Data Science at N26

Case Study

Abhinav Ralhan
abhinav.ralhan1@gmail.com

Motivation

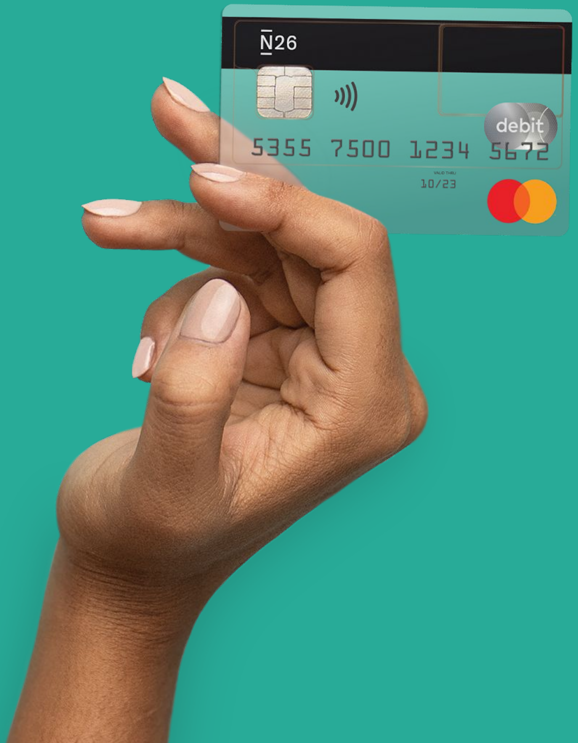
N26 relies on insights generated from customer data to offer users access to the best banking products possible and the greatest level of insight into their own finances.

One recurring need is to assess user creditworthiness & offer customers predictions of how well off they will be financially in future months based on their current transactions.

Your Task

Predict income and expenses for a holdout sample of ~10k users for the month of August based on a training sample of ~10k users from February through July.

Based on your judgement of the usefulness of the results, either aggregate the data into incoming & outgoing flows, or predict based on the transaction type / category level.



Dataset

- Anonymised customer transaction data (random sample of 10000 users, random subset of their transactions)
- User id
- Transaction date
- Transaction type
- Transaction amount (n26’s internal currency)
- mcc_group (Mastercard transaction category, for card transactions)
- Lookup table of transaction types
- Lookup table of credit card categories

Requirements

- Properly packaged source code with running instructions. It should run on Mac / Linuxlike OS
- Commented, clean analysis as Rmarkdown / iPython / iJulia script (or otherwise as appropriate)
- Function that accepts August transaction data in same format as provided here and evaluates performance of algorithm based on this holdout set
- A few visualisations (ggplot, matplotlib, d3.js, etc.) highlighting key results
- A few slides summarising assumptions, results, accuracy, & suitability of predictions for the task
- Outline steps which would be required to integrate this analysis into live production in our app

Analysis Questions

- How confident can we be in the results? Are they useful for the purposes of our original task?
- What performance metrics would you use to evaluate the model?

Contact

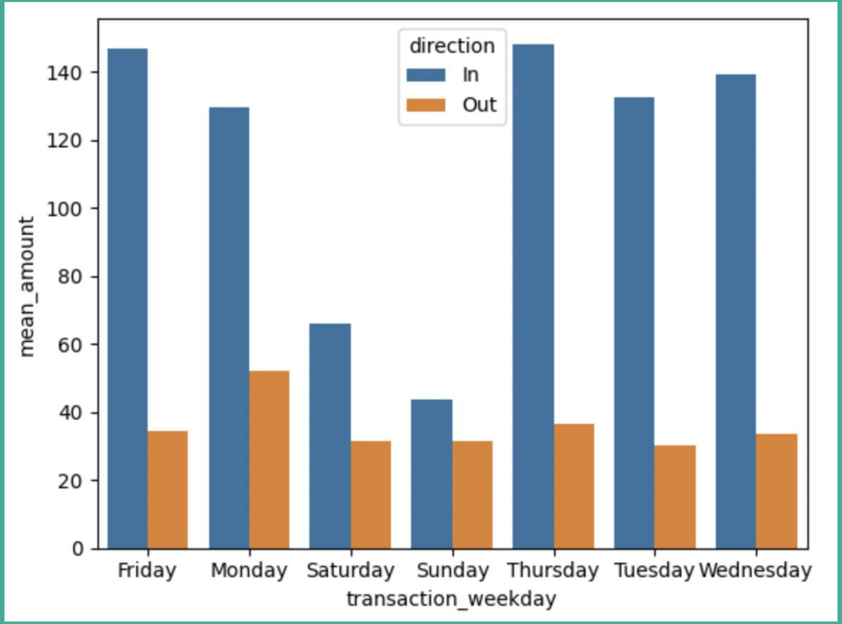
Feel free to send any questions to your Tech Recruiting Partner. We want to ensure everyone has an equal shot at this task, so we will respond to all applicants with compiled task clarifications / corrections based on your feedback.

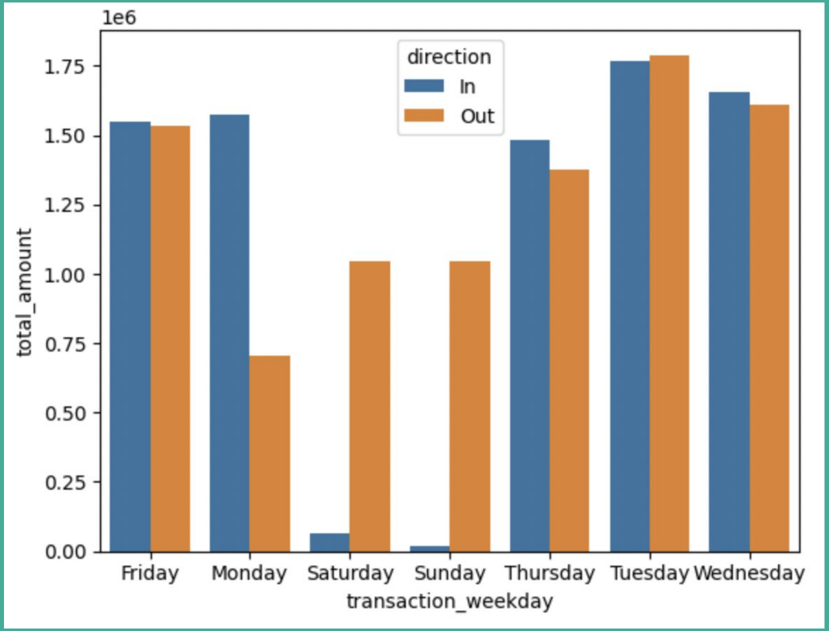
Instructions

- All content is inside n26_AbhinavRalhan.zip.
- The relevant code for insights / analysis / modelling is inside a Jupyter notebook called task.ipynb.
- The notebook can be run by runnings cells sequentially without problems.
- Code is in Python and comments are added.
- Function to test August transaction data is also present.
- Visualisations also present in the notebook.

Transaction Trends

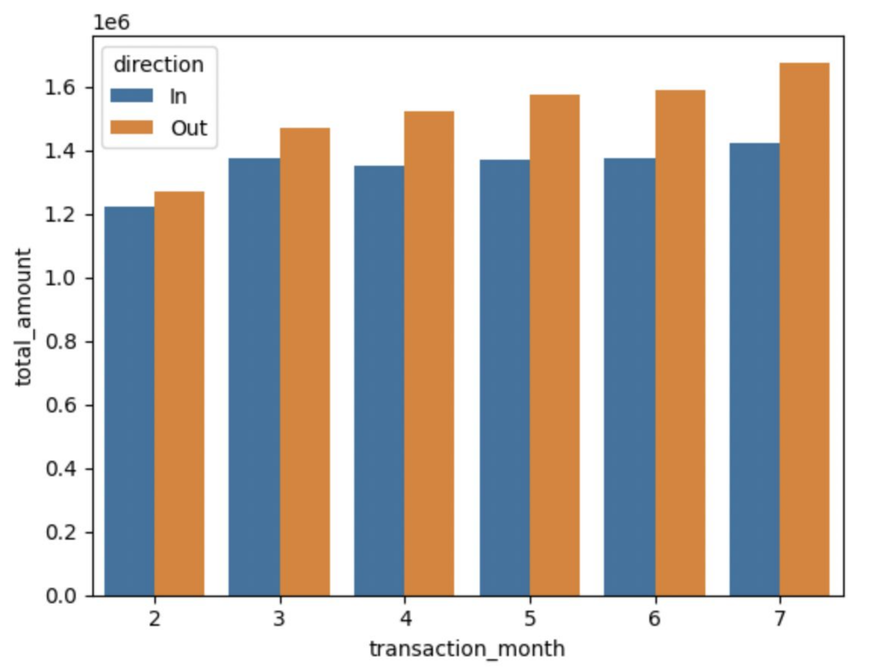
- Tuesday has highest value of expenditures recorded over time.
- Saturday & Sunday have least least value of income reported.
- Monday accounts for the lowest amount spent, but average spend rate on Monday is the highest.

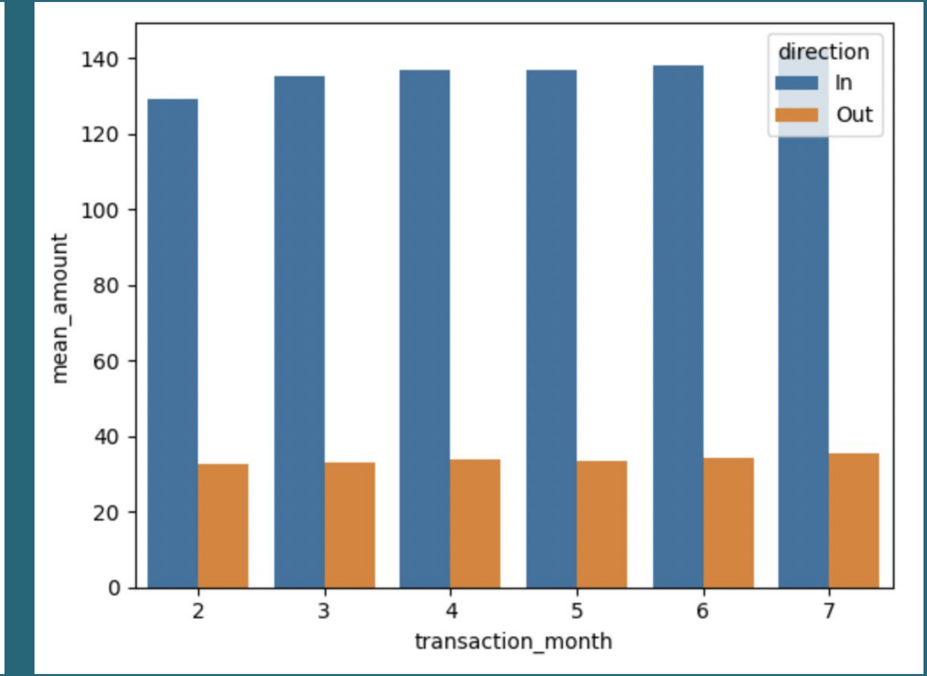




Transaction Trends

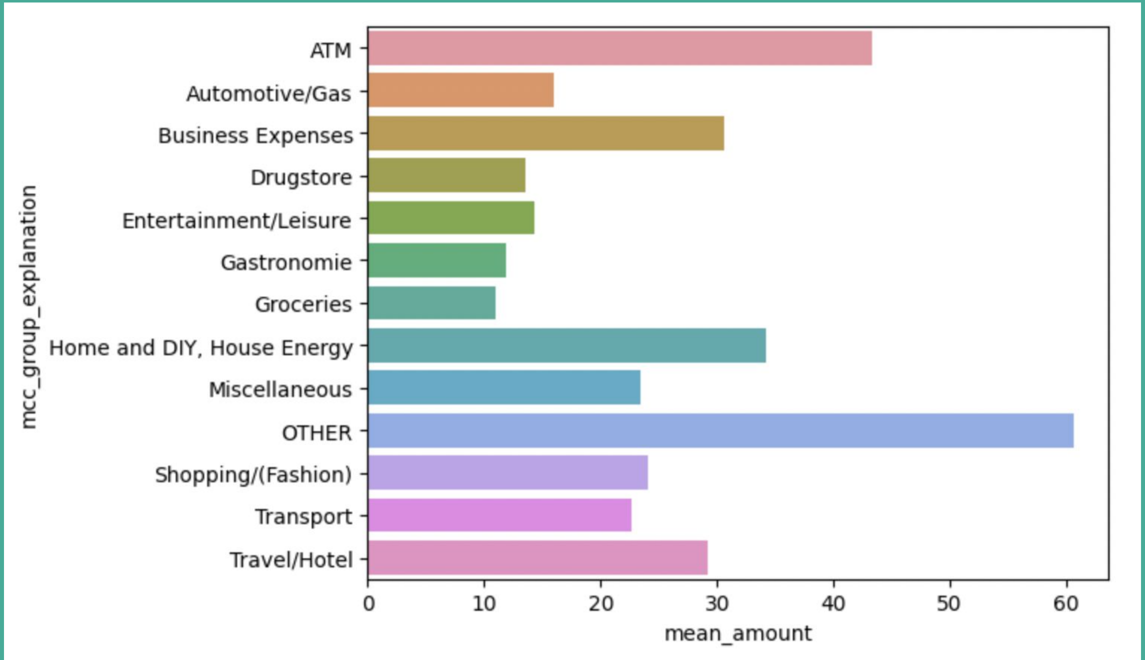
- Transaction value (both Income and Spend) are growing rapidly.
- Mean transaction value is also growing, but at a slower pace.





Transaction Trends

- Average Withdrawals/Expenditure is highest through ATM, Transferwise transfers and Home / Home Energy expenses.
- Least expenditure is done on Leisure / Gastronomie / Groceries.



Assumptions

- **Motivation**

  The key problem is to understand the financial well being of n26 consumers and their financial future.

- **Methodology**

  Regression is a way to predict income and expenditure in future. This helps us solve our problem and reach closer to our goal of understanding consumers. We can also use RFM / Clustering analysis to understand the consumers better.

- **Features provided**

  We are provided with categorical features about how a user transacts (agent, transaction type) and what the user spends on (mcc category), along with the transaction amounts.

- **Features engineered**

  Transactional features, time based features, user activity, user platform age are some categories of features developed.

RFM Analysis

- To be a top tier customer, you should have spent recently (<2 days), transact regularly (>41 times) and have spent a good amount already (>1264 n26 currency).
- N26 should target high affluent dormant users, ie. users who have not transacted in the recent times (<44 days), and have shown high frequency and monetary aspects.
- N26 should be aware of potentially risky high affluent users in the system. These are users who have been transacting recently, at high volumes and at high frequency.

Results & Performance Metrics

- KMeans

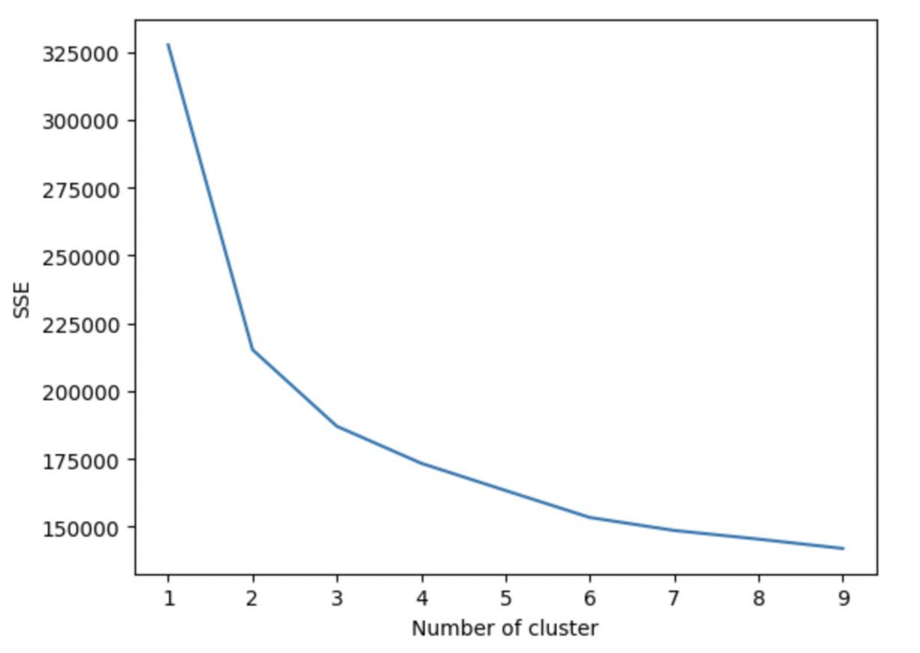
  To find out more about the customers and see if any common pattern exists in the transactional pattern of given users.

  Metrics: SSE, Inertia, Model score

- Regression

  To find predict user transactional behaviour (income / expenditure over the coming months)

  Metrics: MSE, RMSE, R2



Steps to integrate results into app

1. To conduct some more analysis

   Additional effort can be put into creating and experimenting new features. Also, can experiment by using additional sources of data and data points related to consumers.

2. Backtest improved results and evaluate other metrics

   These features once reevaluated should be backtested on historical data to assess model performance.

3. Refactoring code and make code production ready

   Ensuring that code is well written without redundancies and well tested.

4. Assessing code performance

   Entire pipeline (python, query, ML model) should be assessed from response times and memory consumption perspective.

5. Deploying microservice and ensuring uptime
6. Provide endpoints from the microservice and send out results
7. Storing relevant data to a data lake for historical purposes

Thank you!

Abhinav Ralhan
abhinav.ralhan1@gmail.com