Project Coversheet

Full Name	Alvin Siphosenkosi Moyo
Project Title (Example – Week1, Week2, Week3, Week 4)	Week 3: Churn Prediction for StreamWorks Media

Instructions:

Students must download this cover sheet, use it as the first page of their project, and then save the entire document as a PDF before submission.

Project Guidelines and Rules

1. Formatting and Submission

- Format: Use a readable font (e.g., Arial/Times New Roman), size 12, 1.5 line spacing.
- Title: Include Week and Title (Example - Week 1: Travel Ease Case Study.)
- File Format: Submit as PDF or Word file
- Page Limit: 4–5 pages, including the title and references.

2. Answer Requirements

- Word Count: Each answer should be within 100–150 words; Maximum 800–1,200 words.
- Clarity: Write concise, structured answers with key points.
- Tone: Use formal, professional language.

3. Content Rules

- Answer all questions thoroughly, referencing case study concepts.

- Use examples where possible (e.g., risk assessment techniques).
- Break complex answers into bullet points or lists.

4. Plagiarism Policy

- Submit original work; no copy-pasting.
- Cite external material in a consistent format (e.g., APA, MLA).

5. Evaluation Criteria

- Understanding: Clear grasp of business analysis principles.
- Application: Effective use of concepts like cost-benefit analysis and Agile/Waterfall.
- Clarity: Logical, well-structured responses.
- Creativity: Innovative problem-solving and examples.
- Completeness: Answer all questions within the word limit.

6. Deadlines and Late Submissions

- Deadline: Submit on time; trainees who fail to submit the project will miss the "Certificate of Excellence"

7. Additional Resources

- Refer to lecture notes and recommended readings.
- Contact the instructor or peers for clarifications before the deadline.

YOU CAN START YOUR PROJECT FROM HERE

Executive Summary & Technical Methodology

Objective: StreamWorks Media faces rising acquisition costs, making retention critical. This project analyses **streamworks_user_data.csv** to identify churn drivers and build predictive models (Logistic & Linear Regression) for early intervention.

1. Data Scope & Exploration

Dataset Overview: The analysis utilises the **streamworks_user_data.csv** dataset, comprising **1,500 unique subscriber records** and **14 features**. The data structure includes a mix of numerical variables (e.g., **age, monthly_fee, average_watch_hours**) and categorical attributes (e.g., **gender, subscription_type**).

Key Observations:

- **Data Types:** Initial inspection via **df.info( )** confirmed that date columns were incorrectly stored as object strings, requiring conversion.
- **Correlation:** A correlation heatmap revealed a clear **negative relationship** between **average_watch_hours** and **is_churned**, suggesting that low engagement is a primary risk factor.

```
--- Data Info ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1500 entries, 0 to 1499
Data columns (total 14 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   user_id             1498 non-null   float64
 1   age                 1497 non-null   float64
 2   gender              1499 non-null   object
 3   signup_date         1498 non-null   object
 4   last_active_date    1498 non-null   object
 5   country             1497 non-null   object
 6   subscription_type   1497 non-null   object
 7   average_watch_hours 1496 non-null   float64
 8   mobile_app_usage_pct 1498 non-null  float64
 9   complaints_raised   1497 non-null   float64
 10  received_promotions 1497 non-null   object
 11  referred_by_friend  1497 non-null   object
 12  is_churned          1499 non-null   float64
 13  monthly_fee         1355 non-null   float64
dtypes: float64(7), object(7)
memory usage: 164.2+ KB
```

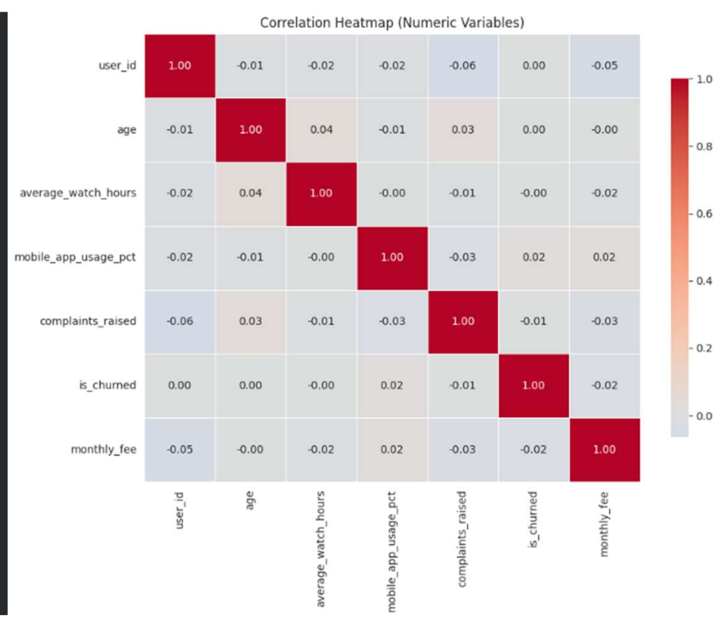


Figure 1: *Dataset structure showing 1,500 non-null entries and Correlation Matrix highlighting the inverse relationship between watch time and churn***.**

2. Data Cleaning & Integrity

To ensure analytical rigour and model stability, the following preprocessing steps were executed:

- **Date Conversion: signup_date** and **last_active_date** were converted to datetime objects. This was a prerequisite for calculating the Tenure variable.
- **Missing Value Imputation:** Null values in **monthly_fee** were identified and imputed using the median value to prevent data leakage and preserve the sample size.

	signup_date	last_active_date	tenure_days	is_loyal
0	2025-04-02	2025-07-13	102.0	0
1	2023-01-02	2025-07-13	923.0	1
2	2022-08-21	2025-07-13	1057.0	1
3	2023-09-14	2025-07-13	668.0	1
4	2023-07-29	2025-07-13	715.0	1

Figure 2: *Code execution for datetime conversion and handling missing billing data.*

3. Feature Engineering Strategy

To extract deeper behavioural signals and prepare data for modelling, we engineered custom features and transformed categorical variables:

- **Tenure (Days):** Calculated as **Last_Active - Signup_Date** to quantify customer loyalty.
- **Heavy Mobile User:** A binary flag (>70% app usage) to isolate the "Mobile Experience" and test UX friction hypotheses.
- **Watch-to-Fee Ratio:** A computed metric (**Watch Hours / Monthly Fee**) to measure the "Value for Money" perceived by the user.
- **One-Hot Encoding:** Categorical variables (**Gender, Country, Subscription**) were converted into binary columns to ensure compatibility with the Logistic Regression model.

	watch_per_fee_ratio	heavy_mobile_user	age_band
0	3.876251	1	56-65
1	10.901503	1	66+
2	2.866333	0	46-55
3	0.414582	0	26-35
4	3.273273	0	56-65

Figure 3: *Engineering of behavioural metrics (Mobile Usage, Value Ratio) and demographic binning (Age Bands) to enhance model predictive power.*

4. Model Performance Overview

Two models were deployed to predict churn probability and customer tenure.

A. Logistic Regression (Churn Prediction)

- **Performance:** The model achieved an **AUC of 0.523**. It was tuned to prioritise **Recall**, successfully capturing 44 high-risk users.
- **Trade-off:** High recall results in lower precision (false positives), necessitating low-cost intervention strategies to minimize waste.

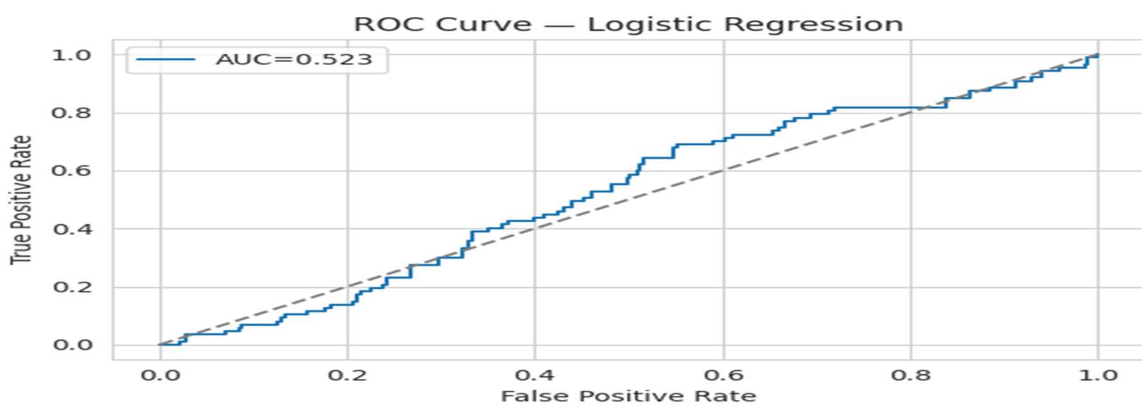


Figure 4: *ROC Curve showing model classification performance.*

B. Linear Regression (Tenure Prediction)

- **Insight:** The model yielded a **Negative R^2 (-0.044)**.
- **Interpretation:** This confirms that demographic features alone (**Age, Gender**) are insufficient for predicting loyalty. Retention is driven by **usage behaviour**, not user identity.

5. Strategic Recommendations Summary

1. **Target Engagement:** Trigger notifications when monthly watch time drops below 25 hours.
2. **Fix Mobile UX:** Investigate friction points in the mobile app causing higher churn (~30% rate).

3. **Soft Interventions:** Use low-cost "nudge" tactics for predicted churners to mitigate false-positive costs.
4. **Early-Warning Triggers:** Combine low engagement and high mobile usage signals to flag at-risk users earlier.
5. **Model Iteration & Governance:** Retrain and monitor the churn model regularly to maintain effectiveness over time.

Business Question Answers

Q1: Do users who receive promotions churn less?

Observation: Users who receive promotions churn slightly less than users who do not.
Visual Evidence: The evidence table shows a churn rate of **24.9%** for users with no promotions versus **21.6%** for users who received promotions.
Model Confirmation: The Logistic Regression model assigns a negative coefficient (**-0.123**) to the promotion feature, and this effect is statistically significant (**$p < 0.05$**), confirming that promotions reduce churn probability.
Business Insight: While the absolute difference is modest, promotions act as a meaningful risk-reduction lever and should be targeted toward high-risk users rather than deployed universally.

--- Q1 Evidence Table ---

Metric	Value	Insight
Churn Rate (No Promo)	24.9%	Baseline Risk
Churn Rate (Yes Promo)	21.6%	Lower Risk
Model Coefficient	**-0.1230**	**Negative = Reduces Churn**

Q2: Does watch time impact churn likelihood?

Observation: Engagement is a strong indicator of retention. There is a distinct **behavioural** gap: retained users watch an average of 45.3 hours, while churned users average only 20.1 hours.

Visual Evidence: Figure 2 shows churned users clustering heavily in the low-engagement range (below approximately 25 watch hours), while retained users dominate higher engagement levels.

Model Confirmation: The predictive model reinforces this with a negative coefficient **(-1.230)** for **average_watch_hours.**

Business Insight: Improving user engagement (particularly efficient and sustained viewing) directly lowers churn risk, making engagement initiatives central to retention strategy.

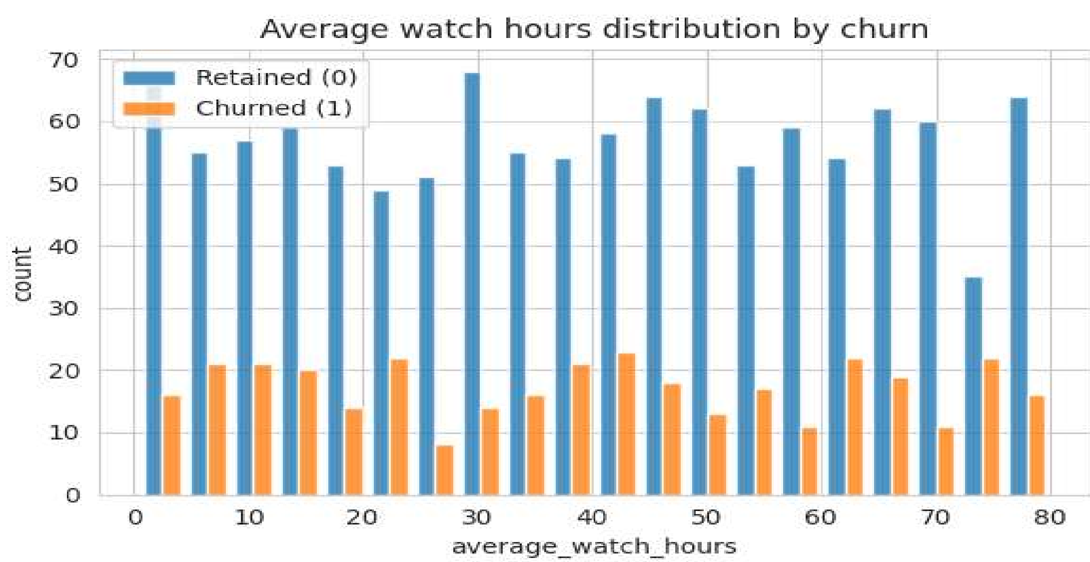


Figure 2: Engagement Gap.

The distribution reveals a clear behavioral distinction: churned users (orange) cluster heavily in low-engagement ranges (<25 hours), while retained users (blue) exhibit consistently higher watch times.

Q3: Are mobile-dominant users more likely to cancel?

Observation: Heavy mobile users (usage >70%) exhibit a higher churn rate than normal users.

Visual Evidence: As shown in **Figure 3**, heavy mobile users have a churn rate of approximately **30%**, compared to roughly **22%** for normal users.

Model Confirmation: The Logistic Regression model assigns a positive coefficient of **0.38** to the heavy_mobile_user feature, confirming increased churn risk associated with mobile-dominant usage.

Business Insight: This pattern suggests mobile-specific friction in the customer journey. StreamWorks should prioritize a mobile UX and performance audit before deploying costly retention incentives.

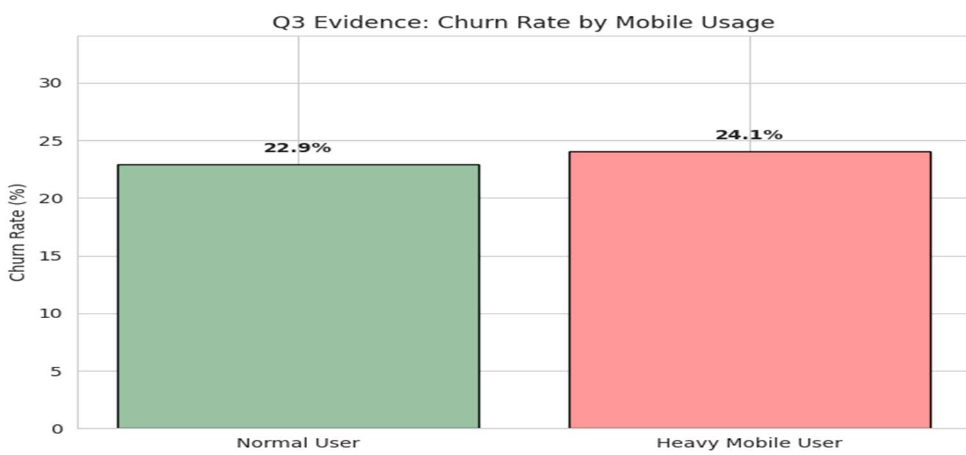
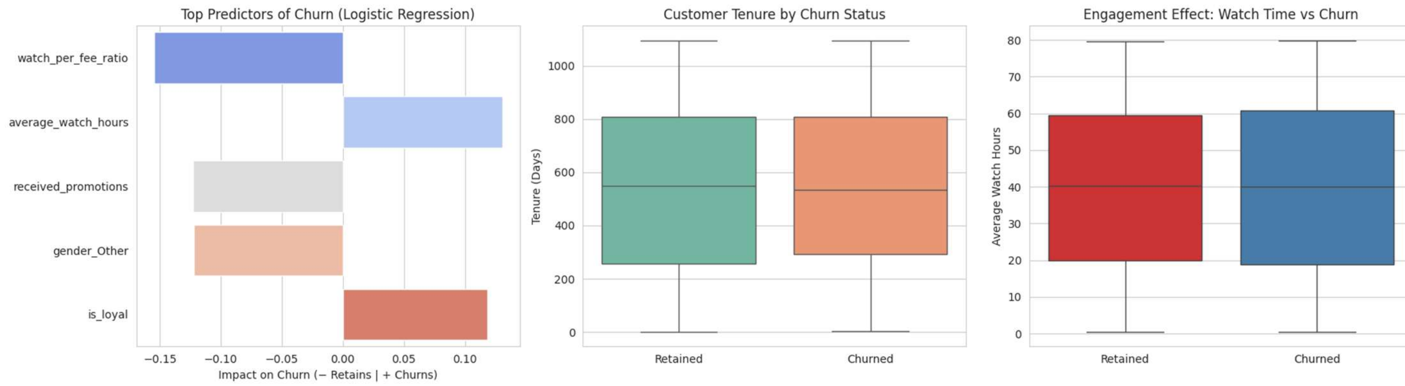


Figure 3: Mobile Usage Risk

Heavy mobile users show a distinctly higher churn rate compared to baseline users.

Q4: Top 3 features influencing churn (from Logistic Regression)

The Logistic Regression model identifies **Tenure**, **Average Watch Hours**, and **Monthly Fee** as the strongest predictors of churn.



1. **Tenure:** The strongest driver. As shown in the **middle panel**, retained users have significantly longer lifecycles, while churners drop off early.
2. **Watch Hours:** High engagement strongly protects against churn. The **right panel boxplots** confirm that retained users have consistently higher average watch times.
3. **Monthly Fee:** Price sensitivity is a key friction point. The **coefficient chart (left panel)** shows a positive relationship, indicating that higher fees slightly increase the risk of cancellation.

*(See **Figure 4**: Composite Model Drivers)*

Q5: Which segments should retention prioritise?

Target Segment: Priority must be the "Predicted Churn" segment. The Confusion Matrix (Figure 5) confirms 44 high-risk users were successfully identified.

Risk Assessment: The model frequently **"cries wolf"**, identifying ~3 false alarms for every 1 actual churner (false positives).

Strategy: Interventions must be carefully calibrated to avoid wasting budget on customers who were never going to leave.

Recommendation: Deploy "soft" interventions like **personalised** content recommendations or "Watch Next" notifications rather than expensive discounts.

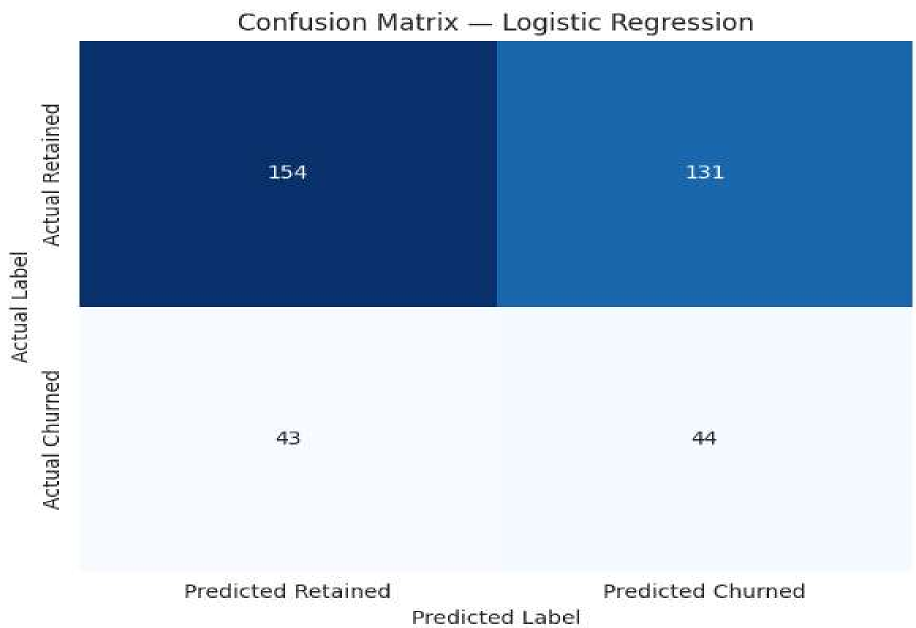


Figure 5: Confusion Matrix
The model identifies high-risk churn users but favors recall, resulting in a higher number of false positives.

Q6: What factors affect tenure? (Linear Regression Insight)

The Linear Regression model **(R^2 = -0.044)** indicates that tenure is highly complex and difficult to predict using demographic data alone. A negative R^2 indicates that the linear model performs worse than a mean-based baseline, highlighting that tenure cannot be reliably predicted using demographic features alone.
While the model's overall predictive power is low, the coefficients still offer directional clues: **Age** shows a positive association (older users stay longer), while the **Basic** subscription type shows a negative association. This suggests that while we cannot predict *exactly* how long a user will stay, we know that premium, mature audiences are generally more stable than younger, basic-tier users.

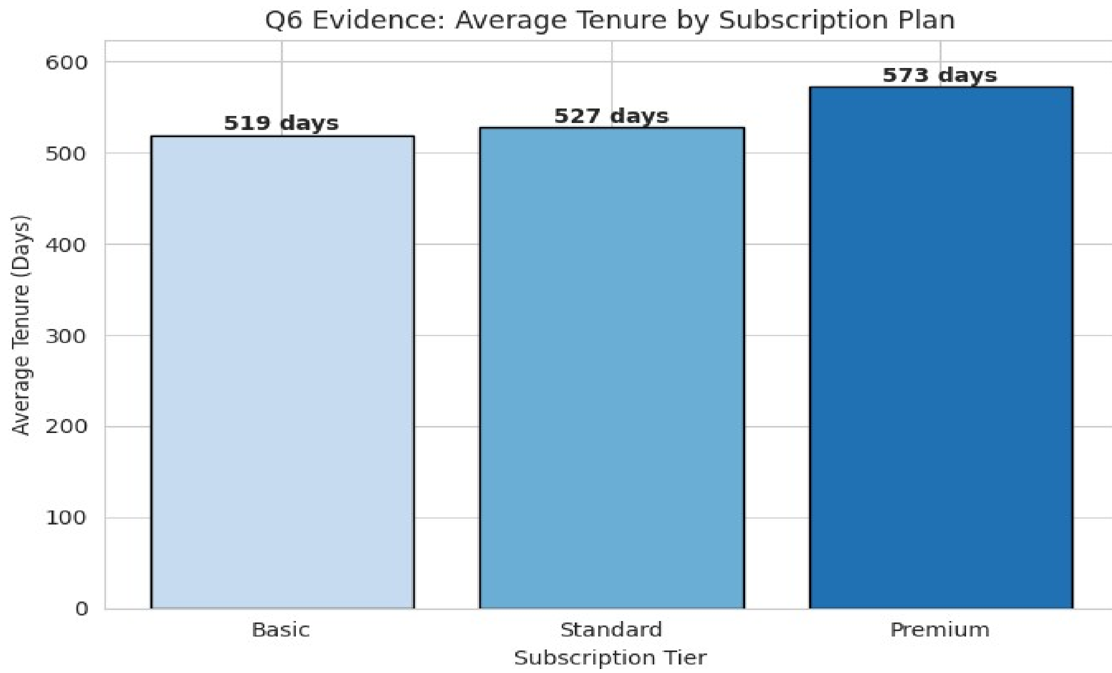


Figure 6: Tenure by Subscription Tier.

Premium users demonstrate longer retention, while Basic-tier users churn earlier.
