STAT 230A Final Project: Bechdel Test Pass Prediction

Connor McCaffrey

May 26, 2025

1 Introduction

> "I have this rule, see. I only go to a movie if it satisfies three basic requirements: One, it has to have at least two women in it... who, two, talk to each other... about, three, something besides a man." – Alison Bechdel, *Dykes to Watch Out For* (1985)[1]

This rule, now known as the Bechdel test, first appeared in Alison Bechdel's comic strip *Dykes to Watch Out For* in 1985. What began as, in Bechdel's words, "a little lesbian joke in an alternative feminist newspaper" gained traction in the 2000s as a widely discussed benchmark for evaluating women's representation in fiction.[2] Though far from a comprehensive measure of gender parity, the test echoes a longstanding critique first articulated by Virginia Woolf: that women in fiction are too often portrayed as one-dimensional characters, defined primarily by their relationships to men.[3] Despite its simplicity, the Bechdel test provides a useful lens for examining broader patterns in representation.

This investigation will analyze film data from IMDb alongside bechdeltest.com's database to determine which types of movies are more or less likely to pass the test. By examining features such as genre, runtime, release year, and audience ratings, this project aims to illuminate evolving trends in filmmaking as well as broader cultural shifts.

2 Data

The data used in this report comes from the Week 11, 2021 installment of the TidyTuesday project,[4] originally sourced from FiveThirtyEight.[5] The dataset was compiled using the bechdeltest.com and IMDb APIs to download movie metadata and user-voted Bechdel test scores. The two sources were then merged using the "imdb_id" field, though we are limited to Bechdel test scores through 2020 and IMDb film data through 2013—representing 1794 movies in total.

In its raw form, the dataset contains several explanatory variables relevant to this analysis, including release year and date, nominal and inflation-adjusted budget and domestic/international gross, age rating, language, production country, Metascore (0-100), IMDb rating (0-10), genre, runtime (min), and IMDb vote count. It also includes a 0-3 integer score based on the three Bechdel criteria, along with a binary pass/fail indicator, which serves as the response variable for this analysis.

However, several variables required preprocessing before they could be used. For instance, the "genre" variable is stored as a list for each movie. To make this usable, I assigned a "gender score" to each major genre based on historical gender representation, then computed the average score across each movie's genres. Similarly, "language" and "production country" are also stored as lists, but with American and English-language films comprising the majority of the dataset, I instead created two binary indicators: "English" (for movies exclusively in English) and "American" (for movies exclusively produced in the United States).

I also transformed the release date into two new variables: a numeric month and a categorical season, to help capture potential seasonal trends in film releasees. Because the age rating variable included both film and TV categories, I mapped each rating to a unified numeric "maturity" scale from 1 to 5, aiming to capture differences in gender representation based on the intended audience.

Altogether, our preprocessing yielded 17 quantitative and dummy-coded variables to be considered as predictors in a logistic regression model.

3 Final Model

We began with 17 potential predictor variables. To address multicollinearity, we conducted a variance inflation factor (VIF) analysis. This led us to remove the three inflation-adjusted financial variables (budget, domestic gross, and international gross) due to their strong correlation with their nominal counterparts. Eliminating these helped ensure greater model stability and interpretability.

With the remaining predictors, we performed model selection using a genetic algorithm designed to minimize the Akaike information criterion (AIC), a metric that would balance model fit and complexity. This approach allowed us to efficiently search a wide range of possible variable combinations without resorting to exhaustive or stepwise procedures.

The resulting final model, shown in Table 1, had an AIC of 1752.7 and included eight quantitative predictors. Of these, seven had statistically significant coefficient etimates at the $\alpha = 0.05$ level, suggesting a meaningful association with the odds that a film passes the Bechdel test. The one remaining predictor, runtime, while not significant on its own (p-value of 0.071), was retained due to its contribution to overall model fit.

Table 1: Coefficient Estimates and Standard Errors of Optimal AIC Model

	Estimate	Std. Error	z value	$Pr(>\|z\|)$	Signif.
(Intercept)	-4.085e+01	1.636e+01	-2.496	1.255e-02	$*$
Year	1.975e-02	8.139e-03	2.427	1.522e-02	$*$
Budget	-6.643e-09	1.996e-09	-3.328	8.756e-04	$***$
Int. Gross	1.503e-09	4.451e-10	3.377	7.331e-04	$***$
Metascore	1.249e-02	4.999e-03	2.498	1.250e-02	$*$
IMDb Rating	-4.785e-01	1.086e-01	-4.407	1.047e-05	$***$
Genre Score	7.150e+00	9.155e-01	7.810	5.736e-15	$***$
Runtime	6.162e-03	3.409e-03	1.807	7.070e-02	.
IMDb Votes	-1.681e-06	7.372e-07	-2.280	2.264e-02	$*$

3.1 Discussion

The coefficients from the logistic regression model represent the expected change in the log-odds of a movie passing the Bechdel test for a one-unit increase in each predictor. To make these results more interpretable, we can exponentiate the coefficients to obtain odds ratios, which reflect the multiplicative change in the odds of passing.

For example, the coefficient for release year is approximately 0.01975, meaning that with each additional year (and all else held constant), the odds of a movie release passing the Bechdel test increase by a factor of $e^{0.001975} \approx 1.02$—or about a 2% increase in odds per year.

Budget and international gross have much smaller coefficients (on the order of 10^{-9}) due to the large scale of these variables. Interpreted practically: a $1 million increase in budget is associated with a 0.66% decrease in passing odds, but a $1 million increase in international gross is associated with a 0.15% increase in passing odds.

Turning to critical reception, a 10-point increase in Metascore (e.g., from 60 to 70) is associated with a 13.3% increase in the odds of passing. Interestingly, a one-point increase in IMDb rating (e.g., from 6.0 to 7.0 on a 0-10 scale) corresponds to a 38.0% *decrease* in the odds of passing, suggesting contradictory associations between movie reception and female representation depending on the film site.

Unsurprisingly, the average genre score, which I constructed to reflect the historical gender representation of a movie's genres, is positively associated with Bechdel test outcomes. A 0.01 increases in this

score (on a 0-1 scale) leads to a 7.4% increase in the odds of passing.

Runtime also shows a positive association: each additional 15 minutes of runtime is associated with a 9.7% increase in the odds of passing. Conversely, for every additional 10,000 IMDb votes, the odds of passing decrease by about 1.7%.

In summary, release year, international gross, Metascore, genre score, and runtime are associated with better female representation, while higher budget, IMDb rating, and vote count are associated with worse representation.

3.2 Limitations

A key limitation of this analysis lies in the nature of the data source. The dataset is curated by users of bechdeltest.com, meaning a movie is only included if it is manually submitted and approved. This process introduces selection bias, favoring high-profile studio releases, critically acclaimed films, or niche titles of particular interest to the site's relatively small user base. As a result, the dataset may underrepresent the broader film landscape, particularly independent or international films that do not attract attention from contributors. The findings of this investigation should be interpreted with this specific cultural and industrial context in mind.

From a methodological standpoint, we prioritized inference over prediction, which led us to choose logistic regression. This model offers relatively interpretable coefficient estimates and aligns with our goal of understanding which factors are most associated with Bechdel test outcomes. Although interaction terms could have potentially improved model fit, we avoid them to preserve interpretability.

Some results, like the opposing effects of Metascore and IMDb rating, highlight complex dynamics that merit further investigation. The descrepancy may reflect fundamental differences between how professional critics (who contribute to Metascore) and general audiences (who rate on IMDb) appraise movies. In terms of performance, our final logistic model achieves an AUC of 0.707 and a classification accuracy of approximately 67%, notably better than random guessing but still modest. Had predictive accuracy been the primary focus, tree-based ensemble methods such as random forest or gradient boosting may have yielded better results.

Future research in this domain could benefit significantly from emerging AI technologies. While IMDb already provides rich metadata, much of it is unstructured text. With the help of LLMs, it would be feasible to extract features such as inferred gender probabilities of directors, screenwriters, or producers. This could enable deeper exploration of how representation behind the camera relates to representation on screen. Additionally, transforming text data (e.g., film titles, plot summaries, cast lists) into meaningful tabular features could significantly enhance both inference and prediction in future models.

4 Conclusion

This investigation reveals several meaningful patterns in how films meet or fail the Bechdel test criteria. The positive association between release year and passing odds suggests a gradual improvement in women's representation over time, though progress remains incremental at only 2% increased odds per year. The conflicting relationship between critical reception metrics—with higher Metascores associated with better representation but higher IMDb ratings linked to worse outcomes—points to a potential disconnect between professional critical assessment and general audience preferences. Budget dynamics further complicate the picture, with higher-budget productions less likely to pass despite generating greater international revenue when they do feature substantive female representation. These findings underscore the complex interaction between commercial pressures, creative choices, and evolving cultural expectations in mainstream filmmaking. While the Bechdel test remains a limited measure of gender representation, it continues to offer valuable insights into persistent patterns of inequality in cinema. As the film industry moves forward, these results suggest that meaningful representation of women requires intentional creative choices that may sometimes challenge conventional audience expectations and traditional production formulas.

5 Additional Work

5.1 Exploratory Data Analysis

Table 2: Movies by Decade

Decade	Count
1970s	21 (0.01)
1980s	101 (0.06)
1990s	147 (0.08)
2000s	699 (0.39)
2010s	826 (0.46)
Total	**1794** (1.00)

Figure 1: Bechdel Pass Rate by Genre and Year

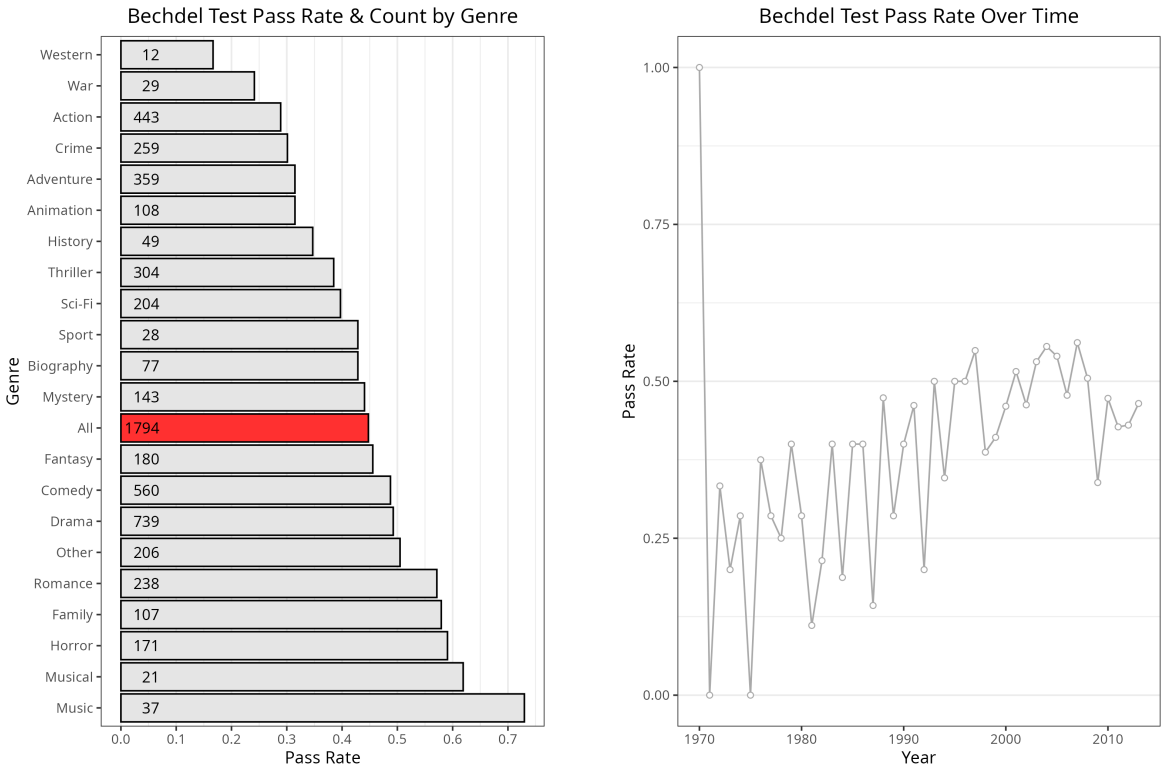
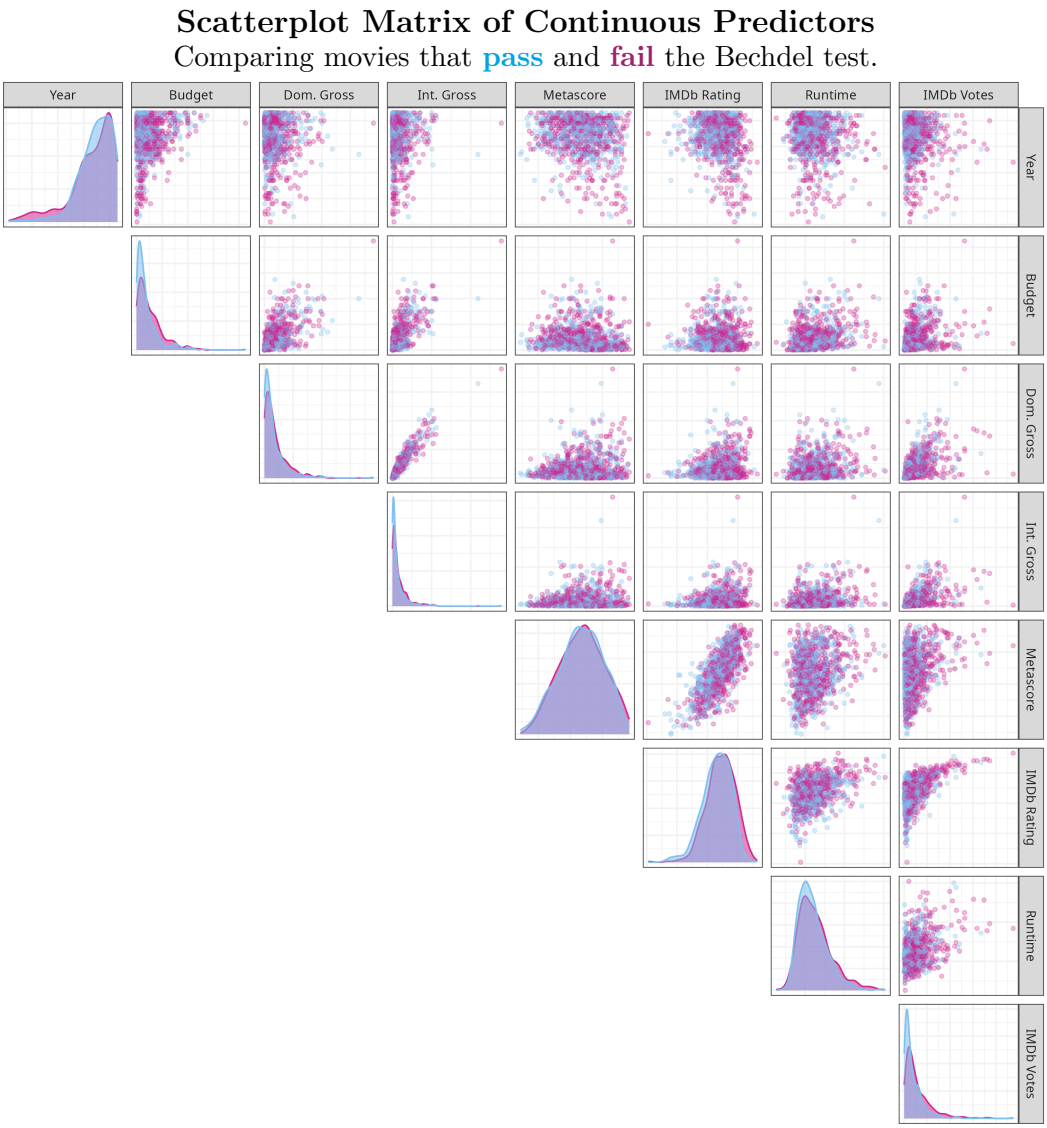


Table 2 shows the distribution of movies by decade, with the vast majority released after 1990.

Figure 1 shows average Bechdel test pass rates by genre and year. Genres like western, war, action, and crime have low pass rates, while music/musical, horror, family, and romance have high rates. Excluding the 1970 outlier ($n = 1$), the pass rate generally increases over time, though with significant year-to-year fluctuations (decades with fewer movies tend to be less stable).

Figure 2 presents a scatterplot matrix of all continuous predictors, based on a balanced sample of 1,000 movies. Passing movies are shown in blue, failing ones in maroon. The scatterplots appear noisy, with little visual separation between the two groups, suggesting that the model's predictive power may be limited.

Figure 2: Scatterplot Matrix Color-Coded by Bechdel Test Outcome



5.2 Model Building

We began by fitting an intercept-only null model and a full model with all 17 predictors. The `vif` function from the `car` package revealed high multicollinearity in the full model, particularly among the inflation-adjusted and nominal dollar amount variables. Table 3 shows the GVIF values above 5 before and adjustments. To address this, we removed the three inflation-adjusted variables, which reduced all GVIF values below 5, as shown in the "after" table.

Table 3: GVIF Values of Selected Variables Before and After Removal

	$GVIF^{(1/2Df)}$ Before
Budget	8.40
Dom. Gross	12.82
Int. Gross	16.30
Budget (2013)	8.22
Dom. Gross (2013)	13.82
Int. Gross (2013)	16.88

	$GVIF^{(1/2Df)}$ After
Budget	1.71
Dom. Gross	3.29
Int. Gross	3.41

After removing the three predictors, the revised "full" model summary is shown in Table 4. Although a few predictors appear highly significant, most are not according to individual Wald tests. Still, the model significantly improves upon the null model, as indicated by the likelihood ratio test in Table 5.

Table 4: Coefficient Estimates and Standard Errors of Full Model

	Estimate	Std. Error	z value	$Pr(>\|z\|)$	Signif.
(Intercept)	-5.113e+01	1.709e+01	-2.991e+00	2.778e-03	**
Year	2.484e-02	8.482e-03	2.928e+00	3.409e-03	**
Budget	-7.770e-09	2.100e-09	-3.699e+00	2.161e-04	***
Dom. Gross	1.138e-09	2.336e-09	4.869e-01	6.264e-01	
Int. Gross	1.067e-09	9.099e-10	1.173e+00	2.409e-01	
Rating Score	-7.200e-02	8.610e-02	-8.363e-01	4.030e-01	
English	4.478e-02	1.235e-01	3.625e-01	7.170e-01	
American	-1.990e-01	1.235e-01	-1.612e+00	1.071e-01	
Metascore	1.255e-02	5.110e-03	2.456e+00	1.403e-02	*
IMDb Rating	-4.613e-01	1.109e-01	-4.159e+00	3.195e-05	***
Month	2.014e-02	2.254e-02	8.935e-01	3.716e-01	
Season Spring	4.004e-01	2.181e-01	1.836e+00	6.635e-02	.
Season Summer	1.733e-01	1.795e-01	9.653e-01	3.344e-01	
Season Winter	-1.268e-03	1.923e-01	-6.595e-03	9.947e-01	
Genre Score	7.090e+00	9.386e-01	7.554e+00	4.208e-14	***
Runtime	7.309e-03	3.639e-03	2.008e+00	4.462e-02	*
IMDb Votes	-1.809e-06	7.816e-07	-2.314e+00	2.067e-02	*

Table 5: Likelihood Ratio Test for Overall Model Fit

Resid. Df	Resid. Dev	Df	Deviance	$Pr(>\chi^2)$
1379	1725.611	16	193.0733	1.983e-32

To further affirm model fit, we conduct the Hosmer-Lemeshow test. As shown in Table 6, the p-value indicates no significant difference between observed and expected counts across deciles.

Table 6: Hosmer and Lemeshow Goodness of Fit Test

χ^2	Df	$Pr(>\chi^2)$
10.135	8	0.2557

To improve the model, we use the `glmulti` package with a genetic algorithm to select models based on AIC and BIC. The best AIC model ($AIC = 1752.75$) is shown in Table 1, while the best BIC model ($BIC = 1787.34$) includes just three predictors: Metascore, IMDb Rating, and Genre Score. We choose the AIC model for its added complexity, but the BIC model highlights the importance of these core predictors. As shown in Table 7, a non-significant likelihood ratio test suggests the AIC model performs as well as the full model.

Note: To ensure comparable IC values during selection, all models were fit on the same subset of rows with no NAs. After selection, the reduced model was refit on the full dataset, including rows containing NAs in dropped predictors, so AIC and estimates will differ slightly. This is reflected in Table 1.

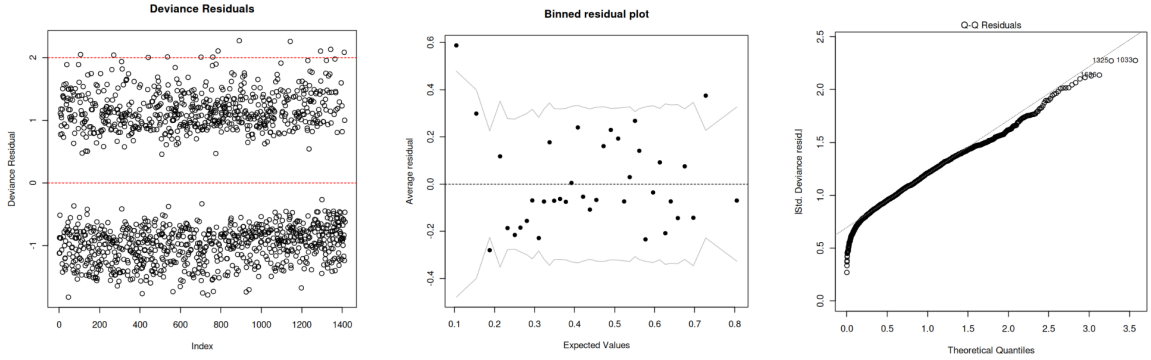
Table 7: Likelihood Ratio Test Comparing Full Model to Reduced

Resid. Df	Resid. Dev	Df	Deviance	$Pr(>\chi^2)$
1379	1725.611	8	9.137107	0.3308622

Figure 3 presents several diagnostic residual plots for the reduced model. In the first panel, the deviance residuals form two homoskedastic bands around zero with no apparent structure, indicating a good

fit. In the second panel, the binned residual plot shows Pearson residuals scattered randomly around zero and mostly within the error bands, reinforcing that conclusion.[6] The third panel's QQ-plot shows approximate normality of the standardized Pearson residuals, especially in the middle range. While normality isn't a logistic regression assumption, this also suggests a reasonably good fit.[7]

Figure 3: Residual Diagnostic Plots



5.3 Performance

Figure 4 shows the ROC curve for the reduced model (Table 1), plotting TPR (sensitivity) against FPR (1 – specificity) across all classification thresholds. The model achieves an AUC of 0.707 with an optimal threshold of 0.457, corresponding to a sensitivity of 0.653 and specificity of 0.679.

Figure 4: ROC Curve Labeled with AUC and Optimal Threshold

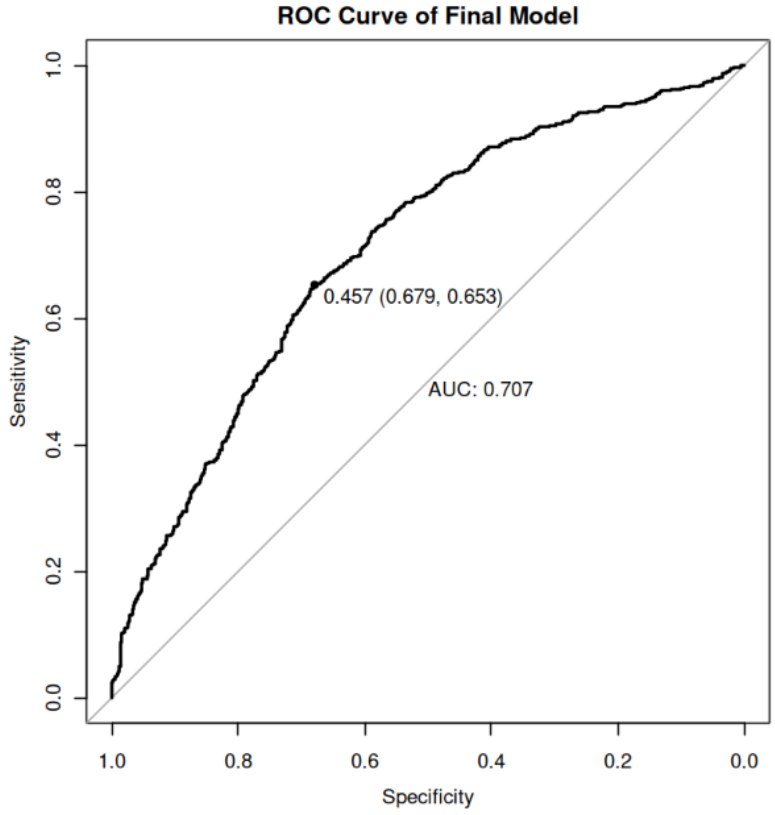


Table 8: Contingency Table Based on Optimal Threshold

	Actual **FAIL**	Actual **PASS**	Total
Pred. **FAIL**	**534** (0.38)	**218** (0.15)	**752** (0.53)
Pred. **PASS**	**252** (0.18)	**410** (0.29)	**662** (0.47)
Total	**786** (0.56)	**628** (0.44)	**1414** (1.00)

Using the classification threshold of 0.457, the model's contingency table is shown in Table 8. The overall accuracy is $\frac{534+410}{1414} \approx 66.8\%$.

6 References

1. Bechdel, Alison. (1985). "The Rule". *Dykes to Watch Out For.*
2. Morlan, Kinsee (2014, July 23). "Comic-Con vs. the Bechdel Test". *San Diego City Beat.* https://web.archive.org/web/20150316161800/http://www.sdcitybeat.com/sandiego/article-13243-comic-con-vs-the-bechdel-test.html
3. Woolf, Virginia (1929). "Chapter V". *A Room of One's Own.* https://gutenberg.ca/ebooks/woolfv-aroomofonesown/woolfv-aroomofonesown-00-h.html
4. Data Science Learning Community. (2024). Tidy Tuesday. https://github.com/rfordatascience/tidytuesday
5. Hickey, Walt. (2014, April 1). *The dollar-and-cents case against Hollywood's exclusion of women.* FiveThirtyEight. https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/
6. Tackett, M. (2019, October 30). *Logistic regression: Model fit & exploratory data analysis* [Lecture slides]. Duke University. https://www2.stat.duke.edu/courses/Fall19/sta210.001/slides/lec-slides/18-logistic-pt3.html
7. Ford, C. (2022, September 28). *Understanding deviance residuals.* University of Virginia Library. https://library.virginia.edu/data/articles/understanding-deviance-residuals