



République Tunisienne

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université de Carthage - École Supérieure de la Statistique et de l'Analyse de l'Information



Rapport de Projet de Fin d'Études présenté pour l'obtention du

Diplôme National d'Ingénieur en Statistique et Analyse de l'Information



Ben Abdessalem Mohamed Amine

Prédiction du churn et rétention optimisée par IA pour Ooredoo

Soutenu le 20 Juin 2024 devant le jury composé de :

- Mr Belmufti Ghazi (**Président**)
- Madame Hamdeni Tasnim (**Rapporteur**)
- Mr Marzouki Zouheir , Chef de Service Reporting, Ooreedo Tunisie (**Encadrant entreprise**)
- Madame Masmoudi Lilia (**Encadrante universitaire**)



Année universitaire 2024/2025

Résumé de travail

Ce rapport décrit le développement d'une solution de prédiction et de recommandation pour améliorer la rétention client chez Ooredoo Tunisie. Après un prétraitement avancé (imputation, transformations Box–Cox, normalisation et encodage), nous avons comparé plusieurs stratégies d'équilibrage de classes (SMOTE, Borderline-SMOTE, ADASYN, SMOTETomek) intégrées à un pipeline de validation croisée emboîtée. Différents classifieurs supervisés ont été explorés, notamment la régression logistique pénalisée (L2, ElasticNet) et XGBoost, avec calibration automatique du seuil sur F1-score. Les performances (AUC, précision, rappel, F1) sont présentées sur les jeux d'entraînement, de validation et de test à l'aide de matrices de confusion et de courbes ROC/PR. Enfin, un module de recommandation dynamique s'appuie sur un large modèle de langage (LLM, p. ex. GPT-4) pour générer en temps réel des offres personnalisées selon le profil client (score de churn, historique d'usage, segmentation), avec boucle de feedback continue afin d'optimiser l'impact des campagnes.

Mots-clés : churn prediction, imputation, normalisation, SMOTE, régression logistique, XGBoost, optimisation du seuil, validation croisée emboîtée, LLM, recommandation dynamique, rétention client, apprentissage profond, apprentissage de machines.

Remerciements

Je tiens à exprimer ma profonde gratitude à toutes celles et ceux qui m'ont soutenu tout au long de ce projet.

Mme Lilia Trabelsi Masmoudi, mon encadrante universitaire, pour son soutien, ses conseils éclairés et sa disponibilité.

M. Marzouki Zouheir, mon encadrant en entreprise chez Ooredoo, pour son accompagnement et son expertise.

Je remercie également ma famille pour son amour et son encouragement, et tout particulièrement **M. Noamen ben Abdessalem**, cousin de mon père, pour son aide précieuse à mon intégration au sein d'Ooredoo.

Merci enfin à tous les collaborateurs et amis qui ont contribué, de près ou de loin, à la réussite de ce travail.

Table des matières

Résumé de travail	2
Remerciements	4
1 Présentation de l'entreprise	10
1.1 Introduction	10
1.2 Aperçu de l'entreprise	10
1.3 Structure organisationnelle	11
1.4 L'organigramme de la direction Technologie	11
1.5 Conclusion	12
2 Le churn, un défi silencieux	13
2.1 Introduction générale au churn	13
2.2 Fidélisation et attrition client dans le secteur B2C des télécommunications	13
2.3 Complexité de la gestion du churn	14
2.4 Problématique et objectifs du projet	15
2.5 Comparaison avec les travaux antérieurs	17
2.6 Contribution de la data science et de l'IA à la fidélisation	17
2.7 Concepts clés : Data Mining, Machine Learning et Deep Learning	18
2.7.1 Data Mining :	18
2.7.2 Machine Learning :	19
2.7.3 Deep Learning	20
2.7.4 Generative IA :	20
2.8 Conclusion	21
3 Cadre théorique du projet	22
3.1 Introduction	22
3.2 Techniques de Prétraitement des Données	22
3.2.1 La normalisation :	22
3.2.2 Traitement des valeurs manquantes	27
3.3 Statistiques descriptives	29
3.4 Analyse des Corrélations	31
3.4.1 Corrélation de Pearson	31
3.4.2 Corrélation de Spearman	31
3.4.3 Corrélation de Kendall	32
3.4.4 Matrice de corrélation (heatmap)	32
3.5 Analyse de Survie	33
3.5.1 Fondements théoriques de l'analyse de survie	33
3.5.2 Méthodes d'estimation de la survie	34

3.5.3	Comparaison de groupes et tests statistiques	35
3.5.4	Applications pratiques pour Ooredoo	35
3.5.5	Conclusion	36
3.6	Équilibrage des Classes	36
3.6.1	Techniques d'Oversampling	36
3.6.2	Sous-échantillonnage	38
3.7	Sélection des Caractéristiques (Feature Selection)	40
3.7.1	Méthodes Filtrantes (Filter Methods)	40
3.7.2	Méthodes Enveloppantes (Wrapper Methods)	41
3.7.3	Méthodes Intégrées (Embedded Methods)	42
3.8	Modélisation du churn	42
3.8.1	Régression Logistique	42
3.8.2	Modèle Random Forest	45
3.8.3	Modèle Gradient Boosting (XGBoost, LightGBM, CatBoost)	46
3.8.4	Modèle SVM (Support Vector Machine)	47
3.8.5	Approche de l'apprentissage profond	49
3.9	Prospective de l'apport de l'IA	51
3.9.1	Génération de Données Synthétiques	51
3.9.2	Explicabilité et Interprétabilité via les Transformers	51
3.9.3	Simulation et Test par Génération de Profils Clients	53
3.9.4	Automatisation de la Communication Client par les LLMs	53
4	Réalisation du projet	54
4.1	Introduction	54
4.2	Prétraitement des données	54
4.3	Analyse exploratoire	60
4.4	Analyse de survie	69
4.5	Sélection des variables explicatives	76
4.6	Modélisation du churn	79
4.6.1	Introduction	79
4.6.2	Modèle de régression logistique avec SMOTETomek	79
4.6.3	Analyse de l'apprentissage – Courbes d'apprentissage	80
4.6.4	Modèle Random Forest	81
4.6.5	Modèle LightGBM avec ADASYN et HalvingRandomSearchCV	87
4.6.6	Modèle Decision Tree	90
4.6.7	Modèle CatBoost	93
4.6.8	Modèle SVM	99
4.6.9	Ensemble Soft Voting	102
4.6.10	Modèle de Réseau de Neurones Profond	104
4.6.11	Modèle TabNet	107
4.7	Tableau de bord interactif de prédiction du churn et génération IA	110
4.7.1	Présentation générale de l'application	110
4.7.2	Affichage des résultats multi-modèles	111
4.7.3	Explication et génération d'email automatique via IA générative	111
4.8	Conclusion(Valorisation métier et perspectives)	113
5	Conclusion et Perspectives	114

Liste des tableaux

1.1	Informations relatives à l'entreprise	10
4.1	Description des variables dans le fichier <code>SAMPLECHURN.csv</code>	55
4.2	Description des variables dans le fichier <code>data_bundle_purchase_july2024_jan2025.csv</code>	56
4.3	Description des variables dans le fichier <code>revenus_data.csv</code>	56
4.4	Description des variables démographiques des abonnés	56
4.5	Meilleures méthodes de transformation par groupe de variables mensuelles	68
4.6	Matrice de confusion du modèle Random Forest	82
4.7	Rapport de classification – CatBoost	94

Table des figures

1.1	La répartition des principaux postes de direction	11
1.2	Répartition des différentes directions au sein de la direction Technique	12
2.1	Types de désabonnement (churn)	14
2.2	Architecture de l'application.	16
3.1	Équilibrage des Classes	36
3.2	Algorithme SMOTE	37
3.3	Schéma de l'architecture LLama	52
4.1	Distribution du churn (Yes/No)	60
4.2	Répartition des abonnés selon le genre et la variable churn	60
4.3	Distribution des abonnés selon l'opérateur d'origine et la variable churn	61
4.4	Taux de churn par type d'appareil utilisé	61
4.5	Répartition des abonnés churners et non-churners selon la tranche d'âge	62
4.6	répartition des abonnées selon l'age	63
4.7	Distribution du churn selon la classe d'ancienneté	63
4.8	Taux de churn (%) par classe d'ancienneté	64
4.9	Distribution du churn selon le statut marital	64
4.10	Visualisation croisée de variables explicatives avec le churn	65
4.11	Les corrélations fortes entre variables numériques moyennes	66
4.12	Exemple de distribution avant et après transformation (groupe <i>data_trafic_volume</i>) .	67
4.13	Projection des données sur les deux premières composantes principales.	69
4.14	ACP selon la variable cible 'churn'.	69
4.15	Courbe de survie globale pour l'ensemble des clients	70
4.16	Courbes de survie des clients selon le genre (Kaplan-Meier)	71
4.17	Courbes de survie selon les groupes d'âge (Kaplan-Meier)	72
4.18	Courbe de survie : Divorcés vs Autres	73
4.19	Courbe de survie : Mariés vs Autres	73
4.20	Courbe de survie : Veufs vs Autres	74
4.21	Courbes de survie des clients selon la classe d'ancienneté (Kaplan-Meier)	74
4.22	Top 20 variables impactant le risque de churn (modèle de Cox)	75
4.23	Probabilité de survie au cours du temps pour un client	76
4.24	Importance des variables selon Random Forest	77
4.25	Importance des variables selon Régression Logistique	78
4.26	Importance par permutation	78
4.27	Matrices de confusion – Régression logistique + SMOTETomek (seuil = 0.460)	79
4.28	Courbe ROC – Train vs Test (seuil = 0.460)	80
4.29	Courbes Précision-Rappel – Train et Test (seuil = 0.460)	80

4.30 Courbes d'apprentissage – Précision, Rappel, F1-score (seuil = 0.460)	80
4.31 Courbe ROC du modèle Random Forest	82
4.32 Courbe Precision-Recall du modèle Random Forest	83
4.33 Top 10 des variables importantes selon Random Forest	83
4.34 Courbes d'apprentissage – Précision, Rappel, F1-score (Random Forest)	84
4.35 Mini arbre de décision sur les 3 variables les plus importantes	84
4.36 Matrice de confusion – XGBoost avec ADASYN (test)	86
4.37 Courbe ROC – XGBoost avec ADASYN (test)	86
4.38 Courbe précision-rappel – XGBoost avec ADASYN (test)	86
4.39 Courbes d'apprentissage – F1, Précision et Rappel (XGBoost + ADASYN)	87
4.40 Matrice de confusion — LightGBM sur jeux train et test	88
4.41 Courbe ROC — LightGBM sur le jeu de test	88
4.42 Courbe Précision-Rappel — LightGBM sur le jeu de test	89
4.43 Courbes d'apprentissage — F1, Précision et Rappel pour LightGBM	89
4.44 Top 20 des variables les plus importantes — LightGBM	90
4.45 Matrice de confusion du modèle Decision Tree (jeu de test)	91
4.46 Courbe ROC du modèle Decision Tree	91
4.47 Courbe précision-rappel du modèle Decision Tree	92
4.48 Courbes d'apprentissage – Decision Tree (F1, précision, rappel)	92
4.49 Top 20 des variables les plus importantes – Decision Tree	92
4.50 Visualisation des premiers niveaux de l'arbre – Decision Tree	93
4.51 Matrices de confusion du modèle CatBoost (Train et Test)	94
4.52 Courbe ROC – CatBoost (AUC = 0.987)	94
4.53 Courbe Précision-Rappel du modèle CatBoost	95
4.54 F1-score en fonction du seuil (CatBoost)	95
4.55 Courbe d'apprentissage du modèle CatBoost (F1, précision, rappel)	96
4.56 Top 20 des variables importantes – SHAP	96
4.57 SHAP Summary Plot – CatBoost	97
4.58 Explication SHAP – prédiction individuelle	98
4.59 Explication LIME – individu #0	98
4.60 Courbe ROC du modèle SVM.	100
4.61 Courbe Précision-Rappel du modèle SVM.	100
4.62 Matrice de confusion du modèle SVM.	101
4.63 Diagramme résumé des valeurs illustrant l'importance des variables pour SVM.	101
4.64 Matrice de confusion du Soft Voting.	102
4.65 Courbe ROC du Soft Voting (AUC = 0,988).	103
4.66 Courbe Précision–Rappel sur l'ensemble de test. <i>Interprétation</i> : À des rappels faibles (détection stricte des churners), la précision est proche de 1. À mesure que l'on recherche à détecter plus de churners (recall croissant), la précision diminue graduellement, illustrant le compromis classique entre couverture et fiabilité.	103
4.67 Courbe d'apprentissage du Soft Voting (F1-score).	104
4.68 Évolution de la perte (gauche) et de l'AUC (droite) en fonction des époques	105
4.69 Matrice de confusion sur le jeu de test	106
4.70 Courbe ROC (AUC = 0,980)	106
4.71 Courbe Précision–Rappel sur le jeu de test	107
4.72 Matrice de confusion du modèle TabNet	108
4.73 Courbe ROC du modèle TabNet (AUC = 0.985)	109

4.74 Courbe Précision–Rappel du modèle TabNet	109
4.75 Évolution de la loss d’entraînement au cours des époques	110
4.76 Évolution de l’AUC d’entraînement et de validation	110
4.77 Dashboard des prédictions multi-modèles pour un abonné sélectionné	111
4.78 Exemple d’explication IA générée dynamiquement pour un client à risque	112
4.79 Email personnalisé généré automatiquement pour la rétention client	112