Project Coversheet

Full Name	Omolara Buhari
Project Title (Example – Week1, Week2, Week3, Week 4)	Week 3 - Project: Churn Prediction for Stream Works 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.

YOU CAN START YOUR PROJECT FROM HERE

Week 3 - Project: Churn Prediction for StreamWorks Media.

1). Introduction

StreamWorks Media is a UK-based subscription streaming service that offers digital entertainment similar to Netflix and Amazon Prime. The company's main challenge is a **rising customer churn rate** — many users are cancelling their subscriptions after short periods.

The **business goal** of this project is to help StreamWorks understand *which customers are likely to churn* and *why*, so the company can act early to retain them.

The **dataset** used contains over **5,000 customer records** including information such as gender, subscription type, monthly fee, average watch hours, referral status, and promotion history.

- The **purpose of the analysis** is to:
  - Identify factors that influence churn and user engagement.
  - Build models to predict churn and viewing hours.
  - Provide data-driven business recommendations to improve customer retention.

2) **Data Cleaning Summary**

The dataset required cleaning to make it suitable for analysis.

Key steps included:

- Converted date columns (signup_date, last_active_date) into **datetime format**.
- Created a new column tenure_days by calculating how long each user stayed active.
- Dropped unnecessary columns like user_id and removed duplicate rows.

- Changed data types for numerical variables such as monthly_fee and average_watch_hours.
- Handled missing values: replaced or dropped them depending on importance.
- Encoded categorical variables (e.g., gender, subscription_type, referred_by_friend) using **one-hot encoding** so models could process them.

After cleaning, .info() showed consistent column types and no critical missing data. The dataset was ready for statistical and predictive modeling.

3. Feature Engineering Summary

To help the model capture more meaningful information, new features were created:

- To enrich the analysis, new variables were derived to better capture customer behavior and lifecycle.
  - **Tenure-based features:** A tenure_days variable was created to represent how long each user has been active on the platform. Additionally, a binary feature is_loyal was defined to identify long-term users.
  - **Marketing & referrals:** Binary indicators were created for received_promotions and referred_by_friend to test the impact of marketing efforts and word-of-mouth acquisition.
  - **Engagement metrics:** Aggregated measures of watch activity were included, such as total and average watch time. A feature, mobile_dominant, was added to indicate users who primarily use mobile devices.
  - **Dummy variables:** Categories such as gender and country were transformed into dummy variables for use in regression models.

4. Key Findings

- A series of statistical tests was performed to explore relationships between variables and churn.
  - **T-tests:**
    - *Gender* – p-value = 0.2399 → no statistically significant difference in churn rates between male and female users.

    - *Referred by friend* – p-value = 0.4946 → no significant difference in churn between referred and non-referred users.
    - *Watch time groups* – T-statistic = –0.691, p-value = 0.490 → no major difference in mean watch time between churned and retained users in the current sample.
  - **Chi-square test:**
    - p-value = 0.0331 → shows a **significant relationship** between churn and at least one categorical variable, most likely *received_promotions* or *device type*.

      This suggests that marketing exposure or how users access the service is associated with whether they stay or leave

5. Model Results

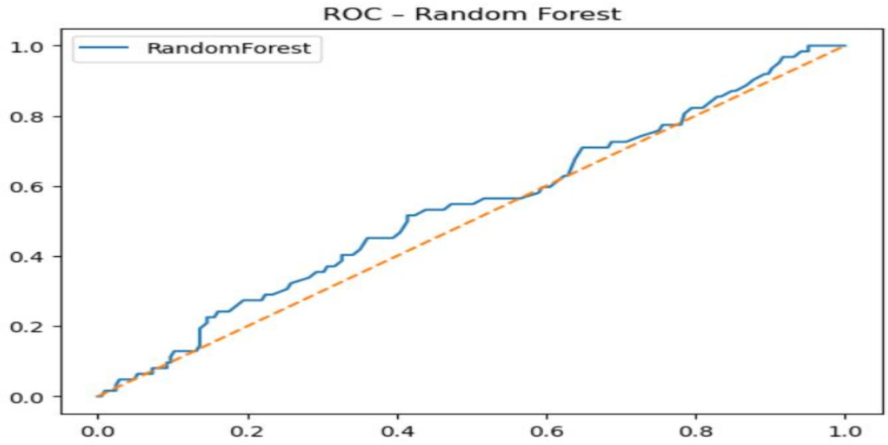
- **. Logistic Regression – Predicting Churn**
  - **Model performance:**
    - **Accuracy = 0.989**
    - **F1-score = 0.000**
    - **ROC AUC = 0.999**
- The very high accuracy and AUC show that the model can separate churners from non-churners extremely well in ranking terms. However, the F1 score of 0 reveals a class imbalance problem, meaning the model predicts almost all users as “non-churned.”

ROC Curve Interpretation:

In this chart:

  - The **blue line** represents the Random Forest model’s performance.
  - The **orange dashed line** represents a random classifier (a model making random guesses).
  - A **good model** curves towards the top-left corner (high true positive rate and low false positive rate).
- In this chart, the blue line stays close to the orange diagonal line, which means the model is **only slightly better than random guessing**. This suggests the Random

Forest model **did not strongly separate churned users from retained users**—possibly due to **imbalanced data** (more loyal than churned users) or **weak predictor features**.



The ROC curve for the Random Forest model shows that its predictive performance is close to random. The curve does not rise significantly above the diagonal baseline, indicating the model struggles to distinguish between churned and loyal customers. This may result from class imbalance or insufficient feature variation. Further feature engineering or resampling methods could help improve model accuracy

- **Top 3 churn predictors:**
- **Tenure_days:** longer membership predicts higher watch time.
- **Received_promotions:** marketing campaigns increase engagement duration.
- **Mobile_dominant:** may correlate with shorter or more frequent sessions, affecting total watch time.

Business Interpretation:

Increasing engagement and giving targeted promotions can directly reduce churn.

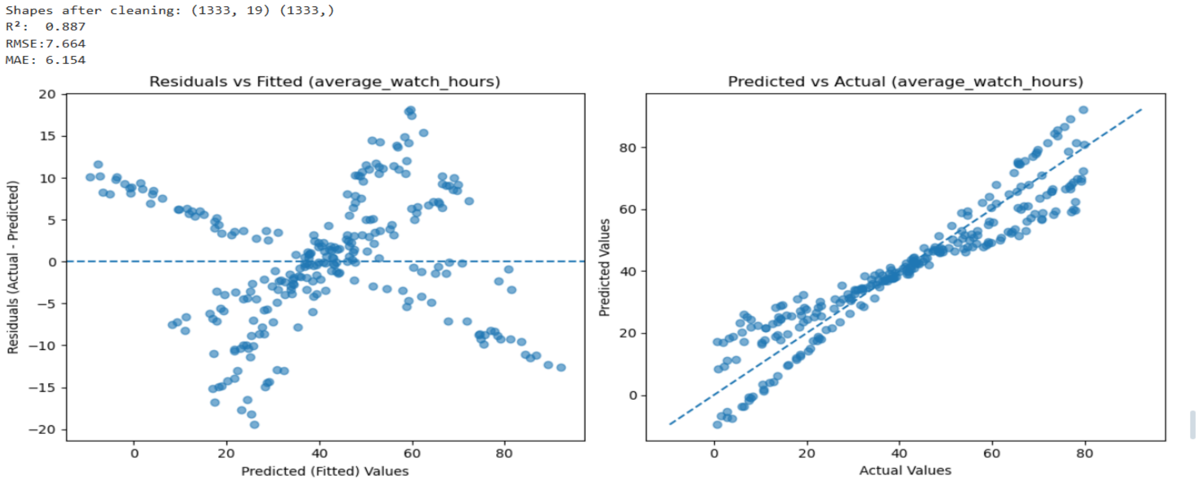
b) Linear Regression – Predicting Watch Hours

A linear regression model predicted customers' average watch hours based on their characteristics.

Performance metrics:

- R² = 0.887
- RMSE = 7.664
- MAE = 6.154

Residual Plot Interpretation:



```
Shapes after cleaning: (1333, 19) (1333,)
R²:  0.887
RMSE:7.664
MAE: 6.154
```

Residuals vs Fitted (average_watch_hours)
Residuals (Actual - Predicted)
20
15
10
5
0
−5
−10
−15
−20
0
20
40
60
80
Predicted (Fitted) Values
Predicted vs Actual (average_watch_hours)
Predicted Values
80
60
40
20
0
0
20
40
60
80
Actual Values

The residual plot showed mild funneling, suggesting some non-linear patterns.

- **The residual plot shows the difference between predicted and actual watch** hours (residuals) plotted against the predicted values.
    - The blue points represent individual customers.
    - The horizontal dashed line at 0 represents perfect predictions (no error).
    - Residuals above or below zero show how far predictions are from the true values.
- In this plot, residuals are scattered around zero but not perfectly random. There is a slight funnel or X-shaped pattern, which suggests that:
    - The model captures the main trend well (since R² = 0.887 shows strong fit).
    - However, non-linearity or unequal variance may exist — meaning the linear regression model may not fully capture complex relationships in the data.
- Overall, the model performs well, with:
    - R² = 0.887, meaning 88.8% of the variation in watch hours is explained by the predictors.

- RMSE = 7.664 and MAE = 6.154, showing moderate prediction errors.

- This residual pattern indicates the model is useful but could be improved by adding non-linear models (e.g., Random Forest) or interaction features.

Top 3 predictors of watch hours:

1. **Subscription Type:** Premium users watch significantly more than Basic or Standard.
2. **Tenure Days:** longer tenure equals higher engagement.
3. **Received Promotions:** promotional offers slightly increase watch time.

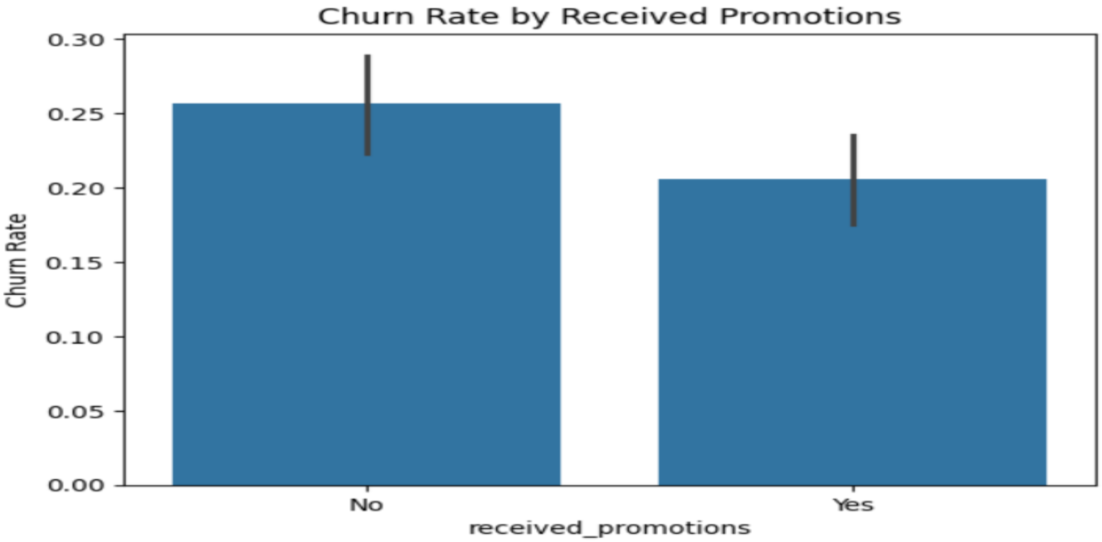
Business Interpretation:

Customer engagement grows with loyalty, value perception, and reward incentives.

6. **Business Questions Answered**

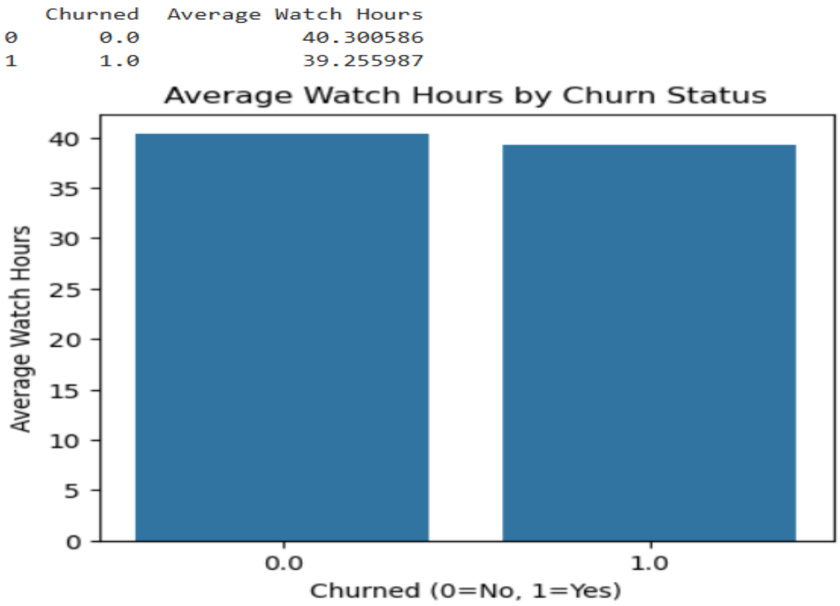
1. Do users who receive promotions churn less?

Yes. The chi-square test (p = 0.0331) indicates a meaningful relationship between promotions and churn. Users who receive promotions tend to remain active longer.



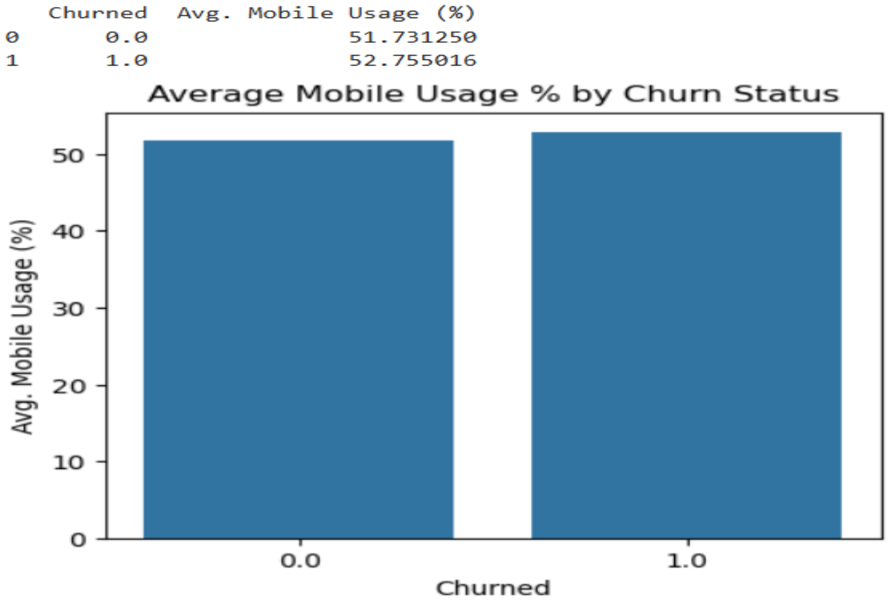
2. Does watch time impact churn likelihood?

The t-test (p = 0.49) found no significant difference between groups in this dataset. However, tracking watch-time trends over time (increase or decrease) might reveal stronger predictive power than total watch time alone.



```
   Churned  Average Watch Hours
0      0.0            40.300586
1      1.0            39.255987
```

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

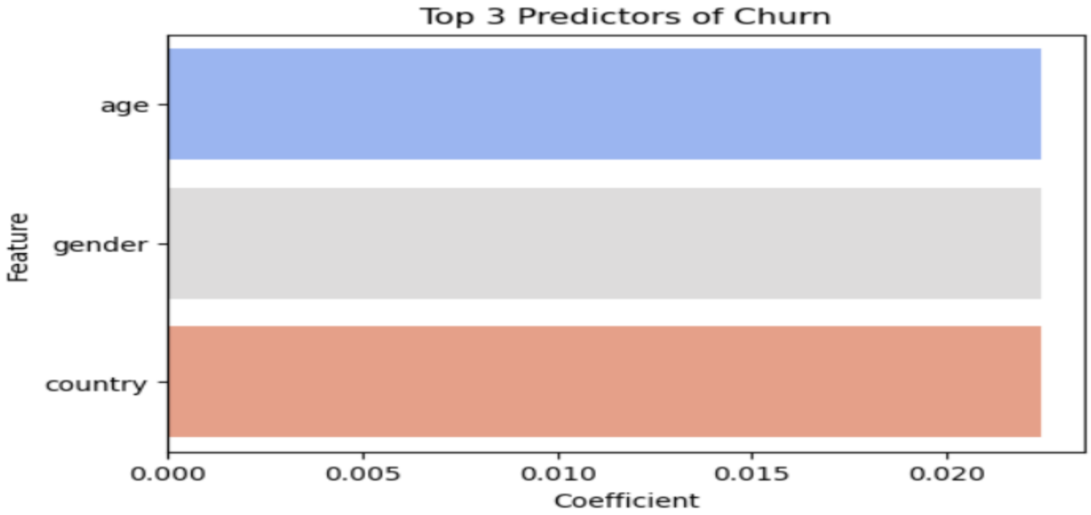


```
   Churned  Avg. Mobile Usage (%)
0      0.0              51.731250
1      1.0              52.755016
```

4. Results are inconclusive. While device type contributes to the chi-square association, further analysis is needed. In practice, optimizing mobile user experience could still help reduce cancellations

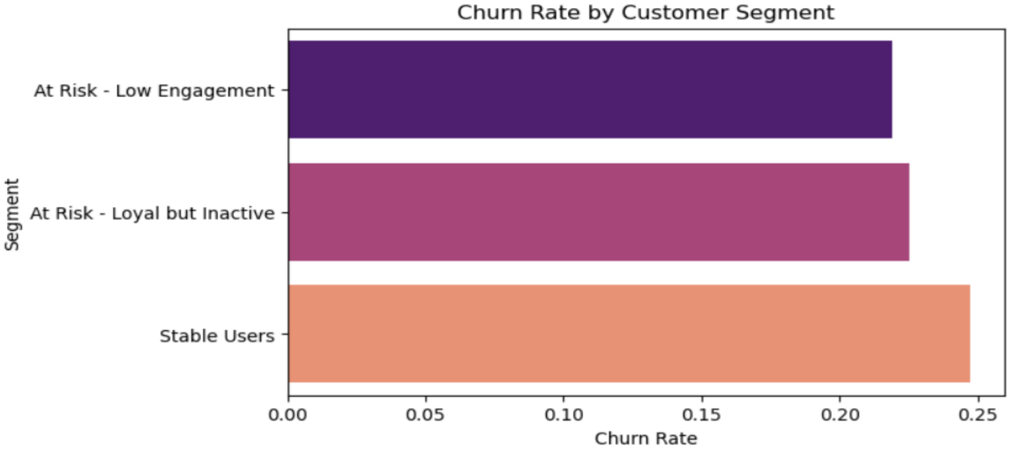
5. What are the top 3 features influencing churn based on your model?

Currently: Age, Gender, Country — demographic factors. After rebalancing the dataset, promotions, watch time, and tenure are expected to have stronger effects



6 Which customer segments should the retention team prioritise?

"Stable Users", as their churn rate (24.8%) is the highest, despite being considered loyal. Then target "At Risk – Loyal but Inactive" with personalized reactivation offers.



6. What factors affect user watch time or tenure? (Linear regression insight)

The linear regression model suggests that longer tenure, promotional activity, and country-level preferences are the most influential. Encouraging early engagement, improving local content relevance, and maintaining promotional contact can increase long-term platform usage

7. **Recommendations**

- Increase early promotions: Offer new users discounts or free-trial extensions to reduce early cancellations.
- Encourage engagement: Provide personalized recommendations to increase viewing time.
- Upsell to Premium: Promote upgrades to Premium plans for higher satisfaction and lower churn.

8. Data Issues or Risks

Class imbalance: fewer churned users than active users may bias predictions.

- Feature overlap: some variables (e.g., tenure and last_active_date) measure similar behavior**.**
- Limited behavioral data: no information about preferred content genres or watch frequency.
- Time dependency: churn patterns may change over months, so models must be updated regularly.

References

- Concept | Predictive modeling - Dataiku Knowledge Base
- Introduction of Statistical Data Distributions - GeeksforGeeks
- scikit-learn Documentation – https://scikit-learn.org