PACE Strategy

PLAN

Business Question: Can we predict whether a newly launched Netflix show will be successful using historical show attributes?

Context: With growing content on Netflix, predicting the success of a show before launch can help optimize content investment, marketing strategy, and audience targeting. Success can be defined by high audience ratings or popularity.

Goal: To build a supervised machine learning model to classify whether a show will be successful or not based on historical attributes such as genre, language, release year, rating, popularity, etc.

Target Variable: A binary variable — 'Success' (1) or 'Not Success' (0), derived from vote_average or popularity thresholds.

Stakeholders: Netflix content and data science teams, marketing analysts, and content strategists (hypothetical).

ANALYZE

1. Data Exploration:

- Identify missing values in columns like director, cast, rating, etc.
- Analyze distribution of target variable (success vs. non-success).
- Understand correlation among features such as popularity, vote_average, genre, etc.
- Handle non-numeric variables using encoding techniques.
- Derive useful features: year bucket, genre category count, etc.

2. Assumption Checks & Ethics:

- Evaluate class imbalance and apply SMOTE or undersampling if needed.
- Ensure fairness: prevent language, gender, or region bias.
- Make the model interpretable using feature importance.
- Consider consequences of false positives (wasted investment) or false negatives (missed opportunity).

3. Criteria for Modeling Approach:

- Binary classification problem with categorical and numerical data.

- Naive Bayes if text-derived features dominate or Decision Tree for interpretability and rule-based logic.

CONSTRUCT

1. Data Preprocessing:

- Clean and impute missing values (drop or fill based on type).
- Encode categorical variables (One-Hot or Label Encoding).
- Feature engineering: derive new fields like 'decade', 'genre_count', etc.
- Create binary target variable based on a vote_average or popularity threshold.

2. Model Building:

- Split data into train and test sets (80/20 or 70/30).
- Train both Naive Bayes or Decision Tree models.
- Tune hyperparameters using GridSearchCV.
- Evaluate performance using accuracy, precision, recall, F1-score, and ROC-AUC.

3. Ethical Implications:

- Avoid using biased features that may lead to discrimination.
- Explain model decisions with SHAP or tree visualizations if required.

EXECUTE

1. Evaluate Model Performance:

- Use Confusion Matrix, ROC-AUC, and F1-score to validate models.
- Select the model with the best trade-off between recall and precision.

2. Dashboard with Power BI:

- Create visuals to show distribution of success rate across genres, languages, years.
- Add KPIs like model accuracy, feature importance, and recommendations for upcoming shows.

3. Conclusion:

- Deliver model, insights, and dashboard to Netflix team.
- Make strategic recommendations: which genres, countries, or show types yield higher success.

4. Recommendation:

- If model performance is satisfactory (>80% F1-score), proceed to integrate into content pipeline.
- If not, consider adding more data (social media buzz, trailer views, etc.) for better prediction.