ISYE-6420 Spring 2024 Final Project

Video Game Metacritic Score Prediction with Bayesian and Deterministic Multivariate Regression

Saad Siddiqui
ssiddiqui60@gatech.edu
GTID: 93956081

I. Abstract

This report details an open-ended investigation into the efficacy of frequentist and Bayesian regression modeling techniques in the context of predicting the Metacritic score for 3,493 video games from the Steam online video game marketplace.

The report begins with a brief discussion of the Steam Marketplace and the motivation for the investigation in the context of hypotheses relating Steam community dynamics and video game quality to Metacritic scores. It then provides a high- level overview of the dataset used to investigate the aforementioned hypotheses followed by a detailed discussion of the preprocessing and feature engineering techniques applied to structured and unstructured predictors in the data. The processed data is then used to create three progressively more complex modeling sub-datasets. The report then presents a discussion of three Bayesian multivariate regression models, specifically Truncated Normal, Poisson, and Negative Binomial regression. The strengths, assumptions, biases, and limitations of each these models in terms of modeling the Metacritic score is analysed. This is followed by a brief discussion of fitting, tuning, and evaluating two popular deterministic regression models, namely Ridge Regression and Decision Trees, as conventional supervised machine learning baselines. The Bayesian models are compared and contrasted with each other as well as with the deterministic models in terms of fitness and time complexity across all three sub-datasets. The report concludes with an assessment of hypotheses in lieu of the investigation and a summary of potential areas for future work.

II. Introduction

A. Steam and Metacritic

Video games have and continue to be a billion dollar business [1], with an estimated revenue of $347 billion in 2022, with PC games accounting for an estimated $50.23 billion of this amount [2]. While there has been a proliferation of many different developer-specific PC video game market places such as the Epic Games store [3] and Electronic Arts (EA)'s proprietary client [4], Valve Corporation's Steam continues to dominate the PC gaming community in terms of market share and volume of concurrent players [5]. In addition to allowing users to purchase video games and related content, Steam also has active user-driven communities built around reviewing and recommending video games. This is in contrast to review aggregation sites such as Metacritic [6] which collates and curates reviews from selected publications and reviewers before weighting them based on a proprietary heuristic to quantify a "critical consensus for games" in the form of a discrete ordinal score ranging from 0 to 100 inclusive.

B. Investigation Motivation and Hypotheses

Based on this information, it is reasonable to assume that Metacritic is driven exclusively by the opinions of a handpicked set of reviewers and publications that may or may not be representative of wider user-driven sentiments around a video game. However, this paper posits that despite its overt exclusion or lack of preference of community or user-driven reviews from its scoring heuristic, Metacritic's score is, in part, driven by the game's community on Steam. Steam's description of the game's genre, developers, price, and other game-specific attributes, in conjunction with user behavior and sentiment regarding the game in the community, is posited to influence the Metacritic score through multiple mechanisms.

Firstly, the level of attention a game receives on Steam can influence buyer perceptions of a game, which could translate to higher engagement with the game as quantified by play time, reviews, and recommendations on Steam. On a second level, Steam user behavior is posited to have a "word-of-mouth" effect on shaping the wider community perception of a game which, in turn, could result in more professional reviews and thus potentially higher scores. Also, while Metacritic claims its score is curated with a preference of individual high-quality reviewers, it is entirely possible the weighted score has a small but non-negligible component that is driven by user opinions on large marketplaces such as Steam. Finally, Steam offers highly detailed information about the game's intrinsic attributes such as its genre(s), developers, themes, recommended age limits, price, and language support, amongst others, which can intrinsically capture the latent quality of a game. This latent game quality is likely to influence Metacritic scores regardless of the selectiveness of the sources of reviews that Metacritic uses in its aggregation.

This investigation therefore aims to validate the hypothesis that it is possible to predict a Metacritic's score, despite its ostensible insulation from Steam, using a combination of the game's intrinsic attributes and its community behavior and perception from Steam.

III. Dataset

A. Overview

The dataset used in this investigation is FronkonGames' Steam Games Dataset [7] available on Hugging Face, an Artificial Intelligence (AI) community with a multi-pronged

focus on Natural Language Processing (NLP), state-of-the-art machine learning models, and datasets that power the former and latter. At the time of writing, the dataset consisted of 85,103 observations of video games from the Steam marketplace with release dates ranging from 30th June, 1997 to as recently as 14th April 2024. Each of these observations has 39 predictors, a full list of which can be found in the dataset's summary card on [7]. Broadly, the predictors can be classified into

- Game Attributes: intrinsic attributes of the game such as its release date, price, supported languages, supported platforms, genre(s), Steam tags, developers, marketing descriptions, to name a few.
- Exogenous Attributes: attributes unrelated to the intrinsic game quality and related to the wider community, such as the concurrent player count, the recommended age of players, peak concurrent player count, curated reviews from other platforms.

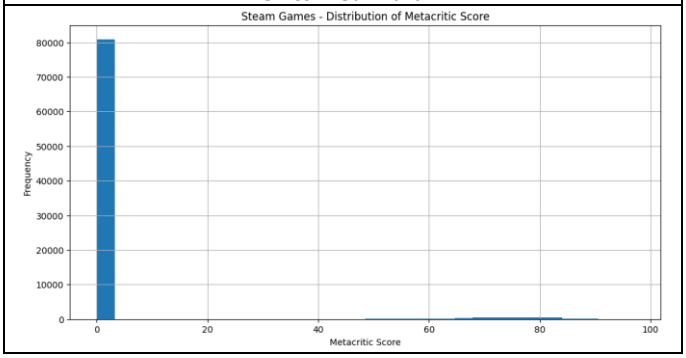
The target or response variable in the dataset is, as posited in II (B), the Metacritic score for the game. As expected, this is an ordinal number ranging from 0 to 100.

B. Preprocessing and Exploratory Analysis

A key assumption underlying this analysis is that the Steam community's volume of users is a critical component in its ability to implicitly or explicitly influence Metacritic scores. As such, the data was filtered for games that were released on or after 1st Jan 2010, based on an (admittedly subjective) assessment of an inflection point in Steam's annual user count [8]. This filtering operation did not have a marked effect on the number of observations, with only 727 games being removed resulting in a ~0.85% decrease.

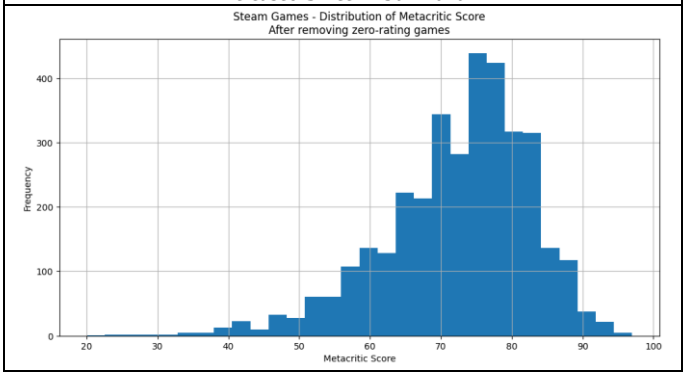
Examining the distribution of Metacritic scores in the resulting dataset showed a highly zero-inflated mixture distribution with more than 95% of the games having a Metacritic score of 0, as shown in Figure 1.

Figure 1: Histogram of Metacritic Scores of Steam Games Released Since 1st Jan 2010



Such a skewed distribution could be modeled with systemic oversampling for deterministic Machine Learning methods as well as zero-inflated likelihoods [9] [10] for Bayesian models. However, it was assumed that the excessive number of zeroes in the Metacritic score was not a true reflection of game quality and more likely to be a data artifact: many of these games may not have had Metacritic scores at all. Given the motivation of the investigation in terms of quantifying the predictive power of a game's Steam attributes with regards to its Metacritic score, it made sense to focus exclusively on games which were known to have a Metacritic score in the first place. This, coupled with the fact that modeling on 89,000 zero-inflated scores would have most likely lead to problems with computational and modeling tractability, motivated a filtration of all observations where the Metacritic score was zero. This operation resulted in a significantly smaller dataset of 3,943 observations whereby the Metacritic scores that resembled a left skewed multi-modal normal.

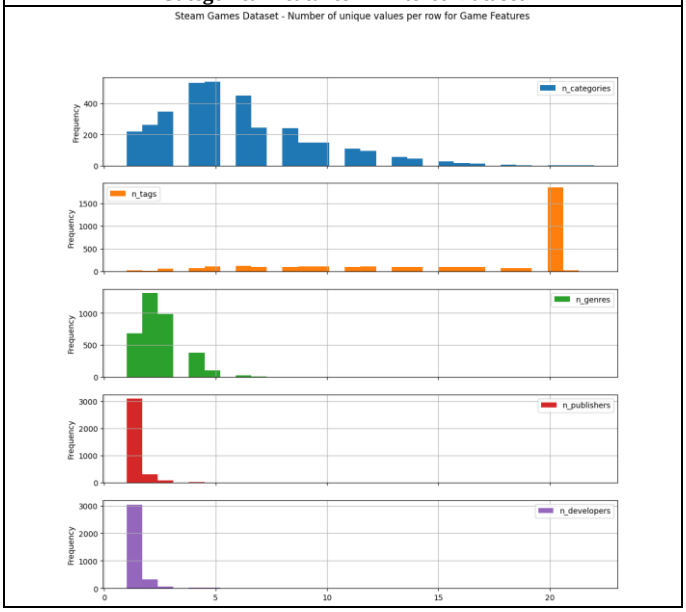
Figure 2: Histogram of Non-zero Metacritic Scores for Steam Games Released Since 1st Jan 2010



C. Feature Engineering – Structured Predictors

Extensive feature engineering has been performed on the dataset to make it as informative as possible for predicting the Metacritic score. Firstly, the game's release year was extracted from its release date attribute. Several predictors, such as `genre`, `supported languages`, `supported audio languages`, `developers`, and `publishers` were comma-separated strings of individual categories. The cardinality of the unique categories per observation for these features was assessed as shown in Figure 3 to assess the utility, if any, of extracting more than one unique category per predictor per observation,

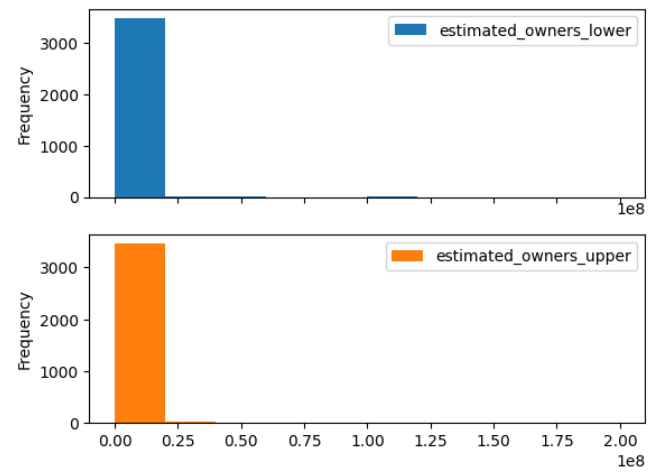
Figure 3: Number of Unique Categories per Observation for Categorical Features in Filtered Dataset



Observation-level cardinality of `developers` and `publishers` showed that most observations had only one

unique category for each of these features, which meant extracting the first category for these predictors would not have led to significant loss of information. While there did seem to be utility in extracting up to the first 2 values of `genre` per observation and the first 3 categories per observation for `categories`, analysis showed that the first 2 genres and first 3 categories were highly correlated: "Action" and "Adventure" were common 2-grams in the `genre` attribute and "Multi-player", "Competitive" were found to be common 2-grams in the `categories` predictor. This, coupled with anticipation of computational and modeling intractability due to the curse of dimensionality [11] with one-hot encoded versions of high cardinality categorical features, motivated the decision to retain only the first category of `genre` for this analysis. The `tags` attribute was discarded as well for similar reasons.

Figure 4: Distribution of estimated owners' bounds



Substrings for the lower and upper bounds of the `estimated owners` attribute were extracted using similar delimiter-based rules. However, they showed a highly skewed zero-inflated distribution with little variance amongst the video games, as shown in in Figure 4. As such, the feature was discarded from future processing.

D. Feature Engineering – Unstructured Text

1) Text Embeddings

"about the game" and "reviews" predictors in the data consisted of variable-length strings of each game's description and collated reviews/praise from official video game review publications respectively. The game description was hypothesized to have meaningful information about the intrinsic qualities of the game such as its gameplay mechanics, the kind of user base it appealed to, unique selling points and technical requirements, to name a few. Likewise, the "reviews" text was expected to provide a proxy for general sentiment around the game that shapes user behavior both prior to and after purchasing the game (setting expectations, post-purchase cognitive rationalization of findings from the review [12]).

As such, both features were considered important in capturing latent qualities of the game and the wider community's perception around it. Consequently, both the reviews and the game description text were transformed into numeric representations of through a two-part feature engineering and transformation process

- Generating 768-dimensional text embeddings using a pre-trained lightweight transformer model, DistilBERT [13]
- Projecting DistilBERT embeddings into lower dimensional space with Principal Components Analysis (PCA) while retaining projecting that account for 90% explained variance.

DistilBERT was used as a quick, fast, and memory-efficient implementation of a state-of-the-art transformer model trained on a relatively large corpus of text that would help extract meaningful semantic information from relatively large pieces of text.

Table 1: Word Count Quantiles for Reviews and the Game Description

	2.5%	5%	25%	50%	75%	97.5%	99%
Reviews	1	1	1	29	64	139	171
About the Game	1	81	100	173	344	555	645

Based on table 1, which shows the quantiles for word-level token counts across all reviews and descriptions, reviews and descriptions tokens up to 120 and 500 words respectively were tokenized using DistilBERT's case-agnostic off-the-shelf tokenizer. Any words greater than the maximum word limit were truncated in the interest of maintaining computational tractability, and any words below the maximum word limit were padded to transform variable length text sequences into a consistently dimensioned array of numerical tokens. The imposition of these token limits ensured that up to 95% of all observations in the dataset were not truncated, preserving as much semantic and contextual information in the fields as possible.

Three new features were created from each 768-dimensional DistilBERT embedding at an observation level

- Mean embedding value
- Max embedding value
- Cumulative embedding value

These features were attempts at distilling the embedding's information to point estimates or single features. While pooling operations on embedding are usually performed through a Pooling neural network layer [14] or similar construct, such an engineering exercise was foregone in favor of a more simpler feature engineering technique and an implicit requirement to expend more energy on modeling than on feature engineering.

2) Principal Components Analysis (PCA)

Table 2: PCA Component Count by Explained Variance Threshold

	90% PCA	95% PCA
Reviews	61	250
About the Game	67	271

Experiments with PCA on the embeddings showed that 90% explained variance threshold resulted in 60 – 70 principal components for each field, whereas increasing the threshold by 5% resulted in an almost 300% increase in projection dimensionality. This suggested that increasing an explained variance threshold beyond 95% may have had diminishing

returns in terms of additional information per unit of additional dimensionality, which is why a 90% threshold was used for projections.

E. Sub-Datasets and Feature Transformation

To explore the robustness of modeling techniques to progressively higher dimensionality and to quantify the incremental benefit of the features engineering by summarizing and projecting the embeddings, three different datasets were created as shown in Table 3.

Table 3: Sub-Datasets and Their Predictors

Dataset Name	Predictors	Number of Predictors	Training Set Size	Test Set Size
D1	release_year, n_supported_languages, n_audio_languages, n_genres, n_tags, n_developers, n_publishers, n_categories, price, achievements, dlc_count, peak_ccu, required age, average playtime forever, average playtime two weeks, median playtime forever, median playtime two weeks, genre_primary (one-hot encoded)	32	2,794	699
D2	All D1 features and mean, max, sum of embedding for reviews and description	38		
D3	All D1 features and PCA projections for reviews and descriptions	160		

Each of the datasets was split into a training-validation and test set with an 80-20 ratio resulting in 2,794 observations used for training and tuning models and 699 used for a final out-of-sample evaluation.

All features, except `genre_primary`, were numeric and were transformed to a standard normal distribution with a mean of 0 and standard deviation of 1 by the equation below

$$\hat{x} = \frac{x - \bar{x}}{\sigma_x}$$

Where

- $\bar{x}$ = Mean from training set
- σx = Standard deviation from training set
- $\hat{x}$ = Transformed predictor

`genre_primary`, the only categorical feature, was converted to a numeric indicator variable through one-hot encoding.

Lastly, a dummy predictor of 1 was added to each dataset to explicitly model an "intercept" coefficient or intercept term.

IV. Bayesian Models

A. Model Overview – Strengths and Weaknesses

Table 4 summarizes three Bayesian models, each with a different set of likelihoods and priors, that have been proposed for modeling the Metacritic scores in the training data.

Table 4: Summary of Proposed Bayesian Models

Model	Likelihood	Strengths	Limitations
M1	Truncated Normal	Independent mean, shared variance Flexibility in modeling skew and kurtosis	Continuous distribution but scores are discrete.
M2	Poisson	Discrete distribution Fewer parameters	Used for modeling counts, not scores Assumes data is not over-dispersed
M3	Negative Binomial	Discrete distribution Flexibility to model over-dispersed data	Used for modeling counts, not scores Additional dispersion parameter to model

These models are similar in that they use a linear combination of the predictors in the training set weighted by regression coefficients β to estimate the expected score at a game level. They differ, however, in the likelihoods used and the assumptions said likelihoods make about the underlying data generating process.

These assumptions are the primary drivers of each model's hypothesized strength and weaknesses in modeling the data. For instance, M1 can model game-level scores without any significant constraints on the variance of the scores. This is contrast with M2, which assumes the mean and variance of the scores are identical given a Poisson likelihood. M1 can also approximate skew and kurtosis effects within the likelihood which M2 and M3 cannot. However, this flexibility comes at the cost of a discrepancy with regards to the generating process: while Metacritic scores are discrete, M1 assumes they are continuous. As such, every single posterior prediction from such a model will intrinsically have a non-zero error relative to a discretized ground truth.

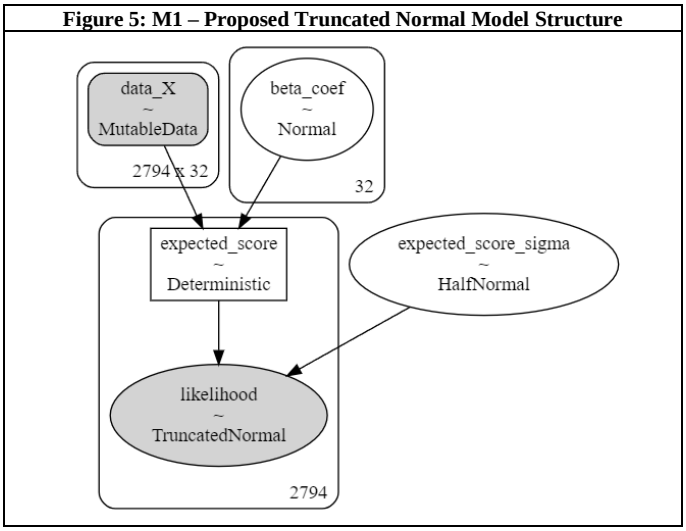
M2 and M3 address this shortcoming by virtue of being discrete distributions. However, they are also, to an extent, misaligned with the nature of the data generating process in the sense that these distributions are generally not used to model scores. Both distributions are used to model count data [15] albeit with different levels of dispersion. It is posited that these likelihoods may still be moderately useful for modeling scores under the assumption that a Metacritic score represents the "count" of perfect scores (100) received over a given number of reviewers, and in doing so are not fundamentally incompatible with the data generating process.

B. M1 – Truncated Normal Model

This model assumes the score for each game can be modeled as a truncated normal distribution with a mean centered around a linear combination of its predictors X

weighted by coefficients shared across all games β. It also assumes the game-level normal likelihood has a pooled or shared standard deviation σ.

Figure 5: M1 – Proposed Truncated Normal Model Structure



$$y_i \sim Truncated\ Normal(\mu_i, \sigma)$$
$$\mu_i = \beta X_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \ldots + \beta_j x_{ij}\ \forall j \in [0, p]\ \forall i \in [1, n]$$
$$\beta_j \sim Normal(\mu = 0, \sigma = 10)\ \forall j \in [0, p]$$
$$\sigma \sim Half\ Normal(\sigma = 10)$$

1) Truncation and Constraints

A truncated normal distribution was necessary to align with the data generating process whereby scores were strictly positive. M2 and M3, by virtue of being distributions generally used for modeling counts, achieve this by design. Not imposing a similar truncation constraint on the lower bound for M1 would therefore not have allowed for an apples-to-apples comparison between the models. No constraint was placed on the upper bound of the likelihood in M1 (or indeed M2 and M3). This was primarily to test the robustness of the models to the scale of the data: the ability posterior predictive scores being ≤ 100 was posited to be a good litmus test of model fit and one of many criteria for model comparison used in section VII.

2) Regularization through Coefficient Priors

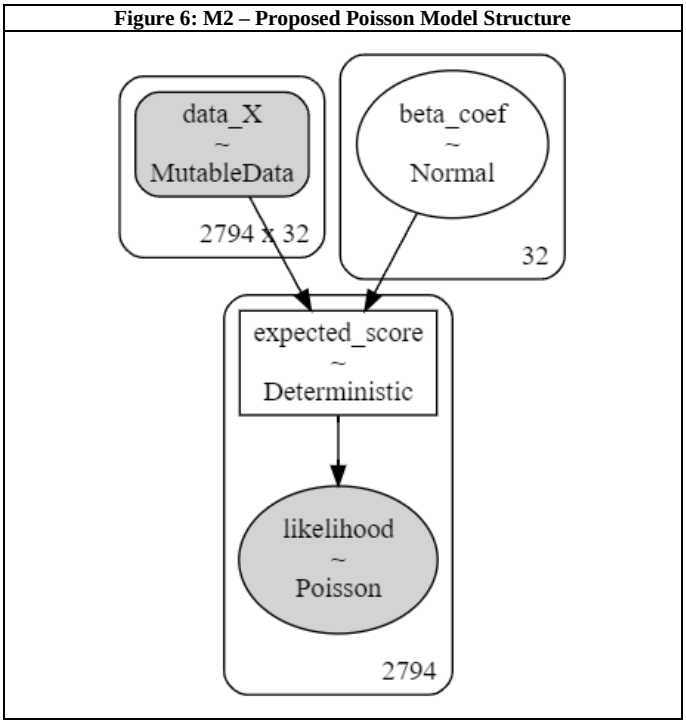
Imposing a normal prior with mean 0 and standard deviation of 10 for the coefficients β was an intentional decision to implement a basic version of L1 regularization [16]: in the absence of evidence to the contrary, the coefficients of redundant features were expected to "shrink" to 0. At the same time, a standard deviation of 10 allowed the model sufficient flexibility to increase or decrease the relative importance of a predictor in estimating the game's expected score while still constraining coefficients to be relatively small, hence ensuring a highly regularized model. The same priors have been used for β across all three models to ensure that each model had approximately the same regularization constraints from a prior perspective.

3) Pooled Standard Deviation

A half-normal standard deviation prior was proposed for the standard deviation in Metacritic scores across all games. Game-level standard deviations were not estimated primarily in the interest of computational tractability. Furthermore, it was assumed that a shared or pooled standard deviation across games would help result in more robust and generalizable estimates of score variability that would not be susceptible to spurious fluctuations in the scores for individual games. The choice of half-normal prior with deviation of 10 was motivated by the lower and upper bounds of the Metacritic score distribution: given minimum and maximum training set scores of approximately 25 and 83 respectively, it was assumed a standard deviation of 10 would adequately account for the correct theoretical range of [0, 100] for game scores based on the 68-95-99.7 rule.

C. M2 – Poisson Model

Figure 6: M2 – Proposed Poisson Model Structure



$$y_i \sim Poisson(\lambda_i)$$
$$\lambda_i = \exp(\mu_i)$$
$$\mu_i = \beta X_i = \beta X_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \ldots + \beta_j x_{ij}\ \forall j \in [0, p]\ \forall i \in [1, n]$$
$$\beta_j \sim Normal(\mu = 0, \sigma = 10)\ \forall j \in [0, p]$$

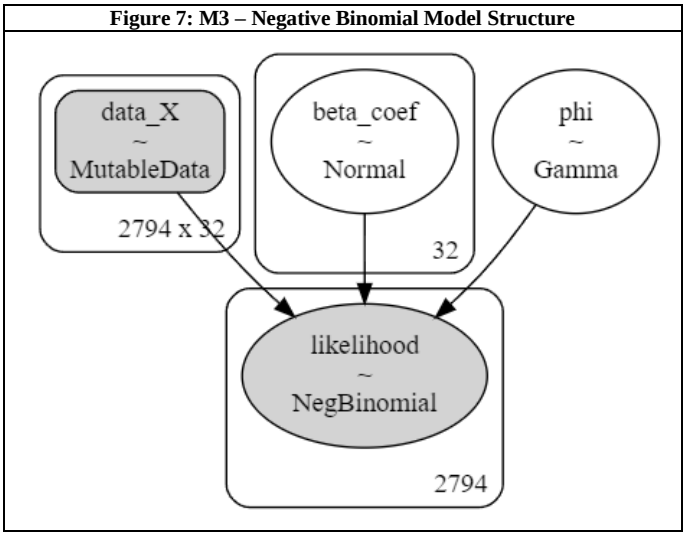
This is a relatively simple regression model which, as discussed in IV(A), tries to model the Metacritic score as a discrete count of perfect scores arriving in a review aggregation process over a fixed number of reviews. The mean number of perfect scores, as a proxy for the Metacritic score, is parameterized by λ, which in turn depends on the predictors X and coefficients β through an exponential link function. M2 serves as a good contrast to M1 in two aspects: firstly in its treatment of the outcome as a discrete, instead of continuous, variable and secondly in its lack of reliance on a shared variance or standard deviation parameter.

D. M3 – Negative Binomial Model

M3 is similar to M2 in that it uses a discrete distribution for the scores under the similar assumptions about the Metacritic score being the count of perfect scores across multiple reviews. However, it improves on M2 by removing

a constraint on the expected score and standard deviation in scores being identical.

Figure 7: M3 – Negative Binomial Model Structure



$$y_i \sim Negative\ Binomial(\mu_i, \phi)$$
$$\mu_i = \exp(\nu_i)$$
$$\nu_i = \beta X_i = \beta X_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \ldots + \beta_j x_{ij}\ \forall j \in [0, p]\ \forall i \in [1, n]$$
$$\beta_j \sim Normal(\mu = 0, \sigma = 10)\ \forall j \in [0, p]$$
$$\phi \sim \Gamma(\alpha = 2, \beta = 2)$$

The expected score, as discussed in IV(A), is still a weighted linear combination of predictors. However, the variance of the expected discrete scores has a Gamma prior with reasonable defaults for exploration. In a sense, the negative binomial model is a hybrid between the truncated normal model M1 in its ability to model shared score variance and the Poisson model M3 in its ability to model discrete scores.

V. Deterministic Models

A. Models Chosen

As a baseline to compare and evaluate Bayesian model performance, two deterministic models were trained, tuned, and evaluated separately on each dataset from Table 3

- M4: Ordinary Least Squares Ridge Regression
- M5: Decision Tree Regressor

M4 was chosen because it presented as a simple, tried-and-tested implementation of linear regression albeit with a slightly different regularization compared to the regression implemented in M1 – M3. While M1 – M3 try to mimic an elementary form of L1 or Lasso regularization to encourage sparsity in the regression coefficients, M4 implements L2-norm based regularization. This was an intentional design choice to ensure that M1 – M3 were functionally different from M4 in more than just their fitting mechanisms.

M5 was chosen to explore the effects and strengths of non-linear relationships between predictors and responses in the sub-datasets. Unlike M4, which imposes a strong preference bias for linear hypotheses functions, M5 is able to recursively partition the feature space into logically consistent regions based on the predictors, optimizing splits to minimize variance in the Metacritic score within regions or maximizing information gain. It was hypothesized that the relationship between Steam-based predictors and Metacritic scores was highly non-linear and. As such, simply contrasting non-linear Bayesian regression models M1 – M3 with a linear deterministic model M4 would not have offered a fair contrast between deterministic and Bayesian regression methods.

B. Hyperparameter Tuning

Table 5: Hyperparameter Grid for Deterministic Models

Model	Parameter	Values Searched
M4 Ridge Regression	Alpha	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0
M5 Decision Tree	ccp_alpha	0, 0.001, 0.003, 0.01, 0.03, 0.1,0.3
	criterion	Squared error, absolute error, friedman MSE
	max_depth	2, 3, 5, 10

Table 5 summarizes the grid of hyperparameters that was searched over using 5-fold cross-validation with R2 score as the objective function.

R2 score was chosen because as the optimization metric / objective function because, in addition to being a scale invariant and interpretable metric of model fit, it directly aligns with the research question motivating the experiment i.e. the proportion of variance in the target (Metacritic score) that can be explained by a model based on Steam predictors.

The alpha parameter, which represents the regularization coefficient, was the only parameter tuned for the Ridge Regression model to explore the extent to which scaling penalty for squared coefficient magnitudes helped achieve better model fits. Likewise, ccp_alpha and max_depth, which control pruning and tree depth of M5 respectively, were tuned to find the simplest possible tree that achieved a good bias-variance tradeoff.

VI. Methodology

A. Bayesian Models

Each model (M1 – M3) was fit to each sub-dataset (D1 – D3)'s training-validation set of 2,794 observations with 2 chains, each with 2000 burn-in or wamups and 2000 samples from the typical set. Random number generators were seeded with the author's GTID, 93956081, whenever possible to ensure reproducibility of results. Log likelihoods were computed after model fitting for each model to quantify goodness of fit in terms of Widely Acceptable Information Criterion (WAIC) [17] and Pareto-smoothed importance sampling leave-one-out cross-validation (LOO) [17], both on the deviance scale. Posterior predictive distributions were generated and visualized for training / in-sample data across all models to visually inspect goodness of fit. This is supplemented by visual inspections of the distribution of r_hat and bulk effective sample size (ess_bulk) statistics across all models and datasets. The distribution of samples with Pareto-K statistics in each of (-inf, 0.5], (0.5, 0.7], (0.7, 1], and (1, inf) is reported as well followed by a comparison of fit times.

Model fit is also evaluated in terms of R2 score, mean absolute error (MAE) and mean squared error (MSE) on both in-sample and out-of-sample data using Bayes estimators for parameters derived from posterior predictive distribution. For M1, these estimates are generated exclusively using Arviz machinery. For M2 and M3, artifacts of poor sampling resulted in exploding λ and erroneous p estimates respectively in the case of posterior predictive outsample generation with Arviz and PyMC machinery. For these models, posterior estimates of regression coefficients are extracted manually and used in conjunction with a NumPy based recreation of the statistical model to generate posterior predictive distributions.

B. Deterministic Models

Models defined in V(A) are tuned as described in V(B) with the same seed as described in VI(A). Fit models were then evaluated in terms of R2 score, MAE, and MSE on both in-sample and out-of-sample data, just as was the case in V(A).

VII. Results And Discussion

A. Bayesian Model Comparison

1) Fit Times

Figure 8: Fitting Times for Bayesian Models Across Sub-Datasets

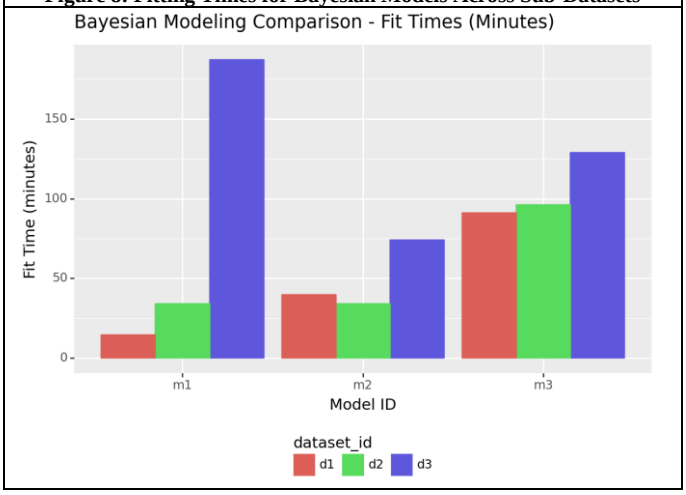


Figure 8 shows the time taken by each Bayesian model (M1 – M3) to fit to each sub-dataset (D1 – D3). In general, fit times increased from D1 to D2 to D3 respectively, owing to the increase in dimensionality from 32 to 36 to 160 with the inclusion of PCA projections of review and description embeddings. Furthermore, the negative binomial model (M3) generally took the longest time to fit across all three datasets. This is most likely a consequence of a significantly more complicated likelihood function which involves estimation of not just the regression coefficients and their corresponding μ or location parameter, but also a dispersion parameter. It is also possible that the longer fit times for the model are a consequence of the $\Gamma(\alpha=2,\beta=2)$ prior on the dispersion parameter not being sufficiently informative, requiring the model to explore a substantially larger subspace of the posterior before converging to the typical set. The Poisson regression model (M2) takes slightly longer to fit than the Truncated Normal model (M1) despite having a likelihood function more closely aligned with the data generating process in terms of outcome cardinality. This is most likely due to the model's assumptions about expected value and expected variance in the likelihood being ill-conditioned: the mean score of the training set was ~67 but the standard deviation was closer to 10. M1 generally fit much faster than M2 and M3 on D1 and D2, but took significantly longer on D3. This suggests the model did not scale as well at higher dimensionality as M2 or M3. It is also possible that the priors of Normal(0, 10) defined for coefficient estimates were good choices for the predictors in D1 and D2, but failed to be informative for joint distribution created by the inclusion of PCA projections of the text embeddings. In general, however, time complexity analysis suggests either the Truncated Normal (M1) or Poisson model (M2) may be more tractable than the negative binomial model (M3).

2) WAIC

Figure 9: WAIC for Bayesian Models Across Sub-Datasets

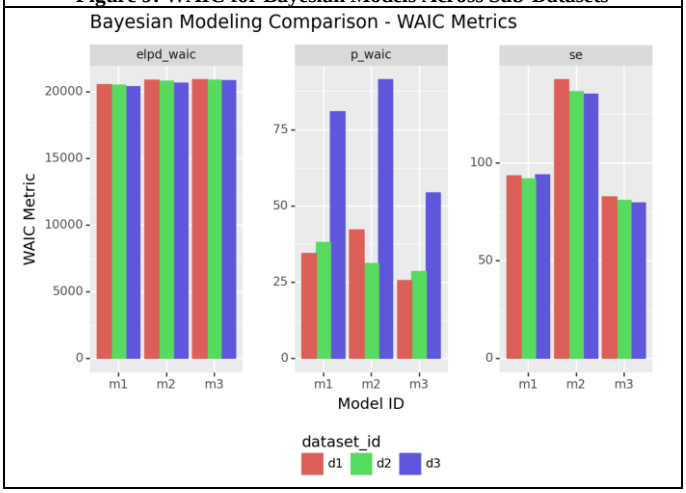


Figure 9 shows the deviance scale WAIC-based expected log predictive density (ELPD) as well as the associated number of effective parameters and ELPD standard error across datasets and Bayesian models. Despite drastic differences in fit times as shown in Figure 8, the models show only marginal differences in ELPD, with M1 generally being incrementally better than M2 which in turn is only slightly better than M3. Model fit generally does improve between D1 and D2 as well as D2 to D3. Concurrently, the effective number of parameter (p_waic) does increase between datasets D1, D2, and D3 which suggests that all models did, to an extent, find the inclusion of point estimates of embeddings as well as PCA projections of embeddings to be somewhat informative in predicting the observed data. The only anomalous trend is for M2, which seems to suggest that the inclusion of embedding point estimates (D2) decreases the effective number of parameters relative to (D1). A decrease in effective number of parameters is generally indicative of a decrease in model complexity and propensity to overfit. This suggests that at least in the case of M2, the inclusion of point estimates of embeddings helped eliminate predictors that were previously resulting in the overfitting. While the raw standard errors in ELPD estimates (se) do look large, they

represent only 0.4 – 0.6% of the average ELPD, suggesting that all models are relatively robust to small changes in the data and the models' ability to predict on new data points is not necessarily dependent on minor differences in data distributions between training and inference data.

3) LOO

Estimates of ELPD, effective parameters and standard error from LOO are very similar to WAIC. LOO investigation did, however, provide an additional summary of the ranges of estimated Pareto-K diagnostic values for individual training set observations, which have been summarized in Table 6.

Table 6: Pareto-K Diagnostic Distribution

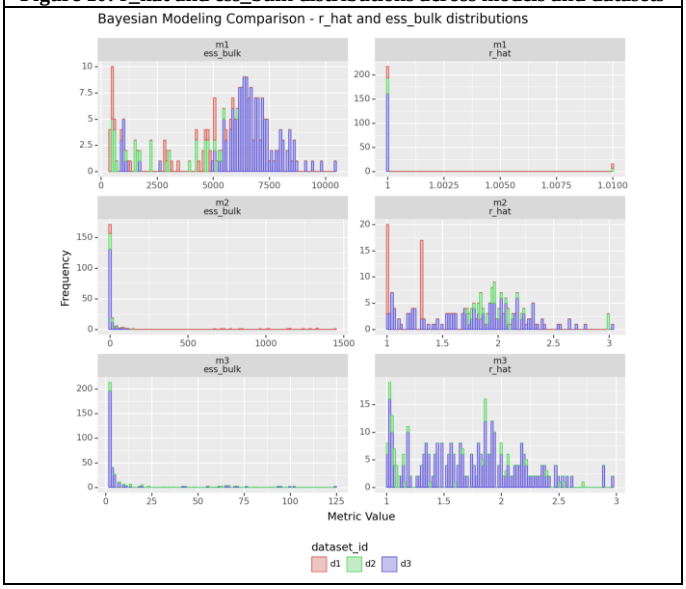
Model	Dataset	(-inf, 0.5]	(0.5, 0.7]	(0.7, 1]	(1, inf)
M1	D1	2785	5	4	0
	D2	2784	5	5	0
	D3	2787	6	1	0
M2	D1	2783	5	5	1
	D2	2752	23	17	2
	D3	2784	6	4	0
M3	D1	2784	5	4	1
	D2	2787	4	2	1
	D3	2786	3	5	0

Comparing these results to diagnostic interpretation for another probabilistic modeling package, STAN [18] indicates that across all models and datasets, there are only a handful of observations with moderate to high k-values. This suggests that MCMC estimates of ELPD, SE, and effective parameters derived through LOO, which are highly correlated with those derived with WAIC, are generally reliable in their assessments of model fit. The presence of 40 observations with k-values > 0.5 for M2 fit to D2 suggests there is an incrementally larger subset of highly influential observations compared to other models. This could be due to a combination of model misspecification and outliers, which aligns with the anomalous trend in effective parameter count estimate for M2 on D2 shown in Figure 9. In general, however, these diagnostics give little reason to doubt the WAIC and LOO-based comparisons of model fit quality. While all models are similar in their fit quality across datasets, the truncated normal likelihood model is incrementally better than the rest.

4) Trace Diagnostics

Figure 10 compares the distributions of the Gelman-Rubin statistic (r_hat) and bulk effective sample size (ess_bulk) [19] of the three models across the three datasets, derived by filtering the Markov Chain Monte Carlo (MCMC) for posterior samples of the regression coefficients β. These diagnostics suggests that of all three models, only M1's posterior estimates for β can be considered reliable. The r_hat distribution for the model is concentrated at 1.0 across all three datasets, suggesting both MCMC chains of the sampler converged to the same typical set. Likewise, bulk_ess values are generally above 5,000, suggesting that after discarding burn-in or warmup samples and highly correlated samples near the end of the chain, the effective number of samples obtained from the posterior still remains relatively high. This in turn suggests that the posterior samples were informative and had a significant influence on determining the posterior estimates for the regression coefficients indicating efficient sampling.

Figure 10: r_hat and ess_bulk distributions across models and datasets



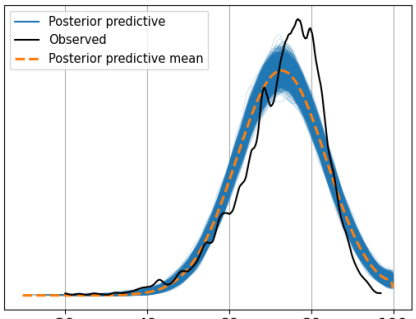
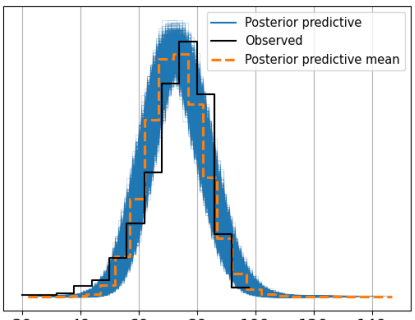
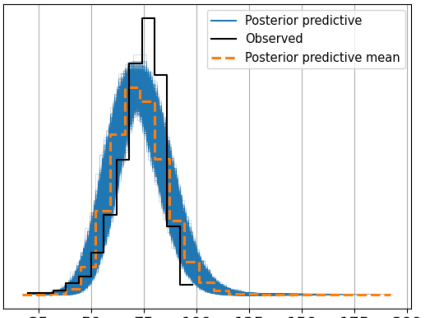
The same cannot be said for M2 and M3, both of which generally have a significant amount of r_hat values associated with regression coefficients well above 1.01 and bulk effective sample sizes closer significantly below half the number of samples from the posterior. These fit issues get progressively worse from D1 to D2 to D3 across M2 and M3, validating prior concerns about the curse of dimensionality. R_hat values for M2 and M3 suggest that the two chains did not converge to the same subspace of the posterior distribution for a substantial portion of parameters, indicating a lack of convergence. Likewise, the low ess_bulk indicates that even though up to 8,000 samples were collected from the posterior for each model, the samples were highly correlated and not providing as much information could be expected given the opportunity to sample the posterior. This in turn means the information content for estimating posterior quantities is relatively low in M2 and M3. The lack of convergence and evidence of inefficient posterior sampling in M2 and M3 could be a consequence of the more complex posterior geometry, especially in the case of the M3 with a negative binomial likelihood model. It could also indicate that priors defined for regression coefficients, while relatively informative and helpful in directing MCMC sampler to the stationary distribution or typical set in the case of M1, were not sufficiently informative for M2 or M3. This is most likely why the model was unable to explore the posterior distribution for M2 and M3 well enough. Given no divergences were observed during sampling for any models, it is reasonable to assume the fit issues in M2 and M3 are driven by model parameterization and prior choices as opposed to problems intrinsic to the sampler such as exploding or vanishing gradients.

5) In-sample Posterior Predictive Distributions

Figure 11 compares posterior predictive distributions on training / in-sample data for all three models on dataset D1. Posterior predictive distributions were found to be very

similar across datasets for the same model, which is why the distribution on D1 has been presented for analysis. Please see Appendix A for a full set of posterior predictive distributions.

Figure 11: Posterior Predictive Distributions – D1

Model 1 – Truncated Normal	
Model 2 - Poisson	
Model 3 – Negative Binomial	

Posterior predictive plots indicate that M1 has generally done the best job of fitting to the data. Even though it did not have an upper bound of 100 imposed on it as part of the model likelihood constraint, it has intrinsically learned that an acceptable upper bound for Metacritic scores is indeed 100, suggesting it is doing a good job of modeling the underlying or latent characteristics of the data generating process. The observed data demonstrates a bimodality with a peak at 67 and another at 73. Under the constraint of remaining symmetrical about a single model, M1 has failed to capture the bimodal nature of the data perfectly and has instead settled on an "intermediate" mean and mode between the two "peaks" in the observed data. Enforcing a lower constraint of 0 instead of 20 has allowed the model to learn coefficients that achieve scores of 0, even though no such observations are found in the observed data. This is not necessarily a problem: it indicates, once again, that the regression coefficients learnt have the capacity to generalize scores to ranges outside of the training set.

M2 and M3 demonstrate a discrete posterior likelihood, as can be expected. M3 has a much higher range in the posterior predictive distribution than M2, although both exceed the maximum theoretical score of 100, suggesting models M2 and M3 were not parameterized to learn the upper bound for Metacritic scores as properly as M1. Forcing M3 to have a different variance or dispersion compared to M2 has not helped the model: the resulting posterior predictive distribution is more diffuse than M2's and does not align as well as M2 does with the observed data, especially near the mode. Based on these posterior predictive distributions, M1 has clearly outperformed M2 and M3 despite its treatment of an inherently discrete response as a continuous one. The fit issues identified in the previous section have clearly affected M3 more than M2, although neither has benefited from its likelihoods' intrinsic alignment with the ordinal nature of the scores.

B. Bayesian vs Deterministic Model Comparisons

Figure 12: Model Comparison – Bayesian vs Deterministic Models

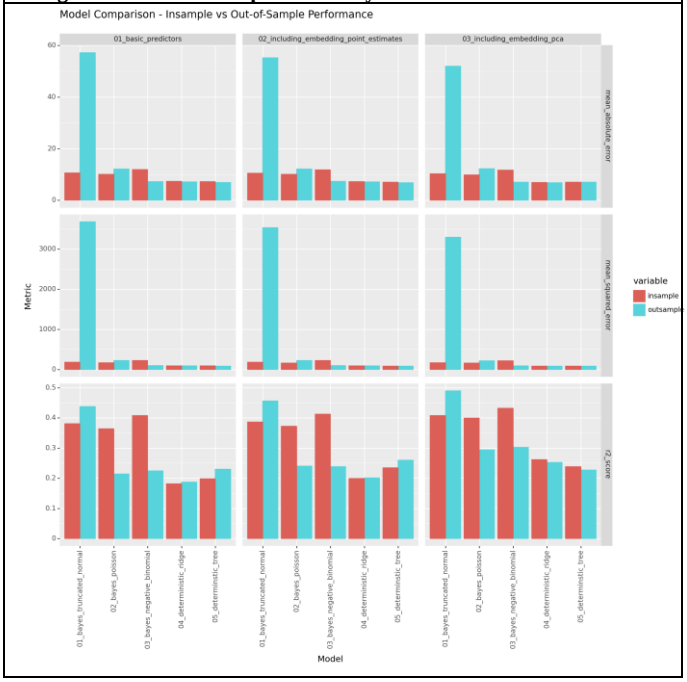


Figure 12 compares the in-sample and out-of-sample performance on R2 score, MAE, and MSE of all 5 models explored in this investigation across all three datasets. The first key finding from these results is that the truncated normal regression model (M1) outperforms all other models, including deterministic models M4 and M5, in terms of R2 score on both in-sample and out-of-sample data. While M2 and M3 outperform deterministic models M4 and M5 in terms of in-sample R2 score, they fail to generalize well and have a significantly lower out-of-sample R2 score that is, at best, comparable to those of M4 and M5. Out-of-sample R2 score for M3 being better than its in-sample counterpart is likely indicative of the model generalizing better to out-of-sample data that is devoid of outliers as opposed to in-sample data that may have outliers. Extremely high MAE and MSE for this model is counterintuitive and anomalous

given previous diagnostics, and is likely to be influenced by one or more anomalous data points which other models are able to predict more reliably by virtue of overfitting to similar observations in the training set. MAE and MSE for the other Bayesian models are comparable to those of the deterministic models, once again showing that while these models do overfit to noise and suffer from convergence and sampling issues, they are performant relative to a conventional supervised learning baseline.

VIII. Conclusions

This investigation shows that it is indeed possible to predict the Metacritic score of video games using attributes of the video games and its community from Steam. It also demonstrates that proposed Bayesian models compensate for their significantly higher fit times than conventional supervised machine learning algorithms in terms of better or comparable performance in terms of multiple fit metrics. This investigation also shows that despite an incongruity in its assumptions about the response variable as continuous score instead of a discrete ordinal one, the truncated normal model generally performs the best amongst the Bayesian models and is properly parameterized. While it does show evidence of scaling poorly with a significantly higher number of predictors in case of dataset 3, this does not affect its ability to generalize to out-of-sample data. In contrast, the Poisson and Negative Binomial likelihood do a good job in terms of MAE, MSE and R2 score, posterior estimates of their regression coefficients are significantly more unreliable and do not seem to have learnt from the posterior well.

IX. Future Work

The high out-of-sample MAE and MSE for the Truncated Normal model is indicative of outliers, which should ideally be removed before assessing model fit. Investigating these outliers and establishing their root cause as a data collection error or genuine artifact of the Metacritic scores dynamic could be useful to evaluate improvements for the model. Given the bimodal nature of the filtered Metacritic score distribution, it may also be useful to experiment with mixture models consisting of Truncated Normal models for individual categories or clusters of games.

The Poisson and Negative Binomial Likelihood models could be improved by tuning priors for regression coefficients and, in case of the Negative Binomial model, the dispersion parameter. This can be done through prior predictive checks or by eliciting priors that are more closely aligned with theoretical properties of the expected likelihood. If prior tuning does not significantly improve model performance and fit issues, it may also be worthwhile to force an upper bound on the likelihood through a truncated distribution. While initial attempts were made at applying all of these ideas to improve models, time and compute constraints necessitate delegating them as potential areas for future work.

In order to further quantify the improvement of Bayesian models over deterministic ones, Bayesian models could be used to generate posterior estimates of latent "game" and "community" attributes based on relevant subsets of features. Posterior predictive inference can be used to generate these latent attributes as new features for data, which can then be used as features for faster and more interpretable deterministic regression models. In the same vein of feature engineering, the utility of the text embeddings could be improved by fine-tuning the DistilBERT embedding with the Metacritic scores as target, although this is likely to require significantly more data and compute.

X. References

[1] J. Clement, "Video game industry - Statistics & Facts," *Statista*, Jan. 10AD. https://www.statista.com/topics/868/video-games/#topicOverview (accessed Apr. 20, 2024).

[2] "GVR Report cover Gaming PC Market Size, Share & Trends Analysis Report By Product Category (Desktop, Laptop), By End-user (Professional Gamers, Casual Gamers), By Price Range (Low-, Mid-range), By Distribution Channel, And Segment Forecasts, 2023 - 2030," *Grand View Research*. https://www.grandviewresearch.com/industry-analysis/gaming-pc-market-report (accessed Apr. 20, 2024).

[3] "Epic Games Store | Download & Play PC Games, Mods, DLC & More – Epic Games," *Epic Games Store*. https://store.epicgames.com/en-US/ (accessed Apr. 20, 2024).

[4] E. Arts, "Download the EA app – Powering next generation of PC gaming - Electronic Arts," *Electronic Arts Inc.*, Apr. 22, 2022. https://www.ea.com/ea-app (accessed Apr. 20, 2024).

[5] R. Yuen, "The Rise of Steam: A Case Study on the Most Dominant Force in Gaming," *www.linkedin.com*, Nov. 12, 2021. https://www.linkedin.com/pulse/rise-steam-case-study-most-dominant-force-gaming-ryan-yuen (accessed Apr. 20, 2024).

[6] "About Us," *www.metacritic.com*. https://www.metacritic.com/about-us/

[7] M. Bustos, "FronkonGames/steam-games-dataset · Datasets at Hugging Face," *huggingface.com*, Jan. 01, 2023. https://huggingface.co/datasets/FronkonGames/steam-games-dataset (accessed Apr. 13, 2024).

[8] J. Clement, "Number of new users on Steam from 2003 to 2017," *Statista*, Aug. 25, 2023. https://www.statista.com/statistics/823944/steam-new-users/ (accessed Apr. 20, 2024).

[9] "pymc.ZeroInflatedPoisson — PyMC dev documentation," *www.pymc.io*. https://www.pymc.io/projects/docs/en/latest/api/distributions/generated/pymc.ZeroInflatedPoisson.html (accessed Apr. 21, 2024).

[10] "pymc.ZeroInflatedNegativeBinomial — PyMC dev documentation," *www.pymc.io*. https://www.pymc.io/projects/docs/en/latest/api/distributions/generated/pymc.ZeroInflatedNegativeBinomial.html (accessed Apr. 21, 2024).

[11] Udacity, "Curse of Dimensionality - Georgia Tech - Machine Learning," *YouTube*. Feb. 23, 2015. [YouTube Video]. Available: https://www.youtube.com/watch?v=QZ0DtNFdDko

[12] "How does game review media affect players? | 5 Answers from Research papers," *SciSpace*. https://typeset.io/questions/how-does-game-review-media-affect-players-isdt09cdp1 (accessed Apr. 20, 2024).

[13] "DistilBERT," *huggingface.co*. https://huggingface.co/docs/transformers/en/model_doc/distilbert (accessed Apr. 14, 2024).

[14] M. Leys, "The Art of Pooling Embeddings 🎨," *Medium*, Jun. 20, 2022. https://blog.ml6.eu/the-art-of-pooling-embeddings-c56575114cf8 (accessed Apr. 13, 2024).

[15] I. Ozsvald and W. Boelrijk, “GLM: Negative Binomial Regression,” *PyMC*, Sep. 01, 2023. https://www.pymc.io/projects/examples/en/latest/generalized_linear_models/GLM-negative-binomial-regression.html (accessed Apr. 15, 2024).

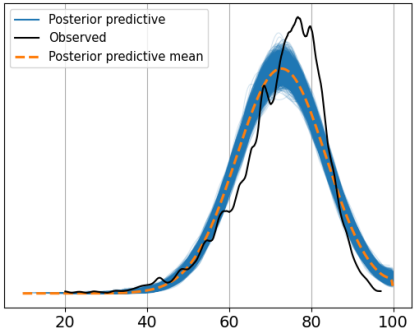
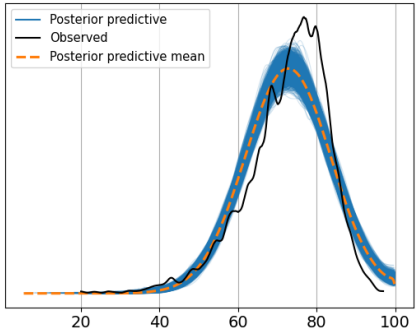
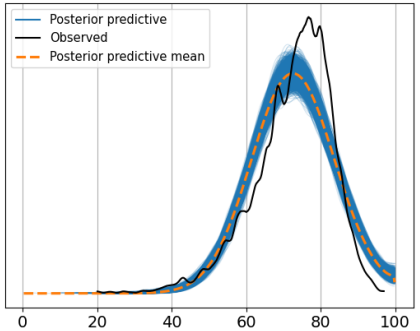
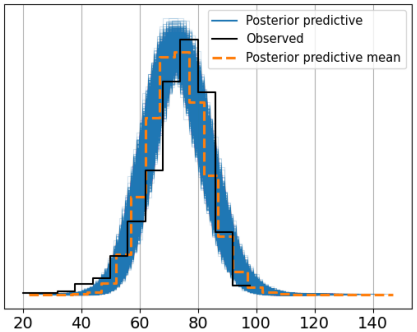
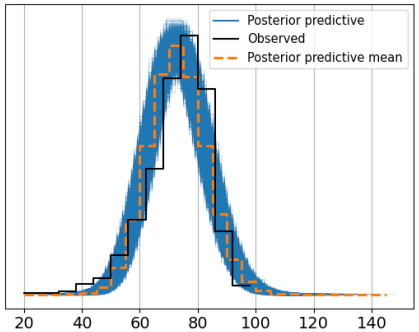
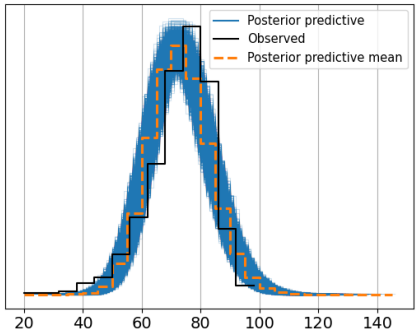
[16] F. Pedregosa, “Linear Models,” *Scikit-Learn - User Guide*. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html (accessed Apr. 15, 2024).

[17] A. Vehtari, A. Gelman, and J. Gabry, “Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC,” *Statistics and Computing*, vol. 27, no. 5, pp. 1413–1432, Aug. 2016, doi: https://doi.org/10.1007/s11222-016-9696-4.

[18] A. Vehtari *et al.*, “LOO package glossary — loo-glossary,” *mc-stan.org*. https://mc-stan.org/loo/reference/loo-glossary.html (accessed Apr. 20, 2024).

[19] J. Guo, J. Gabry, B. Goodrich, A. Johnson, and S. Weber, “Convergence and efficiency diagnostics for Markov Chains — Rhat,” *mc-stan.org*. https://mc-stan.org/rstan/reference/Rhat.html (accessed Apr. 20, 2024).

XI. Appendix A: Posterior Predictive Distributions

	Dataset 1 – Basic Predictors	Dataset 2 –Inc. Embedding Estimates	Dataset 3 – Including Embedding PCA
Model 1 – Truncated Normal			
Model 2 - Poisson			
Model 3 – Negative Binomial