**The Problem:**

The film industry has come under fire recently for its depiction of female characters in various films. Traditional stereotypes of female characters as weak, timid, and reliant on men have come under attack in the past few decades, and Hollywood has been pressured to move away from these stereotypes and portray more diverse, well-rounded female characters. The Bechdel test has become a popular measure used to assess these efforts. Simply put, a movie passes the Bechdel test if it has at least 1 scene with 2 women talking about something other than men or relationships. While the test seems simple enough, a surprisingly large number of movies simply fails to pass this test. The pass rate varies substantially across different genres, however.

It is clear that examining the trends in the Bechdel data and trying to predict if an upcoming movie will pass the Bechdel test would help show which aspects of a film influence whether it will pass the Bechdel test. Another task is to examine the effects of passing the Bechdel test on box office revenues and popularity of a given movie. This is important because Hollywood, a collection of for-profit studios, will be better motivated by financial incentives to depict more well-rounded female characters. The clients will consist of the film industry, especially the one based in Hollywood. Directors and writers are often the people most responsible for creating dialogue and casting characters. Film studios also play a role in determining the focus and direction of a movie. By influencing these players, the film industry will hopefully become more equitable and fair in its portrayal of women.

**Data**

The dataset contains data on over 50 different attributes for more than 7000 films, including IMDB data like the title, the year, the writer, and the director. Additionally, the data includes fields for films that indicate whether or not they pass the Bechdel test and the reason why they failed, such as having no women or no women with dialogue. The dataset also includes data on the film’s genre, box office revenue data, budget data and facebook likes. Finally, the dataset contains the genders for the writer, key producer, and director of each film, along with the past pass percentages for each of these people.

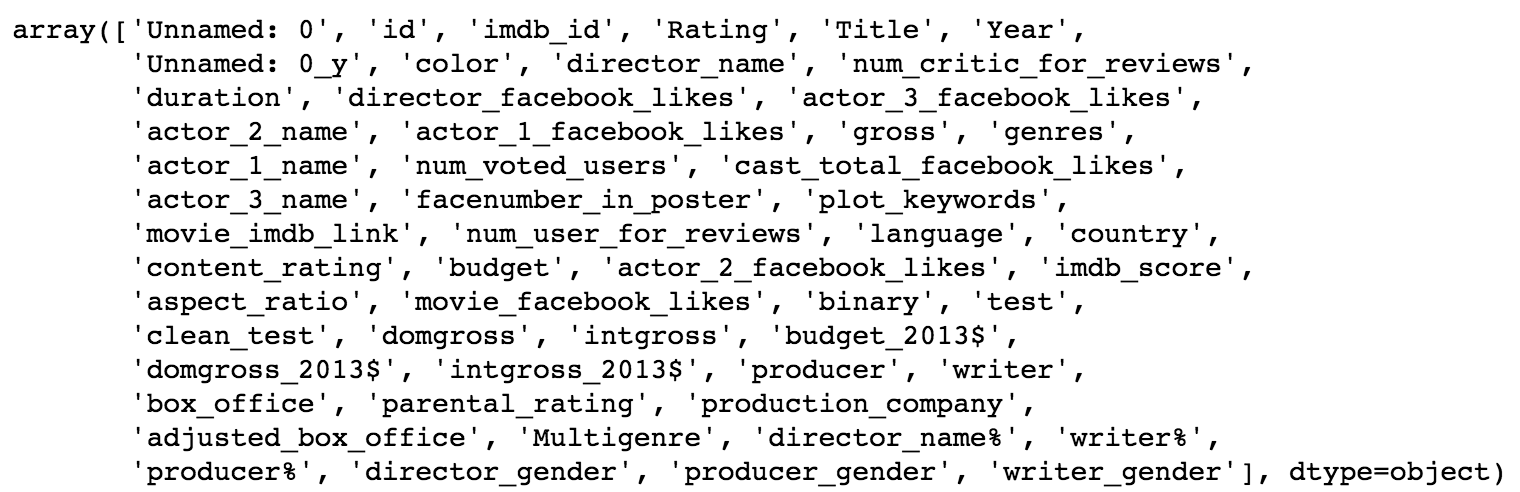
I did a considerable amount of wrangling and cleaning to get the dataset ready. I decided to use a dataset of over 7000 movies from bechdeltest.com, enabling me to have a large dataset of movies from a variety of genres and time periods ready to go. I also used the omdb API to get box office revenue data for hundreds of films and used the behindthename API to get the gender data for producers, writers, and directors based on their first names. If no data could be retrieved from behindthename, I assumed that the person was a male due to the fact these occupations tend to be dominated by men.

I had to wrangle the data intensively to get it ready for analysis. I had to fix issues with certain fields having incorrectly formatted or missing values. Additionally, I merged the original dataset I found on data.world with the expanded bechdeltest dataset containing the information on pass/fail, enabling me to have more data and construct a stronger analysis.

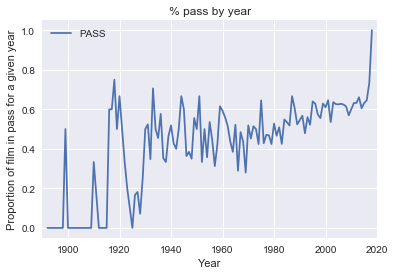
There are some limitations on what the data can be used to answer. It cannot answer questions on the gender makeup of the casts of these films and the salary differences between actors and actresses in these films. Additionally, the dataset does not contain all the box office revenue data and does not give a complete picture of the effect of passing the Bechdel test on revenue. The bechdeltest API also contains potential bias because it consists of user submitted data, which is likely to be biased in favor of movies that pass the Bechdel test. There is also a limit on the machine learning algorithms that can be applied to predict which films pass the Bechdel test since most of the attributes used are categorical. One dataset that would be useful for this analysis would be a dataset containing box office revenue data, DVD/VHS revenues, and streaming views for a variety of films; this would enable a much more comprehensive analysis of the Bechdel test on a film’s popularity and its revenues.

**Exploratory Analysis**

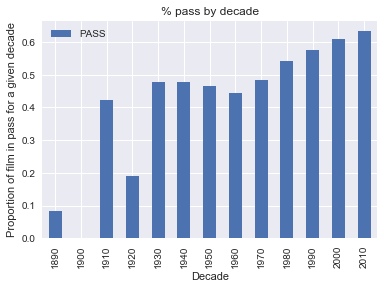
The dataset contains a variety of interesting relationships and trends. Here is a list of the column names used:



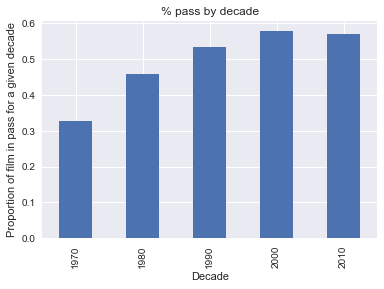
A natural trend to examine is to look at the pass rate of each film by decade. Here is a graph visualizing just such a trend:



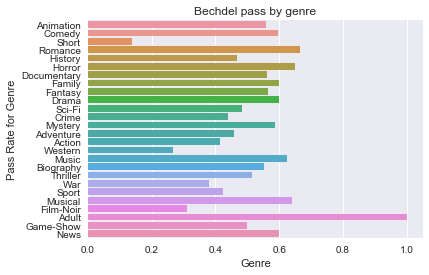
The trend is very choppy, especially in the early years of the 20th century. Let us analyze by decade to see a clearer trend:



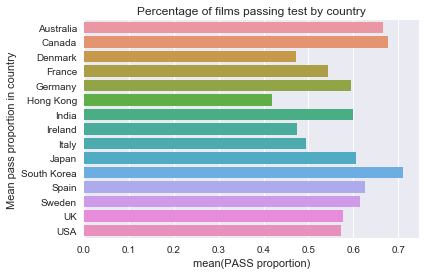
Note that the value is 0 for the 1900s. This indicates that no movies were recorded as passing the Bechdel dataset from that decade. Note also that the 1920s and the 1910s also have vastly differing pass rates, again caused by the low amount of movies found in these decades. The trend seems to be that the pass rate has been steadily increasing since the 1970s, but the Bechdel pass rate is still slightly above 60%. Let us look at the trend by decade for each of the films in the original dataset:



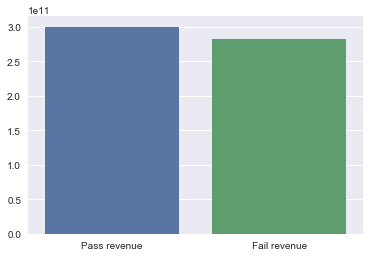
It looks like the pass rates are higher in this case, but still support the idea of an upward trend in the pass rate. Hence it appears that the overall trend in the pass rate is getting better throughout each decade.



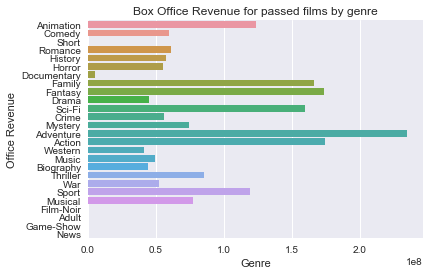
It looks like the Bechdel test pass rate varies by genre considerably. The genre could definitely prove to be a very valuable predictor in future machine learning models. Let us also look at the country of origin for each movie and see how that affects the Bechdel pass rate:

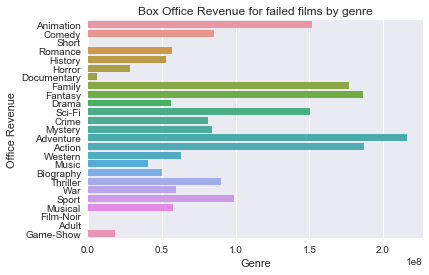


It appears the pass rate seems to between .5 and .7 in these countries, reflecting a general trend towards displaying well-rounded female characters in global films. The US is about in the middle, with a pass proportion of around 57%. Let us look at the overall revenue data too. This data is adjusted to 2013 dollars.



The revenue appears to be higher for the pass revenue dataset, but it is not clear if this difference is statistically significant. Let us look at the genre breakdown of revenue here for passing and failing films:





It seems that for genres like animation and drama, there is a difference between the pass rates for each group. Looking at median revenues, the result is a bit shocking:



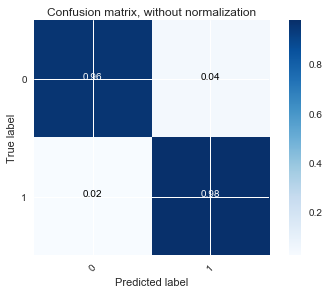
It seems that there is a difference between the median pass and fail revenues here. This suggests that it might not be beneficial in general to write films that pass the Bechdel test. To simplify our analysis, we will focus on analyzing the revenue for post-1990 American films since their characteristics reflect modern mindsets and values. The initial analysis led me to focus on two things: using machine learning to predict if a film passes the Bechdel test and analyzing box office revenue data.

**Machine Learning: Approach and Results**

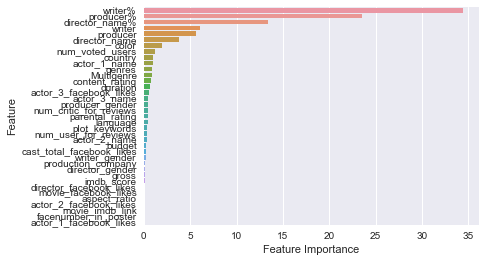
I decided to use an ensemble method of multiple decision trees using boosting. Boosting methods tend to be very powerful and easy to apply, and they focus training on examples that are particularly tricky to classify. They are very focused on this as the objective and the basis for the training loss, creating a classifier that is robust and able to handle difficult data points. I felt that this behavior would work well with my dataset, consisting of various unrelated and complex features. Since most of the features in the dataset are categorical, the decision tree is fundamentally the right choice for such a classification task since it relies on creating rules about the categorical values.

I used the CatBoost library for the machine learning due to it being designed to deal with datasets heavy with categorical features. It enabled me to specify which features were categorical, and it performed the necessary transformations on these features to make them work with the boosting trees. I did not have to manually convert the categorical features, and this saved a lot of time and effort during the construction of a machine learning model. I decided to exclude revenue-related features, the title, the year, and the imdb\_id, realizing that these features would either lead to a classifer that overfits or would not matter much in a model.

The first task was a simple binary classification problem; the task was to determine if a movie passed or failed the Bechdel test. The CatBoostClassifier was run over 16 iterations and with a depth of 12. The accuracy was 96.78% on the test data and the confusion matrix detailing which instances were misclassified was as follows:



The classifier appears to perform very well on the binary classification task, and there is no bias towards one label being classified more over the other. The feature importances were also calculated:



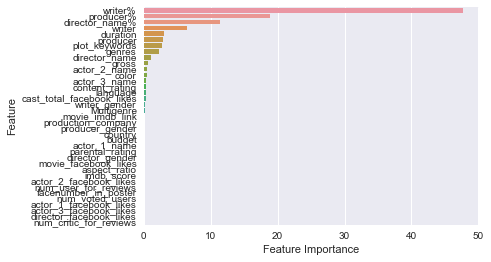
It looks like the most important attributes by far were the overall pass rates for the writers, directors, and producers with the other features being largely insignificant in determining which films passed vs. failed. The feature importance graph suggests that writers play the largest role in determining which films pass the Bechdel test, which makes sense given that they write the actual scripts. Producers and directors also have influence over the results, but it appears their influence is more indirect.

The dataset also contains more labels corresponding to why a film failed the Bechdel test. The labels are 0- no women, 1- no dialogue, 2- dialogue between women focuses on men, and 3- passes the test. A Multiclass Classifier was created to try and solve a more complex classification problem of trying to predict these labels. The Hyperparameters for this classifier were: number of iterations, tree depth, learning rate, and l2 regulation. The loss function was also determined to be a MultiClass loss function, complex but effective here. The number of iterations, the tree depth, and the l2 regulation played little role in the accuracy: the accuracies found were between 67% and 70% throughout the parameter fitting. For each parameter, the values were 16, 8, and 4 respectively. The depth was limited since larger values of depth led to much longer training times for no significant gains in accuracy. The learning rate was crucial to the model’s accuracy. A learning rate of .4 led to an accuracy around 78%, indicating that a small learning rate was crucial to the model’s success. This value indicates how fast or slow we converge towards optimal weights, and we want to converge more slowly.

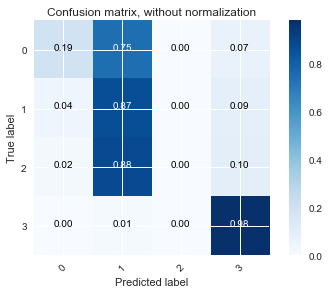
The final result was as follows:

Accuracy 77.65%

Feature Importances:



Confusion Matrix:



The feature importances seem to show that duration and genre become more important, but the writer still is the most influential in determining the classification. The confusion matrix for the data indicates the classifier is good at predicting if a film passes or if a film has no female characters speaking to each other. The other two labels have a very high rate of error, indicating the classifier does not distinguish these films very well at all. Overall, the binary and multiclass classification tasks yield themselves well to boosting, and any future film’s test result can be predicted by feeding relevant attributes to a classifier trained on this type of data.

Revenue Analysis:

I focused on a sbuse