Athens University of
Economics and Business

MSc in Data Science

Statistics for Big Data

Credit Card Assignment 1

MOUSELINOS SPYRIDON

NOVEMBER 2020

Problem Formulation

A bank wants to issue a new credit card. It is going to have small daily limits and no subscription.

It is an innovative product, and they are frustrated how to go with this.

Proposed Approaches:

- **Approach 1**: Use the historical data base with all credit cards of the bank to predict what is the expected usage and risks from that.
- **Approach 2**: Give the card to 5000 clients randomly and see their behavior for two months and then decide.

Let’s analyze both sides and present a viable solution schema … Hopefully

Approach 1 Analysis

Positive Traits / Key Factors:

- Prior Field Knowledge, in the form of a Database already exists, with data of different kinds of credit cards issued by the bank.

- The number of examples / points in our dataset will be large enough for a complete cross-validating training schema.

- Data should be enough to aggregate on different factors /scenarios thus simulating held out datasets of different patterns.

- Virtually zero-cost solution, that can be implemented on the fly and be maintained as a long-term service, refined on new data when necessary.

Approach 1 Analysis

Negative Traits / Obstacles:

- Since the card holds new features, there should be a degree of dissimilarity with previous data, to the point where the set of those features will be distributed over many different cards and different scenarios. That introduces a degree of intrinsic heterogeneity between examples.

- Since the acquisition of a card is low-frequent phenomenon, data collected from previous customers will be biased against behavioral patterns and economical aspects different from that of contemporary users. That can translate into a failure to forecast correctly under confidence.

- Finally, one must consider what is the question in quantitive terms. What does it mean for a card to be effective? What are the indicators and latent variables associated with this? Since a plethora of features (Big Data Era – Big Data DB's) can be collected and stored, a thorough variable selection and dimensionality reduction schema should be designed, for maximizing the effectiveness of the analysis.

Approach 2 Analysis

Positive Traits / Key Factors:

- In this approach, instead of heterogenous data we can design from scratch the variables that should be measured, and the questions that should be asked.

- As a result, the analysis is based on primary data that are of higher quality than sparsely connected features of similar cards.

- Furthermore, we can now model the current trend and increase the confidence of our decision-making process.

Approach 2 Analysis

Negative Traits / Obstacles:

- High-cost experiment with the quantity of the results being probably inadequate in a two-month time span for a high-confidence prediction analysis.

- A new card might not necessarily replace the old ones in terms of payment habits, where some users might still use their old cards thus not contributing to the data acquisition schema.

- Any misconduct or failure in the procedure will only mean more trials to be conducted, thus aggravating the costs of a survey, losing two months for a fallback solution.

Overview

Considering the two approaches given, I would recommend approach 2 for the following reasons:

- By understanding the underlying factors and designing carefully a custom-tailored data-acquisition schema, one should be able to obtain the true current trend leading to a reliable decision making.

- Primary data collected through experiment are easier to analyze and interpret in comparison with the huge amount of historical data, especially when there is no guarantee that the same features should be present in both scenarios, rather a subset of them .

- Surely running an experiment is costly, requires time to get an outcome and involves risk. However, the outcome is reliable avoiding more costly mistakes in the future.

The above solution holds under the following constraints:

- The experiment is designed in a proper way (Important indicators / variables, are collected)

- The number of clients is sufficient to detect the required information for decision making. (High Number of examples)

- The subjects/clients are randomly selected. (Unbiased measurements)