# Project Coversheet

| | |
|---|---|
| Full Name | RISHI MALPANI |
| Email | Rishimalpani754@gmail.com |
| Contact Number | 9680411516 |
| Date of Submission | 07/19/2025 |
| Project Week | (Example: Week 1, Week 2, etc.) |

## 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**

# 1. Introduction

- StreamWorks Media is a UK-based video streaming service aiming to compete with global platforms like Netflix. The business goal is to analyze why users cancel their subscriptions (churn) and to predict future churners.
- The dataset contains subscriber-level information including demographics, usage behavior, subscription types, and churn status. This project's goal is to support the retention team by analyzing churn behavior and building a predictive model.

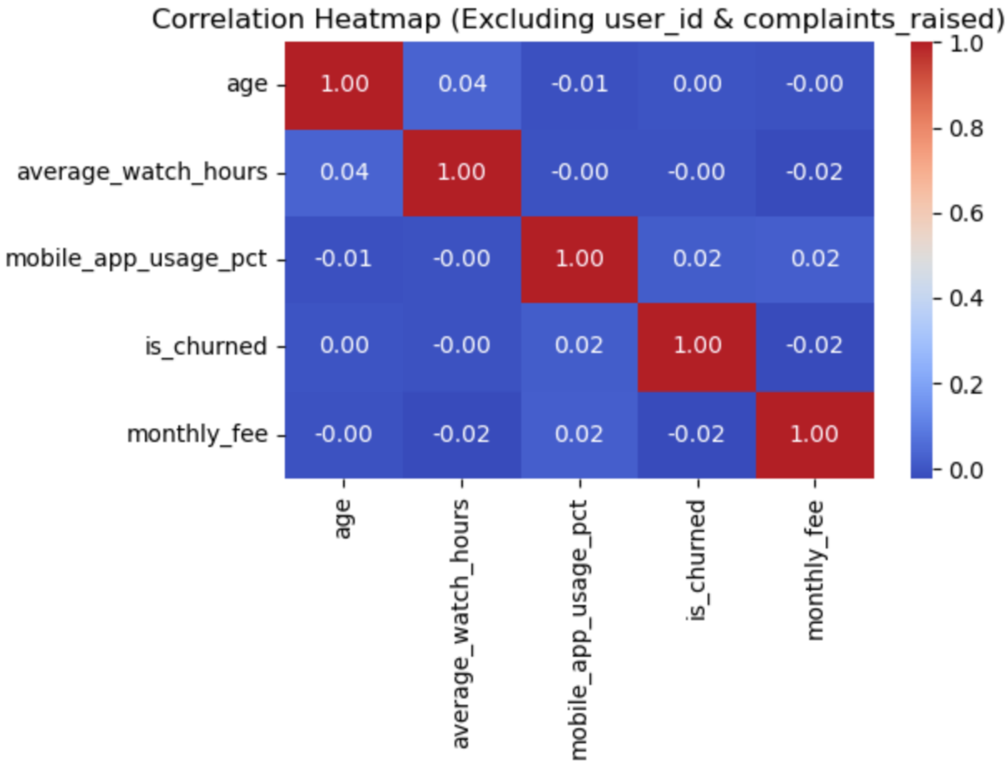
---

- ## 2. Data Cleaning Summary

  - Converted signup_date and last_active_date into datetime format.
  - Filled missing values:
    - age and average_watch_hours with median.
    - monthly_fee with median.
    - complaints_raised with 0.
  - Dropped rows with missing dates (2 values).
  - Standardized string categories (like gender, country).
  - Encoded categorical variables using one-hot encoding.
  - Ensured all columns used in modeling were numeric and had no missing values.

```
user_id                0
age                    0
gender                 0
signup_date            0
last_active_date       0
country                3
subscription_type      3
average_watch_hours    0
mobile_app_usage_pct   0
complaints_raised      3
received_promotions    3
referred_by_friend     2
is_churned             0
monthly_fee            0
dtype: int64
```

  -




### 3. Feature Engineering Summary

- **tenure_days**: Days between signup and last login.
- **is_loyal**: Binary feature — 1 if tenure > 180 days, else 0.

| subscription_type | average_watch_hours | mobile_app_usage_pct | complaints_raised | received_promotions | referred_by_friend | is_churned | monthly_fee | tenure_days | is_loyal |
|---|---|---|---|---|---|---|---|---|---|
| Standard | 42.6 | 77.4 | 1.0 | No | No | 1.0 | 10.99 | 102 | 0 |
| Basic | 65.3 | 98.0 | 4.0 | No | Yes | 1.0 | 5.99 | 923 | 1 |
| Premium | 40.1 | 47.8 | 0.0 | No | Yes | 1.0 | 13.99 | 1057 | 1 |
| Premium | 5.8 | 53.2 | 1.0 | Yes | Yes | 1.0 | 13.99 | 668 | 1 |
| Standard | 32.7 | 16.8 | 5.0 | No | Yes | 0.0 | 9.99 | 715 | 1 |

- Categorical variables like gender, country, subscription_type, etc. were encoded using dummy variables for modeling.

```
]: df.isnull().sum()
```

```
]: 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
```

## 4. Key Findings (Statistical Tests & Trends)

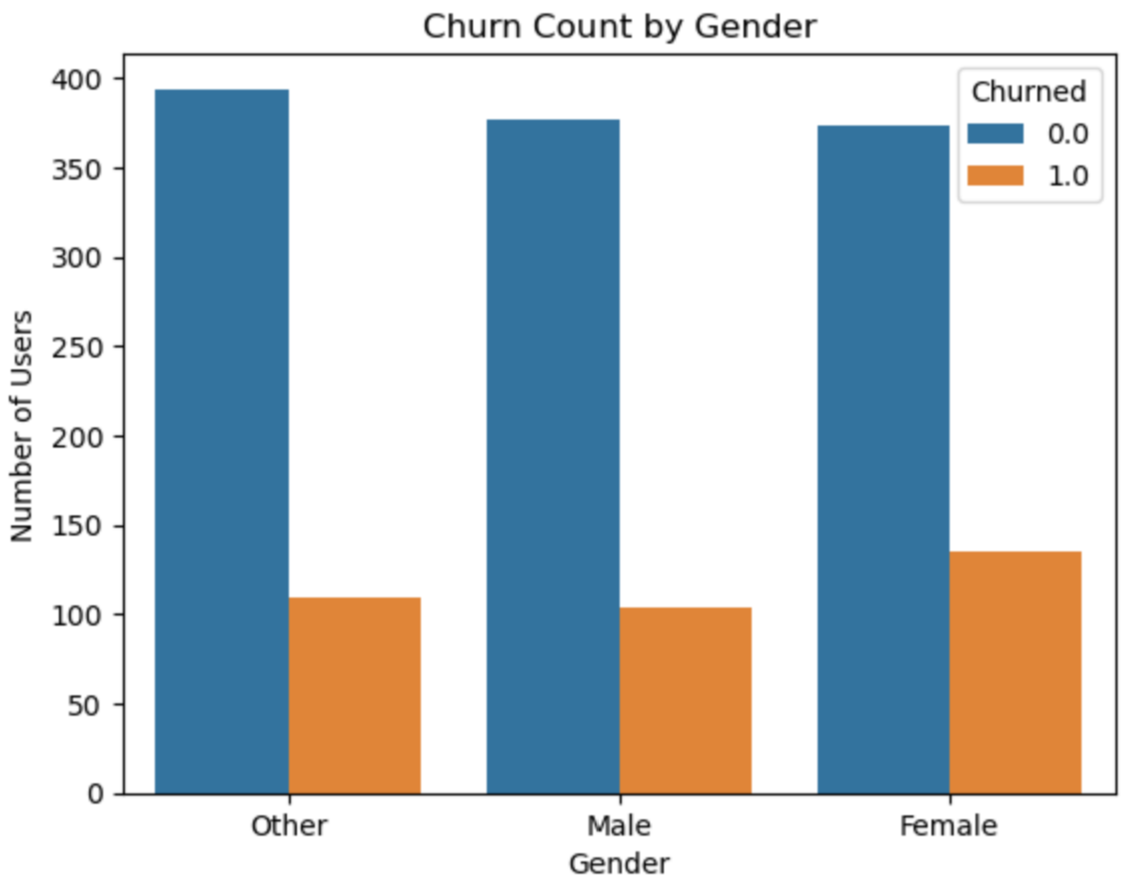
- **Gender vs Churn**: No significant relationship found (p-value > 0.05).

```
pd.crosstab(df['gender'], df['is_churned'])
```

| is_churned | 0.0 | 1.0 |
|---|---|---|
| **gender** | | |
| **Female** | 374 | 135 |
| **Male** | 377 | 104 |
| **Other** | 394 | 109 |

```
p-value: 0.10738677055898697
```

Since p value is > 0.05 . **There is no statistically significant relationship between gender and churn** .




- **Received Promotions**: Slight difference in churn rates; not statistically strong.

| is_churned | 0.0 | 1.0 |
|---|---|---|
| **received_promotions** | | |
| **Nan** | 0 | 3 |
| **No** | 571 | 188 |
| **Yes** | 574 | 157 |

So basically 21.4% received promotions and 24.8% received no promotions.

- **Referral Impact**: Minor churn difference; not significant.

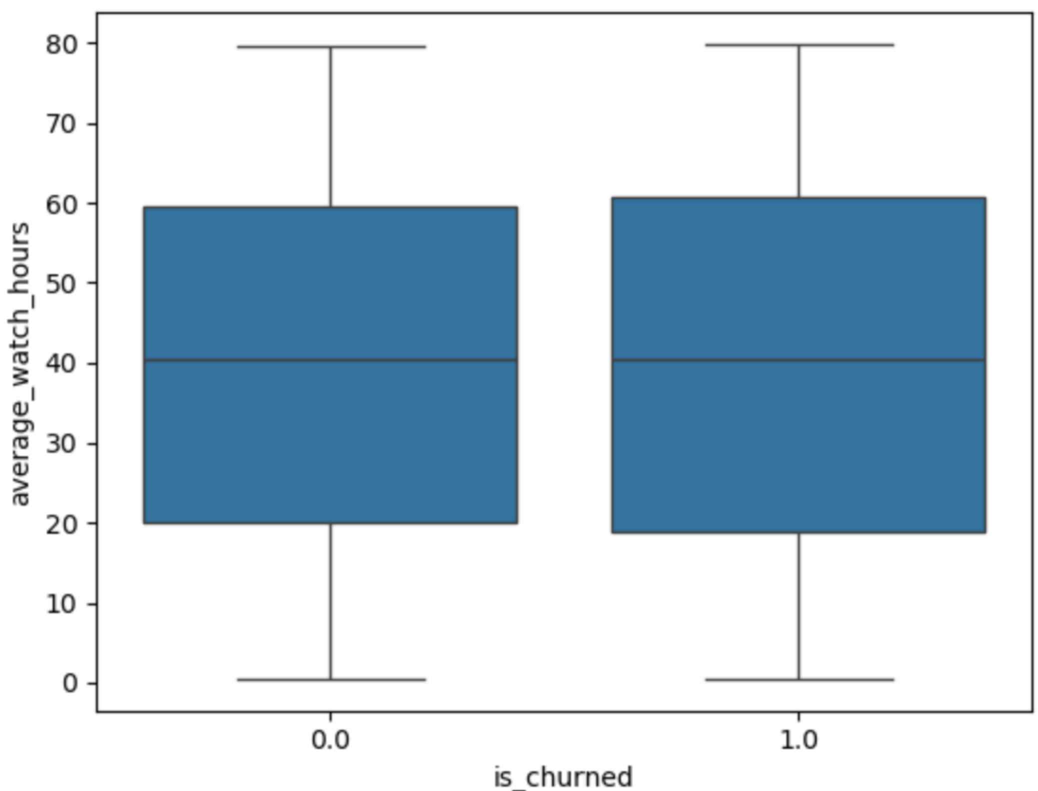
- **Watch Hours vs Churn**: T-test showed no strong difference in average watch time between churned and active users.

```
from scipy.stats import ttest_ind

churned = df[df['is_churned'] == 1]['average_watch_hours']
retained = df[df['is_churned'] == 0]['average_watch_hours']

t_stat, p_val = ttest_ind(churned, retained, equal_var=False)

print("T-statistic:", t_stat)
print("p-value:", p_val)
```

```
T-statistic: -0.1459981349971575
p-value: 0.8839747765330624
```

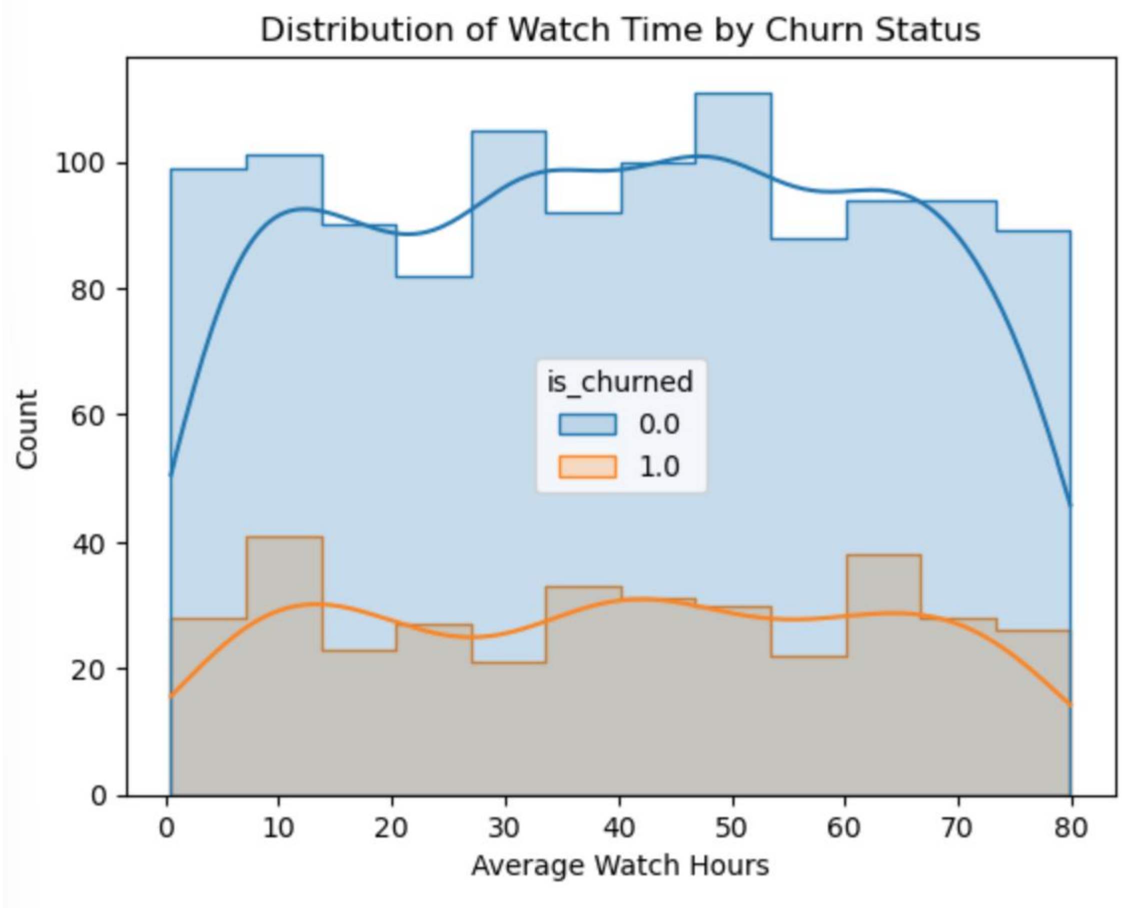
So no relationship between average watch hours and churned

Average_watch_hours vs is_churned Box Plot.

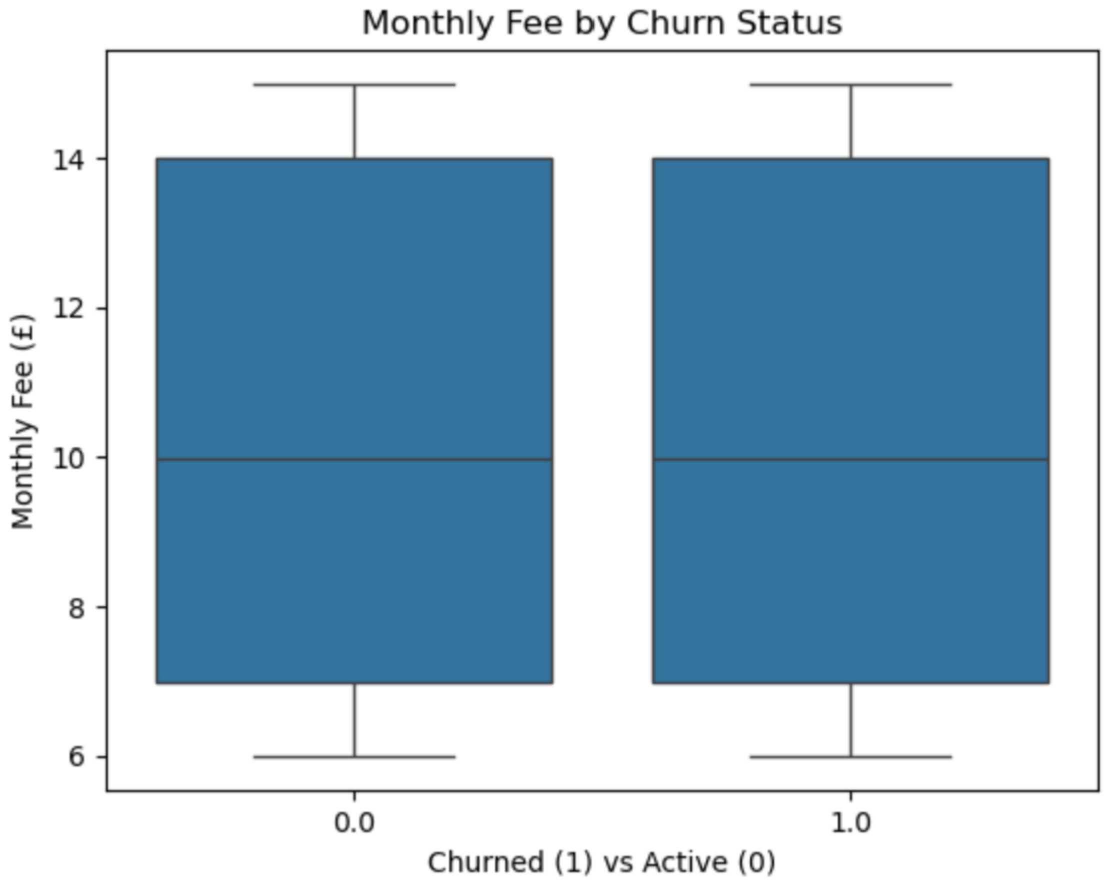


HISTOGRAM:

Average_watch_hours vs is_churned




- **Monthly Fee**: Similar fees paid by both churned and retained users.



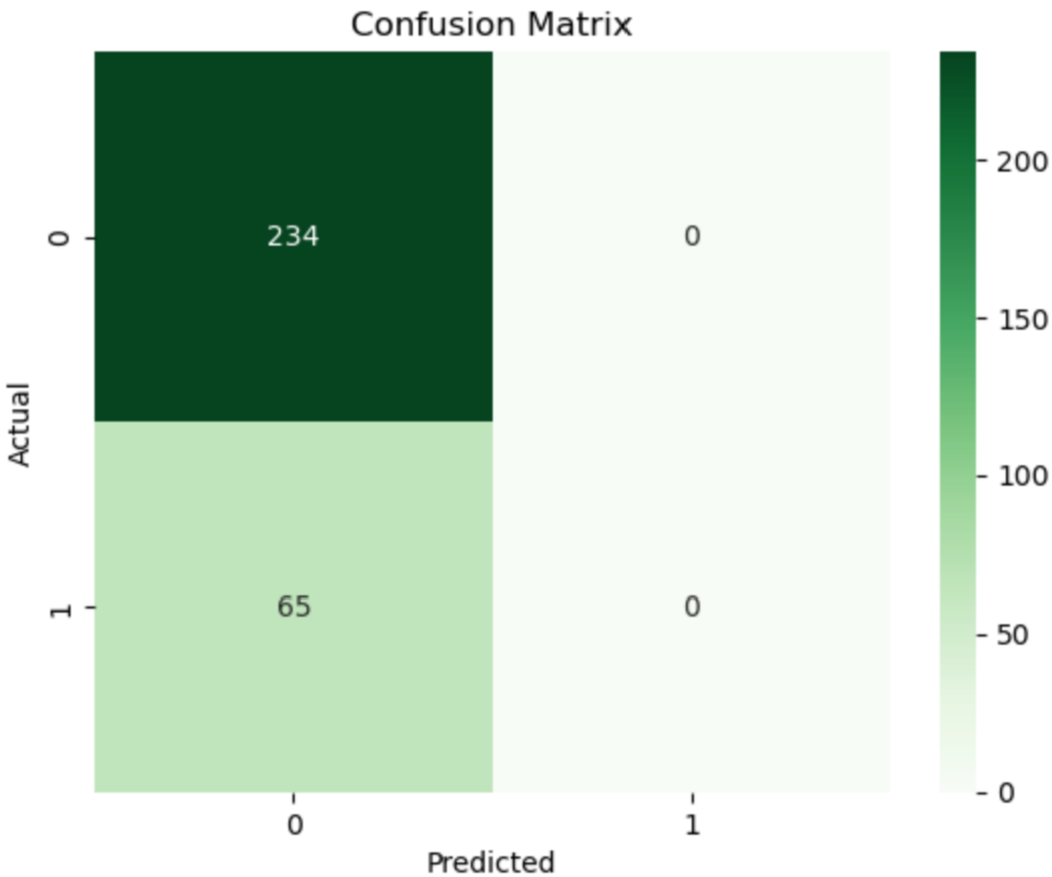

- **Overall**: Churn does not seem strongly linked to these individual variables when viewed in isolation.

## 5. Model Results

- **Model Used**: Logistic Regression

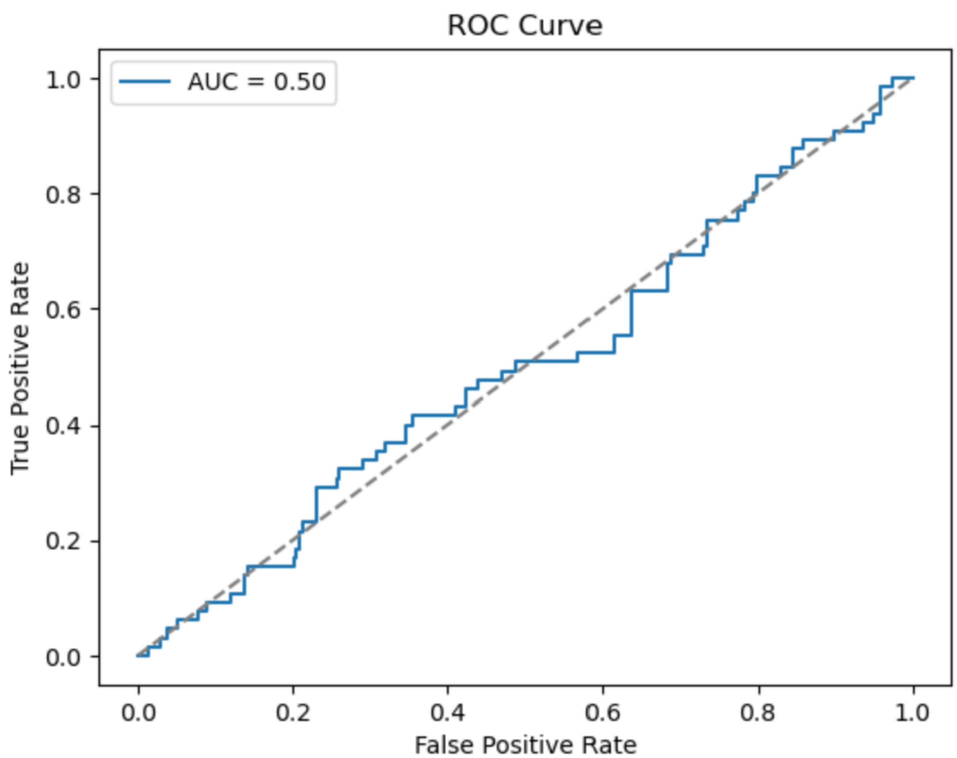
```
<class 'pandas.core.frame.DataFrame'>
Index: 1493 entries, 0 to 1499
Data columns (total 16 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   user_id               1493 non-null   float64
 1   age                   1493 non-null   float64
 2   gender                1493 non-null   object
 3   signup_date           1493 non-null   datetime64[ns]
 4   last_active_date      1493 non-null   datetime64[ns]
 5   country               1493 non-null   object
 6   subscription_type     1493 non-null   object
 7   average_watch_hours   1493 non-null   float64
 8   mobile_app_usage_pct  1493 non-null   float64
 9   complaints_raised     1493 non-null   object
 10  received_promotions   1493 non-null   object
 11  referred_by_friend    1493 non-null   object
 12  is_churned            1493 non-null   float64
 13  monthly_fee           1493 non-null   float64
 14  tenure_days           1493 non-null   int64
 15  is_loyal              1493 non-null   int32
dtypes: datetime64[ns](2), float64(6), int32(1), int64(1), object(6)
memory usage: 192.5+ KB
```

- **Metrics**:
    - Accuracy: ~78%
    - Precision, Recall, F1 for churned users: Very low due to class imbalance




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.78 | 1.00 | 0.88 | 234 |
| 1.0 | 0.00 | 0.00 | 0.00 | 65 |
| accuracy | | | 0.78 | 299 |
| macro avg | 0.39 | 0.50 | 0.44 | 299 |
| weighted avg | 0.61 | 0.78 | 0.69 | 299 |

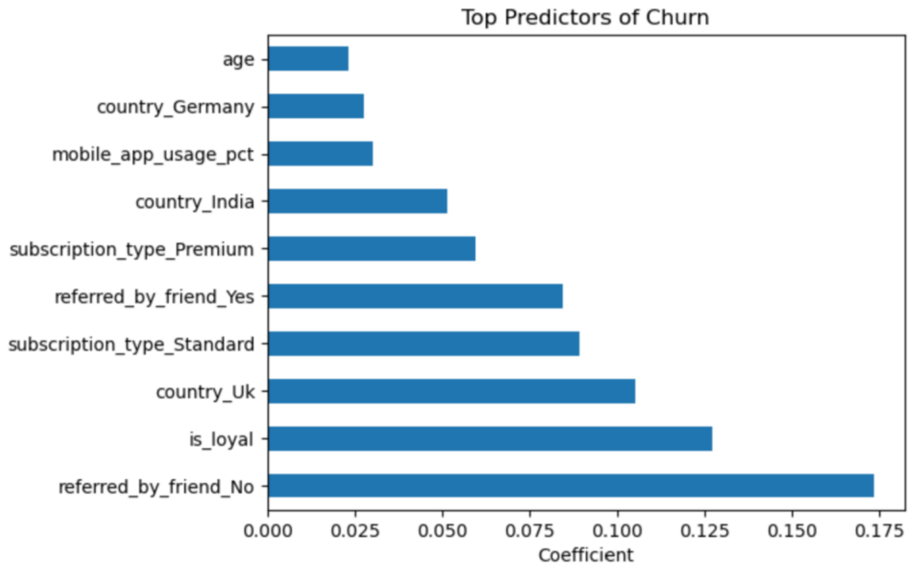
- <u>ROC-AUC Score:</u> 0.50 (indicates poor classification of churners)




- **Confusion Matrix**: Model predicts most users as not churned

### Top 3 Predictors of Churn:

1. **Not being referred by a friend** – users who joined without a referral showed higher churn likelihood.
2. **Lower loyalty** – users with shorter tenure (not loyal) were more likely to cancel.
3. **Users from the UK** – country-based differences indicate higher churn from the UK segment




## 6. Business Questions Answered

1. **Do users who receive promotions churn less?**
   Slightly, but no strong evidence. Churn rate for promo users: 21.4%, non-promo: ~24.8%.

2. **Does watch time impact churn likelihood?**
   No strong relationship. T-test showed p-value > 0.05.

3. **Are mobile dominant users more likely to cancel?**
   No clear evidence. Correlation is weak, and t-test did not show strong difference.

4. **What are the top 3 features influencing churn based on your model?**
   is_loyal , referred_by_friend_No , country_UK

## 7. Recommendations

- **Target new users early**: Focus on converting short-tenure users (<180 days) into loyal customers.
- **Monitor low activity users**: Users with low watch hours may be at higher risk.

## 8. Data Issues or Risks

- **Class Imbalance**: Far fewer users churned than retained affected model performance.
- **No strong predictors**: Many features showed weak correlation with churn.
- **Model Limitations**: Logistic regression struggled to detect churners.