ESRB Video Game Ratings

CSCI 5523 Data Mining Final Report (Spring 2021)

Anirudh Agarwal, Sabrina Arif, Ishan Joshi, Justin Nelson

1 Introduction

1.1 Overview

ESRB ratings are standardized classifications assigned by the Entertainment Software Rating Board for the purpose of providing content and age ratings to video games. Established in 1994, the ESRB scale utilizes an anonymous set of trained reviewers to dictate a game's rating. The ratings suggest age appropriateness and are classified as follows: E (everyone), ET/E10+ (everyone 10 years of age and older), T (teen, 13 and older), M (mature, 17 years and older), and AO (adults only, 18 years and older). According to the organization's website, 77% of parents use these ratings as the regular determinant for determining game safety for children [1], underscoring the influence ESRB maintains on game audience and marketing.

1.2 Descriptors

Despite their ubiquitous use, ESRB ratings are assigned based on both objective and subjective factors. Several of these factors are truly binary, such as the presence/absence of blood, alcohol and drug references, and nudity. Others are of a relative scale, for instance "violence" vs "intense violence", "comic mischief" vs "crude humor", "language" vs "strong language" and "suggestive themes". Although the ESRB website notes that these relative ratings are intended to differentiate mild from moderate or severe occurrences of these factors, the subjectivity of such assignments begs the question of rating strictness.

1.3 Change of ESRB scale over time

There has been a notable increase in the popularity of first-person shooter video games in the last two decades [4], as well as increased tolerance of violence and sexuality in video games. Of the top 10 best-selling video games in 2019, four were rated "M" [3] compared to only one out of the top 10 best-selling video games in 2006 [5]. Given that the ESRB rating system has been employed for nearly 25 years, we were also curious how rating strictness had changed to reflect this cultural shift.

2 Aims

The Video Games Rating By 'ESRB' dataset provided by Mohammed Alhamad on Kaggle [2] is a collection of nearly 2000 ESRB-rated game titles. The attributes included in this dataset are the same content descriptors used by reviewers to assign ratings. Aside from the ESRB rating itself, all of the thirty-four features are binary. Additionally, a test set of nearly 500 titles was provided by the author. Given these pieces of information, our primary goal was to use this dataset to assess ESRB rating strictness, an objective characterized by three aims:

(i) Conduct exploratory analysis of the dataset itself and the potential relationship between release data and ESRB ratings.

(ii) Use a training set to build a classifier to assign game ratings based on the content features.

(iii) Test the model's capacity to accurately predict the ratings of a test set.

3 Data Augmentation

3.1 The Algorithm

The dataset our group used was missing a key metric that we wanted - release dates. We created an algorithm that first checked for non-ASCII characters in a video game title. The approach of using non-ASCII characters to determine English vs. non-English video game titles is not perfect. This is because the English language borrows words from other languages (e.g. café). The next step takes all the white-space in the video game title and converts it to hyphens and all the text to lowercase (e.g. Video Game Title A becomes video-game-title-a). The algorithm then queries an online database with this basic string. If the query was successful then capture the video game release date in the response and move on to the next video game title. Approximately 80% of the video game titles in the dataset fell into this category. If the query was unsuccessful then the algorithm attempted to sanitize the string.

3.2 Sanitization Techniques

The algorithm checks for digits at the end of the video game title and if found, condenses it in the string. This technique was used because a number of the video game titles were numeric sequels. The next technique is to check for an apostrophe in the video game title. If found, the apostrophe and hyphens before and after the apostrophe in the string are removed. Next, the algorithm checks for a single character in the string. It is likely if an apostrophe was found in the video game title that a single character got abandoned in the string. However single characters that are a part of the video game title should be left alone. After all the prior sanitization steps, the newly sanitized string is used to query the online database again. If the query still fails then the last step is to incrementally remove sections of the string, starting from the end, until either success is returned from the query or the algorithm reaches a predetermined minimum size. The algorithm would take a long string like "this-is-a-video-game-title4" and in the next query would use "this-is-a-video-game". There are numerous issues with this approach. The biggest issue is that video games in a series got classified as the same video game. This was definitely the case if the video game titles differed by only a numerical digit at the end. Finally, based upon the string for a video game title the algorithm drops any duplicates in the dataset.

3.3 Manual Verification

After the algorithm completed, we manually inspected the results. We verified a significant number of the release dates, with most being accurate to within a single day. The single day inaccuracy wasn't an issue because we were concerned with the release year of the game only. The inaccuracy was caused by video game releases happening at the same time worldwide. After verifying the retrieval of valid release dates, we manually updated the dataset for the video game titles that the algorithm was unable to find release dates for.

3.4 Exploratory Data Analysis

3.4.1 Features

To begin exploratory analysis, the ESRB rating distribution was assessed. "T" was found to be over-represented in the dataset, accounting for approximately 37% of titles, whereas the remaining ratings ("E", "ET", "M") each accounted for approximately 19-21% of titles (Figure 1A). "Console", a content descriptor indicating if the game was released on the PlayStation 4 ("0") or both the PlayStation 4 and Xbox One ("1"), yielded a nearly equal distribution between 1's and 0's, no matter the ESRB rating (Figure 1B). This was not the case for the remaining dataset features: every attribute, other than "console", displayed a class imbalance in which TRUE ("1") was present at a frequency of <10% for nearly every category. Looking broadly at the number of attributes present, it was obvious that the dataset was skewed right (Figure 1C) indicating that the majority of games contained TRUE values for only a small number of attributes. In fact, nearly 34% of games contained only two attributes with a TRUE value.

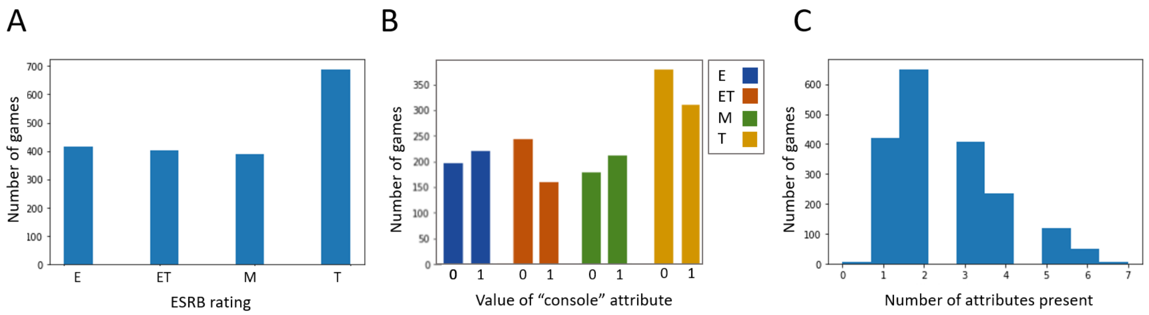


Figure 1: (A) Number of games per ESRB rating (B) Number of games per ESRB rating and console. 0= Playstation 4 only, 1= Playstation 4 and Xbox One. (C) Number of attributes with TRUE ("1") value among all games.

3.4.2 ESRB Rating Variables

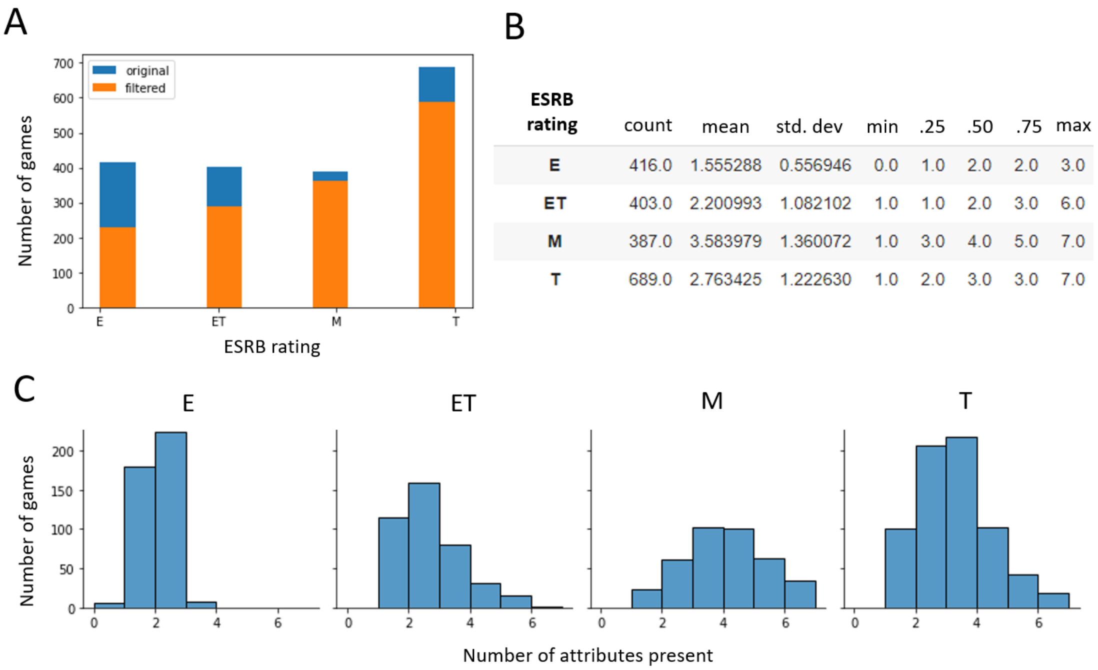
In an attempt to filter and clean the dataset, we tried setting a threshold value for the minimum number of attributes and began with an arbitrary value of 5%. This reduced the dataset to 1469 total games. To check how this altered the distribution of ESRB rating, we plotted the original and filtered rating distribution in a stacked bar chart (Figure 2A) and noticed an interesting trend: “E” rated games were reduced much more than “T” and “M” rated games. Specifically, “E” was reduced from comprising 22% of the dataset to only 12%, whereas “M” was barely affected (20% reduced to 19%). This indicated that the number of attributes alone was an indicator of ESRB rating. We confirmed this suspicion by calculating descriptive statistics (Figure 2B) which revealed that “E” rated games had the lowest average number of descriptors (1.56) followed by “ET” (2.2), “T” (2.76) and “M” (3.58). The opposite trend was true for standard deviation and data-spread (Figure 2C). In lieu of this discovery, we opted not to filter the dataset based on a minimum-attribute cutoff.

3.4.3 Attributes & Correlation

Several content features intuitively appeared to be redundant, for example “blood” and “blood_and_gore”. We proceeded to correlation analysis to explore if any dependent relationships existed between attributes or among the attributes and ESRB ratings. Several metrics were tested, and among Jaccard, Manhattan, Kendall, Pearson, Spearman, Accuracy and Cosine, no strong correlations were found among features. In terms of attribute-ESRB rating relationship, associations (Pearson) were found among “strong_language” and the “M” rating, “blood_and_gore” and “M”, “fantasy violence” and “ET”, and “no_descriptors” and “E” (Figure 3). The latter was not surprising given earlier findings.

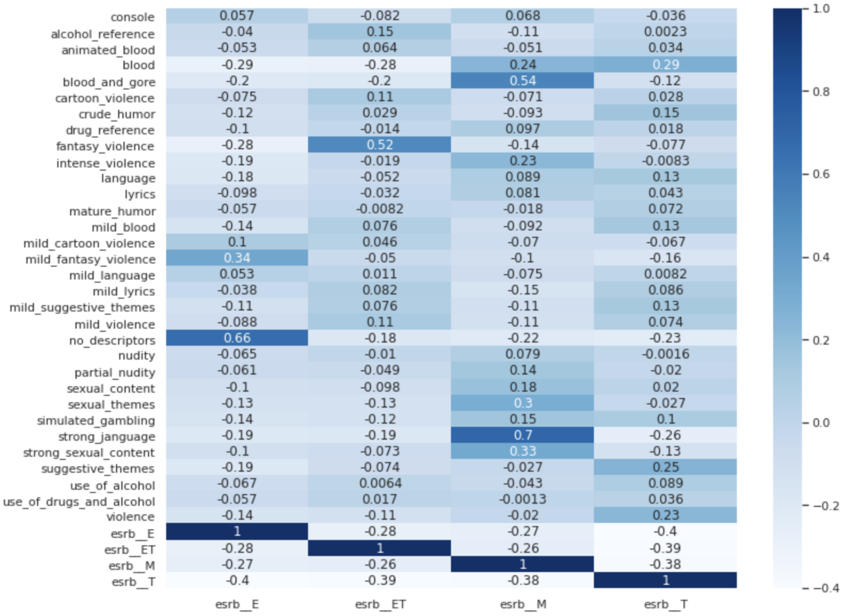
3.4.4 Time-Series

A time-series analysis was also done as well to view how our dataset was distributed over time (Figure 4). Here we can see the number of games released per year in our dataset, split up based on the ESRB rating of the games. It should be noted that most of our dataset falls within the 2015-2020 timeframe, with almost no games before this time period. As a result, our secondary objective of performing a historical comparison of the ESRB rating scale and its progression over time was put on hiatus. Additionally, this time-series analysis reaffirmed that there was a larger number of T-rated games within our dataset than any



ESRB rating	count	mean	std. dev	min	.25	.50	.75	max
E	416.0	1.555288	0.556946	0.0	1.0	2.0	2.0	3.0
ET	403.0	2.200993	1.082102	1.0	1.0	2.0	3.0	6.0
M	387.0	3.583979	1.360072	1.0	3.0	4.0	5.0	7.0
T	689.0	2.763425	1.222630	1.0	2.0	3.0	3.0	7.0

Figure 2: (A) Number of games per ESRB rating: filtered by number of attributes (orange) or unfiltered (blue) (B) Descriptive statistics by ESRB rating: number of attributes with value of TRUE ("1"). (C) Number of attributes with TRUE ("1") value among all ESRB ratings.



	esrb__E	esrb__ET	esrb__M	esrb__T
console	0.057	-0.082	0.068	-0.036
alcohol_reference	-0.04	0.15	-0.11	0.0023
animated_blood	-0.053	0.064	-0.051	0.034
blood	-0.29	-0.28	0.24	0.29
blood_and_gore	-0.2	-0.2	0.54	-0.12
cartoon_violence	-0.075	0.11	-0.071	0.028
crude_humor	-0.12	0.029	-0.093	0.15
drug_reference	-0.1	-0.014	0.097	0.018
fantasy_violence	-0.28	0.52	-0.14	-0.077
intense_violence	-0.19	-0.019	0.23	-0.0083
language	-0.18	-0.052	0.089	0.13
lyrics	-0.098	-0.032	0.081	0.043
mature_humor	-0.057	-0.0082	-0.018	0.072
mild_blood	-0.14	0.076	-0.092	0.13
mild_cartoon_violence	0.1	0.046	-0.07	-0.067
mild_fantasy_violence	0.34	-0.05	-0.1	-0.16
mild_language	0.053	0.011	-0.075	0.0082
mild_lyrics	-0.038	0.082	-0.15	0.086
mild_suggestive_themes	-0.11	0.076	-0.11	0.13
mild_violence	-0.088	0.11	-0.11	0.074
no_descriptors	0.66	-0.18	-0.22	-0.23
nudity	-0.065	-0.01	0.079	-0.0016
partial_nudity	-0.061	-0.049	0.14	-0.02
sexual_content	-0.1	-0.098	0.18	0.02
sexual_themes	-0.13	-0.13	0.3	-0.027
simulated_gambling	-0.14	-0.12	0.15	0.1
strong_language	-0.19	-0.19	0.7	-0.26
strong_sexual_content	-0.1	-0.073	0.33	-0.13
suggestive_themes	-0.19	-0.074	-0.027	0.25
use_of_alcohol	-0.067	0.0064	-0.043	0.089
use_of_drugs_and_alcohol	-0.057	0.017	-0.0013	0.036
violence	-0.14	-0.11	-0.02	0.23
esrb__E	1	-0.28	-0.27	-0.4
esrb__ET	-0.28	1	-0.26	-0.39
esrb__M	-0.27	-0.26	1	-0.38
esrb__T	-0.4	-0.39	-0.38	1

Figure 3: Heatmap of pearson correlation (features vs various ESRB Ratings)

other ESRB rating, which could result in an imbalance. To deal with this imbalance, our dataset was then resampled using undersampling. This removed games from the ‘E’, ‘ET’, and ‘M’ categories of our dataset until there was an equal number of games within each dataset. Both the imbalance and re-balanced training data were used in the modelling section to determine the impact resampling had on our dataset overall.

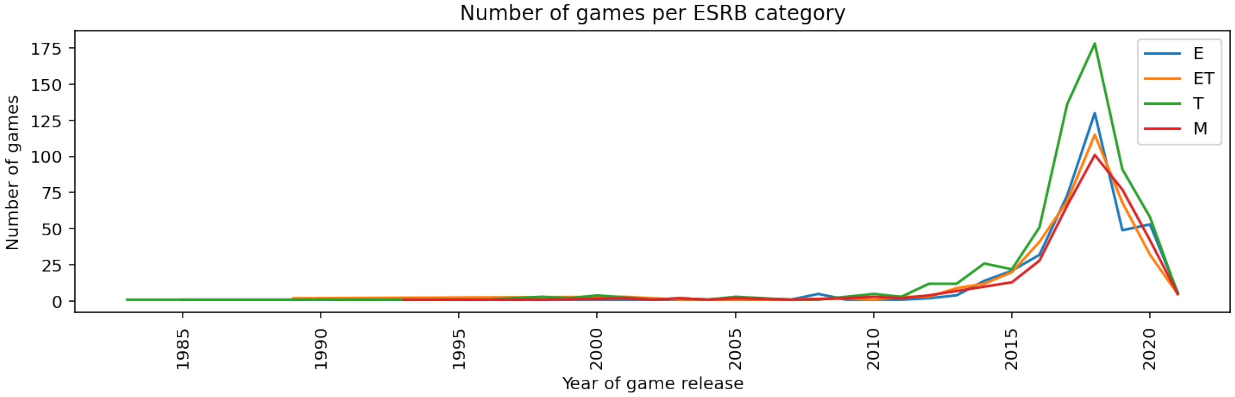


Figure 4: Games released each year by ESRB category

3.5 Model Training and Validation

Our dataset has 32 binary features (excluding game titles, and final target ESRB ratings). Our intuition was that there might exist certain visible patterns in the features that could be exploited to aid the process of building more accurate predictive models. Thus, we tried some variations in terms of feature manipulation, based on intuition and findings from EDA, to enhance our model performance.

3.6 Preprocessing

3.6.1 Feature Selection

After analyzing counts of all the features, we found that only console feature had an equal distribution between 0 and 1 values. Also, the 0 and 1 values were approximately equally divided in all the target values. As mentioned earlier, Figure 2B displays a plot of 0 and 1 values of console feature for each of the ESRB ratings in our dataset. Since the console feature was not crucial in making a decision, and was just adding noise during model training. Thus, we removed the feature for model training.

3.6.2 Dimensionality Reduction

Based on our analysis on the average number of features that are true, we realized that there is high sparsity in our data. Figure 5 shows a sparsity plot of our training dataset. Thus, with the intuition that all features might not be crucial to the decision making process, we employed the unsupervised technique of dimensionality reduction. We used PCA to try variations of reducing features from 32 attributes to (10, 15, 20, 25). In all of the above cases, the performance of our model deteriorated on validation set, hence we decided not to move forward with this variation and keep all the features for the training. We hoped to rely on internal feature selection mechanisms by the ML models. With this intuition, we created a new feature, "total_true_features" (row-wise sum) and included it with other features during model training. The results showed a degradation in validation accuracy and F-1 across all models. Only the Random Forest Classifier model showed a boost in train accuracy and F-1, but even after preventing its over-fitting, the validation metrics were worse than before inclusion of this feature. Thus, we dropped the engineered feature from our final model.

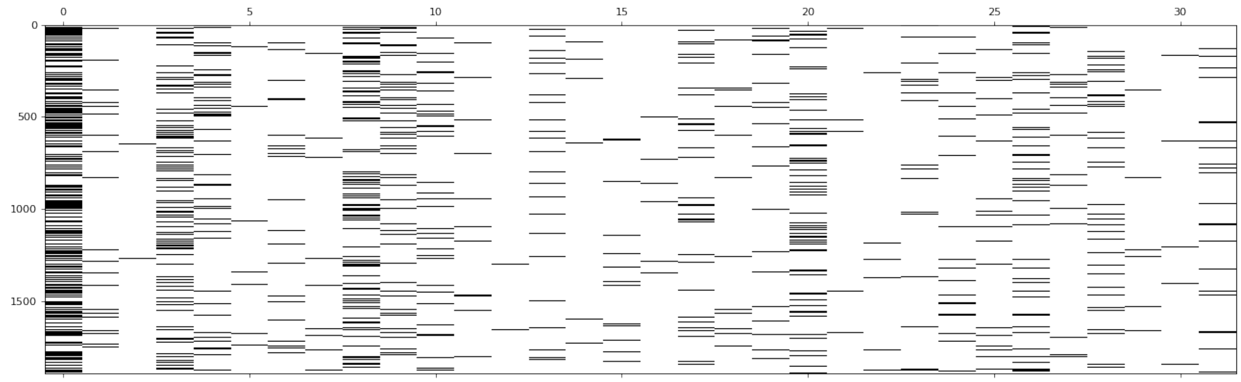


Figure 5: Sparsity plot of training data

3.7 Training

In order to try multiple variations of models and data, we created a generic framework for model training and validation. For primary modelling, and to figure out the right hypothesis space, we analyzed a collection of ML models, including linear, Bayesian, non-linear and even ensemble models. The models that we analyzed for various variations included Multinomial Naive Bayes, Bernoulli Naive Bayes, SGDclassifier (Stochastic Gradient Descent linear SVM classifier), K-Nearest Neighbours, XGBoost Classifier, RBF-Kernel SVM, Linear-Kernel SVM and RandomForestClassifier. The following choice of models covered most spectrums of model complexity and missed ML model complexity is most likely represented by the above collection of models. Based on our understanding of data via EDA, the dataset was binary, highly sparse, with no significant visible patterns, and consisted of non-correlated features with respect to the target variables. The intuition was for non-linear models or ensemble models to work the best, as separating decision boundaries had to be non-trivial for a dataset of this nature.

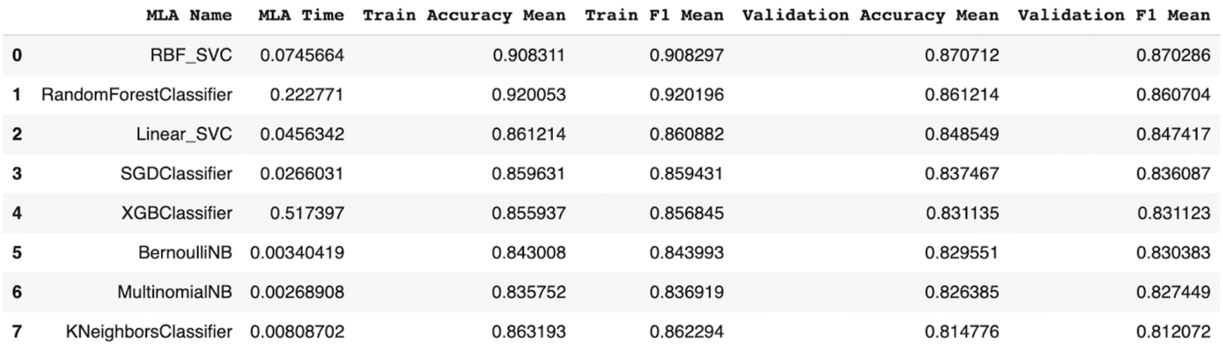
3.8 Validation

3.8.1 Results

For validation, we used a 5-cross validation technique and recorded metrics across various ML models. The final metrics were then averaged out over all the folds. In terms of metrics we recorded both accuracy and F1 scores, due to an imbalance in our data. Each of the metrics were 'weighted' as per the number of samples of each class, so that getting a rarer target wrong was more expensive than getting a frequent one. From the above metrics, we noticed that our intuition about which models would perform the best was approximately right. Given the binary nature of the data, RandomForestClassifier was among the top-performers. Also, with a non-linear decision boundary, RBF-Kernel performed the best in terms of both validation accuracy and F-1. Thus, we decided to move forward with these two models. We also recorded the MLA training time as we wanted to avoid models with high time complexity. The purpose of analyzing the training data metrics was to aid in the process of hyper-parameter tuning. We noticed that RandomForestClassifier was over-fitting to a small degree as it had considerably higher training metrics than RBF kernel SVM. Thus, we narrowed down our efforts to tuning the Random Forest model.

3.8.2 Re-sampling

As reported in previous sections, our dataset has a medium skew in terms of targets such that the category 'Teen' is over-represented. The rest of the categories have approximately equal representation. Thus, in theory all ML models might develop a bias towards predicting the higher represented target. As a result, we tried under-sampling the over-represented targets, to nullify the skew. After training and



	MLA Name	MLA Time	Train Accuracy Mean	Train F1 Mean	Validation Accuracy Mean	Validation F1 Mean
0	RBF_SVC	0.0745664	0.908311	0.908297	0.870712	0.870286
1	RandomForestClassifier	0.222771	0.920053	0.920196	0.861214	0.860704
2	Linear_SVC	0.0456342	0.861214	0.860882	0.848549	0.847417
3	SGDClassifier	0.0266031	0.859631	0.859431	0.837467	0.836087
4	XGBClassifier	0.517397	0.855937	0.856845	0.831135	0.831123
5	BernoulliNB	0.00340419	0.843008	0.843993	0.829551	0.830383
6	MultinomialNB	0.00268908	0.835752	0.836919	0.826385	0.827449
7	KNeighborsClassifier	0.00808702	0.863193	0.862294	0.814776	0.812072

Figure 6: Models Training and Validation (5-cross validation)

validating ML models on our newly re-balanced data, we found that the performance of models as a result of under-sampling decreased for all models across both metrics. This led to the conclusion that the negative of losing training records due to under-sampling outweighed the benefits of solving the class imbalance issue in the context of our problem.

3.8.3 Hyper-parameter tuning

From training and validating ML models across various variations in data, we noticed that RandomForestClassifier performed best in terms of learning the training data, but was over-fitting slightly as per the difference in training and validation metrics. As a result, we decided to focus our efforts on hyper-parameter tuning for RandomForestClassifier. To prevent over-fitting, we focused on parameters including 'n_estimators' (number of decision trees) [10, 30, 50, 70, 100], 'criterion' [gini, entropy], 'max_depth' (maximum of depth of each tree) [5, 10, 15, None], 'max_features' (max features sampled for decision tree construction) [log2, auto, sqrt]. We used GridSearchCV to choose the best parameter values, over the training set, through cross-validation. Below are the best parameter values we received using 'weighted_f1' scores as the selection criteria.

4 Model Intuition

As discussed earlier, we did expect either a non-linear or an ensemble model to perform best for our dataset. The reason being that there were no features with high-correlation to our final target that could be picked up by a linear model. Also, given the binary nature of our dataset, we expected a decision tree based classifier to give the best performance, but a single decision tree would have led to very high variance, due to sparsity in our data. We also did not find much correlation among the features, and the ineffectiveness of dimensionality reduction proved that we did not have much noisy or insignificant features. Thus, a latent combination of all features is what was required to learn a good decision boundary for our use case. The ensemble of decision tree learners for our non-correlated, Boolean dataset, reduced the overall model variance and also solved the issue of the limited training set through 'Bagging'. Thus, we determined this to be the best model in theory and in practice for our use-case.

5 Model Testing

Along with a training set of approximately 1900 data objects, we also obtained an unseen test set of around 400 data items to report our final metrics. After trying multiple iterations of feature manipulation, re-sampling and analyzing validation metrics to narrow down the right hypothesis space, we finalised our training pipeline and ML model. We chose ‘Random Forest‘ to be our final model. We trained a RandomForestClassifier over 31 features (excluding ‘target‘ ESRB ratings, ‘game titles‘, and ‘console‘). The classifier was instantiated using the best hyper-parameter values obtained from GridSearchCV. Due to the ineffectiveness of dimensionality reduction, under-sampling and engineered feature ‘total_true_features‘ during validation, we decided to exclude them from our training pipeline. After training the Random Forest model on the entire training dataset, we used the unseen test set to make final predictions. Below are our final results that we were able to achieve for the problem of ESRB ratings prediction using content features.

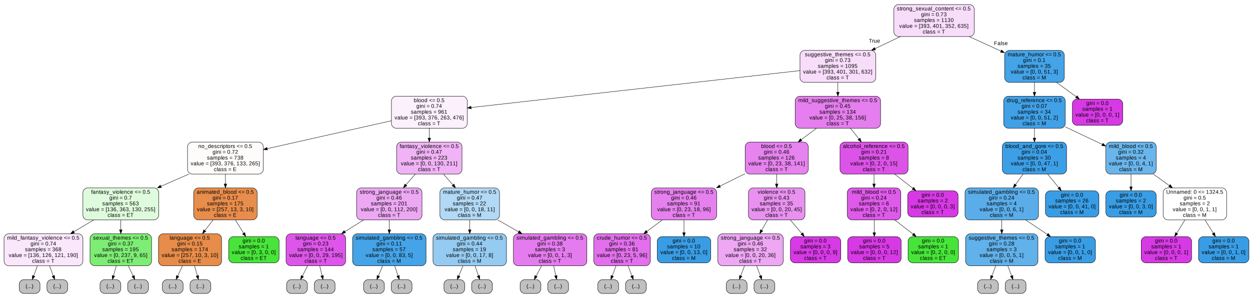
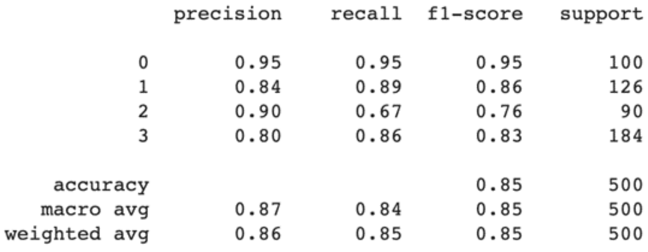


Figure 7: Decision Tree Visualization (max_depth=5)



```
              precision    recall  f1-score   support

           0       0.95      0.95      0.95       100
           1       0.84      0.89      0.86       126
           2       0.90      0.67      0.76        90
           3       0.80      0.86      0.83       184

    accuracy                           0.85       500
   macro avg       0.87      0.84      0.85       500
weighted avg       0.86      0.85      0.85       500
```

Figure 8: Classification report for final model (test set)

6 Conclusions

6.1 Summary & Findings

The goal of this project was to be able to accurately classify games on the ESRB scale, determine the overall rigidity of the scale as a whole and do an analysis of how the scale has changed over time. The primary findings of this project were that we could indeed predict a game’s ESRB rating to a high degree of accuracy using its game descriptors. This in turn also tells us that these ratings hold some predictable/quantifiable properties that allowed us to properly classify games. Thus, we conclude that the ESRB rating overall is not subjective. In particular, we found relationships between attributes and a resultant rating as well as a theory that the number of descriptors also plays a role in determining the rating. These findings could be rather substantial for the gaming industry as all major platforms are reliant on an ESRB rating before a game can be published on that platform. The ESRB rating process, as discussed previously, is completely manual and is reliant on human supervision. The work done on this project could hopefully be a stepping stone to help automate the ESRB rating process for which so many games are reliant on and give game developers an idea for who their audience will be.

6.2 Future Goals

Given the time and dataset constraints associated with our project, we were not able to do a proper comparison of how the ESRB scale has changed or potentially evolved over time. Our first objective moving forward is to expand our dataset to include more games and specifically older games. Adding more games to our dataset could give us a better representation of the ESRB scale as we are currently are reliant on the sample of games in our dataset, which could potentially miss some patterns or trends that are genre-specific or time-sensitive. By adding older games to our dataset, we would get a better idea of how games have changed in rating according to the scale rather than just the 2015-2020 timeframe games we currently have. Another avenue to continue exploring in our work is further increasing the accuracy of our models. While adding more data to our dataset could certainly improve our model accuracy, it is definitely worth exploring other models and approaches as our experiments were not exhaustive.

7 Contributions

7.1 Sabrina Arif

Along with other group members, Sabrina Arif aided in dataset choice and helped formulate research questions. She contributed to exploratory data analysis, including the exploration of attribute TRUE/FALSE distribution and the relationship of "TRUE" attribute count and ESRB rating. Additionally, she took the lead in drafting the Proposal, Interim, and Final reports.

7.2 Anirudh Agarwal

Anirudh contributed towards basic and advanced exploratory data analysis, specifically towards correlation, dimensionality reduction and redundancy analysis. His major contribution was towards feature manipulation & engineering, building various ML models training pipeline, hyperparameter tuning, their validation & evaluation. In terms of reports, he helped with scoping of the research question, porting drafts to latex templates and adding to the technical aspects of the Proposal, Interim and Final reports.

7.3 Ishan Joshi

Ishan Joshi performed basic exploratory data analysis, resampled the dataset, created a time-series analysis of the dataset, and assisted in both research formulation and concluding analyses. Additionally, he helped generate the Proposal, Interim, and Final report drafts into LaTeX.

7.4 Justin Nelson

Justin Nelson helped choose a research question and searched for a dataset to use. He also searched for other datasets that contained release dates. He wrote the initial release date retrieval algorithm and subsequent updates to the algorithm. During exploratory data analysis he found the presence of non-English game titles and duplicated rows in the dataset. He contributed to the final presentation slide deck and edited the final video presentation together. Additionally, he contributed to the Proposal, Interim, and Final reports. Finally, and most importantly to him, he had a lot of fun with the group. They were the best group members he's had in a long time.

8 Associated Code

https://github.com/joshi304/ESRB-Rating-Predictor

References

[1] : *Ratings Guides, Categories, Content Descriptors*. Jun 2020. – URL https://www.esrb.org/ratings-guide/

[2] Alhamad, Mohammed: *Video Game Ratings by 'ESRB'*. https://www.kaggle.com/imohtn/video-games-rating-by-esrb

[3] Kain, Erik: *The 20 Best-Selling Video Games Of 2019*. Jan 2020. – URL https://www.forbes.com/sites/erikkain/2020/01/17/the-20-best-selling-video-games-of-2019/?sh=51f25cb873da

[4] Kain, Erik: *'Call of duty' totally dominated the 20 best-selling video games of the decade*. Jan 2020. – URL https://www.forbes.com/sites/erikkain/2020/01/17/the-20-best-selling-video-games-of-the-decade/?sh=68a73cc2f6db

[5] Seff, Micah: *Best-Selling Games: December 2006*. Jun 2012. – URL https://www.ign.com/articles/2007/01/12/best-selling-games-december-2006