Viral by Design: The Role of Social Networks in IMDB Ratings

Justin Lee
Department of Computer Science and Information Systems
University of Melbourne
juslee1@student.unimelb.edu.au

1 Abstract

In today's era, personal opinion - in the form in the form of Facebook likes or movie ratings - aggregates at massive scale due to a maximally interconnected Web, and generate heavy-tailed distributions due to the positive feedback loops inherent in recommendation algorithms. I propose that the very notion of virality should be viewed strictly as a social phenomena, as it is becomes increasingly irrelevant to any substantive qualities.

This research chronicles how a winning submission to a movie-rating Kaggle competition led me to this conjecture, the multiplicative effect of human subjectivity making itself visible in the dataset's low-dimensional spaces.

2 Methodology

This IMDB dataset's features include 1) highly right-skewed and sparse continuous variables, such as Facebook likes; 2) categorical variables with high cardinality, including names of individuals and locations; and 3) vector representations of textual data which exhibited comparatively low signal strength (see Figure 2).

Noteworthy: Adversarial validation revealed significant distributional differences between the training and test sets. A Random Forest Classifier trained to distinguish between the two datasets achieved an ROC-AUC score of 0.876 (see cell [11]).

2.1 Feature Engineering

For 1), the decision on data transformation techniques was non-trivial. The options considered included a logarithmic transformation and a Box-Cox transformation. An approach contemplated was discretisation into binary categories such as `regular_like_count` and `viral`. This decision was complicated by the discovered correlation between the number of likes and the target variable, which were significant in both the lower and upper ends of the distribution (Figure 1).

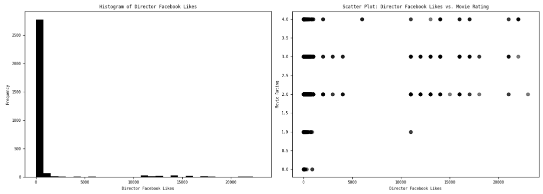

Figure 1: Correlations

For category 2), an exhaustive analysis of feature importance was undertaken via Pearson correlation and as a byproduct of testing Random Forest Classifiers (see Figure 2). *"The random forest classifier with its associated Gini feature importance, on the other hand, allows for an explicit feature elimination"* (Bjoern H Menze, 2014). This exposed categorical features like `num_faces_in_poster` which (clearly) only contributed noise, and useful ones like `rating` were discretised and subsequently transformed via one-hot encoding to enhance their representational utility in predictive modelling. Such methodology to remove unwanted features ensures that our final model does not learn noise, crucial in a relatively small dataset ($\sim$ 3000 instances).

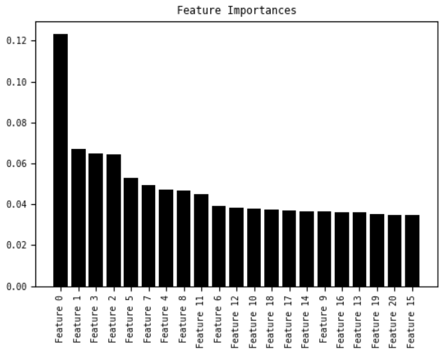

Figure 2: Feature Importance from an RFC (indexes ¿ 14 signal vectorised features)

Regarding 3), new textual embeddings were generated using BeRT (et al., 2019) to attempt a more rich encoding of a movie's semantic substance. These were compressed from the original 768 dimensions into a 2-dimensional space using t-SNE, *"...better than existing techniques at creating a single map that reveals structure at many different scales"* (van der Maaten and Hinton, 2008). The efficacy of these encoded features in predicting movie ratings was revealed later during natural feature selection via XG Boost (Figure 9).

No imputation was necessary for the dataset. Other noteworthy steps taken included binarising the `country` attribute (Figure 3) based on an arbitrary threshold. It was found that `cast_total_facebook_likes` served as a proxy for `actor_x_facebook_likes`; however, empirical ranking of feature importance (see Figure x) identified `actor_1_facebook_likes` as the most predictive, leading to the exclusion of the others. Incorporation of multiplicative features between director and actor likes in order to capture interactions were prototyped but ultimately discarded. Further experimentation included the feature `num_actor_appearances`, based on the frequency of actor names in the training dataset, but this proved unstable and probably unsuitable for scaling to larger real-world datasets.

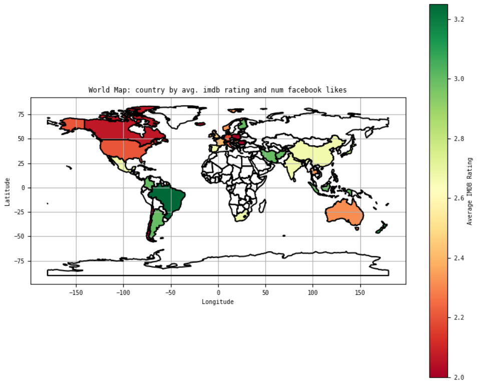

Figure 3: Geopandas, avg. ratings by country

Additionally, techniques to leverage the dataset's graph structure beyond the provided `average_degree_centrality`. This included community detection (Dhananjay Kumar Singh, 2023) and quantifying the number of collaborations between directors and actors. However, these features did not provide a sufficiently robust signal to be viable (cell [26]).

2.2 Aside: target label

This discretisation of `imdb_score_binned` raises an important methodological question: *should this task be framed as classification or regression?* The binning process obscures the original distribution of scores, which likely approximated a normal distribution (dataisbeautiful, 2018), thereby complicating straightforward regression applications (Brownlee, 2017). Nevertheless, treating the task as regression could preserve the ordinal nature of the scores, where for example, the difference in error between predicting a score of 1 versus 3 when the true label is 4 is more heavily penalised. Despite these considerations, I elected to treat this as a classification problem. This approach simplifies the modelling process and ability to handle the evident class imbalance with techniques specifically suited for categorical outcomes (Figure 4).

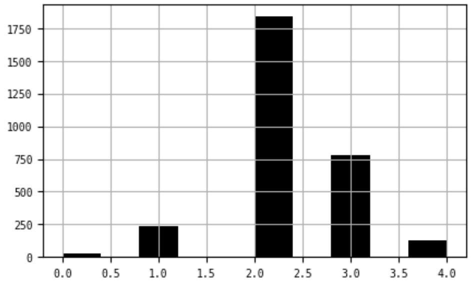

Figure 4: Distribution of `imdb_score_binned`

Selecting an appropriate accuracy measure was crucial. As mentioned, the target label, `imdb_score_binned`, has ordinal significance suggesting a potential application of mean squared error (MSE). However, given the reduced bin size, MSE might provide misleading results.

Traditional accuracy can overestimate model performance due to its bias towards majority classes (He and Garcia, 2009). Therefore, I opted for micro-accuracy, which has shown superior performance on imbalanced datasets, providing a more balanced measure by aggregating the contributions of all classes (Nur Suhailayani Suhaimi, 2022).

2.3 Learners

During critical analysis (Figure 5), it became evident that the feature set contained low discriminatory power. 2 strong learners - XG Boost and MLP compose the final ensemble, since it was necessary to extract rich representations from a relatively low-signal dataset.

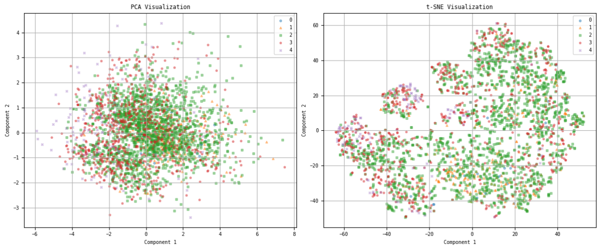

Figure 5: Dimensionality reduction of preprocessed features

During the critical analysis (Figure 5), it was concluded that the feature set possessed limited discriminatory power. To address this and other issues, the final ensemble was composed of three learners: Multinomial Logistic Regression (MLR), XG Boost and a Multilayer Perceptron (MLP). The choice of meta-classifier was searched as a hyper-parameter, with logistic regression found to be the most accurate empirically.

10-fold cross-validation was employed, as recommended in literature for its balance between bias and variance trade-off (Varma, 2006). However, for more precise error analysis, traditional hold-out partitions were used. This approach did lead to some recycling of training data in testing, resulting in over-optimistic estimates of generalisation. Future work should avoid this by strictly separating training and testing datasets, perhaps through nested cross-validation schemes (Tibshirani and Tibshirani, 2009).

The approach to **class imbalance** and **further feature selection** is outlined in the Discussion & Analysis Section.

3 Results

predicted

		0	1	2	3	4
actual	0	0	1	2	0	0
	1	0	10	43	0	0
	2	0	4	342	16	0
	3	0	0	65	93	1
	4	0	0	0	9	15

Figure 6: Confusion Matrix, Predicted vs. Actual

Label	Precision	Recall	F1-Score	Support
0	0.00	0.00	0.00	3
1	0.67	0.19	0.29	53
2	0.76	0.94	0.84	362
3	0.79	0.58	0.67	159
4	0.94	0.62	0.75	24
Accuracy			0.77	601
Macro Avg	0.63	0.47	0.51	601
Weighted Avg	0.76	0.77	0.74	601

Table 1: Precision, Recall, F1-Score, and Support for each label

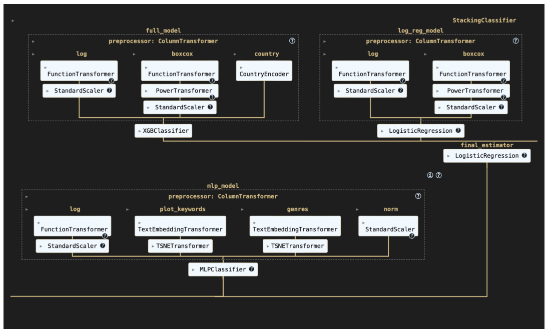

Figure 7: `sk-learn`'s Schematic of the Ensemble Classifier

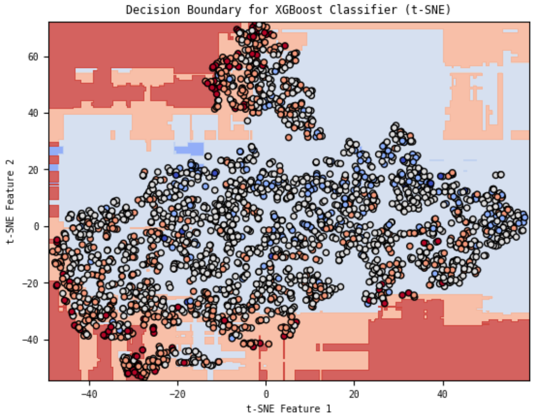

Figure 8: Decision Boundary from Initial XG Boost Classifier

Model	F1 Accuracy
k-Nearest Neighbors (kNN)	0.65
Gaussian Naive Bayes (GNB)	0.60
Linear Regression	0.55
Random Forest Classifier (RFC)	0.75
XG Boost	0.78

Table 2: F1 Accuracy of Various Models

4 Discussion & Critical Analysis

Initially, a regularised XG Boost classifier was employed, achieving 70% accuracy.

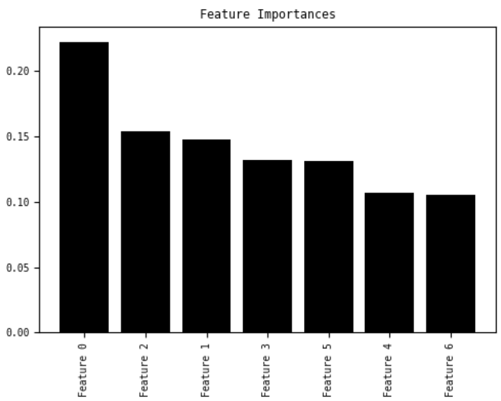

Figure 9: Features ranked by XG Boost `gain`

This score, while not exceptional, should be contextualised by considering that we are utilising quite extrinsic features such as `cast_total_facebook_likes` and `director_facebook_likes` to model an attribute intrinsically linked to the movie's content. This performance indicates a reasonable degree of predictability given the feature limitations (see Conclusion for feature interpretations).

As observed in Table 2, XG Boost outperforms non-boosting models, due to its ability to iteratively correct previous errors, which greatly aids in this highly non-linear and complex feature space - there are no definite linear relationships in our data, and highly adaptive models like SVM with RBF or polynomial regression were empirically found to be prone to overfitting. XG Boost also handles sparse data efficiently through its implementation of trees, though this is less relevant in this dataset due to the omission of the majority one-hot encoded features.

4.1 Baselines

To contextualise the cross-validation accuracy of 0.77, a stratified random classifier was used as a baseline classifier.

Model	Accuracy
Baseline Classifier	0.45 (± 0.03)

Table 3: Baseline Classifier Performance

The stratified random sample accuracy of 0.45 (± 0.03) serves as a reasonable benchmark for evaluating subsequent complex models. It is important to note that Figure 4 shows that more than 50% of the target labels are 2. Thus, a 0-rule classifier, always predicting 2, could achieve an even higher accuracy, highlighting the risk of bias towards the majority class. Nonetheless, this baseline suggests that we can do better through more sophisticated algorithms and refined feature selection.

4.2 Error Analysis

Per our confusion matrix (Figure 6), our ensemble model has expectedly strong performance on label 2, however it is not exactly discriminatory; inspecting the cells above and below reveals that the model struggled to delineate 2 from 1 and 3. (reference the t-SNE).

Examining Figure 5 (right side) suggests that we may be bottlenecked at the feature-level. There was decent separation of label 4 but we see essentially uniform dispersion of 2 through feature space. There was also a lack of representation (only 3 movies with rating 0 in the hold-out set) - synthetic duping techniques would not work (see 4.4); we simply need more data (and perhaps harsher critics!).

Performance was particularly poor on label 1, with over 20% of classifications being incorrect and predominantly biased towards label 2. This indicates a significant challenge in distinguishing the minority class, further emphasising the necessity of making decisions based on balanced evaluation metrics like recall. For example, the recall for label 1 was a mere 0.19.

Inspection of misclassified instances revealed that XG Boost failed to adapt to extremes in continuous values. For example, the following movie was misclassified as 3:

Feature	Value
ID	1060
Budget	26000
Lead Actor	Jason Statham
Lead Actor Facebook Likes	1000
Lead Actress	Qi Shu
Lead Actress Facebook Likes	447
Supporting Actor	Matt Schulze
Supporting Actor Facebook Likes	0.002326279
Gross	27572
Rating	PG-13
Country	France
Duration	255
Director	Louis Leterrier
Director Facebook Likes	92
Actor 2 Facebook Likes	0
Genre	Action—Crime—Thriller
Total Facebook Likes	25296447
Language	English

Table 4: A Misclassified Instance (`1060`)

However, a clear improvement can be seen in XG Boost's attempts to avoid biasing towards the majority class.

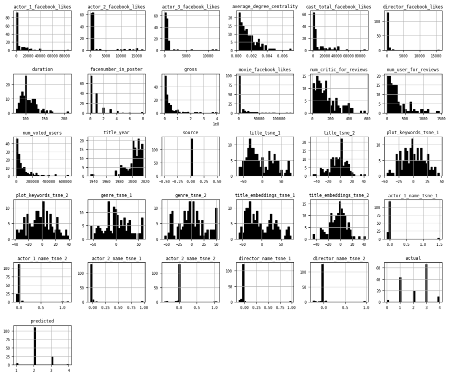

Figure 10: Feature Distribution of Misclassified Instances

4.3 Ensemble

The decision to transition from a single XG Boost model to a three-model ensemble was informed by both theoretical considerations and empirical evidence. Table 5 shows the means of key features for the entire dataset compared to instances where XG Boost failed to make accurate predictions.

The feature set possesses limited discriminatory power as illustrated in Figure 5. To address this challenge, the ensemble was composed of XG Boost, a Multilayer Perceptron (MLP), and Multinomial Logistic Regression (MLR). These models were selected due to their complementary abilities to extract rich representations from datasets with inherently low signal.

XG Boost, known for its efficiency in handling sparse data, utilises gradient boosting to sequentially correct errors from previous models (Chen and Guestrin, 2016). However, its tendency to focus on reducing overall prediction errors may lead to an inadvertent bias towards larger classes. This bias is reflected in the lower mean values for `actor_1_facebook_likes` and `director_facebook_likes` in instances XG Boost failed on, indicating potential oversights in capturing the nuances of less represented classes. This could explain the neglect in learning the decision boundaries for 1, 3. Given the observed pitballs using XG Boost with the selected features, a stacking model was ultimately employed.

On a high-level, MLP was chosen due to its efficacy in learning representations on the high-dimensional, textual embeddings, which simpler models fell short at.

Multinomial logistic regression, known for its robustness in multi-class classification scenarios, offers more nuanced decision boundaries for minority classes, such as those separating the minority classes 0 and 4; a decent degree of discrimination can be seen in Figure 5. These characteristics make it an ideal complement to the tree-based approach of XG Boost.

Hence, features were selectively distributed to the three classifiers based on the characteristics of the individual models. XG Boost also offered feature selection ability, providing greater affordance.

Another side effect is that the meta-classifier in the stacking framework can rectify biases towards majority classes, providing a more balanced and accurate prediction on `imdb_rating_binned`. Constructing an ensemble of these 3 unique models proved crucial on extracting flexible

4.4 Class Imbalance

To address the issue of class imbalance, synthetic data generation methods such as SMOTE were avoided as they have a diminished effect on strong classifiers (Mai, 2023). Instead, weightings were assigned to the training data during XG Boost training using inverse weights to counteract class imbalance. I empirically observed inherent trade-offs in doing this; accuracy was sacrificed on 2, leading to decreased overall accuracy but a more discriminatory model.

4.5 Feature Omission

XG Boost's built-in `gain` functionality was used to understand which features led to the highest training loss reduction. As shown in Figure 9, the feature space is extremely noisy, with no accessible decision boundaries. The data preprocessing pipeline was refactored in order to 1) lower dimensionality by omitting features based on XG Boost's `gain` ranking and 2) apply more aggressive log-transforms (Box-Cox) to the continuous features and remove outliers.

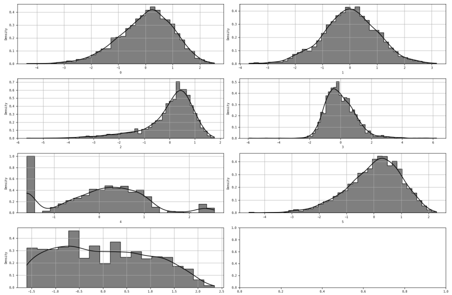

Figure 11: Distribution of Preprocessed Features

Feature	Whole Dataset Mean	XG Boost Failures Mean
Actor 1 Facebook Likes	7654.9	6429.3
Director Facebook Likes	778.8	276.0
Duration (minutes)	110	107

Table 5: Mean values of key features across the whole dataset versus. subsets where XG Boost failed.

Features such as gross, duration, director_facebook_likes, and num_critic_for_reviews contribute to understanding the overall reception and scale of a movie but may not directly correlate with extreme movie ratings (very high or very low). For instance, high-grossing films are generally not extremely poorly rated, making it difficult for the model to accurately predict low ratings (classes 0 and 1).

Furthermore, inspecting feature importance charts revealed that the original textual embeddings had low predictive power. All models consistently scored higher when these embeddings were omitted. Performing naive t-SNE dimensionality reduction and plotting the results revealed no useful structures within the embeddings. Clusters that emerged were not discriminatory, prompting the generation of fresh embeddings with BERT - *"language models utilise a similar embedding space for representing concepts"* (Kumar, 2023) and thus may provide useful implicit categorisation of the nuanced semantics of movies. Even then - they are loosely correlated, thus these embedded features are input only to one of the three base classifiers.

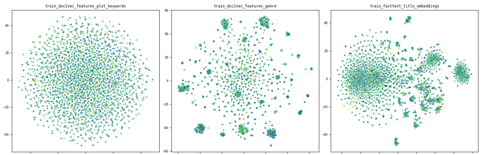

Figure 12: t-SNE of Original Embeddings

4.6 Regularisation

While the three-model ensemble reduced variance, to mitigate the potential introduction of bias through extra parameters (approximately 2000 parameters), regularisation was implemented across all three learners through an extensive random grid search combined with cross-validation techniques. This approach was chosen due to the combinatorial complexity involved in tuning multiple models (Bergstra and Bengio, 2012).

Additionally, probing the Kaggle test set helped refine the model parameters further. The resultant model exhibited a semi-healthy learning curve without signs of over-fitting (see Figure 13). However, the rapid plateau observed on the validation set, coupled with a declining trend on the training set, suggests a limitation in the feature set, referring again to Figure 5 (right side).

This limitation may stem from the attributes themselves or suboptimal feature engineering, rather than the model's inability to discern patterns, given the complexity of the model (Figure **??**).

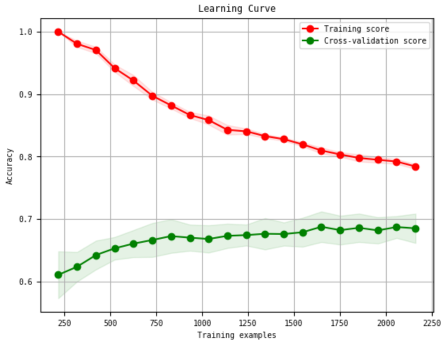

Figure 13: Learning Rates

5 Conclusions

The implementation of an uncharacteristically strong ensemble model effectively equalised performance on classes and reconciled feature under-utilisation seen with XG Boost alone.

This task was a valuable exercise in data mining: movie ratings, though ostensibly indicative of movie quality, are better predicted by viewer interactions rather than direct content attributes (at least given my feature representations 2).

This potentially underscores how subjective ratings are shaped more by public perception and collective bias than by the substantive content of the movies. This novel insight offers a promising interpretation in other domains requiring prediction on human-interaction data.

References

James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*.

Stefan Bauer Bjoern H Menze, Andras Jakab. 2014. *The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)*.

Jason Brownlee. 2017. Regression vs. classification in machine learning. *Machine Learning Mastery*.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

dataisbeautiful. 2018. Imdb score distributions of different movie genres [oc].

Prasenjit Choudhury Dhananjay Kumar Singh. 2023. Journal of network and computer advances in computers.

Devlin et al. 2019.

Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*.

Sai Kumar. 2023. *Bias Modelling and Mitigation in Diffusion Models*. Ph.D. thesis, Monash University.

Wei-Chao Cheng Tan-Ha Mai. 2023. From smote to mixup for deep imbalanced classification. *Department of Computer Science and Information Engineering, National Taiwan University,*.

Zalinda Othman Mohd Ridzwan Yaakub Nur Suhailayani Suhaimi. 2022. *Comparative Analysis Between Macro and Micro-Accuracy in Imbalance Dataset for Movie Review Classification*. Springer.

R. Tibshirani and R.J Tibshirani. 2009. *A bias correction for the minimum error rate in cross-validation. Annals of Applied Statistics.*

van der Maaten and Hinton. 2008. Visualizing data using t-sne.

Simon Varma. 2006. *Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics.*