

**HỌC VIỆN CHÍNH SÁCH VÀ PHÁT TRIỂN**

**KHOA KINH TẾ SỐ**



**BÁO CÁO**

**QUẢN TRỊ DỮ LIỆU LỚN VỚI APACHE SPARK**

**ĐỀ TÀI**

**PHÂN TÍCH KHÁCH HÀNG RỜI BỎ DỊCH VỤ VIỄN THÔNG**



*Hà Nội - Năm 2024*

**MỤC LỤC**

[I. GIỚI THIỆU BÀI TOÁN 4](#_Toc163161009)

[II. NGHIÊN CỨU LIÊN QUAN 6](#_Toc163161010)

[III. KHÁI QUÁT PHƯƠNG PHÁP 14](#_Toc163161011)

[***3.1. Các kỹ thuật nền tảng*** 14](#_Toc163161012)

[***3.2. Đề xuất phương pháp*** 14](#_Toc163161013)

[IV. THỰC NGHIỆM 16](#_Toc163161014)

[***4.1. Miêu tả dữ liệu*** 16](#_Toc163161015)

[***4.2. Tiền xử lý dữ liệu*** 19](#_Toc163161016)

[***4.3. Các độ đo đánh giá hiệu năng*** 23](#_Toc163161017)

[***4.4. Các tham số và môi trường cài đặt*** 24](#_Toc163161018)

[***4.4.1. Tham số*** 24](#_Toc163161019)

[***4.4.2. Môi trường cài đặt*** 24](#_Toc163161020)

[***4.5. Các phương pháp cơ sở*** 24](#_Toc163161021)

[***4.6. Phân tích, so sánh các kết quả*** 24](#_Toc163161022)

[***4.6.1. Mô hình phân loại*** 24](#_Toc163161023)

[***4.6.2. Decision Tree*** 27](#_Toc163161024)

[***4.6.3. Random Forest*** 28](#_Toc163161025)

[***4.6.4. XGBoost*** 29](#_Toc163161026)

[V. KẾT LUẬN 31](#_Toc163161027)

[VI. TÀI LIỆU 33](#_Toc163161028)

***Tóm tắt nội dung*- Thị trường viễn thông đang phát triển mạnh mẽ với nhiều xu hướng mới như 5G, IoT, điện toán đám mây, chuyển đổi số. Tuy nhiên ngày càng cạnh tranh, việc giữ chân khách hàng là một thách thức lớn. Phân tích khách hàng rời bỏ dịch vụ viễn thông đóng vai trò quan trọng trong việc hiểu rõ nguyên nhân khiến khách hàng hủy bỏ dịch vụ, từ đó đưa ra được các chiến lược giữ chân hiệu quả. Đề tài nghiên cứu này ứng dụng sử dụng Apache Spark- một nền tảng xử lý dữ liệu lớn, dùng để phân tích dữ liệu khách hang của các nhà mạng viễn thông. Mục đích là xác định các yếu tố ảnh hưởng đến hành vi rời bỏ dịch vụ của khách hàng từ đó đề xuất các giải pháp giữ chân phù hợp. Trong nghiên cứu này, tôi đề xuất 2 phương pháp phân loại: LightGBM, Logistic** **Regression và 3 mô hình học máy: Decision Tree, Random Forestvà XGBoost . Mô hình đề xuất đã thu được kết quả đánh giá với chỉ số lần lượt gồm mô hình RDF và XGBoost có chỉ số accuracy là 95% cao hơn các mô hình còn lại lần lượt là 1% Decision Tree, 16% LightGBM và 4% Logistic Regression. Hai chỉ số precision cao hơn từ 1% đến 40%, đặc biệt mô hình LightGBM có chỉ số precision và F1 score thấp hơn nhiều so với các mô hình còn lại đạt tỷ lệ lần lượt là 63% và 55%. Qua nghiên cứu nhận ra rằng XGBoost và Random Forest sẽ là hai mô hình tối ưu nhất cho việc dự đoán tỷ lệ khách hàng rời bỏ dịch vụ viễn thông.**

***Từ khóa*: Viễn thông, phân tích khách hàng, khách hàng rời bỏ, rời bỏ dịch vụ, spache spark, machine learning.**

# GIỚI THIỆU BÀI TOÁN

Khách hàng thân thiết là một vấn đề quan trọng đối với các doanh nghiệp viễn thông vì nó ảnh hưởng trực tiếp đến doanh thu và lợi nhuận. Dự đoán rời bỏ là quá trình dự báo hành vi của khách hàng đến việc sử dụng dịch vụ, sản phẩm,.. Trong những năm gần đây nó đã phát triển thành một chủ đề nghiên cứu gây tranh cãi, vì nghiên cứu này sử dụng tập dữ liệu về khách hàng viễn thông để dự đoán khả năng khách hàng rời bỏ dịch vụ. Sự phổ biến ngày càng tăng của các công nghệ phân tích dữ liệu lớn đã giúp giải quyết được nhiều bài toán khác nhau và sự phát triển của ngành khoa học dữ liệu vào những năm gần đây đã dẫn đến việc chúng được sử dụng rộng rãi hơn ở nhiều doanh nghiệp viễn thông. Hơn nữa, nó đặc biệt liên quan vì các doanh nghiệp viễn thông hoạt động trong môi trường toàn cầu hóa, chuyển đổi số đồng nghĩa với việc giữ chân khách hàng ngày càng trở nên khó khăn hơn. Khó khăn cơ bản mà các chủ doanh nghiệp gặp phải là tự động hóa quy trình của doanh nghiệp và phân tích lượng dữ liệu khổng lồ một cách chính xác. Việc xử lý luồng dữ liệu liên tục từ nhiều nguồn là một thách thức lớn. Khả năng dự đoán mức độ tiêu hao của khách hàng dựa trên dữ liệu và tương tác là một yêu cầu quan trọng. Yêu cầu này càng cao đối với các tổ chức có lượng khách hàng lớn, chẳng hạn như tập đoàn viễn thông Viettel. Và mặc dù các doanh nghiệp đầu tư đáng kể vào các hoạt động tiếp thị để có thêm người dùng mới nhưng việc giữ chân khách hàng cũ thường ít tốn kém hơn so với việc có được một khách hàng mới. Vì những lý do này, vấn đề tránh tình trạng khách hàng rời bỏ dịch vụ đã trở thành mối quan tâm lớn đối với các doanh nghiệp viễn thông. Trong bài viết tôi ứng dụng sử dụng Apache Spark- một nền tảng xử lý dữ liệu lớn đã trở nên phổ biến vì hiệu suất vượt trội và hiệu quả mang lại tốt hơn trong việc xử lý lượng dữ liệu khổng lồ nhanh hơn. Mục đích của nghiên cứu này là phân tích dữ liệu khách hang nhằm mục đích xác định được các yếu tố ản hưởng đến hành vi rời bỏ dịch vụ của khách hành, tìm ra được phương pháp tốt nhất để có thể dự báo được khả năng rời bỏ dịch vụ của khách hàng từ đó có thể giúp các tập đoàn viễn thông có biện pháp kịp thời để giữ chân được khách hàng gắn bó với dịch vụ của mình.

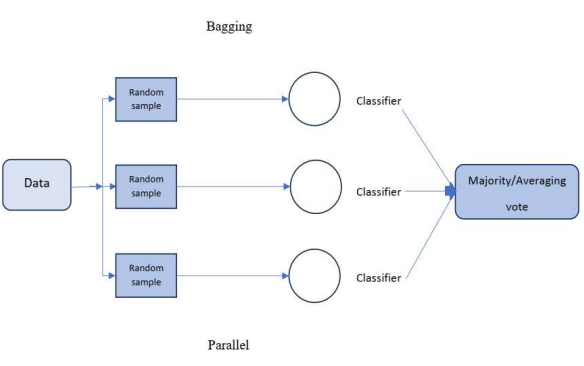
Trong những năm qua lĩnh vực nghiên cứu dời bỏ dịch vụ viễn thông rất được quan tâm tiêu biểu như các nghiên cứu “Khung dự đoán rời bỏ và phân khúc khách hàng tích hợp cho doanh nghiệp Telco” của nhóm tác giả Shuli Wu and Wei-Chuen Yau and Thian Song Ong and Siew-Chin Chong đã đề xuất một khung phân tích khách hàng tích hợp để quản lý tỷ lệ rời bỏ, nghiên cứu “Dự đoán tỷ lệ rời bỏ khách hàng trong ngành viễn thông bằng cách sử dụng Deep Learning” của nhóm tác giả Samah Wael Fujo và Suresh Subramanian và Moaiad Ahmad Khder đề xuất sử dụng hai phương pháp lựa chọn tính năng là Ngưỡng phương sai và Hồi quy Lasso, nghiên cứu “Phân tích tỷ lệ rời bỏ khách hàng bằng cây quyết định XGBoosted” của nhóm tác giả Muthupriya Vasudevan and Revathi Sathya Narayanan and Sabiyath Fatima Nakeeb and Abhishek Abhishek đề xuất một thuật toán học máy có giám sát đã được triển khai bằng Python để thực hiện phân tích bằng cách xây dựng mô hình cây quyết định,… và nhiều nghiên cứu khác nữa.

Tiếp theo một số nghiên cứu liên quan trình bày trong Phần II, Phần III giới thiệu các kỹ thuật và đề xuất phương pháp, Phần IV trinh bày các kết quả thực nghiệm và Phần V trình bày kết luận của nghiên cứu.

# NGHIÊN CỨU LIÊN QUAN

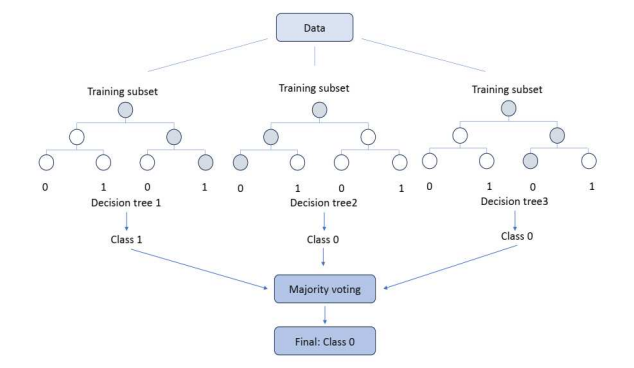
Cho đến nay có rất nhiều nghiên cứu đã đề xuất các kỹ thuật nhằm giải quyết bài toán khách hàng rời bỏ dịch vụ. Trong tài liệu, nhiều kỹ thuật đã được đưa ra để dự báo sự thay đổi trong các doanh nghiệp viễn thông. Học máy và khai thác dữ liệu chủ yếu được sử dụng trong các nghiên cứu này. Trong khi một số nhà nghiên cứu tập trung vào việc sử dụng một kỹ thuật khai thác dữ liệu duy nhất để trích xuất thông tin và hành vi người tiêu dùng, thì những nhà nghiên cứu khác lại chuyên tâm so sánh các phương pháp khác nhau để dự đoán tỷ lệ rời bỏ khách hàng.

Trong nghiên cứu [1] nhóm tác giả khám phá việc ứng dụng các kỹ thuật máy học khác nhau để dự đoán tỷ lệ khách hàng rời bỏ trong lĩnh vực viễn thông. Sử dụng tập dữ liệu công khai để triển khai các mô hình học máy: Mạng nơ-ron nhân tạo (ANN), Cây quyết định (DT), Máy vectơ hỗ trợ (SVM), Rừng ngẫu nhiên (RF), Hồi quy logistic (LR) và các mô hình tăng cường độ dốc (XGBoost, LightGBM và CatBoost).

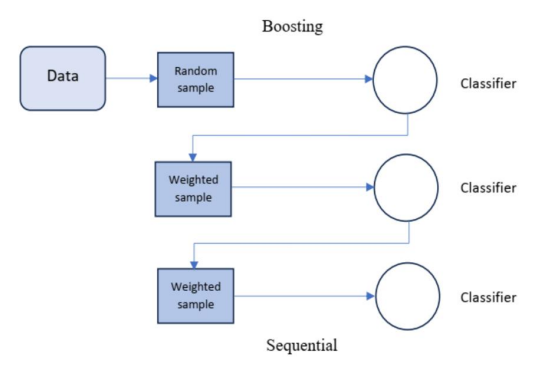


Hình 1. Trực quan hóa phương pháp Bagging

Áp dụng các kỹ thuật lấy mẫu dữ liệu (SMOTE, SMOTE kết hợp Tomek Links, SMOTE kết hợp KNN đã chỉnh sửa) để giải quyết vấn đề mất cân bằng dữ liệu. Sử dụng điều chỉnh siêu tham số (Optuna) để tối ưu hóa hiệu suất mô hình. Đánh giá mô hình bằng các số liệu: Precision, Recall, F1-score và ROC-AUC. Thu về kết quả CatBoost đạt hiệu suất cao nhất về điểm F1 (93%) sau khi tối ưu hóa siêu tham số bằng Optuna. XGBoost và CatBoost đạt hiệu suất cao nhất về ROC-AUC (91%). XGBoost đạt được kết quả này khi kết hợp SMOTE với Tomek Links. CatBoost đạt được kết quả này sau khi tối ưu hóa siêu tham số bằng Optuna.

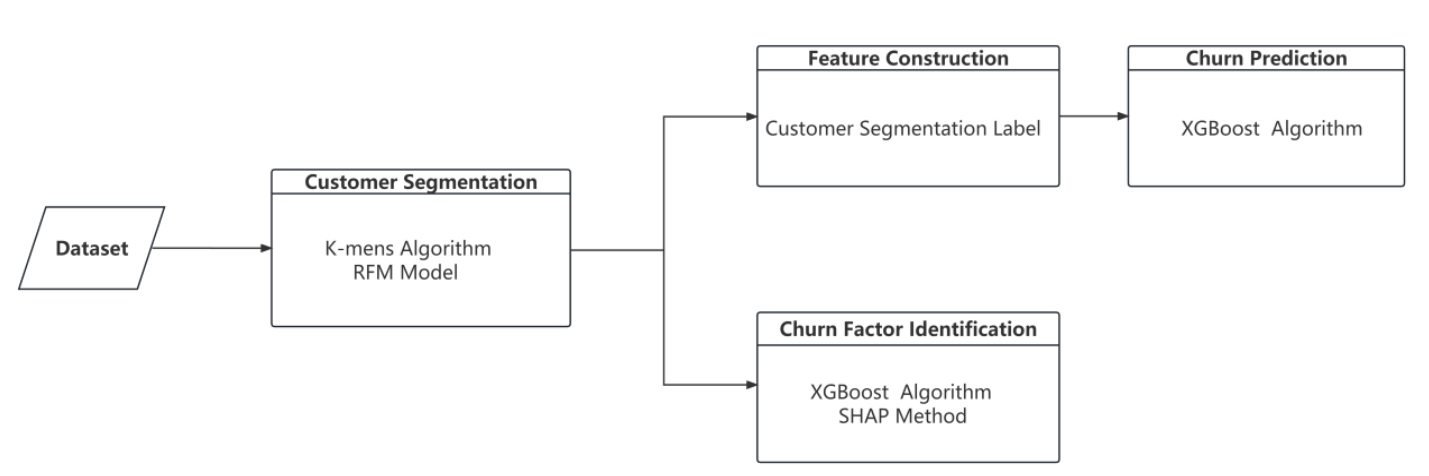


Hình 2. Trực quan hóa bộ phân loại Rừng ngẫu nhiên



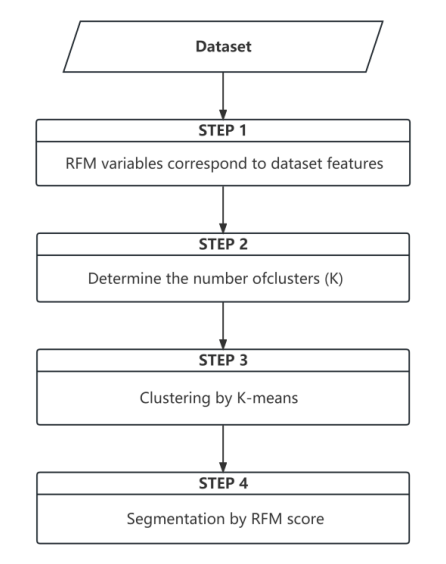
Hình 3. Trực quan hóa phương pháp Boosting

Trong [2] nhóm tác giả đề xuất hệ thống phân tích tỷ lệ rời bỏ khách hàng viễn thông mới dựa trên mô hình RFM để cải thiện độ chính xác dự đoán và giúp các công ty viễn thông giảm tỷ lệ rời bỏ khách hàng. Hệ thống bao gồm 3 phần: phân khúc khách hàng, dự đoán tỷ lệ rời bỏ và xác định yếu tố rời bỏ.



Hình 4. Cấu trúc hệ thống phân tích tỷ lệ rời bỏ khách hàng

Sử dụng mô hình RFM và thuật toán K-means để phân đoạn khách hàng. Sử dụng cấu trúc tính năng dựa trên RFM và thuật toán XGBoost với phương pháp SHAP để dự đoán tỷ lệ rời bỏ và xác định yếu tố rời bỏ. Tập dữ liệu khách hàng rời bỏ nguồn mở bao gồm 7043 phiên bản và 21 tính năng.

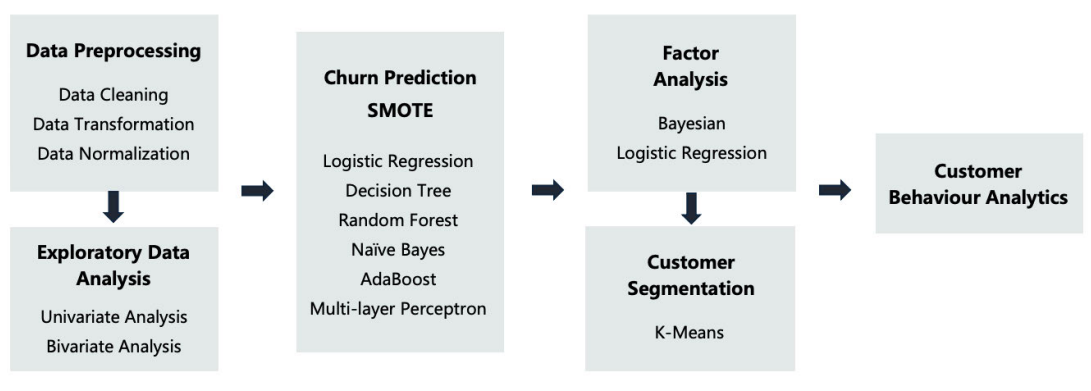


Hình 5. Quy trình phân khúc khách hàng

Hệ thống đạt độ chính xác dự đoán 79,3% trên tập dữ liệu gốc và 81,05% trên dữ liệu mới. Hệ thống mới giúp phân tích xác định được các yếu tố quan trọng góp phần khiến khách hàng rời bỏ cho từng phân khúc khách hàng. Dự đoán tỷ lệ khách hàng rời bỏ chính xác hơn. Tận dụng giá trị khách hàng từ dữ liệu để đề xuất các chiến lược kinh doanh phù hợp.

Trong [3] nhóm tác giả nghiên cứu một mô hình có thể dự đoán được dành cho khách hàng sẽ cho phép các công ty giữ được khách hàng hiện tại và có được khách hàng mới. Đề xuất mô hình Deep-BP-ANN sử dụng hai phương pháp lựa chọn tính năng: Ngưỡng phương sai (Variance Thresholding) và Hồi quy Lasso (Lasso Regression). Áp dụng kỹ thuật dừng sớm (early stopping) để ngăn chặn việc trang bị quá mức. So sánh hiệu quả của mô hình với các kỹ thuật học máy khác (XG\_Boost, Logistic\_Regression, Naïve\_Bayes, KNN) và các kỹ thuật học sâu hiện có. Sử dụng kỹ thuật lấy mẫu ngẫu nhiên để cân bằng tập dữ liệu. Đánh giá hiệu quả mô hình bằng phương pháp giữ lại và xác thực chéo 10 lần. Kết quả cho thấy mô hình Deep-BP-ANN hoạt động tốt nhất với hồi quy Lasso để lựa chọn tính năng, kỹ thuật dừng sớm để chọn kỷ nguyên và số lượng nơ-ron lớn (250) trong lớp đầu vào và lớp ẩn. Mô hình Deep-BP-ANN cho độ chính xác cao hơn các kỹ thuật học máy khác và các kỹ thuật học sâu hiện có trong việc dự đoán tỷ lệ rời bỏ khách hàng.

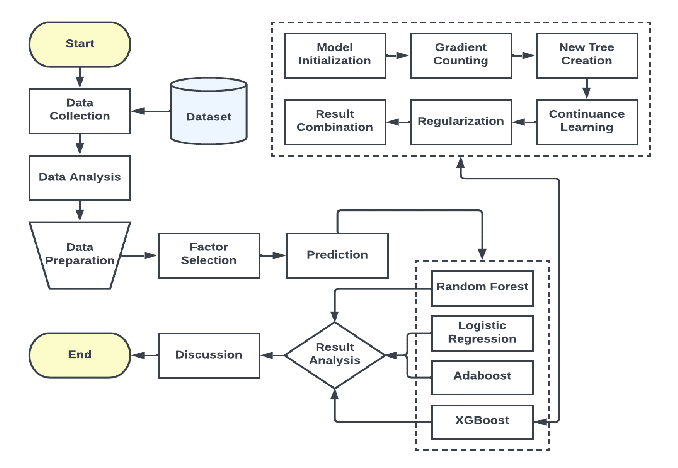
Trong [4] đề xuất một khung phân tích khách hàng tích hợp để quản lý tỷ lệ rời bỏ. Sử dụng 3 bộ dữ liệu và 6 bộ phân loại học máy để dự đoán tỷ lệ rời bỏ.



Hình 6. Khung phân tích khách hàng viễn thông tích hợp

Áp dụng kỹ thuật SMOTE để xử lý tập dữ liệu không cân bằng. Sử dụng xác thực chéo 10 lần để đánh giá mô hình. Sử dụng Hồi quy logistic Bayesian để phân tích nhân tố và tìm ra các tính năng quan trọng cho phân khúc khách hàng. Kết quả: Phân khúc khách hàng bằng thuật toán K-means. AdaBoost hoạt động tốt nhất trong Tập dữ liệu 1 với độ chính xác 77,19% và điểm F1 63,11%. Rừng ngẫu nhiên hoạt động tốt nhất trong Tập dữ liệu 2 với độ chính xác 93,6% và điểm F1 77,20%. Rừng ngẫu nhiên có độ chính xác cao nhất (63,09%) trong Tập dữ liệu 3, Perceptron nhiều lớp có điểm F1 cao nhất (42,84%). Phân tích nhân tố xác định các tính năng quan trọng cho phân khúc khách hàng. Phân khúc K-means giúp phân chia khách hàng thành các nhóm khác nhau để áp dụng chiến lược giữ chân phù hợp.

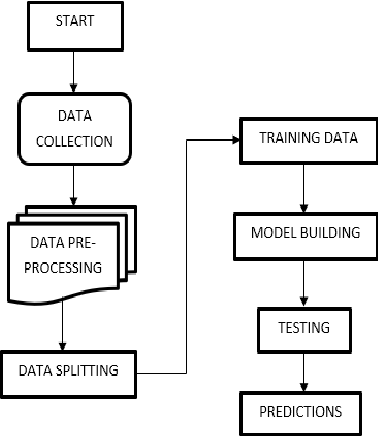
Trong [5] tác giả so sánh hiệu suất của các thuật toán học máy trong việc dự đoán tỷ lệ rời bỏ khách hàng. Xác định các yếu tố ảnh hưởng đến quyết định rời bỏ của khách hàng. Sử dụng tập dữ liệu thông tin khách hàng của ngân hàng.



Hình 7. Luồng nghiên cứu

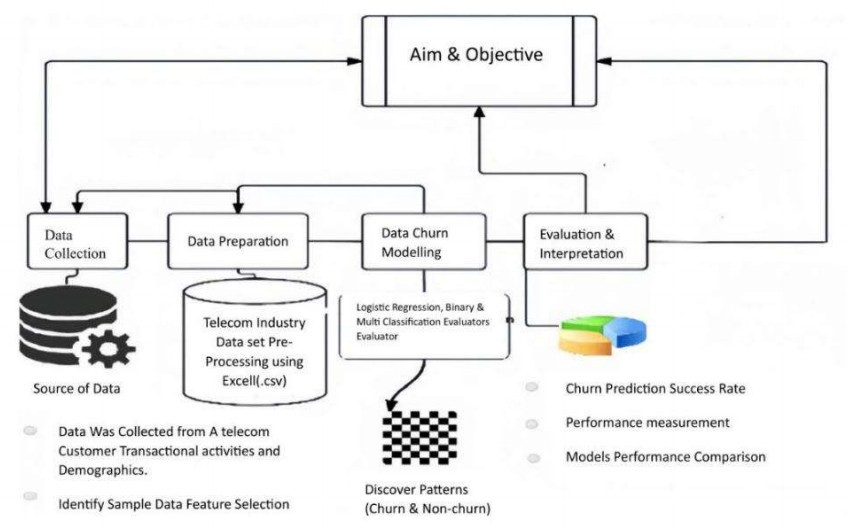
Áp dụng các thuật toán học máy: Rừng ngẫu nhiên, Hồi quy logistic, Adaboost, Tăng cường độ dốc cực cao (XGBoost). Đánh giá hiệu suất bằng độ chính xác và ma trận nhầm lẫn. XGBoost là thuật toán hiệu quả nhất với độ chính xác dự đoán cao nhất. Các yếu tố quan trọng ảnh hưởng đến tỷ lệ rời bỏ: MonthlyCharges, PaperlessBilling, SeniorCitizen, PaymentMethod, MultipleLines, PhoneService.

Trong [6] nhóm tác giả so sánh hiệu suất của các thuật toán học máy khác nhau trong việc dự đoán tỷ lệ rời bỏ khách hàng viễn thông. Đề xuất thuật toán hiệu quả nhất để giúp các công ty viễn thông giảm tỷ lệ rời bỏ khách hàng.



Hình 8. Biểu đồ khối

Sử dụng 4 thuật toán phân loại: Cây quyết định (DT), Rừng ngẫu nhiên (RF), Hồi quy logistic (LR) và Thuật toán kết hợp (HA). Áp dụng các thuật toán cho tập dữ liệu thông tin khách hàng viễn thông bao gồm thông tin nhân khẩu học, xu hướng sử dụng và dữ liệu dịch vụ khách hàng. Đánh giá hiệu suất mô hình bằng các thước đo như accuracy, precision, recall và recall rate. Rừng ngẫu nhiên (RF) đạt độ chính xác cao nhất (95%) trong việc dự đoán tỷ lệ rời bỏ khách hàng. Thuật toán kết hợp (HA) cũng đạt hiệu suất cao (94%). Từ kết quả nghiên cứu thấy được rằng rừng ngẫu nhiên (RF) và Thuật toán kết hợp là những thuật toán hiệu quả để dự đoán tỷ lệ rời bỏ khách hàng trong ngành viễn thông. Các công ty viễn thông có thể sử dụng những thuật toán này để ước tính mức độ tiêu hao khách hàng và thực hiện các biện pháp giữ chân khách hàng hiệu quả.



Hình 9. Sơ đồ dự đoán khách hàng rời bỏ

# KHÁI QUÁT PHƯƠNG PHÁP

***3.1. Các kỹ thuật nền tảng***

* Nền tảng dữ liệu lớn:
* Apache Hadoop là một nền tảng có sẵn dưới dạng nguồn mở được thiết kế để lưu trữ và xử lý các bộ dữ liệu lớn, có kích thược từ Gigabits đến Petabits. Thay vì dựa vào một máy tính lớn duy nhất, Hadoop cho phép phân cụm nhiều máy tính, cho phép phân tích song song các bộ dữ liệu mở rộng hiệu quả hơn
* Spark được tạp ra với mục đích xử lý lượng lớn dữ liệu khổng lồ. Mục tiêu chính của nó là cung cấp cho các ứng dụng dữ liệu lớn với tốc độ tính toán, khả năng mở rộng và khả năng lập trình tuyệt vời.
* Jupyter Notebook có thể tạo và chia sẻ tài liệu với mã trực tiếp, phương trình, đồ họa và văn bản bằng công cụ trực tuyến mã nguồn mở và miễn phí. Cung cấp nhiều chức năng trong các lĩnh vực bao gồm mô hình thống kê, xử lý dữ liệu, học máy và trực quan hóa dữ liệu.
* Xử lý dữ liệu: chuyển đổi dữ liệu, xóa dữ liệu thiếu, xóa cột, làm sạch dữ liệu.
* Phân tích và khám phá dữ liệu: phân tích tương quan, trực quan hóa dữ liệu.
* Kỹ thuật SMOTE để cân bằng dữ liệu.
* Xây dựng và huấn luyện mô hình.
* Đánh giá mô hình dựa trên các chỉ số: accuracy, precision, recall, f1-score.
* Tinh chỉnh siêu tham số hyperparameter tuning.

***3.2. Đề xuất phương pháp***

Trong nghiên cứu này tôi sử dụng thuật toán phân loại và kết hợp giữa các mô hình học máy: Decision Tree, Random Forest và XBoost để dự đoán tỷ lệ khách hàng rời bỏ dịch vụ viễn thông. Qua nghiên cứu các dự án cùng chủ đề tôi nhận thấy rằng mô hình XGBoost và Random Forest có ưu điểm là mang lại hiệu suất dự đoán cao vượt trội so với các phương pháp học máy khác như KNN. Đặc điểm của đề tài teleco customer churn là dữ liệu thường không cân bằng với số lượng khách hàng rời bỏ và số lượng khách hàng không rời bỏ, cả hai mô hình Random Forest và XGBoost đều có khả năng xử lý tốt các bài toán này bằng cách áp dụng các kỹ thuật như Bootstrapping và Weighted Sampling để tạo ra các mô hình mạnh mẽ hơn cho lớp thiểu số. So với mô hình hồi quy tuyến tính hay Decision Tree thì XGBoost và Random Forest có khả năng xử lý các mối quan hệ không tuyến tính và phức tạp tạo điều kiện thuận lợi cho việc phát hiện ra các mẫu phức tạp trong tập dữ liệu. Cả hai mô hình trên đều cung cấp các công cụ để tối ưu hóa siêu tham số tự động giúp giảm thiểu thời gian và công sức trong quá trình tinh chỉnh mô hình, có thể tích hợp nhiều thư viện học máy phổ biến như Scikit-learn và được hỗ trợ trong nhiều ngôn ngữ lập trình khác nhau. Ngoài ra, XGBoost là mô hình rất phù hợp với ngành viễn thông trong việc dự đoán tỷ lệ khách hàng rời bỏ dịch vụ vì có thể phân tích được các yếu tố ảnh hưởng đến quyết định của khách hàng. Mô hình Decision Tree có ưu điểm là có thể tạo ra một cấu trúc cây quyết định giúp người dùng có cái nhìn trực quan hơn, hiễu rõ về dữ liệu và có thể đưa ra các quyết định đúng đắn. Decision Tree có thể xử lý cả dữ liệu phân loại và số mà không cần một quá trình tiền xử lý dữ liệu phức tạp và các dữ liệu thiếu mà không cần các biện pháp điền thêm giá trị. Kỹ thuật phân loại cũng được đánh giá rất phù hợp với bài toán teleco customer churn bởi khả năng dự đoán nhãn của một quan sát dựa vào các biến đầu vào, cung cấp các thông số để đo lường hiệu suất của mô hình trên tập dữ liệu. Mô hình phân loại cũng có thể xử lý các dữ liệu không cân bằng dựa vào các kỹ thuật như Resampling... Có nhiều mô hình phân loại khác nhau có thể được sử dụng tùy thuộc vào đặc điểm của dữ liệu và bài toán, ở trong nghiên cứu này chúng tôi sử dụng hai mô hình phân loại bao gồm Logistic Regression và Light Gradient Boosting Machine (LightGBM) để dự đoán tỷ lệ khách hàng rời bỏ dịch vụ viễn thông.

***Phương pháp được triển khai theo tiến trình:***

*Bước 1: Khai thác dữ liệu*

Quan sát tổng quan, miêu tả dữ liệu, thống kê dữ liệu.

*Bước 2: Chuẩn bị dữ liệu*

* Kiểm tra và xử lý dữ liệu: xóa dữ liệu thiếu, xóa cột, chuyển đổi và làm sạch dữ liệu.
* Xử lý mất cân bằng dữ liệu

*Bước 3: Khám phá dữ liệu*

Trực quan hóa dữ liệu bằng các biểu đồ phù hợp với bài toán.

*Bước 4: Xây dựng và huấn luyện mô hình*

* Chia dữ liệu thành hai tập gồm tập huấn luyện 80% và tập kiểm tra 20%
* Xây dựng và huần luyện mô hình

*Bước 5: Sử dụng các chỉ số để đánh giá, so sánh mô hình.*

# THỰC NGHIỆM

***4.1. Miêu tả dữ liệu***

Bởi vì quy định pháp luật về việc bảo vệ thông tin cá nhân, vậy nên các công ty viễn thông sẽ không được quyền tiết lộ thông tin về tài khoản, giao dịch của khách hàng. Vì vậy trong tài liệu nghiên cứu này sẽ thực nghiệm trên bộ dữ liệu “Telco-Customer-Churn.csv” được tải xuống từ trang Kaggle.com [9] vào ngày 25/01/2024. Tập dữ liệu được tạo hoàn toàn không phụ thuộc bất cứ công ty nào, chỉ mang yếu tố mô tả dữ liệu có thể có trong doanh nghiệp. bộ dữ liệu có mục đích nghiên cứu, thực nghiệm, trong bộ dữ liệu có 19 trường dữ liệu được miêu tả như trong Bảng 1 và 7043 quan sát trong 19 trường dữ liệu.

* Ngữ cảnh: Tập dữ liệu bao gồm hồ sơ thông tin khách hàng và tỷ lệ rời bỏ dịch của họ tại một công ty viễn thông.
* Mục tiêu: Dự đoán những khách hàng có thể ngưng sử dụng dịch vụ viễn thông để đề ra các giải pháp nhằm giảm thiểu tỷ lệ khách hàng rời bỏ dịch vụ viễn thông.
* Bộ dữ liệu bao gồm thông tin:
* Khách hàng đã rời đi trong tháng trước được miêu tả tại cột churn, trong đó có 1869 quan sát là rời bỏ được gắn nhãn là “Yes” chiếm 26,54% và có 5174 quan sát là không rời bỏ được gán nhãn là “No” chiếm 74,46%.
* Các dịch vụ khách hàng đã đăng ký như điện thoại, đường dây, internet, bảo mật trực tuyến, sao lưu trực tuyến, bảo vệ thiết bị, hỗ trợ kỹ thuật và truyền hình trực tuyến.
* Thông tin tài khoản khách hàng: họ đã đã là khách hàng trong bao lâu, hợp đồng, phương thức thanh toán, thanh toán không cần giấy tờ, phí hàng tháng và tổng chi phí đã sử dụng dịch vụ.
* Thông tin nhân khẩu học về khách hàng: giới tính, độ tuổi...

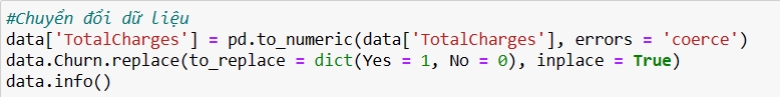
**Bảng 1. Mô tả bộ dữ liệu**

|  |  |  |
| --- | --- | --- |
| **Trường dữ liệu** | **Mô tả** | **Kiểu dữ liệu** |
| CustomerID | Mã khách hàng | object |
| Gender | Giới tính khách hàng | object |
| SeniorCitizen | Khách hàng có phải người cao tuổi (1, 0) | Int64 |
| Partner | Khách hàng có đối tác hay không (Có, Không) | object |
| Dependents | Khách hàng có người phụ thuộc hay không (Có, Không) | object |
| Tenure | Số tháng khách hàng đã gắn bó với công ty | Int64 |
| PhoneService | Khách hàng có dịch vụ điện thoại hay không (Có, Không) | object |
| MultipleLines | Khách hàng có nhiều line hay không (Có, Không, Không có dịch vụ điện thoại) | object |
| InternetService | Nhà cung cấp dịch vụ Internet của khách hàng (DSL, Fiber, Optic, No) | object |
| OnlineSecurity | Khách hàng có bảo mật trực tuyến hay không (Có, Không, Không có dịch vụ Internet) | object |
| OnlineBackup | Khách hàng có sao lưu trực tuyến hay không (Có, Không, Không có dịch vụ Internet) | object |
| DeviceProtection | Khách hàng có bảo vệ thiết bị hay không (Có, Không, Không có dịch vụ Internet) | object |
| TechSupport | Khách hàng có được hỗ trợ kỹ thuật hay không (Có, Không, Không có dịch vụ Internet) | object |
| StreamingTV | Khách hàng có truyền hình trực tuyến hay không (Có, Không, Không có dịch vụ trực tuyến) | object |
| StreamingMovies | Khách hàng có đăng ký phim hay không (Có, Không, Không có dịch vụ trực tuyến) | object |
| Contract | Thời hạn hợp đồng của khách hàng | object |
| MonthlyCharges | Số tiền khách hàng phải trả hàng tháng | float64 |
| TotalCharges | Tổng số tiền phải trả | object |
| Churn | Khách hàng có rời bỏ hay không | object |

***4.2. Tiền xử lý dữ liệu***

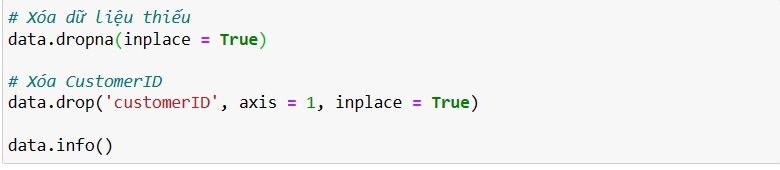
Xử lý trước dữ liệu là một cách tiếp cận quan trọng để chuyển đổi dữ liệu thô sang định dạnh mà thuật toán học máy có thể sử dụng. Nó đóng vai trò là bước đầu tiên trong việc sử dụng các thuật toán học máy. Trong quá trình nghiên cứu học máy, điều bắt buộc là phải làm sạch và cấu trúc lại dữ liệu, đồng thời quá trình làm sạch và định dạng này là bắt buộc đối với mọi thao tác dữ liệu. Tập dữ liệu bị thiếu giá trị, dữ liệu nhiễu, hoặc trình bày một cách lộn xộn cản trở việc sử dụng trực tiếp đến việc áp dụng các mô hình học máy. Việc xử lý dữ liệu trước trở nên cần thiết để làm sạch và sắp xếp dữ liệu, từ đó chuẩn bị cho việc sử dụng hiệu quả mô hình học máy. Quá trình này cuối cùng sẽ năng cao tính chính xác và hiệu quả của mô hình. Yêu cầu quan trọng nhất để phát triển mô hình học máy là tập dữ liệu vì chức năng của mô hình hoàn toàn dựa vào dữ liệu được cung cấp. Và mỗi tập dữ liệu có những đặc điểm riêng.

* Chuyển đổi dữ liệu:





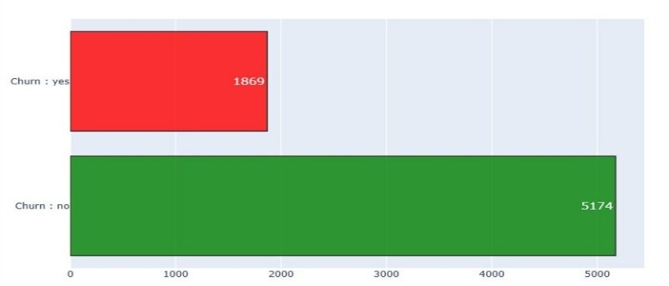
* Xóa dữ liệu:



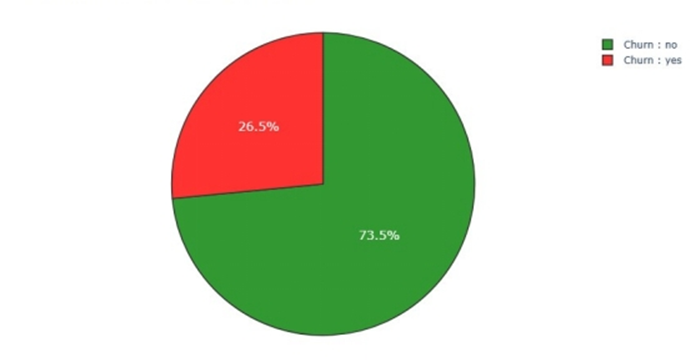
* Tách dữ liệu:



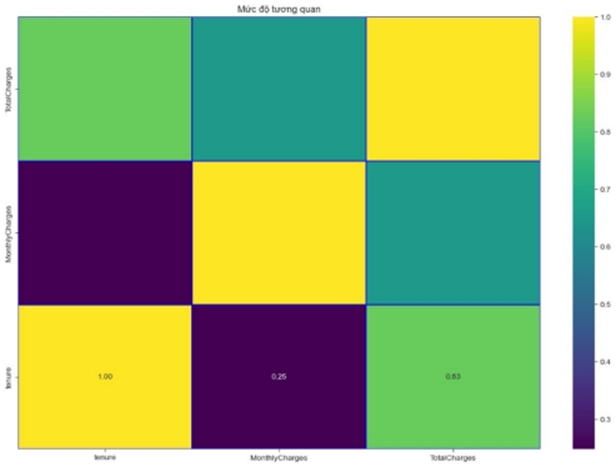
* Khám phá dữ liệu EDA:
* Biểu đồ cột đếm số khách hàng có rời bỏ hay không



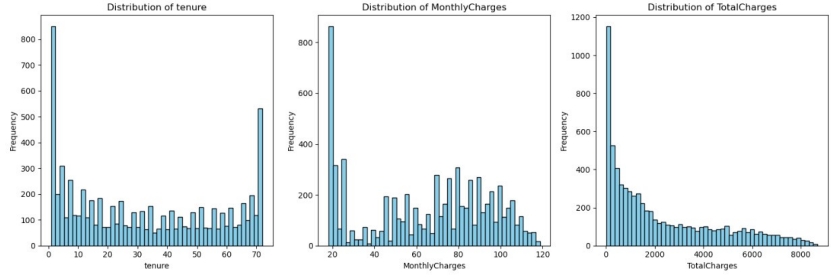
* Biểu đồ thể hiện phần trăm churn



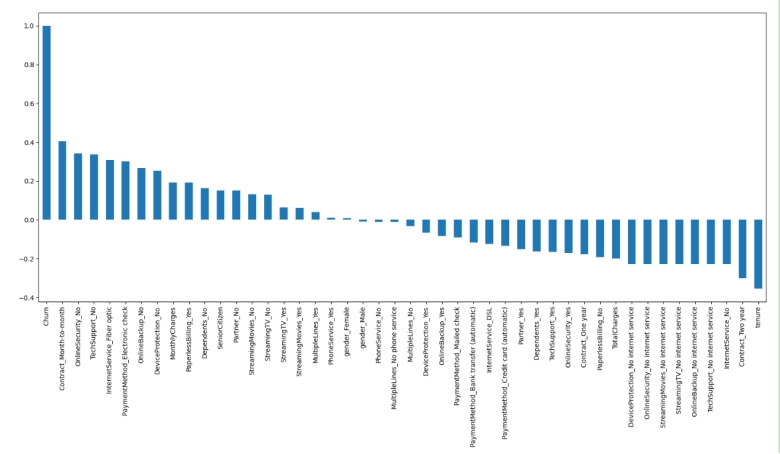
* Biểu đồ thể hiện sự tương quan giữa 3 biến

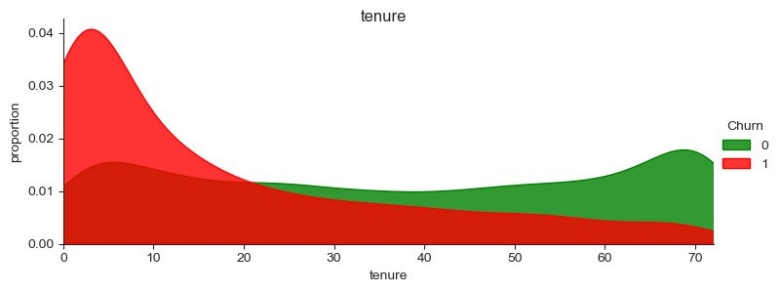
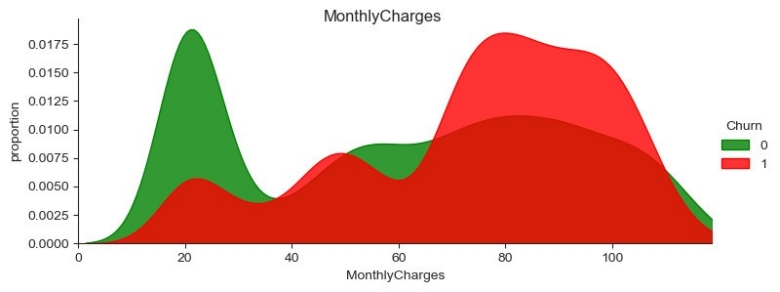


* Biểu đồ histogram

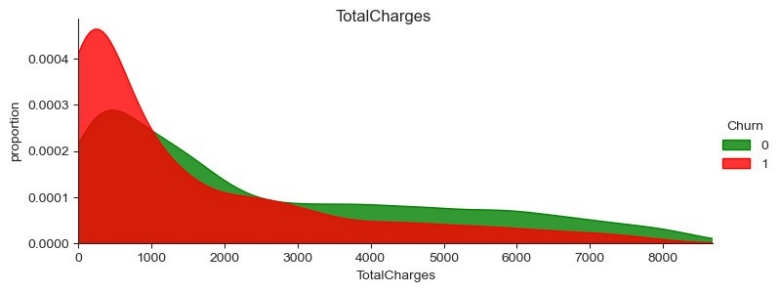


* Sự tương quan giữa các biến



* ******Biểu đồ phân phối thể hiện tỷ lệ khách hàng rời bỏ dựa vào tenure (số tháng đã gắn bó)
* Tỷ lệ khách hàng rời bỏ dựa vào số tiền khách hàng phải trả hàng tháng

* Tỷ lệ khách hàng rời bỏ dựa vào tổng số tiền phải trả



***4.3. Các độ đo đánh giá hiệu năng***

Trong nghiên cứu này chúng tôi sử dụng bốn chỉ số bao gồm accuracy, precision, recall và F1-score để đánh giá mô hình.

* Precision là độ chính xác đo tần suất bộ phân loại đúng cho cả trường hợp dương tính đúng và âm sai.



* Recall là độ nhạy cảm hoặc hồi tưởng do số lần bộ phân loại nhận được kết quả dương tính đúng. Chỉ số này càng cao sẽ càng tốt.



* Accuracy là độ chính xác của mô hình, là tỷ lệ giữa số lượng mẫu được phân loại chính xác trên cho tổng số mẫu của tập dữ liệu huấn luyện.



* F1 score là trung bình trọng số của Precision và Recall.



Trong đó:

* TP (True Positive): tích cực đúng, tổng số trường hợp dự báo khớp mẫu đúng.
* TN (True Negative): tiêu cực đúng, tổng số trường hợp dự báo khớp mẫu sai.
* FP (False Positive): tích cực sai, tổng số trường hợp dự báo các quan sát thuộc mẫu đúng nhưng tính thành sai.
* FN (False Negative): tiêu cực sai, tổng số trường hợp dự báo các quan sát thuộc mẫu sai nhưng tính thành đúng.

***4.4. Các tham số và môi trường cài đặt***

***4.4.1. Tham số***

* Sử dụng tham số parameters, model, criterion, random\_state, max\_depth, min\_samples\_leaf
* Siêu tham số hypererparamete
* Tham số n\_estimators, learning\_rate, subsamrat, subplots.

***4.4.2. Môi trường cài đặt***

* Ngôn ngữ lập trình Python.
* Thư viện Scikit-learn: cung cấp công cụ xây dựng và huấn luyện mô hình.
* Thư viện Pandas: cung cấp cấu trúc dữ liệu và công cụ phân tích dữ liệu
* Thư viện Numpy: hỗ trợ các phép tính số học.
* Thư viện mMatplotlib và seaborn: trực quan hóa dữ liệu bằng các biểu đồ.
* XGBoost và LightGBM: xử lý dữ liệu.

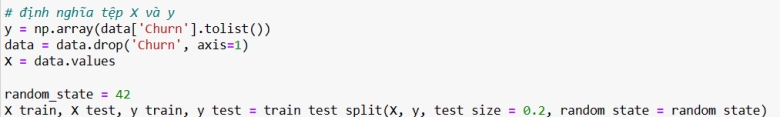
***4.5. Các phương pháp cơ sở***

* Phương pháp phân loại: LightGBM, Logistic Regression.
* Kết hợp các mô hình học máy:
* Decision Tree.
* Random Forest.
* XGBoost.

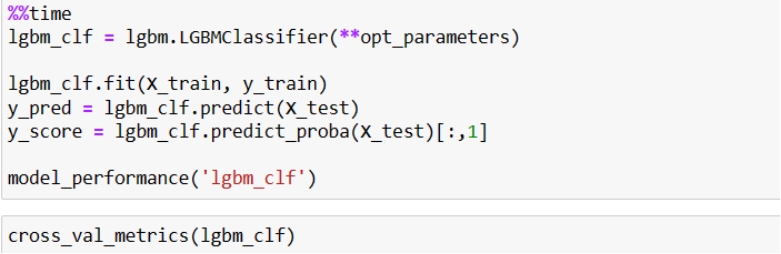
***4.6. Phân tích, so sánh các kết quả***

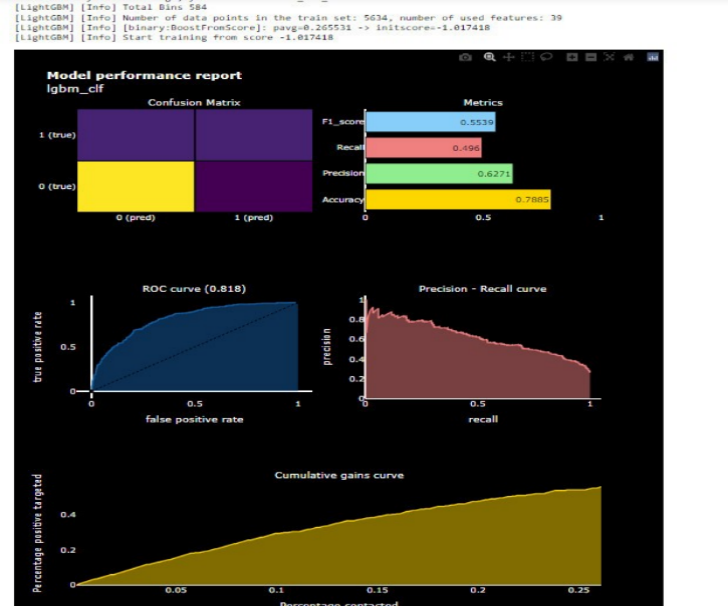
***4.6.1. Mô hình phân loại***

* Mô hình LightGBM:
  + Chia tập dữ liệu:



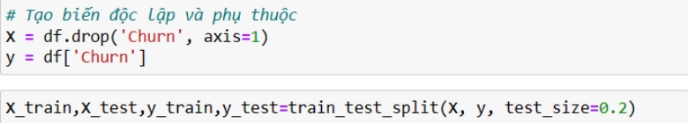
* + Huấn luyện mô hình:



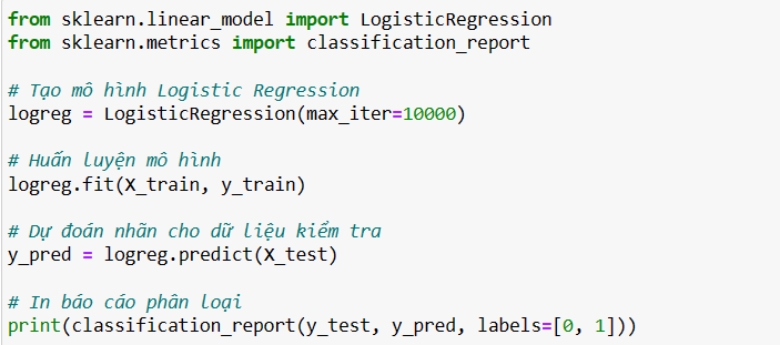


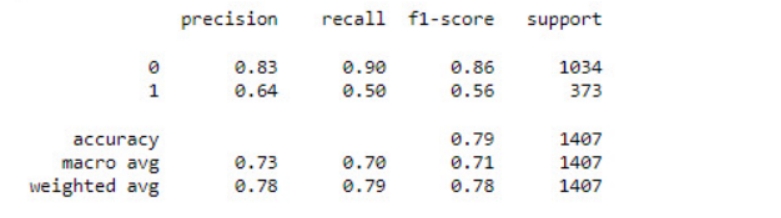
Bộ dữ liệu được chia theo tỷ lệ 80:20 với tập huấn luyện 80% và tập kiểm tra 20%, tỷ lệ này giúp mô hình đạt được hiệu quả tốt vì tỷ lệ giúp dữ liệu được chia không có phương sai quá cao.

* Logistic Regression
  + Chia tập dữ liệu:



* + Xây dựng và huấn luyện mô hình:



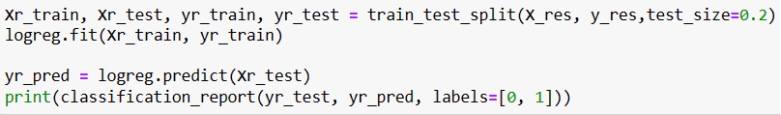


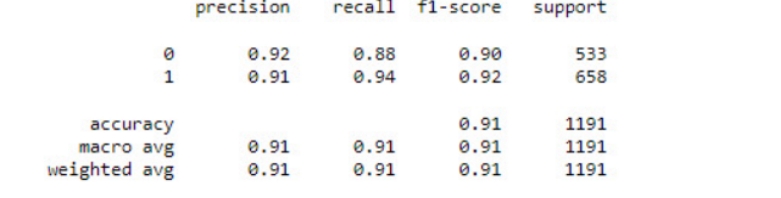
=> Mô hình Logistic Regression dự đoán đúng 83%% các trường hợp khách hàng không rời bỏ với recall khoảng 90% nghĩa là mô hình có khả năng tốt trong việc nhận diện những khách hàng không rời bỏ dịch vụ viễn thông. Tuy nhiên tỷ lệ dự đoán khách hàng rời bỏ dịch vụ viễn thông lại thấp hơn khoảng 64% với recall 50% điều này cho thấy mô hình cần cải thiện khả năng dự đoán tỷ lệ khách hàng rời bỏ dịch vụ viễn thông vì đây mới là mục tiêu của dự án. Độ chính xác của mô hình là 0,81 tương đương khoảng 81% cho thấy tỷ lệ dự đoán đúng của mô hình khá cao.

* + Xử lý dữ liệu mất cân bằng:



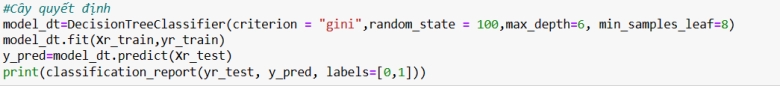
* + Xây dựng và huấn luyện mô hình sau khi xử lý dữ liệu mất cân bằng:

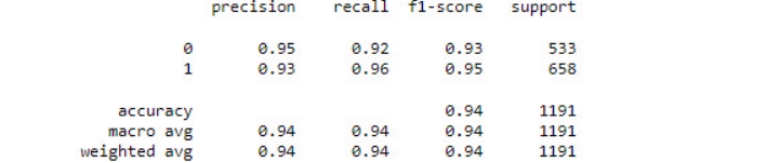




=> Sau khi xử lý dữ liệu mất cân bằng chúng tôi thấy rằng độ chính xác của mô hình đã tăng lên đạt khoảng 91%. Precision của mô hình dự đoán tỷ lệ khách hàng rời bỏ dịch vụ viễn thông là 91% vẫn thấp hơn so với 92% tỷ lệ khách hàng không rời bỏ dịch vụ viễn thông. Tuy nhiên recall của lớp 1 tức là tỷ lệ khách hàng rời bỏ dịch vụ viễn thông cao hơn so với recall của lớp 0 là 88%. F1 score của mô hình khoảng 91% đảm bảo mô hình không chỉ tập trung vào việc dự đoán chính xác một lớp mà còn phát hiện được nhiều mẫu thuộc lớp đó.

***4.6.2. Decision Tree***

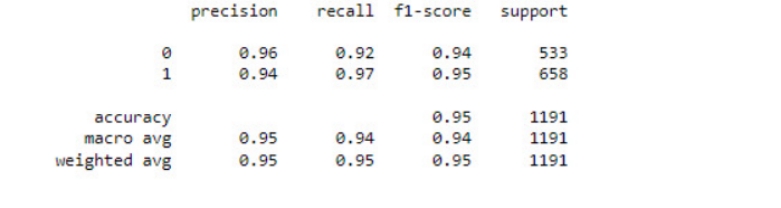




=> Mô hình Decision Tree dự đoán đúng 95% các trường hợp khách hàng không rời bỏ với recall khoảng 92% nghĩa là mô hình có khả năng tốt trong việc nhận diện những khách hàng không rời bỏ dịch vụ viễn thông. Tuy nhiên tỷ lệ dự đoán khách hàng rời bỏ dịch vụ viễn thông lại thấp hơn khoảng 93% với recall 96% nhưng đây vẫn là một tỷ lệ khá cao.

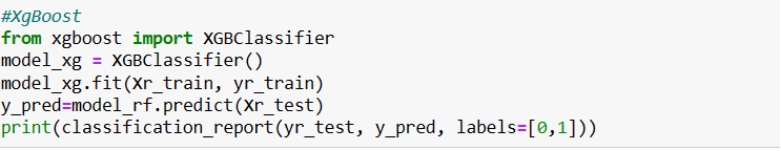
***4.6.3. Random Forest***

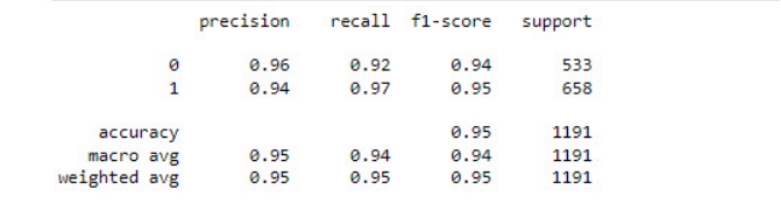




=> Mô hình Random Forest dự đoán đúng 96% các trường hợp khách hàng không rời bỏ với recall khoảng 92% nghĩa là mô hình có khả năng tốt trong việc nhận diện những khách hàng không rời bỏ dịch vụ viễn thông. Tỷ lệ dự đoán khách hàng rời bỏ dịch vụ viễn thông lại thấp hơn khoảng 94% với recall 97% nhưng đây vẫn là một tỷ lệ khá cao. Với chỉ số precision thì mô hình có độ chính xác trong dự đoán tỷ lệ khách hàng không rời bỏ cao hơn so với tỷ lệ khách hàng rời bỏ dịch vụ viễn thông. Tuy nhiên mô hình này lại có recall trong việc dự đoán khách hàng rời bỏ cao hơn có nghĩa mô hình sẽ phát hiện được nhiều mẫu positive. Độ chính xác của mô hình đạt tỷ lệ cao khoảng 95%.

***4.6.4. XGBoost***





=> Mô hình XGBoost cũng có kết quả tương tự như Random Forest dự đoán đúng 96% các trường hợp khách hàng không rời bỏ với Recall khoảng 92% nghĩa là mô hình có khả năng tốt trong việc nhận diện những khách hàng không rời bỏ dịch vụ viễn thông. Tỷ lệ dự đoán khách hàng rời bỏ dịch vụ viễn thông lại thấp hơn khoảng 94% với Recall 97% nhưng đây vẫn là một tỷ lệ khá cao. Với chỉ số Precision thì mô hình có độ chính xác trong dự đoán tỷ lệ khách hàng không rời bỏ cao hơn so với tỷ lệ khách hàng rời bỏ dịch vụ viễn thông. Tuy nhiên mô hình này lại có Recall trong việc dự đoán khách hàng rời bỏ cao hơn có nghĩa mô hình sẽ phát hiện được nhiều mẫu positive. Độ chính xác của mô hình đạt tỷ lệ cao khoảng 95%.

Kết quả: Dưới đây là bảng so sánh các chỉ số đánh giá hiệu năng giữa các mô hình RDF, Decision Tree, XGBoost, LighGBM, Logistic Regression.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **RDF** | **Decision Tree** | **XGBoost** | **LightGBM** | **Logistic Regression** |
| **Accuracy** | 0.95 | 0.94 | 0.95 | 0.79 | 0.91 |
| **Recall** | 0.97 | 0.96 | 0.97 | 0.5 | 0.94 |
| **Precision** | 0.94 | 0.93 | 0.94 | 0.63 | 0.91 |
| **F1-score** | 0.95 | 0.95 | 0.95 | 0.55 | 0.92 |

=> Mô hình RDF và XGBoost có chỉ số Accuracy là 95% cao hơn các mô hình còn lại lần lượt là 1% Decision Tree, 16% LightGBM và 4% Logistic Regression. Hai chỉ số precision cao hơn từ 1% đến 40%, đặc biệt mô hình LightGBM có chỉ số precision và F1-score thấp hơn nhiều so với các mô hình còn lại đạt tỷ lệ lần lượt là 63% và 55%.

# KẾT LUẬN

Dựa vào kết quả của nghiên cứu, tôi đã thực hiện một phân tích sâu về mô hình hóa tỷ lệ khách hàng rời bỏ dịch vụ trong ngành viễn thông. Trong quá trình này, chúng tôi đã tiến hành đánh giá hiệu suất của các mô hình phổ biến như LightGBM, Logistic Regression, Decision Tree, XGBoost và Random Forest.

Đối với hai mô hình phân loại truyền thống LightGBM và Logistic Regression, chúng tôi đã phát hiện rằng mô hình phân loại LightGBM dù có tốc độ huấn luyện nhanh, khả năng xử lý các biến đầu vào phức tạp, xử lý dữ liệu mất cân bằng để đưa ra các quyết định chính xác nhưng trong nghiên cứu này LightGBM lại không mang lại hiệu suất cao cho bài toán, các chỉ số đánh giá của LightGBM không đạt mức độ như chúng tôi mong muốn khi chỉ dao động từ 50% đến 80%. Logistic Regression là một mô hình mặc dù đơn giản và dễ hiểu tuy nhiên mô hình này không thể hoạt động tốt trên dữ liệu không cân bằng nên cần trải qua quá trình xử lý dữ liệu mất cân bằng. Hơn nữa Logistic Regression có thể phản ánh không đúng mối quan hệ giữa các biến và tỷ lệ khách hàng rời bỏ dịch vụ viễn thông. Decision Tree, mặc dù đơn giản và dễ giải thích tuy nhiên mô hình này lại không phù hợp với các tập dữ liệu để lớn, đa chiều và phức tạp như teleco customer churn, nếu cần thiết phải sử dụng Decision Tree cần được kết hợp với nhiều phương pháp khác để cải thiện hiệu suất.

Qua nghiên cứu chúng tôi nhận ra rằng XGBoost và Random Forest sẽ là hai mô hình tối ưu nhất cho việc dự đoán tỷ lệ khách hàng rời bỏ dịch vụ viễn thông. XGBoost có khả năng xử lý tốt các tập dữ liệu lớn, các tham số có thể điều chỉnh để xử lý dữ liệu mất cân bằng, làm việc tốt với các biến đầu vào phức tạp, đặc biệt là có tốc độ và huật suất cao giúp giảm thời gian huấn luyện mô hình và tăng cường hiệu suất dự đoán. Trong khi đó Random Forest cũng là một mô hình có khả năng xử lý xử liệu lớn và phúc tạp, Random Forest không đòi hỏi nhiều công đoạn cấu hình và tinh chỉnh, hiệu suất dự đoán của mô hình này cũng rất cao mà không cần một quá trình quá phức tạp.

Trong ngành viễn thông, việc giảm tỷ lệ khách hàng rời bỏ là một mục tiêu quan trọng để tối ưu hóa chi phí, duy trì và phát triển doanh nghiệp. Trong nghiên cứu này, chúng tôi đã thực hiện một cuộc khảo sát toàn diện về việc lựa chọn các công cụ và phương pháp để dự đoán tỷ lệ khách hàng rời bỏ, từ đó đề xuất các biện pháp giữ chân khách hàng hiệu quả. Qua quá trình phân tích, chúng tôi đã nhận ra rằng việc sử dụng một mô hình dự đoán chính xác là chìa khóa để dự đoán tỷ lệ khách hàng rời bỏ dịch vụ từ đó đưa ra các biện pháp giữ chân khách hàng hiệu quả. Các công cụ và mô hình được đánh giá trong nghiên cứu bao gồm LightGBM, Logistic Regression, Decision Trees, XGBoost và Random Forest. Trong số này, XGBoost và Random Forest là hai mô hình tối ưu nhất có thể xử lý những tập dữ liệu phức tạp và mang lại hiệu suất dự đoán cao. Ngoài ra, sự linh hoạt trong việc điều chỉnh các tham số của Random Forest và XGBoost cũng là điểm mạnh giúp tối ưu hóa hiệu suất của mô hình. Việc lựa chọn các công cụ và phương pháp dự đoán tỷ lệ khách hàng rời bỏ trong dịch vụ viễn thông đòi hỏi sự cân nhắc kỹ lưỡng và sự hiểu biết sâu sắc về dữ liệu và mục tiêu của doanh nghiệp. Random Forset và XGBoost được đề xuất là hai lựa chọn tối ưu nhất cho bài toán, kết quả của hai mô hình này có thể mang lại hiệu quả cao trong việc dự đoán tỷ lệ khách hàng rời bỏ để đưa ra các biện pháp giữ chân khách hàng, giảm tỷ lệ khách hàng rời bỏ và tăng cường sự trung thành của khách hàng với doanh nghiệp.

# TÀI LIỆU

[1] Mehdi Imani and Hamid Reza Arabnia, “Hyperparameter Optimization and Combined Data Sampling Techniques in Machine Learning for Customer Churn Prediction: A Comparative Analysis”, *Technologies,* 2023.

[2] Tianpei Xu and Ying Ma and Changyu Ao and Min Qu and XiangHong Meng, “*A Novel Telecom Customer Churn Analysis System Based on RFM Model and Feature Importance Ranking*”, Interdisciplinary Journal of Information, Knowledge, and Management, 2023

[3] Samah Wael Fujo and Suresh Subramanian and Moaiad Ahmad Khder, “Customer Churn Prediction in Telecommunication Industry Using Deep Learning”, *Information Sciences Letters*, 2022.

[4] Shuli Wu and Wei-Chuen Yau and Thian Song Ong and Siew-Chin Chong. “Integrated Churn Prediction and Customer Segmentation Framework for Telco Business”, *IEEE Access*, 2021.

[5] Chandra Lukita, “Predictive and Analytics using Data Mining and Machine Learning for Customer Churn Prediction”, *Journal of Applied Data Sciences*, 2023.

[6] Navienkumar R and Lalithamani N, “Machine Learning Methods for Predictive Customer Churn Analysis in the Telecom Industry”, *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2023.*

[7] Verma Priyanshu and Sharma Ishan and Deshmukh Sonia and Vashisht Rohit, “Customer Churn Analysis using Spark and Hadoop”, *Int. J. Perform. Eng.*, 2023.

[8] Van-Hieu-Vu, Nâng hiệu hiệu quả dự báo khách hàng rời bỏ dịch vụ ngân hàng sử dụng kỹ thuật học kết hợp mô hình học sâu, 2022.

[9] “to churn or not to churn”, <https://www.kaggle.com/code/caerno/to-churn-or-not-to-churn/notebook>.

|  |  |  |
| --- | --- | --- |
| **id** | **Cite** | **Abstract** |
| 1 | @article{Wu2021IntegratedCP,  title={Integrated Churn Prediction and Customer Segmentation Framework for Telco Business},  author={Shuli Wu and Wei-Chuen Yau and Thian Song Ong and Siew-Chin Chong},  tạp chí={Truy cập IEEE},  year={2021},  âm lượng={9},  pages={62118-62136}  } | [Integrated Churn Prediction and Customer Segmentation Framework for Telco Business](https://www.semanticscholar.org/paper/Integrated-Churn-Prediction-and-Customer-Framework-Wu-Yau/1526084beca5dbebbee1924df751e3b447935008)  This framework integrates churn prediction and customer segmentation process to provide telco operators with a complete churn analysis to better manage customer churn and proposes an integrated customer analytics framework for churn management.  Abstract  In the telco industry, attracting new customers is no longer a good strategy since the cost of retaining existing customers is much lower. Churn management becomes instrumental in the telco industry. As there is limited study combining churn prediction and customer segmentation, this paper aims to propose an integrated customer analytics framework for churn management. There are six components in the framework, including data pre-processing, exploratory data analysis (EDA), churn prediction, factor analysis, customer segmentation, and customer behaviour analytics. This framework integrates churn prediction and customer segmentation process to provide telco operators with a complete churn analysis to better manage customer churn. Three datasets are used in the experiments with six machine learning classifiers. First, the churn status of the customers is predicted using multiple machine learning classifiers. Synthetic Minority Oversampling Technique (SMOTE) is 2applied to the training set to deal with the problems with imbalanced datasets. The 10-fold cross-validation is used to assess the models. Accuracy and F1-score are used for model evaluation. F1-score is considered to be an important metric to measure the models for imbalanced datasets since the premise of churn management is to be able to identify customers who will churn. Experimental analysis indicates that AdaBoost performed the best in Dataset 1, with accuracy of 77.19% and F1-score of 63.11%. Random Forest performed the best in Dataset 2, with accuracy of 93.6% and F1-score of 77.20%. Random Forest performed the best in Dataset 3 in terms of accuracy, at 63.09%, while Multi-layer Perceptron performed the best in terms of F1-score, at 42.84%. After implementing churn prediction, Bayesian Logistic Regression is used to conduct the factor analysis and to figure out some important features for churn customer segmentation. Churn customer segmentation is then carried out using K-means clustering. Customers are segmented into different groups, which allows marketers and decision makers to adopt retention strategies more precisely. |
| 2 | @article{Fujo2022CustomerCP,  title={Customer Churn Prediction in Telecommunication Industry Using Deep Learning},  author={Samah Wael Fujo and Suresh Subramanian and Moaiad Ahmad Khder},  journal={Information Sciences Letters},  year={2022}  } | [Customer Churn Prediction in Telecommunication Industry Using Deep Learning](https://www.semanticscholar.org/paper/Customer-Churn-Prediction-in-Telecommunication-Deep-Fujo-Subramanian/fc7b65724ea8b285cb3abf99938695c142774d02)  In predicting customer churn, the Deep-BP-ANN model’s accuracy outperforms the existing deep learning techniques that use holdout or 10fold CV for the same datasets and outperforms ML techniques: XG\_Boost, Logistic\_Regression, Naïve\_Bayes, and KNN.  Abstract  Without proper analysis and forecasting, industries will find themselves repeatedly churning customers, which the telecom industry in particular cannot afford. A predictable model for customers will allow companies to retain current customers and to obtain new ones. Deep-BP-ANN implemented in this study using two feature selection methods, Variance Thresholding and Lasso Regression, in addition, our model strengthened by early stopping technique to stop training at right time and prevent overfitting. We compared the efficiency of minimizing overfitting between dropout and activity 3regularization strategies for two real datasets: IBM Telco and Cell2cell. Different evaluation approaches used: Holdout, and 10-fold cross-validation to evaluate the model’s efficiency. To solve unbalanced issue, the Random Oversampling technique was used to balance both datasets. The results show that the model implemented performs well with lasso regression for feature selection, early stopping technique to pick the epochs, and large numbers of neurons (250) into the input and hidden layers, and activity regularization to minimize overfitting for both datasets. In predicting customer churn, our findings outperform ML techniques: XG\_Boost, Logistic\_Regression, Naïve\_Bayes, and KNN. Moreover, our Deep-BP-ANN model’s accuracy outperforms the existing deep learning techniques that use holdout or 10fold CV for the same datasets. |
| 3 | @article{Zhang2022ADA,  title={A Data-Driven Approach to Improve Customer Churn Prediction Based on Telecom Customer Segmentation},  author={Tianyuan Zhang and S{\'e}rgio Moro and Ricardo F. Ramos},  journal={Future Internet},  year={2022},  âm lượng={14},  pages={94}  } | [A Data-Driven Approach to Improve Customer Churn Prediction Based on Telecom Customer Segmentation](https://www.semanticscholar.org/paper/A-Data-Driven-Approach-to-Improve-Customer-Churn-on-Zhang-Moro/b9480825c593c84f9287cf983e3f1671e72e344e)  This study aimed to develop a churn prediction model to predict telecom client churn through customer segmentation to help telecom companies efficiently predict the possibility of and take targeted measures to avoid customer churn, thereby increasing their profits.  Abstract  Numerous valuable clients can be lost to competitors in the telecommunication industry, leading to profit loss. Thus, understanding the reasons for client churn is vital for telecommunication companies. This study aimed to develop a churn prediction model to predict telecom client churn through customer segmentation. Data were collected from three major Chinese telecom companies, and Fisher discriminant equations and logistic regression analysis were used to build a telecom customer churn prediction model. According to the results, it can be concluded that the telecom customer churn model constructed by regression analysis had higher prediction accuracy (93.94%) and better results. This study will help telecom companies efficiently predict the possibility of and take targeted measures to avoid customer churn, thereby increasing their profits. |
| 4 | @article{Bilal2022AnEB,  title={An ensemble based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry},  author={Syed Fakhar Bilal and Abdulwahab Ali Almazroi and Saba Bashir and Farhan Hassan Khan and Abdulaleem Ali Almazroi},  journal={PeerJ Computer Science},  year={2022},  âm lượng={8}  } | [An ensemble based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry](https://www.semanticscholar.org/paper/An-ensemble-based-approach-using-a-combination-of-Bilal-Almazroi/bff69c31b4d76b60443467ae8540361892352655)  The proposed churn prediction model is a hybrid model that is based on a combination of clustering and classification algorithms using an ensemble that performed significantly better than state of the art churn prediction models.  Abstract  Mobile communication has become a dominant medium of communication over the past two decades. New technologies and competitors are emerging rapidly and churn prediction has become a great concern for telecom companies. A customer churn prediction model can provide the accurate identification of potential churners so that a retention solution may be provided to them. The proposed churn prediction model is a hybrid model that is based on a combination of clustering and classification algorithms using an ensemble. First, different clustering algorithms (i.e. K-means, K-medoids, X-means and random clustering) were evaluated individually on two churn prediction datasets. Then hybrid models were introduced by combining the clusters with seven different classification algorithms individually and then evaluations were performed using ensembles. The proposed research was evaluated on two different benchmark telecom data sets obtained from GitHub and Bigml platforms. The analysis of results indicated that the proposed model attained the highest prediction accuracy of 94.7% on the GitHub dataset and 92.43% on the Bigml dataset. State of the art comparison was also performed using the proposed model. The proposed model performed significantly better than state of the art churn prediction models. |
| 5 | @article{Nalatissifa2021CustomerDP,  title={Customer Decision Prediction Using Deep Neural Network on Telco Customer Churn Data},  author={Hiya Nalatissifa and Hilman Ferdinandus Pardede},  journal={Jurnal Elektronika dan Telekomunikasi},  year={2021}  } | [Customer Decision Prediction Using Deep Neural Network on Telco Customer Churn Data](https://www.semanticscholar.org/paper/Customer-Decision-Prediction-Using-Deep-Neural-on-Nalatissifa-Pardede/f55f71062dffeec27aa70afc40a721bef0d8ae67)  This study proposes a better architecture for the DNN algorithm by using a hard tuner to obtain more optimal hyperparameters, which is better than modeling using k-nearest neighbor (K-NN), random forest (RF), and decision tree (DT) as comparison algorithms.  Abstract  Customer churn is the most important problem in the business world, especially in the telecommunications industry, because it greatly influences company profits. Getting new customers for a company is much more difficult and expensive than retaining existing customers. Machine learning, part of data mining, is a sub-field of artificial intelligence widely used to make predictions, including predicting customer churn. Deep neural network (DNN) has been used for churn prediction, but selecting hyperparameters in modeling requires more time and effort, making the process more challenging for the researcher. Therefore, the purpose of this study is to propose a better architecture for the DNN algorithm by using a hard tuner to obtain more optimal hyperparameters. The tuning hyperparameter used is random search in determining the number of nodes in each hidden layer, dropout, and learning rate. In addition, this study also uses three variations of the number of hidden layers, two variations of the activation function, namely rectified linear unit (ReLu) and Sigmoid, then uses five variations of the optimizer (stochastic gradient descent (SGD), adaptive moment estimation (Adam), adaptive gradient algorithm (Adagrad), Adadelta, and root mean square propagation (RMSprop)). Experiments show that the DNN algorithm using hyperparameter tuning random search produces a performance value of 83.09 % accuracy using three hidden layers, the number of nodes in each hidden layer is [20, 35, 15], using the RMSprop optimizer, dropout 0.1, the learning rate is 0.01, with the fastest tuning time of 21 seconds. Better than modeling using k-nearest neighbor (K-NN), random forest (RF), and decision tree (DT) as comparison algorithms. |
| 6 | @article{Vasudevan2022CustomerCA,  title={Customer churn analysis using XGBoosted decision trees},  author={Muthupriya Vasudevan and Revathi Sathya Narayanan and Sabiyath Fatima Nakeeb and Abhishek Abhishek},  journal={Indonesian Journal of Electrical Engineering and Computer Science},  year={2022}} | [Customer churn analysis using XGBoosted decision trees](https://www.semanticscholar.org/paper/Customer-churn-analysis-using-XGBoosted-decision-Vasudevan-Narayanan/1417e2d7a4100783d6f593a8f8682b4a69294022)  A supervised machine learning algorithm has been implemented using Python to perform customer churn analysis on a given data-set of Telco, a mobile telecommunication company by building a decision tree model based on historical data provided by the company on the platform of Kaggle.  Abstract  Customer relationship management (CRM) is an important element in all forms of industry. This process involves ensuring that the customers of a business are satisfied with the product or services that they are paying for. Since most businesses collect and store large volumes of data about their customers; it is easy for the data analysts to use that data and perform predictive analysis. One aspect of this includes customer retention and customer churn. Customer churn is defined as the concept of understanding whether or not a customer of the company will stop using the product or service in future. In this paper a supervised machine learning algorithm has been implemented using Python to perform customer churn analysis on a given data-set of Telco, a mobile telecommunication company. This is achieved by building a decision tree model based on historical data provided by the company on the platform of Kaggle. This report also investigates the utility of extreme gradient boosting (XGBoost) library in the gradient boosting framework (XGB) of Python for its portable and flexible functionality which can be used to solve many data science related problems highly efficiently. The implementation result shows the accuracy is comparatively improved in XGBoost than other learning models. |
| 7 | @article{Zhou2023EarlyWO,  title={Early warning of telecom enterprise customer churn based on ensemble learning},  author={Yancong Zhou and Wenyue Chen and Xiaochen Sun and Dandan Yang},  journal={PLOS ONE},  year={2023},  volume={18}} | [Early warning of telecom enterprise customer churn based on ensemble learning](https://www.semanticscholar.org/paper/Early-warning-of-telecom-enterprise-customer-churn-Zhou-Chen/2fa34f42b5823c9ccb761a4100066644d9bcacd5)  The prediction results of customer churn provide strong data support for telecom companies to adopt appropriate retention strategies for pre-churn customers and reduce customer churn.  Abstract  Analyzing customers’ characteristics and giving the early warning of customer churn based on machine learning algorithms, can help enterprises provide targeted marketing strategies and personalized services, and save a lot of operating costs. Data cleaning, oversampling, data standardization and other preprocessing operations are done on 900,000 telecom customer personal characteristics and historical behavior data set based on Python language. Appropriate model parameters were selected to build BPNN (Back Propagation Neural Network). Random Forest (RF) and Adaboost, the two classic ensemble learning models were introduced, and the Adaboost dual-ensemble learning model with RF as the base learner was put forward. The four models and the other four classical machine learning models-decision tree, naive Bayes, K-Nearest Neighbor (KNN), Support Vector Machine (SVM) were utilized respectively to analyze the customer churn data. The results show that the four models have better performance in terms of recall rate, precision rate, F1 score and other indicators, and the RF-Adaboost dual-ensemble model has the best performance. Among them, the recall rates of BPNN, RF, Adaboost and RF-Adaboost dual-ensemble model on positive samples are respectively 79%, 90%, 89%,93%, the precision rates are 97%, 99%, 98%, 99%, and the F1 scores are 87%, 95%, 94%, 96%. The RF-Adaboost dual-ensemble model has the best performance, and the three indicators are 10%, 1%, and 6% higher than the reference. The prediction results of customer churn provide strong data support for telecom companies to adopt appropriate retention strategies for pre-churn customers and reduce customer churn. |
| 8 | @article{Wu2023CustomerCP,  title={Customer Churn Prediction in the Telecommunication Industry},  author={Shaohua Wu},  tạp chí={Những tiến bộ trong Kinh tế, Quản lý và Khoa học Chính trị},  year={2023}} | [Customer Churn Prediction in the Telecommunication Industry](https://www.semanticscholar.org/paper/Customer-Churn-Prediction-in-the-Telecommunication-Wu/2474c66e81694e654542a122b770afbf8ed9a3b2)  This essay uses contrastive analysis to find out the factors of the customer churn and logistic regression to analyze the degree of influence of different factors, which indicates phone charge, quality and diversity of services have impacts on customer churn.  Abstract  Customer churn is essential for telecom fields because it reduces income when a customer switches from one service provider to another. It is important for managers to help them find the factors influencing customer churn so that they can make decisions and optimize services. In this essay, SPSS will be used to analyze the factors that influence customer churn. We use contrastive analysis to find out the factors of the customer churn and use logistic regression to analyze the degree of influence of different factors. The result indicates phone charge, quality and diversity of services have impacts on customer churn. |
| 9 | @article{Pan2023CustomerCP,  title={Customer churn prediction in telecom based on random forest algorithm},  author={Aimin Pan and Zhendan Sun},  journal={2023 8th International Conference on Information Systems Engineering (ICISE)},  year={2023},  pages={189-192}} | [Customer churn prediction in telecom based on random forest algorithm](https://www.semanticscholar.org/paper/Customer-churn-prediction-in-telecom-based-on-Pan-Sun/d34b77d62f1ec557b743dba9b964ce62a749a286)  In order to reduce customer churn, the machine learning algorithm of Python analyzes the characteristics of outbound users, establishes customer churn model prediction, locates transfer subscribers in advance, and maintains stability retention, effectively reducing the probability ofOutbound users.  Abstract  In recent years, with the in-depth implementation of the telcom's Number Portability work, there are more and more Number Portability users, and the Number Portability users have become a major factor of customer churn. In order to reduce customer churn, the machine learning algorithm of Python analyzes the characteristics of outbound users, establishes customer churn model prediction, locates transfer subscribers in advance, and maintains stability retention, effectively reducing the probability of outbound users. |
| 10 | @article{Azzam2023CustomerCP,  title={Customer Churn Prediction Using Apriori Algorithm and Ensemble Learning},  author={Diaa Azzam and Manar Hamed and Nora Kasiem and Yomna Eid and Walaa Medhat},  journal={2023 5th Novel Intelligent and Leading Emerging Sciences Conference (NILES)},  year={2023},  pages={377-381}} | [Customer Churn Prediction Using Apriori Algorithm and Ensemble Learning](https://www.semanticscholar.org/paper/Customer-Churn-Prediction-Using-Apriori-Algorithm-Azzam-Hamed/e7bbd91a5f5b872db9a364844792cbf5b6fa1291)  This research conducted an extensive study aimed at developing a viable customer churn prediction method that utilizes the Apriori algorithm’s strength to identify the key causes of customer churn.  Abstract  Customer churn poses a formidable challenge within the Telecom industry, as it can result in significant revenue losses. In this research, we conducted an extensive study aimed at developing a viable customer churn prediction method. Our method utilizes the Apriori algorithm’s strength to identify the key causes of customer churn. In the pursuit of this goal, we utilized multiple machine learning predictive models. All of which were developed from the insights gleaned from the Apriori algorithm’s feature extraction for churning customers. This extensive analysis encompassed a spectrum of machine learning techniques that include Logistic Regression, Naive Bayes, Support Vector Machines, Random Forests, and Decision Trees. Furthermore, we utilized an ensemble learning approach to enhance the predictive accuracy of our models. We also used a voting classifier refined with the best features within our dataset. The voting classifier yielded an accuracy rate of 81.56%, underscoring the effectiveness of our approach in addressing the critical issue of customer churn in the Telecom industry. |
| 11 | @article{Priya2023CUSTOMERCP,  title={CUSTOMER CHURN PREDICTION ANALYSIS},  author={E Padma Priya and Janashree S S and Sivaganga P},  journal={INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT},  year={2023}} | [CUSTOMER CHURN PREDICTION ANALYSIS](https://www.semanticscholar.org/paper/CUSTOMER-CHURN-PREDICTION-ANALYSIS-Priya-JanashreeS/f574d78b472e541bd3c8d619aff316a6a8fe8a0b)  These medium firms are safeguarded with effective strategies for reducing the speed of churn because of machine learning algorithmic program techniques and the present work makes use of a big-data platform and machine learning methodology.  Abstract  The Customers are the base of many successful businesses; thus, all the sectors are starting to understand how important it is to gain client satisfaction. The technical infrastructure has expanded quickly, changing how businesses operate. Due to growing business competition, the importance of marketing techniques, and customers' increasingly aware behaviour in recent years, leaving the organization is a crucial issue and it is one of the most crucial worries for most of the sectors. Different approaches must be developed by organizations to address the churn problems affecting the services they provide. To gain a deeper understanding of customer churn, this projectsummarizes churn prediction techniques. It also demonstrates that hybrid models, as opposed to single algorithms, provide the most accurate churn predictions, allowing telecom industries to better understand the needs of high- risk customers. Reduced client turnover is turning into a demand for service suppliers as a result of the potential impact that client neglect may wear business profitability. Prediction helps to search out users' World Health Organization area units doubtlessly to change from one organization to a different. The ever-rising churn rate in the medium could be a drawback. In lightweight of this, the present work makes use of a big-data platform and machine learning methodology. These medium firms are safeguarded with effective strategies for reducing the speed of churn because of machine learning algorithmic program techniques. Keywords: Churn, Machine Learning, Random Forest,telecommunication, |
| 12 | @article{Sidiq2023AnalysisAC,  title={Analysis and Classification of Customer Churn Using Machine Learning Models},  author={Muhammad Maulana Sidiq and Dyah Anggraini},  journal={Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)},  year={2023}} | [Analysis and Classification of Customer Churn Using Machine Learning Models](https://www.semanticscholar.org/paper/Analysis-and-Classification-of-Customer-Churn-Using-Sidiq-Anggraini/95b194a4c94e22e0fa428101b63620d9b9727000)  The results of the analysis show that the classification of churn customers using the XGBoost algorithm has the best level of accuracy compared to other algorithms, and the oversampling method with SMOTE tends to reduce the accuracy value of each classification algorithm.  Abstract  Analysis studies of customer loss (customer churn) have been used for years to increase profitability and build customer relationships with companies. Customer analysis using exploratory data analysis (EDA) for visualizing data and the use of machine learning for the classification of customer churn are often used by past analysts. This study uses several machine learning models that can be used for customer churn classification, namely Logistic Regression, Random Forest, Support Vector Machine (SVM), Gradient Boosting, AdaBoost, and Extreme Gradient Boosting (XGBoost). However, there is a class imbalance factor in the dataset, which is the biggest challenge that is usually faced by analysts to get good results in the classification of machine learning models. The Synthetic Minority Over-sampling Technique (SMOTE) method is a popular method applied to deal with class imbalances in datasets. The results of the analysis show that the classification of churn customers using the XGBoost algorithm has the best level of accuracy compared to other algorithms, with an accuracy value of 0.829424, and the oversampling method with SMOTE tends to reduce the accuracy value of each classification algorithm. The Permutation Feature Importance (PFI) technique from the XGBoost model gets the result that tenure, monthly contracts, and TV streaming are the features that affect customer churn the most. |
| 13 | @article{R2023MachineLM,  title={Machine Learning Methods for Predictive Customer Churn Analysis in the Telecom Industry},  author={Navienkumar R and Lalithamani N},  journal={2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)},  year={2023},  pages={1-7}} | [Machine](https://www.semanticscholar.org/paper/Machine-Learning-Methods-for-Predictive-Customer-in-Navienkumar-Lalithamani/908d93e417a37c2a43318c8d93f67efa31d92151) Learning Methods for Predictive Customer Churn Analysis in the Telecom Industry  It is shown that the Random-Forest (RF) model performed best and had the highest degree of accuracy and may be used by telecom businesses to estimate customer attrition and aggressively retain clients.  Abstract  To examine customer churn in the telecom sector, we used four classification algorithms: Decision-Tree(DT), Random-Forest(RF), Logistic-Regression(LR), and a Hybrid-Algorithm(HA). The dataset contained information about telecom users, including demographics, use trends, and customer service data. By resolving missing values and outliers during the preprocessing of the data, we were able to comprehend the variable distribution and connections. The four classification models were then applied to the dataset by dividing it into training and testing sets. The models' accuracy outcomes were as follows: The accuracy rates for Logistic-Regression (LR), Decision-Tree (DT), Random-Forest (RF), and the Hybrid Algorithm(HA), which combines DT and RF capabilities, were 85%, 91%, 94%, and 95%, respectively. Utilising measures like Accuracy, Precision, Recall and recall rate, we evaluated the models' performance. Our research showed that the Random-Forest (RF) model performed best and had the highest degree of accuracy. The hybrid approach greatly increased classification accuracy, underscoring its usefulness. The Hybrid Algorithm(HA) may be used by telecom businesses to estimate customer attrition and aggressively retain clients. These findings highlight how crucial it is to use the right algorithms for accurate customer attrition prediction, with the Hybrid Algorithm (HA) offering a practical way to raise classification accuracy. |
| 14 | @article{Arai2023MethodFC,  title={Method for Characterization of Customer Churn Based on LightBGM and Experimental Approach for Mitigation of Churn},  author={Kohei Arai and Ikuya Fujikawa and Yusuke Nakagawa and Ryo Momozaki and Sayuri Ogawa},  journal={International Journal of Advanced Computer Science and Applications},  year={2023}} | [Method for Characterization of Customer Churn Based on LightBGM and Experimental Approach for Mitigation of Churn](https://www.semanticscholar.org/paper/Method-for-Characterization-of-Customer-Churn-Based-Arai-Fujikawa/a1c1dde1311de9b7614955dd10026f9637729ea2)  A method for customer churn characterization based on LightBGM (Light Gradient Boosting Machine) is proposed and it is found that the first visit date is the most significant factor for churn customer character.  Abstract  A method for customer churn characterization based on LightBGM (Light Gradient Boosting Machine) is proposed. Additionally, experimental approaches for mitigation of churn are conducted through churn prediction. The experiments reveal several churn characteristics such as age dependency, gender dependency (with a high divorce rate among female customers), number of visits dependency (with a higher churn rate for customers with fewer visits), unit price (per hair salon visit) dependency (with a higher withdrawal rate for lowerpriced services), date of first visit dependency (with a high churn rate for recent customers), date of last visit dependency, and menu dependency (with low attrition rates for gray hair dye and high attrition rates for school and child cuts) and so on. Through the experiments, these dependencies are clarified. It is found that the first visit date is the most significant factor for churn customer character. Also, it is found that “distance to hair salon” dependency may be related to the availability of parking lots, although this factor has insignificant impact on the churn rate. Keywords—Churn; LightBGM; churn characteristics; linear regression |
| 15 | @article{Xu2023ANT, title={A Novel Telecom Customer Churn Analysis System Based on RFM Model and Feature Importance Ranking}, author={Tianpei Xu and Ying Ma and Changyu Ao and Min Qu and XiangHong Meng}, journal={Interdisciplinary Journal of Information, Knowledge, and Management}, year={2023}, url={https://api.semanticscholar.org/CorpusID:263646257} } | [A Novel Telecom Customer Churn Analysis System Based on RFM Model and Feature Importance Ranking](https://www.semanticscholar.org/paper/A-Novel-Telecom-Customer-Churn-Analysis-System-on-Xu-Ma/b36abfd4db289a57463a785b6ca83f5d9774443f)  The proposed system outperforms current advanced customer churn prediction methods in the same dataset, with a higher prediction accuracy, and can help telecom companies implement effective CRM and marketing strategies to reduce customer churn.  Abstract  Aim/Purpose: In this paper, we present an RFM model-based telecom customer churn system for better predicting and analyzing customer churn. Background: In the highly competitive telecom industry, customer churn is an important research topic in customer relationship management (CRM) for telecom companies that want to improve customer retention. Many researchers focus on a telecom customer churn analysis system to find out the customer churn factors for improving prediction accuracy. Methodology: The telecom customer churn analysis system consists of three main parts: customer segmentation, churn prediction, and churn factor identification. To segment the original dataset, we use the RFM model and K-means algorithm with an elbow method. We then use RFM-based feature construction for customer churn prediction, and the XGBoost algorithm with SHAP method to obtain a feature importance ranking. We chose an open-source customer churn dataset that contains 7,043 instances and 21 features. Contribution: We present a novel system for churn analysis in telecom companies, which encompasses customer churn prediction, customer segmentation, and churn factor analysis to enhance business strategies and services. In this system, we leverage customer segmentation techniques for feature construction, which enables the new features to improve the model performance significantly. Our experiments demonstrate that the proposed system outperforms current advanced customer churn prediction methods in the same dataset, with a higher prediction accuracy. The results further demonstrate that this churn analysis system can help telecom companies mine customer value from the features in a dataset, identify the primary factors contributing to customer churn, and propose suitable solution strategies. Findings: Simulation results show that the K-means algorithm gets better results when the original dataset is divided into four groups, so the K value is selected as 4. The XGBoost algorithm achieves 79.3% and 81.05% accuracy on the original dataset and new data with RFM, respectively. Additionally, each cluster has a unique feature importance ranking, allowing for specialized strategies to be provided to each cluster. Overall, our system can help telecom companies implement effective CRM and marketing strategies to reduce customer churn. Recommendations for Practitioners: More accurate churn prediction reduces misjudgment of customer churn. The acquisition of customer churn factors makes the company more convenient to analyze the reasons for churn and formulate relevant conservation strategies. Recommendation for Researchers: The research achieves 81.05% accuracy for customer churn prediction with the Xgboost and RFM algorithms. We believe that more enhancements algorithms can be attempted for data preprocessing for better prediction. Impact on Society: This study proposes a more accurate and competitive customer churn system to help telecom companies conserve the local markets and reduce capital outflows. Future Research: The research is also applicable to other fields, such as education, banking, and so forth. We will make more new attempts based on this system. |
| 16 | @article{Lukita2023PredictiveAA,  title={Predictive and Analytics using Data Mining and Machine Learning for Customer Churn Prediction},  author={Chandra Lukita},  journal={Journal of Applied Data Sciences},  year={2023}} | [Predictive and Analytics using Data Mining and Machine Learning for Customer Churn Prediction](https://www.semanticscholar.org/paper/Predictive-and-Analytics-using-Data-Mining-and-for-Lukita/15c5b65ab28f9c49b64961e975d515da2b3003f4)  This research compares the performance of several machine learning algorithms, including Random Forest, Logistic Regression, Adaboost, and Extreme Gradient Boosting, to predict customer churn and showed that XGBoost proved to be the best algorithm in predicting customer churn with high accuracy.  Abstract  This research aims to predict and analyze customer churn using Data Mining and Machine Learning methods. The background of this research is based on the importance of understanding the factors that influence customer decisions to churn, as well as improving the effectiveness of customer retention strategies in a business context. The method used in this research involves the use of a customer bank dataset that includes information about customers who left in the past month, services registered by customers, customer account information, and demographic info about customers. The factors most influential to churn were identified through heatmap analysis, including MonthlyCharges, PaperlessBilling, SeniorCitizen, PaymentMethod, MultipleLines, and PhoneService. This research compares the performance of several machine learning algorithms, including Random Forest, Logistic Regression, Adaboost, and Extreme Gradient Boosting (XGBoost), to predict customer churn. Accuracy metrics and confusion matrix results are used to evaluate the performance of these algorithms. The results showed that XGBoost proved to be the best algorithm in predicting customer churn with high accuracy. The factors that have been correctly identified do not provide missed precision, showing a significant influence on customer churn decisions. The novelty and uniqueness of this research lies in focusing on the factors that have the most influence on customer churn and comparing the performance of machine learning algorithms. This research provides more specific and relevant insights for companies in developing effective customer retention strategies. However, this research has some limitations. One of them is the use of a dataset limited to a customer bank, so the generalizability of the findings of this research may be limited to that business context. In addition, other factors that are not the focus of this research may also contribute to the prediction of customer churn. |
| 17 | @article{Bilisik2023AnalysisOC,  title={Analysis of Customer Churn in Telecommunication Industry with Machine Learning Methods},  author={{\"O}zge Nalan Bilisik and Damla Tuğba Sarp},  journal={D{\"u}zce {\"U}niversitesi Bilim ve Teknoloji Dergisi},  year={2023}} | [Analysis of Customer Churn in Telecommunication Industry with Machine Learning Methods](https://www.semanticscholar.org/paper/Analysis-of-Customer-Churn-in-Telecommunication-Bilisik-Sarp/0685cefc50e64f9bf8ab0a43522727dc32b0f5b8)  This study aims to make predictions by developing models for customer churns through data mining and machine learning methods in the telecommunication sector using Random Forest, Support Vector Machines, and Multilayer Artificial Neural Networks modeled in open-source Phyton environment.  Abstract  In today's conditions, customer loyalty has gained importance with the increase in the competitive environment between companies, the development of marketing strategies and the improvement of companies. Therefore, it is essential to acquire customers for a company to survive. Retaining an existing customer in the telecommunication sector is less costly than gaining a new customer. Customer churn analysis is the process of predicting customers with high abandonment requests by examining the offers and utilizable behaviors. Customer churn analysis provides services to develop various campaigns aiming to increase the company’s loyalty by predicting the customers who are planning to move to another company. In this way, it gives the company a competitive advantage. This study aims to make predictions by developing models for customer churns through data mining and machine learning methods in the telecommunication sector. In addition, we believe that the application in this article will contribute to data analysts and academicians who will want to analyze customer churn with different data sets in telecommunication and other sectors in the future. The analysis in this study is carried out on a data set obtained from an open-access database, including 20 transaction records for the customer from 7043 customers and whether the customer left the company. Among the data mining methods, Random Forest (RF), Support Vector Machines (SVM) and Multilayer Artificial Neural Networks (ANN) are modeled in open-source Phyton environment. The results have shown that ANN has fared better at classifying customers than other machine learning methods. |
| 18 | @article{Tang2023ComparisonOM,  title={Comparison of machine learning methods for estimating customer churn in the telecommunication industry},  author={Thomas Tang},  journal={Applied and Computational Engineering},  year={2023}} | [Comparison of machine learning methods for estimating customer churn in the telecommunication industry](https://www.semanticscholar.org/paper/Comparison-of-machine-learning-methods-for-customer-Tang/5c8ba8a9278fe851df4402e73c4e200ed8953d32)  This article examines several machine learning algorithms for predicting customer churn, including Random Forest, Decision Tree, and Naive Bayes to examine and compare the effectiveness of various models according to accuracy, which can be applied in real-life scenarios.  Abstract  With the rising competition in business, particularly in the telecommunications industry, there has been a growing emphasis on churn prediction. This is attributed to the higher cost involved in attracting new customers than retaining the remaining ones. In telecom churn analysis, the primary goal is to accurately estimate churn behavior by identifying customers who are at risk of leaving. Another objective is to determine the primary reasons for customer churn. Manually predicting the churn in telecommunications is expensive, tedious, and time-consuming. To relieve the burden, machine learning algorithms are introduced to tackle this problem. This article examines several machine learning algorithms for predicting customer churn, including Random Forest, Decision Tree, and Naive Bayes. By constructing and comparing these classification models, the effectiveness of these algorithms is demonstrated. The aim of this work is to examine and compare the effectiveness of various models according to accuracy, which can be applied in real-life scenarios. The result shows that Naive Bayes outperforms the other models. |
| 19 | @article{Maan2023CustomerCP,  title={Customer Churn Prediction Model using Explainable Machine Learning},  author={Jitendra Maan and Harsh Maan},  journal={ArXiv},  year={2023},  volume={abs/2303.00960}} | [Customer Churn Prediction Model using Explainable Machine Learning](https://www.semanticscholar.org/paper/Customer-Churn-Prediction-Model-using-Explainable-Maan-Maan/ec9e6e8e61c0f8a9556985280345bfdf6244aa7c)  A novel approach to calculate Shapley values for possible combination of features to explain which features are the most important/relevant features for a model to become highly interpretable, transparent and explainable to potential customers is proposed.  Abstract  It becomes a significant challenge to predict customer behavior and retain an existing customer with the rapid growth of digitization which opens up more opportunities for customers to choose from subscription-based products and services model. Since the cost of acquiring a new customer is five-times higher than retaining an existing customer, henceforth, there is a need to address the customer churn problem which is a major threat across the Industries. Considering direct impact on revenues, companies identify the factors that increases the customer churn rate. Here, key objective of the paper is to develop a unique Customer churn prediction model which can help to predict potential customers who are most likely to churn and such early warnings can help to take corrective measures to retain them. Here, we evaluated and analyzed the performance of various tree-based machine learning approaches and algorithms and identified the Extreme Gradient Boosting XGBOOST Classifier as the most optimal solution to Customer churn problem. To deal with such real-world problems, Paper emphasize the Model interpretability which is an important metric to help customers to understand how Churn Prediction Model is making predictions. In order to improve Model explainability and transparency, paper proposed a novel approach to calculate Shapley values for possible combination of features to explain which features are the most important/relevant features for a model to become highly interpretable, transparent and explainable to potential customers. |
| 20 | @article{Upadhyay2023CustomerCP,  title={Customer Churn Prediction using Machine Learning},  author={Samiksha Upadhyay and Rajalakshmi M},  journal={2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)},  year={2023},  pages={1-7}} | [Customer Churn Prediction using Machine Learning](https://www.semanticscholar.org/paper/Customer-Churn-Prediction-using-Machine-Learning-Upadhyay-Rajalakshmi/0d25c1050c86379dfbb903e1c6dd6a90982a9b16)  This study proposes a combination of estimating model for customer turnover and performs practical research on the model's efficacy and concludes that the combined prediction model outperforms the single customer churn prediction model in terms of accuracy and predictive impact.  Abstract  One significant problem that businesses face is customer attrition. It has become crucial for corporate operations and growth to prevent customer churn and work to keep clients. It is challenging to effectively estimate customer turnover because the majority of the existing projections use a single prediction model. Concentrating on the results of predictions of the models of machine learning, this study proposes a combination of estimating model for customer turnover and performs practical research on the model's efficacy. The combined prediction model outperforms the single customer churn prediction model in terms of accuracy and predictive impact, according to the findings of the predictions. It can also more naturally express the fundamental traits of the churn consumers. |
| 21 | @article{Zaidan2023ApplicationOT,  title={Application of the Beta Distribution Model to the Customer Churn Rate},  author={Henrique Pinto dos Santos Zaidan},  journal={Socioeconomic Analytics},  year={2023}} | [Application of the Beta Distribution Model to the Customer Churn Rate](https://www.semanticscholar.org/paper/Application-of-the-Beta-Distribution-Model-to-the-Zaidan/e4f29c90cadbcd13ccc979b319717a1216d88452)  This research fits the beta distribution model for mod- eling a recurring problem confronted for many businesses called the customer churn rate (or churn rate), which represents the proportion of customers who cancel their subscriptions after a given time.  Abstract  The beta distribution model has been applied in many different research environments due to the flexibility of its two parameters. In this research, we fit this probabilistic model for mod- eling a recurring problem confronted for many businesses called the customer churn rate (or churn rate). It represents the proportion of customers who cancel their subscriptions after a given time. We use data from a Brazilian media service company to develop the modeling. The parameters are estimated by the maximum likelihood estimation (MLE) technique. Finally, we perform the MLE technique by considering two programming languages; Ox and R. |
| 22 | @article{Zhao2023CustomerCP,  title={Customer Churn Prediction Based on the Decision Tree and Random Forest Model},  author={Shi Zhao},  journal={BCP Business \& Management},  year={2023}} | [Customer Churn Prediction Based on the Decision Tree and Random Forest Model](https://www.semanticscholar.org/paper/Customer-Churn-Prediction-Based-on-the-Decision-and-Zhao/a4157ccf0f8039cccafca22ff61a764d8fdcd813)  This study deals with a dataset about customer behaviors in a bank and building churn prediction models using machine learning algorithms, and shows that both models get most predictions correct while random forest model has a better performance due to its higher accuracy score.  Abstract  The rate at which customers discontinue utilizing a company's services during a predetermined time period is known as the customer churn rate, also known as the attrition rate. Hence, developing a prediction model to predict the potential churn customers will generate an early alert for the company to provide them with better service. This study is divided into two main parts: dealing with a dataset about customer behaviors in a bank and building churn prediction models using machine learning algorithms. The data preprocessing part includes dataset description and some adjustments on original dataset to make it accessible for analysis, including deleting unimportant feature and adjusting feature names. Then the study apportions the modified dataset into train set and test set with an 80-20 split. Next, the study imports two kinds of machine learning algorithms, random forest classifier and decision tree classifier, to build churn prediction models. In each model, the study first performs feature selections and visualizes feature importance in bar graphs. Then the study tests each model on testing set and visualizes model performances using confusion matrices and accuracy scores. The results show that both models get most predictions correct while random forest model has a better performance due to its higher accuracy score of 91%. |
| 23 | @article{SnehalRathi2023ASA,  title={A Survey and Implementation of Machine Learning Algorithms for Customer Churn Prediction},  author={Et al. Snehal Rathi},  journal={International Journal on Recent and Innovation Trends in Computing and Communication},  year={2023}} | [A Survey and Implementation of Machine Learning Algorithms for Customer Churn Prediction](https://www.semanticscholar.org/paper/A-Survey-and-Implementation-of-Machine-Learning-for-Rathi/c99f00e62e078ef587fcc3667ad5d20b88a8e62f)  A machine-learning model for predicting customer churn using historical customer data is developed and it is revealed that monthly fees, customer tenure, contract type, and payment method are the factors that have the most impact on forecasting customer churn.  Abstract  Estimating customer traffic is an important task for businesses because it helps them identify customers who are most likely to leave and take preventative measures to retain them by improving customer satisfaction and further increasing their own revenue. In this article, we focus on developing a machine-learning model for predicting customer churn using historical customer data We performed engineering operations on the data, addressed the missing digits, coded the categorical variables, and preprocessed the data before evaluating it using a variety of performance indicators, including accuracy, precision, recall, f1 score, and ROC AUC\_Score. Our feature significance analysis revealed that monthly fees, customer tenure, contract type, and payment method are the factors that have the most impact on forecasting customer churn. Finally, we conclude the best-performing model, the Soft Voting Classifier, implemented on the four best-performing classifiers with a good accuracy of 0.78 and a relatively better ROC AUC\_Score of 0.82. |
| 24 | @article{Khan2023PrivacyPreservingBT,  title={Privacy-Preserving Based Technique for Customer Churn Prediction in Telecom Industry},  author={Gul Zaman Khan and Ikram Ulhaq and Ihsan Adil and Sajad Ulhaq and Inam Ullah},  journal={VFAST Transactions on Software Engineering},  year={2023}} | [Privacy-Preserving Based Technique for Customer Churn Prediction in Telecom Industry](https://www.semanticscholar.org/paper/Privacy-Preserving-Based-Technique-for-Customer-in-Khan-Ulhaq/8acf5b303fa41175623aa15c240994240fd9b6c6)  The proposed research used the BAT-ANN classification model with the BigML dataset to predict customer churn in the telecom industry and achieved 89.2% testing accuracy.  Abstract  In recent years, customer churn has been one of the most prominent topics, especially in the telecom industry. The telecommunications industry is producing massive amounts of data every minute. Thus, the telecom industry is seeking more ways to analyze and predict their potential and churn customers. According to telecom analysis, acquiring a new customer is costlier than keeping a current one. To lessen customer churn, it is very compulsory for industries to detect an increase in customer churn factors. The number of service suppliers is increasing daily, especially in the telecom industry. Phishing attacks and fraud are crucial points in customer churn. The aim of this study is to predict customer churn with profitable churn models for retention campaigns to satisfy the business requirement of profit maximization. The proposed research used the BAT-ANN classification model with the BigML dataset to predict customer churn in the telecom industry. The proposed model achieved 89.2% testing accuracy. |
| 25 | @article{SBrinthakumari2023mSVMAA,  title={mSVM: An Approach for Customer Churn Prediction using modified Support Vector Machine and various Machine Learning Techniques},  author={Priyanka Dhiraj Sananse Punam Ba S Brinthakumari},  journal={Proceeding International Conference on Science and Engineering},  year={2023}} | [mSVM: An Approach for Customer Churn Prediction using modified Support Vector Machine and various Machine Learning Techniques](https://www.semanticscholar.org/paper/mSVM%3A-An-Approach-for-Customer-Churn-Prediction-and-S-Brinthakumari/be6b03e5dbc808a95dec6ca560f87004cef93e08)  This research work has proposed a customer churn prediction using a modified Support Vector Machine Learning (mSVM) classifier which produces higher accuracy over the traditional machine learning algorithm for different cross-validation methods.  Abstract  Consumer attrition is a major issue across the world companies and that is one of their primary worries. Organizations, specifically in the telecommunication sector, are working to build technologies to forecast future customer churn due to the obvious direct impact on profits. It's essential to define the variables that lead to customer churn before making the required efforts to minimize churn. An important impact was the creation of an attrition estimation method that aids telecommunication corporations in predicting which consumers are willing to churn. The methodology proposed in this study employs mathematical methodologies on a big data framework to provide a unique strategy to feature design and evaluation. This research work has proposed a customer churn prediction using a modified Support Vector Machine Learning (mSVM) classifier. The significant contribution of this research is that we have changed the distance function of SVM in both training and testing. Similar machine learning algorithms are also validated on similar datasets such as Naïve Bayes (NB), K-nearest Neighbor (KNN) and Decision Tree (DT) with J48 classifiers. The BigML dataset has been used for detecting churn in the telecom industry in a real-time scenario. In an extensive experimental analysis, we demonstrate mSVM which produces higher accuracy over the traditional machine learning algorithm for different cross-validation methods. |
| 26 | @article{Kumar2023AutomationOC,  title={Automation of Credit Card Customer Churn Analysis using Hybrid Machine Learning Models},  author={R. P. Ram Kumar and B. Sahithi and K. Neeharika and M. Shivaleela and Digvijay Singh and K. Rama and Krishna Reddy},  journal={E3S Web of Conferences},  year={2023}} | [Automation of Credit Card Customer Churn Analysis using Hybrid Machine Learning Models](https://www.semanticscholar.org/paper/Automation-of-Credit-Card-Customer-Churn-Analysis-Kumar-Sahithi/280384cfc5354a704bb2f2528c6ac67358c47f1b)  The aim of the paper is to create a model that accurately predicts customers who are most likely to stop using CC, and gathering and analyzing customer information from Kaggle, including transaction history, demographics and credit card usage patterns for prediction.  Abstract  Credit Card Customer Churn Analysis (C4A) is a phenomenon where customers stop using a specific business credit card service. Predicting customer churn is crucial for Credit Card (CC) companies because it enables them to spot at-risk customers and take precautions to retain them. The aim of the paper named C4A is to create a model that accurately predicts customers who are most likely to stop using CC. The paper involves gathering and analyzing customer information from Kaggle, including transaction history, demographics and credit card usage patterns for prediction. Machine learning algorithms namely, Logistic Regression, KNN, XGBoost Classifier, Decision Tree and Hybrid Models integrating Logistic Regression and KNN, Logistic Regression and Decision Tree are used to train to find patterns and correlations that point to customer churn. The accuracy of the proposed method is 0.846 with LR, 0.849 with KNN, 0.90 with a hybrid model integrating LR and KNN, 0.928 by integrating LR and DT, 0.91 with DT, and 0.93 with XGBoost. |
| 27 | @article{Arai2023MethodFC,  title={Method for Characterization of Customer Churn Based on LightBGM and Experimental Approach for Mitigation of Churn},  author={Kohei Arai and Ikuya Fujikawa and Yusuke Nakagawa and Ryo Momozaki and Sayuri Ogawa},  journal={International Journal of Advanced Computer Science and Applications},  year={2023}} | [Method for Characterization of Customer Churn Based on LightBGM and Experimental Approach for Mitigation of Churn](https://www.semanticscholar.org/paper/Method-for-Characterization-of-Customer-Churn-Based-Arai-Fujikawa/a1c1dde1311de9b7614955dd10026f9637729ea2)  A method for customer churn characterization based on LightBGM (Light Gradient Boosting Machine) is proposed and it is found that the first visit date is the most significant factor for churn customer character.  Abstract  A method for customer churn characterization based on LightBGM (Light Gradient Boosting Machine) is proposed. Additionally, experimental approaches for mitigation of churn are conducted through churn prediction. The experiments reveal several churn characteristics such as age dependency, gender dependency (with a high divorce rate among female customers), number of visits dependency (with a higher churn rate for customers with fewer visits), unit price (per hair salon visit) dependency (with a higher withdrawal rate for lowerpriced services), date of first visit dependency (with a high churn rate for recent customers), date of last visit dependency, and menu dependency (with low attrition rates for gray hair dye and high attrition rates for school and child cuts) and so on. Through the experiments, these dependencies are clarified. It is found that the first visit date is the most significant factor for churn customer character. Also, it is found that “distance to hair salon” dependency may be related to the availability of parking lots, although this factor has insignificant impact on the churn rate. Keywords—Churn; LightBGM; churn characteristics; linear regression |
| 28 | @article{Imani2023HyperparameterOA,  title={Hyperparameter Optimization and Combined Data Sampling Techniques in Machine Learning for Customer Churn Prediction: A Comparative Analysis},  author={Mehdi Imani and Hamid Reza Arabnia},  journal={Technologies},  year={2023}} | [Hyperparameter Optimization and Combined Data Sampling Techniques in Machine Learning for Customer Churn Prediction: A Comparative Analysis](https://www.semanticscholar.org/paper/Hyperparameter-Optimization-and-Combined-Data-in-A-Imani-Arabnia/1f1b0482b600973f4b0d621647071ab9b4208926)  CatBoost demonstrates superior performance compared to other machine learning models, achieving an outstanding 93% following the application of Optuna hyperparameter optimization, and both XGBoost and CatBoost exhibit exceptional performance, recording remarkable scores of 91%.  Abstract  This paper explores the application of various machine learning techniques for predicting customer churn in the telecommunications sector. We utilized a publicly accessible dataset and implemented several models, including Artificial Neural Networks, Decision Trees, Support Vector Machines, Random Forests, Logistic Regression, and gradient boosting techniques (XGBoost, LightGBM, and CatBoost). To mitigate the challenges posed by imbalanced datasets, we adopted different data sampling strategies, namely SMOTE, SMOTE combined with Tomek Links, and SMOTE combined with Edited Nearest Neighbors. Moreover, hyperparameter tuning was employed to enhance model performance. Our evaluation employed standard metrics, such as Precision, Recall, F1-score, and the Receiver Operating Characteristic Area Under Curve (ROC AUC). In terms of the F1-score metric, CatBoost demonstrates superior performance compared to other machine learning models, achieving an outstanding 93% following the application of Optuna hyperparameter optimization. In the context of the ROC AUC metric, both XGBoost and CatBoost exhibit exceptional performance, recording remarkable scores of 91%. This achievement for XGBoost is attained after implementing a combination of SMOTE with Tomek Links, while CatBoost reaches this level of performance after the application of Optuna hyperparameter optimization. |
| 29 | @article{Babatunde2023ClassificationOC,  title={Phân loại mô hình dự đoán tỷ lệ rời bỏ khách hàng cho ngành viễn thông bằng cách sử dụng phân tích phương sai},  author={Ronke Seyi Babatunde and Sulaiman Olaniyi Abdulsalam and Olanshile Abdulkabir Abdulsalam and Micheal Olaolu Arowolo},  journal={IAES International Journal of Artificial Intelligence (IJ-AI)},  year={2023}} | [Classification of customer churn prediction model for telecommunication industry using analysis of variance](https://www.semanticscholar.org/paper/Classification-of-customer-churn-prediction-model-Babatunde-Abdulsalam/892af2476f8cdc6409400bf7a02be6fc8d23b3c7)  It may be determined that SVM provides the most accurate forecast of customer turnover (95%).  Abstract  Customer predictive analytics has shown great potential for effective churn models. Thriving in today's telecommunications industry, discerning between consumers who are likely to migrate to a competitor is enormous. Having reliable predictive client behavior in the future is required. Machine learning algorithms are essential to predict customer turnovers, and researchers have proposed various techniques. Churn prediction is a problem due to the unequal dispersal of classes. Most traditional machine learning algorithms are ineffective in classifying data. Client cluster with a higher risk has been discovered. A support vector machine is employed as the foundational learner, and a churn prediction model is constructed based on each analysis of variance. The separation of churn data revealed by experimental assessment is recommended for churn prediction analysis. Customer attrition is high, but an instantaneous support can ensure that customer needs are addressed and assess an employee's capacity to achieve customer satisfaction. This study uses an analysis of variance (ANOVA) with a support vector machine (SVM), classification in analyzing risks in telecom systems It may be determined that SVM provides the most accurate forecast of customer turnover (95%). The projected outcomes will allow other organizations to assess possible client turnover and collect customer feedback. |
| 30 | @inproceedings{Kingly2023ComparativeAO,  title={Comparative Analysis of Customer Churn Prediction},  author={Anithaa Kingly},  year={2023}} | [Comparative Analysis of Customer Churn Prediction](https://www.semanticscholar.org/paper/Comparative-Analysis-of-Customer-Churn-Prediction-Kingly/2e8a8f4eb2dcf773a4bc6b359df2d0f2693da499)  Not just in banking and telecommunications, but also in other industries that are heavily reliant on customer engagement, this study examines the most well-known machine learning methods for churn prediction.  Abstract  — The wide variety of service providers are being elevated very swiftly in each business enterprise. For service providers, a rapidly expanding market in each location is leading to a larger subscriber base. Customer acquisition costs are rising as a result of increased competition, new and innovative business models, and better-suited goods. Service providers have learned the value of keeping existing clients happy in such a short period of time. Because of this, it is critical that service providers reduce churn, the occurrence when consumers of a business stop buying from or engaging with the business. Not just in banking and telecommunications, but also in other industries that are heavily reliant on customer engagement, this study examines the most well-known machine learning methods for churn prediction. |
| 31 | @article{Barham2023ARO,  title={A Review on Machine Learning-Based Customer Churn Prediction in the Telecom Industry},  author={Sawsan Barham and Nowfal Aweisi and Ala' F. Khalifeh},  journal={2023 9th International Conference on Control, Decision and Information Technologies (CoDIT)},  year={2023},  pages={2659-2664}} | [A Review on Machine Learning-Based Customer Churn Prediction in the Telecom Industry](https://www.semanticscholar.org/paper/A-Review-on-Machine-Learning-Based-Customer-Churn-Barham-Aweisi/1635a2cfab4ad7e0b5b3c8fea58b4d8c79f33fe5)  The paper summarizes the different machine learning algorithms that are used for customers' churn prediction, with a particular focus on the telecom industry, as well as their accuracy, to provide insights into the effectiveness of these techniques in addressing patients' churn in the telecommunications industry.  Abstract  In today's business landscape, companies are facing significant difficulties in achieving positive client interactions and maintaining customers' satisfaction. As a result, businesses are striving to focus on every aspect related to customers and their behaviors to compete in the industry. This has become increasingly important for building customers' loyalties, given the many opportunities available to customize services or products for each customer. To prevent customers from becoming dissatisfied and leaving, businesses are deploying various techniques to efficiently predict customers' behaviors and identify those who may churn or stop using the company's services or products. The rise of machine learning applications has significantly contributed to addressing the challenge of predicting customers' churn rates. Researchers worldwide are now moving towards applying machine learning techniques in this area. This paper aims to present a review of various studies conducted from 2019 to 2022 that utilized machine-learning techniques to predict customers' churn in the telecom industry. The paper summarizes the different machine learning algorithms that are used for customers' churn prediction, with a particular focus on the telecom industry, as well as their accuracy, to provide insights into the effectiveness of these techniques in addressing customers' churn in the telecom industry. |
| 32 | @article{Priyanshu2023CustomerCA,  title={Customer Churn Analysis using Spark and Hadoop},  author={Verma Priyanshu and Sharma Ishan and Deshmukh Sonia and Vashisht Rohit},  journal={Int. J. Perform. Eng.},  year={2023},  âm lượng={19},  pages={663},} | [Customer Churn Analysis using Spark and Hadoop](https://www.semanticscholar.org/paper/Customer-Churn-Analysis-using-Spark-and-Hadoop-Priyanshu-Ishan/63f7c1f67e8ea8928c314d701113c5bc75a6a56e)  Predicting Customer churn is one of the telecommunication industry's biggest challenges. Why did their customers quit using their product, site, service, or subscription? Machine learning with Spark and Hadoop has considerably increased the ability to predict customer behaviours. The most popular predictive models, such as logistic regression, Binary Classification Evaluator, and Multi Classification Evaluator, have been used in the prediction process. Enhancing and outfit approaches are used on the training dataset to examine the impact on model effectiveness. Additionally, to further optimize the hyperparameters and produce the models, a K-fold cross-validation method is utilized to train the dataset. Finally, the test data were examined by the AUC-ROC curve and confusion matrix. In this research, an adaptation of Spark and Hadoop frameworks is made to predict customer churn. The data is pre-processed, feature analyses are performed, and the feature selection is carried out using the Vector Assembler algorithm. This study aims to analyse customer behaviors by using a dataset. |
| 33 | @article{Suh2023TrendsAF,  title={Trends and Future Directions in Artificial Intelligence-based Customer Churn Prediction Technology},  author={Young-Jung Suh},  journal={Journal of Digital Contents Society},  year={2023} | [Trends and Future Directions in Artificial Intelligence-based Customer Churn Prediction Technology](https://www.semanticscholar.org/paper/Trends-and-Future-Directions-in-Artificial-Customer-Suh/9821e31de75eb18d778893cb451c9e07629e17e6)  Customer churn prediction has emerged as a pivotal strategy for survival in the fiercely competitive markets of major companies in developed countries. With the growth and spread of big data technology, companies have access to vast and diverse customer data. Additionally, the rapid growth of machine learning technologies has enabled companies to leverage big data in a significantly more effective and systematic way to address customer churn. In this study, we analyzed churn prediction technologies used in various business fields, such as marketing, IT, telecommunications, finance, and games. Based on this, for companies to apply practical churn prediction modeling in terms of business, we have suggested future directions for technical factors that need to be considered, from labeling to inference performance indicators of models actually served. |
| 34 | @article{Mohajon2023ChurnPW,  title={Churn Prediction With Explainability for the Customers of Telecom Industry},  author={Joydwip Mohajon and Abu Shamim Mohammad Arif},  journal={2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)},  year={2023},  pages={179-183}} | [Churn Prediction With Explainability for the Customers of Telecom Industry](https://www.semanticscholar.org/paper/Churn-Prediction-With-Explainability-for-the-of-Mohajon-Arif/8d958b7ff31ce7c3f0ca49f32308fd940a2d21fb)  A churn prediction methodology for the telecom industry that aims at better identifying the customers who might churn is presented and new feature crosses are explored in this research which are found to be impactful in predicting customer churn.  Abstract  Customer churn prediction is an essential task in any market segment that deals with customers. The Telecom industry is one such market sector where customer management plays a crucial role. From existing research, it is clear that the cost of retaining the current customers is much lower than the cost of gaining new customers. Hence, the prior detection of the churn customers becomes very important. This paper presents a churn prediction methodology for the telecom industry that aims at better identifying the customers who might churn. Seven machine learning models are trained using three publicly available datasets. The experimental results of this study reveal great performances compared to existing works in terms of recall, F1-score, and AUC-score. Moreover, new feature crosses are explored in this research which are found to be impactful in predicting customer churn. Finally, the best models are explained using SHapley Additive exPlanations (SHAP), an explainable AI technique. |