

**Module:** CMP-7023B Data Mining  
**Assignment:** Data Mining the Lending Club Loan Dataset.  
**Set by:** Dr Tahmina Zebin, <t.zebin@uea.ac.uk>  
**Date set:** 05/03/2021  
**Value:** 65%  
**Date due:** Wednesday, 19<sup>th</sup> May 2021, 3 pm  
**Returned by:** 16<sup>th</sup> June 2021  
**Submission:** Blackboard

### Learning outcomes

- Competence in using KDD software tools in medium to large databases.
- Competence in applying relevant techniques at each stage of the KDD process
- Ability to evaluate the suitability of software tools in the context of different data analysis tasks.
- Competence in combining data manipulation and analysis approaches to improve the quality of input data.
- Understanding and identification of problems in input data such as outliers, missing data, unreliable data, differences in granularity, and others, and identify an adequate strategy to deal with the problem data.
- Presentation of knowledge induced in a format suitable for the target audience and for the particular application.

### Specification

#### Overview

#### Aim

- To obtain an overall view of the complex process of Knowledge Discovery in Databases and understand the need for a methodical approach to KDD.
- To explore tools and algorithms available to each stage of the KDD process.
- To gain experience of using KDD software tools in a medium sized database.
- To learn to combine data manipulation and analysis approaches to improve the quality of input data.
- To produce a suitable report describing the methods applied and the discussion of the findings

## Description

To complete this coursework, you will be using the Lending Club Loan Dataset (<https://www.kaggle.com/ethon0426/lending-club-20072020q1>). A curated version of the dataset is uploaded on Blackboard as 'LendingClubLoans2018-2020.xlsx', which contains the loan applications related data the company received in years 2018 to 2020.

The file has 77160 observations and 108 variables (memory usage: 50+ MB). If your computer has memory restrictions feel free to complete the experiment with a smaller portion of the provided data.

In the given data file, there are various information related to loan applicants status (homeownership status, annual income, purpose of taking the loan, debt-to-income ratio) and related information on loan amount, loan grade, various dates among others. A further description of the fields can be found in the Data Dictionary tab in the excel file for the dataset. Your task is to accurately classify the loan status (Current, Fully Paid, charged off, late etc.) from the given fields and then help the lending club predict applications that may default/be late in paying or can be identified as potential bad loans.

To accomplish your task, you need to perform the following operations:

1. Download the dataset and prepare a **summary of the features** available on the dataset including data type (**numerical/ categorical**), amount of missing data and outliers in individual fields.
2. Undertake any **cleansing or pre-processing** you think is necessary on the dataset. In your report, explain clearly what you have done and why you have done it. Some cleaning could be to remove any feature/column if 60% missing values or holds a NULL, constant, NaN values, or to remove duplicate and highly correlated information.
3. Split the data into a training set and a test set once cleansing is done. Use suitable toolkit and libraries (Python, Orange, Weka, or R whichever platform you are comfortable with) to train models (e.g. Decision Tree, Random Forest or SVM) from the training set to build the **Loan Status Classifier**. Note that you should deal with any **class imbalance**, do feature selection and other adjustments/tuning to improve the quality of the models obtained. You will need to test the performance of your model on your test set. As part of your final report, please describe and justify the decisions you have made, the results, how the models have been validated/evaluated and discuss the model's effectiveness in terms of precision and recall performances.
4. In the next stage, use an **unsupervised clustering algorithm** (K-means, or hierarchical) on the data. Use Scatter plots or t-SNE plots on the clusters to see if there are clusters formed for the various types of loan status (Current, Fully Paid, charged off, Late etc.). The Loan Status field should be omitted during clustering. Discuss your observations on the clusters in your report. Is there any suitable clustering for good loans that are paid regularly and bad loans that late/defaulted based on Loan Status Types?

## Marking scheme

### Assessment criteria

Marks will be distributed as follows:

Part 1: Summary of features	10
Part 2: Data Pre-processing	25
Part 3: Supervised Model Training and Evaluation	30
Part 4: Unsupervised Clustering	20
Overall presentation, references, and conclusions	15
	100%

## Deliverables

Please collate all the answers to the above questions in a report. The report should follow the structure/sections according to the components of the marking scheme and must not exceed 15 pages including bibliography and references. Also write an abstract of the report as well summarizing your findings. It should be written in a clear and professional manner, using good English. You should also submit your cleaned data and the code/workflow produced to accomplish your tasks.

## Handing in procedure

Please submit your piece of coursework electronically on the blackboard dropbox associated to the coursework. You should upload using the following format studentID-StudentName.zip for your file submission.

## Resources

You can use the weekly Lab documentations, Lecture notes, Library resources and other sources to accomplish your tasks. Don't forget to cite any external and online resources used. Students are expected to work independently, and any plagiarism or collusion will be heavily penalised.

## Plagiarism

Plagiarism is the copying or close paraphrasing of published or unpublished work, including the work of another student without the use of quotation marks and due acknowledgement. Plagiarism is regarded a serious offence by the University and all cases will be reported to the Board of Examiners. Work that contains even small fragments of plagiarised material will be penalised.