SF2930: Regression Analysis
Lecture 13: Logistic Regression

Ekaterina Kruglov
Group Data Analyst

February 16, 2018

intrum

Logistic Regression

Why?

Linear Regression:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} + \varepsilon_i$$

Given a set of observations $\sum_{j=1}^{K}(\boldsymbol{x}_j, y_j)$, where $\boldsymbol{x}_j$ = $(x_{j1}, x_{j2}, \ldots, x_{jn})$ are n predictors and response value, y_i.
We build our model by estimating the coefficients, β_i.

In the case above, we were dealing with quantitative response variable.

Many situations, however, require us to predict a qualitative response.
Examples:

- Patient has or does not have the disease
- Credit Seeker will default or not on the loan
- Stock Market will go up or down

Idea: Instead of modeling the response directly, we would like to model the probability that a response function belongs to a certain category.

Logistic Regression

Why Linear Regression will not work

Outcome variable is given by

$$P(Y_i = k) = \begin{cases} p_i, & \text{for k=1} \\ 1 - p_i, & \text{for k=0} \end{cases}$$

where the observations are independent.
Thus, Y is a RV from Bernoulli distribution (special case of Binomial distribution where $n = 1$) with pmf

$$f(y) = p^y (1-p)^{1-y}$$

- Outcome variable is not continuous
- Cannot calculate probabilities with linear regression because we are bounded to $[0,1]$.
- Furthermore, in linear regression we assume

$$E(\varepsilon_i) = 0$$

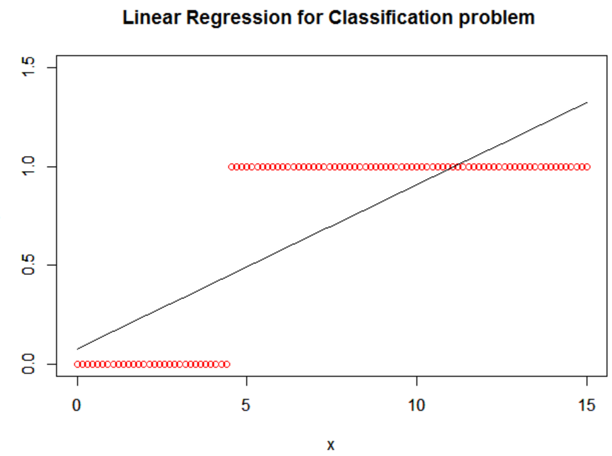
then

$$E(Y_i) = p_i \cdot 1 + (1 - p_i) \cdot 0 = p_i$$
$$E[Y_i - E(Y_i)]^2 = (1 - p_i)^2 p_i + (0 - p_i)^2 (1 - p_i) = p_i(1 - p_i)$$

Since y_i can be either 0 or 1, then

$$\varepsilon_i = \begin{cases} 1 - \boldsymbol{\beta}^T \boldsymbol{x}_i, & \text{for } y_i = 1 \\ -\boldsymbol{\beta}^T \boldsymbol{x}_i, & \text{for } y_i = 0 \end{cases}$$

- Errors are not normal
- Error variance is a function of the mean (p_i), hence not constant



Logistic Regression
Logit function

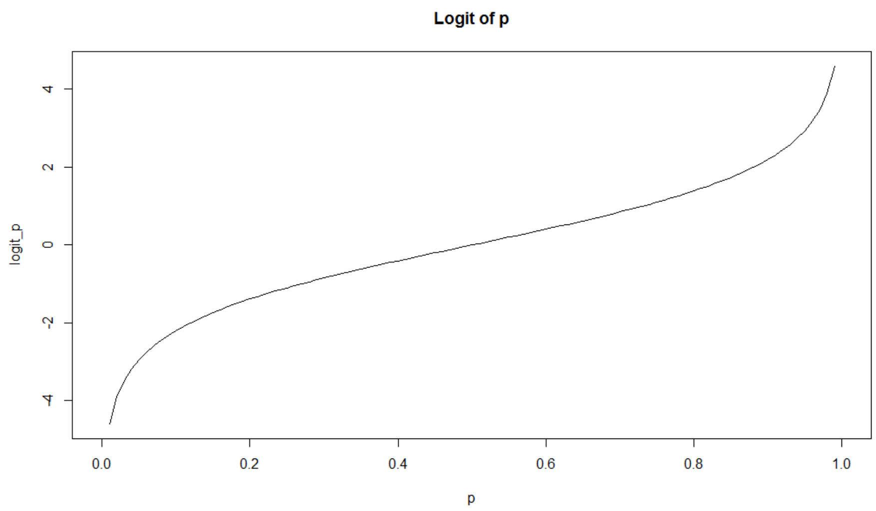
The **odds of an event** is the probability of observing the event divided by the probability not observing it, i.e. Consider event A, then

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

Let $p(A) = p$

Logarithmic odds of success (often referred to as ***logit of p***) is

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$



Logistic Regression

Logit function

Now, we can write our problem in a linear form using the logit of p

$$\begin{cases} logit(p) = \boldsymbol{\beta}^T \boldsymbol{x}_i \text{ where } logit(\text{p}) \in \{-\infty, \infty\} \\ \text{where } \boldsymbol{x}_j = (1, x_{i1}, x_{i2}, \ldots, x_{in}) \text{ and } \boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_n) \end{cases}$$

Let $\theta = logit(p)$, then

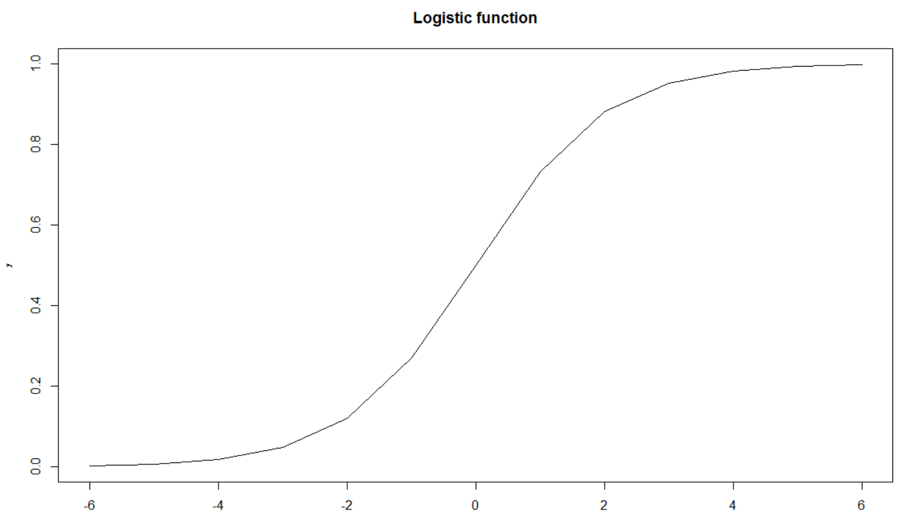
$$p = logit^{-1}(\theta) = \frac{e^\theta}{1 + e^\theta} = \frac{1}{1 + e^{-\theta}}$$

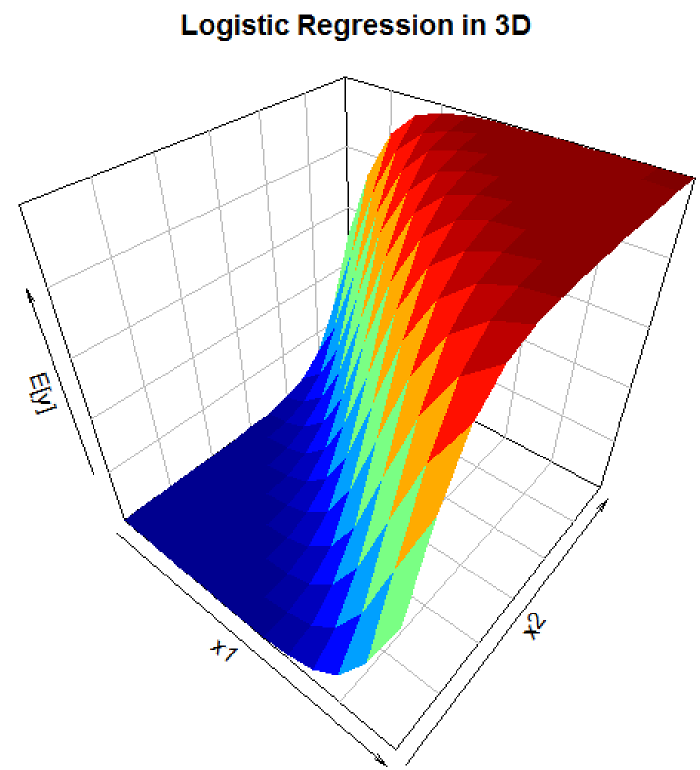
Logistic Regression

Definition

Logistic Response Function

$$E(\boldsymbol{Y}) = \boldsymbol{p} = \frac{e^{logit(p)}}{1 + e^{logit(p)}} = \frac{e^{\boldsymbol{\beta}^T\boldsymbol{x}}}{1 + e^{\boldsymbol{\beta}^T\boldsymbol{x}}} = \frac{1}{1 + e^{-\boldsymbol{\beta}^T\boldsymbol{x}}}$$





intrum

Logistic Regression

Variables

The predictor variables $x_1, x_2, \ldots, x_p$ can be binary, ordinal, categorical, or continuous.

Estimation of parameters: Maximum Likelihood

intrum

Logistic Regression

Estimation of parameters using Maximum Likelihood

Since $y \sim B(1,p)$, we get

$$
\begin{aligned}
f(y) &= p^y(1-p)^{1-y} \\
&= e^{\ln(p^y(1-p)^{1-y})} \\
&= e^{\mathrm{y}\ln(p)+(1-y)\ln(1-p)} \\
&= e^{\ln\left(\frac{p}{1-p}\right)y+\ln(1-p)} \\
&= e^{logit(p)y+\ln(1-p)}
\end{aligned}
$$

Assume that the observations are independent.

Using the training set we want to estimate $\boldsymbol{\beta}' = (\beta_0, \beta_1, \dots, \beta_n)$

Maximum Likelihood function

$$L(y_1, y_2, \dots, \boldsymbol{\beta}) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} {p_i}^{y_i}(1-p_i)^{1-y_i}$$

intrum

Logistic Regression

Estimation of parameters using Maximum Likelihood

Instead of working directly with function $L(y_1, y_2, \ldots, \boldsymbol{\beta})$, we work with $\ln L(y_1, y_2, \ldots, \boldsymbol{\beta})$ because:

- Natural log is an increasing function
- Often $\ln L(y_1, y_2, \ldots, \boldsymbol{\beta})$ has a much simpler form that is easier to differentiate

$$\begin{aligned}
\ln L(y_1, y_2, \ldots, \boldsymbol{\beta}) = \ln \textstyle\prod_{i=1}^{n} f_i(y_i) \\
&= \cdots \\
&= \sum_{i=1}^{n} y_i\, logit(p_i) + \sum_{i=1}^{n} \ln(1 - p_i) \\
&= \sum_{i=1}^{n} y_i\, \boldsymbol{\beta}^T \boldsymbol{x}_i - \sum_{i=1}^{n} \ln\left(1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}\right)
\end{aligned}$$

We want to solve the optimization problem

$$\text{mininimize } -\ln L(y_1, y_2, \ldots, \boldsymbol{\beta}) \text{ w.r.t } \beta$$

Note, the objective function is (see the board)

– Twice continuously differentiable
– (Strictly) Convex

➡ There exist standard optimization algorithms to solve this optimization problem

intrum

Logistic Regression

Estimation of parameters using Maximum Likelihood

Let $Y \in \{-1,1\}$

Then

$$\begin{cases} P(Y = 1|x) = \frac{1}{1+e^{-\boldsymbol{\beta}^T \boldsymbol{x}}} \\ P(Y = -1|x) = 1 - P(Y = 1|x) = \cdots = \dfrac{1}{e^{\boldsymbol{\beta}^T \boldsymbol{x}} + 1} \end{cases}$$

Hence, we can write:

$$P(Y|\boldsymbol{x};\boldsymbol{\beta}) = \frac{1}{1 + e^{-y\boldsymbol{\beta}^T \boldsymbol{x}}}$$

Then the likelihood function:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} P(y_i|\boldsymbol{x}_i;\boldsymbol{\beta})$$

Logistic Regression

Estimation of parameters using Maximum Likelihood

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} P(y_i|\boldsymbol{X}_i;\boldsymbol{\beta})$$

Since we prefer to work with natural log

$$\text{-ln } L(\boldsymbol{\beta}) = \textstyle\sum_{i=1}^{n} -\ln(P(y_i|X_i;\beta)$$

$$= \sum_{i}^{n} -\ln(\frac{1}{1+e^{-y_i\boldsymbol{\beta}^T\boldsymbol{X}_i}})$$

$$= \sum_{i=1}^{n} \ln(1+e^{-y_i\boldsymbol{\beta}^T\boldsymbol{X}_i})$$

intrum

Logistic Regression

Estimation of parameters using Maximum Likelihood

We want to minimize

$$\text{-}\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ln(1 + e^{-y_i \boldsymbol{\beta}^T \boldsymbol{x}_i})$$

We want to find $\beta = (\beta_0, \beta_1, \dots, \beta_m)$ such that

$$\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \beta_i} = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_i} \ln(1 + e^{-y_i \boldsymbol{\beta}^T \boldsymbol{x}_i}) = 0$$

$$\frac{\partial}{\partial \beta_0} \ln(1 + e^{-y_i \boldsymbol{\beta}^T \boldsymbol{x}_i}) = -y_i \left(\frac{e^{-y_i \boldsymbol{\beta}^T \boldsymbol{x}_i}}{1 + e^{-y_i \boldsymbol{\beta}^T \boldsymbol{x}_i}} \right) = -y_i (1 - \frac{1}{1 + e^{-y_i \boldsymbol{\beta}^T \boldsymbol{x}_i}})$$

$$\frac{\partial}{\partial \beta_k} \ln(1 + e^{-y_i \boldsymbol{\beta}^T \boldsymbol{x}_i}) = -y_i \boldsymbol{x}_i \left(\frac{e^{-y_i \boldsymbol{\beta}^T \boldsymbol{x}_i}}{1 + e^{-y_i \boldsymbol{\beta}^T \boldsymbol{x}_i}} \right) = -y_i \boldsymbol{x}_i (1 - \frac{1}{1 + e^{-y_i \boldsymbol{\beta}^T \boldsymbol{x}_i}})$$

Logistic Regression

Estimation of parameters using Maximum Likelihood

$$\frac{\partial}{\partial \beta_0} \ln(1 + e^{-y_i \boldsymbol{\beta}^T \boldsymbol{x}_i}) = -y_i \left(1 - \frac{1}{1 + e^{-y_i \boldsymbol{\beta}^T \boldsymbol{x}_i}}\right) = y_i(1 - P(Y_i | \boldsymbol{x}; \boldsymbol{\beta}))$$

$$\frac{\partial}{\partial \beta_k} \ln(1 + e^{-y_i \boldsymbol{\beta}^T \boldsymbol{x}_i}) = -y_i \boldsymbol{x}_i \left(1 - \frac{1}{1 + e^{-y_i \boldsymbol{\beta}^T \boldsymbol{x}_i}}\right) = y_i \boldsymbol{X}_i(1 - P(Y_i | \boldsymbol{x}; \boldsymbol{\beta}))$$

➡ No closed form of solution w.r.t $\beta_0, \beta_1, \dots, \beta_m$.

Logistic Regression

Newton Raphson Numerical Method for ML

Consider a function of one variable, $f(x)$. We want to find x^* that minimized $f(x^*)$.

We do not have an analytical solution for $\frac{df}{dx}$, then we approximate it using Taylor expansion.

Guess a point x_0, then Taylor expansion is

$$f(x) \approx f(x_0) + (x - x_0)\frac{df}{dx_0} + \frac{1}{2}(x - x_0)^2\frac{d^2f}{dx_0^2}$$

We want to want to solve for $\frac{df}{dx} = 0$. Hence, at some point x_1, we get

$$\frac{df}{dx_1} = \frac{df}{dx_0} + (x_1 - x_0)\frac{d^2f}{dx_0^2} = 0$$

Solving for x_1, (let $\frac{df}{dx} = f'(x)$)

$$x_1 = x_0 - \frac{f'(x_0)}{f''(x_0)}$$

Here we get a point x_1 that is closer to x^*.
Repeat for

$$x_n = x_{n-1} - \frac{f'(x_{n-1})}{f''(x_{n-1})}$$

until $|x_n - x_{n-1}| < \varepsilon$, where ε is suffieciently small.

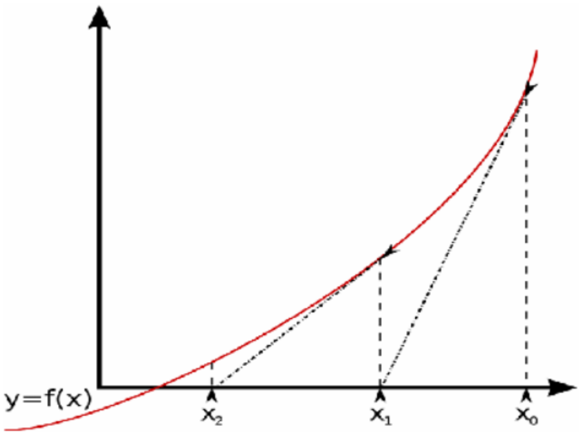


Figure: https://astarmathsandphysics.com/a-level-maths-notes/fp1/3441-the-newton-raphson-method-of-finding-roots-of-equations.html

intrum

Logistic Regression

Newton Raphson Numerical Method for ML

For $\boldsymbol{x} = (x_1, x_2, .., x_k)$

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - H^{-1}(\boldsymbol{x}_n)\nabla f(\boldsymbol{x}_n)$$

Where

$\nabla f(\boldsymbol{x}_n) = (\frac{\partial f}{\partial x_{n1}}, \frac{\partial f}{\partial x_{n2}}, \dots, \frac{\partial f}{\partial x_{nk}})$ (referred to as the gradient of f)

and

$H(\boldsymbol{x}_n) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_{n1}\partial x_{n1}} & \cdots & \frac{\partial^2 f}{\partial x_{nk}\partial x_{n1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_{n1}\partial x_{nk}} & \cdots & \frac{\partial^2 f}{\partial x_{nk}\partial x_{nk}} \end{pmatrix}$ (referred to as the Hessian of f)

Logistic Regression

Maximum Likelihood Vs. Least Squares Method

Least Squares Method

$$\min \sum_{i=1}^{n} (y_i - f(x_i, \beta))^2$$

Maximum Likelihood

$$L(y_1, y_2, \ldots, \boldsymbol{\beta}) = \prod_{i=1}^{n} f_i(y_i)$$

When dealing with binary logistic regression, we have information about the distribution of outcome variable (i.e. Bernoulli).

intrum

Logistic Regression

Interpretation of Parameters

Consider

- Single feature problem
- Numerically estimated parameters $\hat{\beta}$

Then, linear predictor (or $logit(p)$), defined as $\hat{\varphi}(x_i)$ is

$$\hat{\varphi}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

If we increase x_i by one unit

$$\hat{\varphi}(x_i + 1) = \hat{\beta}_0 + \hat{\beta}_1 (x_i + 1)$$

Then

$$\hat{\varphi}(x_i + 1) - \hat{\varphi}(x_i) = \ln\left(\frac{odds_{x_i+1}}{odds_{x_i}}\right) = \hat{\beta}_1$$

Hence, $\hat{\beta}_1$ is the estimated increase of logit function with one unit increase in x_i.

To find the estimated increase in probability, we take the antilog, i.e.

$$\hat{O}_R = \frac{odds_{x_i+1}}{odds_{x_i}} = e^{\hat{\beta}_1}$$

For d units:

$$\hat{O}_R = \frac{odds_{x_i+d}}{odds_{x_i}} = e^{d\hat{\beta}_1}$$

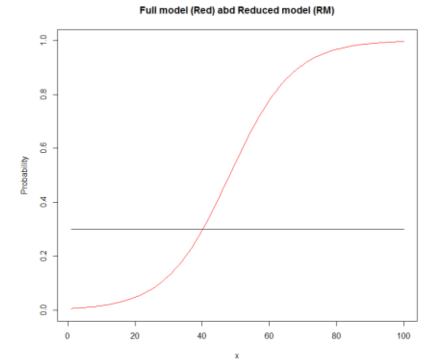
intrum

Model Assessment Methods

intrum

Logistic Regression

Likelihood Ratio Test



Compares “full” model (FM) with a “reduced” model (RM) of interest

$$LR = 2\ \ln\frac{L(FM)}{L(RM)}$$

Likelihood Ratio Test as a test for significance of regression in logistic regression:
FM: Model that we want to assess
RM: Model with constant probability of success ($p = \frac{y}{n}$)

$$\begin{aligned}\ln L(RM) &= \ln\prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} \\ &= \textstyle\sum_{i=1}^{n}(y_i \ln p + (1-y_i)\ln(1-p)) \\ &= y\ln\left(\frac{y}{n}\right) + (n-y)\ln(\frac{n-y}{n}) \\ &= y\ln(y) - y\ln(n) + (n-y)\ln(n-y) - (n-y)\ln(n) \\ &= y\ln(y) + (n-y)\ln(n-y) - n\ln(n)\end{aligned}$$

$$\ln L(FM) = \ln\prod_{i=1}^{n} {p_i}^{y_i}(1-p_i)^{1-y_i}$$

Likelihood Ratio Test

$$LR = 2\left\{\sum_{i=1}^{n} y_i \ln p_i + \sum_{i=1}^{n}(n_i - y_i)\ln(1-p_i) - [y\ln(y) + (n-y)\ln(n-y) - n\ln(n)]\right\}$$

Understanding the results:
Large values indicate that at least one of the variables in the logistic regression model is important

intrum

Logistic Regression

Goodness of Fit: Deviance

This test compares the full model to a saturated model.

FM: The model we have developed
SM: Model where each observation is allowed to have its own parameter (i.e. there are as many predictors as there are data points, which is basically overfitting).

Deviance:

$$D = 2\ \ln \frac{L(saturated\ model)}{L(FM)}$$

Understanding the results:
Small values, imply that the model fits well the data
Large values, imply that the model is inadequate

Logistic Regression

Goodness of Fit: Pearson chi-square

The test compares the observed and expected probabilities of success and failure at each group of observations.

- Expected number of successes: $n_i\hat{\pi}_i$
- Expected number of failures: $n_i(1-\hat{\pi}_i)$

The Pearson chi-square statistic is :

$$\chi^2 = \sum_{i=1}^{r} \frac{(obs.freq_i - exp.freq_i)^2}{exp.freq_i} = \sum_{i=1}^{r} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i}$$

Where

- n_i is the number of observations in the i^{th} group.
- $\hat{\pi}_i$ is the average estimated success probability in the i^{th} group.
- y_i is the number of observed successes in the i^{th} group.

Understanding the results:

Small values, imply that the model fits well the data

Large values, imply that the model is inadequate

Logistic Regression
Goodness of Fit: Hosmer-Lemeshow test

- No replicates on the regressor variables

Observations are classified into g groups based on estimated probability of success

$$\boxed{\bar{\pi}_j = \sum_{i \in group\ j} \frac{\hat{\pi}_i}{N}}$$

For each group j with N_j observations
- Observed number of successes O_j
- Observed number of failures $N_j - O_j$
- Expected number of successes $N_j \bar{\pi}_j$
- Expected number of failures $N_j(1 - \bar{\pi}_j)$

Hosmer – Lemeshow statistic

$$HL = \sum_{j=1}^{g} \frac{\left(O_j - N_j \bar{\pi}_j\right)^2}{N_j \bar{\pi}_j \,(1 - \bar{\pi}_j)}$$

Understanding the results:
Large value of HL imply that the model is not adequate fit to the data.

intrum

Logistic Regression
Model Assessment

		Observed	
		True	False
Predicted	True	True Positive (TP)	False Positive (FP)
	False	False Negative (FN)	True Negative (TN)

True Positive Rate:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})}$$

False Positive Rate:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{(\mathrm{TN}+\mathrm{FP})}$$

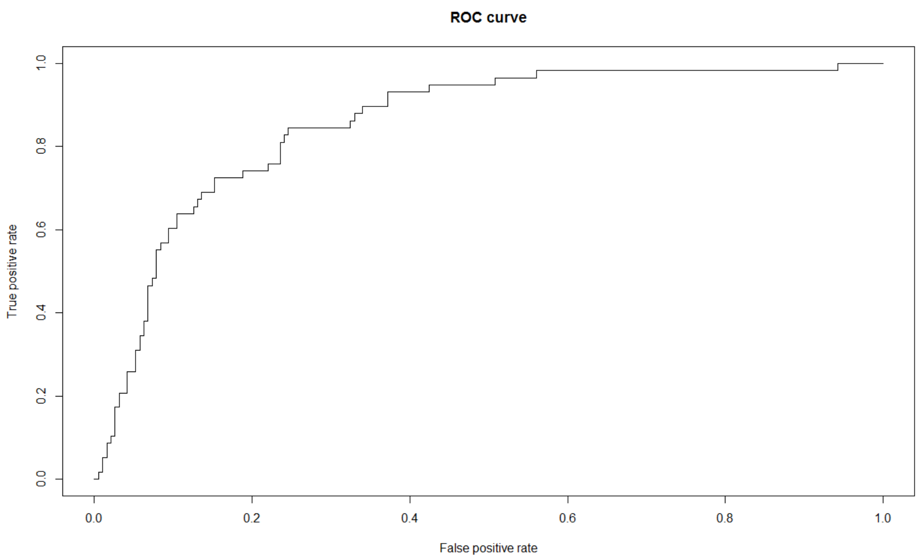
Accuracy:

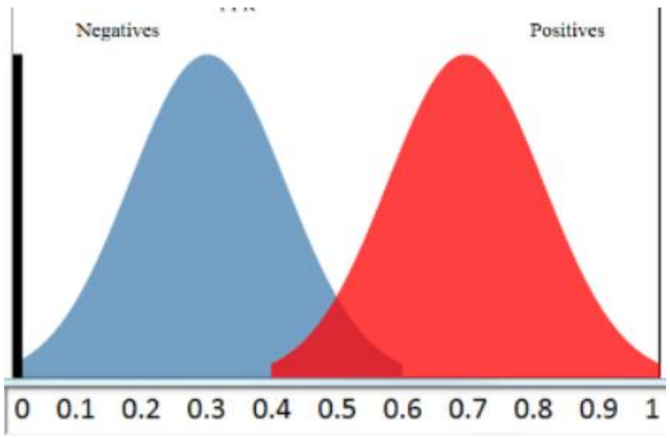
$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})}$$

Logistic Regression

Model Assessment

Reciever Operating Characteristic (ROC curve)
A plot for various thtresholds of false positive rate (FPR) as a function of true positive trate (TPR)





Area under the ROC curve
A measure of accuracy of how well our model separates the classes

- < 0,6 Fail
- 0,60-0,70 Weak separation
- 0,70-0,80 Fair separation
- 0,80 < Good separation

Logistic Regression

Nominal and Ordinal Logistic Regression

One can use logistic regression for classification of response variable into more than two classes.

Nominal example:

- Based on symptoms classifiy a patient in ER to one of the below categories
  - Stroke
  - Drug overdose
  - Epileptic seizure

Each class has a unique set of variables and corresponding coefficients.

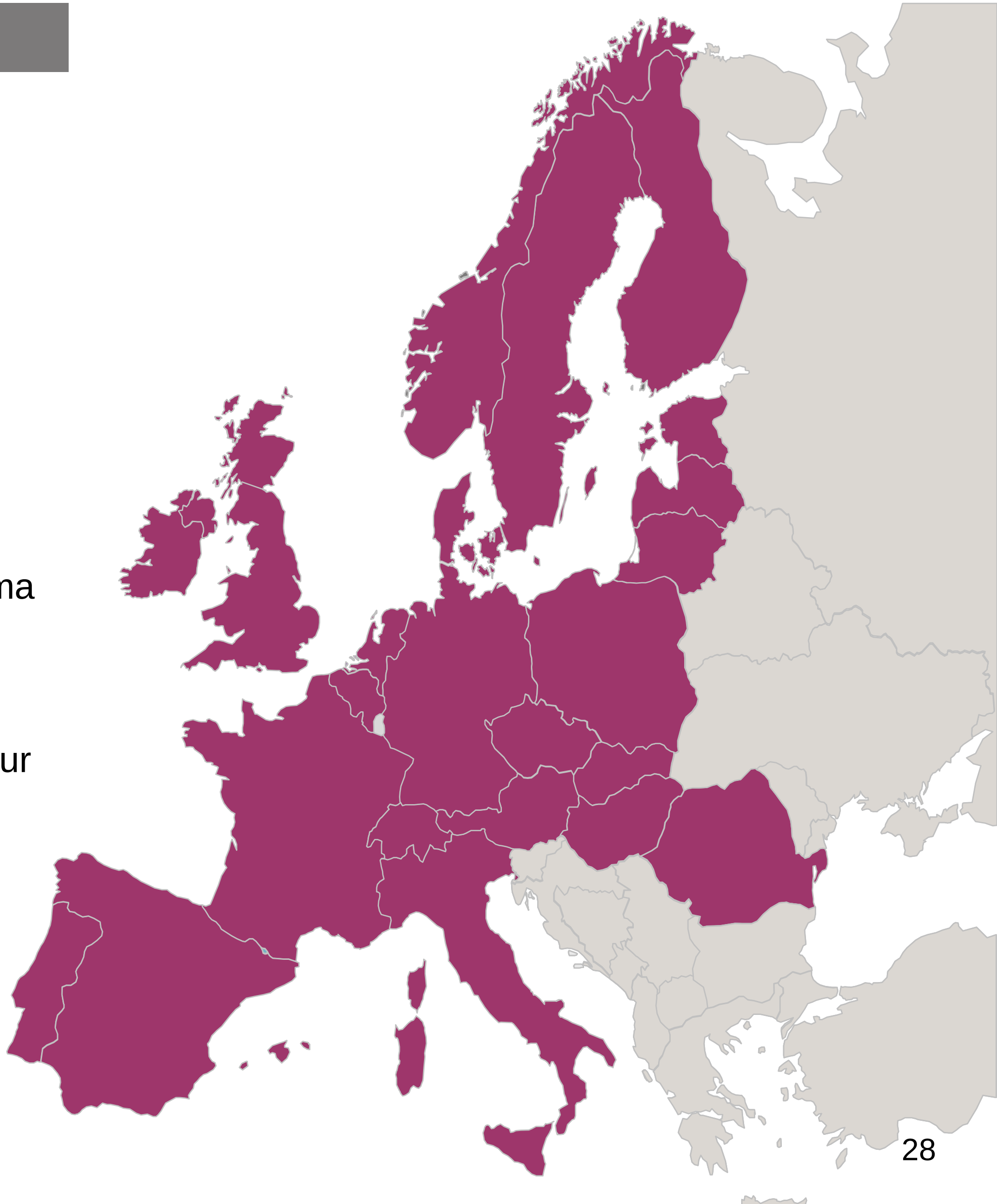
The Role of Logistic Regression in Strategy Development process at Intrum

intrum

Intrum: Who we are?

Key facts about us

- Industry-leading provider of Credit Management Services with presence in 24 markets in Europe
- Offering credit management- and financial services including; payment services, collection services and purchased debt
- We have more than 8,000 dedicated and empathic employees
- In the YTD ending September 2017, pro-forma income amounted to SEK 9.1 billion (EUR 0.94bn)
- Headquartered in Stockholm, Sweden and our share is listed on the Nasdaq Stockholm exchange

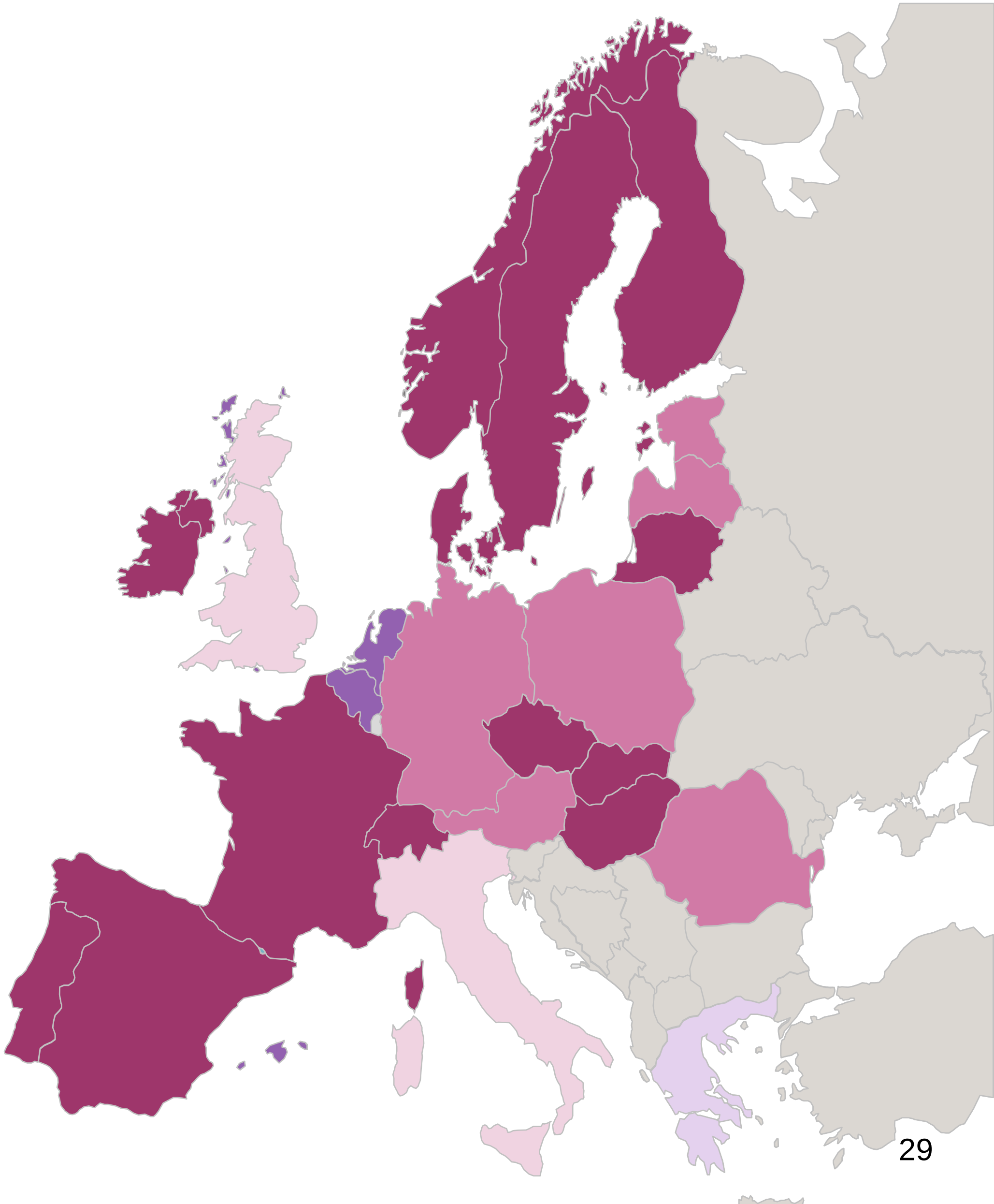


Strong position across entire footprint

- Market leader
- Top five
- Other

Market leader in most of the 24 European countries where Intrum is present

We have around 80,000 clients, most of them are found in sectors such as telecom, energy, banking and retail.



intrum


Leading the way
to
a sound
economy
intrum

The purpose explains the many wins that our business creates.

Individuals get rid of their debts and feel better.

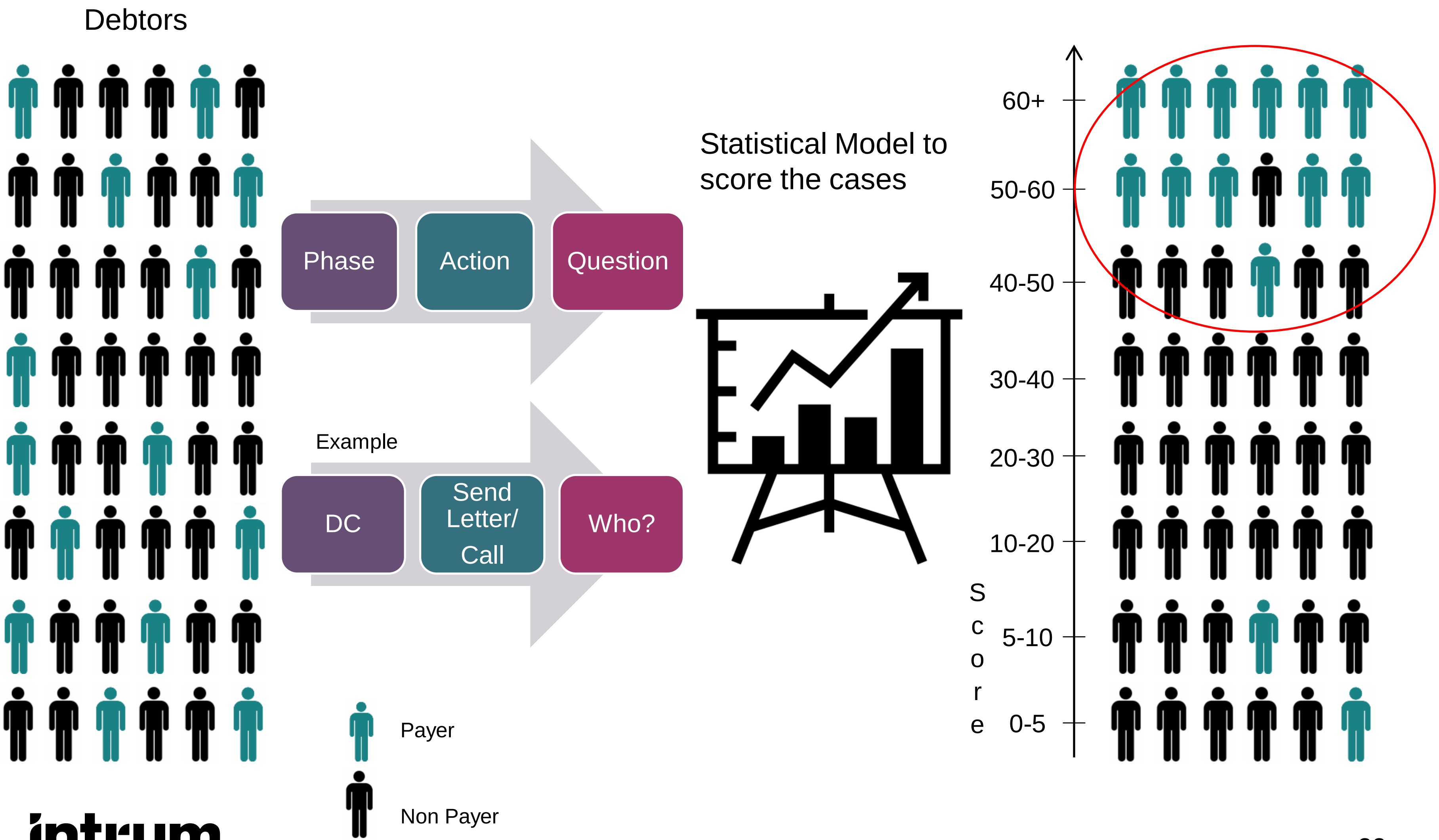
Companies can grow, invest, employ and flourish…

…which in turn is positive for the whole **economy.**

Employees get the chance to grow while doing good.

Logistic Regression: Application

Statistical Modelling for strategy development



Logistic Regression: Application

Data Selection and Data Analysis

Definition:
Depends on the Question

- Starting point
- Target
- Outcome Period

Data Selection
Requirements

- Full outcome period
- Industry/client specific?
- Enough payers and non payers

Data Check
Can past represent the future

- Data Quality (missing values, correct mapping, etc)
- Irregularities (crisis, seasonality, tax return, etc)
- Exclude irrelevant (clients we no longer work with)

Most important step! The model performance directly depends on the data.

Logistic Regression: Application

Example

Data set:
1000 observations, 8 variables

Variable	Variable Code
Target {Did not pay = 0, Paid = 1}	Target
Age of Debt {continuous}	TSD
Time to First payment {continuous}	Time_to_pay
Number of legal cases {continuous}	N_legal
Has contact information {binary}	Has_ci
Amount paid in the last 24 months {continuous}	Sum_pay_24m
Debt size {continuous}	Debt_size

Define new variable Response

```
data$Response=as.factor(data$Target)
```

intrum

Logistic Regression: Application

Example

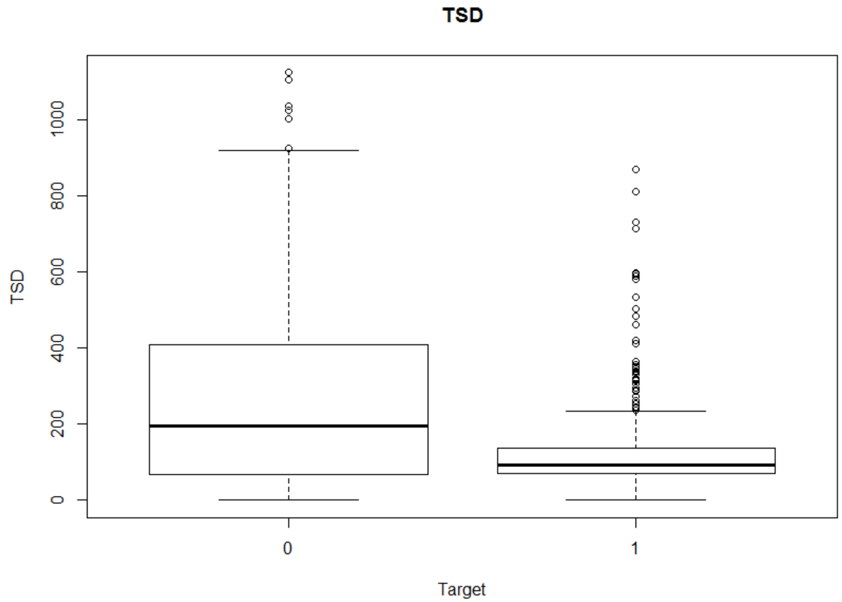
Original Data set

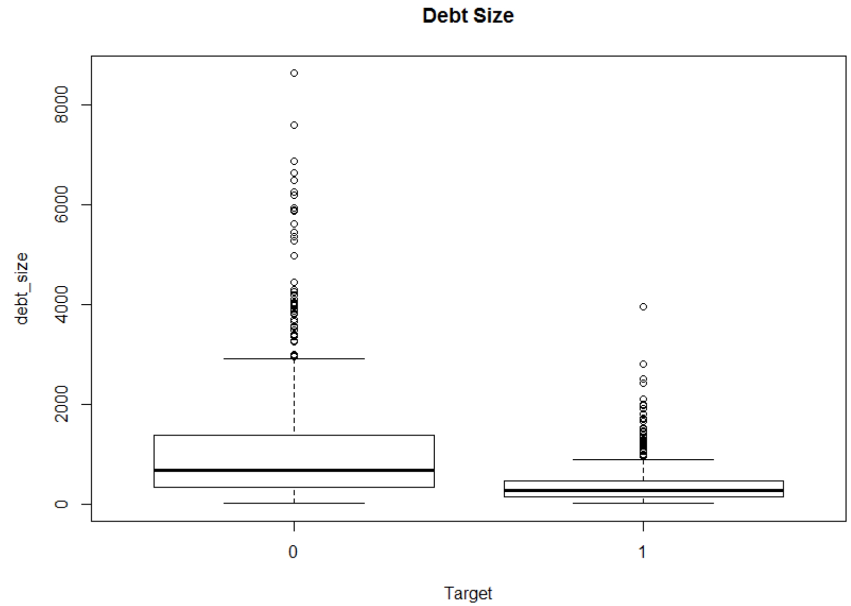
```
> summary(data_orig)
    case_id              Target            TSD            sum_pay_24m        time_to_pay       debt_size           N_legal          has_ci
 Min.   :1.058e+11   Min.   :0.000   Min.   :   0.0   Min.   :    0.0   Min.   :   0.0   Min.   :  18.94   Min.   :0.000   Min.   :0.000
 1st Qu.:1.751e+11   1st Qu.:0.000   1st Qu.:  69.0   1st Qu.:    0.0   1st Qu.:  18.0   1st Qu.: 224.34   1st Qu.:0.000   1st Qu.:1.000
 Median :7.004e+11   Median :0.000   Median : 113.5   Median :    1.1   Median :  47.0   Median : 432.17   Median :0.000   Median :1.000
 Mean   :6.094e+11   Mean   :0.404   Mean   : 207.2   Mean   :  396.9   Mean   : 225.4   Mean   : 824.28   Mean   :0.439   Mean   :0.948
 3rd Qu.:9.680e+11   3rd Qu.:1.000   3rd Qu.: 269.0   3rd Qu.:  342.9   3rd Qu.: 315.5   3rd Qu.: 975.99   3rd Qu.:1.000   3rd Qu.:1.000
 Max.   :9.957e+11   Max.   :1.000   Max.   :1123.0   Max.   :30919.0   Max.   :2054.0   Max.   :8630.29   Max.   :4.000   Max.   :1.000
```

Note:
With logistic regression, we do not need to normalize the data because if the variable has very large values, then the numerical method will make corresponding coefficient very small.

Logistic Regression: Application

Variables





intrum

Logistic Regression: Application
Training and Test Populations

In this example:
Randomly divided the population into training set (75% of all observations) and test set (25% of all observations)

Example of alternative/additional Approach: Cross Validation
Given n observations,

1. Select K (usually 5 or 10)
2. Randomly split the observations into K sets
3. Fit the model using K-1 sets and test on the remaining set. Perfrom K times and for each calculate MSE_i for $i = 1, \ldots, K$.
4. Calculate the average of MSE_i to obtain one estimate

Logistic Regression: Application

Models

```
> logit1_orig=glm(Response~.-Target-case_id, data = train_orig, family=binomial)
> summary(logit1_orig)

Call:
glm(formula = Response ~ . - Target - case_id, family = binomial,
    data = train_orig)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1643  -0.8796  -0.2284   0.8811   2.7757

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.8245173  0.3816693   2.160   0.0308 *
TSD         -0.0023079  0.0005789  -3.987 6.70e-05 ***
sum_pay_24m  0.0001340  0.0001061   1.263   0.2064
time_to_pay -0.0007566  0.0003724  -2.032   0.0422 *
debt_size   -0.0012168  0.0002077  -5.859 4.67e-09 ***
N_legal     -1.1436786  0.2276813  -5.023 5.08e-07 ***
has_ci       0.4194412  0.3746572   1.120   0.2629
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1011.90  on 749  degrees of freedom
Residual deviance:  767.84  on 743  degrees of freedom
AIC: 781.84

Number of Fisher Scoring iterations: 6
```

```
> logit2=glm(Response~.-Target-case_id-has_ci-sum_pay_24m, data = train_orig, family=binomial)
> summary(logit2)

Call:
glm(formula = Response ~ . - Target - case_id - has_ci - sum_pay_24m,
    family = binomial, data = train_orig)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6858  -0.8762  -0.2323   0.8902   2.7781

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.2519989  0.1508715   8.298  < 2e-16 ***
TSD         -0.0023820  0.0005808  -4.101 4.12e-05 ***
time_to_pay -0.0007543  0.0003733  -2.021   0.0433 *
debt_size   -0.0011901  0.0002064  -5.767 8.08e-09 ***
N_legal     -1.1063475  0.2261395  -4.892 9.97e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1011.90  on 749  degrees of freedom
Residual deviance:  770.69  on 745  degrees of freedom
AIC: 780.69

Number of Fisher Scoring iterations: 6
```

Notes:

Estimate = β

AIC: Akaike Information Criterion (type of model assessment) Lower AIC means better model.

Logistic Regression: Application

Calculating the scores

```
predict_model2 <- predict(logit2,train_orig)
probs_model2 = c(exp(predict_model2)/(1+exp(predict_model2)))
score <-ceiling(probs_model2*100)
```

Predict(model, data) – gives us the logit values i.e. $\boldsymbol{\beta}^T\boldsymbol{x}_i$
To calculate the score, we plug the values of predict into

$$\text{probability} = \frac{1}{1 + e^{-\boldsymbol{\beta}^T\boldsymbol{x}_i}}$$

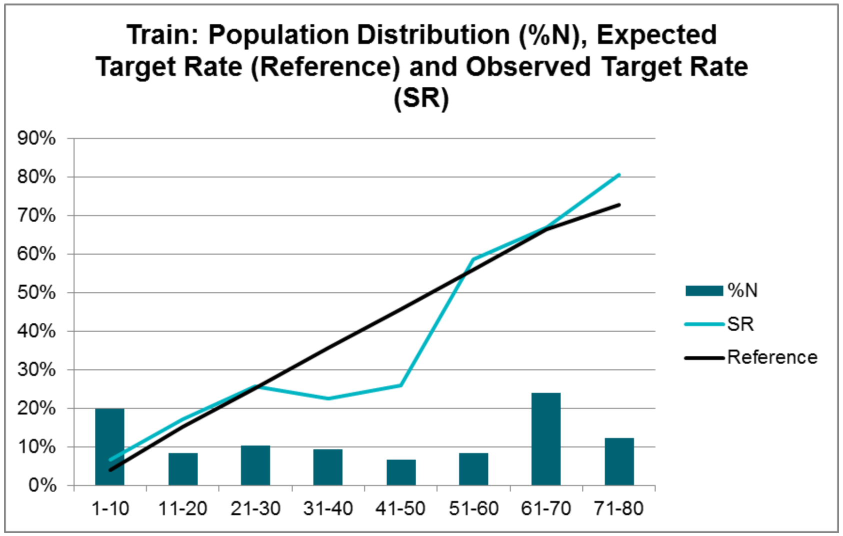
Then the score:

$$\text{score} = ceil(\text{probability} \cdot 100)$$

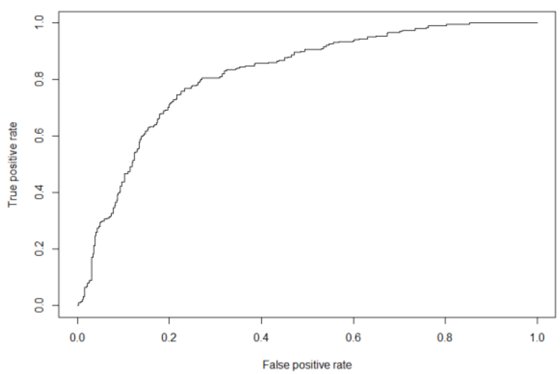
Logistic Regression: Application

Model 2

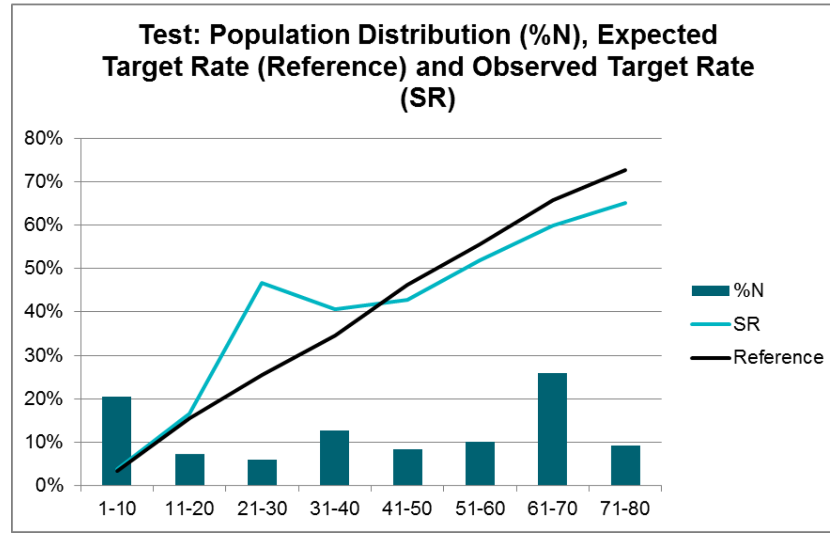
Train



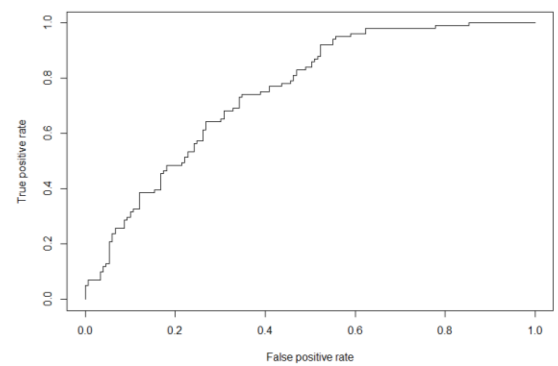
AUR: 0.817



Test



AUR : 0.75



intrum

Logistic Regression: Application

Model 3: Binning

Instead of working with variables directly, we create dummy binary variables based on the combination of optimal binnings (smbinning function) and business reasoning.

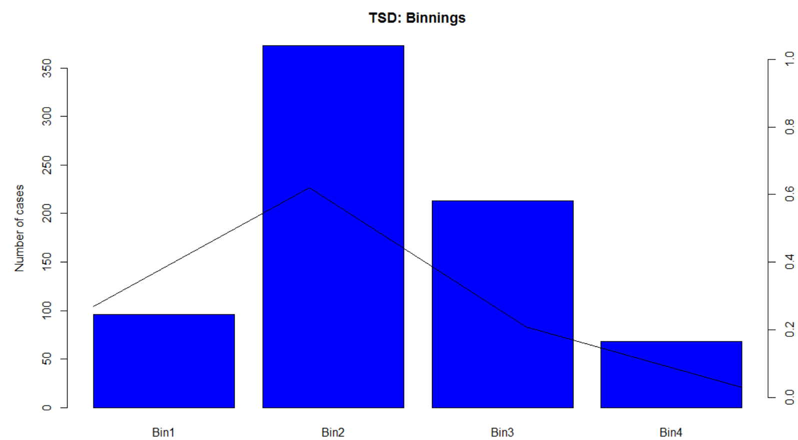
Example: Binning of TSD
result=smbinning(train_orig, "Target","TSD", p = 0.05)

Optimal Binnings:

	Cutpoint	CntRec	CntGood	CntBad	CntCumRec	CntCumGood	CntCumBad	PctRec	GoodRate	BadRate	Odds	LnOdds	WoE	IV
1	<= 0	96	26	70	96	26	70	0.1280	0.2708	0.7292	0.3714	-0.9904	-0.6016	0.0426
2	<= 184	373	231	142	469	257	212	0.4973	0.6193	0.3807	1.6268	0.4866	0.8754	0.3893
3	<= 594	213	44	169	682	301	381	0.2840	0.2066	0.7934	0.2604	-1.3457	-0.9569	0.2228
4	> 594	68	2	66	750	303	447	0.0907	0.0294	0.9706	0.0303	-3.4965	-3.1077	0.4383
5	Missing	0	0	0	750	303	447	0.0000	NaN	NaN	NaN	NaN	NaN	NaN
6	Total	750	303	447	NA	NA	NA	1.0000	0.4040	0.5960	0.6779	-0.3888	0.0000	1.0930

Business Decision

Bin 1 : (0, 184)
Bin 2 : [184,594)
Bin 3 : (594,∞]



intrum

Logistic Regression: Application

Model 3: Binning

Dummy Variables	Bin	
TSD1	1	If $TSD \in [0,184]$ then 1 else 0
TSD2	2	If $TSD \in (184,594]$ then 1 else 0
TSD3	3	If $TSD \in (594, \infty)$ then 1 else 0
ds1	1	If $debt\_size \in (0,311.9]$ then 1 else 0
ds2	2	If $debt\_size \in (311.9, 525.96]$ then 1 else 0
ds3	3	If $debt\_size \in (525.96,744.12]$ then 1 else 0
ds4	4	If $debt\_size \in (744.12,1998.53]$ then 1 else 0
ds5	5	If $debt\_size \in (1998.53, \infty)$ then 1 else 0
Time_to_pay1	1	If $time\_to\_pay \in [0, 27]$ then 1 else 0
Time_to_pay2	2	If $time\_to\_pay \in (27, 118]$ then 1 else 0
Time_to_pay3	3	If $time\_to\_pay \in (118, \infty)$ then 1 else 0
N_legal1	1	If $N\_legal = 0$ then 1 else 0
N_legal2	2	If $N\_legal > 0$ then 1 else 0
Sum_pay_24m1	1	If $sum\_pay\_24m \in [0, 150]$ then 1 else 0
Sum_pay_24m2	2	If $sum\_pay\_24m \in (150, \infty)$ then 1 else 0

Logistic Regression: Application

Model 3: Binning

```
> logit_bin2=glm(train_orig.Target~.-train_orig.case_id-hsi1-hsi2, data = train_bin, family=binomial)
> summary(logit_bin2)

Call:
glm(formula = train_orig.Target ~ . - train_orig.case_id - hsi1 -
    hsi2, family = binomial, data = train_bin)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9333  -0.7467  -0.2662   0.7373   2.8062

Coefficients: (5 not defined because of singularities)
            Estimate Std. Error z value       Pr(>|z|)
(Intercept)  -4.9898     0.7930  -6.292 0.000000000313 ***
lc1           1.1769     0.2859   4.116 0.000038485009 ***
lc2               NA         NA      NA             NA
sp1          -0.5376     0.1975  -2.722       0.006483 **
sp2               NA         NA      NA             NA
ds1           2.3489     0.5211   4.508 0.000006557526 ***
ds2           1.6098     0.5325   3.023       0.002500 **
ds3           1.2013     0.5677   2.116       0.034346 *
ds4           0.7127     0.5353   1.331       0.183043
ds5               NA         NA      NA             NA
TSD1          2.2454     0.6282   3.575       0.000351 ***
TSD2          1.4920     0.6431   2.320       0.020335 *
TSD3              NA         NA      NA             NA
ttp1          0.9198     0.3012   3.054       0.002261 **
ttp2          0.6453     0.3028   2.131       0.033115 *
ttp3              NA         NA      NA             NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1011.90  on 749  degrees of freedom
Residual deviance:  720.93  on 739  degrees of freedom
AIC: 742.93

Number of Fisher Scoring iterations: 6
```

These are reference bins with $\beta = 0$
Each other bin is compared by SR to the reference bin

Each binary variables gets its own β.
Example: TSD

	β	β_i	SR
Bin 1 : (0, 184)	-0,0023	2,24	55%
Bin 2 : [184,594)		1,49	21%
Bin 3 : (594,∞]		0	3%

Logistic Regression: Application

Difference in score

Target	TSD	sum_pay_24m	time_to_pay	debt_size	N_legal	has_ci
0	0	39.51	7	2742.71	3	1

Score when no binning:

```
Coefficients:
               Estimate
(Intercept)   1.2519989
TSD          -0.0023820
time_to_pay  -0.0007543
debt_size    -0.0011901
N_legal      -1.1063475
```

Score = ceil ($\frac{100}{1+e^{-(1.252-0.0024*0-0.0007*7-0.0012*2743-1.1063*3)}}$) =48

Score when binning:

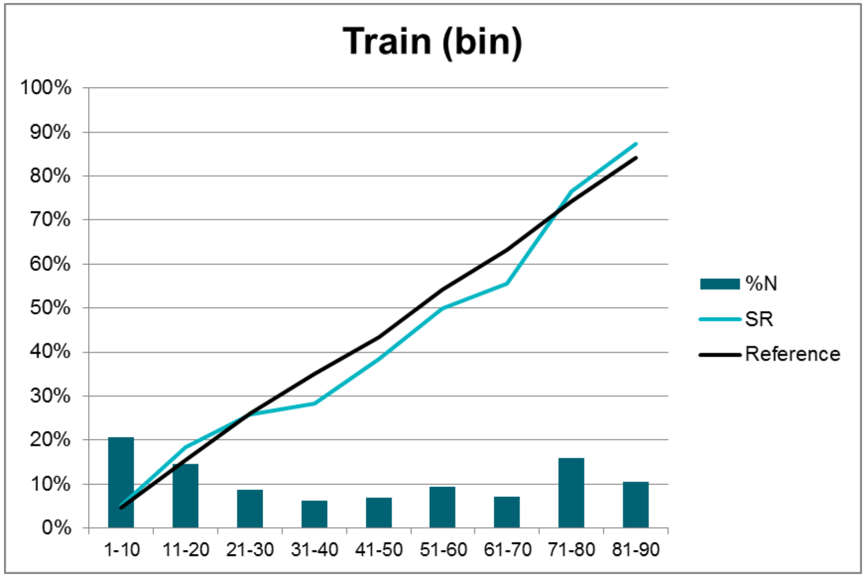
```
Coefficients: (5 not
              Estimate
(Intercept)    -4.9898
lc1             1.1769
lc2                 NA
sp1            -0.5376
sp2                 NA
ds1             2.3489
ds2             1.6098
ds3             1.2013
ds4             0.7127
ds5                 NA
TSD1            2.2454
TSD2            1.4920
TSD3                NA
ttp1            0.9198
ttp2            0.6453
ttp3                NA
```

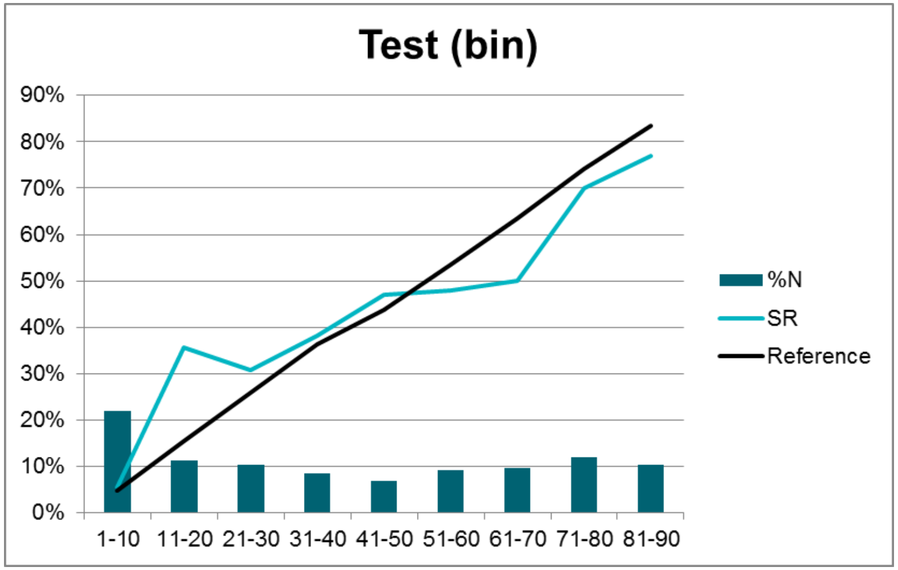
Score = ceil ($\frac{100}{1+e^{-(-4.9898+0-0.5376+2.2454+0.9198)}}$) =9

intrum

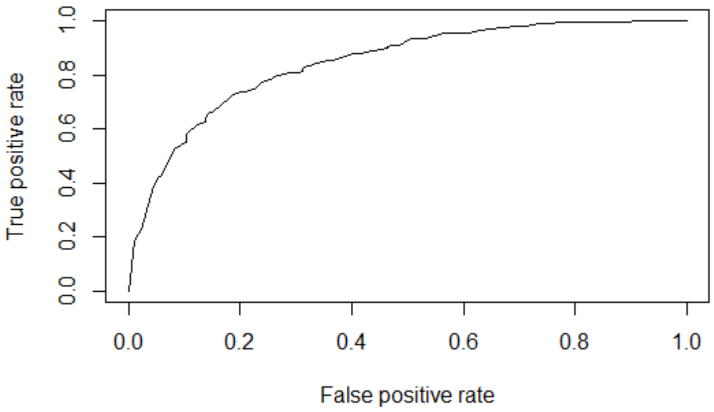
Logistic Regression: Application

Model 3: Binning

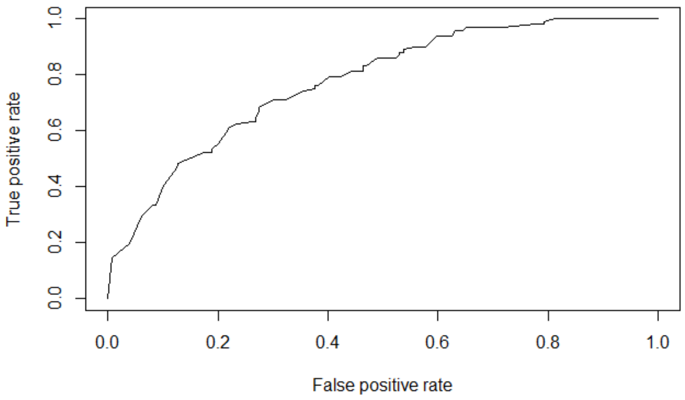




AUR: 0.844



AUR : 0.776



intrum

Logistic Regression: Application
Variable Selection

In this example, I manually selected the cases (not the best way)

In order to pick the best combination of variables one should consider different techniques. For example,

- **Backward**: start with all the available variables, and step by step take away the variable that is contributing the least., until there are no more variables. Each time measure the model performance. Select the model with best performance.
- **Forward:** Opposite to backward, start with no variables and at each step, add a variable that contributes the most. Repeat until all the variables are in the model.
- **Hybrid**: Combination of forward and backward. Start with no variables. At each step select the most contributing variable to the model, then check if there exist a variable that does not contribute, remove if exists.

Logistic Regression: Application
References

These lecture notes are based on
Timo Koski Lecture Notes from previous years
Introduction to Linear Regression Analysis by Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining

Newton Method:
http://www.stat.cmu.edu/~cshalizi/350/lectures/26/lecture-26.pdf

Good explanation of ROC and AUR:
http://www.dataschool.io/roc-curves-and-auc-explained/

Cross Validation and Vartiable Selection Methods overview:
Introduction to Statistical Learning with Applications in R (2013) Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani

Thank you!

Contact Information:
Ekaterina Kruglov: e.kruglov@se.intrum.com