MKTG/STAT 476/776:
Applied Probability Models in Marketing

Spring 2022 (Monday/Tuesday/Wednesday 3:30-6:30PM, JMHH 270)

Professor Peter Fader and TAs: Sarah Ye, Aryan Nagariya, Daniel McKeon, Junbin Huang, Wyatt Currie (group email: mktg476776ta@wharton.upenn.edu)

Motivations and Objectives

The most important questions of life are indeed, for the most part, really only problems of probability.
It is remarkable that a science which began with the consideration of games of chance
should have become the most important object of human knowledge.

Pierre-Simon Laplace, *Théorie Analytique des Probabilités*, 1812

Over the past six decades, statisticians have developed a number of models that have proven to be highly effective in their ability to explain and predict empirical patterns within many areas in business, the social sciences, and other domains in which individual behaviors can be tracked over time. These models use some basic "building blocks" from probability theory to offer behaviorally plausible perspectives on different types of timing, counting, and choice processes. Researchers in marketing have actively contributed to (and benefited from) these models for a wide variety of applications, such as new product sales forecasting, analyses of media usage, targeted marketing programs, estimation of customer lifetime value, and even overall corporate valuation. Other disciplines have seen equally broad utilization of these techniques.

As new forms of information technology provide increasingly rich descriptions of individual-level shopping/purchasing behavior, these models offer great value to practicing managers, particularly those interested in pursuing CRM ("customer relationship management") activities. Furthermore, as more managers become comfortable with non-linear optimization techniques (using, for example, the "Solver" feature within Microsoft Excel), the specification and interpretation of these models can become a regular part of the manager's toolkit. Taken as a whole, the methodological approaches covered in this course are well-suited to address the types of questions that are being asked with increasing frequency and interest by investors and managers of today's data-intensive businesses.

The principal objectives of this course are:

- To familiarize students with probability models and their role in marketing, information systems, supply chain management, actuarial science, operations research, public policy, human resources management, and many other areas.

- To provide students with the analytical and empirical skills required to develop probability models and apply them to problems of genuine managerial interest.

- To have students develop good instincts to judge the appropriateness, performance, and value of different kinds of models in a variety of settings.

- To encourage students to think critically about statistical methods and managerial perspectives that arise in many domains but are not always the best ways to approach data-oriented decision problems.

Prerequisites

This course is open to students at any level (undergraduate, MBA, other masters, PhD) who have sufficient curiosity and raw mathematical aptitude to handle the new methods that will be introduced and featured here. It is essential that students have some familiarity with basic integral calculus. Furthermore, a mid-level probability/statistics course would be helpful, but one's ability to learn and fully understand new concepts/methods is far more important than mere exposure to them. Finally, there is no need to have taken any marketing (or business) courses before this one.

Smart and highly motivated students are encouraged to take the course sooner (e.g., sophomore year or first-year MBA) rather than later. The course is very helpful for summer internships, and it provides an excellent foundation for other advanced modeling/data science courses that can be taken after this one.

Course Organization and Materials

Every session will be lecture-based, with an emphasis on real-time problem solving, including mathematical derivations and numerical investigations using Microsoft Excel. Central to the development of these skills is hands-on experience. To this end, a set of homework exercises will be assigned for most sessions. It is very important to carefully work through these exercises in order to learn the material. By themselves, they play a small role for the course grade, but will have a huge impact on genuine learning through the semester as well as performance on the final exam.

There is no formal textbook for the course (since no suitable book exists), but lecture notes covering most of the material presented in class will be posted on Canvas. All Excel spreadsheets used in class will be made available to the students, and some journal articles, popular press pieces, and blog posts will be suggested as illustrations/applications of the techniques discussed. But most of these readings are just recommended – there will be no formal pre- or post-class reading assignments for any session.

Professor Fader's two books ("Customer Centricity: Focus on the Right Customers for Strategic Advantage" and "The Customer Centricity Playbook: Implement a Winning Strategy Using Customer Lifetime Value") are recommended for students who are interested in seeing how the methods developed in the course can translate into business strategy. They are not required, and will be of little help for the methodological aspects featured in the course. But they do provide a useful frame to motivate and better appreciate many applications of the models.

Teaching Approach

The methods covered in this course will be quite unfamiliar to most students at the start of the semester. As such, it is important to ensure that the first exposure is impactful and that there are opportunities to work with the material multiple times and through multiple formats. To make this possible we will utilize a fairly unique "heads up" learning approach in the classroom. The basic elements include:

- Strongly recommended (although not mandatory) classroom attendance. (But note that class participation is a substantial component of the course grade).
- The use of laptops in the classroom is discouraged; same for detailed note-taking.
- Each session will be recorded and posted in a multi-media format (Panopto); students are expected to carefully review these recordings – that's when note-taking and Excel work should take place. Significant learning occurs as students go through the material for the second time.

- Because there are multiple sections of the course, students are encouraged to go through the recording(s) from one of the sessions that they didn't attend live.

These steps are intended to help students keep their "heads up" to focus on the main points in each session. Students are encouraged to ask questions about key conceptual issues, managerial applications, and the overall modeling philosophy; however, questions about minor clarifications should be addressed by reviewing the presentation decks and recordings after class (and utilizing the online discussion platform to post and answer questions).

Students are expected to create their own complete set of class notes after attending each session and working through the decks/recordings. It is fine for students to collaborate on this task, but it's best for each student to actively participate in the process and create their own notes. Any kind of "divide and conquer" approach will be counterproductive for the student (particularly with regard to the final exam).

Attending Different Sections

The three sections of the course are basically identical and interchangeable: the same material will be covered in each one. Students can freely switch sections from week to week, and there is no need to ask (or notify) anyone in advance. We encourage students to attend any of the sections but then watch the videos from one (or both) of the sections they didn't attend – slight differences from one session to another can be a helpful way to learn the material better.

Auditing, Waitlist, and Pass/Fail

- Anyone is welcome to audit the course; use the survey link below to get full access to Canvas. Auditors are allowed to attend classes and participate in discussions. If they are on the waitlist they can (and should) submit homework, until their status is fully clarified.

- To join the waitlist or for audit access, go to https://goo.gl/YRWBk4 and fill in the form. Professor Fader will notify the most deserving students as spaces open up.

- Students cannot take the course pass/fail: given high demand for it, students must be committed to doing their best throughout the semester; striving to merely pass is not good enough.

Evaluation

Homework (10% of final grade): These exercises will be both analytical and numerical. It is fine for students to communicate about specific problems, but every student must write up each problem independently and hand in their own work. Completed assignments are due on Wednesday at 3:30PM a week after they're assigned, and must be uploaded to Gradescope (via Canvas) – more details below. Homework will be assigned weekly for the first half of the course, and then sporadically afterwards.

Class Participation (15%): Although there are no formal case discussions, students are expected to be actively engaged in the lectures. Active involvement on the online discussion platform is also expected (and will count towards the participation grade) as well.

Project 1 (20%, due 2/23): For the first paper, students will be asked to find a specific type of dataset

and analyze it carefully. Papers will be evaluated using an innovative collaborative grading system, the Wharton Online Ordinal Peer Performance Evaluation Engine (WHOOPPEE). Details about the assignment and grading process will be discussed in class.

Project 2 (25%, due 4/6): The second paper will be more standardized – all students will be given a common dataset to analyze (and WHOOPPEE will be used again for grading).

Final Exam (30%, date TBA): The final exam will be a structured set of questions to assess students' conceptual understanding of the course material. It will not require any complex mathematical derivations or extensive numerical calculations, but it will be one of the most challenging exams you take at Wharton/Penn, so you must prepare for it throughout the semester.

All relevant University of Pennsylvania policies regarding academic integrity must be followed. Students may not submit work that has been prepared by (or in conjunction with) someone else, without explicit instructor permission. Any student who in any way misrepresents somebody else's work as their own will face severe disciplinary consequences.

Homework Guidelines

Throughout the semester, we will be using both **Gradescope** for homework submission and evaluation. For each assignment, you will make two separate submissions:

1. A PDF writeup of the homework solutions – this will be submitted through Gradescope.
2. The excel spreadsheet developed for the assignment – this will be submitted through Canvas.

After you submit on Gradescope, make sure to utilize the **Select Pages** tool to match your response to each question with the correct page(s) in your document. Gradescope also offers a **Regrade** tool for questions that you believe were mis-graded. Feel free to use it, but keep in mind that homework is only worth 10% of your total grade. So please think about re-grades more from a learning standpoint rather than a grade-improvement one.

The PDF writeup can be typed or handwritten (or a mix of the two), but make sure that everything is clear and legible. It is important that it is a fully **self-contained document**, i.e., the reader should not need to check your spreadsheet to understand your solutions. For example, we highly recommend that you include a table of estimated parameter values in your answers to modeling questions.

We allow for submissions of late assignments, with a penalty, for up to **one day** after its original deadline. After that it will not be accepted (and don't ask for exceptions!). Please reach out to the TA team with any questions, concerns, or comments about any homework-related issues.

Course Schedule

Note that the university's academic calendar starts on Wednesday 1/12, so Session 1 will be covered on 1/12, 1/18, and 1/19. As noted above, students can attend any of the sessions, regardless of which section they are formally assigned to. The regular weekly schedule will begin on Monday 1/24.

There are two weeks in which there are no scheduled MBA classes. Nevertheless, MBA students will be responsible for materials covered during those weeks – they must watch the videos on their own (or they are welcome to attend class if they wish to do so).

Session 1 (W 1/12, T 1/18, W 1/19): Introduction to probability models

Motivating problem: forecasting customer retention. Comparisons to traditional regression-based models: "curve-fitting" vs. "model-building." Careful derivation of a parametric mixture model (the beta-geometric). Coverage of maximum likelihood estimation and the Microsoft Excel Solver tool. Discussion about the philosophy and objectives of probability modeling.

Session "1A" (Date/time TBA): Math/stat review

Optional Q&A session on any concepts/methods covered in Session 1.

Session 2 (M 1/24, T 1/25, W 1/26): Models for count data

Motivating problem: projecting media exposure patterns. Introduction to the Poisson process and its extension to the negative binomial distribution. Understanding reach curves.

Session 2A (M 1/31, T 2/1, W 2/2): More on count models

Evaluating goodness-of-fit. Generalizing the model to allow for "spikes" at 0 (and elsewhere). Likelihood ratio test. Dealing with problems of limited/missing data: truncated and shifted NBD models.

Session 3 (M 2/7, T 2/8, W 2/9): Even more on count models

Alternative estimation approaches ("Means and Zeroes" and "Method of Moments"). Making various model inferences, e.g., the "80:20 rule." Applications to Facebook and other real-world datasets.

Session 4 (M 2/14, T 2/15, W 2/16): Repeated choice processes and empirical Bayes methods

Choice vs. counting. The binomial distribution and the beta-binomial mixture model. Parameter estimation. Bayes Theorem. Conditional distributions and expectations. Combining population information ("priors") with observed data for individuals. Regression-to-the-mean.

Session 5 (M 2/21, T 2/22, W 2/23): Continuous-time duration models
Project 1 due

Motivating problem: forecasting new product adoption. Implementing and evaluating different timing models, particularly the Pareto(II). Dealing with grouped data and right censoring. Introducing hazard functions. Discussion of other timing models (e.g., Weibull), and the linkages among them. Exploring the interplay between timing and counting processes.

Session 6 (M 2/28, T 3/1, W 3/2): Customer-base analysis

First-half course review. Understanding customer lifetime value (CLV). Combining the basic building blocks to estimate CLV. Introducing the beta-discrete-Weibull model.

Session 7 (M 3/14, T 3/15, W 3/16): Customer-base analysis (cont.)
More CLV-oriented applications.

Session 8 (M 3/21, T 3/22, W 3/23): Introducing covariates

Poisson regression and NBD regression for count models. Proportional hazard methods and covariate effects for timing models. General discussion about the different role of covariates from the perspective of an econometrician vis-à-vis a probability modeler. Applications.

Session 9 (M 3/28, T 3/29, W 3/30): Finite mixture/latent class methods

Looking at non-parametric (discrete) approaches to capturing heterogeneity. Interpreting support points versus cluster characteristics. Estimation issues. Overview of selection criteria for non-nested models. Other uses for FM/LC methods.

Session 10 (M 4/4, T 4/5, W 4/6): Multi-item choice models
Project 2 due

The multinomial choice process and the Dirichlet mixing distribution. Interplay between the beta and Dirichlet distributions.

Session 11 (M 4/11, T 4/12, W 4/13): Fun with Dirichlet!

Further examination of the Dirichlet-multinomial choice model and its astonishing patterns. Discussion of Ehrenberg's "empirical laws."

Session 12 (M 4/18, T 4/19, W 4/20): Integrated models

Combined models of counting, timing, and/or choice. Particular focus on the BB/NBD as a working example.

Session 13 (M 4/25, T 4/26, W 4/27): Nonstationary processes

Overview and comparison of techniques such as renewal processes, learning models, hidden Markov methods, and other approaches to capture dynamics over time.