Statistical Analysis of CreditScore for Loan Risk Classification

Tran Quoc Thai
Student ID: 2370759

April 9, 2025

1 Introduction

In this section, we compute the variance and entropy of the *CreditScore* feature separately for both risk classes (High and Low). These statistical measures will provide insights into how different machine learning models handle the distribution of this feature.

2 Variance Calculation

Variance is given by:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2, \tag{1}$$

where x_i represents individual credit scores within each risk class, and μ is the mean credit score for that class.

2.1 Low Risk Class

Given the CreditScores for Low Risk: {720, 750, 780, 710},

$$\mu_{\text{Low}} = \frac{720 + 750 + 780 + 710}{4} = 740, \tag{2}$$

$$\sigma^2_{\text{Low}} = \frac{(720 - 740)^2 + (750 - 740)^2 + (780 - 740)^2 + (710 - 740)^2}{4} \tag{3}$$

$$= \frac{400 + 100 + 1600 + 900}{4} = 750. \tag{4}$$

2.2 High Risk Class

Given the CreditScores for High Risk: {650, 600, 630, 640},

$$\mu_{\text{High}} = \frac{650 + 600 + 630 + 640}{4} = 630, \tag{5}$$

$$\sigma^2_{\text{High}} = \frac{(650-630)^2 + (600-630)^2 + (630-630)^2 + (640-630)^2}{4} \tag{6}$$

$$= \frac{400 + 900 + 0 + 100}{4} = 350. \tag{7}$$

3 Entropy Calculation

Entropy is given by:

$$H(X) = -\sum p(x) \log_2 p(x). \tag{8}$$

We estimate entropy based on the frequency of CreditScore ranges within each risk class.

$$H_{\text{Low}} = -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) \tag{9}$$

$$= -4 \times \frac{1}{4}\log_2\frac{1}{4} = 2. \tag{10}$$

Similarly,

$$H_{\text{High}} = -4 \times \frac{1}{4}\log_2\frac{1}{4} = 2. \tag{11}$$

4 Interpretation for Machine Learning Models

- **Decision Trees:** Decision trees use entropy reduction to determine splits. Since both classes have the same entropy, CreditScore alone may not be a strong discriminating feature unless combined with another attribute.
- **Logistic Regression:** Given the variance differences, logistic regression can still effectively model CreditScore using a linear decision boundary.
- **Neural Networks:** With sufficient training data, neural networks can learn nonlinear patterns in CreditScore, potentially improving classification performance.
- **Bayesian Methods:** Since the variance of CreditScore is higher in Low Risk, a Naive Bayes classifier might make incorrect independence assumptions, reducing performance.

5 Conclusion

The variance and entropy calculations show that CreditScore alone may not be the best split criterion for decision trees but can still contribute to other models like logistic regression and neural networks.