[DA_Project TIMA] Task 9_Phân tích dự báo

Quy trình phân tích dự báo

1. 1. **Xác định mục tiêu dự báo** → 2. **Thu thập dữ liệu lịch sử** → 3. **Làm sạch và chuẩn bị dữ liệu** → 4. **Khám phá và phân tích dữ liệu** → 5. **Chọn mô hình dự báo** → 6. **Xây dựng và huấn luyện mô hình** → 7. **Đánh giá mô hình** → 8. **Dự báo và lập kế hoạch** → 9. **Giám sát và cập nhật mô hình**.

Quy trình chi tiết

1. Xác định mục tiêu dự báo

Mục tiêu dự báo giúp xác định rõ điều cần dự đoán dựa trên bộ dữ liệu này. Ví dụ:

- Dự báo khả năng **vay tiền thành công** của khách hàng dựa trên các yếu tố như điểm tín dụng, thu nhập, lịch sử vay nợ.
- Dự báo số tiền **giải ngân** hoặc **số tiền còn lại** cần thanh toán sau một thời gian.
- Dự báo **khả năng trễ hạn** thanh toán dựa trên các thông tin như "LongestOverdue", "HasBadDebt", "HasLatePayment", "Salary".

2. Thu thập dữ liệu lịch sử

Để phân tích dự báo, cần thu thập dữ liệu lịch sử của các khoản vay từ các nguồn có sẵn. Các trường dữ liệu trong bộ dữ liệu cung cấp các thông tin cần thiết, ví dụ:

- **Số tiền đăng ký vay ban đầu** (`SoTienDKVayBanDau`), **Tiền giải ngân** (`TienGiaiNgan`), **Số tiền còn lại** (`SoTienConLai`) giúp theo dõi quá trình vay và thanh toán.
- **Điểm tín dụng** (`TS_CREDIT_SCORE_V2`) và các chỉ số tài chính khác như **lương** (`Salary`), **Thu nhập** (`ReceiveYourIncomeSalary`) sẽ cung cấp thông tin quan trọng để dự đoán khả năng vay tiền và trả nợ.
- **Trạng thái khoản vay** (`Trạng thái`), **Số lượng khoản vay trước đó** (`NumberOfLoans`), **Có nợ xấu** (`HasBadDebt`), **Có trễ hạn** (`HasLatePayment`) để theo dõi quá trình thanh toán và tình trạng nợ.

3. Làm sạch và chuẩn bị dữ liệu

Sau khi thu thập dữ liệu, cần xử lý và làm sạch bộ dữ liệu:

- **Loại bỏ dữ liệu thiếu hoặc không đầy đủ**: Kiểm tra các giá trị thiếu trong các trường dữ liệu quan trọng như **Số tiền đăng ký vay** và **Điểm tín dụng**.
- **Chuyển đổi kiểu dữ liệu**: Đảm bảo các trường dữ liệu như **Ngày tháng** (ví dụ: `CheckTime`, `Thời gian đã sống`) được chuyển đổi thành định dạng ngày tháng hợp lý.
- **Xử lý các giá trị ngoại lệ**: Ví dụ, nếu có các giá trị **tiền giải ngân** vượt quá mức bình thường, cần xác định và xử lý chúng.
- **Tạo các biến mới**: Có thể tạo các chỉ số mới, ví dụ: tỷ lệ giải ngân thành công (so sánh giữa **Số tiền đăng ký vay** và **Tiền giải ngân**).

4. Khám phá và phân tích dữ liệu

Khám phá dữ liệu giúp hiểu rõ hơn về các mối quan hệ giữa các trường dữ liệu và các yếu tố ảnh hưởng đến việc vay tiền và khả năng trả nợ. Các hoạt động có thể thực hiện:

- **Thống kê mô tả**: Xem các chỉ số như trung bình, độ lệch chuẩn cho các trường như **Số tiền vay**, **Điểm tín dụng**, **Thu nhập**.
- **Trực quan hóa dữ liệu**: Sử dụng biểu đồ (histogram, boxplot, scatter plot) để kiểm tra sự phân bố của các trường dữ liệu quan trọng như **Điểm tín dụng** và **Tiền giải ngân**.

- **Phân tích mối quan hệ**: Sử dụng bảng tương quan hoặc kiểm tra các mối quan hệ giữa các biến (ví dụ: **Số tiền đăng ký vay** và **Tiền giải ngân**).

5. Chọn mô hình dự báo

Dựa trên mục tiêu của dự báo, chọn mô hình phù hợp. Các mô hình có thể áp dụng bao gồm:

- **Mô hình hồi quy tuyến tính** (Linear Regression): Dùng để dự báo các giá trị liên tục, ví dụ như **Tiền giải ngân** hoặc **Số tiền còn lại**.
- **Mô hình phân loại** (Logistic Regression, Random Forest, XGBoost): Dùng để dự đoán các giá trị phân loại, ví dụ như khả năng **trễ hạn** (trả tiền đúng hạn hay không).
- **Mô hình chuỗi thời gian** (ARIMA, Prophet): Nếu dữ liệu có tính chu kỳ hoặc theo thời gian, có thể áp dụng mô hình chuỗi thời gian để dự báo **Số tiền giải ngân** theo tháng/quý/năm.

6. Xây dựng và huấn luyện mô hình

- **Chia dữ liệu thành tập huấn luyện và tập kiểm tra**: Chia bộ dữ liệu thành hai phần (70-80% cho huấn luyện và 20-30% cho kiểm tra).
- **Huấn luyện mô hình**: Sử dụng các thuật toán học máy (Machine Learning) như hồi quy tuyến tính, cây quyết định, hoặc mạng nơ-ron để huấn luyện mô hình.
- **Tuning tham số**: Sử dụng các kỹ thuật như Grid Search hoặc Random Search để tối ưu hóa các tham số của mô hình.

7. Đánh giá mô hình

Đánh giá mô hình dựa trên các chỉ số phù hợp với mục tiêu dự báo:

- **Đối với mô hình hồi quy**: Sử dụng các chỉ số như **R-squared**, **MSE (Mean Squared Error)**, hoặc **RMSE (Root Mean Squared Error)** để đánh giá độ chính xác của mô hình.
- **Đối với mô hình phân loại**: Sử dụng các chỉ số như **Accuracy**, **Precision**, **Recall**, **F1-score**, và **ROC Curve** để đánh giá hiệu quả của mô hình phân loại.

8. Dự báo và lập kế hoạch

Sau khi mô hình được huấn luyện và đánh giá, có thể bắt đầu dự báo và lập kế hoạch dựa trên kết quả:

- **Dự báo số tiền giải ngân**: Dự báo số tiền sẽ được giải ngân cho các khách hàng trong tương lai.
- **Dự báo rủi ro trễ hạn**: Dự báo khả năng trễ hạn hoặc không trả nợ của các khách hàng dựa trên các yếu tố như **Điểm tín dụng**, **Số lượng khoản vay trước đó**, và **Thu nhập**.

9. Giám sát và cập nhật mô hình

Dữ liệu và các yếu tố liên quan có thể thay đổi theo thời gian, do đó cần giám sát và cập nhật mô hình thường xuyên:

- **Giám sát hiệu suất**: Theo dõi mô hình trong thời gian thực để xem xét độ chính xác của các dự báo.
- **Cập nhật mô hình**: Dựa trên các thay đổi trong dữ liệu và các yếu tố tác động, có thể cần cập nhật mô hình thường xuyên hoặc huấn luyện lại mô hình với dữ liệu mới.

---

10. Hồi quy tuyến tính (Linear Regression) và Hồi quy logistic (Logistic Regression) là hai kỹ thuật phổ biến trong phân tích dữ liệu và dự báo. Dưới đây là cách mà từng mô hình có thể được áp dụng đối với bộ dữ liệu tín dụng bạn cung cấp, nhằm dự báo các chỉ tiêu liên quan đến số tiền vay, tình trạng tín dụng, v.v.

1. **Hồi quy tuyến tính (Linear Regression)**

Hồi quy tuyến tính là một mô hình dùng để dự báo các giá trị liên tục. Nó cố gắng tìm ra một hàm số tuyến tính giữa biến phụ thuộc (biến cần dự báo) và các biến độc lập (các yếu tố ảnh hưởng).

Áp dụng trong dự báo số tiền vay

- **Mục tiêu dự báo**: Dự báo **số tiền vay** mà khách hàng sẽ đăng ký trong tương lai (dựa trên các yếu tố như thu nhập, thời gian đã sống, điểm tín dụng, số tiền vay ban đầu, v.v.).
- **Biến phụ thuộc (Dependent variable)**: **Số tiền vay** (`SoTienDKVayBanDau` hoặc `TienGiaiNgan`), vì đây là giá trị liên tục cần dự báo.
- **Biến độc lập (Independent variables)**: Các yếu tố ảnh hưởng đến số tiền vay, chẳng hạn:
  - **Thu nhập của khách hàng** (`Salary`, `ReceiveYourIncomeSalary`).
  - **Điểm tín dụng** (`TS_CREDIT_SCORE_V2`).
  - **Tình trạng gia đình** (`RelativeFamilyName`, `FullNameFamily`).
  - **Số năm sống tại địa chỉ hiện tại** (`Thời gian đã sống`).
  - **Loại công việc và mức lương** (`JobName`, `DescriptionPositionJob`).
  - **Thông tin khoản vay trước đó** (`LoanID`, `Tiền giải ngân`).

Công thức hồi quy tuyến tính:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_n \cdot X_n + \epsilon$$

Trong đó:

- Y là số tiền vay (biến phụ thuộc).
- $X_1, X_2, \ldots, X_n$ là các yếu tố ảnh hưởng (biến độc lập).
- $\beta_0, \beta_1, \ldots, \beta_n$ là các tham số cần học (hệ số hồi quy).
- ϵ là sai số ngẫu nhiên.

Quy trình

1. **Xây dựng mô hình**: Sử dụng các biến độc lập như thu nhập, điểm tín dụng, và các đặc điểm khách hàng để dự đoán số tiền vay.
2. **Đánh giá mô hình**: Kiểm tra độ chính xác của mô hình thông qua các chỉ số như R^2, RMSE (Root Mean Squared Error) để đánh giá mức độ phù hợp của mô hình.
3. **Dự báo**: Dự báo số tiền vay cho các khách hàng mới dựa trên mô hình đã huấn luyện.

---

4. **Hồi quy logistic (Logistic Regression)**

Hồi quy logistic là một mô hình dùng để dự báo các biến phân loại, nơi kết quả là một trong các giá trị rời rạc, thường là "0" hoặc "1". Mô hình này được sử dụng khi biến phụ thuộc là một biến phân loại, chẳng hạn như **tình trạng tín dụng** (có nợ xấu hay không, trả nợ đúng hạn hay không).

Áp dụng trong dự báo tình trạng tín dụng

- **Mục tiêu dự báo**: Dự báo **tình trạng tín dụng** của khách hàng, ví dụ như liệu khách hàng có nợ xấu hay không.
- **Biến phụ thuộc (Dependent variable)**: **Tình trạng tín dụng** (`HasBadDebt`, `HasLatePayment`). Đây là các biến phân loại với giá trị "0" (không nợ xấu) hoặc "1" (có nợ xấu).
- **Biến độc lập (Independent variables)**: Các yếu tố có thể ảnh hưởng đến tình trạng tín dụng, chẳng hạn:
  - **Điểm tín dụng** (`TS_CREDIT_SCORE_V2`).
  - **Số năm sống tại địa chỉ hiện tại** (`Thời gian đã sống`).
  - **Số tiền vay ban đầu** (`SoTienDKVayBanDau`).
  - **Thu nhập** (`Salary`, `ReceiveYourIncomeSalary`).
  - **Quá hạn thanh toán** (`LongestOverdue`).
  - **Thông tin công ty và công việc** (`JobName`, `NameCompany`, `Salary`).

Công thức hồi quy logistic:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_n \cdot X_n)}}$$

Trong đó:

- $P(Y = 1)$ là xác suất khách hàng có nợ xấu (tình trạng tín dụng xấu).
- $X_1, X_2, \ldots, X_n$ là các yếu tố ảnh hưởng (biến độc lập).
- $\beta_0, \beta_1, \ldots, \beta_n$ là các tham số cần học (hệ số hồi quy).
- e là cơ số tự nhiên.

Quy trình

1. **Xây dựng mô hình**: Dựa trên các đặc điểm của khách hàng (điểm tín dụng, thu nhập, tình trạng vay) để dự đoán khả năng có nợ xấu.
2. **Đánh giá mô hình**: Đánh giá mô hình dựa trên các chỉ số như **accuracy** (độ chính xác), **confusion matrix** (ma trận nhầm lẫn), **ROC-AUC** (Area Under Curve) để đánh giá chất lượng mô hình phân loại.
3. **Dự báo**: Dự báo xác suất khách hàng có nợ xấu dựa trên các yếu tố đã cho.

Kết luận

- **Hồi quy tuyến tính** thường dùng để dự báo các giá trị liên tục như **số tiền vay** hoặc **số tiền giải ngân** trong các bài toán tín dụng.
- **Hồi quy logistic** dùng để dự báo các vấn đề phân loại như **tình trạng tín dụng** (có nợ xấu hoặc không) hay **khả năng thanh toán đúng hạn**.

Giải thích các chỉ số đánh giá mô hình

- **Ma trận nhầm lẫn với 4 nhãn**

	Predicted A	Predicted B	Predicted C	Predicted D
Actual A	**TP(A)**: 50	**FP(B)**: 10	**FP(C)**: 5	**FP(D)**: 2
Actual B	**FP(A)**: 8	**TP(B)**: 60	**FP(C)**: 4	**FP(D)**: 1
Actual C	**FP(A)**: 3	**FP(B)**: 2	**TP(C)**: 70	**FP(D)**: 6
Actual D	**FP(A)**: 1	**FP(B)**: 3	**FP(C)**: 7	**TP(D)**: 75

Giải thích các phần trong ma trận nhầm lẫn:

TP (True Positive): Số lượng mẫu mà mô hình phân loại chính xác vào lớp đó. Ví dụ: **TP(A)** là số lượng mẫu thực sự thuộc lớp A và được mô hình dự đoán đúng là A.

FP (False Positive): Số lượng mẫu mà mô hình dự đoán nhầm sang một lớp khác. Ví dụ: **FP(B)** là số lượng mẫu thực sự thuộc lớp A nhưng mô hình lại dự đoán là lớp B.

- **Chỉ số ACC (Accuracy)**

Chỉ số ACC (Accuracy) là một chỉ số trong học máy (machine learning) dùng để đo lường độ chính xác của một mô hình phân loại. Nó tính toán tỷ lệ dự đoán đúng trên tổng số dự đoán mà mô hình đưa ra.

Công thức tính chỉ số ACC như sau:

$$\text{ACC} = \frac{\text{Số lượng dự đoán đúng}}{\text{Tổng số mẫu}}$$

Trong đó:

- **Số lượng dự đoán đúng** là số mẫu mà mô hình dự đoán chính xác.
- **Tổng số mẫu** là tổng số mẫu trong tập dữ liệu.

- ## Chỉ số RMSE (Root Mean Squared Error)

Chỉ số RMSE (Root Mean Squared Error) là một chỉ số phổ biến trong học máy và thống kê để đánh giá mức độ sai số giữa giá trị thực tế và giá trị dự đoán của mô hình hồi quy. RMSE cho biết mức độ sai lệch trung bình giữa các dự đoán của mô hình và các giá trị thực tế trong dữ liệu.

Công thức tính RMSE như sau:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Trong đó:

- n là số lượng mẫu dữ liệu.
- y_i là giá trị thực tế của mẫu thứ i.
- $\hat{y}_i$ là giá trị dự đoán của mẫu thứ i.
- $(y_i - \hat{y}_i)^2$ là bình phương sai số giữa giá trị thực tế và giá trị dự đoán.

Ý nghĩa của RMSE:

- **RMSE nhỏ** cho thấy mô hình dự đoán chính xác và gần với giá trị thực tế.
- **RMSE lớn** cho thấy sai số giữa các dự đoán và giá trị thực tế cao, tức là mô hình không chính xác.