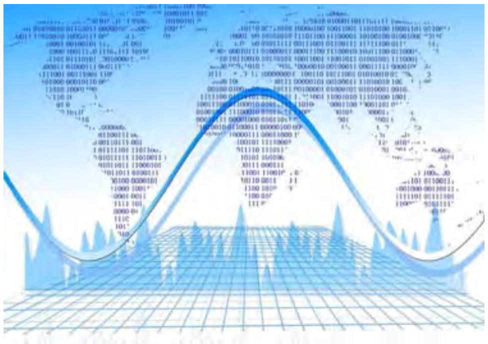
SOA Predictive Analytics Initiatives

KEN GUTHRIE
Managing Director, Education
Society of Actuaries

Predictive Analtyics

- Also known as
  - Data Science
  - Big Data
  - Statistics, but with better marketing



What has Changed? New Data Sources

- Information about insured individuals
  - Driving patterns via telematics
  - Exercise habits via fitness monitors

- Ancillary data
  - Social media
  - Credit reports

What has Changed? New Questions

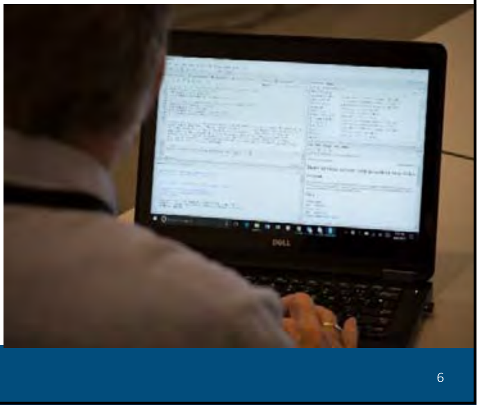
- Previous
  - How many people do we expect to die/lapse/be hospitalized…?
- Now
  - What factors relate to persistence?
  - What factors relate to purchasing additional products?
  - What factors drive claims experience?

What Has Changed? Software

- Previous
  - Expensive
  - Proprietary
- Now
  - Free
  - Open Source
  - Specialized (e.g., ChainLadder package)

Challenges for Actuarial Societies

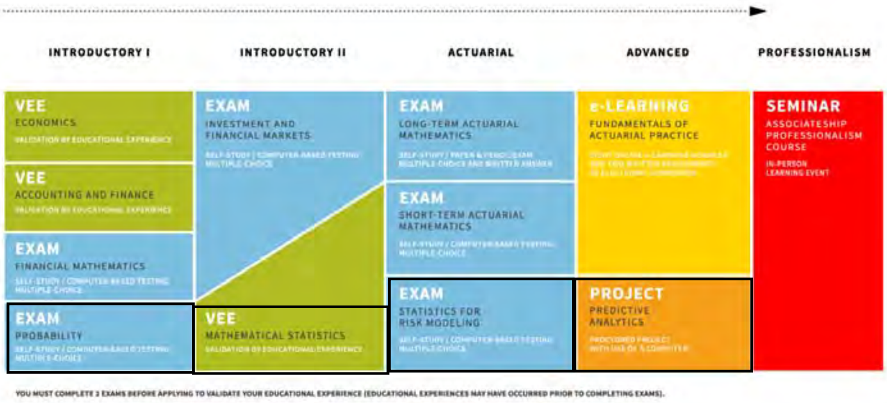
- Methods/Models to Cover
  - GLM, PCA, SVM, GAM, MARS, …
- Related skills
  - Data preparation
  - Exploratory data analysis
  - Feature selection
  - Communication



Challenges for Actuarial Societies

- Educating candidates and members
  - Read a book? OR
  - Have directed practice
- Assessing candidates and members
  - Need to assess more than recollection of facts
  - Must verify that an individual can conduct an analysis from start (data and understanding the problem) to finish (a formal report of findings)

Society of Actuaries Solution - Candidates



INTRODUCTORY I

VEE
ECONOMICS
VALIDATION BY EDUCATIONAL EXPERIENCE

VEE
ACCOUNTING AND FINANCE
VALIDATION BY EDUCATIONAL EXPERIENCE

EXAM
FINANCIAL MATHEMATICS
SELF-STUDY / COMPUTER-BASED TESTING
MULTIPLE-CHOICE

EXAM
PROBABILITY
SELF-STUDY / COMPUTER-BASED TESTING
MULTIPLE-CHOICE

INTRODUCTORY II

EXAM
INVESTMENT AND FINANCIAL MARKETS
SELF-STUDY | COMPUTER-BASED TESTING
MULTIPLE-CHOICE

VEE
MATHEMATICAL STATISTICS
VALIDATION BY EDUCATIONAL EXPERIENCE

ACTUARIAL

EXAM
LONG-TERM ACTUARIAL MATHEMATICS
SELF-STUDY / PAPER & PENCIL EXAM
MULTIPLE-CHOICE AND WRITTEN ANSWER

EXAM
SHORT-TERM ACTUARIAL MATHEMATICS
SELF-STUDY / COMPUTER-BASED TESTING
MULTIPLE-CHOICE

EXAM
STATISTICS FOR RISK MODELING
SELF-STUDY / COMPUTER-BASED TESTING
MULTIPLE-CHOICE

ADVANCED

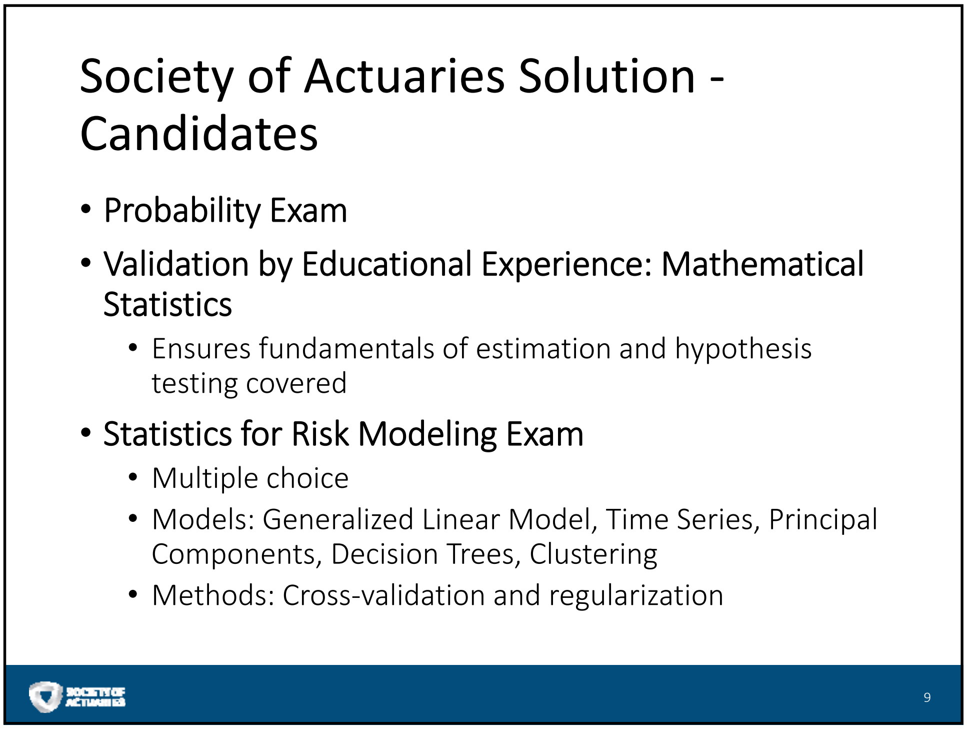
e-LEARNING
FUNDAMENTALS OF ACTUARIAL PRACTICE

PROJECT
PREDICTIVE ANALYTICS
PROCTORED PROJECT
WITH USE OF A COMPUTER

PROFESSIONALISM

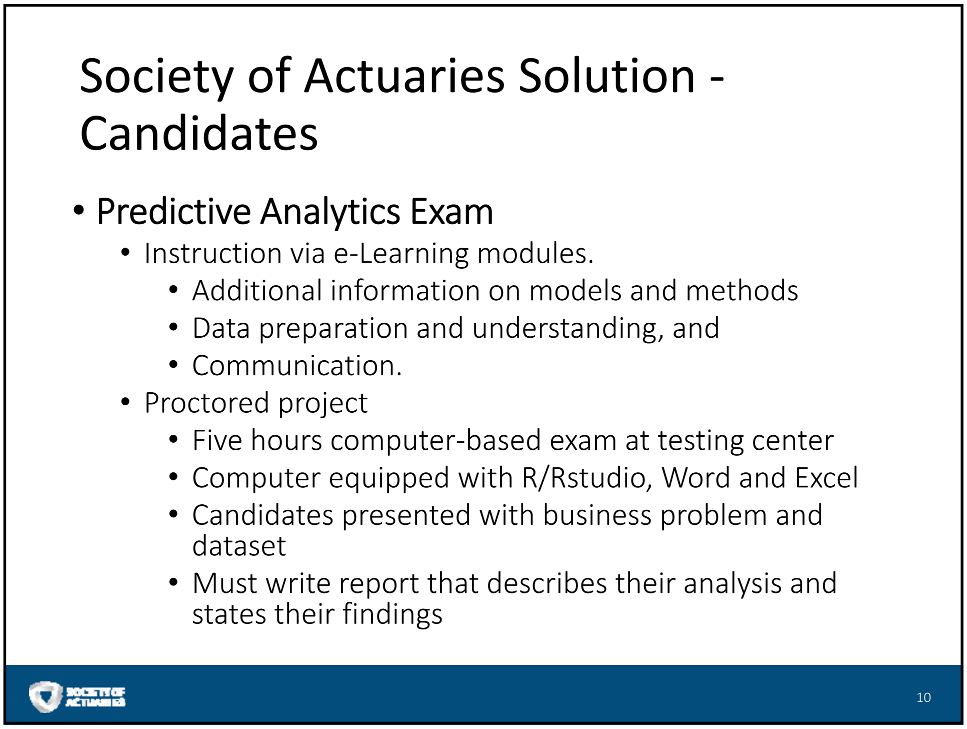
SEMINAR
ASSOCIATESHIP PROFESSIONALISM COURSE
IN-PERSON
LEARNING EVENT

YOU MUST COMPLETE 2 EXAMS BEFORE APPLYING TO VALIDATE YOUR EDUCATIONAL EXPERIENCE (EDUCATIONAL EXPERIENCES MAY HAVE OCCURRED PRIOR TO COMPLETING EXAMS).



Society of Actuaries Solution - Candidates

- Probability Exam
- Validation by Educational Experience: Mathematical Statistics
  - Ensures fundamentals of estimation and hypothesis testing covered
- Statistics for Risk Modeling Exam
  - Multiple choice
  - Models: Generalized Linear Model, Time Series, Principal Components, Decision Trees, Clustering
  - Methods: Cross-validation and regularization



Society of Actuaries Solution - Candidates

- Predictive Analytics Exam
  - Instruction via e-Learning modules.
    - Additional information on models and methods
    - Data preparation and understanding, and
    - Communication.
  - Proctored project
    - Five hours computer-based exam at testing center
    - Computer equipped with R/Rstudio, Word and Excel
    - Candidates presented with business problem and dataset
    - Must write report that describes their analysis and states their findings

Society of Actuaries Solution - Candidates

- Predictive Analytics Exam (cont')
  - Fully proctored
    - No internet access
    - Same security measures as for all SOA exams
  - Fully graded
    - Same protocol as fellowship written-answer exam

Society of Actuaries Solution - Members

Certificate Program

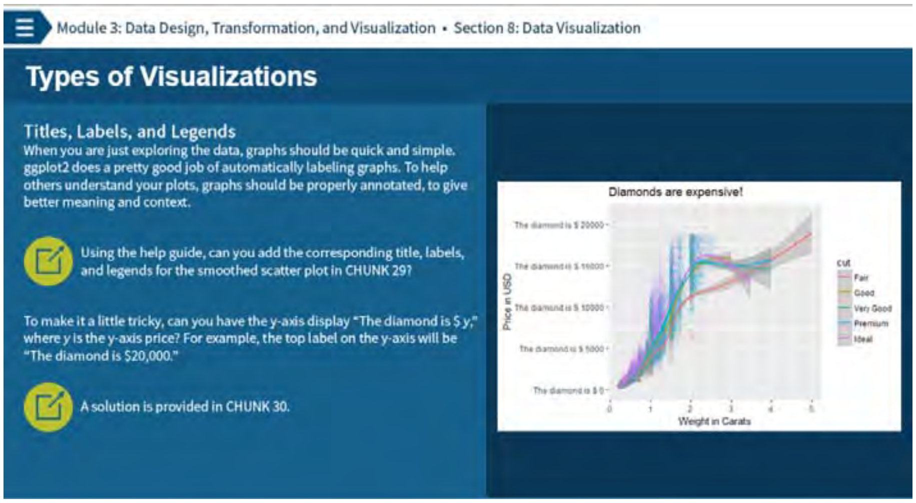
- Open to any credentialed actuary
- Six e-Learning modules
  - 30-40 hours of study time per module
- Two-day seminar with project based assessment
- Self-study portion
  - Discussion forum interactions
  - Exercises
  - Offline Readings
  - Practice
  - End of Module Tests

Society of Actuaries Solution - Members

Certificate Program (cont')

- e-Learning content
  - Predictive Analytics Tools
  - Effective Problem Definition and Project Management
  - Data Design, Transformation and Visualization
  - Data Exploration
  - Feature Generation and Selection
  - Model Development and Validation

Certificate Program Sample Content



Module 3: Data Design, Transformation, and Visualization • Section 8: Data Visualization

Types of Visualizations

Titles, Labels, and Legends

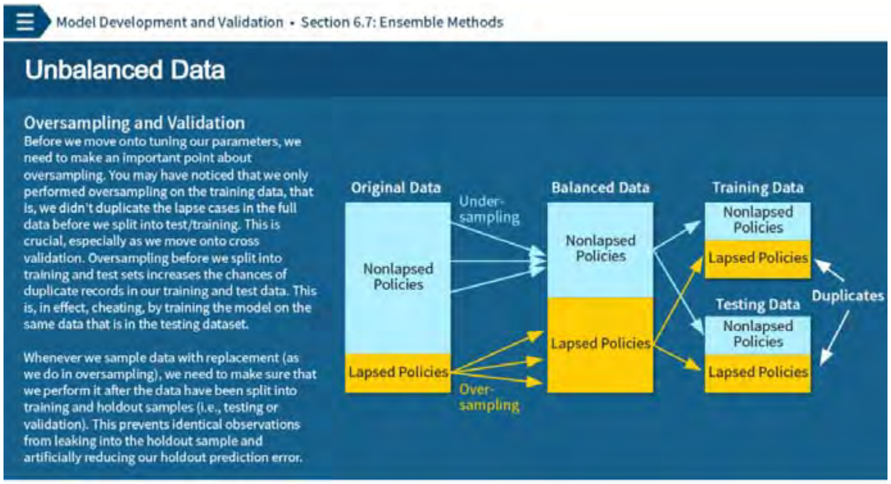
When you are just exploring the data, graphs should be quick and simple. ggplot2 does a pretty good job of automatically labeling graphs. To help others understand your plots, graphs should be properly annotated, to give better meaning and context.

Using the help guide, can you add the corresponding title, labels, and legends for the smoothed scatter plot in CHUNK 29?

To make it a little tricky, can you have the y-axis display "The diamond is $ *y*," where *y* is the y-axis price? For example, the top label on the y-axis will be "The diamond is $20,000."

A solution is provided in CHUNK 30.

Certificate Program Sample Content



Model Development and Validation • Section 6.7: Ensemble Methods

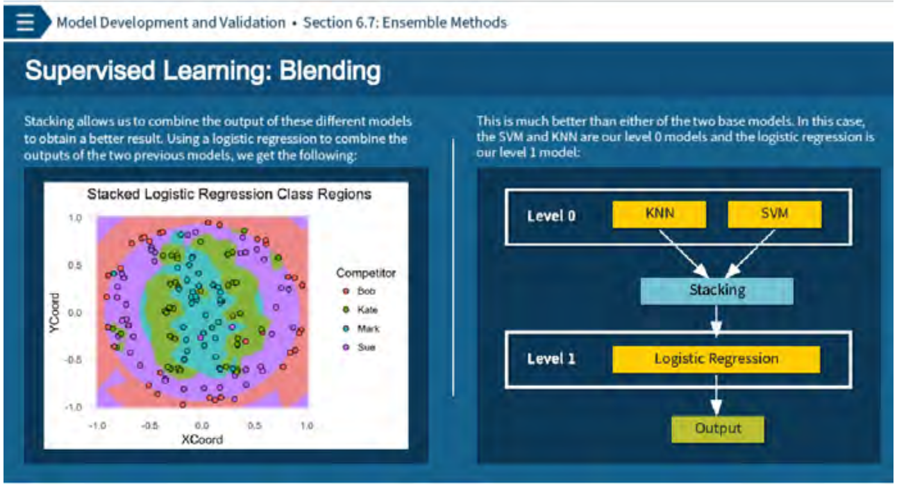
Unbalanced Data

Oversampling and Validation

Before we move onto tuning our parameters, we need to make an important point about oversampling. You may have noticed that we only performed oversampling on the training data, that is, we didn't duplicate the lapse cases in the full data before we split into test/training. This is crucial, especially as we move onto cross validation. Oversampling before we split into training and test sets increases the chances of duplicate records in our training and test data. This is, in effect, cheating, by training the model on the same data that is in the testing dataset.

Whenever we sample data with replacement (as we do in oversampling), we need to make sure that we perform it after the data have been split into training and holdout samples (i.e., testing or validation). This prevents identical observations from leaking into the holdout sample and artificially reducing our holdout prediction error.

Certificate Program Sample Content

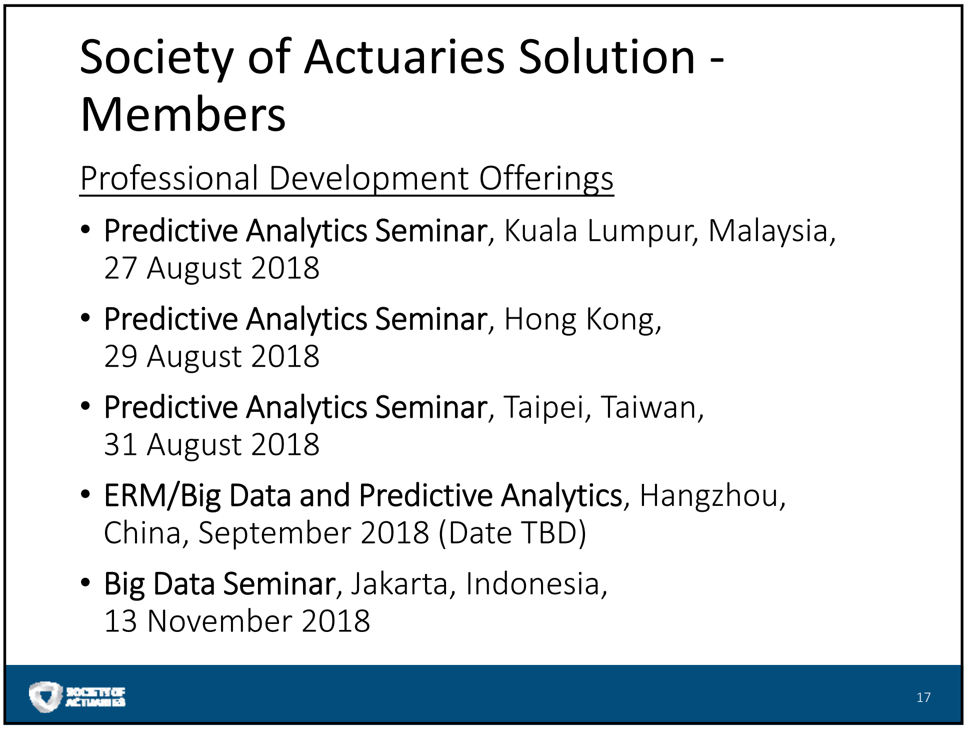


Model Development and Validation • Section 6.7: Ensemble Methods

Supervised Learning: Blending

Stacking allows us to combine the output of these different models to obtain a better result. Using a logistic regression to combine the outputs of the two previous models, we get the following:

This is much better than either of the two base models. In this case, the SVM and KNN are our level 0 models and the logistic regression is our level 1 model:



Society of Actuaries Solution - Members

Professional Development Offerings

- **Predictive Analytics Seminar**, Kuala Lumpur, Malaysia, 27 August 2018
- **Predictive Analytics Seminar**, Hong Kong, 29 August 2018
- **Predictive Analytics Seminar**, Taipei, Taiwan, 31 August 2018
- **ERM/Big Data and Predictive Analytics**, Hangzhou, China, September 2018 (Date TBD)
- **Big Data Seminar**, Jakarta, Indonesia, 13 November 2018