Indian Institute of Technology, Dharwad

CS 209/214 - Artificial Intelligence Term Project Report

Course Instructor:

Dr. Dileep A.D.

Mentor:

Ms. Neha Rudrappa Pudakalakatti

Team 21

Raghuveer Verma   CS23BT041

Annavarapu Vijwal   MC23BT001

Shaikh Aariz   MC23BT030

Abhinandan Jain   CS23BT012

Abstract

In this project, we explore and compare various Machine Learning models for three tasks: Classification, Regression, and Clustering. The data set used contains features that describe the financial situation and personal attributes of individuals.

In classification, the goal is to predict whether a loan is approved or not. For the regression task, we use models to estimate an individual's risk score. Clustering involves identifying clusters in the data, where we use techniques such as the Elbow Method and Silhouette Analysis to get the optimal number of clusters.

Contents

The Dataset

We use a synthetic dataset comprising of 20,000 records of personal and financial data. The original dataset can be found here.

Features

It contains the following thirty-six columns:

1. Demographic and Employment Information

   - Age
   - Marital Status
   - Number of Dependents
   - Education Level
   - Employment Status
   - Experience
   - Job Tenure

2. Financial and Credit Information

   - Annual Income
   - Monthly Income
   - Total Assets
   - Total Liabilities
   - Net Worth
   - Home Ownership Status
   - Savings Account Balance
   - Checking Account Balance
   - Debt-to-Income Ratio
   - Total Debt-to-Income Ratio
   - Monthly Debt Payments
   - Utility Bills Payment History

3. Credit History and Loan Information

   - Credit Score
   - Credit Card Utilization Rate
   - Number of Open Credit Lines
   - Number of Credit Inquiries
   - Bankruptcy History
   - Length of Credit History
   - Previous Loan Defaults
   - Payment History

4. Loan Attributes

   - Loan Amount
   - Loan Duration
   - Loan Purpose
   - Base Interest Rate
   - Interest Rate
   - Monthly Loan Payment
   - Application Date

5. Target Variables

   - Loan Approved
   - Risk Score

Analysis and Preprocessing

Duplicate and Missing Values: No duplicate rows or rows with missing values were found.

Columns with multiple classes: The following columns described multiple classes as strings which were later converted to integer labels (given on the left) and one-hot encodings in separate datasets:

- **Employment Status**

    - 0: Employed
    - 1: Self-Employed
    - 2: Unemployed

- **Education Level**

    - 0: Master
    - 1: Associate
    - 2: Bachelor
    - 3: High School
    - 4: Doctorate

- **Marital Status**

    - 0: Married
    - 1: Single
    - 2: Divorced
    - 3: Widowed

- **Home Ownership Status**

    - 0: Own
    - 1: Mortgage
    - 2: Rent
    - 3: Other

- **Loan Purpose**

    - 0: Home
    - 1: Debt Consolidation
    - 2: Education
    - 3: Other
    - 4: Auto