What features influence the lifespan and continuing popularity of movies among consumers?

Team 30: Zhiqi Fu, Benji Kan, Yizheng Wang, Bo Zhang

Highlights

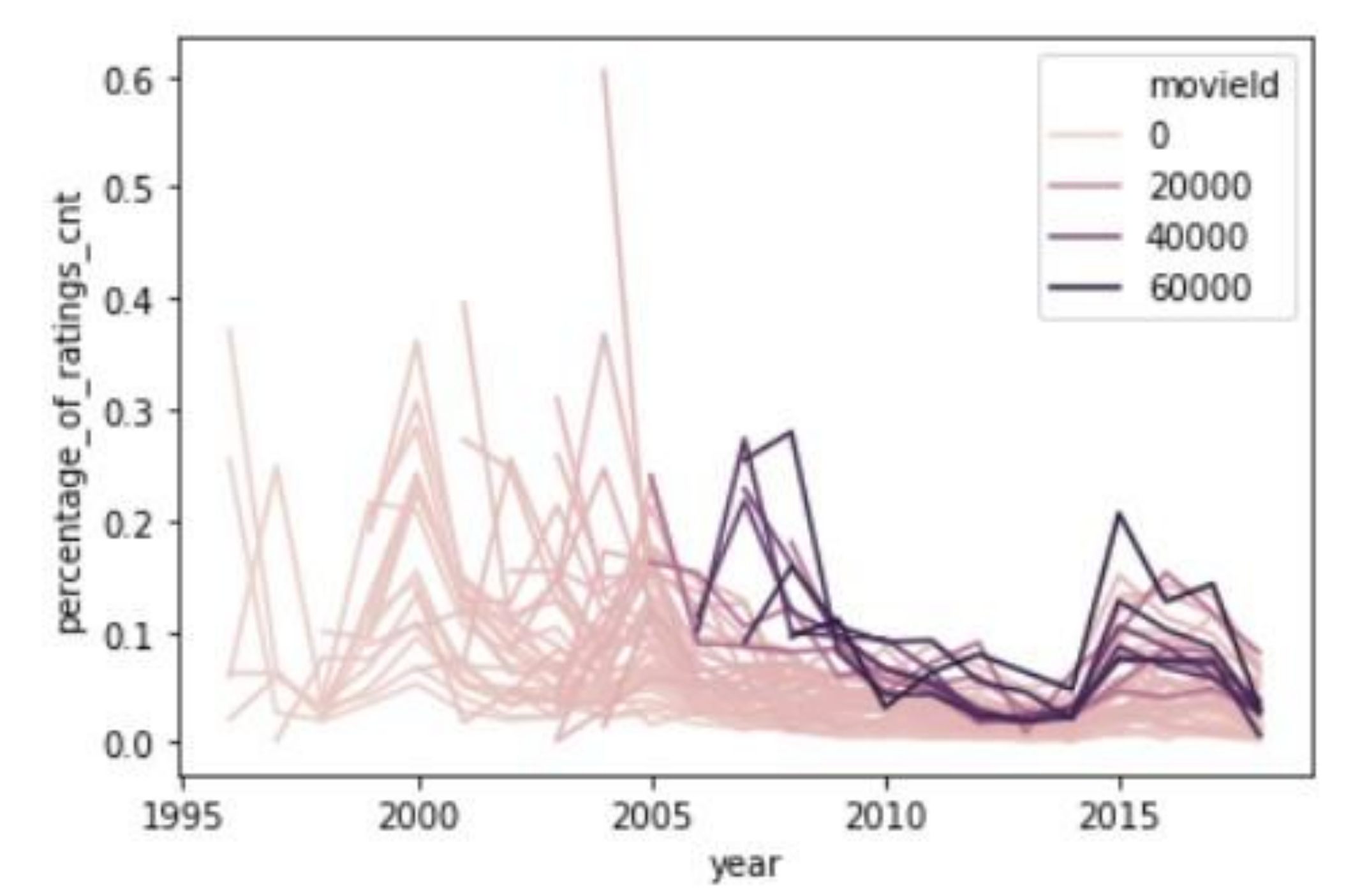
- We developed a model that extracts the most universal qualities by genome tag of movies, loved by consumers across the world and across time.
- Our model uses logistic regression to classify movies into four categories, based on sustainability across time and average rating.
- This model classifies movies that are well-loved over a long period of time with 92% accuracy, and those that are poorly rated yet still survive for al ong time with 71% accuracy.

Background

Whereas movie trends are reflective of trends in popular culture at large, the film industry has the difficult task of predicting future trends in consumer preference. Given the unpredictability of culture and society in the modern age, we present a model that focuses on answering the question: What features influence the lifespan of movies, determining whether a movie has just a brief burst of popularity before fading out or surviving as an enduring classic?

Data

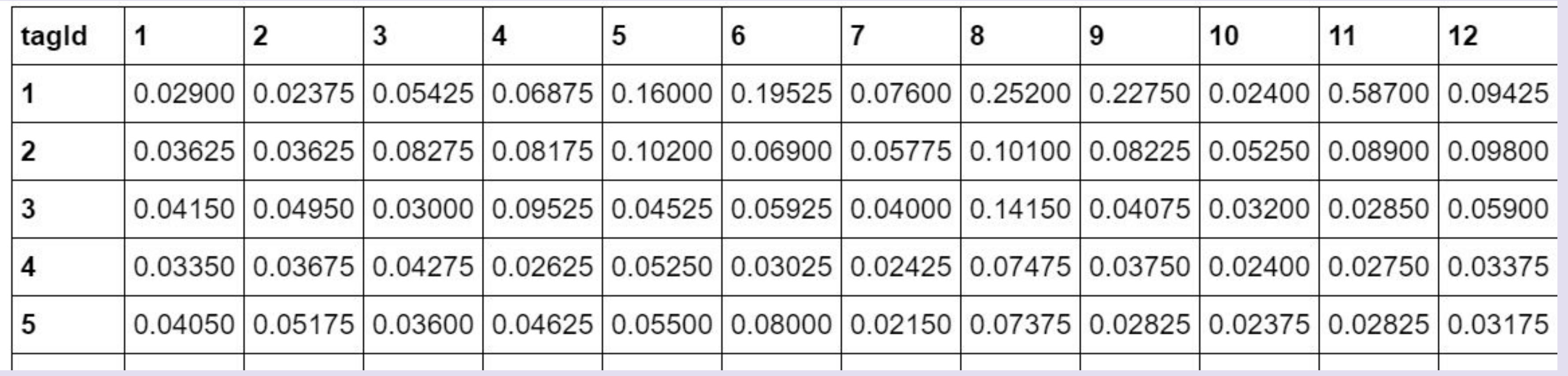
To best understand consumer preferences over time, we utilize the most direct feedback offered by viewers via the rating data. We assign labels to movies based on their consumer approval rating as well as their long-term sustainability. We utilize genome tag metadata, along with relevance scores, to classify common features of movies that perform well across time.



Movie retention of viewers across time, as a percentage of number of ratings received in its first year in the dataset.

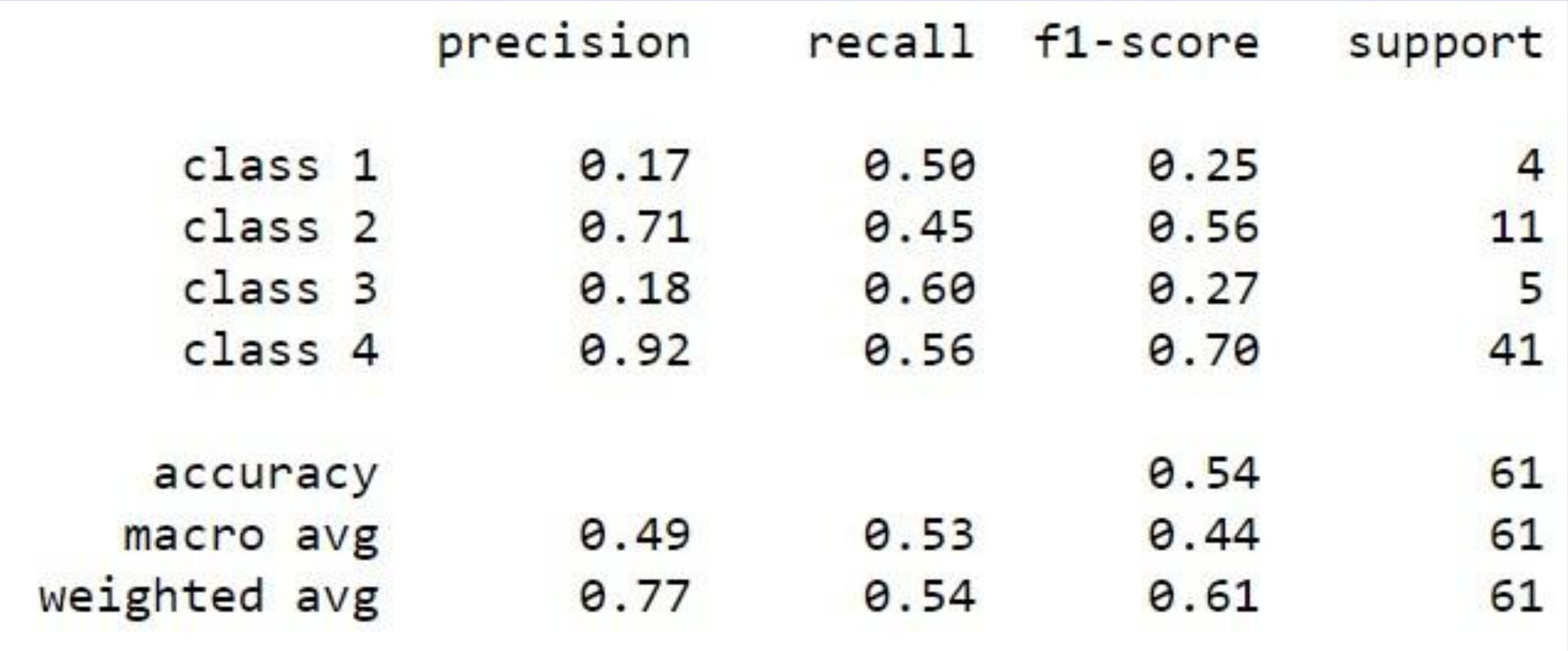
Model

After dividing our dataset into distinct training and test datasets, we train our model on the training dataset with a logistic regression. Since the content tag metadata is inherently made up of categorical variables, we considered three potential modelling methods: logistic regression, random forest, and an SVM classifier. While SVM relies n Euclidean distances, for which we have no good representation of for the dense matrix of genome tag assignments, and our matrix is much denser than can be reasonably handled by a random forest model, we utilize a logistic regression to form our model.



tagId	1	2	3	4	5	6	7	8	9	10	11	12
1	0.02900	0.02375	0.05425	0.06875	0.16000	0.19525	0.07600	0.25200	0.22750	0.02400	0.58700	0.09425
2	0.03625	0.03625	0.08275	0.08175	0.10200	0.06900	0.05775	0.10100	0.08225	0.05250	0.08900	0.09800
3	0.04150	0.04950	0.03000	0.09525	0.04525	0.05925	0.04000	0.14150	0.04075	0.03200	0.02850	0.05900
4	0.03350	0.03675	0.04275	0.02625	0.05250	0.03025	0.02425	0.07475	0.03750	0.02400	0.02750	0.03375
5	0.04050	0.05175	0.03600	0.04625	0.05500	0.08000	0.02150	0.07375	0.02825	0.02375	0.02825	0.03175

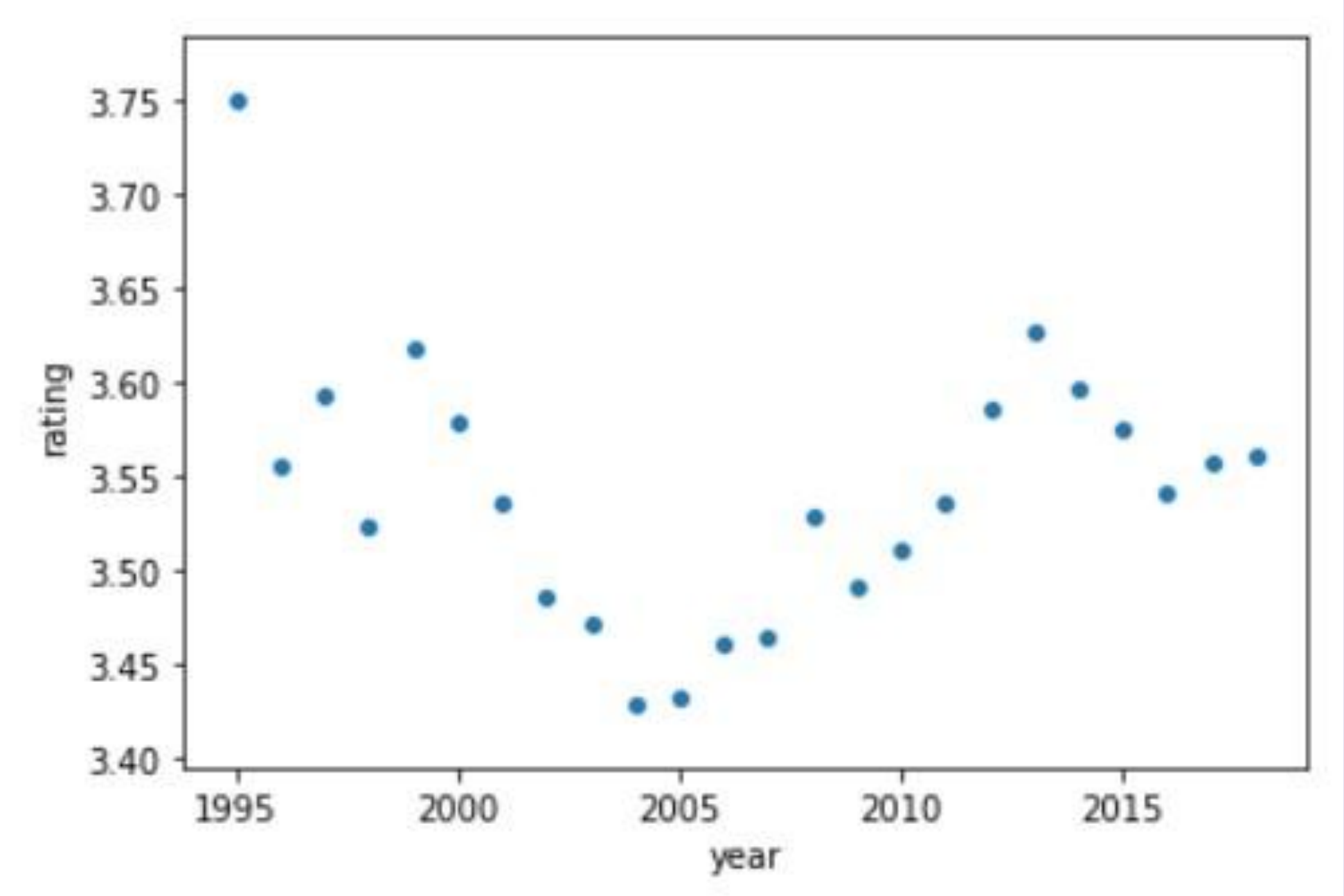
Results



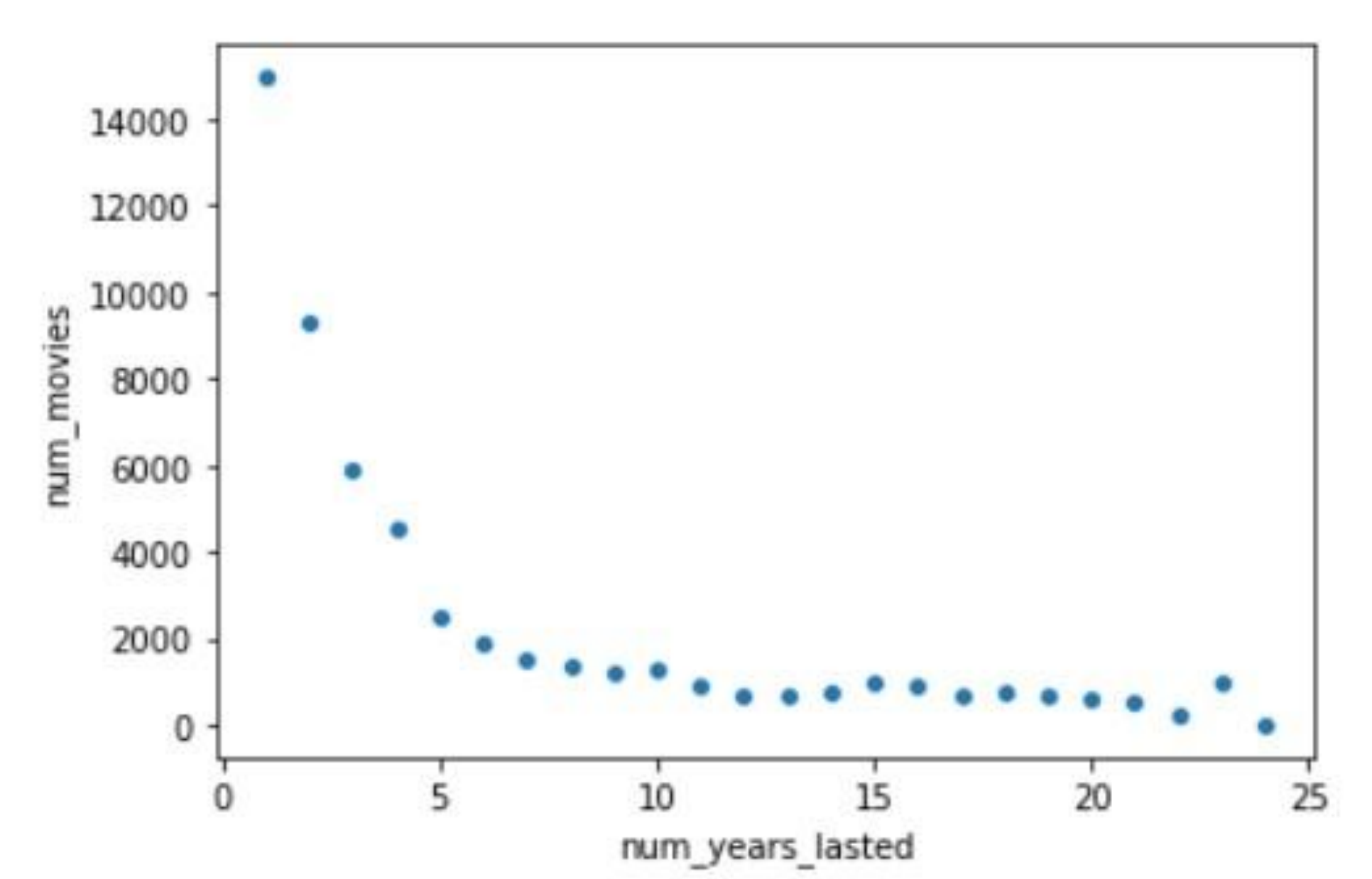
```
                precision    recall  f1-score   support

     class 1         0.17      0.50      0.25         4
     class 2         0.71      0.45      0.56        11
     class 3         0.18      0.60      0.27         5
     class 4         0.92      0.56      0.70        41

    accuracy                             0.54        61
   macro avg         0.49      0.53      0.44        61
weighted avg         0.77      0.54      0.61        61
```

Our model predicts class 4 movies, those that are both sustainable across time and consistently highly rated, with 92% accuracy.

Data Validation via ratings and popularity data across time



Average rating assigned to movies in each year over time.



Number of movies with ratings for number of years in the dataset.

Our model predicts movies that are highly sustainable and highly rated across time with 92% accuracy.

Applications

These conclusions about consumer preferences have far-reaching applications in movie recommendation systems for streaming services, film company executive decisions, and industry understanding the predictability of movie popularity regardless of contemporary trends.