## **Mini Project DT2210L**

**Credit EDA**

**1. Giới thiệu**

* Phân tích rủi ro trong các dịch vụ tài chính ngân hàng, đồng thời hiểu cách sử dụng dữ liệu để giảm thiểu rủi ro mất tiền khi cho khách hàng vay. .

**2. Business Understanding**

* Các công ty cung cấp khoản vay gặp khó khăn trong việc cho người dân vay do lịch sử tín dụng không đầy đủ hoặc không tồn tại. Do đó, một số người tiêu dùng lợi dụng nó để làm lợi thế cho mình bằng cách trở thành người vỡ nợ. Giả sử bạn làm việc cho một công ty tài chính tiêu dùng chuyên cho vay các loại hình cho khách hàng thành thị. Bạn phải sử dụng EDA để phân tích các mẫu có trong dữ liệu. Điều này sẽ đảm bảo rằng những người nộp đơn có khả năng trả khoản vay không bị từ chối.
* Khi công ty nhận được đơn xin vay vốn, công ty phải quyết định phê duyệt khoản vay dựa trên hồ sơ của người nộp đơn. Hai loại rủi ro gắn liền với quyết định của ngân hàng:
* Nếu người nộp đơn có khả năng hoàn trả khoản vay thì việc không phê duyệt khoản vay sẽ dẫn đến việc công ty thua lỗ trong hoạt động kinh doanh
* Nếu người nộp đơn không có khả năng hoàn trả khoản vay, tức là họ có khả năng vỡ nợ, thì việc phê duyệt khoản vay có thể dẫn đến tổn thất tài chính cho công ty.
* Dữ liệu được cung cấp dưới đây chứa thông tin về đơn xin vay tại thời điểm đăng ký khoản vay. Nó chứa **hai loại kịch bản**:
* **Khách hàng gặp khó khăn trong việc thanh toán:** họ đã chậm thanh toán hơn **X ngày** đối với ít nhất một trong các đợt vay **Y đầu tiên** trong mẫu của chúng tôi,
* **Tất cả các trường hợp khác:** Tất cả các trường hợp khác khi thanh toán được thanh toán đúng hạn.
* Khi khách hàng đăng ký vay, khách hàng/công ty có thể đưa ra **bốn loại quyết định** sau:
* **Phê duyệt:** Công ty đã phê duyệt hồ sơ vay vốn
* **Đã hủy:** Khách hàng đã hủy đơn đăng ký trong thời gian phê duyệt. Hoặc khách hàng đã thay đổi quyết định về khoản vay hoặc trong một số trường hợp do khách hàng gặp rủi ro cao hơn nên anh ta nhận được mức giá tồi tệ hơn mà anh ta không mong muốn.
* **Từ chối:** Công ty đã từ chối khoản vay (vì khách hàng không đáp ứng được yêu cầu của họ, v.v.).
* **Ưu đãi chưa sử dụng:**  Khoản vay đã bị khách hàng hủy nhưng ở các giai đoạn khác nhau của quy trình.
* Trong trường hợp nghiên cứu này, bạn sẽ sử dụng EDA để hiểu các thuộc tính của người tiêu dùng và khoản vay ảnh hưởng như thế nào đến xu hướng vỡ nợ.

## **3. Business Objectives**

Nghiên cứu điển hình này nhằm mục đích xác định các mô hình cho biết liệu khách hàng có gặp khó khăn trong việc trả góp hay không, điều này có thể được sử dụng để thực hiện các hành động như từ chối khoản vay, giảm số tiền cho vay, cho vay (đối với những người nộp đơn rủi ro) với lãi suất cao hơn, v.v. Điều này sẽ đảm bảo rằng những người tiêu dùng có khả năng trả nợ không bị từ chối. Mục đích của nghiên cứu trường hợp này là xác định những người nộp đơn như vậy bằng cách sử dụng EDA.

**Nói cách khác, công ty muốn hiểu các yếu tố thúc đẩy (hoặc các biến số điều khiển) đằng sau việc vỡ nợ, tức là các biến số là chỉ báo mạnh về khả năng vỡ nợ. Công ty có thể sử dụng kiến ​​thức này cho danh mục đầu tư và đánh giá rủi ro của mình.**

**Để phát triển sự hiểu biết của bạn về lĩnh vực này, bạn nên nghiên cứu độc lập một chút về phân tích rủi ro - hiểu rõ các loại biến số và tầm quan trọng của chúng là đủ.**

**4. Data Understanding**

**Tập dữ liệu này có 3 file như được giải thích bên dưới:**

1. 'application\_data.csv' chứa tất cả thông tin của khách hàng tại thời điểm đăng ký. Dữ liệu cho biết liệu khách hàng có gặp khó khăn trong thanh toán hay không.

2. 'previous\_application.csv' chứa thông tin về dữ liệu khoản vay trước đây của khách hàng. Nó chứa dữ liệu về việc đơn đăng ký trước đó đã được phê duyệt, hủy bỏ, từ chối hay ưu đãi không được sử dụng.

3. 'columns\_description.csv' là từ điển dữ liệu mô tả ý nghĩa của các biến.

**Results Expected by Learners**

**1. Trình bày cách tiếp cận tổng thể của phân tích trong một bài thuyết trình. Đề cập ngắn gọn đến vấn đề và cách tiếp cận phân tích.**

**2. Xác định dữ liệu còn thiếu và sử dụng phương pháp thích hợp để xử lý nó. (Xóa cột/hoặc thay thế bằng giá trị phù hợp)**

**3. Xác định xem có sự mất cân bằng dữ liệu trong dữ liệu hay không. Tìm tỷ lệ mất cân bằng dữ liệu.**

**4. Giải thích kết quả phân tích đơn biến, phân đoạn đơn biến, hai biến, v.v. bằng thuật ngữ kinh doanh.**

Bạn cần gửi:

* **Ipython notebook** trong đó giải thích rõ ràng quá trình suy nghĩ đằng sau phân tích của bạn **(in comments of markdown text)**, code và các biểu đồ có liên quan.
* File trình bày cần phải ở định dạng **word** và phải chứa các điểm được thảo luận ở trên cùng với các hình ảnh trực quan cần thiết. Ngoài ra, tất cả các hình ảnh trực quan và sơ đồ phải được thực hiện bằng Python **(Ipython notebook)**, có thể sử dụng hình vẽ bằng **power bi.**
* **File báo cáo powerpoint.**
* Bảng phân công công việc

# **Evaluation Rubrics**

| **Criteria** | **Meets expectations** | **Does not meet expectations** |
| --- | --- | --- |
| **Data understanding (20%)** | All data quality issues are correctly identified and reported.    Wherever required, the meanings of the variables are correctly interpreted and written either in the comments or text. | Data quality issues are overlooked or are not identified correctly such as missing values, outliers and other data quality issues.    The variables are interpreted incorrectly or the meaning of variables is not mentioned. |
| **Data Cleaning and Manipulation (10%)** | Data quality issues are addressed in the right way (missing value imputation analysis and other kinds of data redundancies, etc.).    If applicable, data is converted to a suitable and convenient format to work with using the right methods.      Manipulation of strings and dates is done correctly wherever required | Data quality issues are not addressed correctly.          The variables are not converted to an appropriate format for analysis.        String and date manipulation is not done correctly or is done using complex methods |
| **Data analysis (50%)** | The right problem is solved which is coherent with the needs of the business. The analysis has a clear structure and the flow is easy to understand.      **Univariate** and segmented univariate analysis is done correctly and appropriate realistic assumptions are made wherever required. The analyses successfully identify at least the 5 important driver variables (i.e. variables which are strong indicators of default).    Business-driven, type-driven and data-driven metrics are created for the important variables and utilised for analysis. The explanation for creating the derived metrics is mentioned and is reasonable.    **Bivariate analysis** is performed correctly and is able to identify the important combinations of driver variables. The combinations of variables are chosen such that they make business or analytical sense.    The most useful insights are explained correctly in the comments.      Appropriate plots are created to present the results of the analysis. The choice of plots for respective cases is correct. The plots should clearly present the relevant insights and should be easy to read. The axes and important data points are labelled correctly. | The analyses do not address the right problem or deviate from the business objectives. The analysis lacks a clear structure and is not easy to follow.        The univariate and bivariate analysis is not performed in sufficient detail and thus some crucial insights are missed out. The analyses are not able to identify enough important driver variables.    New metrics are not derived wherever appropriate. The explanation for creating the derived metrics is either not mentioned or the metrics are not reasonable.    Derived metrics are not analysed correctly/are insufficiently utilised.    .        Important insights are not mentioned in the report or the Python file. Relevant plots are not created. The choice of plots is not ideal and the plots are either difficult to interpret or lack clarity or neatness. Relevant insights are not clearly presented by the plots. The axes and important data points are not labelled correctly/neatly. |
| **Presentation and Recommendations (10%)** | The presentation has a clear structure, is not too long, and explains the most important results concisely in simple language.    The recommendations to solve the problems are realistic, actionable and coherent with the analysis.    If any assumptions are made, they are stated clearly. | The presentation lacks structure, is too long or does not put emphasis on the important observations. The language used is complicated for business people to understand.    The recommendations to solve the problems are either unrealistic, non-actionable or incoherent with the analysis.      Contains unnecessary details or lacks the important ones.    Assumptions made, if any, are not stated clearly. |
| **Conciseness and readability of the code (10%)** | The code is concise and syntactically correct. Wherever appropriate, built-in functions and standard libraries are used **instead of writing long code (if-else statements, for loops, etc.)**.    Custom functions are used to perform repetitive tasks.    The code is readable with appropriately named variables and detailed comments are written wherever necessary. | Long and complex code used instead of shorter built-in functions.    Custom functions are not used to perform repetitive tasks resulting in the same piece of code being repeated multiple times.    Code readability is poor because of vaguely named variables or lack of comments wherever necessary. |