**DATS6103 Introduction of Data Mining - Final Project**

**Prepared by: Mrunalini Devineni, Aihan Liu and Sara Sanchez – Group 3**

**Introduction:**

A Credit Score is a statistical model that classifies a client according to their risk profile in one of the following categories: good or bad. The score classification (a number) is based on the probability of default. The latter is calculated as the sum of the estimated coefficients for each variable that profile the client. Next, these coefficients are transformed into scores, obtaining, as a result, a score that assesses the customer's risk.

The main objective of the Score is to generate a statistical model that allows forecasting the probability of default of a group of clients, based on their profile information, and ensure that credit is granted to a person who can meet the payment.

**Problem Statement and Dataset:**

The primary purpose of this research is to distinguish between two profiles: those who comply and do not comply with a vehicle loan payment. Therefore, this company may have a better tool for granting credit and thus reducing errors. This problem is interesting since, in a consumer society mainly driven by credit, it is necessary to have tools that allow avoiding non-payment through the efficient recognition of non-delinquent profiles. Furthermore, this filter reduces the portfolio default, and problems are avoided for some people not to acquire a loan they will not fulfill.

For this study, we have chosen to analyze an L&T company data set, in charge of the sale of cars, from kaggle.com. This dataset is similar to the one that financial institutions have to build the scoring models that allow them to forecast the approval or rejection of customers. This dataset does not require any cleaning and is equipped to fuel the analysis of this project. The base consists of 40 variables and 233 154 observations assessing a person's attributes ranging from demographic data (date of birth, etc.) and bureau data, like the amount of loan disbursed and the asset's cost.

Due to the abundance of data recorded mainly from the credit aspects of an individual's life, this dataset enables us to deep dive using decision trees or random forest techniques to estimate the determinants of vehicle loan default.

**Methodology**:

To pursue the project's goal, we will use anaconda libraries with a combination of feature selection methods. Using correlation enables us to analyze the strength of the relationship between the variables in the dataset. At the same time, random forests and decision trees will build a model that will output default rates. The supervised machine learning algorithms will be the main for this project due to the possibility of training models, influencing decisions, and producing binary results (approval or reject, for example).

As we have mentioned, this tool is similar to the one used by financial institutions to build scoring models that allow them to forecast the approval or rejection of clients and thus reduce the delinquency of the institution and the financial system. In addition, this tool also helps people since the refusal of specific profiles with potential insolvents prevents them from acquiring unpayable debts.

The packages we will use for this project including pandas, numpy, sklearn, and some others that might help us to find a better solution.

**Evaluating Performance:**

The accuracy of the model will be validated using the confusion matrix. The use of this methodology allows us to evaluate the presence of false positives, and thus make a bad prediction. In practice, the scoring model, together with decision-making, which depend on the company's risk appetite, will select the cutoff point for the score at which a portfolio is approved or rejected. Another indicator will be the GINI coefficient to measure the inequality, so this is a dispersion coefficient.

The references that we will use are Baesens et al. (2016) and Siddiqi (2006). These sources are some of the most important when it comes to the generation of scoring models. Thanks to the analysis described in these documents, together with our experience, we will create the context we need to develop our model.

**Schedule:**

**Week10 proposal submission, start data cleaning**

**Week11 start feature selection**

**Week12 start modelling**

**Week13 modeling and start writing report**

**Week14 format report and working with the presentation (Exam II)**

**Week15 Final presentation**

A picture containing timeline

Description automatically generated

**References:**

Baesens, B., Roesch, D., & Scheule, H. (2016). *Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS*. John Wiley & Sons, Incorporated.

Siddiqi, N. (2006). *Credit risk scorecards : developing and implementing intelligent credit scoring*. Wiley.