

# Predicting Categories of Bank Transaction Data

## Introduction

At MoneyLion, we aim to transform how individuals track, analyse and understand their spending patterns on different categories to help individuals gain control over their financial health. For example, if a user spends \$100 monthly on loan repayments but only has \$50 monthly income, we would like to be able to advise them on how to manage their debt so that their total debt in monthly loan repayments is not greater than their income.

To help us do this, we want to label our users' bank transactions data with a specific category such as Loans or Payroll. We can then aggregate the bank transactions by category to find out how much they spent or earned from each category to find out important insights such as the example above and advise them accordingly.

## Assignment

In this challenge, we want you to develop a model that predicts the category of each transaction in the provided bank transactions dataset (*bank\_transaction.csv*). The list of expected categories is provided in the category column (e.g. Loans, Restaurants, Supermarkets and Groceries, Insurance etc) of *bank\_transaction.csv*.

You are provided with the following files to get you started on this challenge:

1. *bank\_transaction.csv*
  - This dataset contains users bank transactions data and expected categories (e.g. Loans, Restaurants, Supermarkets and Groceries, Insurance etc)
2. *user\_profile.csv*
  - This dataset contains a selected set of users and their financial profiles
3. *data\_dictionary.pdf*
  - Detailed descriptions and definitions of each column within the datasets provided is available in this file.

We are looking for the following items for this challenge:

- Data analysis on the datasets and how it relates to the development of the model or choice of model used to complete this task.
- Reasoning and justification on what type of architecture or techniques will be used to develop this model.
- Functional code to create this model and reasoning for each step taken to develop the model.
- Discuss the performance of the model and effectiveness of using it to predict the expected categories
- Discuss potential ideas and future development plans for this model. For example, what are the next steps to improve on this model in the next 1 month versus the next 3 months.

## Rules for this challenge

- **Python should be used to complete this challenge.** An IPython notebook might be the best way to show your code and write your comments/thoughts to follow along.
- You may refer to online resources and tools but DO NOT directly use any help from other people or online sources. Your submission should be solely your own ideas and work.

- **You may develop the algorithm using any architecture or technique.** Please ensure you provide justification for your choice.
- **The deadline to submit your answer to this challenge is 1 week from receiving the challenge.** There is no hard limit on the amount of time that can be spent on this challenge within that week. **Please disclose how much time you spent** on the challenge when you've completed it.

## Deliverables

A zip file containing any of the following:

- **Python model development notebook** that can run from end to end. **Please also include a HTML file of this notebook.**
- **Any additional documentations**
- Any other **external data used** (only if you added any new ones; DO NOT send back the original assessment data!)
- **Any model files** (if any)
- **Related Python files and folders** (if you use any custom Python files)
- Any other documentation, **external diagrams/visualisations**

## Tips to make your submission stand out

This challenge is intended to be open-ended.

- **All plans and documentations** should be **comprehensive** and presented in a concise manner.
- **Code quality matters** and is expected in the implementation of the model.
- **Creativity**, originality and practicality in approach and presentation will be highly valued.