Predicting Darth Vader Favorability Using Supervised Learning

One day hackathon project using Star Wars dataset from fivethirtyeight.com

By Jon Godin

Background

- For this project, we had to locate a dataset that we wanted to work with from among the myriad datasets stored on fivethirtyeight.com's GitHub page, then clean and analyze the data, and present rough findings from our initial modeling, all to be completed within a six-hour window of time
- From the various datasets available, I selected a primary research study conducted among Star Wars fans, which included favorability ratings of many of the main characters from the Star Wars saga
- I thought it might be interesting to see whether I could classify respondents into those favorable towards Darth Vader vs. those who were not, so the primary modeling would be based on one or more supervised learning classification models
- Besides trying to optimize predictive model accuracy, I also wanted to gain some insights about what was more or less predictive of favorability towards Darth Vader

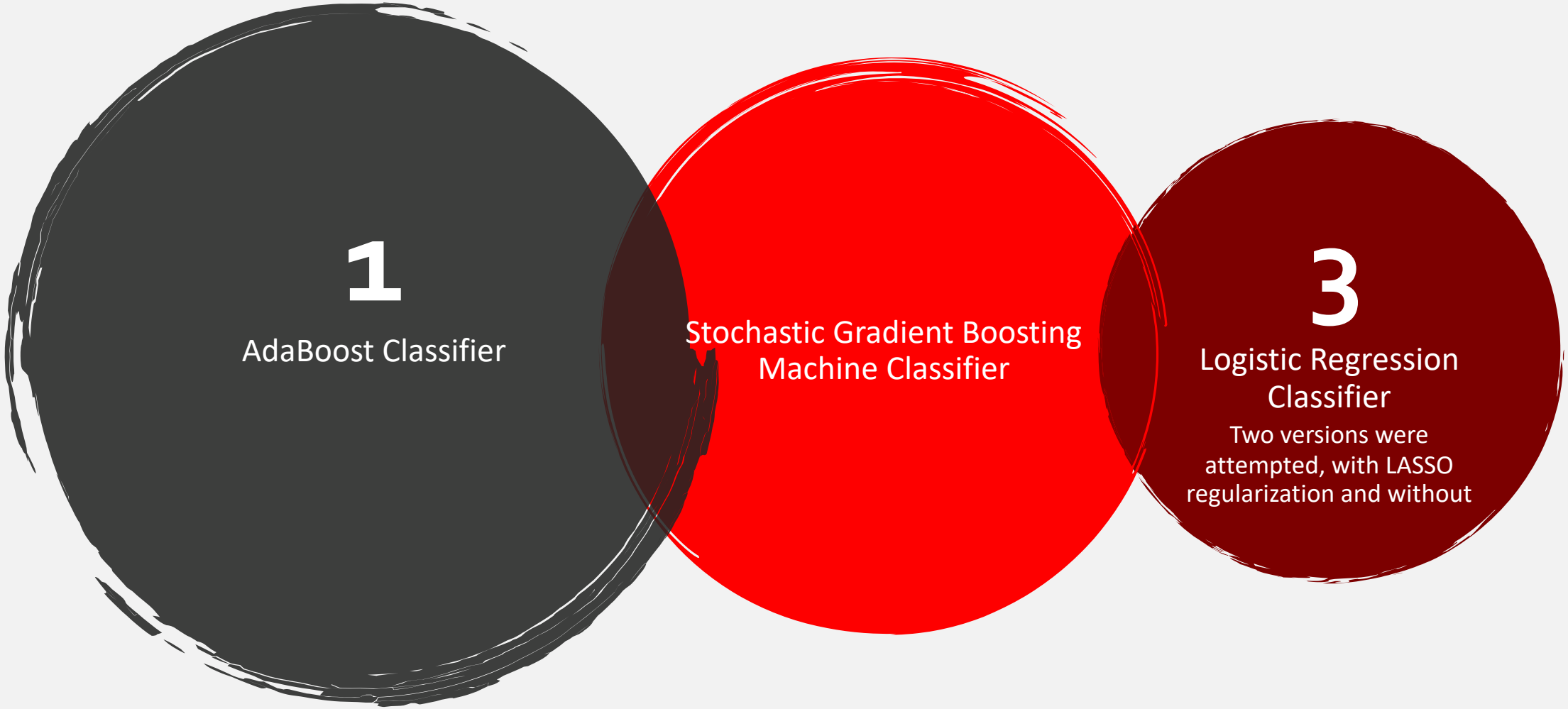
The Data

- fivethirtyeight.com's Star Wars survey included questions on Star Wars series viewership, a ranking of preferences among the first six episodes, favorability scores for a number of characters, opinions about the Star Wars expanded universe and Star Trek, plus some demographics
- Favorability scores were captured on a five-point, fully-labeled scale:
  - Very favorably
  - Somewhat favorably
  - Neither favorably nor unfavorably (neutral)
  - Somewhat unfavorably
  - Very unfavorably
  - A sixth option for Unfamiliar was also included
- The dependent variable (Darth Vader favorability) was recoded so that 1 = Very favorably, 0 = all else
- The full set of predictors is shown to the right

Star Wars Fan?	EP6 Rank	Padme Favorability
Saw EP1 Phantom Menace	Han Solo Favorability	Yoda Favorability
Saw EP2 Attack of the Clones	Luke Skywalker Favorability	Who Shot First, Han or Greedo?
Saw EP3 Revenge of the Sith	Princess Leia Favorability	Familiarity with expanded Star Wars universe
Saw EP4 A New Hope	Anakin Skywalker Favorability	Fan of expanded Star Wars universe
Saw EP5 The Empire Strikes Back	Obi-Wan Kenobi Favorability	Star Trek Fan
Saw EP6 Return of the Jedi	Emperor Palpatine Favorability	Gender
EP1 Rank	Lando Calrissian Favorability	Age
EP2 Rank	Boba Fett Favorability	Income
EP3 Rank	C3P0 Favorability	Education
EP4 Rank	R2D2 Favorability	Region
EP5 Rank	Jar-jar Binks Favorability	

Three types of supervised learning models were attempted

For each, the target variable was 0/1 Darth Vader Favorability Score



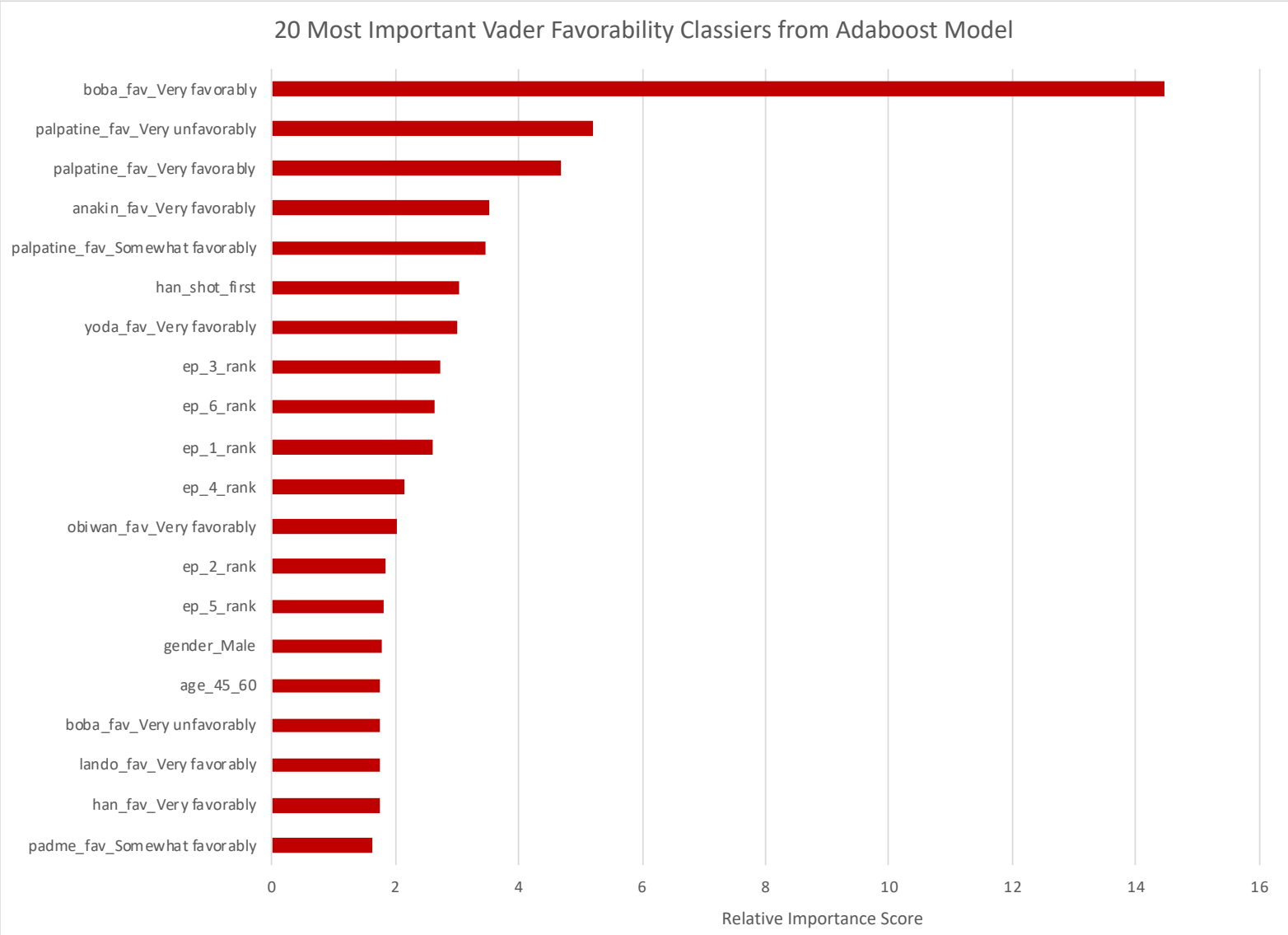
Model	CV Accuracy	Train Accuracy	Test Accuracy	Test Sensitivity	Test Specificity	Test Precision	Test F1 Score	Test ROC AUC Score
AdaBoost Machine	0.7842	0.9812	0.7410	0.5161	0.8734	0.7059	0.5963	0.7928
Stochastic Gradient Boosting Machine	0.7791	0.8442	0.7371	0.5161	0.8671	0.6957	0.5926	0.7927
Logistic Regression (w/LASSO)	0.7670	0.8134	0.7410	0.5269	0.8671	0.7000	0.6012	0.7893
Logistic Regression (no regularization)	0.7346	0.8527	0.7092	0.5591	0.7975	0.6190	0.5876	0.7334
Baseline Accuracy	0.6300							

Accuracy Improvement (best model) 24%

Model Results/Comparison

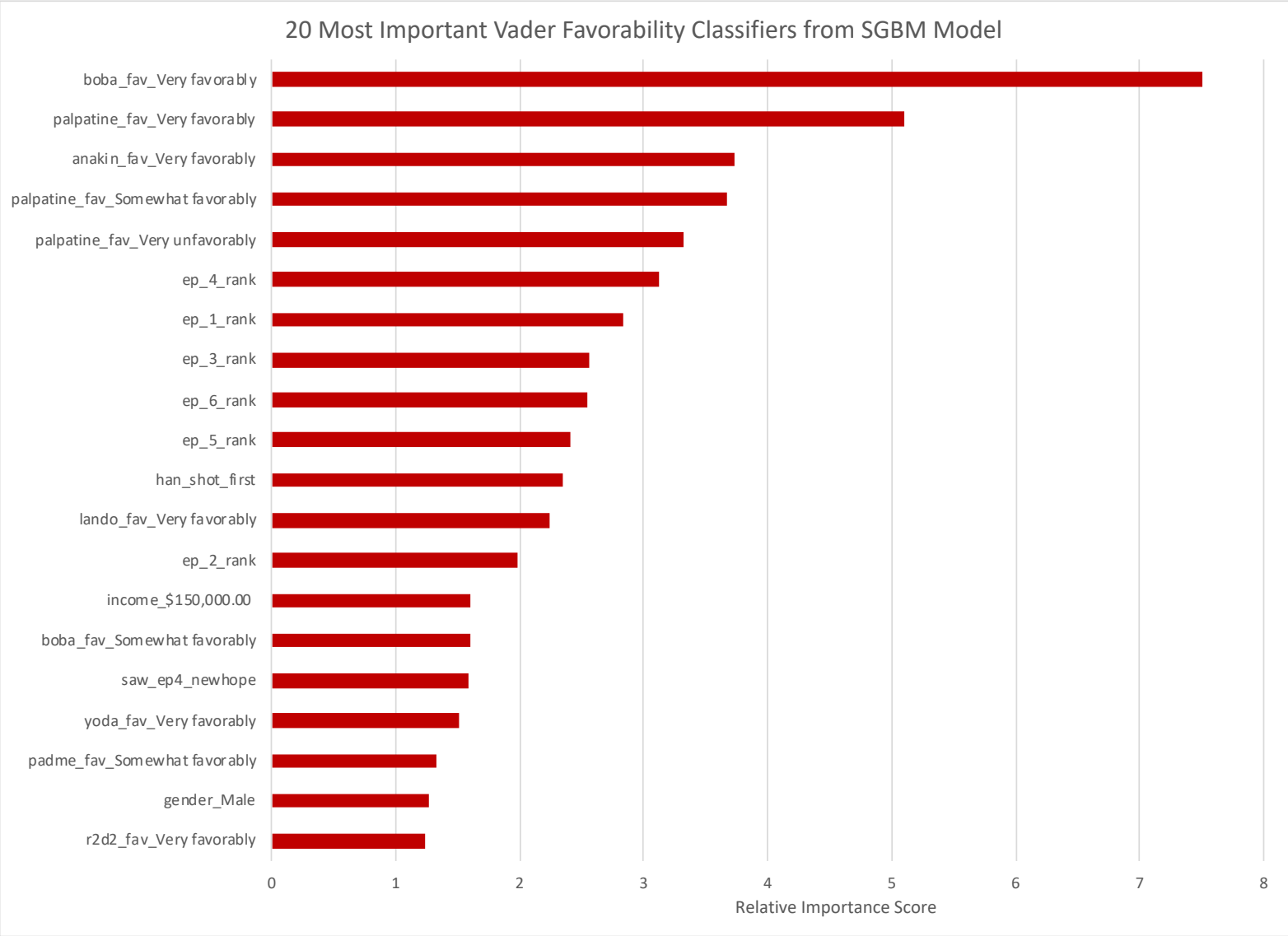
- All of the models attempted beat the baseline model accuracy, but all also show evidence of overfitting, with much higher training than testing or CV data accuracy; since this was a down-and-dirty hackathon project, I did the best I could under the time constraints we were working under
- That said, the AdaBoost model achieved the highest CV accuracy and Test ROC AUC score, narrowly edging the GBM model
- Note that all four models achieved very similar F1 Scores
- LASSO regularization did reduce overfitting vs. the unregularized logistic regression model, though it did not eliminate it completely

Adaboost Model Variable Importance Scores



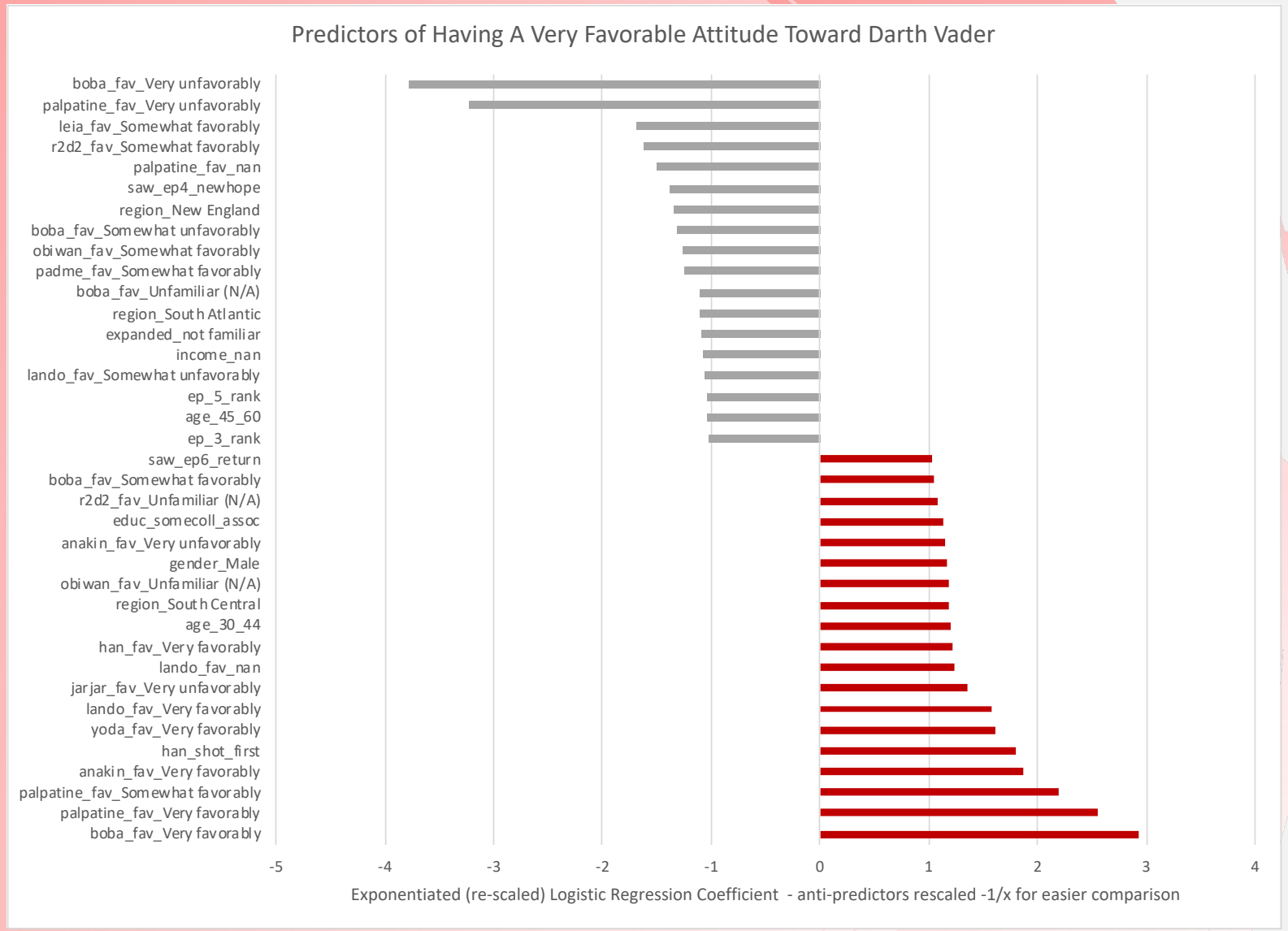
- Ignoring the valence of the predictions, the favorability (or lack thereof) of other “Dark Side” characters were most predictive of whether a respondent was favorable towards Darth Vader
- Most notably, opinions of Boba Fett were by far the most predictive
- Interestingly, those who believe Han Solo shot Greedo first in Episode IV were also more likely to have a favorable view of Darth Vader

Gradient Boosting Machine Model Variable Importance Scores



- In contrast to the Adaboost model, the GBM model had a little more balance across the predictors, though Boba Fett favorability was still the strongest predictor

Logistic Regression Model (w/LASSO) Coefficients



- Here, the valence of the coefficients can be taken into account, so unfavorable opinions of Boba Fett and Emperor Palpatine, along with favorable opinions of Princess Leia and R2D2, are indicative of low Darth Vader favorability
- Han Shot Firsters are also more indicative of Darth Vader favorability in this model