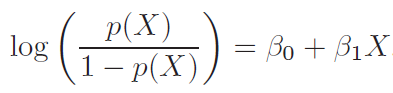
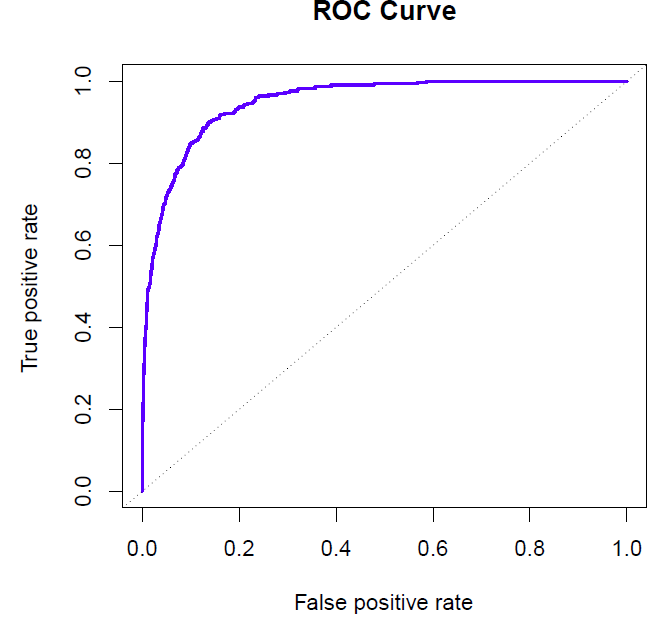
* Curse of dimensionality – non-parametric (KNN) perform worse as number of parameters (p) increases
  + Selecting subset of each feature 🡪 (i.e. subset = 90% and # parameters = 100) 🡪 amount of data used is 0.9100 = 0.002%
* Linear Regression fails with categorical variables b/c:
  + For given variables label encoded from [1,10] assumes value encoded as [5] in between 1 and 10. Not the case usually.
  + Need a natural ordering of categorical variables.
  + Only two values 🡪 binary dummy variable OK for Linear Regression
* Dummy variables
  + Only 2 variables (otherwise ordering makes it unusable)
  + Used for linear regression
* Logistic Regression: Pr(Y = k | X = x)
  + Logit (left-side): 
* Maximum Likelihood Function – find B0 and B1 such that plugging into p(X) yields number close to 1 for p(X) = 1 and close to zero for p(X) = 0
* Confounding –
  + beta is negative (positive) for a single variable regression
  + beta becomes positive (negative) for a multi-variable regression
  + Caused by correlation between variables
* Multi-response-variable logistic regression – Linear Discriminant Analysis (LDA)
  + Reasons logistic regression fails for multi-response-variables:
    - Parameter estimates unstable
* Trade-off of which value (P(X) = 0 or P(X) = 1) to lower:
  + Set threshold below or above P(X=1 | X = x) > 0.5
    - 0.5 🡪 optimal overall error rate
  + 
* 
* KNN outperforms Logistic Regression and LDA: decision boundary highly non-linear
  + LDA and Logistic Regression – decision boundary = linear
* KNN – variable scales
  + KNN looks at absolute difference between variables:
    - Solution 🡪 standardize variables
* KNN (k = 1) 🡪 training error rate = 0