95866: Advance Business Analytics

Fall 2020, Mini I.

Class Schedule: 8-9:20pm, Remote

Instructor: Prof. Rahul Telang
HBH 3040
Ph: 81155
Email: rtelang@andrew.cmu.edu

TA: Yangfan Liang,
HBH 3002
Email: yangfanl@andrew.cmu.edu

Syllabus:

Growth in Web 2.0 coincides with the growth in firms' ability to collect large customer data. Firms know micro level data about customer transactions and have an ability to correlate such data with other datasets. In this course, we will learn powerful but simple probability/statistical models that can be applied to fit these data to generate useful predictions. The course will go beyond pattern detections, clustering, or correlation in data to build models of plausible consumer behavior that generates the data. Thus the goal is to build a "model" of consumer behavior and apply this model to data to test how accurate it is and tweak if necessary. Most importantly, with such a model in hand, we want to predict how outcomes will change if the firm changed its strategy. Thus a key goal of the course is to teach students a model based approach to prediction.

In particular, we will focus on two key aspects of user (or product) behavior; timing process and counting process. Timing process will focus on ***when*** a user does something (when will a user churn from a firm, when will a product drop out of bestseller list, when will a user adopt a new product and so on). Counting process will focus on ***how many*** items are purchased or how many users are adopting and so on. Time permitting we will also introduce ***choice*** process (Out of a number of choices, which product is chosen). We will also examine issues of sales concentration and models of long tail using these processes.

The course is highly hands-on. With each lecture, we will also be using real world datasets to apply the learning in lectures to a practical problem facing a manager. In some cases, datasets would require students to store it and retrieve the relevant information before any analysis can be performed. In other cases, students may have to collect the data themselves. We will be predominantly using Excel for data analysis. All reading material will be provided before each class.

The course will assume some basic understanding of probability/statistics. While we will review some basic probability models, I expect students are familiar with probability distribution functions and basic calculus. An introductory course in probability is helpful though not required.

Syllabus:

Week 1:

Introduction and a quick overview of some probability concepts including the common distributions that we plan to use throughout the class.

In particular, I would like students to understand the conditional and marginal distributions clearly and apply them into some examples.

Reading material: Any standard probability books should have all these details.

Week 2 and 3:

Understanding the distribution skewness and applying the concepts to policy and business problems. For example, a few items sell a lot while a lot of items sell a little. A smaller proportion of users have significant wealth while most have a little. This is also similar to the idea of popular term called Long-tail.

In many instances, we only know the rank for an outcome variable. For example, we may only know that a particular item in 100^{th} best selling item. Or someone is ranked 30^{th} in income distribution. Amazon, Apple App stores etc are good examples of firms publishing the "rank" information but not providing any details on sales or downloads.

1. We will then try to predict the "sale" of an item if we know the "rank".

Reading Material:
"*Inferring App Demand from Publicly available data*",
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1924044

Datasets: We will be using Amazon book sale rank and App Store data for hands on exercises on how to predict demand.

Week 3-4
First, we will spend time on log likelihood and how to write the likelihood to estimate parameters of distribution. We will do couple of example including one example of book rank to see usual regression and likelihood function estimates.

Timing Process:

Our focus will be to model customer's behavior using probability models and then using those models to predict future behavior.

We will start with customers' decision to churn. Every year, firm loses some customers. Our goal will be to predict how long a customer stays with the firm? Or, what is the probability that the customer will churn in a given period. While we start with an individual level customer, the data is usually available only at the aggregate level. So we aggregate the individual level models and create a model that will predict customer survival over time.

The same models can be applied to customer adoption decision as well. Firms releasing a new product or service and use these models to predict adoption rates over time.

Finally, we will extend these models to more general "duration dependence" or "hazard" models. These models are widely used to apply to variety of context. For example, how long does an album last in top 100 billboard? Or, how quickly does a software vendor release a patch for a vulnerability?

Reading Material:

Peter Fader, Bruce Hardie (2007), "How to project Customer Retention", Journal of Interactive Marketing, 21(1). Paper is available on the blackboard.

Pradeep K. Chintagunta, Xiaojing Dong, "Hazard Models in Marketing" (paper is attached).

Dataset:

Customer iPhone adoption data.
Album survival data on Billboard top 100.

Week 5-6

Count Process

Our focus will now be on the count process. Many instances in real world involve count data; number of accidents, number of products purchased, number of tweets sent.

As before, we will first start with customer decision making process to generate a model and then aggregate at the market level to derive some predictions.

We will do couple of hands-on example. The first one will involve customer calls at a call center and the second one will involve automobile insurance rates.

So far, we have only used data at aggregate level. I will try to introduce a framework for using individual specific data at this point. For example, we will try to include gender or age information in insurance example to understand how the premiums rates could be different across these segments.

Reading Material:
Vanasse Charles and George Dionne, “A Generalization of Automobile Insurance Rating Models: The Negative Binomial Distribution with a Regression Component” (Paper attached on blackboard).

Datasets:

Customer Calls to a customer service call center of a large firm.
Number of Accidents and Insurance premium (simulated dataset).

Week 7

Choice process:
If time permits we will cover choice process where people choose one or more given the choice set.

We will try to complete the remaining discussion from previous lectures and spend the last day of the class for review session.