MACQUARIE University
SYDNEY·AUSTRALIA

STAT8107 / COMP8107

Statistical (and Machine) Learning (Methods)

Case Study 2025


# Predict Prevent Preserve: Tackling Attrition at Tifosi Bank

`Analyttica` is a leading data analytics firm that delivers advanced analytical solutions across multiple industries, including finance, healthcare, and insurance. With a strong emphasis on data-driven decision-making, `Analyttica` empowers organisations to optimise operations, enhance customer experiences, and effectively manage risk. The company's mission is to transform business challenges into analytical solutions that drive measurable impact.

By harnessing the power of data, `Analyttica` supports transformative growth through tailored strategies and end-to-end analytics solutions. Their expert team of analysts and data scientists collaborates closely with clients to understand each organisation's unique needs, craft customised analytics strategies, and implement solutions that align with key business objectives.

Through its **strategic analytics consulting services**, `Analyttica` employs a broad suite of techniques – including statistical analysis, machine learning, and predictive modelling – to uncover actionable insights from complex datasets. Two of the company's standout capabilities are:

- Data Visualisation: Translating intricate data into clear, actionable insights and compelling visual narratives.
- Advanced Analytics & Modelling: Using predictive analytics, statistical modelling, and machine learning to forecast trends, optimise decision-making, and foster innovation.

# Decoding Attrition: A Data Scientist's First Challenge

You have recently joined `Analyttica` as a **Data Scientist**, and your first project involves working with `Tifosi Bank`, a financial institution seeking to better understand customer behaviour in its credit card division. The bank has observed a growing concern about customer churn – that is, when existing clients stop using their credit card services and ultimately end their relationship with the bank. Churn can have a significant impact on long-term profitability, especially when it involves high-value customers who contribute more in terms of fees, spending, and loyalty.

`Tifosi Bank` aims to identify the key drivers behind customer attrition and develop strategies to retain high-value clients. Your role is to analyse their credit card customer dataset – `bank_churners.csv`, build predictive models to forecast churn, and deliver actionable insights that inform retention efforts.

By enabling `Tifosi Bank` to anticipate which customers are at risk of leaving, your analysis will help them take proactive measures – such as personalised outreach or improved service offerings – to reverse potential churn and strengthen customer loyalty.

## Turning Insight into Action: Why This Case Study Matters

This case study represents a real-world application of data science to a high-impact business problem. Unlike structured assignments with clearly defined tasks and outcomes, this case study offers a more open-ended and realistic problem-solving scenario – one that mirrors the complexity and ambiguity analysts encounter in industry settings. Rather than following a prescriptive set of steps, you are encouraged to make judgement calls, iterate on your approach, and explore multiple pathways to generate meaningful insights.

This mirrors real-world practice, where problems rarely come with instructions, and multiple valid approaches can exist. There is no single "correct" model or approach; rather, this case study invites creative thinking, iteration, and critical evaluation at every stage of the data science workflow. From initial data exploration and feature engineering, through predictive modelling and evaluation, to the communication of actionable insights.

Your analysis will not only support `Tifosi Bank`'s strategic objectives but also demonstrate how data science can be applied to solve pressing business challenges with measurable impact. Customer attrition represents a critical risk to service-based institutions like `Tifosi Bank`, where retaining existing high-valued customers is both more cost-effective and more strategically valuable than acquiring new ones. Identifying the factors that influence churn, and predicting which customers are likely to leave, allows the bank to take timely and targeted actions that improve retention and foster long-term customer loyalty.

This case study also reinforces the importance of domain understanding, stakeholder alignment, and clear communication – qualities that distinguish impactful analysis from purely technical output. Ultimately, the work undertaken here reflects the broader role of analytics in today's data-driven economy: not only to interpret what has happened, but to anticipate what might happen next, and to inform decisions that drive meaningful change.

## The Final Stretch: Storytelling with Data

Your task is to analyse the provided dataset and build models to predict the likelihood of credit card churn at `Tifosi Bank`. Using statistical and machine learning techniques, you will explore which features – such as spending behaviour, account activity, or demographic characteristics – are most predictive of customer attrition. Beyond identifying key risk indicators, you are also expected to offer practical recommendations the bank could implement to retain valuable customers – for example, through personalised outreach or revised service offerings.

Unlike a traditional scientific report, this case study does not require a literature review or an in-depth methodology section. The bulk of your technical work, including data tidying, model fitting, and performance evaluation – should be handled "behind the scenes". Your primary focus is on communicating the results: translating analysis into a clear, coherent narrative that is both engaging and informative for non-technical stakeholders at `Tifosi Bank`.

Refer to our full data science workflow in Figure 1. While the Assignment led you through each phase of the process step by step, this case study invites you to take a more open-ended, end-to-end approach. You will need to decide how to approach the data, what techniques to apply, and how best to interpret and present your findings. This is your opportunity to demonstrate not just your technical abilities, but also your capacity to think critically, draw meaningful conclusions, and communicate insights that drive business value. This case study places particular emphasis on the final stage of the workflow: transforming complex analysis into insights that inform decisions – a skill that lies at the heart of data storytelling.

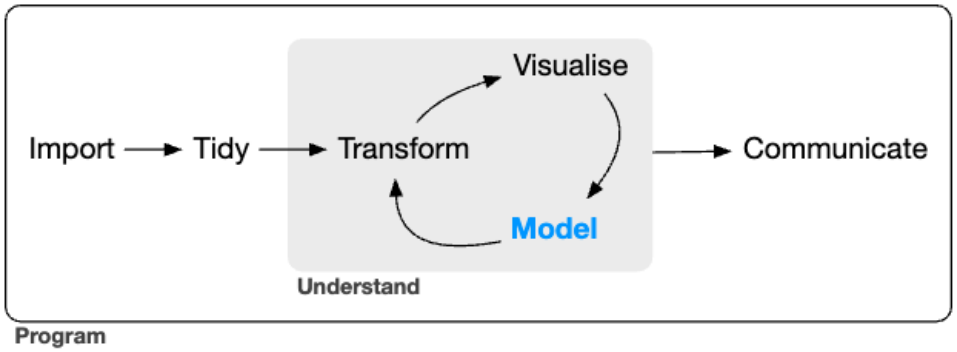


Figure 1: Data Science Workflow from R4DS

# Your Deliverables

You are to produce a **Stakeholder Insight Report** of up to **12 pages** plus **1-page Executive Summary** at the beginning. Use either `Quarto` or `R Markdown` to create a single PDF document. The report should be informative, engaging, and accessible to stakeholders at `Tifosi Bank` – including managers and decision-makers who may not have a deep background in statistics, data science, or finance.

Visual elements such as graphs, tables, and infographics are strongly encouraged to help communicate your story. Avoid including `R` code unless it directly supports your narrative. References are optional but should be limited to five.

### Stakeholder Insight Report Structure Guidelines

This report is designed to mirror how data analytics is communicated in business settings. A clear, logical structure will help your audience quickly grasp the core message and its implications. Please follow the structure outlined below and adapt it as needed to best support your analysis:

1. **Title** A catchy and informative title that captures the essence of your analysis.
2. **Executive Summary** A brief (1-page) overview of your purpose, key findings, and recommendations. This should be understandable on its own, giving busy stakeholders a snapshot of your entire report.
3. **Introduction** Set the scene by introducing the issue of credit card churn and why it matters to stakeholders at `Tifosi Bank`. Briefly highlight the business context and relevance of your analysis.
4. **Analytical Overview** Introduce the dataset and provide a high-level summary of your approach. Keep technical detail minimal – focus on what was done, not how. Use plain language to explain your process.
5. **Key Insights** Present your most important findings clearly and concisely. Use visualisations to bring the data to life and highlight patterns, trends, and risk factors.
6. **Strategic Implications** Discuss what these insights mean for the business. What actions should `Tifosi Bank` consider? How might these findings shape retention strategies or customer engagement?
7. **Conclusion** Wrap up your analysis by reinforcing your key messages and their business value. You may also reflect briefly on limitations or suggest future directions for further investigation.

*This case study is marked out of a total of 100. For a clear breakdown of mark distribution and detailed expectations, please consult the accompanying marking rubric. It outlines the criteria used to assess each component of your work and offers guidance on how to achieve higher marks.*

## Variable Descriptions

| Variable | Description |
|---|---|
| CLIENTNUM | Unique identifier assigned to each credit card account holder |
| Attrition_Flag | Indicates whether the customer has attrited (closed the account) or is an existing customer |
| Customer_Age | Age of the customer in years |
| Gender | Customer's gender: M = Male, F = Female |
| Dependent_count | Number of dependents reported by the customer |
| Education_Level | Highest level of education attained (e.g., High School, Graduate, Post-Graduate, etc.) |
| Marital_Status | Marital status of the customer (e.g., Married, Single, Divorced, Unknown) |
| Income_Category | Estimated annual income bracket (e.g., < \$40K, \$40K–\$60K, > \$120K, etc.) |
| Card_Category | Credit card type issued to the customer (e.g., Blue, Silver, Gold, Platinum) |
| Months_on_book | Duration of the customer's relationship with the bank (in months) |
| Total_Relationship_Count | Total number of accounts or products held with the bank |
| Months_Inactive_12_mon | Number of months the customer was inactive in the past 12 months |
| Contacts_Count_12_mon | Number of customer service contacts made in the past 12 months |
| Credit_Limit | Maximum credit limit assigned to the customer's credit card |
| Total_Revolving_Bal | Outstanding revolving balance on the customer's card |
| Avg_Open_To_Buy | Average available credit amount calculated as Credit Limit minus Revolving Balance |
| Total_Amt_Chng_Q4_Q1 | Relative change in the total transaction amount from the first quarter (Q1) to the fourth quarter (Q4) |
| Total_Trans_Amt | Total dollar value of transactions made over the last 12 months |
| Total_Trans_Ct | Total number of transactions made over the last 12 months |
| Total_Ct_Chng_Q4_Q1 | Relative change in the number of transactions from the first quarter (Q1) to the fourth quarter (Q4) |
| Avg_Utilization_Ratio | Average ratio of credit card balance to credit limit |

# Practical Tips for a Smoother Workflow

- **Expand Your Modelling Toolbox**
  - The `tidymodels` framework provides a consistent and user-friendly interface for training a wide range of models. Beyond the methods introduced in class, you are encouraged to explore more advanced options, such as gradient boosted trees (e.g., using the `xgboost` engine). A full list of supported models and engines is available here: https://www.tidymodels.org/find/parsnip/
- **Demystify Your Predictions**
  - While the `tidymodels` framework does not include built-in tools for model explanation, it integrates smoothly with companion packages like `vip` and `DALEX` for interpreting model results. We've already explored `vip`, which supports model-based explanation methods. These approaches leverage the internal structure of the model – such as coefficients, splits, or weights – making them efficient and tailored to specific model types. For model-agnostic interpretation, try `DALEX` / `DALEXtra`. These methods treat your model as a black box and can be applied to any algorithm, making them a flexible option when structure-specific tools are unavailable. Try to incorporate at least some level of interpretability in your workflow. Knowing why your model makes a prediction is just as critical as knowing what it predicts. For practical examples and more detail, see: https://www.tmwr.org/explain.
- **Accelerate Computation with Parallel Processing**
  - As your modelling becomes more complex, training multiple models or tuning hyperparameters can be time-consuming. To speed things up, consider using parallel processing. In `R`, the `doParallel` and `foreach` packages integrate smoothly with `tidymodels`, enabling you to run computations across multiple CPU cores. For more details, check out SGTA 06.
- **Use Caching to Save Time**
  - For code chunks that take a long time to run – such as those involving `tune_grid()`, enable caching by setting the chunk option `cache = TRUE`. This allows your report to reuse previously computed results, significantly reducing rendering time during draft development. The cache is automatically invalidated and refreshed when the code or chunk options change. For deeper insights, see Yihui Xie's excellent blog post on cache invalidation.

*Due date: Friday 16 May 2025, 11:55pm*

## Submission Instructions

- This case study is worth **35%** of your total unit mark.
- The overall total for this case study is **100 marks**.
- `R` is the only programming language allowed for this case study.
- You will need to make two submissions on iLearn:
  1. Submit a single `PDF` file containing your entire work, including plots, discussions, and results.
     - Other formats (e.g., Word, HTML) will NOT be accepted.
  2. Submit your `Quarto` (`.qmd`) or `R Markdown` (`.Rmd`) file to the Source Code Submission link. This file should serve as the fully reproducible backend for your report and contain the entire `tidymodels` workflow.
- It is strongly recommended that you use `Quarto` or `R Markdown` to typeset your case study.
  - A Small Tutorial on `Quarto` & `R Markdown` is available on iLearn.
  - These tools enable a **literate programming** approach, integrating narrative text, executable code, and dynamic output in a cohesive, reproducible workflow.
  - This ensures clarity, transparency, and reproducibility in your work, making your analysis easier to follow and replicate.
- Keep your explanations concise and include only necessary `R` output. Avoid excessive or redundant code/output.
- Page Limit: Your submission must not exceed **13 pages**.
  - A 10% penalty will apply for every additional page beyond the limit.
- Late Submissions:
  - Standard late penalties apply unless you have an approved Special Consideration request.
  - Refer to iLearn or the Unit Guide for details on late penalties.
- Submission Method: Submit your case study report via Turnitin link and source code to the source code submission link on iLearn.