defour analytics

Logistic Regression

Understanding Regression

Regression

- What is logistic regression?
- Types of logistic regression analysis
- Applications of logistic regression analysis
- Prerequisite / when & why binary logistic regression
- Generalized Linear Model (glm)

Logistic Regression

- Form of regression that allows the prediction of discrete variables by a mix of continuous and discrete predictors.

- Addresses the same questions that discriminant function analysis and multiple regression do but with no distributional assumptions on the predictors (the predictors do not have to be normally distributed, linearly related or have equal variance in each group)

Logistic Regression

- In logistic regression the dependent variable is binary, and the purpose of the analysis is to assess the effects of multiple independent variables, which can be numeric and/or categorical, on the dependent variable.
- If an independent variable is nominal level and not categorical, we need to dummy code the variable.
- The regression equation is : -

$$\text{Log}(p/(1-p)) = a + b1\ x1 + b2\ x2 + \ldots.. + bn\ xn + e$$

Types of logistic regression

defour analytics

- **BINARY LOGISTIC REGRESSION**

  - It is used when the dependent variable is dichotomous.

- **MULTINOMIAL LOGISTIC REGRESSION**

  - It is used when the dependent or outcomes variable has more than two categories.

Typical Applications of Regression Analysis

Building of models to ascertain the pattern/behaviour of certain performance measures

- Asset performance
  - Identification of bad actors
- Supply chain performance
  - Vendor Reliability based on parts supplied
  - Contractor performance
- Customer attrition
- Employee attrition rate
- Human reliability

Who uses it in Plain words.

- Binary Logistic Regression can be used in the following situations.
  - A catalog company wants to increase the proportion of mailings that result in sales.
  - A doctor wants to accurately diagnose a possibly cancerous tumor.
  - A loan officer wants to know whether the next customer is likely to default.
- Using the Binary Logistic Regression procedure, the catalog company can send mailings to the people who are most likely to respond, the doctor can determine whether the tumor is more likely to be benign or malignant, and the loan officer can assess the risk of extending credit to a particular customer.

Prerequisite for Logistic Regression

defour analytics

The Following need to be specified :

1. An outcome variable with two possible categorical outcomes. (1=success; 0=failure).
2. A way to estimate the probability P of the outcome variable.
3. A way of linking the outcome variable to the explanatory variables.
4. A way of estimating the coefficients of the regression equation, as well as their confidence intervals.
5. A way to test the goodness of fit of the regression model.

When and Why Binary Logistic Regression?

- When the dependent variable is non parametric and we don't have homoscedasticity. (variance of DV and IV not equal)
- Used when the dependent variable has only two levels. (Yes/no, male/female, taken/not taken)
- If multivariate normality is suspected.
- If we don’t have linearity.

Sample Size

- Very small samples have so much sampling errors.
- Very large sample size decreases the chances of errors.
- Logistic requires larger sample size than multiple regression.
- Hosmer and Lamshow recommended sample size greater than 400.

- **Sample size per category of the independent variable**
  - The recommended sample size for each group is at least 10 observations per estimated parameters.

Assumptions

- No assumptions about the distributions of the predictor variables.
- Predictors do not have to be normally distributed
- Does not have to be linearly related.
- Does not have to have equal variance within each group.
- There should be a **minimum of 20 cases per predictor**, with a **minimum of 60 total cases**. These requirements need to be satisfied prior to doing statistical analysis.

Generalized Linear Models (GLMs)

- The **generalized linear model** (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have other than a normal distribution.
- The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.
- It uses an iteratively reweighted least squares method for *maximum likelihood estimation* of the model parameters.
- The GLM consists of three elements:
  1. A probability distribution from the exponential family.
  2. A linear predictor $\eta = X\beta$.
  3. A link function g such that $E(Y) = \mu = g\text{-}1(\eta)$.

Performance of the model

- Unlike R^2 and adjusted R^2 in Linear Regression we have null deviance and AIC in Logistic Regression.
- **Null Deviance** : - It is a difference between actual and predicted value when only intercept is used for predicting the value.
- **Residual Deviance** : - When the entire equation is used to predict the value.
- **AIC (Akaike Information criteria)**

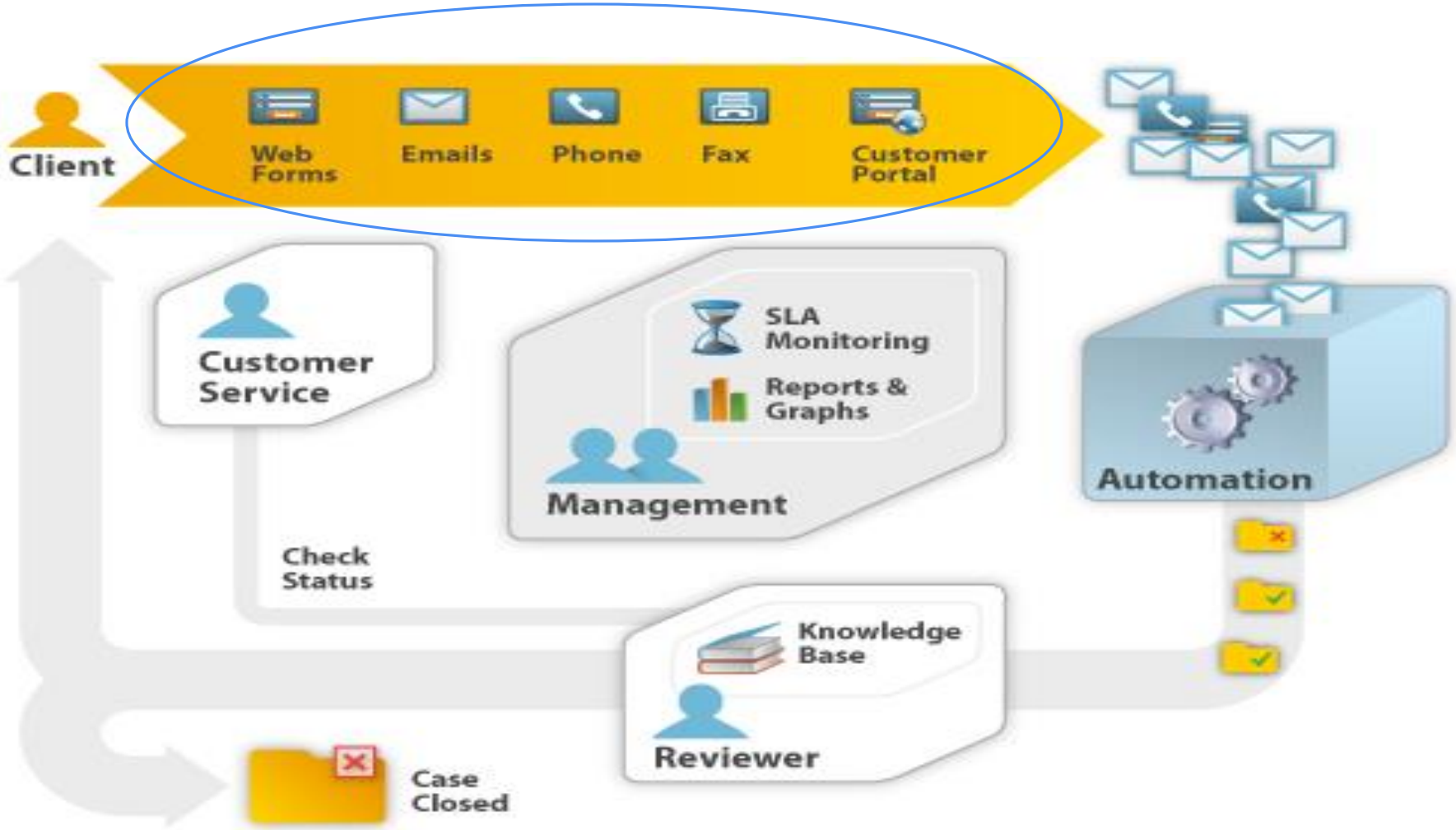
AIC = Residual deviance + 2 * no. of variables

- It is a measure which will help to decide which model to choose.
- So lower the AIC better is the model (provided it passes all t test and chi square test)

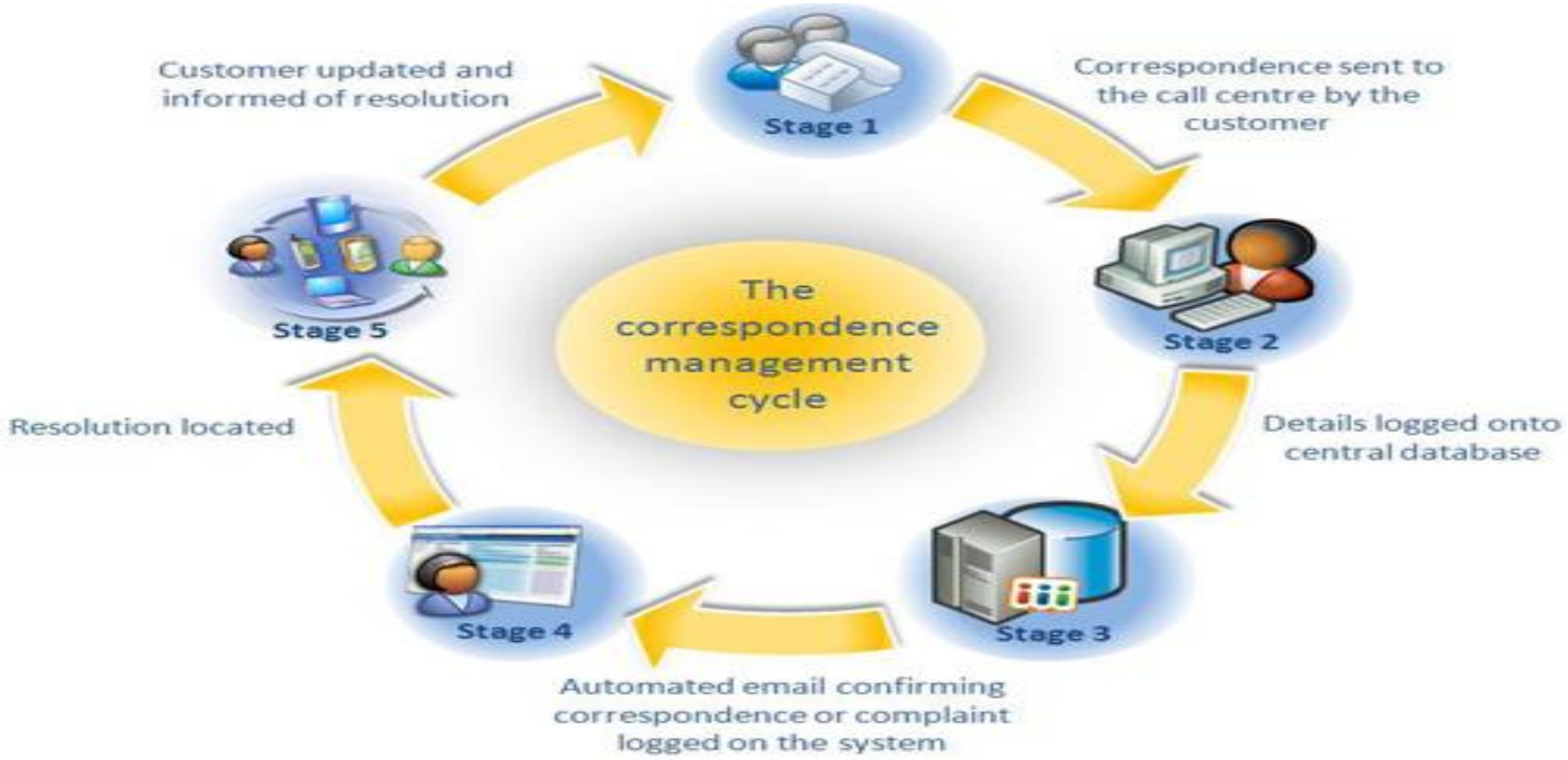
Customer Complaint Management

- It’s an online system where customer’s can log their complaints enabling them to see the resolution of their complaints.
- This system helps the customer satisfaction index.
- This system also measure the stipulated time to close the complaint.
- Through analysis of this system service provider can measure the cross functional department’s performance, SLA monitoring, identify hidden opportunities, spot the critical issues and optimize the process.
- Through which the binding factor for customer engagement is improved.

Types of complaint management system



Complaint management process



Problem Statement

- What are the chances that the complaint call time will get violated?
- Based upon the call closure time, is the performance good or bad?

i . e . If call is getting closed within 5 to 10 hours – performance is good.

If call takes more than 10 hours to close the complaint – performance is bad.

R script

```
setwd("C://R")

cm=read.csv("cm_lor.csv")

View(cm)

colnames(cm)

cm1=cm[,-c(1)]

View(cm1)

FitLogReg=glm(VIOLATED_DUMMY~.,family = binomial("logit"),data=cm1)

summary(FitLogReg)

FitLogReg=glm(VIOLATED_DUMMY~WTG_DUMMY11+WTG_DUMMY12+STATE_DUMMY2+STATE_DUMMY3+STATE_DUMMY6+STATE_DUMMY8+ITEM_DUMMY1+RT_DUMMY1+RT_DUMMY8+RT_DUMMY10+RT_DUMMY20+RT_DUMMY21+RT_DUMMY22+RT_DUMMY24+RT_DUMMY28+RT_DUMMY29+RT_DUMMY35+CR_DUMMY1+CR_DUMMY2+CR_DUMMY3,family = binomial("logit"),data=cm1)
```

R script

summary(FitLogReg)

FitLogReg=glm(VIOLATED_DUMMY~WTG_DUMMY11+WTG_DUMMY12+STATE_DUMMY2+ITEM_DUMMY1+RT_DUMMY1+RT_DUMMY8+RT_DUMMY10+RT_DUMMY20+RT_DUMMY21+RT_DUMMY22+RT_DUMMY28+RT_DUMMY29+CR_DUMMY1+CR_DUMMY2+CR_DUMMY3,family = binomial("logit"),data=cm1)

summary(FitLogReg)

library(car)

vif(FitLogReg)

cm$Predicted=predict(FitLogReg,type="response")

write.csv(cm,"Predicted_cm_lor.csv")

############

2 syntax line need to be added

Type of questions answered

1. Which are the area's where the chances of call getting violated more?

(Area can be state, item code, closure range and so on..)

2. From the probabilities calculated we can suggest to provide date's to the clients in such a way that the complaint gets resolved in time.

3. We can also suggest to increase the work force in the complaint area where chances of complaint not getting resolved in time is more.

Design Thinking

Ask the participants what kind of problems related to their industry can be solved using this concept.

Question & Answers

defour analytics

THANK YOU