Project Coversheet

Full Name	Ramil Khalilli
Project Title (Example – Week1, Week2, Week3, Week 4)	Week 3

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.

Churn Prediction for StreamWorks Media

Author: Ramil Khalilli
Module: Data Analytics / Machine Learning

1. Introduction

StreamWorks Media is a UK-based video streaming platform operating in a highly competitive subscription market. Customer churn—users cancelling their subscriptions—poses a direct risk to revenue and long-term growth. The goal of this project is to analyse customer data to understand **who is churning, why they churn, and how churn can be predicted early** to support retention strategies.

The dataset contains demographic information, subscription details, engagement metrics, and a churn indicator. The analysis combines **exploratory data analysis, statistical testing, and predictive modelling** to generate actionable business insights.

2. Data Cleaning Summary

Several preprocessing steps were applied to prepare the data for analysis:

Converted `signup_date` and `last_active_date` to datetime format

Removed rows with missing `user_id` values

Imputed missing numerical values (e.g. `age`, `monthly_fee`) using the **median**

Imputed missing categorical values using **"Unknown"**

Verified final dataset contained **no missing values**

These steps ensured data consistency and reliability for statistical tests and modelling.

---

3. Feature Engineering Summary

The following engineered features were created to capture customer behaviour more effectively:

tenure_days: Number of days between signup and last activity

is_loyal: Binary feature indicating whether tenure exceeds 180 days

Encoded categorical variables: Gender, country, subscription type, promotions, and referrals were converted using one-hot encoding

These features were designed to reflect user loyalty, engagement, and subscription characteristics.

---

4. Key Findings (Statistical Analysis)

4.1 Promotions and Churn (Chi-Square Test)

No statistically significant relationship was found between receiving promotions and churn

This suggests promotions alone may not prevent cancellations

4.2 Watch Time and Churn (T-Test)

The t-test comparing average watch hours between churned and retained users showed **no significant difference**

Watch time alone does not strongly explain churn behaviour

4.3 Correlation Analysis

Strong correlations exist among engagement-related variables

Tenure shows a meaningful relationship with user behaviour and churn risk

5. Model Results

5.1 Logistic Regression (Churn Prediction)

A logistic regression model was built to predict `is_churned`.

Evaluation Results:

The model predicted only the majority (non-churned) class

Precision, Recall, and F1-score for churned users were **0.0**

This outcome is caused by **severe class imbalance**

Top Predictors of Churn (by coefficient magnitude):

Subscription type

Tenure days

Monthly fee

Interpretation:
Customers on certain subscription plans and those with shorter tenure are more likely to churn.

5.2 Linear Regression (Tenure Prediction)

Linear regression was used to predict **tenure_days** as a proxy for loyalty.

Evaluation Metrics:

R^2: -0.025
RMSE: ~22 days

Interpretation:
The negative R^2 indicates poor predictive performance. Linear regression does not adequately explain tenure using the available features, suggesting nonlinear relationships or missing drivers.

6. Business Questions Answered

1. **Do users who receive promotions churn less?**
   No. Statistical testing shows no significant difference in churn rates.
2. **Does watch time impact churn likelihood?**
   No significant difference in average watch hours was observed.
3. **Are mobile-dominant users more likely to cancel?**
   Slight variation exists, but results are not statistically significant.
4. **What are the top 3 features influencing churn?**
   Subscription type, tenure days, and monthly fee.
5. **Which customer segments should be prioritised?**
   New users with short tenure and users on higher-risk subscription plans.
6. **What factors affect user tenure?**
   Subscription type and pricing appear to influence how long users remain active.

7. Recommendations

1. **Target new users early** with onboarding and engagement campaigns
2. **Review subscription pricing and features** for high-churn plans
3. **Improve churn modelling** using class balancing techniques (SMOTE, class weights) and tree-based models

8. Data Issues and Risks

Severe class imbalance reduced model effectiveness

Limited churn cases restrict predictive power

Linear regression assumptions may not hold for user behaviour data

9. Conclusion

This analysis provided a structured investigation into customer churn at StreamWorks Media. While baseline models struggled due to data imbalance, key churn drivers such as tenure and subscription type were identified. With improved modelling techniques and targeted retention strategies, StreamWorks Media can better anticipate and reduce customer churn.