Project Report: Predicting Success of Netflix Shows using Supervised Learning Techniques

Date: 30th July, 2025
To: Stakeholders
From: Shijin Ramesh, Data Analyst
---

1. Title and Business Problem

Title: Predicting Success of Netflix Shows using Supervised Learning Techniques
Business Question: Can we predict whether a newly launched Netflix show will be successful using historical show attributes?

Netflix continuously invests in content across regions and genres. A data-driven approach to forecast the success of shows before launch could improve decision-making related to content acquisition, marketing, and budgeting.

2. Data Description

We used a comprehensive Netflix Shows dataset (2025), consisting of metadata for over 10,000 shows, with features such as:

- Title, Type, Release Year, Date Added
- Genres, Country, Language
- Popularity, Vote Count, Vote Average, Rating
- Duration

Preprocessing:

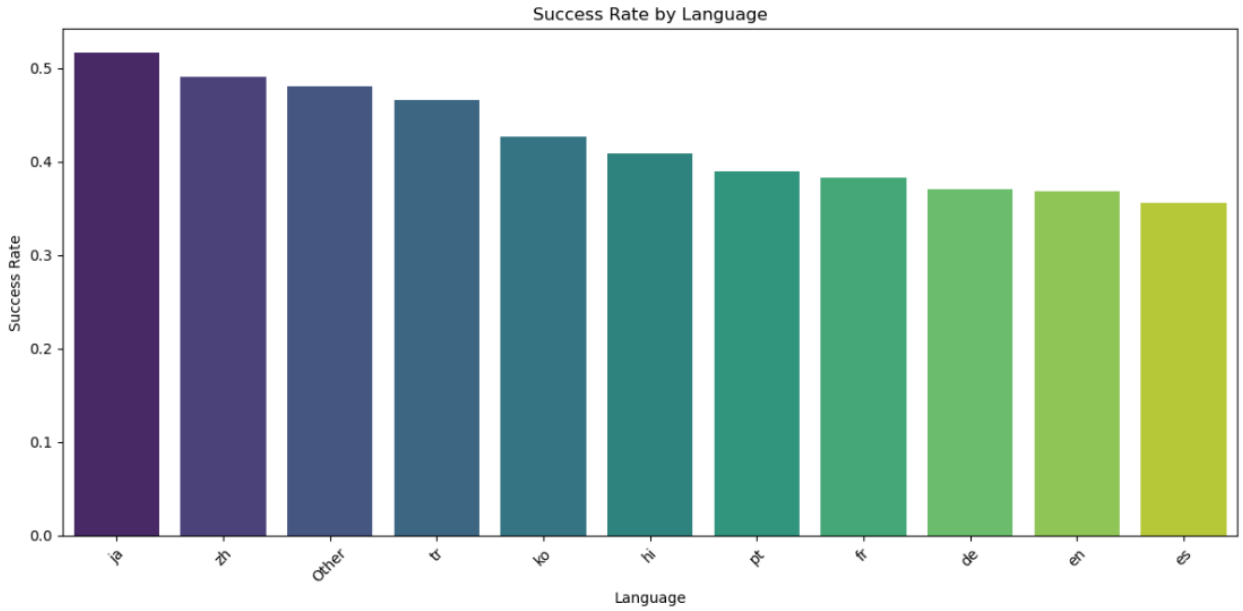
- Cleaned null/missing values and irrelevant fields
- Created binary flags for multi-valued fields (genres, country)
- One-hot encoded categorical features
- Derived new fields such as **success** (our target variable) from rating

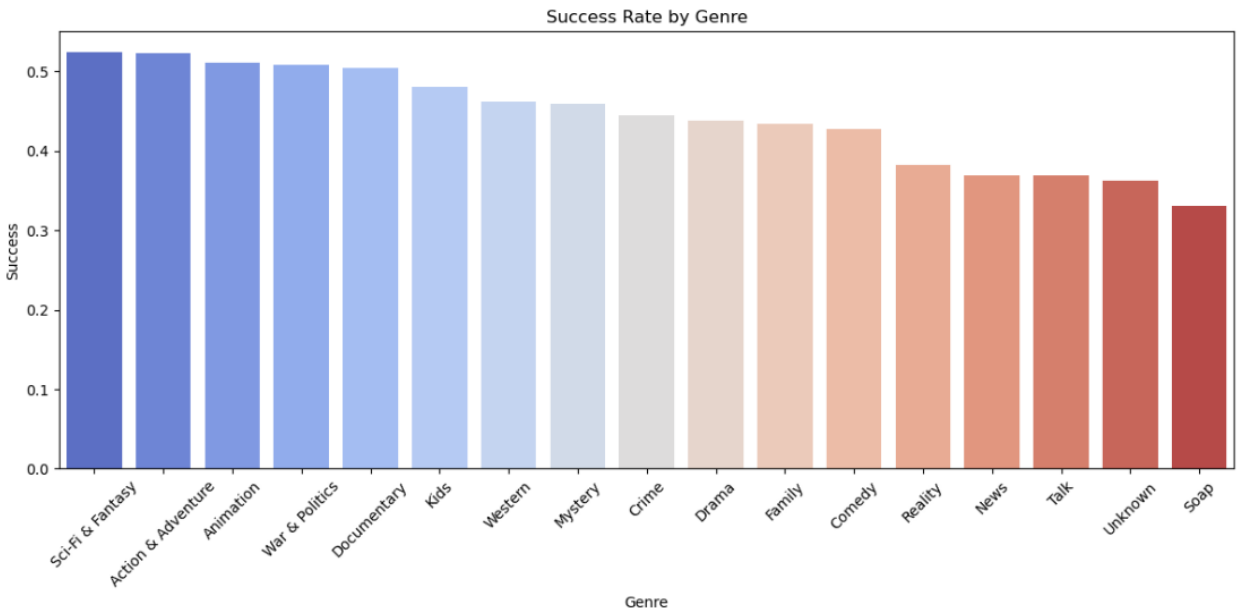
	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	genres	language	description	popularity	vote_co
0	33238	TV Show	Running Man	안재철	Yoo Jae-suk, Jee Seok-jin, Kim Jong-kook, Haha...	South Korea	7/11/2010	2010	8.241	1 Seasons	Comedy, Reality	ko	A reality and competition show where members a...	1929.898	
1	32415	TV Show	Conan	NaN	Conan O'Brien, Andy Richter	United States of America	11/8/2010	2010	7.035	1 Seasons	Talk, Comedy, News	en	A late night television talk show hosted by C...	1670.580	
2	37757	TV Show	MasterChef Greece	NaN	NaN	Greece	10/3/2010	2010	5.600	1 Seasons	Reality	el	MasterChef Greece is a Greek competitive cooki...	1317.092	
3	75685	TV Show	Prostřeno!	NaN	Václav Vydra, Jana Boušková	Czech Republic	3/1/2010	2010	6.500	1 Seasons	Reality	cs	The knives (and forks) are out as a group of s...	1095.776	
4	33847	TV Show	The Talk	NaN	Amanda Kloots, Jerry O'Connell, Akbar Gbaja-Bi...	United States of America, Ireland	10/18/2010	2010	3.400	1 Seasons	Talk	en	A panel of well-known news and entertainment p...	712.070	

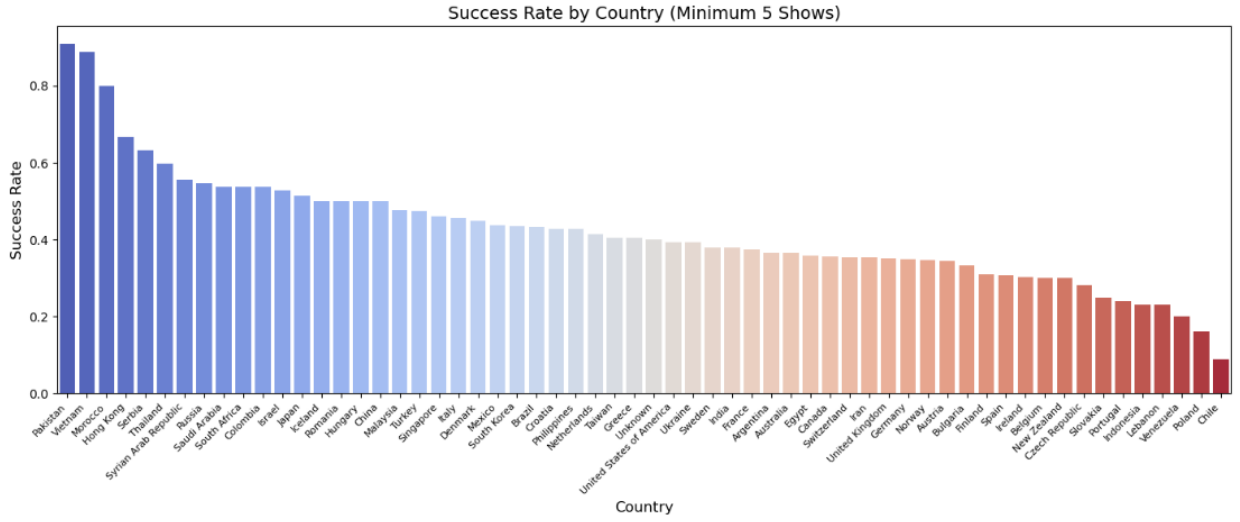
3. Exploratory Data Analysis and Key Patterns

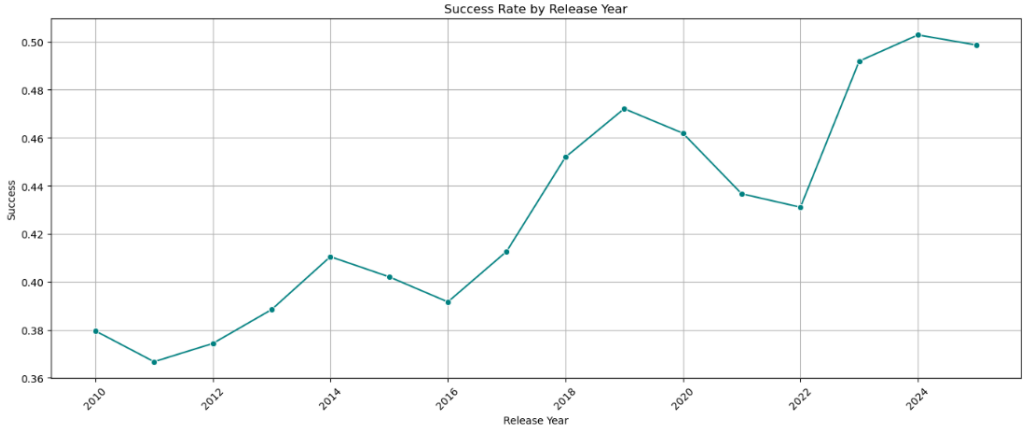
We explored distribution and success rate by:

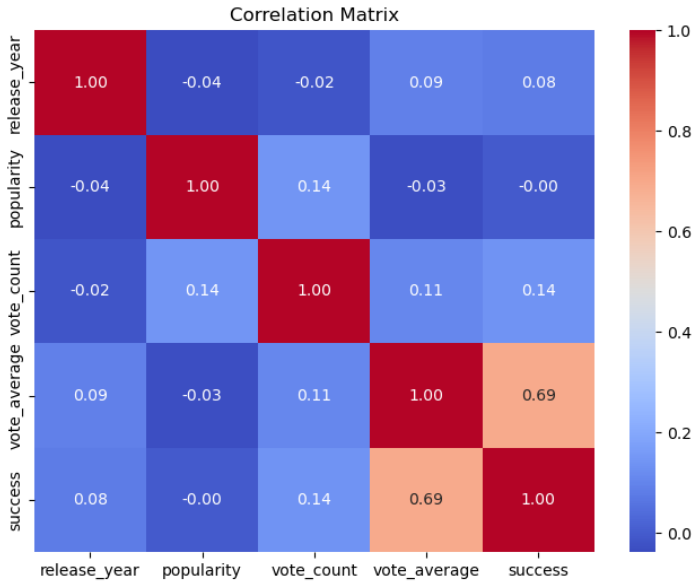
- **Language**: The highest success rates were observed in Japanese and Chinese content — not English or Spanish as initially assumed.
- **Genre**: Shows under Sci-Fi & fantasy, Action & Adventure, Animation, War & Politics genres had notably higher success rates.




Success Rate by Genre
Success
0.0
0.1
0.2
0.3
0.4
0.5
Sci-Fi & Fantasy
Action & Adventure
Animation
War & Politics
Documentary
Kids
Western
Mystery
Crime
Drama
Family
Comedy
Reality
News
Talk
Unknown
Soap
Genre

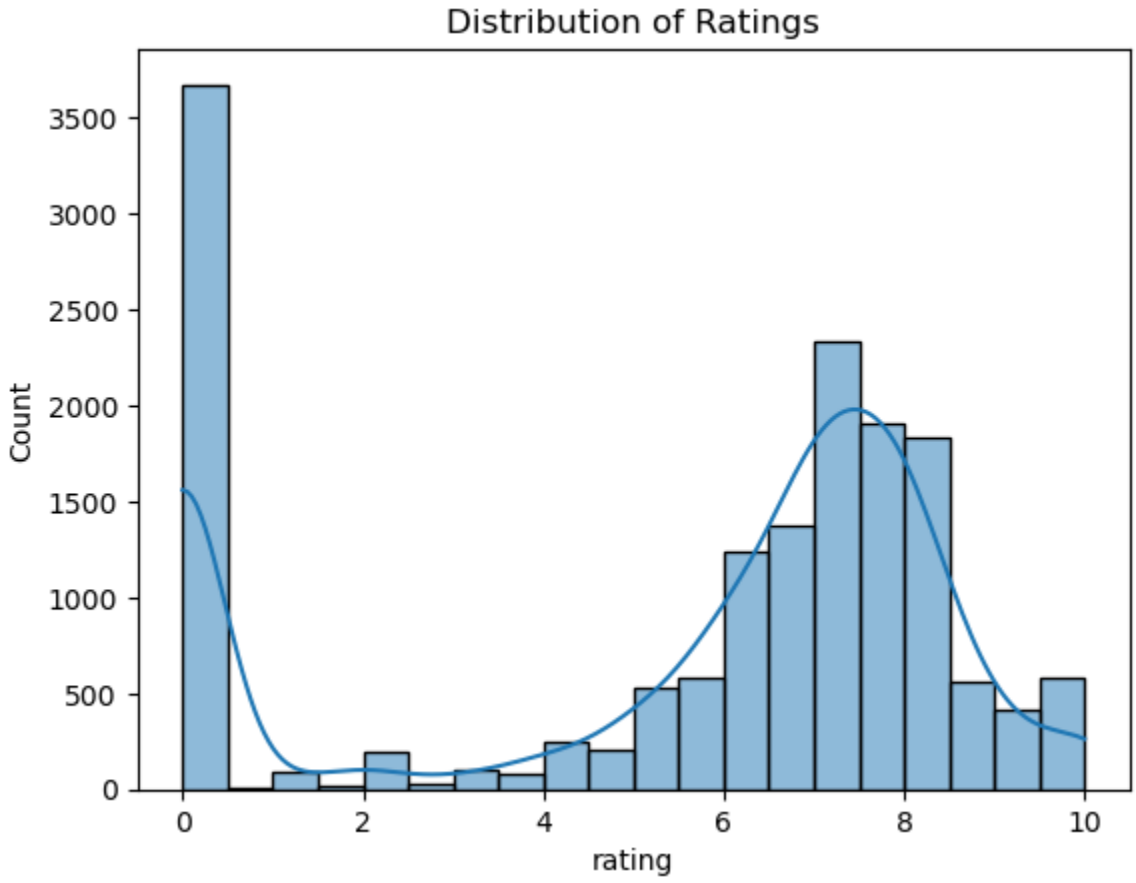

Success Rate by Country (Minimum 5 Shows)
Success Rate
0.0
0.2
0.4
0.6
0.8
Pakistan
Vietnam
Morocco
Hong Kong
Serbia
Thailand
Syrian Arab Republic
Russia
Saudi Arabia
South Africa
Colombia
Israel
Japan
Iceland
Romania
Hungary
China
Malaysia
Turkey
Singapore
Italy
Denmark
Mexico
South Korea
Brazil
Croatia
Philippines
Netherlands
Taiwan
Greece
Unknown
United States of America
Ukraine
Sweden
India
France
Argentina
Australia
Egypt
Canada
Switzerland
Iran
United Kingdom
Germany
Norway
Austria
Bulgaria
Finland
Spain
Ireland
Belgium
New Zealand
Czech Republic
Slovakia
Portugal
Indonesia
Lebanon
Venezuela
Poland
Chile
Country

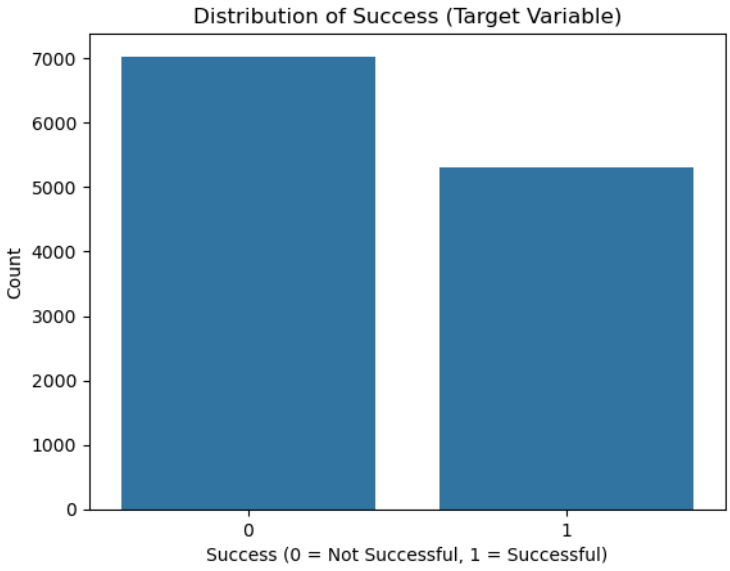

Success Rate by Release Year
Success
0.36
0.38
0.40
0.42
0.44
0.46
0.48
0.50
2010
2012
2014
2016
2018
2020
2022
2024
Release Year



4. Target Engineering

We engineered the target variable **success** using the rating column, as it represents an internal or platform-based classification. A show was labeled successful based on thresholds defined from exploratory analysis.





Why We Excluded vote_average:

- While **vote_average** was strongly correlated with **success**, it's a post-release metric, introducing data leakage if used for prediction. Hence, we excluded it from final model training.
- We engineered **success** using rating and trained models on features that would be available prior to release and using other features like **popularity** and **vote_count**.

5. Modeling Strategy

We structured our modeling in three phases:

- Model A (with vote_average) – benchmark with leakage
- Model B (without vote_average) – realistic model
- Tuned Random Forest with important feature subset and full feature set

We chose Random Forest Classifier for its:

- Robustness to multicollinearity
- Tolerance to outliers
- Handling of nonlinear feature interactions

Outlier Handling: A Strategic Decision

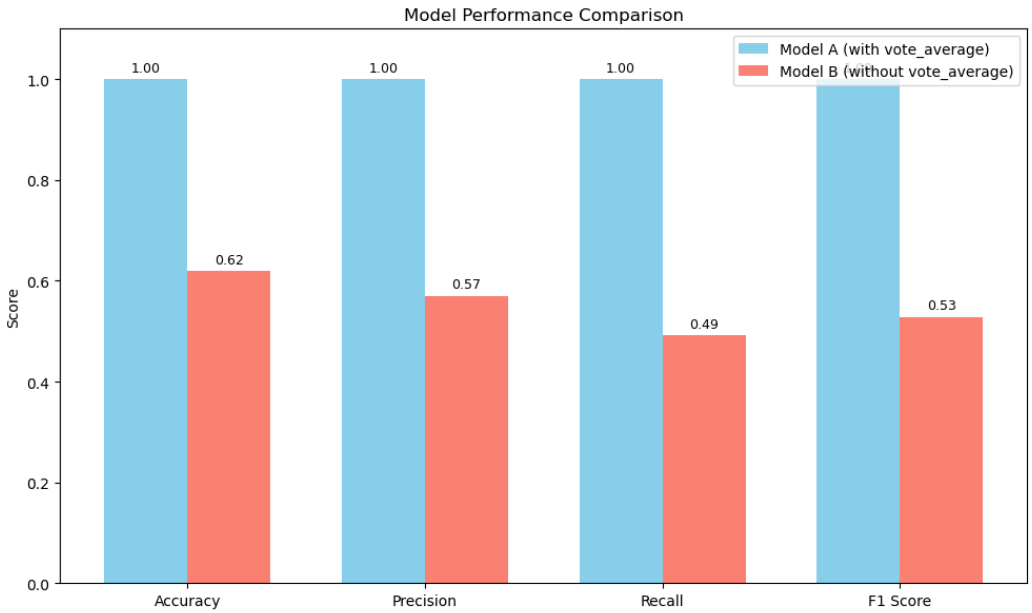
- Popularity: 1,269 outliers found using IQR method
- Vote Count: 1,872 outliers found

Rather than removing them, we retained these outliers due to their business significance (e.g., viral shows) and because Random Forest is inherently resistant to the influence of outliers. Had

we used a model sensitive to outliers (e.g., Logistic Regression), we would have capped or removed them.

6. Performance Comparison

Model	Accuracy	Precision	Recall	F1 Score	AUC
Model A (with vote_average)	1.000	1.000	1.000	1.000	1.000
Model B (without vote_average)	0.6188	0.5703	0.4925	0.5286	0.690
Tuned RF on X_train_b	0.6179	0.5620	0.4918	0.5242	0.681
Tuned RF on Full X_train	0.6559	0.6235	0.4982	0.5538	0.719
Final Model (Test Set)	0.6825	0.6655	0.5393	0.5958	0.7317



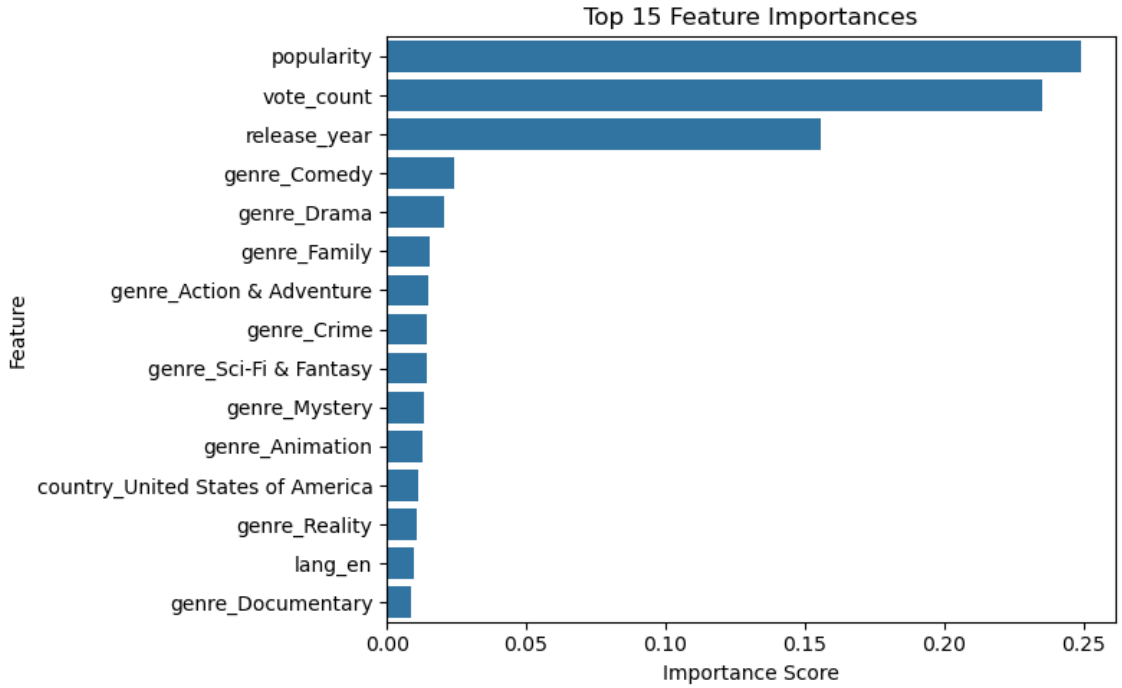
7. Feature Importance

Top predictors from Random Forest:

- vote_count
- popularity
- release_year

- genre_documentary, genre_family, genre_comedy, genre_drama

These aligned with the findings from EDA, especially language and genre.

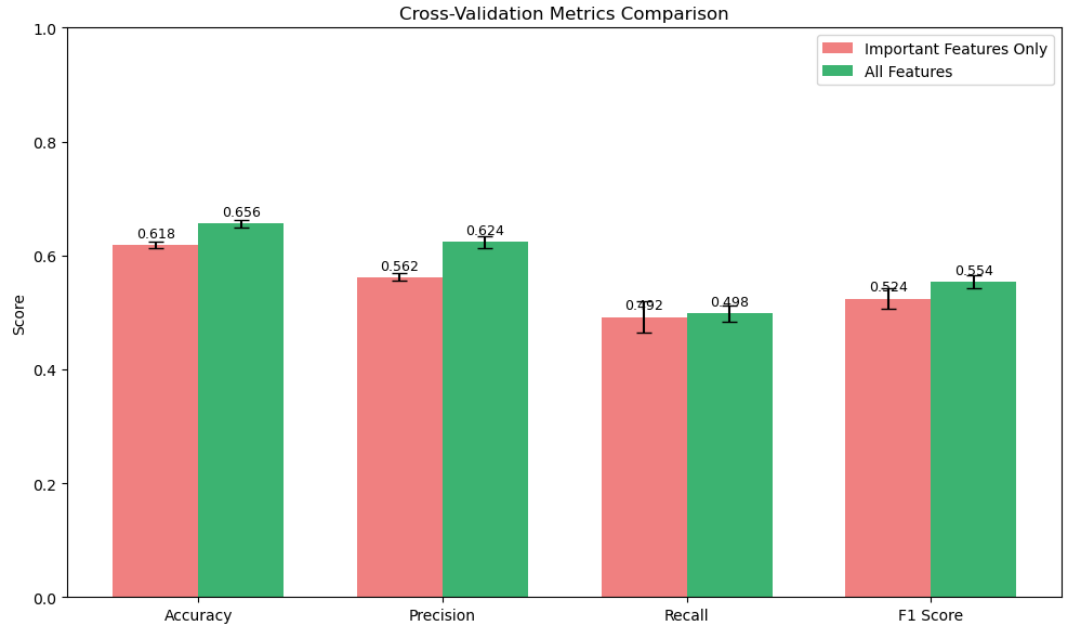


8. Cross-Validation Strategy

We used RandomizedSearchCV with 5-fold cross-validation on both:

- Subset features (X_train_b): Lower generalizability
- Full features (X_train): Better cross-val metrics → Selected for final model

This approach ensured our model wasn't overfitting and had robust performance across unseen data.



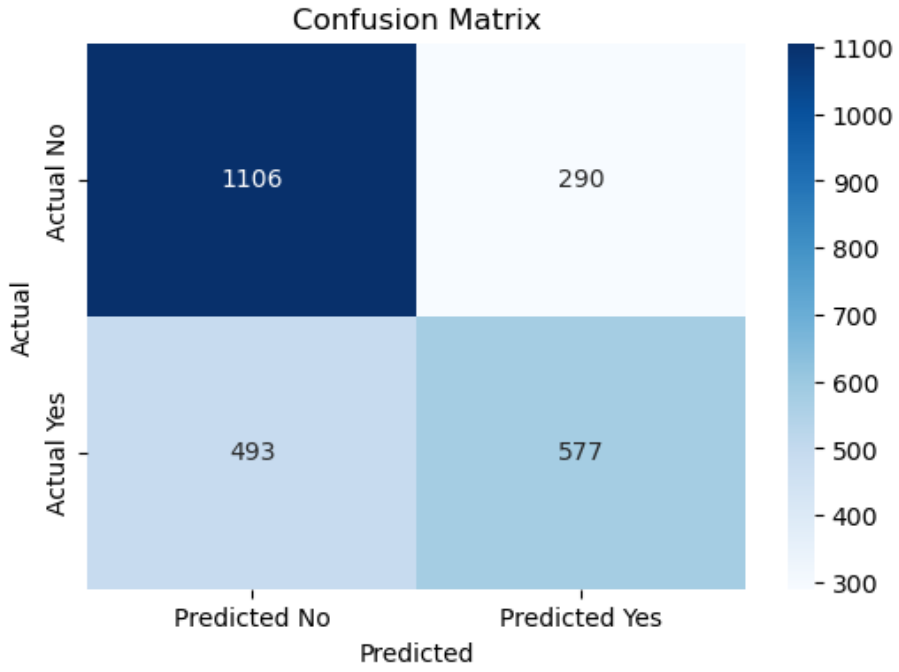
9. Final Model Results

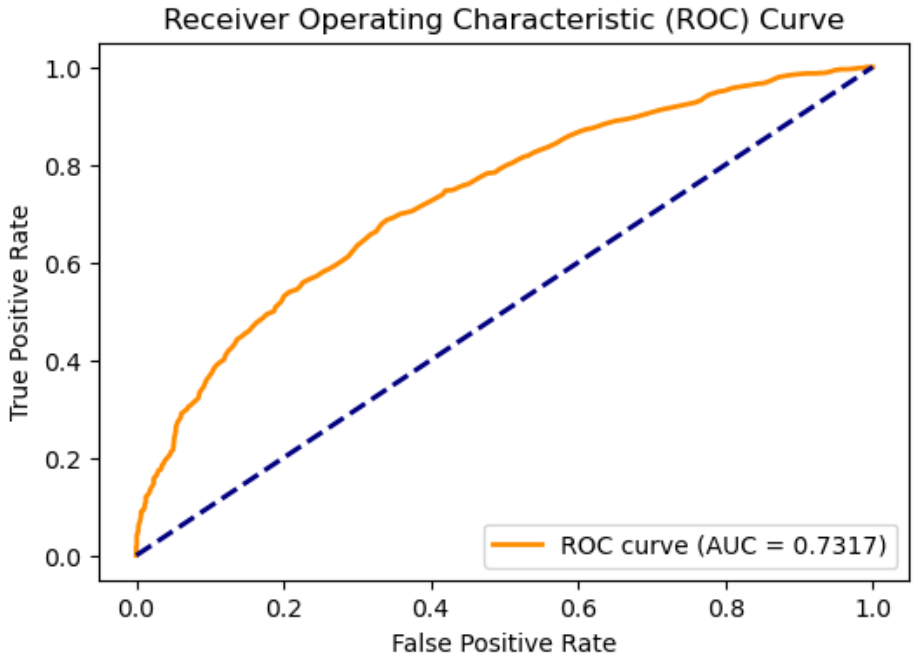
On the hold-out test set, the final tuned model achieved:

- Accuracy: 0.6825
- Precision: 0.6655
- Recall: 0.5393
- F1 Score: 0.5958
- AUC Score: 0.7317

Confusion Matrix:

- TN: 1106 | FP: 290
- FN: 493 | TP: 577





10. Conclusion & Recommendations

- We built a deployable classification model using supervised learning to predict success of Netflix shows before release, with an AUC of 0.73.
- The model uses only features available pre-launch, making it suitable for business forecasting.

Recommendation for Stakeholders:

A. **Integrate the Model into the Pre-Production Pipeline**
   Before approving or investing heavily in a new show, the company can use this predictive model to assess its likelihood of success based on historical data. Since the model uses only information available before a show is launched (like language, genre, country, and show type), it's ideal for guiding early decisions such as content green-lighting, budgeting, and scheduling.
   Business Value: Make smarter investment decisions earlier, reduce the risk of launching underperforming content.

B. **Strategically Allocate Marketing Budgets**
   The model can help identify shows that have a high probability of succeeding, even if they aren't initially popular (e.g., niche genres or non-English content). These shows could benefit the most from targeted marketing efforts, helping them reach their potential audience more effectively.
   Business Value: Optimize marketing spend by focusing resources on promising but less-visible shows, maximizing ROI.

C. **Prioritize Content Based on Language and Genre Trends**
   Our analysis found that success rates varied significantly by language and genre combinations. For example, Japanese and Chinese shows performed particularly well in recent years. This insight can be used when deciding which regions or content types to prioritize in the upcoming production slate.
   Business Value: Align content strategy with emerging viewer preferences to increase the likelihood of success across regions.

D. **Keep the Model Updated with Fresh Data**
   Viewer preferences evolve rapidly — what's trending this year might not work next year. To maintain accuracy, we recommend retraining the model regularly (e.g., quarterly or bi-annually) using the latest Netflix show data.
   Business Value: Stay ahead of changing trends and keep decision-making data-relevant and timely.