**Requirements**

**Main goal:** We'll focus on the mindset of a conservative investor who only wants to invest in the loans that have a good chance of being paid off on time. To do that, we'll need to first clean the dataset so as to understand the features in the dataset and then experiment with building machine learning models (let’s just keep it simple) that reliably predict if a loan will be paid off or not.

**Mục tiêu chính:** Chúng ta sẽ tập trung vào tư duy của một nhà đầu tư thận trọng, người chỉ muốn đầu tư vào các khoản vay có cơ hội trả nợ đúng hạn. Để làm điều đó, trước tiên, chúng tôi cần làm sạch tập dữ liệu để hiểu các tính năng trong tập dữ liệu, sau đó thử nghiệm xây dựng các mô hình máy học (hãy đơn giản hóa nó) để dự đoán một cách đáng tin cậy liệu khoản vay có được trả hết hay không.

* **Dự đoán khả năng trả nợ đúng hạn của các khoản vay.**

Based on domain knowledge (hiểu biết ngành nghề) and the attention to data leakage problem, I suggest choosing criteria to prepare the features as follow:

Dựa trên kiến ​​thức về domain và sự quan tâm đến vấn đề rò rỉ dữ liệu, tôi khuyên bạn nên chọn các tiêu chí để chuẩn bị các tính năng như sau:

1. disclose information from the future (after the loan has already been funded)
2. don't affect a borrower's ability to pay back a loan (e.g. a randomly generated ID value)
3. need to be cleaned up and are formatted poorly
4. require more data or a lot of processing to turn into a useful feature
5. contain redundant information
6. tiết lộ thông tin từ tương lai (sau khi khoản vay đã được tài trợ): các đặc trưng được tạo ra chỉ dựa trên thông tin có sẵn tại thời điểm vay. Điều này giúp tránh tình trạng rỏ rỉ dữ liệu, trong đó thông tin từ tương lai được sử dụng trong quá trình tạo đặc trưng.
7. không ảnh hưởng đến khả năng trả lại khoản vay của người vay (ví dụ: giá trị ID được tạo ngẫu nhiên): đặc trưng không nên chứa thông tin có thể ảnh hưởng đến khả năng của người vay trả lại khoản vay, như 1 giá trị ID được tạo ngẫu nhiên. Điều này đảm bảo rằng các đặc trưng không chứa thông tin có thể dẫn đến kết quả mô hình không chính xác hoặc thiên vị.
8. cần phải được làm sạch và được định dạng kém: các đặc trưng cần được xử lý để loại bỏ dữ liệu không hợp lệ, thiết sót hoặc định dạng không chính xác. Điều này đảm bảo rằng các dữ liệu được sử dụng để tạo ra các đặc trưng là chính xác và đáng tin cậy.
9. yêu cầu nhiều dữ liệu hơn hoặc nhiều quá trình xử lý để biến thành một tính năng hữu ích: Các đặc trưng cần được cân nhắc xem liệu chúng có đáng đầu tư thời gian và công sức để thu thập them dữ liệu hoặc xử lý phức tạp để biến chúng thành đặc trưng hữu ích. Điều này giúp tối ưu hóa quá trình tạo đặc trưng và tránh các công việc không cần thiết.
10. chứa thông tin dư thừa: các đặc trưng không nên chứa thông tin trùng lặp, tức là thông tin đã được bao gồm trong các đặc trưng khác. Điều này giúp giảm kích thước dữ liệu và tránh sự lặp lại không cần thiết trong các đặc trưng.

Therefore, these are recommended columns that should be dropped.

Do đó, đây là những cột được khuyến nghị nên bỏ.

* id
* member\_id
* funded\_amnt
* funded\_amnt\_inv
* grade
* sub\_grade
* emp\_title
* issue\_d
* zip\_code
* out\_prncp
* out\_prncp\_inv
* total\_pymnt
* total\_pymnt\_inv
* total\_rec\_prncp
* total\_rec\_int
* total\_rec\_late\_fee
* recoveries
* collection\_recovery\_fee
* last\_pymnt\_d
* last\_pymnt\_amnt
* pub\_rec\_bankruptcies
* earliest\_cr\_line
* last\_credit\_pull\_d
* desc
* url

We should use loan\_status as target column.

Chúng ta nên sử dụng cột loan\_status như cột mục tiêu

Because of the main goal requirements, Let's remove all the loans that don't contain either Fully Paid or Charged Off as the loan's status.

Bởi vì yêu cầu của mục tiêu chính, hãy xóa bỏ tất cả các khoản vay mà không chứa trạng thái Fully Paid hoặc Charged Off trong cột loan’s status

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, Xanh điện

Mô tả được tạo tự động

1. Hiểu yêu cầu dự án: <https://towardsdatascience.com/predicting-loan-repayment-5df4e0023e92>
2. Thu thập dữ liệu: Xác định và thu thập dữ liệu liên quan đến khả năng trả nợ. Đảm bảo rằng dữ liệu được thu thập chính xác, đầy đủ và đại diện cho mẫu dữ liệu mà bạn muốn dự đoán. Có thể cần xử lý và làm sạch dữ liệu để chuẩn bị cho quá trình phân tích.
3. Khám phá dữ liệu: Tiến hành khám phá dữ liệu để hiểu các đặc trưng, mô hình quan hệ và tiềm năng của chúng. Sử dụng các công cụ và kỹ thuật phân tích dữ liệu để tìm hiểu mối quan hệ giữa các biến và tìm ra các thông tin hữu ích.

* Xử lý dữ liệu trùng lặp: Kiểm tra xem có dữ liệu trùng lặp trong DataFrame không. Nếu có, bạn có thể loại bỏ các hàng trùng lặp bằng cách sử dụng phương thức drop\_duplicates().
* Chuyển đổi kiểu dữ liệu: Kiểm tra kiểu dữ liệu của các cột trong DataFrame và đảm bảo chúng được chuyển đổi đúng. Ví dụ, các cột ngày tháng nên được chuyển đổi sang kiểu dữ liệu ngày tháng (datetime), các cột số nên được chuyển đổi sang kiểu dữ liệu số (float, int).
* Kiểm tra dữ liệu bị thiếu: Xem xét các cột trong DataFrame và kiểm tra xem có giá trị bị thiếu (NaN, null) không. Nếu có, bạn có thể xử lý bằng cách điền giá trị thiếu hoặc loại bỏ các hàng hoặc cột chứa giá trị thiếu, tùy thuộc vào quyết định của bạn và tính chất của dữ liệu.
* Xử lý outliers: Kiểm tra và xử lý các giá trị ngoại lệ (outliers) trong dữ liệu. Outliers có thể là các giá trị đáng ngờ hoặc không thể lý giải được trong ngữ cảnh dữ liệu. Bạn có thể quyết định giữ lại, điều chỉnh hoặc loại bỏ các outliers tùy thuộc vào mục đích và tính chất của dữ liệu.
* Chuẩn hóa dữ liệu: Nếu có sự khác biệt về đơn vị, tỷ lệ hoặc phạm vi giữa các cột dữ liệu, bạn có thể cân nhắc chuẩn hóa dữ liệu để đảm bảo tính nhất quán và dễ dàng so sánh.
* Thêm hoặc xóa cột: Dựa trên mục đích phân tích của bạn, bạn có thể thêm hoặc xóa các cột dữ liệu trong DataFrame. Đảm bảo các cột cung cấp thông tin hữu ích và có ý nghĩa trong phân tích của bạn.
* Xử lý dữ liệu categogical (phân loại): Nếu có các cột dữ liệu phân loại, bạn có thể xử lý chúng bằng cách chuyển đổi sang dạng số hoặc sử dụng mã hóa (encoding) như "one-hot encoding" để sử dụng trong các mô hình học máy.
* Kiểm tra và xử lý dữ liệu bất thường: Kiểm tra các giá trị bất thường hoặc không hợp lý trong dữ liệu và xử lý chúng một cách thích hợp. Điều này có thể bao gồm việc kiểm tra các giá trị ngoại lai, giá trị không hợp lý, hoặc xử lý các giá trị không phù hợp.

1. Xác định đặc trưng và mô hình: Dựa trên khám phá dữ liệu, xác định các đặc trưng quan trọng và phù hợp cho việc dự đoán khả năng trả nợ. Chọn mô hình máy học phù hợp để xây dựng dự đoán, ví dụ như mô hình hồi quy logistic hoặc cây quyết định.
2. Tiền xử lý dữ liệu: Thực hiện các bước tiền xử lý dữ liệu như xử lý dữ liệu thiếu, chuẩn hóa và mã hóa đặc trưng, và chia tách dữ liệu thành tập huấn luyện và tập kiểm tra.
3. Xây dựng và đánh giá mô hình: Sử dụng tập huấn luyện để huấn luyện mô hình dự đoán khả năng trả nợ. Đánh giá mô hình bằng cách sử dụng các phép đo đánh giá phù hợp như độ chính xác, độ phủ, hay F1-score. Điều chỉnh và tối ưu hóa mô hình để đạt được kết quả tốt nhất có thể.
4. Triển khai và kiểm tra: Triển khai mô hình vào môi trường thực tế và kiểm tra hiệu suất của nó trên tập kiểm tra độc lập. Đảm bảo rằng mô hình hoạt động một cách đáng tin cậy và đáp ứng yêu cầu dự án.
5. Theo dõi và tinh chỉnh: Theo dõi hiệu suất của mô hình theo thời gian và tinh chỉnh nếu cần để đảm bảo rằng nó vẫn đáp ứng yêu cầu và mang lại kết quả tốt.