Movie Mining

Caleb Stillman
Graydon Sinclair
Stephen Tynan

Our Questions

- Is it possible to predict the runtime of a film based on its genre?
- Is it possible to predict the runtime of a film based on its country of origin?
- Is it possible to predict the runtime of a film based on its date of release?
- Can an ideal film length (may be a range) be identified which corresponds to high ratings from viewers?

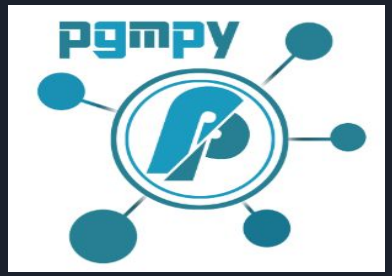
Data Preparation Work

- Outlier handling
- Other data cleaning (removing unnecessary features, etc.)
- Data integration
- Data transformation (one-hot encoding, etc.)

One-Hot Encoding

	id	genre_Action	genre_Adventure	genre_Animation	genre_Comedy	genre_Crime	genre_Documentary	genre_Drama
0	1000001	0	1	0	1	0	0	0
1	1000002	0	0	0	1	0	0	1
2	1000003	1	1	0	1	0	0	0
3	1000004	0	0	0	0	0	0	1
4	1000005	0	0	0	1	0	0	1

Tools Utilized



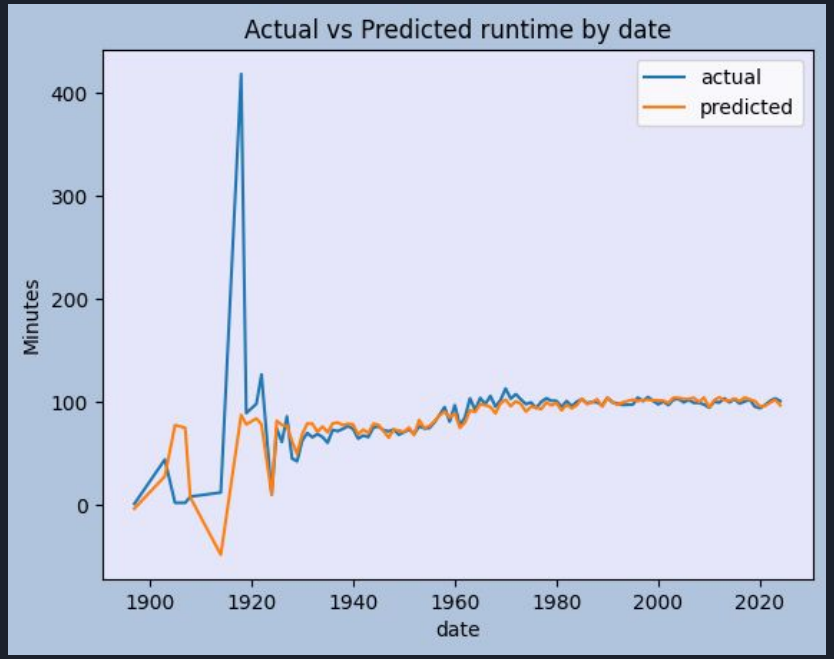
Supervised Learning Algorithms

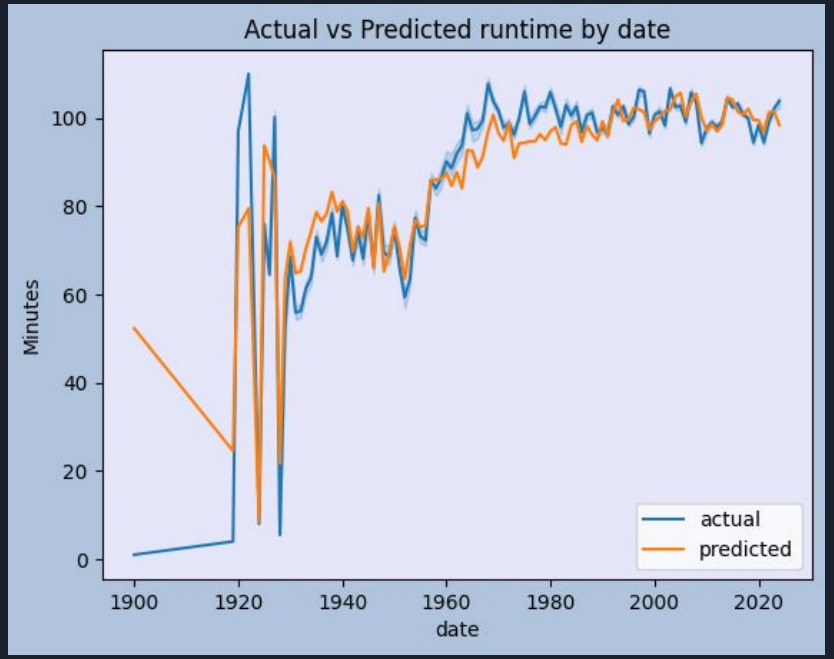
- Linear regression
- Random forest regression
- Support vector machines
- Bayesian Belief Network
- K-nearest neighbor

Insights Gleaned

- Runtime is oft-overlooked
- Can predict runtime to varying degree by other features
  - Rating
  - Genre
  - Country of origin
  - Language
  - Studio

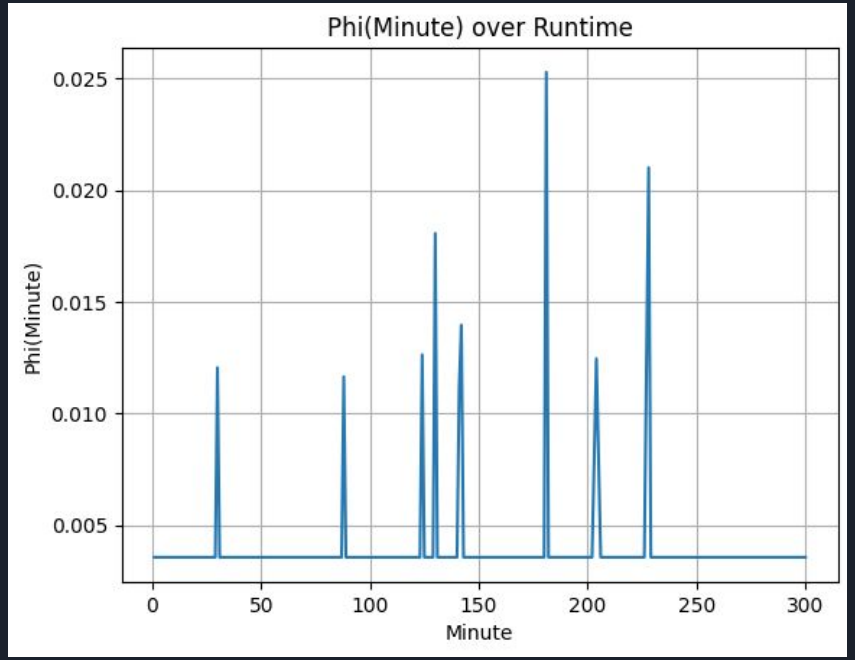
Linear Regression:



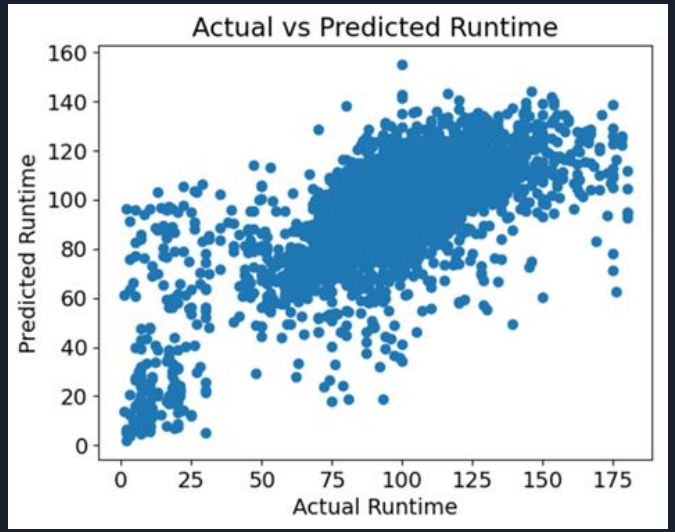


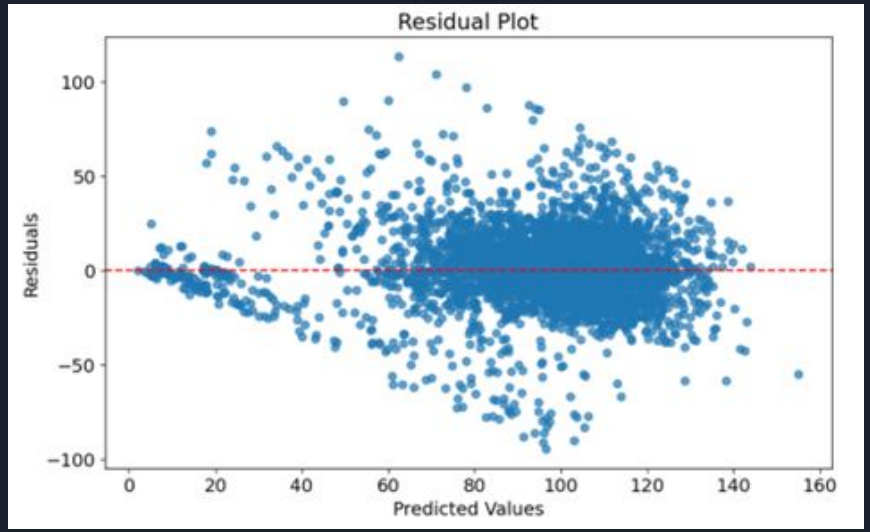
Bayesian Belief Network

Effective for predicting likely runtimes given other variables - can graph conditional probabilities to find optimal ranges.



Random Forest





Application of Insights

- Leverage cultural preferences
- Resonate with target audiences
  - Ratings (popular and/or critical reception)
  - Awards
  - Box Office