# 2. Preface

This report documents the work undertaken for my internship project at Religare Broking Ltd., as a part of the MBA Digital Transformation program at SBM Narsee Monjee Institute of Management Studies (NMIMS). The project focused on a critical business challenge within the financial services sector: the proactive prediction of customer churn.

In today's competitive broking industry, understanding and mitigating customer attrition is paramount for sustained growth and profitability. The primary aim of this project was to leverage data analytics and machine learning methodologies to develop a robust predictive model capable of identifying Religare clients at risk of future inactivity. This involved an end-to-end process, from understanding diverse data sources and comprehensive feature engineering to model development, evaluation, and the interpretation of results for actionable business insights.

This document provides a detailed account of the solution design, the analytical techniques employed, the key findings derived from the models, and the potential strategic impact for Religare. It represents a significant learning experience in applying advanced analytics to real-world business problems and aims to offer valuable insights for the organization's customer retention initiatives.

The opportunity to contribute to such a meaningful project at Religare Broking Ltd. has been an invaluable component of my internship experience, providing practical exposure to data-driven strategy in the financial services domain.

# 3. Acknowledgements

I wish to express my sincere gratitude to the individuals and institutions whose guidance, support, and encouragement were instrumental in the successful completion of this internship project.

First and foremost, I am deeply indebted to my mentor at Religare Broking Ltd., Mr. Jasbir Singh (AVP, Business Excellence Department). I extend my heartfelt thanks to him for entrusting me with this challenging and highly relevant project on customer churn prediction, an area of significant importance in today's competitive market. His insightful guidance, continuous support, and the opportunity to work on such impactful research have been invaluable to my learning and professional development.

My sincere appreciation extends to key members of the Business Excellence team. I am particularly grateful to Mr. Saquib Hayat (Manager) for his consistent mentorship and assistance throughout the project. My thanks also go to Mr. Deepak Rastogi (Executive) for his crucial support, especially in providing system access that facilitated project continuity during technical challenges. The collaborative spirit of the entire Business Excellence team was also highly valued.

I extend my gratitude to NMIMS, and specifically to my Program Head, Mr. Binesh Nair, for providing the platform and the opportunity to engage in this internship. This experience allowed me to bridge academic learnings with real-world industry challenges and understand the dynamics of the corporate world.

Finally, on a personal note, I would like to thank my brother, Mr. Ankit Vashisth (Data Scientist), for his insightful advice regarding the selection of appropriate tools and machine learning models, which greatly benefited the technical direction of this project.

The collective support from all these individuals has been fundamental to the outcomes presented in this report and has made my internship experience at Religare Broking Ltd. exceptionally rewarding.

# 4. Executive Summary

In the highly competitive stockbroking industry, effective customer retention is critical for sustained profitability. This report details a internship project at Religare Broking Ltd. that successfully developed a predictive modeling framework to proactively identify customers at risk of churn, enabling more targeted and efficient retention strategies. In the competitive stockbroking industry, customer retention is paramount due to high acquisition costs and the value of sustained client engagement. The primary challenge addressed was Religare's need for a systematic, data-driven approach to forecast churn, enabling more effective and timely retention interventions.

The core objective of this project was to leverage historical client data—encompassing client details, trading activity, platform logins, funding transactions (deposits and payouts), AUM, and cash balances—to build machine learning models capable of predicting churn within 90-day and 270-day future windows. Churn was specifically defined as the cessation of *both* trading and login activities after a period of recent engagement.

A robust Analytical Base Table (ABT) was constructed using PySpark, involving snapshots of client data from January 2021 to April 2023. This process included extensive feature engineering to create approximately 78 predictive features capturing client tenure, recency, frequency, and monetary aspects of their various interactions, alongside trend-indicating delta features. Exploratory Data Analysis revealed critical insights, including the high frequency of typical trading, a significant "infant mortality" pattern where most long-term inactivity begins at activation, and the dominance of combined trade/login cessation as the primary churn pathway.

Machine learning models, primarily Logistic Regression (as a baseline) and Random Forest, were developed and evaluated. Class imbalance in the churn labels was addressed for Logistic Regression using class weighting. For Random Forest, hyperparameter tuning was performed using TrainValidationSplit on a data sample, optimizing for Area Under the Precision-Recall Curve (AUC-PR).

**Key Findings & Model Performance (Best Tuned Random Forest Models):**

**270-Day Churn Model:** Achieved an excellent **AUC-ROC of 0.989** and a strong **AUC-PR of 0.723**. At the default 0.5 threshold, this model identified **71.1% of actual 270-day churners (Recall)** with **63.85% precision** (correctly identifying churners when a churn prediction is made).

* Key drivers included longer-term login and trade frequency/recency (e.g., Login\_Txns\_Count\_270D, Days\_Since\_Last\_Login).

**90-Day Churn Model:** Achieved an **AUC-ROC of 0.991** and an **AUC-PR of 0.608**. At the default 0.5 threshold, this model identified **45.1% of actual 90-day churners (Recall)** with **63.6% precision**.

* Threshold analysis for this model indicated that adjusting the prediction threshold to 0.4 could improve **recall for 90-day churners to ~63.3%** with a precision of ~56.7%, optimizing the F1-score for this early-warning scenario.
* Key drivers were more focused on *recent* login and trade activity (e.g., Login\_Txns\_Count\_90D, Days\_Since\_Last\_Login).

The differing feature importances underscore that distinct behavioral patterns may precede short-term versus longer-term disengagement. Login activity consistently emerged as a top predictor across both models.

Furthermore, combining churn predictions with client value (using proxies like Historical\_Tag) allows for strategic segmentation, pinpointing high-value, high-risk clients for priority retention with observed churn rates in this segment exceeding 60-70%.

**Potential Business Impact & Recommendations:**

The developed models offer Religare a significant opportunity to transition to proactive, data-driven customer retention. Potential impacts include improved retention rates (illustratively, retaining even 10% of correctly identified at-risk 270-day churners could preserve approximately **₹84.7 lakhs in potential annual revenue** (based on conservative industry benchmarks for client value), underscoring the financial benefits), optimized allocation of retention budgets by focusing on high-risk/high-value segments, and enhanced customer understanding.

It is recommended that Religare:

1. Consider utilizing both the 90-day model (potentially with an adjusted threshold for higher recall) as an early warning system for timely, broad-reach interventions, and the 270-day model for identifying more deeply disengaged clients. Crucially, these churn predictions should be combined with client value segmentation (e.g., based on historical brokerage or AUM) to prioritize retention efforts on high-value, high-risk clients, thereby maximizing the ROI of intervention strategies.
2. Act on the feature insights to develop strategies that encourage consistent platform engagement, particularly login activity.
3. Initiate a pilot program to operationalize one of the models, applying predictions to a client segment and measuring the impact of interventions via A/B testing.
4. Plan for future iterations, including incorporating additional data sources (e.g., payout risk features already in ABT, customer service interactions), dedicated tuning for all target windows, and establishing a model monitoring and retraining schedule.

This project successfully demonstrates the feasibility and value of applying machine learning to predict customer churn at Religare, providing a strong foundation for enhancing customer loyalty and business profitability. Further contributions during the internship included the automation of client segmentation for MTF marketing, design of dynamic Google Forms, automation of email reporting, and development of a system for daily market turnover data collection, showcasing a broad application of analytical and automation skills.

# Table Of Content

[2. Preface 2](#_Toc199341500)

[3. Acknowledgements 3](#_Toc199341501)

[4. Executive Summary 4](#_Toc199341502)

[5. Table Of Content 7](#_Toc199341503)

[6. List of Tables & Figures 10](#_Toc199341504)

[7. About the Organization: Religare Broking Ltd. 11](#_Toc199341505)

[8. Internship Role, Responsibilities, and Contributions 12](#_Toc199341506)

[8.1 Primary Role and Responsibilities: Churn Prediction Project 12](#_Toc199341507)

[8.2: Reporting Structure, Team Collaboration, and Project Management 12](#_Toc199341508)

[8.3 Other Key Contributions and Skill Application 13](#_Toc199341509)

[8.3.1 Client Segmentation and Prioritization for Margin Trading Facility (MTF) Marketing (Task A) 13](#_Toc199341510)

[8.3.2 Enhancement and Design of Google Forms for Data Collection (Task B) 14](#_Toc199341511)

[8.3.3 Automation of Email Reporting and Follow-Ups (Task C) 14](#_Toc199341512)

[8.3.4 Automation of Daily Market Turnover Data Collection and Consolidation (Task D) 15](#_Toc199341513)

[9: Introduction 16](#_Toc199341514)

[9.1 Background: The Importance of Customer Retention in the Broking Industry 16](#_Toc199341515)

[9.2 The Challenge of Customer Churn at Religare Broking Ltd. 16](#_Toc199341516)

[9.3 Problem Statement 16](#_Toc199341517)

[9.4 Project Objectives 17](#_Toc199341518)

[9.5 Scope of the Project 17](#_Toc199341519)

[10. Solution Design and Implementation 18](#_Toc199341520)

[10.1 Data Understanding and Sources 18](#_Toc199341521)

[10.2 Data Preprocessing & Analytical Base Table (ABT) Generation 18](#_Toc199341522)

[10.2.1 Initial Data Extraction Strategy: 18](#_Toc199341523)

[10.2.2 Snapshot Methodology: 19](#_Toc199341524)

[10.2.3 Churn Definition for Predictive Modeling: 19](#_Toc199341525)

[10.2.4 Feature Engineering: 19](#_Toc199341526)

[10.2.5 ABT Generation Strategy and Output: 20](#_Toc199341527)

[10.3 Exploratory Data Analysis (EDA) - Key Insights 20](#_Toc199341528)

[10.3.1 Analysis of Inter-Trade Intervals 20](#_Toc199341529)

[10.3.2 Analysis of Time to First N-Day Inactivity 21](#_Toc199341530)

[10.3.3 Analysis of Inactivity Patterns (Trade vs. Login) 21](#_Toc199341531)

[10.4 Modeling Approach 22](#_Toc199341532)

[10.4.1 Target Variables Modeled: 22](#_Toc199341533)

[10.4.2 Feature Selection for Modeling: 22](#_Toc199341534)

[10.4.3 Data Splitting Strategy: 22](#_Toc199341535)

[10.4.4 Feature Preparation for ML Models: 23](#_Toc199341536)

[10.4.5 Machine Learning Algorithms Evaluated: 23](#_Toc199341537)

[10.4.6 Handling Class Imbalance: 23](#_Toc199341538)

[10.4.7 Hyperparameter Tuning (Random Forest): 24](#_Toc199341539)

[10.4.8 Model Evaluation Metrics: 24](#_Toc199341540)

[10.5 Model Results and Discussion 24](#_Toc199341541)

[10.5.1 Performance of the Tuned Random Forest Model for 270-Day Churn 24](#_Toc199341542)

[10.5.2 Performance of the Tuned Random Forest Model for 90-Day Churn 26](#_Toc199341543)

[10.5.3 Comparative Discussion and Overall Insights: 28](#_Toc199341544)

[10.5.4 Strategic Client Segmentation: Combining Churn Risk and Client Value 28](#_Toc199341545)

[11. Impact (Potential Quantitative and Qualitative) 31](#_Toc199341546)

[11.1: Potential Quantitative Impact 31](#_Toc199341547)

[11.1.1 Improved Customer Retention and Revenue Preservation: 31](#_Toc199341548)

[11.1.2 Optimization of Retention Budgets: 32](#_Toc199341549)

[11.1.3 Long-Term Reduction in Effective Customer Acquisition Costs (CAC): 32](#_Toc199341550)

[11.1.4 Increased Client Lifetime Value (CLV): 32](#_Toc199341551)

[11.2 Potential Qualitative Impact: 32](#_Toc199341552)

[11.3 Measuring Actual Impact (Future Recommendation): 33](#_Toc199341553)

[12. Key Learnings 33](#_Toc199341554)

[12.1 Strategic Application of Data Analytics & Machine Learning: 33](#_Toc199341555)

[12.2 Customer Behavior, Segmentation, and Lifetime Value: 34](#_Toc199341556)

[12.3 Technical Acumen and Project Execution in an Analytical Context: 34](#_Toc199341557)

[12.4 Cross-Functional Insights and Communication: 35](#_Toc199341558)

[13. References 35](#_Toc199341559)

[13.1 Industry Articles & Market Analysis: 35](#_Toc199341560)

[13.2 Academic Papers & Methodological Resources: 35](#_Toc199341561)

[Section 14: Annexures 36](#_Toc199341562)

[Annexure A: Detailed ABT Feature Dictionary 36](#_Toc199341563)

[Annexure B: Top 20 Feature Importances - Tuned RF for Is\_Churned\_Engage\_270Days 43](#_Toc199341564)

[Annexure C: Top 20 Feature Importances - Tuned RF for Is\_Churned\_Engage\_90Days 45](#_Toc199341565)

[Annexure E: SQL Queries for Raw Data Extraction 47](#_Toc199341566)

[E.1 Query 1: Extract Base Client Details 47](#_Toc199341567)

[E.2 Query 2: Extract Raw Trade Data 47](#_Toc199341568)

[E.3 Query 3: Extract Raw Payout Data 48](#_Toc199341569)

[E.4 Query 4: Extract Raw Deposit (Cash Margin) Data 48](#_Toc199341570)

[E.5 Query 5: Extract and Parse Official Dormancy Data (Used for Initial Analysis Only) 49](#_Toc199341571)

# 6. List of Tables & Figures

[Table 1: Primary Data Sources 18](#_Toc199341572)

[Table 2: Random Forest (270 Day) 25](#_Toc199341573)

[Table 3: Metric for Churners (270 Day) 25](#_Toc199341574)

[Table 4: Confusion Matrix (270 Day) 25](#_Toc199341575)

[Table 5: Random Forest (90 Days) 26](#_Toc199341576)

[Table 6: Metrics for Churners (90 Day) 27](#_Toc199341577)

[Table 7: Confusion Matrix (90 Day) 27](#_Toc199341578)

[Table 8: 90-Day Churn - Client Snapshot Counts per Segment 29](#_Toc199341579)

[Table 9: 90-Day Churn - Actual Churners per Segment 29](#_Toc199341580)

[Table 10: 90-Day Churn - Observed Churn Rate (%) within each Segment 30](#_Toc199341581)

[Table 11: 270-Day Churn - Client Snapshot Counts per Segment 30](#_Toc199341582)

[Table 12: 270-Day Churn - Actual Churners per Segment 30](#_Toc199341583)

[Table 13: 270-Day Churn - Observed Churn Rate (%) within each Segment 31](#_Toc199341584)

[Table 14: ABT Features 43](#_Toc199341585)

[Table 15: Top 20 Feature (270 Day) 44](#_Toc199341586)

[Table 16: Top 20 Feature (90 Day) 45](#_Toc199341587)

[Figure 1: Prediction Threshold (270 Day) 26](#_Toc199341698)

[Figure 2: Prediction Threshold (90 Day) 28](#_Toc199341699)

[Figure 3: Feature Importance (270 Day) 44](#_Toc199341700)

[Figure 4: Feature Importance (90 Day) 46](#_Toc199341701)

# 7. About the Organization: Religare Broking Ltd.

Religare Broking Ltd. (RBL) is a prominent full-service retail broking house in India, and a wholly owned subsidiary of Religare Enterprises Limited (REL), a diversified Indian financial services group. With a legacy in the broking business spanning over two decades (operations evolving from Religare Securities Ltd., established earlier), RBL has grown into a significant financial services provider. The company offers a comprehensive suite of services including equity trading, derivatives, commodities, depository services, mutual fund and bond distribution, insurance, and e-governance solutions.

RBL serves a substantial client base, reported to be over 1 million broking clients with approximately 2.5 lakh active clients, through an extensive distribution network. This network combines company-owned branches (around 69) with a large number of broking business partners (over 1,400) and an expanding e-governance franchisee network (approximately 51,000) across more than 400 cities in India. This multi-channel approach enables RBL to cater to diverse client segments across urban, semi-urban, and rural markets.

The company's service portfolio extends beyond traditional broking to include research capabilities for clients and a significant digital presence through its "Religare Digital" vertical, offering a wide array of citizen e-services (PAN, TAN applications, e-TDS returns) and financial inclusion services (banking correspondent services, bill payments, ticketing). RBL is registered with key regulatory bodies including NSE, BSE (for various segments including cash and F&O), NSDL as a depository participant, IRDA as a corporate agent for insurance, and PFRDA/SEBI for NPS and RTA services respectively.

Recent financial performance (FY24) indicated strong growth in average daily turnover (ADTO), total income (₹369.01 crore, a 28% increase YoY), and profit after tax (₹33.34 crore, a significant increase YoY). Strategic developments, such as the Burman group entities being classified as promoters of the parent company Religare Enterprises Limited, are anticipated to enhance RBL's financial flexibility and capital support.

Religare's mission focuses on empowering clients to achieve their financial goals through reliable and innovative services, guided by core values of customer-centricity and ethical practices. For an organization of RBL's scale and diverse service offerings, particularly in the competitive Indian broking market, understanding client behavior and proactively managing customer retention is strategically vital for sustained growth, revenue preservation, and market leadership. This churn prediction project directly addresses this imperative.

# 8. Internship Role, Responsibilities, and Contributions

## 8.1 Primary Role and Responsibilities: Churn Prediction Project

As an Intern, my primary responsibility was the end-to-end development of a predictive model for customer churn at Religare Broking Ltd. This encompassed data extraction and preprocessing from Oracle databases and diverse text-based sources, extensive feature engineering using PySpark, exploratory data analysis to understand client behavior, training and evaluation of machine learning models (Logistic Regression and Random Forest), hyperparameter tuning, and the interpretation of model results to derive actionable business insights. The project also involved conceptualizing a deployment strategy for the developed models.

## 8.2: Reporting Structure, Team Collaboration, and Project Management

My internship was situated within the **Business Excellence Department** at Religare Broking Ltd., a central function responsible for data management and analytics across the organization.

My direct supervisor and mentor for the duration of the internship was **Mr. Jasbir Singh, Assistant Vice President (AVP) of the Business Excellence Department.**

The Business Excellence team with whom I primarily interacted and collaborated comprised members with diverse roles and expertise. Key team members included:

* **Mr. Saquib Hayat** (Manager)
* **Mr. Mukhar Goel** (Manager)
* **Mr. Vaibhav Krishnatrey** (Deputy Manager)
* **Mr. Hoshiyar Gusain** (Deputy Manager)
* **Mr. Kartick Sardar** (Deputy Manager)
* **Mr. Deepak Rastogi** (Executive)

While the churn prediction initiative was designated as my individual project, I received guidance and support from various team members, particularly Mr. Saquib Hayat for insights into the data infrastructure, and Mr. Deepak Rastogi and Mr. Vaibhav Krishnatrey for operational problem-solving. The project involved independent research, iterative development, and regular progress updates to my AVP.

A crucial aspect of this project was inter-departmental collaboration for data acquisition. Notably, the **client login activity data, a key input for the churn model, was sourced in collaboration with the Risk Management Department.** This collaboration ensured access to comprehensive datasets necessary for robust feature engineering.

## 8.3 Other Key Contributions and Skill Application

### 8.3.1 Client Segmentation and Prioritization for Margin Trading Facility (MTF) Marketing (Task A)

* **Objective:**  
  Segment clients for targeted MTF service marketing and develop an initial framework for churn prediction to estimate dormant client percentages.
* **Methodology & Tools:**
  + Used **SQL** extensively for data extraction from Religare’s **Oracle database**.
  + Segmentation based on client activity types:
    - Active non-MTF traders
    - Dormant traders
    - MTF-enabled but inactive clients
  + Developed **prioritization scores** using:
    - Trading frequency
    - Revenue contribution
    - MTF enablement
    - Client age and service zone
  + Defined dormancy through **year-over-year activity status**.
  + Proposed modeling techniques: **Regression** and **Exponential Smoothing**.
* **Outcome & Solutions:**
  + Designed a **SQL-driven framework** for MTF targeting.
  + Outlined a **data-driven dormancy forecasting approach**.
  + Overcame SQL-related challenges via systematic **query validation** and **data type troubleshooting**.

### 8.3.2 Enhancement and Design of Google Forms for Data Collection (Task B)

* **Objective:**  
  Design a complex Google Form with conditional logic and input validation to streamline data entry workflows.
* **Methodology & Tools:**
  + Leveraged **Google Forms** features:
    - **Sections** and **Conditional Branching**
    - **Response Validation** using **Regular Expressions**
  + Explored alternatives: **Google Apps Script**, **AppSheet**
  + Example Regex for Client Code validation: ^[A-Z]{2}\d{5}$
* **Outcome & Solutions:**
  + Built a **multi-section Google Form** with advanced conditional logic.
  + Addressed limitations (e.g., lack of pre-submission preview) with achievable logic paths.
  + Implemented regex-based validation to enforce input format compliance.

### 8.3.3 Automation of Email Reporting and Follow-Ups (Task C)

* **Objective:**  
  Automate internal reporting and task follow-up processes to reduce manual effort and improve tracking accuracy.
* **Methodology & Tools:**
  + Used **Google Sheets** as the data source.
  + Built automation with **Google Apps Script**:
    - MailApp / GmailApp for emails
    - **Google Drive API** for Excel attachments
    - **HTML/JavaScript** to mimic spreadsheet styling in emails
* **Outcome & Solutions:**
  + Created scripts to:
    - Email formatted summary tables
    - Attach styled Excel reports
    - Send **personalized reminders** with embedded tracker links
  + Resolved technical challenges (permissions, rendering, API use) with customized **HTML/CSS**, **scope settings**, and **API parameterization**.
* **Impact:**
  + **Reduced manual workload**, ensured timely follow-ups, and improved internal communication professionalism.
  + Built a **scalable automation base** for future reporting needs.

### 8.3.4 Automation of Daily Market Turnover Data Collection and Consolidation (Task D)

* **Objective:**  
  Build an automated system to collect and consolidate **daily market turnover data** from multiple exchanges to support MIS and trend analysis.
* **Methodology & Tools:**
  + Primary language: **Python**
  + Libraries used:
    - pandas, openpyxl, requests (NSE APIs)
    - selenium + BeautifulSoup (BSE scraping)
  + Script Features:
    - Fetches and refreshes **last 4 trading days** for accuracy
    - Merges data into a **master Excel file ("Master\_data")**
    - Handles **deduplication**, **sorting**, and **error management**
    - Integrated with **Excel Power Query** for transformation
* **Outcome & Solutions:**
  + Established a **centralized, reliable system** for daily data.
  + Ensures **historical completeness** and **accuracy of recent entries**.
  + Modular design supports easy expansion for new sources or metrics.
* **Impact:**
  + **Minimized manual intervention**, reduced risk of error.
  + Enabled **timely access to accurate data** for critical decisions in trading, compliance, and market analysis.

# 9: Introduction

## 9.1 Background: The Importance of Customer Retention in the Broking Industry

The stockbroking industry operates within a dynamic and highly competitive financial services landscape. In such an environment, while customer acquisition is essential for growth, customer retention has emerged as a paramount strategic focus for firms like Religare Broking Ltd. Industry analysis consistently indicates that acquiring a new customer can be significantly more costly—often cited as up to five times more—than retaining an existing one. In sectors like broking, characterized by substantial marketing and onboarding expenditures for new clientele, the cost of retention is comparatively lower, focused on maintaining and enhancing existing relationships. Sustained customer engagement translates directly to more stable revenue streams through consistent brokerage and the potential for increased Assets Under Management (AUM), making effective retention strategies a key differentiator and a driver of long-term profitability. Conversely, unaddressed customer churn can notably erode profitability and market standing.

## 9.2 The Challenge of Customer Churn at Religare Broking Ltd.

For Religare Broking Ltd., proactively understanding and mitigating customer churn is a critical lever for optimizing business performance and maximizing customer lifetime value. For the purpose of this project, customer churn is defined behaviorally: a client ceasing both trading activities and platform logins after a period of engagement. Identifying clients at high risk of such disengagement enables Religare to deploy targeted retention initiatives, address underlying causes of dissatisfaction, and tailor offerings to bolster loyalty before a client becomes entirely inactive. This predictive capability is vital for strategic resource allocation in retention efforts and for safeguarding the active client base, which is a cornerstone of brokerage volume and revenue.

## 9.3 Problem Statement

The core problem addressed by this project is the inability to systematically and proactively identify clients of Religare Broking Ltd. who are at a high risk of churning within specific future timeframes (e.g., 90 days and 270 days). Without a predictive mechanism, retention efforts may be reactive, inefficient, or misdirected, leading to suboptimal resource allocation and missed opportunities to retain valuable customers. This project aims to leverage historical client data and machine learning techniques to develop a predictive model that can accurately forecast the likelihood of customer churn, thereby enabling data-driven and timely interventions.

## 9.4 Project Objectives

The primary objectives of this project were to:  
1. Construct a comprehensive Analytical Base Table (ABT) by engineering relevant features from diverse historical data sources including client details, trading patterns, login frequency, funding activities, AUM, and cash balances.  
2. Establish and implement a precise, behavioral definition of customer churn (cessation of both trades and logins) for predictive modeling across 90-day and 270-day windows.  
3. Develop, evaluate, and tune machine learning models, focusing on Random Forest, to accurately forecast client churn likelihood within these specified windows.  
4. Identify and analyze the key client behaviors and features most indicative of impending churn.  
5. Propose a conceptual framework for the operationalization of these predictive models to support proactive customer retention strategies at Religare.

## 9.5 Scope of the Project

This project utilized historical data primarily from January 2021 through April 2024 (for raw activity), with client snapshots for the ABT generated up to April 2023. Key data inputs included Religare's internal records for client master information, trades, logins, deposits, payouts, AUM, and cash balances. All feature engineering and model development were conducted using PySpark within a Google Colab environment. The churn definition adopted was strictly behavioral. Initial analysis of official dormancy data was performed, but this data was excluded from the predictive model's target definition. Full-scale, real-time deployment and the development of specific retention campaign materials were outside the scope of this phase.

# 10. Solution Design and Implementation

## 10.1 Data Understanding and Sources

The foundation of this predictive modeling project was a comprehensive set of historical data provided by Religare, encompassing various facets of client interactions and account status. The primary data sources are detailed below:

|  |  |  |
| --- | --- | --- |
| **Data Source (Original Table/File)** | **Key Information Extracted for ABT** | **Primary Key(s) Used** |
| Client detail table | Client unique identifier (CLIENTCODE), Account Activation Date | CLIENTCODE |
| Brokerage table | CLIENTCODE, Trade Date, Daily Gross Brokerage | CLIENTCODE, Date |
| Login Data (LOGIN\_YYYY-MM.txt) | CLIENTCODE, Login Date/Timestamp | CLIENTCODE, Date |
| Cash Margin Collected table | CLIENTCODE, Deposit Date, Realized Deposit Amount | CLIENTCODE, Date |
| Payout Request table | CLIENTCODE, Payout Date, Approved Payout Amount | CLIENTCODE, Date |
| AUM.txt (Derived) | CLIENTCODE, Month (Start Date), Monthly AUM, Running Total AUM | CLIENTCODE, Month |
| CASHBAL.txt (Derived) | CLIENTCODE, Date (Month-End), Cash Balance | CLIENTCODE, Date |
| ACCOUNT ADDRESS DETAIL table (Excluded) | Client, Parsed Dormancy Date - Used for initial analysis only | CLIENTCODE |

Table 1: Primary Data Sources

## 10.2 Data Preprocessing & Analytical Base Table (ABT) Generation

A robust data preprocessing and feature engineering pipeline was developed using PySpark in a Google Colab environment to transform raw data into a model-ready Analytical Base Table (ABT).

### 10.2.1 Initial Data Extraction Strategy:

Initial data extraction from Religare's Oracle databases involved five curated SQL queries to pull relevant raw data into delimited text files (.txt). This approach was chosen over a single complex ABT generation query to manage complexity, improve performance, and allow for modular data loading in PySpark.

### 10.2.2 Snapshot Methodology:

The ABT was designed around a client-snapshot concept. Snapshots were generated at a monthly frequency (month-end dates) for each active client within the modeling period (January 2021 - April 2023). Each row in the ABT represents a unique client at a specific snapshot in time.

### 10.2.3 Churn Definition for Predictive Modeling:

A critical step was defining churn. After initial explorations, a behavioral definition was adopted for the predictive models, focusing on the cessation of key engagement activities. A client-snapshot was labeled as 'Churned' (Is\_Churned\_Engage\_XXXDays = 1) for a given XXX-day window if both of the following conditions were met:

* "**Condition A (Recent Engagement):** The client had at least one trade OR at least one login in the XXX days *leading up to* the SnapshotDate.
* **Condition B (Subsequent Inactivity):** The client had NO trades AND NO logins in the XXX days *following* the SnapshotDate.

Otherwise, the client-snapshot was labeled as 'Not Churned' (0).

Churn labels were generated for four distinct prediction windows: 60, 90, 270, and 365 days. The primary modeling efforts reported herein focus on the 90-day and 270-day windows.

### 10.2.4 Feature Engineering:

An extensive set of features was engineered to capture various dimensions of client behavior and history leading up to each snapshot date. These can be broadly categorized as:

* **Base Features:**

Tenure\_Days: Days between ActivationDate and SnapshotDate.

* **Recency Features:** Calculated as days between the SnapshotDate and the last occurrence of an activity.
  + Days\_Since\_Last\_Trade, Days\_Since\_Last\_Login, Days\_Since\_Last\_Deposit, Days\_Since\_Last\_Payout.
* **Frequency Features:** Counts of distinct activity days and total transactions within various lookback periods (30, 90, 180, 270, 365 days prior to SnapshotDate).
  + E.g., Trade\_Days\_Count\_90D, Login\_Txns\_Count\_30D.
* **Monetary Features:** Sums of financial values within lookback periods.
  + E.g., Trade\_Sum\_90D (Gross Brokerage), Deposit\_Sum\_180D.
* **Funding Flow Features:** Derived from deposit and payout sums.
  + E.g., Net\_Funding\_Flow\_90D, Payout\_To\_Deposit\_Ratio\_90D.
* **AUM Features:** Based on the provided monthly AUM data.
  + AUM\_SnapshotMonth\_Monthly (AUM for the month of the snapshot).
  + AUM\_SnapshotMonth\_RunningTotal (Cumulative AUM up to the snapshot month).
* **Payout Risk Features (for analytical enrichment):**
  + Total\_Payout\_In\_Snapshot\_Month.
  + CashBalance\_EOM\_PreviousMonth.
  + Payout\_As\_Pct\_Of\_CashBalance: Percentage of previous month's EOM cash balance paid out in the snapshot month.
  + Payout\_Risk\_Flag: "CHURNRISK" if the above percentage exceeded 70%.
* **Delta Features:** Month-over-month change in key 90-day rolling metrics.
  + E.g., Trade\_Days\_90D\_Delta, Login\_Days\_90D\_Delta, Brokerage\_Sum\_90D\_Delta.
* **Historical Excel-Based Classification (for analytical enrichment):**
  + The logic from an existing Excel-based client classification system was replicated historically.
  + Features added: Historical\_Total\_Score, Historical\_Tag (e.g., "Classic", "Silver", "Gold").
  + The Status Score component of this classification was dynamically derived using the Is\_Churned\_Engage\_365Days label from the ABT (100 if not churned; 75/0 if churned, based on past trading). "36M" Excel inputs were proxied by 12-month (365D) ABT features.
  + While these Historical\_Tag and Historical\_Total\_Score features were not used as direct inputs for the initial churn models in this iteration, the Historical\_Tag proved valuable for subsequent client segmentation analysis when combined with model predictions (see Section 10.5.4).
* "Null values in engineered features were handled systematically, typically by filling with 0 for counts/sums, or a large value (e.g., 9999) for recency features where no prior activity existed.

*(A more detailed list of all final features can be found in Annexure A.)*

### 10.2.5 ABT Generation Strategy and Output:

Due to the large volume of data and the complexity of feature calculations, an iterative ABT generation strategy was employed. Intermediate stages of the ABT were written to and read from Parquet files on Google Drive to manage memory constraints within the Colab environment.

The final ABT, containing approximately 33.2 million client-snapshot records and 94 columns (including features, identifiers, and churn labels), was saved in Parquet format.

## 10.3 Exploratory Data Analysis (EDA) - Key Insights

Prior to comprehensive feature engineering and modeling, several exploratory data analyses were conducted to understand fundamental client activity patterns and the nature of disengagement. These insights were instrumental in shaping the churn definition and focusing the feature engineering efforts.

### 10.3.1 Analysis of Inter-Trade Intervals

* An analysis of the time duration between consecutive trading days for all clients revealed a highly right-skewed distribution.
* **Key Finding:** While the overall average time between trades was approximately 8.14 days, the **median was only 2.00 days.**
* Furthermore, 75% of inter-trade periods were 4 days or less, and 95% were 23 days or less.
* **Implication:** This indicated that for a majority of active trading periods, clients trade quite frequently. Consequently, even relatively short periods of trading inactivity (e.g., >30 days) represent a significant deviation from typical behavior and could be early indicators of disengagement from trading.

### 10.3.2 Analysis of Time to First N-Day Inactivity

* This analysis investigated how long it took for clients, from their activation date, to experience their first continuous N-day spell of *complete inactivity* (no trades AND no logins), for N = 60, 90, 270, and 365 days.
* **Key Finding:** For a very large proportion of clients who eventually exhibited such N-day inactivity, this period of inactivity **began on their activation day itself.** For example, over 75% of clients who experienced a 60-day or 90-day inactivity spell started this spell from day zero of their tenure. This pattern was even more pronounced for the 270-day and 365-day inactivity windows, where over 90% of those experiencing such long inactivity started it at activation.
* **Implication:** This highlighted a significant "infant mortality" or immediate disengagement pattern. It underscores the critical importance of successful client onboarding and early engagement in the first few days and weeks post-activation. It also suggested that features capturing very early lifecycle activity (or lack thereof) would be crucial for the predictive models.

### 10.3.3 Analysis of Inactivity Patterns (Trade vs. Login)

* This analysis examined client-snapshots to determine, for future N-day windows, whether clients: a) stopped only trading, b) stopped only logging in, c) stopped both trading and logging in, or d) remained active in both.
* **Key Finding:** Across all tested future windows (60, 90, 270, 365 days), the most prevalent pattern of disengagement was **"Stopped\_Both"** (no trades AND no logins). For instance, for a 60-day future window, approximately 68% of client-snapshots that showed any form of inactivity fell into the "Stopped\_Both" category. For a 90-day window, this figure was approximately 64%.
* The next most common pattern was "Stopped\_Trading\_Only" (clients stopped trading but continued to log in), with its proportion increasing for longer windows.
* **Implication:** This finding strongly validated the project's refined churn definition, which requires the cessation of *both* trading and login activities. It confirmed that this combined inactivity is the most common way clients fully disengage, providing a robust target for prediction. It also highlighted the "Stopped\_Trading\_Only" segment as a distinct at-risk group.

## 10.4 Modeling Approach

Following the creation of the ABT and informed by EDA insights, a structured modeling approach was adopted to predict customer churn for the defined 90-day and 270-day windows. The primary tool for model development was PySpark's MLlib library.

### 10.4.1 Target Variables Modeled:

* Two primary target variables from the ABT were selected for this phase of modeling:
  + Is\_Churned\_Engage\_90Days: Indicating churn within 90 days.
  + Is\_Churned\_Engage\_270Days: Indicating churn within 270 days.
* Models were developed and evaluated independently for each of these target variables.

### 10.4.2 Feature Selection for Modeling:

* From the comprehensive ABT, a subset of 78 features was selected for model training. This selection excluded:
  + Identifier columns (ClientCode, SnapshotDate, ActivationDate).
  + Date columns used to derive recency features (e.g., Last\_Trade\_Date).
  + Churn labels other than the specific target being modeled for a given run.
  + Descriptive analytical columns not intended as direct predictive inputs in this iteration (e.g., Payout\_Risk\_Flag, Historical\_Tag).
* The selected features primarily consisted of numerical data representing client tenure, recency of various activities, frequency counts and monetary sums over multiple lookback periods, funding flow metrics, AUM details, and delta features indicating trends.

### 10.4.3 Data Splitting Strategy:

* A time-based train/test split was implemented to simulate a realistic prediction scenario where models are trained on past data to predict future outcomes.
* A specific SnapshotDate (2023-03-01) was chosen as the split point. All client-snapshots *before* this date constituted the training set, and snapshots *on or after* this date formed the test set.
* This resulted in approximately 92% of the data for training (approx. 30.7 million snapshots) and 8% for testing (approx. 2.5 million snapshots).
* To manage memory during model development in the Colab environment, the training and test DataFrames were written to and read from Parquet files on Google Drive after the split.

### 10.4.4 Feature Preparation for ML Models:

A standard two-stage feature preparation pipeline was applied using PySpark ML transformers:

1. **VectorAssembler**: All selected numerical feature columns were combined into a single vector column named rawFeatures. The handleInvalid="skip" option was used, meaning rows with any null values in the selected feature columns would be skipped during assembly (though extensive fillna in the ABT generation aimed to minimize such instances).
2. **StandardScaler**: The rawFeatures vector was then scaled to have zero mean and unit standard deviation, producing a scaledFeatures vector. This standardization is beneficial for algorithms like Logistic Regression and can sometimes aid the convergence of tree-based models.

### 10.4.5 Machine Learning Algorithms Evaluated:

The following classification algorithms were implemented and evaluated:"

1. **Logistic Regression:** Chosen as a robust and interpretable linear baseline model.
2. **Random Forest Classifier:** Selected for its ability to capture non-linearities and feature interactions, and its general robustness.
3. ***Gradient-Boosted Trees (GBT)*** *were also considered, but initial training attempts indicated significant computational demands exceeding the available resources for full dataset training in this project phase.*

### 10.4.6 Handling Class Imbalance:

* The target churn labels exhibited significant class imbalance (e.g., for 90-day churn, approx. 1:42 churned vs. not-churned in training data; for 270-day churn, approx. 1:22).
* For **Logistic Regression**, this imbalance was explicitly addressed by applying class weights using the weightCol parameter. Weights were inversely proportional to class frequencies (e.g., Weight\_k = TotalSamples / (2 \* SamplesInClass\_k)).
* For **Random Forest**, initial models were trained without explicit class weights to assess baseline performance, as tree ensembles can sometimes inherently handle moderate imbalance. (Note: The option to use weightCol with Random Forest exists for future iterations).

### 10.4.7 Hyperparameter Tuning (Random Forest):

* To optimize the Random Forest model, hyperparameter tuning was conducted using PySpark's TrainValidationSplit.
* Due to the large training dataset size (~30 million snapshots), tuning was performed on a 5% random subsample of the training data to make the process computationally feasible.
* A small grid of key hyperparameters was explored: numTrees (number of trees), and maxDepth (maximum tree depth).
* The areaUnderPR (Area Under the Precision-Recall Curve) was used as the primary evaluation metric for selecting the best hyperparameter combination during tuning, as it is more informative than AUC-ROC for imbalanced datasets.
* The best parameters identified on the sample were then used to train a final Random Forest model on the *full* training dataset.

### 10.4.8 Model Evaluation Metrics:

Model performance was assessed using a comprehensive set of metrics on the unseen test set:

* + Area Under ROC Curve (AUC-ROC)
  + Area Under Precision-Recall Curve (AUC-PR)
  + Overall Accuracy
  + Weighted Precision, Weighted Recall, Weighted F1-Score
  + Confusion Matrix
  + Specific Precision, Recall, and F1-Score for the churner class (Class 1), calculated from the confusion matrix. This was a key focus for understanding the model's ability to correctly identify actual churners.

## 10.5 Model Results and Discussion

This section presents the performance results of the developed churn prediction models, focusing on the tuned Random Forest algorithm, which demonstrated superior performance compared to the Logistic Regression baseline. Results are detailed for both the 90-day and 270-day churn prediction windows.

### 10.5.1 Performance of the Tuned Random Forest Model for 270-Day Churn

After hyperparameter tuning (using a 5% sample of training data for TrainValidationSplit optimizing for AUC-PR) and subsequent training on the full training dataset, the Random Forest model for predicting 270-day churn achieved the following performance on the unseen test set (using a default prediction threshold of 0.5):

(Class 1 = Churn)

|  |  |
| --- | --- |
| **Metric** | **Value** |
| **AUC-ROC** | 0.9893 |
| **AUC-PR** | 0.7230 |
| **Overall Accuracy** | 0.9782 |
| **Weighted F1-Score** | 0.9788 |

Table 2: Random Forest (270 Day)

|  |  |
| --- | --- |
| **Metric** | **Value** |
| **Recall (Sensitivity)** | 0.7110 |
| **Precision** | 0.6385 |
| **F1-Score (calculated)** | 0.6728 |

Table 3: Metric for Churners (270 Day)

|  |  |  |
| --- | --- | --- |
|  | **Predicted Churn** | **Predicted Not Churn** |
| **Actual Churn (Positive)** | 56,483 (TP) | 22,957 (FN) |
| **Actual Not Churn (Negative)** | 31,977 (FP) | 2,414,254 (TN) |

Table 4: Confusion Matrix (270 Day)

**Discussion of 270-Day Model Performance:**

The model demonstrated excellent overall discriminatory power (AUC-ROC) and a strong AUC-PR, indicating good performance on the imbalanced dataset. The F1-score for the churner class (0.6728) reflects a solid balance between identifying actual churners (71.1% recall) and the accuracy of those churn predictions (63.85% precision).

**Feature Importances (270-Day Model):**

Analysis of feature importances revealed that longer-term login and trade activity patterns were primary drivers. The top predictors included:

1. Login\_Txns\_Count\_270D
2. Login\_Days\_Count\_270D
3. Login\_Days\_Count\_90D (indicating mid-term patterns also matter)
4. Days\_Since\_Last\_Login
5. Trade\_Txns\_Count\_270D

This suggests that sustained engagement over several months, particularly consistent platform logins, is highly indicative of client retention for this longer 270-day window. *(Refer to Annexure B for the full feature importance plot/table if needed).*

**Threshold Analysis (270-Day Model):**

A graph of different colored lines

AI-generated content may be incorrect.Threshold analysis indicated that the default 0.5 provides a good F1-score for churners. Adjusting the threshold (e.g., to 0.4) could increase recall to approximately 80% at the cost of reducing precision to around 58%, offering flexibility based on business strategy.

Figure 1: Prediction Threshold (270 Day)

### 10.5.2 Performance of the Tuned Random Forest Model for 90-Day Churn

A similar Random Forest model was trained for the 90-day churn target:

|  |  |
| --- | --- |
| **Metric** | **Value** |
| **AUC-ROC** | 0.9914 |
| **AUC-PR** | 0.6077 |
| **Overall Accuracy** | 0.9868 |
| **Weighted F1-Score** | 0.9857 |

Table 5: Random Forest (90 Days)

(Class 1 = Churn)

(Class 1 = Churn)

(Class 1 = Churn)

(Class 1 = Churn)

|  |  |
| --- | --- |
| **Metric** | **Value** |
| **Recall (Sensitivity)** | 0.4507 |
| **Precision** | 0.6363 |
| **F1-Score (calculated)** | 0.5277 |

Table 6: Metrics for Churners (90 Day)

|  |  |  |
| --- | --- | --- |
|  | **Predicted Churn** | **Predicted Not Churn** |
| **Actual Churn (Positive)** | 18,587 (TP) | 22,650 (FN) |
| **Actual Not Churn (Negative)** | 10,623 (FP) | 2,473,811 (TN) |

Table 7: Confusion Matrix (90 Day)

**Discussion of 90-Day Model Performance:**

The 90-day model also exhibited excellent overall AUC-ROC. Its AUC-PR (0.6077), while good, was lower than the 270-day model, suggesting the precision-recall balance for this shorter-term churn is more challenging with the current setup. At the default threshold, the model identified 45.1% of actual 90-day churners with a precision of 63.6%.

**Feature Importances (90-Day Model):**

Feature importances for the 90-day model highlighted the criticality of more *recent* activity patterns:

1. Login\_Txns\_Count\_90D
2. Login\_Days\_Count\_90D
3. Days\_Since\_Last\_Login
4. Trade\_Txns\_Count\_90D
5. Login\_Days\_Count\_180D

This shift towards shorter-term lookbacks and recency underscores that predicting near-term churn (90 days) relies heavily on very current engagement signals. *(Refer to Annexure C for the full feature importance plot/table if needed).*

**Threshold Analysis (90-Day Model):**

While the default 0.5 threshold yielded a recall of 45.1% for churners, lowering the threshold could significantly improve this.A graph of different colored lines

AI-generated content may be incorrect.  
This demonstrates that the 90-day model can be tuned via its prediction threshold to serve as a more sensitive early warning system if the business prioritizes capturing a larger proportion of potentially churning clients, accepting a higher rate of false positives for this shorter window.

Figure 2: Prediction Threshold (90 Day)

### 10.5.3 Comparative Discussion and Overall Insights:

* Both the 90-day and 270-day Random Forest models provide strong predictive capabilities. The 270-day model, at its default threshold, offers higher recall for its target group (longer-term churners).
* The 90-day model, especially when its prediction threshold is adjusted (e.g., to 0.4), can effectively identify a majority of clients at risk of near-term churn with reasonable precision, making it suitable for timely interventions.
* The differing feature importances highlight that distinct behavioral patterns may precede short-term versus longer-term disengagement, reinforcing the value of having models for different prediction horizons.
* Login activity (both frequency and recency) consistently emerged as a top predictor across both models, emphasizing its critical role in client engagement beyond just trading.

### 10.5.4 Strategic Client Segmentation: Combining Churn Risk and Client Value

To further enhance the actionability of the churn predictions, an analysis was conducted by segmenting clients in the test set based on their predicted churn risk (from the tuned Random Forest models at a 0.5 threshold) and a proxy for client value. Client value was categorized as 'High Value' or 'Low Value' based on the Historical\_Tag feature (derived from the replicated Excel classification logic, where 'Platinum' and 'Gold' tags were considered High Value). This segmentation provides a 2x2 matrix for prioritizing retention efforts:

|  |  |  |
| --- | --- | --- |
|  | **High Churn Risk *(Predicted by Model)*** | **Low Churn Risk *(Predicted by Model)*** |
| **High Client Value** | **Quadrant 1: PRIORITY 1 - Intensive Retention** - Proactive, personalized outreach - High-value retention offers - Dedicated support/RM assigned - Address specific pain points | **Quadrant 2: NURTURE & UPSELL** - Maintain high satisfaction - Explore cross-sell/upsell opportunities - Loyalty programs - Monitor for any risk shift |
| **Low Client Value** | **Quadrant 3: PRIORITY 2 - Automated/Low-Cost Retention** - Automated re-engagement (emails, platform nudges) - Cost-effective, broad offers - Or, strategically accept churn if intervention cost > potential value | **Quadrant 4: STANDARD SERVICE / MONITOR** - Standard service levels - Monitor for any changes in risk profile - Minimal proactive intervention |

The distribution of client-snapshots from the test set into these quadrants, along with the actual churn rates observed within each segment, is presented below:

**For the Is\_Churned\_Engage\_90Days Model (Tuned RF, 0.5 Threshold):**

|  |  |  |
| --- | --- | --- |
| **Value\_Segment** | **Churn\_Risk\_Segment** | **Count\_Client\_Snapshots** |
| High Value | High Churn Risk | 24 |
| High Value | Low Churn Risk | 36,335 |
| Low Value | High Churn Risk | 29,186 |
| Low Value | Low Churn Risk | 2,460,126 |

Table 8: 90-Day Churn - Client Snapshot Counts per Segment

|  |  |  |
| --- | --- | --- |
| **Value\_Segment** | **Churn\_Risk\_Segment** | **Actual\_Churners\_In\_Segment** |
| High Value | High Churn Risk | 15 |
| High Value | Low Churn Risk | 1,460 |
| Low Value | High Churn Risk | 18,572 |
| Low Value | Low Churn Risk | 21,190 |

Table 9: 90-Day Churn - Actual Churners per Segment

|  |  |  |
| --- | --- | --- |
| **Value\_Segment** | **Churn\_Risk\_Segment** | **Churn Rate (%)** |
| High Value | High Churn Risk | 62.50% |
| High Value | Low Churn Risk | 4.02% |
| Low Value | High Churn Risk | 63.63% |
| Low Value | Low Churn Risk | 0.86% |

Table 10: 90-Day Churn - Observed Churn Rate (%) within each Segment

**Insights for 90-Day Segmentation:** The 'High Value / High Churn Risk' segment, though small (24 snapshots), exhibits a very high actual churn rate (62.5%), making these clients prime candidates for immediate, focused retention. The model also identifies a large pool of 'Low Value / High Churn Risk' clients with a similar high churn rate, suitable for automated retention campaigns. The ~4% churn rate in the 'High Value / Low Churn Risk' segment highlights an area where further model refinement or threshold adjustment (as discussed in 10.5.2 - Threshold Analysis) could improve capture of at-risk valuable clients.

**For the Is\_Churned\_Engage\_270Days Model (Tuned RF, 0.5 Threshold):**

|  |  |  |
| --- | --- | --- |
| **Value\_Segment** | **Churn\_Risk\_Segment** | **Count\_Client\_Snapshots** |
| High Value | High Churn Risk | 7 |
| High Value | Low Churn Risk | 36,352 |
| Low Value | High Churn Risk | 88,453 |
| Low Value | Low Churn Risk | 2,400,859 |

Table 11: 270-Day Churn - Client Snapshot Counts per Segment

|  |  |  |
| --- | --- | --- |
| **Value\_Segment** | **Churn\_Risk\_Segment** | **Actual\_Churners\_In\_Segment** |
| High Value | High Churn Risk | 5 |
| High Value | Low Churn Risk | 793 |
| Low Value | High Churn Risk | 56,478 |
| Low Value | Low Churn Risk | 22,164 |

Table 12: 270-Day Churn - Actual Churners per Segment

|  |  |  |
| --- | --- | --- |
| **Value Segment** | **Churn Risk Segment** | **Churn Rate (%)** |
| High Value | High Churn Risk | 71.43% |
| High Value | Low Churn Risk | 2.18% |
| Low Value | High Churn Risk | 63.85% |
| Low Value | Low Churn Risk | 0.92% |

Table 13: 270-Day Churn - Observed Churn Rate (%) within each Segment

**Insights for 270-Day Segmentation:** The 'High Value / High Churn Risk' segment is extremely targeted (7 snapshots) with a very high actual churn rate (71.43%), emphasizing the critical nature of retaining these few clients. The model maintains good precision (~63.85%) in identifying longer-term churners in the 'Low Value / High Churn Risk' segment. The actual churn rate among 'High Value / Low Churn Risk' clients is lower (2.18%) compared to the 90-day model, reflecting the 270-day model's higher overall recall for its target class.

# 11. Impact (Potential Quantitative and Qualitative)

## 11.1: Potential Quantitative Impact

The deployment of the developed churn prediction models has the potential to yield significant quantitative benefits for Religare Broking Ltd. by enabling more effective customer retention. While specific internal financial metrics are proprietary, the following illustrations utilize industry benchmarks for Indian discount and full-service brokers to demonstrate potential impact:

### 11.1.1 Improved Customer Retention and Revenue Preservation:

The models identify clients at high risk of ceasing activity. For instance, the tuned Random Forest model for 270-day churn identified 56,483 True Positives (clients correctly predicted to churn who indeed churned) in the test set with a recall of 71.1% (at a 0.5 threshold). The 90-day model (at a 0.4 threshold, achieving ~63% recall) identified a different cohort of early-stage at-risk clients.

**Illustrative Impact:** If targeted interventions based on these predictions could successfully retain even a conservative 10% of these correctly identified at-risk clients:

* "For the 270-day window, this would mean retaining approximately 5,648 clients (0.10 \* 56,483)."
* "Industry estimates for annual revenue per active client for discount brokers range from ₹1,000 to ₹2,500. Assuming a conservative average annual revenue of ₹1,500 per client, retaining these 5,648 clients could preserve approximately **₹84.7 lakhs in annual revenue** (5,648 \* ₹1,500). For full-service brokers, this figure could be higher due to typically larger revenue per client."
* "Over a client's lifetime (with CLV estimates for active traders ranging from ₹10,000 to ₹25,000), the long-term value preserved would be substantially greater.

### 11.1.2 Optimization of Retention Budgets:

Predictive scoring allows for more focused allocation of retention resources. Instead of broad campaigns, efforts can be concentrated on clients who are both high-value and high-risk.

The tuned 270-day Random Forest model achieved a precision of ~63.85% for churners (at a 0.5 threshold). This means that for every 100 clients flagged for retention efforts, approximately 64 would be correctly identified churn risks. This level of precision significantly reduces expenditure on clients who were unlikely to churn, leading to a more efficient use of retention budgets compared to untargeted approaches.

### 11.1.3 Long-Term Reduction in Effective Customer Acquisition Costs (CAC):

Industry estimates for CAC for discount brokers in India range from ₹500 to ₹1,500 per client. By improving client retention, the churn rate is reduced. A lower churn rate diminishes the constant pressure to acquire new customers merely to replace those lost.

Over time, successful retention strategies fueled by predictive insights can lead to a more stable client base, thereby reducing the overall proportion of the marketing budget spent on replacing churned clients and effectively lowering the net cost of maintaining a desired active client volume.

### 11.1.4 Increased Client Lifetime Value (CLV):

By identifying and intervening with at-risk clients earlier (e.g., using the 90-day model), Religare can resolve issues, re-engage clients, and extend their active trading lifetime. Each client retained translates directly into an extended period of revenue generation, thereby increasing the average CLV.

## Potential Qualitative Impact:

**Enhanced Customer Understanding:**

* The feature importance analysis (detailed in Section 10.5) provides deep insights into the behaviors and characteristics most strongly associated with churn (e.g., declining login frequency, reduced trading activity, recency of interactions). This understanding can inform product development, service improvements, and communication strategies.

**Proactive vs. Reactive Customer Management:**

* The models enable a shift from a reactive approach (addressing churn after it happens) to a proactive one (intervening before the client is lost), which is generally more effective and fosters better customer relationships.

**Improved Customer Experience:**

* Targeted interventions can be designed to address specific pain points or needs of at-risk clients, potentially improving their overall experience and satisfaction with Religare's platform and services.

**Data-Driven Decision Making:**

* The project establishes a framework for data-driven decision-making within retention and client management functions. Performance metrics from the models provide a clear basis for evaluating the effectiveness of different intervention strategies.

**Strategic Resource Allocation:**

* Insights from the model can help prioritize development efforts or service enhancements in areas that most impact client engagement and reduce churn drivers.

**Competitive Advantage:**

* Leveraging advanced analytics for customer retention can provide Religare with a competitive edge in a crowded broking market.

## 11.3 Measuring Actual Impact (Future Recommendation):

To quantify the true impact, it is recommended to implement the model predictions in a pilot program using A/B testing. A segment of at-risk clients (identified by the model) would receive retention interventions, while a control group (also at-risk but receiving no special intervention) would be monitored. Comparing the churn rates between these groups over time would provide a direct measure of the model's effectiveness and the ROI of retention campaigns.

# 12. Key Learnings

## 12.1 Strategic Application of Data Analytics & Machine Learning:

**Data-Driven Decision Making:**  Gained profound insight into how large-scale data processing (using PySpark) and machine learning can be practically applied to solve critical business problems like customer churn, moving beyond theoretical concepts to tangible model development and evaluation.

**Translating Technical Outputs into Business Value:** Learned to interpret complex model outputs (e.g., feature importances, precision-recall trade-offs) and translate them into actionable business insights and strategic recommendations for customer retention.

**Understanding ROI of Analytical Projects:** Developed an appreciation for how predictive models, such as the churn model, can contribute to quantifiable business impact through improved customer retention, optimized resource allocation, and potentially reduced customer acquisition costs.

## 12.2 Customer Behavior, Segmentation, and Lifetime Value:

**Deep Dive into Customer Engagement Drivers:**  The project provided a data-backed understanding of key behavioral drivers of engagement and disengagement in the broking industry. The consistent importance of login frequency, trading activity, and recency highlighted critical touchpoints in the customer lifecycle.

**Identifying At-Risk Segments:**  Recognized how predictive modeling allows for the proactive identification and segmentation of at-risk customers, enabling tailored and more effective intervention strategies compared to one-size-fits-all approaches.

**Link to Customer Lifetime Value (CLV):**  The project underscored the direct link between churn reduction and the enhancement of CLV, a core metric for sustainable business growth.

**Significance of Early Lifecycle Management:**  The EDA finding on 'infant mortality' (early disengagement) was particularly striking, emphasizing the strategic importance of effective customer onboarding and early-stage engagement programs to maximize long-term value.

## 12.3 Technical Acumen and Project Execution in an Analytical Context:

**Foundational Understanding of Big Data Technologies:**  Acquired practical experience with PySpark for handling and processing large datasets, a crucial skill in today's data-rich business environments. This included addressing technical challenges like memory management through iterative processing and disk-based checkpointing.

**Principles of Predictive Modeling:**  Solidified understanding of the end-to-end machine learning pipeline, from data preprocessing and feature engineering to model training (Logistic Regression, Random Forest), hyperparameter tuning (using TrainValidationSplit on samples), and robust evaluation using metrics appropriate for imbalanced datasets (AUC-PR, class-specific recall/precision).

**Iterative Project Management:**  The project reinforced the value of an iterative development cycle, particularly in data science projects, allowing for adaptation and refinement based on emerging insights and technical hurdles.

**Problem-Solving in Complex Systems:**  Successfully navigated the complexities of integrating diverse data sources and troubleshooting technical issues, developing resilience and a systematic approach to problem resolution.

## 12.4 Cross-Functional Insights and Communication:

**Bridging Technical and Business Perspectives:** The necessity of defining churn behaviorally and discussing model trade-offs (e.g., precision vs. recall with business implications) highlighted the importance of effectively communicating technical outcomes to non-technical stakeholders.

**Importance of Clear Documentation:**  The process underscored the value of meticulous documentation for project reproducibility, knowledge sharing, and for presenting findings to diverse audiences, including senior leadership.

# 13. References

## 13.1 Industry Articles & Market Analysis:

1. Business Standard. (2024, April 15). *Brokerage firm Groww cements top spot with nearly a quarter market share.* Retrieved from [https://www.business-standard.com/companies/news/brokerage-firm-groww-cements-top-spot-with-nearly-a-quarter-market-share-124041500796\_1.html](https://www.google.com/url?sa=E&q=https%3A%2F%2Fwww.business-standard.com%2Fcompanies%2Fnews%2Fbrokerage-firm-groww-cements-top-spot-with-nearly-a-quarter-market-share-124041500796_1.html)
2. Samco Securities. (n.d.). *Broking Stocks.* Retrieved from [https://www.samco.in/knowledge-center/articles/broking-stocks/](https://www.google.com/url?sa=E&q=https%3A%2F%2Fwww.samco.in%2Fknowledge-center%2Farticles%2Fbroking-stocks%2F) *(Note: If there's a publication date on the article, it's better to use it instead of "n.d." - no date)*
3. Dalal Street Investment Journal. (n.d.). *Analyzing how top Indian stock brokers performed in 2023 and their client market share.* Retrieved from [https://www.dsij.in/dsijarticledetail/analyzing-how-top-indian-stock-brokers-performed-in-2023-and-their-client-market-share](https://www.google.com/url?sa=E&q=https%3A%2F%2Fwww.dsij.in%2Fdsijarticledetail%2Fanalyzing-how-top-indian-stock-brokers-performed-in-2023-and-their-client-market-share) *(Note: If there's a publication date, please include it.)*
4. The Economic Times. (2024, May 15). *Zerodha, 4 other discount brokers widen lead in active client vs traditional brokers in April.* Retrieved from [https://economictimes.indiatimes.com/markets/stocks/news/zerodha-4-other-discount-brokers-widen-lead-in-active-client-vs-traditional-brokers-in-april/articleshow/110111491.cms](https://www.google.com/url?sa=E&q=https%3A%2F%2Feconomictimes.indiatimes.com%2Fmarkets%2Fstocks%2Fnews%2Fzerodha-4-other-discount-brokers-widen-lead-in-active-client-vs-traditional-brokers-in-april%2Farticleshow%2F110111491.cms)

## Academic Papers & Methodological Resources:

1. Brito, J. B. G., Bucco, G. B., Heldt, R., Becker, J. L., Silveira, C. S., Luce, F. B., & Anzanello, M. J. (2023). A framework to improve churn prediction performance in retail banking. *Journal of Big Data, 10*(1), 1-28. Retrieved from [https://link.springer.com/content/pdf/10.1186/s40854-023-00558-3.pdf](https://www.google.com/url?sa=E&q=https%3A%2F%2Flink.springer.com%2Fcontent%2Fpdf%2F10.1186%2Fs40854-023-00558-3.pdf)
2. Mutanen, T., Nousiainen, S., & Ahola, J. (Year of Publication - *if available, e.g., from the ResearchGate page*). Customer churn prediction - A case study in retail banking. *[Conference or Journal Name, if available from the ResearchGate page or paper itself]*. Retrieved from [https://www.researchgate.net/publication/47749836\_Customer\_churn\_prediction\_-\_A\_case\_study\_in\_retail\_banking](https://www.google.com/url?sa=E&q=https%3A%2F%2Fwww.researchgate.net%2Fpublication%2F47749836_Customer_churn_prediction_-_A_case_study_in_retail_banking)
3. Selvaraj, N. (n.d.). *How to Build a Customer Churn Prediction Model in Python?* 365 Data Science. Retrieved from [https://365datascience.com/tutorials/python-tutorials/how-to-build-a-customer-churn-prediction-model-in-python/](https://www.google.com/url?sa=E&q=https%3A%2F%2F365datascience.com%2Ftutorials%2Fpython-tutorials%2Fhow-to-build-a-customer-churn-prediction-model-in-python%2F)

# Section 14: Annexures

## Annexure A: Detailed ABT Feature Dictionary

This annexure lists all features present in the final Analytical Base Table (ABT) used for modeling, along with their descriptions and data types.

|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Description / Derivation Logic** | **Data Type** |
| ClientCode | Unique identifier for the client. | String |
| SnapshotDate | The specific month-end date for which the client's features and churn status are recorded. | Date |
| ActivationDate | The date on which the client's account was activated. | Date |
| Tenure\_Days | Number of days between ActivationDate and SnapshotDate. | Integer |
| Last\_Trade\_Date | Most recent trade date on or before SnapshotDate. | Date |
| Days\_Since\_Last\_Trade | Days between SnapshotDate and Last\_Trade\_Date. (Filled with Tenure+1 or 9999 if no trade). | Integer |
| Last\_Login\_Date | Most recent login date on or before SnapshotDate. | Date |
| Days\_Since\_Last\_Login | Days between SnapshotDate and Last\_Login\_Date. (Filled with Tenure+1 or 9999 if no login). | Integer |
| Last\_Deposit\_Date | Most recent deposit date on or before SnapshotDate. | Date |
| Days\_Since\_Last\_Deposit | Days between SnapshotDate and Last\_Deposit\_Date. (Filled with Tenure+1 or 9999 if no deposit). | Integer |
| Last\_Payout\_Date | Most recent payout date on or before SnapshotDate. | Date |
| Days\_Since\_Last\_Payout | Days between SnapshotDate and Last\_Payout\_Date. (Filled with Tenure+1 or 9999 if no payout). | Integer |
| Trade\_Days\_Count\_30D | Unique trading days in the last 30 days. | Long |
| Trade\_Txns\_Count\_30D | Number of trade transactions in last 30 days. | Long |
| Trade\_Sum\_30D | Total brokerage generated in last 30 days. | Double |
| Trade\_Days\_Count\_90D | Unique trading days in the last 90 days. | Long |
| Trade\_Txns\_Count\_90D | Number of trade transactions in last 90 days. | Long |
| Trade\_Sum\_90D | Total brokerage generated in last 90 days. | Double |
| Trade\_Days\_Count\_180D | Unique trading days in the last 180 days. | Long |
| Trade\_Txns\_Count\_180D | Number of trade transactions in last 180 days. | Long |
| Trade\_Sum\_180D | Total brokerage generated in last 180 days. | Double |
| Trade\_Days\_Count\_270D | Unique trading days in the last 270 days. | Long |
| Trade\_Txns\_Count\_270D | Number of trade transactions in last 270 days. | Long |
| Trade\_Sum\_270D | Total brokerage generated in last 270 days. | Double |
| Trade\_Days\_Count\_365D | Unique trading days in the last 365 days. | Long |
| Trade\_Txns\_Count\_365D | Number of trade transactions in last 365 days. | Long |
| Trade\_Sum\_365D | Total brokerage generated in last 365 days. | Double |
| Login\_Days\_Count\_30D | Unique login days in last 30 days. | Long |
| Login\_Txns\_Count\_30D | Number of login events in last 30 days. | Long |
| Login\_Days\_Count\_90D | Unique login days in last 90 days. | Long |
| Login\_Txns\_Count\_90D | Number of login events in last 90 days. | Long |
| Login\_Days\_Count\_180D | Unique login days in last 180 days. | Long |
| Login\_Txns\_Count\_180D | Number of login events in last 180 days. | Long |
| Login\_Days\_Count\_270D | Unique login days in last 270 days. | Long |
| Login\_Txns\_Count\_270D | Number of login events in last 270 days. | Long |
| Login\_Days\_Count\_365D | Unique login days in last 365 days. | Long |
| Login\_Txns\_Count\_365D | Number of login events in last 365 days. | Long |
| Deposit\_Days\_Count\_30D | Unique deposit days in last 30 days. | Long |
| Deposit\_Txns\_Count\_30D | Number of deposit transactions in last 30 days. | Long |
| Deposit\_Sum\_30D | Total deposits in last 30 days. | Double |
| Deposit\_Days\_Count\_90D | Unique deposit days in last 90 days. | Long |
| Deposit\_Txns\_Count\_90D | Number of deposit transactions in last 90 days. | Long |
| Deposit\_Sum\_90D | Total deposits in last 90 days. | Double |
| Deposit\_Days\_Count\_180D | Unique deposit days in last 180 days. | Long |
| Deposit\_Txns\_Count\_180D | Number of deposit transactions in last 180 days. | Long |
| Deposit\_Sum\_180D | Total deposits in last 180 days. | Double |
| Deposit\_Days\_Count\_270D | Unique deposit days in last 270 days. | Long |
| Deposit\_Txns\_Count\_270D | Number of deposit transactions in last 270 days. | Long |
| Deposit\_Sum\_270D | Total deposits in last 270 days. | Double |
| Deposit\_Days\_Count\_365D | Unique deposit days in last 365 days. | Long |
| Deposit\_Txns\_Count\_365D | Number of deposit transactions in last 365 days. | Long |
| Deposit\_Sum\_365D | Total deposits in last 365 days. | Double |
| Payout\_Days\_Count\_30D | Unique payout days in last 30 days. | Long |
| Payout\_Txns\_Count\_30D | Number of payout transactions in last 30 days. | Long |
| Payout\_Sum\_30D | Total payout in last 30 days. | Double |
| Payout\_Days\_Count\_90D | Unique payout days in last 90 days. | Long |
| Payout\_Txns\_Count\_90D | Number of payout transactions in last 90 days. | Long |
| Payout\_Sum\_90D | Total payout in last 90 days. | Double |
| Payout\_Days\_Count\_180D | Unique payout days in last 180 days. | Long |
| Payout\_Txns\_Count\_180D | Number of payout transactions in last 180 days. | Long |
| Payout\_Sum\_180D | Total payout in last 180 days. | Double |
| Payout\_Days\_Count\_270D | Unique payout days in last 270 days. | Long |
| Payout\_Txns\_Count\_270D | Number of payout transactions in last 270 days. | Long |
| Payout\_Sum\_270D | Total payout in last 270 days. | Double |
| Payout\_Days\_Count\_365D | Unique payout days in last 365 days. | Long |
| Payout\_Txns\_Count\_365D | Number of payout transactions in last 365 days. | Long |
| Payout\_Sum\_365D | Total payout in last 365 days. | Double |
| Net\_Funding\_Flow\_30D | Deposit\_Sum\_30D - Payout\_Sum\_30D. | Double |
| Payout\_To\_Deposit\_Ratio\_30D | Payout\_Sum\_30D / Deposit\_Sum\_30D (with division by zero handled). | Double |
| Net\_Funding\_Flow\_90D | Deposit\_Sum\_90D - Payout\_Sum\_90D. | Double |
| Payout\_To\_Deposit\_Ratio\_90D | Payout\_Sum\_90D / Deposit\_Sum\_90D (with division by zero handled). | Double |
| Net\_Funding\_Flow\_180D | Deposit\_Sum\_180D - Payout\_Sum\_180D. | Double |
| Payout\_To\_Deposit\_Ratio\_180D | Payout\_Sum\_180D / Deposit\_Sum\_180D (with division by zero handled). | Double |
| Net\_Funding\_Flow\_270D | Deposit\_Sum\_270D - Payout\_Sum\_270D. | Double |
| Payout\_To\_Deposit\_Ratio\_270D | Payout\_Sum\_270D / Deposit\_Sum\_270D (with division by zero handled). | Double |
| Net\_Funding\_Flow\_365D | Deposit\_Sum\_365D - Payout\_Sum\_365D. | Double |
| Payout\_To\_Deposit\_Ratio\_365D | Payout\_Sum\_365D / Deposit\_Sum\_365D (with division by zero handled). | Double |
| AUM\_SnapshotMonth\_Monthly | Client's AUM for the snapshot month. | Double |
| AUM\_SnapshotMonth\_RunningTotal | Cumulative AUM till snapshot month. | Double |
| Total\_Payout\_In\_Snapshot\_Month | Total payout amount in the snapshot month. | Double |
| PreviousMonthEOM | End-of-month date prior to SnapshotDate. | Date |
| CashBalance\_EOM\_PreviousMonth | Cash balance at previous month-end. | Double |
| Payout\_As\_Pct\_Of\_CashBalance | Total payout in snapshot month as a % of cash balance of previous month. | Double |
| Payout\_Risk\_Flag | "CHURNRISK" if payout > 70% of cash balance, else "UNKNOWN\_RISK" or null. | String |
| Trade\_Days\_90D\_Delta | Change in trade days (90D) from previous snapshot. | Double |
| Login\_Days\_90D\_Delta | Change in login days (90D) from previous snapshot. | Double |
| Brokerage\_Sum\_90D\_Delta | Change in brokerage (90D) from previous snapshot. | Double |
| Is\_Churned\_Engage\_60Days | 1 if churned within 60 days post-snapshot based on engagement. | Integer |
| Is\_Churned\_Engage\_90Days | 1 if churned within 90 days post-snapshot based on engagement. | Integer |
| Is\_Churned\_Engage\_270Days | 1 if churned within 270 days post-snapshot based on engagement. | Integer |
| Is\_Churned\_Engage\_365Days | 1 if churned within 365 days post-snapshot based on engagement. | Integer |
| Excel\_Status\_Score\_S\_Dynamic | Status score based on churn and trade date logic. | Double |
| Historical\_Total\_Score | Score based on Excel logic. | Integer |
| Historical\_Tag | Classification tag based on Historical\_Total\_Score. | String |

Table 14: ABT Features

## Annexure B: Top 20 Feature Importances - Tuned RF for Is\_Churned\_Engage\_270Days

|  |  |
| --- | --- |
| **Feature** | **Importance** |
| Login\_Txns\_Count\_270D | 0.144595 |
| Login\_Days\_Count\_270D | 0.112715 |
| Login\_Days\_Count\_90D | 0.050469 |
| Days\_Since\_Last\_Login | 0.050028 |
| Trade\_Txns\_Count\_270D | 0.044476 |
| Login\_Days\_Count\_365D | 0.044352 |
| Trade\_Days\_Count\_270D | 0.041011 |
| Days\_Since\_Last\_Trade | 0.036131 |
| Login\_Txns\_Count\_90D | 0.035087 |
| Trade\_Sum\_270D | 0.033856 |
| Trade\_Sum\_365D | 0.030946 |
| Trade\_Txns\_Count\_365D | 0.029412 |
| AUM\_SnapshotMonth\_RunningTotal | 0.028926 |
| Login\_Txns\_Count\_365D | 0.026878 |
| Login\_Txns\_Count\_30D | 0.026504 |
| Login\_Days\_Count\_180D | 0.025526 |
| Trade\_Days\_Count\_365D | 0.023332 |
| Trade\_Days\_Count\_180D | 0.023170 |
| Login\_Days\_Count\_30D | 0.022586 |
| Days\_Since\_Last\_Payout | 0.019361 |

Table 15: Top 20 Feature (270 Day)

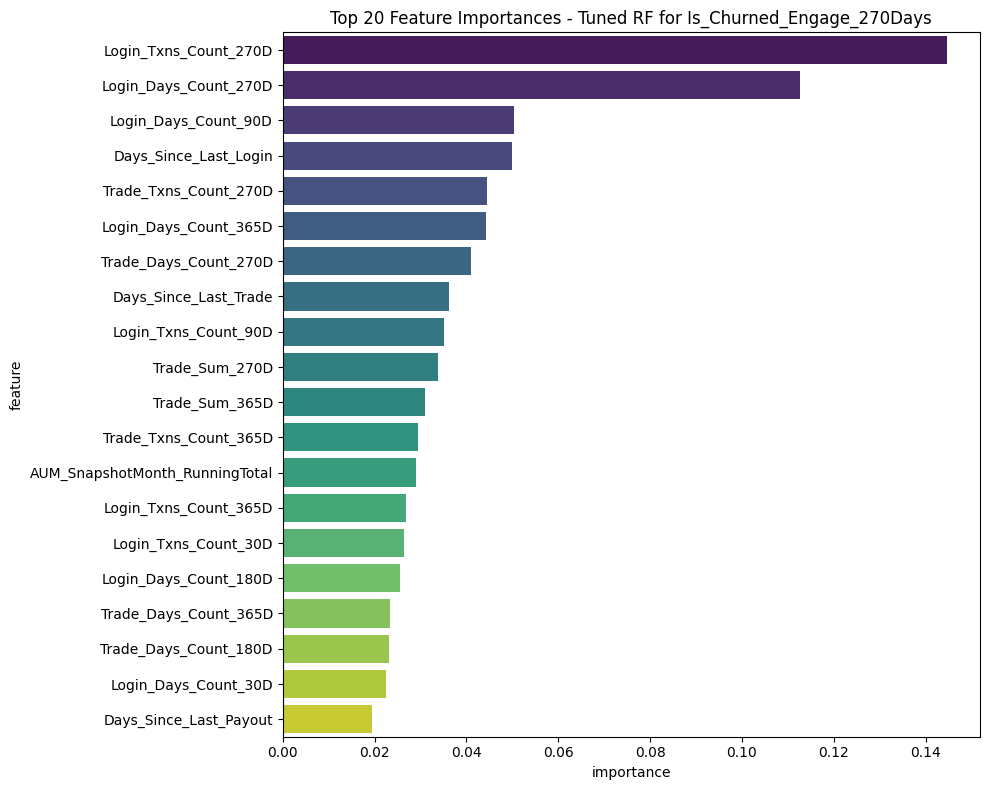


Figure 3: Feature Importance (270 Day)

## Annexure C: Top 20 Feature Importances - Tuned RF for Is\_Churned\_Engage\_90Days

|  |  |
| --- | --- |
| **Feature** | **Importance** |
| Login\_Txns\_Count\_90D | 0.106688 |
| Login\_Days\_Count\_90D | 0.080181 |
| Days\_Since\_Last\_Login | 0.069171 |
| Trade\_Txns\_Count\_90D | 0.057278 |
| Login\_Days\_Count\_180D | 0.054245 |
| Trade\_Sum\_90D | 0.049626 |
| Login\_Days\_Count\_30D | 0.043799 |
| Login\_Txns\_Count\_270D | 0.043380 |
| Trade\_Days\_Count\_90D | 0.042418 |
| Login\_Txns\_Count\_30D | 0.039893 |
| Login\_Days\_Count\_270D | 0.037111 |
| Trade\_Txns\_Count\_180D | 0.034968 |
| Login\_Txns\_Count\_180D | 0.030861 |
| Trade\_Txns\_Count\_270D | 0.027927 |
| Trade\_Days\_Count\_365D | 0.026167 |
| Trade\_Txns\_Count\_365D | 0.023250 |
| Trade\_Sum\_180D | 0.023114 |
| Trade\_Days\_Count\_180D | 0.021898 |
| Trade\_Days\_Count\_270D | 0.020526 |
| Login\_Days\_Count\_365D | 0.017909 |

Table 16: Top 20 Feature (90 Day)

A graph with different colored bars

AI-generated content may be incorrect.

Figure 4: Feature Importance (90 Day)

## Annexure E: SQL Queries for Raw Data Extraction

### E.1 Query 1: Extract Base Client Details

* **Source Table:** biuser.tblclientdetail
* **Purpose:** To retrieve fundamental client information including activation date and overall last activity markers.

SELECT

CLIENTCODE,

TRUNC(ACTIVATIONDATE) AS ACTIVATIONDATE,

TRUNC(LASTLEDGERCREDITDATE) AS Overall\_Last\_Ledger\_Credit\_Date,

LASTLEDGERCREDITAMOUNT,

TRUNC(LASTTRADEDATE) AS Overall\_Last\_Trade\_Date

FROM

biuser.tblclientdetail;

### E.2 Query 2: Extract Raw Trade Data

* **Source Table:** biuser.tblfactbrokcube
* **Purpose:** To gather records of client trading activity and associated gross brokerage.

SELECT

CLIENTCODE,

TRUNC(IMPORTDATE) AS Trade\_Date,

GROSSBROK AS TOTAL\_GROSS\_BROKERAGE\_DAY -- Aliased to match expected txt file header

FROM

biuser.tblfactbrokcube

WHERE

IMPORTDATE IS NOT NULL;

### E.3 Query 3: Extract Raw Payout Data

* **Source Table:** BIUSER.LD\_VWPAYOUTREQUEST
* **Purpose:** To obtain records of processed and approved client payout requests.

SELECT

CCLIENTCODE AS CLIENTCODE,

TRUNC(DAPPROVEDDATE) AS PAYOUT\_DATE, -- Aliased to match expected txt file header

NAPPROVEDCREDIT AS PAYOUT\_AMOUNT -- Aliased to match expected txt file header

FROM

BIUSER.LD\_VWPAYOUTREQUEST

WHERE

CFIRMNUMBER = 'FOR-000001'

AND CREALIZATIONSTATUS = 'Y'

AND CREQUESTPROCESSED = 'Y'

AND CSTATUS = 'ACCEPTED'

AND DAPPROVEDDATE IS NOT NULL;

### E.4 Query 4: Extract Raw Deposit (Cash Margin) Data

* **Source Table:** biuser.VWCASHMARGINCOLLECTED
* **Purpose:** To collect records of realized client deposits.

SELECT

"Client\_Code" AS CLIENTCODE,

TRUNC("Transaction\_Date") AS DEPOSIT\_DATE, -- Aliased to match expected txt file header

"RealizedAmount" AS DEPOSIT\_AMOUNT -- Aliased to match expected txt file header

FROM

biuser.VWCASHMARGINCOLLECTED

WHERE

"RealizationStatus" = 'Y'

AND "Transaction\_Date" IS NOT NULL;

### E.5 Query 5: Extract and Parse Official Dormancy Data (Used for Initial Analysis Only)

* **Source Table:** BIUSER.LD\_VWACCOUNTADDRESSDETAIL
* **Purpose:** To extract and parse dates related to official account dormancy declarations. This data was used for preliminary analysis but was **not** used for defining the target variable in the predictive churn models.

SELECT

oowncode AS CLIENTCODE,

CASE

WHEN REGEXP\_LIKE(SUBSTR(PAGER, -8), '^[0-9]{8}$')

THEN TO\_DATE(SUBSTR(PAGER, -8), 'DDMMYYYY')

ELSE NULL

END AS Dormancy\_Date\_Parsed

FROM

BIUSER.LD\_VWACCOUNTADDRESSDETAIL

WHERE

pager LIKE 'DRM%'

AND FIRMNUMBER = 'FOR-000001'

AND REGEXP\_LIKE(SUBSTR(PAGER, -8), '^[0-9]{8}$');