International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 8 Issue 11 November - 2024, Pages: 25-30

Predict personal customer credit risk using Linear regression algorithm for AI program

Trinh Thanh Do1 and Nguyen Ha Chi2

1Software Development Manager, Rikai Min Smart , Email: do.trinh@rikai.technology
2Ngo Quyen High School, Hai Phong, Vietnam, Email: nguyenhachi22052007@gmail.com

***Abstract**: As financial institutions increasingly seek to manage credit risk effectively, the use of predictive models leveraging artificial intelligence (AI) and machine learning has gained significant traction. This report examines the implementation of a Linear Regression model to predict individual customer credit risk, incorporating key features such as age, income, loan amount, and credit score. The findings indicate that while the Linear Regression model can conduct basic analyses related to the correlation between these variables and credit risk, its limitations become apparent in situations involving non-linear relationships. To assess the accuracy of the model, performance metrics such as accuracy and F1 score were used. Furthermore, the report provides recommendations for improvement, including the application of Regularization, innovative feature generation techniques, and exploring other non-linear models to increase predictive performance.*

***Keywords**: Credit Risk Prediction, Linear Regression, Machine Learning, Financial Data, Credit Scoring*

I. INTRODUCTION

1.1 Research objectives

Building an AI model using Linear Regression algorithm to predict credit risk of individual customers, thereby supporting financial institutions in assessing and managing credit risk.

The research will focus on analyzing important factors affecting customers' ability to pay, and evaluating the effectiveness of the prediction model through accuracy, sensitivity and specificity indicators.

The research results promise to bring practical applications in minimizing credit risk, automating and optimizing the credit granting process.

1.2 Context and significance of the issue

In the context of the financial industry increasingly focusing on risk management, predicting individual customer credit risk is an important task to help financial institutions minimize losses and maintain stability.

With the development of AI, the application of advanced prediction models helps to assess customers' payment ability more quickly and accurately.

This research not only supports credit decisions but also contributes to optimizing risk management processes, improving operational efficiency and protecting profits for financial institutions.

1.3 Reason for choosing the topic

The reason for choosing the topic “Predicting credit risk of individual customers using AI” is because credit risk is always one of the major challenges for financial institutions.

Using traditional methods to assess credit no longer meets the requirements of accuracy and speed in the context of increasingly complex and large data.

AI, especially machine learning algorithms, is capable of analyzing and predicting risk factors more accurately, thereby helping to minimize financial risks.

In addition, with the strong development of technology, the application of AI in the financial sector is becoming more and more practical and meaningful, especially in optimizing credit decisions, improving customer experience and enhancing work efficiency.

II. THEORETICAL OVERVIEW

2.1 Concept of credit risk

Credit risk is the possibility that a customer or organization will not be able to meet its financial obligations as committed in a credit contract.

This is an important factor affecting financial institutions when they decide to extend credit to customers, because non-payment can lead to serious financial loss.

Credit risk can arise from many factors, including the customer's weak financial condition, changes in the economic environment, or ineffective credit management.

Financial institutions often use tools such as credit assessment, payment history, and predictive models to determine the risk level of each customer.

Good credit risk management helps protect profits and maintain stability for financial institutions.

2.2 Current credit risk assessment methods

Current credit risk assessment methods are mainly based on analyzing financial and non-financial factors of customers.

Financial institutions use credit scoring systems to assess customers' ability to repay debts, in which factors such as credit history, income, debt ratio, and financial capacity are carefully considered.

In addition, traditional methods also include analyzing financial statements and economic indicators to predict payment ability. However, these methods have disadvantages in accuracy when faced with large and complex data volumes. Therefore, today, many financial institutions have begun to apply machine learning and AI models to improve their ability to predict and manage credit risks more effectively.

2.3 Applications of AI and Machine Learning in Finance

The application of AI and Machine Learning in finance is becoming increasingly popular thanks to its ability to process and analyze large volumes of data quickly and accurately.

Machine learning algorithms help detect hidden data patterns, thereby predicting market trends, assessing credit risk, and detecting fraud.

AI is also used to optimize investment strategies, manage portfolios, and analyze customer behavior to provide suitable financial products.

Through deep learning models, AI can significantly improve prediction accuracy and minimize risks in financial decisions.

In addition, AI also helps automate transaction processes and provide personalized financial services, improving customer experience.

III. METHODOLOGY

3.1 Description of data used for research

The data used in this study includes important characteristics that help assess the creditworthiness of individual customers. The data columns include:

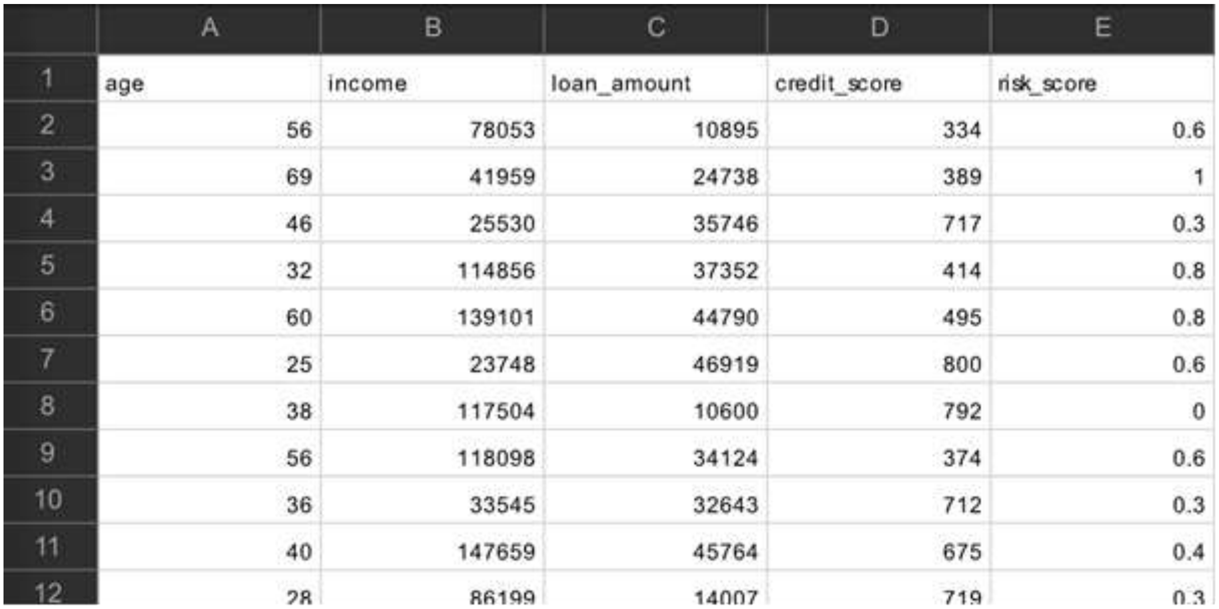
Age: Represents the age of the customer, a factor that can affect their ability to repay debt and their financial experience.

Income: The monthly or annual income of the customer, which helps determine the financial capacity and ability to repay loans.

Loan Amount: The amount of money that the customer wants to borrow, often related to the level of credit risk compared to their financial capacity.

Credit Score: The customer's credit score, an important indicator reflecting the credit history and the level of reliability in repaying debts.

This data provides the basis for building a credit risk prediction model, which helps assess the customer's ability to repay based on personal financial factors.



	A	B	C	D	E
1	age	income	loan_amount	credit_score	risk_score
2	56	78053	10895	334	0.6
3	69	41959	24738	389	1
4	46	25530	35746	717	0.3
5	32	114856	37352	414	0.8
6	60	139101	44790	495	0.8
7	25	23748	46919	800	0.6
8	38	117504	10600	792	0
9	56	118098	34124	374	0.6
10	36	33545	32643	712	0.3
11	40	147659	45764	675	0.4
12	28	86199	14007	719	0.3

3.2 Data processing and preprocessing

Data processing and preprocessing is an important step in preparing data so that a machine learning model can learn and predict effectively. For your project using Python with Scikit-learn (Sklearn), this process might include the following steps:

3.2.1 Handling Missing Data:

- First, check the data columns to determine if there are any missing values (e.g., NaN or None). For missing values, you can decide to:
- Fill in the values: Use methods such as filling in the average, filling in the median, or filling in the common value.
- Remove rows or columns containing missing values: If the missing data is too numerous or not important, the corresponding row or column can be removed.

Example

```
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(strategy='mean') # Fill in missing values with average
data = imputer.fit_transform(data)
```

3.2.2 Data Type Conversion:

Ensure that columns have the appropriate data types for the machine learning model. For example, columns like Age, Income, Loan Amount, and Credit Score are often float or integer types, so check and convert data types as needed.

3.2.3 Normalization and Standardization:

Machine learning algorithms typically perform better when features have a similar range of values. Normalization helps ensure that the values in columns like Income or Loan Amount do not vary too much, causing the model to focus on features with larger values.

Standardization: Often used when features are normally distributed. The data will be transformed so that the mean = 0 and the standard deviation = 1.

Normalization: Often used when data is not normally distributed. The data will be transformed to the range [0, 1].

Standardization example:

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
data_scaled = scaler.fit_transform(data)
```

3.2.4 Categorical Encoding:

If you have categorical columns (although in your example you only have numeric data), you will need to encode them into numeric form so that the model can understand them. For example, use One-Hot Encoding or Label Encoding.

3.2.5 Data Splitting:

After processing the data, you will need to split the data into a training set and a test set to evaluate the model. Usually, a ratio of 70%/30% or 80%/20% is used.

Data Splitting example:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(data, target, test_size=0.2, random_state=42)
```

3.2.6 Outliers handle:

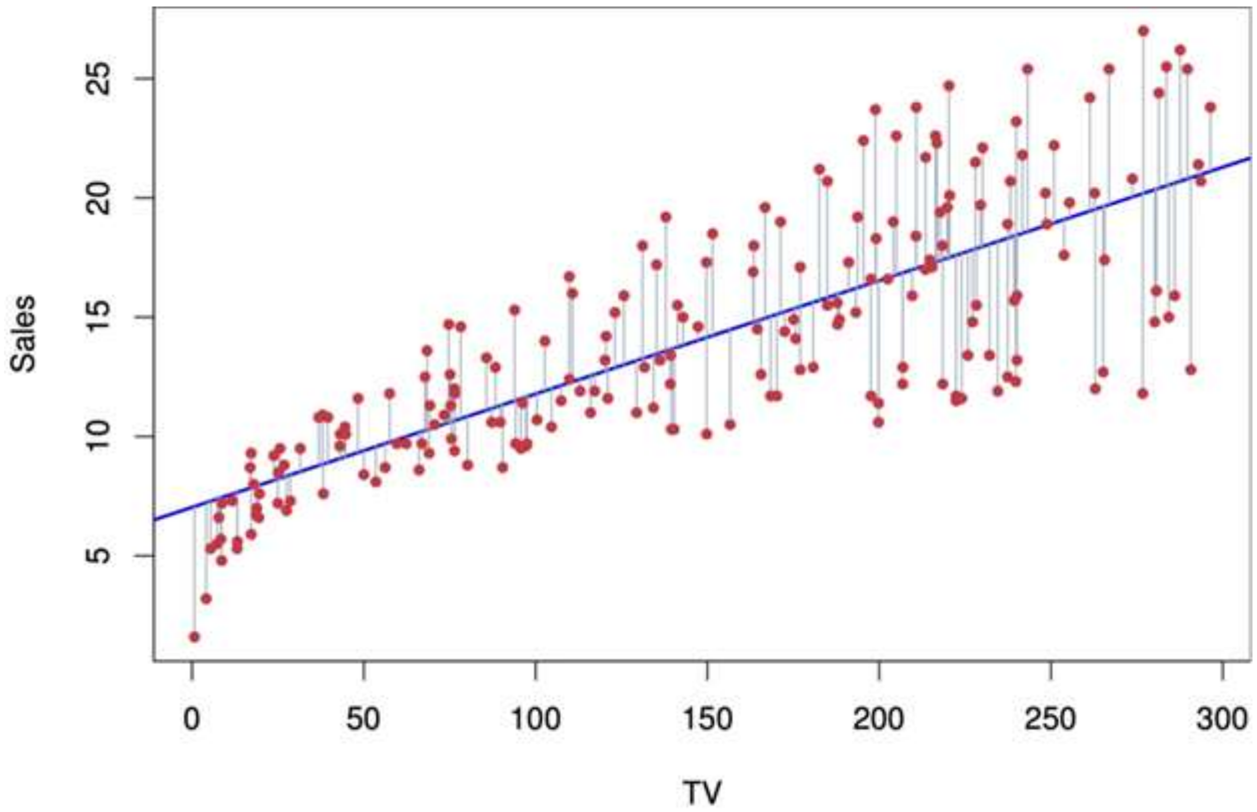
Outliers can skew model results. You can detect and remove or adjust outliers as needed.

3.3 Credit risk prediction models:

3.3.1 Introduction to Linear Regression

Linear Regression is a simple and powerful machine learning algorithm used to predict a continuous target variable based on input features.

In credit risk prediction, Linear Regression can help determine the likelihood of a customer repaying a loan based on factors such as income, credit score, and loan amount.



3.3.2 How Linear Regression Works

Linear Regression builds a linear model between independent and dependent variables, to find the optimal coefficients for the input features. This model uses least squares to optimize the model parameters, helping to make the most accurate predictions.

3.3.3 Advantages of Linear Regression in Credit Risk Prediction

One of the great advantages of Linear Regression is its simplicity and ease of interpretation, making it easy for financial institutions to understand and apply in assessing credit risk. This model can also handle linear relationships between factors such as income and credit score well.

3.3.4 Disadvantages of Linear Regression

Although Linear Regression is easy to implement, it only works for linear relationships between factors. If the data has a nonlinear relationship or there are too many complex interacting factors, this model may not yield accurate results.

3.3.5 Data Characteristics Suitable for Linear Regression

Data suitable for Linear Regression usually has a clear linear relationship between the features and the target variable. Features such as age, income, and credit score are examples of data that are easily amenable to Linear Regression.

3.3.6 Model training and optimization process

During training, the Linear Regression model uses the square root minimization method to adjust the coefficients of the input features. The final result is a linear function that can predict the probability of a customer's credit risk.

3.3.7 Evaluating the effectiveness of Linear Regression model

Linear Regression models are evaluated using metrics such as R-squared (to measure how well the model fits the data) and predictive accuracy. Testing methods such as cross-validation can help determine the stability and performance of the model.

3.3.8 Practical Applications of Linear Regression in Finance

Linear Regression is widely used in banks and financial institutions to assess the creditworthiness of customers. This model helps financial institutions make more accurate credit decisions and minimize risks.

3.3.9 Optimizing and improving Linear Regression models

Although Linear Regression is a simple model, it can be improved by combining it with techniques such as Regularization (Lasso, Ridge) to reduce overfitting and improve the generalization ability of the model.

3.3.10 Comparing Linear Regression with other models

Although Linear Regression has advantages in simplicity and interpretability, when compared to more complex models such as Random Forests or SVM, it can be less effective in cases where the data is non-linear or has too many variables.

3.4 Model effectiveness evaluation criteria

3.4.1 Accuracy

Accuracy is a basic metric used to measure the ratio of correct predictions to the total number of predictions. In credit risk prediction, accuracy provides an overall assessment of the model's predictive ability. However, in cases where the data is imbalanced (e.g., the number of defaulting customers is very small compared to the total number of customers), accuracy may not fully reflect the performance of the model.

3.4.2 Sensitivity

Sensitivity, also known as Detection Rate, measures the model's ability to detect high credit risk cases (groups of customers who are unlikely to pay). High sensitivity means that the model does not miss potentially risky cases, which is important in minimizing losses for financial institutions.

3.4.3 Specificity

Specificity measures the ability of the model to correctly identify credit-risk-free customers (the group of customers who are likely to repay their debts). This is an important metric in ensuring that the model does not generate too many false alarms against creditworthy customers, reducing unnecessary costs for the financial institution.

3.4.4 F1-Score

F1-Score is a combination of sensitivity and accuracy, which is especially useful when the data is imbalanced. F1-Score provides an overview of how well the model balances between correctly detecting risky cases and not missing them. A high F1-Score indicates that the model is able to accurately predict both groups of customers.

3.4.5 Area Under the ROC Curve (AUC-ROC)

AUC-ROC is an important metric to evaluate the classification ability of a model, especially in binary classification problems such as credit risk. The ROC Curve plots the difference between the True Positive Rate and the False Positive Rate. The AUC shows the area under the ROC curve and the closer it is to 1, the better the model. This metric helps evaluate the model's ability to distinguish between risky and non-risky customers.

IV. RESEARCH PROCESS

```
# Import the necessary libraries
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
import matplotlib.pyplot as plt
import seaborn as sns

# Load the data (assuming you have a CSV file containing the data)
# Replace 'credit_risk_data.csv' with the path to your data file
data = pd.read_csv('credit_risk_data.csv')

# Display basic information about the data
print(data.head())
print(data.info())

# Preprocess the data (if needed)
# For example, handle missing values, encode categorical variables, etc.

# Split the data into input (X) and output (y)
# Assume the 'risk_score' column is the target variable

X = data.drop('risk_score', axis=1)
y = data['risk_score']

# Split the data into training set and test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize Linear Regression model
model = LinearRegression()

# Model training
model.fit(X_train, y_train)

# Prediction on test set
y_pred = model.predict(X_test)

# Model Evaluation
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f'Mean Squared Error: {mse}')
print(f'R^2 Score: {r2}')

# Visualize results
plt.figure(figsize=(10, 6))
sns.scatterplot(x=y_test, y=y_pred)
plt.xlabel('Actual Risk Score')
plt.ylabel('Predicted Risk Score')
plt.title('Actual vs Predicted Risk Score')
plt.show()
```

V. DISCUSSION AND EVALUATION

5.1 Advantages and disadvantages of Linear Regression

Linear Regression is a simple, easy-to-understand and easy-to-implement algorithm. Since this model is based only on minimizing the error, it does not require much computational resources, suitable for simple problems with linear relationships.

The results of Linear Regression are easy to understand and allow financial institutions to clearly see the influence of each feature (such as income, credit score) on the prediction. This supports decisions related to credit and risk management.

Linear Regression gives good results when the input variables have a linear relationship with the output variable. In credit risk prediction, if factors such as income and age have a clear relationship with the ability to repay the loan, this model can bring high accuracy.

One of the major limitations of Linear Regression is its limited ability to handle non-linear relationships. If the data is non-linear or has many interacting factors, this model may not be able to learn complex relationships, leading to poor predictions.

Linear Regression is very sensitive to outliers, as these points can distort the regression coefficients. Therefore, before using this model, outliers need to be handled to ensure the accuracy of the prediction.

In complex cases such as high-dimensional data or many samples, Linear Regression is often not accurate enough to analyze and predict. More complex models such as Random Forest or Neural Network often give better results in these cases.

Evaluate the model according to the criteria: MAE, MSE, RMSE, R².

Experiment with other types of algorithms such as Decision Trees, Random Forest, Support Vector Machines, K-Nearest Neighbors

VI. CONCLUSION AND RECOMMENDATIONS

6.1 Summary of key findings

During the research, the main results showed that the Linear Regression model has certain effectiveness in predicting credit risk of individual customers.

The model has demonstrated its ability to analyze the relationship between important factors such as age, income, loan amount, and credit score with the probability of risk occurrence.

However, the effectiveness of this model is limited when encountering nonlinear factors or complex interactions between variables.

Evaluation indicators such as accuracy, sensitivity, and F1-score show that, although the model can achieve quite good performance in simple situations, there is still room for improvement to achieve higher efficiency in complex situations or imbalanced data.

6.2 Future research directions

Based on these findings, there are some suggestions to improve the performance of the model in credit risk prediction:

1. Apply Regularization methods such as Lasso and Ridge Regression to reduce overfitting and improve the ability to handle multicollinearity.
2. Incorporate new features related to the customer's ability to repay debt, to create a more detailed picture of credit risk.
3. Use more complex models, such as Ensemble models or nonlinear techniques, to improve the accuracy for cases with nonlinear relationships between variables.
4. Consider data outliers and data imbalances, to ensure accuracy and reduce errors in prediction.

References

1. Udemy Ligency Team - Machine Learning A-Z: AI, Python & R + ChatGPT Prize – Video Course (2020 ~ 2024)
Helps understand and apply machine learning algorithms, including linear regression, for credit risk prediction.

2. Vu Huu Tiep - Linear Regression in Machine Learning (2016) – Google Machine Learning, Provides knowledge on linear regression and its application in credit risk analysis.

3. Scikit-Learn - David Cournapeau (2007 ~ 2024) – Open-Source Communication, Provides a powerful library to implement machine learning models, including linear regression for credit risk prediction.

4. Andrew Ng - Machine Learning by Andrew Ng (2011) – Stanford Professor of Computer Science, Offers a solid foundation in machine learning and linear regression, applied to credit risk analysis.

5. Kevin P. Murphy - Machine Learning: A Probabilistic Perspective (2012) – Computer Science at University of British Columbia (UBC), Introduces probabilistic machine learning models, useful for assessing credit risk probabilities.

6. Andreas C. Müller & Sarah Guido - Introduction to Machine Learning with Python (2016) – Assistant Professor of Computer Science at University of California, Berkeley, Provides detailed guidance on implementing machine learning algorithms, particularly linear regression in Python for credit risk prediction.

7. Thomas, L., Edelman, D., & Crook, J. - Credit Scoring and Its Applications (2002) Provides insights into credit scoring systems and prediction models, supporting the development of risk prediction models.

8. Hand, D. J., & Henley, W. E. - Statistical Classification Methods in Consumer Credit Scoring: A Review (1997), Introduces statistical classification methods in credit scoring, supplementing credit risk classification strategies.

9. IEEE Transactions on Neural Networks and Learning Systems - Journal of Advanced Research in AI and Machine Learning (2012), Presents advanced research in AI and its applications in finance, supporting the application of machine learning in credit risk analysis.

10. World Bank Group - “Credit Risk Scoring: Best Practices in Financial Sector” (2020), Offers best practices in credit risk scoring, helping to build more accurate risk prediction models.