[DA_Project TIMA] Task 8_Phân tích Chẩn đoán

Phân tích Chẩn đoán

1. Kiểm tra sự liên quan giữa "Lãi suất" và "Số tiền vay"

- **Mục đích:** Để xác định mối quan hệ giữa lãi suất và số tiền vay.
- **Hàm cần sử dụng:** sns.regplot(), corr()
- **Ý nghĩa của hàm:** sns.regplot() vẽ đường hồi quy để xác định mối quan hệ, corr() tính toán hệ số tương quan.

2. Phân tích ảnh hưởng của "Giới tính" tới "Số tiền vay"

- **Mục đích:** Kiểm tra xem giới tính có ảnh hưởng đến số tiền vay hay không.
- **Hàm cần sử dụng:** groupby(), mean()
- **Ý nghĩa của hàm:** groupby() nhóm theo giới tính và mean() tính số tiền vay trung bình cho từng giới tính.

3. Phân tích sự ảnh hưởng của "Tuổi" đối với "Tiền giải ngân"

- **Mục đích:** Xác định ảnh hưởng của độ tuổi đến tiền giải ngân.
- **Hàm cần sử dụng:** sns.regplot(), corr()
- **Ý nghĩa của hàm:** sns.regplot() vẽ đồ thị hồi quy và corr() tính toán hệ số tương quan giữa tuổi và tiền giải ngân.

4. Kiểm tra sự thay đổi của "Lãi suất" theo các nhóm "Ngành nghề"

- **Mục đích:** Phân tích sự thay đổi của lãi suất theo các ngành nghề khác nhau.
- **Hàm cần sử dụng:** groupby(), mean()
- **Ý nghĩa của hàm:** groupby() nhóm theo ngành nghề và mean() tính giá trị trung bình của lãi suất trong từng nhóm.

5. Kiểm tra sự ảnh hưởng của "Thời gian đã sống" tới "Điểm tín dụng"

- **Mục đích:** Để phân tích ảnh hưởng của thời gian đã sống đến điểm tín dụng.
- **Hàm cần sử dụng:** sns.regplot(), corr()
- **Ý nghĩa của hàm:** sns.regplot() vẽ đồ thị hồi quy, corr() tính toán mối tương quan giữa thời gian đã sống và điểm tín dụng.

6. Xác định mối quan hệ giữa "Khu vực" và "Tỷ lệ nợ xấu"

- **Mục đích:** Kiểm tra mối quan hệ giữa khu vực và tỷ lệ nợ xấu.
- **Hàm cần sử dụng:** groupby(), mean()
- **Ý nghĩa của hàm:** groupby() nhóm theo khu vực và mean() tính tỷ lệ nợ xấu trong từng khu vực.

7. Phân tích mối quan hệ giữa "Giới tính" và "Lãi suất"

- **Mục đích:** Kiểm tra sự khác biệt giữa giới tính và mức lãi suất.
- **Hàm cần sử dụng:** groupby(), mean()
- **Ý nghĩa của hàm:** groupby() nhóm theo giới tính và mean() tính mức lãi suất trung bình cho từng giới tính.

8. Tìm hiểu ảnh hưởng của "Số tiền vay" đến "Nợ xấu"

- **Mục đích:** Để xem xét số tiền vay có ảnh hưởng đến nợ xấu hay không.
- **Hàm cần sử dụng:** sns.regplot(), corr()

- **Ý nghĩa của hàm:** sns.regplot() vẽ đồ thị hồi quy và corr() tính hệ số tương quan giữa số tiền vay và nợ xấu.

9. Kiểm tra mối quan hệ giữa "Thu nhập" và "Số tiền vay"

- **Mục đích:** Để xem xét mối quan hệ giữa thu nhập và số tiền vay của khách hàng.
- **Hàm cần sử dụng:** sns.regplot(), corr()
- **Ý nghĩa của hàm:** sns.regplot() vẽ đồ thị hồi quy và corr() tính hệ số tương quan giữa thu nhập và số tiền vay.

10. Phân tích mối quan hệ giữa "Điểm tín dụng" và "Thu nhập"

- **Mục đích:** Kiểm tra sự ảnh hưởng của thu nhập đến điểm tín dụng.
- **Hàm cần sử dụng:** sns.regplot(), corr()
- **Ý nghĩa của hàm:** sns.regplot() vẽ đồ thị hồi quy, corr() tính toán hệ số tương quan giữa điểm tín dụng và thu nhập.

11. Phân tích ảnh hưởng của "Thành phố" tới "Số tiền vay"

- **Mục đích:** Để xem thành phố có ảnh hưởng đến số tiền vay hay không.
- **Hàm cần sử dụng:** groupby(), mean()
- **Ý nghĩa của hàm:** groupby() nhóm theo thành phố và mean() tính giá trị trung bình của số tiền vay trong từng thành phố.

12. Kiểm tra ảnh hưởng của "Phường" tới "Lãi suất"

- **Mục đích:** Để phân tích sự ảnh hưởng của phường đến mức lãi suất.
- **Hàm cần sử dụng:** groupby(), mean()
- **Ý nghĩa của hàm:** groupby() nhóm theo phường và mean() tính giá trị trung bình của lãi suất.

13. Kiểm tra sự thay đổi "Số tiền vay" theo "Mức lương"

- **Mục đích:** Để xác định mối quan hệ giữa số tiền vay và mức lương của khách hàng.
- **Hàm cần sử dụng:** sns.regplot(), corr()
- **Ý nghĩa của hàm:** sns.regplot() vẽ đồ thị hồi quy và corr() tính toán hệ số tương quan giữa số tiền vay và mức lương.

14. Phân tích ảnh hưởng của "Job Name" đến "Tiền vay"

- **Mục đích:** Để kiểm tra ảnh hưởng của công việc đối với số tiền vay.
- **Hàm cần sử dụng:** groupby(), mean()
- **Ý nghĩa của hàm:** groupby() nhóm theo "Job Name" và mean() tính giá trị trung bình của số tiền vay.

15. Phân tích mối quan hệ giữa "Hình thức cư trú" và "Lãi suất"

- **Mục đích:** Để phân tích mối quan hệ giữa hình thức cư trú và lãi suất.
- **Hàm cần sử dụng:** groupby(), mean()
- **Ý nghĩa của hàm:** groupby() nhóm theo "Hình thức cư trú" và mean() tính giá trị trung bình của lãi suất trong mỗi nhóm.

16. Phân tích sự ảnh hưởng của "Thời gian sống" tới "Điểm tín dụng"

- **Mục đích:** Để kiểm tra sự thay đổi của điểm tín dụng khi thời gian sống thay đổi.
- **Hàm cần sử dụng:** sns.regplot(), corr()
- **Ý nghĩa của hàm:** sns.regplot() vẽ đồ thị hồi quy và corr() tính toán mối tương quan giữa thời gian sống và điểm tín dụng.

17. Kiểm tra ảnh hưởng của "Tình trạng hôn nhân" đến "Nợ xấu"

- **Mục đích:** Để tìm hiểu liệu tình trạng hôn nhân có ảnh hưởng đến nợ xấu không.

- **Hàm cần sử dụng:** groupby(), mean()
- **Ý nghĩa của hàm:** groupby() nhóm theo "Tình trạng hôn nhân" và mean() tính tỷ lệ nợ xấu trong mỗi nhóm.

18. Phân tích mối quan hệ giữa "Số tiền vay" và "Số lần quá hạn"

- **Mục đích:** Để tìm hiểu mối quan hệ giữa số tiền vay và số lần khách hàng quá hạn.
- **Hàm cần sử dụng:** sns.regplot(), corr()
- **Ý nghĩa của hàm:** sns.regplot() vẽ đồ thị hồi quy và corr() tính hệ số tương quan giữa số tiền vay và số lần quá hạn.

19. Phân tích sự ảnh hưởng của "Công ty" tới "Khoản vay"

- **Mục đích:** Để tìm hiểu công ty có ảnh hưởng đến số khoản vay của khách hàng không.
- **Hàm cần sử dụng:** groupby(), mean()
- **Ý nghĩa của hàm:** groupby() nhóm theo công ty và mean() tính giá trị trung bình của số khoản vay.

20. Kiểm tra sự thay đổi của "Lãi suất" theo "Thành phố"

- **Mục đích:** Để phân tích sự thay đổi của lãi suất ở các thành phố khác nhau.
- **Hàm cần sử dụng:** groupby(), mean()
- **Ý nghĩa của hàm:** groupby() nhóm theo thành phố và mean() tính giá trị trung bình của lãi suất.

21. Kiểm tra mối quan hệ giữa "Thu nhập" và "Số lần trễ hạn"

- **Mục đích:** Kiểm tra ảnh hưởng của thu nhập đối với số lần trễ hạn thanh toán.
- **Hàm cần sử dụng:** sns.regplot(), corr()
- **Ý nghĩa của hàm:** sns.regplot() vẽ đồ thị hồi quy và corr() tính toán hệ số tương quan giữa thu nhập và số lần trễ hạn.

22. Tính toán sự phân bố của "Lãi suất" theo nhóm "Giới tính"

- **Mục đích:** Xác định sự khác biệt về phân phối lãi suất giữa các giới tính.
- **Hàm cần sử dụng:** sns.boxplot()
- **Ý nghĩa của hàm:** sns.boxplot() vẽ biểu đồ hộp giúp phân tích sự phân bố của lãi suất theo giới tính.

23. Kiểm tra ảnh hưởng của "Số tiền vay" đối với "Số lần quá hạn"

- **Mục đích:** Xem xét liệu số tiền vay có ảnh hưởng đến số lần quá hạn không.
- **Hàm cần sử dụng:** sns.regplot(), corr()
- **Ý nghĩa của hàm:** sns.regplot() vẽ đồ thị hồi quy và corr() tính toán hệ số tương quan giữa số tiền vay và số lần quá hạn.

24. Kiểm tra sự ảnh hưởng của "Vị trí công ty" đến "Thu nhập"

- **Mục đích:** Xác định mối quan hệ giữa vị trí công ty và thu nhập của khách hàng.
- **Hàm cần sử dụng:** groupby(), mean()
- **Ý nghĩa của hàm:** groupby() nhóm theo vị trí công ty và mean() tính giá trị trung bình của thu nhập.

25. Phân tích sự thay đổi của "Thu nhập" theo "Tình trạng nợ xấu"

- **Mục đích:** Phân tích mối quan hệ giữa thu nhập và tình trạng nợ xấu.
- **Hàm cần sử dụng:** groupby(), mean()
- **Ý nghĩa của hàm:** groupby() nhóm theo tình trạng nợ xấu và mean() tính thu nhập trung bình trong từng nhóm.

Tóm tắt luồng phân tích

Xác định vấn đề → Thu thập dữ liệu liên quan → Phân tích tương quan → Kiểm tra mối quan hệ nhân quả → Chẩn đoán nguyên nhân gốc rễ → Kiểm tra giả thuyết → Cung cấp giải pháp hành động

1. **Xác định vấn đề**

   Vấn đề: Phát hiện sự trì hoãn trong việc thanh toán hoặc tình trạng nợ xấu trong một số khoản vay của khách hàng. Một số khách hàng có thể có nợ xấu hoặc thanh toán chậm trong khi các khách hàng khác không gặp vấn đề này.

   Câu hỏi cần giải quyết: Tại sao một số khách hàng có nợ xấu, trong khi các khách hàng khác không bị ảnh hưởng?

2. **Thu thập dữ liệu liên quan**

   Các trường dữ liệu liên quan:

   Số tiền đăng ký vay ban đầu (`SoTienDKVayBanDau`): Để xác định giá trị khoản vay ban đầu.

   Tiền giải ngân (`TienGiaiNgan`): Số tiền thực tế đã được giải ngân cho khách hàng.

   Tiền gốc còn lại (`SoTienConLai`): Số tiền còn lại mà khách hàng chưa thanh toán.

   Trạng thái (`TrangThai`): Trạng thái của khoản vay, có thể là "Đã thanh toán", "Chậm trả", "Nợ xấu", v.v.

   Lịch sử thanh toán (`HasLatePayment`): Cột này cho biết liệu khách hàng có thanh toán trễ không.

   Thu nhập (`Salary`): Thu nhập của khách hàng, có thể ảnh hưởng đến khả năng thanh toán.

3. **Phân tích tương quan**

   Phân tích mối quan hệ giữa các yếu tố:

   - **Tương quan giữa Thu nhập (`Salary`) và Tình trạng nợ xấu (`TrangThai`)**:
     - Mục tiêu: Kiểm tra xem thu nhập của khách hàng có ảnh hưởng đến khả năng trả nợ không.
     - Sử dụng **`corr()`** hoặc **`sns.heatmap()`** để kiểm tra sự tương quan giữa các biến này.
   - **Tương quan giữa Số tiền vay và Tiền giải ngân**:
     - Mục tiêu: Kiểm tra xem số tiền đăng ký vay có tương quan với số tiền đã giải ngân cho khách hàng hay không.
     - Sử dụng **`corr()`** hoặc **`sns.heatmap()`** để kiểm tra sự tương quan giữa các biến này.
   - **Mối quan hệ giữa Tiền gốc còn lại và Trạng thái khoản vay**:
     - Mục tiêu: Kiểm tra liệu khách hàng còn bao nhiêu tiền gốc chưa thanh toán có liên quan đến trạng thái khoản vay của họ (nợ xấu, trả nợ đúng hạn).
     - Sử dụng **`sns.barplot()`** hoặc **`sns.boxplot()`** để trực quan hóa mối quan hệ này.

4. **Kiểm tra mối quan hệ nhân quả**
   - Mục tiêu: Kiểm tra liệu **Thu nhập** và **Số tiền vay ban đầu** có thực sự ảnh hưởng đến khả năng trả nợ của khách hàng (mối quan hệ nhân quả).
   - Các bước:
     - **Kiểm tra mối quan hệ giữa Thu nhập và Thanh toán trễ**:
       - Hãy kiểm tra liệu thu nhập có phải là nguyên nhân khiến khách hàng thanh toán trễ hay không.
       - Sử dụng **`sns.regplot()`** hoặc **`sns.scatterplot()`** để vẽ biểu đồ và tìm mối quan hệ giữa Thu nhập và việc thanh toán trễ.
     - **Mối quan hệ giữa Tiền vay và Trạng thái nợ xấu**:
       - Sử dụng **`sns.regplot()`** để kiểm tra xem số tiền vay ban đầu có liên quan đến tình trạng nợ xấu của khách hàng.

5. **Chẩn đoán nguyên nhân gốc rễ**
   - Mục tiêu: Xác định nguyên nhân chính dẫn đến tình trạng khách hàng có nợ xấu hoặc thanh toán trễ.
   - Các phân tích:

- **Khách hàng có thu nhập thấp dễ có nợ xấu hơn**:
  - Dựa trên phân tích tương quan, nếu thấy rằng khách hàng có thu nhập thấp có khả năng bị nợ xấu hoặc thanh toán trễ, có thể rút ra kết luận về nguyên nhân.
- **Khách hàng vay số tiền quá lớn có thể gặp khó khăn trong thanh toán**:
  - Nếu thấy có sự tương quan mạnh giữa số tiền vay ban đầu và tình trạng nợ xấu, có thể kết luận rằng việc vay một số tiền quá lớn so với khả năng tài chính của khách hàng là nguyên nhân chính.

6. **Kiểm tra giả thuyết**
   - Giả thuyết:
     - **Giả thuyết 1**: Khách hàng có thu nhập thấp sẽ dễ gặp tình trạng nợ xấu.
     - **Giả thuyết 2**: Khách hàng vay số tiền lớn có khả năng nợ xấu cao hơn.
   - Các kiểm tra:
     - **Kiểm tra giả thuyết 1**: Sử dụng **sns.boxplot()** hoặc **sns.violinplot()** để kiểm tra phân phối thu nhập của khách hàng có liên quan đến tình trạng nợ xấu.
     - **Kiểm tra giả thuyết 2**: Sử dụng **sns.scatterplot()** để kiểm tra mối quan hệ giữa số tiền vay ban đầu và tình trạng nợ xấu.
7. **Cung cấp giải pháp hành động**
   - Giải pháp:
     - **Giải pháp 1**: Đề xuất tăng cường kiểm tra khả năng tài chính của khách hàng trước khi giải ngân khoản vay lớn, đặc biệt đối với khách hàng có thu nhập thấp.
     - **Giải pháp 2**: Xem xét điều chỉnh mức vay đối với những khách hàng có thu nhập không ổn định, giảm thiểu tỷ lệ nợ xấu.
     - **Giải pháp 3**: Đưa ra các chương trình hỗ trợ khách hàng có thể gặp khó khăn trong việc thanh toán như gia hạn thời gian trả nợ hoặc điều chỉnh lãi suất.

Danh sách hàm chính cần sử dụng:

1. Nhóm hàm tóm tắt thông tin thống kê cơ bản

Hàm: describe(), mean(), std(), sum(), median()

- **Ý nghĩa**: Nhóm hàm này được sử dụng để tóm tắt các đặc điểm cơ bản của dữ liệu số trong DataFrame. Các hàm này giúp bạn nắm bắt các chỉ số thống kê cơ bản, cung cấp cái nhìn tổng quan về phân phối và sự phân tán của dữ liệu.
- **Chức năng**:
  - **describe()**: Trả về các chỉ số thống kê tổng quát như số lượng giá trị (count), giá trị trung bình (mean), độ lệch chuẩn (std), giá trị nhỏ nhất (min), các phân vị (25%, 50%, 75%), và giá trị lớn nhất (max).
  - **mean()**: Tính giá trị trung bình của một cột.
  - **std()**: Tính độ lệch chuẩn, đo mức độ phân tán của dữ liệu xung quanh giá trị trung bình.
  - **sum()**: Tính tổng các giá trị trong một cột.
  - **median()**: Tính giá trị trung vị, giá trị giữa của dữ liệu khi đã được sắp xếp theo thứ tự.

2. Nhóm hàm phân tích phân phối dữ liệu

Hàm: value_counts(), hist(), boxplot()

- **Ý nghĩa**: Nhóm hàm này được sử dụng để phân tích phân phối của dữ liệu. Chúng giúp bạn hiểu sự phân bổ của các giá trị trong cột, phát hiện các giá trị ngoại lệ, và nhận diện các đặc điểm phân phối như độ lệch và sự phân bố đồng đều.
- **Chức năng**:

- **value_counts()**: Đếm số lần xuất hiện của mỗi giá trị trong một cột. Thích hợp cho việc phân tích dữ liệu phân loại.
- **hist()**: Vẽ biểu đồ histogram để hiển thị phân phối của một biến số. Biểu đồ này giúp nhận diện sự phân bố của dữ liệu, ví dụ: có lệch trái, lệch phải hay phân phối đều.
- **boxplot()**: Vẽ biểu đồ hộp để phân tích sự phân bố và phát hiện các giá trị ngoại lệ. Biểu đồ này cung cấp cái nhìn rõ ràng về các phân vị và phạm vi của dữ liệu.

3. Nhóm hàm phân tích mối quan hệ giữa các biến

Hàm: corr(), sns.regplot(), sns.scatterplot()

- **Ý nghĩa**: Nhóm hàm này được sử dụng để phân tích và xác định mối quan hệ giữa các biến trong dữ liệu. Chúng giúp phát hiện các mối quan hệ tuyến tính hoặc phi tuyến tính, từ đó xây dựng các mô hình dự báo hoặc tìm hiểu sự tương quan giữa các yếu tố.
- **Chức năng**:
  - **corr()**: Tính toán ma trận tương quan giữa các biến số. Hệ số tương quan cho biết mức độ mạnh yếu của mối quan hệ giữa các biến.
  - **sns.regplot()**: Vẽ biểu đồ hồi quy tuyến tính giữa hai biến, kết hợp với đường hồi quy. Giúp phân tích mối quan hệ tuyến tính giữa các biến.
  - **sns.scatterplot()**: Vẽ biểu đồ phân tán (scatter plot) giữa hai biến để nhìn nhận trực quan mối quan hệ giữa chúng.

4. Nhóm hàm nhóm và phân tích dữ liệu theo các nhóm

Hàm: groupby(), pivot_table(), sns.barplot()

- **Ý nghĩa**: Nhóm hàm này cho phép bạn nhóm dữ liệu theo các đặc tính nhất định và áp dụng các phép toán tổng hợp để phân tích sự khác biệt giữa các nhóm. Chúng rất hữu ích trong việc so sánh và phân tích các nhóm con của dữ liệu.
- **Chức năng**:
  - **groupby()**: Nhóm dữ liệu theo một hoặc nhiều cột và tính toán các giá trị thống kê cho mỗi nhóm, như trung bình, tổng, v.v.
  - **pivot_table()**: Tạo bảng tổng hợp (pivot table) từ dữ liệu, cho phép tính toán các chỉ số tổng hợp như trung bình, tổng, v.v., theo các nhóm và phân loại.
  - **sns.barplot()**: Vẽ biểu đồ cột để so sánh giá trị trung bình của các nhóm dữ liệu, hỗ trợ so sánh giữa các nhóm một cách trực quan.

5. Nhóm hàm trực quan hóa dữ liệu

Hàm: sns.distplot(), sns.heatmap(), sns.pairplot()

- **Ý nghĩa**: Nhóm hàm này được sử dụng để trực quan hóa dữ liệu dưới dạng đồ họa, giúp bạn dễ dàng hiểu và phân tích các mối quan hệ và phân phối trong dữ liệu.
- **Chức năng**:
  - **sns.distplot()**: Vẽ biểu đồ phân phối kết hợp với đường mật độ, giúp bạn hiểu rõ hơn về phân phối của dữ liệu.
  - **sns.heatmap()**: Vẽ heatmap, một biểu đồ thể hiện mối quan hệ giữa các biến thông qua bảng màu. Thường được sử dụng để trực quan hóa ma trận tương quan hoặc các bảng số liệu phức tạp.
  - **sns.pairplot()**: Vẽ biểu đồ phân tán của từng cặp biến trong DataFrame, giúp bạn dễ dàng phát hiện các mối quan hệ giữa các biến trong dữ liệu.

6. Nhóm hàm phân tích phân phối và phát hiện giá trị ngoại lệ

Hàm: sns.boxplot(), sns.violinplot()

- **Ý nghĩa**: Nhóm hàm này được sử dụng để phân tích sự phân bố của dữ liệu và phát hiện các giá trị ngoại lệ (outliers). Các biểu đồ này giúp bạn nhận diện những điểm bất thường trong dữ liệu và phân tích các phân phối một cách trực quan.
- **Chức năng**:
  - **sns.boxplot()**: Vẽ biểu đồ hộp (box plot) để phân tích sự phân bố của dữ liệu và phát hiện các giá trị ngoại lệ. Biểu đồ này cho thấy các phân vị (quartile) và các giá trị ngoài phạm vi bình thường.
  - **sns.violinplot()**: Vẽ biểu đồ violin để hiển thị sự phân phối của dữ liệu. Nó kết hợp giữa biểu đồ hộp và biểu đồ mật độ, giúp cung cấp cái nhìn sâu sắc hơn về phân phối của dữ liệu.