PREDICTIVE ANALYTICS

2024

MARCH 19

Contents

1.0 Introduction

This report encapsulates the essence and objectives of a collaborative initiative undertaken within BlackRock, a prominent figure in global investment management. Our endeavor revolves around crafting tailored investment proposals for a specific subset of BlackRock's client base, primarily focusing on retail clients. Throughout this project, we've had access to historical data and received weekly updates containing new client leads. However, the sheer volume of leads has presented a challenge, as we grapple with the task of effectively prioritizing outreach efforts within our operational constraints.

Our primary aim is to devise a systematic approach to assist BlackRock in determining which clients warrant priority for weekly engagement. This involves the development of predictive models to guide our decision-making process, leveraging insights obtained from data analysis. This report provides insights into the resources at our disposal, including the historical dataset and weekly client leads, and outlines the methodologies employed in our predictive modelling.

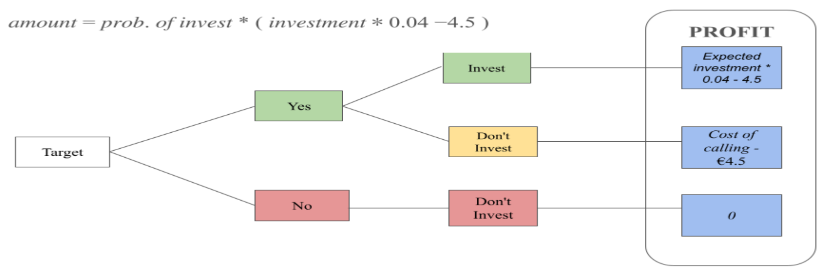
Our overarching objective is to refine BlackRock's client engagement strategies by harnessing the power of data-driven insights. Join us as we navigate the intricacies of client outreach optimization and strive to enhance the efficacy and profitability of BlackRock's interactions with retail clients.

2.0 Business Understanding

BlackRock is a leading global investment management firm, managing about $9.4 trillion in assets. It is known for its extensive offerings in mutual funds and as a major player in the global ETF market. As a new member of the client advisory team, you will work with advisors to create investment proposals for retail clients, utilizing a historical dataset of client profiles and receiving new client leads weekly.

As we know, we can only call a maximum of 1000 clients per week. We need to identify a pool of clients and their likelihood to invest, how much they would invest and generate profits for Blackrock while considering the marketing cost per client call, which is €4.5; and the historical statistics related to the clients' allocation of their current balance to investment products, which overall, it rounds about 28%; we also know that Blackrockmakes an average profit of 4.0% of the amounts that the clients invest.

We started by analyzing the historical dataset to identify patterns and trends in client behavior, focusing on factors that influence investment decisions. We then used this information to develop predictive models that can assist us in identifying which clients to contact each week. We used the client lead files provided by Blackrock each week, and we applied our predictive models to identify the client's likelihood to invest (Classification Model) and their expected investment amount (Regression Model). We also reviewed the previous week's client solutions to expand our historical dataset and improve the accuracy of our predictive models. We came up with a formula to help us predict the average expected amount that the targeted clients would invest, which is the following:



2.1 AI Canvas

In understanding this Business problem, we decided to use the AI canvas model. The AI Canvas offers a structured approach for BlackRock to optimize client outreach using AI, focusing on predicting which clients will invest, evaluating the importance of accurate predictions, deciding on actions based on those predictions, and defining success metrics. It involves training the AI with historical data, using specific client data for predictions, and assessing how AI implementation will change current workflows, all aimed at enhancing efficiency and investment returns

AI Canvas

The task/decision being analyzed
The task is to develop a strategy for BlackRock to selectively contact 1,000 out of 7,500 weekly client leads to maximize investments and profits.

Prediction	Judgment	Action	Outcome	Training
Identify which of the 7,500 weekly leads are most likely to invest in BlackRock's investment products, thereby maximizing the expected profit from the 1,000 clients that can be reached due to resource constraints. Will client X Invest in Invest in BlackRock's product	Compare the Investment success of the prospect choice using machine learning against the current practice	Call Clients based on their likelihood to invest in the Blackrock's product	The primary measure of performance is the profit generated from the clients contacted. Secondary measures include the response rate, the investment rate, and the accuracy of the predictive model in forecasting client investment behavior.	Data on past client investments, responses to marketing contacts, demographics, account balances, and other attributes from the historical dataset will be used to train the predictive model.

Input	Feedback	How will this AI impact the overall workflow?
Data needed to generate predictions will include client age, job, marital status, preferred contact channel, current balance, whether they have housing or personal loans, number of past marketing contacts, duration of previous calls, and amount previously invested.	Feedback comes from the results of the outreach each week, indicating which clients were contacted and how much they invested. This information can be used to refine the predictive model and improve future lead selection.	The AI's role is to optimize the client outreach process, which could significantly increase the efficiency of the advisory team. It's not expected to replace staff but rather aid them in making more informed decisions. The need for retraining might involve understanding and using the predictions to guide daily activities.

3.0 Preamble

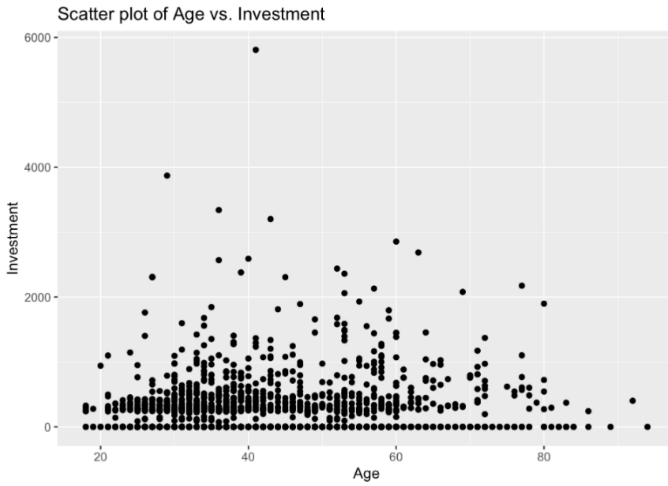
In the initial section of our report, we set the stage for our analysis by loading necessary libraries and datasets. This step is crucial as it ensures that we have all the tools and data required to conduct a thorough investigation. By initializing libraries such as tidyverse for data manipulation and ggplot2 for visualization, we lay the groundwork for our subsequent analysis. Additionally, we load two datasets: period_0.csv, which serves as our historical data, and period_1_prediction.csv, which we will use for predictive modeling.

3.1 Data Exploration

Exploring the historical dataset (tb.hist) is crucial for our analysis. Utilizing techniques like summary statistics, data structure examination, and visualizations, we gain insights into data characteristics and patterns. Variables with missing data, like job and education, are noted for special attention. Understanding dataset dimensions and types is essential. Key findings include job category distribution, marital status, and age demographics. Scatter plots unveil investment behavior patterns across age groups and professions, highlighting potential outliers and trends.

3.2 Analysis of Investment Patterns

Our analysis dives deeper into investment patterns by exploring relationships between variables such as age, number of previous marketing contacts, and balance with investment amounts. Through visualizations such as scatter plots faceted by job categories, we uncover nuanced insights into investment behavior within specific demographic and professional groups.



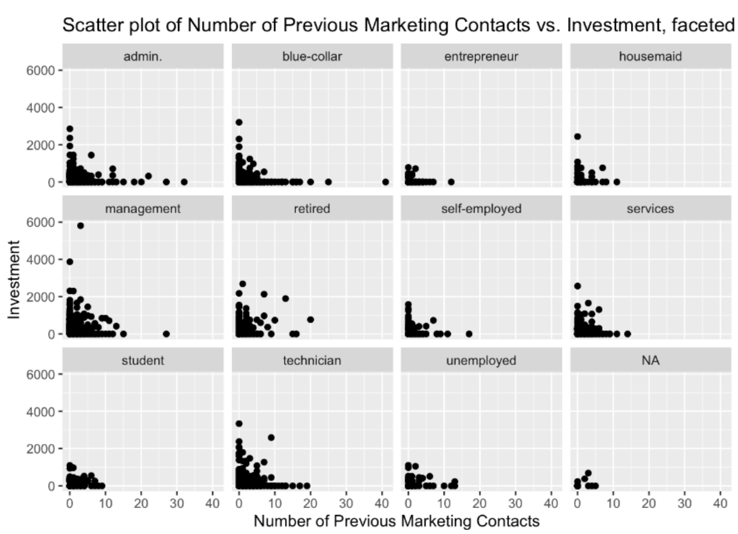
For instance, the scatter plot analysis of age versus investment highlights potential outliers and trends, providing valuable information for understanding investment behavior across different age groups. Similarly, examining the number of previous marketing contacts versus investment reveals insights into the effectiveness of past marketing efforts and their impact on investment decisions.

The scatter plot of age versus investment, faceted by job, provides valuable insights into investment behavior across different professions. Here are some interesting observations derived from the plot:

- Missing Jobs: Individuals with missing job information appear to have minimal investment activity, suggesting that they may not be employed and therefore less likely to invest. Consequently, they can be disregarded in the search for potential investors.

- Entrepreneurs: Interestingly, entrepreneurs exhibit lower investment levels, with none of them investing above €1000. This observation indicates that entrepreneurship may not necessarily correlate with high investment propensity.
- Top Investing Professions: Certain professions, such as management, blue-collar, technician, and administration, stand out for their higher investment rates. This finding suggests that individuals in these roles are more inclined to invest and may represent lucrative targets for investment campaigns.
- Retirees: The presence of retirees in their 20s and below 50 is unexpected and may require further investigation. It is notable that none of these early retirees have made investments, suggesting that they may not be suitable targets for investment initiatives.
- Students and Service Industry: While individuals in professions such as students, services, and self-employed show some inclination towards investing, their investment levels are relatively lower, especially among students. This observation highlights the importance of understanding demographic and professional nuances when targeting potential investors.

Analyzing the relationship between the number of previous marketing contacts and investment amounts provides additional insights into investment behavior. Here are notable observations from the plot:



- Lower Marketing Contacts, Higher Investment: Initially, there appears to be a trend suggesting that individuals with fewer previous marketing contacts tend to make higher investments. This observation indicates that excessive marketing outreach may not necessarily lead to increased investment activity and could potentially have diminishing returns.

- Investment Threshold at 10 Marketing Contacts: Interestingly, with the exception of individuals in management, retirees, and technicians, there are very few instances where investments surpass €1000 among those with more than 10 previous marketing contacts. This finding suggests that after a certain threshold of marketing contacts, the likelihood of high-value investments decreases significantly.
- Blue-collar Clients with High Marketing Contacts: Notably, all blue-collar clients with more than 10 previous marketing contacts did not make any significant investments. This observation underscores the importance of targeted marketing strategies tailored to specific demographic and professional segments.

The analysis of the number of previous marketing contacts against investment amounts reveals valuable insights into investment behavior and the effectiveness of marketing outreach efforts. By understanding the impact of marketing contacts on investment decisions, organizations can refine their marketing strategies to focus on quality rather than quantity. Additionally, identifying thresholds beyond which marketing contacts yield diminishing returns can help optimize resource allocation and improve the efficiency of marketing campaigns.

a marketer would expect that the higher the marketing potential, the greater the investment required. This is because marketing channels with higher potential are likely to reach a larger audience and generate more leads or sales.

4.0 Predictive Modeling

Our analysis aims to develop a predictive model for targeted marketing by leveraging features like employment status, age, past interactions, and loan status. The model filters individuals likely to respond positively to campaigns, excluding certain job categories and creating detailed groups based on loan statuses for targeted actions. We analyzed historical data over seven weeks, systematically building and refining machine learning models. We experimented with Logistic Regression, K-nearest neighbor, SVM, Decision Trees, Random Forests, and boosting methods to predict client behavior effectively, aligning with our objective of identifying individuals with high marketing potential.

4.1 Model Development Process

In the pursuit of crafting a resilient predictive model post-week 1 Prediction, we adhered to a methodical framework encompassing distinct stages. Our dataset, sourced from various periods, served as the foundation for model training and calibration. This dataset encapsulated essential features including age, marital status, occupation, and investment patterns. Initial preparatory steps involved data cleaning, merging, and harmonization to ensure data integrity. Subsequently, categorical variables such as age groups and marital status were engineered to facilitate deeper analytical insights and model refinement.

Exploratory Data Analysis (EDA): Our journey commenced with an in-depth exploration of the historical dataset. Through EDA, we sought to uncover underlying patterns and trends within the data, shedding light on client demographics, investment behaviors, and the efficacy of previous marketing initiatives.

Feature Engineering: Building upon the insights gleaned from EDA, we embarked on feature engineering endeavors. This involved crafting new features such as age groups and marital status combined with job titles. By creating categorical variables that encapsulate nuanced client characteristics, we aimed to enrich the predictive power of our model.

Model Experimentation: In the next phase of model development, we rigorously experimented with various machine learning algorithms, including Decision Trees, Random Forests, Boosted Trees, and Logistic Regression. Each underwent meticulous training and evaluation to gauge predictive performance. Ultimately, boosted trees and random forests emerged as focal points due to their effectiveness in classification tasks. Leveraging their robust capabilities, we aimed to strengthen our predictive model's accuracy in identifying potential investors.

Hyperparameter Tuning: To boost model performance, we conducted hyperparameter tuning using grid search and cross-validation, refining parameters to maximize predictive accuracy. This iterative process optimized our models significantly. We further improved accuracy and reliability through calibration, fitting logistic regression to predicted probabilities from primary models. Calibrated probabilities refined predictions, ensuring robust and dependable results. This strategic measure enhanced model performance, acknowledging the importance of accurate predictions in our analysis.

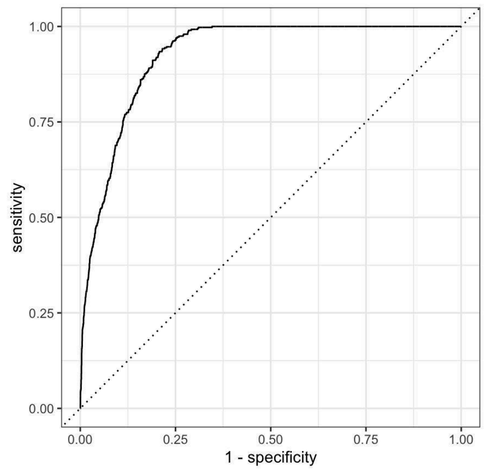
Evaluation Metrics: Throughout the model development journey, we placed significant emphasis on the evaluation of model performance using a plethora of metrics. These included ROC curves, lift curves, confusion matrices, accuracy, sensitivity, and specificity. By closely monitoring these metrics, we gained invaluable insights into the strengths and weaknesses of each model, enabling informed decision-making.

4.2 Best Performing Model

After extensive experimentation, the Random Forest model emerged as the top performer due to its exceptional predictive accuracy and robustness. It incorporates various demographic variables like age, job category, marital status, etc., to enhance its predictive capabilities. Random Forest operates as an ensemble method, amalgamating multiple decision trees to capture intricate data relationships while mitigating overfitting risks, thereby improving generalization performance.

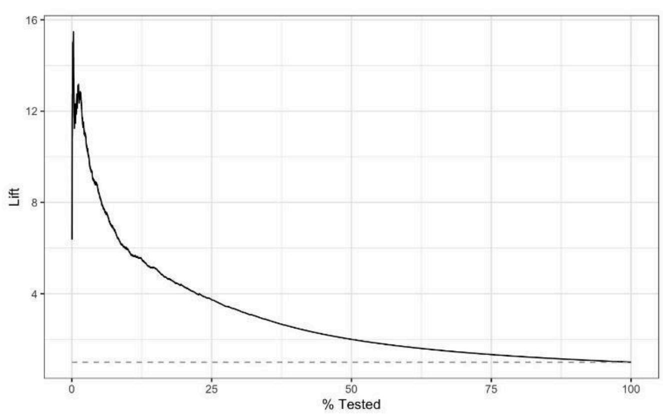
4.3 Model Evaluation

ROC Curve



The ROC curve illustrates the trade-off between sensitivity and specificity across different threshold levels. A higher area under the curve (AUC) indicates better model performance in distinguishing between positive and negative cases.

Lift Curve



The lift curve measures the effectiveness of the model in predicting positive outcomes compared to random chance. It demonstrates how much better the model performs compared to random selection, particularly at different segments of the population.

4.4 Final Evaluation/Deployment

Once the models were trained, calibrated, and evaluated, they were deployed to make predictions for the current period (Week 7). The predictions were made based on the available data, and the results were used to identify potential clients for investment targeting.

Our best-performing Random Forest model achieved an AUC score of 0.926, indicating strong discriminatory power in distinguishing between clients likely to invest and those who are not. Additionally, the model exhibited a high accuracy of 0.801, sensitivity of 0.929, and specificity of 0.794, demonstrating its effectiveness in predicting client behavior.

Evaluation Metric	Score
Accuracy	0.801
Sensitivity	0.929
Specificity	0.794
Precision	0.200
ROC_AUC	0.926

4.5 Expected Profit Calculation

An expected profit calculation was performed using the predictions from the calibrated models. This calculation involved estimating the profit generated by each predicted investment decision. The total expected profit was then computed based on these estimations.

$$amount = prob.\ of\ invest * (investment * 0.04 - 4.5)$$

5.0 Conclusion

Selecting the 1000 clients with the highest expected profit will give us the best results and the sum of the expected profit of these clients will give us the overall expected profit of the selected clients. Our predictive models identified the clients most likely to invest and generate profits for Blackrock, and we provided weekly proposals of client targets to the Call Center. We met the weekly deadlines. Our plan took into account the available resources, including the historical dataset and client lead files, and leverage predictive models to assist in the task.