Pattern Recognition

**Milestone (1) report**

Movie Revenue Prediction

horizontal line

**Preprocessing**

1. Missing Data Filling:
   1. Missing data were filled in using TMDB API. These data are revenue, genres, directors, release dates and MPAA ratings.
2. Encoding:
   1. Rating: Three encoding types were used; Label encoding, One-hot encoding, and Ordinal encoding, which was discarded due to the rating not having a logical order. The encoder with the best results was the Label encoder.
   2. Genre: Three encoding types were used; Label encoding, One-hot encoding and a modified Ordinal encoding, The weight of each genre was calculated by finding the mean revenue of each genre.
   3. Directors: Two encoding types were used; Target encoding and a modified One-hot encoding. The best result was achieved by applying Target encoding.
   4. Movie Title: Label Encoding
   5. Voice actors: Label Encoding
   6. Character: Label Encoding

1. Feature Selection:
   1. Dropped characters, voice actors and movie title columns due to low correlation with revenue (> ~19%) and insufficient data.
2. Data Cleaning:
   1. Dropped duplicate data.
   2. Dropped unfillable data; some movie titles were incomplete and there were a couple of tv series and tv episodes, therefore not all revenues and/or directors of some data could be found.
3. Feature Scaling:
   1. Applied min-max scaler.
4. Feature Extraction:
   1. Date column was expanded into a day, month and weekday columns, tried year seasons but didn’t give good MSE.
   2. Created a “new movie” column based on the year of the release date; to indicate whether the movie was released later than 2005.

Regression Techniques:

1. Polynomial Regression: We tried degrees in the range of (2, 5). The best results yielded an MSE of:  
   {TRAIN} 1.3691602990770952e-11  
   {TEST} 1.4712136428963402e-11  
    with degree of 3.
2. Ridge Regression: the model yielded an MSE of  
   {Train} 4151743279357848.   
   {Test} 3151608274064461.5.

Training time of each model was nearly instantaneous.

Used/Discarded Features

In the final model, the following features were handled accordingly:

|  |  |
| --- | --- |
| Used | Discarded |
| Directors (Target encoded) | Characters |
| Release date (Created five columns: day, month, year, weekday and new\_movie) | Voice actors |
| Rating (Ordinal encoded) | Movie title |
| Genre (Mean-weighted) |  |
| Genre (Ordinal encoded) |  |

Train-Test split: A distribution of 80%-20% was applied.  
  
Correlation Heat Map

Chart

Description automatically generated with medium confidence

Ridge Model Results:

Text

Description automatically generated

Poly Model Results

Text

Description automatically generated

Conclusion:

Firstly, we suspected that rating would be ordinal but the analysis turns the result otherwise and was best encoded by label.

Secondly, genre was also a candidate for ordinal encoding and the order was calculated based on the genre weight in the data which the analysis proved to be sufficient

Thirdly, Director was planned to take a subset from it that represent top directors (i.e., who produced 3 or more movies) to give them a reasonable weight, but the analysis disproved that.

Fourthly, movie title was suspected not yield a good correlation with movie revenue, which was proven.

Lastly, the voice-actors and characters features data were insufficient that we suspected it wouldn’t correlate with the revenue, which the analysis proved.