# Chapter 10 – Logistic Regression

## Data Mining for Business Intelligence

Shmueli, Patel & Bruce

# Logistic Regression

- Extends idea of linear regression to situation where outcome variable is categorical
- Widely used, particularly where a structured model is useful to explain (=*profiling*) or to predict
- We focus on binary classification
  i.e. $Y$=0 or $Y$=1

# The Logit

**Goal:** Find a function of the predictor variables that relates them to a 0/1 outcome

- Instead of *Y* as outcome variable (like in linear regression), we use a function of Y called the ***logit***
- Logit can be modeled as a linear function of the predictors
- The logit can be mapped back to a probability, which, in turn, can be mapped to a class

# Step 1: Logistic Response Function

$p$ = probability of belonging to class 1

Need to relate $p$ to predictors with a function that guarantees $0 \leq p \leq 1$

Standard linear function (as shown below) does not:

$$p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_q x_q$$

$q$ = number of predictors

# The Fix:
# use *logistic response function*

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_q x_q)}}$$

Equation 10.2 in textbook

# Step 2: The Odds

The odds of an event are defined as:

eq. 10.3

$$Odds = \frac{p}{1-p}$$

$p$ = probability of event

Or, given the odds of an event, the probability of the event can be computed by:

eq. 10.4

$$p = \frac{Odds}{1+Odds}$$

# We can also relate the Odds to the predictors:

eq. 10.5

$$Odds = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q}$$

To get this result, substitute 10.2 into 10.4

# Step 3: Take log on both sides

This gives us the logit:

$$\log(Odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q$$

*log(Odds)* = *logit* (eq. 10.6)

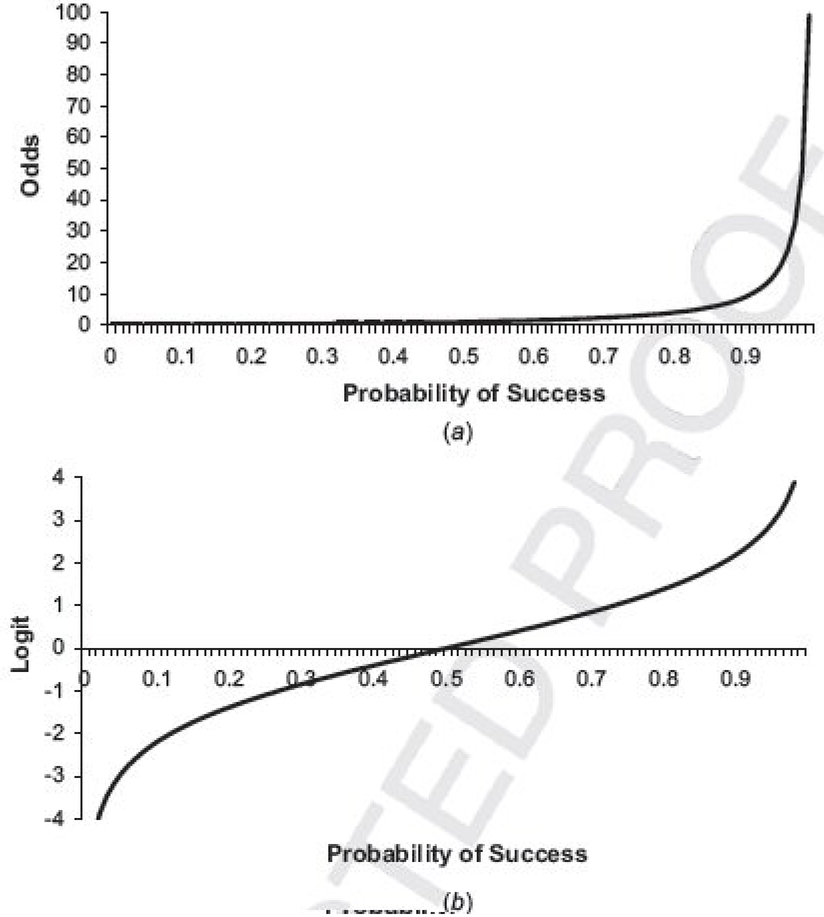
# Logit, cont.

So, the logit is a linear function of predictors $x_1$, $x_2$, …

- Takes values from -infinity to +infinity

Review the relationship between logit, odds and probability

# Odds (a) and Logit (b) as function of *P*



# Example

# Personal Loan Offer

**Outcome variable**: accept bank loan (0/1)

**Predictors:** Demographic info, and info about their bank relationship

# Data preprocessing

- Partition 60% training, 40% validation
- Create 0/1 dummy variables for categorical predictors

$$EducProf = \begin{cases} 1 \text{ if education is } \textit{Professional} \\ 0 \text{ otherwise} \end{cases}$$

$$EducGrad = \begin{cases} 1 \text{ if education is at } \textit{Graduate} \text{ level} \\ 0 \text{ otherwise} \end{cases}$$

$$Securities = \begin{cases} 1 \text{ if customer has securities account in bank} \\ 0 \text{ otherwise} \end{cases}$$

$$CD = \begin{cases} 1 \text{ if customer has CD account in bank} \\ 0 \text{ otherwise} \end{cases}$$

$$Online = \begin{cases} 1 \text{ if customer uses online banking} \\ 0 \text{ otherwise} \end{cases}$$

$$CreditCard = \begin{cases} 1 \text{ if customer holds Universal Bank credit card} \\ 0 \text{ otherwise} \end{cases}$$

# Single Predictor Model

Modeling loan acceptance on income (*x*)

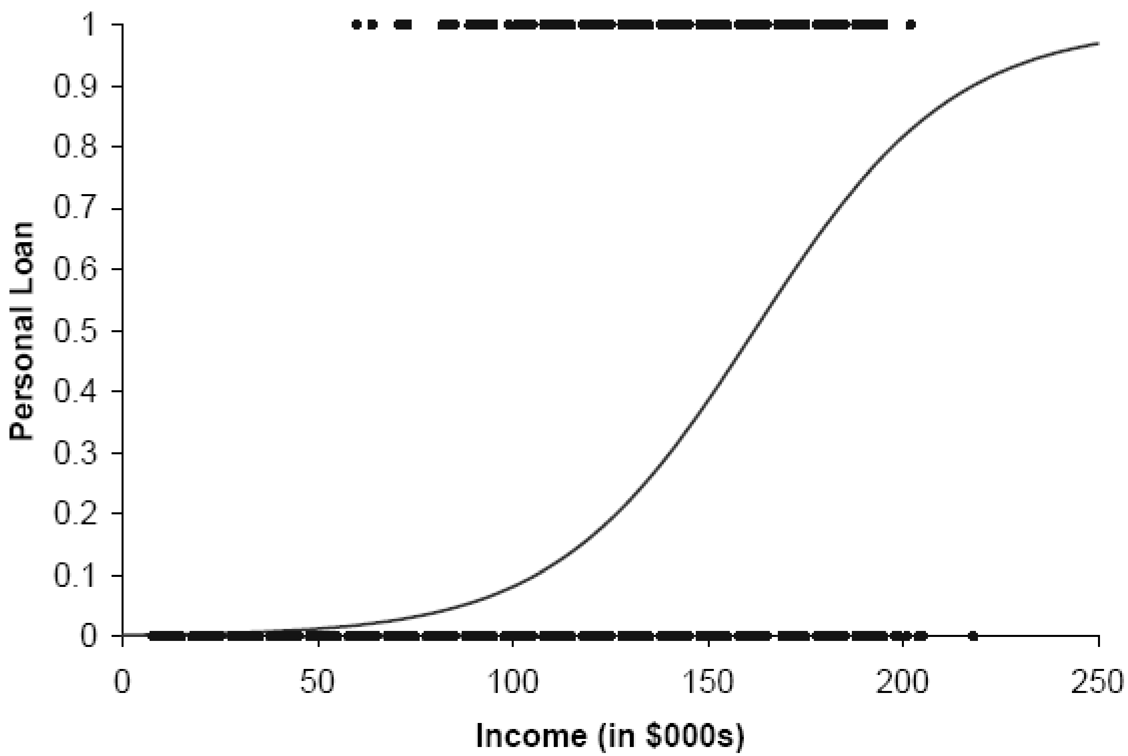
$$\text{Prob}(Personal\ Loan = Yes \mid Income = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Fitted coefficients (more later): $b_0$ = -6.3525, $b_1$ = -0.0392

$$P(Personal\ Loan = Yes \mid Income = x) = \frac{1}{1 + e^{6.3525 - 0.0392x}}$$

# Seeing the Relationship

$$P(Personal\ Loan = Yes \mid Income = x) = \frac{1}{1 + e^{6.3525 - 0.0392x}}$$



# Last step - classify

Model produces an estimated probability of being a “1”

- Convert to a classification by establishing cutoff level
- If estimated prob. > cutoff, classify as “1”

# Ways to Determine Cutoff

- 0.50 is popular initial choice

- Additional considerations (see Chapter 5)
  - Maximize classification accuracy
  - Maximize sensitivity (subject to min. level of specificity)
  - Minimize false positives (subject to max. false negative rate)
  - Minimize expected cost of misclassification (need to specify costs)

# Example, cont.

- Estimates of $\beta$'s are derived through an iterative process called *maximum likelihood estimation*

- Let's include all 12 predictors in the model now

- XLMiner's output gives coefficients for the logit, as well as odds for the individual terms

## The Regression Model

| Input variables | Coefficient | Std. Error | p-value | Odds |
|---|---|---|---|---|
| Constant term | -13.20165825 | 2.46772742 | 0.00000009 | * |
| Age | -0.04453737 | 0.09096102 | 0.62439483 | 0.95643985 |
| Experience | 0.05657264 | 0.09005365 | 0.5298661 | 1.05820346 |
| Income | 0.0657607 | 0.00422134 | 0 | 1.06797111 |
| Family | 0.57155931 | 0.10119002 | 0.00000002 | 1.77102649 |
| CCAvg | 0.18724874 | 0.06153848 | 0.00234395 | 1.20592725 |
| Mortgage | 0.00175308 | 0.00080375 | 0.02917421 | 1.00175464 |
| Securities Account | -0.85484785 | 0.41863668 | 0.04115349 | 0.42534789 |
| CD Account | 3.46900773 | 0.44893095 | 0 | 32.10486984 |
| Online | -0.84355801 | 0.22832377 | 0.00022026 | 0.43017724 |
| CreditCard | -0.96406376 | 0.28254223 | 0.00064463 | 0.38134006 |
| EducGrad | 4.58909273 | 0.38708162 | 0 | 98.40509796 |
| EducProf | 4.52272701 | 0.38425466 | 0 | 92.08635712 |

Figure 10.3: Logistic regression coefficient table for personal loan acceptance as a function of 12 predictors.

# Estimated Equation for Logit (Equation 10.9)

$$\begin{aligned}\text{logit} = &-13.201 - 0.045 Age + 0.057 Experience + 0.066 Income + 0.572 Family \\ &+0.18724874 CCAvg + 0.002 Mortgage - 0.855 Securities + 3.469 CD \\ &-0.844 Online - 0.964 Credit\ Card + 4.589 EducGrad + 4.523 EducProf\end{aligned}$$

# Equation for Odds (Equation 10.10)

$$\begin{aligned}\text{odds}(Personal\ Loan = \text{Yes}) \;=\; & e^{-13.201}(0.956)^{Age}\ (1.058)^{Experience}\ (1.068)^{Income} \\ & \cdot\ (1.771)^{Family}\ (1.206)^{CCAvg}\ (1.002)^{Mortgage} \\ & \cdot\ (0.425)^{Securities}\ (32.105)^{CD}\ (0.430)^{Online} \\ & \cdot\ (0.381)^{CreditCard}(98.405)^{EducGrad}\ (92.086)^{EducProf}\end{aligned}$$

# Converting to Probability

$$p = \frac{Odds}{1 + Odds}$$

# Interpreting Odds, Probability

For predictive classification, we typically use probability with a cutoff value

For explanatory purposes, odds have a useful interpretation:

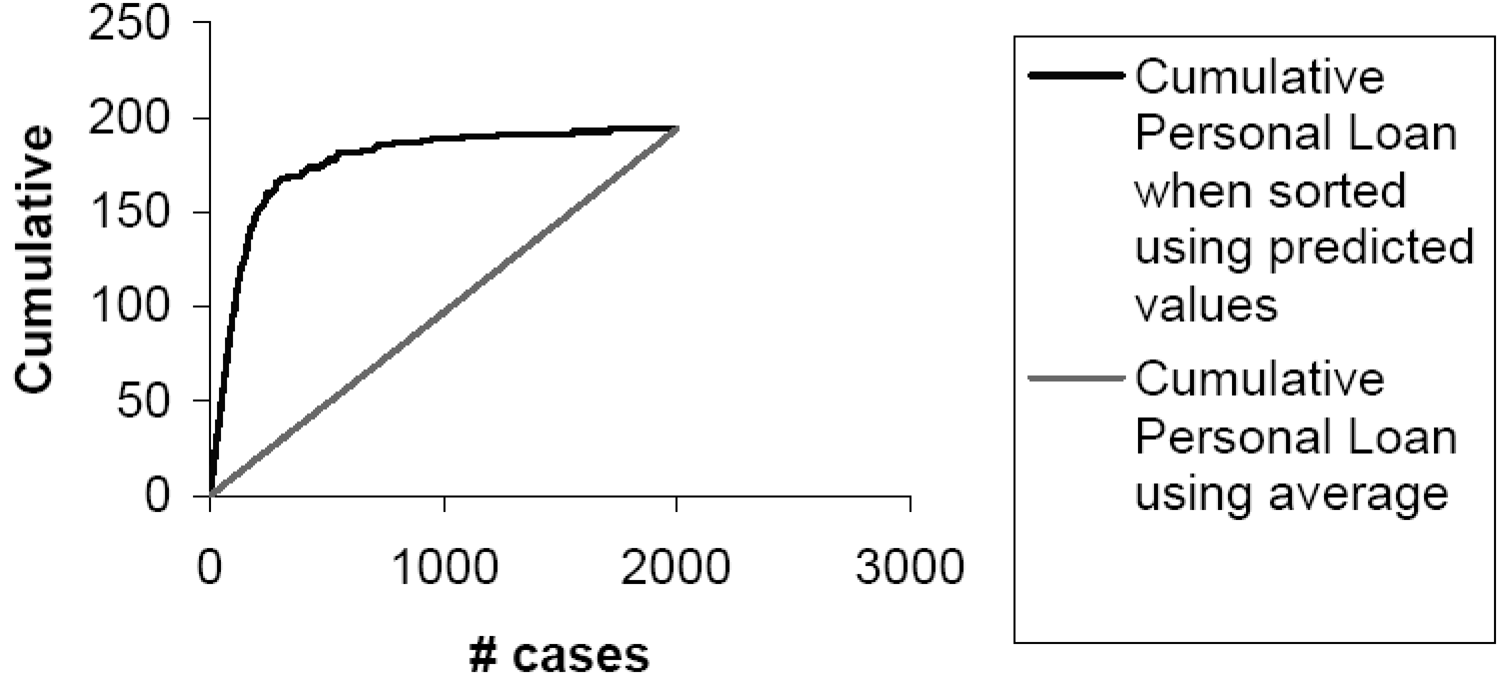
- If we increase $x_1$ by one unit, holding $x_2$, $x_3$ … $x_q$ constant, then
- $b_1$ is the factor by which the odds of belonging to class 1 increase

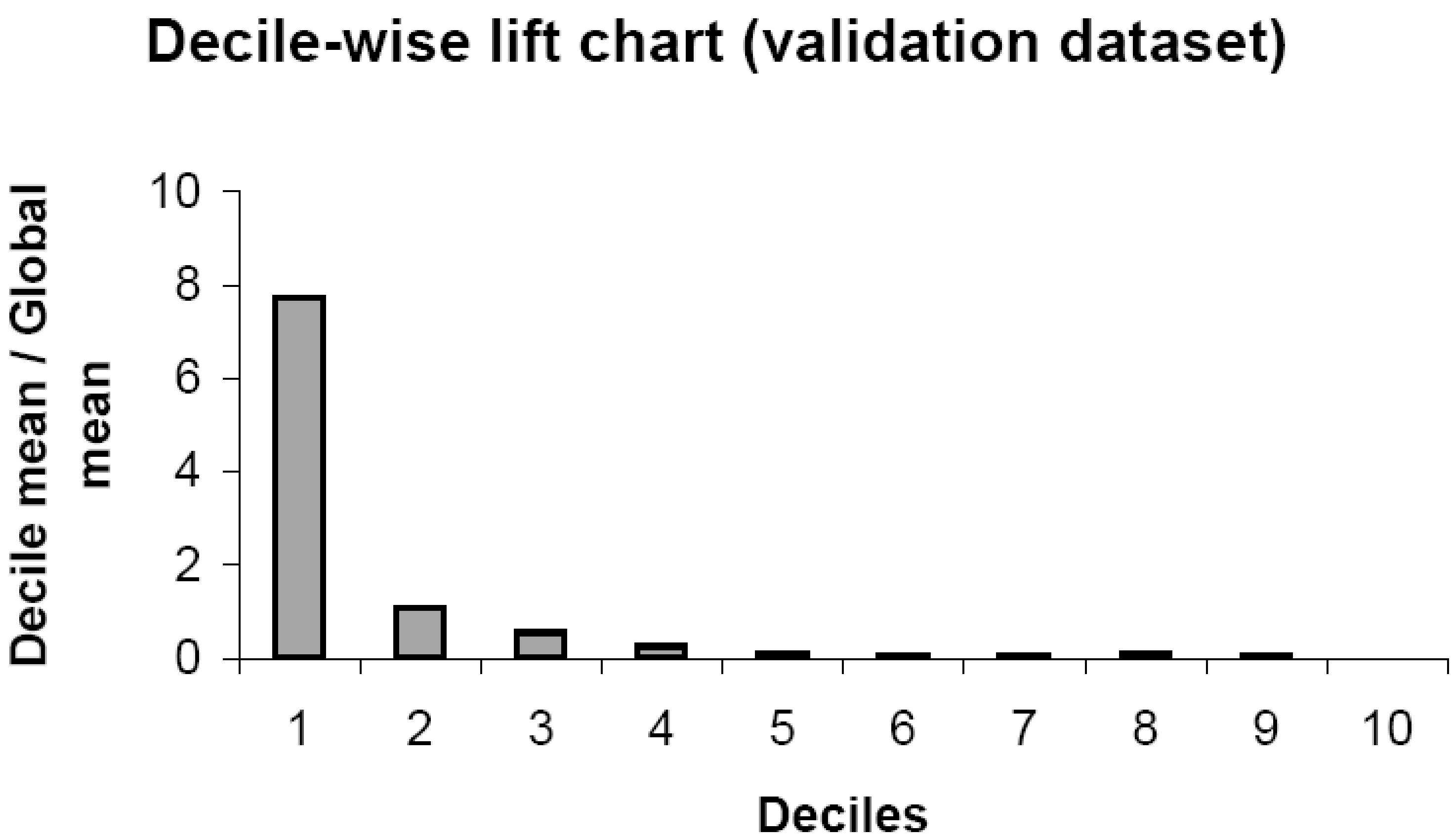
# Loan Example: Evaluating Classification Performance

Performance measures: Confusion matrix and % of misclassifications

More useful in this example: **lift**

# Lift chart (validation dataset)





Decile-wise lift chart (validation dataset)
Decile mean / Global mean
10
8
6
4
2
0
1
2
3
4
5
6
7
8
9
10
Deciles

# Multicollinearity

**Problem:** As in linear regression, if one predictor is a linear combination of other predictor(s), model estimation will fail

- Note that in such a case, we have at least one redundant predictor

**Solution:** Remove extreme redundancies (by dropping predictors via variable selection – see next, or by data reduction methods such as PCA)

# Variable Selection

This is the same issue as in linear regression

- The number of correlated predictors can grow when we create derived variables such as **interaction terms** (e.g. *Income x Family)*, to capture more complex relationships

- Problem: Overly complex models have the danger of overfitting

- Solution: Reduce variables via automated selection of variable subsets (as with linear regression)

# P-values for Predictors

- Test null hypothesis that coefficient = 0
- Useful for review to determine whether to include variable in model
- Key in profiling tasks, but less important in predictive classification

# Complete Example:
# Predicting Delayed Flights DC to NY

# Variables

**Outcome:** delayed or not-delayed

**Predictors:**

- Day of week
- Departure time
- Origin (DCA, IAD, BWI)
- Destination (LGA, JFK, EWR)
- Carrier
- Weather (1 = bad weather)

# Data Preprocessing

Create binary dummies for the categorical variables

Partition 60%-40% into training/validation

# The Fitted Model (not all 28 variables shown)

| Input variables | Coefficient | Std. Error | p-value | Odds |
|---|---|---|---|---|
| Constant term | -2.76648855 | 0.60903645 | 0.00000556 | * |
| Weather | 16.94781685 | 472.3040772 | 0.97137541 | 22926812 |
| ORIGIN_BWI | 0.31663841 | 0.407509 | 0.43715307 | 1.37250626 |
| ORIGIN_DCA | -0.52621925 | 0.37920129 | 0.1652271 | 0.59083456 |
| DEP_TIME_BLK_0700-0759 | 0.17635399 | 0.52038968 | 0.73469388 | 1.19286025 |
| DEP_TIME_BLK_0800-0859 | 0.37122276 | 0.4879483 | 0.44678667 | 1.44950593 |
| DEP_TIME_BLK_0900-0959 | -0.2891154 | 0.61024719 | 0.6356656 | 0.74892575 |
| DEP_TIME_BLK_1000-1059 | -0.84254718 | 0.65849793 | 0.20072155 | 0.4306123 |
| DEP_TIME_BLK_1100-1159 | 0.26919952 | 0.62188113 | 0.66510242 | 1.30891633 |
| DEP_TIME_BLK_1200-1259 | 0.39577994 | 0.47712085 | 0.40681183 | 1.48554242 |
| DEP_TIME_BLK_1300-1359 | 0.23689635 | 0.49711299 | 0.63368666 | 1.26730978 |
| DEP_TIME_BLK_1400-1459 | 0.94953001 | 0.4257178 | 0.02571949 | 2.58449459 |
| DEP_TIME_BLK_1500-1559 | 0.81428736 | 0.47320139 | 0.08528619 | 2.25756645 |
| DEP_TIME_BLK_1600-1659 | 0.73656398 | 0.46096623 | 0.11007198 | 2.08874631 |
| DEP_TIME_BLK_1700-1759 | 0.80683631 | 0.42013136 | 0.05480258 | 2.24080753 |
| DEP_TIME_BLK_1800-1859 | 0.65816337 | 0.56922781 | 0.2475834 | 1.93124211 |
| DEP_TIME_BLK_1900-1959 | 1.40413988 | 0.47974923 | 0.00342446 | 4.07202291 |
| DEP_TIME_BLK_2000-2059 | 0.94785261 | 0.63308424 | 0.1343417 | 2.580163 |
| DEP_TIME_BLK_2100-2159 | 0.76115495 | 0.45146817 | 0.09180449 | 2.14074731 |
| DEST_EWR | -0.33785093 | 0.31752595 | 0.28732395 | 0.7133016 |
| DEST_JFK | -0.66931868 | 0.2657896 | 0.01179471 | 0.5120573 |
| CARRIER_CO | 1.81500936 | 0.53502011 | 0.0006928 | 6.14113379 |
| CARRIER_DH | 1.25616693 | 0.52265555 | 0.016242 | 3.51193428 |
| CARRIER_DL | 0.41380161 | 0.33544913 | 0.21736139 | 1.51255703 |
| CARRIER_MQ | 1.73093832 | 0.32989427 | 0.00000015 | 5.64594936 |

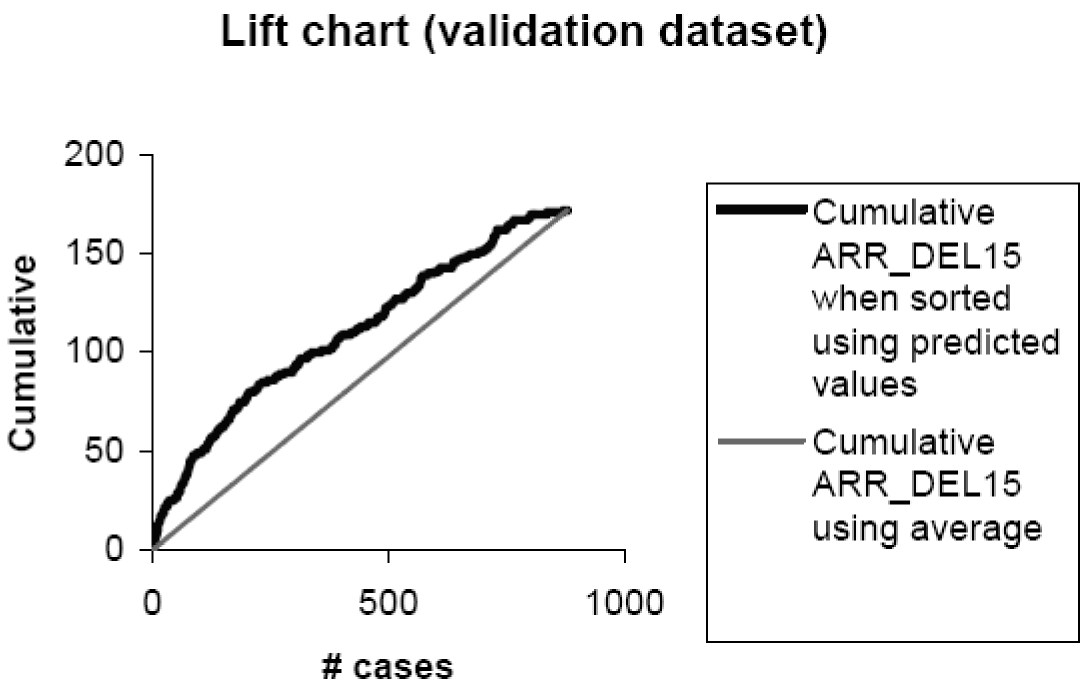
# Model Output (Validation Data)

| Cut off Prob.Val. for Success (Updatable) | **0.5** |
|---|---|

| **Classification Confusion Matrix** | | |
|---|---|---|
| | **Predicted Class** | |
| **Actual Class** | delayed | non-delayed |
| delayed | 18 | 154 |
| non-delayed | 3 | 705 |

| **Error Report** | | | |
|---|---|---|---|
| **Class** | **# Cases** | **# Errors** | **% Error** |
| delayed | 172 | 154 | 89.53 |
| non-delayed | 708 | 3 | 0.42 |
| **Overall** | 880 | 157 | 17.84 |

# Lift Chart



Lift chart (validation dataset)
200
150
100
50
0
Cumulative
0
500
1000
# cases
Cumulative ARR_DEL15 when sorted using predicted values
Cumulative ARR_DEL15 using average

# After Variable Selection (Model with 7 Predictors)

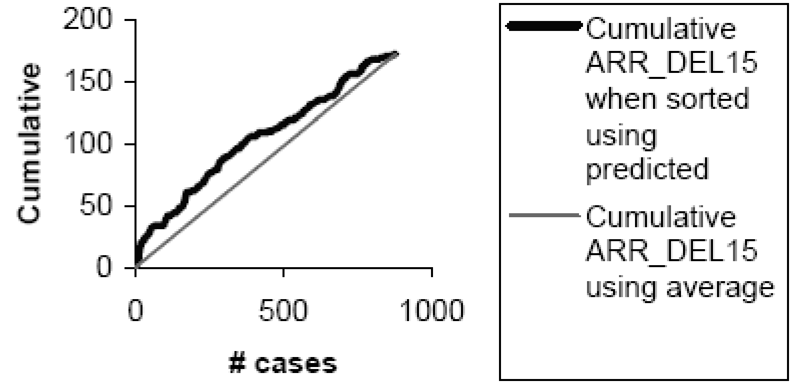
**Validation Data scoring - Summary Report**

| | |
|---|---|
| Cut off Prob.Val. for Success (Updatable) | **0.5** |

| Classification Confusion Matrix | | |
|---|---|---|
| | Predicted Class | |
| Actual Class | 1 | 0 |
| 1 | 13 | 159 |
| 0 | 0 | 708 |

| Error Report | | | |
|---|---|---|---|
| Class | # Cases | # Errors | % Error |
| 1 | 172 | 159 | 92.44 |
| 0 | 708 | 0 | 0.00 |
| Overall | 880 | 159 | 18.07 |

**Lift chart (validation dataset)**



# 7-Predictor Model

| Input variables | Coefficient | Std. Error | p-value | Odds |
|---|---|---|---|---|
| Constant term | -1.76942575 | 0.11373349 | 0 | * |
| Weather | 16.77862358 | 479.4146118 | 0.97208124 | 19358154 |
| DEP_TIME_BLK_0600-0659 | -0.62896502 | 0.36761174 | 0.08709048 | 0.53314334 |
| DEP_TIME_BLK_0900-0959 | -1.26741421 | 0.47863296 | 0.00809724 | 0.28155872 |
| DEP_TIME_BLK_1000-1059 | -1.37123489 | 0.52464402 | 0.00895813 | 0.25379336 |
| DEP_TIME_BLK_1300-1359 | -0.6303032 | 0.3188065 | 0.04803356 | 0.53243035 |
| Sun-Mon | 0.52237105 | 0.15871418 | 0.00099736 | 1.68602061 |
| Carrier_CO_OH_MQ_RU | 0.68775123 | 0.15049717 | 0.00000488 | 1.98923719 |

Note that Weather is unknown at time of prediction (requires weather forecast or dropping that predictor)

# Summary

- Logistic regression is similar to linear regression, except that it is used with a categorical response
- It can be used for explanatory tasks (=profiling) or predictive tasks (=classification)
- The predictors are related to the response Y via a nonlinear function called the *logit*
- As in linear regression, reducing predictors can be done via variable selection
- Logistic regression can be generalized to more than two classes (not in XLMiner)