# Data Science and Machine Learning in mathematics education: Highschool students working on the Netflix Prize

Sarah Schönbrodt, Martin Frank

# Data Science and Machine Learning in mathematics education: High-school students working on the Netflix Prize

Sarah Schönbrodt[1] and Martin Frank[2]

[1]Karlsruhe Institute of Technology, Germany; sarah.schoenbrodt@kit.edu

[2]Karlsruhe Institute of Technology, Germany; martin.frank@kit.edu

*One goal of contemporary mathematical modeling classes in schools should be to include up-to-date problems or interesting, new technologies from the everyday life of students - especially if these allow the didactical reduction to elementary (school-)mathematical knowledge and thus have the potential to enrich mathematics education. Data Science and Machine Learning is applied in numerous areas of science and technology and used in many applications in our everyday life. Using movie recommender systems and the so-called Netflix Prize as an example, this paper discusses how mathematics education can be enriched by modeling real-world, student-centered problems from the field of Machine Learning in school. For this purpose, we describe tested digital learning material from guided modeling projects and share our experience with giving the problem of developing a recommender system as a completely open problem to upper secondary students.*

*Keywords: Data Science, Machine Learning, mathematical modeling, recommender system, digital learning material.*

## Motivation

Many technologies and applications that students use in their everyday life are based on methods from the fields of Data Science (DS) and Machine Learning (ML). An important tool for answering questions from these areas is (mathematical) modeling. Therefore, DS problems offer a great opportunity to design mathematical modeling activities on student-oriented, up-to-date problems. In this way, not only modeling competencies but also the handling of data can and should be trained. Especially in today's world, it is essential to deal with data in an understanding, responsible and critical way. Some profitable approaches to implement problems from the field of DS and ML in the classroom are advanced by the Paderborn project ProDaBi (Opel et al., 2019). In addition, Sube (2019) and Schönbrodt et al. (2021) already worked out how mathematical modeling lessons on DS questions can be designed using real-world problems.

We developed digital learning material in which upper secondary students can acquire an understanding of essential steps in solving data-heavy problems based on methods from DS and ML by actively working on an authentic, and relevant problem (Schönbrodt & Frank, 2021). The learning material presented in the following can be implemented within a mathematical modeling day or be distributed over several school lessons.

## The problem – Authentic and relevant

Netflix, Amazon, and many other e-commerce companies that the students encounter every day mainly rely on one thing for customer loyalty: personalized recommendations for new products, movies, etc. For this purpose, recommender systems are developed, which should predict what the respective user might like. The developed learning material is based on the Netflix Prize which

Netflix launched in 2006 to further improve their own recommendation system: The team, that could predict at least 10% more accurately than Netflix's system, which movies a user would like had a chance to win the grand prize of one million USD (Feuerverger et al., 2012, p. 203).

## Mathematical modeling days – Guided digital learning

The key questions of the developed learning material are: "How can we model user preferences by exploiting given user ratings?" and "How can we predict unknown user ratings in the best possible way to then use the predicted ratings to suggest relevant new movies to users?" (Schönbrodt et al., 2021).

In the learning material the students work with the original dataset of the Netflix Prize. On digital worksheets they explore the dataset and collect their own ideas for the development of a recommender system. Afterwards, they develop a mathematical model and apply it to the Netflix dataset. The implementation of the data exploration and the modeling process is discussed in more detail below. The digital tool used to develop the digital worksheets is Jupyter Notebook[1]. Jupyter Notebooks are widely used in numerous branches of industry, research, and economy. Thus, not only the learning content but also the digital tool provides an authentic insight into current problem-solving strategies and technical implementations in applied mathematics or more generally in STEM fields. Put in simple terms, Jupyter Notebooks are digital documents which can be used to write text, execute code (to compute mathematical terms), and can be edited interactively by the students. Also, Biehler & Fleischer (2021) and Opel et al. (2019) highlight the possibilities of Jupyter Notebooks for designing learning material on ML methods. In these papers the focus is stronger on statistics and computer science teaching whereas we focus more on the process of mathematical modeling and related competencies.

The developed learning material exemplifies how data-heavy problems from the students' everyday life can be prepared and carried out in the context of modeling projects, both in distance and face-to-face learning. The material is available on a cloud-platform of the Karlsruhe Institute of Technology for direct use in class and can both be accessed and edited in a web browser (see www.cammp.online/english/214.php). The material can be used in heterogeneous learning groups by means of digital differentiation material (scaffolded tips, optional tasks for more advanced students) and individual, automated feedback on the students' solutions. The dataset and the main modeling steps of the learning material are briefly described below.

For the processing of the learning material, it is assumed that the students are familiar with mean values and standard deviations. In addition, an understanding of functional relationships is also central. Knowledge of vectors and the scalar product is preferable. Matrices, however, are not a prerequisite for the comprehensible processing of the learning material. They are rather introduced in a problem-oriented manner as explained below. Furthermore, no programming knowledge is required on the part of the students. Instead, a fill in the gap approach is implemented. The students simply need to replace a placeholder within the code (specifically *NaN*, which stands for Not a Number) with

[1] For further information see https://jupyter.org, last accessed: 05 August 2021

the relevant solution to the task (a number, formula, or function). They receive task-specific feedback on their input and, if necessary, a tip on how to correct their solution.

**Unit 1: Understanding and analyzing the data**

The starting point of the learning material is the Netflix dataset. This consists of 17,700 movies, 480,189 (anonymized) users, and 100,480,507 ratings from one (worst) to five (best) that users gave to the movies (Feuerverger et al., 2012, p. 204). The title and release year of the movies are known.

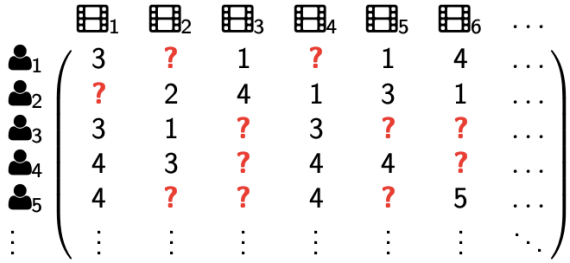
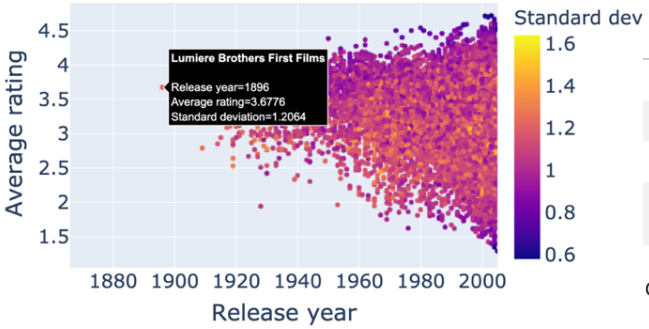


**Figure 1: Rating matrix / table containing users' ratings for the movies**

The students examine the dataset using various forms of representation such as the rating matrix / table (see Figure 1) or interactive scatter diagrams and tables (see Figure 2). In doing so, they answer questions such as “For which movie do the ratings differ the least from the mean rating?” or “Which conclusions can you make regarding the distribution of the data over time?”. The students can modify the interactive plots shown on the worksheets to answer the questions. For example, they can zoom into the scatterplot or sort the table in Figure 2 by any column. They decide independently which representation of the data is suitable for answering which questions.




| Titles | Number of ratings | Average rating | Standard deviation | Release year |
|---|---|---|---|---|
| Miss Congeniality | 232944 | 3.3613 | 1.1117 | 2000 |
| Independence Day | 216596 | 3.7242 | 1.0391 | 1996 |
| The Patriot | 200832 | 3.7839 | 1.097 | 2000 |
| The Day After Tomorrow | 196397 | 3.4422 | 1.0993 | 2004 |
| Pirates of the Caribbean: The Curse of the Bla... | 193941 | 4.1539 | 0.9071 | 2003 |

**Figure 2: Visualizations of the Netflix dataset on the digital worksheets**

When introducing the tabular representation of the rating data (see Figure 1), the concept of a matrix is avoided at first, as this should not be a prerequisite for the processing of the learning material.

**Unit 2: Developing a mathematical model**

After exploring the data, a brainstorming takes place. Whilst discussing with their fellow students the students collect their own ideas on how to use the known ratings in the rating table to predict the unknown ones. During the lessons conducted with secondary students so far, the students mentioned diverse approaches. Among them:

- Find similar users and then find the movies the similar users liked.

- Identify users' areas of interest, such as a specific genre.
- Determine the genre of the movies a user liked. Then suggest movies from that genre.

Next, using small rating tables, the students start with the development of a selected mathematical model: the decomposition of the rating table into a user-feature table (short user table) and a movie-feature table (short movie table, see Figure 3). The core of the described model is a matrix factorization. This was also the basis of the model used by the winning team of the Netflix Prize (Koren et al., 2009, p. 32). Thus, not only the problem but also the mathematical methods used are authentic (Vos, 2011). The user table indicates how much a user likes certain properties, such as the action or comedy genres. The movie table specifies the extent to which a film contains these features (see Figure 3). The students independently develop a formula to calculate the known ratings from the rows of the user table and the columns of the movie table and transfer this to the calculation of unknown ratings. This is basically achieved by the scalar product of the row vector of the user table and the column vector of the movie table.

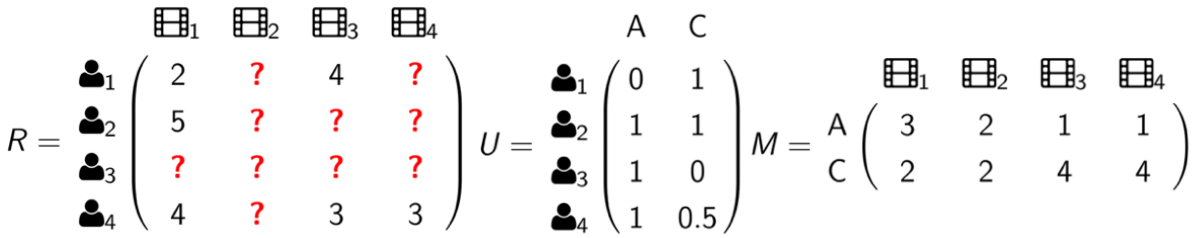



**Figure 3: Exemplary rating table R and corresponding user table U and movie table M. The user and movie tables reflect the interests of the users and the characteristics of the movies regarding the features "action" (A) and "comedy" (C)**

Motivated by the fact that the Netflix dataset does not provide any information on the features of the movies and thus both the user and the movie table must be calculated appropriately, the students should first determine a suitable decomposition for a 2x2 rating table themselves using pen and paper.

*Note:* Another modeling approach for which learning material for guided modeling projects was developed and tested is the modeling of similarities - either between users or between movies. Students can be creative in their choice of similarity measures, but also learn about common similarity measures such as (adjusted) cosine similarity and Pearson correlation (Sarwar et al., 2001, p. 287). This approach falls in the class of so-called neighborhood methods.

**Unit 3: Error measure and optimization**

Calculating a decomposition "by hand" is not feasible for large rating matrices. Therefore, the goal is to leave the calculation to the computer. For this purpose, a step-by-step optimization procedure is applied, which provides a sufficiently good decomposition. The students first define an error measure by which they (and later the computer) can evaluate whether a found decomposition is already sufficiently good. The error measure is defined by looking at a small example of a given rating table R and predicted ratings given in the table P (see Figure 4).

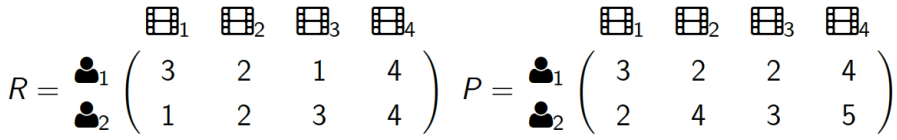


**Figure 4: Exemplary rating table R and table P containing the predicted ratings**

The self-defined error measures are then compared with the mean absolute error and the mean squared error. Then an alternating optimization algorithm is used to compute a decomposition that leads to the smallest possible error on the known data. The algorithm does not have to be developed by the students themselves. Instead, an optimization package of the used programming language, in our case Julia (Python would also be possible), is applied as a black box.

**Unit 4: Application to the Netflix dataset and critical discussion**

So far, we have only evaluated how well a found decomposition is suited to represent known ratings, but unknown ratings are to be predicted. We apply an essential strategy of supervised ML: Dividing the known rating data into those used to compute a decomposition (training data) and those used to afterwards validate how well the prediction works on "unknown" data (test data). Students apply this strategy to the Netflix dataset. Finally, they conduct a parameter study (varying the number of features considered) and try to improve the results on the test data.

In a final discussion, limitations of the developed model are discussed, such as: "How do we deal with new users who have not yet submitted ratings?". A critical discussion on the possibilities of manipulating recommender systems, the problems that could arise from de-anonymizing[2] users, and the extent to which so-called filter bubbles could be problematic takes place. The opinions and the students' experiences with recommender systems are particularly included in this discussion.

## Experiences with students

The digital learning material was already implemented with more than 90 students from Grade 10 to Grade 13 (this corresponds to an age range of approx. 15 to 19 years) within school lessons (4-5 lessons 90 minutes each) or in the framework of a mathematical modeling day (approx. 5 hours) in Germany. During the implementations, the diverse ideas of the students for the development of a recommender system but also the numerous arguments in the critical reflection of such systems stood out. The very lively discussions also showed that the students were highly interested in solving the problem.

While working with the learning material numerous school mathematical contents are used; among them mean values, standard deviation, vectors, scalar product, and functions. Therefore, the problem setting is not only extremely student-centered but especially real and authentic and thus provides an answer to the question "What's the point of math?".

[2] In 2008, two researchers at the University of Texas showed that de-anonymization of the dataset was partially possible by combining it with another publicly available film dataset (Narayanan & Shmatikov, 2006).

## Open modeling within a mathematical modeling week

The Netflix Challenge was given to a group of six upper secondary students within a so-called modeling week. The task of the students was to develop a movie recommender system for a smaller subset of the Netflix dataset from scratch. The students developed the recommender system and thus the mathematical model all by themselves. In contrast to the more guided modeling days described above within the modeling week the students did not receive any learning material with pre-structured subtasks. They only received a description of the datasets and the problem as well as a short introduction in using Jupyter Notebook. The students implemented code themselves and filled the Notebooks with content on their own. They worked on the problem for four full days and were supported by a scientific advisor. The advisor only helped with programming related questions but did not intervene in the modeling process of the students at all. The students not only developed a recommender system but also documented their modeling process in a report and gave a presentation on their results at the end of the modeling week. The core idea of the model the students developed was to measure similarities between users by computing the mean absolute deviation of the ratings of different users. Once they determined the similarity measure and computed the similarities between any pair of users the students predicted the rating of a specific user by taking a weighted sum of the ratings of the most similar users to the selected one. The model developed by the students without any guidance, can be linked to the class of neighborhood methods mentioned above. These are methods that are used by experts in this field (Sarwar, 2001, p. 285).

During the modeling process, the students discovered a relevant trade-off in the field of ML:

“Furthermore, we concluded that the more data we use in the algorithm, the more accurate our result will be. One problem that arose was the long, very rapidly increasing, computation time that the program needs to get results with the large amount of data” (quote from the students’ report).

## Summary and Outlook

Through the learning material, secondary students collect experience in dealing with large amounts of data and gain an active insight into essential strategies of mathematical modeling and ML. We already implemented the learning material on the Netflix Challenge with numerous learning groups as part of stronger guided modeling projects (modeling days). Jupyter Notebooks were used as a digital tool to structure and guide the modeling process through smaller subtasks. The feedback from the students on the modeling activities, which were so far only carried out virtually, showed that there is a great interest in this type of problem.

On top, the Netflix Challenge was posed as a problem in the context of an open modeling project (modeling week) once. The results of the students at the end of the modeling week showed that they were indeed able to work independently in a small team and to develop and validate a recommender system on a real-world dataset. To make a more general statement about the extent to which this problem is suitable for open modeling projects, further testing with a more heterogeneous student group would be necessary. Here, the participating students were interested in mathematics and some of them already had programming skills.

The tool Jupyter Notebook provides different design possibilities, variability of the programming language and the neat combination of code and text in one document. Nevertheless, it would also be feasible to provide the data (or a part of the data) in the form of CSV files and to have the neighborhood methods described above developed with the help of a table calculation program.

In addition to the problem presented here, various other applications based on ML methods offer a good opportunity to design modeling activities on real and up-to-date problems. The following are examples of possible ML methods which can be reduced to school-mathematical concepts: The so-called support vector machine (SVM), which is widely used for classification problems. The SVM is based on minimizing distances between points and hyperplanes. Vectors, straight lines, and planes as well as the scalar product can find an authentic application (Schönbrodt et al., 2021). Starting from a classification problem with two- or three-dimensional data, the method could first be developed with students in the visual case using rich visualizations implemented in Jupyter Notebooks. After that the method can be abstracted and applied to higher dimensional problems (e. g., for the classification of images with faces). Another example is represented by so-called n-gram models, which are used to predict words when typing a message on a smartphone. In this case conditional probabilities and absolute and relative frequencies play a central role. Finally, Neural Networks, which are the basis for countless applications in our everyday life (Schlichtig et al., 2019), rely on various school mathematics contents such as the chain rule, vectors, and the scalar product.

It seems important to prepare a selection of these methods to give meaning to (school) mathematical content and to offer students an insight into methods and problem-solving strategies that increasingly influence technologies and our society now and in the future. Therefore, the development of a curriculum that precisely addresses these kinds of methods and meaningfully combines content from mathematics and computer science education is necessary. Important steps in this direction are already accomplished by the projects ProDaBi and the International Data Science in Schools Project (Biehler et al., 2018; IDSSP, 2019). These projects as well as the learning material on the Netflix Prize presented in this paper underline that addressing problems from DS and ML with high-school students is feasible and necessary.

## References


Biehler, R., & Schulte, C. (2018). Perspectives for an interdisciplinary data science curriculum at German secondary schools. In R. Biehler, L. Budde, D. Frischemeier, B. Heinemann, S. Podworny, C. Schulte & T. Wassong (Eds.), *Paderborn Symposium on Data Science Education at School Level* 2017: *The Collected Extended Abstracts* (pp. 2–14). Paderborn University. http://doi.org/10.17619/UNIPB/1-374

Feuerverger, A., He, Y., & Khatri, S. (2012). Statistical Significance of the Netflix Challenge. *Statistical Science*, *27*(2), 202–231. https://doi.org/10.1214/11-STS368

Biehler, R., & Fleischer, Y. (2021). Introducing students to machine learning with decision trees using CODAP and Jupyter Notebooks. *Teaching Statistics, 43, 133*–142.

IDSSP Curriculum Team (2019). *Curriculum Frameworks for Introductory Data Science* (Framework). http://idssp.org/files/IDSSP_Frameworks_1.0.pdf

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. *IEEE Computer Society Press*, *42*(8), 30–37. https://doi.org/10.1109/MC.2009.263

Narayanan, A., & Shmatikov, V. (2006). How To Break Anonymity of the Netflix Prize Dataset. The University of Texas at Austin. https://arxiv.org/abs/cs/0610105

Opel, S., Schlichtig, M., Schulte, C., Biehler, R., Frischemeier, D., Podworny, S., & Wassong, T. (2019). Entwicklung und Reflexion einer Unterrichtssequenz zum Maschinellen Lernen als Aspekt von Data Science in der Sekundarstufe II [Development and reflection of a teaching sequence on machine learning as an aspect of data science in upper secondary education]. In A. Pasternak (Ed.), *Proceedings zur 18. GI-Fachtagung Informatik und Schule* (pp. 285–294). Gesellschaft für Informatik. https://doi.org/10.18420/infos2019-c14

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In V. Y. Shen (Ed.), *Proceedings of 10th International Conference on the World Wide Web* (pp. 285–295). ACM. https://doi.org/10.1145/371920.372071

Schlichtig, M., Opel, S., Schulte, C., Biehler, R., Frischemeier, D., Podworny, S., & Wassong, T. (2019). Maschinelles Lernen im Unterricht mit Jupyter Notebook [Machine learning in the classroom with Jupyter Notebook]. In A. Pasternak (Ed.), *Proceedings zur 18. GI-Fachtagung Informatik und Schule* (p. 385). Gesellschaft für Informatik. http://doi.org/10.18420/infos2019

Schönbrodt, S. (2019). *Maschinelle Lernmethoden für Klassifizierungsprobleme: Perspektiven für die mathematische Modellierung mit Schülerinnen und Schülern* [Machine learning methods for classification problems: Perspectives for mathematical modeling with high-school students]. Springer Spektrum. https://doi.org/10.1007/978-3-658-25137-6

Schönbrodt, S., & Frank, M. (2021). Digitales Lernmaterial zur Netflix Challenge [Digital learning material on the Netflix challenge]. In K. Hein, C. Heil, S. Ruwisch & S. Prediger (Eds.), *Beiträge zum Mathematikunterricht 2021* (pp. 299–303). WTM-Verlag. http://dx.doi.org/10.17877/DE290R-22328

Schönbrodt, S., Camminady, T., & Frank, M. (2021). Mathematische Grundlagen der Künstlichen Intelligenz im Schulunterricht: Chancen für eine Bereicherung des Unterrichts in linearer Algebra [Mathematical foundations of artificial intelligence in school lessons: Opportunities for enriching the teaching of linear algebra]. *Mathematische Semesterberichte*. Springer. https://doi.org/10.1007/s00591-021-00310-x

Sube, M. (2019). *Entwicklung und Evaluation von Unterrichtsmaterial zu Data Science und mathematischer Modellierung mit Schülerinnen und Schülern* [Development and evaluation of teaching materials on data science and mathematical modeling with students] [Doctoral dissertation, RWTH Aachen University], Universitätsbibliothek RWTH Aachen.

Vos, P. (2011). What is "authentic" in the teaching and learning of mathematical modelling? In G. Kaiser, W. Blum, R. Borromeo Ferri & G. Stillman (Eds.), *Trends in Teaching and Learning of Mathematical Modelling (ICTMA14)* (pp. 713–722). Springer. https://doi.org/10.1007/978-94-007-0910-2_68