Version 2.0

Advanced Analytics with R Training Plan (May 2025)

This document provides a comprehensive training program for developing Predictive Analytics, aligned with Jube's R Advanced Analytics Guidance.

JUBE

JUBE

Amendments

Date	Author	Version	Description
1st May 2025	Richard Churchman	2.0	Rewrite of legacy materials for publishing.

JUBE

Contents

JUBE

Introduction

Jube is an open-source transaction monitoring platform that automates machine learning while emphasizing the importance of understanding foundational ML concepts to maintain quantitative competence and demystify fraud prevention systems.

This practical training course focuses on implementing analytical techniques in real business environments, using real-world case studies cantered on transaction monitoring while also exploring other datasets like housing prices, credit risk, and fraud simulation.

Participants will develop skills in data analysis, predictive analytics (regression, decision trees, neural networks), and model optimization using R - the industry-standard, cost-effective statistical language - along with its powerful ecosystem of packages and RStudio's intuitive interface for comprehensive advanced analytics without licensing costs. The course bridges theory and practice, equipping analysts to deploy advanced analytics across financial use cases while maintaining transparency in machine learning processes.

Program Outcomes

By the end of the training, participants will:

- Understand foundational ML & statistical techniques, including linear/logistic regression, decision trees, Bayesian classifiers, neural networks, feature engineering, and model optimization.
- Gain practical implementation skills in R, covering data manipulation, visualization (ggplot2), statistical analysis, and building/evaluating predictive models for fraud, credit risk, and numeric forecasting.
- Explore advanced topics like unsupervised learning (anomaly detection with Bayesian networks & SVMs), Monte Carlo simulation for risk modelling, and model deployment (APIs, reports, operational integration).

Training Plan

The three-day training program covers advances analytics from foundational to advanced topics. On the first day, participants learn the basics of R, data structures, and linear regression, applying them to case studies like the Boston Housing Dataset and stock portfolio analysis. The second day focuses on classification and machine learning, including logistic regression for fraud detection, decision trees (C5.0) for credit risk, and Naive Bayes for handwriting recognition. The final day delves into advanced techniques, such as neural networks, anomaly detection (Bayesian networks, SVMs), Monte Carlo simulations for fraud modelling, and deploying models via APIs (Plumber) and reports. The course provides a comprehensive, hands-on approach to predictive analytics.

Day 1: Foundations of Advanced Analytics

The first day of the workshop will establish core competencies in numeric prediction through applied linear statistical techniques using real market datasets. Participants will develop practical advanced analytics skills focused on foundational linear methods for numeric value forecasting, with an emphasis on hands-on implementation rather than theoretical discussion.

The training will cover essential linear regression techniques and basic statistical modelling approaches, all directly applied to actual data scenarios to ensure immediate, tangible skills development in market data analysis and numeric prediction. This practical foundation will enable attendees to competently generate and interpret numeric forecasts while understanding the underlying statistical principles.

Time	Topic	Description
09:00 – 09:40	**Module 1:** Descriptive, Machine Learning, Predictive and Prescriptive Analytics Introduction.	An introduction to advanced analytics to provide foundation for the techniques to be covered on the course: • Creating the Data Driven Organisaton. • Predictive Analytics? • Origins. • Quantitative \ Frequentist. • Acclimatisation. • Process. • Limits. • Types of Predictive Analytics. • Problem Framing and Elicitation. • Review. By the end of the session the participants will have an overview of analytical techniques and a clear methodology to approach projects.
09:40 – 10:10	**Case Study 1:** Elicitation of Variables.	Create an exhaustive list of reasons that a customer may leave a telecommunications provider for the purposes of these variables being modelled by the company's data analysts, then be carried through to predictive analytics.
10:10 – 10:30	Break	Break
10:30 – 11:10	**Module 2:** Getting Started with R.	An R Primer, introducing some of the key concepts in R such that it can be made use of as a centralised advanced analytics application: • Advanced Analytics Tools. • What is R? • R Console and Command Line. • RStudio. • Packages. • Data Types. • Review. By the end of the session participants ill be able to interact with R and understand the core principles of its extensibility.

11:10 – 11:50	**Module 3:** Data Structures.	R has a variety of data structures which underpin the entire Advanced Analytics endeavour. This module will provide a comprehensive overview of these data structures: • Vectors. • Lists. • Data Frames. • Save and Load. • Review. By the end of the session participants will understand the most crucial R data structures and be able to state datasets for complex analysis and advanced analytics.
11:50 – 12:10	Break	Break
12:10 – 12:50	**Module 4:** Loading, Shaping and Merging Data.	R has an extremely wide reach in terms of the data it can fetch and aggregate. This module introduces a variety of data fetch techniques followed by showcasing a means to interact with the data in R: • Methodology. • Numeric Functions. • String and Logic Functions. • Date Functions. • Importing Data. • dplyr Shaping and Moulding. • Exporting. • Review. By the end of the session the participants will be able to look up R and dplyr as a tool for complex data manipulation.
12:50 – 13:50	Lunch	Lunch
13:50 – 14:40	**Module 5:** Summary Statistics and Basic Plots in R.	Advanced Analytics is rooted almost entirely in statistics and an understanding of basic statistics is vital to fully understand the linear and nonlinear predictive analytics techniques to be taught in the course. This module will present techniques to use R to create summary statistics and plots: • What are Statistics. • Data Description. • Frequency. • Histogram: • Range.

		• Tendency. • Spread. • Shape. • Range. • Tendency. • Spread. • Shape. • What is Probability. • Z Score to Probability. By the end of the session participants will be confident using R to rapidly perform summary statistics and perform dataset summarisation.
14:40 – 15:00	**Case Study 5:** Stock Portfolio Selection	Using the one-day price history for the companies on the NYSE, perform an analysis and pick ten stocks which meet an investment criterion to the greatest extent. Allocate cash resources between these stocks at your discretion – being on notice to present \ defend your investment decision.
15:00 – 15:20	Break	Break
15:20 – 15:50	**Module 6:** Abstraction and Transformations (aka Feature Engineering).	Statistics underpin predictive analytics; however, the real power is the shaping and moulding of data for a given problem domain to create datasets that are more enriched and meaningful and thus lead to better predictive analytics models, measured by better predictive or classification accuracy: • Dataset Terminology. • Data Types. • Domain and Data. • IoT. • Databases. • Data Creativity. • Vertical Abstraction. • Horizontal Abstraction. • Statistical Abstraction. • Momentum Abstraction. • Point Step Abstraction. • Standardisation. • Sampling. • Review. By the of the session the participants will have a full comprehension of the importance of feature engineering and have a methodology to generically approach it.

15:50 – 16:30	**Module 7:** Linear Regression.	Linear Regression is an approach to modelling a relationship between variables and a continuous value. Linear Regression can form a simple yet extremely powerful form of predictive analytics: • Relationship Estimation. • Correlation Analysis. • "Stepwise" Correlation. • Line Fitting and Extrapolation. • Linear Regression Output. • Deployment. • Confidence Intervals. • "Stepwise" Linear Regression. • Multicollinearity. • Visual Representation. • Review. By the end of the session the participants will understand how to make predictions on numeric continuous values using an explainable predictive analytics technique.
16:30 – 17:20	**Case Study 7:** Linear Regression.	Use Linear Regression to create a predictive model to predict the house prices in Boston based on a range of factors in the Boston Housing Market Dataset.
17:20 – 17:30	Summing Up	Summing Up

Day 2: Classification & Machine Learning:

The second day of the workshop advances into predictive analytics using machine learning, shifting focus from numeric prediction to classification techniques. While building on linear foundations with logistic regression, the day primarily explores nonlinear methods including decision trees and Bayesian classifiers. Participants will progress from data visualization with ggplot2 through increasingly sophisticated algorithms, applying each technique to real-world cases like fraud detection and credit risk analysis. The curriculum balances theoretical understanding with practical implementation, using R to develop models for handwriting recognition, fraud prevention optimization, and financial risk assessment - equipping attendees with both the technical skills and business context to deploy these methods effectively.

Time	**Topic**	**Description**
09:00 – 09:20	**Recap:** Day 1	A review of day one.
09:20 – 10:00	**Module 8:** Pretty Plots with ggplot2 and rapid, visual, dataset exploration.	Plots are revisited using more advanced R graphics functions. • Introduction to ggplot2. • Line. • Bar.

		• Pie. • Scatter. • Adding a Trend Line. • Colours. • Opposing Datasets. By the end of the session participants will be able to rapidly explore datasets and produce visualisations that summaries datasets and how vectors relate to one another. Visual relationships between vectors provide an important grounding for the same being achieved in statistical measures (e.g. correlations).
10:00 - 11:40	**Case Study 8:** Describing Fraud using Summary Statistics and Charts.	Use Summary Statistics and Charts to identify the difference between fraud and genuine transactions.
11:40 – 12:20	**Module 9:** Logistic Regression.	Regression is revisited, as is the concept of curve fitting, before reviewing how these are ported to R: • Behavioural Analytics. • "Binary Regression" \ Classification. • Logistic Curve. • Symmetric Sampling. • Skew in Behavioural Analytics. • Categorical Data Pivoting. • Profile Abstraction Deviation. • Logistic Regression Output. • "Stepwise" Logistic Regression. • Logistic Regression Deployment. • Logistic Regression Output. • Probability Threshold Function. • ROC Curves. • Review. By the end of the session participants will have extended regression knowhow to facilitate classification.
12:20 – 13:20	**Case Study 9:** Fraud Prevention Model Performance Improvement.	A group case study where each group will attempt to improve the performance of their model.
13:20 – 14:20	Lunch	Lunch

14:20 – 14:50	**Module 10:** R Naive Bayesian Classifiers and Laplace Estimator.	The Bayesian Network packages in R are a particularly powerful means to create automated models. Bayesian Networks, created via automated structure, are a stalwart to machine learning of classification problems and R has some extremely sophisticated packages available. This module aims to introduce the machine learning aspects of Bayesian Networks as explained in the foundation course, ported to R: • Naive Bayesian Network. • Calibration. • Laplace Estimator. • Review. By the end of the session the participants will have used probabilistic methods to perform classification. In addition to classification, probabilistic techniques provide an important grounding for the use in anomaly detection.
14:50-15:20	**Case Study 10:** Slate Fray Handwriting Recognition.	Use Bayesian Networks to create a handwriting recognition classifier.
15:20 – 15:40	Break	Break
15:40 – 16:20	**Module 11:** Splits, Probability and Decision Trees.	Continuing probability-based techniques, this module will produce Decision Trees. Homogenising the concept of Decision Trees with that of Regression, Regression trees will be presented as an underused tool for numerical prediction, albeit predicting within a range: • Factors. • Decision Trees. • Activation Utility. • Reducing Standard Deviation. • Regression Trees. • Maximum Entropy. • C5 Decision Trees. • Boosting \ Ensemble Activation. • Boosting Activation. • Overfitting and Test Datasets. • Review. By the end of the session participants will have extended tooling to create models using

		probabilistic methods that provide high explanatory value given their logical statement evaluation.
16:20 – 17:20	**Case Study 11:** Decision Trees for Credit Risk Analysis.	Using the Credit Risk Dataset, the class will create a decision tree using c5.
17:20 – 17:30	Summing Up	Summing Up

Day 3: Advanced Techniques & Deployment

On the final day, the workshop culminates with advanced predictive analytics in R, focusing particularly on neural networks and model deployment. Participants will explore automated machine learning tools while diving deep into neural network architectures and their applications. The day introduces simulation techniques to generate actionable prescriptions from trained models, bridging the gap between analysis and decision-making. A core emphasis will be placed on operationalizing models - covering the entire pipeline from development to production deployment for real-time inference. This includes best practices for model serving, performance monitoring, and integration with existing systems. Through hands-on exercises, attendees will learn to transition models from experimental prototypes to production-grade solutions capable of handling live transaction monitoring scenarios, completing their journey from statistical foundations to deployable AI solutions.

Time	Topic	Description
09:00 – 09:20	**Recap:** Day 2	A review of day two.
09:20 – 10:00	**Module 12:** Neural Networks.	Neural Networks is a catch all technique for a variety of non-linear machine learning methods. This module explores the most advanced packages and compares their utility: • Artificial Neural Networks. • Visualisation of Regression. • Visualisation of ANN. • Learning. • Training Numeric Prediction. • Symmetric Sampling. • Training Classification. • Review. By the end of the session participants will be able to see how regression topology can evolve into more complex classification models capable of their own abstraction – feature engineering – to the end of bringing about improvements in classification accuracy, especially in the context on nonlinear datasets.
10:00 – 10:40	**Case Study 12:** Improving Logistic	Using a variety of Neural Network based predictive analytics tools for which there are R packages,

	Regression with a variety of Neural Networks in R.	create a predictive model improving upon all preceding models created in the workshop. The groups will present the improvement that they have achieved on their previous models.
10:40 – 10:50	Break	Break
10:50 – 11:30	**Module 13:** Unsupervised Learning with Bayesian Network Anomaly Detection	Up to this point in the course the datasets have all presented a class or dependent variable that is available to train models for future predictions. In transaction monitoring use cases the class variable may not be anything like frequent or available enough, and instead the requirement becomes the reliable identification of anything unusual in the dataset. This session will present the use of probabilistic methods for the purpose of anomaly detection: • Visualization of Anomalies. • The humble Z-Score. • Bayesian Networks. • Bayesian Network Anomaly Detection. By the end of the session participants will be able to detect anomaly on a probabilistic basis.
11:30 – 12:10	**Case Study 13:** Detecting anomaly in Credit Risk applications.	Using a Bayesian Network, create an anomaly detection model for a consumer credit lending, ignoring the class or dependent variable. Detect unusual credit applications.
12:10 – 12:50	**Module 14:** Unsupervised Learning with Bayesian Network Anomaly Detection	A classification model that does not get much attention in this course, but is quite powerful, is the Support Vector Machine (SVM). Support Vector Machines are introduced here given that they can also be applied to one class problems, which can be used for the purpose of Anomaly Detection: • Support Vector Machines. • Create a Support Vector Machine. • Creating a multi class Support Vector Machine with Slate Fray Letters. • One Class Support Vector Machines. • Anomaly detection One Class Support Vector Machine By the end of the session participants will understand an alternative class that is particularly good at nonlinear problems, both for where a class exists, and for the purpose of anomaly detection.

12:50 – 13:30	**Case Study 14:** Detecting anomaly in financial transactions.	Create a two class and one class support vector machine on for transaction monitoring. Contrast classification accuracy in both approaches.
13:20 – 14:20	Lunch	Lunch
14:20 – 15:00	**Module 15:** Monte Carlo Model Search and Prescriptive Analytics.	Monte Carlo Search takes certain statistical assumptions and creates random simulations between these assumptions. Monte Carlo has three uses for the purposes of this course, firstly simulating models before placing them into production, secondarily identifying optimal and overly speculative scenarios and finally creating datasets to create predictive analytics models where data would otherwise not be available: • What is Monte Carlo Simulation? • Mitigation. • Distribution Fitting. • Random Vectors. • Optimization and Prescription. • Probability Density in Activation. By the end of the session participants will be able to perform monte-carlo simulation on models for the purpose of optimization and prescriptions.
15:00 – 15:30	**Case Study 15:** Describe the fraud environment.	Create an accurate Neural Network model in fraud prevention, using H2O and use Monte Carlo Simulation to describe the environment.
15:30 – 15:50	Break	Break
15:50 – 16:50	**Module 16:** Integration to the Operation (Outputting Reports and API's)	This module will explore how to integrate the R and Exhaustive work product inside the enterprise by recalling models to reports or invoking via API for real-time recall. • Exhaustive File Recall. • R File Recall. • Exhaustive API Recall. • R API Recall with Plumber. By the end of the session participants will be able to deploy work product as microservices and embed into real-time processes.
16:50 – 17:10	**Case Study 16:** Pushing Code	Pushing Code to bring together Abstraction and Model Recall.

17:10 – 17:30	Summing Up	Summing Up

Resources

https://jube.io/AdvancedAnalyticsWithR.pdf

Conclusion

This Jube training course on Advanced Analytics for Transaction Monitoring delivers a robust and practical learning experience, equipping participants with essential machine learning skills tailored for real-world financial applications. Over three immersive days, attendees develop a strong command of predictive modelling techniques using R, from foundational statistical methods like linear and logistic regression to advanced approaches such as neural networks and anomaly detection. The hands-on, case-study-driven format ensures that theoretical concepts are immediately applied to practical scenarios, including fraud detection, credit risk analysis, and transaction monitoring.

Beyond model building, the course emphasizes operational readiness, teaching participants how to deploy models effectively through APIs and reporting tools. By combining technical expertise with business context, the training enables professionals to transform raw data into actionable insights that enhance fraud prevention and risk management strategies. Designed for analysts, data scientists, and risk professionals, this program bridges the gap between theory and implementation, ensuring learners leave with the confidence and skills to apply advanced analytics in their organizations.

In an increasingly data-driven financial landscape, mastering these techniques is not just valuable—it's essential. This course provides the tools and knowledge needed to stay ahead in fraud detection and risk analytics, making it a critical investment for professionals looking to leverage machine learning for real-world impact.