ITEC 621 Predictive Analytics

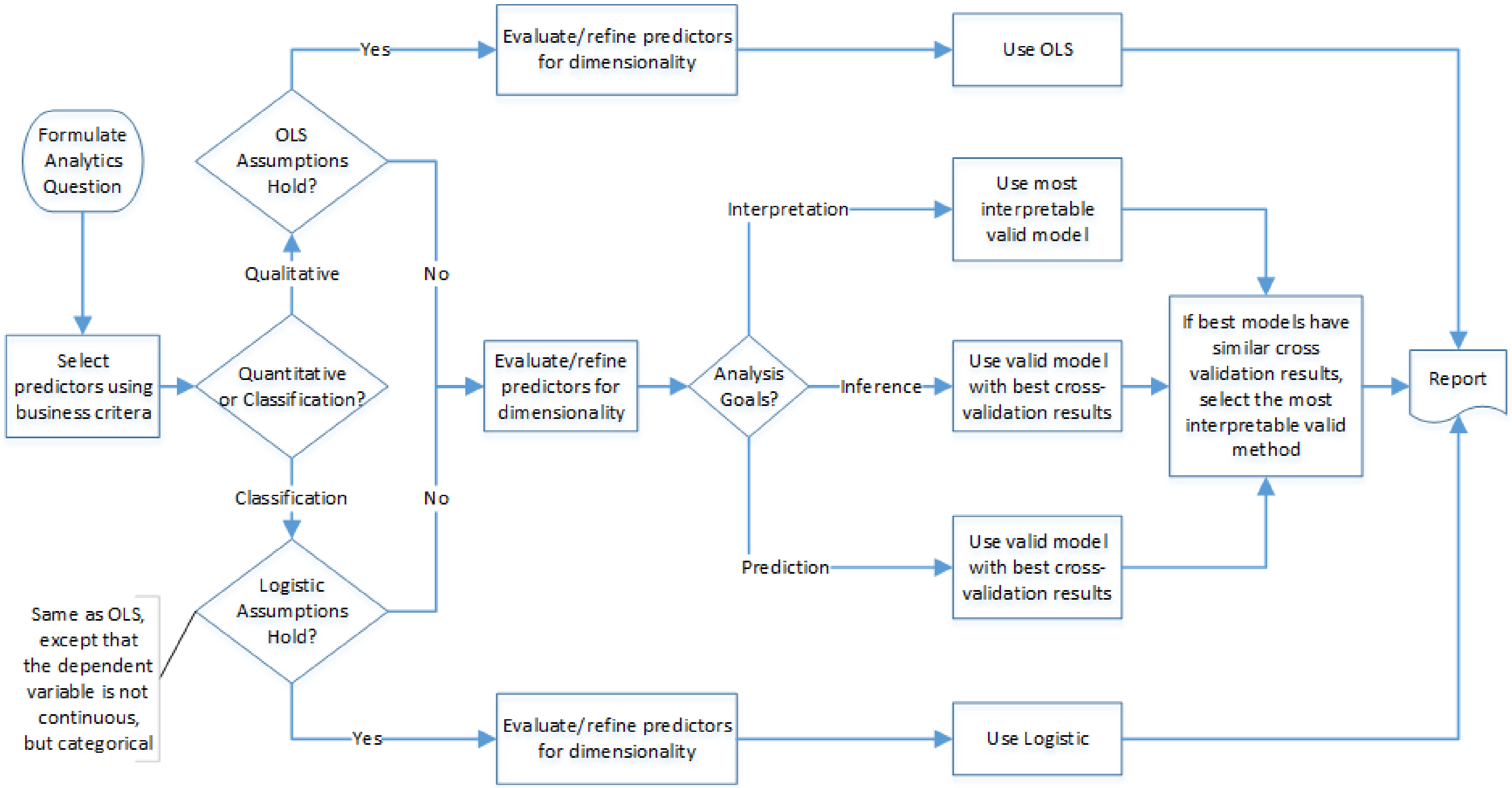
Model and Method Summary

Prof. Alberto Espinosa *(last revised 10/3/2016)*

General Predictive Model Selection Approach in 10 Steps:

1. Review and carefully understand your analytics question(s). This involves understanding:
   a. Your response variable(s) – what are you trying to predict?
   b. Whether you have a quantitative or classification problem – note: this may change as you complete your analysis, as we sometimes transform quantitative variables into categorical and vice versa.
   c. What are some of the likely predictors you may use – naturally, you are only speculating at the beginning, but you should have some idea from a business perspective about what are the possible predictors based on business knowledge or intuition.
2. Identify, review and describe your data:
   a. Inspect your data sources for omitted data and inconsistencies
   b. Do the necessary descriptive analytics review, including but not limited to: descriptive statistics, correlation matrix, correlation plots, other plots to evaluate things like normality, linearity, etc. No need to do everything, but only what is needed to do your predictive modeling
   c. Do all the necessary data pre-processing and transformations
3. Try first, either:
   a. An OLS model, if you are addressing a quantitative problem.
   b. A Logistic model, if you are addressing a classification problem.
   c. Use the full data set on your first exploratory OLS or Logistic models
4. If you have many predictors, you may want to rationalize the best set of predictors from a business perspective. You should then determine the appropriate number of predictors using a variable selection method.
5. Test key OLS assumptions as needed – e.g., normality of the response variable, multi-collinearity, heteroscedasticity, serial correlation (only for time related data), linear relationship between predictors and the response variable, etc. – OLS is BLUE if all assumptions hold (Logistic is BLUE for classification problems).
6. Correct for OLS assumption violations when possible (e.g., transformations, variable selection, weighting, etc.)
7. State your analysis goals – interpretation (explanation of effects), inference (hypotheses testing), and/or prediction; and narrow down your choice of models accordingly.
8. Fit the resulting models using the full data set and evaluate these viable models using fit statistics (e.g., adjusted R square, MSE, AIC, error rates, etc.).
9. Do cross-validation of all the viable models by sampling/re-sampling training data to fit the model and computing fit statistics (i.e., MSE, classification error, deviance, sensitivity, specificity, ROC curves, etc., as appropriate) on the test data. Please rationalize and justify your cross-validation and re-sampling method (i.e., subset partition, K-Fold validation or Leave One Out validation).
10. Select the model with best cross-validation test results. For similar cross-validation results, select the most interpretable model.

Analytics Model Selection Process


Formulate Analytics Question
Select predictors using business critera
Quantitative or Classification?
Qualitative
OLS Assumptions Hold?
Yes
Evaluate/refine predictors for dimensionality
Use OLS
No
Classification
Logistic Assumptions Hold?
No
Yes
Evaluate/refine predictors for dimensionality
Use Logistic
Same as OLS, except that the dependent variable is not continuous, but categorical
Evaluate/refine predictors for dimensionality
Analysis Goals?
Interpretation
Use most interpretable valid model
Inference
Use valid model with best cross-validation results
Prediction
Use valid model with best cross-validation results
If best models have similar cross validation results, select the most interpretable valid method
Report

Method/Model	*Type		** OLS Assumptions	*** When to Use/Comments
Ordinary Least Squares (OLS)	R	Q	YC (✓) Y is continuous YN (✓) Y is normally distributed XI (✓) X's are independent (uncorrelated) LI (✓) Y and X's have linear relationship OI (✓) Observations are independent EI (✓) Errors are independent EA (✓) The error average is 0 EV (✓) The error variance is constant	• If OLS assumptions hold → **OLS is BLUE** (best linear unbiased estimator → Gauss-Markov Theorem) • It means it is the most efficient → Lowest variance • Need to test the assumptions • Excellent for **inference/interpretation** (same is true for most regression models) • Good for **prediction**, but other models with smaller variance may have better cross-validation results • Need large samples with 30+ degrees of freedom (i.e., N-P-1 > 30)
Weighted Least Squares (WLS)	R	Q	EV (×) The error variance is not constant	• OLS is unbiased, but the variance is wrong • WLS is more efficient • Use when heteroscedasticity is present (i.e., EV does not hold – i.e., error variance varies with Y • Good for **inference/interpretation** and **prediction**
Standardized OLS	R	Q	All OLS assumptions (✓)	• Scale of X's vary widely • Scale of X's not easily interpreted mathematically (e.g., survey ratings) • Good for **inference/interpretation** and **prediction**
Log Transformed OLS	R	Q	YN (×) Y is not normally distributed (e.g., distribution is skewed	• Use when YN does not hold • Once logged transformed, all OLS assumptions should hold • You can log-transform some X's if you are interested in the effect of a % increase, rather than a unit increase • You must log-transform X's if not normally distributed and the sample is small (no need for large samples) • Can't log transform Y or any X's if they contain negatives or 0 (can't log these values) • Good for **inference/interpretation** and **prediction**
Log Transformed GLM	R	Q	YC (×) Y is not continuous EV(×) Error variance is not constant	• Use when the response variables contains **count data** • Data is discrete (not continuous); truncated at 0; and with uneven errors (low near 0 and increasing as counts get larger • A popular model for count data is to use the Generalized Linear Method (**GLM**) with a log-transformed Y and a Poisson distribution.
Rank Transformed OLS	R	Q	YN (×) Y is not normally distributed (e.g., distribution is skewed	• Useful with small samples • Also, when data is not normally distributed and the distribution doesn't seem to have a pattern • And when there is very little variance in a variable.

			Note: any rank-transformed variable becomes non-parametric and has a uniform distribution	• OLS assumptions no longer apply to rank transformed variables • OK for **inference/interpretation**; good for **prediction**
Causal (Lag) Models	R	Q	EI (×) Errors are not independent → the vary systematically across time (e.g., serial correlation)	• Typically used when the X's include a time variable • And when the Durbin-Watson test finds auto correlated residuals (e.g., serial correlation) • Good for **inference/interpretation** and **prediction**
Ridge Regression	R	Q	All OLS assumptions (✓)	• Increases predictor bias • But provide better predictive accuracy and cross validation when there are too many predictors (i.e., curse of dimensionality). • A desirable method when there are too many predictors in the model and the goal is accurate predictions • Small coefficients are shrunk with "penalized" modeling, but they don't become 0 • How much shrinkage will depend on the selected tuning parameter λ (=0 → OLS; =∞ → no coefficients, just intercept) • OK for **inference/interpretation**; good for **prediction**
LASSO Regression	R	Q	All OLS assumptions (✓)	• Similar to Ridge, but some coefficients do become 0 • OK for **inference/interpretation**; good for **prediction**
Principal Components Regression (PCR)	R	Q	XI (×) Some X's are not independent (correlated)	• Useful when there is a large number of predictors P • And there is high multi-collinearity and multiple variables that appear to measure similar things (e.g., size of car, weight of car) • So, so for **inference/interpretation**; good for **prediction**
Factor Analysis Models	R	Q	XI (×) Some X's are not independent (correlated)	• Similar to PCR, but principal components are rotated further to find stronger commonality among predictors, which are then grouped by averaging. • Very popular with survey data, where there are hundreds of questions, many of which elicit similar answers (e.g., is this a good course? Do you like this course?) • OK for **inference/interpretation**; good for **prediction**
Partial Least Squares (PLS)	R	Q	XI (×) Some X's are not independent (correlated)	• Similar to PCR, but components are also evaluated for their correlation with Y. • A popular method when the number of predictors P is high, compared to N • And also when the X's are highly correlated

				• It is also a very popular and effective method applied to structural equation modeling. • So, so for **inference/interpretation**; good for **prediction**
Interaction Models	R	Q	LI (×) Y does not have a linear relationship with all X's	• Useful when you suspect that the value of one predictor influences the effect of another predictor (e.g., the effect of an antibiotic is diminished if you drink alcohol). • Very good for **inference/interpretation**; very good for **prediction**
Polynomial Models	R	Q	LI (×) Y does not have a linear relationship with all X's EI (×) Errors not entirely independent EV (×) The error variance is not constant	• If the X's have a curvilinear relationship with Y and you fit a straight line, then sections of the data will be above the line and others below the line and thus chunks of errors will be above 0 and other chunks below 0. • Quadratic, Cubic and other polynomial models will fit the data better. • The value for **inference/interpretation** diminishes as the power of the polynomial goes up; but the **prediction** value goes up (except at both tail ends).
Step Regression Models	R	Q	Same as Polynomial Models	• Good when data seems to have different patterns in different sections • Preferred to polynomial models when polynomials don't fit the data well at both ends of the curve (i.e., waging the tail) • Regression lines are horizontal and they change the intercept in different sections of the data • Good for **inference/interpretation** but the value diminishes as the number of steps increases; but the **prediction** value goes up.
Piecewise Models	R	Q	Same as Polynomial Models	• Similar to Step models, but more effective when the data in each section appears to slope linearly upwards or downwards. • Good for **inference/interpretation** but the value diminishes as the number of piecewise sections increases; but the **prediction** value goes up. It generally performs better than polynomial regression at the tails.
Spline Regression (MARS)	R	Q	Same as Polynomial Models	• Similar to Piecewise models, but more effective when the data appears to have a curvilinear relationship with Y • Not good for **inference/interpretation** especially as the power of the spline goes up; very good for **prediction.** It generally performs better than polynomial regression at the tails.
Smoothing Splines	R	Q	Same as Polynomial Models	• Similar to Spline models, but the transition of the curves from one section to another is "smoothed"

				• Not good for **inference/interpretation**; very good for **prediction** and it generally performs better than polynomial regression at the tails.
Binomial Logistic Regression	R	C	YC (×) Y is not continuous YN (×) Y is not normally distributed	• Most popular regression model when Y is binary (e.g., yes/no, approve/decline, etc.) • Y can only have two classes (hence binary) • It tends to outperform LDA and QDA below when observations don't come from a normal distribution • The model predicts the probability of falling into one of the classes, not the actual value of Y • OK for **inference/interpretation** if the goal is to interpret likelihoods or probabilities, not values; good for **prediction**
Multinomial Logistic Regression	R	C	Same as Binary Logistic Models	• Use instead of Binary Logistic Regression when Y has more than two classes (e.g., Rural, Suburban, Urban)
Linear/Quadratic Discriminant Analysis (LDA/QDA)	R	C	Same as Binary Logistic Models	• If N is large, Logistic Regression and Discriminant Analysis produce similar results, but Logistic is preferred because the model is more interpretable. • If N is small and X's are normally distributed Discriminant Analysis models are more stable and Discriminant Analysis is preferred if **prediction** is the goal and **inference/interpretation** is not.
K Nearest Neighbors (KNN)	na	C	Non-parametric. OLS assumptions and other parametric restrictions don't apply	• Generally outperforms logistic and LDA/QDA when the decision non-linear and has no clear pattern. • Not good for **inference/interpretation**; may be good for **prediction** depending on cross-validation statistics.
Regression Trees	T	Q	Non-parametric. OLS assumptions and other parametric restrictions don't apply	• Not good for **inference/interpretation**; may be good for **prediction** depending on cross-validation statistics.
Classification Trees	T	C	Non-parametric. OLS assumptions and other parametric restrictions don't apply	• Not good for **inference/interpretation**; may be good for **prediction** depending on cross-validation statistics.
Bootstrap Aggregation (Bagging)	T	C	Non-parametric. OLS assumptions and other parametric restrictions don't apply	• Not good for **inference/interpretation**; much better than classification or regression trees for **prediction** but cross-validation statistics need to be checked to select the model with highest predictive accuracy.
Random Forests	T	C	Non-parametric. OLS assumptions and other parametric restrictions don't apply	• Not good for **inference/interpretation**; much better than classification or regression trees for **prediction** but cross-validation statistics need to be checked to select the model with highest predictive accuracy.

Boosting	T	C	Non-parametric. OLS assumptions and other parametric restrictions don't apply	• Not good for **inference/interpretation**; much better than classification or regression trees for **prediction** but cross-validation statistics need to be checked to select the model with highest predictive accuracy.
Support Vector Machines	T	C	Non-parametric. OLS assumptions and other parametric restrictions don't apply	• Not good for **inference/interpretation**; much better than classification or regression trees for **prediction** but cross-validation statistics need to be checked to select the model with highest predictive accuracy.
Neural Networks	R	Q	XI (×) Some X's are not independent (correlated)	• Useful when there is a very large number of predictors P • And there is high multi-collinearity and multiple variables that appear to measure similar things (e.g., size of car, weight of car) • Not good for **inference/interpretation**; good for **prediction**
Structural Equation Models	R	Q	XI (×) Some X's are not independent (correlated)	• Useful when multi-collinearity is very or extremely high, but it is important to retain all or most predictors in the model. • Or when structural relations are hypothesized (e.g., some variables predict some outcomes, and these outcomes predict other outcomes in turn, and so on. • Very good for **inference/interpretation**; good for **prediction**

***Type:** R or T (Regression or Tree)/Q or C (Quantitative Value or Classification Outcome)
**** OLS Assumptions:** (✓) Holds; (×) Does not Hold;
***** Notation:** Y = Outcome Variable; X's = Predictor Variables; N = Number of Observations; P = Number of Predictors