

# Predictive Analytics Certificate Program

## Learning Objectives

### Module 1: Predictive Analytics Tools

- Determine whether a problem should be addressed with predictive modeling or traditional analysis techniques
- Practice layouts and execute basic commands in R
- Practice writing basic R scripts

### Module 2: Problem Definition and Project Management

- Translate a vague question into one that can be analyzed with data, statistics and machine learning to solve a business problem
- Use case design and evaluation/prioritization based on available data and technology, significance of business impact and/or implementation considerations to define the problem
- Implement technology to efficiently utilize statistical and machine learning techniques taking into account problem objectives and implementation constraints
- Explain the key principles involved in creating and managing an effective predictive modeling team which can successfully manage your project from problem definition to implementation

### Module 3: Data Design, Transformation & Visualization

- Identify structured, unstructured and semi-structured data types
- Identify the types of variables and terminology used in predictive modeling
- Evaluate the quality of appropriate data sources for a problem
- Identify the regulations, standards, and ethics surrounding predictive modeling and data collection
- Utilize lapse, mortality and health datasets in predictive modeling exercises
- Implement effective data design: time frame, sampling, and granularity

- Use common data blending techniques
- Describe and apply common data visualization techniques
- Describe and apply common data transformation techniques
- Reinforce the concepts of Module 3 through practice with a real-world dataset

## Module 4: Data Exploration

- Identify data anomalies and outliers using univariate exploration techniques
- Use bivariate exploration to determine relationships, calculate correlation and investigate conditional means.
- Identify relationships among multiple variables using principle component analysis
- Identify relationships and structure among multiple variables using clustering techniques

## Module 5: Feature Generation & Selection

- Explain the differences between features and variables and apply prior knowledge to create features
- Describe various filter-based selection techniques and then select and apply the most appropriate one
- Apply algorithm-based selection and data mining techniques to generate and select features in a dataset
- Apply various text mining methods in order to generate appropriate features for modeling text data, and list the advantages, disadvantages and limitations of using each method
- Reinforce the feature generation and selection techniques learned in Module 5 through practice.

## Module 6A: Model Development & Validation

- Differentiate types of business problems and understand their impact on model development and validation
- Explain the concepts of bias, variance and model complexity; the bias variance tradeoff and its implications for building robust models
- Learn approaches to assessing model performance and selection of the best model
- Understand the advantages and disadvantages of cross validation and apply it during modeling

- Understand and apply decision trees in the contexts of classification and regression
- Understand and apply generalized linear models to regression problems
- Understand the advantages, disadvantages and limitation of ensemble techniques
- Understand the benefits, limitations and requirements of  $k$ -nearest neighbors

*Note: Completion of Modules 1-6A along with a passing result on the assessment are required to earn the certificate. Module 6B is optional advanced material.*

## Module 6B: Advanced Topics in Model Development & Validation

- Understand the advantages and disadvantages of blending/stacking as an ensemble method and how to avoid using test data during training when using such methods
- Describe the use cases, data requirements and results associated with hyperparameter tuning
- Understand the types of problems addressed by support vector machines, their approach to such problems, and the advantages and disadvantages with respect to alternative models
- Understand the advantages and disadvantages of Bayesian Learning, complete a Bayesian analysis of a basic problem, and discuss the differences between Bayesian and frequentists models
- Understand how Gaussian mixture models provide an alternative to  $k$ -means clustering and how the expectation-maximization algorithm is used to fit these models.
- Understand the types of problems addressed by additive models, how they approach such problems and discuss the advantage and disadvantages with respect to alternative models
- Understand the types of problems addressed by neural networks, their approach to such problems, and the advantages and disadvantages with respect to alternative models
- Understand the types of problems addressed by genetic algorithms, their approach to such problems, and the advantage and disadvantages with respect to alternative optimization models