Comparing Supervised Learning Models to predict Netflix Customer Churn

Boram Gaudet, Chris Gu, Jason Jia, Srita Kothuri

Introduction & Research Question

Research Question:
What behavioral characteristics are strongly associated with churn among Netflix customers?

Why is this Important?:
By understanding the characteristics that can affect customer retention, Netflix can tailor future digital experiences, personalize offers, and optimize revenue strategies to keep customers engaged.

Data Collection & Preparation

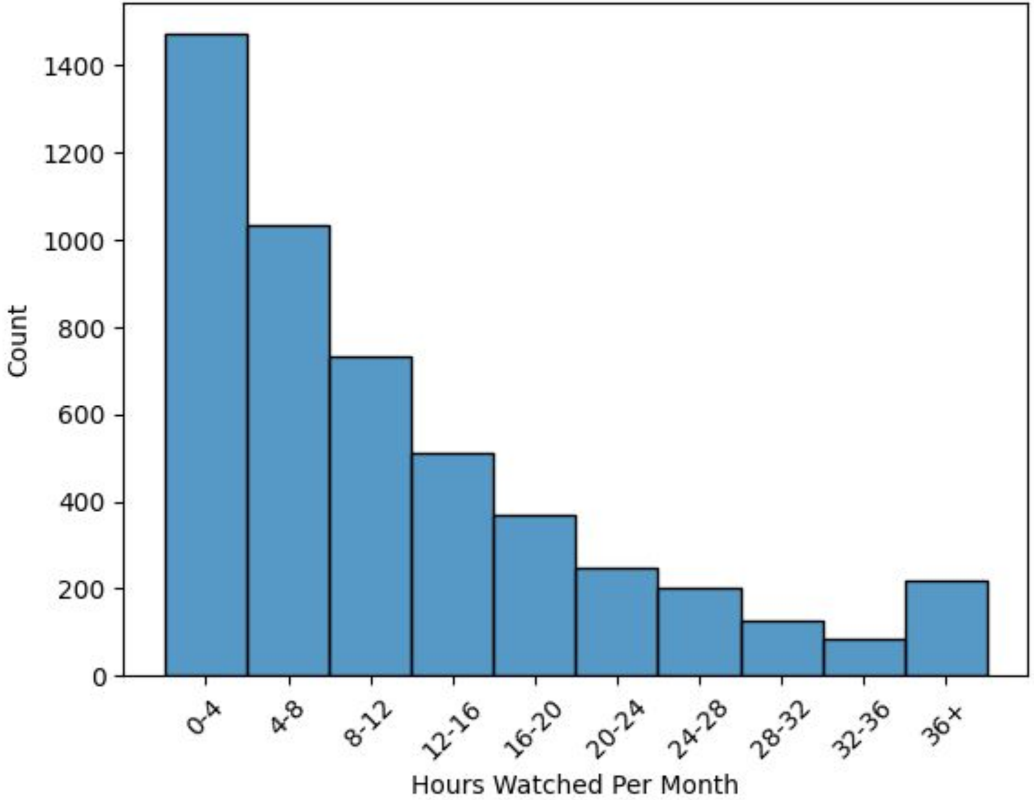
- Netflix Customer Churn dataset from Kaggle.
  - Synthetic dataset of 5,000 unique Netflix customers

- 14 columns of data:
  - Age
  - Gender
  - Average watch time per day
  - Last login
  - Watch Hours
  - Number of Profiles

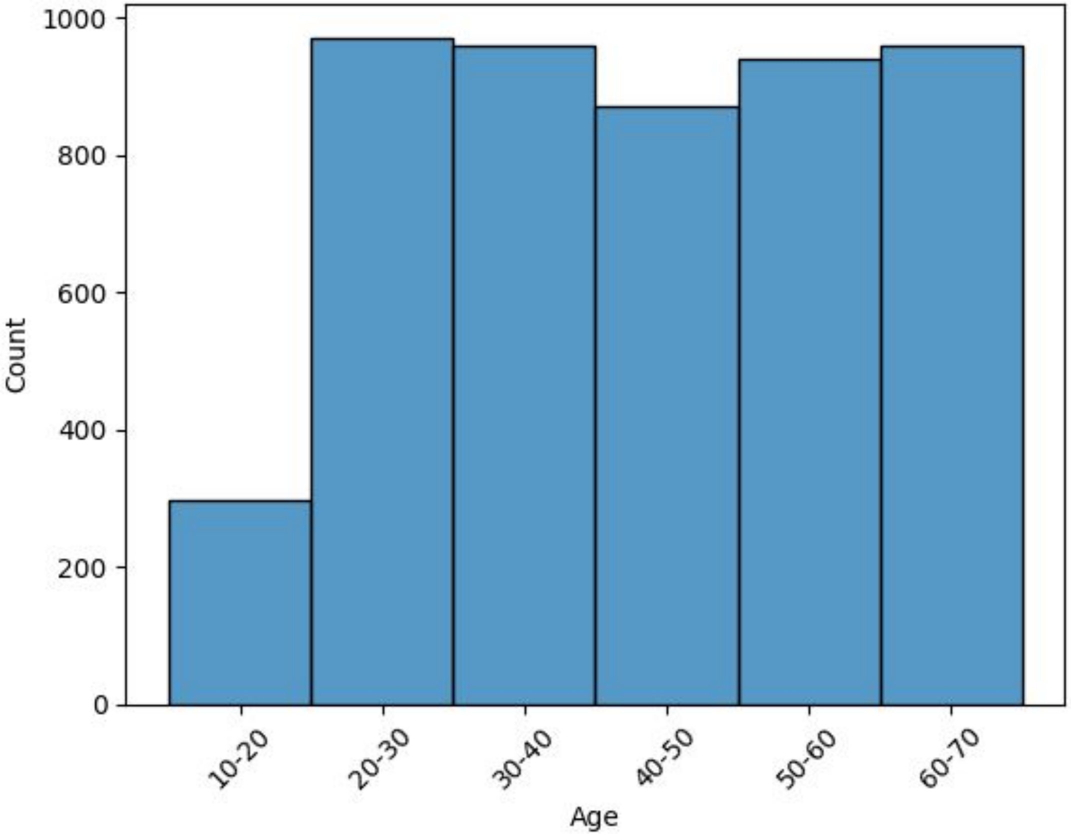
- Cleaned data and prepared it for supervised learning models

Data Description

Distribution of Monthly Watch Hours



Distribution of User Age



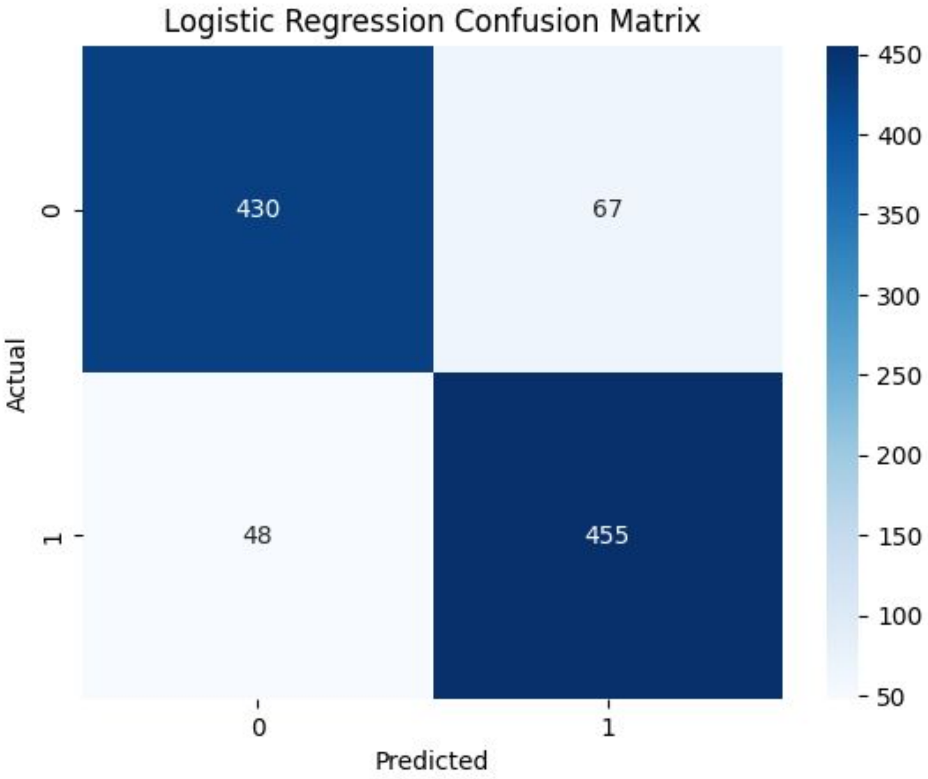
Distribution of User Gender

Logistic Regression

Why Logistic Regression?

- Logistic Regression models are used to predict binary outcomes
  - Churn vs No Churn
- Predicts the probability that something will happen
  - Predicts (%) probability that a customer will churn
- Can handle both numerical and categorical independent variables

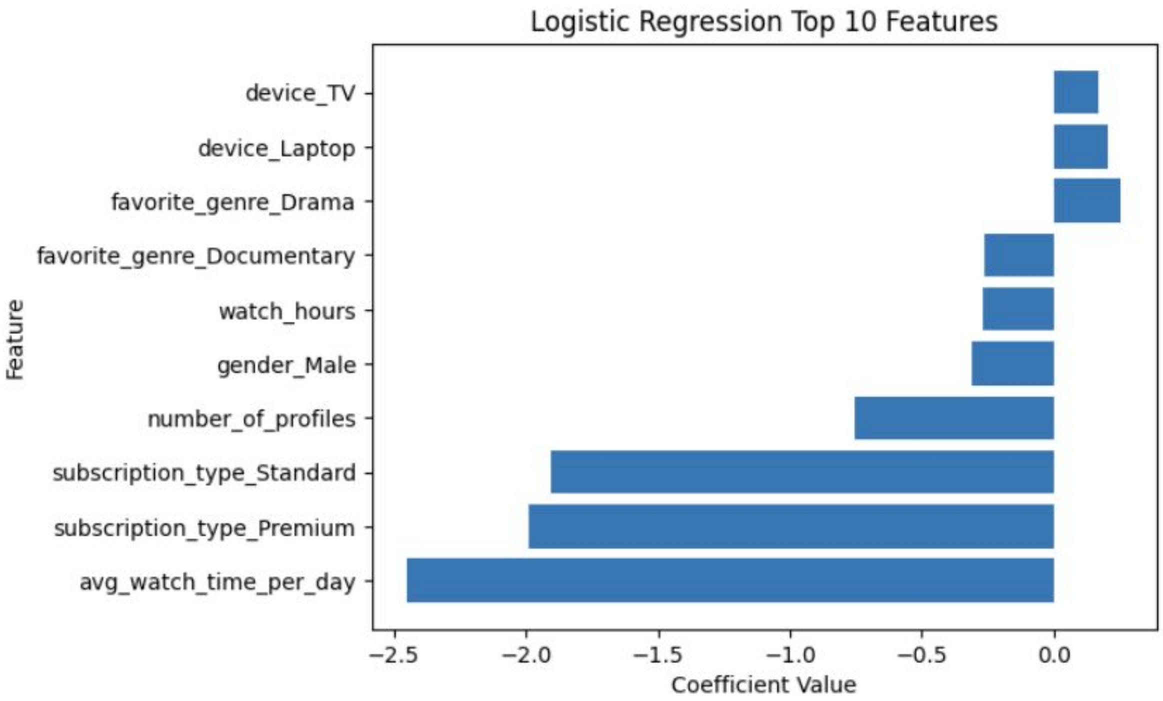
Logistic Regression: Confusion Matrix & Accuracy



Accuracy	0.885
Precision	0.872
Recall	0.905
F1 Score	0.888

*0 = Not churn; 1 = Churn

Logistic Regression: Top Predictors



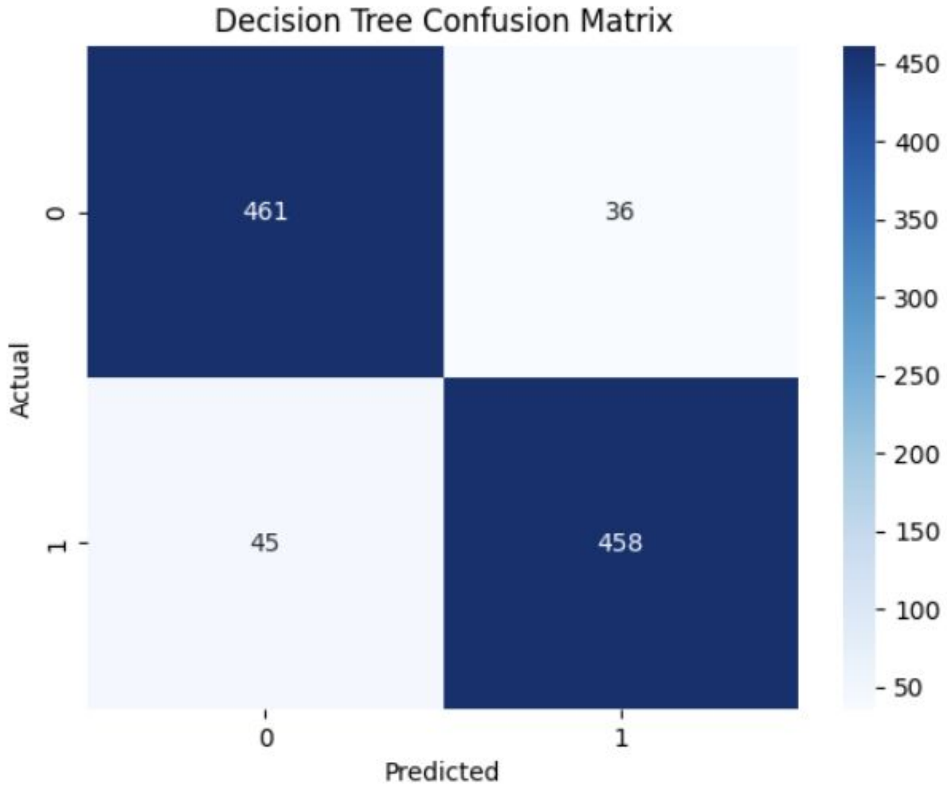
- For each 1-unit increase in average watch time per day, the odds of churn decrease by about 92%.

Decision Tree

Why Decision Tree?

- Splits data into branches based on true/false conditions
- Creates series of decision rules that classify users into Churn or Not churn
- Can capture non-linear relationships
- Highly interpretable (visual decision paths)

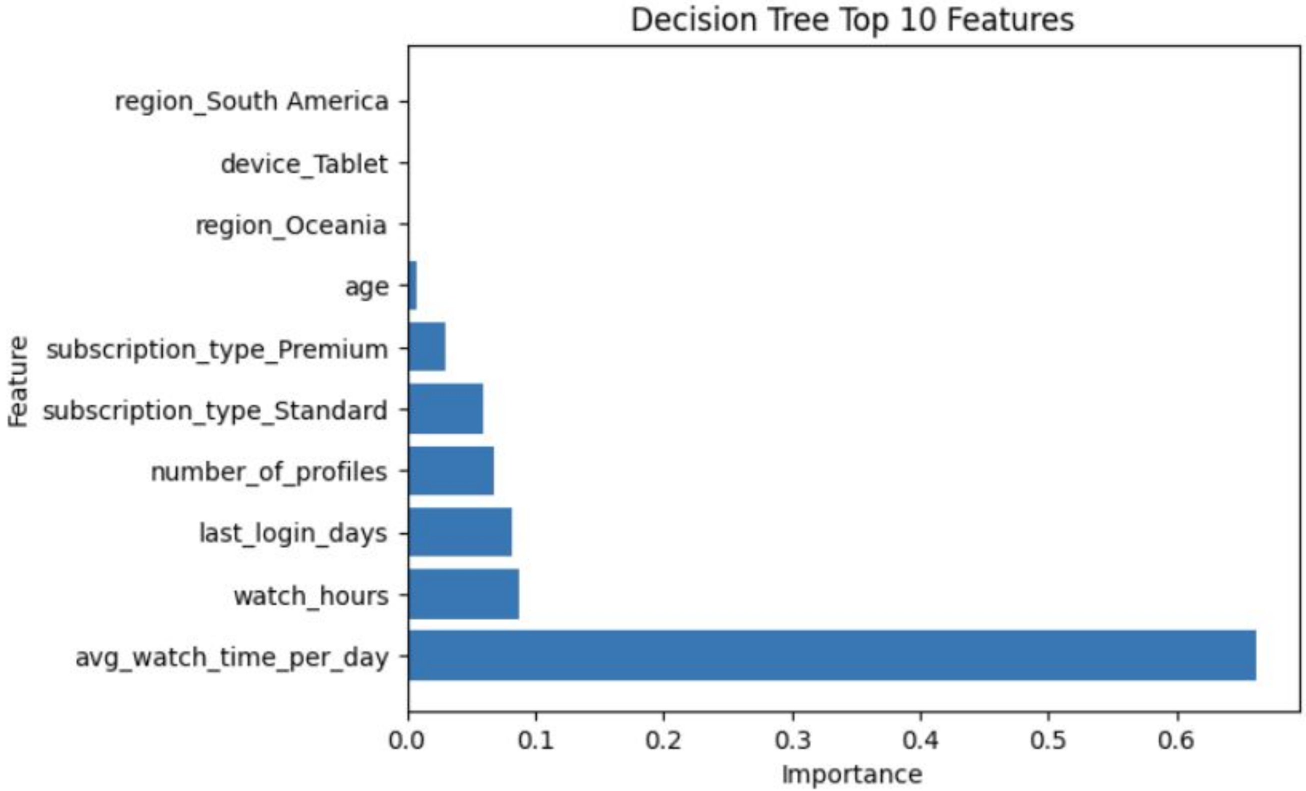
Decision Tree: Confusion Matrix & Accuracy



Accuracy	0.927
Precision	0.919
Recall	0.911
F1 Score	0.919

*0 = Not churn; 1 = Churn

Decision Tree: Top Predictors



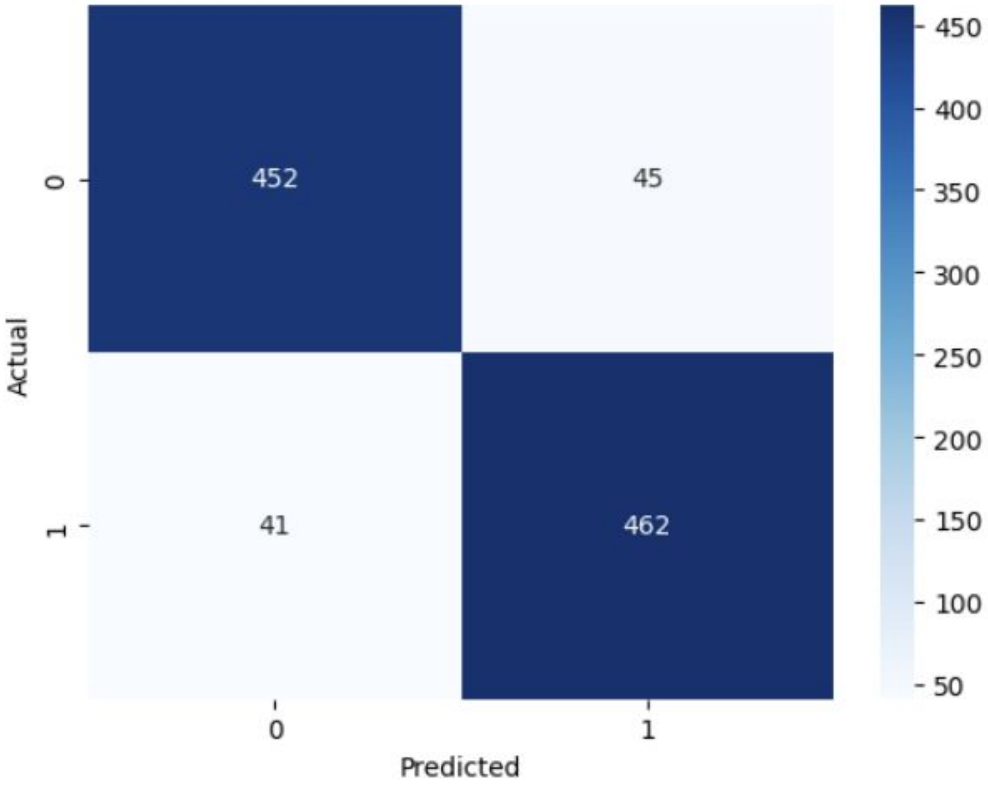
- Average daily watch contributes roughly 67% of the total predictive power of the model,

Random Forest

Why Random Forest?

- Builds many decision trees and averages their predictions
  - Reduces overfitting, more accurate and robust than single Decision Tree
- Captures more complex interactions between features than a single tree
- Handles large datasets and noisy features well

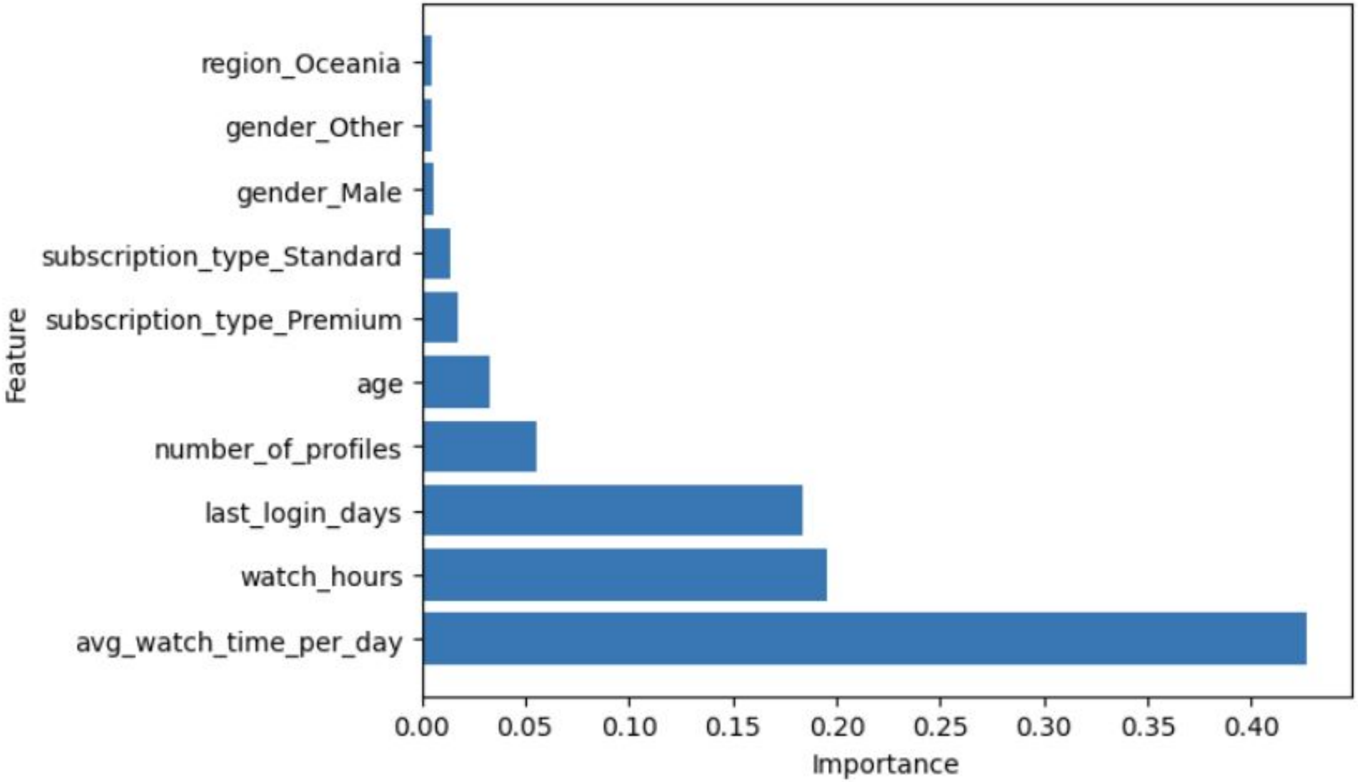
Random Forest: Confusion Matrix & Accuracy



Accuracy	0.914
Precision	0.911
Recall	0.919
F1 Score	0.915

*0 = Not churn; 1 = Churn

Random Forest: Top Predictors



- The importance of *average watch time per day* decreased compared to the decision tree (from ~0.67 to ~0.43)

Model Comparison Summary

Model	Accuracy	Precision	Recall	F1 Score	Notes
Logistic Regression	0.885	Moderate	High	Moderate	Simple linear baseline
Decision Tree	0.927	High	High	High	Highest accuracy but may overfit
Random Forest	0.914	High	Highest	High	More robust; generalizes better to unseen data

Conclusions

- Top predictors of churn are related to user engagement

- **avg_watch_time_per_day** = Strongest predictor across all models
  - Appeared in top splits of decision tree
  - Top ranked feature in all models
  - Low watch time strongly signals churn risk

- **last_login_days**
  - Used in early tree splits
  - Longer gaps since last login = higher churn likelihood

- **watch_hours** and **number_of_profiles** also contributed
  - Higher total watch hours = lower churn
  - More profiles = shared/household accounts are more retained

Business Implications & Suggestions

- Use engagement as an early indicator of at-risk churn customers
- Target low-engagement users with personalized content, reminders, or incentives
- Promote family plans to increase retention
- Deploy Random Forest model for continuous monitoring of churn prediction

Limitations

1) Data procured from a public data set
    a) Lacks completeness
    b) Not real-world applicable

2) Logistic Regression: Assumes predictor variables are independent

3) Decision Tree: High risk of overfitting

4) Random Forest: Computationally expensive

Thank You