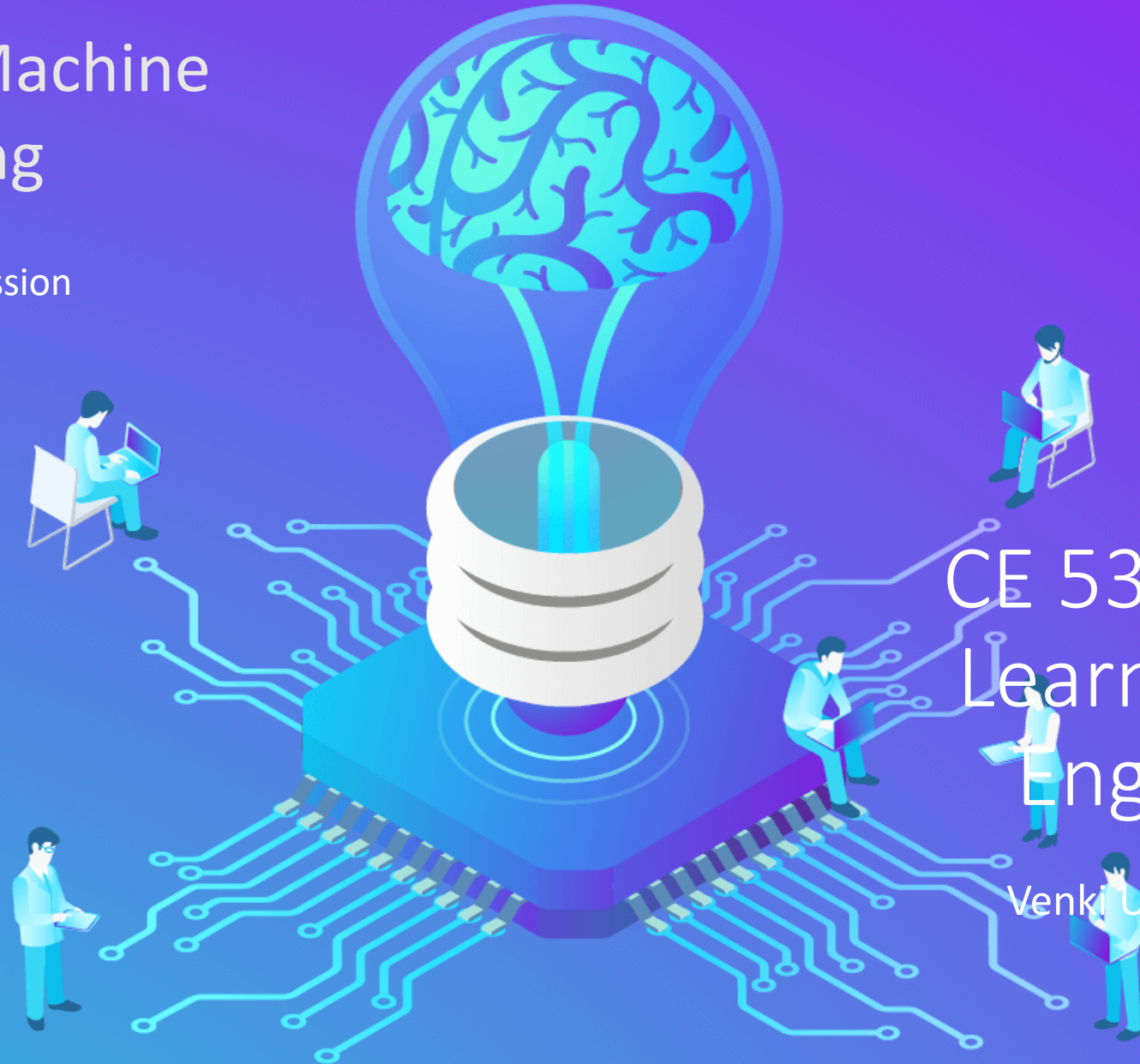


Python for Machine Learning

Logistic Regression

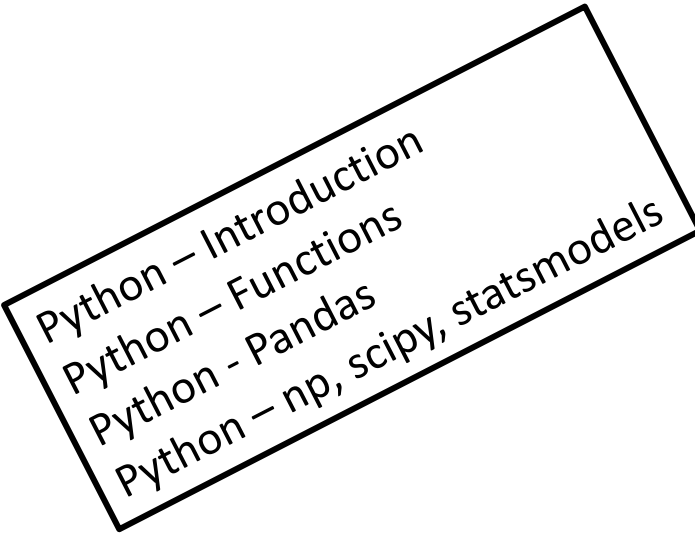


CE 5331 Machine Learning for Civil Engineers

Venki Uddameri, Ph.D. , P.E.

Recap

- What is Machine Learning
- How is it useful for Civil Engineers
- Overview of Machine Learning Methods
- Linear Regression
 - Bivariate
 - Regression interpretation
 - Multivariate



Python – Introduction
Python – Functions
Python – Pandas
Python – np, scipy, statsmodels

In this module we shall look at Logistic Regression

Logistic Regression

- Logistic Regression is used when the output variable is dichotomous
 - Y can only take values of 0 or 1
 - Y only has two states it can be in
- Dichotomous variables appears in many civil engineering applications
 - Flow, no flow
 - Accident, no accident
 - Damage, No Damage
- We typically can use thresholds to make continuous variables dichotomous
 - Use a threshold to categorize the variables
 - If ($\text{NO}_3 \leq 10$) THEN Safe ELSE Contaminated

Logistic Regression

- Let Y be a dichotomous variable $\{0,1\}$
 - Y is related to several inputs
 - These inputs can be continuous or discrete
- What logistic regression seeks to do is estimate the probability of $Y = 1$
- The odds ratio is the probability of $Y=1$ (i.e., $P(Y=1)$) to the probability that $Y = 0$ (i.e., $P(Y=0)$)

$$O(Y) = \frac{P(Y = 1)}{P(Y = 0)}$$

} Odds Ratio

$$P(Y = 1) + P(Y = 0) = 1$$

$$O(Y) = \frac{P(Y = 1)}{P(Y = 0)} = \frac{P(Y = 1)}{1 - P(Y = 1)}$$

Odds ratio tells us how probable
is seeing $Y=1$ over $Y=0$

While the dichotomous variable only assumes values of 0 or 1 the odds ratio varies from 0 to ∞ and can be regarded continuous

By convention odds ratio is written for $P(Y=1)$

Logit

- The log of odds ratio is called the logit
- The logit varies from $-\infty$ to ∞
 - Continuous variable
- Logit (logistic) regression where logit is regressed against independent variables

$$P(Y = 1) = \frac{1}{1 + \text{EXP}[-(\beta_o + \beta_1 X_1 + \dots + \beta_{k-1} X_{k-1})]}$$

Logit

↓

$$L(O) = \log \left[\frac{P(Y = 1)}{P(Y = 0)} \right]$$

$$L(O) = \log \left[\frac{P(Y = 1)}{P(Y = 0)} \right] = \beta_o + \beta_1 X_1 + \dots + \beta_{k-1} X_{k-1}$$

$$\left[\frac{P(Y = 1)}{P(Y = 0)} \right] = \text{EXP}(\beta_o + \beta_1 X_1 + \dots + \beta_{k-1} X_{k-1})$$

$$\left[\frac{P(Y = 1)}{1 - P(Y = 1)} \right] = \text{EXP}(\beta_o + \beta_1 X_1 + \dots + \beta_{k-1} X_{k-1})$$

$$\left[\frac{1 - P(Y = 1)}{P(Y = 1)} \right] = \text{EXP}[-(\beta_o + \beta_1 X_1 + \dots + \beta_{k-1} X_{k-1})]$$

$$\frac{1}{P(Y = 1)} = 1 + \text{EXP}[-(\beta_o + \beta_1 X_1 + \dots + \beta_{k-1} X_{k-1})]$$

Using logistic regression we can estimate the probability of $Y = 1$

Logistic Regression Maximum Likelihood

- The unknown model parameters are estimated using the method of maximum likelihood
- A Y is a dichotomous variable it follows Bernoulli's distribution
- Y_i has a probability of p is the value = 1 and $(1-p)$ if the value is equal to zero

$$L_i = p^{Y_i} (1 - p)^{1-Y_i} \left\{ \begin{array}{l} \text{Likelihood of an single value} \\ \text{IF } Y_i = 1 \text{ then } p \text{ ELSE } (1-p) \end{array} \right.$$

$$L_i = \beta_o + \sum_{j=1}^{k-1} \beta_j X_j \quad \forall i = 1, \dots, N \left\{ \begin{array}{l} \text{Likelihood a function of} \\ \text{independent variables} \end{array} \right.$$
$$p_i = \frac{1}{1 + e^{-L_i}}$$


$$L = \prod_{i=1}^N L_i = \prod_{i=1}^N p_i^{Y_i} (1 - p_i)^{1-Y_i} \left\{ \begin{array}{l} \text{Likelihood function to be} \\ \text{maximized} \end{array} \right.$$

$$\log(L) = \sum_{i=1}^N \log(L_i) = \sum_{i=1}^N Y_i \log(p_i) + (1 - Y_i) \log(1 - p_i) \left\{ \begin{array}{l} \text{Maximize} \\ \text{the log} \\ \text{likelihood} \end{array} \right.$$

Estimate unknown model coefficients by
maximizing likelihood

Parameter Estimation

- We can perform the maximum likelihood using the minimize function is scipy
 - Remember maximization of L is the same as minimization of $-L$
- There are a few packages that allow us to perform logistic regression in Python
 - 'statsmodels'
 - 'Scikitlearn'
 - This applies regularization so does not use conventional maximum likelihood
 - Tries to minimize the number of input parameters while maximizing likelihood
- We can also use R to perform logistic regression



Most flexible but also
needs some detailed
understanding

Evaluating Logistic Regression

- Contingency Tables are used for evaluating logistic regression

		Actual class	
		Cat	Non-cat
Predicted class	Cat	True Positives	False Positives
	Non-cat	False Negatives	True Negatives

sensitivity, recall, hit rate, or true positive rate (TPR)

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} = 1 - FNR$$

specificity, selectivity or true negative rate (TNR)

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} = 1 - FPR$$

precision or positive predictive value (PPV)

$$PPV = \frac{TP}{TP + FP} = 1 - FDR$$

negative predictive value (NPV)

$$NPV = \frac{TN}{TN + FN} = 1 - FOR$$

condition positive (P)

the number of real positive cases in the data

condition negative (N)

the number of real negative cases in the data

true positive (TP)

eqv. with hit

true negative (TN)

eqv. with correct rejection

false positive (FP)

eqv. with false alarm, Type I error

false negative (FN)

eqv. with miss, Type II error

sensitivity, recall, hit rate, or true positive rate (TPR)

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} = 1 - FNR$$

specificity, selectivity or true negative rate (TNR)

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} = 1 - FPR$$

precision or positive predictive value (PPV)

$$PPV = \frac{TP}{TP + FP} = 1 - FDR$$

negative predictive value (NPV)

$$NPV = \frac{TN}{TN + FN} = 1 - FOR$$

miss rate or false negative rate (FNR)

$$FNR = \frac{FN}{P} = \frac{FN}{FN + TP} = 1 - TPR$$

fall-out or false positive rate (FPR)

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} = 1 - TNR$$

false discovery rate (FDR)

$$FDR = \frac{FP}{FP + TP} = 1 - PPV$$

false omission rate (FOR)

$$FOR = \frac{FN}{FN + TN} = 1 - NPV$$

Threat score (TS) or Critical Success Index (CSI)

$$TS = \frac{TP}{TP + FN + FP}$$

accuracy (ACC)

$$ACC = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN}$$

F1 score

is the harmonic mean of precision and sensitivity

$$F_1 = 2 \cdot \frac{PPV \cdot TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$$

Matthews correlation coefficient (MCC)

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Informedness or Bookmaker Informedness (BM)

$$BM = TPR + TNR - 1$$

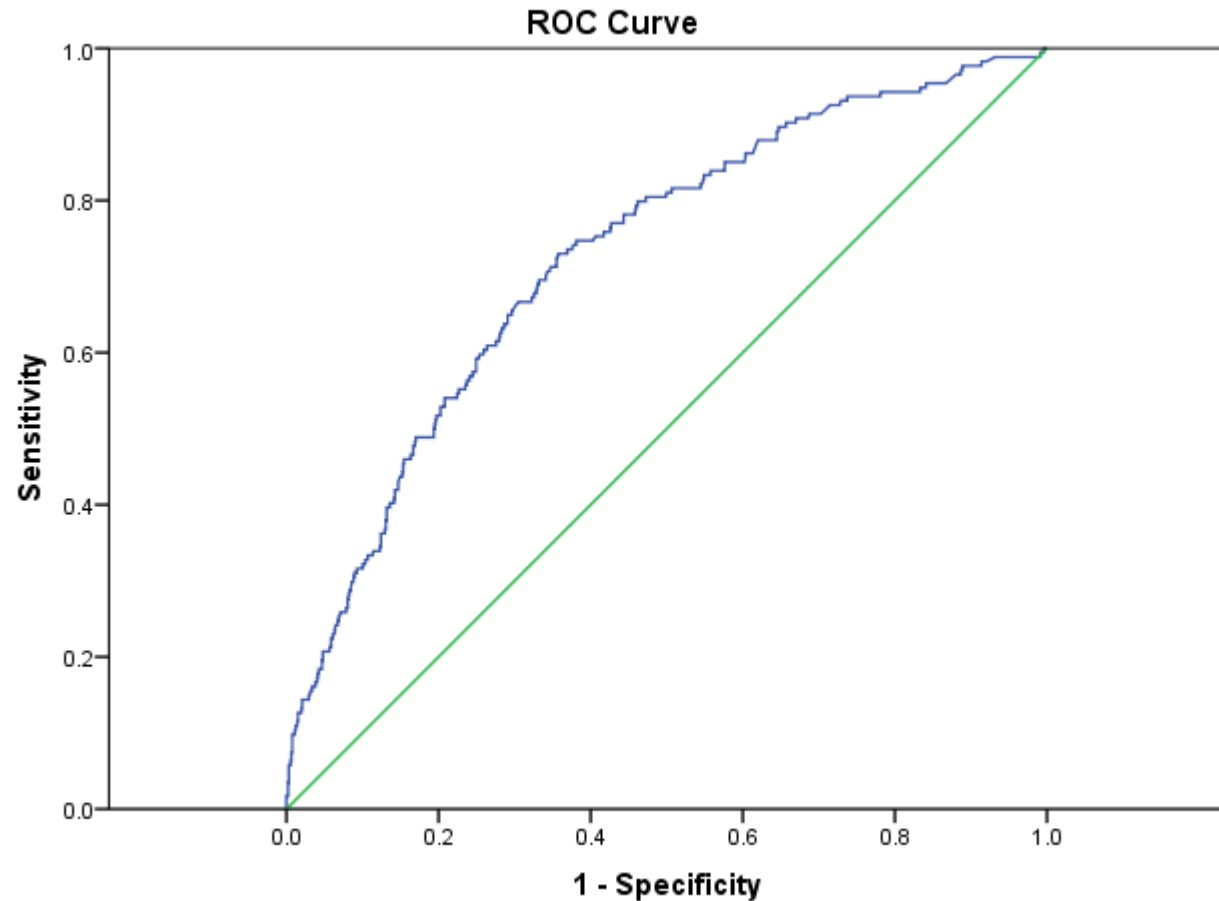
Markedness (MK)

$$MK = PPV + NPV - 1$$

Sources: Fawcett (2006),^[1] Powers (2011),^[2] Ting (2011),^[3] and CAWCR^[4] Chicco & Jurman (2020)^[5].

ROC Curve

- Receiver operator characteristics
 - The area under the ROC curve is useful for assessing models
 - A perfect classifier has a value of 1
 - A random classifier has a value of 0.5



Diagonal segments are produced by ties.

Illustrative Example

- Predicting damage to culverts in Texas



Satisfactory	Code	Meaning	Description
	9	Excellent	As new
	8	Very Good	No problems noted.
	7	Good	Some minor problems.
	6	Satisfactory	Structural elements show some minor deterioration.
Unsatisfactory	5	Fair	All primary structural elements are sound but may have minor section loss, cracking, spalling or scour.
	4	Poor	Advanced section loss, deterioration, spalling or scour.
	3	Serious	Loss of section, deterioration, spalling or scour has seriously affected primary structural components. Local failures are possible. Fatigue cracks in steel or shear cracks in concrete may be present.
	2	Critical	Advanced deterioration of primary structural elements. Fatigue cracks in steel or shear cracks in concrete may be present or scour may have removed substructure support. Unless closely monitored it may be necessary to close the bridge until corrective action is taken.
	1	Imminent Failure	Major deterioration or section loss present in critical structural components or obvious vertical or horizontal movement affecting structure stability. Bridge is closed to traffic but with corrective action may put back in light service.
	0	Failed	Out of service, beyond corrective action.

Source: United States Department of Transportation. Recording and Coding Guide for the Structure Inventory and Appraisal of the Nation's Bridges. Washington, D.C., 1995, page 38.

Data

- Data was taken from Federal Highway Administration (FHWA)
 - <https://www.fhwa.dot.gov/bridge/nbi/ascii.cfm>
 - Year 2018 data release (updated 4/22/2019)
 - Downloaded csv file and cleaned up the dataset
- Previous studies indicate
 - Reconstruction Record (1 = Reconstructed; 0 = No Reconstruction)
 - Age (Inspection Date – (re)Construction Year)
 - Age²
 - ADT (Average Daily Traffic per lane)
- The % of Truck Traffic on the Culvert could also be important
 - PTRCK
 - A measure of higher loads → greater damage potential

Culvert type could also be important but not considered here due to lack of data

Logistic Regression in Python

- We shall use scikit.learn package
 - https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
- This package does not actually maximize log likelihood but also adds a regularization term
 - Change the default 'C' value to a large number to minimize this regularization effect

Python Algorithm

- Load libraries
- Read data
 - Pandas
- Split data into training and testing
 - Scikit.learn
- Fit the model using scikit learn linear logistic regression function
 - Change c value to a large number to remove regularization
- Extract model coefficients
- Perform model evaluation on testing data
 - AUC (Area under the curve)
 - Contingency tables
 - Precision, Accuracy, Recall

Results

$$P(Y = 1) = \frac{1}{1 + EXP[-(\beta_o + \beta_1 X_1 + \dots + \beta_{k-1} X_{k-1})]}$$

Model Coefficients	Values	Remarks
Intercept	-2.07047029	
SVCYRS	4.33642499e-02	Culver age
ADT	5.18423613e-06	Average Daily Traffic
Reconst	7.86477941e-01	Whether there was reconstruction
PTRUCK	-1.28189271e-02	Percent of Truck Traffic in ADT
SVCYR ²	-6.81931225e-05	Sq. of Age

Contingency Table and Metrics

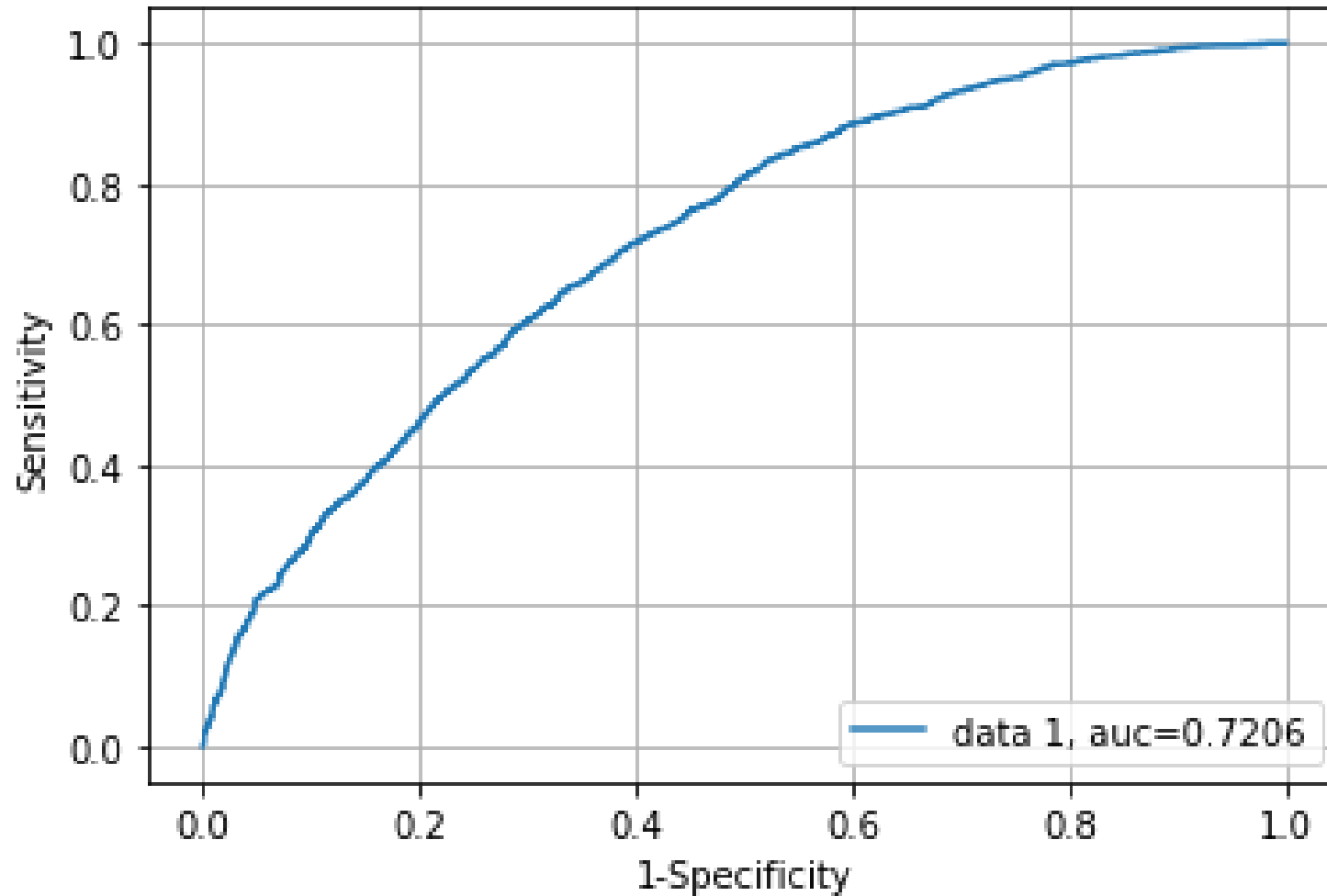
		Predicted	
		0	1
Observed	0	2716	789
	1	1230	1251

Metric	Value	Remarks
Accuracy	0.663	Fraction of ALL correct predictions
Precision	0.613	Fraction of correct 0 state predictions
Recall	0.504	Fraction of correct 1 state predictions

Model is able to better predict 'Satisfactory' states better than 'Unsatisfactory' States

One reason could be there are more satisfactory state records in the dataset then there are 'unsatisfactory' states (Unbalanced dataset)

Area Under the ROC Curve



A 0.72 AUC on
testing data points
to a reasonable
model

Overall Summary

- The 'basic' 6 parameter logistic regression model appears to be reasonable to detect culvert damage
- Contingency and AUC analysis indicate the model is reasonable based on its performance on unseen data
- The model does a better job on predicting 'Satisfactory' State than 'Unsatisfactory' state
 - Unbalanced dataset
 - Look at 'SMOTE' (Oversampling of minority data) to minimize this problem
 - Python has a package "Imbalanced" to do this - <https://imbalanced-learn.readthedocs.io/en/stable/>

You should know

- What is logistic regression
- What is the theoretical basis
- How the parameters are estimated using log-likelihood
 - Some idea of regularization that can be added to this approach
- How evaluate logistic regression classifier
 - Contingency tables
 - ROC analysis
- Use of Python libraries to do these calculations