Project Coversheet

Full Name	Rory Scott
Email	Rdscott40ksx@hotmail.co.uk
Contact Number	+447590235745
Date of Submission	25/07/2025
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.

YOU CAN START YOUR PROJECT FROM HERE

Uptrail Internship Program Week 2 Project

Churn Prediction for StreamWorks Media

By Rory Scott

Introduction

I've joined the **Data Strategy Team at StreamWorks Media**, a fast-growing UK-based video streaming platform competing with global giants like Netflix and Amazon Prime. With rising customer acquisition costs and increased market competition, **retaining existing users has become a critical focus** for the business.

In this project, I will investigate **customer churn** - users who cancel their subscriptions - with two main goals:

- **Analyse churn behaviour** to understand who is churning and why.
- **Build a predictive model** to identify users likely to churn, so the retention team can take proactive steps.

To achieve this, I will conduct **statistical analysis** (including correlation and hypothesis testing), followed by **predictive modelling** using logistic regression. I will also evaluate the model's performance using metrics such as **precision, recall**, and **ROC-AUC**, ensuring the solution is both reliable and actionable.

Section 1: Data Cleaning Summary

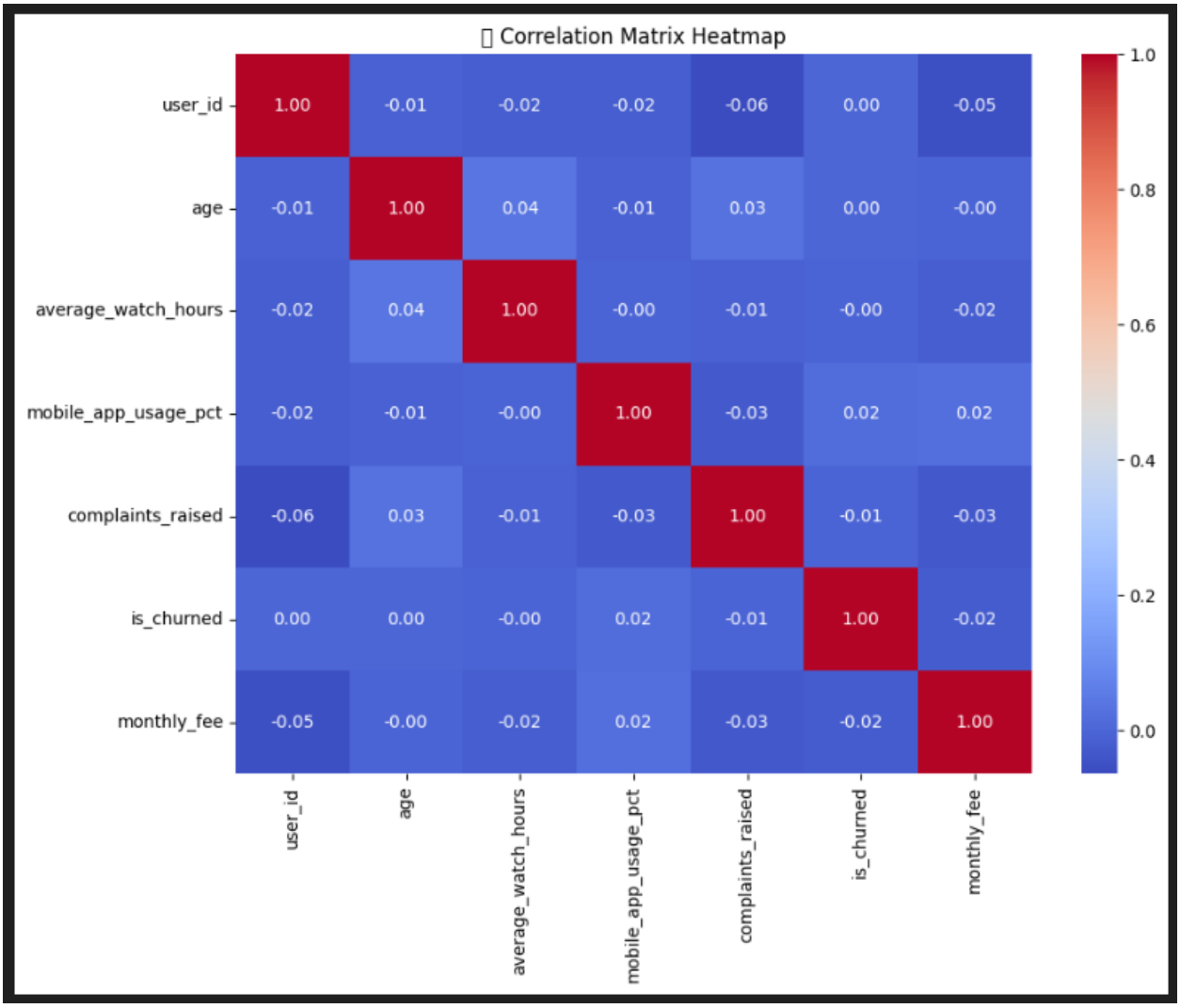
📌 Part 1: Identifying Data Frame Structure

- **1500 rows, 14 columns.**
- >5 missing values per column, except **145 missing values in monthly_fee**.
- Signup_date and last_active_date need to be changed to date-time.

```
<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
```

📌 Part 2: Correlation Matrix Heatmap (Peek)

A correlation matrix heatmap was plotted to explore linear relationships between numerical features and to identify variables that may be strongly associated with customer churn.



Finding: There's no strong linear signal driving churn from any **single numeric feature**.

📌 Part 3: Handling Missing and Duplicate Fields

- Rows with **missing values were dropped across all columns except for 'monthly_fee'**, as fewer than 5 records out of 1,500 contained missing values - posing minimal risk to data integrity.

- The **'monthly_fee'** column was handled by filling **143 missing entries** (~10%) with the **mode**, reflecting the most common pricing tier, avoiding skewing the data with mean or median imputation.
- **No duplicate records** were found based on the **'user_id'** field.

```
Handled Missing Values - Dropped Rows = 24
Handled Missing Values - Filled Monthly Fee: 143 records filled with mode value.

Remaining Missing Values:

user_id                 0
age                     0
gender                  0
signup_date             0
last_active_date        0
country                 0
subscription_type       0
average_watch_hours     0
mobile_app_usage_pct    0
complaints_raised       0
received_promotions     0
referred_by_friend      0
is_churned              0
monthly_fee             0
tenure_days             0
is_loyal                0
dtype: int64
0 duplicate rows dropped based on 'user_id'.
```

- All missing values have been handled.

Section 2: Feature Engineering Summary

📌 Introduction

Feature engineering plays a critical role in preparing data for machine learning. It includes techniques to improve data quality, expressiveness, and compatibility with specific algorithms.

📌 Part 1: New Features Created / Why

Feature Name	Description
tenure_days	Days between signup and last active date; measures user longevity.
is_loyal_binary	Flags users as loyal (1) if tenure > 180 days.
loyal_and_referred	Combines loyalty and referral status to highlight strong retention signals.
loyal_x_watch_hours	Captures how loyal users engage with the platform through content consumption.
referred_x_tenure	Measures long-term engagement of referred users.
promos_x_complaints	Evaluates if promotions correlate with increased dissatisfaction.
watch_per_fee	Normalizes content consumption relative to price.
complaints_per_day	Measures complaint frequency over the user's lifespan.
loyal_x_tenure	Combines loyalty status with actual tenure in days.
fee_per_day	Average fee paid per active day.
engagement_score	Composite feature: watch hours × mobile app usage percentage.
tenure_bucket	Binned tenure into: New, Recent, Mid-Term, Long-Term categories.

📌 Part 2: Feature Encoding / Scaling

Transforming raw data into a model-ready format involved the following:

One-Hot Encoding:

Applied to categorical features to avoid imposing ordinal relationships. This preserved the neutrality of groups like countries, subscription types, and tenure buckets.

Standard Scaling:

Applied to numeric features using StandardScaler. This ensures fair treatment of variables by bringing them to the same scale—essential for distance-based algorithms and regularized linear models like logistic regression.

```
gender_Female  gender_Male  gender_Other  country_Canada  country_France  \
          0.0          0.0           1.0             0.0             1.0
          0.0          1.0           0.0             0.0             0.0
          0.0          1.0           0.0             0.0             0.0

country_Germany  country_India  country_UK  country_USA  \
            0.0            0.0         0.0          0.0
            0.0            1.0         0.0          0.0
            0.0            0.0         1.0          0.0

subscription_type_Basic  ...  received_promotions_binary  \
                    0.0  ...                           0
                    1.0  ...                           0
                    0.0  ...                           0

referred_by_friend_binary  is_loyal_binary       age  average_watch_hours  \
                        0                0  0.809870             0.116605
                        1                1  1.672616             1.104960
                        1                1  0.146219             0.007755

mobile_app_usage_pct  complaints_raised  monthly_fee  tenure_days  \
            0.908786          -0.880114     0.352295    -1.137463
            1.629441           0.877732    -1.124941     1.040758
           -0.126720          -1.466063     1.238636     1.527445

is_churned
       1.0
       1.0
       1.0
```

- **'model_df'** above contains fully processed, numeric features and is ready for input into machine learning models.

Section 3: Key Findings

📌 Part 1: Is churn related to gender, received_promotions, or referred_by_friend?

Chi-square tests of independence were performed to assess whether churn was associated with each categorical variable.

Hypotheses:

For each variable:

- **H_0**: There is **no association** between the variable and churn.
- **H_1**: There **is an association** between the variable and churn.

Results:

Feature	p-value	Result
gender	0.1336	Fail to reject H_0 (no relationship)
received_promotions	0.1130	Fail to reject H_0 (no relationship)
referred_by_friend	0.5397	Fail to reject H_0 (no relationship)

Conclusion:

No statistically significant relationships with churn were found among the tested features (all p > 0.05). Although the p-values for **'gender'** and **'received_promotions'** fall between 0.1 and 0.15, therefore hinting at weak evidence against the null

hypothesis, these features do not appear to be meaningfully associated with churn in this sample. They may, however, warrant further investigation with a larger or more targeted dataset.

📌 Part 2: Does 'watch time' differ significantly between churned and retained users?

Independent Samples T-Test evaluates whether the mean watch time differs between users who churned and those who didn't.

Hypotheses:

- **H_0**: There is **no difference** in average watch time between churned and retained users.
- **H_1**: There **is a difference** in average watch time between churned and retained users.

Results:

- **Mean watch time (churned)**: 39.62 hours
- **Mean watch time (retained)**: 40.01 hours
- **p-value**: 0.7856

Conclusion: The difference in average watch time is **not statistically significant** ($p > 0.05$), so we **fail to reject H_0**. This suggests that **watch time is not strongly associated with churn** in this dataset.

Section 4: Model Results

📌 Part 1: Our Problem

In a competitive streaming landscape, retaining users is crucial. The goal is to identify which users are most at risk of churning and uncover the key drivers behind their decisions.

📌 Part 2: Building a Logistic Regression Model

Logistic regression was chosen for its simplicity, interpretability, and effectiveness in binary classification. The model aims to **predict churn based on user behavior and profile attributes**.

📌 Part 3: Train-Test Split

The dataset was split into **80% training and 20% testing** using train_test_split(). This ensures model evaluation is performed on unseen data, supporting unbiased performance estimates.

📌 Part 4: Feature Scaling

StandardScaler was applied to **numeric variables** to ensure equal contribution across features. **Categorical variables were one-hot encoded** but not scaled, as their values represent categories rather than magnitude.

📌 Part 5: Model Training

The logistic regression model was trained on scaled numeric and encoded categorical features. Since churned users made up only ~23% of the dataset (shown below), we used class_weight = 'balanced' to address class imbalance and prevent the model from ignoring the minority class.

is_churned	proportion
0.0	0.769648
1.0	0.230352

📌 Part 6: Model Predictions

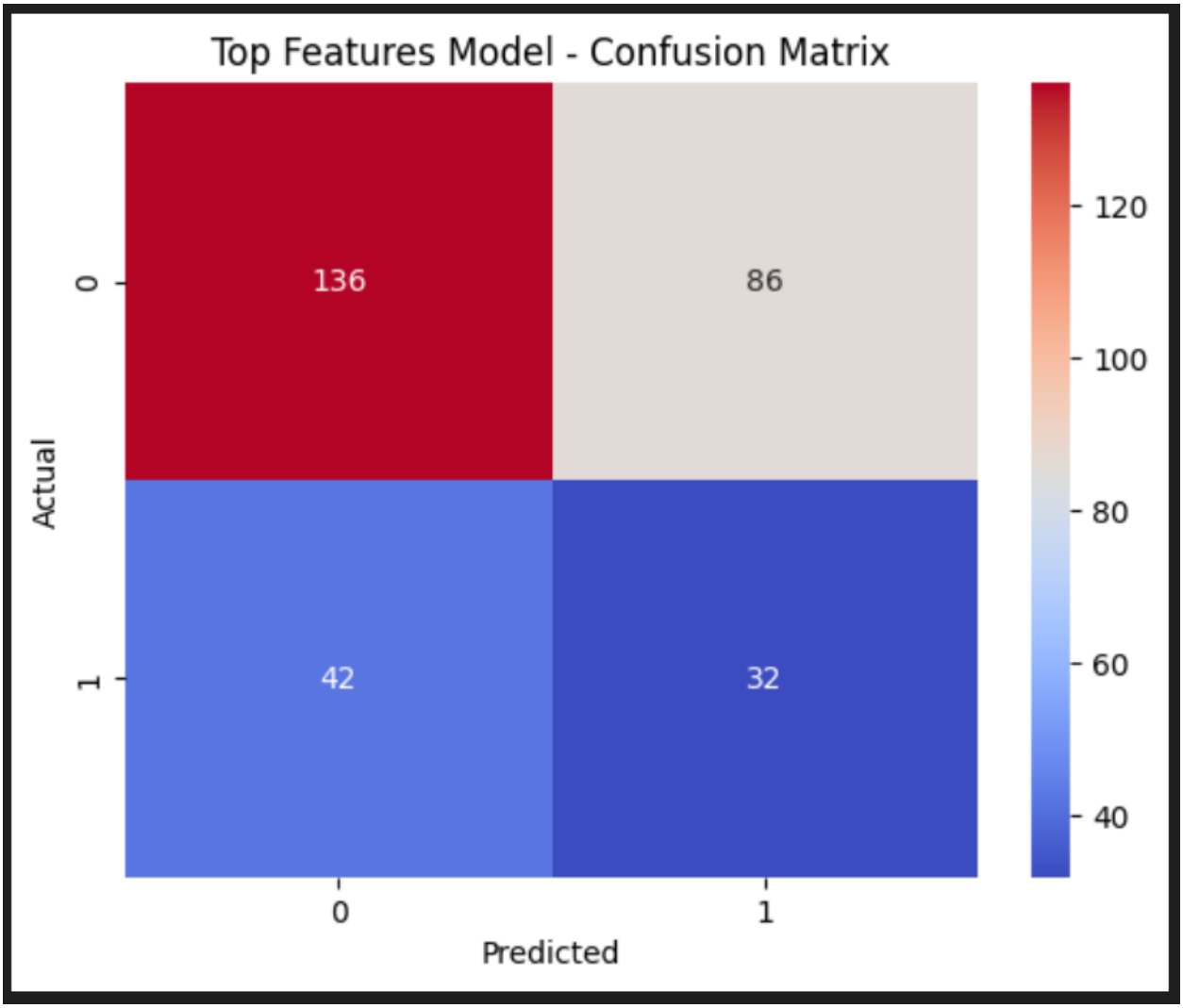
The model was used to predict churn on the test set, producing:

- A **probability score indicating churn risk**
- A **binary prediction** (0 = retain, 1 = churn)

These predictions formed the basis for performance evaluation and business insight.

📌 Part 7: Model Evaluation

The **confusion matrix** (shown below) provides a clear view of how well the model distinguishes between churned and retained users, highlighting its ability to correctly identify churners while also showing the trade-offs in prediction accuracy.

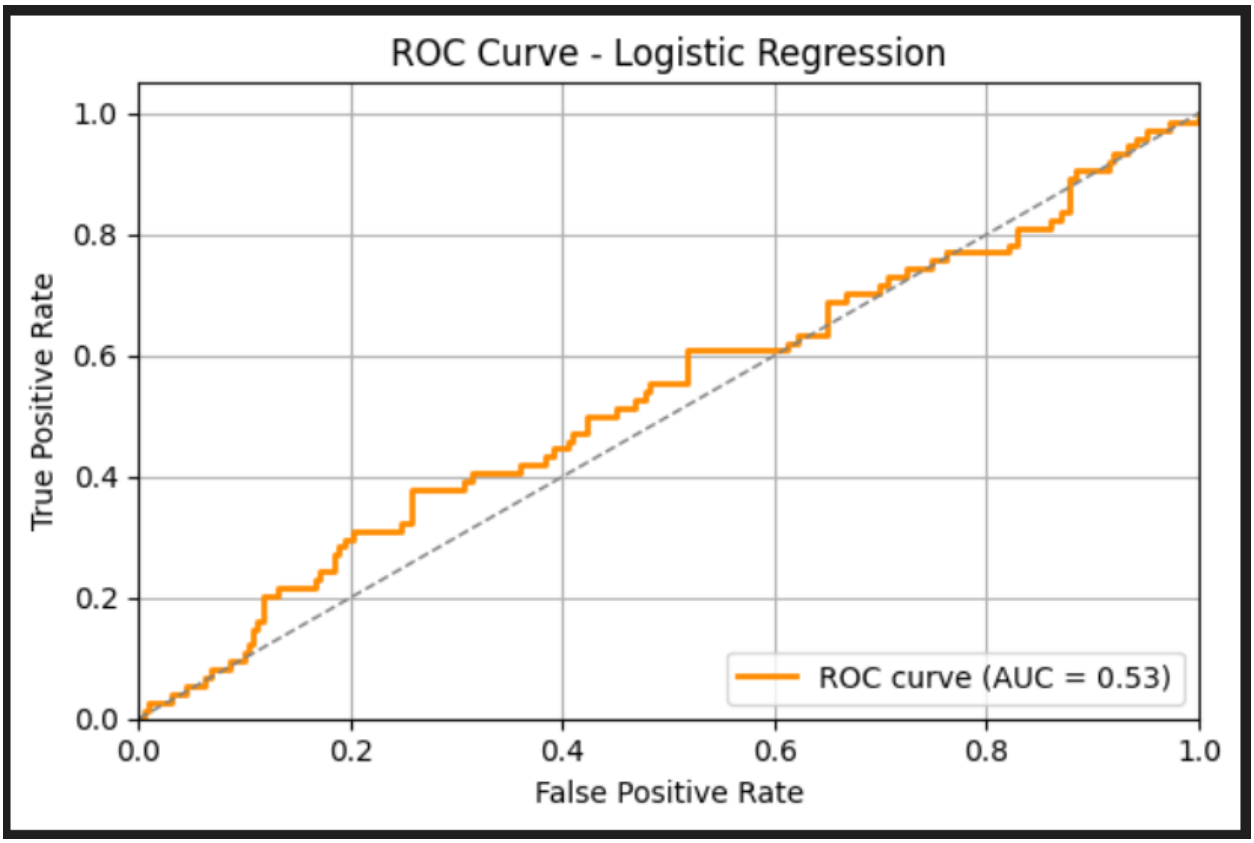


- **136 users** were correctly predicted to **stay** (**true negatives**).
- **32 users** were correctly predicted to **churn** (**true positives**).
- **86 users** were predicted to churn but stayed (**false positives**).
- **42 users** churned but the model failed to detect them (**false negatives**).

The model was evaluated on multiple metrics:

- **Accuracy**: 56.76%
- **Precision**: 0.27
- **Recall**: 0.44
- **F1 Score**: 0.33

ROC curves are plotted to assess a model's ability to differentiate between classes at various thresholds, allowing analysts to visualize the trade-off between recall and false positive rate, and compare model performance using the AUC score.



- **AUC Score:** The ROC curve shows an AUC of 0.53, indicating only marginally better performance than random guessing.
- **Churn Detection:** The model achieves 44% recall, meaning it successfully identifies nearly half of churners.
- **Business Trade-Off:** Although precision is low, the misclassification cost may be acceptable if retention efforts are inexpensive and churn is costly.

📌 Part 8: Feature Importance

Top 3 features based on model coefficients:

• **tenure_bucket_Mid-Term** (+1.323): Customers in the mid-term tenure bracket are significantly more likely to churn, suggesting disengagement may peak during this period.

• **received_promotions_binary** (−1.210): Receiving promotions is strongly associated with lower churn, highlighting the potential value of well-timed offers.

• **referred_by_friend_binary** (−1.082): Referred customers are also less likely to churn, suggesting higher loyalty in this segment.

These insights suggest that mid-term users are a churn risk and may benefit from engagement nudges, while promotions and referral programs are effective retention levers.

Section 5: Business Insights and Answers

📌 Question 1: Do users who receive promotions churn less?

Users who received promotions **were significantly less likely to churn**, as shown by a strong negative coefficient (−1.210) in the logistic regression model. This suggests that promotions may play a key role in encouraging customer retention. Targeted, well-timed offers could be a valuable strategy for reducing churn.

📌 Question 2: Does watch time impact churn likelihood?

The feature average_watch_hours has a **negative coefficient (-0.21)**, meaning higher watch time is weakly associated with lower churn. However, **the effect is weak**, so while watch time helps, it's not a decisive predictor on its own.

📌 Question 3: Are mobile dominant users more likely to cancel?

The feature mobile_app_usage_pct has a **negative coefficient (-0.30)**, suggesting that users who engage more through mobile are **less likely to churn**. This contradicts any assumption that mobile-dominant users are more likely to cancel.

📌 Question 4: What are the top 3 features influencing churn based on your model?

These features provide the strongest linear signals in the model and can guide targeted retention strategies.

1. **tenure_bucket_Mid-Term** (1.32): Users in the mid-term stage of tenure (90–180 days) are the most likely to churn, indicating a potential drop in engagement during this period.
2. **received_promotions_binary** (−1.21): Users who received promotions are less likely to churn, reinforcing the importance of marketing campaigns.
3. **referred_by_friend_binary** (−1.08): Referred users show a significantly lower churn risk, suggesting referral-based users may be more engaged or loyal.

These insights suggest retention strategies should focus on mid-tenure users and leverage referral and promotion campaigns.

📌 Question 5: Which customer segments should the retention team prioritise?

Users in the **mid-term** and **long-term** tenure buckets are significantly more likely to churn based on model coefficients. Those who **haven't received promotions** or **weren't referred by friends** are also at higher risk. The retention team should prioritise these segments with targeted offers and referral incentives to reduce churn.

Section 7: Business Recommendations, Data Issues and Risks

📌 Part 1: Business Recommendations

1. Refocus Promotional Strategy

Promotions appear to reduce churn (coefficient −1.21), but model uncertainty remains.

- **Action:** Validate effectiveness via **A/B testing** to avoid assumptions based on weak linear signals.

2. **Support Mid-Term and At-Risk Users**

'tenure_bucket_Mid-Term' had the highest positive coefficient (~1.32), suggesting **higher churn risk**.

- **Action:** Create **targeted retention programs** for users in months 3–6.

3. **Leverage the Referral Channel Carefully**

Referred users were **less likely to churn** (coefficient −1.08), but general churn is still high.

- **Action:** Invest in **onboarding and reward structures** to boost referred-user retention further.

📌 Part 2: Data Issues and Risks

1. Model Performance Limitations

Despite feature engineering, SMOTE, and threshold tuning, the logistic model's F1 score plateaued (~0.40), indicating weak predictive power.

- **Risk**: May lead to **ineffective churn interventions.**
- **Recommendation**: Explore ensemble models and time-based or qualitative features for deeper insights.

2. Synthetic Balancing (SMOTE) vs. Reality

SMOTE boosted recall but significantly lowered precision, leading to **high false positives.**

- Risk: Could result in **targeting users unlikely to churn.**
- Recommendation: Validate predictions against real churn outcomes before deploying.

4. **Lack of Temporal or Behavioral Signals**

The dataset lacks time-based or dynamic engagement data (e.g., session trends or watch streaks).

- Risk: Static features may miss key churn triggers.
- Recommendation: Integrate temporal/event-level data in future models to improve accuracy.

[END OF PROJECT]