# Santander Product Recommendation

**Maria Alice Vieira and Aritra Ray**

## Problem Statement

Santander Bank's current product recommendation system is inefficient, leading to an uneven customer experience. Some customers receive an abundance of recommendations, while others receive very few or none at all. This disparity results in missed opportunities for both the bank and its customers. The challenge is to develop a more sophisticated and personalized recommendation model that can accurately predict which products each customer is most likely to need or want in the last month of the dataset, based on their historical behavior and transaction data. In other words, they need to know where the demand is so they are able to supply. We will create a machine learning model that will assist on which customers to target along with which products from the bank they are more likely to buy.

To illustrate, I will use age as one of the factors that determines what customers are looking for: customers in their 20s would likely be interested in student loan refinancing options and high-yield savings accounts that could help manage debt and start building wealth. As they enter their 30s, the focus could shift to mortgage products and family-oriented credit cards with rewards. The examples keep changing as they get older, but there are other important factors to take into account as well, such as gender, income and others.

## Articulation of Value

Implementing an improved product recommendation system for Santander Bank offers significant value across multiple dimensions within the bank:

- **Enhanced Customer Experience**: By providing more relevant and timely product recommendations, customers will feel that the bank understands and anticipates their needs. This approach can lead to increased customer satisfaction and loyalty, which in turn, becomes a **competitive advantage**: In the highly competitive banking sector, offering superior personalized services can differentiate Santander from its competitors, potentially attracting new customers and retaining existing ones.

- **Increased Product Adoption**: More accurate recommendations are likely to result in higher conversion rates, as customers are presented with products that genuinely align with their financial needs and behavior patterns, and as a result, higher revenue for the bank.
- **Operational Efficiency**: A well-tuned recommendation system can reduce the workload on customer service representatives and financial advisors, allowing them to focus on more complex customer needs.
- **Data-Driven Decision Making**: The insights gained from the recommendation model can inform product development and marketing strategies, allowing the bank to align its offerings even more closely with customer needs and preferences.
- **Risk Mitigation**: Better understanding of customer needs and behaviors can help in early identification of potential churn risks or opportunities for upselling, allowing for proactive measures.

Overall, by addressing the current limitations in product recommendations, Santander Bank could improve its customer relationships, increase its revenue, and strengthen its market position. This project aims to transform raw customer data into actionable insights, driving both customer satisfaction and business growth.

**Potential Economic Value**

The potential economic value of implementing the improved recommendation system is estimated at $20,000,000 for the first year based on the below assumptions. This significant increase in revenue justifies the investment in developing and implementing the new machine learning-based recommendation system. This is a simplified calculation and does not account for implementation costs, maintenance, or other factors that might affect the actual return on investment. However, it provides a clear indication of the potential value that can be generated through this project.

Assumptions:

1. Current customer base: 1,000,000 customers
2. Conversion rate of recommendations: 5%
3. Average revenue per product sold: $200
4. Expected increase in conversion rate after implementing the new system: 10%

5. Duration of impact: 1 year

Calculation

Step 1: Calculate the current annual revenue from recommendations

- Current annual revenue

Current customers × Current conversion rate × Average revenue per product

1,000,000 × 0.05 × $200 = $10,000,000

Step 2: Calculate the projected annual revenue with the new system

- New conversion rate = Current conversion rate + Expected increase

5% + 10% = 15%

- Projected annual revenue = Current customers × New conversion rate × Average revenue per product

1,000,000 × 0.15 × $200 = $30,000,000

Step 3: Calculate the potential economic value (increase in revenue)

- Potential economic value = Projected annual revenue - Current annual revenue

$30,000,000 - $10,000,000 = $20,000,000

**Project plan**

| Week No. | Expected Outcome | Deliverables |
|---|---|---|
| W1 | 1.Identify problem statement 2.Find an appropriate dataset 3.Articulate the value of solving the problem and develop a business case 4.Setup environment | 1. Written problem statement<br>2. Github repository setup |
| W2 | 1.Ingest the Dataset 2.Identify dependent and independent variables | 1. Jupyter notebook with basic exploration |

| | | |
|---|---|---|
| | 3.Impute missing values and conduct basic analysis | 2. Code documentation upload on GitHub |
| W3 | 1.Split the dataset 2.Conduct deep EDA | 1. Jupyter notebook with EDA results |
| W4 | 1.Preprocess dataset(handle missing values, scale/normalize, encode categorical variables) 2.Ensure data consistency across training, validation, and test sets | 1. Cleaned dataset and preprocessing code<br>2. Update Jupyter notebook |
| W5 | 1.Engineer features 2.Reduce dimensionality 3.Finalise data for training and evaluation | 1. Updated notebook<br>2. Code merged to GitHub |
| W6 | 1.Build a baseline model 2.Hyperparameter tuning 3.Select evaluation metric | 1. Updated notebook<br>2. Model evaluation results.<br>3. All artifacts merged into GitHub. |
| W7 | 1.Develop a more sophisticated model(2nd model) 2.Continue hyperparameter tuning 3.Evaluate model performance | 1. Jupyter notebook with second model<br>2. Performance comparison with baseline |
| W8 | 1.Develop a more sophisticated model(3rd model) 2.Fine-tune hyperparameters 3.Compare results | 1. Jupyter notebook with third model<br>2. Performance comparison with baseline |
| W9 | 1.Compare all models based on key metrics 2.Select the best-performing model 3.Calculate final performance on the test dataset | 1. Updated docs<br>2. Winning model selected |
| W10 | 1.Improve results but appending | 1. Improved model results |

| | | |
|---|---|---|
| | datasets 2.Retrain the model | 2. Updated documents |
| W11 | 1.Perform feature importance analysis to explain model behavior 2.Identify potential biases in the data or model predictions | 1. Report explaining the model and identifying risks<br>2. Updated documents |
| W12 | 1.Save trained model 2.Develop monitoring plan | 1. Model packaged and saved<br>2. Monitoring plan documented<br>3. Update GitHub |
| W13 | 1.Compile all previous weeks' work into a comprehensive report 2.Ensure all code, documentation, and results are well-organized in GitHub 3.Finalize and submit the project. | 1. Final report written<br>2. Final code uploaded<br>3. GitHub updated |

**The Dataset**

The dataset we will be using for this project is the Santander Product Recommendation dataset, obtained from Kaggle. This dataset was originally part of a Kaggle competition in late 2016, providing a rich source of real-world banking data for our analysis.

Santander has provided 1.5 years of sample customer behavior data held between January 2015 to May 2016. There are approximately 13.5 million records for the training dataset and a bit over 900k for the test set. The first 24 columns are variables that give us information about the customer's demographics and relationship with the bank (e.g., tenure). The last 24 columns (for a total of 48 columns) are binary indicators for the products offered by the bank.

### Type of modeling

To handle the challenge of developing a personalized recommendation system for Santander Bank, supervised learning is the best modeling technique. Specifically, the purpose of this classification task is to predict which products each client is most likely to be interested in during the dataset's last month. This is a multi-class classification problem since each consumer may be interested in a variety of goods. For example, we will forecast target variables such as ind_ahor_fin_ult1 (savings account), ind_cco_fin_ult1 (current accounts), and ind_tjcr_fin_ult1 (credit card), among others. Each product type is assigned to a class in the classification issue, and the model is trained to predict which of these product classes each consumer is most likely to interact with in the dataset's final month. Using historical data on client behavior and transaction patterns as characteristics, the model will assign probabilities to each product class, allowing the bank to make more tailored recommendations and enhance target accuracy.

This dataset is highly suitable for addressing our problem statement of improving Santander's product recommendation system as the comprehensive customer profiles will enable us to segment customers and identify patterns in product preferences across different groups; the historical product ownership data spanning 1.5 years allows us to track changes in customer portfolios over time, which is crucial for predicting future product needs; we can develop a recommendation system that accounts for various financial needs and cross-selling opportunities; and the monthly data points allows us to capture seasonality and trends in product adoption, enhancing the accuracy of our predictions. Moreover, being derived from an actual bank, this dataset reflects genuine customer behavior, making our model more applicable to real-world scenarios.