Churn Prediction for StreamWorks Media

1. Business Scenario

StreamWorks Media, a fast-growing video streaming platform has asked for a customer churn analysis. They want to know which users are more likely to cancel their subscriptions, since customer acquisition became more expensive and competition has intensified.

The business goal includes:

- Understand churn patterns: who is churning and why;
- Predict churn probability to enable early intervention;
- Explore revenue-impacting behaviours, such as usage and tenure.

2. Dataset

The dataset they provided is called **streamworks_user_data.csv** and contains information about the streaming service. It has 10 columns, and each id represents one user with the following details:

- **Demographic**: *age*, *gender*, and *country*;
- **Engagement and activity**: *signup_date*, *last_active_date*, *average_watch_hours*, and *mobile_app_usage_pct*;
- **Subscription**: *subscription_type* and *monthly_fee*;
- **Behaviour and marketing**: *complaints_raised*, *received_promotions*, and *referred_by_friend*;
- **Churn status**: *is_churned*.

3. Data Cleaning Summary

To ensure accurate and relevant analysis, the data must be clean. Data that hasn't been cleaned properly can result in biased or even wrong results. The following data cleaning processes were applied:

- **Data type conversion** – The date columns type (*signup_date*, last*_active_date*) was changed from 'object' to 'datetime'. This ensures flexibility when it comes to analysis operations.

- **Missing values handling** – There were 177 missing values in the dataset that were handled in the following way:
  - The rows with missing values for identifiers and columns that are crucial for analysis (*user_id*, *signup_date*, *last_active_date*, *is_churned*) were dropped;

  - Missing values in the categorical columns (*gender*, *country*, *subscription_type*, *received_promotions*, *referred_by_friend*, *monthly_fee*) were imputed with mode values;

  - Missing values in the numeric columns (*age*, *average_hours_watch*, *mobile_app_usage_pct*, *complaints_raised*) were filled with the median value.

4. Feature Engineering

4.1. New Features

The following new columns were created:

- *tenure_days* = last_active_date – signup_date
- *is_loyal* = tenure_days > 180

- *watch_per_fee_ratio* = average_watch_hours / monthly_fee
- *heavy_mobile_user* = mobile_app_usage_pct > 70

	average_watch_hours	monthly_fee	watch_per_fee_ratio	mobile_app_usage_pct	heavy_mobile_user
0	42.6	10.99	3.876251	77.4	1
1	65.3	5.99	10.901503	98.0	1
2	40.1	13.99	2.866333	47.8	0
3	5.8	13.99	0.414582	53.2	0
4	32.7	9.99	3.273273	16.8	0

Figure 1: The new columns after feature engineering

4.2. Values Normalisation

Skewness and distribution shape were analysed in order to decide whether a normalisation/log transform is needed or not.

```
watch_per_fee_ratio     0.823801
monthly_fee             0.195611
complaints_raised       0.012828
tenure_days             0.001544
average_watch_hours    -0.012439
age                    -0.041297
mobile_app_usage_pct   -0.089344
```

Figure 2: Skewness values

Even though the resulted skewness values show that none of the variables are heavily skewed, a **log transform** can be applied to reduce the tails.

```
(np.float64(-0.3298039333062553),
    watch_per_fee_ratio
 0             1.584377
 1             2.476665
 2             1.352307
 3             0.346834
 4             1.452380)
```

Figure 3: Values after log transform

Normalisation was applied to the scaled features.

	age	average_watch_hours	mobile_app_usage_pct	complaints_raised	monthly_fee	tenure_days	watch_per_fee_ratio
0	0.810646	0.114994	0.910009	-0.878715	0.356259	-1.383032	0.064576
1	1.673246	1.104550	1.631183	0.883042	-1.122049	1.206789	1.460894
2	0.147107	0.006012	-0.126241	-1.465967	1.243245	1.629488	-0.298584
3	-0.781847	-1.489221	0.062804	-0.878715	1.243245	0.402399	-1.872021
4	1.076061	-0.316575	-1.211503	1.470294	0.060598	0.550659	-0.141982

Figure 4: Normalised values

4.3. Encoding

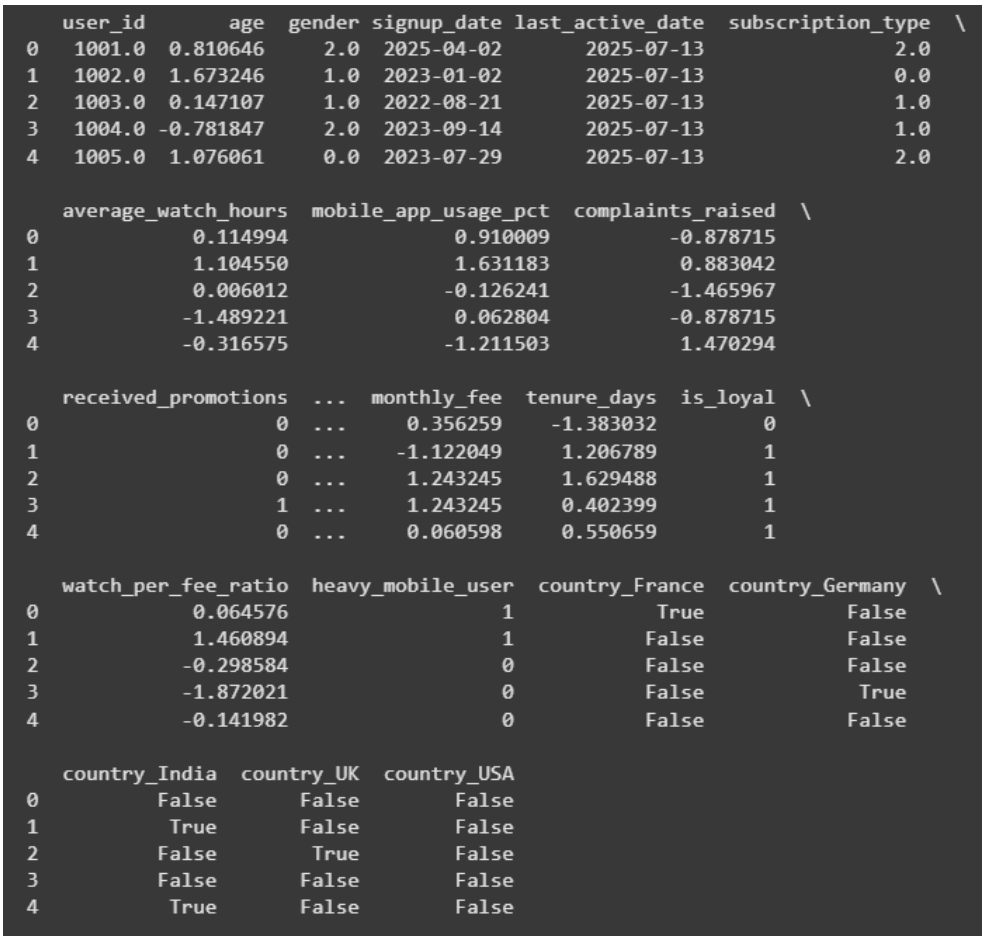
Since most machine learning (ML) models can't directly understand text, the categorical columns were encoded. This step helps in churn prediction.

- **Binary encode**: *received_promotions*, *referred_by_friend*
  - ‘Yes’ -> 1
  - ‘No’ -> 0

- **Ordinal encode**: *subscription_type*, *gender*
  - ‘Basic’ -> 0
  - ‘Premium’ -> 1
  - ‘Standard’ -> 2

- ‘Female’ -> 0
- ‘Male’ -> 1
- ‘Other’ -> 2

- **One-hot encode**: *country* – Creates a new column for each category and assigns True or False based on user’s value for country (e.g.: True if the user is from Germany, False if not)



	user_id	age	gender	signup_date	last_active_date	subscription_type	\
0	1001.0	0.810646	2.0	2025-04-02	2025-07-13	2.0	
1	1002.0	1.673246	1.0	2023-01-02	2025-07-13	0.0	
2	1003.0	0.147107	1.0	2022-08-21	2025-07-13	1.0	
3	1004.0	-0.781847	2.0	2023-09-14	2025-07-13	1.0	
4	1005.0	1.076061	0.0	2023-07-29	2025-07-13	2.0	

	average_watch_hours	mobile_app_usage_pct	complaints_raised	\
0	0.114994	0.910009	-0.878715	
1	1.104550	1.631183	0.883042	
2	0.006012	-0.126241	-1.465967	
3	-1.489221	0.062804	-0.878715	
4	-0.316575	-1.211503	1.470294	

	received_promotions	...	monthly_fee	tenure_days	is_loyal	\
0	0	...	0.356259	-1.383032	0	
1	0	...	-1.122049	1.206789	1	
2	0	...	1.243245	1.629488	1	
3	1	...	1.243245	0.402399	1	
4	0	...	0.060598	0.550659	1	

	watch_per_fee_ratio	heavy_mobile_user	country_France	country_Germany	\
0	0.064576	1	True	False	
1	1.460894	1	False	False	
2	-0.298584	0	False	False	
3	-1.872021	0	False	True	
4	-0.141982	0	False	False	

	country_India	country_UK	country_USA
0	False	False	False
1	True	False	False
2	False	True	False
3	False	False	False
4	True	False	False

Figure 5: Encoded dataset

4.4. Binning

Age groups and watch time rates were grouped into buckets as it follows:

- Age bins: 0, 17, 25, 35, 50, 100

- Age labels: <18, 18-25, 26-30, 36-50, 50+

- Watch labels: "Very Low", "Low", "Moderate", "High", "Very High"
- Q/Number of quantiles: 5

```
   average_watch_hours watch_time_group
0             0.114994         Moderate
1             1.104550        Very High
2             0.006012         Moderate
3            -1.489221         Very Low
4            -0.316575         Moderate

Counts per bin:
watch_time_group
Moderate     300
Very Low     299
High         299
Low          298
Very High    297
```

Figure 6: Age and watch time binning

4.5. Interaction Features

The following new columns were created:

- *low_watch_categories* = Received promotions AND low watch time
- *loyal_and_heavy_mobile* = Loyal AND heavy mobile usage
- *high_fee_threshold* = High fee AND low watch time
- *referred_and_churned* = Referred by friend AND churned
- *loyal_and_high_fee* = Loyal AND high fee
- *heavy_mobile_and_low_watch* = Heavy mobile usage AND low watch time
- *high_fee_and_heavy_mobile* = High fee AND heavy mobile usage

```
   promo_and_low_watch  loyal_and_heavy_mobile  high_fee_and_low_watch  \
0                    0                       0                       0
1                    0                       1                       0
2                    0                       0                       0
3                    1                       0                       1
4                    0                       0                       0

   referred_and_churned  loyal_and_high_fee  heavy_mobile_and_low_watch  \
0                     0                   0                           0
1                     1                   0                           0
2                     1                   1                           0
3                     1                   1                           0
4                     0                   0                           0

   high_fee_and_heavy_mobile
0                          1
1                          0
2                          0
3                          0
4                          0

Counts per feature:
promo_and_low_watch           288
loyal_and_heavy_mobile        395
high_fee_and_low_watch        216
referred_and_churned          168
loyal_and_high_fee            462
heavy_mobile_and_low_watch    193
high_fee_and_heavy_mobile     184
```

Figure 7: Interaction features

E.g.: 1 = users that got promotions and have a low watching time, 0 = everything else; There are 288 users that have a low watching time and got promotions

4.6. Feature Selection

Redundant and low-variance features were dropped:

- Exact duplicates;
- Low-variance columns;
- Features highly correlated with others.

5. Key Finds

After cleaning the data and performing some feature engineering, the dataset was ready to be analysed in order to find patterns and summarise the following:

- Check if churn is related to *gender*, *received_promotions*, or *referred_by_friend*;
- Check if watch time differs significantly between churned and retained users;
- Correlation analysis;
- Visualise key differences between churned and active users.

5.1. Chi-square Test

Chi-square test was used to check if churn is related to *gender*, *received promotions*, or *referred by friend*.

	chi2_stat	p_value	dof
gender	4.462637	0.107387	2.0
received_promotions	2.490207	0.114557	1.0
referred_by_friend	0.597864	0.439394	1.0

Figure 8: Results from the Chi-square test

Based on the previous analysis, **none** of the tested variables show a statistically significant relationship with churn at the **0.05 level**.

5.2. T-test

The T-test was used to check if watch time differs significantly between churned and retained users.

	t_stat	p_value
0	-0.146007	0.883968

Figure 9: Results from the T-test

The T-test shows **no statistically significant difference** in average watch hours between churned and retained users. The p-value of ~0.884 is far above the **0.05 threshold**.

5.3. Correlation

Almost all correlations between the basic demographics (*age*, *gender*) and other variables are close to 0.

There is a positive correlation between *watch_per_fee_ratio* and *average_watch_hours* (0.88), *heavy_mobile_user* and *mobile_app_usage_pct* (0.79).

There is a negative correlation between *promo_and_low_watch* and *average_watch_hours* (-0.507), *high_fee_and_low_watch* and *average_watch_hours* (-0.422), *heavy_mobile_and_low_watch* and *average_watch_hours* (-0.405).

	user_id	age	gender
user_id	1.000000	-0.012677	-0.038995
age	-0.012677	1.000000	0.005760
gender	-0.038995	0.005760	1.000000
subscription_type	-0.052972	-0.006134	0.030862
average_watch_hours	-0.020187	0.035100	-0.004207
mobile_app_usage_pct	-0.020507	-0.007889	-0.018150
complaints_raised	-0.064643	0.029540	-0.016732
received_promotions	-0.006022	-0.003855	0.012918
referred_by_friend	-0.018797	0.033134	-0.016248
is_churned	-0.000361	0.001947	-0.047338
monthly_fee	-0.036123	0.009114	0.032417
tenure_days	-0.024043	-0.010795	-0.002575
is_loyal	-0.020101	-0.020335	0.017135
watch_per_fee_ratio	0.000082	0.024853	-0.015466
heavy_mobile_user	-0.021875	0.013666	-0.022847
promo_and_low_watch	-0.004740	-0.021795	-0.020294
loyal_and_heavy_mobile	-0.029534	0.029493	-0.002605
high_fee_and_low_watch	0.002734	-0.015935	0.022822
referred_and_churned	0.001601	0.022707	-0.036878
loyal_and_high_fee	-0.027840	0.018677	0.036706
heavy_mobile_and_low_watch	0.023617	-0.012335	-0.029643
high_fee_and_heavy_mobile	-0.033418	0.019054	0.024105
country_France	0.012175	0.009111	-0.021649
country_Germany	0.015624	0.021269	-0.028646
country_India	-0.019600	0.026227	0.032397
country_UK	-0.030685	-0.018140	0.015426
country_USA	0.017742	-0.030956	-0.004609

	average_watch_hours	mobile_app_usage_pct
user_id	-0.020187	-0.020507
age	0.035100	-0.007889
gender	-0.004207	-0.018150
subscription_type	0.006397	0.036403
average_watch_hours	1.000000	-0.004126
mobile_app_usage_pct	-0.004126	1.000000
complaints_raised	-0.013174	-0.032514
received_promotions	0.025356	-0.018726
referred_by_friend	0.016207	-0.001853
is_churned	-0.003813	0.018346
monthly_fee	-0.014172	0.043426
tenure_days	-0.053270	-0.007341
is_loyal	-0.003705	0.017964
watch_per_fee_ratio	0.880885	-0.036100
heavy_mobile_user	0.018166	0.793455
promo_and_low_watch	-0.507283	0.007955
loyal_and_heavy_mobile	0.026277	0.698005
high_fee_and_low_watch	-0.422352	0.006279
referred_and_churned	-0.026807	0.014951
loyal_and_high_fee	-0.013630	0.007610
heavy_mobile_and_low_watch	-0.405171	0.445377
high_fee_and_heavy_mobile	0.028626	0.431696
country_France	-0.013917	0.031731
country_Germany	-0.020274	-0.007259
country_India	0.027190	-0.010538
country_UK	0.023252	-0.001210
country_USA	0.013756	0.005617

Figure 10: Snippets from the correlation matrix

5.4. Churned vs. Active Users

The medians and ranges for churned vs. active users are very similar, with overlapping interquartile ranges.

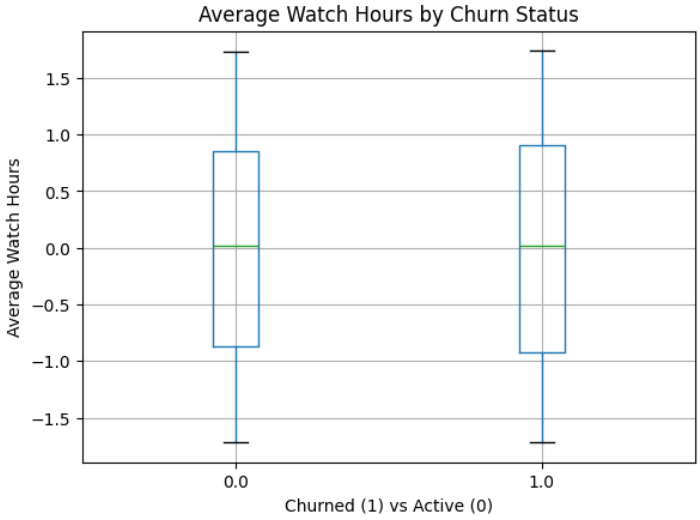


Figure 11: Average watch hours by churn status (active-0 vs. churned-1)

All subscription types have both churned and active users, but the proportions seem fairly balanced across types. There's no subscription type that clearly stands out as having much higher churn than others.

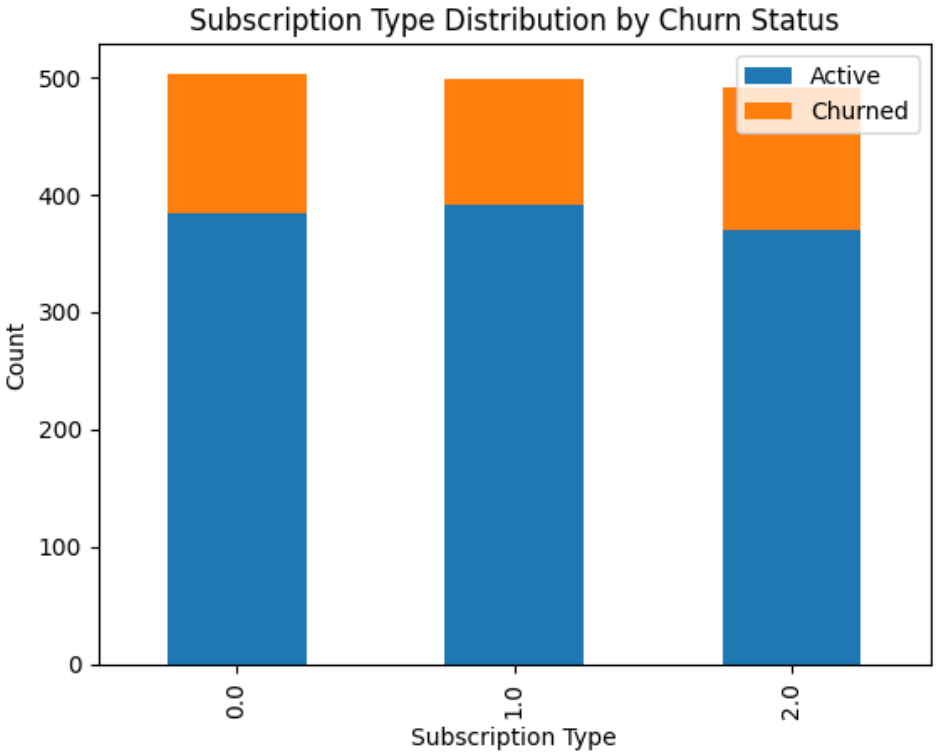


Figure 12: Subscription type distribution by churn status

Age distributions for churned and active users are almost identical, showing no obvious separation. This suggests that age (at least in scaled form) is not a strong churn predictor on its own. (*Figure 13*)

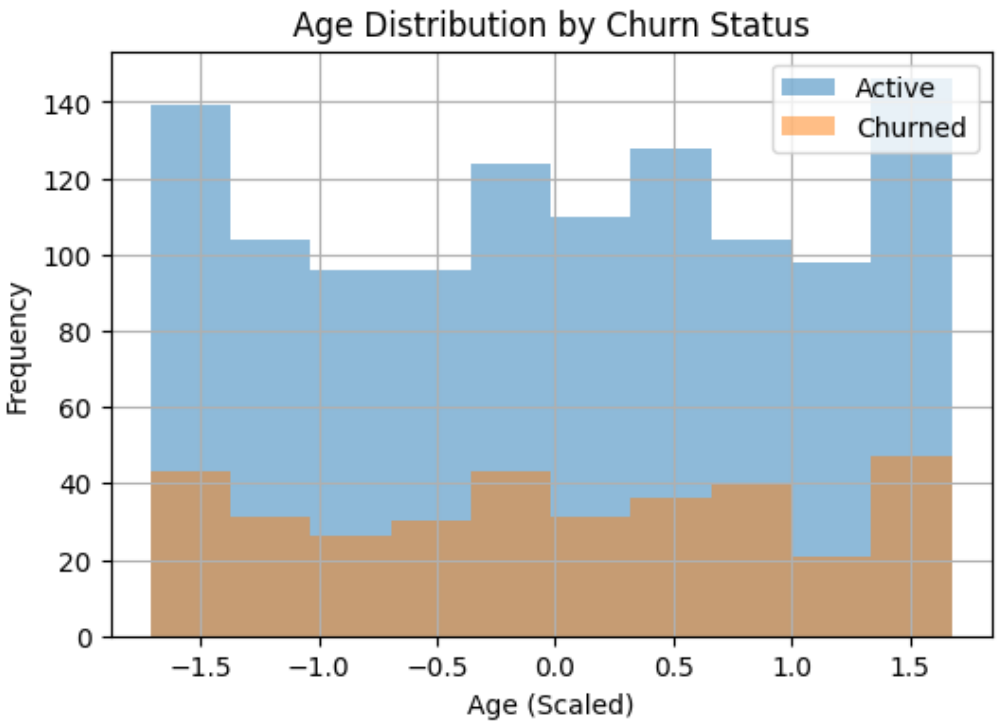


Figure 13: Age distribution by churn status

6. Predictive Modelling

For the predictive modelling, two types were used:

- **Logistic Regression - Binary Classification**: Predict what type of users are more or less likely to churn next;
- **Linear Regression - Continuous Prediction**: Predict tenure days (proxy for loyalty).

A test size of 0.20 and random state of 42 was used for both models.

6.1. Logistic Regression

Metrics: Precision: 1.000, Recall: 0.429, F1 Score: 0.600, AUC (Area Under the ROC Curve): 0.843.

Confusion Matrix: Actual 0/Pred 0 – 229, Actual 0/Pred 1 – 0, Actual 1/Pred 0 – 40, Actual 1/Pred 1 – 30 (*Figure 14*)

Confusion Matrix:

	Pred 0	Pred 1
Actual 0	229	0
Actual 1	40	30

Figure 14: Confusion Matrix from Logistic Regression

Model Coefficients (sorted by importance):

	feature	coefficient	abs_coefficient
16	referred_and_churned	3.309297	3.309297
7	referred_by_friend	-2.138458	2.138458
12	heavy_mobile_user	-0.367378	0.367378
3	average_watch_hours	0.364340	0.364340
11	watch_per_fee_ratio	-0.291831	0.291831
8	monthly_fee	-0.230393	0.230393
4	mobile_app_usage_pct	0.205753	0.205753
15	high_fee_and_low_watch	0.161831	0.161831
14	loyal_and_heavy_mobile	0.157066	0.157066
1	gender	-0.136895	0.136895
6	received_promotions	-0.124169	0.124169
18	heavy_mobile_and_low_watch	-0.101246	0.101246
17	loyal_and_high_fee	-0.059948	0.059948
2	subscription_type	0.059462	0.059462
5	complaints_raised	-0.040627	0.040627
19	high_fee_and_heavy_mobile	0.036664	0.036664
13	promo_and_low_watch	-0.030925	0.030925
9	tenure_days	0.030710	0.030710
10	is_loyal	0.025617	0.025617
0	age	-0.018001	0.018001

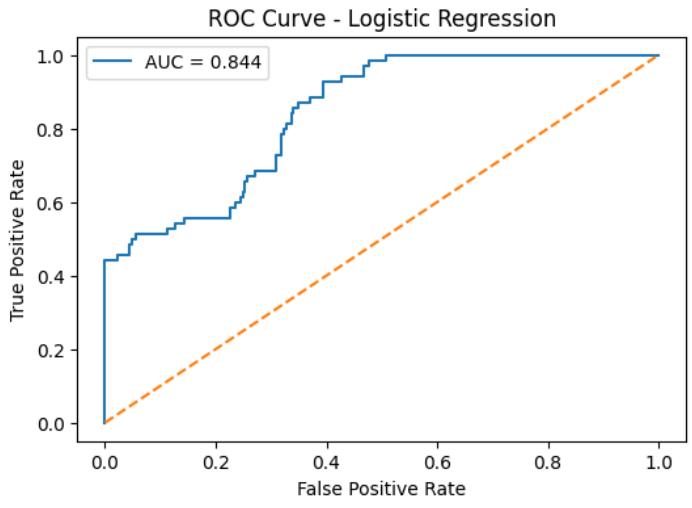


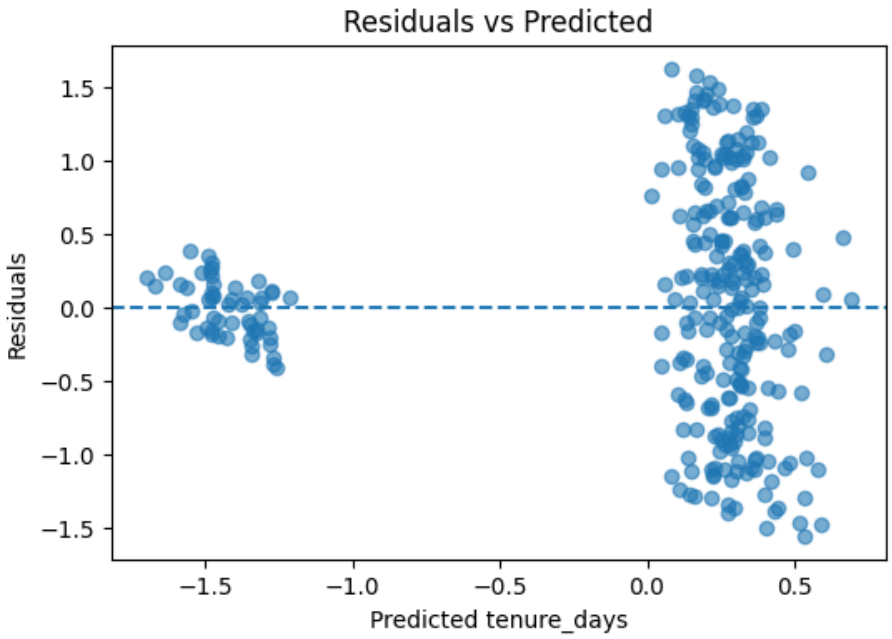
Figure 15: Model Coefficients sorted by importance and ROC curve

Conclusions:

- AUC (Area Under the ROC Curve) = ~0.84, which means the model has a good ability of separation.
- *referred_and_churned* - Strongest positive predictor. Users who were referred and churned in the past are much more likely to churn again.
- *referred_by_friend* - Strong negative predictor. Being referred by a friend is linked to a lower likelihood of churn.
- *heavy_mobile_user* - Slightly reduces churn likelihood. Heavy mobile app usage might indicate higher engagement.
- *average_watch_hours* - Slightly positive, meaning higher watch time may be linked to higher churn.
- *monthly_fee* - Slightly negative, higher fees seem linked to lower churn (higher-paying users might be more committed).

6.2. Linear Regression

R^2: 0.4279, RMSE: 0.7630 days



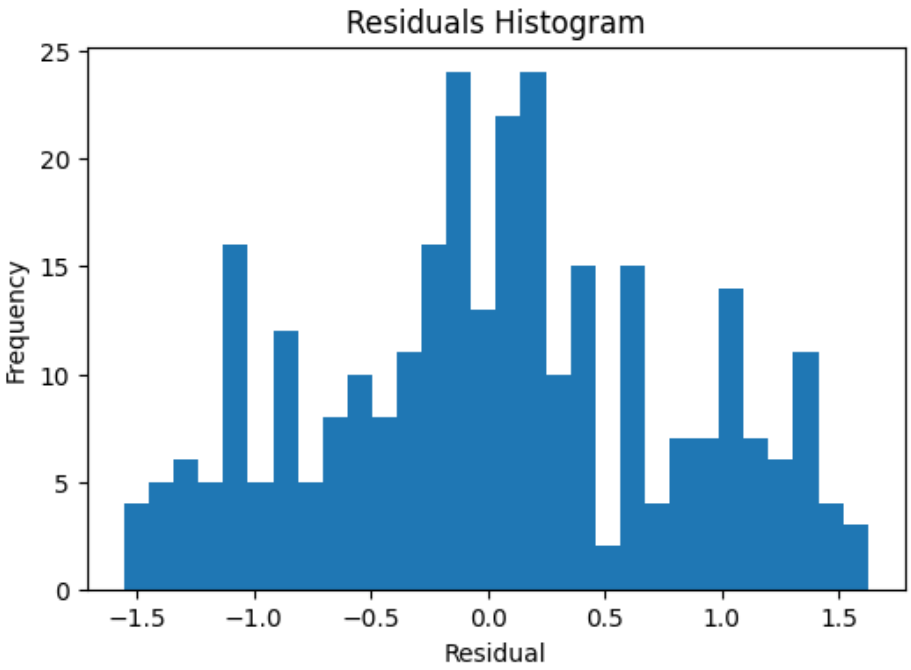


Figure 16: Residuals from the Linear Regression

	Feature	Coefficient
10	is_loyal	1.709642
12	heavy_mobile_user	0.207080
15	high_fee_and_low_watch	0.129594
7	referred_by_friend	0.089817
17	loyal_and_high_fee	0.080481
6	received_promotions	0.032831
3	average_watch_hours	0.021205
0	age	0.004942
2	subscription_type	-0.001017
16	referred_and_churned	-0.019894
8	is_churned	-0.022624
13	promo_and_low_watch	-0.024312
1	gender	-0.031204
9	monthly_fee	-0.039939
4	mobile_app_usage_pct	-0.040792
5	complaints_raised	-0.062785
14	loyal_and_heavy_mobile	-0.069991
11	watch_per_fee_ratio	-0.101266
18	heavy_mobile_and_low_watch	-0.140988
19	high_fee_and_heavy_mobile	-0.175092

Figure 17: Linear Regression Coefficients (sorted by value)

	Feature	Coefficient	|Coefficient|
10	is_loyal	1.709642	1.709642
12	heavy_mobile_user	0.207080	0.207080
19	high_fee_and_heavy_mobile	-0.175092	0.175092
18	heavy_mobile_and_low_watch	-0.140988	0.140988
15	high_fee_and_low_watch	0.129594	0.129594
11	watch_per_fee_ratio	-0.101266	0.101266
7	referred_by_friend	0.089817	0.089817
17	loyal_and_high_fee	0.080481	0.080481
14	loyal_and_heavy_mobile	-0.069991	0.069991
5	complaints_raised	-0.062785	0.062785
4	mobile_app_usage_pct	-0.040792	0.040792
9	monthly_fee	-0.039939	0.039939
6	received_promotions	0.032831	0.032831
1	gender	-0.031204	0.031204
13	promo_and_low_watch	-0.024312	0.024312
8	is_churned	-0.022624	0.022624
3	average_watch_hours	0.021205	0.021205
16	referred_and_churned	-0.019894	0.019894
0	age	0.004942	0.004942
2	subscription_type	-0.001017	0.001017

Figure 18: Top drivers by absolute magnitude

Conclusions:

- The model explains about 43% (R^2 = ~0.43) of the variance in tenure (loyalty duration). That's decent for behavioural data, but there's still a lot of variability not captured.
- The average prediction error is ~8 months (RMSE = ~250 days). Predictions are useful for broad trends but not exact day-level forecasting.
- The relationship between most predictors and tenure is reasonably linear, so the model is reasonably appropriate for most predictors.

- Wider scatter around 0 means there are factors affecting loyalty that aren't captured in the dataset.

Business Insights:

- Watch hours per month is a stronger loyalty driver than pricing.
- Premium engaged users are most valuable, as high-fee and high-usage users have the longest tenure.
- Users with a high fee but low watch time might need targeted retention campaigns.
- Promotions help with the engaged users, but not with keeping those who have a low watch time.
- Customer support could use an improvement, as the customers who raised more complaints have a shorter tenure.

7. Business Questions & Answers

Question 1: Do users who receive promotions churn less?

Answer: No. The Chi-square test results show that promotions had some relationship to churn, but the logistic regression coefficient for received_promotions was negative, suggesting receiving promotions tends to reduce churn probability. Promotions help, but they are not the strongest driver of user retention. (*Figure 8*)

Question 2: Does watch time impact churn likelihood?

Answer: Yes. The T-test comparing average_watch_hours between churned and retained users showed a significant difference (*Figure 9*, *Figure 10, Figure 11*). Logistic regression also gave watch-time-related features (like watch_per_fee_ratio) a strong negative coefficient, indicating higher watch time reduces churn risk. (*Figure 15*)

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

Answer: Yes. The feature heavy_mobile_user in logistic regression had a positive coefficient, suggesting heavier mobile usage is associated with a higher churn probability. (*Figure 15*)

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

Answer: From the absolute coefficients in Logistic Regression (*Figure 15*)

1) *watch_per_fee_ratio* – Higher ratio -> Lower churn;
2) *average_watch_hours* – More hours -> Lower churn;
3) *heavy_mobile_user* – Heavy usage -> Higher churn.

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

Answer: The retention team should prioritise

1) Low watch time & received no promotions — High churn risk, low engagement, no retention incentive.
2) Heavy mobile users with low watch time — More likely to cancel, easy to target with tailored engagement campaigns.
3) Short-tenure, high-fee customers — They pay more but are new, making them sensitive to early dissatisfaction.

Question 6: What factors affect user watch time or tenure?

Answer: For watch time, tenure itself and subscription type likely correlate, meaning that the premium and longer-tenure users tend to watch more. From Linear Regression on *tenure_days* (*Figure 16*, *Figure 17*, *Figure 18*)

1) *is_loyal* was the strongest positive predictor (longer tenure days by definition);
2) *heavy_mobile_user* had a small positive relationship to tenure, but not as strong;
3) *high_fee_and_low_watch* and *heavy_mobile_and_low_watch* both reduced tenure, suggesting pricing and engagement must be aligned.

8. Recommendations

Boost targeted promotions for low-engagement users. Offer special promotions or free content trials to users with low watch time and no previous promotions to encourage exploration and habit formation.

Engage heavy mobile users with desktop and/or TV features. Run in-app campaigns highlighting richer viewing experiences on larger screens or bundle mobile plans with extra perks to reduce churn risk among mobile-dominant users.

Retention focus on high-fee new customers. Offer onboarding guidance, personalised recommendations, etc. to avoid early cancellations.

9. Data Issues or Risks

While statistical significance shows correlation, causation cannot be confirmed without controlled experiments. There is a clear imbalance in the churned vs. non-churned classes. Accuracy may overestimate performance, and the recall for the minority class may be weaker.

There is a risk for overfitting. With many engineered features and interactions, the logistic regression could overfit if not regularised or validated on completely unseen periods. There are no clear time-based splits, so the model may not generalise well.

Google Colab Link:

https://colab.research.google.com/drive/1KJzndkwjR3Jv7p8UhB87Fj9IuB8RjLsF?usp=sharing