1. **Tóm tắt**

Dự đoán sự rời đi của khách hàng giúp cho doanh nghiệp đưa ra các chiến lược, giải pháp hợp lý cho ngân hàng hạn chế sự rời đi của khách hàng, đáp ứng nhu cầu sử dụng của khách hàng. Nhóm sẽ sử dụng Pyspark để xây dựng 3 model Machine Learning đó là: Logistic Regression, Random Forest, Gradient Boosting Tree và đưa ra mô hình phù hợp nhất cho doanh nghiệp có thể áp dụng để có thể hạn chế sự rời đi của khách hàng trong ngân hàng của mình. Nhóm sử dụng 2 phương pháp để đánh giá model, phương pháp thứ nhất là sử dụng độ đo accuracy, phương pháp thứ hai là sử dụng cross-validation chọn siêu tham số cho mô hình. Kết quả đạt được sau khi thử nghiệm trên hai phương pháp thì mô hình thuật toán Random Forest cho kết quả tốt nhất cho bài toán.

1. **Giới thiệu**

Sự rời đi của khách hàng (còn gọi là sự tiêu hao khách hàng) xảy ra khi khách hàng ngưng sử dụng sản phẩm hoặc dịch vụ của công ty. Khách hàng rời đi sẽ ảnh hưởng đến lợi nhuận của công ty. Các nhà phân tích cho rằng việc có được khách hàng mới sẽ tốn kém gấp 4 – 5 lần so với việc giữ chân khách hàng hiện tại. Do đóc việc phân tích sự rời đi của khách hàng giúp cho doanh nghiệp có thể xác định các vấn đề trong dịch vụ của mình, đưa ra các giải pháp dẫn đến sự hài lòng của khách hàng và giúp giữ chân khách hàng cao hơn.

Input của bài toán là tập hợp các thông tin của khách hàng như {mã khách hàng, điểm tín dụng, quốc gia, giới tính, tuổi, thời gian vay, số dư, số sản phẩm (dịch vụ) sử dụng, thẻ tín dụng, thành viên thường xuyên giao dịch, mức lương ước tính}. Nhóm chúng em sử dụng các thuật toán {Logistic Regression, Random Forest, , Gradient Boosting Tree} để dự đoán {sự rời đi của khách hàng}

1. **Dữ liệu**

Tập dữ liệu của nhóm em được lấy từ Kaggle với định dạng file là csv gồm có 10000 dòng và 12 cột bao gồm các thuộc tính:

CustomerID: mã khách hàng

Credit Score: điểm tín dụng Country: quốc gia

Gender: giới tính

Age: tuổi

Tenure: thời gian vay

Balance: số dư

Products Numbers: số sản phẩm (dịch vụ) sử dụng

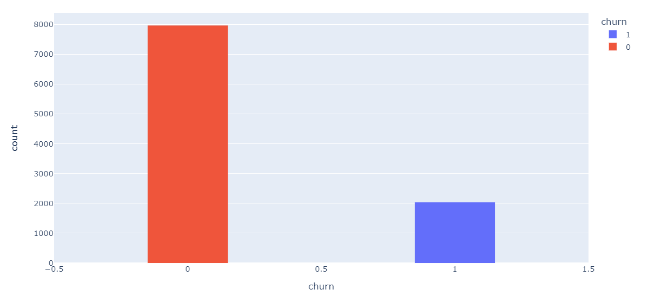
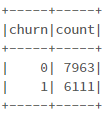
Credit Card: thẻ tín dụng

Active Member: thành viên thường xuyên giao dịch

Estimated Salary: mức lương ước tính

Churn: sự rời đi của khách hàng

Qua phần phân tích EDA, nhóm nhận thấy rằng biến Churn bị mất cân bằng giữ giá trị 0 (khách hàng không rời đi) và 1 (khách hàng rời đi), do đó trong phần tiền xử lý dữ liệu nhóm em có sử dụng phương pháp oversampling để cân bằng lại giá trị của biến Churn trong tập dữ liệu.

**6. Kết luận**

Sau khi áp dụng 3 thuật toán vào tập dữ liệu và dung độ đo accuracy để đánh giá mức độ hiệu quả của thuật toán thì kết quả mà nhóm thu được kết quả như sau:

**Dùng tham số mặc định để tính accuracy**

Logistic Regression: accuracy = 0.70447

Random Forest: accuracy = 0.7739 (numTrees=5, maxDepth=4)

Gradient Boosting Tree: accuracy = 0.7928 (maxIter: int = 20, maxDepth: int = 5)

**Chọn tham số để tính accuracy**

Logistic Regression: accuracy = 0.70702 (regParam=0.001, elasticNetParam=0.75)

Random Forest: accuracy = 0.9404 (numTrees=50,maxDepth=20)

Gradient Boosting Tree: accuracy = 0.9015 (maxIter: int = 10, maxDepth: int = 20)

Thuật toán Random Forest sẽ cho kết quả tốt nhất do có giá trị accuracy lớn nhất trong 3 thuật toán mà nhóm sử dụng.

**8. Đóng góp**

|  |  |
| --- | --- |
| Thành viên | Công việc |
| Lê Phương Nam | Thuật toán Logistic Regression |
| Võ Thành Đạt | Thuật toán Gradient Boosting Tree |
| Nguyễn Phạm Duy Khiêm | Thuật toán Random Forest |
| Nguyễn Thị Nhả Thư | Phân tích EDA |

**9. Tham khảo**

[1] Tạ Quốc Bảo, *Thư viện Matplotlib – Thư viện Python dùng để vẽ đồ thị,* 23/07/2019

<https://code24h.com/thu-vien-matplotlib-8211-thu-vien-python-dung-de-ve-do-thi-d30561.htm>

[2] W3School, *NumPy Tutorial*

<https://www.w3schools.com/python/numpy/numpy_intro.asp>

[3] Igor Radovanovic, *Sklearn – An Introduction Guide to Machine Learning,* 15/07/2022

<https://algotrading101.com/learn/sklearn-guide/>

[4] W3School, *Pandas Tutorial*

<https://www.w3schools.com/python/pandas/default.asp>

[5] GeeksForGeeks, *Python Seaborn Tutorial,* 22/11/2022

<https://www.geeksforgeeks.org/python-seaborn-tutorial/>

[6] *PySpark Tutorial*

<https://www.tutorialspoint.com/pyspark/index.htm>