# Project Coversheet

| | |
|---|---|
| Full Name | Khalid Takar |
| Project Title (Example – Week1, Week2, Week3, Week 4) | Week3 |

**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**

# Week 3 Project: Churn Prediction for StreamWorks Media

---

## 1. Introduction

This project focuses on analysing customer behaviour data from StreamWorks Media to understand churn and explore whether churn can be predicted using historical user data. Churn is a key business problem for subscription-based services, as retaining existing customers is typically more cost-effective than acquiring new ones. The objective of this analysis is to identify behavioural patterns linked to churn and assess the effectiveness of basic statistical methods and predictive models in identifying at-risk users.

---

## 2. Dataset Overview

The dataset contains approximately 1,500 user records and includes demographic information, subscription details, engagement metrics, and customer experience indicators. Key variables include age, subscription type, monthly fee, average watch hours, mobile app usage, promotions received, complaints raised, and a churn indicator.

Before modelling, one record with a missing churn label was removed, as supervised learning requires a known outcome. The final dataset used for modelling therefore contained 1,499 users. The observed churn rate in the dataset was approximately 23%, indicating a moderately imbalanced target variable.

---

## 3. Data Cleaning and Preparation

Initial inspection revealed several data quality issues that required attention before analysis could be performed.

Key cleaning actions included:

- Conversion of signup and last active dates into datetime format to support tenure calculations.
- Standardisation of categorical variables such as gender, country, subscription type, and promotion indicators to ensure consistent grouping.
- Handling of missing values using simple and defensible approaches, including median imputation for numeric fields and default category labels such as "Unknown" for missing categorical values.
- Cleaning and standardisation of the churn variable into a strict binary format (0 = retained, 1 = churned).

These steps ensured the dataset was consistent and suitable for statistical analysis and modelling.

---

**4. Feature Engineering**

Several additional features were created to better capture customer behaviour and engagement:

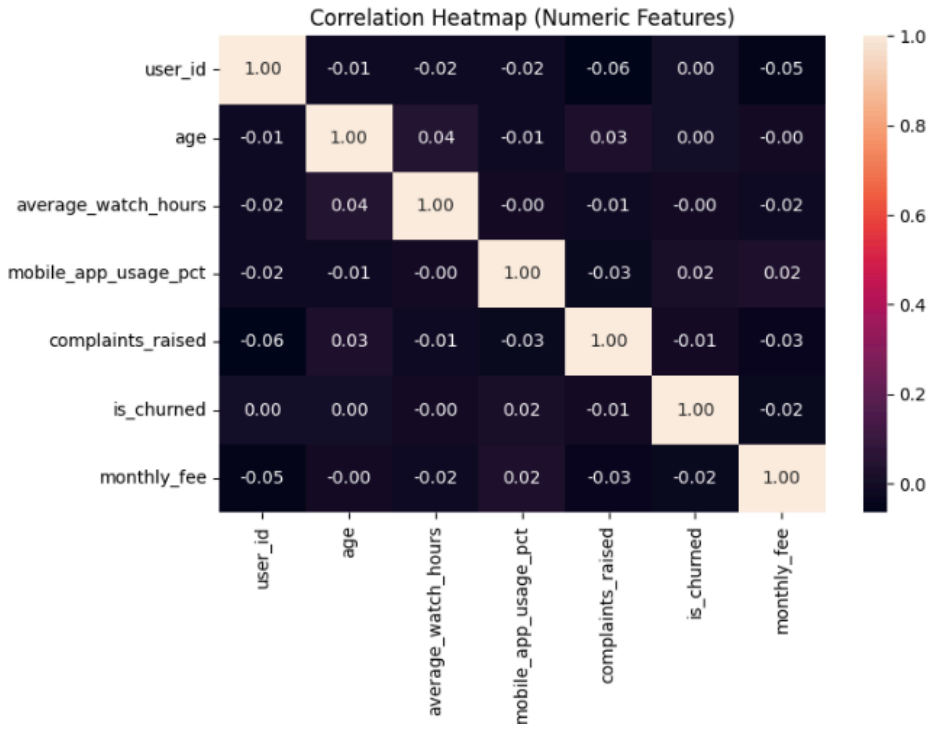
- **Tenure (days):** the number of days between signup and last activity.
- **Loyalty indicator:** a binary flag identifying users with long tenure.
- **Watch-to-fee ratio:** a value-for-money measure combining engagement and pricing.
- **Heavy mobile usage flag:** identifying users who primarily engage via mobile devices.

These engineered features were designed to reflect engagement depth and customer value rather than relying solely on raw activity measures.

---

**5. Exploratory Analysis**

Exploratory analysis was used to understand relationships between key numeric variables and to guide later modelling choices.




The heatmap suggested that engagement and pricing-related variables were more strongly related to each other than demographic variables, indicating that behavioural signals may be more useful than static user attributes when analysing churn.

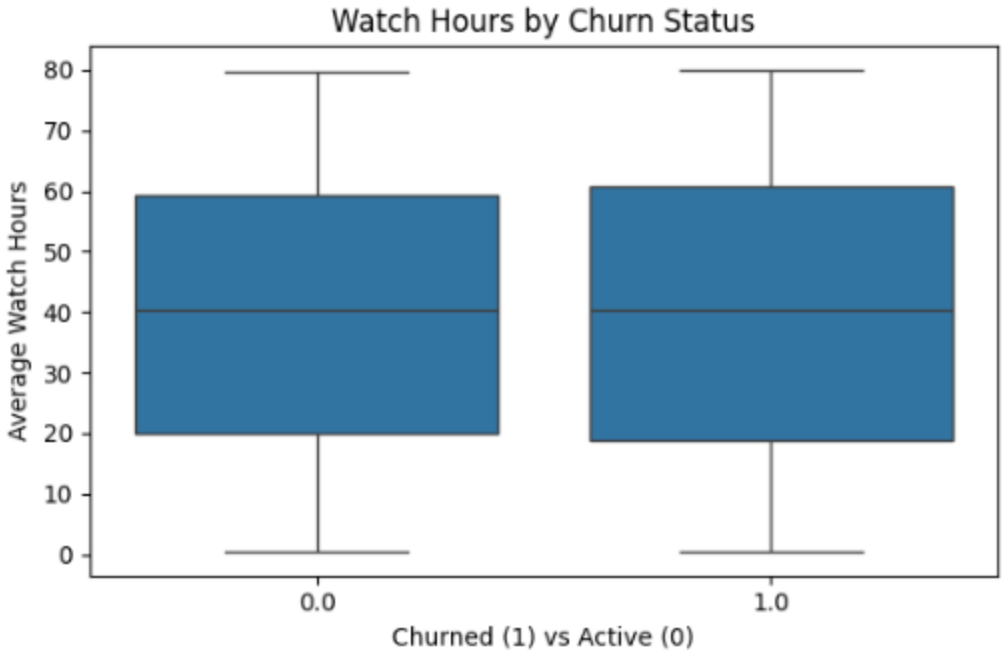
---

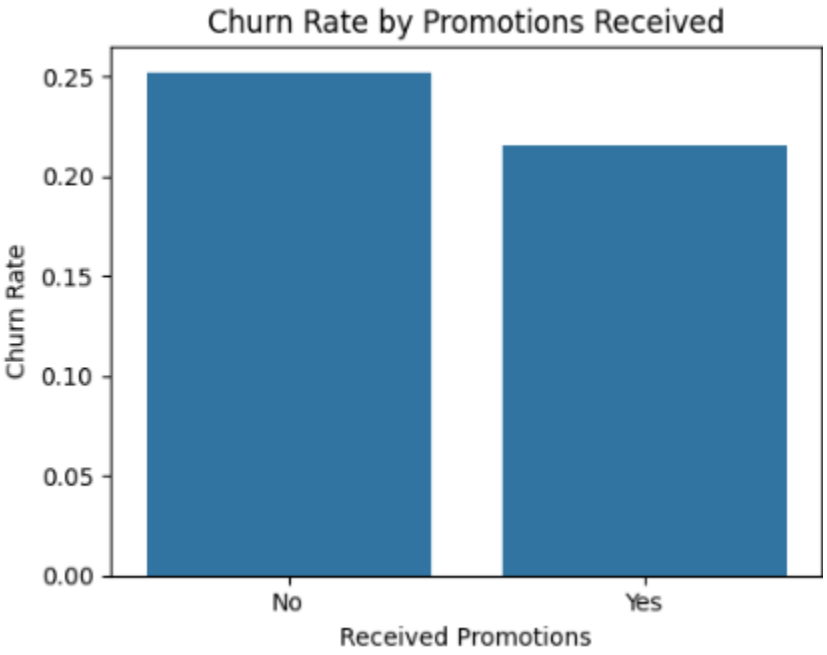
**6. Statistical Analysis**

Statistical tests were applied to assess whether churn was associated with selected customer attributes.

Chi-square tests were conducted to examine relationships between churn and categorical variables such as gender, promotions received, and referral status. None of these tests produced statistically significant results at the 5% significance level, suggesting weak or inconsistent relationships between these categorical features and churn.

A t-test was used to compare average watch hours between churned and retained users. The results showed no meaningful difference between the two groups, indicating that overall average engagement alone is not a strong differentiator of churn behaviour.




---

### 7. Predictive Modelling

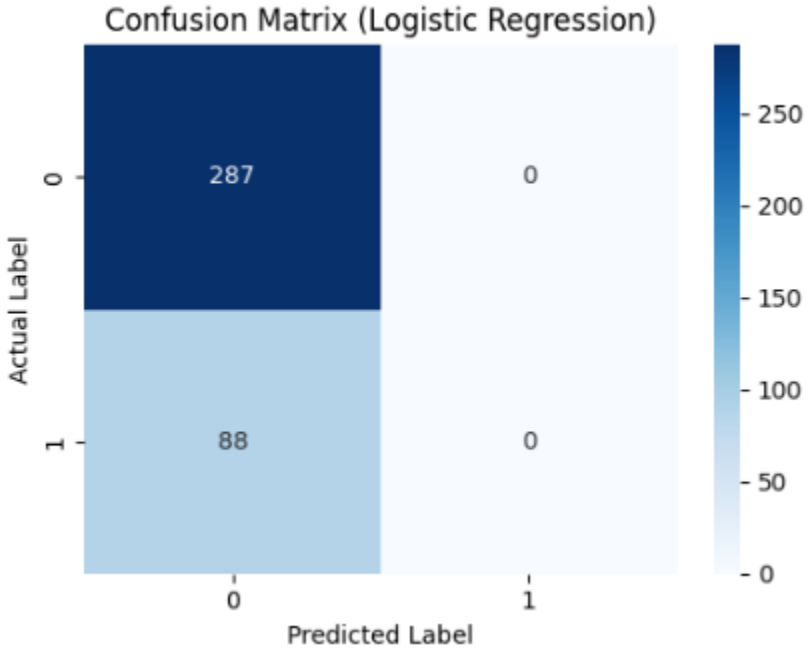
Two predictive models were developed to explore different analytical objectives.

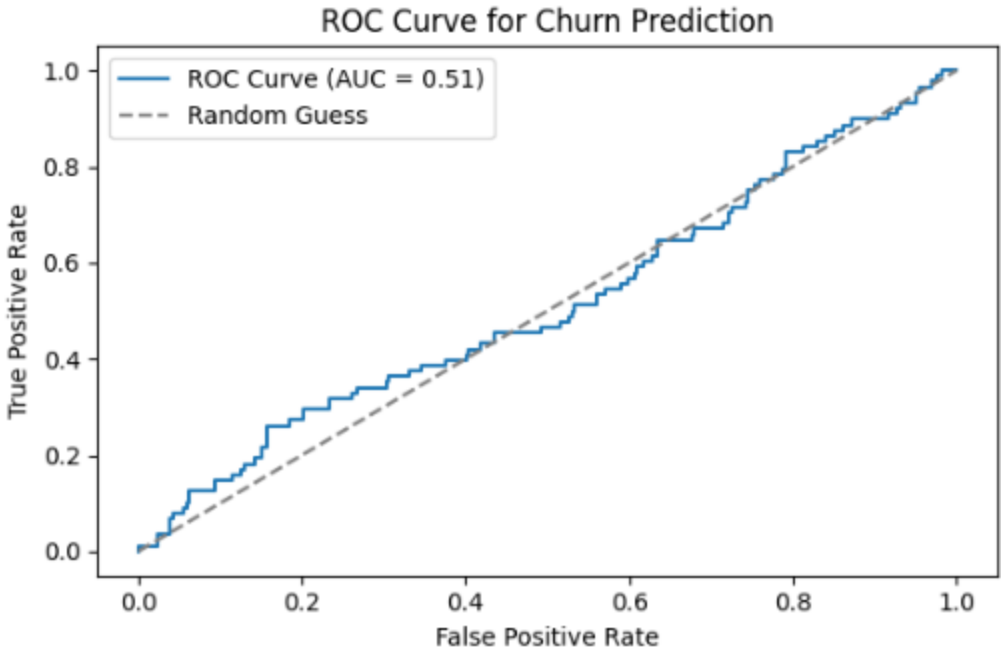
#### 7.1 Logistic Regression: Churn Prediction

A logistic regression model was trained to predict whether a user would churn. The model used a train-test split and included both original and engineered features. Categorical variables were one-hot encoded and numeric features were standardised.

The model achieved an accuracy of approximately 77% on the test set. However, further inspection showed that the model primarily predicted non-churned users and struggled to correctly identify churned users. This behaviour is consistent with class imbalance and limited churn-specific signal in the available features.

The ROC-AUC score was close to 0.5, indicating that the model performed only slightly better than random guessing when distinguishing between churned and retained users.





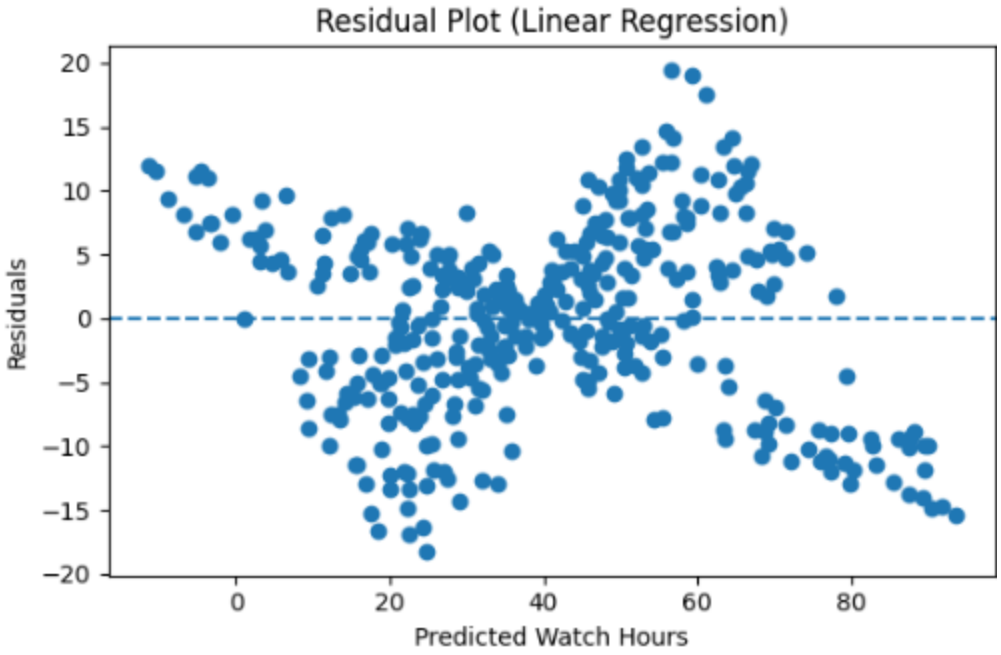

While the model's predictive performance was limited, analysis of model coefficients suggested that value-for-money indicators and missing profile information may influence churn risk. These findings should be interpreted cautiously.

---

### 7.2 Linear Regression: Predicting Engagement

A linear regression model was developed to predict average watch hours. This model performed significantly better than the churn classifier, achieving a high $R^2$ value and relatively low error metrics.

The results suggest that engagement levels are more predictable from available features than churn itself. Pricing context, subscription type, and engagement-related features were among the strongest predictors of watch behaviour.




---

**8. Key Insights**

- Churn is not strongly explained by simple demographic or high-level engagement metrics in this dataset.
- Behavioural value indicators, such as watch-to-fee ratio, appear more informative than raw watch time.
- Predicting engagement levels is more feasible than directly predicting churn using the available data.
- Class imbalance significantly affects churn model performance and should be addressed in future work.

---

**9. Recommendations**

StreamWorks Media should focus on collecting richer behavioural signals, such as recent engagement trends, customer support interactions, and cancellation reasons, to improve churn prediction. Early warning systems should prioritise changes in behaviour over static averages. Improving data completeness at signup will also enhance future modelling and segmentation efforts.

---

## 10. Data Issues and Risks

Key data risks include missing values in pricing fields, reliance on imputed data, and incomplete churn labelling. These issues limit the effectiveness of predictive models and may lead to misleading conclusions if not addressed. Implementing stronger data validation and ongoing data quality monitoring would improve the reliability of future analyses.