**Loan full schema data set**

1. Data exploration

* Về tập dữ liệu:
* Bộ dữ liệu gồm 55 cột và 10000 dòng.
* Một số cột có dữ liệu null
* Không cột nào bị sai kiểu dữ liệu.
* Xử lý dữ liệu null:
* Xóa bỏ các cột có trên 50% giá trị là null
* Giá trị null trong cột nghề nghiệp được fill bằng ‘unemployed’
* Giá trị null trong các cột dạng số khác fill bằng trung vị
* Xóa cột 'num\_accounts\_120d\_past\_due' do chỉ chứa 1 giá trị
* Do mục tiêu của bài là dự đoán lãi suất nên các cột xuất hiện sau khi biết lãi suất sẽ bị xóa.

1. Data visualization

* Kiểm tra tương quan:
* Các cột dữ liệu dạng số không có mối tương quan chặt chẽ với tỷ lệ lãi suất, cao nhất là paid\_interest chỉ hơn 0,35.
* Một số cột có mối tương quan rất lớn với nhau, trong khi không có mối tương quan rõ ràng với tỷ lệ lãi suất có thể được hợp nhất hoặc xóa.
* Vì hầu hết các mối tương quan với lãi suất quá nhỏ, ngoài 2 yếu tố có tương quan lớn nhất, các yếu tố liên quan trực tiếp đến lãi suất như: annual\_income, loan\_amount, sẽ được giữ lại để đưa vào mô hình.
* Distribution plot

|  |  |
| --- | --- |
| Chart, histogram  Description automatically generated | Hầu hết các giá trị của lãi suất là thấp hơn 20%. Phổ biến nhất là khoảng 10%. Các giá trị quá lớn có thể là outliers và có thể cân nhắc loại bỏ nếu ảnh hưởng đến kết quả dự đoán. |

* Box plot:

|  |  |
| --- | --- |
| Chart, line chart, box and whisker chart  Description automatically generated | Không có ngoại lệ đáng kể nào được ghi nhận. Các khoản vay tiền mặt trung bình có tổng số tiền nhỏ hơn các khoản vay DirectPay nhưng không đáng kể, cho thấy tổ chức cho vay tương đối uy tín và các bên ít lo lắng về tính bảo mật của các hình thức thanh toán khác nhau. |
|  | Chủ nhà không giàu hơn người đi thuê nhà, và thậm chí còn nghèo hơn người thế chấp. Đa số người vay có thu nhập bình quân dưới 100.000. Tuy nhiên, có rất nhiều người có thu nhập cao (so với những người còn lại) vẫn lựa chọn hình thức vay tại cơ sở này. |
| Chart, box and whisker chart  Description automatically generated | Các khoản cho vay thu nhập chưa được xác minh thường có lãi suất thấp hơn các khoản đã được xác minh. Tuy nhiên, có thể chỉ ra rằng trạng thái xác minh dường như có mối quan hệ với lãi suất. |
| Hình thức sở hữu nhà và phương thức cho vay dường như không ảnh hưởng nhiều đến lãi suất. | |
|  | Kỳ hạn vay càng dài thì lãi suất càng lớn |
|  | Lãi suất của những khoản vay bằng tiền mặt thường lớn hơn rõ rệt so với các khoản DirectPay. |

* + Qua các boxplot có thể thấy các lãi suất từ trên 25% có thể coi là outliers. Trước tiên, tiến hành loại bỏ các hàng có lãi suất >30% và đo độ chính xác của mô hình dự đoán.
  + Các cột verified\_income, disbursement\_method, term có thể đưa vào mô hình
* Scatter plot:

|  |  |
| --- | --- |
|  | Rõ ràng có một mối quan hệ tuyến tính chặt chẽ giữa xếp hạng và lãi suất. Xếp hạng khoản vay càng tốt thì lãi suất trung bình càng nhỏ. |
|  | Có vẻ như có một mối quan hệ tuyến tính giữa mục đích vay và lãi suất. |
| Công việc vả địa chỉ của người đi vay không ảnh hưởng nhiều đến lãi suất. | |

* + Các cột subgrade, loan\_purpose có thể đưa vào mô hình

1. Model development

Bộ features được sử dụng để dự đoán lãi suất bao gồm: annual\_income, loan\_amount, term, total\_debit\_limit, verified\_income, disbursement\_method, sub\_grade, loan\_purpose. Dữ liệu được chia thành tập train và test (75%-25%) và scaled để giảm mức độ ảnh hưởng khác nhau do thang đo khác nhau của các cột.

* Linear Regression Model: Mô hình dự đoán giá trị y dựa trên giá trị xi sẵn có theo một hàm tuyến tính. (y=f(xi))

Kết quả so sánh giá trị dự đoán và giá trị thực tế được thể hiện ở biểu đồ:

Chart, scatter chart

Description automatically generated

Chart, line chart, histogram

Description automatically generated

Train Test

Thời gian chạy mô hình là: 18.0 microseconds

Train score: 0.9978

Test score: 0.9812

* SVR: dựa trên thuật toán SVM (thuật toán sử dụng vector để phân lớp dữ liệu) nhưng được áp dụng trong bài toán hồi quy.

Kết quả so sánh giá trị dự đoán và giá trị thực tế được thể hiện ở biểu đồ:

Chart, scatter chart

Description automatically generated

Chart, histogram

Description automatically generated

Train Test

Thời gian chạy mô hình là: 981.3 microseconds

Train score: 0.9979

Test score: 0.9778

**Kết quả:** cả hai mô hình đều cho kết quả khá tốt ở cả tập train và tập test. Tuy nhiên, mô hình Linear Regression tối ưu hơn về cả thời gian thực hiện và kết quả trên tập test.