Predictive Analytics Module Topics

Note: Not everything in the required texts for this Exam is covered in the modules. Candidates are responsible for all the text material.

Module 1 – Predictive Analytics Problem Definition

1. Assess whether descriptive, predictive, or prescriptive analytics applies to a business problem
    a. Descriptive: Insights from the past, focus on "what happened?"
    b. Predictive: Focus on the future, "what might happen next?"
    c. Prescriptive: Focus on decision options, "What would happen if I do this?"
2. Describe the characteristics of predictive modeling problems
3. Explain the concepts of bias, variance, model complexity, and the bias-variance tradeoff
    a. Simulate data from a known underlying process and then measure bias and variance using various models
    b. This illustrates the formula from ISLR that splits error into bias, variance, and process error
4. Translate a vague question into one that can be analyzed with statistics and predictive analytics to solve a business problem
5. Consider factors such as available data and technology, significance of business impact, and implementation challenges to define the problem
6. Assess what additional information and next steps would improve the ability to apply predictive analytics to a business problem

Module 2 – Data Design and Visualization

1. Identify structured and unstructured data types
2. Identify the types of variables and terminology used in predictive modeling
    a. Target versus predictor variables
    b. Types of values: Categorical/factor, string, continuous, discrete
    c. Characteristics: Dimensionality, granularity, and ordered
3. Evaluate effective data design with respect to time frame, sampling, and granularity
    a. Time frame: Balancing using older values, which may not be representative
    b. Sampling methods: Random, stratified, systemic, oversampling, undersampling
    c. Granularity: Number of levels for a factor variable
4. Apply the key principles of constructing graphs
    a. This section is based on Chapters 1-4 of the text *Data Visualization: A Practical Introduction*, by Kieran Healy
    b. Graphs are created using ggplot2
    c. Use of the tidyverse and tibbles is not required

5. Apply univariate data exploration techniques
    a. Numerical statistics: Mean, variance, frequencies, percentiles, etc.
    b. Graphic representations: Histograms, boxplots, bar charts for frequency, etc.
6. Apply bivariate data exploration techniques
    a. Categorical versus categorical: Stacked and split bar charts by category
    b. Categorical versus numeric: Stacked and split boxplots and histograms by category
    c. Numeric versus numeric: Scatterplots (that can also be split by category)

Module 3 – Data Transformations and Unsupervised Learning Technics

1. Create features from existing data that may add value
    a. Understand the difference between variables and features
    b. Why feature generation is needed
    c. Transformation: Single function (e.g., log), one variable into many (e.g., binarization, bucketing), many variables into one (e.g., a single sex/smoker variable, clustering)
2. Apply principal components analysis to transform data
    a. Explain the value of PCA
    b. Understand and interpret loadings and calculate PCs
    c. Understand the importance of standardization
    d. Use a scree plot to select the number of components to use
    e. Interpret a biplot
3. Apply *K*-means and hierarchical clustering to transform data
    a. Explain the value of clustering
    b. Understand the two algorithms and interpret output
    c. Understand the importance of standardization
    d. Visualizing the output of clustering (e.g., boxplots for a specific feature split by clusters, scatterplots with clusters identified by color)

Module 4 – Generalized Linear Models

1. Select and validate a GLM as appropriate for a business problem
    a. Selecting the right model (e.g., binary data, count data, non-negative continuous data)
    b. Understand and interpret GLM output
    c. Understand the types of residuals and interpret plots (e.g., residual v fitted, q-q, scale-location, residual v leverage)
2. Apply offsets and weights as appropriate
    a. Understand the role of offsets, how they are incorporated in R, interpreting output and making predictions
    b. Understand the role of weights (and how they differ from offsets), how they are incorporated in R, and interpreting output

3. Interpret model coefficients, including interaction terms
    a. Understand how interaction terms change model interpretation
    b. Make predictions with models that include interaction terms
4. Select appropriate hyperparameters for regularized regression
    a. Understand the purpose and implementation of regularization
    b. Understand the difference between ridge, lasso, and elastic net regression
    c. Use cross-validation to select alpha and lambda

Module 5 – Tree-Based Models

1. Construct, prune, and validate regression and classification trees
    a. Calculate entropy and employ it to create a tree
    b. Calculate Gini impurity and employ it to create a tree
    c. Interpret the summary output from R (rpart)
    d. Understand the control parameters for creating a tree
    e. Understand and employ the complexity parameter and tree pruning
    f. Calculate and interpret performance measures from a confusion matrix (e.g., accuracy, precision, sensitivity)
    g. Calculate and interpret a Receiver Operating Curve and the area under the curve
2. Apply bagging and random forests as appropriate
    a. Understand the advantages of these models and how random forests improve bagging
    b. Fit a random forest model (in this section and many others, fitting against a test set is used to build the model)
    c. Understand the issue with unbalanced data and methods to account for it (undersampling and oversampling)
    d. Tune random forest parameters using cross-validation (caret package)
    e. Interpret feature importance measures
    f. Understand and interpret partial dependence plots
3. Apply boosting as appropriate
    a. Understand the basics of boosting (but rely on software to do the fitting)
    b. Understand the role of the control parameters in a gradient boosting machine (GBM)
    c. Train a GBM (using xgboost and caret)