**Slide 1: Title Slide – Introduction**

**Script:**

Good morning/afternoon, everyone.  
My name is Vaibhav, and I am an MBA Intern from SVKM's Narsee Monjee Institute of Management Studies.  
Today, I’m excited to present the core project of my internship at Religare Broking Ltd.: the development of a **Customer Churn Prediction Model**—a tool designed to proactively identify clients at risk of disengagement and help improve client retention strategies.

**Slide 2: Introduction & Problem Statement**

**Script:**

In the highly competitive world of stockbroking, retaining clients is not just important—it’s critical for long-term profitability.  
Studies show that acquiring a new customer can be up to five times more expensive than retaining an existing one. For a firm like Religare, where acquisition and onboarding costs are significant, focusing on retention can deliver strong ROI through stable brokerage revenue and Assets Under Management.  
However, the problem we tackled was this:  
Religare lacked a **data-driven, proactive method** to identify clients likely to churn in the near future.  
Our objective was to build a **predictive machine learning model** that not only forecasts churn probability but also uncovers key churn drivers—providing a foundation for timely and targeted retention interventions.

**Slide 3: Data & Methodology Overview**

**Script:**

To address the problem, we first collected and unified data from multiple internal sources, including:

* Client master information
* Historical trades
* Login records (newly included in the analysis)
* Deposits, payouts, AUM, and cash balances

The core of our approach was building a **monthly Analytical Base Table (ABT)**. This captured time-specific snapshots of client behavior, enabling true forward-looking predictions.

All processing, feature engineering, and modeling were done using **PySpark**—which allowed us to efficiently handle large volumes of data across multiple timeframes.

**Slide 4: Key Exploratory Data Analysis (EDA) Insights**

**Script:**

Before modeling, we conducted extensive EDA to understand client engagement patterns and define churn more precisely.  
Key insights included:

* **Inter-trade intervals**: Active clients traded frequently, with a median gap of only 2 days.
* **Early disengagement**: Many clients showed signs of ‘infant churn’ within their first few months.
* **Activity drop-off**: Churn was best defined as **simultaneous inactivity in both trades and logins**—a powerful indicator of disengagement.

These insights were critical in designing our churn definition and selecting meaningful features.

**Slide 5: Modeling Approach**

**Script:**

With our ABT ready and churn defined, we built models focused on two timeframes:

* **90-day churn** (short-term risk)
* **270-day churn** (longer-term risk)

We applied a **time-based train-test split** to simulate real-world prediction: training on past data and testing on future unseen data.  
After comparing multiple algorithms, **Random Forest** emerged as the best performer—especially after hyperparameter tuning.

**Slide 6: Model Performance**

**Script:**

Here are the results:

* **90-Day Model**:
  + AUC-ROC: **0.989**, AUC-PR: **0.723**
  + At a 0.5 threshold, we achieved:
    - **71.1% Recall** (capturing actual churners)
    - **63.85% Precision** (accuracy of predicted churners)
* **270-Day Model**:
  + AUC-ROC: **0.989**, AUC-PR: **0.684**
  + Recall: **62.9%**, Precision: **64.7%**

Both models demonstrate strong discriminative power and practical reliability across different horizons.

**Slide 7: Key Churn Drivers**

**Script:**

Beyond predictions, the models provided **valuable behavioral insights**:

* **Recent engagement** (last 30-90 days) was the strongest predictor—covering both trading and login activity.
* **Recent brokerage values** were significant monetary indicators.
* **Activity deltas**—changes in engagement levels over time—also proved highly predictive.

This shows that **declining or inconsistent engagement** is a major red flag and offers a strategic area to focus retention efforts.

**Slide 8: Strategic Application – Prioritizing Retention**

**Script:**

We went a step further by combining churn risk predictions with **client value** using Religare’s own classification logic.

This formed a **2x2 priority matrix**:

* High Risk / High Value
* High Risk / Low Value
* Low Risk / High Value
* Low Risk / Low Value

The most critical segment?  
**High Value / High Risk clients**—with actual churn rates often exceeding **60-70%**.

This prioritization enables **smart allocation of retention resources**—focusing on saving the most valuable at-risk clients.

**Slide 9: Potential Business Impact**

**Script:**

Although the model isn’t live yet, its **potential impact is substantial**:

* **Cost savings**: Proactive retention of high-value clients is significantly cheaper than reacquiring lost ones.
* Even a **small improvement in retention** could translate into substantial ROI.
* **Qualitative benefits** include improved understanding of client behavior, better resource allocation, and a more customer-centric strategy.

A logical next step would be a **controlled pilot** to quantify this impact through targeted interventions.

**Slide 10: Key Learnings**

**Script:**

From an MBA perspective, this project offered several critical takeaways:

* **Bridging data science with strategy**: Seeing how machine learning directly supports business outcomes.
* **Data quality and engineering**: Learning the importance of feature relevance and domain context.
* **Stakeholder communication**: Explaining complex models in simple, decision-useful terms.
* **Project management**: Executing a technical project with multiple moving parts under time constraints.

It reinforced how **data-driven decisions can transform financial services**.

**Slide 11: Operationalization & Future Steps**

**Script:**

Though model deployment was beyond the internship scope, we outlined a roadmap for operationalization:

* Build a **regular scoring pipeline** using the same ABT logic
* Run churn predictions weekly or monthly
* Feed scores into business processes—such as CRM systems or retention campaigns

Ongoing **monitoring and retraining** will be critical as market conditions evolve and client behavior shifts.

**Slide 12: Conclusion & Acknowledgements**

**Script:**

To conclude:  
We successfully developed high-performing, business-ready churn prediction models for Religare Broking Ltd.  
These models identify key drivers of churn, segment clients for intervention, and lay the groundwork for proactive retention strategies.

I’d like to thank **Mr. Jasbir Singh**, the **Business Excellence team at Religare**, and **NMIMS** for this invaluable opportunity.

Thank you for your attention. I’m happy to take any questions.

**1. AUC-ROC (Area Under the Receiver Operating Characteristic Curve)**

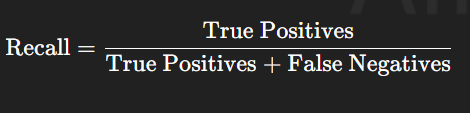
* **What it measures**: How well the model distinguishes between the two classes (e.g., churners vs. non-churners).
* **How it works**: It plots the **True Positive Rate (Recall)** vs. **False Positive Rate** at various threshold settings.
* **Why it’s useful**: It gives an overall view of model performance across all classification thresholds.
* **Interpretation**:
  + AUC = 1 → Perfect model
  + AUC = 0.5 → Random guessing
  + The higher the AUC, the better the model is at distinguishing between classes.

**2. AUC-PR (Area Under the Precision-Recall Curve)**

* **What it measures**: The trade-off between **Precision** and **Recall** for the positive class (e.g., churners).
* **Why it’s important for imbalanced data**: More informative than AUC-ROC when the positive class is rare.
* **Interpretation**:
  + High AUC-PR means the model is good at identifying the minority class (churners) without too many false positives.

**3. Recall (Churners)**

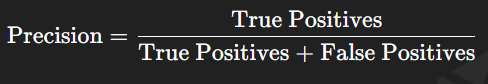
* **Also called**: Sensitivity or True Positive Rate.
* **What it measures**: Of all actual churners, how many did the model correctly predict?
* **Formula**:



* **Interpretation**:
  + High recall means the model is good at catching most of the churners.
  + Important when missing a churner is costly (e.g., losing a customer).

**4. Precision (Churners)**

* **What it measures**: Of all customers the model predicted as churners, how many actually did churn?
* **Formula**:

​

* **Interpretation**:
  + High precision means the model makes fewer false churn predictions.
  + Important when acting on a churn prediction is expensive (e.g., giving incentives).

|  |  |  |
| --- | --- | --- |
| **Metric** | **Focuses On** | **Best For** |
| AUC-ROC | Overall class separation | General model quality |
| AUC-PR | Precision vs Recall | Imbalanced classification |
| Recall | Capturing actual churners | When missing churners is costly |
| Precision | Correct churn predictions | When false alarms are expensive |

**🔹 90-Day Churn Model**

**Key Metrics:**

* **AUC-ROC = 0.989**  
  → This is **excellent**. It means the model is very good at distinguishing between customers who will churn within 90 days and those who won’t.
* **AUC-PR = 0.723**  
  → This is a **strong** result, especially given the **class imbalance** (where churners are much fewer than non-churners). It means the model is effective at identifying actual churners with relatively few false alarms.
* **Recall = 71.1% (at 0.5 threshold)**  
  → The model correctly identifies about **71 out of every 100** churners.
* **Precision = 63.85% (at 0.5 threshold)**  
  → Of all customers predicted to churn, about **64 out of every 100 actually did churn** within 90 days.

**What it means:**

* The model is **very good at ranking and identifying churners**.
* It strikes a good balance between **catching most churners** and **avoiding too many false positives**.
* You might still consider **adjusting the threshold** if your business prefers either higher recall (to catch more churners) or higher precision (to reduce costs of false alarms).

**🔹 270-Day Churn Model**

**Key Metrics:**

* **AUC-ROC = 0.989**  
  → Again, this is **excellent**, showing strong discriminative ability over a longer timeframe.
* **AUC-PR = 0.684**  
  → Still very good, though slightly lower than the 90-day model — which makes sense because predicting churn farther in advance is generally more difficult.
* **Recall = 62.9%**  
  → The model identifies about **63 out of every 100 churners** who will churn within 270 days.
* **Precision = 64.7%**  
  → Of all customers flagged as future churners, about **65 out of every 100 actually churned** within that longer window.

**What it means:**

* This model still performs very well, though it's a bit less aggressive in identifying churners compared to the 90-day model.
* It gives your business **more lead time** to take action and try to retain at-risk customers.

**✅ Overall Interpretation**

Both models are:

* **Highly effective** at distinguishing between churners and non-churners.
* **Well-calibrated**, even on imbalanced data.
* **Useful for different business strategies**:
  + **90-Day model** is more responsive and can trigger short-term retention actions.
  + **270-Day model** offers a **longer horizon** to implement strategic interventions.