Logistic Regression Review

What we are to do when trying to solve a logistic Regression Problem:

1. Identify Target variable: Binary or ordinal
   ~~Binary or ordinal~~
2. Check for Rare event ●5% ≥ X
   If we have a Rare event we can do the following:
   - OverSampling: *duplicating the rare event observations
   - underSampling: Randomly sample the non-rare events to make them smaller
   Techique to correct the bois from the above methods:

	Mode correct	Model misspecified
Small Sample $n \leq 1000$	Adjust intercept	weighted observations
Large sample $n > 1000$	Either	weighted observations

3. Split into Train, validation, & Test sets

4. Identify the predictor variables:
   - If Continous:
     → build individual Logistic Regression models for each variable. (MLE)
     → Evaluate if variable is statistcaly significant (p-values)
   - If Ordinal:
     → Check for Seperation Concerns
     50 0 / 0 50 — Linear seperation: When a combo perfectly predicts every outcome
     50 23 / 0 22 — Quasi Seperation: When a combo perfectly predicts a subset of outcomes
     ↳ To correct the above cases we can fix it through thresholding
     → Check if variable is significant w/ Mantel-Haenszal X^2 test.
   - If Nominal:
     → Check for Seperation Concerns
     ↳ To correct issues we Cluster categories with issues to their most common category. Ex: A 8 3 / B 3 11 / C 5 0 → A/C 13 3 / B 3 11
     → Pearson X^2 test to see if variable is significant.

5. Variable Selection:
   → Forward, backward, & Stepwise
   ↳ AIC, BIC, p-value

Logistic Regression Review

"large" β w/ large p-value → signal something is wrong

6. Adding Interactions: (be carefull with seperation issues)

7. Check Model Diagnostics:
   Deviance = measure of how far our model is from the saturated model.
   Through Deviance we can calculate: Cooks D, DFBetas, DifDev

8. Model Evaluation:
- Likelihood based calculations: (AIC, BIC, Generalized R^2)
- Probability Metrics: (Rank order observations)
  -> concordance, discordance, # Ties
  -> Discrimination Slope = $(\bar{P}_1 - \bar{P}_0)$ bigger is better
- Classification Metrics: $TP/(TP+FN)$ ← $\frac{TN}{TN+FP}$
  -> Sensitivity (TPR) vs. Specificity (TNR)
  -> ROC Curve (we want the curve to go as far from the diagonal as possible)
  -> KS Statistic (we want the two lines to as far as possible from each other)
  -> Precision = # true positive / (# true positive + # false positive)
  -> Precision vs. Recall (Sensitivity (TPR))
    ↳ F_1 score is how we select optimal values for the above.
    The highest F_1 score is the optimal F_1 score.

Optimal Point is Youden's Index ↕ PR+(TNR-1)

-> Lift curve, Response Curve, Gain Chart
(Lift (Resp Rate (Cap Rate

↳ Lift curve: represents how much better the model is at predicting responses compared to random selection. Higher is better
Ex: "In the top 10% of data our model is X times more likely to identify responses compared to random selection"

buckets are sorted by our models predicted probability

↳ Response Curve: the proportion of responders in each bucket w/o of the data
Ex: "In the top 10% of our data our models prescion is X"

↳ Gain Chart; Shows how well our model captures the total response across all buckets.
Ex: "By targeting the top X percent of the data X% of the total responses can be captured using our model"

← Don't use for model selection, only Reporting!

- Accuracy & Error: 1 = Accuracy + Error
  - ↳ Accuracy = % of predictions our model got correct
  - ↳ Error = % of predictions our model got wrong.

(9.) Interpreting Coefficients:
- Odds Ratios: $e^{\hat{\beta}}$ or $100 \times (e^{\hat{\beta}} - 1)\%$
  - ↑ X times more likely...
  - ↑ x% higher odds of...

(10.) Make Final Cutoff: Cost/profit should always be how we choose cutoffs!
- If cost/profit is equal, we choose cutoff from: Youden, KS, F_1

(11.) Create the Report for Stakeholder!