Prepared by:

Data science consultant

Yuriy Gajev

Subject:

The goal of the project is to create a model that can predict the success of a TV show based on information about it: genre, year of production, director, actors, number of episodes, votes etc.

It is also necessary to identify the characteristics of the show, that have the greatest impact on success.

To this end, work was done on the collection and processing of information. More than 5,000 TV shows were analyzed. The collected data is presented in a wide range: 26 genres, year of issue between 1950 - 2017, rating from 1.5 to 9.9.

Information was collected from two sources: imdb.com and tvmaze.com. Work was done on cleaning and preparing information for further analysis. The following techniques were used: NLP, ensemble methods and decision tree models, models to search over specified parameters.

The modeling goal was to predict whether a show would be above or below 8.0 average IMDB rating. After setting that threshold, our shows broke down in 70% below and 30% above.

On the basis of the final data, a large number of models were used, from which five were selected based on the best result. The best model showed a result of 76% accuracy score, in comparison with baseline score 70%.

Using the results of five models, ten features were identified that most significantly influence the success of a TV show. Among them: number of votes, number of episodes, total reviews, genres: documentary, comedy and animation.

Next steps:

Build a model that can predict tv show success using only leading characteristics (*for example: summary of the show*) of the show which are independent of lagging Indicators (*votes, reviews and likes*).