Good movie prediction

Data Science

Simon Picard

Predicting good movies

Using IMDB data, the goal is to predict movies with an average rating above 7.5, i.e. a good movie.

In this presentation, you will find the key takeaways of the exercise including model performance, insights and suggested next steps.

For the modeling approach and the choices made in the process, please refer to the markdown and comments in the notebook.

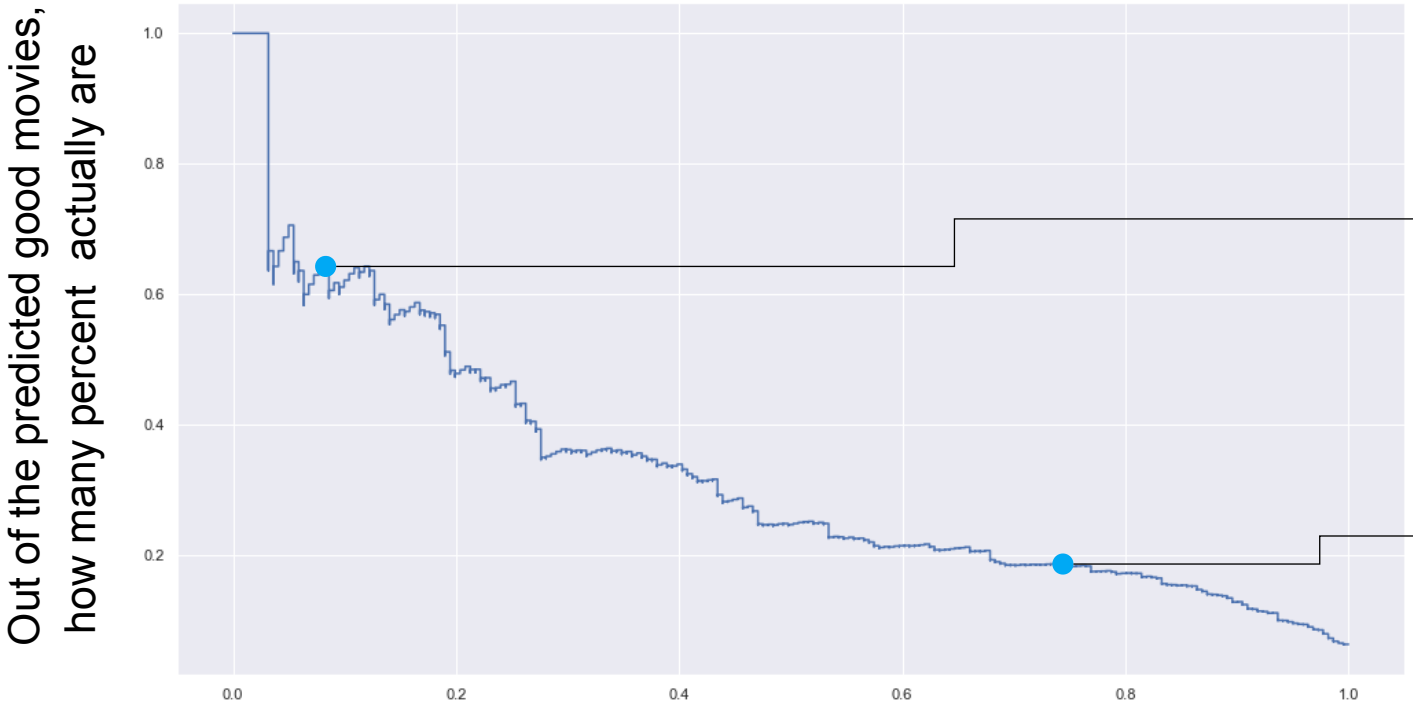
Business

Technical

Selecting the prediction threshold depends on the business use case

Precision-Recall curve

Percent over percent

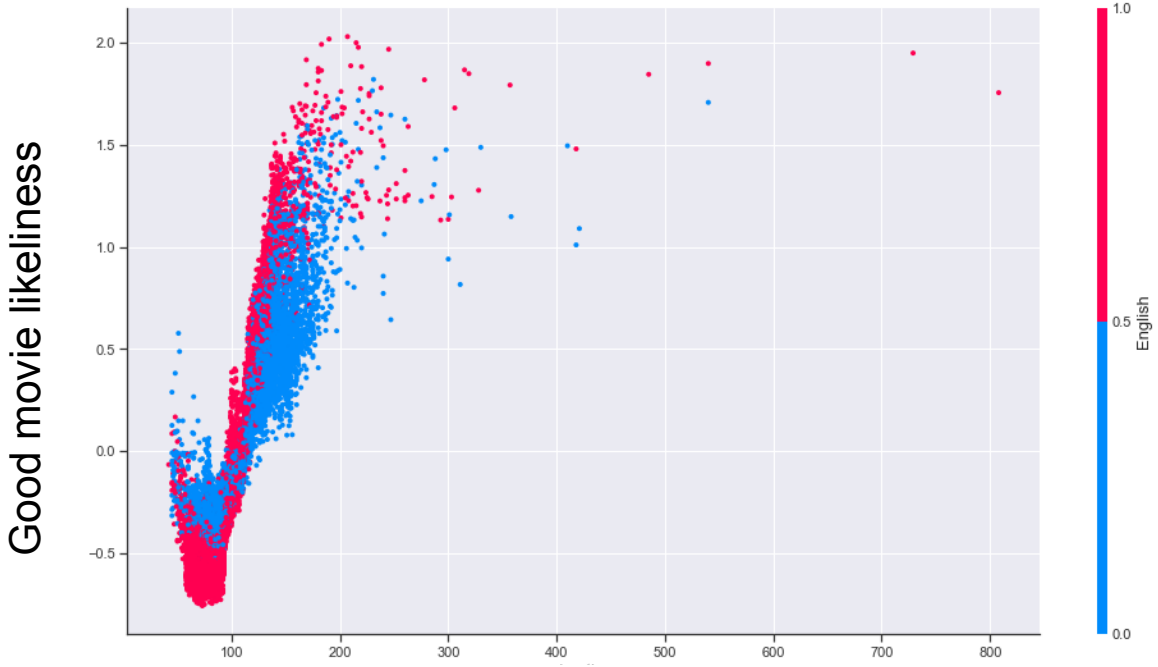


Selecting the right prediction threshold will allow to tailor the good movie model to the use case:

- A more confident prediction, 65% of correct predictions for 15% of the good movies, could lead to recommending to announcers on which movies to broadcast ads at the theater
- A more general prediction, 20% of correct predictions for 75% of the good movies, could be used to recommend next movies to discover for a media platform broadcaster

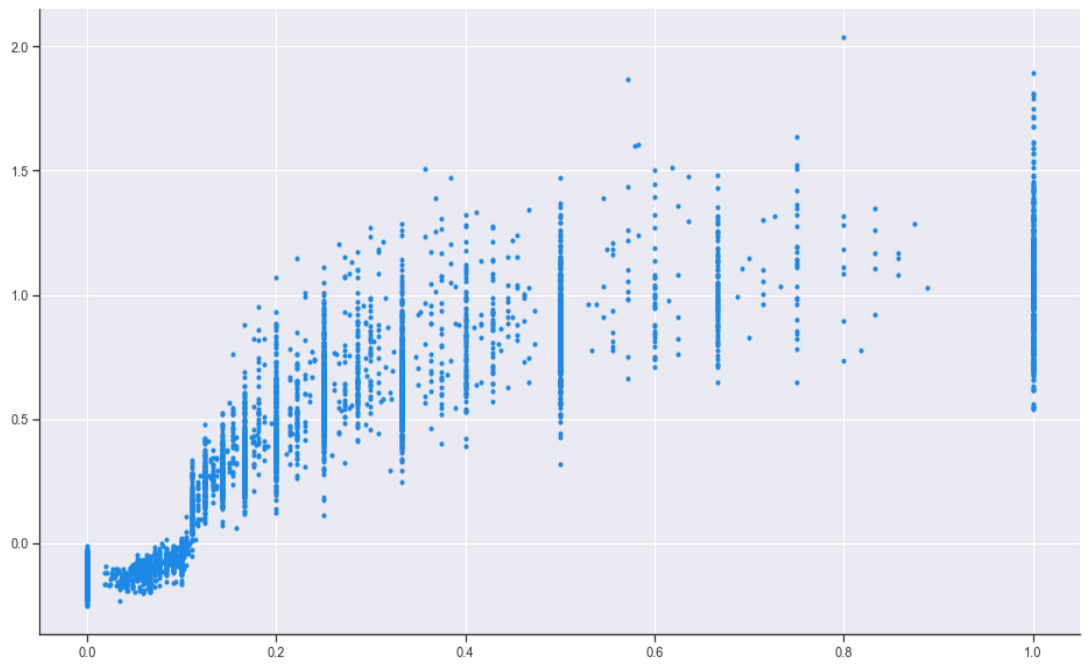
The golden movie recipe is to have an English, 2 hours+ movie directed by someone renowned

Shapely values allow to understand what are driving the good movie prediction



Movie duration

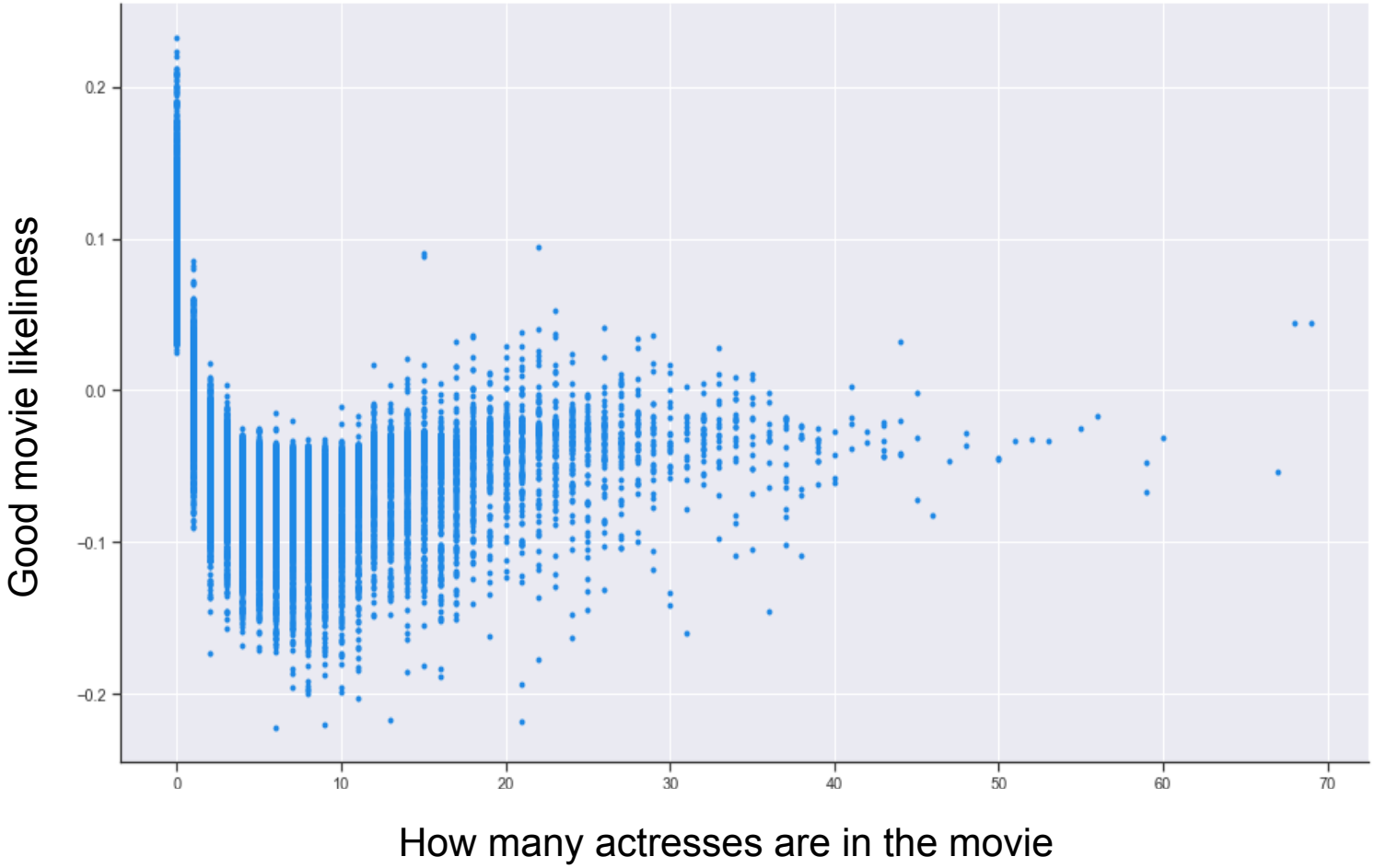
A longer movie is more likely to be a good one, especially if it is in English



Director's previous movies ratings average

The better the director's previous movies, the better his/her next movie is going to be

Ensuring fairness in the model is required to productionize it



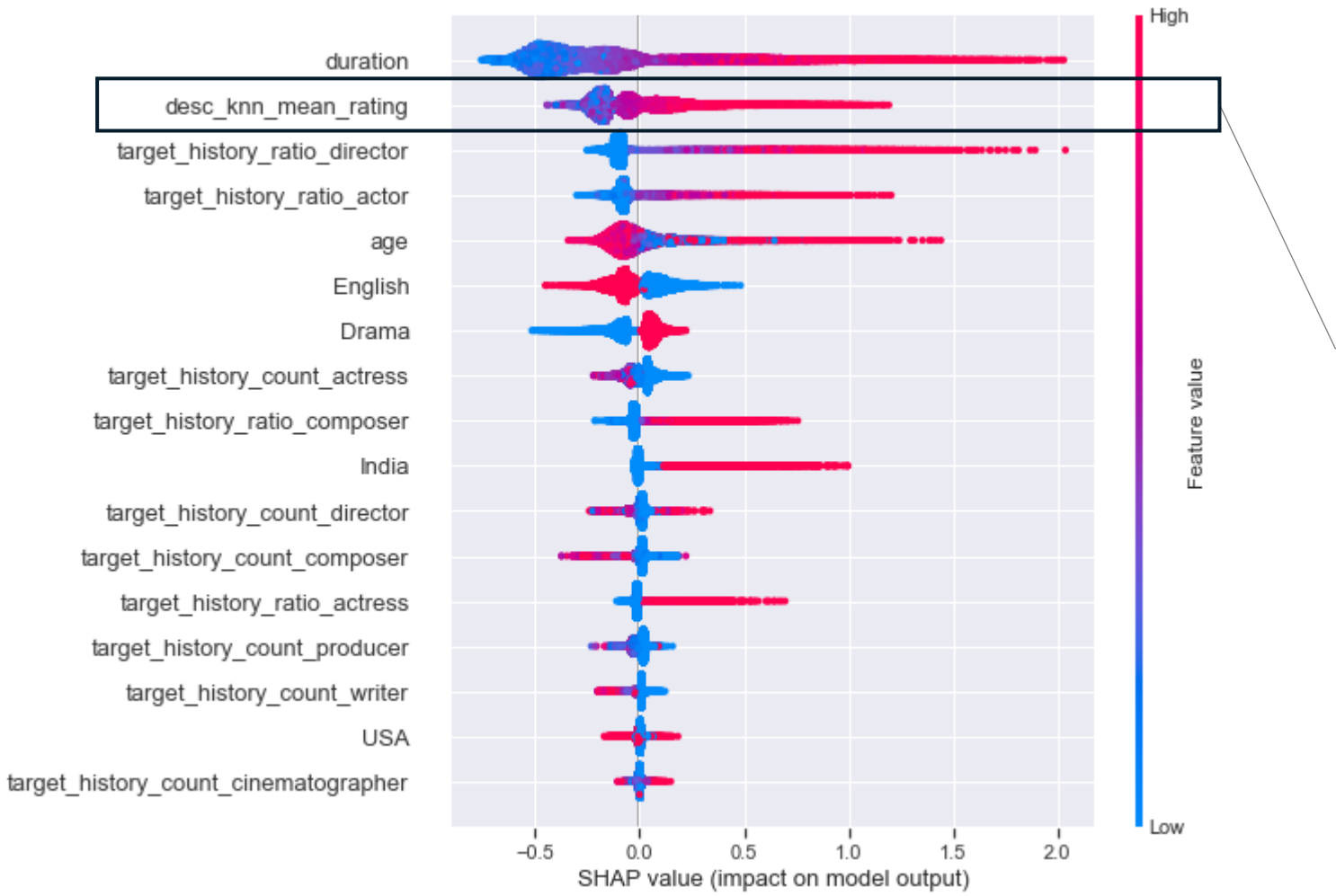
In its current state, the model will predict that a movie with one or more actresses is less likely be a good one.

Although such pattern has been found in the data, it is relevant to prevent it in order to avoid discrimination in the model, eventually guiding future decisions.

Different methods exist to prevent this behavior, such as the *Intersectional Fairness Framework*[1]

1. Morina, Giulio, et al. "Auditing and Achieving Intersectional Fairness in Classification Problems." arXiv preprint arXiv:1911.01468 (2019).

The movie content is a strong predictor of its future rating



The movie description

Using word embeddings, it is possible to find older movies having a similar description to the one being predicted.

The average rating of those similar movies turns out to be the top two feature of the model.

Such result calls for more investigations in the NLP landscape:

1. Understand which intent are driving the good ratings, allowing to discover what topics are leading to good movies
2. Explore other text data, such as the movie subtitles

Good movie prediction next steps

1

Define the use case of the good movie prediction, in order to tailor the model to the need

2

Improve the model performance by:

a. integrate more data sources (e.g. actor data, reward data, tv show data)
b. compute new features (movie crew network, subtitles NLP, image recognition)

3

Productionize the model for reproducibility deployment, using a pipeline framework such as Kedro[1]

1. https://kedro.readthedocs.io/en/stable/index.html