Project Coversheet

Full Name	Robert May
Project Title (Example – Week1, Week2, Week3, Week 4)	Week 3

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.

YOU CAN START YOUR PROJECT FROM HERE

1. Introduction

This report, for Streamworks Media, aims to analyse customer churn and the reasons behind it, especially with customer acquisition becoming more expensive. The aim is to understand who is churning and why, predict churn probability and enable intervention.

The dataset has 1500 rows and 14 columns.

2. Data Cleaning Summary

Null values were changed to ‘Missing’ or ‘Unknown’ depending on whether it was thought that the information could be found elsewhere. Dates were left empty. Average watch hours were changed to the median for each subscription type, and mobile app pct to median. Complaints raised were changed to 0.0. For any Missing values, it is recommended that other teams investigate and find the correct value.

Column types were changed to numeric where appropriate and dates to datetime.

The data was in a good state, the worst column monthly_fee having 1% null values.

Dates were in the wrong column – signup_date and last_active_date and needed to be swapped in some instances. This was discovered when creating tenure_days feature.

	Missing values	% missing	duplicates	inconsistent data	%
user_id	2	0.13%		0	
age	3	0.20%		0	
gender	1	0.07%		0	
signup_date	2	0.13%		0	
last_active_date	2	0.13%		0	0.00%
country	3	0.20%		0	0.00%
subscription_type	3	0.20%		0	0.00%
average_watch_hours	4	0.27%		0	0.00%
mobile_app_pct	3	0.20%		0	0.00%
complaints_raised	4	0.002667		0	0
received promotions	3	0.002		0	
referred_by_freind	3	0.60%		0	
is_churned	1	0.20%		0	0.00%
monthly_fee	5	1.00%		0	0.00%

For the string columns, one-hot encoding was run to prepare the data for prediction and statistical insights.

3. Feature Engineering

New columns were created –

tenure_days = last active date – signup date

watch_per_fee = average hours watched / monthly fee

is_loyal = tenure_days > 180

4. Key findings

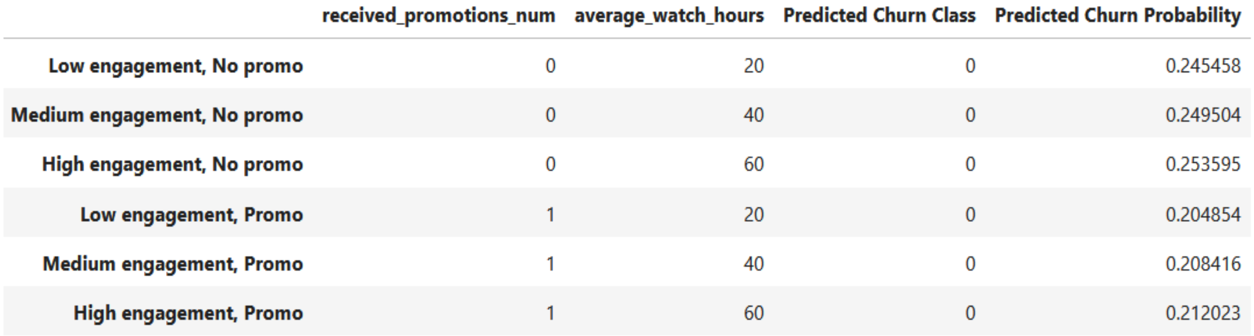
Chi-square tests were run against is_churned: gender, p value 0.059, received_promotions, p-value 0.002 (significant), referred_by_friend: p value 0.49

A two group t-test was run for average watch hours vs is_churned. The stat was -0.175 which is insignificant, with a small difference between the two means. There was a p value of 0.8613.

The only p value with significance here was between received promotions and churn.

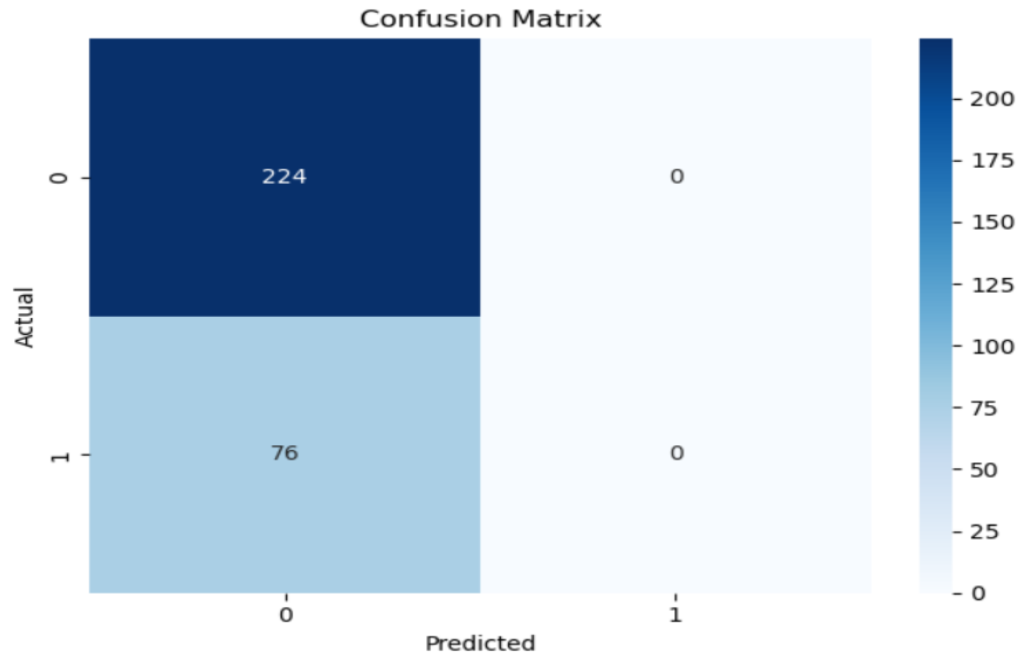
5. Model Results

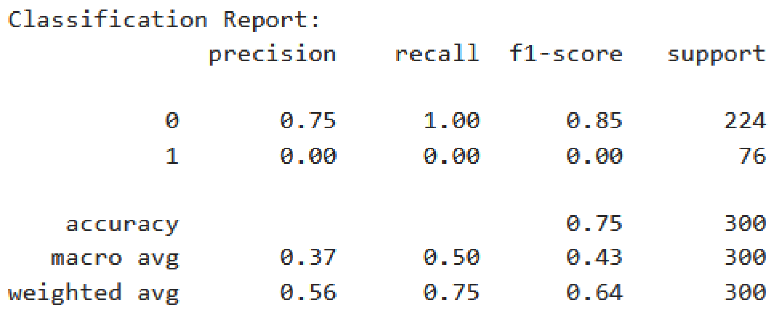
A Logistic regression model was created to measure average watch hours and receiving marketing promotions impact on predicted churn. Average watch hours, was divided into low, medium and high users. This table shows that receiving promotions reduces the probability of churning by approximately 4%.



	received_promotions_num	average_watch_hours	Predicted Churn Class	Predicted Churn Probability
Low engagement, No promo	0	20	0	0.245458
Medium engagement, No promo	0	40	0	0.249504
High engagement, No promo	0	60	0	0.253595
Low engagement, Promo	1	20	0	0.204854
Medium engagement, Promo	1	40	0	0.208416
High engagement, Promo	1	60	0	0.212023

A Confusion Matrix was created. There were no ‘positives’ (predictions of churn).





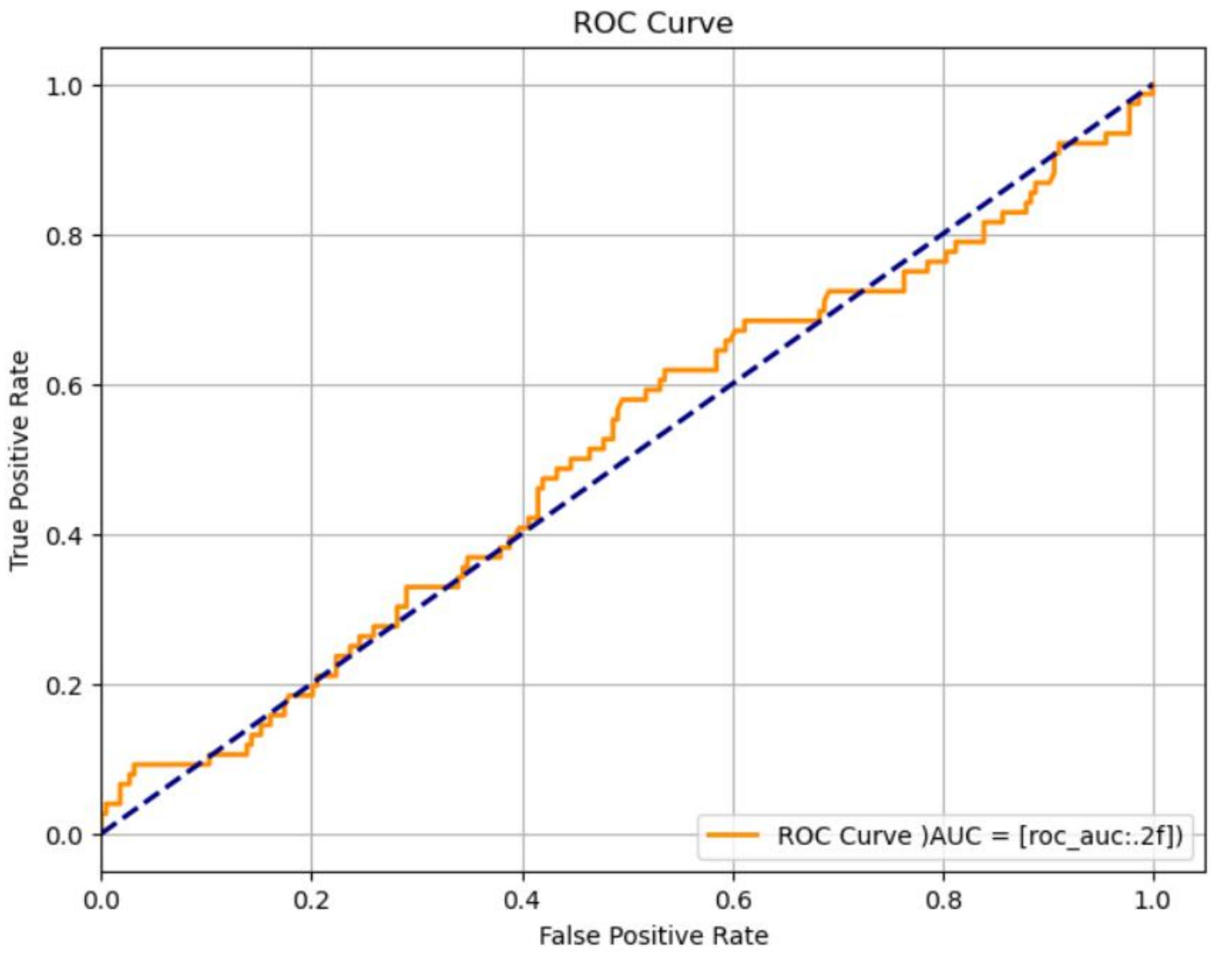
```
Classification Report:
              precision    recall  f1-score   support

           0       0.75      1.00      0.85       224
           1       0.00      0.00      0.00        76

    accuracy                           0.75       300
   macro avg       0.37      0.50      0.43       300
weighted avg       0.56      0.75      0.64       300
```

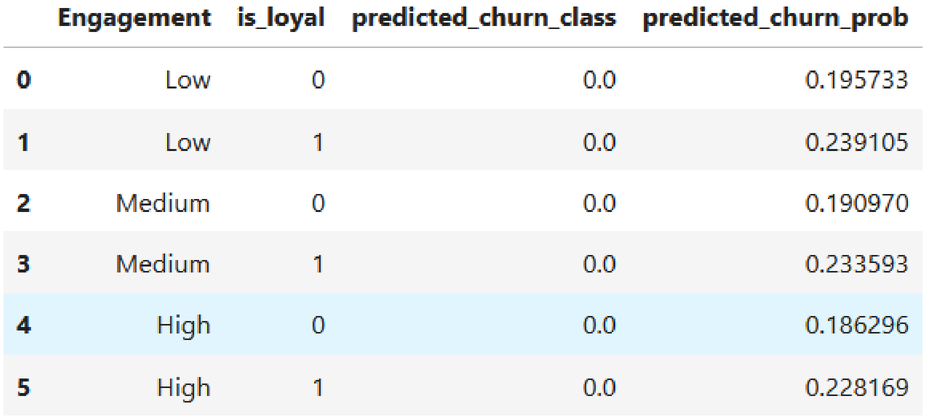
The model was decent at predicting no churn, but terrible at predicting churn. (F1 score=0). No precision and recall for predicting churn. With the average probability of churn being under 25%, it does make sense that for each data point, the model would predict no_churn.

ROC curve shows almost diagonal line with AUC score of 0.511. (0.5 is a random classifier). So only marginally better than that.



In terms of the coefficients, promotions was -0.1178 and watch hours 0.025. This implies that for every hour watched, the probability of churn is slightly higher. Although the coefficient is so minimal it is almost negligible.

A model was also run with is_loyal instead of promotions, giving the following table:



	Engagement	is_loyal	predicted_churn_class	predicted_churn_prob
0	Low	0	0.0	0.195733
1	Low	1	0.0	0.239105
2	Medium	0	0.0	0.190970
3	Medium	1	0.0	0.233593
4	High	0	0.0	0.186296
5	High	1	0.0	0.228169

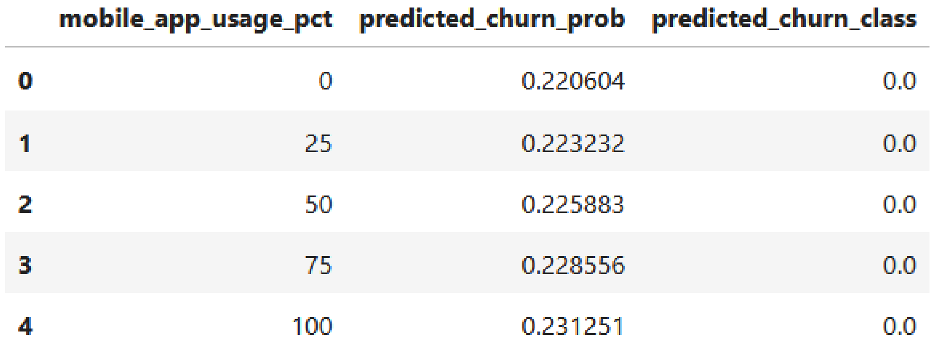
Once again, there was a 4% difference. This showed that being loyal means actually a 4% higher chance of churning. It suggests that after a longer period, people have seen what they want, and don’t feel the need to stay continously.

Is_loyal gave a p value of 0.461 from a chi-square test, showing it is not a strongly significant relationship. The coefficient for is_loyal was 0.255 and watch hours was negligible and negative. Being loyal increases the odds of churn by 29%, very considerable.

Tenure days gave a p value of 0.000 from a chi-square test, showing a very significant relationship. The coefficient was only -0.0002 though.

Complaints raised was also run in a logistic model. This gave a coefficient of 0.0107, average watch hours -0.0016. This means each extra complaint increases the odds of churning by 11.3%. p value is 0.286 in a chi squared test though.

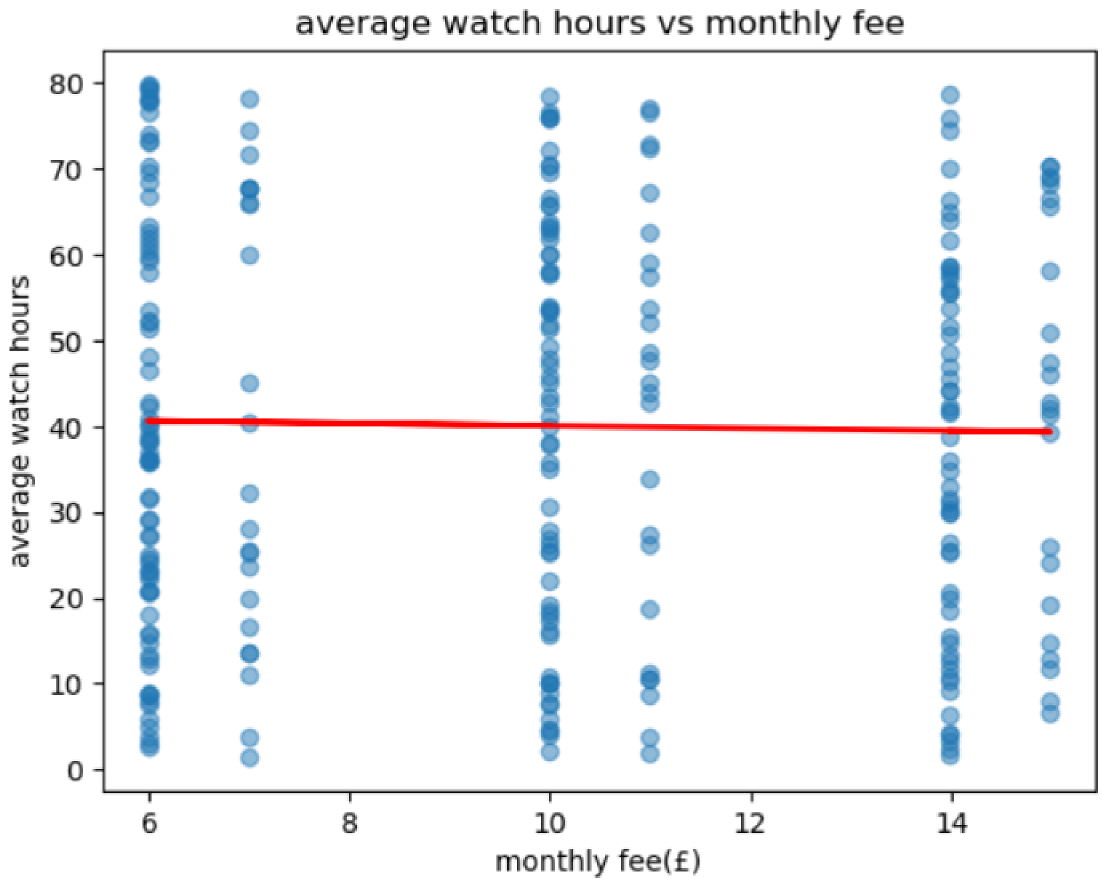
Mobile app pct gave a p value from a chi square of 0.000, but the coefficient from regression was 0.0006. For each 10% increase in mobile phone app pct, the probability of churn increases by 0.6%.



	mobile_app_usage_pct	predicted_churn_prob	predicted_churn_class
0	0	0.220604	0.0
1	25	0.223232	0.0
2	50	0.225883	0.0
3	75	0.228556	0.0
4	100	0.231251	0.0

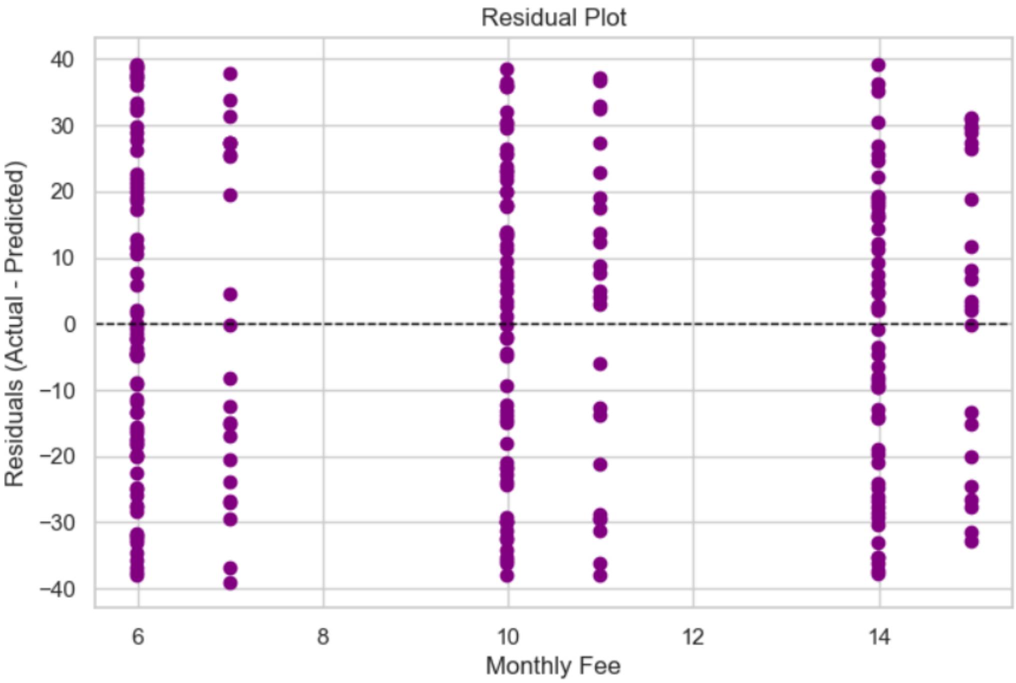
Three predictors of churn are whether or not customers receive marketing promotions, whether they are loyal (have been a customer at least 180 days), and how many complaints they raise.

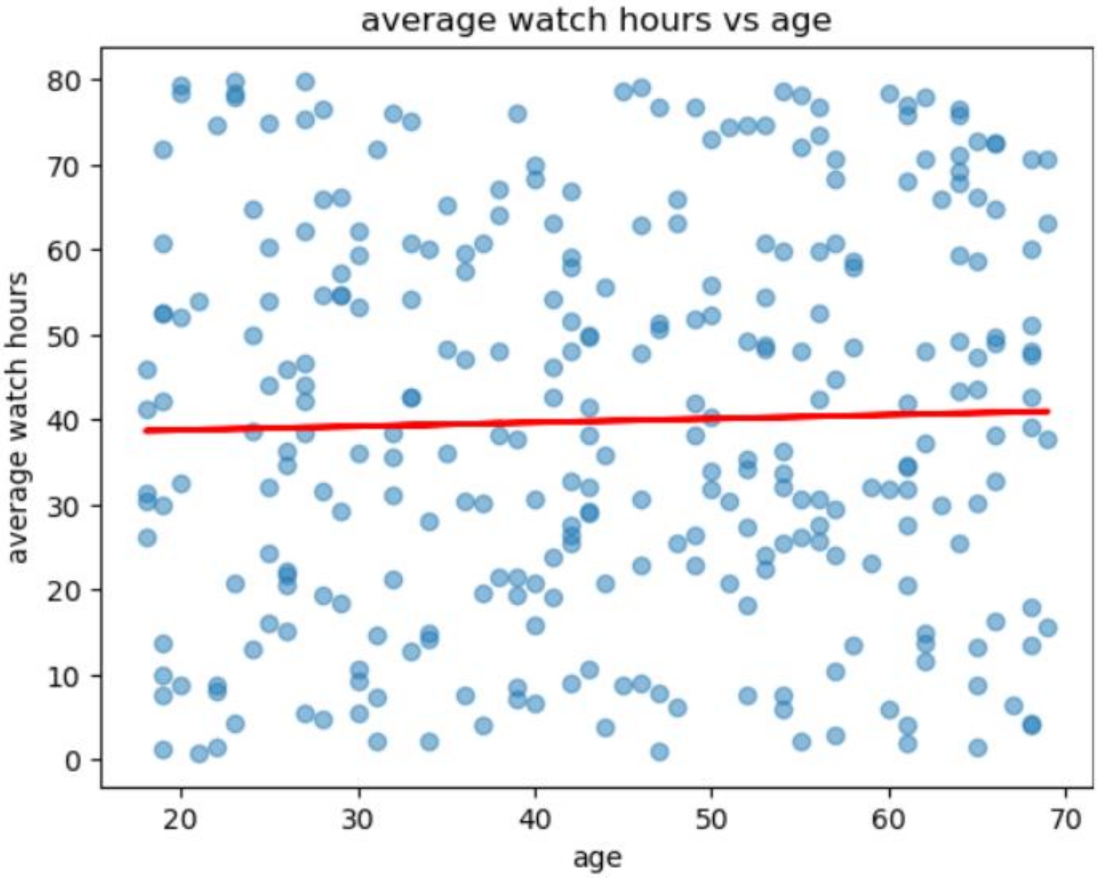
To predict average watch hours, a linear regression model was run to see if they could be predicted by monthly_fee. As you can see, there is no obvious correlation.



The r” is 0.0005 and the RMSE is 23.2791. The coefficient is -0.145. The MAE (Mean Absolute Error) is 20.28. There is a slight negative correlation. For every £1 increase in monthly fee, watch hours decrease by 0.145 hours.

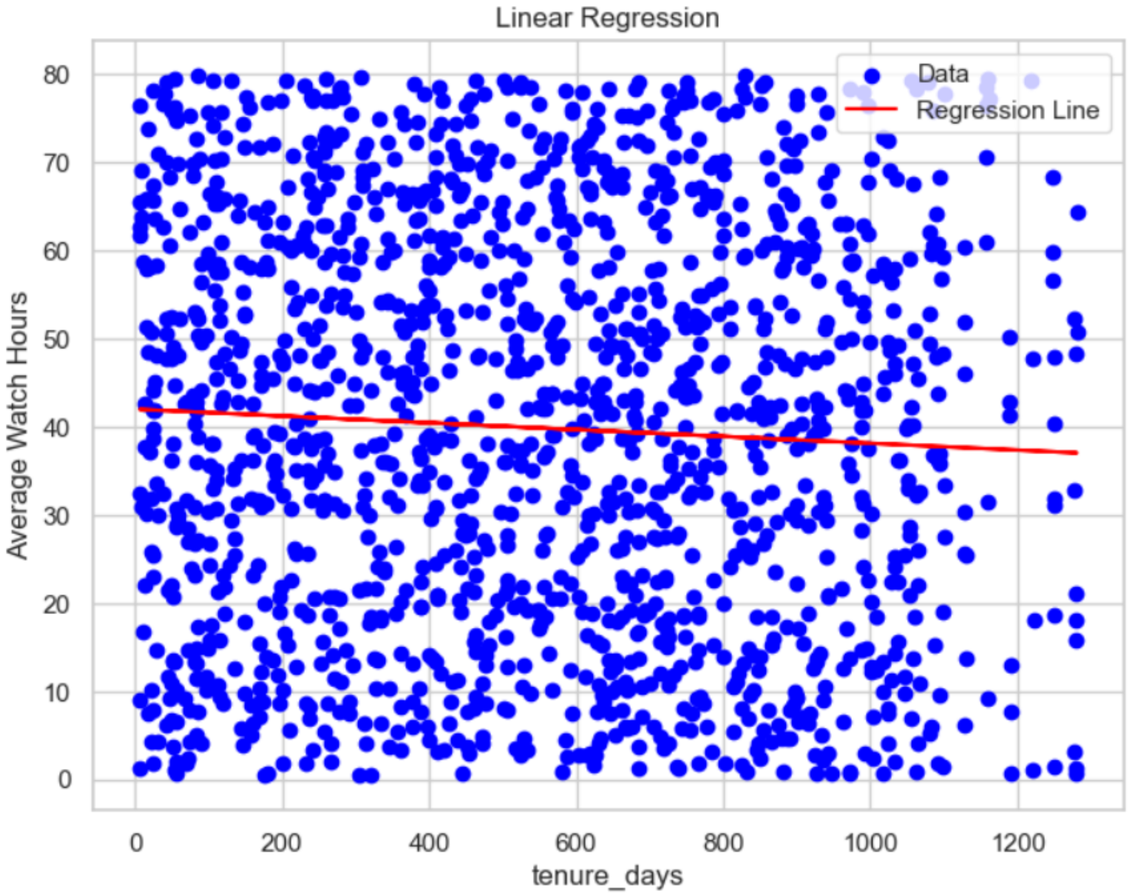
A residual plot is shown below. These show the actual – predicted value. Again there is no real trend.


Residual Plot
Residuals (Actual - Predicted)
40
30
20
10
0
−10
−20
−30
−40
6
8
10
12
14
Monthly Fee

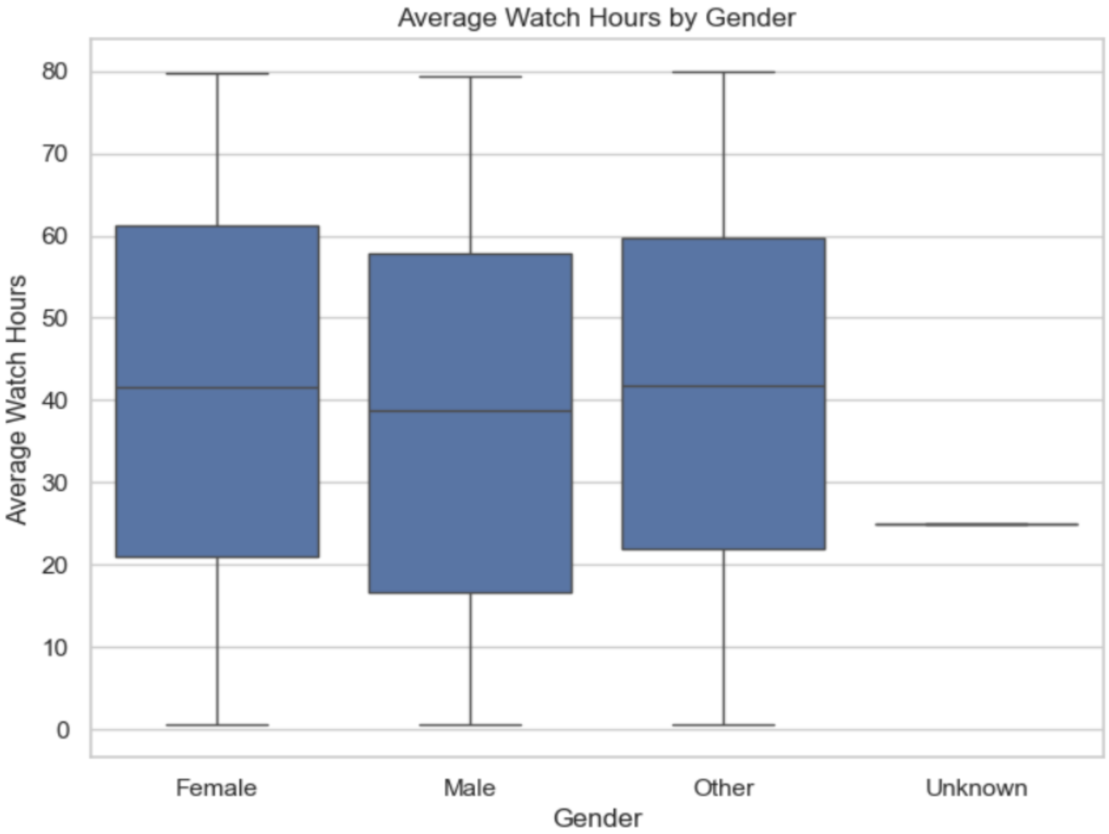

average watch hours vs age
average watch hours
80
70
60
50
40
30
20
10
0
20
30
40
50
60
70
age

Age was also run to see if predicted watch hours. There was a slight positive trend. Watch hours = 37.89 + 0.04 x age. The r2 was 0.003 and RMSE 22.959. The effect is small though – for every 25 years you might watch for 1 more hour.

Tenure days was plotted against average watch hours, showing a negative correlation. Again it's small, a coefficient of -0.004. r2 of 0.001. For each 100 days, people may watch for 0.4 hours less. The RMSE was 21.502 – average error of this.



Gender was also tested, although this would not suit linear regression. Therefore, a boxplot was produced. This showed that Men watch on average less than other genders.



The top 3 predictors of watch time can be said to be age, tenure days and gender.

6. Business questions

Do users who receive promotions churn less?
Yes, they do. There is a 4% lower churn probability for users who receive them.

Does watch time impact churn likelihood?
No, not in any significant way.

Are mobile dominant users more likely to cancel?
Yes.. For each 10% increase in app % usage, churn odds increase by 0.6%.

What are the top 3 features influencing churn, based on the model?
The top 3 factors are receiving promotions, whether the customer is loyal, and how many complaints they raise.

Which customer segments should the retention team prioritise? They should prioritise sending promotions to customers, giving them no reason to complain, and keeping loyal/long term customers. Perhaps this could include additional benefits once they have been a member for a certain time.

What factors affect customer watch time? Age and tenure days are the only variables with any meaningful impact, and that is small. Gender also has an impact, with men watching less on average.

7. Recommendations

- It is recommended to make sure customers receive promotions, whether new or existing, as it gives a 4% reduction in churn probability.
- It is recommended to incentivise loyal customers to stay, perhaps giving additional benefits or discounts.
- It would be a good idea to ensure customers don’t raise complaints, perhaps with better customer support and instant messaging service.

8. Data issues

The data was in a good state, however, care needs to be taken to ensure signup_date and last_active dates are the correct way around. Missing values also need to be populated correctly.

References:

Uptrail (2026). Week 3 recording.

Chatgpt

Uptrail (2025). Week 3 – Statistics and Predictive Modelling.pdf

Have loaded data and filled null values. I changed user id to ‘Missing’. I changed age to median. I changed gender to unknown. I left empty dates. I changed country to Unknown. I changed subscription type to Missing. I changed average watch hours to the median for each subscription type. I filled mobile app percentage usage to median. I changed complaints raised to 0.0. I changed recieived promotions to ‘Unknown’. Referred by friend to Unknown and ‘is churned’ to Missing. I changed monthly fee to Missing.

Now checking inconsistent values. Changing column types to numeric, dates to datetime. Ran correlation matrix but no correlations. When creating new feature tenure days, I found that some dates were the wrong way round in signup date and last active date columns so swapped them around.

I then started doing encoding, doing one hot encoding for the string columns.

Used chi-square test to check if gender, received_promotions or referred_by_friend were related to churn. Received-promotions came out with a p value of 0.002, below the 0.05 threshold, making it related.

I ran a two group t – test of is churned and average watch hours. The P value was 0.86 so no correlation. The t statistic -0.175 so insignificant, it needed to be 2 or more to be significant and suggest a statistically meaningful relationship. The group means, the mean watch hours for is churned was slightly less, 39.71 compared to 37.96.

I decided to compare churned users by subscription type. I ran a contingency table with churn rates, and a chi square test, which said they weren’t related.

I ran some bar charts to show numbers, preferring absolute numbers rather than churn rates. Churn rates could be affected by small sample sizes. I created one more feature ‘watch per fee’ which was average hours watched / monthly fee.

I then built a model, deciding to focus on average monthly hours and received_promotions as the vairables. Received promotions was the only one that was statistically significant earlier, that is why I chose it.

Ran model, which produced table.

Observations with missing values in the explanatory variables were removed prior to model fitting, as logistic regression does not support missing feature values.”

I then did Confusion matrix, which found model gave no positives. Also Recall and scores. I plotted ROC curve and AUC score. I found coefficients, - for promotions and + for

Linear regression. I tried monthly fee to predict average watch hours, and then age. Both had such little effects as to be negligible. Monthly fee negative, age positive. I also tried subscription type, which had no effect at all.