Predictive Analytics Exam—December 2018

UPDATED: November 7, 2018

The Predictive Analytics exam is administered as a five-hour project requiring analysis of a data set in the context of a business problem and submission of a written report. Candidates will have access to a computer equipped with Microsoft Word, Microsoft Excel, R, and RStudio.[1] The report will be submitted electronically. For additional details, please refer to the Exam PA home page.

Exam PA assumes knowledge of probability, mathematical statistics, and selected analytical techniques as covered in Exam P (Probability), Exam SRM (Statistics for Risk Modeling), and VEE Mathematical Statistics. Credit for Exam SRM is required to take the PA exam. Candidates who have received transition credit for Exam SRM will be eligible to take Exam PA but should note that knowledge of the Exam SRM learning objectives will be assumed.

Please check the Updates section on this exam's home page for any changes to the exam or syllabus.

The learning objectives and outcomes provided on the following pages follow the nine modules of the e-Learning support provided. Weights are not provided. Rather, weights will be attached to the grading rubric that appears later in this syllabus.

Recognized by the Canadian Institute of Actuaries

[1] The Prometric computers will have the 2016 versions of Microsoft Word and Excel, version 3.5.0 of R, and version 1.1453 of RStudio.

LEARNING OBJECTIVES

1. **Predictive Analytics Problems and Tools**
Learning Objectives
The Candidate will be able to articulate the types of problems that can be addressed by predictive modeling and be able to work with RStudio to implement basic R packages and commands.
Learning Outcomes
The Candidate will be able to: a) Understand the different types of predictive modeling problems. b) Write and execute basic commands in R using RStudio.

2. **Topic: Problem Definition**
Learning Objectives
The Candidate will be able to identify the business problem, how the available data relates to possible analyses, and use the information to propose models.
Learning Outcomes
The Candidate will be able to: a) Translate a vague question into one that can be analyzed with statistics and predictive analytics to solve a business problem. b) Consider factors such as available data and technology, significance of business impact, and implementation challenges to define the problem.

3. **Topic: Data Visualization**
Learning Objectives
The Candidate will be able to create effective graphs in RStudio.
Learning Outcomes
The Candidate will be able to: a) Understand the key principles of constructing graphs. b) Create a variety of graphs using the ggplot2 package.

4. **Topic: Data Types and Exploration**
Learning Objectives
The Candidate will be able to work with various data types, understand principles of data design, and construct a variety of common visualizations for exploring data.
Learning Outcomes
The Candidate will be able to: a) Identify structured, unstructured, and semi-structured data. b) Identify the types of variables and terminology used in predictive modeling. c) Understand basic methods of handling missing data. d) Implement effective data design with respect to time frame, sampling, and granularity. e) Apply univariate and bivariate data exploration techniques.

5. **Topic: Data Issues and Resolutions**

Learning Objectives

The Candidate will be able to evaluate data quality, resolve data issues, and identify regulatory and ethical issues.

Learning Outcomes

The Candidate will be able to:

a) Evaluate the quality of appropriate data sources for a problem.

b) Identify opportunities to create features from the basic data that may add value.

c) Identify outliers and other data issues.

d) Handle non-linear relationships via transformations.

e) Identify the regulations, standards, and ethics surrounding predictive modeling and data collection.

6. **Topic: Generalized Linear Models**

Learning Objectives

The Candidate will be able to describe and select a Generalized Linear Model (GLM) for a given data set and regression or classification problem.

Learning Outcomes

The Candidate will be able to:

a) Implement ordinary least squares regression in R and understand model assumptions.

b) Understand the specifications of the GLM and the model assumptions.

c) Create new features appropriate for GLMs.

d) Interpret model coefficients, interaction terms, offsets, and weights.

e) Select and validate a GLM appropriately.

f) Explain the concepts of bias, variance, model complexity, and the bias-variance trade-off.

g) Select appropriate hyperparameters for regularized regression.

7. **Topic: Decision Trees**
Learning Objectives
The Candidate will be able to construct decision trees for both regression and classification.
Learning Outcomes
The Candidate will be able to: a) Understand the basic motivation behind decision trees. b) Construct regression and classification trees. c) Use bagging and random forests to improve accuracy. d) Use boosting to improve accuracy. e) Select appropriate hyperparameters for decision trees and related techniques.

8. **Topic: Cluster and Principal Component Analyses**
Learning Objectives
The candidate will be able to apply cluster and principal components analysis to enhance supervised learning.
Learning Outcomes
The Candidate will be able to: a) Understand and apply *K*-means clustering. b) Understand and apply hierarchical clustering. c) Understand and apply principal component analysis.

9. **Topic: Communication**
Learning Objectives
The Candidate will be able to effectively communicate the results of applying predictive analytics to solve a business problem.
Learning Outcomes
The Candidate will be able to: a) Develop and justify a recommended analytics solution. b) Communicate in a clear and straightforward manner using common language that is appropriate for the intended audience. c) Structure a report in an effective manner. d) Follow standards of practice for actuarial communication.

REQUIRED RESOURCES:

e-Learning Modules

Candidates will have access to a series of nine e-Learning modules providing instruction in the objectives stated above. The modules will also provide guidance with regard to knowledge and approaches that will be expected in the assessment. Sample assessments will be included within the modules along with additional readings beyond those listed here.

R and RStudio

Candidates will be expected to be able to work with R within the RStudio environment. For those unfamiliar with the environment, instruction is provided in the first e-Learning Module. Instruction is also provided for setting up the environment to match the packages and versions that will be available on the Prometric computers. For reference, the following packages (and all dependencies) will be available on the Prometric computers. It is not expected that candidates are familiar with each and every one of them. It is expected that candidates can use a selection of these packages to perform the tasks covered in the supporting modules.

boot	data.table	ggplot2	pdp	rpart
broom	devtools	Glmnet	pls	rpart.plot
caret	dplyr	gridExtra	plyr	tidyverse
cluster	**e1071**	ISLR	pROC	Xgboost
coefplot	Gbm	MASS	randomForest	

Note: e1071 was added with the October update to this syllabus. Candidates who have previously set up RSudio with packages frozen at June 1, 2018 should check the updates page for instructions on adding this package.

Study Notes

PA-21-18 Chapter 24, *Healthcare Risk Adjustment and Predictive Modeling*, Second Edition (A link to this study note is found in Module 9 of the e-Learning Predictive Analytics curriculum.)

Textbooks

There are four texts required for the course. Two of them are the texts for the Statistics for Risk Modeling Exam. These are listed below this paragraph. It is assumed that candidates are familiar with all this material. Explicit reference to parts of these texts may be made from time to time within the modules. However, that does not imply the other sections are unimportant.

Regression Modeling with Actuarial and Financial Applications, Edward W. Frees, 2010, New York: Cambridge. ISBN: 978-0521135962.

Chapter 1 – Background only
Chapter 2 – Sections 1-8
Chapter 3 – Sections 1-5
Chapter 5 – Sections 1-7
Chapter 6 – Sections 1-3
Chapter 7 – Sections 1-6
Chapter 8 – Sections 1-4
Chapter 9 – Sections 1-5
Chapter 11 – Sections 1-6
Chapter 12 – Sections 1-4
Chapter 13 – Sections 1-6

An Introduction to Statistical Learning, with Applications in R, James, Witten, Hastie, Tibshirani, 2013, New York: Springer. A PDF of the text can be downloaded at www.statlearning.com.

Chapter 2 – Sections 1-3
Chapter 3 – Sections 1-6
Chapter 4 – Sections 1-3 and parts of 6 (referenced in Module 6, not on SRM)
Chapter 5 – Sections 1 and 3 (excluding 5.3.4)
Chapter 6 – Sections 1-7
Chapter 8 – Sections 1-3
Chapter 10 – Sections 1-6

While the exercises in the texts are not considered required readings, candidates are encouraged to work them as part of the learning experience.

The two additional texts required for this exam are shown below. The indicated chapters are referenced in the modules. However, not all of each chapter is required. The module content will provide further guidance.

R for Everyone, 2nd ed. Lander, 2017, Boston: Addison-Wesley, ISBN 978-0-13-454692-6.

Chapters 1-10, 14, 20, 23, 26, and 28

Data Visualization: A Practical Introduction, Healy, 2018, Princeton University Press. This book is expected to be available for purchase in late 2018. For now, it is available as web pages at http://socviz.co/. Note that this version can be viewed only on the web, there is no PDF version to download.

Chapters 1-4

"Cheat Sheets"

Two such sheets will be available at the examination. They will be available onscreen for viewing at any time. Candidates cannot bring copies into the examination room nor will hard copies be provided. The two sheets are:

- Base R
- Data Visualization with ggplot2

Sample Project

A sample project named "Student Success" has been constructed and it can be downloaded using the following links.

- Student Success project statement
- Student Success data file
- Student Success report template
- Student Success rmd template
- Student Success sample solution
- Student Success sample solution rmd file

Because a wide variety of topics and models are presented in this course, neither the exam nor any sample project can cover them all. It would not be appropriate to infer from this sample project that similar items will be covered on the exam. This sample project is designed to be illustrative of the nature of the materials that will be provided and the length and structure of a high-quality solution.

A Note About Shortcut Keys

Due to security issues, many shortcut key combinations will be disabled on the Prometric computers. Click here to obtain a list of disabled shortcut keys.

Grading Rubric

This rubric is designed to provide candidates with a general idea of what is expected on a successful project submission. It also includes weight ranges indicating the relative emphasis the grading process will place on each section. Note that the division is with respect to the modeling process, not with respect to individual modeling techniques, which will vary from session to session as it will not be possible to construct a project that requires every technique to be applied. Finally, while accurate work is important, graders will place more emphasis on the modeling process employed and how it is described and justified than they will on the correctness of the execution of the analysis.

Communication (30-40%)

- Executive summary – clearly and concisely written summary that is appropriate for someone who reads nothing else
- Problem statement – clearly defines the problem and its business context
- Use of tables and graphs – clearly constructed, labeled, and referenced
- Interpretation of model results – relates the results of the modeling process to the problem statement
- Audience – sections tailored to the audience as described in the project statement
- Appendices – appropriate material relegated to the appendices and properly referenced in the main document
- Code – easy to follow, using intuitive variable names and sufficient comments

Data Exploration and Feature Selection (15-25%)

- Description of the data – summary statistics and graphs with interpretation
- Identification of issues and corrective steps – includes handling missing data and possible transformations
- Selection of features for use in the model – includes creating new features through transformations, clustering, or principal component analysis as appropriate.
- Code – successfully runs and produces output presented in this part of the report

Model Selection and Construction (40-50%)

- Selection and justification of model type – relates model choice to the business problem and the available data
- Estimation of model parameters and hyperparameters, with explanation – calibrates the selected model, including selecting features from the list previously established
- Validation of the selected model – documents that an appropriate validation method was used and provides an estimate of model accuracy using previously unseen data
- Description of selected model – describes the model in appropriate terms for the stated audience
- Code – successfully runs and produces output presented in this part of the report