Bechdel Test and Movie Revenue

Krissy Groom

June 8, 2019

DSC 540: Advanced Machine Learning Applications

Spring Semester, 2019

ABSTRACT

The purpose of this analysis and Machine Learning project is to understand if and how the Bechdel test, a test that is typically used to measure Hollywood bias, has an influence on a movie's success as determined by international revenue. This project utilized various Machine Learning regression models and other advanced ML learning techniques such as Feature Selection, in addition to an initial Exploratory Data Analysis, to further understand revenue trends for films that pass the Bechdel test and for those that do not. Along with understanding the general trends of budget and revenue given the Bechdel test, the purpose of this project is also to evaluate model performance of multiple Machine Learning models on this dataset, such as Decision Tree Regression, Random Forest Regression, Gradient and Ada Boosting, Neural Network Regressor model, and Support Vector Machine (rbf and linear kernel models) by comparing test evaluation metrics such as RMSE, Explained Variance, and utilizing 5-fold Cross-Validation. Exploratory Data Analysis and Feature Selection processes confirmed that Bechdel test features were not considered to be important features in determining the target variable of international film revenue, and that budget and domestic revenue were the most significant features for predicting this target. Each of the Machine Learning models performed well according to the test evaluation metrics and comparably for this data, however, Gradient Descent and SVM (linear kernel) performed the fastest. Issues with the handling of missing data, the dataset having a small feature set, time constraints, and the general simplicity of the Bechdel test itself may limit the general applicability of the models in this project.

INTRODUCTION

Recently, a review of the new Captain Marvel movie brought my attention to a potentially more gender-inclusive and evolving Hollywood. The review described the film as "empowering for girls (and everyone!) from start to finish. It features women in STEAM, an interracial female friendship, a woman of color as a model parent, and women as effective, compassionate leaders. Female action heroes are often hyper-sexualized, even though research [Heldman, Frankel, & Holmes, 2016] shows this diminishes their power and agency, but *Captain Marvel* thankfully avoids even a single moment of sexual objectification. *Captain Marvel* is presented as a full and complex human being" **[3]**. I celebrated what I saw as media beginning to reflect the need for and value of strong girls growing into future leaders and technologists. After reading the review and considering shifts toward greater diversity in shows and movies over the past few years, I sought out to discover if trends were beginning to change and if so how are these possible changes influencing film popularity.

Right away I found a dataset of interest at https://data.world/carolee/women-in-movies with the original data found at: https://github.com/fivethirtyeight/data/tree/master/bechdel.

The dataset references a 2014 article from FiveThirtyEight.com – found at: https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/. -- which analyzes data about movies that either pass or do not pass the Bechdel Test. The Bechdel Test is a test that measures Hollywood bias in that it specifies that "if a movie can satisfy three criteria — there are at least two named women in the picture, they have a conversation with each other at some point, and that conversation isn't about a male character — then it passes "The Rule," whereby female characters are allocated a bare minimum of depth" **[7].** Although this test is simplistic as a way of authentically determining whether women are adequately featured, it is considered one of the most "enduring tools to measure Hollywood's gender bias" **[7]**, and was originally promoted by cartoonist Alison Bechdel in a 1985 comic strip "The Rule"**(figure 1)**. Also, researchers note that "Whether a movie passes the test is nothing more than an anecdote, but the systematic analysis of a set of movies can reveal the gender bias of the movie industry" **[4]**.

This dataset consists of data for films from the years 1970 to 2013, and it has 15 features, 1796 instances, and there are approximately 200 missing values. The features include label features such as IMDB id and the title of the film, categorical features such as test values with a cleaned test values feature, as well as a binary feature with PASS/NO PASS, year and period categories, as well as numeric features such as budget and revenue. After reviewing the features and considering which target values were most interesting to my research, I chose international revenue as my target variable because I noticed in the original article from fivethirtyeight.com, the researchers described Hollywood as maintaining the belief that female featured films will not be well-received by international audiences **[7].**

My question for this project was: Do movies that pass the Bechdel test differ in their success from movies that fail the test as measured by international movie revenue. My interest in this dataset was also to learn to utilize various Machine Learning regression models and other advanced ML learning techniques such as Feature Selection that had been introduced throughout this course so that I may further understand the revenue trends for these films. I was also interested in discovering which models perform the best for predicting international revenue, as indicated by test evaluation metrics such as RMSE, Explained Variance, and 5-fold Cross-Validation.

LITERATURE REVIEW

I began my analysis by reviewing previous research regarding gender bias in Hollywood and gender bias in Machine Learning, as well as the specific Machine Learning techniques that I would be using for my project. Prior research on gender independence in Hollywood tells us that "Starting from an equal approach to male and female independence in movies, we verified the existence of a generalized bias in which female characters are shown as dependent on male characters" **[4].** This research furthered my curiosity as to whether this initial bias that seems to prevent women from being cast with independence and autonomy would influence a movie's revenue potential prior to even being produced at the level of budget. Researchers from the original analysis from fivethirtyeight.com found that "the median budget of movies that passed the test — those that featured a conversation between two women about something other than a man — was substantially lower than the median budget of all films in the sample. What's more, [they] found that the data doesn't appear to support the persistent Hollywood belief that

films featuring women do worse at the box office. Instead, [they] found evidence that films that feature meaningful interactions between women may in fact have a better return on investment, overall, than films that don't" **[7]**. Researchers also discovered that more movies are passing the Bechdel test since 1970 but that the number of films passing the test have leveled off since the late 1990's **(figure 2).**

As for Machine Learning bias, which was informative but less practically applicable for this specific project, I discovered that gender bias can be introduced into Machine Learning algorithms by the researchers themselves **[1], [2]**. This can be highly problematic as researchers attempt to find valuable information from the data which influences decision-makers and those reliant on those decision makers. I agreed with the research that reminds us that "data is not a panacea. Where data is used predictively to assist decision making, it can affect the fortunes of whole classes of people in consistently unfavorable ways" **[1]**, and that the "blind application of machine learning runs the risk of amplifying biases present in data" **[2]**. I recognized my responsibility as a Machine Learning Researcher to be as aware as possible of my own inherent biases, and to move forward with my analyses and ML procedures with care to avoid introducing further inequities into my models particularly as I explore inequalities.

In addition to gender bias, I researched various Machine Learning models that I will be using throughout my project including Decision Trees and Random Forests **[8]**, Neural Networks **[9], [6]**, and Support Vector Machines **[6], [10].** I will be describing these techniques and my methods for utilizing them in the following sections.

METHODS

Prior to running Machine Learning models on the dataset, I performed Exploratory Data Analysis (EDA) to understand the raw data, such as the features, the observations, and the distributions of the data. I removed columns that would be unnecessary for the Machine Learning predictions, such as the ID and movie title columns and checked the data for missing values. I discovered that many of these values could be easily filled so I added the missing period code and decade code values for movies from 1970 - 1989. I then removed any remaining rows with missing values because there were less than 20 rows and I decided that this was a small enough number of observations to likely not affect the data very significantly, although I recognized, as with any preprocessing, that I could be introducing bias into the dataset. I chose to keep only the clean_test and binary features which appeared to have more clear and descriptive values and I created dummy variables for these features. I then visualized the distributions of the remaining variables and noted that they were not normally distributed **(figure 3)**. It was during my EDA, that I decided to use international gross revenue as my target feature as I noted the differences between the means and medians of passing vs failing movies for budget, domgross, and intgross. I checked correlations to see if there are linear relationships among the variables and I noticed that budget and domestic revenue, and budget and international revenue are positively correlated, domestic and international revenue have a strong positive relationship, and I see little to no linear relationship between revenue and pass/fail of the Bechdel test **(figure 4)**. I realized after I initially ran a Decision Tree model on the data that I would need to normalize the data given that revenue is so much larger than the binary values of the dummy variables and the period and decade code so I used the sklearn

preprocessing module to perform min/max normalization on the data, although I did not normalize the period and decade code values. I used the normalized dataset for my Machine Learning models at this point.

After cleaning and normalizing the dataset, I began running my models. Since my target variable is numeric, I focused on Regressor models. I started with Decision Tree and Random Forest models. A Decision tree is "primarily a method of constructing a set of decision rules on the predictor variables" **[8].** Whereas, Random Forests are similar to Bagging Trees in that "bootstrap samples are drawn to construct multiple trees; the difference is that the each tree is grown with a randomized subset of predictors, hence the name ''random'' forests. A large number of trees (500 to 2,000) are grown, hence a ''forest'' of trees. The number of predictors used to find the best split at each node is a randomly chosen subset of the total number of predictors" **[8].**

I ran the Decision tree and Random Forest models on my preprocessed data first with no cross-validation. In order to determine the goodness of a model, I was looking for RMSE (Root mean-squared-error) to be very low which tells me that the model was accurate and predicted values very close to the actual values, and for Explained Variance to be high, similarly to the R2 metric which tells us the percent of variability of the target that can be explained by the model. I ran the Decision Tree and Random Forest models with cross-validation which, as we learned in class, is the gold-standard of determining the robustness of the model. I noted the results with cross-validation and without cross validation as well as the runtimes for both Decision Trees and Random Forests. I initially thought that I would run the RF model with the number of trees at 100, 200, 300, and 1000 in addition to the original 500, however I ended up only running with 100 estimators because the results were comparable to 500 but with much faster runtime**.**

I also ran the models with Feature Selection using a wrapper procedure with a Random Forest Regressor to determine the best subset of features and I compared the results to those without Feature Selection. Additionally, I looked at Feature Importance using a Random Forest Regressor **(figure 5).**

Next, I ran Gradient Boosting and Ada Boosting models. Boosting models work by utilizing iterations of weighted learners that are weak learners in order to avoid overfitting. I ran the Gradient Boosting model, paying attention to the parameters: n_estimators = 100, which tells me the number of iterations; the learning rate which determines the step size and was set to 0.1; the loss function which was set to 'ls' which is least squares regression; max_depth which was set to 3, and determines the maximum depth of each individual regression estimators. For Ada Boosting, I set n_estimators = 100, and learning_rate = 0.5. I ran both models with CV=5 and viewed the results with and without feature selection.

I then ran my Neural Network Regressor with parameters: activation function set to logistic sigmoid function, solver (for weight optimization) was set to 'lbfgs' which is best for smaller datasets, alpha (regularization parameter) was set to the default value of 0.0001, and max_iter (in order to tell the model to cut-off at some point) was set to 1000, hidden_layer_sizes was set to (10,) - which determines the number of nodes in the ith layer. I learned that "Neural networks (NN) are known to be biologically inspired analytical techniques, capable of modeling extremely complex nonlinear functions" **[9].** I ran the Neural Network Regressor with and without feature selection to compare results.

The last model that I ran on my dataset was Support Vector Machine. Hajek and Olej (2010), describe SVM's as having a design that "depends on the nonlinear projection of the input space Ξ into multidimensional space Λ, and on the construction of an optimal hyperplane. This operation is dependent on the estimation of inner product kernel referred to as kernel function" **[6]**. I ran the SVM Regressor model with the following parameters: kernel initially set to 'rbf' - as I mentioned, kernel refers to the way in which the data will be transformed into a different dimensionality and here I began with rbf, gamma = 0.1, and C=1.0. Gamma is the kernel coefficient and according to the scikit-learn API, "Intuitively, the gamma parameter defines how far the influence of a single training example reaches, with low values meaning 'far' and high values meaning 'close'". Also the C value is the error variable: "The C parameter trades off correct classification of training examples against maximization of the decision function's margin" (https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html). I first ran my model with the kernel set to rbf, then changed this parameter to linear to compare results. I also ran the SVM Regressor with feature selection using the wrapper method with SVM linear kernel model.

RESULTS

From the initial Exploratory Data Analysis and by visualizing the differences in budget and revenues, I noticed an average difference between films that pass and those that fail the Bechdel test. Although given that movies that pass the test, on average receive less of a budget, and that budget and revenue are fairly strongly correlated, I recognized that budget may have more to do with a movie's success than perhaps the number and complexity of the women in the movie. I would take this into account as I reviewed results from Feature Selection for each of the models.

Decision Tree (**figure 6)** and Random Forest results varied slightly when running the models without Cross Validation but overall both Decision Tree model and Random Forest models performed well with very low RMSE for all models at ~0.02 and a high Explained Variance with all values ranging from approximately 0.84 - 0.92. The very low RMSE and pretty high Explained Variance tell me that both of these models could be good models for the dataset, although Decision Trees do not generalize well, so I likely would not use this model for prediction of international revenue. With Cross Validation, the results tell me that Decision Tree and Random Forest models perform equally well given RMSE with both at 0.03 (+/- 0.01 - 0.02), although the Random Forest model may perform slightly better when comparing Explained Variance between the two models with DT at 0.85 (+/- 0.07) and RF at 0.87 (+/- 0.10). However, Random Forest Runtime was much longer, with DT at approximately 0.03 seconds and RF at ~3.96 seconds. Feature Selection improved runtimes for both DT and RF, but only slightly, with comparable RMSE and Explained Variance. Random Forest with the number of estimators reduced to 100, maintained the high Explained Variance and low RMSE but reduced the runtime significantly to ~0.75 seconds.

At this point and importantly, I found by implementing Feature Selection using a wrapper with Random Forest Regression, it was determined that the two significant features for predicting international gross revenue were domestic gross revenue and budget **(see Feature Importance graph - figure 5).** This confirmed the results from the original

fivethirtyeight.com analysis, that determined that Bechdel related variables were not significant in determining movie success.

Gradient Boosting and Ada Boosting performed similarly well to DT and RF, with Gradient Boosting results: RMSE - 0.03 (+/- 0.01), Explained Variance - 0.88 (+/- 0.03) and Ada Boosting results: RMSE - 0.03 (+/- 0.01), Explained Variance - 0.89 (+/- 0.05). Gradient Boosting seems to perform very slightly better than the other models based on RMSE with an Explained variance that is very similar to both Random Forest and Ada Boosting, although, GB performed faster with a runtime ~0.267 seconds. Utilizing Feature Selection with a wrapper method and Gradient Boosting model, I noticed a slight drop in the performance metrics for Gradient Boosting model and an improvement in the runtime and only one feature was selected. I changed the Feature Selection method to Random Forest which improved Gradient Boosting RMSE (0.02 (+/- 0.01) and Explained Variance(0.92 (+/- 0.03), and it also improved the runtime slightly(~0.165). Up to this point, the results are similarly as good without Feature Selection, the model implementing Feature Selection would be preferable.

Neural Network with CV=5 and without Feature Selection performed similarly to Random Forest, Gradient Boosting, and Ada Boosting, but with a very slightly lower Explained Variance (0.90 (+/- 0.02) and surprising to me, a faster runtime (~0.1551). I learned that Neural Networks perform well with normalized data, so that could explain the faster runtime. With Feature Selection, the performance metrics remained the same except for runtime which was very fast (~0.0553).

Support Vector Machine model using 'rbf' kernel with no Feature Selection performed very fast (~0.035), but seemed to have a much lower Explained Variance 0.64 (+/- 0.54), although when I view the spread I see there is a very large range of values, possibly indicating that this model is less robust and consistent than the other models. A SVM model with a linear kernel results show somewhat more consistency given the spread: RMSE:: 0.04 (+/- 0.02), Explained Variance: 0.81 (+/- 0.15),Runtime: ~0.0825. The results are similar to the other models but with the large spread, again this may be a less consistent model. With Feature Selection, the model seems to improve slightly in consistency and performance metrics: RMSE: 0.03 (+/- 0.01), Expl Var: 0.86 (+/- 0.08), Runtime: ~0.010. Linear SVM with Feature Selection is the fastest model.

Generally, all models performed very well and similarly. A comparison of results from all ML models with Feature Selection and CV=5 can be viewed in **figure 7.**

CONCLUSION

In reviewing my EDA and ML results, I found that there is insufficient evidence given this dataset to support the idea that Bechdel test passing films will not perform comparably to films that do not pass the test when the differences in budgets are accounted for. My results confirmed what was stated in the original analysis: that budget played a very strong roll in the gross revenue for both domestic and internationally audiences, and as I noted in my EDA, movies that pass the Bechdel test do receive on average less of a budget than movies that do not pass the Bechdel test. However, variable selection procedures for my models concluded that passing or not passing the Bechdel test was not a significant feature to include in determining international revenue. Removing the Bechdel variables had no impact on the evaluation metrics for the models in that the models performed just as well or in some cases

appeared to perform slightly better, and the majority had faster runtimes. I would suggest the use of Gradient Boosting as a model with the best results for this dataset, given its equally good RMSE and Explained Variance, but with faster runtime and more consistent results than some of the other models, although Random Forest, Neural Network, SVM with a linear kernel all produced very good results.

I am aware of the limitations that may prevent these results from generalizing to other datasets, such as preprocessing that could have introduced bias into the models, including the removal of missing values and the normalization of the data. I also recognize that the small number of features and somewhat small number of observations could be a limitation. I would have liked to include additional features such as star ratings as well as more current films. I have seen more diversity in television and movies than ever in recent years and and I am curious to see if this would be reflected in the data and if this would influence my results in any way.

Time limitations were also an issue for me in that I would have liked to have run the models with the target variable switched to domestic gross revenue, and then also budget to compare results, although, given the high correlation between domestic gross revenue and international revenue, the results may have been somewhat similar. I saw from my EDA that budget did seem to be influenced by pass/no pass of Bechdel test, however, it would have been interesting to see if my models would have confirmed this in testing domestic revenue and budget as the target variable.

Additionally, I recognize the limitations of the Bechdel test itself, in that there are movies with strong female roles that do not pass and other films that would likely be seen as terribly sexist that do pass. For now, this is deemed the best test to generally examine gender bias, although I hope even the need for such a simplistic determinant of female representation will change in the near future with all movies meeting such basic criteria. "There are anecdotal signs that there's a shift in thinking when it comes to movies featuring women and female relationships. Recently, Hollywood has been able to boast about the success of female-dominated films in the marketplace. [Some producers believe that] younger viewers tend to have a 'more equal view of men and women'" **[7]**, which is my hope, and that this will be reflected in all movies to come, from Captain Marvel and beyond.

TABLES AND FIGURES

Figure 1:

Figure 2:

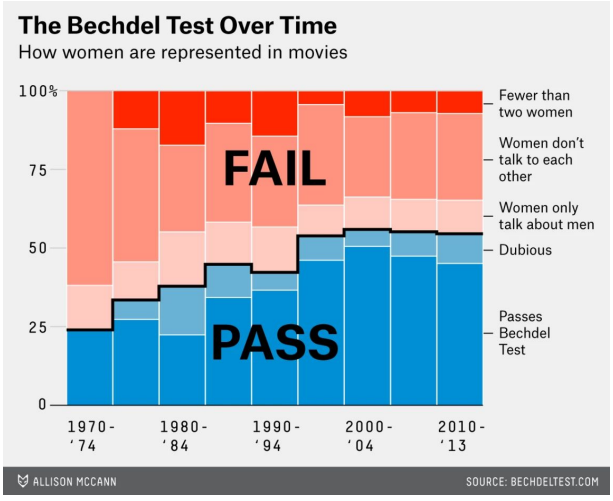

Figure 3:

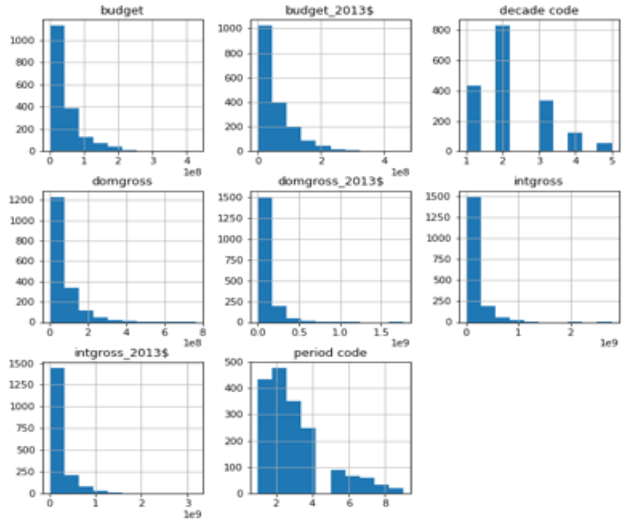

Figure 4:

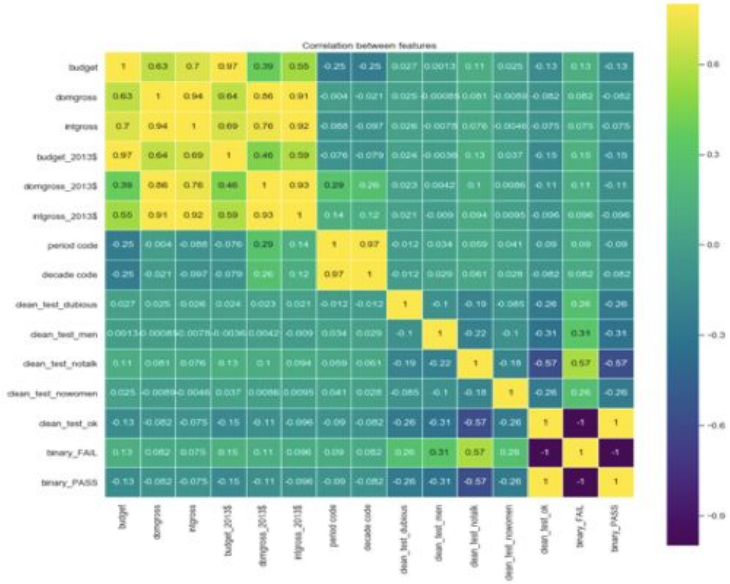

Figure 5:

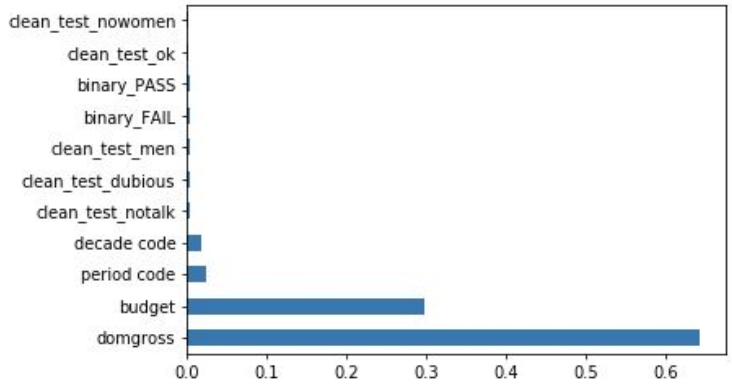

Figure 6:

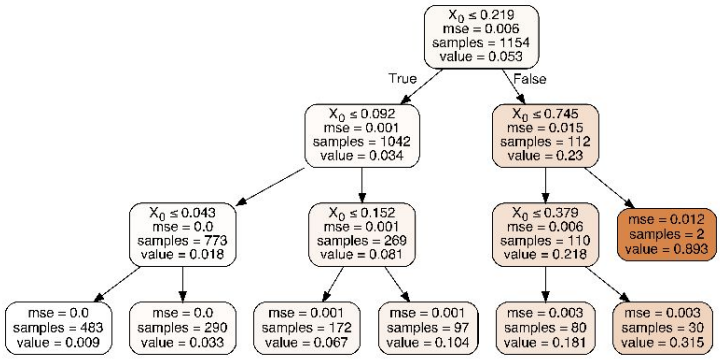

Figure 7:

	Results with CV=5 and Feature Selection	
Model	Test Metric	Diabetes
Decision Tree	RMSE	*0.03 (+/- 0.01)*
	Explained Variance	*0.85 (+/- 0.05)*
	Runtime	*0.017571*
Random Forest	RMSE	*0.02 (+/- 0.01)*
	Explained Variance	*0.87 (+/- 0.10)*
	Runtime	*0.748247*
Gradient Boosting	RMSE	*0.02 (+/- 0.01)*
	Explained Variance	*0.92 (+/- 0.03)*
	Runtime	*0.165877*
Ada Boosting	RMSE	*0.03 (+/- 0.01)*
	Explained Variance	*0.89 (+/- 0.07)*
	Runtime	*0.336013*
Neural Network	RMSE	*0.02 (+/- 0.01)*
	Explained Variance	*0.90 (+/- 0.02)*
	Runtime	*0.055305*
SVM (linear kernal)	RMSE	*0.03 (+/- 0.01)*
	Explained Variance	*0.86 (+/- 0.08)*
	Runtime	*0.010782*
SVM (rbf kernal)	RMSE	*0.07 (+/- 0.01)*
	Explained Variance	*0.83 (+/- 0.07)*
	Runtime	*0.015981*

REFERENCES

[1] Barocas, & Selbst (2016). Big data's disparate Impact. *California Law Review(671).* Retrieved from https://poseidon01.ssrn.com/delivery.php?ID=711119082125007113089094091086064030054021093008061013104027067125117022103070093064058001029022012102023080089103110124089113103029074046010090002086124114000121074050044015120085116124103127093115000118083078122005086088094098072122123067088088090122&EXT=pdf

[2] Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *NIPS'16 Proceedings of the 30th International Conference on Neural Information Processing Systems.* Retrieved from https://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf

[3] Captain Marvel Makes Film Herstory. (2019). Retrieved from http://therepresentationproject.org/captain-marvel-makes-film-herstory/

[4] Garcia, D., Weber, I., Rama Kiran Garimell, V. (2014) Gender asymmetries in reality and fiction: The bechdel test of social media. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media.* Retrieved from https://arxiv.org/pdf/1404.0163.pdf

[5] Heldman, C. & Frankel, L. & Holmes, J. (2016). "Hot, Black Leather, Whip": The (De)evolution of Female Protagonists in Action Cinema, 1960-2014. Sexualization, Media, and Society. 2. 10.1177/2374623815627789. Retrieved from https://journals.sagepub.com/doi/full/10.1177/2374623815627789

[6] Hajek, P., Olej, V. (2010) Municipal revenue prediction by ensembles of neural networks and support vector machines. *WSEAS Transactions on Computers*(I. 11, V.9). Retrieved from https://www.researchgate.net/profile/Petr_Hajek8/publication/228945890_Municipal_revenue_prediction_by_ensembles_of_neural_networks_and_support_vector_machines/links/55daf31c08aeb38e8a8a2f02/Municipal-revenue-prediction-by-ensembles-of-neural-networks-and-support-vector-machines.pdf

[7] Hickey, W. (2014). The Dollar-And-Cents Case Against Hollywood's Exclusion of Women. Retrieved from https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/

[8] Prasad, A., Iverson, L., and Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems(9).* Retrieved from https://www.researchgate.net/profile/Louis_Iverson/publication/225909333_Newer_Classification_and_Regression_Tree_Techniques_Bagging_and_Random_Forests_for_Ecological_Prediction/links/02e7e51e420d000b52000000/Newer-Classification-and-Regression-Tree-Techniques-Bagging-and-Random-Forests-for-Ecological-Prediction.pdf Retrieved from http://therepresentationproject.org/captain-marvel-makes-film-herstory/

[9] Sharda, R., Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications(Volume 30, Issue 2).* Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.85.679&rep=rep1&type=pdf

[10] Yu, W., Liu, T., Valdez, R., Gwinn, M., Khoury, M., (2010). Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Medical Informatics and Decision Making (10:16).*