Machine Learning Mina Milanovic 74308

Mini Project 3 Report

1. Motivation and goal

In this project, I will use two predictive models, logistic regression and SVM to build a model which will try to predict will the movie, which is about to be released, be successful. I will use a database which I found on the internet (database is without copyrights) which contains 28 columns describing features of the movies, and has more than 5000 rows.

2. Approach

First, I've deleted all the empty cells from the database. Then I started analyzing the columns and the data I have in the database, so I can exclude all the unnecessary columns.

I will use the data for the years 2000 to 2014 to make a model, and years 2015 and 2016 for the test data.

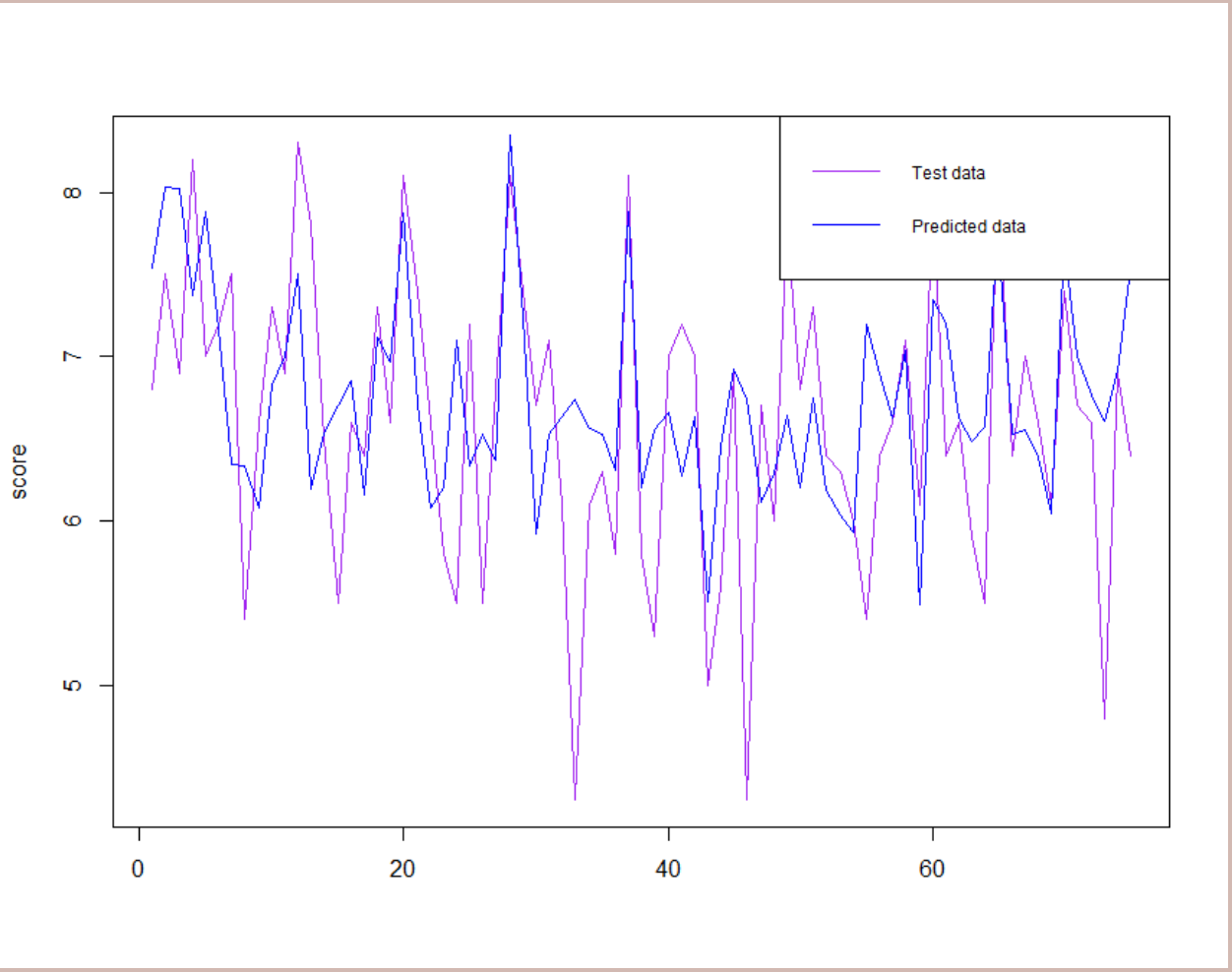
I will predict will the movie be successful with imdb score, counting that the movie is successful if the imdb score is more than 6.

I will exclude columns containing names, because they have usually 1-2% and they make the model slower. Also, columns country, color, genres, imdb link, movie title and number of the faces in the poster shows to be irrelevant when plotted. Also, some other data resulted to be useless while testing the models.

The most relevant information resulted to be duration, num_critic_for_reviews, budget, cast_total_facebook_likes and movie_facebook_likes, so I used them for my model.

For my logistic regression model, I have calculated the means and the auc value, which showed respectfully, 81% and 70%, which means that the model is very good.

For the SVM model, I have calculated the root mean square deviation, which is 1.2 for my model, meaning it is very accurate. Also, I have made a plot to compare some of the predicted results with the actual test data, showing that in the most of the cases it predicts good, it struggles only with the cases of the movies that are really bad, but that happens in the real life predictions also.



3. Discussion

The hardest part of this project was formatting the database, so it can be used and the problem can be solved. Both of the models were easy to implement, once the data is analyzed so the formula can be made.

Logistic regression model is probabilistic, it creates calibrated probabilities and allows smooth objective. The farther data lays from the separating hyperplane the better; it maximizes the probability of data.

SVM model is good for the generalization; it's deterministic and faster than LR because it only uses support vectors. It tries to find the separating hyperplane that maximizes the distance of the closest points to the margin.

4. Running the code

For this project, I've used R-Studio. The use is very simple, there are two scripts in the project, which the user can open using the standard environment procedure (File-Open File) or just make a new R script and then copy the code. Database should be stored in the same folder where the script is (if the first method is used), or added into R-Studio using Import Dataset on the right side of the R-Studio.

The libraries used are e1071 for SVM, and caret for plotting. Every library that is reported missing can be installed using install.packages('name_of_the_library').

The code is simply run with the R Studio R command. It runs one command at the time.