Estimating Risk Level for Missing Education Value

Tran Quoc Thai
Student ID: 2370759

March 23, 2025

1 Introduction

In predictive modeling, missing values are a common challenge. In this report, we estimate the probability of T_2 (ID = 3) being classified as High Risk given that its Education value is missing. We utilize patterns in the available features, `Age` and `CreditScore`, to make this determination. Furthermore, we propose methods to handle similar missing values in future cases.

2 Data Analysis and Risk Patterns

The training dataset consists of the following records:

ID	Age	CreditScore	Education	RiskLevel
1	35	720	16	Low
2	28	650	14	High
3	45	750	Missing	?
4	31	600	12	High
5	52	780	18	Low
6	29	630	14	High
7	42	710	16	Low
8	33	640	12	High

Table 1: Training Dataset with Missing Education Value for T_2 (ID = 3)

To estimate the risk level for T_2, we examine the relationship between `Age`, `CreditScore`, and `RiskLevel`.

2.1 Observing Risk Patterns

We categorize the training data into two risk groups:

- **High Risk** cases: `Age` ≤ 35 and `CreditScore` ≤ 650.
- **Low Risk** cases: `Age` > 35 and `CreditScore` > 700.

From the dataset:

- High Risk individuals: (ID = 2, 4, 6, 8) have Age ≤ 35 and CreditScore ≤ 650.
- Low Risk individuals: (ID = 1, 5, 7) have Age > 35 and CreditScore > 700.

For T_2 (ID = 3):

- Age = 45 (Falls in the Low Risk category)
- CreditScore = 750 (Falls in the Low Risk category)

Since all individuals with Age > 35 and CreditScore > 700 belong to the Low Risk group, we infer that T_2 is most likely Low Risk.

3 Probability Estimation Using Bayesian Approach

The probability of an instance being High or Low Risk given `Age` and `CreditScore` can be estimated using conditional probability:

$$P(R|X) = \frac{P(X|R)P(R)}{P(X)}. \tag{1}$$

We estimate:

$$\begin{aligned} P(\text{High Risk}|\text{Age} > 35) &= 0\% \\ P(\text{Low Risk}|\text{Age} > 35) &= 100\% \\ P(\text{High Risk}|\text{CreditScore} > 700) &= 0\% \\ P(\text{Low Risk}|\text{CreditScore} > 700) &= 100\% \end{aligned}$$

Since both conditions indicate a 100% probability of Low Risk, we conclude:

$$P(T_2 = \text{High Risk}) \approx 0\%, \quad P(T_2 = \text{Low Risk}) \approx 100\%. \tag{2}$$

4 Handling Missing Values

The missing `Education` value for T2 does not significantly affect our classification since `CreditScore` and `Age` patterns already indicate a `Low Risk` classification. However, missing values must be addressed systematically in predictive modeling. The following strategies can be employed:

1. Mean/Median Imputation: One of the simplest methods is replacing missing values with the mean or median of the available data in the same column. However, this may introduce bias if the missing values are not missing at random.

2. K-Nearest Neighbors (KNN) Imputation: The missing value can be estimated based on the average `Education` values of the K-nearest instances with similar `Age` and `CreditScore`. This method maintains data consistency but requires careful selection of K.

3. Predictive Modeling: A machine learning model, such as a regression model, can be trained to predict `Education` values based on features like `Age`, `CreditScore`, and `RiskLevel`. This approach is more robust but requires a well-structured dataset for training.

4. Deletion of Incomplete Records: If missing values are sparse and appear randomly, deleting such records may not significantly impact model performance. However, this is not advisable when data loss is considerable.

For future cases, KNN imputation is recommended since it preserves relationships between features while effectively handling missing values without significant distortion.

5 Conclusion

Based on the patterns in `Age` and `CreditScore`, the probability of T_2 being High Risk is approximately 0%, while the probability of being Low Risk is 100%. This conclusion is supported by observed trends where all individuals with Age > 35 and CreditScore > 700 belong to the Low Risk category. To handle missing values in future cases, we recommend predictive imputation and probabilistic estimation techniques to ensure reliable classification.