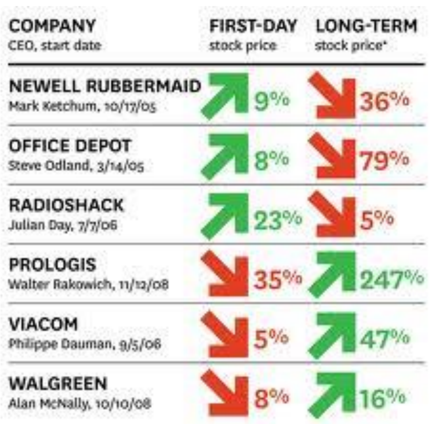
Binary Forecasts and Logistic Regression

Prof Galit Shmueli

Forecasting Analytics

Binary Forecasts



COMPANY CEO, start date	FIRST-DAY stock price	LONG-TERM stock price*
NEWELL RUBBERMAID Mark Ketchum, 10/17/05	9%	36%
OFFICE DEPOT Steve Odland, 3/14/05	8%	79%
RADIOSHACK Julian Day, 7/7/06	23%	5%
PROLOGIS Walter Rakowich, 11/12/08	35%	247%
VIACOM Philippe Dauman, 9/5/06	5%	47%
WALGREEN Alan McNally, 10/10/08	8%	16%

Forecast the **direction** of a numerical series (up/down)

Time series: numerical

Forecast: binary

Will an event **occur or not** at time *t+k*?

Time series: binary values

Forecast: binary

Will a value **cross a threshold** of interest at time *t+k*?

Time series: numerical/binary values

Forecast: binary value

Example: Daily Rainfall in Melbourne

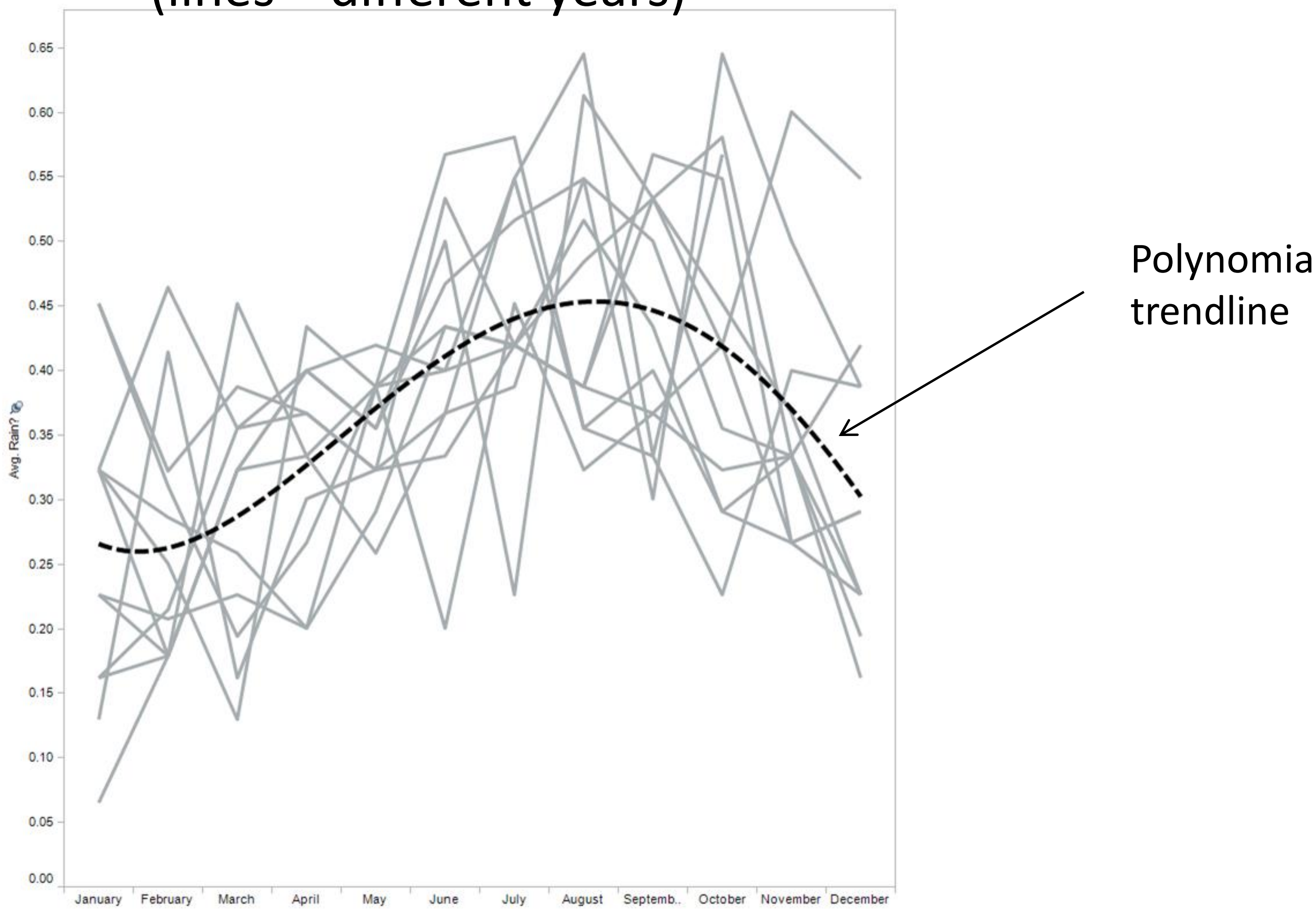
Date	Rainfall amount (millimetres)	Rain?
1/1/2000	0.4	1
1/2/2000	0	0
1/3/2000	0	0
1/4/2000	3.4	1
1/5/2000	1.4	1
1/6/2000	0	0
1/7/2000	0	0
1/8/2000	0	0
1/9/2000	0	0
1/10/2000	2.2	1
1/11/2000	0	0
1/12/2000	0	0
1/13/2000	0	0
1/14/2000	0	0
1/15/2000	0	0
1/16/2000	0.4	1
1/17/2000	0.8	1
1/18/2000	0	0
1/19/2000	0	0
1/20/2000	0	0
1/21/2000	0	0
1/22/2000	0.6	1
1/23/2000	9.4	1

"MelbourneRainfall.xlsx"
Daily rainfall amounts between Jan 1, 2000 and Oct 31, 2011, as reported by Melbourne Regional Office station
www.bom.gov.au/climate/data

Goal: next-day forecasts of rain/no-rain

Visualizing binary data: temporal aggregation

% rainy days per month
(lines = different years)



Naïve forecasts

Binary value of previous period

“Majority vote” of previous periods
(=most popular value)

Data Partitioning

Training:

Jan 1, 2000 to Dec 31, 2009

3653 days

1298 rainy, 2355 no rain

Validation:

Jan 1, 2010 to Oct 31, 2011

668 days

274 rainy, 394 no rain

MelbourneRainfallNew.xlsx
MASmoothingOutput, MASmoothingOutput1

Naïve Forecasts: previous day value

Roll-forward

MA(1)

Fitted Model

Date	Actual	Forecast	Residuals
1/1/2000	1	*	*
1/2/2000	0	1	-1
1/3/2000	0	0	0
1/4/2000	1	0	1

1/1/2010	1	0	1
1/2/2010	1	1	0
1/3/2010	0	1	-1
1/4/2010	0	0	0
1/5/2010	0	0	0
1/6/2010	0	0	0
1/7/2010	0	0	0
1/8/2010	0	0	0
1/9/2010	0	0	0
1/10/2010	0	0	0
1/11/2010	0	0	0
1/12/2010	0	0	0
1/13/2010	1	0	1
1/14/2010	0	1	-1

Predictions on Dec 31, 2009

MA(1) with "Give Forecast On Validation"

Date	Forecast	LCI	UCI
1/1/2010	0	-1.12755	1.127552
1/2/2010	0	-1.5946	1.594599
1/3/2010	0	-1.95298	1.952977
1/4/2010	0	-2.2551	2.255103
1/5/2010	0	-2.52128	2.521282
1/6/2010	0	-2.76193	2.761926
1/7/2010	0	-2.98322	2.983221
1/8/2010	0	-3.1892	3.189197
1/9/2010	0	-3.38265	3.382655
1/10/2010	0	-3.56563	3.565631
1/11/2010	0	-3.73967	3.739665
1/12/2010	0	-3.90595	3.905953
1/13/2010	0	-4.06544	4.065445
1/14/2010	0	-4.21891	4.218912
1/15/2010	0	-4.36699	4.366988
1/16/2010	0	-4.51021	4.510206
1/17/2010	0	-4.64901	4.649014

Evaluating Predictive Performance

What type of forecast errors can we get?

Error Measures (Training)

MAPE	45.703374
MAD	0.3324074
MSE	0.3324074

Do these make any sense with binary forecasts?

Error Measures (Validation)

MAPE	0
MAD	0.5898204
MSE	0.5898204

Output from binary forecasting method

Cut off Prob.Val. for Success (Updatable)	0.5

Row Id.	Predicted Class	Actual Class	Prob. for 1 (success)
1	0	1	0.450417696
2	0	0	0.449504648
3	0	0	0.206752251
4	0	1	0.206176152
5	0	1	0.446895992
6	0	0	0.446070732
7	0	0	0.204541535
8	0	0	0.204028139
9	0	0	0.203530597
10	0	1	0.203048976

Cutoff Value
(default=0.5)

Summarizing Forecast Errors: Classification Matrix

Training Data Scoring - Summary Report

Cutoff probability value for success (UPDATABLE)

Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	492	806
0	375	1980

Classification Matrix and error rates: training and validation

Training Data Scoring - Summary Report

Cutoff probability value for success (UPDATABLE)

Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	492	806
0	375	1980

Validation Data Scoring - Summary Report

Cutoff probability value for success (UPDATABLE)

Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	102	173
0	74	320

When one outcome (C_1) is more important

Sensitivity of a classifier =
its ability to **correctly detect** C_1 periods
= % correctly classified C_1 periods

Specificity of a classifier =
its ability to **correctly rule out** C_2 periods
= % correctly classified C_2 periods

FORECASTING WITH LOGISTIC REGRESSION

Regression model for Melbourne rainfall

Potential Predictors:

- annual seasonality (sine, cosine)
- Previous day(s) Rain indicator or rainfall amount

How about linear regression?

$$\text{Rain}_t = \beta_0 + \beta_1 \text{Rain}_{t-1} + \beta_2 \, \mathit{sin}(2\pi t/365.25) + \beta_3 \, \mathit{cos}(2\pi t/365.25) + \varepsilon$$

Create predictors in spreadsheet

Date	Rainfall amount (mm)	Rainfall lag-1	Rain?	Lag1	t	Seasonal_sine	Seasonal_cosine
1/1/2000	0.4	1.8	1	1	1	0.017201575	0.999852042
1/2/2000	0	0.4	0	1	2	0.034398061	0.999408212
1/3/2000	0	0	0	0	3	0.051584367	0.99866864
1/4/2000	3.4	0	1	0	4	0.068755408	0.997633547
1/5/2000	1.4	3.4	1	1	5	0.085906104	0.996303238
1/6/2000	0	1.4	0	1	6	0.103031379	0.994678106
1/7/2000	0	0	0	0	7	0.120126165	0.992758634
1/8/2000	0	0	0	0	8	0.137185404	0.990545388
1/9/2000	0	0	0	0	9	0.154204048	0.988039023
1/10/2000	2.2	0	1	0	10	0.17117706	0.985240283
1/11/2000	0	2.2	0	1	11	0.188099418	0.982149993
1/12/2000	0	0	0	0	12	0.204966114	0.97876907
1/13/2000	0	0	0	0	13	0.221772158	0.975098513
1/14/2000	0	0	0	0	14	0.238512575	0.971139409
1/15/2000	0	0	0	0	15	0.255182413	0.966892929
1/16/2000	0.4	0	1	0	16	0.271776738	0.96236033
1/17/2000	0.8	0.4	1	1	17	0.288290641	0.957542953
1/18/2000	0	0.8	0	1	18	0.304719233	0.952442223
1/19/2000	0	0	0	0	19	0.321057654	0.947059651
1/20/2000	0	0	0	0	20	0.337301069	0.941396829

Note: In this example we are using an extrapolation model (only uses its past history)

We can use external predictors in addition/instead

Partition data (on-the-fly), and run linear regression

Regression Model

Input Variables	Coefficient	Std. Error	t-Statistic	P-Value
Intercept	0.262573	0.009457518	27.76341079	4.4399E-154
Lag1	0.260855	0.015976534	16.32740444	6.78548E-58
Seasonal_sine	-0.04676	0.010743837	-4.35241212	1.38339E-05
Seasonal_cosine	-0.05954	0.010764875	-5.53104688	3.40685E-08

Training Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
762.0157649	0.456727	-1.85416E-15

What is the problem with this approach?

Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
148.1441394	0.470575	0.039106295

MelbourneRainfallNew.xlsx
Worksheet *MLR_Output*

Logistic regression

- Common for modeling **cross-sectional data**
- **Predict** a binary outcome, given a set of predictors
  - fraud/non-fraud, buyer/non-buyer
  - called "**classification"** in data mining
- **Explain/describe** difference between classes as a function of input variables
  - (male/female, online/offline users)
- Provides output similar to linear regression
  - Coefficients
  - statistical significance

Logistic Regression

$$\text{Rain}_t = \beta_0 + \beta_1 \text{ Rain}_{t-1} + \beta_2 \; \textit{sin}(2\pi t/365.25) + \beta_3 \, \textit{cos}(2\pi t/365.25) + \varepsilon$$

Replace with a **function of "Rain"** that guarantees forecasts in range [0,1] and give probability of rain

The logit function

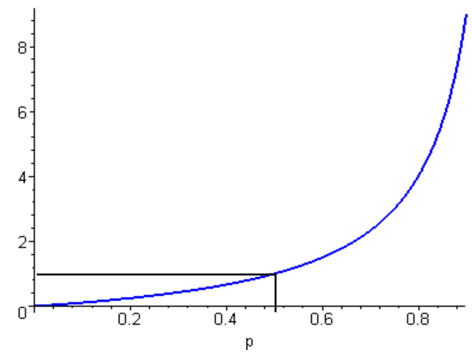
Output variable:
Rain_t (binary variable)

p =Prob(Rain_t =1)

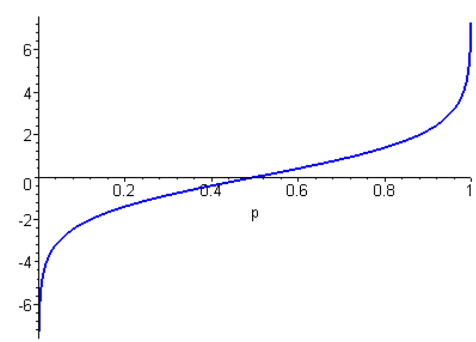
odds(Rain_t =1) = $\frac{p}{1-p}$

logit (Rain_t=1) = log(odds)

odds



logit



Logistic regression formula

$$\text{Logit}(\text{Rain}_t\text{=1}) = \beta_0 + \beta_1\ \text{Rain}_{t\text{-}1} + \beta_2\ \textit{sin}(2\pi t/365.25) + \beta_3\ \textit{cos}(2\pi t/365.25)$$

$$\text{Odds}(\text{Rain}_t\text{=1}) = e^{\beta_0 + \beta_1\ \text{Raint}_{-1} + \beta_2\ \textit{sin}(2\pi t/365.25) + \beta_3\ \textit{cos}(2\pi t/365.25)}$$

$$\text{Prob}(\text{Rain}_t\text{=1}) = \frac{1}{1+e^{-\{\beta_0 + \beta_1\ \text{Raint}_{-1+}\beta_2 \textit{sin}(2\pi t/365.25) + \beta_3 \textit{cos}(2\pi t/365.25)\}}}$$

$$Prob(Rain_t = 1) = \frac{1}{1 + e^{-logit}}$$

Logistic regression estimation

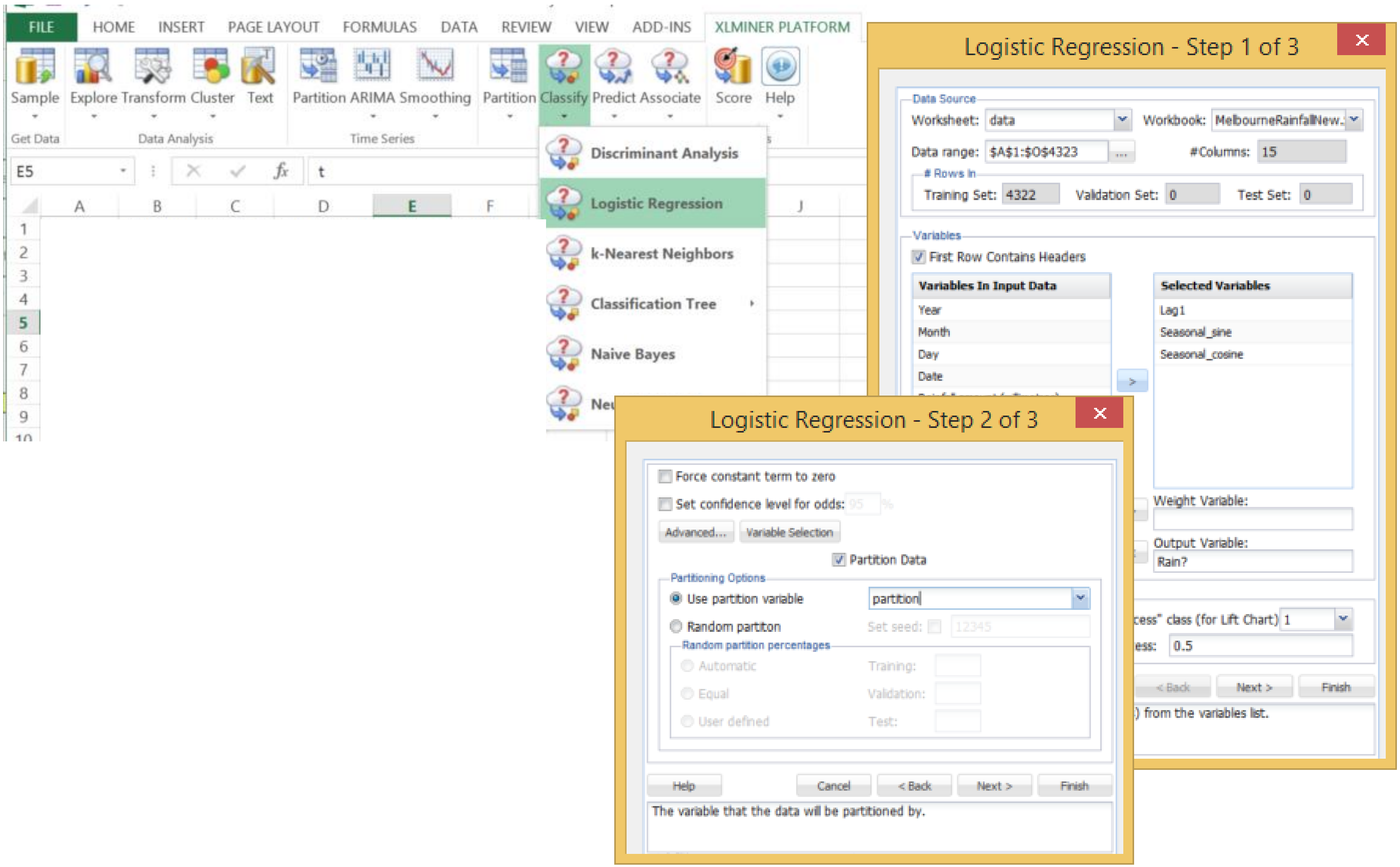
$$\text{Logit}(\text{Rain}_t=1) = \beta_0 + \beta_1 \text{ Rain}_{t-1} + \beta_2 \; \mathit{sin}(2\pi t/365.25) + \beta_3 \, \mathit{cos}(2\pi t/365.25)$$

Least Squares impossible

Instead: **Maximum Likelihood Estimation**

(find estimates that maximize the chance of obtaining the data that we see); done iteratively

Running logistic regression in XLMiner



Regression Model

Input Variables	Coefficient	Std. Error	Chi2-Statistic	P-Value
Intercept	-1.05077	0.047483	489.7064167	1.6508E-108
Lag1	1.138336	0.073761	238.1692585	9.86041E-54
Seasonal_sine	-0.22183	0.051336	18.67178847	1.55263E-05
Seasonal_cosine	-0.28278	0.051543	30.09876221	4.10593E-08

Training Data Scoring - Summary Report

Cutoff probability value for success (UPDATABLE)	

Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	492	806
0	375	1980

Error Report			
Class	# Cases	# Errors	% Error
1	1298	806	62.09553159
0	2355	375	15.92356688
Overall	3653	1181	32.32959212

Validation Data Scoring - Summary Report

Cutoff probability value for success (UPDATABLE)	0.5

Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	102	173
0	74	320

Error Report			
Class	# Cases	# Errors	% Error
1	275	173	62.90909091
0	394	74	18.78172589
Overall	669	247	36.92077728

Using the model to forecast 1/1/2010

Regression Model

Input Variables	Coefficient	Std. Error	Chi2-Statistic	P-Value	Odds
Intercept	-1.05077	0.047483	489.7064167	1.6508E-108	0.34967
Lag1	1.138336	0.073761	238.1692585	9.86041E-54	3.12157
Seasonal_sine	-0.22183	0.051336	18.67178847	1.55263E-05	0.801055
Seasonal_cosine	-0.28278	0.051543	30.09876221	4.10593E-08	0.753688

Row Id.	Date	Rain?	Lag1
3652	12/30/2009	0	0
3653	12/31/2009	0	0
3654	1/1/2010	1	0

Software generates forecasts for training & validation

	Predicted Class	Actual Class	Success Probability	Log Odds	Lag1	Seasonal_sine	Seasonal_cosine
1/1/2010	0	1	0.207646189	-1.3392	0	0.025800772	0.999667105
1/2/2010	0	1	0.449056738	-0.2045	1	0.042992804	0.999075382
1/3/2010	0	0	0.448176166	-0.208	1	0.060172113	0.998188017
1/4/2010	0	0	0.20589473	-1.3499	0	0.077333617	0.997005272
1/5/2010	0	0	0.205342005	-1.3532	0	0.094472236	0.995527497
1/6/2010	0	0	0.204804958	-1.3565	0	0.1115829	0.993755129
1/7/2010	0	0	0.204283661	-1.3597	0	0.128660544	0.991688693
1/8/2010	0	0	0.203778185	-1.3628	0	0.145700115	0.989328801

To forecast **future** values, re-combine training/validation and refit logistic regression

Try to improve the model

Change the cutoff (in XLMiner, interactively)

Try other/additional predictors - what are some options here?

Trying lots of models and comparing performance on validation period: beware of **over-fitting**!

Example with External Predictors: Forecasting *Powdery Mildew* epidemic in mango

Big problem in Uttar Pradesh!

Epidemic hits week 3 or 4 of March each year

Airborne disease, affected by temperature, humidity, wind velocity, dews, wind direction…

Goal: in the 2nd week of March, forecast an outbreak

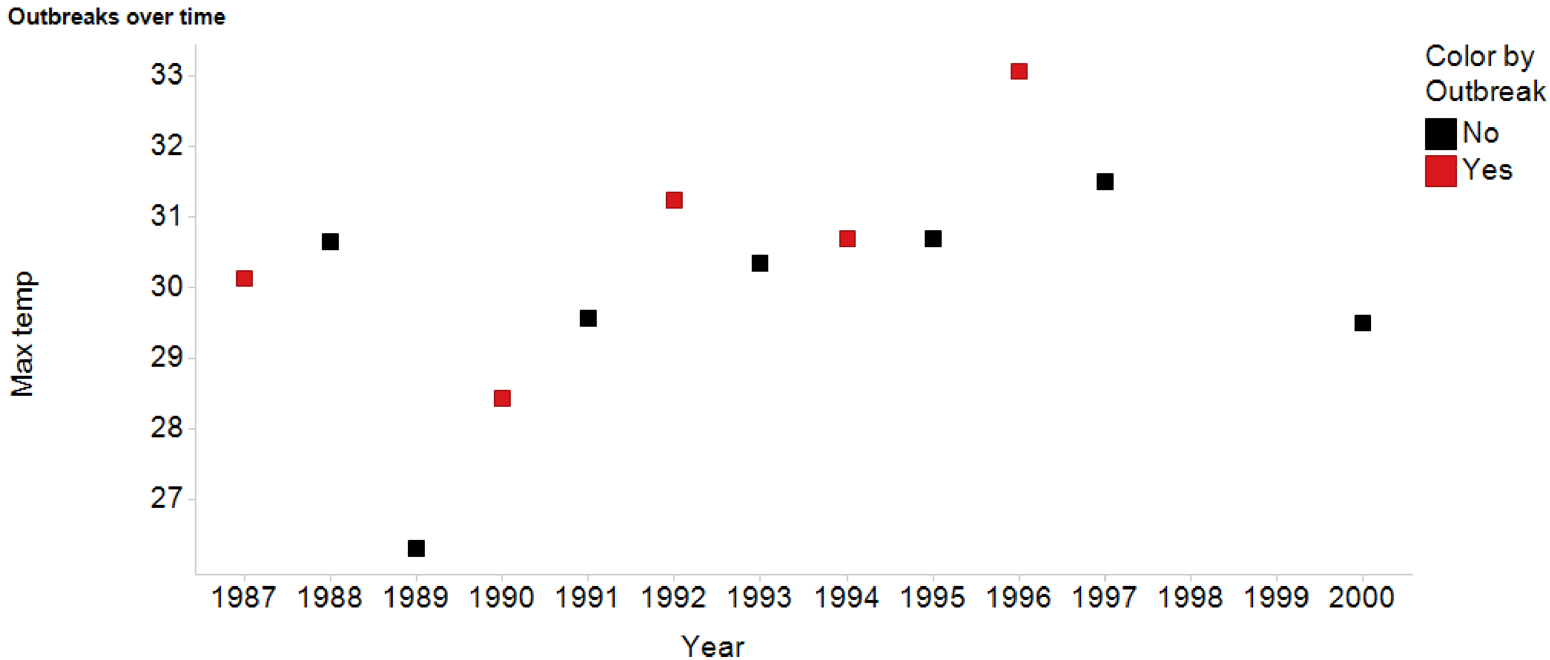
Forecasting *Powdery Mildew* epidemic in mango

"PowderyMildewEpidemic.xlsx"

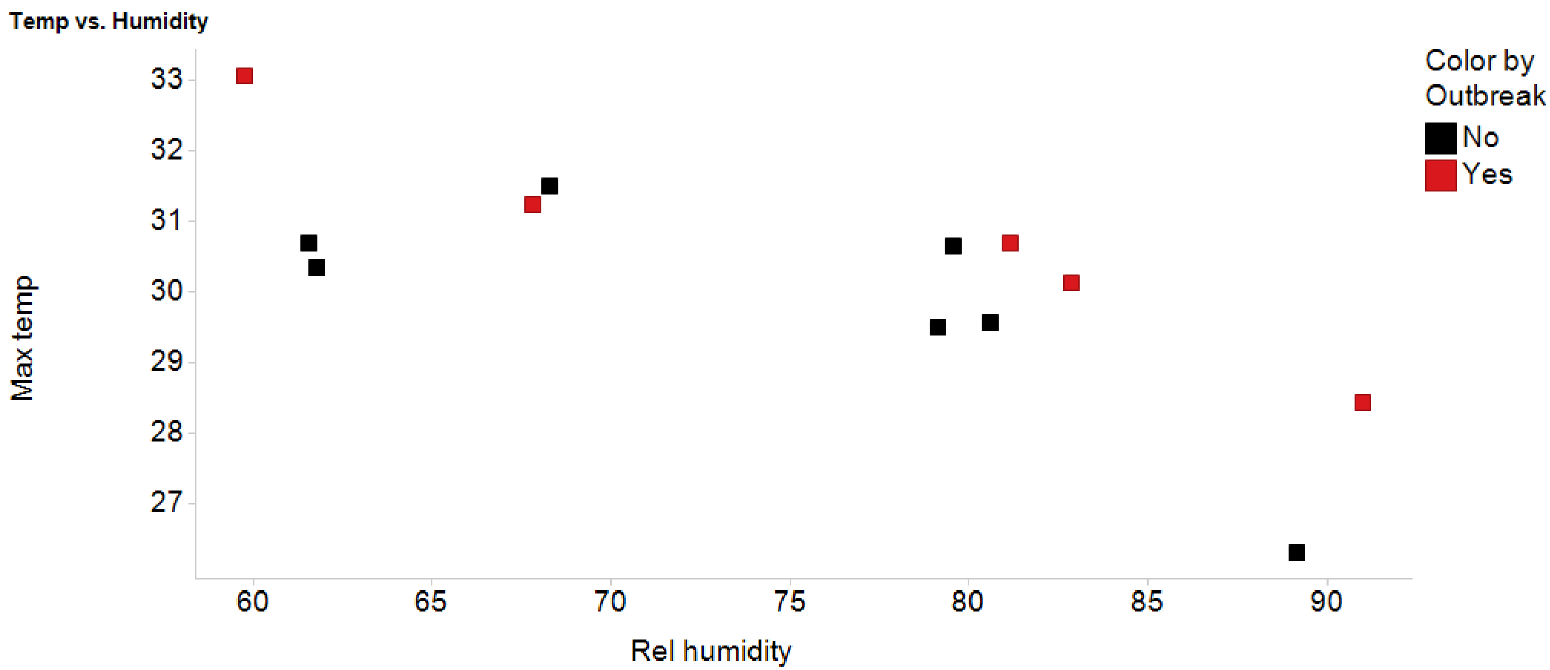
Annual outbreak and weather information on Powdery Mildew epidemic in Uttar Pradesh

Predictors: max temp and relative humidity in second week of March

Year	Outbreak?	Max temperature	Relative humidity
1987	Yes	30.14	82.86
1988	No	30.66	79.57
1989	No	26.31	89.14
1990	Yes	28.43	91.00
1991	No	29.57	80.57
1992	Yes	31.25	67.82
1993	No	30.35	61.76
1994	Yes	30.71	81.14
1995	No	30.71	61.57
1996	Yes	33.07	59.76
1997	No	31.50	68.29
2000	No	29.50	79.14



Outbreaks over time
Max temp
33
32
31
30
29
28
27
1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000
Year
Color by Outbreak
No
Yes



Temp vs. Humidity
Max temp
33
32
31
30
29
28
27
60 65 70 75 80 85 90
Rel humidity
Color by Outbreak
No
Yes

Naïve roll-forward forecasts

Forecasted year	Available data	Naïve last-value Forecast	Naïve majority vote (50%)
1996	1987-1995	No	No (5-4)
1997	1987-1996	Yes	No (5-5)
2000	1987-1997 (no data for 98-99)	No	No (6-5)

Year	Outbreak?	Max temperature	Relative humidity
1987	Yes	30.14	82.86
1988	No	30.66	79.57
1989	No	26.31	89.14
1990	Yes	28.43	91.00
1991	No	29.57	80.57
1992	Yes	31.25	67.82
1993	No	30.35	61.76
1994	Yes	30.71	81.14
1995	No	30.71	61.57
1996	Yes	33.07	59.76
1997	No	31.50	68.29
2000	No	29.50	79.14

Powdery Mildew Epidmic.xlsx
LR_Output, LR_ValidationScore

Logistic Regression

Training: 1987-1994, Validation:1995-2000

Regression Model

Input Variables	Coefficient	Std. Error	Chi2-Statistic	P-Value	Odds
Intercept	-56.1522	44.45559	1.595441736	0.20655	4.11E-25
Max temp	1.384798	1.140566	1.474117141	0.224697	3.99402
Rel humidi	0.187657	0.157789	1.414419855	0.234324	1.20642

Training Data Scoring - Summary Report

Cutoff probability value for success (UPDATABLE)	0.5

Confusion Matrix

	Predicted Class	
Actual Clas	Yes	No
Yes	3	1
No	1	3

Error Report

Class	# Cases	# Errors	% Error
Yes	4	1	25
No	4	1	25
Overall	8	2	25

XLMiner : Logistic Regression - Classification of Validation Data

Predicted Class	Actual Class	Success Probability	Log Odds	Max temp	Rel humidity
No	No	0.111948	-2.071	30.71	61.57
Yes	Yes	0.702131	0.8575	33.07	59.76
Yes	No	0.570539	0.2841	31.5	68.29
No	No	0.389488	-0.4495	29.5	79.14

Validation Data Scoring - Summary Report

Cutoff probability value for success (UPDATABLE)	0.5

Confusion Matrix

	Predicted Class	
Actual Clas	Yes	No
Yes	1	0
No	1	2

Error Report

Class	# Cases	# Errors	% Error
Yes	1	0	0
No	3	1	33.33333333
Overall	4	1	25

Roll-forward: re-run regression

Regression Model Training: 1987-1994

Input Variables	Coefficient	Std. Error	Chi2-Statistic	P-Value	Odds
Intercept	-56.1522	44.45559	1.595441736	0.20655	4.11E-25
Max temp	1.384798	1.140566	1.474117141	0.224697	3.99402
Rel humidi	0.187657	0.157789	1.414419855	0.234324	1.20642

Regression Model Training: 1987-1995

Input Variables	Coefficient	Std. Error	Chi2-Statistic	P-Value	Odds
Intercept	-62.1715	44.96623	1.911653701	0.16678	9.98E-28
Max temp	1.508743	1.178528	1.638895022	0.200477	4.521045
Rel humidi	0.215877	0.14926	2.091811685	0.14809	1.240949

Regression Model Training: 1987-1996

Input Variables	Coefficient	Std. Error	Chi2-Statistic	P-Value	Odds
Intercept	-72.4698	46.81975	2.395826213	0.121659	3.36E-32
Max temp	1.845462	1.201978	2.357312054	0.124697	6.331026
Rel humidi	0.21982	0.152958	2.065323572	0.150683	1.245852

Roll forward data partitioning (1-step ahead forecasting)

fit logistic regression three times

Training period	Validation period	Naïve validation forecast	Logistic validation forecast
1987-1994	1995 (No)	Yes	No
1987-1995	1996 (Yes)	No	Yes
1987-1996	1997 (No)	Yes	Yes

Predicted Class	Actual Class	Success Probability	Log Odds	Max temp	Rel humidity
No	No	0.111947966	-2.071	30.71	61.57

Predicted Class	Actual Class	Success Probability	Log Odds	Max temp	Rel humidity
Yes	Yes	0.651006367	0.6235	33.07	59.76

Predicted Class	Actual Class	Success Probability	Log Odds	Max temp	Rel humidity
Yes	No	0.66235304	0.6738	31.5	68.29

Year	Outbreak?	Max temperature	Relative humidity
1987	Yes	30.14	82.86
1988	No	30.66	79.57
1989	No	26.31	89.14
1990	Yes	28.43	91.00
1991	No	29.57	80.57
1992	Yes	31.25	67.82
1993	No	30.35	61.76
1994	Yes	30.71	81.14
1995	No	30.71	61.57
1996	Yes	33.07	59.76
1997	No	31.50	68.29
2000	No	29.50	79.14

Roll-forward validation performance

Training period	Validation period	Naïve validation forecast	Logistic validation forecast
1987-1994	1995 (No)	Yes	No
1987-1995	1996 (Yes)	No	Yes
1987-1996	1997 (No)	Yes	Yes

Classification matrix

	Predicted epidemic	Predicted no epidemic
Epidemic	1	0
No epidemic	1	1

How about extrapolation model?

$$\text{Logit}(\text{Epidemic}_t = 1) = \beta_0 + \beta_1 \text{Epidemic}_{t-1}$$

Can we forecast year 2000?

Summary

Logistic regression can be used to forecast binary values based on

- Previous binary values
- Previous numerical values
- Trend and seasonality predictors
- External predictors
- Interaction terms

Like linear regression, it is model-driven

- Estimate the model from the training period
- Evaluate on validation period
- Re-run model on complete series to create forecasts