Buy Now, Pay Later Project
Applied Data Science (MAST30034)

Authors: Akira Takihara Wang, Hamish Gunasekara, Simon Hahn, James Priest, Liam Hodgkinson

Industry Project Overview

Picking which Merchants to Onboard

Welcome to the future of payments! A generic Buy Now, Pay Later (BNPL) firm has begun offering a new "Pay in 5 Installments" feature. Merchants (also known as retailers) are looking to boost their customer base by forming a partnership with this firm and in return, the BNPL firm gets a small percentage of revenue (take rate) to cover operating costs. Since this is a great Win-Win opportunity, there are X number of merchants who wish to partner up!

However, the BNPL firm can only onboard at most $100 < X$ number of merchants every year due to limited resources. This industry project aims to provide students with a unique opportunity to generate a robust ranking system with insights to assist this BNPL firm in finding which merchants they should accept.

To do so, students should aim to answer the following questions in some form:

1. *What are the* 100 *best merchants according to your ranking system?*
2. *What features or heuristics did you discover that greatly separated merchants that should and shouldn't be accepted?*

Students are expected to justify all recommendations by overlaying features such as Customer Demographics, Merchant Locations, and Transaction Data.

Project Details

Students are expected to utilise Python to:

A) Rank all merchants according to their Ranking System. Datasets will be provided such as the list of merchants, transaction data, and consumer data. These datasets will be synthetically generated.

B) Create an automated ingestion pipeline to extract the datasets, apply business rules, and output a curated dataset(s) for insights.

C) Deliver recommendations and insights of merchants that are "interesting". For example, there may be merchants with:

* a small consumer base, but regularly make large purchases;
* or a large consumer base whose customers are prone to higher fraud rates;

    * or a new merchant with little information on how they will perform.

D) Finally, students will need to present the recommendations with appropriate justification and clarity. Notebooks must be readable such that a non-technical user can interpret and verify the findings. Groups that cannot produce sound recommendations or develop a presentable notebook will lose several marks for this project.

Expected Skillset from Student Groups

We expect student groups to be proficient at writing Python in Jupyter Notebooks. Groups will also need to have at least one member who is familiar with requesting APIs, using PySpark, and are comfortable working with database concepts and creating table schemas.

Data Sources

Excluding ABS data, all datasets can be obtained on Canvas under the Modules tab. Be sure to place all data within the `tables` folder in your GitHub repo.

- Synthetically generated merchant data;
- Synthetically generated transaction data;
- Synthetically generated consumer data;
- ABS datasets (public):
    * *Groups should determine which ABS dataset(s) to use and find out where to get them.*
    * SA2 District Boundaries (ABS);
    * Total population by SA2 Districts;
    * Income by SA2 Districts;
    * Census data by SA2 Districts;
- Any other publicly available dataset or data provided by an API.

Groups should determine which ABS dataset(s) to use and find out where to get them.

Weekly Checkpoint Assessment

Groups that do not meet weekly checkpoints will lose the mark(s) allocated for that specific checkpoint. If your tutorial is earlier in the week, then you should show sufficient progress or plans to have it completed by the next week.

Whilst Sprints are usually 2-3 weeks long, we will keep them 1 week long for this subject. Checkpoints may be adjusted over the coming weeks depending on progress or updates from the authors.

- **By the End of Sprint 1 (Semester Week 6):**
  Clone the template repository which contains the datasets. Conduct preliminary analysis to get comfortable with it and begin writing a generalised ETL script for this dataset. Additionally, aim to find at least one external dataset that will be beneficial to the analysis. We recommend looking at the ABS census data that contain Statistical Area fields (i.e SA2).

- **By the End of Sprint 2 (Semester Week 7):**
  Use the ETL script from the previous week on the new datasets. Now, conduct an in-depth outlier analysis and generate insights for the datasets. We expect groups to create a very short summarised notebook detailing the following topics (and any other interesting findings):

  * If you found NULL values after you joined the datasets, what did you do with them? How many were there?
  * Was there any missing data that shouldn't be missing after joining to your external dataset? If anything was missing, how much was there and what did you do about it?
  * If you decided to omit outliers, what does the distribution look like prior and after?

  This week's checkpoint marks will be given based on the answers above, plus, any other insights and findings. Groups are strongly recommended to include geospatial visualisations where possible.

- **By the End of Sprint 3 (Semester Week 8):**
  By now, the ETL script should be fully automated and run end-to-end with no issues. Students should aim to finalise the previous week's work and now aim to curate a dataset that can be used for the Ranking System.

  * A delta file of transactions to be treated as fraud is available. It is up to your group to remove these transactions or add a new field denoting fraud i.e an `is_fraud` boolean field.
  * Groups may see the distribution of the delta file containing fraud transactions and create simple models to predict whether a transaction may be fraud in future (i.e high volume transactions for a consumer who rarely purchases).

- **By the End of Sprint 4 (Semester Week 9):**
  Determine or create features and heuristics that rank your merchants based on the curated dataset above. An initial test set may be released which can be used to see how well your ranking system performs.

  * Groups can train a model to predict a certain feature that will be used in rankings, such as forecasting merchant revenue or the number of customers over time based on historical data.

- **By the End of Sprint 5 (Mid-Semester Break):**
  By this week, an initial ranking system should be in working order. Groups should now segment the merchants themselves to identify the top 10 merchants within each segment.

  * Groups should identify at least three and at most five different segments to classify merchants. For example, clothing vs technology vs groceries.
  * Once the segments have been decided upon, students should curate their ranking systems for each segment.

- **By the End of Sprint 6 (Semester Week 10):**
  Determine the Top 100 merchants overall and the Top 10 merchants for the chosen segments. Additionally, prepare a final summary notebook and walk your tutor through your findings and additional insights (3-5 minutes max). Groups are strongly recommended to include visualisations and prepare a draft of their presentations.

This week's checkpoint marks will be allocated based on the quality of insights and justification over the final ranking system.

- **Presentations begin from Semester Week 11 onward**. Your slides must be uploaded by the Friday before Semester Week 11 by 1700 AEST.
- **By the End of Sprint 7 (Semester Week 11):**
  Deliverable Assessment Submissions Due. This final sprint should mainly be tidying and collating the repository for submission.

Sprints 1 & 2 will be marked together in Week 7. Since Sprint 5 is during the Mid-Semester Break, Checkpoints 4 and 5 will be marked together the following week. You may treat this as your manager (tutor) being away on annual leave (holidays).

Deliverable Assessment

Summary of Outcomes and Repository - 10% of final mark

This project aims to give students the freedom to choose how to approach this as industry work is quite open-ended where you are given full control over the project methodology. For the summary of outcomes, we would like readable Jupyter Notebooks summarising your key findings which can be presented to clients and stakeholders.

- 5% for a readable Jupyter Notebook summarising the overall approach taken, issues that you may have run into, and the limitations/assumptions you made along the way. As this is open-ended, there is no right or wrong answer and we will assess you based on your overall approach.
- 5% for the readability and reproducibility of your group's GitHub repository/code quality. This will be personally assessed by the tutor team.

If you are unsure of how to write a readable Jupyter Notebook, your tutors have been permitted to help you along the way. For example, the first few tutorials on this subject consist of documents that are well documented using a combination of markdown cells, and appropriate code cells with comments. Also, keep your GitHub repository private and only make it public where necessary.