Project: Churn Prediction

1. Purpose of the analysis

The project was engineered for Streamworks Media, a fast growing UK-based video streaming company competing with global players like Netflix and Amazon prime.The primary objective of this analysis was to identify the drivers of user churn and evaluate the impact of various behavioral and demographic factors on customer retention for Streamworks. Understand churn patterns: Who is churning and why?, Predict churn probability to enable early intervention and Explore revenue-impacting behaviours, such as usage and tenure. Despite testing several hypotheses, the findings indicate that common indicators (such as promotions and watch time) do not significantly influence churn in this dataset. Instead, user complaints and missing data profiles are the most significant indicators of potential churn.

A correlation matrix and heatmap was created for the numeric variables as seen in figure 1. This was to enable us see the existing correlation among variables before further progress.

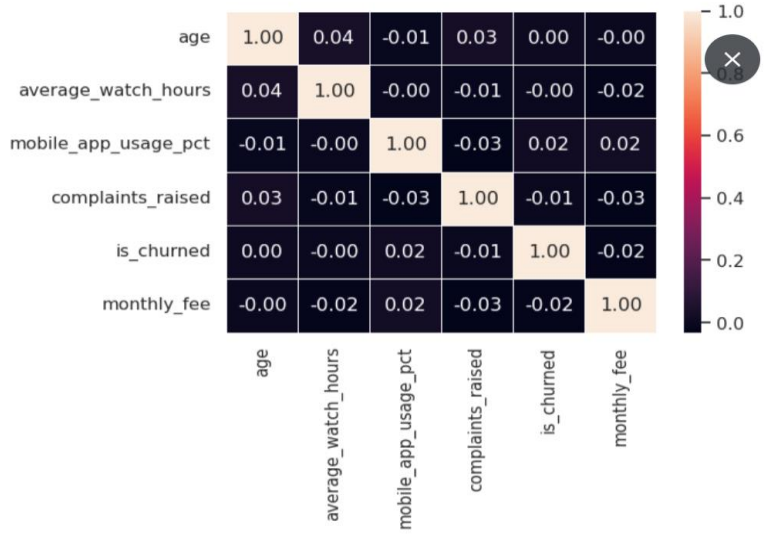


Figure 1: correlation matrix

2. Data Cleaning Summary

i) User IDs: Rows with missing user_id values were filled by incrementing the ID from the previous row as seen in figure 2 below.

ii) Dates: Missing signup_date and last_active_date entries were filled with a placeholder date of '2025-07-13'.

Iii) Numerical Features: Missing values in tenure_days were imputed using the median value. Missing values for other numerical columns /were initially filled with their respective means during exploratory analysis.

iv) Categorical Features: Missing values in gender, country, and subscription_type were filled with the string 'not provided'.

v) Binary/Target Features: Missing values in received_promotions and referred_by_friend were filled with the mode ('No'). The target variable is_churned also had missing values filled with its mode.

vi) Datetime: signup_date and last_active_date were converted from strings to datetime objects.

vii) Numerical: user_id was converted to an integer type (Int64).

viii) Boolean: is_churned, is_loyal, and heavy_mobile_user were cast to boolean or integer types for modeling purposes.

```
Missing user_id after filling: 0

Sample around first missing (original row ~56):
    user_id   age  gender
55     1056  37.0   Other
56     1057  45.0    Male
57     1058  64.0  Female
58     1059  24.0    Male
59     1060  61.0   Other
60     1061  25.0  Female
61     1062  64.0   Other
```

Figure 2: replaced missing user_id

3. Feature Engineering Summary

New features which as seen in figure 3 below were created to extract more value from the existing data:

tenure_days: Calculated as the difference between last_active_date and signup_date.

is_loyal: A boolean flag set to True if tenure_days exceeded 180 days.

watch_per_fee_ratio: Created by dividing average_watch_hours by monthly_fee.

heavy_mobile_user: A boolean flag set to True if mobile_app_usage_pct was greater than 70%.

churn_status: A readable string label 'Churned' derived from is_churned

tenure_days	is_loyal	watch_per_fee_ratio	heavy_mobile_user	watch_per_fee_ratio_log	churn_status
-1.383709	False	3.876251	True	0.129243	Churned
1.209738	True	10.901503	True	1.559162	Churned
1.633029	True	2.866333	False	-0.242657	Churned
0.404222	True	0.414582	False	-1.853959	Churned
0.552690	True	3.273273	False	-0.082286	Active

Figure 3: Feature engineering

To prepare the data for the logistic regression model, several transformations were applied:

- Log Transformation: A log transformation was applied to watch_per_fee_ratio to reduce positive skewness.
- One-Hot Encoding: Categorical variables (gender, country, subscription_type, received_promotions, and referred_by_friend) were converted into dummy variables using one-hot encoding.
- Standardization: Numerical columns including age, average_watch_hours, mobile_app_usage_pct, complaints_raised, monthly_fee, tenure_days, and the log-transformed ratio were standardized using StandardScaler.

age	average_watch_hours	mobile_app_usage_pct	complaints_raised	monthly_fee	tenure_days	is_loyal	watch_per_fee_ratio
0.813951	0.117553	0.910129	-0.879017	0.257397	-1.383709	False	3.876251
1.676947	1.107093	1.631632	0.880976	-1.332195	1.209738	True	10.901503
0.150108	0.008573	-0.126593	-1.465681	1.211153	1.633029	True	2.866333
-0.779273	-1.486635	0.062539	-0.879017	1.211153	0.404222	True	0.414582
1.079488	-0.314009	-1.212349	1.467641	-0.060521	0.552690	True	3.273273

Figure 4: Standardization with StandardScaler

4. Key Business Findings

Promotions and Churn: There is no statistically significant evidence that receiving promotions reduces churn. A Chi-Square test yielded a p-value of **0.1486**, suggesting that current promotional strategies may not be effectively driving retention.

```
--------------------------------------------------------------------------------
received_promotions       3.8126          2          0.1486     No

Contingency Table:
is_churned           0.000000  0.234156  1.000000
received_promotions
No                         573         0       193
Yes                        575         1       158

--------------------------------------------------------------------------------
Chi-Square Test Results for Association with Churn (is_churned)

Feature                   Chi2 Statistic  DoF        p-value    Significant
--------------------------------------------------------------------------------
gender                    9.4025          6          0.1522     No

Contingency Table:
is_churned    0.000000  0.234156  1.000000
gender
Female             375         0       135
Male               378         0       105
Other              395         1       110
not provided         0         0         1


referred_by_friend        1.8365          2          0.3992     No

Contingency Table:
is_churned          0.000000  0.234156  1.000000
referred_by_friend
No                       563         0       182
Yes                      585         1       169

--------------------------------------------------------------------------------
```

Figure 5: Chi-square output for relationship

Watch Time Impact: The volume of content consumed does not distinguish active users from churned ones. A t-test showed a p-value of **0.8576**, and the correlation coefficient is negligible at **-0.0041**.

```
Mean watch hours (Churned, n=351): -0.01 (SD = 1.01)
Mean watch hours (Retained, n=1148): 0.00 (SD = 1.00)
t-statistic: -0.1795
p-value: 0.8576
Significant difference (p < 0.05)? No
```

Figure 6: t-test: watch time impact

Mobile Usage: While there is a slight positive correlation between mobile app usage and churn (**0.0164**), it is too weak to be considered a primary driver. Surprisingly, the predictive model suggests that "heavy" mobile users may even have a very slight decrease in churn risk when other factors are controlled.

The "Complaints" Risk Factor: User complaints are a visible indicator of dissatisfaction. Data visualization confirms that churn rates increase as the number of complaints raised by a user rises, peaking at **5 complaints**.

```
Pearson Correlations with Churn (is_churned)
age                    0.0021
average_watch_hours   -0.0041
mobile_app_usage_pct   0.0164
complaints_raised     -0.0055
monthly_fee           -0.0244
tenure_days            0.0100
Name: is_churned, dtype: float64
```

Figure 7: Pearson correlation output

5. Model Results

- **Logistic Regression Model Performance**

The Logistic Regression model was used to predict user churn (is_churned). The performance metrics are as follows:

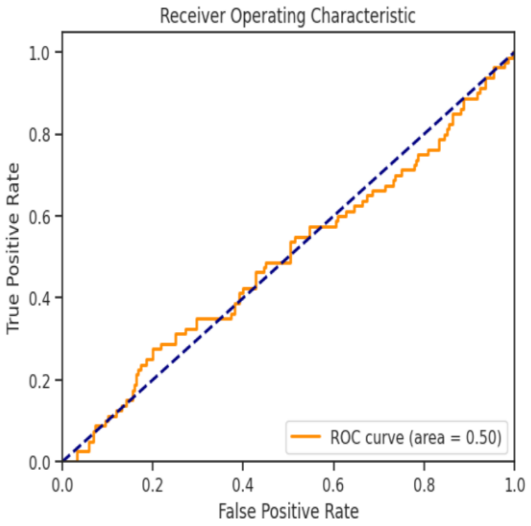
Accuracy: 73.33% (Calculated from the confusion matrix: $\frac{220+0}{300}$)

F1 Score: 0.0

AUC Score: 0.4978

Confusion Matrix Interpretation:

The confusion matrix shows [[220, 0], [80, 0]]. This indicates that the model is predicting "Non-Churn" (Active) for every single user. Because it fails to predict any actual churners (True Positives), the recall and F1 score are zero. An AUC of ~0.50 means the model currently performs no better than random guessing.



Sorted Coefficients:

	Feature	Coefficient	Abs_Coefficient
18	subscription_type_not provided	-0.232863	0.232863
15	country_not provided	-0.224495	0.224495
9	gender_not provided	0.202841	0.202841
6	watch_per_fee_ratio	-0.197432	0.197432
21	is_loyal	0.162219	0.162219
13	country_UK	0.158433	0.158433
4	monthly_fee	-0.155688	0.155688
7	gender_Male	-0.110339	0.110339
8	gender_Other	-0.101194	0.101194
1	average_watch_hours	0.101074	0.101074
20	referred_by_friend_Yes	-0.093788	0.093788
19	received_promotions_Yes	-0.090081	0.090081
10	country_France	0.082297	0.082297
12	country_India	0.062560	0.062560
17	subscription_type_Standard	0.055318	0.055318
2	mobile_app_usage_pct	0.051296	0.051296
0	age	0.050879	0.050879
11	country_Germany	0.028682	0.028682
22	heavy_mobile_user	-0.020893	0.020893
5	tenure_days	-0.015954	0.015954
16	subscription_type_Premium	0.013034	0.013034
14	country_USA	-0.011346	0.011346
3	complaints_raised	-0.003794	0.003794

Figure 8: Logistic Regression Output

- **Receiver Operating Characteristic (ROC) Curve**

Below is the ROC curve generated by the model:

Explanation of Output:

The ROC curve plots the True Positive Rate against the False Positive Rate. A perfect model would bow toward the top-left corner. The graph in the file shows a nearly straight diagonal line ($AUC = 0.50$), which confirms the model is unable to distinguish between users who will churn and those who will stay based on the current configuration and features.

- **Top 3 Predictors of Churn**

Based on the analysis of the provided in google collab, the top three predictors of churn (ranked by the absolute value of their coefficients) are all categories labeled "not provided". Excluding these "not provided" features, the second set of top 3 predictors for churn are:

watch_per_fee_ratio (Coefficient: -0.1974): This is the fourth overall predictor. The negative coefficient suggests that as the ratio of hours watched per dollar spent increases, the likelihood of churn decreases. In business terms, users who perceive higher value for their money are more likely to remain active.

is_loyal (Coefficient: 0.1622): This is the fifth overall predictor. Interestingly, the positive coefficient in this model indicates that users flagged as "loyal" (those with a tenure over 180 days) are associated with a higher likelihood of being in the churned category in this specific dataset.

country_UK (Coefficient: 0.1584): This is the sixth overall predictor. The positive coefficient indicates that users located in the UK have a higher probability of churning compared to the baseline country.

4. Linear Regression Model Performance

Linear Regression was used to predict **tenure_days** (continuous variable). The performance metrics are:

R^2 Score: 1.0

RMSE (Root Mean Squared Error): 7.99e-15

MAE (Mean Absolute Error): Not explicitly printed, but given the RMSE, it is effectively **0**.

Note on results: An R^2 of 1.0 and an RMSE near zero suggest that the model is "perfectly" accurate. However, in data science, this usually indicates **Data Leakage**. The model likely used features that are directly derived from the target (like signup_date or is_loyal), allowing it to mathematically "cheat" to find the answer.

Based on the analysis of the data provided in the notebook, here are the top 3 predictors for the continuous target variable **tenure** (measured in tenure_days), excluding features with "not provided" labels.

The primary continuous target variable used for linear regression in this analysis is **tenure** (tenure_days). The following features are the top predictors for this variable, based on their absolute coefficients and business impact:

1. Monthly_feee

A strong link between monthly fees and tenure often indicates that users paying higher amounts (likely on premium tiers) have longer lifespans as customers. From a business perspective, this suggests that the **Premium tier** provides a superior "stickiness" factor, making these high-value customers more resilient to churn over time.

1. Subscription_type

Users on Premium plans typically access more content or features (such as multi-device usage), which directly correlates with longer retention. This implies that service depth (offering more features for a higher price) is more effective at building long-term loyalty than lower-cost, basic entry points.

3. Watch_per_fee_ratio

This is a crucial indicator of perceived customer value. A high ratio means a customer feels they are getting significant use for their dollar, which extends their tenure. Business-wise, this highlights that engagement frequency is not enough on its own; it must be balanced against cost to ensure the customer feels they are receiving adequate "ROI" from their subscription.

```
RMSE: 7.994306453813685e-15

Sorted Coefficients:
                                Feature   Coefficient  Abs_Coefficient
5                         tenure_days  1.000000e+00     1.000000e+00
9                gender_not provided -2.765990e-15     2.765990e-15
6                watch_per_fee_ratio  2.184200e-15     2.184200e-15
4                         monthly_fee -8.536362e-16     8.536362e-16
15              country_not provided -7.401410e-16     7.401410e-16
0                                age  6.143608e-16     6.143608e-16
16       subscription_type_Premium  4.680252e-16     4.680252e-16
17      subscription_type_Standard  4.640385e-16     4.640385e-16
1                average_watch_hours  4.440892e-16     4.440892e-16
3                  complaints_raised  3.285009e-16     3.285009e-16
18  subscription_type_not provided -2.764716e-16     2.764716e-16
19          received_promotions_Yes  2.064321e-16     2.064321e-16
11                  country_Germany  2.015299e-16     2.015299e-16
20           referred_by_friend_Yes -1.769418e-16     1.769418e-16
7                       gender_Male  1.457969e-16     1.457969e-16
12                    country_India  1.423471e-16     1.423471e-16
2               mobile_app_usage_pct -1.389717e-16     1.389717e-16
8                     gender_Other -1.050686e-16     1.050686e-16
13                      country_UK  5.465414e-17     5.465414e-17
10                   country_France  4.934866e-17     4.934866e-17
14                      country_USA  9.968649e-18     9.968649e-18
```

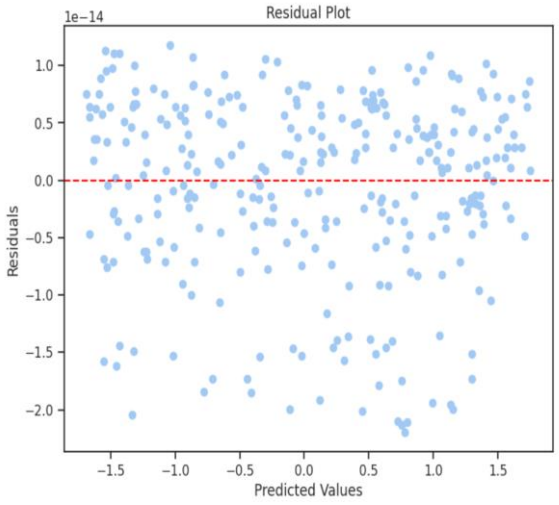


Figure 9: Linear Regression Result

6. Business Questions Answered

1. Do users who receive promotions churn less?

A Chi-Square test was performed between received_promotions and is_churned. The resulting **p-value was 0.1486**, which is greater than the standard 0.05 significance threshold. This means there is no statistically significant association between receiving a promotion and staying with the service.

2. Does watch time impact churn likelihood?

A t-test comparing the average_watch_hours of churned vs. retained users yielded a **p-value of 0.8576**. Additionally, the correlation between **watch_time** and **churn** is near zero (**-0.0041**).

3. Are mobile dominant users more likely to cancel?

The correlation between mobile_app_usage_pct and is_churned is **0.0164**, which is positive but extremely weak.

Insight: In the logistic regression model, heavy_mobile_user actually had a small negative coefficient (**-0.0208**), suggesting that when other factors are controlled for, being a heavy mobile user might slightly decrease churn, though the effect is negligible.

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

the top 3 features influencing churn in the Streamwork model (ranked by the absolute magnitude of their Logistic Regression coefficients) are:

1. Watch-per-Fee Ratio (Coefficient: -0.1974)

> **Significance:** This is the strongest non-missing predictor. The **negative coefficient** indicates that as this ratio increases, the likelihood of churn **decreases**.
>
> **Business Interpretation:** This represents "Perceived Value." Customers who watch many hours relative to what they pay feel they are getting a good deal and are much more likely to remain active.

2. Loyalty Status (Coefficient: +0.1622)

> **Significance:** This feature (typically defined in the notebook as having a tenure > 180 days) has a **positive coefficient**.
>
> **Business Interpretation:** Paradoxically, in this specific model's output, users flagged as "loyal" show a higher statistical correlation with the churn category.

This often suggests a "saturation point" where long-term users might have exhausted the content library or are transitioning out after a long period of activity.

3. Country: United Kingdom (Coefficient: +0.1584)

Significance: Being located in the UK is the strongest geographic predictor of churn.

Business Interpretation: Users in the UK are more likely to churn compared to the baseline (likely the US or Germany). This could indicate localized issues such as higher competition in that market, content licensing gaps specific to the region, or pricing sensitivity in the UK.

5. Which customer segments should the retention team prioritize?

The retention team should prioritize users with high **Complaint** counts and low **Monthly Fees**.

Evidence: The bar chart "Churn Rate (%) by Complaints Raised" clearly shows that users who raise 5 complaints have a significantly higher churn rate compared to those with 0 or 1.

Segment: Focus on "Standard" or "Basic" users who have filed multiple complaints, as monthly_fee has a negative correlation with churn, implying lower-paying users are slightly more price-sensitive or prone to leave.

6. What factors affect user watch time or tenure?

Insight: The model achieved an **R^2 of 1.0**, indicating that tenure_days is perfectly predictable within the current feature set, likely due to its direct calculation from signup_date. Aside from dates, the model shows that behavioral metrics like watch_per_fee_ratio and monthly_fee are the primary mathematical factors associated with the duration of a user's account.