



CSXX0265: Predictive Analysis

Course Credit: 3

Prerequisites:

- Fundamental knowledge of statistics and probability
- Basic linear algebra and calculus
- Introduction to data science or machine learning
- Proficiency in using analytical tools (Python/R)

Course Objectives:

1. To understand the process of developing and validating predictive models.
2. To explore the role of data preprocessing in improving model performance.
3. To examine various linear and nonlinear modeling techniques in depth.
4. To interpret predictive performance metrics and address common modeling pitfalls.
5. To investigate variable importance and model simplification strategies.

Course Outcomes

CO No.	Course Outcome	Mapped POs
CO1	Understand the end-to-end process of predictive modeling.	PO1, PO2
CO2	Apply linear, regularized, and tree-based models to predictive tasks.	PO2, PO3
CO3	Evaluate and tune models using appropriate validation and performance metrics.	PO2, PO4
CO4	Interpret results and address modeling challenges including class imbalance.	PO4, PO6, PO7
CO5	Analyze time series data and build appropriate forecasting models.	PO1, PO2, PO4

Syllabus

Unit 1: Predictive Modeling Framework and Data Preprocessing (9 Hours)

- Predictive modeling process: goal, data, models, validation
- Overfitting and the bias-variance trade-off
- Sources of data variation and their impact on modeling
- Data splitting strategies: resampling, cross-validation
- Data preprocessing techniques:
 - Zero-variance and near-zero variance predictors
 - Data transformation (centering, scaling, Box-Cox)
 - Handling skewness and outliers
- Principles of feature engineering

Unit 2: Linear and Regularized Models for Prediction (9 Hours)

- Linear regression theory and modeling process
- Regularization techniques: Ridge, Lasso, Elastic Net
- Variable selection and multicollinearity
- Logistic regression: concepts and interpretation
- Residual analysis and diagnostic evaluation

Unit 3: Nonlinear and Tree-Based Predictive Models (9 Hours)

- Nonlinear regression strategies
- MARS (Multivariate Adaptive Regression Splines)
- Regression and classification trees
- Ensemble methods: Bagging, Random Forests, Boosting
- Introduction to k-NN and SVMs
- Handling class imbalance in classification

Unit 4: Model Evaluation, Tuning, and Comparison (9 Hours)

- Performance metrics: RMSE, R^2 , ROC, AUC, Kappa
- Resampling techniques for evaluation: k-fold, bootstrap
- Model tuning using grid search and cross-validation
- Variable importance and dimensionality reduction
- Model comparison and interpretability trade-offs
- Stability, reproducibility, and deployment considerations

Unit 5: Time Series Prediction (6 Hours)

- Nature of time series data: trend, seasonality, noise
- Stationarity and transformations
- Autocorrelation and partial autocorrelation
- AR, MA, ARMA, and ARIMA models
- Exponential smoothing methods (ETS, Holt-Winters)
- Time series forecasting accuracy: MAE, MSE, RMSE, MAPE

Books

1. James, Gareth, et al. *An Introduction to Statistical Learning with Applications in R*, Springer.
2. Kuhn, Max, and Kjell Johnson. *Applied Predictive Modeling*, Springer.

Provost, Foster, and Tom Fawcett. *Data Science for Business*, O'Reilly Media.