

Week 3 - Project: Churn Prediction for StreamWorks Media.

Please write the report in the 'Project Coversheet' and refer to the dataset provided for completing the tasks.

Deliverables

1. **Jupyter Notebook (.ipynb):** A clean, well-organised notebook that includes: Data loading and exploration, Data cleaning and preprocessing, Feature engineering (e.g. new variables, encodings), Summary tables and visualisations, Statistical analysis and predictive modelling, Explanations and insights written in markdown throughout the notebook
2. **PDF Report (max 1500 words):** A professional summary including key insights, tables and charts, business questions answered, and clear recommendations (optional screenshots of outputs. No code screenshots required).

Business Scenario

You've joined the Data Strategy Team at StreamWorks Media, a fast growing UK-based video streaming platform competing with global players like Netflix and Amazon Prime.

With rising customer acquisition costs and increasing competition, your manager wants you to investigate customer churn, users who cancel their subscriptions.

The business goal is twofold:

- Analyse churn behaviour: Identify who is churning and why.
- Build a model to predict future churners, so the retention team can take proactive steps.

This project will involve:

- Statistical analysis (correlation, hypothesis testing)
- Predictive modelling (logistic regression)
- Model evaluation using metrics like precision, recall, ROC-AUC

Dataset

streamworks_user_data.csv

Each row represents a unique subscriber and includes:

Column Name	Description
-------------	-------------

user_id	Unique user identifier
age	Age of the user
gender	Male, Female, Other
signup_date	Date user joined
last_active_date	Date of last login
country	User's country
subscription_type	Basic, Standard, Premium
monthly_fee	Amount paid monthly (£)
average_watch_hours	Avg. monthly watch time
mobile_app_usage_pct	% of viewing via mobile app
complaints_raised	No. of complaints submitted
received_promotions	Whether user received offers (Yes/No)
referred_by_friend	Yes/No
is_churned	1 if user cancelled in past 30 days, else 0

Tasks

1. Load & Explore the Data

- Use pandas to load the dataset
- Use `.info()`, `.describe()`, `.value_counts()`, `.isnull().sum()` to understand structure and missing values
- Create a correlation matrix and heatmap (e.g. `sns.heatmap()`) for numeric variables

2. Clean & Prepare the Data

- Convert `signup_date`, `last_active_date` to datetime
- Create new features:
 - `tenure_days` = days between `signup` and `last_active_date`
 - `is_loyal` = `tenure_days` > 180
- Encode categorical features (e.g. `LabelEncoder`, `pd.get_dummies()`)
- Fill or drop missing values, depending on context

3. Statistical Analysis & Insights

Perform and summarise the following:

- Use Chi-square test to check if churn is related to gender, `received_promotions`, or `referred_by_friend`
- Use a t-test to check if watch time differs significantly between churned and retained users
- Use charts (boxplots, bar plots, histograms) to visualise key differences between churned and active users

4. Predictive Modelling

Build a logistic regression model to predict `is_churned`:

- Split into training and test sets (`train_test_split()`)
- Scale features (`StandardScaler`)
- Fit a `LogisticRegression()` model
- Predict probabilities and classes
- Evaluate the model using:
 - Confusion Matrix
 - Precision, Recall, F1 Score
 - ROC Curve and AUC Score
- Identify the most important predictors of churn from model coefficients

5. Business Questions to Answer (in PDF Report)

1. Do users who receive promotions churn less?
2. Does watch time impact churn likelihood?
3. Are mobile dominant users more likely to cancel?
4. What are the top 3 features influencing churn based on your model?
5. Which customer segments should the retention team prioritise?

6. Optional Stretch Goals

- Use `GridSearchCV` to tune logistic regression hyperparameters
- Try a second model (e.g. Random Forest) and compare performance
- Segment churn by country or subscription type

Report Structure (Submit as a PDF file)

Please write your answer/ full report in the '**Project Coversheet**' and submit a concise and professional PDF report.

1. Introduction

- Describe the business goal and dataset
- State the purpose of your analysis

2. Data Cleaning Summary

- Mention changes made: column types, missing values, encoding
- Optional: include screenshot of `.info()` or `.isnull().sum()`

3. Feature Engineering Summary

- Briefly list and explain new features created (e.g., `tenure_days`, `is_loyal`, dummy variables)

4. Key Findings

- Summarise statistical findings from t-tests and chi-square tests
- Highlight any correlations or behavioural trends

5. Model Results

- Report model performance (accuracy, F1, AUC)
- Include ROC curve screenshot and explain output
- List top 3 predictors of churn and their business interpretation

6. Business Questions Answered

- Answer the 5 business questions above with evidence (tables or charts)

7. Recommendations

- Suggest 2-3 actionable strategies (e.g., target users with low watch time, boost promotions to new users)

8. Data Issues or Risks

- Mention any limitations or risks (e.g., data imbalance, feature leakage)

Note: Include screenshots from your jupyter notebook where required.

Submission Checklist

Before you submit, ensure you have:

- ✓ Your completed Jupyter Notebook (.ipynb)
- ✓ A PDF report following the structure above, written in the '**Project Coversheet**' as instructed

Final Tip: Explain your findings like you're advising a non-technical manager. Focus on what the data says about customer behaviour and what StreamWorks can do next.