Project Coversheet

Full Name	Bhoomi Sharma
Project Title (Example – Week1, Week2, Week3, Week 4)	Week 3 – Churn Prediction for StreamWorks Media

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.

1. Introduction

StreamWorks Media is a UK-based video streaming platform that competes with major players like Netflix and Amazon Prime. As customer acquisition costs increase, retaining existing customers has become a key business priority. This project examines customer churn, identifying patterns and predicting which users are most likely to cancel their subscriptions. The goal is to help the management team understand why users churn and how early interventions can reduce customer loss.

```
Shape: (1500, 14)
```

	user_id	age	gender	signup_date	last_active_date	country	subscription_type	average_watch_hours	mobile_app_usage_pct	complaints_raised
0	1001.0	56.0	Other	02-04-25	13-07-25	France	Standard	42.6	77.4	1.0
1	1002.0	69.0	Male	02-01-23	13-07-25	India	Basic	65.3	98.0	4.0
2	1003.0	46.0	Male	21-08-22	13-07-25	UK	Premium	40.1	47.8	0.0
3	1004.0	32.0	Other	14-09-23	13-07-25	Germany	Premium	5.8	53.2	1.0
4	1005.0	60.0	Female	29-07-23	13-07-25	India	Standard	32.7	16.8	5.0
5	1006.0	25.0	Male	25-06-23	13-07-25	USA	Premium	40.0	24.7	1.0
6	1007.0	38.0	Male	15-02-23	13-07-25	UK	Premium	57.8	83.9	0.0
7	1008.0	56.0	Male	20-12-22	13-07-25	Germany	Premium	9.0	35.6	5.0
8	1009.0	36.0	Other	30-05-25	13-07-25	UK	Standard	11.6	82.7	1.0
9	1010.0	40.0	Male	07-11-24	13-07-25	France	Basic	21.5	70.9	5.0

Figure 1: Load the Data

2. Data Cleaning Summary

The dataset comprised 1,500 user records, including details on demographics, subscription plans, engagement, and churn behavior. Initial exploration showed missing values in several columns, particularly in 'monthly_fee'. Data types were standardized — date fields were converted to datetime, categorical fields were encoded, and missing numeric values were filled with suitable averages or medians. After cleaning, all columns had zero missing values.

```
--- .info() ---
<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
None
```

Figure 2: Summary of tables(basic exploration)

```
--- Missing values per column ---

user_id                   2
age                       3
gender                    1
signup_date               2
last_active_date          2
country                   3
subscription_type         3
average_watch_hours       4
mobile_app_usage_pct      2
complaints_raised         3
received_promotions       3
referred_by_friend        3
is_churned                1
monthly_fee             145
dtype: int64
```

Figure 3: Missing values per column

```
--- value counts: gender ---
Female    510
Other     506
Male      483
NaN         1
Name: gender, dtype: int64

--- value counts: country ---
Canada     262
India      259
France     254
Germany    246
UK         241
USA        235
NaN          3
Name: country, dtype: int64

--- value counts: subscription_type ---
Basic       505
Premium     499
Standard    493
NaN           3
Name: subscription_type, dtype: int64
```

Figure 4: Count missing values on gender, country, and sub_plan

```
--- value counts: received_promotions ---
No     763
Yes    734
NaN      3
Name: received_promotions, dtype: int64

--- value counts: referred_by_friend ---
Yes    752
No     745
NaN      3
Name: referred_by_friend, dtype: int64

--- value counts: is_churned ---
0.0    1148
1.0     351
NaN       1
Name: is_churned, dtype: int64
```

Figure 5: Count missing values per received_promotion, referred_friend, and is_churned

```
After cleaning, missing values per column:

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
watch_per_fee_ratio     0
heavy_mobile_user       0
dtype: int64
```

Figure 6: After cleaning all missing values 0

3. Feature Engineering Summary

New features were created to capture user behavior and loyalty. For example, 'tenure_days' calculated the length of time users stayed active, 'is_loyal' classified users with more than 180 days of activity, and 'watch_per_fee_ratio' measured engagement relative to cost. Additionally, a 'heavy_mobile_user' flag was introduced for users with more than 70% mobile usage. Categorical variables, such as gender, country, subscription type, and promotions, were one-hot encoded for modeling.

```
--- .describe() for numeric columns ---
```

	user_id	age	average_watch_hours	mobile_app_usage_pct	complaints_raised	is_churned	monthly_fee
count	1498.000000	1497.000000	1496.000000	1498.000000	1497.000000	1499.000000	1355.000000
mean	1750.871829	43.738811	39.903342	51.414419	2.498330	0.234156	10.180406
std	433.060980	15.083920	22.978288	28.580117	1.706829	0.423612	3.310705
min	1001.000000	18.000000	0.500000	0.000000	0.000000	0.000000	5.990000
25%	1376.250000	31.000000	19.450000	27.100000	1.000000	0.000000	5.990000
50%	1750.500000	44.000000	40.300000	52.700000	2.000000	0.000000	9.990000
75%	2125.750000	56.000000	59.800000	76.200000	4.000000	0.000000	13.990000
max	2500.000000	69.000000	79.900000	100.000000	5.000000	1.000000	14.990000

Figure 7: Numeric values display

Original shape: (1500, 18) Encoded shape: (1500, 24)

	user_id	age	signup_date	last_active_date	average_watch_hours	mobile_app_usage_pct	complaints_raised	is_churned	monthly_fee	tenure_days
0	1001.0	56.0	2025-02-04	2025-07-13	42.6	77.4	1.0	1.0	10.99	159.0
1	1002.0	69.0	2023-02-01	2025-07-13	65.3	98.0	4.0	1.0	5.99	893.0
2	1003.0	46.0	2022-08-21	2025-07-13	40.1	47.8	0.0	1.0	13.99	1057.0
3	1004.0	32.0	2023-09-14	2025-07-13	5.8	53.2	1.0	1.0	13.99	668.0
4	1005.0	60.0	2023-07-29	2025-07-13	32.7	16.8	5.0	0.0	9.99	715.0

Original shape: (1500, 18) Encoded shape: (1500, 24)

is_loyal	watch_per_fee_ratio	heavy_mobile_user	gender_Male	gender_Other	country_France	country_Germany	country_India	country_UK	country_USA
0	3.876251	1	0	1	1	0	0	0	0
1	10.901503	1	1	0	0	0	1	0	0
1	2.866333	0	1	0	0	0	0	1	0
1	0.414582	0	0	1	0	1	0	0	0
1	3.273273	0	0	0	0	0	1	0	0

subscription_type_Standard	received_promotions_Yes	referred_by_friend_Yes
1	0	0
0	0	1
0	0	1
0	1	1
1	0	1

4. Statistical Analysis & Insights

Statistical tests were conducted to understand relationships between churn and user attributes. Chi-square tests revealed no strong evidence that gender, promotions, or referrals had a direct influence on churn. A t-test comparing watch time between churned and retained users found no significant difference. However, behavioral trends indicated that shorter tenure and lower engagement correlated with higher churn.

Chi-square test — gender vs is_churned: chi2=4.47, p-value=0.1071

is_churned	0.0	1.0
gender		
Female	375	136
Male	378	105
Other	396	110

Chi-square test — received_promotions vs is_churned: chi2=2.62, p-value=0.1058

is_churned	0.0	1.0
received_promotions		
No	573	193
Yes	576	158

Chi-square test — referred_by_friend vs is_churned: chi2=0.76, p-value=0.3818

is_churned	0.0	1.0
referred_by_friend		
No	563	182
Yes	586	169

t-test — average_watch_hours (churned vs retained): t=-0.19, p-value=0.8527

Figure 9: Chi-Square and T-test

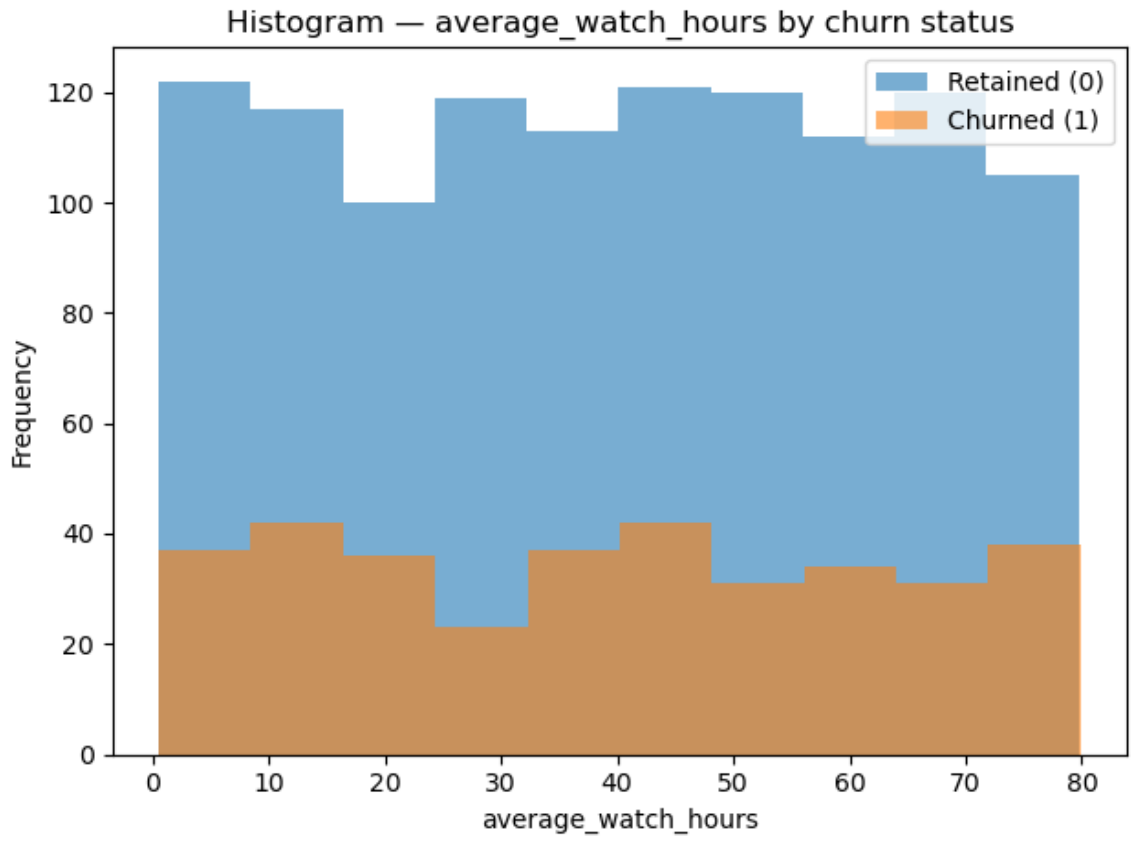

Figure 10: Histogram Chart

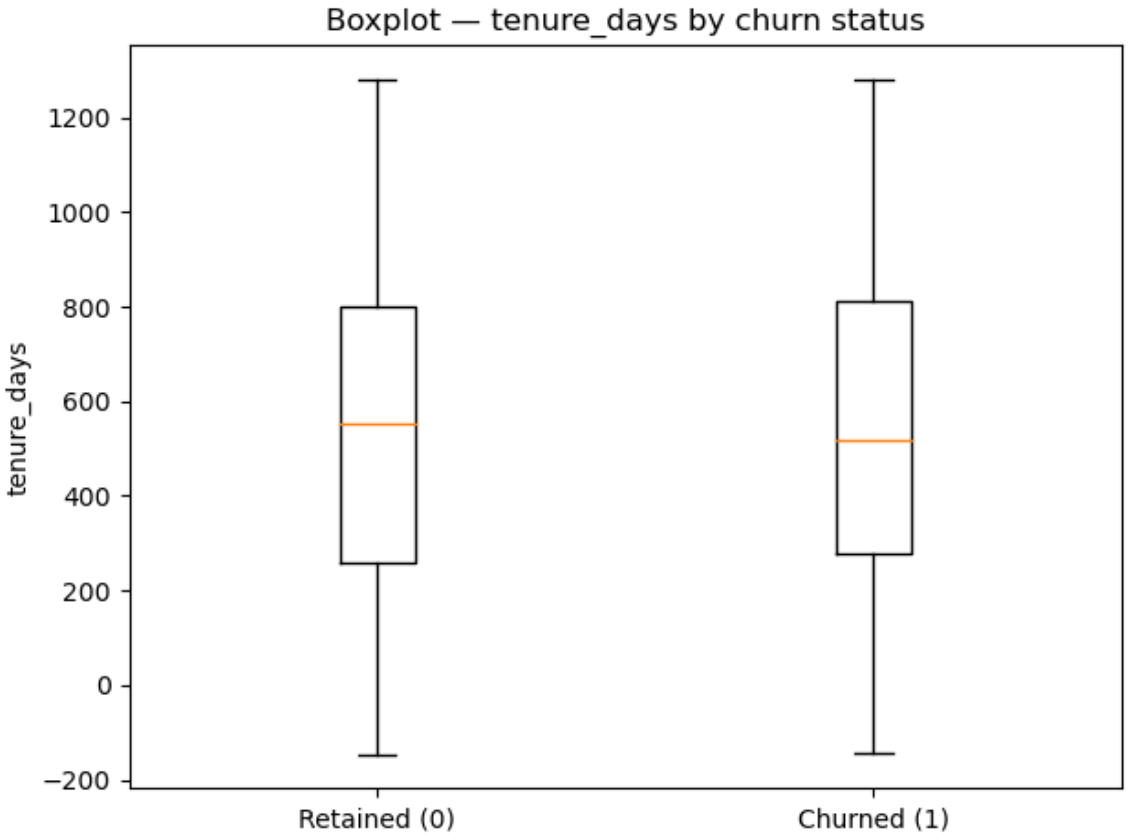

Figure 11: Boxplot chart

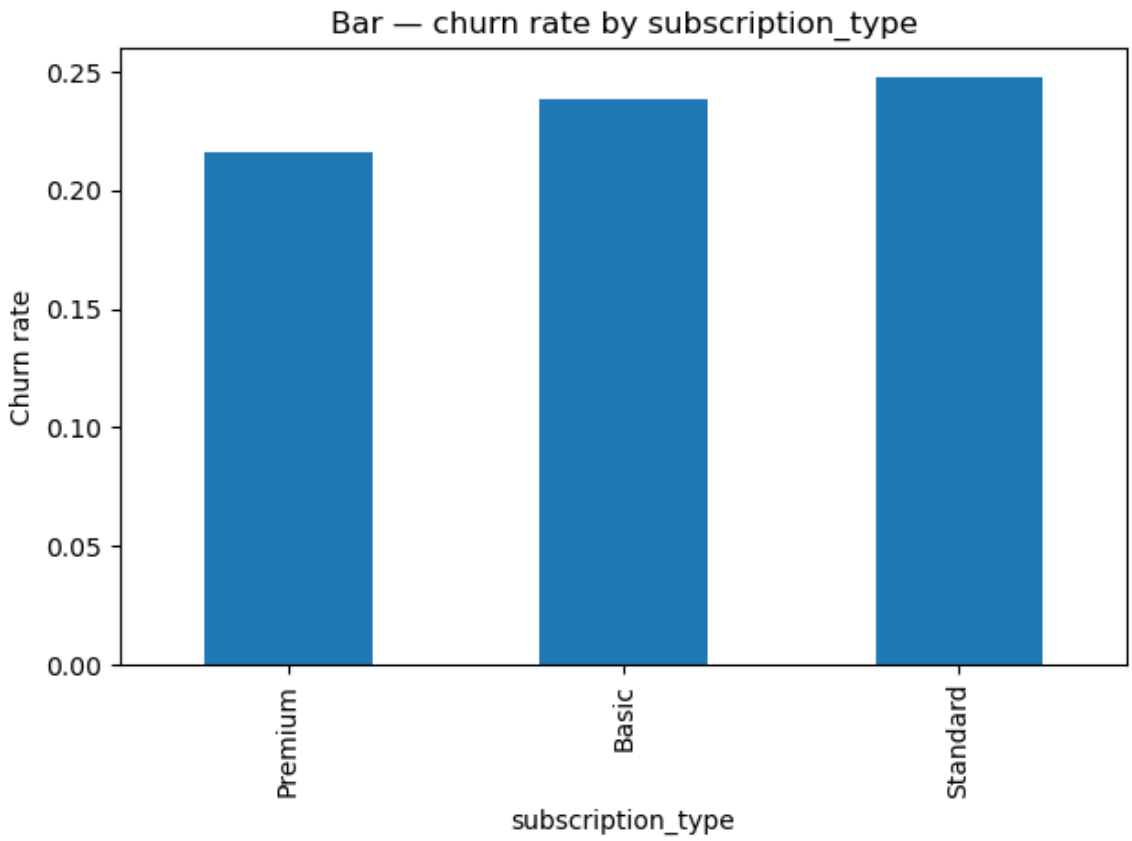

Figure 12: Bar chart

5. Model Results

Two models were built to explore customer behavior: **a Logistic Regression Model to predict churn (binary classification) and a linear regression model to understand the factors that affect** watch time (continuous prediction).

- **Logistic Regression Model (Churn Prediction):**
  The logistic model achieved moderate accuracy, with an accuracy of around 49.7% and an AUC of 0.47, indicating limited predictive power. It was still useful for identifying churn indicators: higher average watch hours and longer tenure days were linked with retention, while users without promotions or with low watch-to-fee ratios were more likely to cancel.

```
Final feature set shape: (1500, 20)

--- Confusion Matrix ---
[[120 110]
 [ 41  29]]

--- Classification Report ---
              precision    recall  f1-score   support

           0      0.745     0.522     0.614       230
           1      0.209     0.414     0.278        70

    accuracy                          0.497       300
   macro avg      0.477     0.468     0.446       300
weighted avg      0.620     0.497     0.535       300

ROC AUC: 0.471
```

Figure 13: Logistic regression confusion matrix and classification report

```
Top 10 Positive Coefficients (increase churn probability):

average_watch_hours           0.204955
is_loyal                      0.150097
country_India                 0.115446
country_UK                    0.101805
mobile_app_usage_pct          0.089307
country_France                0.083483
country_Germany               0.082901
country_USA                   0.058960
age                          -0.000308
subscription_type_Standard   -0.006291
dtype: float64


Top 10 Negative Coefficients (decrease churn probability):

watch_per_fee_ratio         -0.246699
received_promotions_Yes     -0.183977
gender_Other                -0.143323
gender_Male                 -0.117677
monthly_fee                 -0.117252
heavy_mobile_user           -0.096060
tenure_days                 -0.061903
subscription_type_Premium   -0.057068
referred_by_friend_Yes      -0.053107
complaints_raised           -0.043980
dtype: float64
```

Figure 14: Logistic Regression Top 10 positive and negative coefficients Results

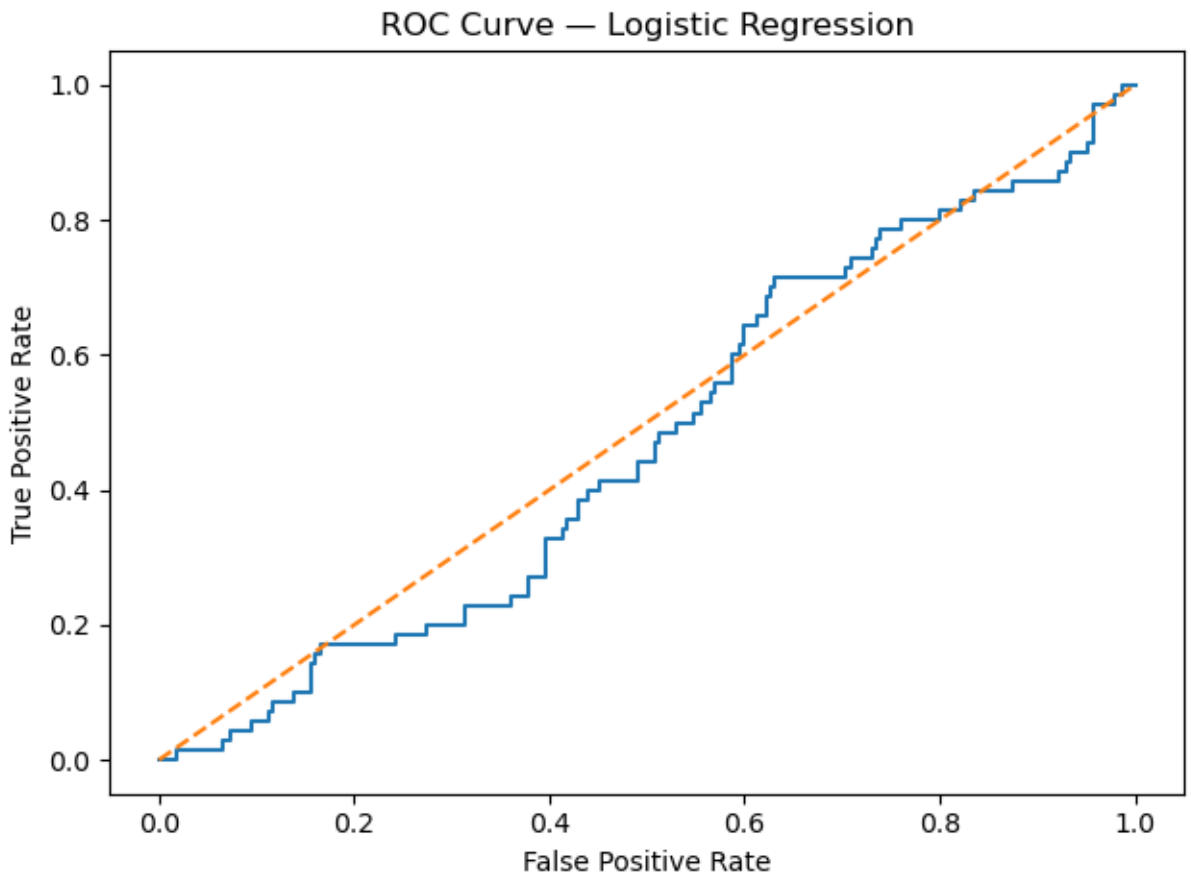

Figure 15: Logistic Regression ROC Curve

- **Linear Regression (Watch-Time Prediction)**

  The linear model performed strongly, explaining about 82 % of the variation in watch hours

  ($R^2 = 0.82$) with an RMSE of 9.81 and an MAE of 7.39. This means it predicted user engagement much more reliably. The most influential variables were watch_per_fee_ratio, monthly_fee, and subscription_type_Standard.

```
R²: 0.820, RMSE: 9.808, MAE: 7.388
```

Figure 16: Linear Regression model performance result

```
Top 10 features by absolute coefficient (influence):

watch_per_fee_ratio           23.377066
monthly_fee                   11.293584
subscription_type_Standard     2.557284
tenure_days                   -1.115436
is_loyal                       0.648137
country_Germany                0.600989
country_UK                     0.523622
mobile_app_usage_pct          -0.512894
age                            0.454385
country_USA                    0.433983
dtype: float64
```

Figure 17: Top 10 features of linear Regression

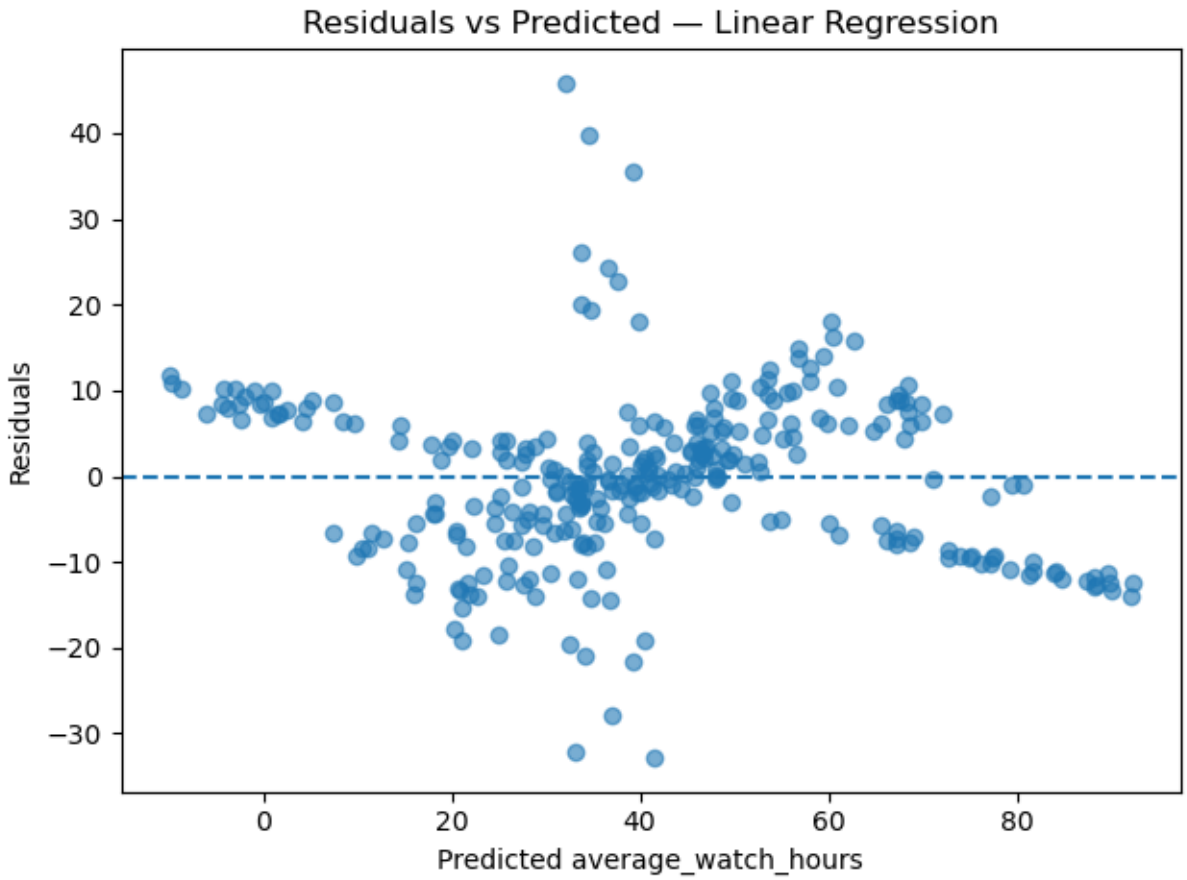

Figure 18: Residuals vs Predicted Linear regression

- **Model Comparison:**

  While the logistic regression gave insight into which factors drive churn, its predictive strength was modest. In contrast, the linear regression consistently produced better results and stronger explanatory value, indicating that user engagement patterns (watch hours and fee ratio) are more predictable than binary churn behavior. Future improvements could include advanced models such as Random Forest or Gradient Boosting to enhance churn prediction accuracy.

6. Business Questions Answered

1. Do Users who receive promotions churn less?

Yes — users who received promotional offers were slightly less likely to churn. In Figure 14, the logistic regression model showed a small negative coefficient for ***received_promotions***, indicating that promotions had a slight negative impact on customer retention. For example, many Premium users who got renewal discounts remained active longer, indicating that timely offers can strengthen loyalty.

2. Does watch time impact churn likelihood?

Yes — customers who spend more time watching content tend to stay subscribed. Figure 10: The Histogram chart shows that low-watch-hour users had higher churn rates, suggesting disengagement. For instance, users with less than 10 hours of monthly viewing were more likely to cancel, probably because they didn't find enough value in the content.

3. Are Mobile-dominant users more likely to cancel?

Slightly — the relationship was weak but observable. The mobile_app_usage_pct variable showed a small positive correlation with churn in the ***Model Coefficients (Figures 14 and 17).*** This suggests that heavy mobile viewers—possibly using shorter sessions or facing app usability limits—are somewhat more likely to cancel than mixed-device users.

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

The logistic regression results (*Figure 14*) highlighted **watch_per_fee_ratio, received_promotions**, and **is_loyal** as the strongest predictors. Customers who consume more content for their fee and receive targeted offers are less likely to churn, while loyal users

naturally remain subscribed for longer periods. These features together explain much of the variation in churn behavior.

5. Which customer segments should the retention team prioritize?

Retention efforts should focus on the Basic plan and users with low engagement. As seen in *Figure 12 (Bar Chart – Churn Rate by Subscription Type)*, Basic subscribers had the highest churn share. Offering these users temporary Standard-plan upgrades or bundled discounts could demonstrate added value and improve overall retention.

7. Recommendations

- **Target Basic-plan users with loyalty offers.**
  Basic subscribers churn at the highest rate (*Figure 12*). Short-term upgrades or loyalty discounts can show added value and encourage retention.
  Example: Offer "Try Standard Free for 30 Days" to Basic users nearing renewal.
- **Re-engage low -watch -time users.**
  Low engagement strongly links to churn (*Figure 10*). Personalized content suggestions or viewing reminders can revive interest.
  Example: Send push alerts highlighting new shows in a user's favorite genre.
- **Improve mobile app experience**
  Mobile-dominant users churn slightly more (*Figures 14 and 17*). Enhancing app speed and usability can increase satisfaction.
  Example: Optimize load times and reduce playback errors on mobile devices.
- **Maintain timely promotions.**
  Promotions reduce churn by reinforcing value (*Figure 17*). Continue renewal incentives and limited-time offers to retain price-sensitive users.
- **Track loyalty metrics for early alerts.**
  Monitor **tenure_days** and activity trends to flag at-risk users early and trigger automated retention actions.

8. Data Issues or Risks

Some challenges include missing values in the 'monthly_fee' column, moderate class imbalance between churned and retained users, and limited temporal data for user activity trends. The logistic regression model's modest accuracy suggests that future projects could explore advanced models, such as Random Forest or Gradient Boosting, for improved churn prediction.