**Churn Analytics in Streaming Services: A Data Science Driven Approach**

**Index**

1. Introduction
2. Statement of Work
3. Project Description
4. Objectives
5. Scope of Work
6. Deliverables
7. Consultant Responsibilities
8. Customer Responsibilities
9. Schedule
10. Payment Schedule
11. Phase 1: Business Understanding
    1. Clarify the Business Context and Problem Definition
    2. Translate Business Problems into Data-Driven Objectives
    3. Stakeholder Engagement
    4. Outcome of Phase 1
    5. Why This Project Matters
12. Phase 2: Data Understanding and Governance
    1. Data Source and Collection
    2. Data Quality Assessment
    3. Iterative Feedback
    4. Outcome of Phase 2
13. Phase 3: Exploratory Data Analysis & Insight Generation
    1. Data Exploration
    2. Hypothesis Testing
    3. Documentation and Iteration
    4. Outcome of Phase 3
14. Phase 4: Data Preparation & Feature Engineering
    1. Data Cleaning and Transformation
    2. Feature Engineering
    3. Advanced Preprocessing and Automation
    4. Train/Test Splitting
    5. Outcome of Phase 4
15. Phase 5: Modeling & Experimentation
    1. Problem Framing
    2. Integrate Experimentation
    3. Algorithm Selection and Experimentation
    4. Interpretability and Explainability
    5. Bias, Fairness, and Ethical Considerations
    6. Iteration and Refinement
    7. Outcome of Phase 5
16. Phase 6: Model Evaluation & Business Review
    1. Performance Metrics
    2. Error Analysis
    3. Business Validation
    4. Go/No-Go Decision
    5. Outcome of Phase 6
17. Phase 7: Deployment & MLOps
    1. Preparation for Deployment
    2. Packaging & Versioning
    3. Integration and Scalability
    4. CI/CD Integration and Testing
    5. Monitoring, Maintenance, and Retraining
    6. Documentation and Knowledge Transfer
    7. Post-Deployment Review
    8. Outcome of Phase 7
18. Strategic Impact of Data Science on Business Decision-Making
    1. How Data Science Drives Strategic Decision-Making
    2. Why Data-Driven Decisions Are Necessary for Business Success
    3. Conclusion

**Introduction**

In the highly competitive streaming industry, customer churn poses a significant challenge to revenue stability and growth. **StreamHub, Inc.**, is a fake company that resemble a real-world leading streaming platform with 10 million subscribers (like Netflix), faced a 5% monthly churn rate, resulting in $5 million in monthly revenue losses. To address this, Data Science “WithMe” LLC executed a six-month data science initiative from June to November 2025, leveraging the CRISP-DM methodology to develop a predictive churn model. This project delivered a production-grade XGBoost model (AUC-ROC = 0.85, recall = 0.74), identifying 74% of at-risk customers and enabling targeted retention strategies that yielded $2.93 million in annual net savings (166% ROI).

This portfolio document chronicles the end-to-end journey, from defining business objectives to deploying a real-time API integrated with StreamHub’s CRM and marketing systems. It showcases technical expertise in data analysis, machine learning, and MLOps, alongside strategic alignment with business goals. Key achievements include precise churn predictions, scalable infrastructure, and ethical considerations (e.g., fairness metrics, GDPR compliance), positioning StreamHub for sustained growth. This work exemplifies the power of data-driven decision-making and serves as a cornerstone of my data science portfolio, demonstrating my ability to deliver impactful, production-ready solutions.

**Statement of Work**

Data Science “**WithMe**” LLC (Consultant) is pleased to submit this Statement of Work (SoW) to **StreamHub, Inc.** (Client) for an engagement commencing on **June 1, 2025**.

**1. Project Description**

**StreamHub, Inc.** operates a streaming platform with approximately 10 million subscribers, offering video and audio content similar to Netflix and Spotify. The platform faces a significant business challenge with a monthly churn rate of approximately 5% (500,000 subscribers), resulting in an estimated revenue loss of $5 million per month, based on an average revenue per user (ARPU) of $10. Customer churn—when subscribers cancel their subscriptions—threatens revenue stability and growth in a highly competitive market.

Data Science “**WithMe**” LLC will conduct a comprehensive six-month data science project to predict customer churn and enable proactive retention strategies. Leveraging the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, the project will involve analyzing subscriber data, developing predictive models, and deploying a scalable solution to identify at-risk customers. The initiative aims to reduce churn by 10% (from 5% to 4.5%), saving $500,000 monthly, and provide actionable insights to enhance customer retention and engagement.

The project will encompass business understanding, data exploration, modeling, and deployment, ensuring alignment with StreamHub’s strategic goals. The solution will integrate seamlessly with existing systems, enabling real-time churn predictions and targeted interventions such as personalized offers or improved content recommendations.

**2. Objectives**

The primary objectives of this engagement are:

* **Business Objective:** Reduce StreamHub’s monthly churn rate by 10% (from 5% to 4.5%) within six months, translating to an estimated annual revenue retention of $6 million (500,000 subscribers × $10 ARPU × 12 months).
* **Technical Objective:** Develop a predictive machine learning model with a minimum AUC-ROC score of 0.85 to identify subscribers at risk of churning, enabling targeted retention campaigns.
* **Analytical Objective:** Uncover key drivers of churn (e.g., low engagement, subscription tenure) through exploratory data analysis to inform strategic decisions.
* **Operational Objective:** Deploy a scalable, real-time churn prediction system integrated with StreamHub’s marketing and customer relationship management (CRM) systems, with continuous monitoring to maintain performance.

These objectives align with StreamHub’s goal of improving subscriber retention and maximizing revenue in a competitive streaming market.

**3. Scope of Work**

The scope of work includes all activities required to deliver a comprehensive churn prediction solution, following the CRISP-DM methodology. The project will cover the following phases:

* **Business Understanding:**
  + Engage with stakeholders (e.g., marketing, customer service, and executive teams) to define churn-related business problems.
  + Establish measurable success criteria, such as a 10% churn reduction and model performance metrics (AUC-ROC ≥ 0.85).
  + Document a problem statement linking technical tasks to business outcomes.
* **Data Understanding and Governance:**
  + Identify and assess data sources, including user demographics, viewing history, subscription details, and customer interactions.
  + Evaluate data quality (e.g., completeness, anomalies) and ensure compliance with privacy regulations (e.g., GDPR).
  + Profile datasets to confirm their suitability for churn prediction.
* **Exploratory Data Analysis (EDA) and Insight Generation:**
  + Conduct univariate, bivariate, and time-series analyses to identify churn patterns (e.g., seasonal trends, low engagement).
  + Test hypotheses (e.g., “Users with <5 hours/week watch time are more likely to churn”).
  + Visualize findings using heatmaps, scatter plots, and trend graphs.
* **Data Preparation and Feature Engineering:**
  + Clean data by addressing missing values, outliers, and duplicates.
  + Engineer features such as average watch time per session, genre diversity, and days since last login.
  + Create automated preprocessing pipelines and a time-based train/test split to prevent data leakage.
* **Modeling and Experimentation:**
  + Frame churn prediction as a binary classification problem.
  + Test algorithms (e.g., logistic regression, random forests, XGBoost) and tune hyperparameters using grid search.
  + Ensure model interpretability with SHAP/LIME and fairness across demographics.
  + Track experiments using MLflow for reproducibility.
* **Model Evaluation and Business Review:**
  + Evaluate models using metrics like AUC-ROC, precision, recall, and business KPIs (e.g., 70% churner identification rate).
  + Conduct error analysis to address model shortcomings.
  + Present findings to stakeholders for deployment approval.
* **Deployment and MLOps:**
  + Deploy the model as a real-time REST API using Docker containers.
  + Integrate with StreamHub’s CRM and marketing systems for automated retention campaigns.
  + Implement monitoring for performance and concept drift, with scheduled retraining.
  + Document all processes for knowledge transfer.

**Out of Scope:**

* Acquisition of new data sources beyond what StreamHub provides.
* Development of retention campaign content (e.g., specific offers or recommendations).
* Hardware procurement or infrastructure upgrades beyond model deployment requirements.

**4. Deliverables**

The Consultant will provide the following deliverables to StreamHub, Inc.:

* **Project Plan (Month 1):**
  + A detailed roadmap outlining tasks, timelines, and milestones for the six-month project.
  + A validated problem statement and success criteria.
* **Data Assessment Report (Month 2):**
  + Documentation of data sources, quality scores (e.g., 95% completeness), and governance practices.
  + Initial findings on data suitability for churn prediction.
* **Exploratory Data Analysis Report (Month 3):**
  + Comprehensive analysis with visualizations (e.g., churn trends, feature correlations).
  + Prioritized features and insights (e.g., “Low watch time predicts 3x higher churn”).
* **Model Development Report (Month 4):**
  + Summary of tested algorithms, hyperparameters, and performance metrics (e.g., AUC-ROC 0.85).
  + SHAP/LIME explanations and fairness assessments.
* **Model Evaluation and Business Case (Month 5):**
  + Final model performance metrics and error analysis.
  + Business case projecting ROI (e.g., $1.75M/year from retaining 50% of identified churners).
  + Stakeholder presentation for deployment approval.
* **Deployed Model and Documentation (Month 6):**
  + A real-time churn prediction API integrated with StreamHub’s systems.
  + Automated preprocessing pipelines and monitoring dashboards.
  + Comprehensive documentation covering data preparation, modeling, deployment, and maintenance processes.
  + A handover report with lessons learned and recommendations for ongoing management.
* **Interactive Dashboard (Month 6):**
  + A Streamlit-based dashboard displaying churn predictions, key features, and business insights for stakeholder use.

All deliverables will be provided in digital format (e.g., PDF, Jupyter notebooks, GitHub repository) and presented during stakeholder meetings.

**5. Consultant Responsibilities**

Data Science “WithMe” LLC will:

* Assign a dedicated data science team, including a project manager, data scientists, and MLOps engineers, to execute the project.
* Conduct all CRISP-DM phases as outlined in the Scope of Work, ensuring high-quality deliverables.
* Engage with StreamHub stakeholders bi-weekly to provide updates, gather feedback, and align on objectives.
* Ensure data privacy and security compliance (e.g., GDPR) during data handling and processing.
* Develop and deploy a scalable, interpretable, and fair churn prediction model.
* Provide thorough documentation for reproducibility and knowledge transfer.
* Train StreamHub’s operations team on model usage and maintenance during the handover phase.
* Address any issues related to model performance or deployment within 30 days post-deployment at no additional cost.

**6. Customer Responsibilities**

StreamHub, Inc. will:

* Provide access to relevant data sources (e.g., subscriber demographics, viewing history) by June 1, 2025, including necessary permissions and documentation.
* Designate a primary point of contact (e.g., data manager or marketing lead) to facilitate communication and data access.
* Participate in bi-weekly meetings to provide feedback, validate findings, and approve deliverables.
* Ensure availability of domain experts (e.g., customer service leads) to clarify data context and business needs.
* Provide access to existing systems (e.g., CRM, cloud infrastructure) for model integration and deployment.
* Review and approve deliverables within five business days of submission to maintain the project timeline.
* Implement retention strategies based on model outputs (e.g., campaigns targeting high-risk churners).

**7. Schedule**

The project will span six months, from **June 1, 2025, to November 30, 2025**, with the following milestones:

|  |  |  |
| --- | --- | --- |
| **Milestone** | **Completion Date** | **Deliverable** |
| Project Kickoff & Business Understanding | June 30, 2025 | Project Plan, Problem Statement |
| Data Understanding & Governance | July 31, 2025 | Data Assessment Report |
| Exploratory Data Analysis | August 31, 2025 | EDA Report with Insights |
| Model Development & Experimentation | September 30, 2025 | Model Development Report |
| Model Evaluation & Business Review | October 31, 2025 | Evaluation Report, Business Case |
| Model Deployment & Handover | November 30, 2025 | Deployed Model, Documentation, Dashboard |

* **Bi-Weekly Meetings:** Scheduled every other Wednesday at 10:00 AM (StreamHub’s local time) to review progress and gather feedback.
* **Final Presentation:** November 25, 2025, to stakeholders for project closeout.

**8. Payment Schedule**

The total project cost is **$150,000**, reflecting the complexity of the six-month engagement, including data analysis, modeling, deployment, and ongoing support. Payment will be invoiced monthly based on milestone completion, as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Milestone** | **Completion Date** | **Payment Amount** | **Percentage** |
| Project Kickoff & Business Understanding | June 30, 2025 | $25,000 | 16.67% |
| Data Understanding & Governance | July 31, 2025 | $25,000 | 16.67% |
| Exploratory Data Analysis | August 31, 2025 | $25,000 | 16.67% |
| Model Development & Experimentation | September 30, 2025 | $25,000 | 16.67% |
| Model Evaluation & Business Review | October 31, 2025 | $25,000 | 16.67% |
| Model Deployment & Handover | November 30, 2025 | $25,000 | 16.67% |
| **Total** |  | **$150,000** | **100%** |

* **Invoicing Terms:** Invoices will be issued on the first business day following each milestone completion, with payment due within 30 days (Net 30).
* **Payment Method:** Electronic transfer to Data Science “WithMe” LLC’s designated bank account (details provided in invoices).
* **Late Payments:** A 1.5% monthly interest rate will apply to overdue payments.
* **Additional Costs:** Any out-of-scope work (e.g., new data source integration) will require a separate SoW and mutual agreement.

**Acceptance**

By signing below, both parties agree to the terms and conditions outlined in this Statement of Work.

**For Data Science “WithMe” LLC:**

Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
Title: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**For StreamHub, Inc.:**

Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
Title: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Notes**

* **Assumptions:** The project assumes timely access to data and systems, stakeholder availability, and no significant changes to StreamHub’s business model during the engagement.
* **Confidentiality:** All data and findings will be treated as confidential, with a separate Non-Disclosure Agreement (NDA) if required.
* **Termination:** Either party may terminate the engagement with 30 days’ written notice, with payment due for completed milestones.

This SoW provides a clear, professional framework for the churn prediction project, ensuring alignment between Data Science “WithMe” LLC and StreamHub, Inc. The structure aims to demonstrate a realistic, industry-relevant engagement.

**Phase 1: Business Understanding**

**Goal:** Align on why we are building the model, what we will measure, and who cares.

In this phase, we establish the foundation for a customer churn prediction project for StreamHub, Inc., a streaming platform. The focus is on understanding the business context, translating it into data-driven objectives, engaging stakeholders, and defining a clear path forward. Below, we address each component of the Business Understanding phase.

**1.1 Clarify the Business Context and Problem Definition**

**Project Objectives Using SMART Goals:**  
To ensure the project delivers actionable business outcomes, we define objectives using the SMART framework:

* **Specific:** Predict which StreamHub subscribers are likely to churn, enabling targeted retention strategies to reduce cancellations.
* **Measurable:** Reduce the monthly churn rate by 10% (from 5% to 4.5%) within six months, saving $500,000 monthly in revenue (assuming 500,000 fewer churners at $10 average revenue per user, ARPU).
* **Achievable:** Build a predictive model with at least 85% accuracy or an AUC-ROC of 0.85 using existing subscriber data (e.g., viewing history, demographics).
* **Relevant:** Lowering churn is vital in the competitive streaming industry, where retaining customers is more cost-effective than acquiring new ones, supporting revenue stability and market share.
* **Time-bound:** Deploy the solution by November 30, 2025, completing the project within six months.

**Relevant Questions to Understand the Business Model:**  
To fully grasp StreamHub’s context and ensure the problem is well-defined, we pose the following questions:

* What is the current customer acquisition cost (CAC), and how does it compare to retention costs?
* What are the main reasons for churn based on customer feedback or exit surveys?
* How does StreamHub’s churn rate stack up against industry benchmarks (e.g., Netflix, Spotify)?
* What retention strategies are currently implemented, and how effective are they?
* How does churn affect the business? Beyond immediate revenue loss, it impacts customer lifetime value (CLV) and increases reliance on costly customer acquisition.
* Are there specific customer segments (e.g., by subscription tier or region) with higher churn rates?

These questions anchor the technical work in StreamHub’s operational realities, ensuring the solution addresses the most pressing business needs.

**1.2 Translate Business Problems into Data-Driven Objectives**

**Formulate Hypotheses:**  
To guide the data science effort, we propose the following hypotheses based on potential churn drivers:

1. **Hypothesis 1:** Customers with low engagement (e.g., less than 5 hours of weekly watch time) are more likely to churn.
2. **Hypothesis 2:** Newer subscribers (e.g., tenure under 3 months) have a higher churn probability.
3. **Hypothesis 3:** Churn rates spike due to seasonal trends, such as post-holiday cancellations.
4. **Hypothesis 4**: Customers who haven’t logged in or streamed content in the last 30 days are more likely to churn.
5. **Hypothesis 5**: Those who downgraded from a premium to a basic plan might signal dissatisfaction or intent to leave.
6. **Hypothesis 6**: A drop-in watch time or playlist activity (e.g., fewer hours streamed, or songs added) could indicate waning interest.
7. **Hypothesis 7**: Younger users (e.g., 18-24) or those in competitive markets might churn more due to price sensitivity or alternatives.

These hypotheses will shape feature engineering and model development, focusing on key predictors of churn.

**Determine if Data Science/ML is the Right Approach:**

* **Why ML?** Churn prediction involves classifying customers as “likely to churn” or “not likely to churn” based on historical patterns, making supervised machine learning an ideal fit.
* **Temporal Dynamics:** The problem includes time-related factors (e.g., subscription tenure, seasonal effects), suggesting potential use of time-series analysis or survival modeling to capture churn behavior over time.

Machine learning is appropriate here because it can uncover complex patterns in large datasets that simpler rule-based systems might miss.

**Define Measurable Success Criteria:**  
Success must be quantifiable from both technical and business perspectives:

* **Technical Metrics:**
  + Achieve an AUC-ROC of ≥ 0.85 to ensure strong model performance.
  + Balance precision and recall to minimize false positives (avoiding wasteful retention efforts) and false negatives (missing at-risk customers).
* **Business Metrics:**
  + Reduce churn rate by 10% (from 5% to 4.5%) within six months.
  + Identify at least 70% of potential churners for targeted interventions.
  + Generate an estimated annual ROI of $1.75 million by retaining 50% of identified churners (350,000 users × $10 ARPU × 12 months).

These metrics bridge the gap between ML performance and business impact, ensuring the model moves the needle on StreamHub’s key priorities.

**1.3 Stakeholder Engagement**

**Identify Key Stakeholders:**  
Engaging the right people early ensures alignment and resource support:

* **Marketing Team:** Designs and executes retention campaigns based on model insights.
* **Customer Service Team:** Offers qualitative insights into churn reasons and customer pain points.
* **Data Engineering Team:** Provides data access and supports model deployment.
* **Executive Team (e.g., CEO, CFO):** Approves the project based on its strategic value and financial return.

**Communication Plan:**  
To keep stakeholders informed and aligned:

* **Bi-Weekly Updates:** Held every other Wednesday at 10:00 AM (StreamHub’s local time) to share progress, address challenges, and gather input.
* **Feedback Sessions:** Conducted after key milestones (e.g., exploratory data analysis, model evaluation) to validate findings and refine goals.
* **Documentation:** Maintained in a shared repository (e.g., Confluence or Google Drive) with meeting notes, reports, and deliverables for transparency and reference.

This structured communication ensures stakeholder expectations are met, and the project stays on track.

**Outcome of Phase 1**

* **Validated Problem Statement:** “Develop a machine learning model to predict customer churn for StreamHub, Inc., enabling proactive retention strategies that reduce churn by 10% and save $500,000 monthly in revenue.”
* **Project Plan:** A six-month roadmap with key milestones, such as data assessment by July 31, 2025, and model deployment by November 30, 2025.
* **Ethics Guardrail:** Ensure predictions are fair across demographics (e.g., no bias by age or region) and comply with data privacy laws (e.g., GDPR).
* **Stakeholder Buy-In:** Achieved through initial meetings, aligning on objectives and clarifying roles.

**Note:** The objectives and hypotheses may evolve as we uncover deeper insights from the data (e.g., new churn drivers during analysis), requiring iterative adjustments to keep the project aligned with business needs.

**Why This Project Matters**

StreamHub’s 5% monthly churn rate threatens long-term profitability, especially in a market where acquiring new customers costs significantly more (e.g., $50–$100 per subscriber) than retaining existing ones. By reducing churn, StreamHub can preserve revenue, improve customer lifetime value, and strengthen its competitive position. While ML metrics like AUC-ROC gauge model quality, the business cares about outcomes—lower churn and higher ROI. This project bridges that gap by tying technical success to tangible financial gains, ensuring the effort delivers real value.

This Business Understanding phase sets a clear, actionable foundation for the churn prediction project, balancing StreamHub’s strategic goals with data-driven strategies.

**Phase 2: Data Understanding and Governance**

**Goal:** Inventory, assess, and govern the data to ensure it is fit for purpose and compliant with relevant regulations.

In this phase, we focus on understanding the data available at StreamHub, Inc. to support the customer churn prediction project. We will identify data sources, assess their quality, and evaluate their potential to address the business problem of reducing churn by 10% (from 5% to 4.5%) within six months. This phase ensures the data is reliable, relevant, and governed appropriately, laying the groundwork for subsequent analyses. The work aligns with the CRISP-DM methodology and builds on the Business Understanding phase completed previously.

**2.1 Data Source and Collection**

**Interaction with Domain Experts and Stakeholders:**  
To identify relevant data sources, we engaged with StreamHub’s data engineering team, marketing team, and customer service leads during bi-weekly stakeholder meetings (initiated June 1, 2025). These discussions clarified the data landscape and its alignment with the churn prediction objective. Key questions included:

* What data is collected on subscriber behavior and demographics?
* How is churn status recorded, and what defines a churn event?
* Are there timestamps or temporal metadata to capture trends or seasonality?

**Identified Data Sources:**  
Based on stakeholder input, StreamHub provides access to the following datasets, stored in a cloud-based data warehouse (e.g., Amazon Redshift):

* **User Demographics (10M rows, ~500MB):**
  + Fields: user\_id, age, gender, location (country/region), signup\_date.
  + Relevance: Demographic factors (e.g., age, region) may correlate with churn risk.
* **Viewing History (100M rows, ~5GB):**
  + Fields: user\_id, content\_id, content\_type (movie, series, music), watch\_time (minutes), device\_type, timestamp.
  + Relevance: Engagement metrics (e.g., watch time) are likely strong predictors of churn.
* **Subscription Details (10M rows, ~600MB):**
  + Fields: user\_id, plan\_type (basic, standard, premium), payment\_method, billing\_date, churn\_status (0 = active, 1 = churned), promotion\_applied (yes/no).
  + Relevance: Plan type and churn status are critical for defining the target variable and understanding pricing sensitivity.
* **Customer Interactions (5M rows, ~300MB):**
  + Fields: user\_id, interaction\_type (support ticket, login, app review), interaction\_date, sentiment\_score (for support tickets).
  + Relevance: Customer complaints or low login frequency may signal churn risk.

**Data Collection and Storage:**

* **Collection Process:** Data is collected in real-time via StreamHub’s platform (e.g., app logs for viewing history) and batch-processed daily into the data warehouse.
* **Storage:** Hosted on Amazon Redshift with access granted via SQL queries and Python APIs (e.g., boto3).
* **Access:** StreamHub’s data engineering team provided credentials and schema documentation by June 15, 2025, ensuring seamless access for analysis.

**Relevance to Business Problem:**  
These datasets cover key aspects of customer behavior (engagement, interactions), attributes (demographics, plan type), and outcomes (churn status), making them highly relevant for predicting churn. The presence of timestamps supports temporal analysis, critical for hypotheses like seasonal churn spikes.

**2.2 Data Quality Assessment**

To evaluate the suitability of these datasets, we conducted a thorough quality assessment using Python libraries (e.g., Pandas, Matplotlib, Seaborn) on a sample of the data (e.g., 1M rows per dataset) extracted on June 20, 2025. The assessment focused on missing values, anomalies, outliers, distributions, relationships, and temporal characteristics.

**Missing Values:**

* **User Demographics:**
  + age: 5% missing (e.g., users skipped input during signup).
  + gender: 8% missing (non-mandatory field).
  + location: 1% missing (geolocation failures).
* **Viewing History:**
  + watch\_time: 2% missing (incomplete session logs).
  + device\_type: 0.5% missing (untracked devices).
* **Subscription Details:**
  + payment\_method: 3% missing (e.g., third-party payment issues).
  + churn\_status: 0% missing (well-maintained).
* **Customer Interactions:**
  + sentiment\_score: 20% missing (only available for support tickets, not logins).

**Anomalies and Outliers:**

* **Viewing History:**
  + Anomalies: Negative watch\_time values (0.1% of records, likely logging errors).
  + Outliers: Extreme watch\_time values (e.g., >10 hours in one session, top 1% of distribution).
* **Subscription Details:**
  + Anomalies: Duplicate user\_id entries (0.05%, due to billing errors).
* **Customer Interactions:**
  + Outliers: Users with >50 interactions/month (top 0.5%, potentially bots or power users).

**Summary Statistics and Visualizations:**

* **User Demographics:**
  + age: Mean = 35, Median = 32, Std = 12 (right-skewed, visualized via histogram).
  + location: 60% of users in North America, 25% in Europe (bar plot).
* **Viewing History:**
  + watch\_time: Mean = 120 min/week, Median = 90 min (skewed, log-transformed for visualization).
  + Correlation: Positive relationship between watch\_time and plan\_type (premium users watch more, scatter plot).
* **Subscription Details:**
  + plan\_type: 50% basic, 30% standard, 20% premium (pie chart).
  + churn\_status: 5% churned monthly, consistent with industry norms.
* **Customer Interactions:**
  + interaction\_type: 70% logins, 25% support tickets, 5% reviews (stacked bar).
  + sentiment\_score: Mean = 0.6 (scale 0–1, slightly positive, histogram).

**Temporal Components:**

* **Viewing History:** Timestamps (timestamp) span January 2023 to June 2025, enabling analysis of trends and seasonality. Preliminary line plots show higher watch times in winter months (November–January).
* **Subscription Details:** billing\_date and churn\_status timestamps allow tracking of churn events over time. A time-series plot indicates potential spikes in January (post-holiday).
* **Customer Interactions:** interaction\_date supports temporal analysis of engagement (e.g., login frequency dropping before churn).

**Significance of Temporal Data:**  
The presence of timestamps is critical for testing hypotheses like seasonal churn (Hypothesis 3: “Churn spikes post-holidays”). Temporal features (e.g., days since last login) will be engineered in Phase 4 to capture these dynamics, potentially justifying time-series or survival models.

**Quality Scores:**

* **Completeness:** 92% (average across datasets, accounting for missing values).
* **Accuracy:** 99% (after flagging anomalies like negative watch times).
* **Consistency:** 98% (minor issues with duplicates resolved).

**2.3 Iterative Feedback**

**Insights and Refinements:**  
The data quality assessment revealed several insights that refine our approach:

* **High Relevance:** The datasets cover engagement (watch\_time), customer attributes (age, plan\_type), and outcomes (churn\_status), aligning well with the churn prediction goal.
* **Temporal Opportunities:** Timestamps enable analysis of seasonal patterns and recency metrics, supporting time-based features and models.
* **Quality Gaps:** Missing sentiment\_score (20%) and age (5%) may limit some analyses, but imputation strategies (e.g., median age) can mitigate this. Anomalies (e.g., negative watch\_time) require cleaning in Phase 4.
* **Potential Enrichment:** Stakeholder feedback suggests adding content ratings (e.g., user reviews of movies) to capture satisfaction, which could enhance prediction. We recommend StreamHub explore collecting this data, though it’s out of scope unless provided.

**Refined Problem Statement:**  
Based on data availability, we refine the problem statement slightly:  
“Develop a machine learning model to predict customer churn for StreamHub, Inc., using subscriber engagement, demographics, and subscription data, enabling retention strategies that reduce churn by 10% and save $500,000 monthly, with a focus on temporal patterns like seasonality.”

This refinement emphasizes the role of temporal data, aligning with insights from the assessment.

**Privacy Safeguards:**

* **Compliance:** All data handling complies with GDPR, as StreamHub operates globally. Personal identifiers (user\_id) are pseudonymized, and sensitive fields (age, location) are aggregated where possible.
* **Governance:** We established a data lineage log tracking access, transformations, and storage. Access is restricted to authorized team members via role-based permissions.

**Outcome of Phase 2**

* **Comprehensive Understanding:** The datasets are well-suited for churn prediction, covering engagement, demographics, and churn outcomes, with timestamps enabling temporal analysis.
* **Data Quality Profile:**
  + Completeness: 92% (addressable via imputation).
  + Anomalies: Minor issues (e.g., negative watch\_time) to be cleaned.
  + Temporal Relevance: Strong, with seasonality and recency metrics viable.
* **Governed Dataset:**
  + Lineage: Documented in a shared repository (e.g., GitHub).
  + Quality Scores: 92% completeness, 99% accuracy, 98% consistency.
  + Privacy: GDPR-compliant with pseudonymized data and restricted access.
* **Deliverable:** **Data Assessment Report** (completed July 31, 2025), summarizing sources, quality, visualizations (e.g., histograms, time-series plots), and recommendations for enrichment (e.g., content ratings).

**Next Steps:**  
The insights from this phase confirm the data’s potential to address the churn problem, with minor quality issues to resolve in Phase 4 (Data Preparation). We will proceed to Phase 3 (Exploratory Data Analysis) to uncover patterns and test hypotheses, using the profiled datasets. Stakeholder approval of the Data Assessment Report was secured on July 30, 2025, ensuring alignment.

**Phase 3: Exploratory Data Analysis & Insight Generation**

**Goal:** Uncover signals, patterns, and actionable insights to refine hypotheses and guide churn prediction for StreamHub, Inc.

In this phase, we perform a deep dive into StreamHub’s datasets to identify patterns, relationships, and anomalies that inform the customer churn prediction project. We test the seven hypotheses outlined in the Business Understanding phase using univariate, bivariate, multivariate, and time-series analyses. Advanced visualizations (e.g., heatmaps, scatter plots, time-series plots) support our findings. This phase builds on the Data Understanding and Governance phase, leveraging the profiled datasets (10M users, covering demographics, viewing history, subscription details, and customer interactions). The outcomes will prioritize predictive features, refine the problem statement, and provide a go/no-go decision for modeling.

**3.1 Data Exploration**

We analyzed the datasets using Python libraries (Pandas, NumPy, Matplotlib, Seaborn, Statsmodels) on a sampled subset (1M users) and the full dataset where feasible, processed on StreamHub’s cloud infrastructure (Amazon Redshift) between August 1 and August 31, 2025. The exploration focused on uncovering patterns, trends, and relationships relevant to churn, with a particular emphasis on temporal dynamics and the seven hypotheses.

**Univariate Analysis**

* **User Demographics:**
  + age: Right-skewed (mean = 35, median = 32, std = 12). Histogram shows a peak at 25–35 years, with fewer users >50.
  + gender: 52% male, 40% female, 8% unspecified (pie chart).
  + location: 60% North America, 25% Europe, 15% other (bar plot).
* **Viewing History:**
  + watch\_time (weekly): Highly skewed (mean = 120 min, median = 90 min). Log-transformed histogram reveals most users watch <200 min/week.
  + content\_type: 50% movies, 40% series, 10% music (stacked bar).
* **Subscription Details:**
  + plan\_type: 50% basic, 30% standard, 20% premium (pie chart).
  + churn\_status: 5% churned monthly, consistent with industry norms (bar plot).
* **Customer Interactions:**
  + interaction\_type: 70% logins, 25% support tickets, 5% reviews (stacked bar).
  + sentiment\_score (support tickets): Mean = 0.6 (scale 0–1, slightly positive, histogram).

**Bivariate Analysis**

* **Churn vs. Watch Time:** Users who churned had lower median watch\_time (60 min/week) than active users (100 min/week). Box plot shows significant difference (p < 0.01, Mann-Whitney U test).
* **Churn vs. Plan Type:** Churn rate is highest for basic plan (6%) vs. standard (4%) and premium (3%) (stacked bar). Chi-square test confirms significance (p < 0.01).
* **Churn vs. Age:** Younger users (18–24) have a higher churn rate (7%) than older users (3% for 35+). Bar plot and t-test (p < 0.05).
* **Churn vs. Sentiment Score:** Churners had lower sentiment\_score (mean = 0.4) vs. active users (0.7). Box plot, significant (p < 0.01).

**Multivariate Analysis**

* **Correlation Heatmap:**
  + Strong negative correlation between watch\_time and churn\_status (-0.45).
  + Moderate positive correlation between plan\_type (basic) and churn\_status (0.30).
  + Weak correlation between age and watch\_time (-0.15), suggesting younger users watch slightly less.
* **Interaction Effects:**
  + Users with low watch\_time (<60 min/week) and basic plans have a churn rate of 8%, vs. 2% for high watch\_time (>200 min/week) and premium plans (3D scatter plot).

**Time-Series Analysis**

* **Churn Rate Over Time:** Monthly churn rates (January 2023–June 2025) show spikes in January (6.5%) and July (6%), suggesting post-holiday and mid-year trends (line plot).
* **Watch Time Trends:** Average watch\_time peaks in December (150 min/week) and dips in August (80 min/week), indicating seasonal engagement (time-series plot).
* **Autocorrelation:** watch\_time shows significant autocorrelation at lag 1 (0.65), suggesting user engagement is consistent month-to-month (ACF plot). This justifies time-series features like rolling averages.

**Anomalies Detected:**

* **Viewing History:** Spikes in watch\_time (>10 hours/session, top 0.1%) likely due to auto-play errors.
* **Customer Interactions:** Users with >50 support tickets/month (0.2%) may indicate system abuse or extreme dissatisfaction.

**Visualization Techniques:**

* Histograms for distributions (e.g., age, watch\_time).
* Box plots for churn vs. continuous variables (e.g., watch\_time, sentiment\_score).
* Bar plots for categorical variables (e.g., plan\_type, location).
* Heatmaps for correlations.
* Time-series plots for temporal trends (e.g., churn rate, watch\_time).
* Scatter plots for multivariate relationships (e.g., watch\_time vs. age vs. churn\_status).

**3.2 Hypothesis Testing**

We tested the seven hypotheses using statistical methods and visualizations, refining them based on findings. Tests were conducted at a 95% confidence level (p < 0.05).

* **Hypothesis 1: Customers with low engagement (e.g., <5 hours/week watch time) are more likely to churn.**
  + **Test:** Compared watch\_time for churned vs. active users (Mann-Whitney U test).
  + **Result:** Churners had significantly lower watch\_time (median = 60 min/week) than active users (100 min/week, p < 0.01). Box plot confirms users with <300 min/week (~5 hours) have a 7% churn rate vs. 3% for >300 min/week.
  + **Conclusion:** Supported. Low engagement is a strong churn predictor.
* **Hypothesis 2: Newer subscribers (e.g., tenure <3 months) have a higher churn probability.**
  + **Test:** Calculated churn rate by tenure (months since signup\_date, chi-square test).
  + **Result:** Users with <3 months tenure have an 8% churn rate vs. 4% for >12 months (p < 0.01). Bar plot shows a steep drop in churn after 6 months.
  + **Conclusion:** Supported. Short tenure is a key risk factor.
* **Hypothesis 3: Churn rates spike due to seasonal trends, such as post-holiday cancellations.**
  + **Test:** Analyzed monthly churn rates (January 2023–June 2025, ANOVA).
  + **Result:** Significant spikes in January (6.5%) and July (6%) vs. average 5% (p < 0.05). Time-series plot confirms post-holiday (January) and mid-year (July) peaks.
  + **Conclusion:** Supported. Seasonal patterns justify time-series features.
* **Hypothesis 4: Customers who haven’t logged in or streamed content in the last 30 days are more likely to churn.**
  + **Test:** Created a binary feature (inactive\_30\_days) based on interaction\_date or timestamp (chi-square test).
  + **Result:** Users with no activity in 30 days have a 10% churn rate vs. 3% for active users (p < 0.01). Bar plot highlights stark difference.
  + **Conclusion:** Supported. Inactivity is a critical predictor.
* **Hypothesis 5: Those who downgraded from a premium to a basic plan might signal dissatisfaction or intent to leave.**
  + **Test:** Identified users with plan\_type changes in subscription\_details (chi-square test).
  + **Result:** Users who downgraded from premium to basic (2% of users) have a 9% churn rate vs. 5% for stable plans (p < 0.01). Bar plot confirms higher risk.
  + **Conclusion:** Supported. Downgrading is a strong churn signal.
* **Hypothesis 6: A drop-in watch time or playlist activity could indicate waning interest.**
  + **Test:** Calculated month-over-month watch\_time change (delta\_watch\_time) and content\_id additions for music (t-test).
  + **Result:** Users with a >20% watch\_time drop have a 7.5% churn rate vs. 4% for stable/increasing (p < 0.01). Music users adding <5 songs/month churn at 6% vs. 3% for >10 songs (p < 0.05). Line plot shows declining watch\_time precedes churn.
  + **Conclusion:** Supported. Declining engagement is predictive.
* **Hypothesis 7: Younger users (e.g., 18–24) or those in competitive markets might churn more due to price sensitivity or alternatives.**
  + **Test:** Compared churn rates by age group and location (North America/Europe as competitive markets, chi-square test).
  + **Result:** Users aged 18–24 have a 7% churn rate vs. 3% for 35+ (p < 0.01). North America (6% churn) and Europe (5.5%) show higher rates than other regions (4%, p < 0.05). Bar plots confirm trends.
  + **Conclusion:** Supported. Younger users and competitive markets are high-risk.

**Refined Hypotheses:**

* Hypothesis 4 refined: “Inactivity >30 days is a stronger predictor than shorter periods (e.g., 7 days), with 10% churn rate.”
* Hypothesis 6 refined: “A >20% month-over-month drop in watch\_time or <5 songs added/month are specific thresholds for waning interest.”

**3.3 Documentation and Iteration**

**Documentation:**  
Findings are documented in a comprehensive **Exploratory Data Analysis Report** (completed August 31, 2025), including:

* Visualizations (histograms, box plots, heatmaps, time-series plots) saved as PNGs in a shared repository (GitHub).
* Statistical test results (p-values, effect sizes) in a Jupyter notebook.
* Narrative summary of insights, prioritized features, and implications for modeling.

**Iteration:**

* **Data Collection:** The strong signal from sentiment\_score (despite 20% missing) suggests collecting more customer feedback (e.g., app ratings) could enhance predictions. Recommended to StreamHub for future data enrichment (out of scope unless provided).
* **Business Assumptions:** High churn in competitive markets (North America, Europe) suggests pricing or competitor analysis may be needed. Marketing team will explore this post-project.

**Stakeholder Feedback:**  
Presented findings to stakeholders (marketing, customer service, executives) on August 30, 2025. Key feedback:

* Marketing team prioritized inactive\_30\_days and delta\_watch\_time for campaign triggers.
* Executives approved proceeding to modeling, citing strong ROI potential ($1.75M/year if 50% of identified churners are retained).

**Outcome of Phase 3**

* **Thorough Understanding of Data Characteristics:**
  + Engagement (watch\_time, inactive\_30\_days), tenure, plan changes, and demographics (age, location) are key churn drivers.
  + Temporal patterns (January/July spikes) justify time-series features.
* **Actionable Insights:**
  + Users with <5 hours/week watch\_time are 2.3x more likely to churn (7% vs. 3%).
  + Inactive users (>30 days) have a 10% churn rate, ideal for urgent retention campaigns.
  + Downgraders (premium to basic) and younger users (18–24) are high-risk, warranting tailored offers.
* **Prioritized Features:**
  + watch\_time (weekly, min).
  + inactive\_30\_days (binary).
  + tenure (months).
  + plan\_type (categorical).
  + delta\_watch\_time (month-over-month % change).
  + age (years).
  + location (region).
  + sentiment\_score (if available).
* **Refined Problem Statement:**  
  “Develop a machine learning model to predict customer churn for StreamHub, Inc., using engagement, tenure, plan changes, and demographic data, with a focus on temporal patterns and inactivity, to enable retention strategies that reduce churn by 10% and save $500,000 monthly.”
* **Go/No-Go Decision:**  
  **Go.** The data shows strong predictive signals (e.g., watch\_time, inactive\_30\_days), and stakeholder approval confirms business alignment. The project is ready for Phase 4 (Data Preparation & Feature Engineering).

**Deliverable:**

* **Exploratory Data Analysis Report** (August 31, 2025), including visualizations, statistical results, prioritized features, and recommendations. Available in the project repository for portfolio inclusion.

**Phase 4: Data Preparation & Feature Engineering**

**Goal:** Produce clean, enriched, reproducible datasets optimized for modeling and deployment to predict customer churn for StreamHub, Inc.

In this phase, we transform StreamHub’s raw datasets into a clean, structured format ready for machine learning, building on insights from the Exploratory Data Analysis (EDA) phase. We address data quality issues, engineer features based on prioritized predictors, automate preprocessing pipelines, and prepare a train/test split that respects the temporal nature of churn. The resulting dataset will be stored in a feature store for scalability and reproducibility. This phase, conducted between September 1 and September 30, 2025, ensures the data is optimized for modeling while maintaining security, compliance, and alignment with business objectives.

**4.1 Data Cleaning and Transformation**

**Collaboration with Domain Experts:**  
We collaborated with StreamHub’s data engineering and customer service teams via bi-weekly meetings to ensure correct interpretation of fields (e.g., churn\_status as a confirmed cancellation). These discussions clarified business rules, such as defining churn as a user who cancels their subscription without reactivation within 30 days.

**Addressing Data Quality Issues:**  
Using the datasets from Phase 2 (10M users across demographics, viewing history, subscription details, and customer interactions), we addressed issues identified in Phase 3 using Python (Pandas, NumPy) on Amazon Redshift.

* **Missing Values:**
  + **User Demographics:**
    - age (5% missing): Imputed with median (32 years), as distribution is skewed.
    - gender (8% missing): Assigned “unspecified” category to preserve data.
    - location (1% missing): Imputed with mode (North America, 60% of users).
  + **Viewing History:**
    - watch\_time (2% missing): Imputed with user’s median watch\_time if available, else dataset median (90 min/week).
    - device\_type (0.5% missing): Assigned “unknown” category.
  + **Subscription Details:**
    - payment\_method (3% missing): Assigned “other” category, as it’s not critical for prediction.
  + **Customer Interactions:**
    - sentiment\_score (20% missing): Left as-is, as it’s only available for support tickets. Models will handle missing values via imputation or exclusion.
* **Duplicates:**
  + **Subscription Details:** Removed 0.05% duplicate user\_id entries (5,000 rows), keeping the latest record based on billing\_date.
  + **Viewing History:** No duplicates found after aggregating sessions by user\_id and timestamp.
* **Outliers:**
  + **Viewing History:** Capped watch\_time at the 99th percentile (600 min/session) to handle extreme values (0.1% of records).
  + **Customer Interactions:** Flagged users with >50 interactions/month (0.2%) as potential anomalies but retained them, as they may indicate dissatisfaction.
* **Transformations for Modeling:**
  + **Normalization/Scaling:**
    - Continuous features (watch\_time, tenure, sentiment\_score) will be standardized (mean = 0, std = 1) in the pipeline to ensure model compatibility (e.g., for logistic regression).
  + **Encoding Categorical Variables:**
    - plan\_type (basic, standard, premium): One-hot encoded to create binary columns (e.g., plan\_basic, plan\_standard).
    - location (North America, Europe, other): One-hot encoded.
    - gender (male, female, unspecified): One-hot encoded.
  + **Dimensionality Reduction:**
    - Deferred to modeling phase, as the number of features (post-engineering) is manageable (<20). PCA or feature selection (e.g., based on SHAP values) may be applied if needed.

**Security and Scalability:**

* Transformations were implemented via automated Python scripts (using scikit-learn’s Pipeline) to ensure reproducibility and scalability.
* Data processing was conducted in StreamHub’s secure cloud environment, with access logged and restricted to authorized team members (GDPR compliance).

**4.2 Feature Engineering**

Based on EDA insights and the seven validated hypotheses, we engineered features to capture churn predictors. Features were developed using domain knowledge and prioritized based on their predictive power (e.g., watch\_time, inactive\_30\_days). All features were computed using SQL queries (for aggregation) and Python (for complex logic) and stored in a feature store (e.g., AWS Feature Store) for reuse.

**Engineered Features:**

* **weekly\_watch\_time (Continuous):**
  + Description: Average weekly watch time (min) per user, aggregated from viewing\_history (watch\_time).
  + Rationale: Hypothesis 1 (low engagement predicts churn). EDA showed churners have lower watch\_time (60 vs. 100 min/week).
  + Predictive Power: High (negative correlation with churn, -0.45).
* **inactive\_30\_days (Binary):**
  + Description: 1 if no logins or streams in the last 30 days (interaction\_date, timestamp), else 0.
  + Rationale: Hypothesis 4 (inactivity predicts churn). 10% churn rate for inactive users vs. 3% for active.
  + Predictive Power: High (chi-square p < 0.01).
* **tenure (Continuous):**
  + Description: Months since signup (signup\_date to latest billing\_date).
  + Rationale: Hypothesis 2 (newer subscribers churn more). 8% churn for <3 months vs. 4% for >12 months.
  + Predictive Power: Moderate (chi-square p < 0.01).
* **plan\_type (Categorical):**
  + Description: Current plan (basic, standard, premium), one-hot encoded.
  + Rationale: Hypothesis 5 (basic plan users churn more). 6% churn for basic vs. 3% for premium.
  + Predictive Power: Moderate (correlation 0.30).
* **delta\_watch\_time (Continuous):**
  + Description: Month-over-month % change in watch\_time.
  + Rationale: Hypothesis 6 (>20% drop indicates waning interest). 7.5% churn for >20% drop vs. 4% for stable.
  + Predictive Power: High (t-test p < 0.01).
* **age\_group (Categorical):**
  + Description: Binned age (18–24, 25–34, 35–44, 45+), one-hot encoded.
  + Rationale: Hypothesis 7 (younger users churn more). 7% churn for 18–24 vs. 3% for 35+.
  + Predictive Power: Moderate (chi-square p < 0.01).
* **location\_competitiveness (Categorical):**
  + Description: Region competitiveness (high: North America/Europe, low: other), one-hot encoded.
  + Rationale: Hypothesis 7 (competitive markets churn more). 6% churn in North America vs. 4% elsewhere.
  + Predictive Power: Moderate (chi-square p < 0.05).
* **sentiment\_score (Continuous):**
  + Description: Average sentiment score from support tickets (0–1).
  + Rationale: Lower scores (0.4) for churners vs. 0.7 for active users.
  + Predictive Power: Moderate (t-test p < 0.01, limited by 20% missing).
* **month\_of\_year (Categorical):**
  + Description: Billing month (1–12), one-hot encoded.
  + Rationale: Hypothesis 3 (seasonal churn spikes). January (6.5%) and July (6%) have higher churn.
  + Predictive Power: Moderate (ANOVA p < 0.05).
* **plan\_downgrade\_flag (Binary):**
  + Description: 1 if user downgraded from premium/standard to basic in the last 6 months, else 0.
  + Rationale: Hypothesis 5 (downgraders signal churn). 9% churn for downgraders vs. 5% for stable.
  + Predictive Power: High (chi-square p < 0.01).

**Evaluation of Features:**

* **Relevance:** All features align with hypotheses and EDA insights, capturing engagement, inactivity, tenure, plan changes, demographics, and seasonality.
* **Predictive Power:** Assessed via statistical tests (e.g., correlation, chi-square) and visualized in EDA (e.g., box plots, heatmaps). High-power features (watch\_time, inactive\_30\_days, delta\_watch\_time, plan\_downgrade\_flag) will be prioritized in modeling.
* **Feature Store:** Features were stored with metadata (e.g., creation date, source table) in AWS Feature Store, enabling reuse and versioning.

**4.3 Advanced Preprocessing and Automation**

**Automation of Preprocessing Tasks:**  
To ensure scalability and reproducibility, we developed automated pipelines using scikit-learn and Airflow:

* **Data Quality Checks:** Scripts to flag missing values (>10%), duplicates, or anomalies (e.g., negative watch\_time) before processing.
* **Transformation Pipeline:**
  + Imputation: Median for continuous (age, watch\_time), mode for categorical (location).
  + Encoding: One-hot encoding for plan\_type, age\_group, location\_competitiveness, month\_of\_year.
  + Scaling: Standardization for continuous features.
  + Capping: Outliers capped at 99th percentile (e.g., watch\_time).
* **Scheduling:** Airflow DAGs run daily to update features with new data, ensuring freshness for deployment.

**Optimal Data Types and Structures:**

* **Data Types:**
  + Continuous: Float32 for watch\_time, delta\_watch\_time, sentiment\_score, tenure.
  + Categorical: String for plan\_type, age\_group, location\_competitiveness, month\_of\_year (pre-encoding).
  + Binary: Int8 for inactive\_30\_days, plan\_downgrade\_flag.
* **Aggregation for Time-Series Data:**
  + Aggregated watch\_time to weekly averages to reduce noise and align with Hypothesis 1.
  + Computed delta\_watch\_time and inactive\_30\_days using rolling windows (30–60 days) to capture temporal trends.

**Output Format:**

* A clean dataset (10M rows, ~15 features post-encoding) stored as a Parquet file for efficiency, with schema documented in the feature store.

**4.4 Train/Test Splitting**

**Splitting Method:**  
Given the temporal nature of churn (e.g., seasonal patterns, recency metrics), we used a **time-based split** to prevent data leakage and align with real-world use (predicting future churn):

* **Training Set:** January 2023–April 2025 (80% of data, ~8M users).
* **Test Set:** May 2025–June 2025 (20% of data, ~2M users).
* **Validation Set:** Within training set, used 5-fold cross-validation with time-based splits (e.g., last 3 months of training as validation fold).

**Rationale:**

* Time-based splitting ensures the model learns from historical patterns and is evaluated on recent data, mimicking production scenarios where predictions are made for future periods.
* The split preserves temporal dependencies (e.g., seasonality, delta\_watch\_time), critical for Hypotheses 3 and 6.

**Implementation:**

* Split performed using SQL queries to filter by billing\_date or timestamp.
* Verified class balance: ~5% churn rate in both train and test sets, ensuring representativeness.

**Outcome of Phase 4**

* **Refined, Clean Dataset:**
  + 10M rows, ~15 features (post-encoding), stored as Parquet in Amazon S3.
  + Quality: 100% complete (post-imputation), no duplicates, outliers capped.
  + Schema: Documented with data types, ranges, and descriptions.
* **Feature-Rich Dataset:**
  + 10 engineered features (weekly\_watch\_time, inactive\_30\_days, tenure, plan\_type, delta\_watch\_time, age\_group, location\_competitiveness, sentiment\_score, month\_of\_year, plan\_downgrade\_flag) aligned with EDA insights and hypotheses.
  + Stored in AWS Feature Store with metadata for reuse.
* **Automated Pipelines:**
  + Preprocessing pipeline (imputation, encoding, scaling) implemented in scikit-learn.
  + Airflow DAGs for daily feature updates and quality checks.
  + Security: GDPR-compliant, with access controls and data lineage logged.
* **Train/Test Split:**
  + Time-based: Training (Jan 2023–Apr 2025), Test (May–Jun 2025).
  + Balanced: ~5% churn rate in both sets.
* **Deliverable:**
  + **Data Preparation Report** (completed September 30, 2025), documenting cleaning steps, feature engineering logic, pipeline code, and train/test split details. Available in the GitHub repository with Jupyter notebooks for portfolio inclusion.

**Next Steps:**  
The clean, feature-rich dataset is ready for Phase 5 (Modeling & Experimentation). Stakeholder approval of the Data Preparation Report was secured on September 29, 2025, confirming readiness for modeling. The automated pipelines and feature store allows scalability and production-readiness.

**Phase 5: Modeling & Experimentation**

**Goal:** Identify, train, and track candidate models that balance predictive power, interpretability, and cost for predicting customer churn at StreamHub, Inc., while ensuring reliability, scalability, maintainability, and adaptability.

In this phase, we develop and evaluate machine learning models to predict customer churn, building on the clean, feature-rich dataset from Phase 4. We frame the problem, select algorithms, track experiments, ensure interpretability, and address fairness, aligning with business objectives (10% churn reduction, $500,000 monthly savings). The modeling process, conducted between October 1 and October 31, 2025, incorporates the four ML design requirements: reliability (consistent performance), scalability (handles 10M users), maintainability (easy updates), and adaptability (adjusts to new patterns). The outcome is a shortlist of validated models ready for evaluation in Phase 6.

**5.1 Problem Framing**

**Problem Type:**

* **Binary Classification:** Predict whether a customer will churn (churn\_status = 1) or not (churn\_status = 0) based on features like weekly\_watch\_time, inactive\_30\_days, and plan\_downgrade\_flag.
* **Rationale:** Churn is a discrete outcome (churn/no churn), and the business needs a probability score to prioritize retention efforts.

**Loss Function Aligned to KPIs:**

* **Primary Loss Function:** Weighted log-loss (binary cross-entropy) to penalize misclassifications of churners more heavily, as false negatives (missing at-risk customers) are costlier than false positives (over-targeting). Weight ratio: 3:1 for churners vs. non-churners (reflecting 5% churn rate).
* **Evaluation Metrics:**
  + **Primary Metric:** AUC-ROC (target ≥ 0.85), as it balances precision and recall across thresholds, aligning with the technical objective.
  + **Secondary Metrics:**
    - Precision (minimize false positives for cost-effective campaigns).
    - Recall (maximize identification of churners, targeting 70% of at-risk users).
    - F1-score (harmonic mean of precision and recall).
  + **Business KPI:** Identify 70% of churners to enable retention campaigns, projecting $1.75M annual savings (50% of 350,000 identified churners retained at $10 ARPU).

**Temporal Considerations:**

* EDA (Phase 3) confirmed seasonal patterns (e.g., January churn spikes) and recency effects (e.g., inactive\_30\_days). While time-series forecasting (e.g., ARIMA) was considered, binary classification with temporal features (e.g., month\_of\_year, delta\_watch\_time) is sufficient, as the goal is to predict churn probability for the next billing cycle, not long-term trends.
* **Future Evolution:** If StreamHub requires multi-month churn forecasts, we could pivot to survival analysis (e.g., Cox regression) in future iterations.

**5.2 Integrate Experimentation**

**Experiment Tracking:**

* **Tool:** MLflow, hosted on StreamHub’s AWS infrastructure, to log:
  + Code version (Git commit hash).
  + Model parameters (e.g., max\_depth for trees).
  + Hyperparameters (e.g., learning rate).
  + Metrics (AUC-ROC, precision, recall, F1-score).
  + Artifacts (e.g., confusion matrices, SHAP plots).
* **Setup:** Each experiment is logged as a run under a project named “StreamHub\_Churn\_Prediction,” with tags for algorithm and date (e.g., “XGBoost\_2025-10-15”).

**Model Registry:**

* **Implementation:** MLflow Model Registry to store and version models (e.g., “XGBoost\_v1.0”).
* **Process:** Top-performing models are registered with metadata (e.g., AUC-ROC, training date) and tagged as “Staging” or “Production” for deployment.

**Benefits:**

* Ensures reproducibility (reliability).
* Enables comparison of models (maintainability).
* Supports versioning for updates (adaptability).

**5.3 Algorithm Selection and Experimentation**

**Model Choice Criteria:**

* **Predictive Power:** AUC-ROC ≥ 0.85, recall ≥ 0.70 to meet technical and business goals.
* **Interpretability:** Models must support SHAP/LIME explanations for stakeholder trust (e.g., why a user is at risk).
* **Scalability:** Must handle 10M users efficiently (e.g., fast inference for real-time API).
* **Maintainability:** Easy to retrain with new data (e.g., minimal manual tuning).
* **Adaptability:** Robust to concept drift (e.g., changing user behavior).

**Selected Algorithms:**  
We evaluated four algorithms based on their suitability for binary classification, scalability, and interpretability:

* **Logistic Regression (Baseline):**
  + Pros: Highly interpretable, fast, scalable.
  + Cons: Limited to linear relationships, may underperform on complex patterns.
* **Random Forest:**
  + Pros: Handles non-linearities, robust to outliers, interpretable via feature importance.
  + Cons: Slower inference, higher memory usage.
* **XGBoost:**
  + Pros: High performance, handles imbalanced data, scalable with GPU support.
  + Cons: Less interpretable, requires tuning.
* **LightGBM:**
  + Pros: Fast training/inference, scalable, handles large datasets.
  + Cons: Similar interpretability challenges as XGBoost.

**Experimentation Process:**

* **Data:** Used the training set (Jan 2023–Apr 2025, 8M users) from Phase 4, with 5-fold time-based cross-validation to assess robustness.
* **Hyperparameter Tuning:**
  + **Logistic Regression:** Grid search on regularization strength (C = [0.01, 0.1, 1, 10]).
  + **Random Forest:** Random search on n\_estimators (50–200), max\_depth (5–20).
  + **XGBoost:** Bayesian optimization (via Optuna) on learning\_rate (0.01–0.3), max\_depth (3–10), n\_estimators (100–500).
  + **LightGBM:** Similar Bayesian optimization for learning\_rate, num\_leaves (20–100), n\_estimators (100–500).
* **Temporal Features:** Included month\_of\_year, delta\_watch\_time, and inactive\_30\_days to capture seasonality and recency, aligning with EDA findings.

**Results (Validation Set):**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **AUC-ROC** | **Recall** | **Precision** | **F1-Score** | **Training Time** | **Inference Time (1M users)** |
| Logistic Regression | 0.78 | 0.65 | 0.60 | 0.62 | 5 min | 1 min |
| Random Forest | 0.83 | 0.68 | 0.65 | 0.66 | 30 min | 5 min |
| XGBoost | **0.86** | **0.72** | **0.70** | **0.71** | 20 min | 2 min |
| LightGBM | 0.85 | 0.70 | 0.68 | 0.69 | 15 min | 1.5 min |

* **Analysis:**
  + XGBoost outperformed others, achieving AUC-ROC = 0.86 and recall = 0.72, meeting the target (≥ 0.85, ≥ 0.70).
  + LightGBM was competitive but slightly less accurate.
  + Logistic Regression underperformed due to non-linear patterns (e.g., watch\_time interactions).
  + Random Forest was robust but slower, less suitable for real-time inference.

**Shortlisted Models:**

* **XGBoost:** Best performance, scalable with GPU, adaptable via retraining.
* **LightGBM:** Backup due to fast inference and similar performance.

**5.4 Interpretability and Explainability**

**SHAP Explanations:**

* **Implementation:** Used SHAP (SHapley Additive exPlanations) to compute feature importance and individual prediction explanations for XGBoost and LightGBM.
* **Findings:**
  + **Global Importance:** Top features: inactive\_30\_days (30% impact), weekly\_watch\_time (25%), plan\_downgrade\_flag (15%), delta\_watch\_time (10%).
  + **Example Prediction:** For a user with inactive\_30\_days = 1 and weekly\_watch\_time = 50 min, SHAP shows +0.35 churn probability due to inactivity and -0.10 due to moderate watch time.
  + **Visualization:** SHAP summary plot (beeswarm) and force plots for individual users, saved as artifacts in MLflow.

**LIME Explanations:**

* **Implementation:** Applied LIME (Local Interpretable Model-agnostic Explanations) to 100 test cases to validate SHAP results.
* **Findings:** Consistent with SHAP, highlighting inactive\_30\_days and watch\_time as key drivers. LIME explanations were simpler, ideal for stakeholder presentations.

**Stakeholder Benefit:**

* Explanations enable marketing to understand why users are flagged (e.g., “low engagement”), supporting targeted campaigns (reliability, maintainability).

**5.5 Bias, Fairness, and Ethical Considerations**

**Fairness Metrics:**

* **Demographic Parity:** Ensured churn predictions have similar positive rates across age\_group and location\_competitiveness.
  + **Result:** XGBoost shows parity within 5% across groups (e.g., 18–24 vs. 35+), acceptable for business use.
* **Equalized Odds:** Checked if true positive rates (recall) and false positive rates are balanced.
  + **Result:** Recall for 18–24 (0.73) vs. 35+ (0.71) is balanced, but false positives slightly higher for younger users (0.25 vs. 0.20). Mitigated by adjusting prediction threshold (0.5 to 0.6).

**Bias Evaluation:**

* **Age Bias:** Younger users (18–24) were flagged more due to higher churn rates (7% vs. 3%, per EDA). This reflects real patterns, not model bias, but we monitored for over-targeting.
* **Location Bias:** North America/Europe predictions aligned with higher churn (6% vs. 4%), justified by competitive markets.

**Ethical Considerations:**

* **Transparency:** SHAP/LIME explanations ensure stakeholders understand predictions, avoiding black-box decisions.
* **Privacy:** All data is pseudonymized (user\_id), and GDPR compliance is maintained.
* **Impact:** Predictions feed into retention campaigns (e.g., discounts), with no adverse social impact (e.g., no credit or hiring decisions).

**Mitigation:**

* Adjusted thresholds to balance false positives across demographics.
* Documented fairness metrics in MLflow for transparency (maintainability).

**5.6 Iteration and Refinement**

**Validation Performance Feedback:**

* Initial XGBoost runs showed high false positives (0.30), reducing precision. Iterated by:
  + Adding feature interactions (e.g., watch\_time × plan\_type).
  + Tuning class weights to prioritize recall.
* LightGBM required fewer iterations due to faster training, but XGBoost retained superior AUC-ROC.

**Business Feedback:**

* Marketing team requested higher recall (≥ 0.75) for broader campaign coverage. Adjusted XGBoost threshold to 0.4, increasing recall to 0.75 at the cost of precision (0.65).
* Stakeholders approved XGBoost as the primary model on October 29, 2025, citing alignment with ROI goals ($1.75M/year).

**Design Requirements:**

* **Reliability:** Consistent AUC-ROC (0.86 ± 0.01 across folds) and SHAP explanations ensure trustworthy predictions.
* **Scalability:** XGBoost/LightGBM handle 10M users with <2 min inference (AWS EC2, GPU-enabled).
* **Maintainability:** MLflow tracks experiments, and pipelines automate retraining.
* **Adaptability:** Temporal features (month\_of\_year, delta\_watch\_time) and scheduled retraining handle evolving patterns.

**Outcome of Phase 5**

* **Shortlisted Models:**
  + **Primary:** XGBoost (AUC-ROC = 0.86, Recall = 0.75, Precision = 0.65, F1 = 0.70).
  + **Backup:** LightGBM (AUC-ROC = 0.85, Recall = 0.70, Precision = 0.68, F1 = 0.69).
* **Tracked Lineage:**
  + All experiments logged in MLflow with parameters, metrics, and artifacts.
  + Models registered as “XGBoost\_v1.0” and “LightGBM\_v1.0” in MLflow Model Registry.
* **Explainability:**
  + SHAP/LIME explanations highlight inactive\_30\_days, weekly\_watch\_time, and plan\_downgrade\_flag as key drivers.
  + Visualizations (beeswarm, force plots) stored for stakeholder review.
* **Bias Assessment:**
  + Fairness metrics (demographic parity, equalized odds) within acceptable ranges.
  + Threshold adjustments mitigate over-targeting of younger users.
* **Plan for Interpretation and Ethical Evaluation:**
  + **Interpretation:** SHAP summary plots and LIME case studies will be presented to stakeholders in Phase 6 to explain predictions (e.g., “User X has 80% churn risk due to 40-day inactivity”).
  + **Ethical Evaluation:** Fairness metrics and GDPR compliance will be reviewed pre-deployment to ensure ethical use.
* **Deliverable:**
  + **Model Development Report** (completed October 31, 2025), detailing problem framing, algorithm selection, experiment results, SHAP/LIME explanations, fairness metrics, and model registry details. Available in the GitHub repository with Jupyter notebooks for portfolio inclusion.

**Next Steps:**  
The shortlisted models (XGBoost, LightGBM) are ready for evaluation on the test set (May–June 2025) in Phase 6 (Model Evaluation & Business Review). Stakeholder approval of the Model Development Report was secured on October 30, 2025, confirming alignment with business objectives. The rigorous experimentation and fairness focus on technical depth and ethical responsibility.

**Phase 6: Model Evaluation & Business Review**

**Goal:** Ensure the selected model meets technical thresholds and delivers real-world business value for StreamHub, Inc.’s customer churn prediction project without undue risk.

In this phase, we evaluate the shortlisted models (XGBoost and LightGBM) from Phase 5 on the test set (May–June 2025, ~2M users) to confirm their performance against technical and business metrics. We conduct error analysis, benchmark against a baseline, and perform a cost-benefit analysis to quantify ROI. The findings are presented to stakeholders to secure a go/no-go decision for deployment. This phase, conducted between November 1 and November 15, 2025, ensures the model is practical, aligned with business objectives (10% churn reduction, $500,000 monthly savings), and mitigates risks. The process adheres to the four ML design requirements: reliability, scalability, maintainability, and adaptability.

**6.1 Performance Metrics**

**Evaluation Setup:**

* **Dataset:** Test set (May–June 2025, ~2M users, 5% churn rate) from Phase 4, unseen during training to ensure unbiased evaluation.
* **Models Evaluated:**
  + Primary: XGBoost (AUC-ROC = 0.86, recall = 0.75 on validation).
  + Backup: LightGBM (AUC-ROC = 0.85, recall = 0.70 on validation).
* **Baseline Model:** Random guess (predicts churn with 5% probability, matching dataset prevalence).

**Standard Metrics (Test Set):**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **AUC-ROC** | **Recall** | **Precision** | **F1-Score** | **Accuracy** |
| XGBoost | **0.85** | **0.74** | **0.66** | **0.70** | 0.92 |
| LightGBM | 0.84 | 0.71 | 0.64 | 0.67 | 0.91 |
| Random Guess | 0.50 | 0.05 | 0.05 | 0.05 | 0.95 |

* **Analysis:**
  + **XGBoost:** Achieves AUC-ROC = 0.85, meeting the target (≥ 0.85). Recall = 0.74 indicates 74% of churners are identified, close to the 70% business KPI. Precision = 0.66 means 66% of flagged users are true churners, balancing campaign costs.
  + **LightGBM:** Slightly lower performance (AUC-ROC = 0.84, recall = 0.71), but still viable.
  + **Baseline:** Random guess performs poorly (AUC-ROC = 0.50, recall = 0.05), confirming the models’ predictive power.

**Business-Specific Metrics:**

* **Churn Identification Rate:** XGBoost identifies 74% of churners (74,000 out of ~100,000 churners in the test set), meeting the 70% KPI.
* **Retention Impact:** Assuming 50% of identified churners are retained via campaigns (e.g., discounts), 37,000 users are saved monthly. At $10 ARPU, this yields $370,000 monthly savings ($4.44M annually).
* **Campaign Efficiency:** With precision = 0.66, 34% of targeted users (25,160) are false positives. Assuming a $5 campaign cost per user, the cost is $125,800/month, far outweighed by savings.

**What-If Scenarios:**

* **Scenario 1: Higher Retention Rate (75%):** If 75% of identified churners are retained (55,500 users), savings increase to $555,000/month ($6.66M/year).
* **Scenario 2: Lower Precision (0.50):** If precision drops to 0.50 (more false positives), campaign costs rise to $185,000/month, but savings remain positive ($370,000 – $185,000 = $185,000/month).
* **Scenario 3: Seasonal Spike (January):** In high-churn months (6.5%), XGBoost identifies 74% of 130,000 churners (96,200), saving $481,000/month at 50% retention.

**P&L Impact:**

* **Revenue Savings:** $370,000/month (base case, 50% retention).
* **Campaign Costs:** $125,800/month (based on precision = 0.66).
* **Net Impact:** $244,200/month ($2.93M/year), exceeding the $1.75M annual target.

**Cost-Benefit Analysis:**

* **Costs:**
  + Model development: $150,000 (from SoW).
  + Campaign costs: $1.51M/year ($125,800 × 12).
  + Infrastructure: ~$10,000/year (AWS EC2, Redshift).
  + Total: $1.67M/year.
* **Benefits:** $4.44M/year (base case).
* **ROI:** ($4.44M – $1.67M) / $1.67M = 166%, justifying investment.

**Alignment with Design Requirements:**

* **Reliability:** Consistent performance (AUC-ROC = 0.85 vs. 0.86 on validation) ensures dependable predictions.
* **Scalability:** Models handle 2M test users in <2 min, scalable to 10M.
* **Maintainability:** MLflow logs and automated pipelines simplify updates.
* **Adaptability:** Robust to seasonal variations (e.g., January performance).

**6.2 Error Analysis**

**Confusion Matrix (XGBoost, Test Set):**

|  |  |  |
| --- | --- | --- |
| **Prediction\Truth** | **Churn (1)** | **No Churn (0)** |
| **Predicted Churn (1)** | 74,000 (TP) | 38,120 (FP) |
| **Predicted No Churn (0)** | 26,000 (FN) | 1,861,880 (TN) |

* **True Positives (TP):** 74,000 churners correctly identified, enabling retention campaigns.
* **False Positives (FP):** 38,120 non-churners flagged, increasing campaign costs ($5 × 38,120 = $190,600/month at lower precision).
* **False Negatives (FN):** 26,000 churners missed, representing $260,000/month in lost revenue (26,000 × $10).
* **True Negatives (TN):** 1.86M non-churners correctly ignored, minimizing unnecessary costs.

**Error Patterns:**

* **False Positives:** Often users with moderate watch\_time (80–100 min/week) but recent inactivity (15–20 days). Likely borderline cases where engagement is declining but not yet critical.
* **False Negatives:** Common among new users (<1 month tenure) with high watch\_time (>200 min/week). These may be trial users who churn unexpectedly, not captured by current features.

**Refinement Recommendations:**

* **Feature Engineering:** Add trial user flag or engagement volatility (e.g., standard deviation of watch\_time) to capture FN cases.
* **Threshold Tuning:** Increase prediction threshold (0.4 to 0.5) to reduce FP, improving precision to 0.70 (trade-off: recall drops to 0.70).
* **Data Enrichment:** Collect trial user metadata or exit survey data to address FN among new users (out of scope unless provided).

**Iteration Decision:**  
The model’s performance (AUC-ROC = 0.85, recall = 0.74) meets targets, and ROI ($2.93M/year) is strong. Errors are manageable, and refinements can be implemented post-deployment via retraining. No immediate iteration is required.

**6.3 Business Validation**

**Stakeholder Presentation (November 10, 2025):**

* **Audience:** Marketing, customer service, data engineering, and executive teams.
* **Content:**
  + **Performance Summary:** XGBoost achieves AUC-ROC = 0.85, recall = 0.74, identifying 74% of churners. Outperforms baseline (random guess, AUC-ROC = 0.50).
  + **Business Impact:** Saves $370,000/month (base case), with net ROI of 166% ($2.93M/year).
  + **What-If Scenarios:** Up to $6.66M/year if retention rate reaches 75%.
  + **Error Insights:** FP (38,120) manageable at $190,600/month cost; FN (26,000) acceptable given ROI.
  + **Fairness:** Demographic parity within 5%, no significant bias by age\_group or location.
  + **SHAP Explanations:** Highlight inactive\_30\_days (30% impact) and weekly\_watch\_time (25%) as key drivers, enabling targeted campaigns (e.g., re-engagement offers for inactive users).
* **Visual Aids:**
  + ROC curve comparing XGBoost, LightGBM, and baseline.
  + SHAP summary plot showing feature impacts.
  + Bar chart of monthly savings vs. campaign costs.

**Stakeholder Feedback:**

* **Marketing:** Excited about 74% churner identification, plans to target inactive users with discounts and personalized recommendations.
* **Executives:** Approved ROI ($2.93M/year) and scalability for 10M users, citing alignment with $1.75M target.
* **Customer Service:** Suggested capturing exit survey data to reduce FN, noted for future iterations.

**Alignment with Business Expectations:**

* **Technical:** AUC-ROC = 0.85 meets the ≥ 0.85 goal.
* **Business:** 74% churner identification exceeds 70% KPI, with $370,000/month savings approaching the $500,000 target.
* **Risk:** Low, as errors are cost-manageable, and fairness metrics ensure ethical use.

**6.4 Go/No-Go Decision**

**Evaluation Insights:**

* **Strengths:**
  + XGBoost meets technical (AUC-ROC = 0.85) and business (74% churner identification) thresholds.
  + Strong ROI (166%, $2.93M/year) justifies investment.
  + Scalable (2 min inference for 2M users), interpretable (SHAP/LIME), and fair (demographic parity).
* **Risks:**
  + False positives increase campaign costs ($190,600/month), but net savings remain positive.
  + False negatives miss 26,000 churners ($260,000/month), addressable via future data enrichment.
* **Mitigation Plans:**
  + Monitor FP/FN rates post-deployment via dashboards.
  + Schedule quarterly retraining to adapt to new patterns (e.g., trial user churn).
  + Explore exit survey data to reduce FN (out of scope unless provided).

**Decision:** **Go for Deployment.**

* The model delivers significant value ($2.93M/year) with manageable risks.
* Stakeholder approval was secured on November 12, 2025, confirming readiness for Phase 7 (Deployment & MLOps).

**Contingency:** If performance degrades post-deployment (e.g., AUC-ROC < 0.80), we will iterate by adding features (e.g., trial user flag) or collecting new data, leveraging MLflow’s experiment tracking.

**Outcome of Phase 6**

* **Formal Approval to Deploy:**
  + XGBoost selected as the primary model, with LightGBM as a backup.
  + Approved by stakeholders on November 12, 2025.
* **Documented Expected ROI:**
  + Base case: $2.93M/year ($370,000/month savings – $125,800/month costs).
  + Potential: Up to $6.66M/year with 75% retention.
* **Risks and Mitigation Plans:**
  + **Risk 1:** High FP costs ($190,600/month). **Mitigation:** Adjust thresholds post-deployment.
  + **Risk 2:** FN miss 26,000 churners. **Mitigation:** Explore exit surveys, retrain quarterly.
  + **Risk 3:** Concept drift (e.g., new user behaviors). **Mitigation:** Monitor AUC-ROC, automate retraining.
* **Deliverable:**
  + **Model Evaluation and Business Case Report** (completed November 15, 2025), including performance metrics, error analysis, cost-benefit analysis, SHAP visualizations, and stakeholder presentation slides. Available in the GitHub repository with Jupyter notebooks for portfolio inclusion.

**Alignment with Design Requirements:**

* **Reliability:** Consistent test performance (AUC-ROC = 0.85) and robust metrics ensure trust.
* **Scalability:** Handles 2M users in <2 min, extensible to 10M.
* **Maintainability:** MLflow and pipelines support easy updates.
* **Adaptability:** Threshold tuning and retraining plans address future changes.

**Next Steps:**  
The project is ready for Phase 7 (Deployment & MLOps), where the XGBoost model will be operationalized as a real-time API. The comprehensive evaluation and business case strengthen this project by showcasing technical rigor and business impact.

**Phase 7: Deployment & MLOps**

**Goal:** Operationalize the XGBoost churn prediction model for StreamHub, Inc. with reliability, scalability, security, and continuous feedback, ensuring seamless integration with business operations.

In this phase, we deploy the XGBoost model (AUC-ROC = 0.85, recall = 0.74) selected in Phase 6 as a real-time API to predict customer churn and enable retention campaigns. We focus on serialization, containerization, integration, CI/CD pipelines, monitoring, and maintenance, aligning with the four ML design requirements: reliability (consistent predictions), scalability (handles 10M users), maintainability (easy updates), and adaptability (adjusts to new patterns). The deployment, conducted between November 16 and November 30, 2025, delivers a production-ready solution with documented processes and a post-deployment review. The outcome supports StreamHub’s goal of reducing churn by 10% ($500,000 monthly savings) and enhances your portfolio with a production-grade MLOps workflow.

**7.1 Preparation for Deployment**

**Model Serialization:**

* **Process:** The XGBoost model was serialized as a pickle file (xgboost\_v1.0.pkl) using Python’s joblib library, including the preprocessing pipeline (imputation, encoding, scaling) from Phase 4.
* **Version Control:** The model and pipeline code were committed to a GitHub repository (StreamHub\_Churn\_Prediction) with a tagged release (v1.0). MLflow Model Registry entry (“XGBoost\_v1.0”) links to the pickle file and metadata (e.g., AUC-ROC = 0.85, training date).

**Deployment Method:**

* **Choice:** Real-time API endpoint using FastAPI, deployed on AWS ECS (Elastic Container Service).
* **Rationale:**
  + **Business Alignment:** StreamHub’s marketing team requires real-time churn probabilities to trigger immediate retention campaigns (e.g., personalized offers for users with >70% churn risk).
  + **Scalability:** API supports high-throughput predictions for 10M users monthly.
  + **Alternative Considered:** Batch processing (e.g., nightly predictions) was rejected, as real-time responses align better with dynamic campaign needs.

**Output:**

* Serialized model (xgboost\_v1.0.pkl) and API specification (input: user features, output: churn probability + SHAP explanations).

**7.2 Packaging & Versioning**

**Containerization:**

* **Tool:** Docker, creating a container image (streamhub-churn: v1.0) with:
  + Python 3.9 environment.
  + Dependencies (xgboost, fastapi, scikit-learn, shap).
  + Serialized model and preprocessing pipeline.
  + FastAPI application to serve predictions.
* **Process:** Dockerfile built and pushed to AWS Elastic Container Registry (ECR) on November 18, 2025.

**Model Registry:**

* **MLflow Integration:** The model was registered as “XGBoost\_v1.0” in MLflow Model Registry, tagged as “Production” with metadata (e.g., AUC-ROC, feature schema).
* **Versioning:** Semantic versioning (e.g., v1.0, v1.1 for future updates) ensures traceability.

**Experiment Tracking in Deployment:**

* **Setup:** MLflow tracking server logs API performance metrics (e.g., latency, prediction volume) and model metrics (e.g., AUC-ROC on new data) post-deployment.
* **Benefit:** Enables continuous monitoring and comparison with future model versions (maintainability, adaptability).

**Output:**

* Docker image (streamhub-churn: v1.0) in AWS ECR.
* MLflow entry with versioned model and tracking setup.

**7.3 Integration and Scalability**

**Integration with Business Systems:**

* **Systems:**
  + **CRM System:** Integrated with StreamHub’s Salesforce instance to push churn predictions (e.g., user\_id, churn\_probability) for campaign targeting.
  + **Marketing Platform:** Connected to StreamHub’s email/SMS platform (e.g., Braze) to trigger retention offers based on churn risk (e.g., >70% probability).
* **Process:**
  + API endpoint (/predict) accepts JSON inputs (e.g., user\_id, weekly\_watch\_time, inactive\_30\_days) and returns JSON outputs (e.g., churn\_probability, shap\_values).
  + Data engineering team provided API keys and OAuth2 authentication for secure access.
  + Integration tested with sample user data on November 20, 2025, ensuring predictions flow to CRM and marketing platforms.

**Scalability Considerations:**

* **Workload:** 10M users/month, ~330K predictions/day, ~230/min during peak hours.
* **Infrastructure:**
  + AWS ECS with auto-scaling (2–10 containers, each handling ~100 predictions/sec).
  + Load balancer (AWS ALB) distributes traffic.
  + Feature store (AWS Feature Store) provides real-time feature access, reducing latency.
* **Performance:** Inference time = 1.8 min for 1M users (Phase 5), scalable to 10M in ~18 min with parallel containers.
* **Cost:** ~$500/month for ECS, ALB, and Feature Store, within budget ($10,000/year, Phase 6).

**Output:**

* Integrated API seamlessly connected to Salesforce and Braze, handling 10M predictions/month (reliability, scalability).

**7.4 CI/CD Integration and Testing**

**CI/CD Pipeline:**

* **Tools:** GitHub Actions for CI/CD, integrated with AWS ECS and ECR.
* **Pipeline Stages:**
  + **Continuous Integration (CI):**
    - Trigger: Code push to main branch.
    - Steps: Run unit tests (e.g., validate preprocessing, API response format), lint code, build Docker image.
  + **Continuous Deployment (CD):**
    - Trigger: Successful CI and tagged release (e.g., v1.1).
    - Steps: Push Docker image to ECR, update ECS service with new image, run integration tests.
* **Testing:**
  + **Unit Tests:** Validate model predictions on synthetic data (e.g., inactive\_30\_days = 1 → high churn probability).
  + **Integration Tests:** Confirm API outputs reach Salesforce and Braze correctly.
  + **Load Tests:** Simulated 500 predictions/sec using Locust, ensuring <200ms latency at peak.

**Pre-Launch Testing:**

* Conducted November 22–24, 2025, in a staging environment (AWS ECS).
* **Results:**
  + 100% test coverage for preprocessing and API logic.
  + Zero errors in integration with CRM/marketing platforms.
  + Latency: 150ms average, supporting real-time use.
* **Sign-Off:** Data engineering team approved on November 25, 2025.

**Output:**

* Automated CI/CD pipeline (GitHub Actions), ensuring reliable and maintainable deployments.
* Fully tested API ready for production.

**7.5 Monitoring, Maintenance, and Retraining**

**Monitoring System:**

* **Tools:**
  + AWS CloudWatch for API metrics (latency, error rates, request volume).
  + MLflow for model performance (AUC-ROC, recall on new data).
  + Custom dashboard (Streamlit) displaying real-time metrics (e.g., churn predictions, campaign triggers).
* **Metrics Tracked:**
  + **Model Performance:** AUC-ROC, recall, precision on daily labeled data (churn status from subscription\_details).
  + **Business Impact:** Number of churners identified, campaign costs, retention rate.
  + **Concept Drift:** Distribution shifts in weekly\_watch\_time, inactive\_30\_days (using Kolmogorov-Smirnov test).
* **Alerts:**
  + Triggered if AUC-ROC < 0.80 or recall < 0.70 (email/Slack to data team).
  + Drift alerts if feature distributions shift significantly (p < 0.01).

**Maintenance and Retraining:**

* **Schedule:** Quarterly retraining (e.g., March 2026) using new data (January–February 2026) to adapt to changing patterns (e.g., new user behaviors).
* **Process:**
  + Automated pipeline (Airflow) pulls new data, retrains XGBoost, and logs results in MLflow.
  + New model version (e.g., XGBoost\_v1.1) registered and tested in staging before promotion to production.
* **Rollback Mechanism:**
  + ECS stores previous model version (XGBoost\_v1.0) in MLflow Registry.
  + If performance degrades (e.g., AUC-ROC < 0.80), rollback to v1.0 via CD pipeline within 1 hour.

**Output:**

* Monitoring dashboard and alerts for performance and drift (reliability, adaptability).
* Automated retraining and rollback mechanisms (maintainability).

**7.6 Documentation and Knowledge Transfer**

**Documentation:**

* **Deployment Guide:** Details API setup, Docker configuration, CI/CD pipeline, and monitoring (PDF, GitHub repository).
* **User Manual:** Instructions for marketing team to query API (e.g., POST requests with user features) and interpret outputs (churn probability, SHAP explanations).
* **Technical Report:** Summarizes model performance, integration, and maintenance processes.

**Knowledge Transfer:**

* **Training Session:** Conducted on November 28, 2025, for StreamHub’s data engineering and marketing teams (virtual, recorded).
  + Topics: API usage, monitoring dashboard, retraining process, rollback procedures.
  + Materials: Slides and Jupyter notebooks in repository.
* **Handover:** Primary contact (StreamHub’s data manager) assigned to manage API and coordinate retraining.

**Output:**

* Comprehensive documentation and trained operations team, ensuring maintainability and smooth handoff.

**7.7 Post-Deployment Review**

**Review Meeting (November 30, 2025):**

* **Attendees:** Data Science “WithMe” LLC, StreamHub’s marketing, data engineering, and executive teams.
* **Agenda:**
  + **Performance:** API deployed on November 26, 2025, handling 50K predictions/day with 150ms latency. Initial AUC-ROC = 0.85 on first week’s data.
  + **Business Impact:** Marketing launched campaigns targeting 10,000 high-risk users (churn probability >70%), with early retention rates (~40%) to be tracked.
  + **Lessons Learned:**
    - Real-time API was critical for campaign agility, validating deployment choice.
    - Feature store reduced latency, but initial setup required extra coordination with data engineers.
    - Stakeholder training improved adoption but needs follow-up for non-technical teams.
  + **Ongoing Alignment:** Quarterly reviews scheduled to track ROI ($2.93M/year target) and refine campaigns.

**Action Items:**

* Monitor retention rates and campaign costs for 30 days to validate ROI.
* Plan follow-up training for marketing team in January 2026.
* Explore exit survey data (per Phase 6 feedback) for future iterations.

**Output:**

* Post-deployment review report capturing performance, lessons learned, and action items, ensuring ongoing business alignment (reliability, adaptability).

**Outcome of Phase 7**

* **Deployed Model:**
  + Real-time FastAPI endpoint (/predict) deployed on AWS ECS, serving churn probabilities and SHAP explanations for 10M users/month.
  + Integrated with Salesforce and Braze for automated retention campaigns.
* **Accessibility and Scalability:**
  + Handles 330K predictions/day with 150ms latency, auto-scaling to peak loads.
  + Feature store ensures low-latency feature access.
* **Maintainability and Reliability:**
  + CI/CD pipeline (GitHub Actions) automates updates.
  + MLflow tracks performance, with rollback to XGBoost\_v1.0 if needed.
  + Monitoring dashboard and alerts ensure consistent AUC-ROC ≥ 0.80.
* **Adaptability:**
  + Quarterly retraining and drift detection address changing user behaviors.
  + Versioning (v1.0, v1.1) supports iterative improvements.
* **Deliverables:**
  + **Deployed API:** Live on November 26, 2025, with documented endpoints.
  + **Monitoring Dashboard:** Streamlit-based, showing predictions, metrics, and alerts.
  + **Documentation:** Deployment guide, user manual, technical report in GitHub.
  + **Training Materials:** Slides and notebooks for operations team.
  + **Post-Deployment Review Report:** Summarizing performance and lessons learned.

**Alignment with Business Goals:**

* Enables real-time churn predictions, supporting $370,000/month savings (Phase 6).
* Scalable, reliable, and maintainable, meeting StreamHub’s operational needs.
* Adaptable to future data or user patterns, ensuring long-term value.

The project is complete, with the model live and monitored. StreamHub will track retention rates and ROI over 30 days, with quarterly retraining planned.

**Final Analysis: Strategic Impact of Data Science on Business Decision-Making**

The six-month data science case study on customer churn for **StreamHub, Inc.**, a streaming platform with 10 million subscribers, demonstrates the transformative power of data science in driving strategic business decision-making. By leveraging the CRISP-DM methodology, we developed a predictive churn model (XGBoost, AUC-ROC = 0.85, recall = 0.74) that identifies at-risk customers, enabling targeted retention strategies. This analysis evaluates the strategic impact of this project on StreamHub’s decision-making and articulates why data-driven decisions are essential for business success, using the churn study as a concrete example.

**How Data Science Drives Strategic Decision-Making**

The churn prediction project illustrates how data science reshapes strategic decision-making by providing actionable insights, optimizing resource allocation, and aligning technical solutions with business objectives. Below, we outline the key strategic impacts:

* **Enhanced Customer Retention through Predictive Insights**
  + **Impact:** The XGBoost model identifies 74% of potential churners (74,000 out of 100,000 monthly churners), enabling StreamHub to target high-risk users with personalized retention campaigns (e.g., discounts, tailored content recommendations). This shifts decision-making from reactive (e.g., post-churn analysis) to proactive, addressing churn before it occurs.
  + **Business Value:** The model delivers $370,000 in monthly revenue savings (assuming 50% retention of identified churners at $10 ARPU), contributing to an annual net ROI of $2.93 million (166% return on $1.67M costs). This empowers executives to prioritize retention over costly acquisition (e.g., $50–$100 per new user), directly impacting profitability.
  + **Strategic Decision-Making:** Marketing teams now base campaign strategies on predictive probabilities (e.g., targeting users with >70% churn risk), optimizing budget allocation and improving customer lifetime value.
* **Data-Driven Resource Optimization**
  + **Impact:** The model’s precision (0.66) ensures 66% of targeted users are true churners, minimizing wasteful spending on false positives (campaign cost: $125,800/month vs. $370,000 savings). Error analysis (Phase 6) identified false positives among moderately engaged users, guiding refinements like threshold tuning to further optimize costs.
  + **Business Value:** By focusing resources on high-impact users, StreamHub avoids blanket retention efforts (e.g., universal discounts), which are less effective and more expensive. The cost-benefit analysis (166% ROI) justifies data science investments, influencing budget decisions.
  + **Strategic Decision-Making:** Executives can allocate marketing budgets based on quantifiable ROI, while operations teams streamline campaign execution using automated API integrations (Phase 7), enhancing efficiency.
* **Informed Strategic Planning with Actionable Insights**
  + **Impact:** Exploratory Data Analysis (Phase 3) uncovered key churn drivers: low engagement (weekly\_watch\_time <60 min/week, 7% churn rate), inactivity (inactive\_30\_days, 10% churn rate), and plan downgrades (9% churn rate). These insights informed hypotheses and feature engineering, grounding the model in business realities.
  + **Business Value:** Insights like seasonal churn spikes (January, 6.5%) and higher churn among younger users (18–24, 7%) enable StreamHub to tailor strategies, such as seasonal promotions or youth-targeted content. The model’s SHAP explanations (Phase 5) clarify why users churn (e.g., 30% impact from inactivity), building stakeholder trust.
  + **Strategic Decision-Making:** Product teams can use these insights to enhance platform features (e.g., re-engagement notifications), while executives align long-term strategies with data-driven trends, such as addressing competitive pressures in North America (6% churn).
* **Competitive Advantage through Real-Time Decision-Making**
  + **Impact:** The real-time API (Phase 7, 150ms latency) integrates with StreamHub’s CRM and marketing platforms (Salesforce, Braze), enabling immediate campaign triggers for at-risk users. This agility is critical in the competitive streaming industry, where rivals like Netflix and Spotify vie for subscriber loyalty.
  + **Business Value:** Real-time predictions support dynamic retention (e.g., offering discounts to inactive users), reducing churn by up to 10% (from 5% to 4.5%), as targeted. This strengthens StreamHub’s market position, with potential savings of $6.66M/year at 75% retention (Phase 6 what-if scenario).
  + **Strategic Decision-Making:** Leadership can respond swiftly to market shifts, using the model’s adaptability (quarterly retraining, drift detection) to maintain performance amid changing user behaviors, ensuring a competitive edge.
* **Ethical and Transparent Decision-Making**
  + **Impact:** Fairness metrics (Phase 5, demographic parity within 5%) and SHAP/LIME explanations ensure predictions are unbiased across age and region, while GDPR-compliant data handling (Phase 2) protects user privacy. This builds trust with stakeholders and customers.
  + **Business Value:** Ethical models mitigate reputational risks, crucial in a consumer-facing industry. Transparent explanations (e.g., “User X flagged due to 40-day inactivity”) enable marketing to justify campaign decisions, aligning with corporate governance.
  + **Strategic Decision-Making:** Executives can confidently deploy data-driven solutions, knowing they meet ethical standards, enhancing brand reputation and customer trust.

**Why Data-Driven Decisions Are Necessary for Business Success**

The StreamHub churn study underscores why data-driven decisions are critical for business success, particularly in competitive, data-rich industries like streaming. Below, we articulate the necessity of data-driven approaches, grounded in the project’s outcomes:

* **Precision in a Competitive Landscape**
  + **Challenge:** Streaming platforms face intense competition, with subscribers easily switching to alternatives (e.g., Netflix, Spotify). StreamHub’s 5% monthly churn rate ($5M revenue loss) demands precise interventions to retain users.
  + **Data-Driven Solution:** The churn model identifies high-risk users with 74% recall, enabling targeted campaigns that save $370,000/month. Without data-driven insights, StreamHub would rely on intuition or broad strategies, wasting resources on low-risk users.
  + **Necessity:** Data-driven decisions provide precision, ensuring resources are allocated to the most impactful actions, maintaining competitiveness in a crowded market.
* **Maximizing ROI in Resource-Constrained Environments**
  + **Challenge:** Budget constraints require StreamHub to optimize marketing spend (e.g., $125,800/month campaign costs). Non-data-driven approaches risk overspending on ineffective campaigns.
  + **Data-Driven Solution:** The model’s cost-benefit analysis (166% ROI, $2.93M/year net savings) quantifies value, guiding budget decisions. Precision (0.66) minimizes false positives, keeping costs manageable.
  + **Necessity:** Data-driven decisions quantify ROI, enabling businesses to justify investments and optimize limited resources, critical for financial sustainability.
* **Adapting to Dynamic Customer Behaviors**
  + **Challenge:** User behaviors evolve (e.g., seasonal churn spikes, declining engagement), and static strategies fail to adapt. StreamHub’s high churn among younger users (7%) and competitive markets (6%) requires agility.
  + **Data-Driven Solution:** The model’s temporal features (month\_of\_year, delta\_watch\_time) and quarterly retraining (Phase 7) adapt to patterns like January churn spikes. Monitoring for concept drift ensures performance (AUC-ROC ≥ 0.80).
  + **Necessity:** Data-driven decisions enable businesses to anticipate and respond to changing customer needs, maintaining relevance and retention in dynamic markets.
* **Reducing Uncertainty and Risk**
  + **Challenge:** Intuition-based decisions carry high uncertainty, risking revenue loss (e.g., missing 26,000 churners, $260,000/month). Ethical risks (e.g., biased targeting) could harm StreamHub’s reputation.
  + **Data-Driven Solution:** The model reduces uncertainty with 74% churner identification and mitigates risks via fairness metrics (5% demographic parity) and GDPR compliance. Error analysis (Phase 6) identifies false negatives for future improvement.
  + **Necessity:** Data-driven decisions minimize uncertainty by grounding strategies in evidence, reducing financial and reputational risks essential for long-term success.
* **Building Stakeholder Trust and Alignment**
  + **Challenge:** Stakeholders (marketing, executives) require transparency to adopt data-driven strategies. Without trust, models risk underutilization.
  + **Data-Driven Solution:** SHAP explanations (e.g., 30% impact from inactive\_30\_days) and stakeholder presentations (Phase 6) clarified predictions, securing approval for deployment. The Streamlit dashboard (Phase 7) provides real-time insights, fostering ongoing trust.
  + **Necessity:** Data-driven decisions build trust through transparency, aligning cross-functional teams and ensuring strategic initiatives are executed effectively.
* **Enabling Scalable, Long-Term Growth**
  + **Challenge:** StreamHub’s 10M-user scale demands solutions that grow without proportional cost increases. Manual churn analysis is unscalable.
  + **Data-Driven Solution:** The real-time API (150ms latency, 10M predictions/month) and automated pipelines (CI/CD, Airflow) scale efficiently, with monitoring ensuring reliability (AUC-ROC ≥ 0.80). The feature store reduces latency, supporting growth.
  + **Necessity:** Data-driven decisions enable scalable solutions, allowing businesses to manage growth cost-effectively, critical for market expansion and profitability.

**Conclusion**

The StreamHub churn prediction project exemplifies the strategic impact of data science on business decision-making. By identifying 74% of churners, the model drives $2.93M in annual net savings, optimizes marketing spend, and provides actionable insights (e.g., targeting inactive users). It empowers StreamHub to make precise, proactive, and ethical decisions, strengthening its competitive position in the streaming industry.

Data-driven decisions are necessary for business success because they provide precision, maximize ROI, adapt to change, reduce risk, build trust, and enable scalability. Without data science, StreamHub would rely on costly, inefficient strategies, risking revenue loss and market share. This case study underscores that in data-rich, competitive environments, businesses that harness data-driven insights—like StreamHub’s churn model—achieve sustainable growth and resilience, making data science a cornerstone of modern strategic success.