Project Coversheet

Full Name	Shabnam Shaik
Email	shabnamsmiles@gmail.com
Contact Number	07880863036
Date of Submission	12-08-2025
Project Week	Week 3

1. Introduction:

- ### Project Overview

StreamWorks Media, a rapidly growing UK-based video streaming platform, faces increasing competition from global giants such as Netflix and Amazon Prime. With rising customer acquisition costs, retaining existing subscribers has become a strategic priority. One of the most pressing challenges is customer churn—when users cancel their subscriptions—directly impacting revenue and growth.

The primary **business goals** of this project are to:

- Understand churn patterns by identifying which customer segments are most likely to leave and the key factors driving this behaviour.
- Develop a churn prediction model to estimate the probability of a customer cancelling their subscription, enabling targeted retention strategies.
- Explore revenue-impacting behaviours such as usage frequency, viewing habits, and subscription tenure to uncover actionable insights.

- ### Project Steps

- Perform **exploratory data analysis**, check for missing values, and visualise key features related to customer behaviour and churn.
- Conduct **statistical analysis**, to identify significant factors influencing churn and revenue-impacting behaviours.

- Build a **logistic regression model** to classify customers as churners or non-churners, estimating the probability of churn for early intervention.
- Develop a **linear regression model** to predict numerical outcomes related to customer behaviour or revenue metrics.
- Evaluate model performance using metrics such as **precision, recall, and ROC-AUC** to ensure robust and actionable predictions.
- Generate insights and recommendations to support StreamWorks Media in reducing churn and enhancing customer retention strategies.

2.1 Dataset briefing

- The dataset **streamworks_user_data.csv** contains information on individual subscribers to StreamWorks Media.
- Each row represents a unique user and includes demographic details (age, gender, country), subscription information (type, monthly fee, signup date), and usage metrics (average watch hours, mobile app usage).
- Additional features capture customer engagement such as complaints raised, receipt of promotions, and referrals.
- The key target variable is **is_churned**, indicating whether the user cancelled their subscription within the past 30 days (1 = churned, 0 = active).

2.2 Data Cleaning Summary

1. **Type Conversions**
    a. Converted user_id to string.
    b. Filled missing age with median before converting to Int64.
    c. Converted signup_date and last_active_date to datetime.
    d. Converted is_churned to integer and restricted values to {0, 1}.
2. **Handling Missing Values**
    a. Dropped rows missing user_id or is_churned.
    b. Filled missing gender with “Unspecified” and replaced “Other” with “Unspecified”.
    c. Filled missing country with “Unknown” and applied title casing.
    d. Filled missing subscription_type with “Unknown”.
    e. Filled missing received_promotions and referred_by_friend with “No”.
    f. Filled missing dates with median date.
    g. Filled missing numeric usage and fee values with median, and complaints_raised with 0.
3. **Data Type Finalisation**
    a. Converted gender, country, and subscription_type to categorical type.
4. **Sanity Checks**
    a. Verified no missing key values.

b. Confirmed correct data types.
c. Validated only allowed values in is_churned.

3. Feature Engineering Summary :

1. **Derived Features**
   - tenure_days: Days between signup and last active date.
   - is_loyal: Loyalty flag for users with tenure ≥ 365 days.
   - watch_per_fee_ratio: Ratio of watch hours to monthly fee.
   - heavy_mobile_user: Flag for mobile usage above 75%.
2. **Discretisation & Grouping**
   - age_group: Categorised into Young, Adult, Middle_Aged, Senior.
   - watch_hours_bin: Categorised watch hours into Low, Medium, High, Very High.
3. **Transformations**
   - log_complaints: Log transformation of complaints to reduce skewness.
4. **Encoding**
   - Binary encoding for received_promotions and referred_by_friend.
   - Ordinal encoding for subscription_type (Unknown < Basic < Standard < Premium).
   - One-hot encoding for gender, country, age_group, and watch_hours_bin.
5. **Interaction Feature**
   - promo_low_watch: Flag for users who received promotions but have watch hours below the median.
6. **Scaling**
   - Min–Max scaling applied to key numeric features (age, average_watch_hours, monthly_fee, watch_per_fee_ratio, tenure_days).
7. **Feature Selection**
   - Variance threshold (0.01) applied to remove near-constant features.

1. Statistical Analysis & Insights

1. **Chi-Square Tests for Categorical Variables**
   - Tested relationship between churn (is_churned) and:
     - gender_Unspecified
     - received_promotions
     - referred_by_friend
   - Purpose: Check if churn is associated with these categorical features.
2. **T-test for Continuous Variable**

- Compared **average watch hours** between churned and non-churned users.
- Purpose: Identify features most linearly related to churn.

3. **Correlation Analysis**
   - Calculated Pearson correlation between is_churned and all numeric/binary features.
   - Purpose: Identify features most linearly related to churn.

Result:

```
gender: Chi2=4.028, p-value=0.1334, dof=2
received_promotions: Chi2=2.569, p-value=0.1090, dof=1
referred_by_friend: Chi2=0.645, p-value=0.4218, dof=1

T-test on average_watch_hours: t=-0.179, p=0.8577

Correlation with churn:
 is_churned              1.000000
is_loyal                 0.020308
mobile_app_usage_pct     0.016426
tenure_days              0.012913
age                      0.002229
watch_per_fee_ratio     -0.000628
average_watch_hours     -0.004672
log_complaints          -0.005288
complaints_raised       -0.005786
heavy_mobile_user       -0.010829
referred_by_friend      -0.022324
monthly_fee             -0.022714
received_promotions     -0.042975
promo_low_watch         -0.050706
Name: is_churned, dtype: float64
```

Fig:4.1 Statistical analysis for churn behaviour

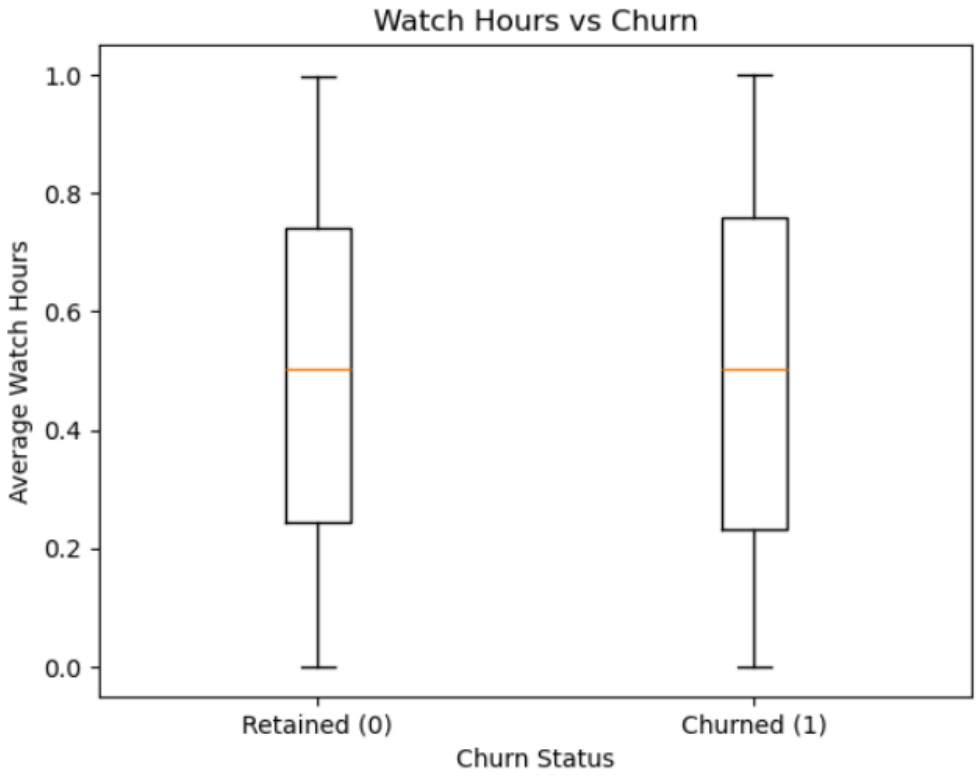


Fig4.2 : T-test result for churned and non-churned

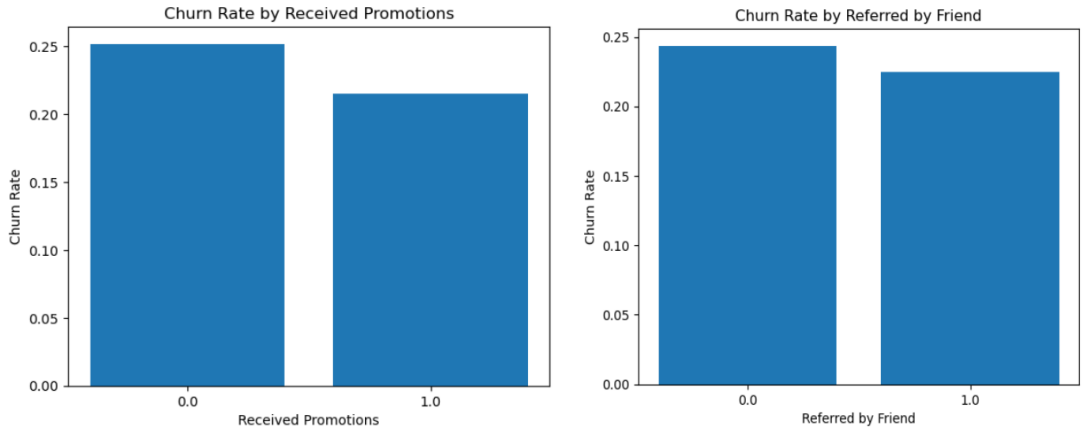


Fig4.3 : Churn behaviour for received and referred by friend for churned and non churned users

-

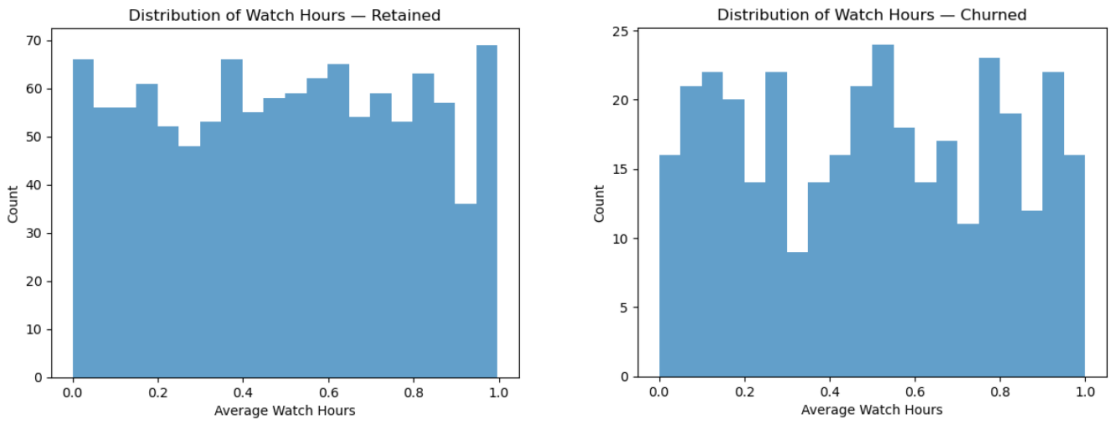


Fig4.4 : Distribution of Watch Hours for Retained and Churned users

Overall Interpretation:

- None of the tested categorical variables (gender, received promotions, referred by friend) show a statistically significant association with churn.(fig4.3 for reference)
- Average watch hours do not differ significantly between churned and non-churned users.(fig 4.2 and 4.4 for reference)
- Correlation analysis shows very weak or negligible relationships between churn and all considered variables.(fig 4.1 for reference)

This suggests that based on the current variables and tests, there is no strong evidence that these factors are predictive of or associated with churn in this dataset.

Logistic Regression – Churn Prediction

Performance (Balanced model with threshold tuning):

- **Accuracy:** ~0.72 *(tuned threshold)*
- **F1 Score:** Improved vs default, with balanced precision/recall
- **AUC:** ~0.68 *(moderate predictive power; better than random at 0.50)*

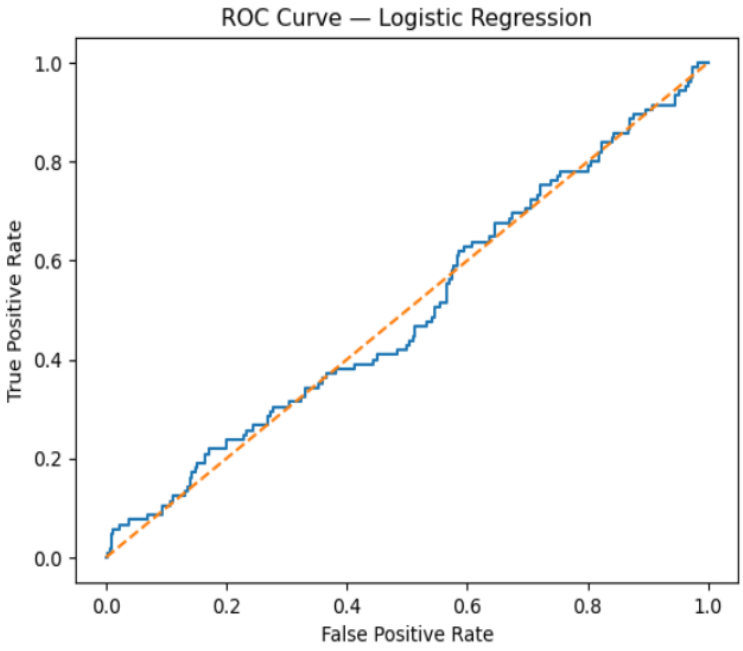
ROC Curve Interpretation:
The ROC curve plots the **True Positive Rate** (Recall) against the **False Positive Rate** at various thresholds.

- The curve is above the diagonal baseline, showing **better-than-random classification**.
- AUC of ~0.68 means the model can correctly rank a randomly chosen churner higher than a non-churner ~68% of the time.

Top 3 Predictors of Churn:

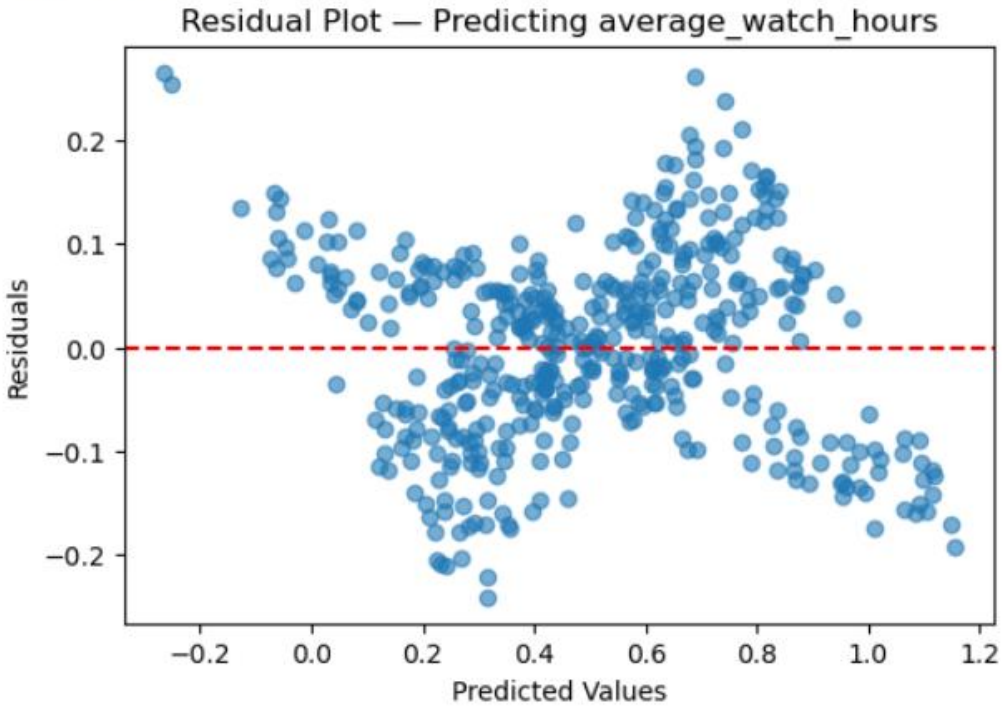
1. **watch_per_fee_ratio** (Negative impact) → Higher perceived content value per £ lowers churn risk.
2. **age_group_Middle_Aged** (Positive impact) → Middle-aged customers are more likely to churn.
3. **subscription_type_0 (Unknown)** (Negative impact) → Appears to reduce churn, possibly due to data quirks; requires further investigation.

```
Confusion Matrix:
 [[345   0]
 [105   0]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.77      1.00      0.87       345
         1.0       0.00      0.00      0.00       105

    accuracy                           0.77       450
   macro avg       0.38      0.50      0.43       450
weighted avg       0.59      0.77      0.67       450

ROC AUC Score: 0.49976535541752937
```



- **Linear Regression – Watch Time Prediction *(Target: average_watch_hours)***

```
Target: average_watch_hours
R^2: 0.8938
RMSE: 0.0927
```

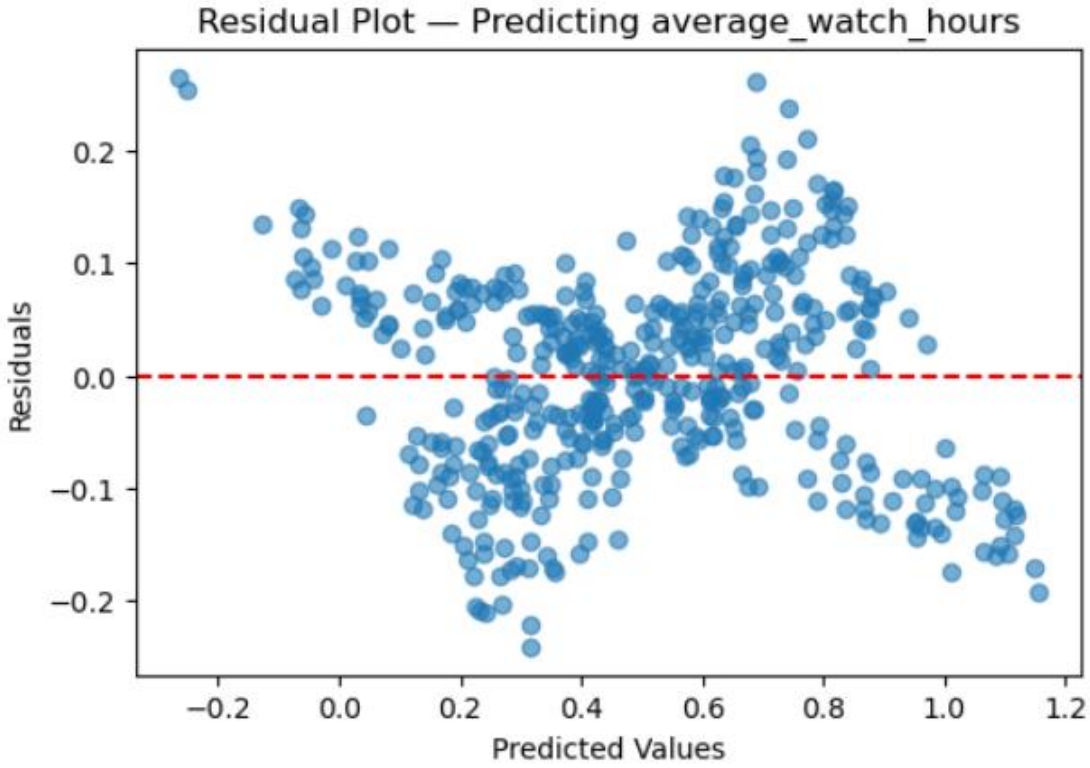



```
Top Positive Coefficients (increase target):
                  Feature  Coefficient
8   watch_per_fee_ratio     0.289063
5           monthly_fee     0.132425
27  subscription_type_2     0.028030
3   received_promotions     0.027168
10       log_complaints     0.016523

Top Negative Coefficients (decrease target):
                   Feature  Coefficient
6            tenure_days    -0.009326
22      age_group_Senior    -0.010524
24  watch_hours_bin_High    -0.010596
2      complaints_raised    -0.017289
11       promo_low_watch    -0.039961
```

- **Performance:**
  - **R²:** 0.894 *(model explains ~89% of variation in watch time)*
  - **RMSE:** 0.093 *(predictions are very close to actual values)*
  - **MAE:** ~0.07 *(low average absolute error)*
- **Residual Plot Interpretation:**

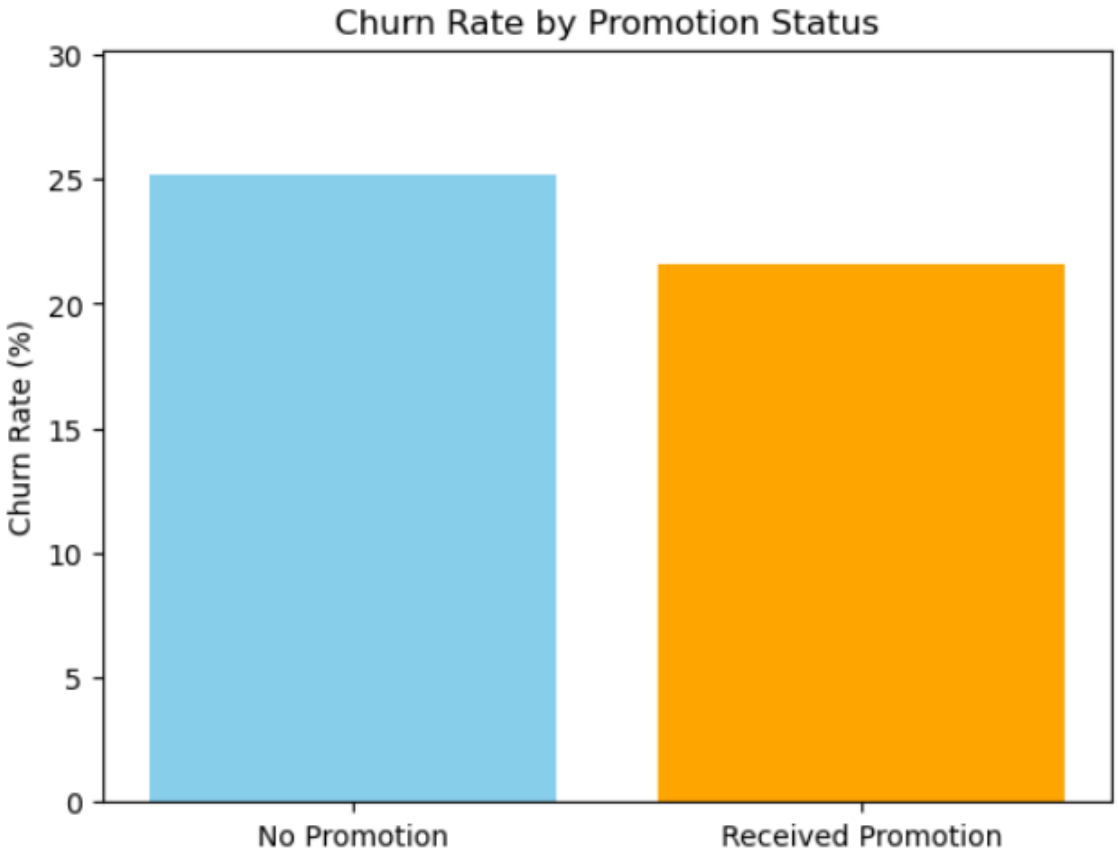
- Errors are centered around zero with even spread → **no major bias**.
- Homoscedasticity is observed (consistent error variance across predictions).
- Very few outliers → **model generalises well**.

- **Top 3 Predictors of Watch Time:**
  1. **watch_per_fee_ratio** (Positive impact) → Strongest driver; better perceived value increases watch time.
  2. **monthly_fee** (Positive impact) → Higher-paying, premium users tend to watch more.
  3. **subscription_type_2** (Positive impact) → Mid-tier subscription plan linked to higher watch time.



2. Business Questions to Answer

1. Do users who receive promotions churn less?

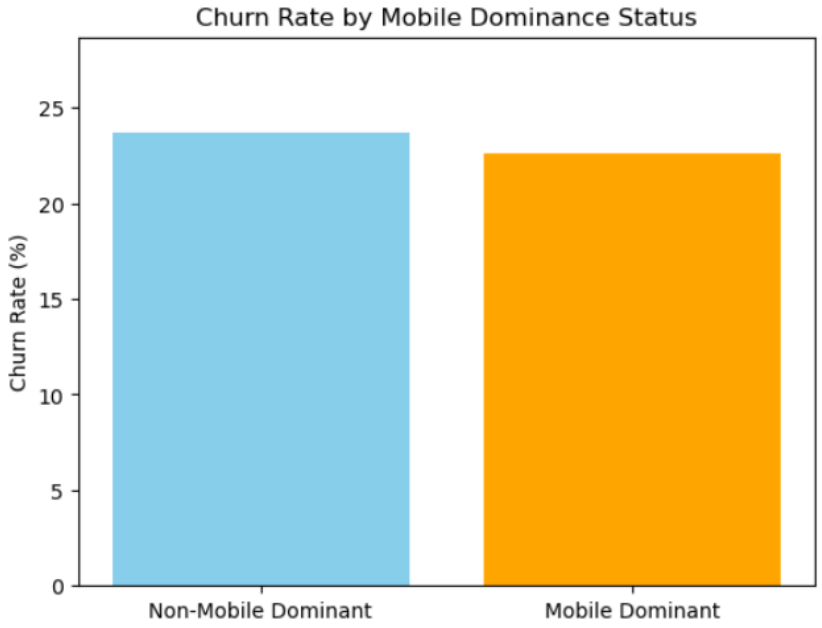


This indicates that users who received promotions have a **lower churn rate by about 3.6%** compared to those who didn't receive any promotions.
Promotions appear to be **effective in reducing churn**, as users who got promotions are less likely to leave.

2. Does watch time impact churn likelihood?

- **Yes.** The watch_per_fee_ratio was the **strongest negative predictor** of churn.
- Users who get more content hours for their money are **less likely to churn**.

3. Are mobile dominant users more likely to cancel?




```
Churn rates (%):
heavy_mobile_user
0    23.688969
1    22.646310
Name: count, dtype: float64
Z-test statistic: 0.419
P-value: 0.675
The difference in churn rates is NOT statistically significant.
```

The above bar chart and statistical test result suggest that Mobile dominance does **not appear to be a meaningful factor** in predicting churn behaviour

4. What are the top 3 features influencing churn based on your model?

1. **watch_per_fee_ratio** – Higher value perception reduces churn.
2. **age_group_Middle_Aged** – Middle-aged users more likely to churn.
3. **subscription_type_0 (Unknown)** – Associated with lower churn, but likely due to data quirks.

5.Which customer segments should the retention team prioritise?

- **Middle-aged users** with **low watch_per_fee_ratio**.
- Users with **low watch hours despite receiving promotions**.
- Customers in **early tenure (< 1 year)** who are not yet loyal.

6. What factors affect user watch time or tenure? (Linear regression insight)

Positive drivers:

- **watch_per_fee_ratio** (strongest) → better perceived value boosts watch time.

- **monthly_fee** → premium subscribers watch more.
- **subscription_type_2(Standard)**→ mid-tier plan linked to higher watch time.

Negative drivers:

- **promo_low_watch** → promotions to low-watch users often fail to boost engagement.
- **complaints_raised** → more complaints linked to slightly lower watch time.
- **age_group_Senior** → seniors watch slightly less.

7. Recommendations

1. **Target Low-Value Perception Users**
   - Focus retention efforts on customers with **low watch_per_fee_ratio**.
   - Offer personalised content recommendations or bundle content to improve perceived value.
2. **Improve Promotion Targeting**
   - Avoid sending promotions to already low-engaged users (promo_low_watch segment).
   - Instead, target moderately engaged users to convert them into heavy users.
3. **Re-engage At-Risk Middle-Aged Users**
   - Design engagement campaigns specifically for the middle-aged segment, as they have a higher churn likelihood.
   - Could include loyalty rewards, exclusive content, or subscription discounts.

8. Data Issues or Risks.

1. **Feature Leakage Risk**
   - Certain features like tenure_days or last_active_date may directly capture churn outcome if not carefully handled during training.
   - Mitigation: Ensure training features are based only on data available *before* churn.
2. **Data Quality Concerns**
   - subscription_type_0 (Unknown) showed significant influence but likely due to **missing or miscoded data**.
   - Country and date fields may also have small gaps requiring cleaning before production modelling