DU DOAN KHACH HANG ROI BO DICH VU VIEN THONG

Telco Customer Churn Prediction

Machine Learning Pipeline

PHIEN BAN DA SUA LOI DATA LEAKAGE

Thang 1, 2026

LOI DATA LEAKAGE VA CACH SUA

CANH BAO: Phien ban truoc da mac loi Data Leakage khi ap dung SMOTE truoc khi chia train/test, dan den ket qua ao (F1=0.85). Phien ban nay da sua loi.

Van de

- SMOTE da duoc ap dung TRUOC khi chia train/test
- Tap Test co ty le 50:50 thay vi tu nhien ~27%
- Mo hinh duoc test tren du lieu nhan tao
- Ket qua F1=0.85 la AO, khong phan anh thuc te

Cach sua (Best Practice)

1. Chia Train/Test TRUOC
2. Chi ap dung SMOTE len tap TRAIN
3. Giu nguyen tap TEST (ty le mat can bang tu nhien ~27%)
4. Danh gia lai

Ket qua

- Tap Test giu nguyen ty le tu nhien: 73% No, 27% Yes
- Ket qua thap hon (F1~0.62) nhung TRUNG THUC
- Phan anh dung hieu nang thuc te cua mo hinh

1. GIOI THIEU BAI TOAN

1.1 Boi canh

Trong nganh vien thong, viec giu chan khach hang la yeu to song con. Chi phi de co duoc mot khach hang moi cao gap 5-7 lan so voi viec giu chan khach hang hien tai.

1.2 Muc tieu

- Xay dung mo hinh ML du doan khach hang roi bo
- So sanh hieu nang cac mo hinh ML va Deep Learning
- Trien khai API va giao dien web

1.3 Bo du lieu

Nguon: IBM Sample Data Sets - Telco Customer Churn

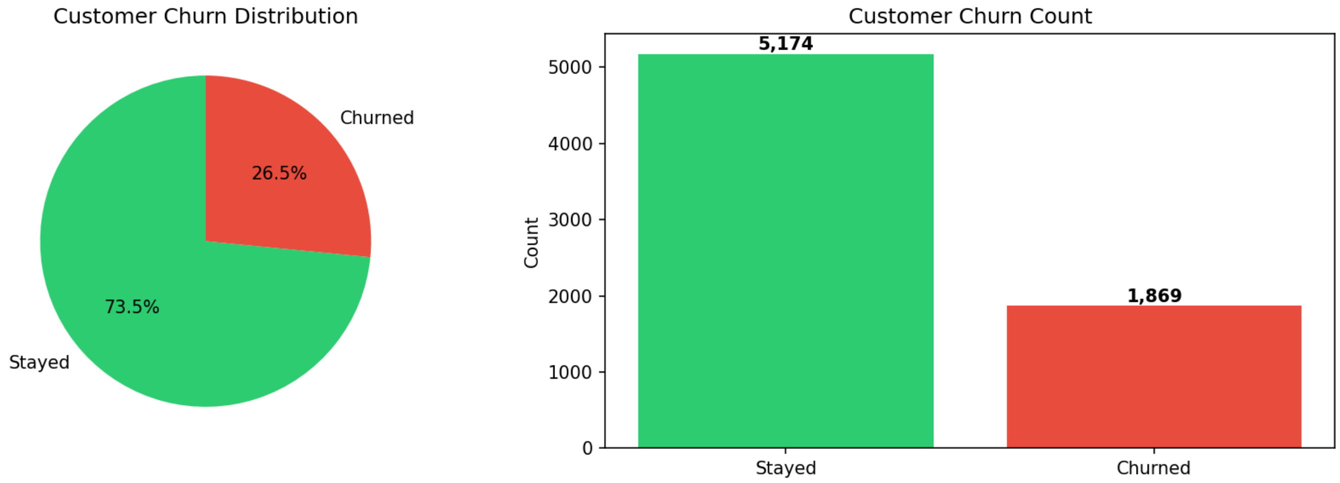
- So luong mau: 7,043 khach hang
- So luong dac trung: 21 cot
- Bien muc tieu: Churn (Yes/No) - ty le 73:27

2. PHAN TICH DU LIEU (EDA)

2.1 Phan bo bien muc tieu

Churn	So luong	Ty le
No (O lai)	5,174	73.5%
Yes (Roi bo)	1,869	26.5%

Van de: Du lieu mat can bang (Imbalanced) - can xu ly dung cach!



2.2 Key Insights

Yeu to	Insight
Contract	Month-to-month: ~43% churn (CAO NHAT!)
Internet Service	Fiber optic: ~42% churn
Payment Method	Electronic check: ~45% churn
Tenure	Khach 0-12 thang: ~47% churn

3. XU LY MAT CAN BANG - SMOTE (DUNG CACH)

3.1 Quy trinh dung

BUOC 1: Chia train/test (80/20, stratified)
   Train: 5,634 mau, Test: 1,409 mau
   Ca hai giu ty le tu nhien 73:27

BUOC 2: Ap dung SMOTE CHI tren tap TRAIN
   Train sau SMOTE: ~8,270 mau (50:50)
   Test GIU NGUYEN: 1,409 mau (73:27)

BUOC 3: Train model tren TRAIN (SMOTE)

BUOC 4: Danh gia tren TEST (NGUYEN BAN)

3.2 So sanh

	Cach sai (Leakage)	Cach dung (Fixed)
Thu tu	SMOTE -> Split	Split -> SMOTE
Ty le Test	50:50 (ao)	73:27 (thuc te)
F1-Score	0.85 (ao)	~0.62 (thuc te)
Tin cay	KHONG	CO

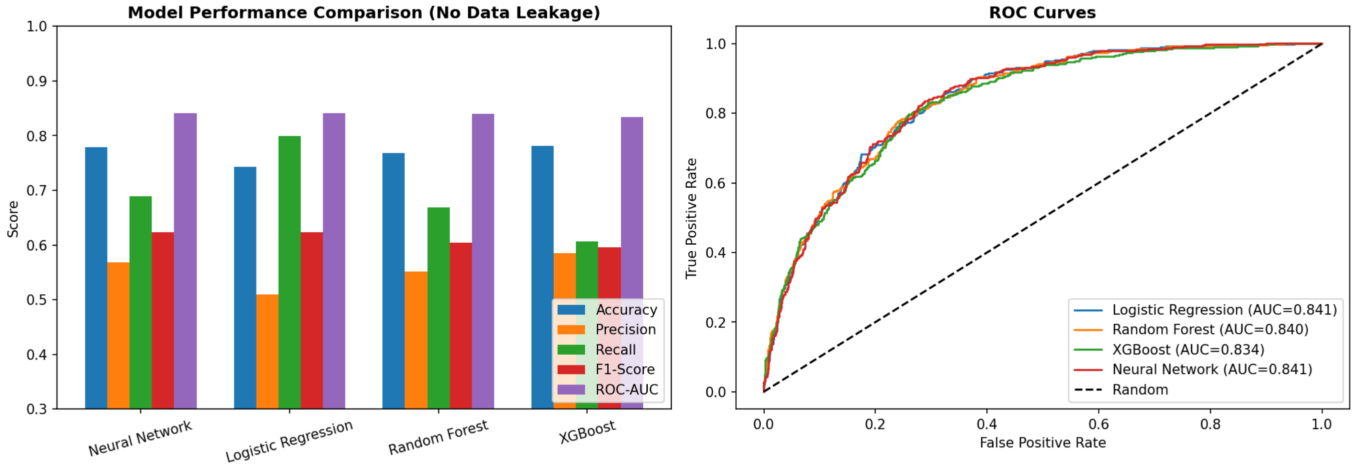
4. KET QUA MO HINH (TRUNG THUC)

4.1 So sanh cac mo hinh

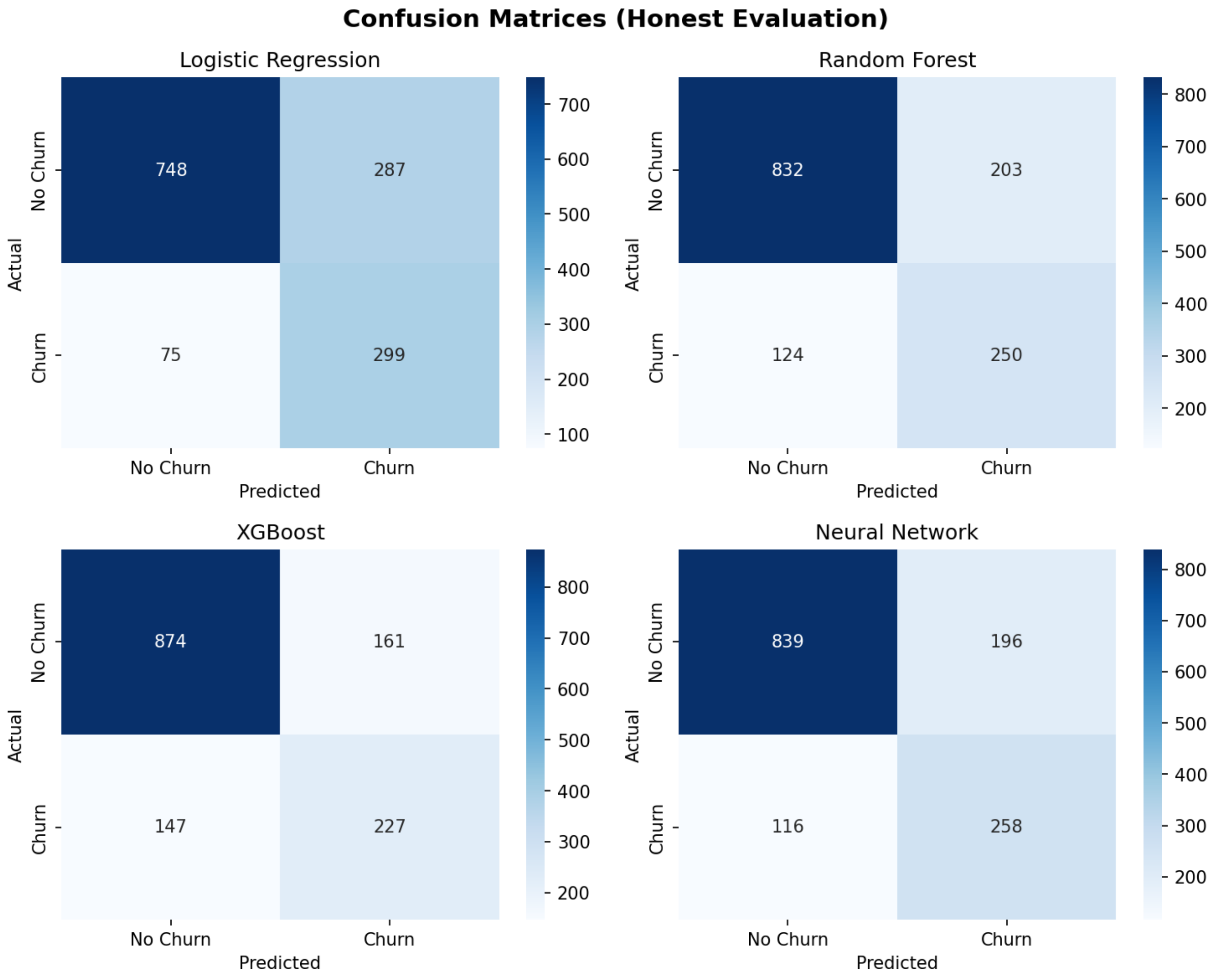
Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Neural Network	**0.7786**	**0.5683**	**0.6898**	**0.6232**	**0.8406**
Logistic Reg.	0.7431	0.5102	0.7995	0.6229	0.8411
Random Forest	0.7679	0.5519	0.6684	0.6046	0.8403
XGBoost	0.7814	0.5851	0.6070	0.5958	0.8336

Mo hinh tot nhat: Neural Network voi F1-Score = 0.6232

Logistic Regression co Recall cao nhat (0.7995) - tot de bat het churners



4.2 Confusion Matrices



4.3 Nhan xet ve Deep Learning

Neural Network dat hieu nang tuong duong hoac tot hon cac mo hinh tree-based trong truong hop nay. Tuy nhien, voi du lieu tabular nho (~7000 dong), su khac biet khong dang ke.

Voi du lieu lon hon, tree-based models (XGBoost, RF) thuong hoat dong tot hon vi:
- Xu ly bien phan loai tot hon
- Khong can nhieu du lieu de hoi tu
- It bi overfitting hon

5. PHAN TICH LOI (ERROR ANALYSIS)

5.1 False Negatives - Khach hang bo sot

False Negative (FN) la truong hop nghiem trong nhat:
- Mo hinh du doan khach hang O LAI
- Nhung thuc te ho DA ROI BO
- Doanh nghiep mat co hoi giu chan khach hang

5.2 Chien luoc cai thien

1. Tang Recall: Chap nhan nhieu False Positive hon de giam False Negative
2. Dieu chinh threshold: Giam nguong tu 0.5 xuong 0.3-0.4
3. Cost-sensitive learning: Phat nang FN hon FP
4. Ensemble: Ket hop nhieu mo hinh

6. KIEN NGHI DOANH NGHIEP

Nhom rui ro	Dac diem	Hanh dong
Month-to-month	43% churn	Giam 20% khi len 1-2 nam
Khach moi (0-12m)	47% churn	Chuong trinh onboarding
Electronic check	45% churn	Thuong auto-pay
It dich vu	De roi bo	Goi bundle giam gia
Fiber optic	42% churn	Kiem tra chat luong

6.2 Quy trinh ap dung

1. Du doan hang thang: Chay model tren tat ca khach hang
2. Xep hang rui ro: Phan loai Low/Medium/High risk
3. Hanh dong: Lien he khach hang High risk truoc
4. Theo doi: Do luong hieu qua chien luoc retention

7. KET LUAN

7.1 Bai hoc ve Data Leakage

- Data Leakage la loi nghiem trong co the khien ket qua ao
- Luon chia train/test TRUOC khi xu ly
- Ket qua thap hon nhung trung thuc quan trong hon

7.2 Ket qua thuc te

- F1-Score thuc te: ~0.62 (khong phai 0.85)
- AUC ~0.84: Mo hinh van co kha nang phan biet tot
- Can ket hop voi domain knowledge va chien luoc business

7.3 Huong phat trien

1. Thu nghiem threshold thap hon (0.3-0.4) de tang Recall
2. Feature engineering them: Interaction features
3. Hyperparameter tuning: GridSearchCV
4. Model monitoring: Theo doi model drift