NUS DATATHON 2025

TLDR

We invite datathon participants to build a model to recommend the most suitable financial advisors (agents) to customers.

Overview

Optimizing Financial Advisor Matching with Data Science

The objective of this datathon is to develop a model that recommends the **best financial advisors** for **individual customers**. This can be achieved through various techniques such as recommendation systems, supervised learning models, or unsupervised learning models. The model will be used to assign the most suitable financial advisors to customers, thereby enhancing engagement and improving the likelihood of successful policy conversions.

Improving Customer Engagement and Revenue Growth

By identifying the optimal match between customers and financial advisors, the model will help increase the likelihood of customers purchasing financial products and services. This will result in higher conversion rates, greater customer satisfaction, and, ultimately, increased revenue.

Data

Background

Financial advisors play a crucial role in guiding customers toward the right financial products, tailoring recommendations to individual needs and goals.

We present a curated sample of 20,000 customer profiles, ~30,000 policies, and ~10,000 agents. With a strong commitment to **privacy and data security**, the dataset has been fully anonymized. All **personally identifiable information (PII)**—including names, NRICs,

addresses, and other sensitive data—has been completely removed. Synthetic identifiers for customers have been generated to prevent any real-world identification.

Additionally, to ensure confidentiality and privacy, product names have been masked, and numerical values have been slightly adjusted, maintaining the integrity of the data while obscuring exact details.

We emphasize the importance of data security in this initiative and encourage innovative, privacy-conscious solutions.

Data Dictionary

1. Agent (financial advisor) information table:

- **`agntnum`**: unique identifier for agent
- **`agent_age`**: agent's age
- **`agent_gender`**: agent's gender
- **`agent_marital`**: agent's marital status
- **`agent_tenure`**: agent's tenure with the company
- **`cnt_converted`**: count of policies converted by agent
- **`annual_premium_cnvrt`**: total annual premium from converted policies by agent
- **`pct_lapsed`**: percentage of policies that are currently lapsed
- **`pct_cancel`**: percentage of policies that are currently cancelled
- **`pct_inforce`**: percentage of policies that are currently in force
- Percentage of customers' gender handled by the agent:
  - **`pct_sx0_unknown`**, **`pct_sx1_male`**, **`pct_sx2_female`**
- Percentage of products sold by agent:
  - **`pct_prod_0_cnvrt`**, **`pct_prod_1_cnvrt`**, ..., **`pct_prod_9_cnvrt`**
- Percentage of customers' age groups handled by the agent:
  - **`pct_ag01_lt20`**, **`pct_ag02_20to24`**, ..., **`pct_ag10_60up`**
- **`cluster`**: an old segment that the agent belongs to
- **`agent_product_expertise`**: a list of products the agent is comfortable selling (based on feedback)

2. Policy information table:

- **`chdrnum`**: unique identifier for policy
- **`agntnum`**: unique identifier for agent
- **`secuityno`**: unique identifier for customer
- **`occdate`**: inception date of policy
- **`annual_premium`**: annual premium
- **`product`**: product of the policy
- **`product_grp`**: product group of the policy
- **`flg_main`**: flag indicating the main policyholder
- **`flg_inforce`**: flag indicating a policy that is in force
- **`flg_cancel`**: flag indicating a policy that is cancelled
- **`flg_expire`**: flag indicating a policy that expired
- **`flg_converted`**: flag indicating a policy that is converted
- **`cust_age_purchase_grp`**: customer's age group at purchase
- **`cust_tenure_at_purchase_grp`**: customer's tenure at purchase

3. Client information table:

- **`secuityno`**: unique identifier for customer

- **cltsex**: gender of customer
- **cltdob**: date of birth of customer
- **marryd**: marital status of customer
- **race_desc_map**: race of customer
- **cltpcode**: customer postal code
- **household_size**: household size (based on postal code)
- **economic_status**: economic status (based on postal code)
- **family_size**: family size (based on postal code)
- **household_size_grp**: discretized household size of customer
- **family_size_grp**: discretized family size of customer

4. Sample final modeling table: (Sample final modeling table is provided as an example of what the final table may look like. Participants are not required to replicate or obtain that exact table.)

- **chdrnum**: unique identifier for policy
- **agntnum**: unique identifier for agent
- **secuityno**: unique identifier for customer
- **occdate**: inception date of policy
- **product_grp**: product group of the policy
- **cust_age_at_purchase_grp**: customer's age group at purchase
- **cust_tenure_at_purchase_grp**: customer's tenure at purchase
- **cltdob**: customer’s date of birth
- **marryd**: marital status of customer
- **race_desc_map**: race of customer
- **household_size_grp**: discretized household size of customer
- **family_size_grp**: discretized family size of customer
- **cluster**: old segment that the agent belongs to (chosen as the target label)

Judging Criteria

Here's a table summarizing the percentage allocation for each judging criterion:

Criterion	Weight (%)	Description
Creativity and Innovation in Approach	**30%**	• Novel techniques and unique problem-solving approaches.
Performance and Efficiency	**30%**	• Model Performance (e.g., Precision@k, Recall@k, NDCG@k, AUC, F1-Score, etc.) and efficiency of the recommendation/classification model. • Participants **must** split the provided dataset into **80% training data** and **20% testing data** before training their models.
Ethical Use of Data & Fairness	**20%**	• Ethical use of data and fairness of the model.
Quality of Report and Explanation	**20%**	• A well-structured report that effectively communicates their approach, methodology, and findings. • Provide clear justifications.

Data Usage Agreement

- **Public sharing** of the dataset is prohibited (e.g., GitHub, portfolio platforms, social media).
- Only individuals **18 years and older** are allowed to participate to ensure enforceable agreements.

Submission Instruction

Participants are required to submit a Google Drive folder link (with **"Anyone with the link can edit"** permissions) to **ensure both the folder and individual files are accessible**.

Folder name format: `NUS_DATATHON_CAT_A_<TEAM NUMBER>`

The folder must contain the following:

1. **Submission Notebook**:
    - File format: `.ipynb`
    - File name: `CAT_A_<TEAM NUMBER>.ipynb`
        - `Replace <TEAM NUMBER> with your actual team number. Avoid spaces and special characters.`
        - `Example: CAT_A_304.ipynb`
    - Ensure your notebook includes all steps of your solution
2. **Requirements File**:
    - File format: `requirements.txt`
    - List all Python packages and versions required to run your notebook.
    - Example:
        - pandas==1.3.3
        - numpy==1.21.2
        - scikit-learn==0.24.2
        - matplotlib==3.4.3
        - joblib==1.0.1
3. **Model File** (if applicable):
    - File format: `eg. model.joblib, model.pkl`
    - Include the saved model if applicable.
4. **README File**:
    - File format: `README.pdf`
    - Clearly explain:
        - Instructions for setting up the environment.
        - How to run your notebook and reproduce results.
        - Any specific instructions required for executing the model.
        - Key insights and findings from your solution.
5. **Report**: eg.
    - **Introduction**
    - **Dataset Overview**
    - **Methodology**
    - **Results**
    - **Insights**
    - **Conclusion**
6. **Important Note**!

- **<u>Do not</u> upload the dataset** in your submission folder.
- **Public sharing** of the dataset is **prohibited** (e.g., GitHub, portfolio platforms, social media).
- Finalists will be informed on **13th Feb** and will have to prepare a **presentation deck**.

FAQ

- **What is the objective of the datathon?**
    - The datathon aims to develop models that can match a customer with the best financial advisor to increase the chances of customers purchasing financial products and services using a more diverse set of features or interesting techniques.
    - Participants should think about what to use as the labels and are free to have their own approaches. Potential approaches can be to predict the individual agents, the clustered agents or individual agents given a product.
    - We want participants to innovate and learn various ways available to solve a business problem
- **How will these models benefit Singlife and its customers?**
    - The model will help increase the chances of customers purchasing financial products and services, resulting in higher conversion rates, and ultimately greater customer satisfaction.
- **What data will participants work with?**
    - Participants will be provided with anonymized customer, policy and agent related data, ensuring privacy and compliance with data protection regulations.
- **What is the significance of this hackathon for the field of data science in insurance?**
    - This event is an opportunity for participants to apply data science in a real-world setting, potentially leading to significant advancements in machine learning data driven revenue generating strategies.