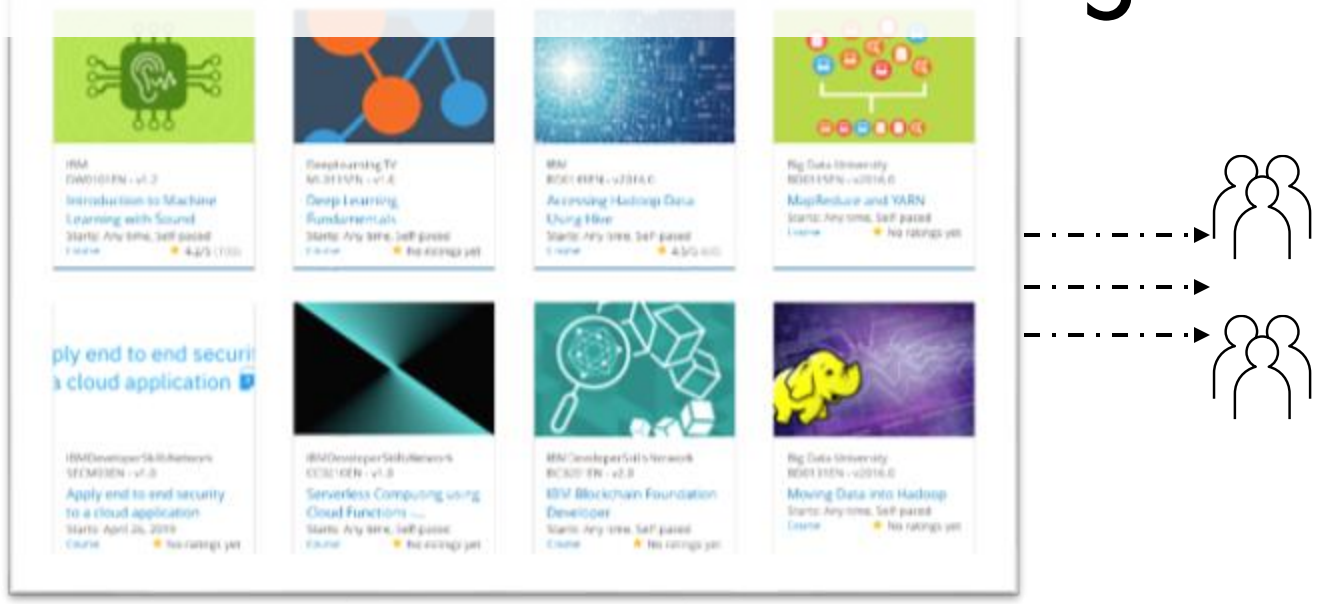
Build a Personalized Online Course Recommender System with Machine Learning

Anthony R. Poggioli

19 September 2024



Outline

- Introduction and Background
- Exploratory Data Analysis
- Content-based Recommender System using Unsupervised Learning
- Collaborative-filtering based Recommender System using Supervised learning
- Conclusion
- Appendix

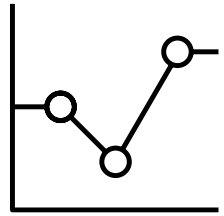
Introduction

AI Training Room is a startup offering massive open online courses (MOOCs) in AI, data science, machine learning, cloud development, et cetera. This firm would like to enhance growth, engagement, and learner satisfaction by developing a course recommendation system.

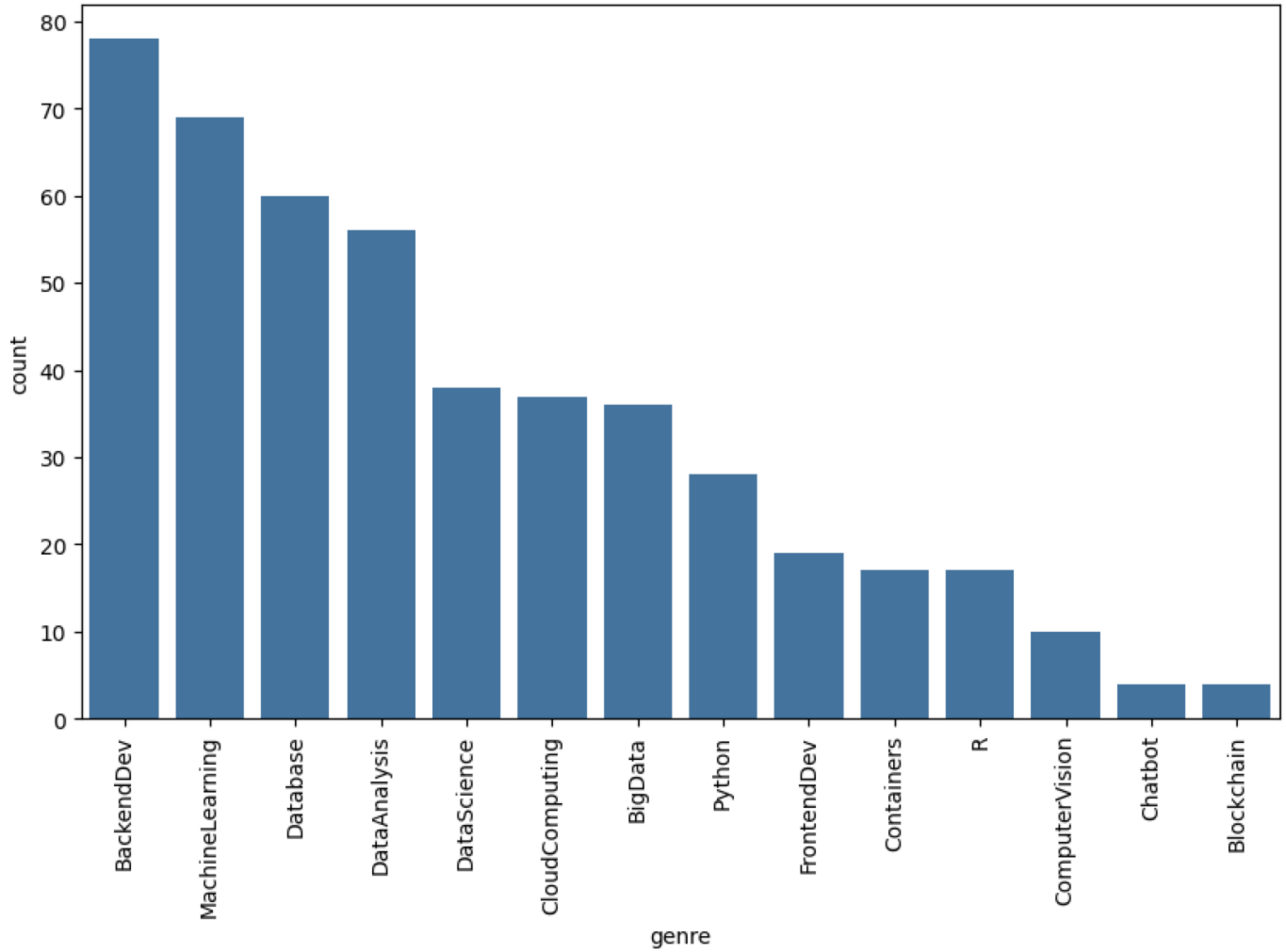
The goal of this project is to develop a memory-based and a model-based recommendation system. These systems will be deployed on separate user segments, and engagement and satisfaction will be measured for the purposes of A/B testing.

The model-based system will be trained on course ratings and evaluated based on its ability to accurately reproduce test ratings.

Exploratory Data Analysis

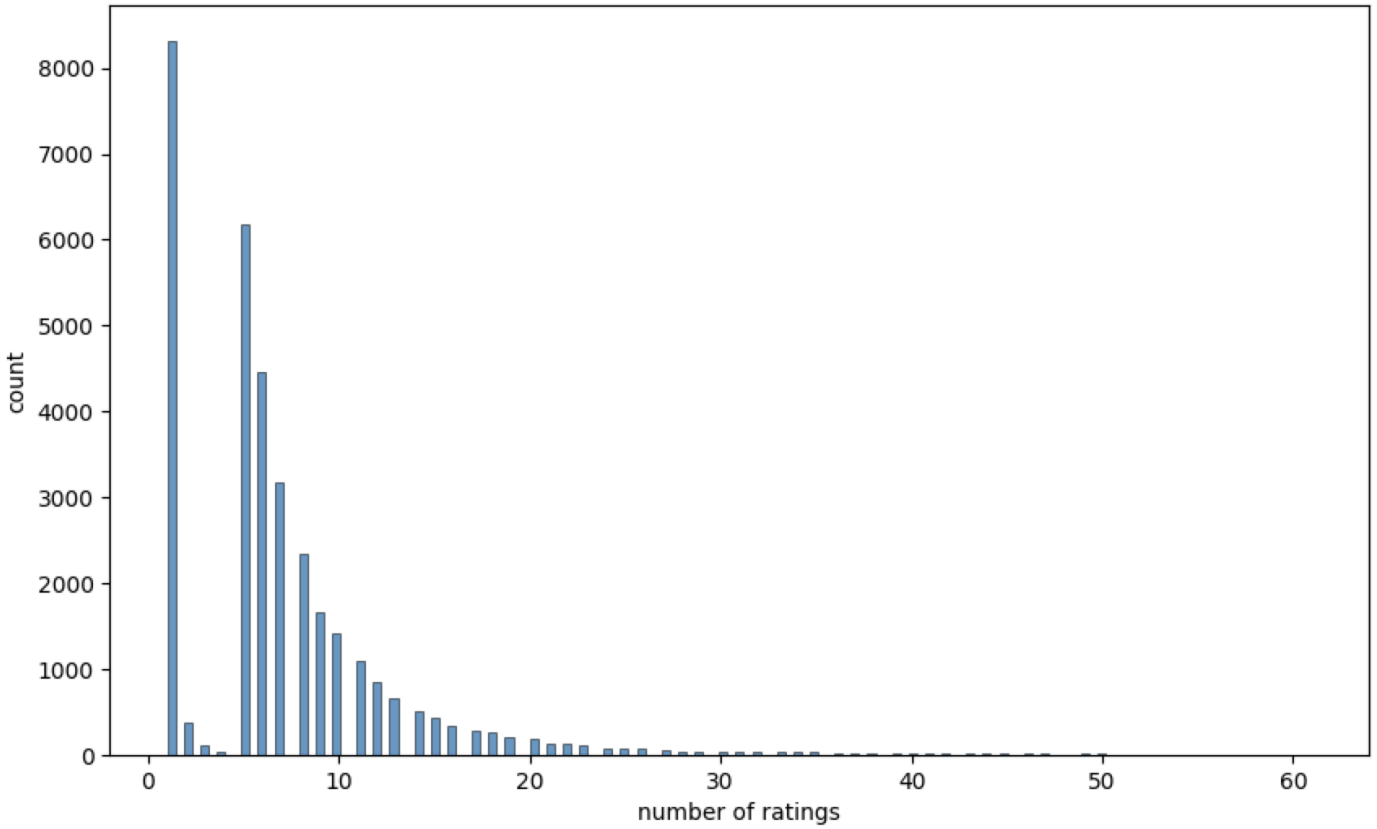


Course counts per genre



This plot shows a bar chart of the number of courses associated with each genre, providing us an indication of the distribution and relative popularity of different course genres.

Course enrollment distribution



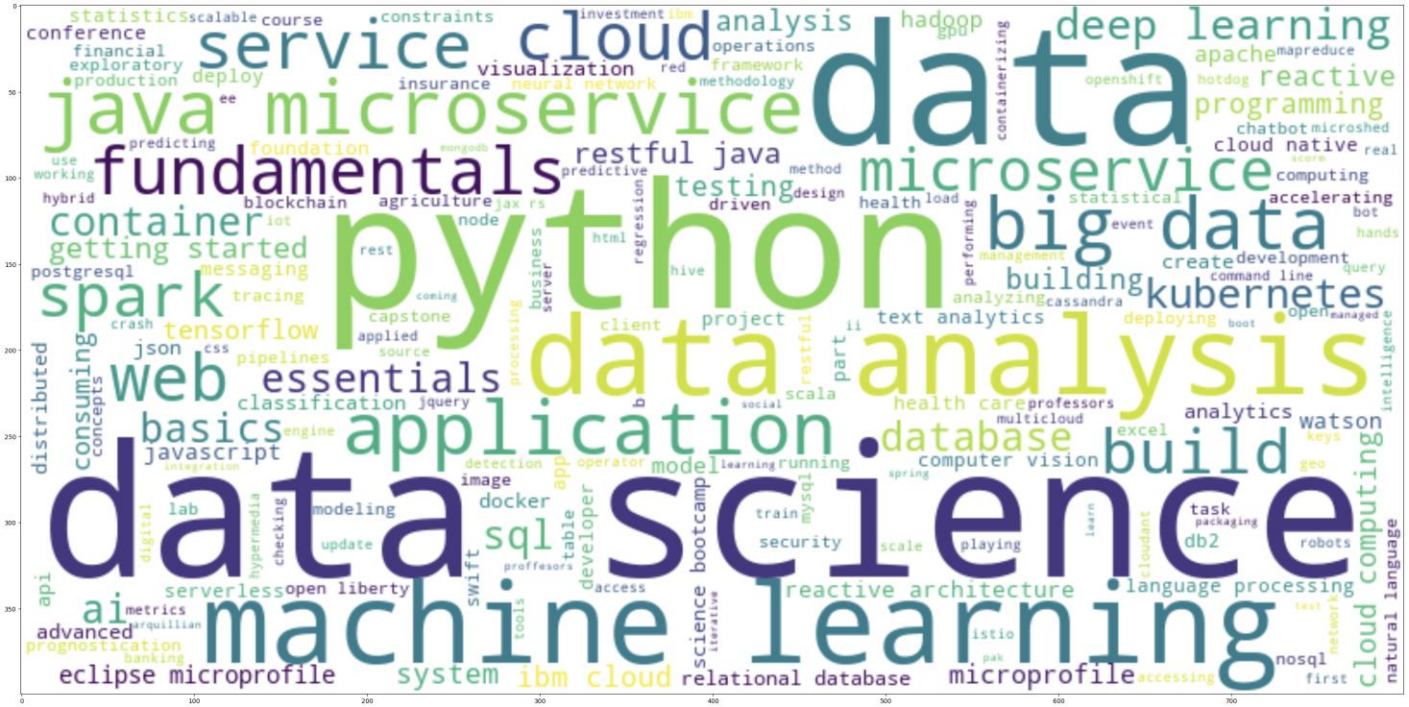
This plot shows a histogram of the number of ratings provided by users, which is equal to their number of enrollments. We see two peaks – the first is associated with a large number of users with one enrollment. There is a second peak at five enrollments with an exponential decay as enrollments increase – likely associated with full-time students.

20 most popular courses

	TITLE	ratings_count
0	python for data science	14936
1	introduction to data science	14477
2	big data 101	13291
3	hadoop 101	10599
4	data analysis with python	8303
5	data science methodology	7719
6	machine learning with python	7644
7	spark fundamentals i	7551
8	data science hands on with open source tools	7199
9	blockchain essentials	6719
10	data visualization with python	6709
11	deep learning 101	6323
12	build your own chatbot	5512
13	r for data science	5237
14	statistics 101	5015
15	introduction to cloud	4983
16	docker essentials a developer introduction	4480
17	sql and relational databases 101	3697
18	mapreduce and yarn	3670
19	data privacy fundamentals	3624

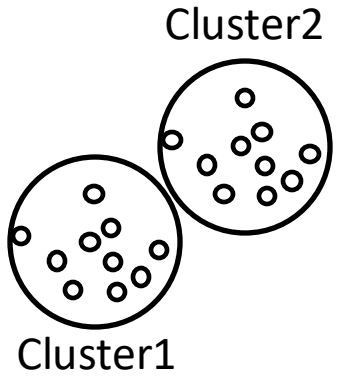
This list shows the 20 most popular courses by enrollment, including both the course title and the total number of times the course was rated (equal to the number of course enrollments).

Word cloud of course titles



This is a word cloud generated based on the course titles in the system. The word sizes are determined by the frequency with which they appear in course titles.

Content-based Recommender System using Unsupervised Learning



User profile and course genres recommender system

This is a simple, deterministic algorithm in which predicted ratings are calculated as the inner product of user profile and course genre vectors. Course recommendations are then returned based on these calculated ratings and according to a minimum threshold, adjusted such that ers are recommend 5-6 courses on averaged

Evaluation results of user profile-based recommender system

The only hyperparameter of this model is the score threshold for recommending courses; this was adjusted to result in 5-6 course recommendations/user on average, resulting in a value of 95.

The score threshold was adjusted to result in approximately 5.5 course recommendations/user on average.

10 most commonly recommended courses:

	course_id	title	times_recommended
0	excourse72	285 foundations for big data analysis with sql	202
1	excourse73	286 analyzing big data with sql	202
2	TMP0105EN	29 getting started with the data apache sp...	137
3	excourse31	244 cloud computing applications part 2 b...	86
4	GPXX0M6UEN	169 using the cql shell to execute keyspace...	53
5	GPXX097UEN	170 performing table and crud operations wi...	53
6	excourse03	216 nosql systems	53
7	excourse05	218 \r\ndistributed computing with spark sql	53
8	excourse42	255 big data analysis hive spark sql dat...	53
9	excourse10	223 database architecture scale and nosql...	53

Course similarity recommender system

Course similarity is determined based on a bag-of-words analysis of the course titles and descriptions. A TF-IDF vector is calculated for each course, and this is then embedded using a pre-trained word embedder, Word2Vec. Course similarities are then calculated based on cosine similarity in this latent feature space, and this is used to construct a course similarity matrix.

Recommendations are returned based on the similarity of unseen courses to the courses a user has previously enrolled in. There is a minimum threshold for course similarity that is again adjusted such that the algorithm returns 5-6 course recommendations.

Evaluation results of course similarity based recommender system

The only hyperparameter of the model was the minimum similarity to recommend a course. This was adjusted to 0.47.

Adjusting the minimum similarity score to 0.47 resulted in 6.2 course recommendations/user on average.

10 most commonly recommended courses:

	course_id	title	times_recommended
0	excourse23	236 data analysis using python	10820
1	excourse36	249 data analysis using python	10820
2	DA0101EN	161 data analysis with python	9883
3	TMP107	60 data science bootcamp with python	8936
4	excourse22	235 introduction to data science in python	7514
5	excourse62	275 introduction to data science in python	7514
6	excourse20	233 python and statistics for financial ana...	7398
7	excourse32	245 introduction to data analytics	6451
8	DS0110EN	131 data science with open data	6405
9	excourse68	281 big data modeling and management systems	4134

Clustering-based recommender system

PCA analysis is first used to reduce the dimensionality of a user profile matrix. K-means clustering is then performed on this reduced-dimension dataset.

Courses common across clusters are identified and ordered based on the total number of enrollments. An enrollment threshold is then specified to ensure that 5-6 courses are recommended to users on average.

Evaluation results of clustering-based recommender system

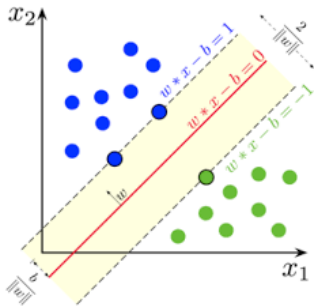
The original data frame contained 14 features. Clustering was performed using k-means with k = 20 clusters on the first 9 principal components (explaining roughly 92.7% of the total variance in the data). Course recommendations were thresholded based on enrollment, with a minimum enrollment of 400 required for a course to be recommended.

The hyperparameters detailed above resulted in an average of 5.9 course recommendations/user.

10 most commonly recommended courses:

	rec_courses	title	times_recommended
0	BD0101EN	147 big data 101	14694
1	BD0111EN	181 hadoop 101	12552
2	BD0211EN	102 spark fundamentals i	12242
3	BD0115EN	97 mapreduce and yarn	9170
4	BD0212EN	37 spark fundamentals ii	8282
5	BD0131EN	96 moving data into hadoop	8195
6	BD0141EN	55 accessing hadoop data using hive	8163
7	BD0121EN	93 apache pig 101	7997
8	DS0101EN	176 introduction to data science	7543
9	PY0101EN	188 python for data science	6640

Collaborative-filtering Recommender System using Supervised Learning



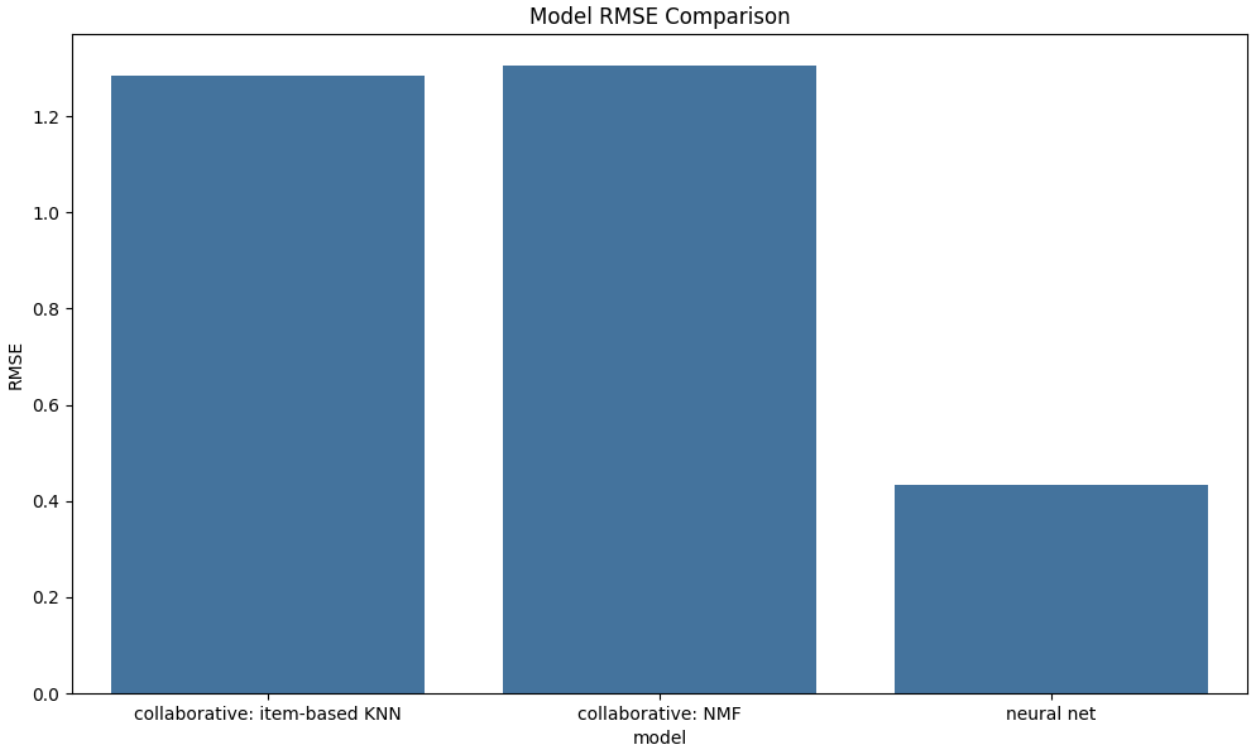
Model-based/collaborative filtering recommendation systems

Three model-based recommendation systems are trained on course ratings with the purpose of predicting how users will rate unseen courses.

The first two are collaborative filtering systems based on the user-item interaction matrix: an item-based k-nearest-neighbors algorithm and a non-negative matrix factorization algorithm.

The third model is a neural network trained on embedded representations of user profile and course description vectors.

Compare the performance of collaborative-filtering models



This bar chart indicates the performance of three tune models: an item-based k-nearest-neighbors (KNN) collaborative filtering model, a non-negative matrix factorization (NMF) collaborative filtering model, and a neural network model based on user and item embeddings.

Performance is indicated by root-mean-squared error in course ratings predictions.

As we can see, the neural network model strongly outperforms the KNN and NMF models

Conclusions

- User-profile-, course-similarity-, and clustering-based recommendation systems will need to be validated/calibrated based on user engagement and satisfaction metrics.
- The neural network model for predicting course ratings outperforms both (optimized) collaborative filter models. If a model-based approach is taken, it should be neural-network-based.
- The neural network model presented here is a proof of concept. The model should likely undergo additional fine-tuning to improve test performance.

Appendix

- GitHub repository containing all Jupyter Notebooks associated with this capstone project can be found at https://github.com/arpogg24/arpogg24.github.io/tree/main/machine-learning/IBM-Machine-Learning/capstone