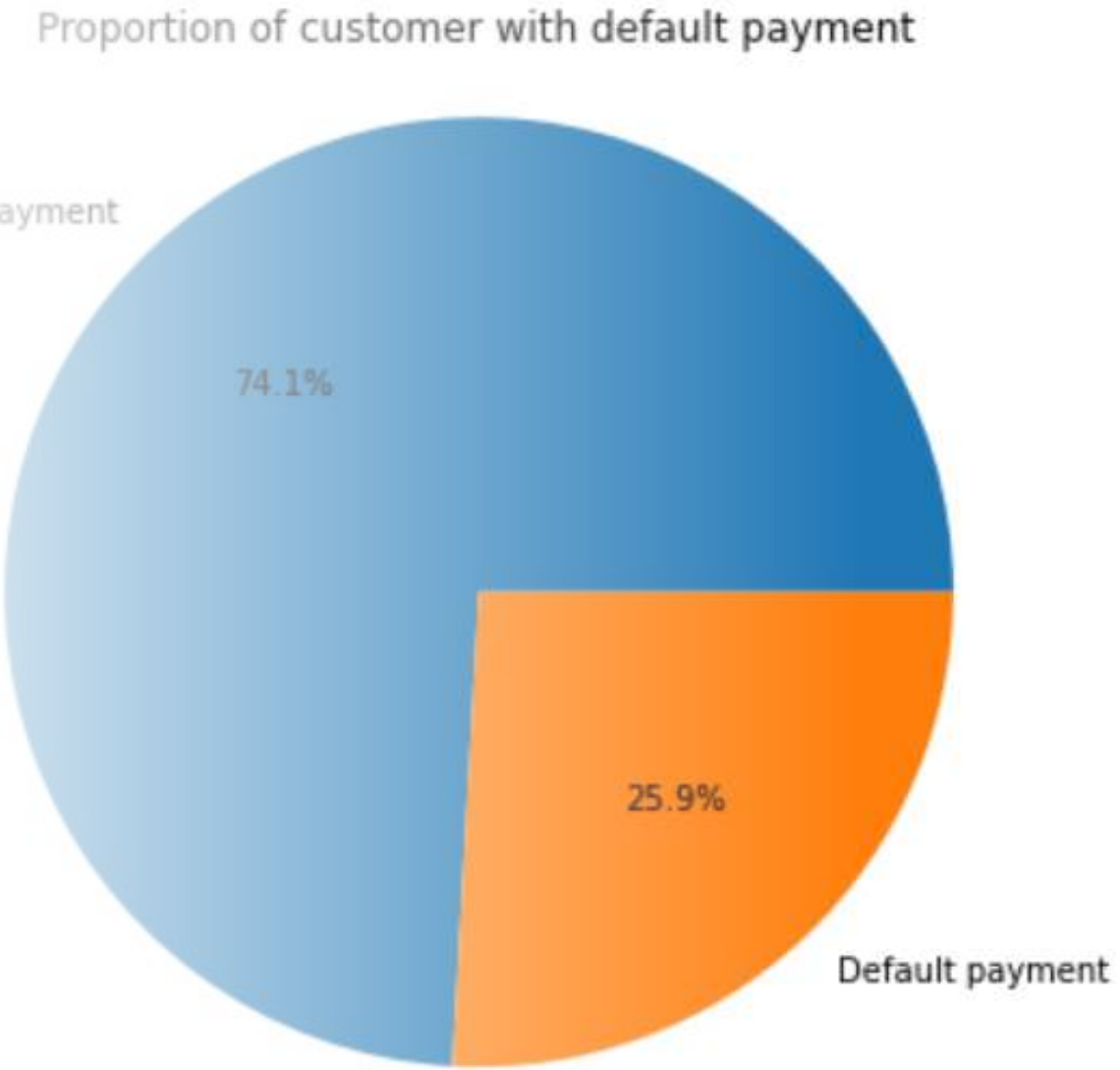
- Prédire la probabilité qu'un client ne rembourse pas le montant du solde de sa carte crédit.
- Variable target pour indiquer un défaut de payement
- Calculée à l'aide des informations des 18 derniers mois
- Si non remboursement après 120 jours, client considéré comme étant en défaut de payement

Métrique d'évaluation

- Métrique particulière mise à disposition
- $M = 0{,}5(G + D)$
- G: coefficient de Gini
- D: taux de défaut capturé à 4%

II. IMPORTATION ET EXPLORATION DES DONNÉES

- Importation à partir de données compressées pour réduire les coûts mémoire
- Beaucoup de valeurs manquantes, probablement lié au cadre de travail
- Pas forcément à considérer comme valeurs manquantes mais comme indicateurs d'évaluation
- Problématique de traitement des valeurs manquantes
- 458913 clients distincts, données bancaires des 13 derniers mois
- Pas de corrélation importante entre les variables



III. MODÈLES DE PRÉDICTION

PRÉPARATION DES DONNÉES

- Données déjà séparées en train/validation/test
- Utilisation de seulement 10% des données (~500 000 entrées) pour les tests afin de réduire les coûts de calcul et d'espace mémoire
- Métriques de comparaison: temps de calcul, Accuracy, métrique fournie par American Express
- Optimisation des hyperparamètres

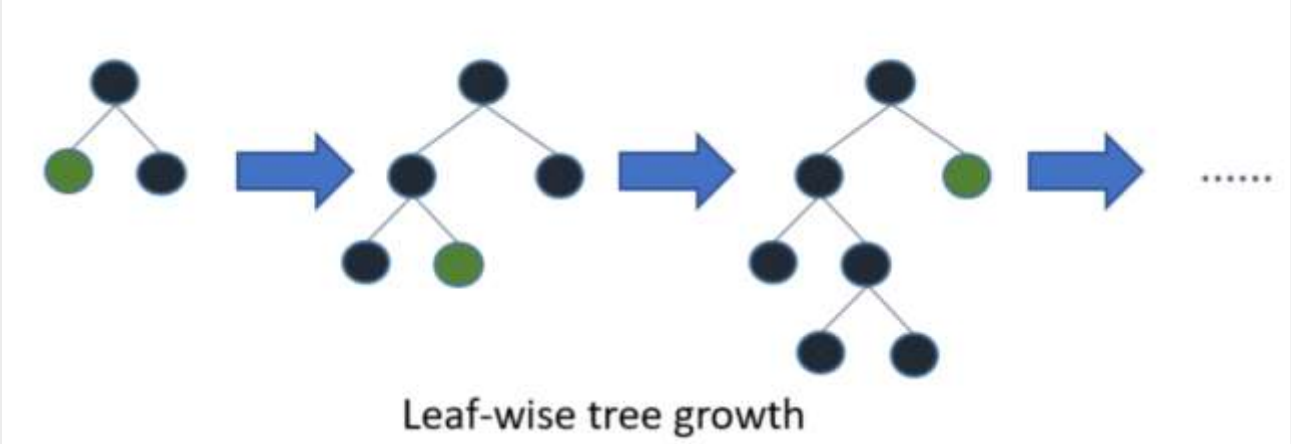
3 types de traitement des valeurs manquantes utilisés:

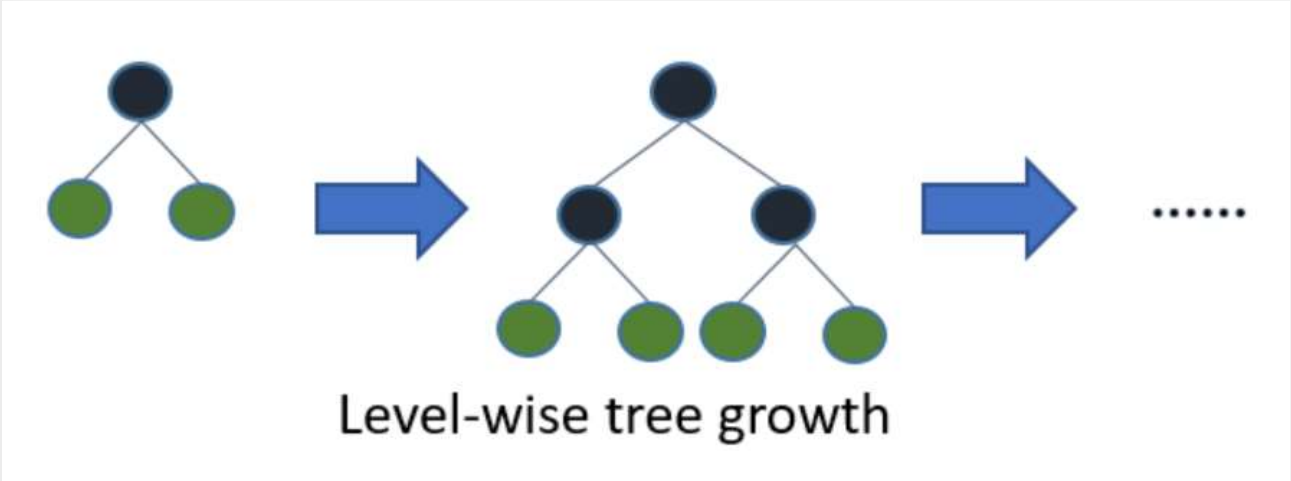
1. Conserver toutes les valeurs manquantes
2. Supprimer toutes les variables avec des valeurs manquantes
3. Conserver et traiter uniquement les variables avec moins de 30% de valeurs manquantes

MODÈLES UTILISÉS

1. LightGBM

- Framework de gradient boosting basé sur les arbres de décision
- Rappel boosting: combinaison d'apprenants faibles sans hypothèse d'indépendance
- Principale caractéristique LGBM: croissance verticale de l'arbre





2. CATBOOST

- Développé par Yandex, géant du numérique Russe
- CatBoost = Categorical Boosting
- Capable de traiter les données catégorielles
- Globalement intéressant sur un jeu de données hétérogènes (variabilité, types différents, valeurs manquantes)

RÉSULTATS

- Résultats similaires mais LGBM plus rapide
- Méthode de suppression de toutes les variables avec des valeurs manquantes à écarter
- Compte tenu du rôle central des banques dans le fonctionnement des sociétés, idéal de conserver toutes les données pour ne pas passer à côté de détails importants
- Modèle retenu: LGBM entraîné sur le jeu original avec toutes les variables et toutes les valeurs manquantes

	LGBM	CatBoost
computation time	80.144679	139.162659
metric score	0.504866	0.506867
accuracy score	0.880013	0.880465

IV. CONCLUSION

- Nouvelles compétences
  - Découverte et participation sur la plateforme Kaggle
- Difficultés rencontrées
  - Limites liées aux capacités mémoire
- Pistes d'améliorations
  - Intégrer une équipe de travail
  - Utiliser le cloud pour gérer la grande quantité de données