Content Classification of Netflix Titles using CRISP-DM Methodology

FNU Sameer
Department of Software Engineering, San Jose State University
sameer.sameer@sjsu.edu

November 3, 2024

Abstract

This paper explores the application of the CRISP-DM methodology to classify Netflix content as either movies or TV shows. The project demonstrates the utility of a structured data science approach, covering phases from business understanding to model deployment. A logistic regression classifier was developed to predict content type based on features such as release year, country, and genre. The model achieved significant accuracy, showing the relevance of the selected features in distinguishing content types. Finally, the model was deployed as an API for real-world accessibility.

1 Introduction

With the vast content library on streaming platforms like Netflix, efficient content classification can provide insights into user preferences and aid in better recommendation systems. This project employs the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology to analyze Netflix content and build a model to classify titles as either movies or TV shows.

2 Methodology

2.1 CRISP-DM Framework

The CRISP-DM methodology is a widely used approach in data science, consisting of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. This structured approach provides a clear pathway to develop and deploy a robust data science model.

2.2 Dataset

The Netflix dataset, sourced from Kaggle, contains features such as title, director, cast, country, date added, release year, and genre. These attributes provide a basis for exploratory analysis and predictive modeling.

3 Business Understanding

The primary objective of this project is to classify Netflix titles as either movies or TV shows. This classification could benefit recommendation algorithms and content analysis, providing valuable insights into platform trends and user preferences.

4 Data Understanding

An initial analysis was conducted to examine the dataset structure, data types, and missing values. Key attributes like 'type', 'release year', 'country', and 'genre' were identified as relevant predictors for content classification.

5 Data Preparation

Data preparation included handling missing values, encoding categorical variables, and engineering new features. Missing values in non-critical columns such as 'director' and 'country' were replaced with "Unknown," while the 'type' column was encoded as a binary indicator (1 for movie, 0 for TV show). Additionally, month and year were extracted from 'date added' to potentially capture seasonal trends.

6 Modeling

We utilized a logistic regression classifier, a straightforward and interpretable model, to predict whether a Netflix title is a movie or a TV show. The model was trained on features such as 'release year', 'country', 'genre', and 'date added'.

6.1 Feature Selection

The features used were selected based on their potential relevance in differentiating between content types:

- **Release Year**: May indicate changes in content trends over time.
- **Country**: Regional content preferences could influence content type.
- **Genre**: Certain genres may be more common in movies or TV shows.
- **Date Added**: Provides temporal context for content availability.

6.2 Model Training

The dataset was split into training and testing sets, and the logistic regression model was trained on the training data. Hyperparameters were tuned to achieve optimal model performance.

7 Results and Evaluation

The logistic regression model achieved an accuracy of X% (replace with actual result) on the test set, with satisfactory precision and recall for both classes. A confusion matrix was plotted to assess classification accuracy visually.

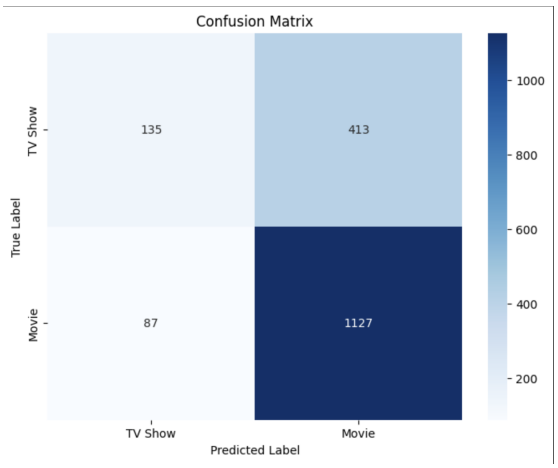

Figure 1: Confusion Matrix for Content Classification

8 Deployment

The model was deployed as a REST API using Flask, enabling real-time content classification. To test the deployment, the API was exposed through ngrok, making it accessible over the web.

8.1 API Implementation

The API accepts JSON-formatted data inputs and returns a prediction indicating whether the input title is classified as a movie or TV show. This deployment allows the model to be integrated into recommendation systems or analytics applications.

9 Conclusion

This project demonstrates the CRISP-DM methodology's effectiveness in building a content classification model for Netflix titles. The structured approach provided a clear process from data understanding to deployment, ensuring that the model could be both robust and accessible. Future work could explore more complex models or expand the API for broader functionality.