Capstone Assignment (all versions combined)

Capstone Project

The Challenge: Design, Measure, Mix, Propose and Justify

This capstone project will bring together the skills you've learned across the four prior courses of the Recommender Systems specialization in a single project. You will be given a data set and a specific scenario, and will be expected to research, propose, and justify a recommender specifically designed to match that data set and scenario. You will carry out this project individually in four parts, all of which will be submitted together as a final project report:

Design. Your first challenge is to understand your data set and scenario and produce a research design. This design will identify a set of metrics and evaluation techniques you will use to evaluate possible algorithms for your recommender, and will outline your plans for exploring both individual and hybrid recommender algorithms. Pay particular attention to how you plan to separate training and test data to ensure that your tests are valid. The plan can be brief (2-3 pages), but must explain how the metrics you choose relate to the business goals in the scenario.

Measure. Next you will work with a set of at least three different base algorithms (drawn from among those you've studied) to understand how they perform on the provided data set, using your selected metrics. As part of this step you may need to tune some of the base algorithms to get reasonable performance.

Mix. With complex objectives, it is likely that no single algorithm will produce a set of recommendations that meet all of the goals of the scenario. Thus, you need to explore hybrid algorithms to provide the best result set. We expect you to explore at least two, and possibly several different hybrids to find the best results.

Proposal and Reflection. Finally, you should present your recommendation for the algorithm (including possibly a hybrid algorithm) that should be used to fulfill the scenario. You should justify the result and the means used to achieve it, and should address a set of questions about your exploration.

One Project, Two Paths

This project was designed to support programmers in the course (as most of our specialization enrollees have taken the honors track with programming), but also has an option for non-programmers.

The programming path will include using LensKit for the base algorithms, for evaluation metrics, and to compute the mixtures. The non-programming path will provide you with some raw result data over a set of base algorithms; you can use spreadsheets or tools of your choice to compute metrics and hybrids.

The Data Set

You will be using a data set derived from Amazon.com with product metadata and ratings data on office products. The data set is provided thanks to Julian McAuley at UCSD, and involves actual data from the period May 1996-July 2014. To make your computation more tractable, we've used a dense subset of the data (called the 5-core subset) that only includes items and users with at least five ratings. [Note that the original datasets are available at http://jmcauley.ucsd.edu/data/amazon/, though these should not be used for this capstone.

Note: There are separate data set extracts for those of you completing the programming (honors) track and the non-programming track. The non-programming track dataset is smaller to make it more feasible for use in spreadsheet computation.

For each item, your meta-data includes:

- An item number
- Amazon's ITEM number ("asin")
- The item's brand name
- The item title
- The item category (both leaf category and full path)
- A price in dollars
- An availability score between 0 and 1 that reflects how widespread the product is in retail stores; higher scores reflect broad availability; lower scores indicate products not found in most big box store. Note that the availability score is synthetic (we created it), but for purposes of this capstone, treat it as if it were real data.

You also are provided with a ratings matrix with a row for each item and columns representing each user (ratings are on a 1-5 star scale). Your ratings matrix includes all the ratings data you will receive (we have not separated out test and training data -- that's your responsibility).

Capstone Scenario (Project Objective)

Your project should assess and recommend a recommender solution for the following scenario:

You work for a large online retailer (we'll call it Nile-River.com) as a recommender systems expert on a team focused on direct sales to consumers in the US.

Your market research team has identified “back to school” as a critical time period for office product sales to consumers in the US. They note that the six weeks including and surrounding the month of August are responsible for 31% of yearly office product sales.

They also report that the surge in office product sales is not limited to traditional school products (such as notebooks, pencils, and erasers). Rather, it appears that once people are buying school products, they also buy other office products (indeed, more document shredders are sold during these six weeks than at any other time of the year, even tax preparation season).

Indeed, most large-dollar office-product purchases include a mix of inexpensive and more expensive products in the same transaction. This data suggests that it may be important to have inexpensive products as an entry point, but more expensive ones that build the transaction size. (For example, I buy paper and pens, and then realize I need several hundred dollars of laser printer toner.) Or alternatively that once someone comes to buy something large, they also fill in smaller items (one I’m buying a new printer, I might as well also buy a calculator and a box of colored paper clips).

They suspect that the surge in office product sales is due to in-person sales and promotion at retail outlets, with two particularly important prompts:

- Visits to office products superstores (chain stores such as Staples and Office Depot) peak during this time of year. Once parents are in the store to buy supplies for their children, they see other products of interest.
- Special displays are set up at “Big-Box” stores such as Wal-mart and Target with both school supplies and other office products.

Problem statement. Unfortunately, your site (Nile-River.com) does not experience as large a surge in office product sales during the back-to-school period. You do experience a surge (about double typical sales, or about 23% of annual consumer sales), but it is far below that of your offline competitors.

These figures include the results of existing promotions such as back-to-school banners and a free next-day shipping promotion for products sold during the two weeks when schools most commonly start classes (late August and early September).

Your challenge, therefore, is to develop a recommender system to increase sales of office products during this important time period. To maximize business value, you also have a set of key goals and constraints.

Given that your site already has a very effective product-association recommender system, you’ve been asked to focus on recommending products based on customer’s overall profiles, not their current browsing or basket.

Your product recommendations will be displayed in two places on the site:

- Five products displayed on the “office products” landing page where customers will land if they click on banner ads (back to school shopping!) or select the office products category (from various menus or navigation aids).
- Five products displayed as part of “other suggestions” that will be displayed as part of the shopping cart display and near the bottom of product pages (primarily will be placed on product pages from the same category, but also related products such as textbooks, school bags, and backpacks).

Research shows that additional sales at this time of year are divided fairly broadly among categories of office products (school supplies, consumable supplies, durable office equipment). Your recommender should respond to this research appropriately.

Your recommender should also address the finding above about having both cheaper and more expensive products available to attract customers.

Finally, Nile-River.com prides itself on having a much deeper product catalog than the typical big-box store. One of the key drivers of repeat business is customer discovery of new products they likely couldn’t buy at a local store. Your recommender should respond to this information appropriately.

Part I: Designing a Measurement and Evaluation Plan (joint)

Your first task is to develop a written capstone project plan. Specifically, this plan should detail a set of objective measures that will be used to evaluate candidate recommender systems as well as a clear plan for the multi-step evaluation process. Among the elements of this plan are:

- Translation of business goals and constraints into metrics and measurable criteria. This critical section (approximately one page) should explain how (using techniques based on what you’ve learned in the specialization) you will evaluate a set of potential recommender algorithm candidates to determine which is the best fit for your capstone scenario. You must be about the goals and specific measures, though you do not need to combine the various into a single measure -- it is acceptable to decide that you will look at the trade-offs among different objectives before selecting a final combination.

- A plan for evaluating a set of base algorithms. This plan must identify at least three different algorithms from different families, explain why these are the algorithms you’ve chosen, and explain how they will be implemented (honors only), tuned (honors only), and evaluated against your dataset and metrics (both tracks). Pay careful attention to making sure that you do not taint your evaluation by training and testing on the same data (or by tuning and testing on the same data). This plan is what you will carry out in part II of the capstone project. The length of this part of the capstone report should be 1-3 pages.

- A plan for constructing and evaluating hybrid algorithms. Your final algorithm will likely need some of the properties of different algorithms, and you should therefore plan to explore at least two, and possibly more, hybrids that mix together different algorithms. You are free to use weighted, switching, hybrid, or other mixture methods depending on what you see as most appropriate for achieving your goals. Because the exact mixtures you choose will likely depend on the specific outcomes of your base algorithm evaluation, this section of the plan can be short -- simply give a description of a half-page to a full-page explaining how you would explore combining different algorithms, and how you would evaluate the combination. You'll report on the details later.

Note: This is a two-part capstone project. As a result, you will be submitting your "plan" as part of your initial and final submissions. If you were carrying this project out in an industry or research setting, you would likely stop here and have your plans reviewed by others before proceeding. In this case, however, we want you to self-evaluate your plan -- take time to make sure you're comfortable with it -- and then proceed with carrying out the project. The plan and first set of results will be peer-evaluated in one batch. Then the rest of the results and the reflection with be reviewed (in the context of the plan and initial results) in a second peer-evaluation.

Reference Material: For this assignment, you may need to go back and review relevant material. Specifically:

Metrics and evaluation can be found in the Evaluation and Metrics course:

- https://www.coursera.org/learn/recommender-metrics

Possible algorithms to combined are spread among three courses:

- https://www.coursera.org/learn/recommender-systems-introduction (content-based)
- https://www.coursera.org/learn/collaborative-filtering (nearest neighbor CF)
- https://www.coursera.org/learn/matrix-factorization (matrix factorization, latent factor)

Hybridization techniques are presented in:

- https://www.coursera.org/learn/matrix-factorization

Part II: Measurement (separate)

Standard (Non-Programming) Version

For the standard version of this assignment, you have been provided with the predicted rating outputs from five different algorithms: a TFIDF content-based recommender, an item-item collaborative filtering recommender, a matrix factorization (gradient descent) recommender, a user-user collaborative filtering recommender, and a baseline recommend that uses product and customer ratings distributions to provide personally-scaled average predictions (PersBias).

The file is below:

office-products.xlsx

Note that these predicted ratings include predictions for all items (including rated ones) as you will need those predictions to compute your evaluation metrics.

For part II of this assignment, you must calculate a set of metrics for at least three of these algorithms (it may be useful to do all five). The metrics are those that you identified above as relevant to the business goals of your algorithm, though in some cases you may need to make compromises to find close-enough metrics that you are able to compute.

In particular, your metrics will cover some subset of the following types of measurements, all of which can be achieved using a spreadsheet and techniques that were used in the four prior courses:

- Accuracy measures -- measures that look at the difference between known “ground truth” data (e.g., provided ratings) and predicted ratings. You are not likely to need more than one accuracy measure for this capstone.
- Rank measures -- measures that look at how often the top-recommended items for each user are actually items the user has rated/liked.
- Top-n characteristic measures -- measures that look at properties of the top-recommended items for each user, such as popularity, diversity along some characteristic, etc.

Deliverable. In your report you should provide a summary table of the statistics, along with the conclusions you draw about the strengths and weaknesses of the different algorithms for this particular data set and objective. You should also identify the specific algorithms you plan to carry forward into the next phase to mix into your final algorithm.

NOTE: After you complete Part II of this project, you should submit Parts I and II as a single report for peer evaluation.

Honors (Programming) Version

For the honors version, you need to select and run a at least three base algorithms, tune the individual algorithms, and select a subset to contribute to your final solution. These algorithms can be non-personalized recommenders, collaborative filters we have studied, or additional algorithms of your design such as content-based recommenders using item metadata.

The provided project template contains several things:

capstone.zip

- A LensKit-compatible version of the rating data.
- Untuned configurations for several of LensKit’s algorithms for the Amazon data. The ‘-E’ variants use rating data (explicit feedback), while the ‘-I’ variants are configured to use the rating data as implicit feedback, looking at whether or not a user rated an item.
- A skeleton experiment that prepares the for the evaluator, and runs the algorithms on the data set.

- A simple category-aware item recommender that limits the number of items per category that can be recommended, as an example of how to make use of that data in LensKit. It is configured in the ‘pop-spread’ and ‘ii-spread’ configurations.

You will need to extend this with your additional algorithm configurations, experiment configurations, and possibly custom algorithms and metrics.

Tune the parameters of the underlying algorithms to produce effective versions, and report on the performance of your various algorithms in accordance with your evaluation plan. This report should be a section of your final report. Describe the logic of any custom algorithms as well.

Depending on your evaluation plan, you may need to generate train-test splits yourself and manually configure them. The LensKit manual documents how to configure such data sets: https://lenskit.gitbooks.io/lenskit-manual/evaluator/train-test.html#input-data

Deliverable. In your report you should provide:

- A summary of your algorithm tuning (showing the tuning parameters tried and metric results for different tunings);
- A summary table of the metrics associated with your tuned algorithms (there may be more than one tuning you choose to keep from a given algorithms, if different tunings have different benefits), along with the conclusions you draw about the strengths and weaknesses of the different algorithms for this particular data set and objective
- Identification of the specific algorithm variants you plan to carry forward into the next phase to mix into your final algorithm.

LensKit Version Note: The data infrastructure makes use of features added since LensKit 3.0-M2, and the latest version of LensKit has slightly changed the interface for writing metrics. If you use your metric class from Module 3 as the starting point for a new metric, you will need to adapt your method interfaces. The two changes are that ‘createContext’ now takes a ‘RecommenderEngine’ instead of a ‘Recommender’, and ‘measureUser’ now takes a ‘Recommender’ as its first parameter. These changes were made in order to make LensKit experiments more efficient by allowing multiple users to be measured in parallel.

NOTE: After you complete Part II of this project, you should submit Parts I and II as a single report for peer evaluation.

Part III: Mixing the Algorithms (separate)

Standard (Non-Programming) Version

Next, you will explore at least four combinations of algorithms to produce the best possible hybrid algorithm for this dataset and objective (based on your defined criteria. Remember to anchor this exploration and evaluation to the metrics of importance -- when using top-n, we recommend focusing on top-5. Note that it is acceptable to include recommendations for products already rated, since office products are a domain where re-purchase is common). Your goal is to produce top-5 lists that provide the best performance against your combined set of criteria.

The specific hybrids you use may include ones that combine prediction scores from different algorithms, then produce a new top-5 list, and ones that combine the top-5 lists generated by individual algorithms. You may also create “single-algorithm mixes” that apply specific criteria to existing algorithms (e.g., diversification) or across multiple algorithms. And you may incorporate switching hybrids that use different algorithms for different customers or products.

In some cases you may not be able to automate these computations in a spreadsheet. Where this is the case, please feel free to either hand-calculate results for just 10 users or to use external tools or programming to generate the mixture (some of you will prefer to do this, for example, using Python).

Deliverable. In the end, you will have evaluated at least three combinations against your criteria, and should provide three things in your final report:

- A list of the hybrids tried with a clear explanation of the specific form of hybrid used. For each hybrid, give a brief explanation of why you chose to use that hybrid (what did you hope it would achieve).
- A table of metric results for each of your hybrids and the results of each of the component algorithms.
- A set of sample outputs for each of your hybrids and component algorithms -- give the top-5 recommendation lists for three different users.

Honors Version

With your individual algorithms prepared and tuned from Part II, you now need to explore different hybrids (at least three of them), evaluate their performance, and provide example recommendations.

The ‘recommend’ task in the provided Gradle configuration will produce recommendations. You need to provide it with options, using the ‘-Pname=value’ syntax in the command line:

- The ‘algorithm’ property specifies the name of an algorithm configuration (e.g. ‘-Palgorithm=pop-spread’)
- The ‘userId’ property specifies the user ID for recommendations (e.g. ‘-PuserId=100’)

You will need to write configuration files to configure your hybrids; the support code (such as RecommenderList) and configurations from the hybrid assignment in Module 4 will help with this.

Deliverable. In the end, you will have evaluated at least three combinations against your criteria, and should provide three things in your final report:

- A list of the hybrids tried with a clear explanation of the specific form of hybrid used. For each hybrid, give a brief explanation of why you chose to use that hybrid (what did you hope it would achieve).
- A table of metric results for each of your hybrids and the results of each of the component algorithms.
- A set of sample outputs for each of your hybrids and component algorithms -- give the top-5 recommendation lists for three different users.

Part IV: Proposal and Reflection (joint)

The final part of your project is a brief “business proposal” and a reflection on what you’ve learned in the project.

Deliverable. The first deliverable is a two-page “business proposal” for the recommender choice you’ve determined is best. It should be written to be accessible to a non-technical reader, and should address the following items:

- Short summary of the business need and opportunity
- Explanation of the evaluation criteria used (in terms understandable to a non-expert)
- Non-expert description of the recommender algorithm being recommended
- Explanation of why this algorithm is recommended over other alternatives
- Example recommendation lists to illustrate how this algorithm meets the business need
- Brief notes on anything you feel would be important to explain about the research process or result

These do not need to be separate sections (e.g., you might choose to use the examples in the explanation, or describe the criteria as part of explaining why your chosen algorithm is recommended).

Deliverable. The second deliverable for this part of the capstone is a reflection document on the process, approximately two pages, but no fixed limit, addressing at least the following five questions.

- Reflect on the process of translating business requirements to metrics. What was easy or hard about that process? Do you feel you had adequate preparation in the specialization to take on such a task?
- Reflect on the process of evaluating individual algorithms and creating hybrids. How easy or hard was it to identify differences in performance of algorithms on different criteria? How easy or hard was is to bring together elements from different algorithms through hybrids? How confident are you that your final result is a good algorithm for the problem.
- Reflect on data management, including the process of separating training and test data. How well were you able to maintain that separation and avoid overfitting your dataset? How confident are you in the generalizability of your result to other similar datasets?

- Reflect on the tools you used for this capstone (whether spreadsheet, LensKit, or external ones). Do you feel you had the experience and skill with the tools you needed for the capstone? Do you feel the tool was a good match for the problem (and if not, what would you have preferred)? Please identify areas where the tools were particularly helpful or particularly challenging.
- Finally, reflect on the capstone experience as a whole. Did it achieve its goal of giving you one project to bring together the diverse set of materials you learned in this specialization? Do you feel more capable of (or confident in your ability for) taking on applications of recommender systems? Other lessons you're willing to share?

NOTE: After you complete Part IV of this project, you should submit the entire project (including Parts I and II again) as a single report for peer evaluation.

Peer Evaluation Note: You must complete peer evaluation of others in order to complete the Capstone project.

Evaluation Rubric

To help you understand how you will be evaluated, here is a summary form evaluation rubric:

700	Total Project	Passing = 560+	
200	Part I -- Plan		
	100	Translation of Business Goals into Metrics	
		20	Identify specific metrics to use in evaluation
		20	Evaluation clearly focused on "five recommendation" opportunity
		20	Addresses diversity (by category and/or price)
		20	Addresses broader availability
		20	Overall match of metrics to business goals

	80	Plan for evaluating base algorithms	
		10	Identifies specific algorithms
		20	Justifies selected algorithms
		20	Plan for implement/evaluate
		10	Addresses data management (test/train)
		20	Overall match to business goals and problem
	20	Plan for hybrids	
		20	Addresses building and evaluating
200	Part II -- Measurement		
	100	Summary Table of Statistics	
		20	At least three alborithms
		20	Evidence of having measured them on datasets
		20	Table showing clear results
		20	Tuning (Honors) or Good Approx when needed (Standard)
		20	Description of any custom algorithm or metric

	70	Conclusions	
		10	Discusses the performance of each algorithm or variant
		20	Clear link between discussion and the metrics identified
		20	Link between discussion and the specific case/problem
		20	Good discussion of strengths/weaknesses of each algorithm for this use
	30	Selection for Hybridizing	
		20	Identification of algorithms to hybridize
		10	Reasonable rationale (if not simply taking all)
150	Part III -- Mixing		
	60	Hybrids	
		20	Clear description of at least three hybrids (must be different)
		20	Justification for why this hybrid seems promising
		20	Reasonable choices given goals, metrics, results
	60	Evaluation	

		20	Table of metrics for hybrids and component algorithms
		20	Includes all relevant metrics
		20	Sanity check on results
	30	Sample Outputs	
		20	Sample top-5 lists for at least three users
		10	Shown in a way that is consistent with illustrating the claimed result
150	Part IV -- Proposal and Reflection		
	80	Proposal	
		10	Clear summary of the business need and opportunity
		20	Clear summary and explanation of evaluation criteria
		20	Clear description of the algorithm
		20	Clear justification in terms of business needs
		10	Includes examples that illustrate
	70	Reflection	
		20	Data management -- evidence of significant effort to manage appropriately

		20	Thoughtful reflection on the project as a whole (including translating RQ to metrics)
		20	Thoughtful reflection on substance of the project (algorithms, metrics, hybrids)
		10	Thoughtful reflection on tools
Notes:	10-point items will generally be graded on a scale of 10/9/8/6/4/0 (outstanding, acceptable, deficient, seriously deficient not worthy of credit) but with detailed descriptions		
	20-point items will generally be graded on a scale of 20/18/16/12/8/0 (outstanding, superior, acceptable, deficient, seriously deficient, not worthy of credit)		