dig data

a Data Inspiration Group programme

itv

Final ITVx challenge

Customer base habits and model predictions

Data Strategy Task

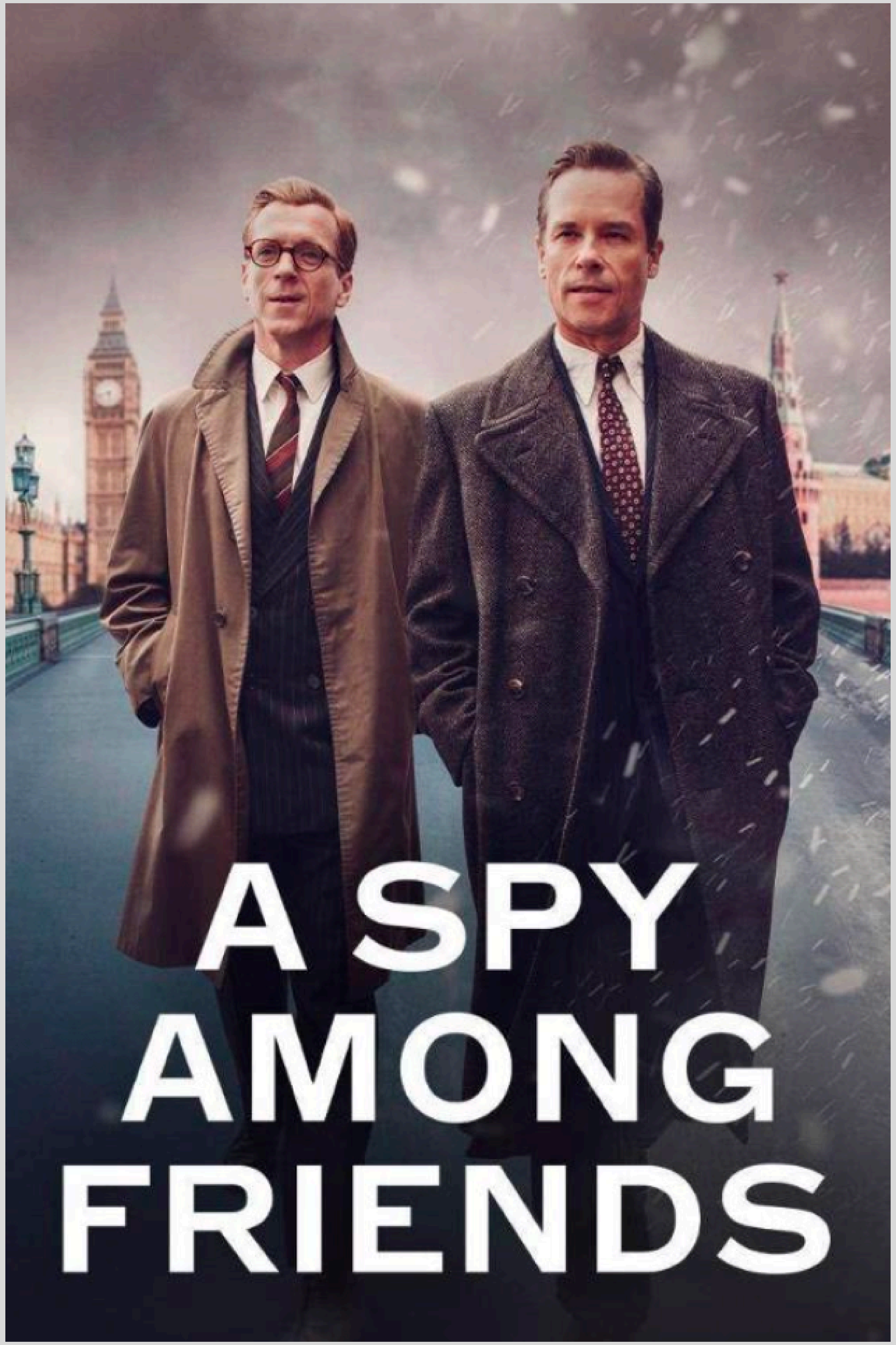
Analysing consumer behavior in relation to new movie ‘A Spy Among Friends’

Data Strategy

We want to know what the customer base viewing habits look like by profiling them

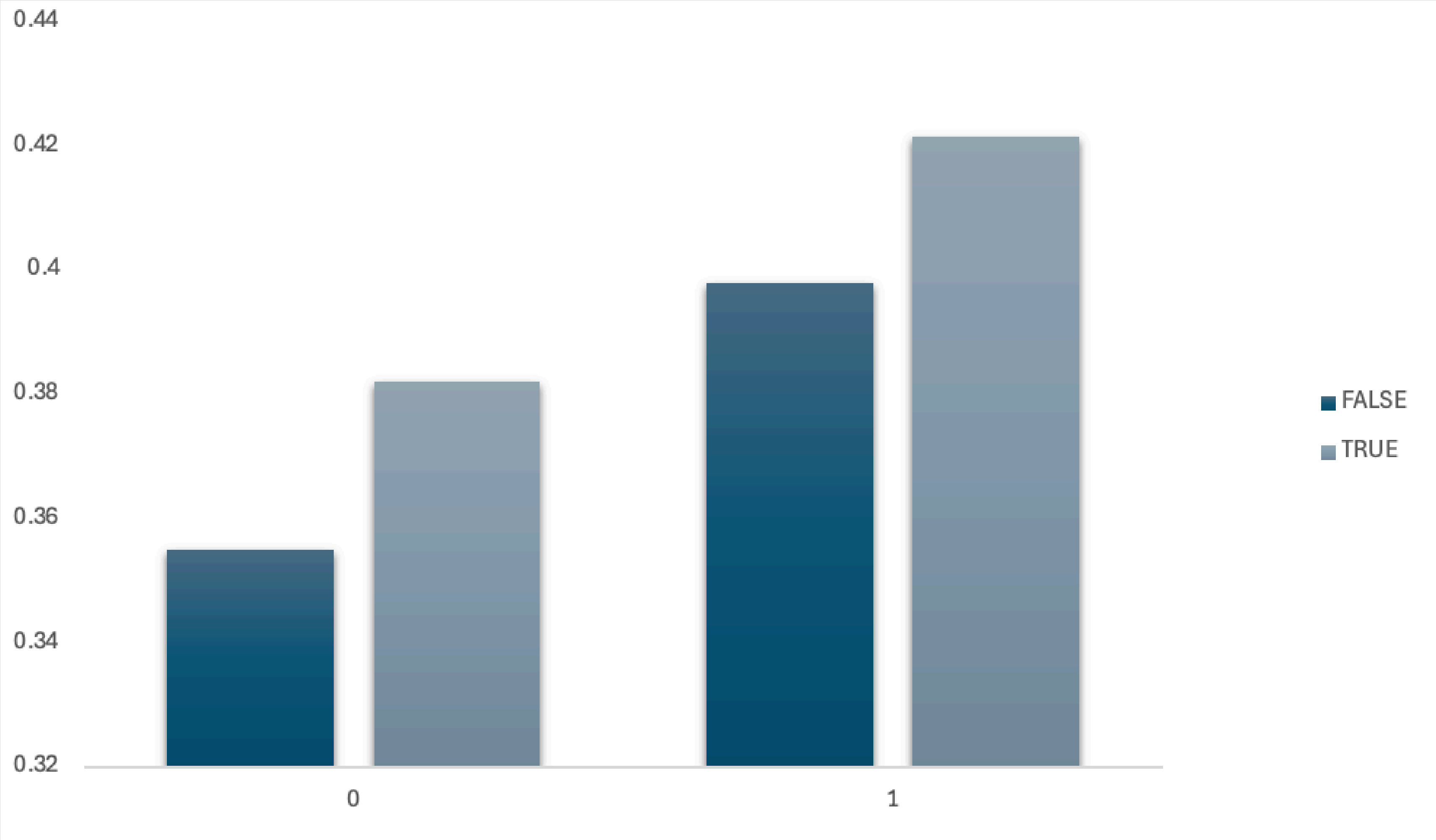
=>

make informed marketing decisions on who ITVX should target to watch the new originals for ITVX

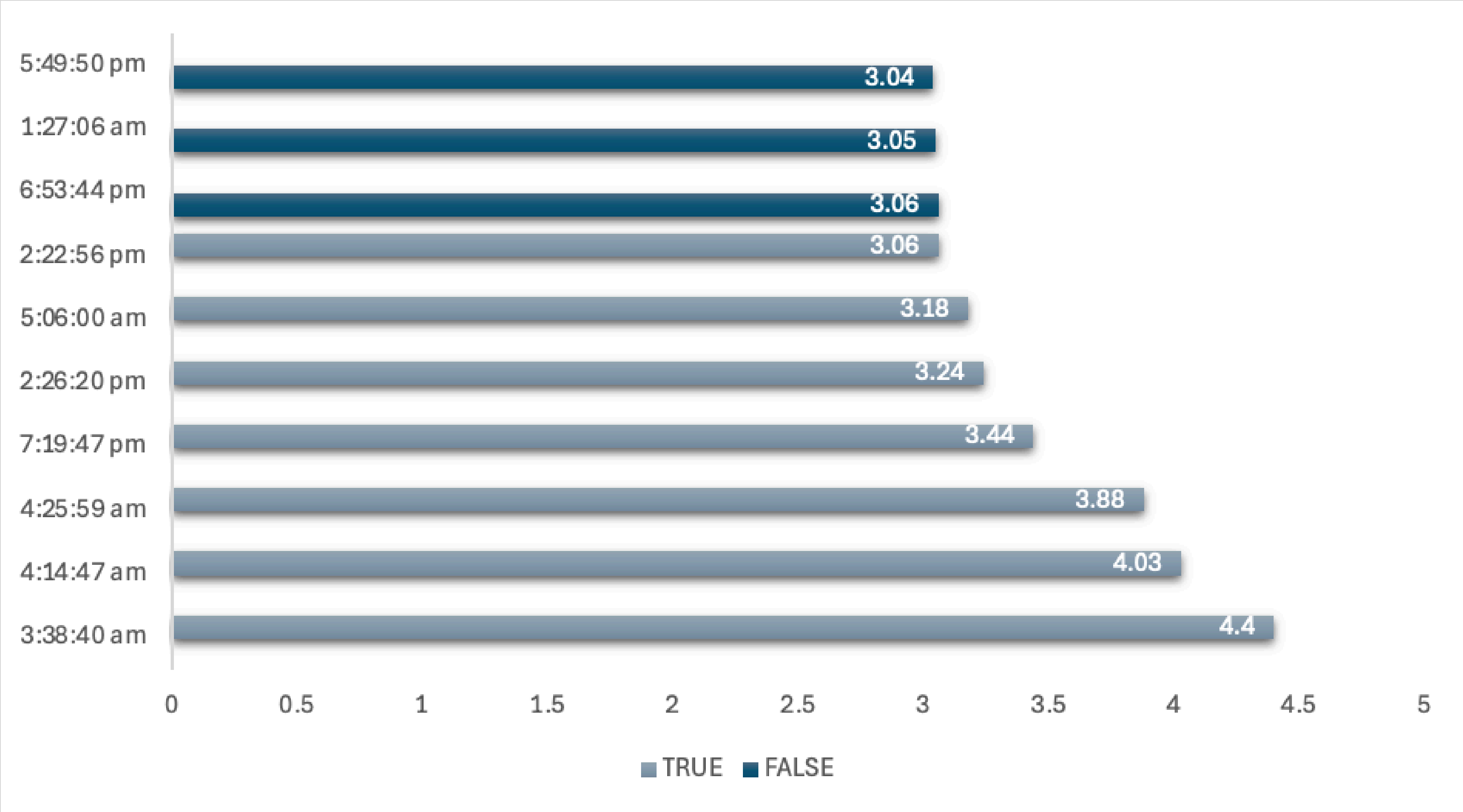


Part I – ITVX subscriber profile

Average Streaming Time in hours according to those who have watched the movie (1) vs those who haven’t (0) during the weekend (TRUE) or not (FALSE)



Common day time for streaming on ITVX during weekend (TRUE) or weekdays (FALSE) for those who have watched the movie

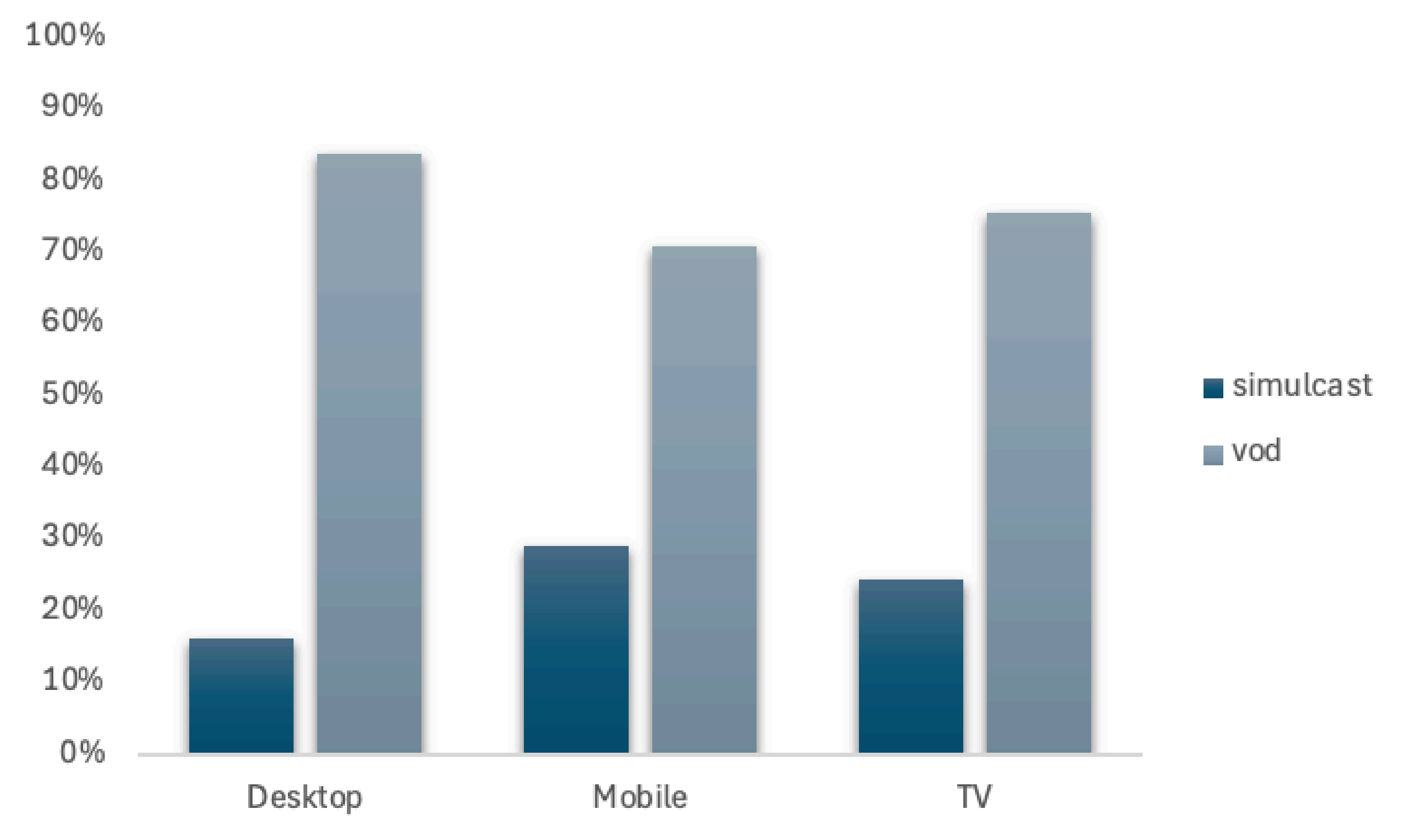


Notes:

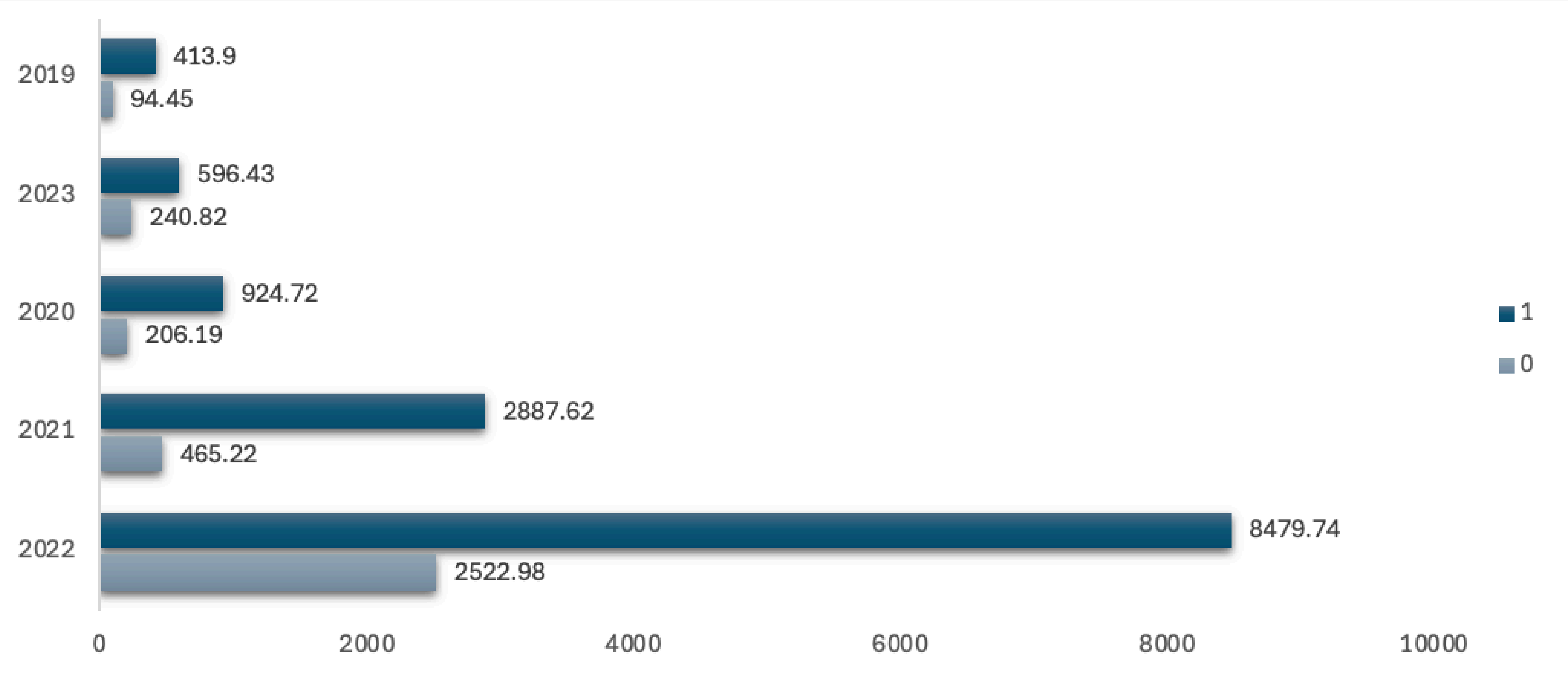
- Those who have watched the movie tend to have a higher overall average for streaming time than those who haven’t
- Subscribers have a higher streaming average during the weekend, with popular times being dispersed during early morning, afternoon, evening

Part I – ITVX subscriber streaming profile (cont)

Percentage of individuals streaming content, according to device type for those who have watched the movie



Total number of hours spent streaming the Top 5 year preferences for subscribers who watched the movie (1) and those who have not (0)

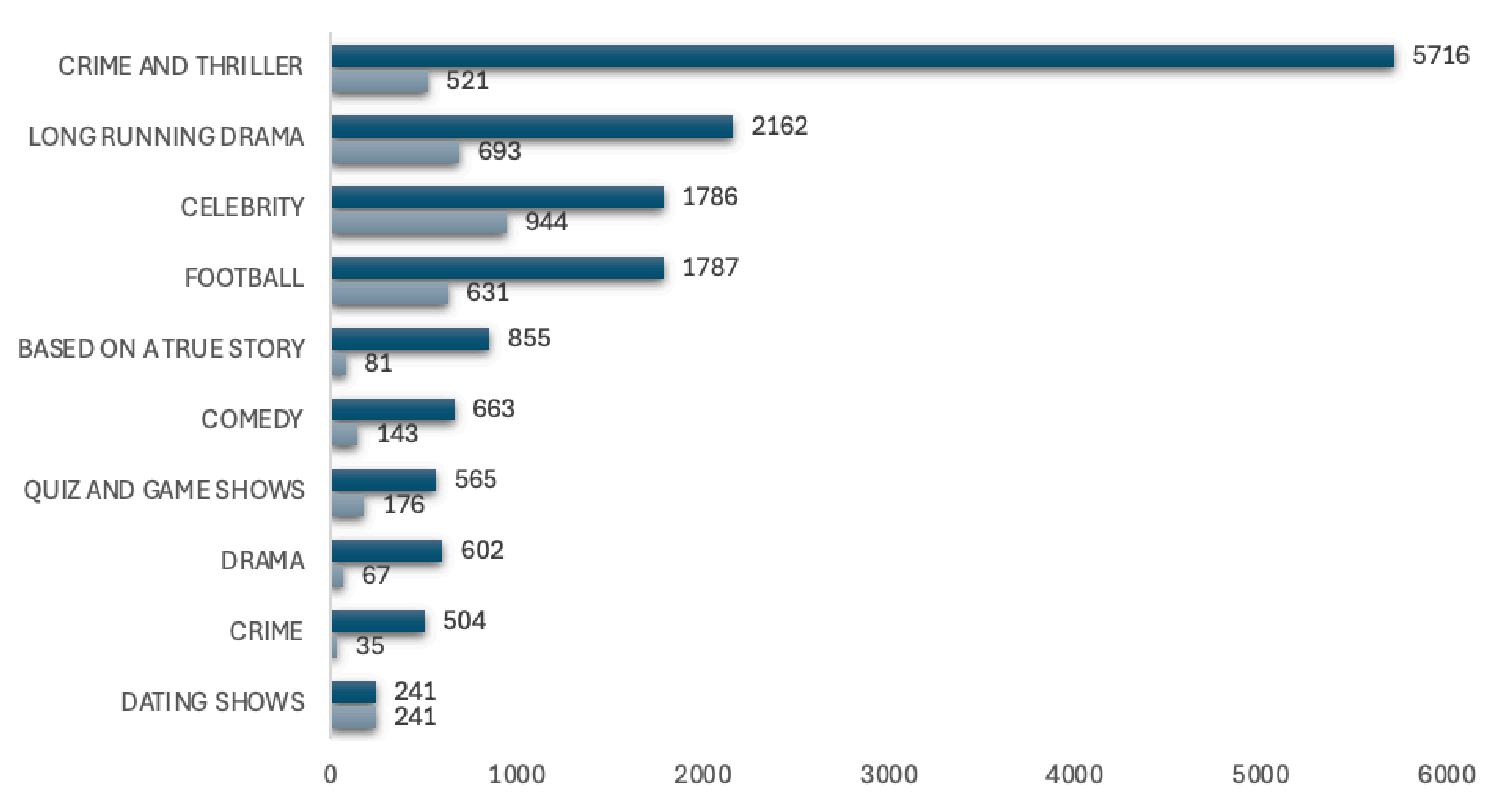


Notes:

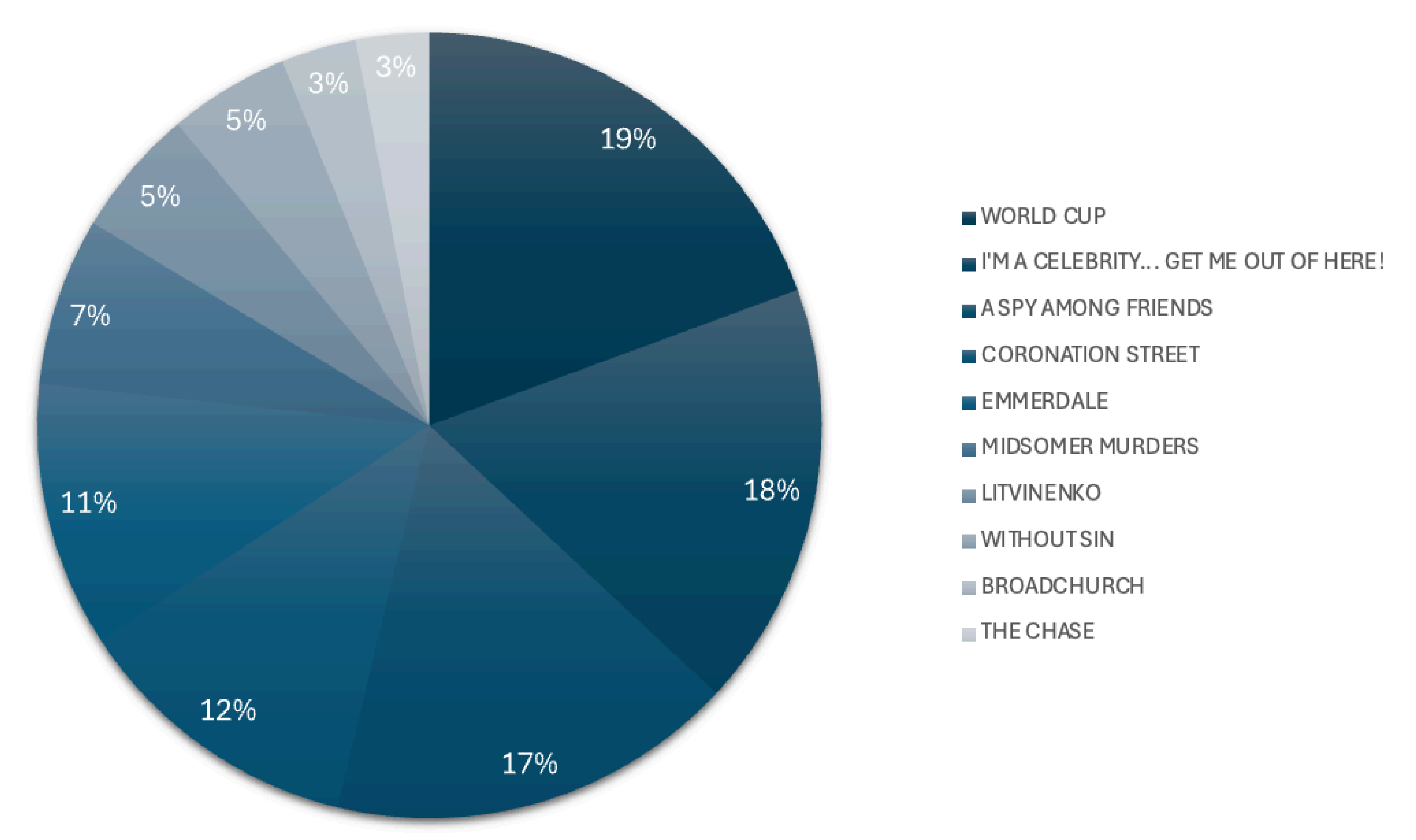
- Desktop seems to be the overall preference for subscribers to watch videos on demand, whereas for live content, mobile seems the go-to choice.
- Individuals prefer content around the same time that the movie was put out, with a clear front runner being year 2022, then 2021.

Part I – ITVX subscriber streaming profile (cont)

Overall Top 10 genre distribution according to total number of streaming hours for those who have watched the movie (1) and those who have not (0)



Overall Top 10 programmes distribution according to total number of streaming hours



Notes:

- For those who have watched the movie, popular genres to stream are the following: Crime and Thriller (38.42%), Long Running Drama (14.53%), Football (12.01%) and Celebrity (12%).
- For those who have not watched the movie, popular genres include: Celebrity (26.74%), Long Running Drama (19.62%), Football (17.86%) and Crime and Thriller (14.75%).
- ITVX audience heavily favour large-scale or long-running programmes, with World Cup, I'm a celebrity and Coronation Street accounting for over half of total streaming hours.

Part II – ML algorithms to predict show streaming

Task:

Design a model which predicts the likelihood of a user watching the new ITVX exclusive 'A Spy Among Friends'

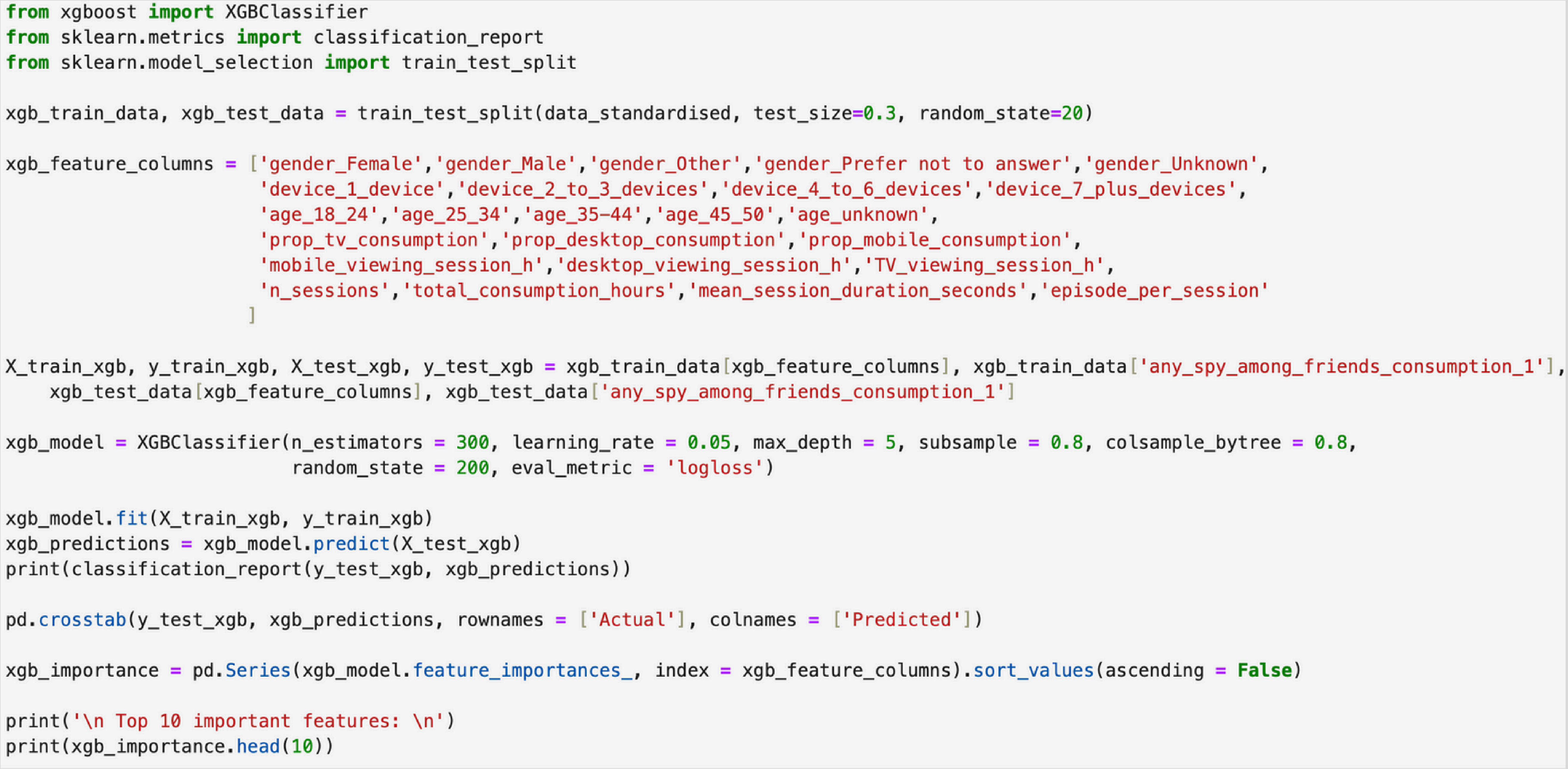
Created 3 different ML models with different feature sets to answer the following:

1. **Out of the three models, which one performed the best?**
2. **Which features played the most significant role in determining whether or not someone was a viewer?**
3. **Are there any additional data points that could be collected to enhance the model's performance?**
4. **Can you explain, in your own words, how each of these models work and their underlying mechanisms? Remember to do some research first to deepen your understanding.**

We'll present and examine each model in the next slides.

Part II – ML algorithms to predict show streaming

A. XGBoost model – demographics (age, gender, platform and device use)



```
from xgboost import XGBClassifier
from sklearn.metrics import classification_report
from sklearn.model_selection import train_test_split

xgb_train_data, xgb_test_data = train_test_split(data_standardised, test_size=0.3, random_state=20)

xgb_feature_columns = ['gender_Female','gender_Male','gender_Other','gender_Prefer not to answer','gender_Unknown',
                       'device_1_device','device_2_to_3_devices','device_4_to_6_devices','device_7_plus_devices',
                       'age_18_24','age_25_34','age_35-44','age_45_50','age_unknown',
                       'prop_tv_consumption','prop_desktop_consumption','prop_mobile_consumption',
                       'mobile_viewing_session_h','desktop_viewing_session_h','TV_viewing_session_h',
                       'n_sessions','total_consumption_hours','mean_session_duration_seconds','episode_per_session'
                      ]

X_train_xgb, y_train_xgb, X_test_xgb, y_test_xgb = xgb_train_data[xgb_feature_columns], xgb_train_data['any_spy_among_friends_consumption_1'],
    xgb_test_data[xgb_feature_columns], xgb_test_data['any_spy_among_friends_consumption_1']

xgb_model = XGBClassifier(n_estimators = 300, learning_rate = 0.05, max_depth = 5, subsample = 0.8, colsample_bytree = 0.8,
                          random_state = 200, eval_metric = 'logloss')

xgb_model.fit(X_train_xgb, y_train_xgb)
xgb_predictions = xgb_model.predict(X_test_xgb)
print(classification_report(y_test_xgb, xgb_predictions))

pd.crosstab(y_test_xgb, xgb_predictions, rownames = ['Actual'], colnames = ['Predicted'])

xgb_importance = pd.Series(xgb_model.feature_importances_, index = xgb_feature_columns).sort_values(ascending = False)

print('\n Top 10 important features: \n')
print(xgb_importance.head(10))
```

One-hot encoded the columns:

- age – 18-24, 25-34, 35-44, 45-50, unknown
- gender – female, male, other, prefer not to answer, unknown
- number of devices used – 1 device, 2-3 devices, 4-6 devices, 7+ devices

Numerical variables:

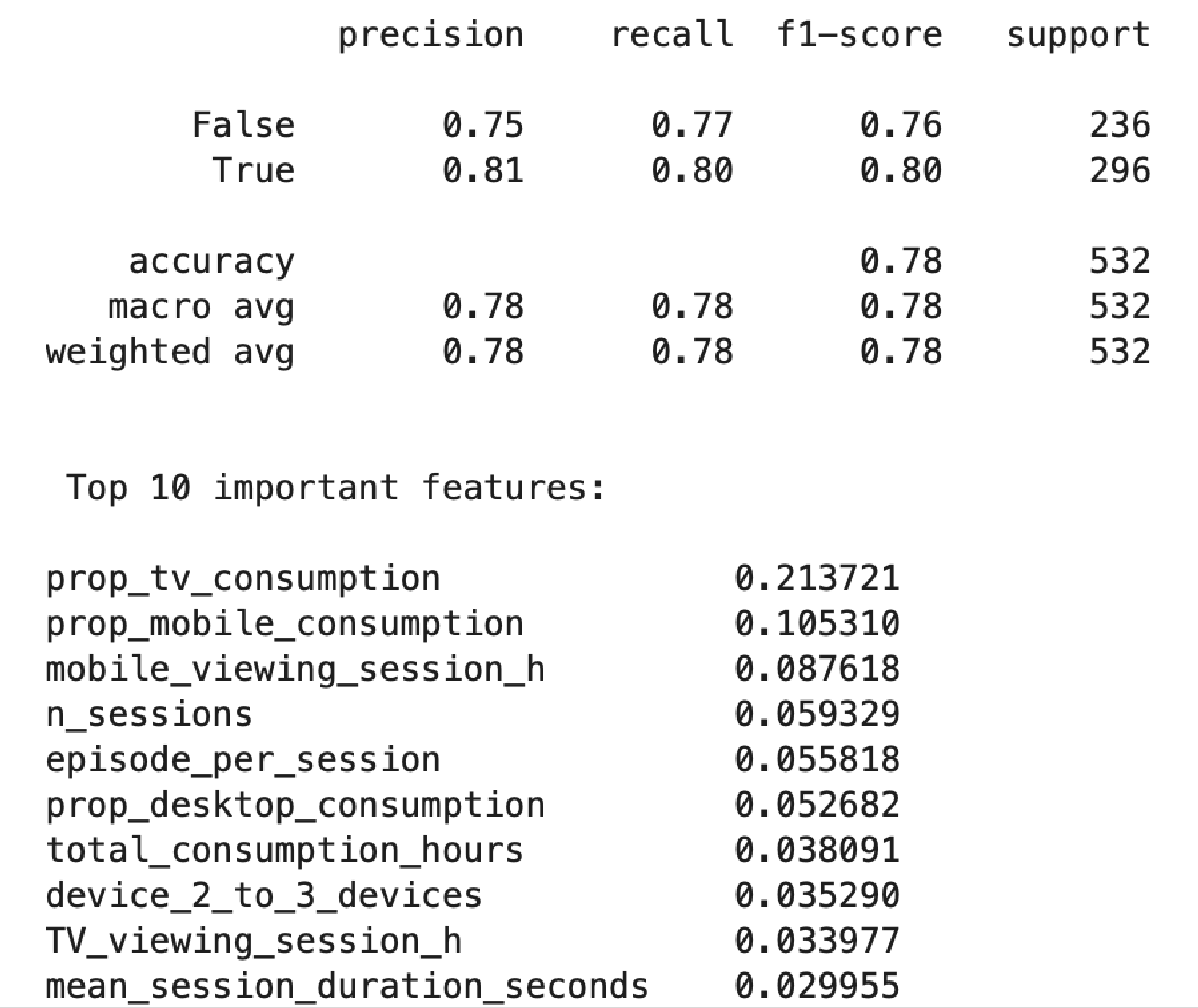
- proportion of consumption according to each platform – TV, desktop, mobile
- the amount of time (h) spent streaming on each platform – TV, desktop, mobile
- number of sessions
- total streaming time (h)
- average section length (s)
- average number of episodes per session

I used XGBoost for the demographics model because it handles high-dimensional categorical data extremely well, learns complex behavioral interactions, and consistently delivered the strongest and most stable predictive performance.

The model is designed to capture how demographic traits and platform + device habits predict whether a user has watched the target show, identifying subtle interaction patterns that simpler models may overlook.

Part II – ML algorithms to predict show streaming

A. XGBoost model – demographics (age, gender, platform and device use)



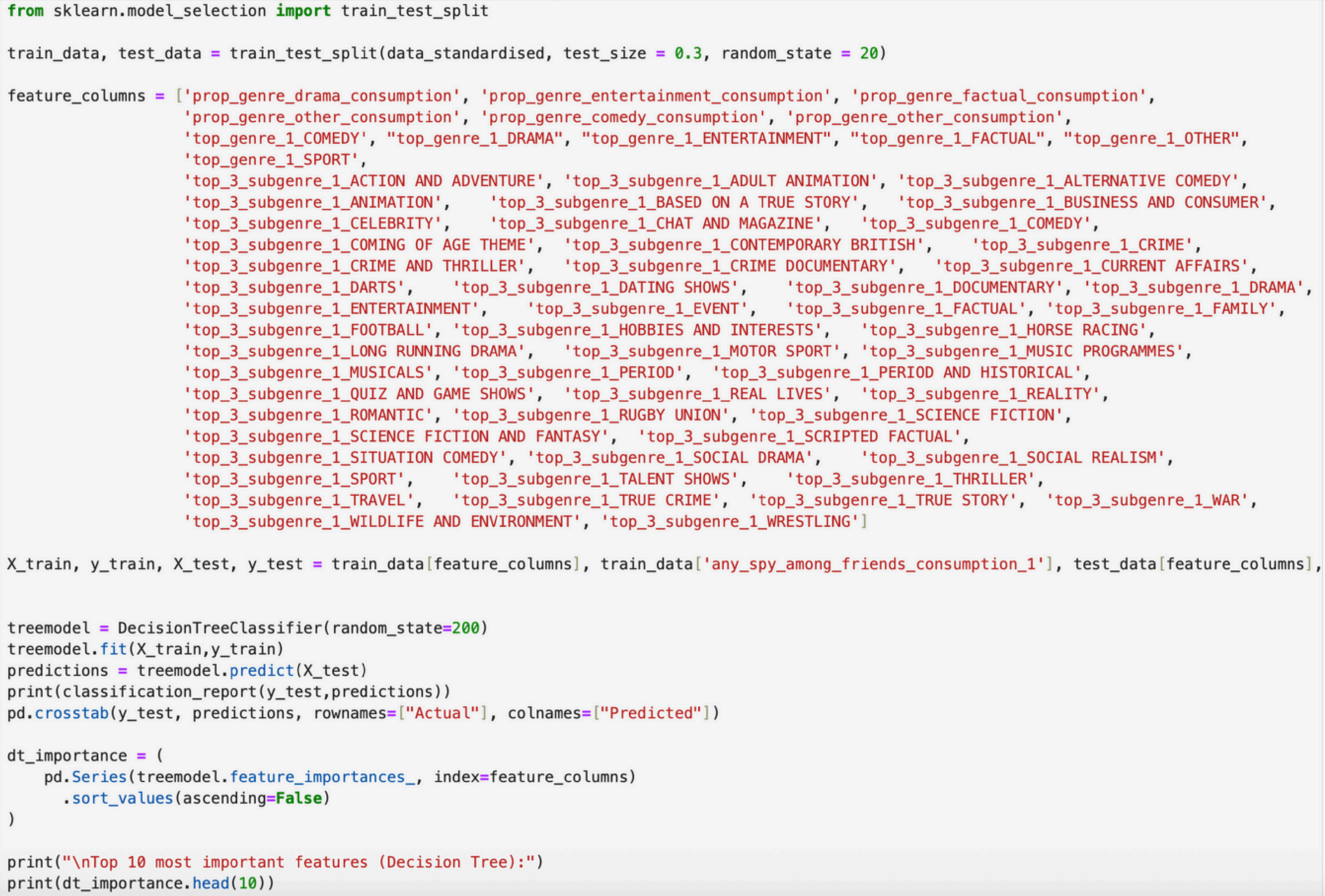
```
              precision    recall  f1-score   support

       False       0.75      0.77      0.76       236
        True       0.81      0.80      0.80       296

    accuracy                           0.78       532
   macro avg       0.78      0.78      0.78       532
weighted avg       0.78      0.78      0.78       532


 Top 10 important features:

prop_tv_consumption             0.213721
prop_mobile_consumption         0.105310
mobile_viewing_session_h        0.087618
n_sessions                      0.059329
episode_per_session             0.055818
prop_desktop_consumption        0.052682
total_consumption_hours         0.038091
device_2_to_3_devices           0.035290
TV_viewing_session_h            0.033977
mean_session_duration_seconds   0.029955
```

The XGBoost model achieved the following:

- **strong model performance –** overall accuracy of 78%, with balanced precision and recall for both viewers (80%) and non-viewers (75%)
- **platform choice is the top predictor –** higher proportions of TV and mobile consumption strongly increase the likelihood of watching the show
- **device count helps –** users with 2-3 devices show different viewing patterns, but the influence is modest compared to platform behavior
- **demographics play a minimal role –** behavioral variables overwhelmingly outperform demographic factors in predicting viewer likelihood.

Part II – ML algorithms to predict show streaming

B. Decision Tree model – genre and subgenre



```
from sklearn.model_selection import train_test_split

train_data, test_data = train_test_split(data_standardised, test_size = 0.3, random_state = 20)

feature_columns = ['prop_genre_drama_consumption', 'prop_genre_entertainment_consumption', 'prop_genre_factual_consumption',
                   'prop_genre_other_consumption', 'prop_genre_comedy_consumption', 'prop_genre_other_consumption',
                   'top_genre_1_COMEDY', "top_genre_1_DRAMA", "top_genre_1_ENTERTAINMENT", "top_genre_1_FACTUAL", "top_genre_1_OTHER",
                   'top_genre_1_SPORT',
                   'top_3_subgenre_1_ACTION AND ADVENTURE', 'top_3_subgenre_1_ADULT ANIMATION', 'top_3_subgenre_1_ALTERNATIVE COMEDY',
                   'top_3_subgenre_1_ANIMATION',    'top_3_subgenre_1_BASED ON A TRUE STORY',   'top_3_subgenre_1_BUSINESS AND CONSUMER',
                   'top_3_subgenre_1_CELEBRITY',    'top_3_subgenre_1_CHAT AND MAGAZINE',   'top_3_subgenre_1_COMEDY',
                   'top_3_subgenre_1_COMING OF AGE THEME',  'top_3_subgenre_1_CONTEMPORARY BRITISH',    'top_3_subgenre_1_CRIME',
                   'top_3_subgenre_1_CRIME AND THRILLER',   'top_3_subgenre_1_CRIME DOCUMENTARY',   'top_3_subgenre_1_CURRENT AFFAIRS',
                   'top_3_subgenre_1_DARTS',    'top_3_subgenre_1_DATING SHOWS',    'top_3_subgenre_1_DOCUMENTARY', 'top_3_subgenre_1_DRAMA',
                   'top_3_subgenre_1_ENTERTAINMENT',    'top_3_subgenre_1_EVENT',   'top_3_subgenre_1_FACTUAL', 'top_3_subgenre_1_FAMILY',
                   'top_3_subgenre_1_FOOTBALL', 'top_3_subgenre_1_HOBBIES AND INTERESTS',   'top_3_subgenre_1_HORSE RACING',
                   'top_3_subgenre_1_LONG RUNNING DRAMA',   'top_3_subgenre_1_MOTOR SPORT', 'top_3_subgenre_1_MUSIC PROGRAMMES',
                   'top_3_subgenre_1_MUSICALS', 'top_3_subgenre_1_PERIOD',  'top_3_subgenre_1_PERIOD AND HISTORICAL',
                   'top_3_subgenre_1_QUIZ AND GAME SHOWS',  'top_3_subgenre_1_REAL LIVES',  'top_3_subgenre_1_REALITY',
                   'top_3_subgenre_1_ROMANTIC', 'top_3_subgenre_1_RUGBY UNION', 'top_3_subgenre_1_SCIENCE FICTION',
                   'top_3_subgenre_1_SCIENCE FICTION AND FANTASY',  'top_3_subgenre_1_SCRIPTED FACTUAL',
                   'top_3_subgenre_1_SITUATION COMEDY', 'top_3_subgenre_1_SOCIAL DRAMA',    'top_3_subgenre_1_SOCIAL REALISM',
                   'top_3_subgenre_1_SPORT',    'top_3_subgenre_1_TALENT SHOWS',    'top_3_subgenre_1_THRILLER',
                   'top_3_subgenre_1_TRAVEL',   'top_3_subgenre_1_TRUE CRIME',  'top_3_subgenre_1_TRUE STORY',  'top_3_subgenre_1_WAR',
                   'top_3_subgenre_1_WILDLIFE AND ENVIRONMENT', 'top_3_subgenre_1_WRESTLING']

X_train, y_train, X_test, y_test = train_data[feature_columns], train_data['any_spy_among_friends_consumption_1'], test_data[feature_columns],


treemodel = DecisionTreeClassifier(random_state=200)
treemodel.fit(X_train,y_train)
predictions = treemodel.predict(X_test)
print(classification_report(y_test,predictions))
pd.crosstab(y_test, predictions, rownames=["Actual"], colnames=["Predicted"])

dt_importance = (
    pd.Series(treemodel.feature_importances_, index=feature_columns)
      .sort_values(ascending=False)
)

print("\nTop 10 most important features (Decision Tree):")
print(dt_importance.head(10))
```

One-hot encoded the columns:

- genre preferences – comedy, drama, entertainment, factual, other, sport
- sub-genre preferences – 50+ categories

Numerical variables:

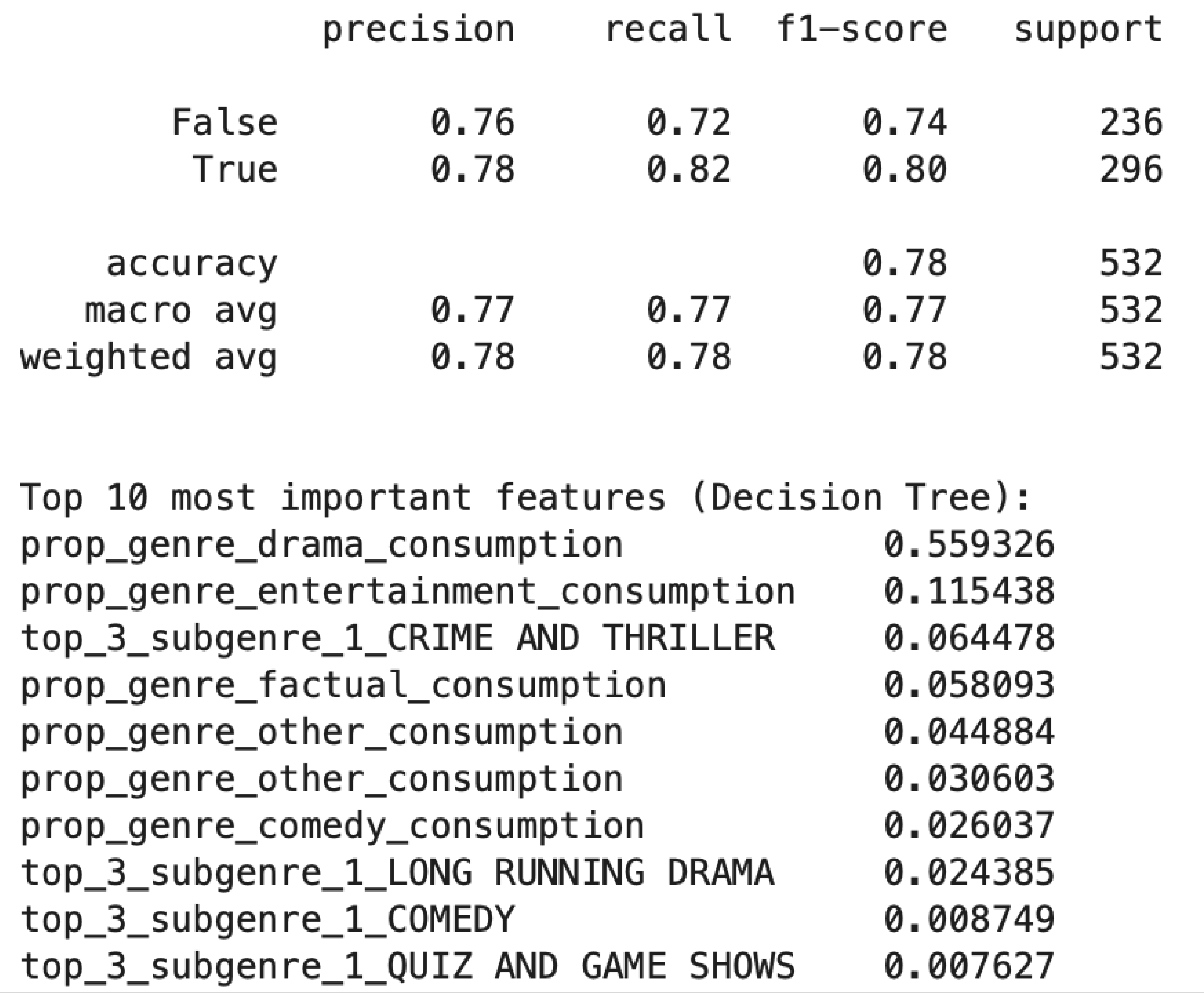
- proportion of streaming according to each genre – comedy, drama, entertainment, factual, other, sport

I used a Decision Tree for the content-preference model because it naturally handles mixed data types – binary genre and sub-genre flags alongside continuous viewing proportions; and produces an interpretable set of 'if-then' rules.

My Decision Tree is designed to capture how a user's preferred genres and subgenres relate to their likelihood of watching the target show, by learning the specific content patterns that differentiate viewers from non-viewers. It identifies which genres and niche subgenres act as strongest predictors, revealing content themes that align most closely with interest in 'A Spy Among Friends'

Part II – ML algorithms to predict show streaming

B. Decision Tree model – genre and subgenre



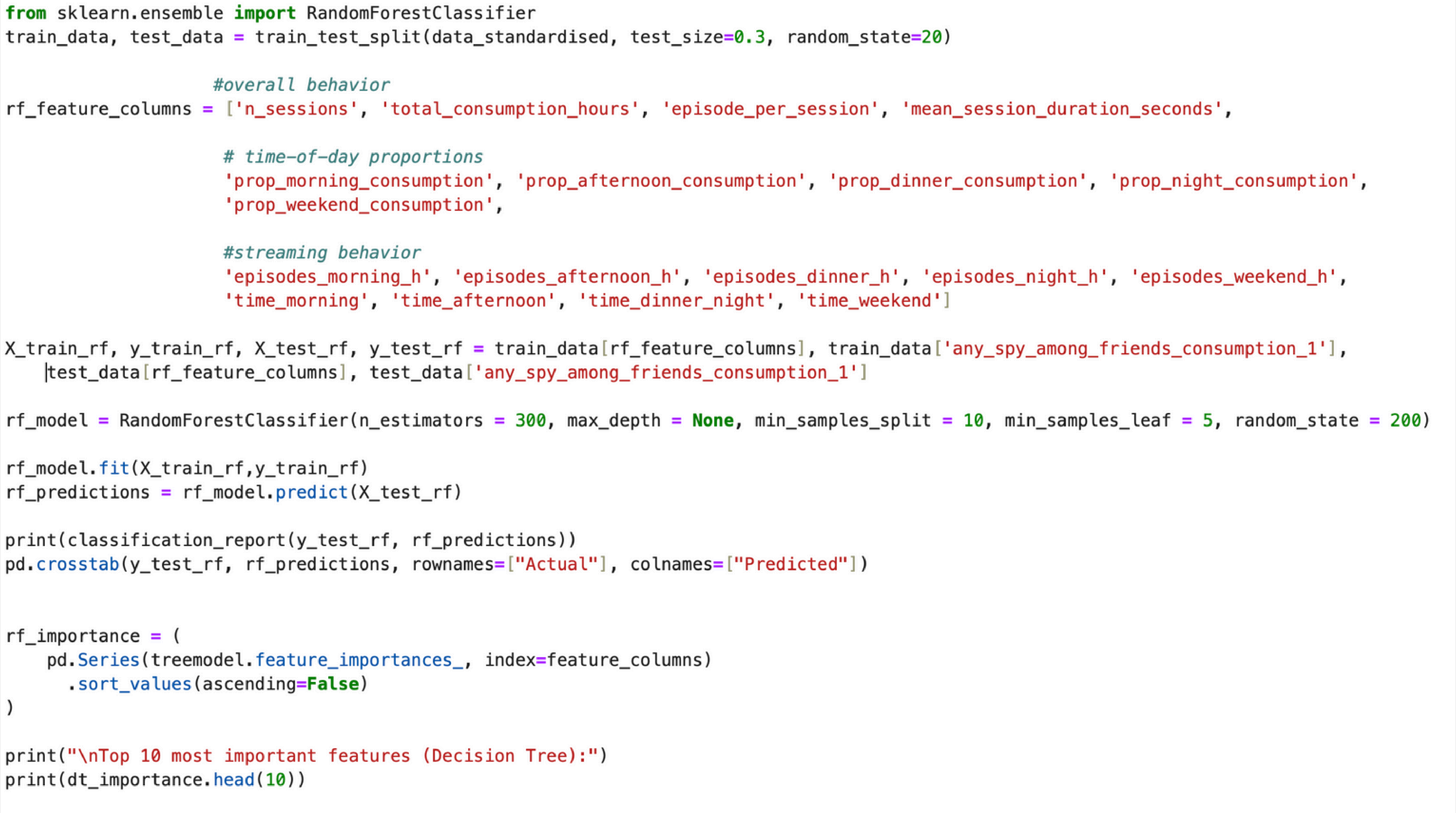
```
              precision    recall  f1-score   support

       False       0.76      0.72      0.74       236
        True       0.78      0.82      0.80       296

    accuracy                           0.78       532
   macro avg       0.77      0.77      0.77       532
weighted avg       0.78      0.78      0.78       532


Top 10 most important features (Decision Tree):
prop_genre_drama_consumption            0.559326
prop_genre_entertainment_consumption    0.115438
top_3_subgenre_1_CRIME AND THRILLER     0.064478
prop_genre_factual_consumption          0.058093
prop_genre_other_consumption            0.044884
prop_genre_other_consumption            0.030603
prop_genre_comedy_consumption           0.026037
top_3_subgenre_1_LONG RUNNING DRAMA     0.024385
top_3_subgenre_1_COMEDY                 0.008749
top_3_subgenre_1_QUIZ AND GAME SHOWS    0.007627
```

The Decision Tree model achieved the following:

- **strong model performance –** overall accuracy of 77% for behavioral prediction, with balanced precision and recall for both viewers (82%) and non-viewers (76%)
- **Drama is the strongest indicator** – viewers who spend a higher share of their time watching Drama are significantly more likely to watch the target show
- **Entertainment is the second strongest signal** – subscribers display interest in lighter, popular, mainstream formats correlates with watching the show
- **Crime & Thriller are highly predictive** – aligns perfectly with the tone and narrative themes of ‘A Spy Among Friends’
- **Comedy & Other genres have smaller, meaningful impact** – ITVX viewers have diverse streaming patterns
- **Long Running Drama is a niche predictor** – viewers who follow extended story arcs might be more engaged with the target series
- **Comedy & Quiz and Game Shows** – unexpected but valuable behavioral cross-over

Part II – ML algorithms to predict show streaming

C. Random Forest model – time preference and streaming behavior



```
from sklearn.ensemble import RandomForestClassifier
train_data, test_data = train_test_split(data_standardised, test_size=0.3, random_state=20)

                    #overall behavior
rf_feature_columns = ['n_sessions', 'total_consumption_hours', 'episode_per_session', 'mean_session_duration_seconds',

                     # time-of-day proportions
                     'prop_morning_consumption', 'prop_afternoon_consumption', 'prop_dinner_consumption', 'prop_night_consumption',
                     'prop_weekend_consumption',

                     #streaming behavior
                     'episodes_morning_h', 'episodes_afternoon_h', 'episodes_dinner_h', 'episodes_night_h', 'episodes_weekend_h',
                     'time_morning', 'time_afternoon', 'time_dinner_night', 'time_weekend']

X_train_rf, y_train_rf, X_test_rf, y_test_rf = train_data[rf_feature_columns], train_data['any_spy_among_friends_consumption_1'],
    test_data[rf_feature_columns], test_data['any_spy_among_friends_consumption_1']

rf_model = RandomForestClassifier(n_estimators = 300, max_depth = None, min_samples_split = 10, min_samples_leaf = 5, random_state = 200)

rf_model.fit(X_train_rf,y_train_rf)
rf_predictions = rf_model.predict(X_test_rf)

print(classification_report(y_test_rf, rf_predictions))
pd.crosstab(y_test_rf, rf_predictions, rownames=["Actual"], colnames=["Predicted"])


rf_importance = (
    pd.Series(treemodel.feature_importances_, index=feature_columns)
      .sort_values(ascending=False)
)

print("\nTop 10 most important features (Decision Tree):")
print(dt_importance.head(10))
```

One-hot encoded the columns:

- time preference – morning, afternoon, dinner&night, weekend

Numerical variables:

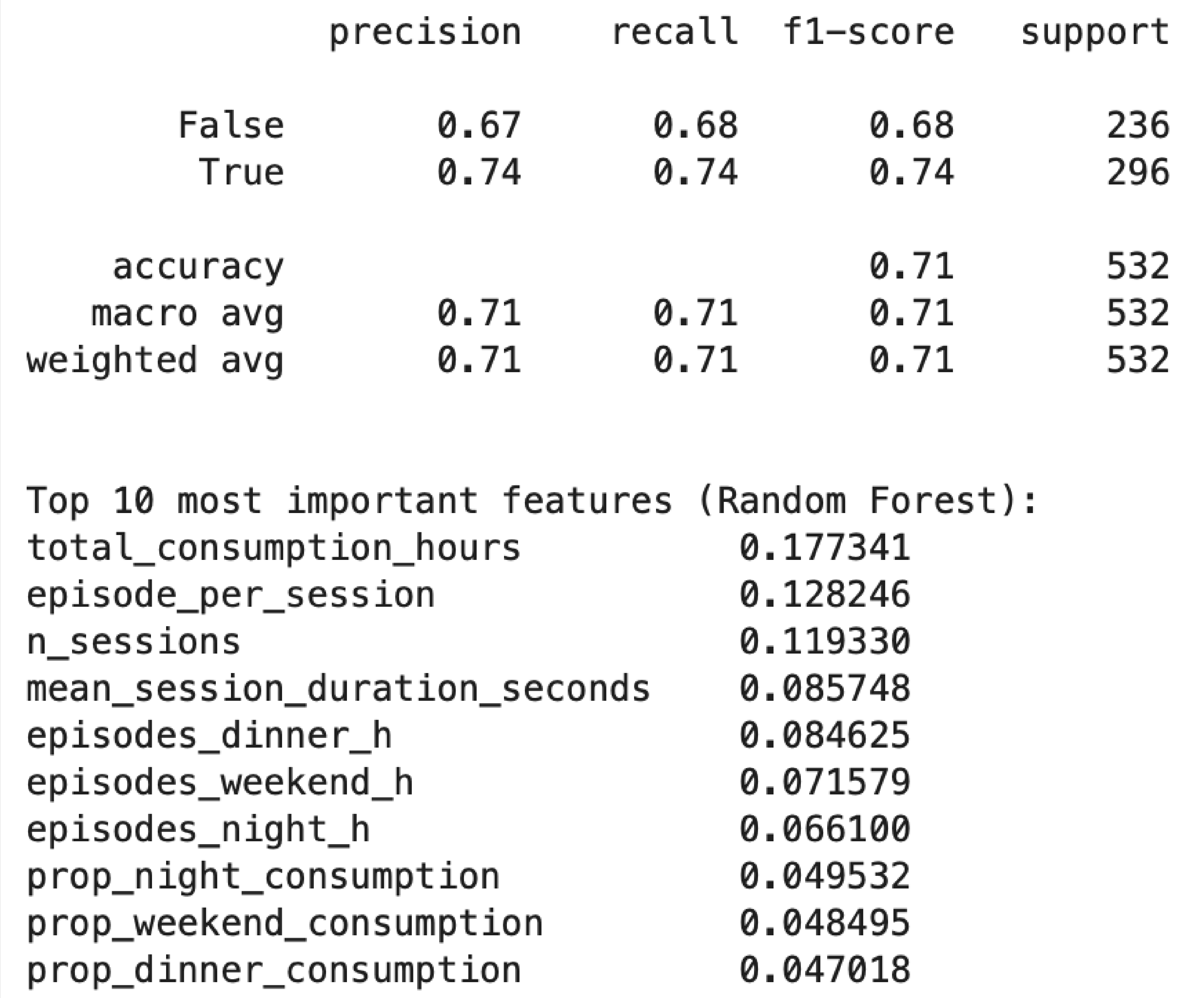
- overall viewing behavior – number of sessions and episodes per session, total number of streaming hours and average session duration (s)
- time-of-day consumption proportions – morning, afternoon, dinner, night, weekend
- session structure by time-of-day – number of episodes during the morning, afternoon, dinner, night, weekend

I used Random Forest for the time-preference and behavioral model because it captures non-linear viewing patterns, handles many correlated numerical features effectively, and reduces the risk of overfitting that a single tree would introduce. It aggregates hundreds of decision trees to learn stable viewing habits, making it ideal for identifying when users prefer to stream and how their session behavior relates to watching the target show.

This model is designed to uncover how a user's time-of-day habits and session behaviors influence their likelihood of watching 'A Spy Among Friends' . It detects patterns such as: which time windows correlate most with viewership, whether heavier & more consistent or more fragmented sessions predict interest and how episode pacing, device behavior and platform engagement affect the likelihood of watching the target show.

Part II – ML algorithms to predict show streaming

B. Decision Tree model – time-of-day preference



```
              precision    recall  f1-score   support

       False       0.67      0.68      0.68       236
        True       0.74      0.74      0.74       296

    accuracy                           0.71       532
   macro avg       0.71      0.71      0.71       532
weighted avg       0.71      0.71      0.71       532


Top 10 most important features (Random Forest):
total_consumption_hours          0.177341
episode_per_session              0.128246
n_sessions                       0.119330
mean_session_duration_seconds    0.085748
episodes_dinner_h                0.084625
episodes_weekend_h               0.071579
episodes_night_h                 0.066100
prop_night_consumption           0.049532
prop_weekend_consumption         0.048495
prop_dinner_consumption          0.047018
```

The Decision Tree model achieved the following:

- **strong model performance –** overall accuracy of 781% for behavioral prediction, with balanced precision and recall for both viewers (74%) and non-viewers (67%)
- **total streaming hours is the strongest predictor –** heavier consumption predicts higher likelihood of viewing the target series
- **longer or more consistent sessions matter** – users who watch for longer periods of time or watch multiple episodes at once show higher interest, representing a strong behavioral signal
- **Evening/Weekend episodes consumption correlate with viewership** – subscribers tend to engage during high-leisure, relaxed time windows, including dinner, night, weekend
- **night-time consumption proportion is a meaningful indicator** – late-evening viewers are more likely to be drawn to suspense or drama-driven films
- **Daytime viewers are weaker predictors** – indicator that ITVX customers may prefer lighter or shorter content in the morning and afternoon

ML model comparison

XGBoost

Decision Tree

Random Forest

Which model performed the best?

The model achieved the strongest overall performance, with an accuracy of .78, with a balanced precision and recall for both classes

The model obtained .78 accuracy, but was less stable and more influenced by dominant predictors

The model performed the weakest – .71 and both precision and recall were noticeably lower

Which features were most significant in predicting viewers?

- proportion of TV consumption
- proportion of mobile consumption
- mobile viewing session length
- total number of sessions = engagement
- episodes per session = binge tendency
- proportion of desktop consumption
- total streaming hours
- number of devices = 2-3 devices

=> device diversity + platform engagement + viewing intensity are powerful indicators of interest

- drama consumption proportion
- entertainment consumption proportion
- Crime & Thriller subgenre
- factual consumption
- Long Running Drama, Comedy, Quiz & Game shows

=> clear thematic alignment between content preferences and interest in ‘A Spy Among Friends’

- total consumption hours
- episodes per session
- number of sessions
- mean session duration
- dinner / night viewing habits
- weekend viewing habits

=> highlights the importance of when users watch and how consistently they stream

ML model comparison

How does each model work?

XGBoost

- builds trees sequentially, each correcting the errors of the previous one
- powerful for high-dimensional, mixed categorical + numerical data
- learns complex interactions between devices, platforms, session patterns, demographics

Decision Tree

- splits users into groups based on ‘if-then’ rules
- finds the single strongest content preferences that separate viewers from non-viewers
- interpretable but can overfit
- best used when you want clarity and transparent rules

Random Forest

- builds hundreds of trees and averages their outcome
- reduces overfitting and captures more nuanced patterns
- good at handling many behavioral metrics
- best used for understanding behavioral consistency and intensity

Additional data that could improve model performance

To enhance prediction accuracy, ITVX could ensure the collection of:

- more granular content preferences – likes and dislikes, historical preferences over the past 1-2 months
- stronger demographics signals – hosuehold size & shared-device patterns; individual data collection for each user for a shared device (similar to how Netflix has different profiles, each profile would be a new user_id entry)
- behavioral intent indicators – search behavior on the platform, trailer views and abandon points, watchlist / favourites
- engagement quality signals – completion rate, rewatch frequency, skipping / fast-forwarding patterns