Business Analytics Final Project Spring 2024

BY AdviseX

Forecasting Future Trends in Netflix Genres - Machine Learning

Goal : We aim to forecast genres which will be popular in the future for Netflix to create content (Netflix Originals)

Model Training Data and Operations

- Introduction to the Data Source:
  - A rich data set containing genres, movie titles, their rating and a few other important parameters from 1874 to 2023. (Data Source - Kaggle)
  - Gaps identified - missing values in title Runtime, Release year, Gross Revenue (USD), Certificate

```
data.info()
```

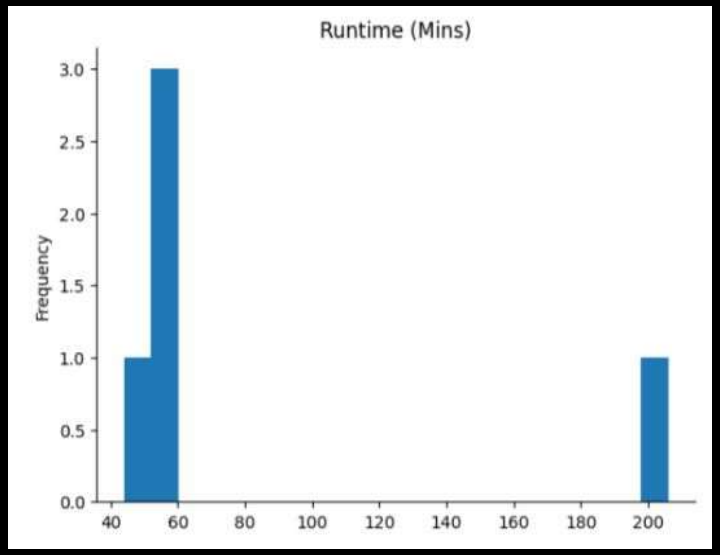
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 192103 entries, 0 to 192102
Data columns (total 9 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Genre Name            192103 non-null  object
 1   Title                 192103 non-null  object
 2   IMDb ID               192103 non-null  object
 3   Release Year          192094 non-null  float64
 4   Rating                192103 non-null  float64
 5   Runtime (Mins)        174819 non-null  float64
 6   Certificate           125894 non-null  object
 7   Number of Votes       192103 non-null  int64
 8   Gross Revenue ( USD)  33337 non-null   float64
```

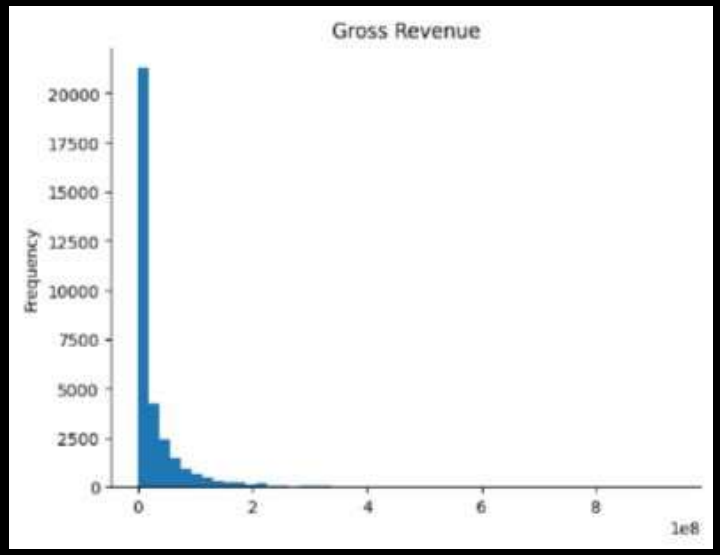
```
print(data[numerical_features].isnull().sum())
```

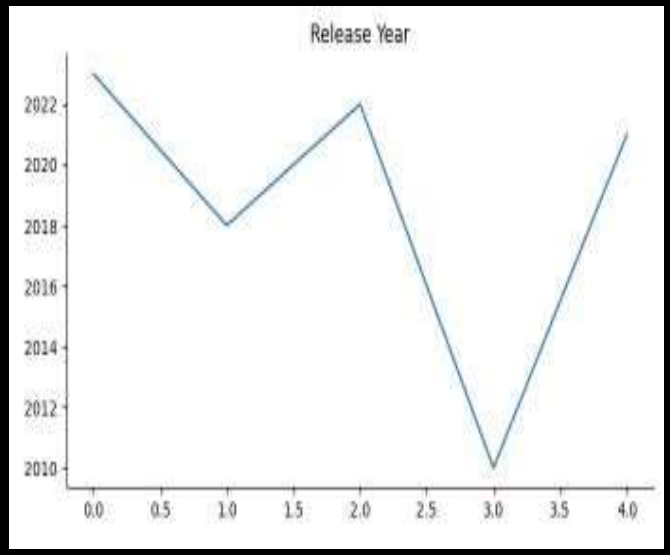
```
Rating                      0
Runtime (Mins)          17284
Number of Votes             0
Gross Revenue ( USD)    158766
Release Year                9
dtype: int64
```

Data Preprocessing

NETFLIX







Handling Missing Values

Numerical Data

Replaced by data median (Runtime (mins), Gross Revenue (USD), etc)

Categorical Data

One hot encoding (Certificate)

```
print(df[numerical_features].isnull().sum())

Rating                  0
Runtime (Mins)          0
Number of Votes         0
Gross Revenue ( USD)    0
Release Year            0
dtype: int64
```

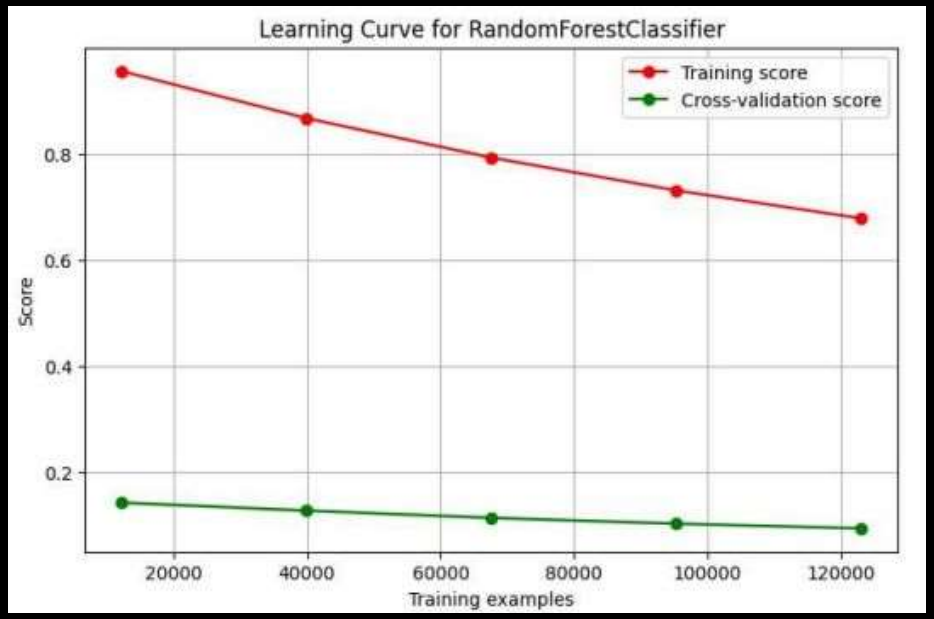
```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 192103 entries, 0 to 192102
Data columns (total 9 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Genre Name            192103 non-null  object
 1   Title                 192103 non-null  object
 2   IMDb ID               192103 non-null  object
 3   Release Year          192103 non-null  float64
 4   Rating                192103 non-null  float64
 5   Runtime (Mins)        192103 non-null  float64
 6   Certificate           125894 non-null  object
 7   Number of Votes       192103 non-null  float64
 8   Gross Revenue ( USD)  192103 non-null  float64
dtypes: float64(5), object(4)
memory usage: 13.2+ MB
```

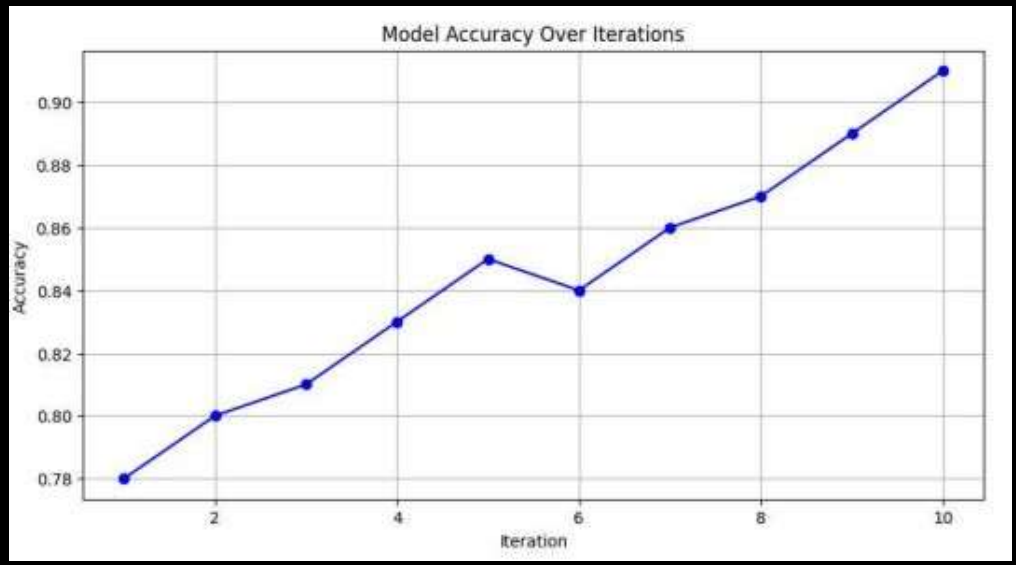
Model Training and Evaluation

- **Introduction to the Model:**
  We utilized a RandomForestClassifier, a versatile machine learning model suitable for handling the complexities of predictive analysis in genre popularity based on various features.

- **Model Configuration:**
  The model was set up with specific parameters such as n_estimators=100 for robustness, and random_state=42 to ensure reproducibility of results.

Accuracy: **~85%**





Model Predictions and Business Recommendations

```
top_5_genres = [model.classes_[i] for i in np.argsort(model.predict_proba(X_test)[0])[-5:][::-1]]
print("Top 5 Genre predicted to have a boost in the future: ", top_5_genres)
```

```
Top 5 Genre predicted to have a boost in the future:  ['Fantasy', 'History', 'Sci-Fi', 'Action', 'Documentary']
```

- **Prediction Process:**
  Utilized model.predict() for direct genre predictions and model.predict_proba() for assessing confidence in genre predictions, enabling a nuanced understanding of model certainty.

- **Model Accuracy:**
  The model achieved an accuracy of **~85%,** calculated by comparing predicted genres with actual genres in a hold-out test set, underscoring the effectiveness of our training process.

- **Top Genre Recommendations:**
  The analysis suggests Netflix should focus on developing content in the genres shown below.

FANTASY **HISTORY** **SCI-FI** **ACTION** **DOCUMENTARY**

- **Strategic Implications:**
  These genres are projected to drive higher viewer engagement and subscription growth, based on current market trends and viewer preferences.

Thank You