INSTITUTO DE TECNOLOGIA E LIDERANÇA – INTELI

OPTIMIZING THE ACHIEVEMENT OF NNM TARGET FOR FINANCIAL ADVISORS

RAPHAEL LISBOA ANTUNES

SÃO PAULO
2025

RAPHAEL LISBOA ANTUNES

OPTIMIZING THE ACHIEVEMENT OF NNM TARGET FOR FINANCIAL ADVISORS

Report presented to the Instituto de Tecnologia e Liderança as an artifact of Module 14 of the Bachelor's Degree in Computer Science, in the corporate track.

Supervisor: Prof. Dr. Rafael Will Macedo de Araujo

São Paulo

2025

ABSTRACT

In the financial market, investment advisors encounter significant challenges in achieving their net fundraising targets, such as *Net New Money* (NNM). This project seeks to solve the lack of predictability in target attainment by implementing an algorithm using *Machine Learning* and Statistical Models. This approach aims to enhance efficiency and strategic decision-making for both advisors and team leaders. To achieve this goal, in this project we will structure and seek a better understanding of the dataset. To do so, we will formulate hypotheses, conduct experiments, and draw conclusions in a way that allows for better refinement in training the models.

Keywords: Net New Money, Predictability, Machine Learning, Investment Advisors, Database Construction, Hypothesis, Validation, Conclusions.

LIST OF ABBREVIATIONS AND ACRONYMS

NNM	Net New Money
AUC	Assets under Custody
KPI	Key Performance Indicator
CGE	Code used to anonymize the advisor's identity.

TABLE OF CONTENTS

1 INTRODUCTION

1.1 Problem Statement

In the B2C segment of the financial market, clients with invested wealth exceeding a minimum threshold gain access to investment advisors who support them in allocating resources according to their financial goals and risk profile. These advisors are assigned a range of monthly targets across different financial products, with Net New Money (NNM) standing out as one of the most challenging, given the requirement for net inflows of new funds.

In addition to advisors, professionals in leadership roles, known as Team Leaders, must have access to a forecast of target achievement at the end of the month based on daily behavior throughout the period. This allows for identifying deviations, helping an advisor who initially would not meet their target to realign their strategy and, therefore, have a chance to achieve the objective by the end of the period.

1.2 Project Objectives

To address the proposed issue, this project aims to solve the pain point of advisors and team leaders who lack an accurate prediction of target achievement at the end of the month. Therefore, an algorithm will be developed using Machine Learning and Statistical Models.

To achieve this goal, this module focuses on understanding the data and identifying the features most correlated with achieving the NNM target. For this purpose, the CRISP-DM methodology will be implemented. According to Shearer (2000), CRISP-DM is an iterative six-phase process used in data mining projects: it begins with business and data understanding, followed by data preparation, modeling, and evaluation, and concludes with the deployment of results. Its objective is to ensure that the analysis project delivers practical value to the business. Throughout the sprints, hypotheses will be formulated, experiments conducted, and conclusions drawn, focusing on the key variables in the dataset.

2 DATASET

In this session, we describe the construction of the dataset which will be used during the development of this project. To achieve this, firstly it was essential to understand some key concepts regarding the behavior of tables in a relation database.

When building the dataset, one of the first steps was to identify the entity tables available in the data lake. These tables focus on a single topic in a deep and structured way, meaning they contain only information related to the subject they represent. In this case, some of the tables used included those related to clients, advisors, AuC, NNM, goals, base coverage, and pipeline. Each of these tables is dedicated exclusively to the topic indicated by its name, but with a high level of detail, allowing them to be used in various scenarios without being limited to a specific purpose.

Once the tables were mapped, it was necessary to understand the granularity of each one. Granularity refers to the level of detail represented by a table. For example, a single advisor may serve multiple clients, just as a single client may perform NNM in several different ways. Therefore, the NNM table has a finer granularity than the client table, which in turn is more granular than the advisor table. More granular tables contain more detailed information about a given topic.

With these two concepts in mind, unique keys were identified, the remaining information was aggregated so that multiple values were reduced to one, and finally, the tables were joined using these unique keys.

2.1 Dataset variables

For the project dataset, we started with the financial subject tables, which are the most granular, since a client may deal with multiple topics in a single day. After aggregating these tables at the client level, we unified the data by adding columns and linking them through a unique identifier. Then, we reduced this unified table to the advisor-level granularity and integrated it with the other tables that operate at the same level. At the end of the process, the resulting table contains the following variables:

Table 1 - Dataset variables

Advisor	Anonymized by a code (CGE), it represents the advisor responsible for all the features that will be mentioned below
Job Title	Advisor's Job Title
Reference Date	Within a range between January 2024 and January 2025, the dataset provides daily granularity for each advisor
AuC	Represents the advisor's total assets, determined by summing the assets of all clients
Number of Clients	The total number of clients for a specific advisor. The number changes based on the advisor's job title
Average AuC	Represents the ratio between the total AuC and the number of clients
KPI	Represents the amount the advisor is expected to achieve during the month
Target	Represents the percentage of the target achieved
Target Achievement	Financial product tied to the target
New Accounts Ratio	Represents the percentage of clients who have been with the advisor for three months or less
Old Accounts Ratio	Represents the percentage of clients who have been with the advisor for more than three months
Base Coverage	The ratio between the number of productive clients[1] and the total number of clients
Pipeline Count	Number of pipelines
Pipeline Volume	The total financial volume of pipelines

We are interested in a table within this granularity because we want our Machine Learning model to learn the behavior of this employee over time. Therefore, it is essential to provide the model with data that aligns with this objective. If we were to supply data with a finer granularity, such as client-level data, the model would struggle to accurately learn how an advisor behaves.

[1] A client is considered productive when they contact the advisor through a call lasting more than 30 seconds, respond to a WhatsApp message from the advisor within 24 hours, or purchase an investment suggestion.

3 HYPOTHESES

KELLEHER, TIERNEY (2018) explain that hypothesis formulation is a central step in the data science process. Hypotheses are initial assumptions about patterns or relationships in the data, usually based on prior knowledge or preliminary observations. Data science seeks to test these hypotheses through statistical analyses, visualizations, and computational models, with the goal of confirming or refuting these ideas based on empirical evidence. This process is iterative: hypotheses are constantly refined or discarded as the data reveal new insights. The authors also highlight the difference between exploratory analysis, which helps to formulate hypotheses, and confirmatory analysis, which tests specific hypotheses. Furthermore, they emphasize the importance of well-defined hypotheses to ensure clarity, focus, and reproducibility in data science projects.

3.1 Base Coverage

We believe that the main metric contributing to advisors reaching their NNM target is productive contact with clients. This shows that, beyond simply reaching out, the advisor must provide the necessary attention and deliver meaningful results through the conversation. To confirm this, we can look at the goal achievement of advisors with the highest client coverage rates, as well as those with the lowest rates. This way, we can understand whether there is a direct correlation between the two variables.

3.2 Impact of Pipeline Count on the Same Volume

Considering that the pipeline represents only a promise of funding and that assets are only counted toward NNM once they are applied, it is believed that diversifying the pipeline — that is, distributing a given amount across several committed clients rather than concentrating it in one — increases the likelihood of conversion. For example, it is more likely to reach the NNM target with five pipelines of R$100,000 than with a single pipeline of R$500,000. This is because, in a scenario of multiple commitments, the default of one client does not fully compromise the

advisor’s results, whereas the failure of a single concentrated pipeline may result in no value being generated at all.

3.3 New accounts ratio

A study conducted prior to the development of this project indicated that the most favorable period for attracting new funds occurs within the first three months of the client's relationship with the advisor. During this time, the client is still becoming familiar with the new environment and is generally more open to exploring alternatives and making new investments. Based on this, it is hypothesized that advisors with a higher proportion of new clients have a significantly greater likelihood of reaching their NNM targets.

4 Validation of the hypotheses

In this section, we will seek to validate the hypotheses discussed in Section 3. Hypothesis validation is a fundamental pillar of the scientific method, as it allows researchers to test assumptions about observable phenomena in an objective, systematic, and reproducible way. According to CRESWELL (2018), the scientific method begins with the formulation of hypotheses grounded in theory or preliminary observations, which are then subjected to rigorous empirical testing. The importance of validating hypotheses lies in the need to ensure that conclusions are not based on guesswork or bias, but rather on statistical evidence that reliably reflects the studied reality with a high degree of confidence. This process involves collecting representative data, using appropriate statistical tests, and defining clear criteria for the acceptance or rejection of the null hypothesis — a proposition that assumes no effect or relationship between the variables being analyzed.

It is important to clarify that, in this section, graphs supporting the statements made are not presented due to a compliance rule from the partner institution involved in the project. This measure is intended to protect sensitive data that is considered the institution’s intellectual property.

4.1 Base Coverage

As we saw in section 3.1, we aim to determine whether the base coverage variable exhibits a directly proportional behavior in relation to the achievement of the NNM target. During the studies regarding this hypothesis, some points were identified about the behavior of this data in the database developed throughout the project. Some of these behaviors include:

4.1.1 Inaccurate data for the 2024 period

Analyzing the volume of distinct data for this variable over the months, starting from January 2024, it was observed that for the period prior to October 2024, there is no data available for this variable. Upon noticing this, we reached out to the institution responsible for the data to understand the cause of this issue.

As previously explained, throughout the year, advisors have goals based on different KPIs and meeting these goals results in earning points that lead to rewards based on top performance. Base coverage is a KPI that was not considered in the advisors' targets in 2024 and was only included starting in 2025. Therefore, since this KPI was not relevant for determining the monthly winners, there was not a high level of precision in measuring data related to this metric as there is today.

This lack of precision occurred due to several factors, one of the main ones being the freedom advisors had in how they contacted clients. In the past, they could reach out through means such as WhatsApp or personal phone calls, which made it impossible for the institution to track this data. Currently, communication is done through a specific platform that allows for the identification and storage of this information.

4.1.2 Removing underrepresented job titles

We aim, with this project, to develop an intelligent way to predict the achievement of NNM targets by advisors based on their behavior. However, in order to reach this goal, we need to train the model using data that reflects the reality in which it will be applied once it is ready.

Throughout the data analysis process in collaboration with the institution, we discovered that not all advisor job titles compete with each other. Depending on the client profile, whether individual or corporate — the advisors serving each type of client only compete within their respective groups. Focusing on the different types of advisors, we noticed two very clear behavioral patterns: first, the number of advisors working with individual clients is significantly larger than those serving corporate clients. Second, the advisors who deal with corporate clients handle much larger transaction volumes, generating results that are often considered outliers.

Therefore, to train the model using data that more closely represents the advisors who will actually use it once it is ready, we focused exclusively on the advisors who serve individual clients.

4.1.3 Outliers treatment

After removing the 2024 data and excluding advisor roles that focus on corporate clients, we observed that some target achievement values still appeared to be highly irregular and significantly distant from the average. This occurs due to circumstances that are difficult to explain based on the current dataset. Possible explanations include seasonal market conditions during the period in which the targets were achieved, or even factors related to other variables that will be explored later in this study.

Therefore, in order to draw concrete conclusions regarding the specific variable of base coverage, we analyzed the distribution of the data. It was found that, despite the presence of outliers, both the base coverage variable and the target achievement behave in a way that is very close to a normal distribution — with the mode, mean, and median showing extremely similar values.

According to KELLER (2015), knowing the mean and standard deviation allows researchers to extract useful information — if the histogram has a bell-shaped curve normal distribution, the Empirical Rule can be applied: approximately 68% of observations fall within one standard deviation of the mean, 95% within two, and 99.7% within three. This knowledge is essential for identifying values that deviate significantly from the center of the distribution and are therefore considered outliers. As a result, for outlier treatment in this project, we considered as valid only the data falling within the

closed interval ranging from the mean minus one standard deviation to the mean plus one standard deviation.

4.1.4 Hypothesis Conclusion

After processing the aforementioned data, we developed a scatter plot to observe the goal achievement of advisors who exhibit high and low base coverage rates. To determine whether there is a directly proportional or inversely proportional correlation, we use linear regression, which identifies the line that best fits the pattern of the scattered data points. If the line shows an upward trend, it indicates a directly proportional relationship; however, if the line trends downward, it suggests an inversely proportional relationship.

Since we aim to demonstrate a directly proportional correlation — meaning that the advisors who most consistently meet their goals are those with the highest base coverage rates — we expect the line to trend upward. Upon applying the regression line, this is precisely the result we obtained. Thus, we are able to confirm the validity of the proposed hypothesis.

5 Conclusion

Throughout the development of this module, we encountered several challenges that provided valuable lessons along the way. In the previous module, we empirically determined that the data should be processed with daily granularity and in chronological order, in order to yield more accurate results for the project, since the model needs to observe day-by-day progress over time. In this module, we realized that maintaining chronological order is still essential; however, such a fine level of granularity is not strictly necessary. While this level of detail may be useful in the future, for the current analysis of variables and understanding of the data, it is not as crucial as we initially assumed.

As we are working with sensitive institutional data, we had to undergo a long and time-consuming process to gain access to it. This led to a reorganization of the initial plan, as our original goal for this stage was to analyze at least three hypotheses. However, we were only able to reach concrete conclusions for one. Therefore, we have

defined that the next steps will include continuing the study of the remaining hypotheses, while maintaining our focus on training time series models, as already planned for the next module.

REFERENCES

SHEARER, Colin. ***The CRISP-DM model: the new blueprint for data mining***. Journal of Data Warehousing, 2000.

KELLEHER, John D.; TIERNEY, Brendan. ***Data Science***. Cambridge, MA: The MIT Press, 2018.

CRESWELL, John W.; CRESWELL, J. David. ***Research Design: Qualitative, Quantitative, and Mixed Methods Approaches***. Thousand Oaks, CA: SAGE Publications, 2018.

KELLER, Gerald. ***Statistics for Management and Economics***. 11. ed. Boston: Cengage Learning, 2015.