MKTG/STAT 4760/7760:
Applied Probability Models in Marketing

Spring 2026 (Monday/Tuesday/Wednesday 3:30-6:30PM)

Professor Peter Fader and TAs (aka the "K Team"): Nicky Kantakom (head TA), Anika Bastin, Otakar Korinek, Kristen Li, and Kun Kei Xiao (group email: mktg476776ta@wharton.upenn.edu)

Motivations and Objectives

The most important questions of life are really only problems of probability. It is remarkable that a science that began with the consideration of games of chance has become the most important object of human knowledge.

Pierre-Simon Laplace, Théorie Analytique des Probabilités, 1812

Almost all of machine learning can be viewed in probabilistic terms, making probabilistic thinking fundamental. It is through this view that we can connect what we do in machine learning to every other computational science. For this reason alone, mastery of probabilistic thinking is essential.

Shakir Mohamed, Google DeepMind, 2018

Statisticians have developed a family of models that have proven highly effective in explaining and predicting empirical patterns within various areas of business, the social sciences, and other domains where individual behaviors can be tracked over time. These models utilize basic "building blocks" from probability theory to provide behaviorally plausible perspectives on various 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. 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 explaining and forecasting a wide range of phenomena in marketing, information systems, supply chain management, corporate finance, epidemiology, public policy, human resources management, and many other areas.

- To introduce students to *generative models* and develop an appreciation of how simple "as-if random" stories can often explain and forecast data patterns better than complex alternatives.

- To provide students with the analytical and empirical skills required to implement probability models, apply appropriate evaluation criteria to judge their suitability/performance/usefulness, and to effectively communicate the results/implications to managers, policy makers, and other leaders.

- To encourage students to think critically about common statistical methods and managerial perspectives that may not be the best ways to approach certain data-oriented decision problems.

Prerequisites

This course is open to students at any level (undergraduate, MBA, other master's, PhD) with sufficient curiosity and raw mathematical aptitude to handle the new methods introduced here. Students must have some familiarity with basic integral calculus (even if it was in the distant past). 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 can be helpful for summer internships and 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, emphasizing real-time problem-solving, including mathematical derivations and numerical investigations using Microsoft Excel. This intense experience will carry over to the weekly homework exercises. It is very important to work through these exercises carefully; by themselves, they carry little weight (10%) on the overall 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 three books ("Customer Centricity," "The Customer Centricity Playbook," and "The Customer-Base Audit") 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 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 essential to ensure that the initial exposure is impactful and that there are opportunities to work with the material multiple times and through various formats. To make this possible, we will utilize a unique "heads up" learning approach in the classroom. The basic elements include:

- Strongly recommended classroom attendance. It is not mandatory and will not be tracked. But note that class participation is a substantial component of the course grade.
- Laptop use in the classroom is discouraged (but not prohibited); same for detailed note-taking.
- Frequent (but friendly) cold calling will take place throughout each session.
- Students must review the classroom recordings – that's when note-taking and Excel work should occur. 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 review the recording(s) from one of the sessions they didn't attend live.

These steps are intended to help each student keep their "head up" and 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 technical issues should be addressed by reviewing the presentation decks and recordings after class (and utilizing TA office hours as well as the online discussion platform to post and answer questions).

We will also utilize a unique LLM, i.e., the "Lecture Recall Chatbot," to better connect the in-class experience with the out-of-class learning opportunity. More details will be provided as the course begins.

Students are expected to create their own complete set of class notes after attending each session and working through the decks/recordings. It is great for students to collaborate on this task by talking through the key "takeaway" points from each session, but it is 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 course's three sections are 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 lecture recordings 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.

Waitlist, Pass/Fail, and Auditing

- To join the waitlist, complete the form at https://goo.gl/YRWBk4. Professor Fader will notify the most deserving students as spaces open up.

- Students can take the course pass/fail, but audits will not be permitted.

Evaluation

Homework (10% of final grade): These exercises will be both analytical and numerical. It is fine (in fact encouraged) for students to discuss specific problems with each other, but everyone must write up their work independently. 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.

Class Participation (15%): Although there are no formal case discussions, students are expected to be actively engaged in the lectures, which will include frequent cold calls to ensure that everyone is following (and participating in) the conversation. Active involvement on the Ed Discussion online platform is also expected (and will count towards the participation grade) as well.

Project 1 (25%, due 2/25): 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 4/8): 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 (25%, dates 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 will be one of the most challenging exams you take at Wharton/Penn, so you must prepare for it throughout the semester. For most students, the exam will take place in early May (during the university's official final exam period), but for MBAs only it will take place in late April. More details to come.

All relevant University of Pennsylvania policies regarding academic integrity must be followed. Students may not submit work prepared by (or in conjunction with) someone else. Any student who misrepresents somebody else's work as their own will face severe disciplinary consequences.

Homework Guidelines

Throughout the semester, we will use **Gradescope** for submitting and evaluating homework. 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.

On Gradescope, you will utilize the **Select Pages** tool to identify page(s) in your document are associated with each problem. Gradescope offers a **Regrade** tool for questions that you believe were mis-graded. Feel free to use it but remember that homework is only worth 10% of your total grade. So, please consider regrades from a learning standpoint rather than a grade-improvement perspective.

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 essential that the document is fully self-contained, i.e., the reader should not need to refer to your spreadsheet to understand your solutions. For example, we highly recommend including 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 1/14 and there are no classes on 1/19 for MLK Day, so Session 1 will be covered on 1/14, 1/20, and 1/21. 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 1/26.

There are two weeks in which there are no scheduled MBA classes, but MBA students will be responsible for everything covered during those weeks – they must watch the lecture recordings (or they are welcome to attend class if they wish to do so).

Session 1 (W 1/14, T 1/20, W 1/21): 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 2 (M 1/26, T 1/27, W 1/28): 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 2/2, T 2/3, W 2/4): 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 2/9, T 2/10, W 2/11): 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 2/16, T 2/17, W 2/18): 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 2/23, T 2/24, W 2/25): Continuous-time duration models
Project 1 due 2/25
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 3/2, T 3/3, W 3/4): 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.

Spring Break

Session 7 (M 3/16, T 3/17, W 3/18): Customer-base analysis (cont.)
More CLV-oriented applications.

Session 8 (M 3/23, T 3/24, W 3/25): 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 3/30, T 3/31, W 4/1): 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 applications for FM/LC methods.

Session 10-11 (M 4/6, T 4/7, W 4/8): Multi-item choice models and the "empirical laws" of customer choice
Project 2 due 4/8
The multinomial choice process and the Dirichlet mixing distribution. Interplay between the beta and Dirichlet distributions. Further examination of the patterns associated with the Dirichlet-multinomial choice model and the "empirical laws" of Ehrenberg/Sharp.

Session 12 (M 4/13, T 4/14, W 4/15): Integrated models
Combined models of counting, timing, and/or choice. Particular focus on the BB/NBD as a working example.

Session 13 (M 4/20, T 4/21, W 4/22): Nonstationary processes
Overview and comparison of techniques such as renewal processes, learning models, hidden Markov methods, and other approaches to capture dynamics over time.

Session 14 (M 4/27, T 4/28, W 4/29): Customer-based corporate valuation
Bringing everything together to forecast company revenue – and extensions to overall corporate valuation.

Final exam (TBA)