Project Coversheet

Full Name	Maria Schiza
Email	mariaschiza005@gmail.com
Contact Number	07899940691
Date of Submission	27/07/2025
Project Week	Week 3

Project Guidelines and Rules

1. Submission Format

- **Document Style**:
  - Use a clean, readable font such as *Arial* or *Times New Roman*, size 12.
  - Set line spacing to **1.5** for readability.
- **File Naming**:
  - Use the following naming format:
    Week X – [Project Title] – [Your Full Name Used During Registration]
    Example: Week 1 – Customer Sign-Up Behaviour – Mark Robb
- **File Types**:
  - Submit your report as a **PDF**.
  - If your project includes code or analysis, attach the **.ipynb notebook** as well.

2. Writing Requirements

- Use formal, professional language.
- Structure your content using headings, bullet points, or numbered lists.

3. Content Expectations

- Answer **all** parts of each question or task.

- Reference tools, frameworks, or ideas covered in the programme and case studies.
- Support your points with practical or real-world examples where relevant.
- Go beyond surface-level responses. Analyse problems, evaluate solutions, and demonstrate depth of understanding.

4. Academic Integrity & Referencing

- All submissions must be your own. Plagiarism is strictly prohibited.
- If you refer to any external materials (e.g., articles, studies, books), cite them using a consistent referencing style such as APA or MLA.
- Include a references section at the end where necessary.

5. Evaluation Criteria

Your work will be evaluated on the following:

- Clarity: Are your answers well-organised and easy to understand?
- Completeness: Have you answered all parts of the task?
- Creativity: Have you demonstrated original thinking and thoughtful examples?
- Application: Have you effectively used programme concepts and tools?
- Professionalism: Is your presentation, language, and formatting appropriate?

6. Deadlines and Extensions

- Submit your work by the stated deadline.
- If you are unable to meet a deadline due to genuine circumstances (e.g., illness or emergency), request an extension **before the deadline** by emailing: **support@uptrail.co.uk**
  Include your full name, week number, and reason for extension.

7. Technical Support

- If you face technical issues with submission or file access, contact our support team promptly at **support@uptrail.co.uk.**

8. Completion and Certification

- Certificate of Completion will be awarded to participants who submit at least two projects.
- Certificate of Excellence will be awarded to those who:
  - Submit all four weekly projects, and
  - Meet the required standard and quality in each.
- If any project does not meet expectations, you may be asked to revise and resubmit it before receiving your certificate.

1. Introduction

This report presents a data quality review and churn analysis of the streamworks_user_data.csv dataset provided by StreamWorks Media, a UK-based video streaming platform. As a member of their Data Strategy Team, I conducted this analysis using Python to support the company's customer retention efforts.

StreamWorks seeks to understand the patterns of driving subscription cancellations (churn) and to develop strategies to retain valuable users. This report supports that goal through a combination of data cleaning, statistical analysis, and predictive modelling.

The dataset used contains 1,500 user records and the following fields:

- user_id: Unique identifier for each user
- age, gender: Demographic information
- signup_date, last_active_date: Subscription lifecycle markers
- country, subscription_type: Geographic and service-level data
- monthly_fee, average_watch_hours: Engagement and payment behavior
- mobile_app_usage_pct: % of watch time via mobile
- complaints_raised: Number of complaints submitted
- received_promotions, referred_by_friend: Marketing and referral data
- is_churned: Binary label indicating whether the user cancelled their subscription in the past 30 days

The objectives of this project are to:

- Identify which users are churning and why, using descriptive and inferential statistics
- Develop a logistic regression model to predict future churners
- Provide actionable insights to help the Retention Team design targeted intervention strategies

The findings from this analysis aim to inform strategic decisions that enhance user satisfaction and reduce churn rates across the platform.

2. Data Cleaning Summary

The dataset provided includes important user information and it is therefore necessary to ensure data accuracy and reliability using the following cleaning steps before conducting meaningful analysis.

Binary Encoding

Converted received_promotions and referred_by_friend columns with entries"Yes"/"No" to 1/0.

Missing Data Handling

- Dropped rows with missing information in critical columns (e.g., user_id, signup_date, last_active_date).
- Filled missing values in categorical columns with their corresponding mode value.
- Applied mean imputation in numerical columns with missing values.

Feature Encoding

Applied one-hot encoding using pd.get_dummies() with drop_first=True, which converts categorical to dummy variables, a useful step before applying a regression model.

Date Formatting

Converted signup_date and last_active_date columns using pd.to_datetime() to the same datetime format.

3. Feature Engineering Summary

The following features were created to enhance analysis:

- **tenure_days:** Calculated as the number of days between signup_date and last_active_date.
- **is_loyal:** Binary feature showing whether a user’s tenure_days exceeds 180 days.
- **Dummy Variables:** Created for categorical features such as gender, and subscription_type.to convert string entries to numerical.
- **Scaling:** Standardised numerical features using StandardScaler to improve logistic regression model’s performance.

4. Key Findings

Statistical Testing

- Chi-Square Tests:
  - No significant relationship between churn and gender or referred_by_friend as their p-values are 0.105 and 0.458 respectively which are higher than the significant level 0.05.
  - Significant association between churn and received_promotions from the very small p-value of 0.002.
- t-Test:
  - No significant difference in average_watch_hours between churned and retained users, as p = 0.8.

Correlation Insights

- complaints_raised: +0.20, showing a moderate positive correlation with churn.
- received_promotions: –0.30, showing a moderate negative correlation; promotions reduce churn.
- tenure_days: –0.24 and is_loyal: –0.35, indicating that longer-tenured users churn less.
- referred_by_friend: –0.10, showing a slight reduction in churn.

Behavioural Trends

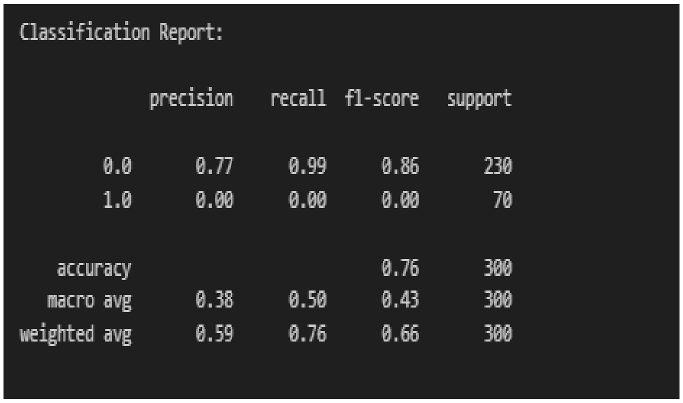
- Higher complaints correlate with higher churn.
- Promotions and referrals are associated with improved retention.
- Watch time has no influence on churn likelihood.
- Mobile usage patterns show inconsistent churn rates and do not present a clear trend.

5. Predictive Model Summary

A logistic regression model on a test set of 300 users.

Model Performance

- Accuracy: 76% – misleading due to class imbalance since most users did not churn.
- Precision and Recall for churners (class 1): Both 0.00 - meaning failure to detect churners.
- F1 Score: 0.43 – reflects weak model performance.



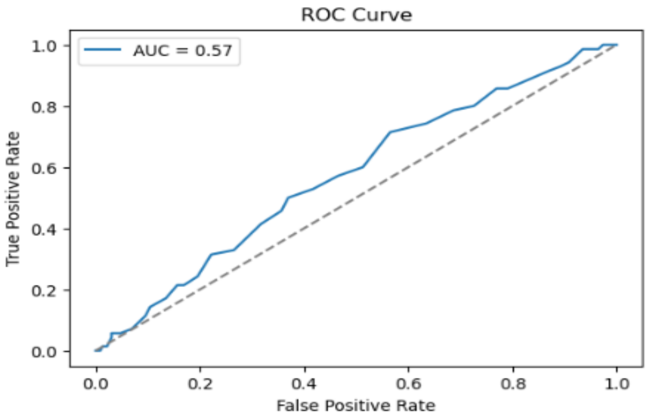
```
Classification Report:

              precision    recall  f1-score   support

         0.0       0.77      0.99      0.86       230
         1.0       0.00      0.00      0.00        70

    accuracy                           0.76       300
   macro avg       0.38      0.50      0.43       300
weighted avg       0.59      0.76      0.66       300
```

- AUC: 0.57 – only slightly better than random, suggesting poor discriminatory power.



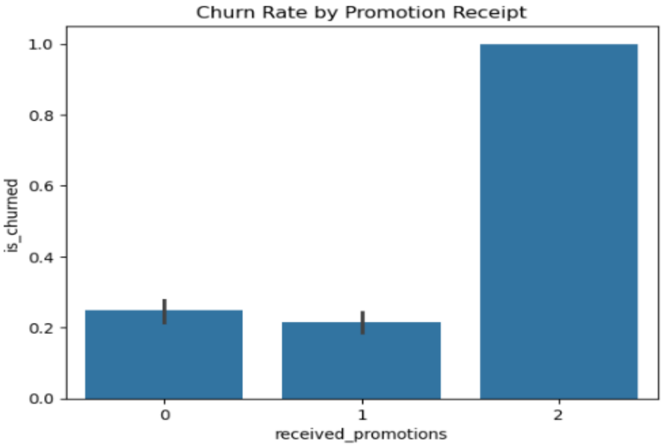
Top 3 Predictors of Churn

- gender_Male and gender_Other with coefficients –0.1846 and –0.1617 respectively show that male and 'Other' gender users are less likely to churn.
- country_UK with coefficient +0.1318 indicates that UK users are more likely to churn compared to others.

6. Business Questions

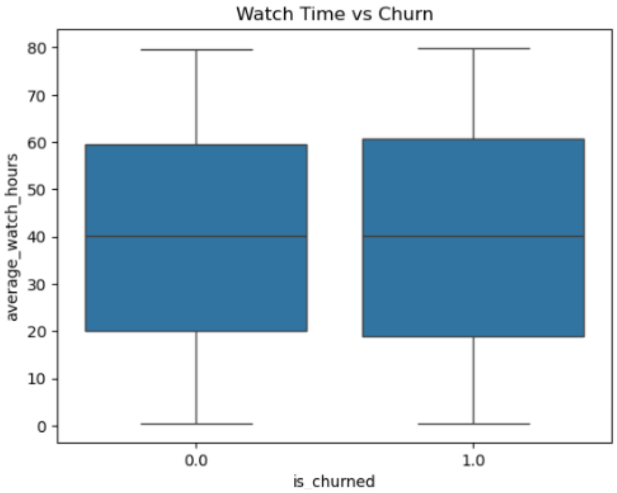
I. Do users who receive promotions churn less?

Yes, users who received promotions have shown a lower churn rate. For example, 24.8% of those who have churned have received no promotions while 21.4% of them have, as we can see from the bar chart to the right.



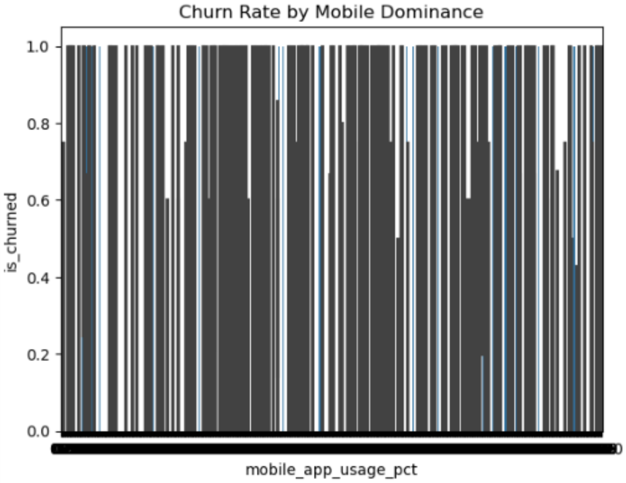
II. Does watch time impact churn likelihood?

No, there is no correlation between watch time and churn as correlation is at -0.00 and p-value of t-test is 0.8.



III. Are mobile-dominant users more likely to cancel?

Churn rates vary across mobile_app_usage_pct with no clear pattern and thus it is difficult to make result from the data gathered.



IV. What are the top 3 features influencing churn?

Based on logistic regression:

- gender_Male and gender_Other with lower churn likelihood
- country_UK with higher churn likelihood

V. Which customer segments should the retention team prioritize?

- 42 of the users who did not receive promotions have churned
- 26 of the users who did receive promotions have churned

7. Recommendations

1. Engage users early in advance with declining activity by sending them personalised content suggestions or surveys to identify pain points.
2. Offer targeted promotions to at-risk users e.g. discount offers for users who raise complaints or reduce engagement.
3. Segment users using tailored strategies. For example, by combining tenure, geography, app usage, and complaints for smarter targeting.

8. Data Issues or Risks

- The dataset contains missing values in essential columns like user_id and last_active_date which are crucial for user tracking and communication. Dropping these rows results in data loss, reducing dataset completeness. Best way to address this is as mentioned before, by making these fields mandatory for users in providing valid information as well as real-time validation, such as email verifications.
- Class imbalance (low churn rate) skewed model performance by making it overly biased toward the majority class (non-churn). As a result, the model achieved high overall accuracy but failed to identify actual churners, reducing its usefulness for retention strategies.