



The Hebrew University of Jerusalem

Syllabus

fintech analytics- credit risk - 55790

Last update 13-02-2020

HU Credits: *1*

Degree/Cycle: *2nd degree (Master)*

Responsible Department: *Business Administration*

Academic year: *0*

Semester: *2nd Semester*

Teaching Languages: *English*

Campus: *Mt. Scopus*

Course/Module Coordinator: *Prof ROGER STEIN*

Coordinator Email: steinr@stern.nyu.edu

Coordinator Office Hours:

Teaching Staff:

Prof ROGER STEIN

Course/Module description:

This short seminar focuses on the practical challenges that arise in implementing a variety of data-driven models for discrete choice problems in finance, as well as several frameworks for thinking about developing FinTech platforms that use these building blocks.

Data-driven methods for predicting discrete choice are now in wide use in both traditional banks and new financial firms platform companies. However, machine learning and statistical algorithms are only a small part of what is involved in building robust analytics. And analytics, and technology more generally, form only one of the building-blocks for successful FinTech businesses.

Although the approaches we describe are applicable to a wide-variety of financial and insurance problems, throughout this short seminar, we will use default (e.g., bankruptcy and default models retail and commercial entities) as a prediction as a working example throughout this short seminar. With a focus on large data sets, we explore a number of data-driven approaches to modeling binary outcomes. I will draw heavily on my experiences building and evaluating some of the most widely used and commercially successful data-driven credit evaluation tools in the industry. This seminar will tend heavily towards discussions of practical model implementations and the “frictions” that make these implementations difficult in real-world settings. We pay special attention to validating discrete-choice models in real-world settings.

We will take the view that an effective, practical modeling framework will sometimes be rough around the edges with the odd inconsistency (usually to deal with available data or the lack thereof). This implies that seemingly incompatible models can each have value in specific contexts, resulting in retention of several models despite their theoretical inconsistency. Because the focus is applied, we will discuss model validation and calibration in detail and highlight data issues in estimation and validation.

Lectures will focus on conceptual themes and practical issues, with much of the technical detail underlying these to be found in the readings.

I will also provide suggested “mini-projects” for those students who are more technically inclined. These projects serve to provide motivation and, if you do them, you will leave the seminar with some very useful tools for applying this subject matter in practice. These “mini-projects” are described using R syntax, though they

may be implemented in any language in which you work (Python, SAS, Matlab, etc.). I will go over a “solution” to at least one of these during the seminar.

Course/Module aims:

To expose students to the practical challenges associated with building and testing data-driven discrete choice models introducing several of the modeling techniques that can be used to build them and to provide a framework for building robust fintech platforms that use these tools.

Learning outcomes - On successful completion of this module, students should be able to:

be familiar with modeling techniques that can be used to build them and to provide a framework for building robust fintech platforms that use these tools.

Attendance requirements(%):

Teaching arrangement and method of instruction:

Course/Module Content:

Day 1

Thursday

March 26,

2020

Introduction to discrete choice models, credit risk modeling concepts, the challenge of data analytics and the nature of FinTech platforms

- How can we add value in developing data-driven analytics

- Data problems and resolutions

- Key components of credit risk – PD, LGD, (EAD), correlation, size

ACPMIP: Chapter 1, pp. 2-16; 19-23; 32-34; 38; 42-43. Chapter 2, pp. 60-62; 72-74.

Supplemental readings:

- Dhar, V. and

Day 2

Friday

March 27,

2020

Introduction to PD model validation

- The role of trust in FinTech
- Validating model power using ROC curves
- Validating model calibration using probability-based measures

ACPMIP: Chapter 7, pp. 361-397.

• Dhar, Vasant. (2016) "When to Trust Robots with Decisions, and When Not To", Harvard Business Review. May.

• Stein, R. M., A. E. Kocagil, J. Bohn and J. Akhavain (2003). "Systematic and Idiosyncratic Risk in Middle-Market Default

Prediction: A Study of the Performance of the RiskCalc and PFM Models." Moody's KMV.

For those interested, see if you can write this tool: Function to calculate the AUC ROC for two different subsets of a single data set.

• Definition: `subROC<-function(x, split.val, split.on, score, outcome,...)` where
o `x` is a dataframe

o `split.val` is a scalar, factor value, date or string used to divide the data frame

o `split.on` is a vector of length `nrow(x)` of the same type as `split.val`; `split.on <= split.val` goes to one data subset

while the remainder goes to the other

o `score` is a numerical vector of length `nrow(x)` for calculating the ROC AUC

o `outcome` is a binary numerical vector of length `nrow(x)` for calculating the ROC AUC

o ... additional parameters

• Return: The function should return a list with three elements:

o A vector of length 2 with the ROC AUC for each data subset.

o A vector of length `nrow(x.subset1)` giving the indices of `x` that for the records included in `x.subset1`

o A vector of length `nrow(x.subset2)` giving the indices of `x` that for the records included in `x.subset2`

• Tasks:

• Implement `subROC`

• Describe how you would make `subROC` more general so that it could take in an arbitrary one or two variable statistic (function) as an input and return the appropriate data

Day 3

Sunday

March 29,

2020

Data driven models for discrete choice problems

• Discrete choice models

• Survival models

Tree-based models

• CART

• RandomForests

ACPMIP: Chapter 4, pp. 183-215, 238-252.

Supplemental readings:

-
- Dhar, V. and Stein, R. (1997), *Seven Methods for Transforming Corporate Data into Business Intelligence*, Ch. 10.
 - Friedman, Hastie and Tibshirani (2013), *Elements of Statistical Learning*, Section 9.2, pp. 305-311.

Day 4

Thursday

April 2,

2020

Introduction to discrete choice model calibration

- *Calibrating to empirical data using calibration curves*
- *Adjusting for differing baseline rates*
- *Mapping between ordinal scales and PDs and back again*
- *The features of successful of FinTech platforms*

ACPMIP: Chapter 4, pp. 215-233.

Supplemental readings:

- Dhar, V. and Stein R. M. (2017) "Economic and Business Dimensions on FinTech Platforms and Strategy.", *Communications of the ACM*, 60, 10, October, pp. 32-35.

• Stein, R. M., A. E. Kocagil, J. Bohn and J. Akhavan (2003). "Systematic and Idiosyncratic Risk in Middle-Market Default

Prediction: A Study of the Performance of the RiskCalc and PFM Models." *Moody's KMV*.

For those interested, see if you can write this tool:

- *Function to create a calibration curve mapping a variable to a default rate.*
- *Function to use the calibration curve to map a variable to a default rate (including interpolation).*

• *Definition: estimateCalibCurve<-function(x, outcome, k, ...) where*

o x is a vector of model scores

o outcome is a binary numerical vector of length length(x)

k is a scalar, denoting the number of "buckets" to use in the mapping

• *Return: The function should return a list with three elements:*

o A list containing

§ map a dataframe of length k with two columns

• *the cutoff (on the same scale as x)*

• *the mapped PD corresponding to the cutoff*

§ baseline a scalar containing the baseline PD for outcome

• *Definition: applyCalibCurve <-function(x, map, baseline=&eq;NULL, ...) where*

o x is a vector of model scores

o map is a dataframe returned in map by buildCalibCurve

baseline is a scalar, to be used if baseline adjustment is to be applied after mapping

• *Return: The function should return a vector of length length(x) containing the mapped PD for each element of x .*

• *Tasks:*

• *Implement estimateCalibCurve*

-
- *Implement applyCalibCurve*

Required Reading:

Bohn, J. R. and R. M. Stein, (2009) Active Credit Portfolio Management in Practice, NY, Wiley. (ACPMIP).

Additional Reading Material:

1-2 papers per course session may be recommended (see outline).

Grading Scheme:

Additional information:

This seminar is a highly compressed version of a full-semester course I give. Much of that course focuses on implementing the ideas through hands-on data science projects with large data sets. While I will allude to this work from time-to-time, much of the technical detail will be omitted. No programming is required for this seminar.