# Opgave 2 Analytics 2019

Naam student: *graag hier invullen*

Data voor deze opgave: **telcotrain.csv** en **telcotest.csv**. Documentatie van de data: **telco.docx**

## Inleveren

Inleveren UITSLUITEND via Blackboard. Uiterste datum van inleveren: 10-3-2019.

Inleveren UITSLUITEND in formaat .PDF, .DOC of .DOCX.

Als je met RStudio werkt kan je de .Rmd file als basis gebruiken. Anders mag je de resultaten in dit document opnemen. Vergeet niet je naam in het document te zetten.

## Inleiding

De opgave gaat over een belangrijke vraag uit de Telecomindustrie: hoe houd ik mijn klanten?

In het databestand staan een aantal gegevens over klanten, zoals het aantal telefoontjes per dagdeel, of ze voicemail hebben, en nog zo het een en ander. De belangrijke variabele is de “Churn”, dat wil zeggen of de klanten bij deze provider zijn vertrokken in de periode waarin deze data is verzameld. (Churn=YES betekent de klant is vertrokken).

De provider wil graag kunnen voorspellen welke klanten waarschijnlijk op het punt staan te vertrekken, zodat hij deze kan benaderen met een goed aanbod.

De onderzoeksvraag is dus: hoe groot is de kans op Churn (vertrek) en welke variabelen spelen daarbij wel en niet een rol. Dit gaan we doen met de methode “logistic regression”.

De werkwijze is grotendeels hetzelfde als bij de videoreeks “Modeling the Expert”.

## Vraag 1

Laad de data in. De data bevat twee kolommen die je beter kan verwijderen (de eerste en de derde).

Maak daarna een Logistic Regression model voor het bepalen van de kans op Churn:

**1a)** Laad de data (**telcotrain.csv**), verwijder de variabelen in kolom 1 en kolom 3, en bouw dan een logistic regression model voor de kans dat de Churn variabele de waarde YES heeft (afhankelijke variabele: Churn), afhankelijk van alle andere variabelen. Druk de summary af van het model om te bepalen welke variabelen significant zijn in het model.

**Gevraagde antwoorden:**

Geef alle gebruikte R-commando’s. Geef ook de R-uitvoer.

**1b)** Verwijder **één voor één** de niet-significante variabelen uit het model, net zolang tot alleen significante variabelen over zijn. Noteer steeds wat de AIC-waarde is (als het goed is wordt deze steeds kleiner).

**Gevraagde antwoorden:**

Geef alle gebruikte R-commando’s. Geef ook de R-uitvoer. Geef de AIC-waarde voor elke tussenstap. Geef aan het eind ook de summary van het model.

**1c)** In het vervolg moet je een aantal malen de Confusion Matrix en kengetallen (accuracy, specificity, sensitivity) bepalen, voor verschillende drempelwaarden. Schrijf een r-script met daarin een functie waarmee je deze zaken automatisch (en correct) uitrekent, waarna je ze kan afdrukken. Als input geef je het model, de (volledige) naam van de afhankelijke variabele en de drempelwaarde. De functie berekent dan de juiste confusion matrix en de waarden voor accuracy, specificity, sensitivity. Zorg dat deze worden afgedrukt op de R-console.

*De spreadsheet “*testprogramma confusionmatrix.xlsx*” kan je gebruiken om te testen of je functie de juiste waarden oplevert.*

**Gevraagde antwoorden:**

Geef het r-script (de r-code).

**1d)** Bepaal de Confusion Matrix voor de gegevens volgens het uiteindelijke model. Daaraan kan je zien in hoeverre dat model de gegevens goed kan voorspellen Doe dit voor een drempelwaarde van 0.1, 0.2, 0.6 en 0.9. Bereken steeds de accuracy, de specificity en de sensitivity.

**Gevraagde antwoorden:**

Geef de vier confusion matrices met daarbij steeds de accuracy, specificity en de sensitivity.

**1e)** Welke drempelwaarde geeft de beste resultaten en waarom?

**Gevraagde antwoorden:**

Beantwoord de vraag.

## Vraag 2

Bepaal de ROC-curve voor dit model.

De ROC-curve geeft aan hoe de specificity en sensitivity afhangen van de gekozen drempelwaarden.

Laad eerst de ROCR-bibliotheek in.

**2a)** Maak een ROC-plot van het logistic regression model uit opgave 1. Gebruik colorize en eventueel print.cutoffs.at om het effect van verschillende drempelwaarden beter te kunnen zien.

**Gevraagde antwoorden:**

Geef alle R-commando’s en de plot.

**2b)** Wat is naar jouw mening de beste waarde voor de drempel-waarde (treshold) ? Houd rekening met de omstandigheden waarom een bepaalde threshold-waarde het beste zou kunnen zijn.

**Gevraagde antwoorden:**

Beantwoord de vraag.

## Vraag 3

Vergelijk het opgestelde model met een testset om te zien hoe goed het model nieuwe gegevens kan voorspellen.

We gaan nu het model uit opgave 1 vergelijken met een testset om te zien hoe goed het model nieuwe gegevens kan voorspellen. De testset heet **telcotest.csv** en bevat extra gegevens volgens dezelfde data structuur als **telcotrain**. Hint: gebruik de parameter *newdata* in de functie predict om de nieuwe resultaten te voorspellen.

Laad deze testdata in een apart dataframe.

**3a**) Gebruik de predict() functie en het model uit opgave 1 om de kans op Churn voor deze testdata te voorspellen.

**Gevraagde antwoorden:**

Geef alle R-commando’s en de R-uitvoer.

**3b)** Bepaal de confusion matrix voor de testset volgens dit model. Doe dit voor de bij 1d en 2a bepaalde optimale drempelwaarde. Bereken ook de accuracy, specificity en de sensitivity.

**Gevraagde antwoorden:**

Geef de confusion matrix en de berekende kengetallen.

**3c**) Bespreek de resultaten. Hoe goed (of slecht) is het bij 1 bepaalde model bij het voorspellen van de testdata?

**Gevraagde antwoorden:**

Beantwoord de vraag.