Learning Objectives Exam 3

Predictive Modeling – Methods and Techniques

A. IDENTIFY THE BUSINESS PROBLEM

TOPIC	LEARNING OBJECTIVES	READINGS	PAGES	R PACKAGES AND INSTRUCTIONS
1. Identifying the Business Problem	a. Understand the limitations of modeling, how to leverage models effectively, and what to look for in a dataset. b. Understand the issues of working with observational data versus randomized experimental data c. Understand matching methodology, including propensity scoring, and know why matching methodology is preferred to other algorithms for observational data d. Be able to apply matching methodology to implement/structure a project e. Understand concepts of fairness in risk assessment f. Understand why multiple concepts of fairness are generally incompatible with one another	Luca Stuart & Rubin, Ch. 11 Kleinberg, et al	50	

B. **SET UP THE PROJECT DESIGN**

TOPIC	LEARNING OBJECTIVES	READINGS	PAGES	R PACKAGES AND INSTRUCTIONS
1. Project Design	a. Understand randomized experimental design, including factorial designs, randomized blocks and covariance design. b. Understand quasi-experimental design, including nonequivalent group design and regression-discontinuity design. c. Understand sampling methodology, what "external validity" is, what the threats to it are, and how it can be improved. d. Understand sampling error and the standard percent rules of sampling. e. Be able to describe and apply various probability sampling methods. f. Be able to describe and implement non-probability sampling methods. g. Understand the difference between supervised and unsupervised models.	Thomki & Manzi Trochim, Ch. 2 Trochim, Ch. 9, Sections 9-1,9-2,9-5 Trochim, Ch. 10, Sections 10-1,10-2	53	

C. **EXPLORATORY DATA ANALYSIS**

TOPIC	LEARNING OBJECTIVES	READINGS	PAGES	R PACKAGES AND INSTRUCTIONS
1. Exploratory data Analysis	a. Be able to perform data preprocessing. i. What transformations of the dependent variable should be made, including trend and development ii. What transformations of the independent variables should be made b. Perform truncation and censoring. c. Be able to apply treatments for missing data. d. Be able to Bin variables. e. Perform additional Exploratory Data Analysis on preprocessed data. f. Create test and holdout data.	Chapman & Feit, Ch. 3	30	packages:car, psych, rworldmap, RColorBrewer functions: read.csv, data.frame, c(), dim, rep(), rm(), str(), factor(), head(0, tail(), set.seed(), rbinom(), set.seeed(), rpois, table, plot, min,max, sd, IQR, mad, quantile, names, rownames, summary, apply, hist, boxplot, qqplots, by(), aggregate()

D. **FEATURE SELECTION AND REGULARIZATION**

TOPIC	LEARNING OBJECTIVES	READINGS	PAGES	R PACKAGES AND INSTRUCTIONS
1. Linear models and assessing their accuracy	a. Estimate parameters in a linear regression, including categorical variables. b. Estimate standard deviation of each parameter in linear regression, and assess their accuracy. c. Calculate R^2 for a linear model. d. Apply forward stepwise selection. e. Calculate confidence intervals for model predictions and assess their accuracy. f. Determine best subset of variables to use in a linear regression model. g. Incorporate non-additive and non-linear relationships.	ESL, Ch. 3, pp. 43-55 and 57-61	18	
2. Shinkage methods to improve model prediction	a. Use ridge regression to fit a linear model, and comment on the selection of lamda. b. Use lasso to fit a linear model, and comment on the selection of lamda. c. Comment on how the penalty works in both ridge and lasso methods	ESL, Ch. 3, pp. 61-79 and 82-83	21	glmnet package
3. Classification with categorical variables	a. Fit a linear model and use it for classification. b. Fit a logistic model and use it for classification. c. Use k-nearest-neighbors for classification. d. Identify when linear discriminant analysis will perform better than regression.	ISL, Ch. 2, pp. 39-42 ISL, Ch. 4, pp. 127-149 (stop at 4.4.4), and 151-154	30	knn in class lda in mass

4. Bias, variance, and their tradeoffs	a. Estimate variance of model estimates. b. Describe why a model may be biased. c. Describe how to build a model to minimize the expected mean squared error. d. Describe the tradeoffs between bias and variance	ESL, Ch. 7, pp. 219-235	17	

E. **MODEL CHOICE/SELECTION & PERFORMANCE MEASURES**

TOPIC	LEARNING OBJECTIVES	READINGS	PAGES	R PACKAGES AND INSTRUCTIONS
1. Linear model diagnostics (See note below on prerequisite assumptions)	a. Perform, interpret, and act upon standard diagnostics on linear models, including assessment and treatment of i. Outliers ii. Appropriateness of model specification iii. Non-normal errors iv. Nonlinear dependencies v. Heteroscedasticty vi. Multicollinearity b. Understand and apply the hat matrix, hat values, residuals (raw, standardized, studentized, and Pearson), and Cook's D to detect outliers and influential observations c. Apply residual plots, marginal model plots, and added variable plots to assess quality of fit and the impact of each predictor d. Use QQ plots to diagnose non-normal errors e. Use F-tests, residual plots, component-plus-residual plots, and CERES plots to identify non-linear dependencies f. Use residual plots and spread-level plots to identify heteroscedasticity; determine when transformation of the target variable (possibly via Box Cox) is an appropriate remedy, and when weighted regression is appropriate	Fox, Ch. 11, up to but not including 11.4.1 (pp. 241-252) Fox, Ch. 11, sections 11.6-11.8.3 (pp 255-263) Fox, Ch. 12, up to and including 12.5.1 (pp. 267-294) Fox, Ch. 12, results stated in exercises 12.3-12.4 (p 302) Fox, Ch. 13, excluding 13.1.1 (pp. 307-313, 322-329) Fox & Weisberg, Ch. 6 (pp. 285-327) but excluding all of the following: (i) the last subsections of 6.4.1 (pp. 307-309), (ii) the last subsection of 6.4.2 (pp. 312-314), (iii) 6.5.2	97	base R (stats package): residuals, rstandard, rstudent, hatvalues, cooks.distance, dfbeta, dfbetas, lm.influence car package: residualPlots, marginalModelPlots, qqPlot, outlierTest, influenceIndexPlot, boxCox, powerTransform, crPlots, ceresPlots, spreadLevelPlot, vif, avPlots, boxCoxVariable base R "prereq": lm and functions that manipulate lm objects (e.g., predict.lm, summary.lm, coef, effects, vcov)

	g. Identify collinearity via variance-inflation factors and generalized variance-inflation factors and discuss possible ways to deal with collinearity	(p. 316), and (iv) 6.6 (pp. 317-324)		
2. Generalized linear models	a. Understand the assumptions behind different forms of the GLM and be able to select the appropriate model. b. Understand the relationship between mean and variance by model family member c. Understand how to select the appropriate link function and distribution for the dependent variable. d. Understand the Tweedie as compound gamma-Poisson and also as the GLM with variance function a power law. e. Describe the reason for a double GLM and two ways in which a double GLM might be fit. f. Describe similarities and differences between a double GLM and a weighted GLM g. Describe the iteratively reweighted least squares algorithm h. Use appropriate diagnostics to evaluate the fit of a GLM i. Fit a logistic regression by penalized ML, and describe when that should be preferred to ML j. Describe the effect of non-canonical link function on bias k. Define deviance and its relationship to a GLM	Fox, Ch. 14, section 14.1 Fox, Ch. 15 Fox & Weisberg, Ch. 5, sections 5.10-5.11 Fox & Weisberg, Ch. 6, section 6.6 Smyth & Jorgensen Heinze & Ploner Allison	99	glm, glm.nb, family, logistf, plus methods for glm model objects of the R packages mentioned in the linear model diagnostics learning objective

3. Goodness-of-fit metrics for generic models	a. Define and apply ROC curves, AUC, Lorenz curves, and Gini index	ISL, pp. 147-149 Two-page study note		
4. Predictor variables and transformations	a. Types of variables b. Transformations of variables c. Categorical vs. continuous explanatory variables d. Interaction terms e. Significance and model comparison statistics f. Residuals and model parameter selection. g. ROC curves, Lift Curves	ESL, Ch. 2, section 2.2 Stat Ed, pp. 580-593	16	
5. Linear Mixed models and Bühlmann Credibility	a. Understand Linear Mixed models and Bühlmann Credibility as Hierarchical Linear Models, repeated measures and subject effects b. Describe sources of correlation in longitudinal data c. Describe the reasons for estimation using REML rather than maximum likelihood d. Use appropriate plots to do EDA of longitudinal data e. Build hierarchical models via the linear mixed-effects approach f. Describe the Bayesian approach to hierarchical models g. Apply Bühlmann credibility theory and describe the connection to linear mixed effects models	Gelman & Hill, Chs. 11, 12, and 13 Fitzmaurice, et al, sections 2.5, 3.3-3.5, 4.1-4.5 Frees, Ch. 18 through 18.3.1; with corrective study note Galecki & Buryzkowski, Ch. 15 (24pp…for lmer reference)	135	In packages lme4: lmer, getME, ranef, simulate,refit,residuals, coef, fixef
6. Non-linear models	a. Be able to capture non-linear relationships using polynomials, splines, local regression, and GAM's b. Use polynomials in regression c. Use step functions in regression	ISL, Ch. 7, pp. 265-287	23	gam in gam bs in splines smooth.spline in base loess in base

	d. Use regression splines, and identify helpful constraints to ensure that model is smooth. e. Use local regression to determine estimates. f. Use GAM's to build a model.			
7. Tree-based methods	a. Develop a general knowledge of tree-based methods. b. Understand and use regression trees c. Build a regression tree with a dataset d. Use a regression tree to determine an estimate for an observation. e. Build a classification tree with a dataset f. Use a classification tree to determine an estimate for an observation g. Discuss reasons for pruning and methods to prune h. Use bagging to get estimates for a new observation i. Discuss why bagging is helpful compared to regular trees j. Use random forest to get estimates for a new observation k. Discuss why random forests are helpful compared to bagging l. Use boosting to get an estimate for a new observation m. Discuss how boosting works	ISL, Ch. 8, pp. 303-323	21	Tree package prune.tree in tree randomForest gbm
8. Resampling methods	a. Understand and explain how cross-validation works	ISL, Ch. 5, pp. 175-186 ISL, Ch. 5, pp. 187-190	23	caret package, cv.glm in boot

	b. For a given dataset and model, use cross-validation to estimate the accuracy of model predictions. c. Understand and explain how bootstrap works d. For a given dataset and model, use bootstrap to determine which variables are significant e. Define and calculate each of i. Accuracy ii. Precision iii. Specificity iv. Sensitivity f. Describe challenges arising from unbalanced designs: bias toward majority class; different cost for different errors g. Describe how unbalanced training datasets can influence classifiers and why that is a problem h. Recognize advantages and drawbacks of oversampling, SMOTE, undersampling, and cost-sensitive learning i. Identify algorithmic solutions to using unbalanced training sets	Sokolova & Lapalme, pp. 427 - 431 (up to but not including '4. Invariance properties of measures') Ganganwar, sections I, II, III, and VI		
9. A/B testing	a. Understand how A/B testing works b. Describe practical considerations in designing AB tests c. Describe power and significance in A/B testing d. Summarize common mistakes and difficulties in A/B testing and techniques to address these e. Recognize alternatives techniques to classical A/B testing	Mount 1 Mount 2 Miller	15	

10. Association models	a. Understand the basics of an association model as used in a market basket analysis b. Identify the types of patterns that can be detected using simplified association rules c. Interpret and evaluate the support, confidence, and lift of a market basket analysis	ESL, Ch. 14, sections 14.2.0-14.2.3, and 14.2.7	12	
11. Propensity Scoring and Matching Methodologies	a. Understand and evaluate methods of estimating causal effects including CEM and propensity score matching b. Critique CEM, propensity score, and model based methods for estimating causal effects c. Discuss the process for using propensity scores and CEM to estimate causal effects d. Distinguish causal effects from predictions e. Explain SATT (sample average treatment effect on the treated)	Rubin (TBA) Iacus, et al, sections 1, 3.0, 3.2, 3.3, and 3.5	28	
12. Unsupervised learning and clustering	a. Understand and apply the unsupervised learning and clustering using k-means, and agglomerative hierarchical clustering b. Differentiate between supervised and unsupervised learning tasks c. Describe the choices involved in using k-means and hierarchical clustering and the implications of them d. Interpret a dendrogram e. Summarize potential issues with using clustering and ways to mitigate them f. Cluster data using kmeans and heirarchical clustering	ISL, Ch. 10, sections 10.1, 10.3, 10.5	19	kmeans hclust dist cutree

13. Choice models and conjoint analysis	a. Understand Choice models b. Ggiven Choice data, be able to fit a Choice model c. Understand the use of conjoint analysis with survey data d. Given some survey data be able to apply a conjoint model	Chapman & Feit, pp. 244-247, 363-383, and 396-399	26	

B. MISCELLANEOUS

TOPIC	LEARNING OBJECTIVES	READINGS	PAGES	R PACKAGES AND INSTRUCTIONS
1. Model summary and presentation	a. Presenting and showing material b. Communicating how the model is applied c. Communicating and showing material associated with the holdout data	TBA		
2. Public Perception	a. Recognize concerns (whether fair or not) that are commonly expressed about predictive models in the press and political discourse.	O'Neil, Ch. 6	18	
3. Fraud Detection	a. Apply Pridit and Random Forests to fraud detection problems	Frees, et al, Ch. 7 (up to and including 7.12)	21	base package: princomp, randonForest package: randomForest

4. Ethics & Practicalities	a. Summarize how threshold classifiers can be discriminatory b. Recognize the implications of threshold selection methods (Max profit, group unaware, demographic parity, and equal opportunity) c. Identify potential issues in scenarios where ethical care and oversight would be advised d. Describe ways models can promote fairness and perpetuate social inequalities e. Recognize potential issues with proxy variables and spurious correlations f. Explain the importance of calibrating models using feedback to reduce bias	Wattenberg, et al Cabinet Office	12	