Credit Card Data Science Project - CRISP-DM Methodology

Topic Selected

Credit Cards

Business Understanding Stage

Problem Description

Our financial institution is experiencing significant financial losses due to credit card defaults. Currently, we approve credit card applications based on basic criteria, but many approved customers fail to make their payments on time or default entirely. This results in:

- High financial losses from unpaid debts
- Increased collection costs
- Reduced profitability
- Poor risk management

The bank needs a more sophisticated approach to evaluate credit card applications to minimize the risk of approving customers who are likely to default while still maintaining a competitive approval rate for qualified customers.

Data Science Question

"Can we predict whether a credit card applicant will be a reliable payer (low default risk) based on their demographic, financial, and credit score historical information?"

CRISP-DM Methodology Implementation

1. Analytic Approach

We will use a predictive analytics approach, specifically binary classification, to predict whether an applicant will be a good payer or not. This supervised machine learning approach will analyze historical customer data to identify patterns that indicate payment reliability.

Credit Card Data Science Project - CRISP-DM Methodology

2. Data Requirements

We need demographic data (age, income, employment), credit history information (credit score, previous loans, payment history), and the target variable indicating whether past customers were reliable payers. The data should be in structured format from internal bank databases and external credit bureaus.

3. Data Collection

Data will be collected from internal bank databases containing customer applications and payment histories, external credit bureaus for credit scores and reports, and government databases for income verification. We'll gather 3 to 5 years of historical data to ensure sufficient examples of both reliable and unreliable payers.

4. Data Understanding and Preparation

We'll explore the collected data to understand its quality, completeness, and distribution of good vs bad payers. Then we'll clean inconsistencies, handle missing values, create new features like debt-to-income ratio, encode categorical variables, and split the data into training and testing sets for model development.

5. Modeling and Evaluation

We'll develop and test multiple classification models (decision trees, logistic regression, random forest) using the training data. Models will be evaluated using the test set to measure accuracy, precision, and recall. We'll use ROC curves to find the optimal balance between approving good customers and rejecting risky ones, then select the best-performing model for deployment.