BI2025 Assignment 3 Interim Report - Group 79

Mehmet Fatih Dogan[*]
TU Wien
Austria

Merve Yilmaz[†]
TU Wien
Austria

Abstract

This report details the data preparation and exploratory analysis phases of a predictive modeling project within a Business Intelligence context. Following the CRISP-DM methodology, we conducted a rigorous data cleaning process on an initial dataset of 8,799 records. By identifying and removing 2,387 erroneous entries—including extreme outliers such as ages reaching 1,808—the dataset was stabilized to 6,412 plausible records. Furthermore, feature engineering was performed through categorical binning of age and income variables to capture non-linear patterns. To ensure model compatibility, we implemented Label Encoding and Z-score Standardization, centering features and reducing bias from differing scales. The resulting preprocessed dataset provides a robust foundation for subsequent machine learning tasks, ensuring data integrity and improved predictive reliability.

CCS Concepts

• **Computing methodologies → Machine learning**; • **Information systems → Data mining**.

Keywords

CRISP-DM, Provenance, Knowledge Graph, Credit Scoring

ACM Reference Format:
Mehmet Fatih Dogan and Merve Yilmaz. 2025. BI2025 Assignment 3 Interim Report - Group 79. In *Proceedings of Business Intelligence (BI 2025).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Business Understanding

1.1 Data Source and Scenario

This project utilizes the Customer Credit Score dataset to automate the risk assessment process for a financial institution. The scenario focuses on identifying high-risk applicants to minimize potential loan defaults. Our data source is: https://www.kaggle.com/datasets/systemdesigner/s and our Scenario provide best model for credit risk calculation

1.2 Business Objectives

Minimize financial losses by accurately predicting loan default probability and identifying high-risk applicants before approval.

2 Data Understanding

Dataset Description: Comprehensive dataset containing customer demographics, financial indicators, credit history, and payment behavior.

The following features were identified in the dataset:

Table 1: Raw Data Features and Descriptions

Feature Name	Data Type	Description
`age`	float	Age of the customer
`annual_income`	float	Total yearly income
`credit_history_age`	integer	Age of credit history (months)
`credit_mix`	string	Type of accounts (Good/Std/Bad)
`id`	string	Unique record identifier
`interest_rate`	float	Average interest rate on loans
`monthly_salary`	float	Net monthly take-home pay
`target`	integer	Credit score classification

3 Data Preparation

3.1 Data Cleaning

To ensure the integrity of our predictive model, we performed a rigorous data cleaning process focused on outlier mitigation and handling missing values.

Outlier Analysis: Before cleaning, an exploratory analysis was conducted to identify data quality issues. As shown in Figure 1, the raw dataset contained extreme anomalies, including a maximum age of 1,808 and annual income values reaching 2.8 million. These physically and financially impossible values were identified as significant noise.

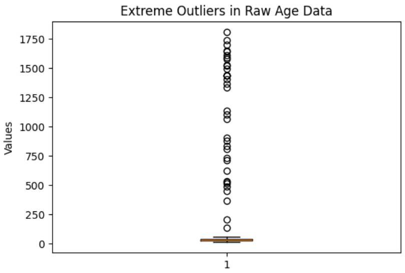


Figure 1: Boxplot analysis showing extreme outliers in the raw Age and Income features prior to cleaning.

Following the identification of outliers, we filtered the dataset to include only plausible records. The original dataset of 8,799 rows was reduced to 6,412 clean rows by removing 2,387 erroneous

[*] Student A, Matr.Nr.: 12437437
[†] Student B, Matr.Nr.: 12536887

BI 2025, December 2025

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

entries. As shown in Figure 2, the resulting age distribution now aligns with realistic biological expectations, capped at 134.

- **Outlier Removal:** We identified and removed 2,387 records that contained physically and financially impossible values (e.g., age of 1,808 and `annual_income` of 2.8M).
- **Final Dataset:** After cleaning, the dataset size was reduced to 6,412 rows.
- **Data Stabilization:** This process stabilized the features, reducing the maximum age to 134 and the maximum `annual_income` to 174,098.

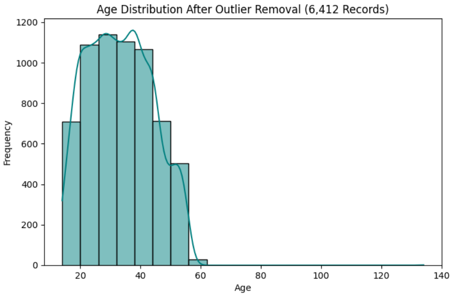
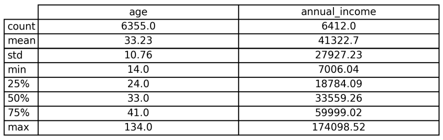


Figure 2: Data distribution after removing extreme outliers (Age and Income).



	age	annual_income
count	6355.0	6412.0
mean	33.23	41322.7
std	10.76	27927.23
min	14.0	7006.04
25%	24.0	18784.09
50%	33.0	33559.26
75%	41.0	59999.02
max	134.0	174098.52

Figure 3: Descriptive statistics of the dataset after removing outliers and invalid entries ($N = 6,412$).

As shown in Figure 3, the dataset's central tendencies and ranges have been stabilized. Specifically, extreme values in the `age` and `annual_income` columns were addressed to ensure the integrity of the subsequent scaling process.

3.2 Feature Engineering (Binning)

We applied a binning strategy to transform continuous numerical features into discrete categorical variables to capture non-linear patterns.

- **Age Groups:** Customers were categorized into 'Young' (0-30), 'Adult' (30-50), and 'Senior' (50+). The resulting distribution is shown in Figure 4.
- **Income Levels:** `annual_income` was divided into 'Low' (<20k), 'Medium' (20k-70k), and 'High' (>70k). The classification of these tiers is visualized in Figure 5.

In addition to individual feature distributions, the interaction between these two binned variables was analyzed. Figure 6 presents a heatmap illustrating the cross-tabulation of Age Groups and Income Levels, providing a comprehensive view of the prepared features.

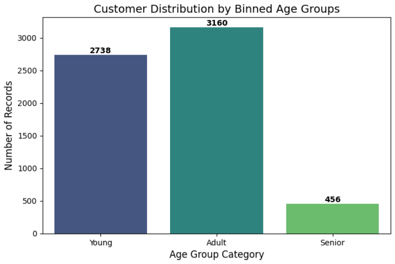


Figure 4: Distribution of customer records across Age Group categories.

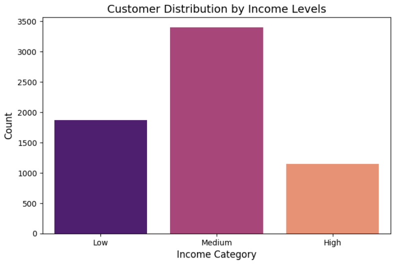


Figure 5: Distribution of customer records across Income Level categories.

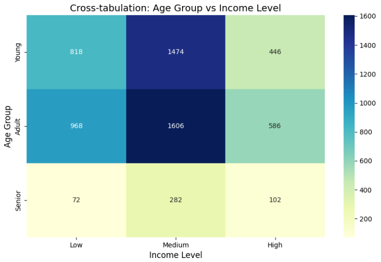


Figure 6: Heatmap of the relationship between discretized Age and Income features.

3.3 Scaling and Encoding

To finalize the data preparation phase and ensure model compatibility, we performed categorical encoding and numerical scaling as highlighted in the preliminary analysis.

Label Encoding: Since the binned features (*Age Groups* and *Income Levels*) represent ordinal data, we applied *Label Encoding* to convert these categories into numerical formats. This allows the machine learning algorithms to process demographic segments as ranked variables without losing their inherent order.

Feature Scaling: To prevent features with larger numerical ranges, such as `annual_income`, from disproportionately influencing the model, we implemented *Standardization (Z-score normalization).* This process rescales the features to have a mean of 0 and a standard deviation of 1. As illustrated in Figure 7, this stabilization ensures that all features contribute equally to the model's weight updates[cite: 8, 68].

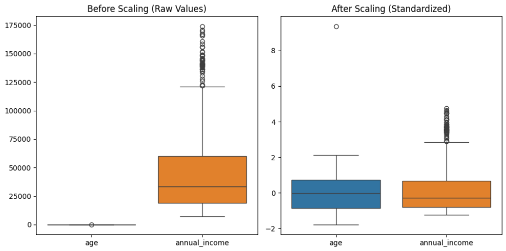


Figure 7: Comparison of numerical feature distributions before and after Z-score Standardization.

4 Availability and Licensing

To ensure reproducibility and transparency, all materials related to this study are publicly available:

- **Repository URI:** https://github.com/merciyilmaz/BI2025_Group79_Assignment3_CreditScoring
- **Software License:** The source code is licensed under the **GNU General Public License v3.0 (GPLv3)**.
- **Documentation License:** The report and visualizations are licensed under the **Creative Commons Attribution 4.0 International (CC-BY 4.0)**.