Predictive Analytics Exam—December 2021

The Predictive Analytics exam is administered as a five hour and fifteen minute exam requiring analysis of a data set in the context of a business problem and submission of written responses to specified tasks. There is no scheduled break for candidates. The additional fifteen minutes is included to allow for breaks, if desired. 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), VEE Mathematical Statistics, and Exam SRM (Statistics for Risk Modeling).

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. The ranges of weights shown in the Learning Objectives below are intended to apply to the large majority of exams administered. On occasion, the weights of topics on an individual exam may fall outside the published range. Candidates should also recognize that tasks often cover multiple learning objectives, including some weight for communication in most tasks.

Recognized by the Canadian Institute of Actuaries

[1] The Prometric computers will have the 2016 versions of Microsoft Word and Excel, version 4.1.0 of R, and version 1.4.1717 of RStudio.

LEARNING OBJECTIVES

1. **Predictive Analytics Problems and Tools (0-10%)**
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 (0-10%)**
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 (0-10%)**
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 (5-15%)**
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 (5-15%)**
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 (20-30%)**
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 (10-20%)**

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 (0-10%)**

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 (15-25%)**
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 regarding knowledge and approaches that will be expected in the exam.

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. 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	

Working with the Prometric versions of R, RStudio, and packages

Candidates are likely to have more recent versions of these items on their computers. As far as learning and practicing is concerned, the versions used should make no difference as long as the packages are compatible with the version of R being used. There are two known exceptions.

Versions 3.6.0 of R or later use a different random number generator than earlier versions. This can affect output from several functions. All of the examples in the e-Learning modules as well as the model solutions for past exams prior to December 2020 were created using version 3.5.0 of R. Candidates

using a later version of R can force it to use the older generator by executing the following command at the start of an R session: RNGkind(sample.kind = "Rounding"). The command will be in effect for the remainder of the session but needs to be run again each time a new session is started.

Versions 4.0.0 of R and later have a change to the read.csv() function. The default is to interpret non-numeric data as character, not factor, variables. All files used in the modules and exams assume the data are interpreted as factor variables. To force this interpretation when running 4.0.0 or later, run the following command at the beginning of an R session: options(stringsAsFactors = TRUE). An alternative is to add the following to the read.csv function: read.csv(file = "filename", stringsAsFactors = TRUE).

The Prometric R environment will have package versions that were in effect on July 1, 2021. If you would like to work in a matching environment, the following is a way to create it using RStudio Cloud. (Depending on when you access Module 1, you may see older instructions that have not been updated. Instructions on how to configure your own computer for this updated environment will not be available within the modules. Also, the link to the appropriate RStudio Cloud project may be outdated. If those module pages have yet to be removed or updated, please ignore them.)

> RStudio Cloud is a web-based option for running R and RStudio. To use it, the first step is to establish an RStudio Cloud account (free) at https://rstudio.cloud/. Then use this link to access the Prometric Environment project: https://rstudio.cloud/project/2726909. To save this project to your account, click on "Save a Permanent Copy". This project will then be available at any future login. The workspace has one data set and one rmd file pre-loaded. Additional files can be added by clicking the upload button on the files tab in the lower right workspace. Files can be downloaded by selecting them, clicking on "More" on the files tab, and then selecting "Export".

The SOA cannot guarantee the continued existence of RStudio Cloud or that it will remain free for limited use. Finally, while every attempt has been made to recreate the Prometric environment, the SOA cannot guarantee that it is an exact match to what will be at the testing center. Questions about using RStudio Cloud can be addressed to rstudio@soa.org.

There is also a way to change your installation to use the older versions. Write to rstudio@soa.org to obtain these instructions. While this approach has been tested on a few computers there is no assurance it will work on your machine and the SOA is limited in the troubleshooting it can provide.

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: 978-0-521-13596-2.

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. There is now a second edition, 2021. PDFs of both editions can be obtained via links found here: https://www.statlearning.com. The material used on this exam is virtually unchanged and so either version may be used. The following table indicates the relevant sections from the two editions.

First Edition	**Second Edition**
Chapter 2 – Sections 1-3	Chapter 2 – Sections 1-3
Chapter 3 – Sections 1-6	Chapter 3 – Sections 1-6
Chapter 4 – Sections 1-3 and parts of 6 (referenced in Module 6, not on SRM)	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 5 – Sections 1 and 3 (excluding 5.3.4)
Chapter 6 – Sections 1-7	Chapter 6 – Sections 1-5
Chapter 8 – Sections 1-3	Chapter 8 – Sections 1-3 (excluding 8.2.4, 8.2.5, and 8.3.5
Chapter 10 – Sections 1-6	Chapter 12 – Sections 1-2, 4-5 (excluding 12.5.2)

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, but not all of each chapter is required. The modules will provide further guidance.

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

Chapters 1-10, 14, 26 and 28

Data Visualization: A Practical Introduction, Healy, 2018, Princeton University Press, ISBN 978-0-691-18162-2. This book may be 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. The author has indicated that at some point the web version may be discontinued

Chapters 1-4

"Cheat Sheets"

Two such sheets will be available onscreen at the examination 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 and Data Visualization with ggplot2

Past Exams and Model Solutions

Past exams and model solutions can be found here. The format change seen in the June 2021 exam will be applied going forward.

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.