**BÁO CÁO CHẤM ĐIỂM TÍN DỤNG CỦA KHÁCH HÀNG**

# GIỚI THIỆU

Ngày nay, các ngân hàng thương mại giữ một khối lượng tín dụng tiêu dùng trị giá vô cùng lớn, tạo nên một ngành kinh tế khổng lồ và vô cùng quan trọng. Số lượng lớn các khoản vay cho thấy rằng ngay cả những cải thiện nhỏ trong độ chính xác của việc đánh giá tín dụng cũng có thể mang lại lợi ích tài chính đáng kể [1]. Vì vậy, việc phát triển các mô hình đánh giá tín dụng chính xác đã trở thành ưu tiên hàng đầu của các tổ chức tài chính, nhằm tối ưu hóa lợi nhuận và quản lý rủi ro một cách hiệu quả. Xếp hạng tín dụng là một phương pháp được sử dụng để đánh giá độ tin cậy của người cho vay tiềm năng, có thể là một cá nhân hoặc một công ty. Mục đích của mô hình chấm điểm tín dụng là giải quyết bài toán phân loại khách hàng vay vốn thành hai loại: khách hàng tốt (những khách hàng được dự đoán sẽ được thanh toán đầy đủ trong khoảng thời gian quy định) và khách hàng xấu (những khách hàng được dự đoán sẽ vỡ nợ). Những khách hàng tốt có nhiều khả năng trả nợ hơn cho vay đúng hạn sẽ mang lại lợi ích cho tổ chức tài chính. Ngược lại, khách hàng xấu sẽ dẫn tới tổn thất tài chính. Vì vậy, các ngân hàng và tổ chức tài chính ngày càng quan tâm đến việc xây dựng các mô hình chấm điểm tín dụng, bởi vì thậm chí chất lượng tín dụng xấu chỉ tăng 1% người nộp đơn sẽ chuyển thành khoản tiết kiệm đáng kể trong tương lai cho các tổ chức tài chính [3,4]. Tuy nhiên, một số vấn đề dần xuất hiện cùng với sự phát triển nhanh chóng của ngành tài chính Internet. Một mặt, số lượng nền tảng cho vay trực tuyến P2P (điểm - điểm) có vấn đề chiếm tỷ lệ đáng kể trong tất cả các nền tảng cho vay trực tuyến. Mặt khác, tỷ lệ nợ xấu trong tài chính Internet cao hơn so với các ngân hàng truyền thống. Vì vậy, việc ngăn ngừa và kiểm soát rủi ro tín dụng đã trở thành một bài toán quan trọng đối với ngành tài chính Internet và việc chấm điểm tín dụng cung cấp một công cụ hữu hiệu để giải quyết vấn đề này [7].

# NGHIÊN CỨU LIÊN QUAN

Nghiên cứu dự đoán Điểm Tín dụng bằng Kỹ thuật Thuật toán Di truyền-LSTM, nghiên cứu này trình bày phương pháp tối ưu hóa để xác định các tham số tối ưu cho thuật toán deep learning. Các tham số LSTM được xác định bằng thuật toán tối ưu hóa. Các tham số LSTM bao gồm kỷ nguyên, kích thước lô, số lượng nơ-ron, tốc độ học tập và tỷ lệ bỏ học. Kết quả cho thấy mô hình LSTM được tối ưu hóa vượt trội hơn cả mô hình phân loại đơn và mô hình tập hợp [25].

Một nghiên cứu gần đây năm 2021 đã đặc biệt tập trung vào việc phân tích ứng dụng chấm điểm tín dụng của các tổ chức tài chính dựa trên mô hình học máy [30] Với sự phát triển của khoa học công nghệ, công nghệ dữ liệu lớn đã thâm nhập vào lĩnh vực tài chính, điều tra tín dụng cá nhân đã bước vào một kỷ nguyên mới. Đánh giá tín dụng cá nhân dựa trên dữ liệu lớn là một trong những chủ đề nghiên cứu nóng. Bài viết này chủ yếu hoàn thành ba công trình. Thứ nhất, theo kịch bản ứng dụng đánh giá tín dụng của dữ liệu tín dụng cá nhân, tập dữ liệu thử nghiệm được làm sạch, dữ liệu rời rạc được mã hóa một HOT và dữ liệu được chuẩn hóa. Do dữ liệu tín dụng cá nhân có kích thước lớn, thuật toán pdC-RF được sử dụng trong bài viết này để tối ưu hóa mối tương quan của các tính năng dữ liệu và giảm dữ liệu 145 chiều thành dữ liệu 22 chiều. Trên cơ sở đó, mã hóa WOE đã được thực hiện trên tập dữ liệu, được áp dụng cho rừng ngẫu nhiên, máy vectơ hỗ trợ và mô hình hồi quy logistic, đồng thời hiệu suất được so sánh. Người ta nhận thấy rằng hồi quy logistic phù hợp hơn với mô hình đánh giá tín dụng cá nhân dựa trên bộ dữ liệu của Lending Club. Cuối cùng, dựa trên mô hình hồi quy logistic với các tham số tốt nhất, mẫu người dùng sẽ được phân loại và thẻ điểm cuối cùng sẽ được đưa ra.

Trong nghiên cứu số [28], tác giả của bài viết đã xây dựng mô hình chấm điểm tín dụng dựa trên việc học sâu và lựa chọn tính năng để đánh giá điểm tín dụng của người nộp đơn từ các đặc điểm đầu vào của người nộp đơn. Hai bộ dữ liệu công khai, dữ liệu tín dụng của Úc và Đức, đã được sử dụng để kiểm tra phương pháp. Kết quả thử nghiệm dữ liệu trong thế giới thực cho thấy phương pháp được đề xuất mang lại tỷ lệ dự đoán cao hơn phương pháp cơ sở đối với một số bộ dữ liệu nhất định và cũng cho thấy hiệu suất tương đương và đôi khi tốt hơn so với các phương pháp lựa chọn tính năng được sử dụng rộng rãi trong chấm điểm tín dụng.

Ngoài ra, một bài nghiên cứu được viết vào năm 2022 là so sánh trực tiếp GBM và DL về độ chính xác dự đoán để phân loại chính xác danh mục mặc định của khách hàng. Ba tập dữ liệu riêng biệt với các tính năng khác nhau đã được sử dụng để giải thích khả năng lựa chọn/ sức mạnh của mô hình dựa trên các đặc điểm của tập dữ liệu cơ bản. Khi làm như vậy, nghiên cứu này làm sáng tỏ khả năng dự đoán và tính hữu ích của cả hai mô hình quản lý rủi ro tín dụng trong thị trường cho vay – đặc biệt là chấm điểm tín dụng [5].

Một nghiên cứu được công bố vào năm 2023 với mục đích là đề xuất một mô hình tín dụng tổng thể không đồng nhất, năng động, có khả năng thích ứng, tích hợp các mô hình XGBoost và Máy Vector hỗ trợ để cải thiện độ chính xác và độ tin cậy của các mô hình chấm điểm tín dụng đánh giá rủi ro. Phương pháp này sử dụng các kỹ thuật học máy để nhận biết các mẫu và xu hướng từ dữ liệu trong quá khứ nhằm dự đoán các sự kiện xảy ra trong tương lai [14].

# PHƯƠNG PHÁP

## 3.1. Các kỹ thuật nền tảng

Trong phần này, nhóm chúng tôi sẽ trình bày và giải thích các kỹ thuật cơ bản để xây dựng một hệ thống chấm điểm tín dụng thường sử dụng. Các kỹ thuật này là các công cụ và phương pháp chính để xây dựng mô hình dự đoán rủi ro tín dụng và đánh giá khả năng trả nợ của khách hàng. Dưới đây là một số kỹ thuật quan trọng:

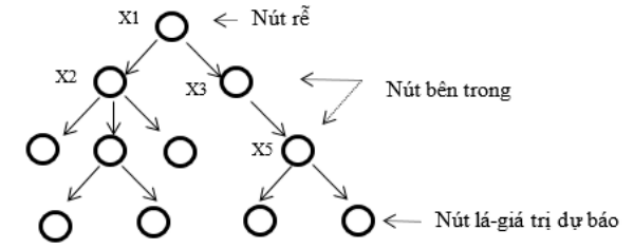
* *Phân tích và xử lý dữ liệu:* Đây là bước quan trọng nhất trong quá trình xây dựng mô hình chấm điểm tín dụng. Phân tích dữ liệu giúp hiểu rõ hơn về các đặc điểm của khách hàng và mối quan hệ giữa các yếu tố dữ liệu và khả năng trả nợ của họ. Xử lý dữ liệu bao gồm việc xử lý các giá trị bị thiếu, chọn lọc đặc trưng và chuẩn hóa dữ liệu bằng Min-Max Scaling (chuẩn hóa Min-Max) hay Standard Scaler .
* *Mô hình hóa dữ liệu:* Các kỹ thuật này bao gồm việc lựa chọn và xây dựng các mô hình phân loại và học máy để dự đoán rủi ro tín dụng. Các mô hình phổ biến bao gồm Logistic Regression, Decision Trees, Random Forests, Adaboost, và Neural Networks.
* Đánh giá mô hình: Kỹ thuật này liên quan đến việc đánh giá hiệu suất của các mô hình dự đoán. Các phương pháp đánh giá bao gồm Confusion Matrix, ROC Curve, Precision - Recall Curve và các độ đo như Accuracy, Precision, Recall và F1-score.
* Xử lý mất cân bằng dữ liệu: Trong các vấn đề chấm điểm tín dụng, thường có sự mất cân bằng giữa các nhóm dữ liệu, ví dụ như số lượng khách hàng trả nợ và không trả nợ. Các kỹ thuật như Oversampling, Undersampling, và SMOTE được sử dụng để giải quyết vấn đề này.
* Tinh chỉnh mô hình: Sau khi xây dựng mô hình, quá trình tinh chỉnh là cần thiết để điều chỉnh các tham số mô hình và tối ưu hóa hiệu suất của nó. Các kỹ thuật như Cross-Validation và Grid Search được sử dụng để tinh chỉnh mô hình.

Các kỹ thuật nền tảng này cùng với các công cụ và phương pháp khác tạo ra cơ sở cho việc phát triển các mô hình chấm điểm tín dụng hiệu quả và chính xác. Trong phần tiếp theo, chúng ta sẽ đề xuất phương pháp cụ thể để áp dụng các kỹ thuật này vào việc giải quyết vấn đề chấm điểm tín dụng khách hàng.

## 3.2. Đề xuất phương pháp

Vấn đề lớn đối với các tổ chức tài chính và ngân hàng là xác định tín dụng xấu, vì nó có thể gây ra các vấn đề nghiêm trọng trong tương lai như thất thoát vốn, giảm doanh thu và tăng tổn thất cho ngân hàng. Hiện nay, có rất nhiều kĩ thuật học máy để chấm điểm tín dụng của khách hàng. Nhóm chúng tôi đã đề xuất sử dụng một tổ hợp các mô hình học máy truyền thống và mô hình học sâu để chấm điểm tín dụng của khách hàng. Cụ thể, chúng tôi đề xuất sử dụng 7 mô hình sau đây:

* Decision Tree (Cây quyết định): Cây quyết định (Decision Tree - DT) là một thuật toán học máy được ứng dụng rộng rãi trong phân loại và dự đoán, bao gồm cả lĩnh vực đánh giá rủi ro tín dụng. Mô hình này hoạt động dựa trên hệ thống các quy tắc logic được biểu diễn dưới dạng cây, giúp phân loại khách hàng và dự đoán khả năng thanh toán khoản vay của họ. Mô hình này bao gồm một nút gốc (Root node), các nút bên trong (Internal node) và nút lá (Leaf node). Cấu trúc của cây quyết định gồm nút gốc - đại diện cho điểm khởi đầu của quá trình phân loại, nút bên trong - thể hiện các thuộc tính (biến) được sử dụng để phân chia dữ liệu và nút lá - đại diện cho kết quả dự đoán (ví dụ: mức độ rủi ro tín dụng) (Hình 1).



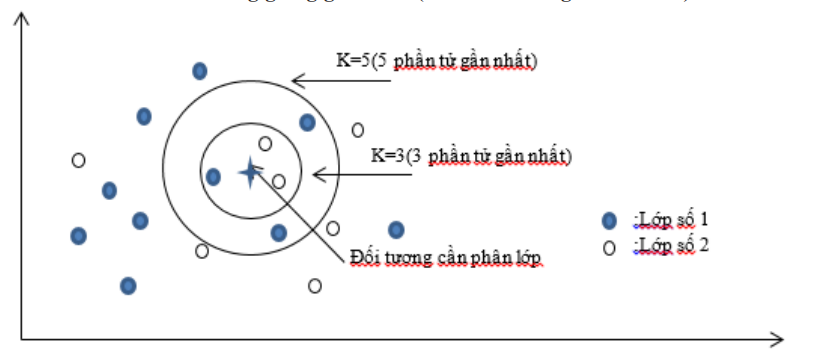
**Hình 1.** Cây quyết định (Decision tree)

*Nguồn: Abdou. (2011)*

* K-Nearest Neighbor (K-Láng giềng gần nhất): KNN là một mô hình dựa trên sự gần gũi của điểm dữ liệu, dựa vào các điểm láng giềng gần nhất để dự đoán nhãn của một điểm dữ liệu mới.

Từ Hình 2 có thể thấy, nếu chọn số điểm gần nhất để phân loại là 3 thì xung quanh điểm cần phân lớp có 3 điểm trong đó 1 điểm thuộc lớp 1 và 2 điểm thuộc lớp 2. Với số điểm thuộc lớp 2 nhiều hơn, điểm cần phân lớp được đưa vào lớp số 1. Tương tự như vậy, nếu lựa chọn số điểm gần nhất k=5 thì có năm điểm xung quanh điểm cần phân lớp, trong đó có 3 điểm thuộc lớp số 1 và 2 điểm thuộc lớp 2. Nếu theo dữ liệu này thì điểm cần phân lớp sẽ thuộc lớp số 1. Trên thực tế dữ liệu cần phân loại có nhiều thuộc tính trong đó mỗi thuộc tính tương ứng với một chiều không gian, do vậy khi tính khoảng cách gần nhất cần tính khoảng cách vector trong không gian đa chiều với công thức khoảng cách Euclidean:

trong đó: x, y là 2 điểm cần xác định khoảng cách; n là số chiều không gian (số thuộc tính của biến mục tiêu).



**Hình 2.** K láng giềng gần nhất (K-Nearest Neighbors - KNN)

*Nguồn: Marinakis và cộng sự (2008)*

* Random Forest (Rừng ngẫu nhiên): RF là một mô hình ensemble của nhiều cây quyết định, có khả năng tạo ra các dự đoán mạnh mẽ và ổn định bằng cách kết hợp kết quả từ nhiều cây quyết định. Random Forest có thể được sử dụng để dự đoán khả năng thanh toán của khách hàng dựa trên các đặc trưng tài chính và hành vi thanh toán. Điều này là do RF thích hợp với việc xử lý dữ liệu có cấu trúc và có khả năng làm việc với các đặc trưng có mức độ quan trọng khác nhau mà không cần phải tiền xử lý đặc biệt.
* XGBoost (Extreme Gradient Boosting): XGBoost là một mô hình ensemble learning, cải tiến từ Gradient Boosting, giúp cải thiện hiệu suất dự đoán bằng cách tối ưu hóa hàm mất mát và xây dựng các cây quyết định tuần tự.
* Naive Bayes: Naive Bayes là một mô hình đơn giản dựa trên xác suất, dựa trên giả định về sự độc lập giữa các biến, nhưng thường cho kết quả tốt trong nhiều bài toán phân loại.
* Convolutional Neural Network (Mạng nơ-ron tích chập sâu): CNN là một loại mạng nơ-ron nhân tạo đặc biệt phù hợp để phân tích dữ liệu dạng lưới, phổ biến nhất là hình ảnh. Nó có khả năng học một lượng lớn các dữ liệu trong khoảng thời gian ngắn hơn nhiều so với mạng nơ ron thông thường. Lý do là nó sử dụng ít trọng số hơn trong khi độ chính xác chỉ kém hơn một phần nhỏ so với kiến trúc truyền thống.
* Recurrent Neural Network (Mạng nơ-ron hồi quy): RNN là một mô hình mạng neural thích hợp cho việc xử lý dữ liệu chuỗi như văn bản hoặc chuỗi thời gian, có khả năng nhớ thông tin từ quá khứ để đưa ra dự đoán tín dụng cho tương lai.

Sử dụng một tổ hợp của các mô hình này giúp tăng cường khả năng dự đoán và đảm bảo tính đa dạng trong quá trình đánh giá tín dụng của khách hàng. Các mô hình truyền thống cung cấp một cơ sở ổn định và dễ hiểu, trong khi các mô hình học sâu có khả năng học được các biểu diễn phức tạp từ dữ liệu, giúp cải thiện hiệu suất dự đoán trong các trường hợp phức tạp.

Phương pháp được thực hiện theo tiến trình:

**Bước 1:** Khai thác dữ liệu

* Quan sát tổng quan dữ liệu (Hệ tương quan, giá trị thống kê, vấn đề mất cân bằng dữ liệu ở biến mục tiêu).

**Bước 2:** Chuẩn bị dữ liệu

* Tiến hành tiền xử lý dữ liệu (xóa trùng lặp, kiểm tra dữ liệu trống), biến đổi các trường dữ liệu định danh với phương pháp mã hóa nhãn (LabelEncoder).
* Xóa các đặc trưng không quan trọng
* Sử dụng phương pháp IQR để loại bỏ các giá trị ngoại lệ
* Sử dụng phương pháp Min-Max Scaling để chuẩn hóa

**Bước 3:** Xây dựng tập dữ liệu huấn luyện và kiểm tra

Bộ dữ liệu được chia thành 2 tập gồm tập huấn luyện 80% và tập dữ liệu kiểm tra 20%.

**Bước 4:** Xây dựng và huấn luyện các mô hình

* Với tập dữ liệu huấn luyện tiến hành xây dựng các mô hình với phương pháp Grid search CV để tìm ra các siêu tham số.

**Bước 5:** Mô hình thử nghiệm với tập dữ liệu kiểm tra đưa ra đánh giá mô hình.

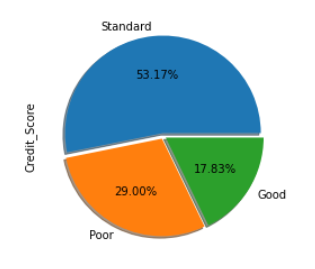
**Bước 6:** So sánh các mô hình dựa trên hiệu suất các chỉ số và đưa ra mô hình tốt nhất

**IV. THỰC NGHIỆM**

## 4.1. Miêu tả dữ liệu

Để tuân thủ các quy định pháp luật hiện hành về bảo mật thông tin cá nhân, các tổ chức tài chính phải tuân thủ nguyên tắc không tiết lộ thông tin chi tiết về tài khoản và các giao dịch của khách hàng. Trong phạm vi nghiên cứu này, chúng tôi tiến hành thực nghiệm trên một tập dữ liệu được công bố công khai có tên là “Credit score classification”, được truy cập và tải về từ trang Kaggle.com. Bộ dữ liệu này được thiết kế để mô phỏng các loại dữ liệu giao dịch tài chính và không liên quan trực tiếp đến bất kỳ ngân hàng cụ thể nào. Tập dữ liệu bao gồm 28 đặc trưng với 100,000 quan sát và chi tiết được mô tả như trong Bảng 1.

Trường dữ liệu “Credit Score”: được sử dụng để đánh giá khả năng thanh toán của khách hàng và phản ánh lịch sử tín dụng của họ. Trường dữ liệu này có ba nhóm giá trị: “Good” (tốt), “Standard” (trung bình), và “Poor” (kém). Nhãn “Good” được gán cho những khách hàng có điểm tín dụng cao, thường là do họ có lịch sử thanh toán nợ đúng hạn và duy trì một hồ sơ tín dụng tích cực, nhãn này có 17828 quan sát chiếm 17.828%. Nhãn “Standard” thường ám chỉ đến những khách hàng có lịch sử tín dụng trung bình, có thể có một số vấn đề hoặc trễ hẹn trong quá khứ, nhãn này có 53174 quan sát chiếm 53.174%. Trong khi đó, nhãn “Poor” được gán cho những khách hàng có điểm tín dụng thấp, thường xuyên trễ hẹn hoặc không thực hiện thanh toán đúng hạn, nhãn này có 28998 quan sát chiếm 28.998%. Thông qua việc phân loại khách hàng thành các nhóm dựa trên biến "Credit\_Score", các tổ chức tài chính có thể chấm điểm tín dụng khách hàng dựa trên lịch sử tín dụng và các yếu tố liên quan.



**Hình 3.** Tỷ lệ các quan sát trong trường dữ liệu “Credit\_Score”

**Bảng 1:** Mô tả dữ liệu thực nghiệm

|  |  |  |
| --- | --- | --- |
| **Trường dữ liệu** | **Mô tả** | **Kiểu dữ liệu** |
| **ID** | Định danh duy nhất của mỗi quan sát | object |
| **Customer\_ID** | Định danh của khách hàng | object |
| **Month** | Tháng | object |
| **Name** | Tên khách hàng | object |
| **Age** | Tuổi khách hàng | object |
| **SSN** | Số An sinh Xã hội | object |
| **Occupation** | Nghề nghiệp của khách hàng | object |
| **Annual\_Income** | Thu nhập hàng năm của khách hàng | object |
| **Monthly\_Inhand\_Salary** | Lương thực nhận hàng tháng | float64 |
| **Num\_Bank\_Accounts** | Số tài khoản ngân hàng | int64 |
| **Num\_Credit\_Card** | Số thẻ tín dụng | int64 |
| **Interest\_Rate** | Lãi suất | int64 |
| **Num\_of\_Loan** | Số khoản vay | object |
| **Type\_of\_Loan** | Loại khoản vay | object |
| **Delay\_from\_due\_date** | Số ngày trễ so với ngày đáo hạn | int64 |
| **Num\_of\_Delayed\_Payment** | Số lượng thanh toán bị trễ | object |
| **Changed\_Credit\_Limit** | Số lượng thay đổi hạn mức tín dụng | object |
| **Num\_Credit\_Inquiries** | Số lượng điều tra tín dụng | float64 |
| **Credit\_Mix** | Tổ hợp tín dụng | object |
| **Outstanding\_Debt** | Nợ chưa thanh toán | object |
| **Credit\_Utilization\_Ratio** | Tỷ lệ sử dụng tín dụng | float64 |
| **Credit\_History\_Age** | Lịch sử tín dụng | object |
| **Payment\_of\_Min\_Amount** | Số tiền thanh toán tối thiểu | object |
| **Total\_EMI\_per\_month** | Tổng EMI hàng tháng | float64 |
| **Amount\_invested\_monthly** | Số tiền đầu tư hàng tháng | object |
| **Payment\_Behaviour** | Hành vi thanh toán | object |
| **Monthly\_Balance** | Số dư hàng tháng | object |
| **Credit\_Score** | Điểm tín dụng | object |

## 4.2. Tiền xử lý dữ liệu

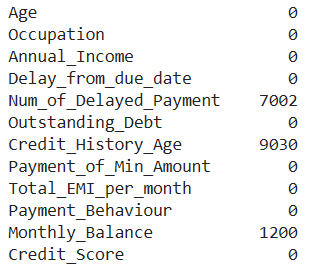
Việc tiền xử lý dữ liệu là một bước quan trọng trong quá trình chuẩn bị dữ liệu trước khi đưa vào quá trình huấn luyện mô hình học máy. Mục tiêu của tiền xử lý dữ liệu là tối ưu hóa dữ liệu để tăng cường hiệu suất của mô hình và cải thiện độ chính xác của mô hình theo mô hình học máy. Chúng tôi đã thựuc hiện tiền xử lý dữ liệu với các bước như sau:

**Bước 1:** *Loại bỏ dữ liệu không cần thiết*

Loại bỏ các cột dữ liệu không sử dụng trong mô hình học máy, 16 trường dữ liệu này bị loại bỏ đi bao gồm là 'ID', 'Customer\_ID', 'Month', 'Name', 'SSN', 'Monthly\_Inhand\_Salary', 'Num\_Bank\_Accounts', 'Num\_Credit\_Card', 'Interest\_Rate', 'Num\_of\_Loan', 'Type\_of\_Loan', 'Changed\_Credit\_Limit', 'Num\_Credit\_Inquiries', 'Credit\_Mix', 'Credit\_Utilization\_Ratio', 'Amount\_invested\_monthly'

**Bước 2:** *Loại bỏ các giá trị NaN*

Xóa các hàng chứa giá trị NaN, cụ thể số lượng hàng bị xóa do các trường dữ liệu có chứa giá trị NaN được thể hiện trong hình 1.



**Hình 4.** Số lượng giá trị lỗi NaN trong các trường dữ liệu

**Bước 3:** *Loại bỏ các hàng chứa giá trị không chính xác và thay thế các dữ liệu sai*

Xóa các hàng có dữ liệu không chính xác trong cột "Occupation" và "Payment\_Behaviour", thay thế ký tự đặc biệt bằng chuỗi rỗng.

**Bước 4:** *Mã hóa và chuẩn hóa dữ liệu*

* Chuyển đổi dữ liệu cột "Credit\_History\_Age" từ chuỗi thành số, cụ thể thay thế chuỗi con ' Years and ' (bao gồm khoảng trắng) bằng dấu chấm ('.') và chuỗi con 'Months' (bao gồm khoảng trắng) bằng chuỗi rỗng (''). Ví dụ: '2 Years and 3 Months' thành '2.3'. Mục đích là chuyển đổi định dạng từ năm và tháng sang dạng thập phân để tính toán dễ dàng hơn.
* Chuyển đổi các nhãn trong cột “Payment\_Behaviour” thành các giá trị theo cấp độ. Sủ dụng hàm ‘.str.replace(...)’ để thay thế gắn nhẵn “Low\_spent\_Small\_value\_payments” thành ‘1’,

“Low\_spent\_Medium\_value\_payments” thành ‘2’,

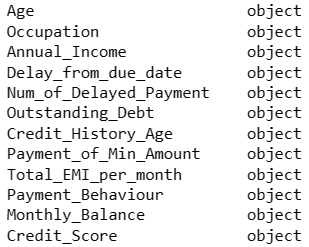
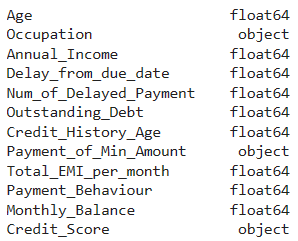
“Low\_spent\_Large\_value\_payments” thành ‘3’,

“High\_spent\_Small\_value\_payments” thành ‘4’,

“High\_spent\_Medium\_value\_payments” thành ‘5’,

“ High\_spent\_Large\_value\_payments” thành ‘6’.

* Chuyển đổi kiểu dữ liệu của các cột 'Age', 'Delay\_from\_due\_date', 'Num\_of\_Delayed\_Payment', 'Outstanding\_Debt', 'Total\_EMI\_per\_month', 'Monthly\_Balance', 'Payment\_Behaviour', 'Credit\_History\_Age', 'Annual\_Income' sang số thập phân

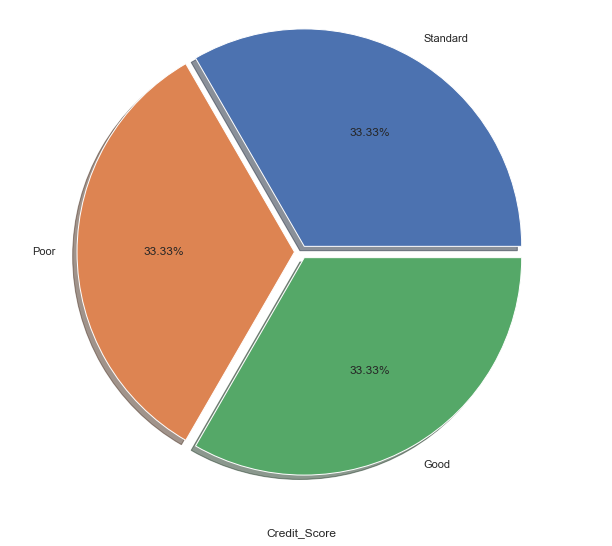
*Trước khi xử lý* *Sau khi xử lý*

**Hình 5.** Dữ liệu trước và sau khi thay đổi kiểu dữ liệu

* Tương tự đối với cột “Credit\_Score” và “Payment\_of\_Min\_Amount”, các nhãn của cột “Credit\_Score” lần lượt là ‘Good’, ‘Standard’, ‘Poor’ sẽ được thay thế thành ‘3’, ‘2’, ‘1’ và các nhãn cột “Payment\_of\_Min\_Amount” lần lượt là ‘NM’. ‘Yes’, ‘No’ được thay thế thành ‘0’, ‘1’, ‘2’, ‘3’. Cuối cùng chuyển kiểu dữ liệu 2 cột này thành số nguyên int64
* Sử dụng phương pháp IQR để loại bỏ các giá trị ngoại lệ. Đây là một phần quan trọng của tiền xử lý dữ liệu, vì ngoại lệ có thể làm méo mó các mô hình học máy và dẫn đến dự đoán không chính xác.

Phương pháp IQR (Interquartile Range) thường được sử dụng để xác định ngoại lệ. IQR là phạm vi giữa quartile thứ nhất (Q1) và quartile thứ ba (Q3) của tập dữ liệu. Các giá trị nằm ngoài khoảng từ Q1 - 1.5IQR đến Q3 + 1.5IQR thường được coi là ngoại lệ và có thể bị loại bỏ khỏi dữ liệu phân tích.

**Bước 5:** Sử dụng module Random Over Sampler xử lý sự mất cân bằng dữ liệu trong tập huấn luyện



**Hình 6.** Tỷ lệ các quan sát trong trường dữ liệu “Credit\_Score” sau khi cân bằng

**Bước 6:** Sử dụng phương pháp Min-Max Scaling để chuẩn hóa các biến số (numerical variables) . Các trường dữ liệu thuộc biến độc lập lúc này sẽ được chuẩn hóa Min - Max đưa dữ liệu về một phân bố trong phạm vi [0, 1], chuẩn hóa được xác định theo Công thức (1).

(1)

trong đó: x là giá trị ban đầu

x\_scaled là giá trị sau khi áp dụng Min-Max Scaling (thường nằm trong khoảng 0 đến 1)

x\_min, x\_max lần lượt là giá trị nhỏ nhất, giá trị lớn nhất của trường dữ liệu trong tập dữ liệu.

Bảng 2 minh họa 3 quan sát đầu của dữ liệu sau khi được chuẩn hóa.

**Bảng 2:** Dữ liệu sau chuẩn hóa

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Age | Annual Income | Delay from due date | Num of Delayed Payment | Outstanding Debt | Credit History Age | Total EMI per month | Monthly Balance |
| 0.07031 | 0.083668 | 0.111111 | 0.002500 | 0.162020 | 0.67062 | 0.000602 | 1.0 |
| 0.10937 | 0.192389 | 0.166667 | 0.000909 | 0.121012 | 0.79228 | 0.000229 | 1.0 |
| 0.10937 | 0.192389 | 0.111111 | 0.000455 | 0.121012 | 0.79525 | 0.000229 | 1.0 |

**Bước 5:** *Mã hóa dữ liệu*

* Sử dụng phương pháp One - Hot Encoding để biến đổi các giá trị của cột "Occupation" thành các biến giả định dạng số.

Kết quả là DataFrame chứa dữ liệu cần biến đổi sẽ bao gồm các biến giả định mới được tạo ra từ One-Hot Encoding, mỗi biến giả định tương ứng với một nhóm nghề nghiệp (Occupation) và chứa các giá trị 0 hoặc 1 để biểu thị sự hiện diện của từng nhóm nghề nghiệp.

**Bước 6:** *Xác định dữ liệu tập huấn và kiểm tra*

Bộ dữ liệu được chia theo tỷ lệ 80:20 với tập huấn luyện (80%) và tập kiểm tra (20%), việc huấn luyện mô hình trên toàn bộ tập dữ liệu có thể tốn nhiều thời gian và tài nguyên, đặc biệt với các tập dữ liệu lớn. Chia dữ liệu thành hai tập nhỏ hơn giúp giảm thời gian huấn luyện và cho phép bạn thử nghiệm nhiều mô hình khác nhau một cách hiệu quả hơn. Mục tiêu là để đảm bảo rằng dữ liệu được chia dẫn tới phương sai không quá cao.

* Tập huấn luyện (80%) cung cấp dữ liệu cho mô hình học và xây dựng mối quan hệ giữa các biến.
* Tập kiểm tra (20%) là tập dữ liệu độc lập, chưa được sử dụng trong quá trình huấn luyện. Việc đánh giá hiệu suất mô hình trên tập kiểm tra giúp đảm bảo tính khách quan, tránh tình trạng "quá khớp" (overfitting) - khi mô hình học thuộc lòng tập huấn luyện nhưng không thể khái quát hóa tốt cho dữ liệu mới.

## 4.3. Các độ đo đánh giá hiệu năng

Trong bài nghiên cứu này, các phương pháp đánh giá nhóm đã sử dụng là các chỉ số đánh giá Accuracy, Precision, Recall và F1 score, dựa trên ma trận nhầm lẫn có dạng như sau:

* Phương pháp Accuracy (độ chính xác) của mô hình là thước đo đơn giản nhất để đánh giá hiệu suất mô hình, tỷ lệ giữa số lượng mẫu được phân loại chính xác trên cho tổng số mẫu của tập dữ liệu theo công thức sau:

|  |
| --- |
|  |

Accuracy cho biết mức độ tin cậy chung của mô hình chấm điểm. Nếu đạt giá trị cao: Mô hình dự đoán chính xác điểm tín dụng của khách hàng và giá trị thấp: Mô hình có thể đưa ra dự đoán sai lệch.

* Phương pháp Precision (độ chính xác dương) là tỷ lệ giữa số lượng mẫu được dự đoán khớp mẫu trên tổng dự đoán khớp mẫu và dự đoán đúng nhưng sai, được biểu diễn theo công thức dưới đây:

|  |  |
| --- | --- |
|  |  |

Precision đánh giá mức độ tin cậy của dự đoán khách hàng có điểm tín dụng cao (tốt). Nếu đạt giá trị cao: Mô hình ít dự đoán sai "tốt" cho khách hàng có điểm thấp và giá trị thấp: Mô hình có thể dự đoán sai "tốt" cho khách hàng có điểm thấp, dẫn đến việc đánh giá rủi ro sai lệch.

* Phương pháp Recall (độ nhạy) là tỷ lệ giữa số lượng mẫu được dự đoán khớp mẫu với tổng dự đoán khớp mẫu và dự đoán sai nhưng đúng được xác định theo công thức sau:

|  |
| --- |
|  |

Trong đó:

* + TP (True Positive) là tổng số trường hợp dự báo khớp mẫu đúng.
  + TN (True Negative) là tổng số trường hợp dự báo khớp với mẫu sai.
  + FP (False Positive) là tổng số trường hợp dự báo các quan sát thuộc mẫu đúng tính thành sai.
  + FN (False Negative) là tổng số trường hợp dự báo các quan sát thuộc mẫu sai tính thành đúng

Recall đánh giá khả năng xác định chính xác khách hàng có điểm tín dụng cao. Nếu đạt giá trị cao: Mô hình ít bỏ sót khách hàng có điểm cao, giúp giảm thiểu việc đánh giá thấp rủi ro và giá trị thấp: Mô hình có thể bỏ sót khách hàng có điểm cao, dẫn đến việc cấp tín dụng sai cho những người có rủi ro cao.

* Phương pháp F1 - Score là trung bình trọng số của Precision và Recall, thể hiện sự cân bằng, trong đó F1 - Score đạt giá trị tốt nhất là 1 và kém nhất là 0, chỉ số F1 - Score được xác định theo công thức sau:

|  |
| --- |
|  |

F1- Score cân bằng giữa việc dự đoán chính xác "tốt" và xác định chính xác khách hàng có điểm cao. Nết đạt giá trị cao: Mô hình có hiệu quả tốt trong việc phân biệt khách hàng có điểm tín dụng cao và thấp và giá trị thấp: Mô hình cần được cải thiện để nâng cao hiệu quả phân loại.

Các chỉ số đánh giá Accuracy, Precision, Recall, và F1-Score là những chỉ số quan trọng để đánh giá hiệu quả mô hình tín dụng. Việc sử dụng các chỉ số này giúp lựa chọn mô hình phù hợp và cải thiện hiệu quả đánh giá tín dụng.

## 4.4. Các tham số và môi trường cài đặt

*Môi trường Phần mềm:*

Ngôn Ngữ Lập Trình: Sử dụng Python trong môi trường Jupyter Notebook, một công cụ phổ biến trong khoa học dữ liệu cho phép thực hiện mã, ghi chú và hiển thị kết quả một cách tương tác.

*Thư Viện Sử Dụng:*

* pandas và numpy cho việc xử lý và phân tích dữ liệu.
* matplotlib và seaborn cho việc trực quan hóa dữ liệu.
* sklearn cho các mô hình học máy như Decision Tree, KNN, Random Forest, SVM và Adaboost, cũng như các công cụ tiền xử lý dữ liệu như MinMaxScaler.
* tensorflow.keras được sử dụng cho việc xây dựng và huấn luyện mô hình mạng nơ-ron tích chập (CNN) và mạng nơ-ron hồi quy (RNN).

*Môi trường Phần cứng:* Môi trường thử nghiệm của nhóm chúng tôi như sau: CPU Intel Core i5-8265U, RAM: 2.80GHz.

*Thiết Lập Thí Nghiệm:* Được thực hiện trên tập dữ liệu đã được tiền xử lý, với việc chia tách tập dữ liệu huấn luyện và kiểm thử mô hình, và sử dụng các phương pháp đánh giá như accuracy, precision, recall, và F1-score để đánh giá hiệu quả của mô hình.

## 4.5. Các phương pháp cơ sở

Trong phần nghiên cứu này, nhóm chúng tôi thực hiện so sánh 7 mô hình để đánh giá tính hiệu quả và chính xác. Tất cả các phương pháp được thực hiện dưới cùng một điều kiện thí nghiệm để đảm bảo công bằng trong so sánh:

* Phương pháp phân lớp: Decision Tree, K-Nearest Neighbor, Random Forest, XGBoost, Naive Bayes
* Mô hình học sâu: Convolutional Neural Network, Recurrent Neural Network

Các thí nghiệm được thực hiện trên cùng một tập dữ liệu, sử dụng cùng một quy trình tiền xử lý và chia dữ liệu để đảm bảo tính khách quan. Mỗi phương pháp được đánh giá dựa trên các chỉ số như accuracy, precision, recall, và F1-score để có cái nhìn toàn diện về hiệu suất của chúng.

## 4.6. Kết quả và phân tích

*4.6.1. Các kỹ thuật học máy truyền thống*

Trong bài nghiên cứu, nhóm chúng tôi đã xây dựng, thử nghiệm các mô hình học máy truyền thống sau: Decision Tree (Cây quyết định), K-Nearest Neighbor (K- láng giềng gần nhất), Random Forest (Rừng ngẫu nhiên), XGBoost, Naive Bayes và có kết quả được trình bày trong bảng dưới đây:

**Bảng 3.** Kết quả đánh giá các mô hình học máy truyền thống

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| 1 | **Decision Tree** | 0.855382 | 0.854924 | 0.855382 | 0.851688 |
| 2 | **K-Nearest Neighbor** | 0.712090 | 0.707868 | 0.712090 | 0.702121 |
| **3** | **Random Forest** | **0.885779** | **0.887773** | **0.885779** | **0.882678** |
| 4 | **XGBoost** | 0.835985 | 0.837708 | 0.835985 | 0.830886 |
| 5 | **Naive Bayes** | 0.617953 | 0.611433 | 0.617953 | 0.585627 |

Qua bảng phân tích của các mô hình học máy cổ khi sử dụng lần lượt các chỉ số đánh giá hiệu suất Accuracy, Precision, Recall, F1-score và kết quả cho thấy mô hình Random Forest (Rừng ngẫu nhiên) có hiệu suất đạt tốt nhất khi các chỉ số đều đạt trên 88%, cao hơn các mô hình còn lại khoảng từ 0-30% tùy vào từng phương pháp so sánh. Bên cạnh đó, Naive Bayes và K-Nearest Neighbor hai mô hình có hiệu suất được nhìn nhận là thấp nhất trong bảng phân tích với kết quả lần lượt là hơn 60% và hơn 70% ở cả bốn chỉ số đánh giá.

*4.6.2 Các kỹ thuật học sâu*

Nhóm chúng tôi xây dựng mô hình Convolutional Neural Network (Mạng neural tích chập sâu) và Recurrent Neural Network (Mạng nơ ron hồi quy), thu được kết quả sau:

**Bảng 4.** Kết quả đánh giá các mô hình học sâu

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** | **Thời gian** |
| 1 | **CNN** | 0.686358 | 0.679902 | 0.686358 | 0.677860 | **1 phút 56s** |
| 2 | **RNN** | 0.675702 | 0.671962 | 0.675702 | 0.663117 |  |

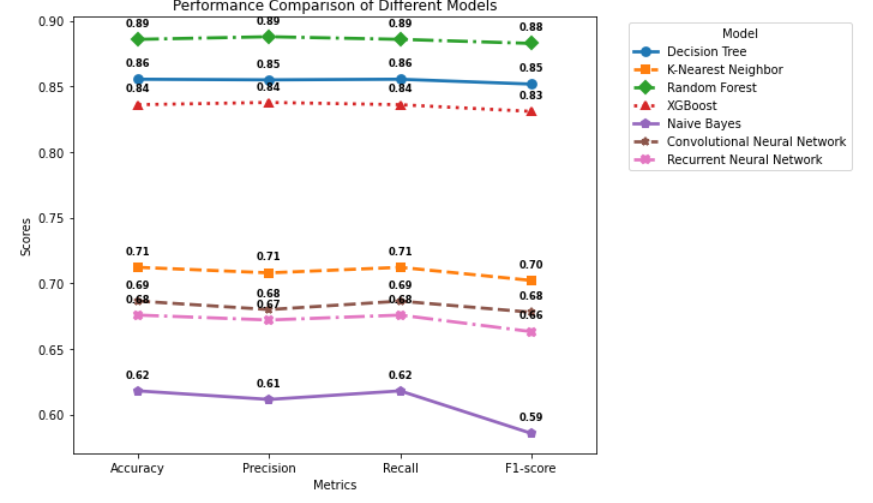
Hai mô hình ở bảng 4 đưa ra kết quả là hiệu suất ở cả 2 phương pháp đánh giá xấp xỉ nhau. Tuy nhiên, mô hình Convolutional Neural Network đã có mức % nhỉnh hơn một chút so với mô hình còn lại ví dụ như Accuracy của CNN chỉ cao hơn 1% so với RNN. Quan sát kết quả thu được có thể thấy RNN đưa ra kết quả không cao và thời gian thực nghiệm dài hơn CNN khoảng

*4.6.3. So sánh các mô hình*

**Bảng 5.** So sánh các mô hình

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| 1 | **DT** | 0.855382 | 0.854924 | 0.855382 | 0.851688 |
| 2 | **KNN** | 0.712090 | 0.707868 | 0.712090 | 0.702121 |
| 3 | **RF** | **0.885779** | **0.887773** | **0.885779** | **0.882678** |
| 4 | **XGB** | 0.835985 | 0.837708 | 0.835985 | 0.830886 |
| 5 | **NB** | 0.617953 | 0.611433 | 0.617953 | 0.585627 |
| 7 | **CNN** | 0.686358 | 0.679902 | 0.686358 | 0.677860 |
| 8 | **RNN** | 0.675702 | 0.671962 | 0.675702 | 0.663117 |

Dựa vào bảng tổng hợp các chỉ số đánh giá hiệu suất của các mô hình trên, ta nhận thấy rằng mô hình Random Forest có tỷ lệ hiệu suất cao nhất trong 7 mô hình còn lại lần lượt với tất cả các chỉ số đánh giá đều trên 88%.



**Hình 7.** Đánh giá hiệu quả của các mô hình

Thông qua biểu đồ trên ta nhận thấy rằng các phương pháp đánh giá trên mô hình RF có hiệu suất cao nhất (khoảng 89%), tiếp đó là mô hình DT với hiệu suất chỉ kém 2-3%. Ngoài ra, mô hình XGB có hiệu suất cao thứ ba (khoảng 84%). Với phương pháp đánh giá hiệu suất bằng 4 chỉ số accuracy, precision, recall, f1-score thì mô hình RNN có kết quả hiệu suất là thấp nhất (trên 60%).

# V. KẾT LUẬN

Tóm lại, bài báo này đã so sánh nhiều cách tiếp cận khác nhau của mô hình chấm điểm tín dụng, mô hình truyền thống với 5 mô hình cơ sở là DT, KNN, RF, XGB và NB; mô hình học sâu CNN và RNN. Việc đánh giá trên nhiều mô hình lựa chọn ra mô hình có khả năng chấm điểm tín dụng chính xác nhất. Các thí nghiệm đã chỉ ra rằng mô hình Random Forest được huấn luyện vượt trội hơn các mô hình khác với các chỉ số đánh giá hiệu năng accury, precision, recall, f1-score đều trên 88%. Đối với phương pháp xây dựng tính năng, mỗi phương pháp thể hiện thế mạnh của mình khi được đào tạo bởi các mô hình khác nhau. Tuy nhiên, cần nhấn mạnh rằng không có mô hình tính điểm tín dụng lý tưởng cũng như không có kỹ thuật tốt nhất được sử dụng để xây dựng mô hình chấm điểm tín dụng; và kỹ thuật tốt cho tập dữ liệu này chưa chắc tốt trên tập dữ liệu khác. Kỹ thuật này hơn kỹ thuật khác khả năng dự đoán trong các trường hợp cụ thể khác nhau. “Kĩ thuật tốt nhất” phụ thuộc vào từng yêu cầu cụ thể, cấu trúc của dữ liệu, các chức năng của ứng dụng, mức độ tách biệt các lớp bằng mục tiêu phân loại.

Trong công việc trong tương lai, để cải thiện hiệu suất của các mô hình được đề xuất, nhiều mô hình và kỹ thuật học máy sẽ được khám phá, chẳng hạn như Deep Forest và xếp chồng mô hình. Ngoài ra, chúng tôi đang có kế hoạch thử nghiệm các mô hình được đề xuất với nhiều bộ dữ liệu chấm điểm tín dụng hơn để xác minh tính mạnh mẽ của chúng.

**TÀI LIỆU THAM KHẢO**

|  |  |  |
| --- | --- | --- |
| **ID** | **CITE** | **ABSTRACT** |
| 1 | @inproceedings{Ha2016CreditSW,  title={Credit scoring with a feature selection approach based deep learning},  author={Van-Sang Ha and Ha-Nam Nguyen},  year={2016}, | [Credit scoring with a feature selection approach based deep learning](https://www.semanticscholar.org/paper/Credit-scoring-with-a-feature-selection-approach-Ha-Nguyen/50ddde187f34a555e9c565b7e046a354ac3c82af)  In financial risk, credit risk management is one of the most important issues in financial decision-making. Reliable credit scoring models are crucial for financial agencies to evaluate credit applications and have been widely studied in the field of machine learning and statistics. Deep learning is a powerful classification tool which is currently an active research area and successfully solves classification problems in many domains. Deep Learning provides training stability, generalization, and scalability with big data. Deep Learning is quickly becoming the algorithm of choice for the highest predictive accuracy. Feature selection is a process of selecting a subset of relevant features, which can decrease the dimensionality, reduce the running time, and improve the accuracy of classifiers. In this study, we constructed a credit scoring model based on deep learning and feature selection to evaluate the applicant’s credit score from the applicant’s input features. Two public datasets, Australia and German credit ones, have been used to test our method. The experimental results of the real world data showed that the proposed method results in a higher prediction rate than a baseline method for some certain datasets and also shows comparable and sometimes better performance than the feature selection methods widely used in credit scoring. |
| 2 | @inproceedings{Zhou2018ASO,  title={A state of the art survey of data mining-based fraud detection and credit scoring},  author={Xun Zhou and Si-Jie Cheng and Mengxia Zhu and Chen Guo and Sida Zhou and Peng Xu and Zhenghua Xue and Weishi Zhang},  year={2018}, | [A state of the art survey of data mining-based fraud detection and credit scoring](https://www.semanticscholar.org/paper/A-state-of-the-art-survey-of-data-mining-based-and-Zhou-Cheng/f942352fc84b71231d71c26e6c3a6d1360fe5028)  Credit risk has been a widespread and deep penetrating problem for centuries, but not until various credit derivatives and products were developed and novel technologies began radically changing the human society, have fraud detection, credit scoring and other risk management systems become so important not only to some specific firms, but to industries and governments worldwide. Frauds and unpredictable defaults cost billions of dollars each year, thus, forcing financial institutions to continuously improve their systems for loss reduction. In the past twenty years, amounts of studies have proposed the use of data mining techniques to detect frauds, score credits and manage risks, but issues such as data selection, algorithm design, and hyperparameter optimization affect the perceived ability of the proposed solutions and it is difficult for auditors and researchers to explore and figure out the highest level of general development in this area. In this survey we focus on a state of the art survey of recently developed data mining techniques for fraud detection and credit scoring. Several outstanding experiments are recorded and highlighted, and the corresponding techniques, which are mostly based on supervised learning algorithms, unsupervised learning algorithms, semi-supervised algorithms, ensemble learning, transfer learning, or some hybrid ideas are explained and analysed. The goal of this paper is to provide a dense review of up-to-date techniques for fraud detection and credit scoring, a general analysis on the results achieved and upcoming challenges for further researches. |
| 3 | @article{Rajamohamed2018ImprovedCC,  title={Improved credit card churn prediction based on rough clustering and supervised learning techniques},  author={R. Rajamohamed and J. Manokaran},  journal={Cluster Computing},  year={2018},  pages={1-13}, | [Improved credit card churn prediction based on rough clustering and supervised learning techniques](https://www.semanticscholar.org/paper/Improved-credit-card-churn-prediction-based-on-and-Rajamohamed-Manokaran/917a5bff467602c0d9d57c3fe6da447661962ecb)  Every process is digitized in the current society. Transfer of money from one account holder to another has become possible in seconds because of advanced technologies in information processing. Not only this in all sectors like railways, insurance, health sector, fashion technology, education sector, sales and business sectors, and advertisement sectors every firm digitized its operations. One such sector is banking where every individual based on his or her financial status will be considered for crediting loan, and credit card etc. If the credit score of the loan availing person is high banks will be ready to provide him with the loan but the availing person can opt for any one of the banks on his or her own willing. Such scenario happens in credit card churn prediction also. Hence the banks should take healthy measures to retain the existing credit card holders without any churn. Withholding existing customers of a firm plays an important role to increase the overall revenue of the firm and retains the good name of the firm in competitive market. Hence every organization takes key measures to withhold existing customers using customer management models. Because customer retention is a crucial task as it reduces the time, money and workforce needed for adding new customers to the firm. Customers retention technique in credit card churn prediction (C3P) was done using only supervised classification techniques. But it could not end with better results. So, through many proven hybrid classification techniques we can bring better accuracy in C3P. Also C3P lags in highly efficient techniques like rough set theory. Hence in this work initially we perform data processing techniques and in second stage we propose modified rough K-means algorithm used for clustering credit card holders and in next stage hold-out method divides the cluster data into testing and training clusters. At last classification is performed using various algorithms like support vector machine, random forest, decision tree, K-nearest neighbor, and Naive Bayes. Finally we evaluate the work using precision, recall (sensitivity), specification, accuracy, and misclassification error. |
| 4 | @article{Zanin2017CreditCF,  title={Credit card fraud detection through parenclitic network analysis},  author={Massimiliano Zanin and Miguel Romance and Santiago Moral and Regino Criado},  journal={ArXiv},  year={2017},  volume={abs/1706.01953}, | [Credit card fraud detection through parenclitic network analysis](https://www.semanticscholar.org/paper/Credit-card-fraud-detection-through-parenclitic-Zanin-Romance/42122ffc1b637e2b267622bcef6cf130dbeb12c4)  The detection of frauds in credit card transactions is a major topic in financial research, of profound economic implications. While this has hitherto been tackled through data analysis techniques, the resemblances between this and other problems, like the design of recommendation systems and of diagnostic/prognostic medical tools, suggest that a complex network approach may yield important benefits. In this paper we present a first hybrid data mining/complex network classification algorithm, able to detect illegal instances in a real card transaction data set. It is based on a recently proposed network reconstruction algorithm that allows creating representations of the deviation of one instance from a reference group. We show how the inclusion of features extracted from the network data representation improves the score obtained by a standard, neural network-based classification algorithm and additionally how this combined approach can outperform a commercial fraud detection system in specific operation niches. Beyond these specific results, this contribution represents a new example on how complex networks and data mining can be integrated as complementary tools, with the former providing a view to data beyond the capabilities of the latter. |
| 5 | @article{Ignatius2018AFD,  title={A fuzzy decision support system for credit scoring},  author={Joshua Ignatius and Adel Hatami-Marbini and Amirah Rahman and Lalitha Dhamotharan and Pegah Khoshnevis},  journal={Neural Computing and Applications},  year={2018},  volume={29},  pages={921-937}, | [A fuzzy decision support system for credit scoring](https://www.semanticscholar.org/paper/A-fuzzy-decision-support-system-for-credit-scoring-Ignatius-Hatami-Marbini/807e2de6cd6031ce9c58fb190d59443536cff2c6)  Credit score is a creditworthiness index, which enables the lender (bank and credit card companies) to evaluate its own risk exposure toward a particular potential customer. There are several credit scoring methods available in the literature, but one that is widely used is the FICO method. This method provides a score ranging from 300 to 850 as a fast filter for high-volume complex credit decisions. However, it falls short in the aspect of a decision support system where revised scoring can be achieved to reflect the borrower’s strength and weakness in each scoring dimension, as well as the possible trade-offs made to maintain one’s lending risk. Hence, this study discusses and develops a decision support tool for credit score model based on multi-criteria decision-making principles. In the proposed methodology, criteria weights are generated by fuzzy AHP. Fuzzy linguistic theory is applied in AHP to describe the uncertainties and vagueness arising from human subjectivity in decision making. Finally, drawing from the risk distance function, TOPSIS is used to rank the alternatives based on the least risk exposure. A sensitivity analysis is also demonstrated by the proposed fuzzy AHP-TOPSIS method. |
| 6 | @article{Li2020ACP,  title={A Comparative Performance Assessment of Ensemble Learning for Credit Scoring},  author={Yiheng Li and Weidong Chen},  journal={Mathematics},  year={2020}, | [A Comparative Performance Assessment of Ensemble Learning for Credit Scoring](https://www.semanticscholar.org/paper/A-Comparative-Performance-Assessment-of-Ensemble-Li-Chen/8fbbcb390f16441bfd6594153dfc4c3022ea3872)  Extensive research has been performed by organizations and academics on models for credit scoring, an important financial management activity. With novel machine learning models continue to be proposed, ensemble learning has been introduced into the application of credit scoring, several researches have addressed the supremacy of ensemble learning. In this research, we provide a comparative performance evaluation of ensemble algorithms, i.e., random forest, AdaBoost, XGBoost, LightGBM and Stacking, in terms of accuracy (ACC), area under the curve (AUC), Kolmogorov–Smirnov statistic (KS), Brier score (BS), and model operating time in terms of credit scoring. Moreover, five popular baseline classifiers, i.e., neural network (NN), decision tree (DT), logistic regression (LR), Naïve Bayes (NB), and support vector machine (SVM) are considered to be benchmarks. Experimental findings reveal that the performance of ensemble learning is better than individual learners, except for AdaBoost. In addition, random forest has the best performance in terms of five metrics, XGBoost and LightGBM are close challengers. Among five baseline classifiers, logistic regression outperforms the other classifiers over the most of evaluation metrics. Finally, this study also analyzes reasons for the poor performance of some algorithms and give some suggestions on the choice of credit scoring models for financial institutions. |
| 7 | @article{Maniraj2019CreditCF,  title={Credit Card Fraud Detection using Machine Learning and Data Science},  author={S. P. Maniraj and Aditya Saini and Shadab Ahmed and Swarna Deep Sarkar},  journal={International Journal of Engineering Research and},  year={2019}, | [Credit Card Fraud Detection using Machine Learning and Data Science](https://www.semanticscholar.org/paper/Credit-Card-Fraud-Detection-using-Machine-Learning-Maniraj-Saini/81cd45d9dd13e9c0a15b076cfd97f7d0768cf238)  Credit card fraud detection is presently the most frequently occurring problem in the present world. This is due to the rise in both online transactions and e-commerce platforms. Credit card fraud generally happens when the card was stolen for any of the unauthorized purposes or even when the fraudster uses the credit card information for his use. In the present world, we are facing a lot of credit card problems. To detect the fraudulent activities the credit card fraud detection system was introduced. This project aims to focus mainly on machine learning algorithms. The algorithms used are random forest algorithm ,linear regression , XGBoost ,KNearest, Support vector classifier, Linear Discriminant Analysis, GaussianNB algorithm. The results of the algorithms are based on accuracy, precision, recall, and F1-score. The ROC curve is plotted based on the confusion matrix. Algorithms are compared and the algorithm that has the greatest accuracy, precision, recall, and F1-score is considered as the best algorithm that is used to detect the fraud. Introduction: Credit card fraud is a significant threat in the BFSI sector. This credit card fraud detection system studies and analyzes user behavior patterns and uses location scanning techniques to identify any unusual patterns. One of The user patterns includes important user behavior like spending habits, usage patterns, etc. The system uses geographic location for identity verification. In case it detects any unusual pattern, the user will be required to undergo the verification. The fraud detection system stores thepast transaction data of each user. Based on this data, it calculates the standard user behavior patterns for individual users, and any deviation from those normal patterns becomes a trigger for the system. In the instance of any unusual activity, the system will not only raise alerts, but it will also block the user after three invalid attempts. 'Fraud' in credit card transactions is unauthorized and unwanted usage of an account by someone other than the owner of that account. Necessary prevention measures can be taken to stop this abuse and the behaviour of such fraudulent practices can be studied to minimize it and protect against similar occurrences in the future.In other words, Credit Card Fraud can be defined as a case where a person uses someone else’s credit card for personal reasons while the owner and the card issuing authorities are unaware of the fact that the card is being used. Fraud detection involves monitoring the activities of populations of users in order to estimate, perceive or avoid objectionable behaviour, which consist of fraud, intrusion, and defaulting. This is a very relevant problem that demands the attention of communities such as machine learning and data science where the solution to this problem can be automated. This problem is particularly challenging from the perspective of learning, as it is characterized by various factors such as class imbalance. The number of valid transactions far outnumber fraudulent ones. Also, the transaction patterns often change their statistical properties over the course of time. These are not the only challenges in the implementation of a real-world fraud detection system, however. In real world examples, the massive stream of payment requests is quickly scanned by automatic tools that determine which transactions to authorize. Machine © 2021 JETIR May 2021, Volume 8, Issue 5 www.jetir.org (ISSN-2349-5162) JETIR2105307 Journal of Emerging Technologies and Innovative Research (JETIR) www.jetir.org c405 learning algorithms are employed to analyse all the authorized transactions and report the suspicious ones. These reports are investigated by professionals who contact the cardholders to confirm if the transaction was genuine or fraudulent. The investigators provide a feedback to the automated system which is used to train and update the algorithm to eventually improve the fraud-detection performance over time. Literature Survey: New methods for credit card fraud detection with a lot of research methods and several fraud detection techniques with a special interest in the neural networks, data mining, and distributed data mining. Many other techniques are used to detect such credit card fraud. When done the literature survey on various methods of credit card fraud detection, we can conclude that to detect credit card fraud there are many other approaches in Machine Learning itself. The research on credit card fraud detection uses both Machine Learning and Deep Learning algorithms. In this section, we enhance the work done in two different points: (i) the methods that are readily available for fraud detection (ii) The techniques that are available to handle the imbalanced data. To handle the imbalanced data some of the techniques are available. They are (a) classification methods (b) sampling methods (c) resembling techniques. Here are some of the Machine Learning algorithms that are used for credit fraud detection are support vector machine(SVM), decision trees, logistic regression, gradient boosting, K-nearest neighbor, etc. In 2019, Yashvi Jain, NamrataTiwari, Shripriya Dubey, Sarika jain have researched various techniques[10] for credit cards fraud detection such as support vector machines(SVM), artificial neural networks(ANN), Bayesian Networks, Hidden Markov Model, K-Nearest Neighbours (KNN) Fuzzy Logic system and Decision Trees. In their paper, they have observed that the algorithms k-nearest neighbor, decision trees, and the SVM give a medium level accuracy. The Fuzzy Logic and Logistic Regression give the lowest accuracy among all the other algorithms. Neural Networks, naive bayes, fuzzy systems, and KNN offer a high detention rate. The Logistic Regression, SVM, decision trees offer a high detection rate at the medium level. There are two algorithms namely ANN and the Naïve Bayesian Networks which perform better at all parameters. These are very much expensive to train. There is a major drawback in all the algorithms. The drawback is that these algorithms don’t give the same result in all types of environments. They give better results with one type of datasets and poor results with another type of dataset. Algorithms like KNN and SVM give excellent results with small datasets and algorithms like logistic regression and fuzzy logic systems give good accuracy with raw and unsampled data. In 2019, Heta Naik, Prashasti Kanikar, has done their research on various algorithms like Naïve Bayes, Logistic Regression, J48, and Adaboost. Naïve Bayes on among the classification algorithm. This algorithm depends upon Bayes theorem. Bayes's theorem finds the probability of an event that is occurring is given. The Logistic regression algorithm is similar to the linear regression algorithm. The linear regression is used for the prediction or forecasting the values. The logistic regression is mostly used for the classification task. The J48 algorithm is used to generate a decision tree and is used for the classification problem. The J48 © 2021 JETIR May 2021, Volume 8, Issue 5 www.jetir.org (ISSN-2349-5162) JETIR2105307 Journal of Emerging Technologies and Innovative Research (JETIR) www.jetir.org c406 is the extension of the ID3 (Iterative Dichotomieser). J48 is one of the most widely used and extensively analyzed areas in Machine Learning. This algorithm mainly works on constant and categorical variables. Adaboost is one of the most widely used machine learning algorithms and is mainly developed for binary classification. The algorithm is mainly used to boost the performance of the decision tree. This is also mainly used for the classification of the regression. The Adaboost algorithm is fraud cases to classify the transactions which are fraud and non-fraud. From their work they have concluded that the highest accuracy is obtained for both the Adaboost and Logistic Regression. As they have the same accuracy the time factor is considered to choose the best algorithm. By considering the time factor they concluded that the Adaboost algorithm works well to detect credit card fraud. In 2019 Sahayasakila V, D.Kavya Monisha, Aishwarya, Sikhakolli Venkatavisalakshiswshai Yasaswi have explained the Twain important algorithmic techniques which are the Whale Optimization Techniques (WOA) and SMOTE (Synthetic Minority Oversampling Techniques). They mainly aimed to improve the convergence speed and to solve the data imbalance problem. The class imbalance problem is overcome using the SMOTE technique and the WOA technique. The SMOTE technique discriminates all the transactions which are synthesized are again resampled to check the data accuracy and are optimized using the WOA technique. The algorithm also improves the convergence speed, reliability, and efficiency of the system. In 2018 Navanushu Khare and Saad Yunus Sait have explained their work on decision trees, random forest, SVM, and logistic regression. They have taken the highly skewed dataset and worked on such type of dataset. The performance evaluation is based on accuracy, sensitivity, specificity, and precision. The results indicate that the accuracy for the Logistic Regression is 97.7%, for Decision Trees is 95.5%, for Random Forest is 98.6%, for SVM classifier is 97.5%. They have concluded that the Random Forest algorithm has the highest accuracy among the other algorithms and is considered as the best algorithm to detect the fraud. They also concluded that the SVM algorithm has a data imbalance problem and does not give better results to detect credit card fraud. Methodology: System development method is a process through which a product will get completed or a product gets rid from any problem. Software development process is described as a number of phases, procedures and steps that gives the complete software. It follows series of steps which is used for product progress. |
| 8 | @article{Chen2017TheTS,  title={The Transparent Self Under Big Data Profiling: Privacy and Chinese Legislation on the Social Credit System},  author={Yongxi Chen and Anne S. Y. Cheung},  journal={LSN: Data Protection (Sub-Topic)},  year={2017}, | [The Transparent Self Under Big Data Profiling: Privacy and Chinese Legislation on the Social Credit System](https://www.semanticscholar.org/paper/The-Transparent-Self-Under-Big-Data-Profiling%3A-and-Chen-Cheung/bd28d67f07d29a02f3c04a86535554e273d64bd7)  China’s Social Credit System (SCS) has captured the imagination and power of big data technology. Launched at the national level in 2014, the system’s aim is to assess the trustworthiness of Chinese citizens in keeping their promises and complying with legal rules, moral norms, and professional and ethical standards. It is essentially a collaborative project between the authorities and the business sectors to rate both individuals and other entities. In certain contexts, each individual will be given a social credit score or rank. Despite the claimed good will to curb escalating dishonesty across societal sectors in China, the worry is that the totality of individuals’ lives will be captured, that citizens will be monitored and that the Orwellian state will become a reality. Individuals risk being reduced to transparent selves before the state in this uneven battle. They are uncertain about what contributes to their social credit scores, how those scores are combined with the state system, and how their data is interpreted and used. In short, the big data-driven SCS is confronting Chinese citizens with major challenges to their privacy and personal data, amongst other interests critical to them. This article first maps out the background to the construction of China’s big data social laboratory and the SCS, and then summarises the legislative history and evolving concept of social credit. It stresses that apart from the conspicuous SCS policy document introduced by the Chinese central government, pilot legislation has already been implemented at the local levels to regulate the collection and processing of social credit data. The third part critically reviews such local legislation (which often uses the term ‘public credit information’) with reference to personal data protection principles. It also highlights the restrictions on the data subjects’ rights that are placed by the uncoordinated legal framework of personal data and the extra-legal regime of personal archive. It argues that existing legislation and proposed regulations require substantial revisions to mitigate the impacts of the SCS on data privacy. This article hopes to lay the groundwork for further legal study related to social credit and big data in China. |
| 9 | @article{Kuppili2020CreditSC,  title={Credit score classification using spiking extreme learning machine},  author={Venkatanareshbabu Kuppili and Diwakar Tripathi and Damodar Reddy Edla},  journal={Computational Intelligence},  year={2020},  volume={36},  pages={402 - 426}, | [Credit score classification using spiking extreme learning machine](https://www.semanticscholar.org/paper/Credit-score-classification-using-spiking-extreme-Kuppili-Tripathi/13f7e89e177fe2eff2bbe7059829060ea1b9c0c0)  Credit score classification is a prominent research problem in the banking or financial industry, and its predictive performance is responsible for the profitability of financial industry. This paper addresses how Spiking Extreme Learning Machine (SELM) can be effectively used for credit score classification. A novel spike‐generating function is proposed in Leaky Nonlinear Integrate and Fire Model (LNIF). Its interspike period is computed and utilized in the extreme learning machine (ELM) for credit score classification. The proposed model is named as SELM and is validated on five real‐world credit scoring datasets namely: Australian, German‐categorical, German‐numerical, Japanese, and Bankruptcy. Further, results obtained by SELM are compared with back propagation, probabilistic neural network, ELM, voting‐based Q‐generalized extreme learning machine, Radial basis neural network and ELM with some existing spiking neuron models in terms of classification accuracy, Area under curve (AUC), H‐measure and computational time. From the experimental results, it has been noticed that improvement in accuracy and execution time for the proposed SELM is highly statistically important for all aforementioned credit scoring datasets. Thus, integrating a biological spiking function with ELM makes it more efficient for categorization. |
| 10 | @article{Hassija2020SecureLB,  title={Secure Lending: Blockchain and Prospect Theory-Based Decentralized Credit Scoring Model},  author={Vikas Hassija and Gaurang Bansal and Vinay Chamola and Neeraj Kumar and Mohsen Guizani},  journal={IEEE Transactions on Network Science and Engineering},  year={2020},  volume={7},  pages={2566-2575}, | [Secure Lending: Blockchain and Prospect Theory-Based Decentralized Credit Scoring Model](https://www.semanticscholar.org/paper/Secure-Lending%3A-Blockchain-and-Prospect-Credit-Hassija-Bansal/28d49e482ac4d8110d05ded3afbb19dfc31dacd9)  Credit scoring is a rigorous statistical analysis carried out by lenders and other third parties to access an individual's creditworthiness. Lenders use credit scoring to estimate the degree of risk in lending money to an individual. However, credit score evaluation is primarily based on a transaction record, payment history, professional background, etc. sourced from different credit bureaus. So, evaluating a credit score is a laborious and tedious task involving a lot of paperwork. In this paper, we propose how blockchain can provide the solution to decentralized credit scoring evaluation and reducing the amount of dependence of paperwork. Lending money is not always objective but subjective to every lender. The decision of lending involves different levels of risk and uncertainty, depending on their perspective. This paper uses the prospect theory to model the optimal investment strategy for different risk vs. return scenarios. |
| 11 | @inproceedings{Kumar2020ReviewOM,  title={Review of Machine Learning models for Credit Scoring Analysis},  author={Madapuri Rudra Kumar and Vinit Kumar Gunjan},  year={2020}, | [Review of Machine Learning models for Credit Scoring Analysis](https://www.semanticscholar.org/paper/Review-of-Machine-Learning-models-for-Credit-Kumar-Gunjan/569766945340bb024fc818b24292dc045943d949)  Introduction: Increase in computing power and the deeper usage of the robust computing systems in the financial system is propelling the business growth, improving the operational efficiency of the financial institutions, and increasing the effectiveness of the transaction processing solutions used by the organizations. Problem: Despite that the financial institutions are relying on the credit scoring patterns for analyzing the credit worthiness of the clients, still there are many factors that are imminent for improvement in the credit score evaluation patterns. There is need for improving the pattern to enhance the quality of analysis. Objective: Machine learning is offering immense potential in Fintech space and determining a personal credit score. Organizations by applying deep learning and machine learning techniques can tap individuals who are not being serviced by traditional financial institutions. Methodology: One of the major insights into the system is that the traditional models of banking intelligence solutions are predominantly the programmed models that can align with the information and banking systems that are used by the banks. But in the case of the machine-learning models that rely on algorithmic systems require more integral computation which is intrinsic. Hence, it can be advocated that the models usually need to have some decision lines wherein the dynamic calibration model must be streamlined. Such structure demands the dynamic calibration to have a decision tree system to empower with more integrated model changes. Results: The test analysis of the proposed machine learning model indicates effective and enhanced analysis process compared to the non-machine learning solutions. The model in terms of using various classifiers indicate potential ways in which the solution can be significant. Conclusion: If the systems can be developed to align with more pragmatic terms for analysis, it can help in improving the process conditions of customer profile analysis, wherein the process models have to be developed for comprehensive analysis and the ones that can make a sustainable solution for the credit system management. Originality: The proposed solution is effective and the one conceptualized to improve the credit scoring system patterns. If the model can be improved with more effective parameters and learning metrics, it can be sustainable outcome. Limitations: The model is tested in isolation and not in comparison to any of the existing credit scoring patterns. Only the inputs in terms of shortcomings from the existing models are taken in to account and accordingly the proposed solution is developed. |
| 12 | @article{Shoumo2019ApplicationOM,  title={Application of Machine Learning in Credit Risk Assessment: A Prelude to Smart Banking},  author={Syed Zamil Hasan Shoumo and Mir Ishrak Maheer Dhruba and Sazzad Hossain and Nawab Haider Ghani and Hossain Arif and Samiul Islam},  journal={TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)},  year={2019},  pages={2023-2028}, | [Application of Machine Learning in Credit Risk Assessment: A Prelude to Smart Banking](https://www.semanticscholar.org/paper/Application-of-Machine-Learning-in-Credit-Risk-A-to-Shoumo-Dhruba/882533f8557a6670be0446b5d52663049d6049a5)  A precise credit risk assessment system is always vital to any financial institution for impeccable and gainful functioning. In such an ever-changing economy as the rate of loan defaults are gradually increasing, authorities of financial institutions are finding it more and more difficult to correctly assess loan requests and tackle the risks of loan defaulters. In light of these events this paper proposes a machine learning model which can precisely assess credit risk and predict possible loan defaulters for credit lending institutions. A comparative analysis has been made using tuned supervised learning algorithms such as Support Vector Machine, Random Forest, Extreme Gradient Boosting and Logistic Regression for identifying defaulters. Recursive Feature Elimination with Cross-Validation and Principal Component Analysis have been used for dimensionality reduction. Metrics such as F1 score, AUC score, prediction accuracy, precision and recall have been used to evaluate each model. Among all the models, the combination of a tuned Support Vector Machine and Recursive Feature Elimination with Cross-Validation have shown great promise in identifying loan defaulters. The proposed model, therefore, can assist financial institutions in accurately identifying loan defaulters and prevent them from incurring further loss. |
| 13 | @article{Hassija2020SecureLB,  title={Secure Lending: Blockchain and Prospect Theory-Based Decentralized Credit Scoring Model},  author={Vikas Hassija and Gaurang Bansal and Vinay Chamola and Neeraj Kumar and Mohsen Guizani},  journal={IEEE Transactions on Network Science and Engineering},  year={2020},  volume={7},  pages={2566-2575}, | [Secure Lending: Blockchain and Prospect Theory-Based Decentralized Credit Scoring Model](https://www.semanticscholar.org/paper/Secure-Lending%3A-Blockchain-and-Prospect-Credit-Hassija-Bansal/28d49e482ac4d8110d05ded3afbb19dfc31dacd9)  Credit scoring is a rigorous statistical analysis carried out by lenders and other third parties to access an individual's creditworthiness. Lenders use credit scoring to estimate the degree of risk in lending money to an individual. However, credit score evaluation is primarily based on a transaction record, payment history, professional background, etc. sourced from different credit bureaus. So, evaluating a credit score is a laborious and tedious task involving a lot of paperwork. In this paper, we propose how blockchain can provide the solution to decentralized credit scoring evaluation and reducing the amount of dependence of paperwork. Lending money is not always objective but subjective to every lender. The decision of lending involves different levels of risk and uncertainty, depending on their perspective. This paper uses the prospect theory to model the optimal investment strategy for different risk vs. return scenarios. |
| 14 | @article{Tripathi2020BinaryBA,  title={Binary BAT algorithm and RBFN based hybrid credit scoring model},  author={Diwakar Tripathi and Damodar Reddy Edla and Venkatanareshbabu Kuppili and Ramesh Dharavath},  journal={Multimedia Tools and Applications},  year={2020},  volume={79},  pages={31889 - 31912}, | [Binary BAT algorithm and RBFN based hybrid credit scoring model](https://www.semanticscholar.org/paper/Binary-BAT-algorithm-and-RBFN-based-hybrid-credit-Tripathi-Edla/edf545237fb920e5109e5b3b87061ce02f862b0d)  Credit scoring is a process of calculating the risk associated with an applicant on the basis of applicant’s credentials such as social status, financial status, etc. and it plays a vital role to improve cash flow for financial industry. However, the credit scoring dataset may have a large number of irrelevant or redundant features which leads to poorer classification performances and higher complexity. So, by removing redundant and irrelevant features may overcome the problem with huge number of features. This work emphasized on the role of feature selection and proposed a hybrid model by combining feature selection by utilizing Binary BAT optimization technique with a novel fitness function and aggregated with for Radial Basis Function Neural Network (RBFN) for credit score classification. Further, proposed feature selection approach is aggregated with Support Vector Machine (SVM) & Random Forest (RF), and other optimization approaches namely: Hybrid Particle Swarm Optimization and Gravitational Search Algorithm (PSOGSA), Hybrid Particle Swarm Optimization and Genetic Algorithm (PSOGA), Improved Krill Herd (IKH), Improved Cuckoo Search (ICS), Firefly Algorithm (FF) and Differential Evolution (DE) are also applied for comparative analysis. |
| 15 | @article{Guo2018ANM,  title={A novel multi-objective particle swarm optimization for comprehensible credit scoring},  author={Yan Guo and Jing He and Libo Xu and Wei Liu},  journal={Soft Computing},  year={2018},  volume={23},  pages={9009 - 9023}, | [A novel multi-objective particle swarm optimization for comprehensible credit scoring](https://www.semanticscholar.org/paper/A-novel-multi-objective-particle-swarm-optimization-Guo-He/7d50cb8255d8cc8c6a6dee9893bb4378718d0cce)  Credit scoring is an important tool for banks and financial institutions to measure credit risk. Linear discriminant analysis (LDA) which according to the score of each credit applicant categorizes these applicants by a cutoff is a comprehensible and robust method in the credit scoring domain. This work presents a novel multi-objective particle swarm optimization for credit scoring (MOPSO-CS), and MOPSO-CS focuses on enhancing credit scoring models based on LDA in three aspects: (i) to construct a higher accuracy credit scoring model which is easy to be interpreted; (ii) to find the most suitable cutoff for discriminating “good credit” customers and “bad credit” customers; and (iii) to improve the sensitivity of the classifier by using multi-objective particle swarm optimization. Finally, through the experiments with two real-world data sets and two benchmark data sets, our proposed MOPSO-CS is compared with 11 counterparts: NaiveBayes, LR, SVM, ANN, DT, CART, bagging-DT, bagging-ANN, RF, MC2 and XGBoost, the results of experiments demonstrate MOPSO-CS outperforms the above-mentioned counterparts in term of sensitivity while maintaining an acceptable accuracy rate. |
| 16 | @article{en2020AHB,  title={A Hybrid Bi-level Metaheuristic for Credit Scoring},  author={Doruk Şen and Cem Çağrı D{\"o}nmez and Umman Mahir Yildirim},  journal={Information Systems Frontiers},  year={2020},  pages={1-11}, | [A Hybrid Bi-level Metaheuristic for Credit Scoring](https://www.semanticscholar.org/paper/A-Hybrid-Bi-level-Metaheuristic-for-Credit-Scoring-%C5%9Een-D%C3%B6nmez/74e9d3839ddd1c8ae9a7443257cf2347727cb9d1)  This research aims to propose a framework for evaluating credit applications by assigning a binary score to the applicant. The score is targeted to determine whether the credit application is ‘good’ or ‘bad’ in small business purpose loans. Even tiny performance improvements in small businesses may yield a positive impact on the economy as they generate more than 60% of the value. The method presented in this paper hybridizes the Genetic Algorithm (GA) and the Support Vector Machine (SVM) in a bi-level feeding mechanism for increased prediction accuracy. The first level is to determine the parameters of SVM and the second is to find a feature set that increases classification accuracy. To test the proposed approach, we have investigated three different data sets; UCI Australian data set for preliminary works, Lending Club data set for large training and testing, and UCI German and Australian datasets for benchmarking against some other notable methods that use GA. Our computational results show that our proposed method using a feedback mechanism under the hybrid bi-level GA-SVM structure outperforms other classification algorithms in the literature, namely Decision Tree, Random Forests, Logistic Regression, SVM and Artificial Neural Networks, effectively improves the classification accuracy. |
| 17 | @article{Dumitrescu2020MachineLO,  title={Machine Learning or Econometrics for Credit Scoring: Let's Get the Best of Both Worlds},  author={Elena Ivona Dumitrescu and Sullivan Hu{\'e} and Christophe Hurlin and Sessi Tokpavi},  journal={Informatics eJournal},  year={2020}, | [Machine Learning or Econometrics for Credit Scoring: Let's Get the Best of Both Worlds](https://www.semanticscholar.org/paper/Machine-Learning-or-Econometrics-for-Credit-Let's-Dumitrescu-Hu%C3%A9/f6607d50c9ff793cd17ee3ae8a46eff60c96a7c9)  Decision trees and related ensemble methods like random forest are state-of-the-art tools in the field of machine learning for credit scoring. Although they are shown to outperform logistic regression, they lack interpretability and this drastically reduces their use in the credit risk management industry, where decision-makers and regulators need transparent score functions. This paper proposes to get the best of both worlds, introducing a new, simple and interpretable credit scoring method which uses information from decision trees to improve the performance of logistic regression. Formally, rules extracted from various short-depth decision trees built with couples of predictive variables are used as predictors in a penalized or regularized logistic regression. By modeling such univariate and bivariate threshold effects, we achieve significant improvement in model performance for the logistic regression while preserving its simple interpretation. Applications using simulated and four real credit defaults datasets show that our new method outperforms traditional logistic regressions. Moreover, it compares competitively to random forest, while providing an interpretable scoring function. JEL Classification: G10 C25, C53 |
| 18 | @article{Agarwal2019FinancialIA,  title={Financial Inclusion and Alternate Credit Scoring: Role of Big Data and Machine Learning in Fintech},  author={Sumit Agarwal and Shashwat Alok and Pulak Ghosh and Sudip Gupta},  journal={FEN: Behavioral Finance (Topic)},  year={2019}, | [Financial Inclusion and Alternate Credit Scoring: Role of Big Data and Machine Learning in Fintech](https://www.semanticscholar.org/paper/Financial-Inclusion-and-Alternate-Credit-Scoring%3A-Agarwal-Alok/8d61e4afc991ed909f4e579c0be9276de28bde90)  We use unique and proprietary data from a large Fintech lender to analyze whether alternative data captured from an individual’s mobile phone (mobile/social footprint) can substitute for traditional credit bureau scores. Variables that measure a borrowers’ digital presence, such as the number and types of apps installed, crude measures of social connections, and measures of borrowers’ “deep social footprints” based on call logs, significantly improve default prediction and outperform the credit bureau score. Using machine learning-based prediction counterfactual analysis, we show that alternate credit scoring based on the mobile and social footprints can expand credit access for individuals who lack credit scores without adversely impacting the default outcomes. Our analysis suggests that the marginal benefit of using alternative data for credit decisions are likely to be higher for borrowers with low levels of income and education, as well as borrowers residing in regions with low levels of financial inclusion. |
| 19 | @article{Tripathi2021ExperimentalAO,  title={Experimental analysis of machine learning methods for credit score classification},  author={Diwakar Tripathi and Damodar Reddy Edla and Annushree Bablani and Alok Kumar Shukla and B. Ramachandra Reddy},  journal={Progress in Artificial Intelligence},  year={2021},  volume={10},  pages={217 - 243}, | [Experimental analysis of machine learning methods for credit score classification](https://www.semanticscholar.org/paper/Experimental-analysis-of-machine-learning-methods-Tripathi-Edla/a8446707770c18d93e07b54fdd9d704946e2bc74)  redit scoring concerns with emerging empirical model to assist the financial institutions for financial decision-making process. Credit risk analysis plays a vital role for decision-making process; statistical and machine learning approaches are utilized to estimate the risk associated with a credit applicant. Enhancing the performance of credit scoring model, particularly toward non-trustworthy “or non-creditworthy” group, may result incredible effect for financial institution. However, credit scoring data may have excess and unimportant data and features which degrades the performance of model. So, selection of important features (or reduction in irrelevant and redundant features) may play the key role for improving the effectiveness and reducing the complexity of the model. This study presents a experimental results analysis of various combinations of feature selection approaches with various classification approaches and impact of feature selection approaches. For experimental results analysis, nine feature selection and sixteen classification state-of-the-art approaches have been applied on seven benched marked credit scoring datasets. |
| 20 | @article{Yu2020DataCF,  title={Data Cleaning for Personal Credit Scoring by Utilizing Social Media Data: An Empirical Study},  author={Xi Yu and Qi Yang and Ruiqi Wang and Runqing Fang and Mingsen Deng},  journal={IEEE Intelligent Systems},  year={2020},  volume={35},  pages={7-15}, | [Data Cleaning for Personal Credit Scoring by Utilizing Social Media Data: An Empirical Study](https://www.semanticscholar.org/paper/Data-Cleaning-for-Personal-Credit-Scoring-by-Social-Yu-Yang/ef35b792861f159ba7d6cbc2e3c578db75f54e82)  With the accumulation of data on personal behavior and the development of machine learning models and algorithms, it is becoming possible to use social media data for personal credit scoring. In this article, we use the systematic sampling method to obtain Douban's social media data. Because there are many abnormal users in these data, they are “real but false data” for personal credit evaluation. In order to better carry out personal credit scoring, we propose three criteria, power exponents of time interval distribution of individual user $\gamma \_i$γi, user activity $A\_i$Ai, and the ratio of out-degree and in-degree $R\_i$Ri of user $i$i, which are used to systematically clean the data. And then, we used the logistic regression method to score the individual credits of users before and after data cleaning, and found that the rank order of personal credit scoring has changed significantly. This change is largely attributed to the changes of network structure after data cleaning. We believe that our work is very important to use the social media data to establish a credible personal credit evaluation system to reduce the credit risk of the current Internet financial industry. |
| 21 | @inproceedings{Dushimimana2020UseOM,  title={Use of Machine Learning Techniques to Create a Credit Score Model for Airtime Loans},  author={Bernard Dushimimana and Yvonne Wambui and Timothy Lubega and Patrick E. McSharry},  year={2020}, | [Use of Machine Learning Techniques to Create a Credit Score Model for Airtime Loans](https://www.semanticscholar.org/paper/Use-of-Machine-Learning-Techniques-to-Create-a-for-Dushimimana-Wambui/c62be3ea7821e5140c6ef635f4b1176b6e84afc9)  Airtime lending default rates are typically lower than those experienced by banks and microfinance institutions (MFIs) but are likely to grow as the service is offered more widely. In this paper, credit scoring techniques are reviewed, and that knowledge is built upon to create an appropriate machine learning model for airtime lending. Over three million loans belonging to more than 41 thousand customers with a repayment period of three months are analysed. Logistic Regression, Decision Trees and Random Forest are evaluated for their ability to classify defaulters using several cross-validation approaches and the latter model performed best. When the default rate is below 2%, it is better to offer everyone a loan. For higher default rates, the model substantially enhances profitability. The model quadruples the tolerable level of default rate for breaking even from 8% to 32%. Nonlinear classification models offer considerable potential for credit scoring, coping with higher levels of default and therefore allowing for larger volumes of customers. |  |
| 22 | @article{Lin2022AnEP,  title={An Efficient Privacy-Preserving Credit Score System Based on Noninteractive Zero-Knowledge Proof},  author={Chao Lin and Min Luo and Xinyi Huang and Kim-Kwang Raymond Choo and De-biao He},  journal={IEEE Systems Journal},  year={2022},  volume={16},  pages={1592-1601}, | [An Efficient Privacy-Preserving Credit Score System Based on Noninteractive Zero-Knowledge Proof](https://www.semanticscholar.org/paper/An-Efficient-Privacy-Preserving-Credit-Score-System-Lin-Luo/c1d48197c6f6203f72ab52a2e71170c0e2179b73)  Credit system is generally associated with the banking and financial institutions, although it has far reaching implications for residents of countries, such as U.S., particularly for those with a poor credit history. Specifically, a credit score computation (CSC) quantifies an individual’s credit value or credit risk, which is used by banking and financial institutions, as well as other entities (e.g., during purchasing of insurance policies and application of rental properties), to facilitate their decision-making (e.g., whether to approve the insurance policy purchase or the level of premium). Although a number of CSC models have been proposed in the literature for supporting different application scenarios, privacy protection of CSC is rarely considered despite the potential for leakage of user private information (e.g., registration, hobbies, credit, relationships, and inquiry). Such information can then be abused for other nefarious activities, such as identity theft and credit card fraud. Thus, in this article, we first analyze the privacy strength of existing CSC models, prior to presenting the formal definition of a privacy-preserving CSC system alongside its security requirements. Then, we propose a concrete construction based on Paillier encryption with three proposed noninteractive zero-knowledge schemes. To demonstrate feasibility of our proposal, we evaluate both its security and performance. |
| 23 | @article{Prassanna2020SecrecyProtectorAN,  title={SecrecyProtector: A Novel Data Analytics based credit score management system},  author={J. Prassanna and Abdul Quadir and J ChristyJackson and R. Prabakaran and Sakkaravarthi Ramanathan},  journal={International Journal of Scientific \& Technology Research},  year={2020},  volume={9},  pages={1215-1218}, | [SecrecyProtector: A Novel Data Analytics based credit score management system](https://www.semanticscholar.org/paper/SecrecyProtector%3A-A-Novel-Data-Analytics-based-Prassanna-Quadir/cc222e00f8cca5eeab4d10e31754987c2565f4ec)  This work gives an account of the Credit Score web service application and the primary purpose of a credit score is to help lenders assess individuals' risk of not repaying a loan. Credit scoring assessment, despite the fact that a moderately new idea in the Indian money related business sector, have increased wide acknowledgment among financial specialists. In the meantime, easy-going and narrative confirmation recommends that there are worries among speculators and controllers about the execution of rating offices in India. This paper looks at financial specialists' mindfulness, discernment, understanding level and use of Credit scoring assessment through a poll-based example overview covering individual and additionally institutional speculators. We find high dissemination of rating use among all class of financial specialists, however, there is a recognizable upsetting with the dependability of appraisals, inclination of ensuing minimizing and opportuneness of rating reconnaissance. The review additionally uncovers that the institutional financial specialists have predominant information and comprehension about evaluations than individual speculators. In this way, the review underlines the requirement for rating offices to take a shot at instructing the basic speculators to engender appropriate comprehension and use of Credit |
| 24 | @article{Rajaleximi2019FeatureSU,  title={Feature Selection using Optimized Multiple Rank Score Model for Credit Scoring},  author={P.Ramila Rajaleximi and Mohammed Zaheer Ahmed and Ahmed Alenezi},  journal={International Journal of Intelligent Engineering and Systems},  year={2019}, | [Feature Selection using Optimized Multiple Rank Score Model for Credit Scoring](https://www.semanticscholar.org/paper/Feature-Selection-using-Optimized-Multiple-Rank-for-Rajaleximi-Ahmed/9a400564a0578fe31242a4129c7bf5ceb35fcde1)  Fraud identification and prevention is becoming the decisive challenge for any financial institutions and financial service industries. Majority of the banks use a credit score, a numerical or statistical value to estimate the customer’s creditworthiness based on their credit history. A credit score model will be built by employing training dataset and further, an analytical process will be conceded to estimate the credit score of each customer. Thus, for any credit score model, the lenders acquire various data about the customer from various external agencies. The collected data may encompass irrelevant parameters which will not help in making any decisions and also decelerates the global performance of the model. Accordingly, feature selection is an imperative process to eliminate less relevant attributes for any dataset, especially in credit scoring. This paper emphasizes on enlightening the performance of the attribute selection process using multiple rank score model. The proposed method accumulates the results using the optimized threshold algorithm and outperforms well in selecting the quality attribute for the underlying credit score model. The experimental evaluation has been carried out and it is proved that the accuracy and performance of the suggested method are comparatively better than the existing single rank techniques. |
| 25 | @article{Adisa2022CreditSP,  title={Credit Score Prediction using Genetic Algorithm-LSTM Technique},  author={Juliana Adeola Adisa and Samuel Olusegun Ojo and Pius Adewale Owolawi and Agnieta Pretorius and S. Ojo},  journal={2022 Conference on Information Communications Technology and Society (ICTAS)},  year={2022},  pages={1-6}, | [Credit Score Prediction using Genetic Algorithm-LSTM Technique](https://www.semanticscholar.org/paper/Credit-Score-Prediction-using-Genetic-Technique-Adisa-Ojo/7f3e79785fd7d78de811f83f52aad512b30ceb77)  In data mining, the goal of prediction is to develop a more effective model that can provide accurate results. Prior literature has studied different classification techniques and found that combining multiple classifiers into ensembles outperformed most single classifier approaches. The performance of an ensemble classifier can be affected by some factors. How to determine the best classification technique? Which combination method to employ? This paper applies Long Short-Term Memory (LSTM), one of the most advanced deep learning algorithms which are inherently appropriate for the financial domain but rarely applied to credit scoring prediction. The research presents an optimization approach to determine the optimal parameters for a deep learning algorithm. The LSTM parameters are determined using an optimization algorithm. The LSTM parameters include epochs, batch size, number of neurons, learning rate and dropout. The results show that the optimized LSTM model outperforms both single classifiers and ensemble models. |
| 26 | @article{en2020AHB,  title={A Hybrid Bi-level Metaheuristic for Credit Scoring},  author={Doruk Şen and Cem Çağrı D{\"o}nmez and Umman Mahir Yildirim},  journal={Information Systems Frontiers},  year={2020},  volume={22},  pages={1009 - 1019}, | [A Hybrid Bi-level Metaheuristic for Credit Scoring](https://www.semanticscholar.org/paper/A-Hybrid-Bi-level-Metaheuristic-for-Credit-Scoring-%C5%9Een-D%C3%B6nmez/74e9d3839ddd1c8ae9a7443257cf2347727cb9d1)  This research aims to propose a framework for evaluating credit applications by assigning a binary score to the applicant. The score is targeted to determine whether the credit application is ‘good’ or ‘bad’ in small business purpose loans. Even tiny performance improvements in small businesses may yield a positive impact on the economy as they generate more than 60% of the value. The method presented in this paper hybridizes the Genetic Algorithm (GA) and the Support Vector Machine (SVM) in a bi-level feeding mechanism for increased prediction accuracy. The first level is to determine the parameters of SVM and the second is to find a feature set that increases classification accuracy. To test the proposed approach, we have investigated three different data sets; UCI Australian data set for preliminary works, Lending Club data set for large training and testing, and UCI German and Australian datasets for benchmarking against some other notable methods that use GA. Our computational results show that our proposed method using a feedback mechanism under the hybrid bi-level GA-SVM structure outperforms other classification algorithms in the literature, namely Decision Tree, Random Forests, Logistic Regression, SVM and Artificial Neural Networks, effectively improves the classification accuracy. |
| 27 | @article{Dumitrescu2020MachineLO,  title={Machine Learning or Econometrics for Credit Scoring: Let's Get the Best of Both Worlds},  author={Elena Ivona Dumitrescu and Sullivan Hu{\'e} and Christophe Hurlin and Sessi Tokpavi},  journal={Informatics eJournal},  year={2020}, | [Machine Learning or Econometrics for Credit Scoring: Let's Get the Best of Both Worlds](https://www.semanticscholar.org/paper/Machine-Learning-or-Econometrics-for-Credit-Let's-Dumitrescu-Hu%C3%A9/f6607d50c9ff793cd17ee3ae8a46eff60c96a7c9)  Decision trees and related ensemble methods like random forest are state-of-the-art tools in the field of machine learning for credit scoring. Although they are shown to outperform logistic regression, they lack interpretability and this drastically reduces their use in the credit risk management industry, where decision-makers and regulators need transparent score functions. This paper proposes to get the best of both worlds, introducing a new, simple and interpretable credit scoring method which uses information from decision trees to improve the performance of logistic regression. Formally, rules extracted from various short-depth decision trees built with couples of predictive variables are used as predictors in a penalized or regularized logistic regression. By modeling such univariate and bivariate threshold effects, we achieve significant improvement in model performance for the logistic regression while preserving its simple interpretation. Applications using simulated and four real credit defaults datasets show that our new method outperforms traditional logistic regressions. Moreover, it compares competitively to random forest, while providing an interpretable scoring function. JEL Classification: G10 C25, C53 |
| 28 | @article{Tripathi2021ExperimentalAO,  title={Experimental analysis of machine learning methods for credit score classification},  author={Diwakar Tripathi and Damodar Reddy Edla and Annushree Bablani and Alok Kumar Shukla and B. Ramachandra Reddy},  journal={Progress in Artificial Intelligence},  year={2021},  volume={10},  pages={217 - 243}, | [Experimental analysis of machine learning methods for credit score classification](https://www.semanticscholar.org/paper/Experimental-analysis-of-machine-learning-methods-Tripathi-Edla/a8446707770c18d93e07b54fdd9d704946e2bc74)  Credit scoring concerns with emerging empirical model to assist the financial institutions for financial decision-making process. Credit risk analysis plays a vital role for decision-making process; statistical and machine learning approaches are utilized to estimate the risk associated with a credit applicant. Enhancing the performance of credit scoring model, particularly toward non-trustworthy “or non-creditworthy” group, may result incredible effect for financial institution. However, credit scoring data may have excess and unimportant data and features which degrades the performance of model. So, selection of important features (or reduction in irrelevant and redundant features) may play the key role for improving the effectiveness and reducing the complexity of the model. This study presents a experimental results analysis of various combinations of feature selection approaches with various classification approaches and impact of feature selection approaches. For experimental results analysis, nine feature selection and sixteen classification state-of-the-art approaches have been applied on seven benched marked credit scoring datasets. |
| 29 | @article{Yu2020DataCF,  title={Data Cleaning for Personal Credit Scoring by Utilizing Social Media Data: An Empirical Study},  author={Xi Yu and Qi Yang and Ruiqi Wang and Runqing Fang and Mingsen Deng},  journal={IEEE Intelligent Systems},  year={2020},  volume={35},  pages={7-15}, | [Data Cleaning for Personal Credit Scoring by Utilizing Social Media Data: An Empirical Study](https://www.semanticscholar.org/paper/Data-Cleaning-for-Personal-Credit-Scoring-by-Social-Yu-Yang/ef35b792861f159ba7d6cbc2e3c578db75f54e82)  With the accumulation of data on personal behavior and the development of machine learning models and algorithms, it is becoming possible to use social media data for personal credit scoring. In this article, we use the systematic sampling method to obtain Douban's social media data. Because there are many abnormal users in these data, they are “real but false data” for personal credit evaluation. In order to better carry out personal credit scoring, we propose three criteria, power exponents of time interval distribution of individual user $\gamma \_i$γi, user activity $A\_i$Ai, and the ratio of out-degree and in-degree $R\_i$Ri of user $i$i, which are used to systematically clean the data. And then, we used the logistic regression method to score the individual credits of users before and after data cleaning, and found that the rank order of personal credit scoring has changed significantly. This change is largely attributed to the changes of network structure after data cleaning. We believe that our work is very important to use the social media data to establish a credible personal credit evaluation system to reduce the credit risk of the current Internet financial industry. |
| 30 | @article{Wu2021ApplicationAO,  title={Application Analysis of Credit Scoring of Financial Institutions Based on Machine Learning Model},  author={Yi Wu and Yuwen Pan},  journal={Complex.},  year={2021},  volume={2021},  pages={9222617:1-9222617:12}, | [Application Analysis of Credit Scoring of Financial Institutions Based on Machine Learning Model](https://www.semanticscholar.org/paper/Application-Analysis-of-Credit-Scoring-of-Financial-Wu-Pan/1427ca5dbd0feba4123f3f95d6ccd8399642cbac)  redit score is the basis for financial institutions to make credit decisions. With the development of science and technology, big data technology has penetrated into the financial field, and personal credit investigation has entered a new era. Personal credit evaluation based on big data is one of the hot research topics. This paper mainly completes three works. Firstly, according to the application scenario of credit evaluation of personal credit data, the experimental dataset is cleaned, the discrete data is one-HOT coded, and the data are standardized. Due to the high dimension of personal credit data, the pdC-RF algorithm is adopted in this paper to optimize the correlation of data features and reduce the 145-dimensional data to 22-dimensional data. On this basis, WOE coding was carried out on the dataset, which was applied to random forest, support vector machine, and logistic regression models, and the performance was compared. It is found that logistic regression is more suitable for the personal credit evaluation model based on Lending Club dataset. Finally, based on the logistic regression model with the best parameters, the user samples are graded and the final score card is output. |