# Monash University

# FIT5202 - Data processing for Big Data (SSB 2024)

Assignment 2A: Building Models to Predict Loan Default

Due: **Friday, 2 February 2024, 11:55 PM**

Worth: 10% of the final marks

# Background

**Mo**nash **Lo**an **Co**rporation(**MoLoCo**, an imaginary company) is an established loan servicing company offering various home loans, consumer credit and other unsecured loans defined as below.

* Home loan: applicants borrow money from a bank or company to purchase a property(house/apartment), usually the property is used as a security (called a mortgage). If an applicant can’t repay the loan, the lender has the right to confiscate the property and sell it.
* Consumer credit: applicants borrow money to purchase goods or services (e.g. an iPhone), and the goods are not secured against the loan. This type of credit usually attracts a higher interest rate. Usually the applicants are required to make a periodical payment set by the contract. As a simple example, if you purchase a $1000 phone with a 20% interest rate over 12 months, you will repay $100 each month. The size of consumer credit is usually small.
* Other unsecured loans: Other types of loans do not require any type of collateral. For example, in an unsecured business loan, applicants borrow money to start a business. If the business fails, the lender could lose money depending on the terms and contracts.

Over many years of operation, they have accumulated many applicants and collected large amounts of data from applicants and applications.

MoLoCo is still using loan assessors and risk managers to process applications, which is slow and inefficient. It plans to execute a digital transformation strategy to optimise operational costs and improve applicant experience. In the first stage, it will utilise big data processing, machine learning and streaming processing to manage loan risk.

# The Problem and Project

One of the main risks of a loan business is **default**, which means applicants borrow money from the company but fail to repay for some reason. We will use big data processing to help the company reduce this risk.

In Assignment 1**,** we analysed historical loan data to help the company better understand their applicants.

In Assignment 2 part A, we are going to build classification models to predict whether an applicant will default.

# The Dataset

**Three files from assignment 1;** the following dataset is **identical** to assignment 1. If you still have the original files(unmodified), you don’t have to re-download them from Moodle.

1. previous\_application.csv: contains data from previous/existing applications
2. application\_data.csv: contains current application and applicant information
3. value\_dict.csv: contains a dictionary of values. Some columns in application tables are stored as integers to save storage, you need to join this table to find the meaning of those values.

**Two new files(available on Moodle):**

1. loan\_default.csv: contains labelled records of loan defaults. This file only has two columns: id\_app and is\_default.
2. metadata.pdf(new): contains column information and data type.

# What you need to achieve

**MoLoCo** requires us to build models to predict whether a loan application will default based on application and applicant information. They provided us with a loan default dataset. Is this supervised or unsupervised learning? (Your thoughts will help you understand the problem context however you don’t have to answer this question in your submission.)

Also, we will use K-means clustering to perform applicant segmentation and analysis.

| Use case 1 | Based on the application and applicant information, predict whether a default will occur. | Classification |
| --- | --- | --- |
| Use case 2 | Perform applicant segmentation using K-Mean clustering | K-Mean |

*\*Applicant segmentation: separate applicants into groups(two or more) with similar attributes.*

# Architecture of Assignment 2 (Parts A and B)

The overall architecture of the assignment setup is represented by the following figure. **Part A** of the assignment consists of preparing the data, performing data exploration and extracting features, and building/persisting the machine learning models.

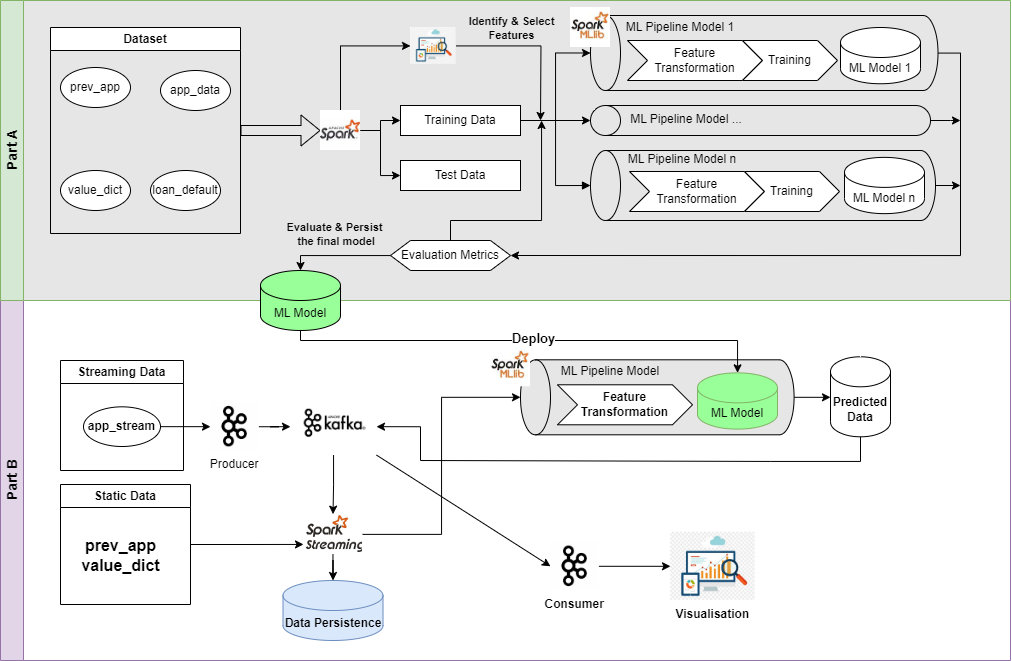


Fig 1: Overall Architecture for Assignment 2 (We work on Part A for now)

In both parts, for the data pre-processing, and the machine learning pipelines, you are required to implement the solutions using PySpark SQL / MLlib / ML packages(excessive usage of Pandas for data processing is discouraged). Please follow the steps to document the processes and write the codes in your Jupyter Notebook.

# Getting Started

● Download the datasets from Moodle.

● Download a template file for submission purposes:

* ***A2A\_template.ipynb*** file in Jupyter Notebook to write your solution. Rename it into the format (for example, ***A2A\_xxxx0000.ipynb*. xxxx0000** is your authcate ID.
* You will be using Python 3+ and PySpark 3.5.0 for this assignment (This environment will be automatically set up if you follow the steps in Moodle ([Unit Information >> Software, Documentation, and Resources](https://lms.monash.edu/course/view.php?id=163247&section=0))).

****

## Part 1. Data Loading, Transformation and Exploration (30%)

In this section, you need to load the given datasets into PySpark DataFrames and use *DataFrame functions* to process the data, usage of Spark SQL is discouraged. You are allowed to use third-party libraries to format the output. For plotting, different visualisation packages can be used, please ensure that you have included instructions to install the additional packages and the installation will be successful in the provided docker container(in case your marker needs to clear the notebook and rerun it).

### 1.1 Data Loading (7%)

1. Write the code to create a SparkSession. For creating the SparkSession, you need to use a SparkConf object to configure the Spark app with a proper application name, to ensure the maximum partition size does not exceed 20MB, and to run locally with **4** CPU cores on your machine[[1]](#footnote-0). (2%)
2. Write code to define the schemas for previous\_application and application\_data, following the data types suggested in the metadata file. (3%)
3. Using your schemas from step 2, write code to load all CSV files into separate data frames(note: value\_dict and loan\_default are simple and don’t need schemas). Print the schemas of all data frames. (2%)

### 1.2 Data Transformation and Create Features (12%)

In this step, we’re going to perform data transformation and create some new features using existing information. (note: you are allowed to use your own code from assignment 1 for some transformation.)

Perform the following tasks on the **application\_data data frame**:

1. Create a new column called **loan\_to\_income\_ratio**(loan to income ratio) defined as amt\_credit/income\_total.
2. Perform age bucketing and create a new string column called **age\_bucket** and set the value below:
   1. age < 25: **Y**
   2. 25 <= age <35: **E**
   3. 35 <= age <45: **M**
   4. 45 <= age < 55: **L**
   5. 55 <= age < 65: **N**
   6. Age >= 65: **R**
3. Create a new string column named **credit\_worthiness**. It takes the average value of credit\_score\_1,2,3 (note: replace null value with 0.5, not 0). If the average >= 0.7, set credit\_worthiness to “high”; 0.4 <= average < 0.7 set to “medium”, < 0.4 set to “low”.
4. Create 4 columns: **num\_of\_prev\_app**(number of previous applications)**, num\_of\_approved\_app** (number of approved applications), **total\_credit** (sum of credit of all **approved** previous applications), **total\_credit\_to\_income\_ratio** (total credit/income). (note: you need to join previous applications to fill in the values.)
5. Replace **education\_type, occupation\_type, income\_type and family\_status** with matching strings from value\_dict (hint: consider reusing code from your A1).
6. Join the loan\_default data frame and add is\_default to application data. We’ll use this column as the label.

Print 10 records from the application\_data data frame.

### 1.3 Exploring the data (11%)

1. With the transformed data frame from 1.2, write code to show the basic statistics **(3%)** (pandas is allowed for this task):
   * a) For each numeric column, show count, mean, stddev, min, max, 25 percentile, 50 percentile, and 75 percentile;
   * b) For each non-numeric column, display the top 5 based on counts in descending order;
   * c) For each boolean column, display the value and count(i.e., two rows in total).
2. Explore the data frame and write code to present **two plots**[[2]](#footnote-1) worthy of presentation to MoLoCo, describe your plots and discuss the findings from the plots. **(8%)**
   * One plot must be related to the default. (e.g. what attribute/attributes are correlated to default, what kind of application/applicant has a higher probability of default, etc.)
   * Feel free to choose any for the other plot.
   * Hint: you can use the basic plots (e.g., histograms, line charts, scatter plots) for the relationship between a column and the label; or more advanced plots like correlation plots. If the data is too large for the plotting, consider using sampling before plotting.
   * 150 words max for each plot’s description and discussion
   * Feel free to use any plotting libraries: matplotlib, seabon, plotly, etc.

## Part 2. Feature extraction and ML training (50%)

In this section, you are **only** allowed to use PySpark DataFrame functions and ML packages for data preparation, model building, and evaluation. Other ML packages, such as scikit-learn, would receive **zero marks**. Excessive usage of Spark SQL is discouraged.

### 2.1 Discuss the feature selection and prepare the feature columns (10%)

1. Based on the data exploration from 1.2 and considering the use case, discuss the importance of those features. (For example, which features may be useless and should be removed, which feature has a great impact on the label column, which should be transformed)Which features you areplanning to use? Discuss the reasons for selecting them and how you create/transform them[[3]](#footnote-2)
   * 300 words max for the discussion
   * Feel free to add/remove features based on your exploration
2. Write code to create/transform the columns based on your discussion above

### 2.2 Preparing Spark ML Transformers/Estimators for features, labels, and models (10%)

1. Write code to create Transformers/Estimators for transforming/assembling the columns you selected above in 2.1, and create ML model Estimators for Random Forest (RF) and Gradient-boosted tree (GBT) model.
   * **Please DO NOT fit/transform the data yet**
2. Write code to include the above Transformers/Estimators into two pipelines(RF and GBT).
   * **Please DO NOT fit/transform the data yet**

### 2.3 Prepare, Train and Evaluate models (30%)

1. Write code to split the data for training and testing purposes. (Note: if the dataset is too large for your machine to train, sampling/sub-sampling is allowed.)
2. Write code to use the corresponding ML Pipelines to train the models on the training data. And then use the trained models to predict the testing data from 2.3[[4]](#footnote-3)
3. For both models(RF and GBT) and testing data, write code to display the count of TP/TN/FP/FN. Compute the AUC, accuracy, recall, and precision for the above-threshold/below-threshold label from each model testing result using pyspark MLlib/ML APIs.
   * Draw a ROC plot.
   * Discuss which **one** is the **better** model and use metrics to support your claim (no word limit, please keep it concise)
4. Save the better model, we will need this for Part B of assignment 2.

(note: You may need to go through a few training loops, adjust features, use more data and other techniques to create a better-performing model.)

## Part 3. Applicant Segmentation and Knowledge Sharing with K-Mean (10%)

In addition to building the previous models, the company would like to learn more about applicant segmentation. The applicant segmentation process clusters/groups applicants into groups based on some criteria. Traditionally, the company segments customer based on their education, income level, marital status, etc. However, they would like us to employ an unsupervised, data-driven approach to see if we can discover new angles to segment applicants.

1. Utilize K-Mean clustering/hyperparameter tuning you have learned in this unit and try to find the optimal K value and train the model.
2. Based on your trained model parameters, make recommendations on how to segment the applicants. Write a paragraph with 300 words maximum.

Note: This is an open question with no right or wrong answers. The evaluation will be based on the quality of your work, feel free to include plots and/or other metrics to support your analysis/claim.

## Part 4. Data Ethics, Privacy, and Security (10%)

In the era of big data, the convergence of vast quantities of information from various sources raises critical questions related to data ethics, privacy, and security. For example, in the case of privacy, many companies are collecting much more data than they need from customers. In our case, we used a real-world data set with real applicant information. How do you utilize those datasets with ethics, privacy and security in mind?

In this part of the assignment, you are tasked to explore these topics within the context of big data processing, drawing on contemporary research, real-world examples, and ethical considerations.

**(word limit: 500 words)**

**(mandatory):** Define the concepts of data ethics, privacy, and security within the big data domain.

**(Choose one or more topics, you can also create your own topic)** Explain the significance of these issues in today’s data-driven world.

**Data Ethics:**

* Analyze how data ethics can influence big data processing;
* Examine real-world examples of how data ethics has been handled, both positively and negatively.
* Analyze the balance between technological advancements and ethical responsibilities

**Data Privacy:**

* Discuss the challenges and importance of maintaining privacy in big data.
* Investigate regulations and laws that govern data privacy, such as GDPR.
* Evaluate tools and techniques used to ensure privacy, and suggest improvements or new methodologies.

**Data Security:**

* Explore the potential security risks associated with big data processing.
* Assess the measures currently in place to secure big data, including encryption, authentication, and authorization.

(**mandatory**) Summarize the key findings of your analysis and include a list of references.

Submission

You should submit your final version of the assignment solution online via Moodle. You must submit the files created:

* Your jupyter notebook file A2A\_authcate.ipynb
* **A pdf file** saved from jupyter notebook with all outputfollowing the file naming format as follows: **A1\_authcate.pdf**

Note that both submitted (ipynb and pdf) files will be scanned using plagiarism detection software. The highest similarity score among students may be interviewed to prove the originality of the task.

Assignment Marking Rubric

A detailed marking rubric is available in Moodle. For complex tasks and explanation questions, you will receive marks based on the quality of your work.

In your submission, the jupyter notebook file should contain the **code and its output**. It should follow *programming standards, readability of the code, and organization of code*. Please find the PEP 8 -- Style Guide for Python Code for your reference. Here is the link: <https://peps.python.org/pep-0008/> Penalty applies if your code is hard to understand with insufficient comments.

Other Information

Where to get help

You can ask questions about the assignment in the Assignments section in the Ed Forum accessible on the unit's Moodle Forum page. This is the preferred venue for assignment clarification-type questions. You should check this forum regularly, as the responses of the teaching staff are "official" and can constitute amendments or additions to the assignment specification. Also, you can attend scheduled consultation sessions if the problem and the confusion are still not solved.

Searching and learning on commercial websites/forums (e.g. Quora, Stack Overflow) is allowed. However, you should not post/ask assignment questions on those forums.

Plagiarism and collusion

Plagiarism and collusion are serious academic offences at Monash University. Students must not share their work with any other students. Students should consult the policy linked below for more information.

<https://www.monash.edu/students/academic/policies/academic-integrity>

See also the video linked on the Moodle page under the Assignment block.

Students involved in collusion or plagiarism will be subject to disciplinary penalties, which can include:

● The work not being assessed

● A zero grade for the unit

● Suspension from the University

● Exclusion from the University

Late submissions

Late Assignments or extensions will not be accepted unless you submit a special consideration form. ALL Special Consideration, including within the semester, is now to be handled centrally. This means that students MUST submit an online Special Consideration form via Monash Connect. For more details, please refer to the **Unit Information** section in Moodle.

There is a **10% penalty per day including weekends** for a late submission. Also, the cut-off date is 7 days after the due date. No submission will be accepted after the cut-off date unless you have a special consideration.

Mark Release and Review

* Mark will be released within 10 business days after the submission deadline.
* Reviews and disputes regarding the mark will be accepted a maximum of 7 days after the release date (including weekends).

**Generative AI Statement**

As per the University’s [policy](https://www.monash.edu/learning-teaching/teachhq/Teaching-practices/artificial-intelligence/policy-and-practice-guidance-around-acceptable-and-responsible-use-of-ai-technologies) on the guidelines and practice pertaining to the usage of Generative AI, all use of generative AI is **restricted** for this assessment. You should **not** use generative artificial intelligence (AI) to generate any materials or content in relation to the assessment task.

The teaching team restricts all use of generative AI to ensure that students apply their own critical thinking and reasoning skills when working on the assessments. In addition, generative AI tools may produce inaccurate content and this could have a negative impact on students’ comprehension of big data topics.

**Data source acknowledgement:**

The dataset is a remix based on several real-world and synthetic datasets. We thank the authors/owners for sharing the original datasets.

1. More information about Spark configuration can be found in <https://spark.apache.org/docs/latest/configuration.html> [↑](#footnote-ref-0)
2. This is an open question, in which you would need to decide what plots to show.

   * You can combine multiple features into one plot, but the plot should be clear to be seen, and do not contain an overwhelming amount of information.
   * If you use subplots, each subplot would be considered as one plot, and the two-plot limit would allow only two subplots for each activity data.

   [↑](#footnote-ref-1)
3. This is an open question, in which you would need to decide what columns to use as features and what transformation(s) would be required for each feature. Include reference when you use arguments from third parties. [↑](#footnote-ref-2)
4. Each model training might take from minutes to hours, depending on the complexity of the pipeline model, the amount of training data, the computing power of your laptop and the code efficiencies [↑](#footnote-ref-3)