**Phân tích và dự đoán khả năng khách hàng rời bỏ dịch vụ viễn thông**

Lê Phước Cường – [cuonglp1713@gmail.com](mailto:cuonglp1713@gmail.com)

Mentor: Ứng Kim Phượng – [phuonguk@viettel.com.vn](mailto:phuonguk@viettel.com.vn)

**Tóm tắt**

Nghiên cứu này tập trung phân tích dữ liệu khách hàng để đưa ra được lý do/giải pháp cho câu hỏi vì sao khách hàng lại rời bỏ dịch vụ của công ty/tập đoàn, cùng với đó xây dựng mô hình dự đoán khách hàng nào sẽ rời bỏ dịch vụ vào tháng tới. Kết quả của nghiên cứu đã cho thấy được một số lý do quan trọng khiến khách hàng rời bỏ dịch vụ và huấn luyện được mô hình LGBM-Classifier đạt hiệu năng cao nhất với 82.09% recall, 54.63% precision đối với dữ liệu chưa từng thấy.

**1. Giới thiệu**

**1.1. Giới thiệu chung**

Hiện nay, việc phân tích lý do vì sao khách hàng rời bỏ dịch vụ là bài toán cần thiết cho những công ty/tập đoàn cung cấp dịch vụ. Kết quả của bài toán sẽ không được công ty/tập đoàn công bố rộng rãi vì điều này liên quan đến trực tiếp lợi nhuận của công ty đó.

Nghiên cứu này được triển khai với mong muốn sẽ được đông đảo mọi người quan tâm và có khả năng áp dụng được vào thực tế và giải quyết bài toán khách hàng rời bỏ dịch vụ được đặt ra.

**1.2. Dữ liệu đầu vào**

Dữ liệu được sử dụng cho bài toán này là dữ liệu thu thập được trong 1 tháng của công ty viễn thông nước ngoài gồm dữ liệu lịch sử của 7043 khách hàng với 21 trường. Trường “Churn” sẽ là trường mục tiêu với việc cho biết khách hàng đó có rời bỏ dịch vụ vào tháng sau hay không.

Bảng dữ liệu đầu vào của mô hình sẽ bao gồm:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| customerID | gender | … | TotalCharges | Churn |
|  |  |  |  |  |

Khi thực hiện huấn luyện và kiểm tra, trường customerID sẽ không được sử dụng.

Mô tả dữ liệu đầu vào:

* Churn: Khách hàng có rời dịch vụ vào tháng sau hay không (label)
* customerID: ID đăng kí của mỗi khách hàng khi đăng kí dịch vụ
* *gender, SeniorCitizen, Partner, Dependents*: Thông tin nhân khẩu học của khách hàng bao gồm: giới tính, nhóm tuổi tác (lớn hơn 62 tuổi hay không), có vợ/chồng hay người phụ thuộc hay không
* *tenure*: Số tháng đăng kí dịch vụ của người dùng
* *MonthlyCharges*: Số tiền mà người dùng đã chi cho dịch vụ vào tháng này
* *TotalCharges*: Tổng số tiền mà người dùng đã chi cho dịch vụ của công ty
* *PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod*: Những dịch vụ mà công ty cung cấp để người dùng đăng kí.

**2. Hướng tiếp cận**

**Explainable Machine Learning**

Với bài toàn này, bài nghiên cứu sẽ chọn ra 4 mô hình học máy gồm: Logistic Regression, Random Forest, XGBClassifier, LightGBM.

Với Logistic Regression, vì đây là bài toán phân loại với trường mục tiêu là binary. Nên việc sử dụng Logistic làm thuật toán cơ bản đầu tiên sẽ là rất hợp lý.

Với 3 thuật toán còn lại gồm Random Forest, XGBClassifier và LightGBM, vì bài toán này có bộ dữ liệu bị mất cân bằng với 26.54% là khách hàng rời dịch vụ (so với 73.46% là khách hàng ở lại). Vì thế, việc lựa chọn những mô hình cây kết hợp sẽ giúp tối ưu hiệu năng hết mức có thể.

Qua quá trình huấn luyện và kiểm tra, LightGBM đã đạt được hiệu năng tốt nhất trên bộ dữ liệu này. Tuy nhiên, các thuật toán còn lại cũng đạt được độ hiệu quả tiệm cận LightGBM.

**3. Dataset và Feature**

**3.1. Dataset**

Bộ dữ liệu được chia thành:

|  |  |
| --- | --- |
|  | Data point |
| Bộ huấn luyện | 5634 |
| Bộ kiểm tra | 1409 |

Với bộ huấn luyện gồm 5634 khách hàng, chúng ta sẽ có 4139 khách hàng ở lại với dịch vụ và 1495 khách hàng rời bỏ dịch vụ (chiếm 26.54%).

Với bộ kiểm tra gồm 1409 khách hàng, chúng ta sẽ có 1035 khách hàng ở lại với dịch vụ và 374 khách hàng rời bỏ dịch vụ (chiếm 26.54%).

**3.2. Tăng cường Feature**

*Hình 1: Mất cân bằng dữ liệu*

Thêm các feature mới từ những feature đã có để có thể trích xuất thêm nhiều đặc trưng từ bộ dữ liệu.

|  |  |
| --- | --- |
| **Feature** | **Feature tăng cường** |
| tenure | LoyalCustomer |
| PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, PaperlessBilling | Num Of Services |

Trường dữ liệu *LoyalCustomer* sẽ được tính toán dựa vào trường *tenure* với công thức: Nếu khách hàng đó có hơn 24 tháng đăng kí thì sẽ được tính là LoyalCustomer và ngược lại.

Trường dữ liệu NumOfService sẽ là tổng số dịch vụ tùy chọn mà một khách hàng có thể đăng kí được. Đây là sự tính tổng giữa 10 trường dữ liệu kể trên.

**3.3. Mất cân bằng dữ liệu**

**Chart, bar chart

Description automatically generated**

Hình trên cho chúng ta thấy đây là bài toán có dữ liệu bị mất cân bằng nhẹ với 26.5% cho class 1.

Để xử lý dataset mất cân bằng, bài nghiên cứu đã đưa ra 3 cách: down-sampling bằng NearMiss, up-sampling bằng SMOTE và class-weight. Thông qua training và testing, phương pháp chính được quyết định cho bài nghiên cứu là fit class-weight bằng GridSearchCV.

Class-weight là một phương pháp gắn trọng số cho những class thiểu số, đối với bài toán sẽ là class 1. Chúng ta sẽ cung cấp một list chỉ số gồm: [‘None’, ‘balanced’, {0:0.1, 1:0.9}, {0:0.2, 1:0.8}, {0:0.3, 1:0.7}, {0:0.4, 1:0.6}] và gắn vào hyper-parameter ‘class-weight’ của mỗi mô hình trong lúc training. Đối với XGBClassifier, chúng ta sẽ thay ‘class-weight’ bằng ‘scale\_pos\_weight’.

**3.4. Chuẩn hóa dữ liệu**

Toàn bộ trường dữ liệu thuộc nhóm nhân khẩu học: *gender, SeniorCitizen, Partner, Dependents*, nhóm dịch vụ tùy chọn: *PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod* và trường dữ liệu tự tạo *LoyalCustomer* sẽ được thực hiện one-hot encoding trước khi đưa vào mô hình

Những trường còn lại bao gồm trường chi tiêu: *MonthlyCharges, TotalCharges*, trường *tenure* và trường tự tạo *NumOfService* sẽ được chuẩn hóa bằng MinMaxScaler để đưa về khoảng [0,1].

**4. Phương pháp**

Tất cả các thuật toán đều được áp dụng GridSearchCV kèm với StratifiedKFold với k = 5 để có thể đạt được hiệu năng tốt nhất.

**4.1. LogisticRegression**

Logistic Regression là thuật toán xác định class y cho một điểm dữ liệu x được xác định bằng cách so sánh 2 biểu thức xác suất *P(y=0|x;w) và P(y=1|x;w).* Từ đó xác định được điểm dữ liệu sẽ thuộc về class nào.

Vì đây là bài toàn phân loại binary nên việc chọn Logistic Regression sẽ rất hợp lý.

Với Logistic Regression, bài nghiên cứu thực hiện tuning với các siêu tham số: ‘C’, ‘solver’, ‘penalty’, ‘**class\_weight**’.

Thời gian chạy Logistic Regression là khá nhanh và đạt được hiệu năng tiệm cận LightGBM với 82.35% Recall and 52.47% Precision.

**4.2. Random Forest**

Random Forest là thuật toán tạo ra hàng loạt Decision Tree với các mẫu dữ liệu được tạo ngẫu nhiên, từ đó đưa ra giải pháp tốt nhất.

Với việc tạo rừng ngẫu nhiên và kết hợp phương pháp Bagging, Random Forest sẽ hoạt động tốt với bộ dữ liệu mất cân bằng như trường hợp của bài toán.

Với Random Forest, bài nghiên cứu thực hiện tuning với các siêu tham số: ‘criterion’, **‘class\_weight’**, ‘max\_depth’, ‘min\_sample\_leaf’, ‘min\_sample\_split’, ‘n\_estimators’.

Thời gian chạy của Random Forest là rất lâu (hơn 15) đòi hỏi phải giảm số lượng params đặt vào GridSearchCV. Vì vậy rất khó để tuning mô hình này.

Hiệu năng mà Random Forest cũng khá cao với 80.21% Recall và 53.29% Precision.

**4.3. XGBoost Classifier**

XGBoost là thuật toán sử dụng kĩ thực Boosting dựa trên Gradient Descent. Với sự kết hợp giữa phần mềm và đẩy mạnh sức mạnh của phần cứng sẽ giúp đạt được những kết quả vượt trội cũng như tiếp kiệm thời gian training và bộ nhớ.

Giống với Random Forest, XGBClassifer sử dụng phương pháp học kết hợp nhưng sử dụng phương pháp Boosting thay cho Bagging giúp cho mô hình huấn luyện tốt trên bộ dữ liệu mất cân bằng.

Với XGBoost, bài nghiên cứu thực hiện tuning với các siêu tham số: ‘learning\_rate’, ‘reg\_lambda’, ‘**scale\_pos\_weight**’, ‘max\_depth’, ‘subsample’, ‘colsample\_bytree’, ‘gamma’, ‘n\_estimators’.

Thời gian chạy đối với XGBoost là nhanh hơn rất nhiều và đạt được hiệu năng đứng thứ 2 chỉ sau LightGBM với 80.48% Recall và 55.13% Precision.

**4.4. LightGBM Classifier**

LightGBM là thuật toán cải thiện của XGBoost với hiệu năng còn tuyệt vời hơn đặc biệt trong việc cải thiện tốc độ training và giảm bộ nhớ cần sử dụng với 2 thuật toán: Gradient Based One Side Sampling (GOSS) và Exclusive Feature Bundling (EFB).

Với LightGBM, bài nghiên cứu thực hiện tuning với các siêu tham số: ‘learning\_rate’, ‘n\_estimators’, ‘boosting\_type’, ‘class\_weight’, ‘colsample\_bytree’, ‘reg\_alpha’, ‘reg\_lambda’.

Thời gian chạy đối với LightGBM là rất nhanh và đạt được hiệu năng cao nhất trong 4 phương pháp được liệt kê với 82.09% Recall và 54.63% Precison, giúp cho LightGBM trở thành phương pháp được chọn chính trong bài nghiên cứu này.

**5. Kết quả và thảo luận**

Kết quả của việc sử dụng tất cả các phương pháp trên là mô hình LightGBM với độ chính xác khá cao khi thực hiện kiểm tra với tập kiểm tra

Chart, treemap chart

Description automatically generated

*Hình 2: Confusion matrix*

Mô hình đạt được 82.09% Recall và 54.63% Precision.

Chart, line chart

Description automatically generated

*Hình 3: ROC AUC trên tập kiểm tra*

Mô hình đạt được 0.86 ROC AUC score.

Table

Description automatically generated

*Bảng 1: So sánh các mô hình dựa trên tập kiểm tra*

LightGBM bắt được những trường dữ liệu quan trọng từ bộ dữ liệu:

* Mạnh: *MonthlyCharges, TotalCharges, tenure*
* Trung bình: *Num Of Service, PaperlessBilling*
* Yếu: *PaymentMethod, InternetService*

**6. Kết luận**

Sau khi thực hiện khai phá dữ liệu, những kết luận đã được đề xuất và thảo luận với mong muốn sẽ được những tập đoàn/công ty triển khai.

**7. Hiệu quả kinh doanh**

Bằng những giải pháp đã đưa ra trong quá trình khai phá dữ liệu, nghiên cứu sẽ phần nào giúp giữ lại một số lượng khách hàng có ý định rời bỏ dịch vụ, cải thiện được chất lượng dịch vụ,… Điều này còn cần được áp dụng vào thực tiễn để có thể thấy được hiệu quả trực tiếp, cụ thể:

Về phần khách hàng, công ty/tập đoàn nên tập trung vào nhóm khách hàng là *SeniorCitizen*, nhóm khách hàng có *Partner, Dependents* vì đây là những nhóm người cung cấp lại doanh thu khá cao cho công ty và họ cũng đang có tỉ lệ rời đi khá cao (đối với nhóm khách hàng là *SeniorCitizen*).

Về mặt dịch vụ, công ty/tập đoàn nên chú trọng vào việc nâng cấp chất lượng của gói *Fiber optic* thuộc dịch vụ *InternetService*, gói *Electronic check* thuộc dịch vụ *PaymentMethod* và dịch vụ *PaperlessBilling*. Những gói dịch vụ này mặc dù mang lại lợi nhuận cao cho công ty cùng với lượng khách hàng đăng kí cao, nhưng cùng với đó, tỉ lệ rời đi của những khách hàng này cũng rất cao. Điều này cho thấy chất lượng mà những gói dịch vụ này mang lại đang thật sự chưa xứng đáng với giá tiền.

Cùng với việc nâng cấp/sửa chữa, công ty/tập đoàn cũng nên có chiến dịch quảng bá rộng rãi đến với khách hàng của họ những gói dịch vụ chất lượng, cung cấp nhiều lợi nhuận và giúp ích trong việc giữ chân người dùng ở lại công ty như: gói *Contract* 2 năm, những gói *automatic* thuộc dịch vụ PaymentMethod và những gói dịch vụ liên quan đến bảo mật bao gồm: *Online Security, Online Backup, Device Protection* và *Tech Support*

**8. Công việc tương lai**

Trong tương lai, để có thể tiếp tục với bài toán này theo tính thực tiễn hơn, chúng ta sẽ cần một bộ dữ liệu phức tạp hơn với nhiều trường dữ liệu hơn: trường địa lý, trường thời gian, lý do rời bỏ dịch vụ,…

Sử dụng các phương pháp xử lý mất cân bằng dữ liệu phức tạp và hiệu quả hơn.

Loại bỏ những trường dữ liệu thiếu liên kết với trường mục tiêu để cải thiện thêm thời gian huấn luyện của mô hình.