Master Scriptie

Afstudeerder:	Jade Dang
VU Begeleiders:	dr. Fetsje Bijma
	prof. Rob van der Mei
	dr. Wojtek Kowalczyk
Bedrijfsbegeleiders:	Maikel van der Hoorn
	Willibrord Ferwerda

Vrije Universiteit
Faculteit der Exacte Wetenschappen
Business Mathematics and Informatics
De Boelelaan 1081a
1081 HV Amsterdam

Lucka Consultancy BV
'WTC Alnovum'
P.J. Oudweg 3
1314 CH Almere

Voorwoord

Met trots presenteer ik hierbij het eindverslag van mijn afstudeerstage voor de masterstudie Business Mathematics and Informatics (BMI). Deze afstudeerstage heb ik gelopen bij Lucka Consultancy B.V. in de periode van 01-06-2008 tot 01-12-2008.

Als afstudeeropdracht heb ik een wiskundig model ontwikkeld dat gebruikt kan worden voor het selecteren van geschikte kandidaten bij de werving van personeel bij Lucka.

Tijdens deze afstudeerstage ben ik begeleid vanuit de Vrije Universiteit (VU) en Lucka Consultancy. Vanuit de VU ben ik begeleid door dr. Bijma, prof. Van der Mei en dr. Kowalczyk. Ik wil mijn begeleiders bij de VU graag bedanken voor alle adviezen en goede ideeën. Mijn dank gaat ook uit naar de heer Maikel van der Hoorn voor het beschikbaar stellen van de dataset en de heer Willibrord Ferweda voor zijn bijdrage aan functionele kant van het onderzoek.

Tot slot wil ik mijn familie en vrienden bedanken die mij tijdens deze stageperiode gesteund en met mij meegeleefd hebben. Hun behulpzaamheid heb ik zeer gewaardeerd. Speciale dank gaat uit naar Thoai Banh, in eerste plaats een zeer goede partner, maar daarnaast ook mijn SPSS vraagbaak.

Jade Dang
November 2008

Samenvatting

<< VERTROUWELIJK >>

Inhoudsopgave

1 Inleiding

In dit hoofdstuk worden verschillende onderwerpen beschreven. Het eerste onderdeel is het introduceren van het bedrijf Lucka Consultancy B.V. Daarna worden de probleemstelling, huidige situatie, onderzoeksvragen, en doelstelling van dit onderzoek beschreven.

1.1 Lucka Consultancy B.V.

Lucka is een ICT dienstverlener. Zij helpt klanten om bedrijfsproblematiek op te lossen. Lucka voert haar opdrachten uit bij organisaties als Philips, Graydon, VTS Politie, KLM, Ministerie van Defensie, Rabobank, ING, ANWB, TNT en Belastingdienst.
Op het gebied van Software Engineering, Enterprise Content Management (ECM) en Testen is Lucka actief bezig.

1.2 Probleemstelling

Dagelijks ontvangen de recruiters van Lucka nieuwe curriculum vitaes (CV's) van kandidaten van website zoals Monsterboard. Via email of per telefoon krijgen ze ook nieuwe aanmeldingen van kandidaten binnen. Op dit moment moeten ze dagelijks handmatig op zoek gaan naar de geschikte prospect kandidaten op de website Monsterboard. Deze kandidaten worden geselecteerd op basis van hun CV op Monsterboard. De geschikte prospect kandidaten hebben als kenmerk dat ze hun kennis en skills hebben doorgegeven via hun CV op de website en voldoen aan de voorwaarden voor minstens één van de drie typen vacatures van Lucka. Deze vacatures zijn de functies Tester, ECM en Java ontwikkelaar (Java).
De website van Monsterboard is niet alleen populair voor mensen die een baan zoeken, maar is ook populair bij bedrijven, uitzendbureaus en werving- en selectiebureaus. Lucka Consultancy B.V. is één van de bedrijven die de website van Monsterboard gebruikt om kandidaten te vinden op het gebied van ICT. Monsterboard.nl is al 10 jaar marktleider op het gebied van online recruitment. Met elke maand meer dan een miljoen bezoekers en ruim 850.000 CV's groeit Monsterboard.nl met de dag, met per dag meer dan 750 CV's (bron: http://hiring.monsterboard.nl/, 11-09-2008). Recruiters, uitzendbureaus, de werving- en selectiebureaus, werkgevers, etc… kunnen toegang krijgen tot de CV database om de perfecte kandidaat voor hun vacatures te vinden.
Met de huidige krapte op de arbeidsmarkt voor ICT is het zeer moeilijk om goede kandidaten te vinden. Vaak worden de mogelijke kandidaten al binnen de eerste week nadat ze hun CV's op Monsterboard hebben geplaatst uitgenodigd voor een gesprek door werkgevers. Deze eerste week is zeer belangrijk, omdat bedrijven door schaarste snel moeten inspringen op het aanbod van kandidaten en door snel met de kandidaten te onderhandelen probeert men op deze manier de juiste kandidaten te vinden voor de te vervullen functies. Men heeft niet genoeg tijd om alle beschikbare CV's van kandidaten door te nemen, daarom zijn er modellen nodig voor de drie typen vacatures. Modellen zullen onderzocht en opgezet moeten worden om met behulp van bepaalde technieken een effectieve en efficiënte manier te ontwikkelen die het selecteren van de juiste kandidaten voor een gegeven functie zoekt door het analyseren van kenmerken die opgegeven werden door de kandidaten in hun profiel.

Voor Lucka is het beantwoorden van de vraag “wanneer is een kandidaat een prospect kandidaat voor een vacature?” van belang. De prospect kandidaat is een kandidaat die bij Lucka de interesse wekt en uitgenodigd wordt voor een oriënterende sollicitatiegesprek. Met andere woorden: welke kenmerken van een kandidaat zijn van belang voor de selectie van de prospect kandidaat voor de functies Tester, Java ontwikkelaar en ECM binnen Lucka? Behalve de vraag wanneer de kandidaat een prospect kandidaat is, is het ook van belang voor Lucka of de kandidaat een geschikte prospect kandidaat is voor de functie Tester, Java ontwikkelaar of ECM. Bovendien zou men graag de mate van geschiktheid voor de verschillende functies willen kwantificeren. Wanneer deze vragen beantwoord worden dan is voor Lucka de beslissing om bepaalde kandidaat snel te interviewen eenvoudig te maken.

1.3 Huidige situatie

<< VERTROUWELIJK >>

1.4 Onderzoeksvragen

Wat zijn de belangrijkste kenmerken van een kandidaat voor de functie Tester, Java en ECM binnen Lucka die van invloed zijn voor het selecteren van de geschikte prospect kandidaten voor deze drie vacatures?
Welk wiskundig model kan gebruikt worden door ICT detacheringsbedrijf Lucka Consultancy B.V. om de geschikte prospect kandidaten te selecteren voor de functie Tester, Java en ECM?
Welke kandidaat van de prospect kandidaten is het meest geschikt? Deze vraag kan beantwoord worden door te kijken naar het geschiktheidspercentage per kandidaat. Anders geformuleerd “Welke formule kan gebruikt worden om de geschiktheid van een kandidaat voor een functie binnen Lucka te berekenen?” “Wat zijn de percentages van geschiktheid van deze geschikte prospect kandidaten voor de functie Tester, de functie Java en de functie ECM?” “Wie is het meest geschikt?”.

1.5 Doelstelling

<< VERTROUWELIJK >>

2 Data analyseren in SPSS

Het programma “Statistical Package for the Social Sciences” (SPSS) wordt met name gebruikt voor statistische doeleinden in de sociale wetenschappen. In dit hoofdstuk wordt SPSS gebruikt om data te analyseren en wordt een voorbeeld samenvatting gegeven van een CV van een kandidaat die op Monsterboard staat. Daarna wordt beschreven hoe de data opgeschoond en gecontroleerd worden en ook hoe deze geanalyseerd worden.

2.1 Formaat samenvatting van de CV op Monsterboard

Samenvatting		
Gewenst salaris/loon:	3,000.00 EUR Per maand	
Carrièreniveau:	Startfunctie (weinig ervaring)	
Totaal aantal jaren werkervaring:	2 tot 5 jaar	
Management ervaring:	Geen ervaring	
Project Management ervaring:	< 1 jaar	
Datum beschikbaar:	Binnen 2 weken	
Werkstatus:	Nederland - Ik ben geautoriseerd in dit land te werken voor elke werkgever.	
Geslacht:	Man	
Nationaliteit:	Nederland	
Geboortedatum:	14-09-1979	
Gewenst bedrijf:	**Branche:**	Automobielbranche - productie Telecommunicatie Computer hardware Computer software Computer/IT diensten
	Beroep	IT/Software Development Computer/Netwerkbeveiliging Software/Systeemarchitectuur Software/Web Development Systeemanalyse - IT Web/UI/UX design
Gewenst werkverband:	**Gewenste functietitel:**	TESTER
	Gewenst dienstverband:	Vast
	Gewenste status:	Full Time
	Bereid om de volgende diensten te werken:	Dagdienst Avonddienst
	Bereid om in het weekend te werken:	Ja
Gewenste locaties:	**Geselecteerde locaties:**	Nederland-GE Nederland-OV
	Verhuizen:	Nee
Talen:	**Talen**	**Vaardigheidsgraad**
	Arabisch	Goed
	Duits	Minimaal
	Engels	Voldoende
	Nederlands	Goed

Tabel 1: Samenvatting van CV op Monsterboard

2.2 Data van de kandidaat opschonen en controleren

<< VERTROUWELIJK >>

2.2.1 Opschonen

De acties ter opschoning van overbodige gegevens bestaan uit de volgende stappen:

(1) verwijderen van 'lege' records (cases): lege records zijn records die onvoldoende gegevens (waarnemingen) op sleutelkenmerken (indicatorkenmerken) bevatten. De sleutelkenmerken zijn:

   (a) nationaliteit
   (b) Nederlandse taalkennis
   (c) werkvergunning
   (d) gewenste functie

(2) controle op replica's van ingevoerde kandidaten: replica's van ingevoerde kandidaten zijn de kandidaten die twee keer of meer zijn opgenomen. Bij replica's van gegevens wordt de versie met de meeste gegevens bewaard en alle andere worden verwijderd.

2.2.2 Controleren

Na het opschonen van de dataset moet de dataset op juistheid gecontroleerd worden. Hieronder staan de vier meest voorkomende situaties benoemd:

1) De invulling van de kolommen vergunning en nationaliteit. Als iemand de Nederlandse nationaliteit heeft, dan wordt er aangenomen dat die ook een werkvergunning heeft. Als een kandidaat wel de Nederlandse nationaliteit heeft, maar geen werkvergunning dan moet de data nog een keer gecontroleerd worden. In de onderstaande tabel staan de mogelijke combinaties tussen de kenmerken Nederlandse nationaliteit en werkvergunning.

Nederlandse nationaliteit	Werkvergunning	Resultaat van combinatie	Reden
1	0	Fout	Als de kandidaat een Nederlandse nationaliteit heeft dan heeft deze automatisch een werkvergunning.
0	1	Goed	Kandidaten van binnen de EU-landen hebben de mogelijkheid om een werkvergunning voor Nederland aan te vragen.
1	1	Goed	Een Nederlandse nationaliteit en een werkvergunning is vanzelfsprekend.
0	0	Goed	Geen Nederlandse nationaliteit en geen werkvergunning. Deze kandidaten zijn mensen van buiten de EU-landen.

Tabel 2: Mogelijke combinaties tussen Nederlandse nationaliteit en werkvergunning

2) De drie mogelijke statussen van elke kandidaat zijn suspect, prospect en lead. Er wordt aangenomen dat elke kandidaat slechts een van deze drie mogelijke statussen heeft. Als een kandidaat meer dan twee statussen heeft, dan moet de status van deze kandidaat opnieuw gecontroleerd worden. Omdat dit niet juist is.

3) Als de kandidaten Nederlandse nationaliteit hebben, dan wordt er aangenomen dat ze ook Nederlandse taalkennis hebben. In tabel 3 wordt weergegeven wat de mogelijke combinaties zijn.

Nederlandse nationaliteit	Nederlandse taalkennis	Resultaat van combinatie	De reden
0	0	Goed	Geen Nederlandse nationaliteit en ook geen Nederlandse taalkennis.
0	1	Goed	Geen Nederlandse nationaliteit spreekt wel Nederlands. Bij deze kandidaat wordt er vervolgens naar zijn werkvergunning gekeken. Als hij/zij een werkvergunning heeft, is er wel de mogelijkheid dat deze kandidaat de geschikte prospect kandidaat zal worden. Anders niet.
1	0	Fout	Wel Nederlandse nationaliteit en spreekt geen Nederlands. Dit klopt niet.
1	1	Goed	Een Nederlandse nationaliteit en spreekt Nederlands is vanzelfsprekend.

Tabel 3: Mogelijke combinaties tussen Nederlandse nationaliteit en Nederlandse taalkennis

4) De “gewenste salaris” variabele is een categorisch kenmerk en bestaat uit gewenste salaris van minder dan 2000 euro per maand, gewenste salaris van meer dan 2000 en minder dan 2500, gewenste salaris van 2500 en minder dan 3000, gewenste salaris van 3000 en minder dan 3500, gewenste salaris van 3500 maar minder dan 9000, gewenste salaris van 9000 tot 14000. Elke kandidaat heeft één indicatie voor het gewenste salaris opgegeven. Als de kandidaat meer dan één indicatie heeft voor het gewenste salaris dan moet er onderzocht worden wat zijn/haar gewenste salaris was.

2.3 Analyse van data

2.3.1 Algemeen

- Populatie en steekproef

<< VERTROUWELIJK >>

- Variabelen

Van elke kandidaat worden bepaalde kenmerken geregistreerd. Deze kenmerken heten verklarende variabelen. In dit verslag worden de variabelen ‘werkervaring’, ‘opleiding’, ‘dienstverband’, ‘status’, ‘gewenst regio’, ‘taalkennis’ en ‘nationaliteit’ gebruikt.

- Meetschalen: Nominale schaal

Als de gegevens zijn ingedeeld in elkaar uitsluitende categorieën, zonder dat er sprake is van een rangorde, dan spreken we van een nominale schaal. De dichotome variabele is een speciale soort van nominale variabele. De dichotome variabele is een variabele die slechts twee waarden kan aannemen, bijvoorbeeld ‘geschikt’ en ‘niet geschikt’, of ‘0’ en ‘1’.

2.3.2 Beschrijving van de dataset

In de bijlage A zijn de variabelen van de beschikbare dataset opgenomen. Kolom 1 is de omschrijving van de variabelen. Kolom 2 is de omschrijving zoals ze gecodeerd worden in SPSS. Kolom 3 is de omschrijving met mogelijke waarden van elke variabele zoals ze gecodeerd worden in SPSS.

3 Samenhang tussen twee variabelen

Voor elk onderzoek is het gevaar dat er tussen de verklarende variabelen onderling een lineaire relatie bestaat. Als dat het geval is, spreken we van multicollineariteit. Dit betekent dat deze verklarende variabelen uit het model gehaald moeten worden. Anders kunnen deze verklarende variabelen in het model blijven. We willen collineariteit onderzoeken omdat een sterke samenhang tussen verklarende variabelen tot slechtere schattingen met een lagere betrouwbaarheid leidt van de regressiecoëfficiënten. Het opsporen van multicollineariteit is mogelijk door het opstellen van een correlatiematrix, waarin de correlatie tussen elk paar van onafhankelijke variabelen in het regressiemodel wordt weergegeven. In deze analyse is gekeken naar de samenhang tussen de variabelen met als doel om te kijken of bepaalde onafhankelijke variabelen weggelaten kunnen worden.
Welke statistische methode wordt gebruikt om de samenhang tussen twee categorische variabelen te berekenen? Mogelijke statistische methoden zijn Cramér's V-statistiek en "Pearson-chisquared Contingency Table". De theorie van deze methoden worden hieronder beschreven.

3.1 Pearson-chisquared Contingency Table

De χ^2*-toets* toetst op onafhankelijkheid. Met behulp van een χ^2*-toets* kan worden nagegaan of twee variabelen van nominaal of ordinaal meetniveau onafhankelijk van elkaar zijn. De nulhypothese (aangeduid met H_o) is dat de twee kenmerken onafhankelijk van elkaar zijn. Dat wil zeggen dat er geen associatie is tussen de categorieën van de rij- en kolom variabele. Deze bekende associatie-test wordt gebruikt bij N >= 50 frequenties. Deze test bepaalt uit het geobserveerde aantal O_{ij} van elke cel de verwachte frequentie E_{ij} en de bijpassende Chi-kwadraat-kritieke-waarde x_r. Een kruistabel van 2 kansvariabelen (kolom, rij) heeft (k-1)*(r-1) vrijheidsgraden.

Het recept voor 2 gerelateerde kansvariabelen is k x r

Gegeven is dat een klassenindeling *k* resp. *r* klassen heeft. Voer de volgende stappen uit:

1. Stel een *k* x *r* -tabel op met de *k* waargenomen frequenties O_{ij} in de steekproef.

   Bereken hieruit de regel- en kolomtotalen.

   Kolomtotaal = som van de aantallen in de cellen van een kolom van een kruistabel.
   Regeltotaal = som van de aantallen in de cellen van een rij uit een kruistabel.

Bereken op basis hiervan de verwachte frequentie:

$$E_{ij} = \frac{regeltotaal\_in\_rij\_i * regeltotaal\_in\_kolom\_j}{steekproefomvang\_N}$$

2. Bepaal de kritieke waarde x_r uit tabel Chikwadraat-verdeling (zie de bijlage I Chikwadraat-verdeling) met statistiek pakket op basis van de gekozen onbetrouwbaarheidsdrempel α, bij (k-1)(r-1) vrijheidsgraden.
3. Bereken de uitkomst van χ^2 over alle cellen van de kruistabel**.**

$$\chi^2(totaal) = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

4. Vergelijk de uitkomst van χ^2 met de kritieke waarde x_r. Is χ^2 groter dan x_r, dan wordt de hypothese verworpen. In dit geval is de conclusie dat de twee kenmerken afhankelijk zijn.

3.2 Cramér's V-statistiek

Hieronder is de tabel die ons laat zien welk statistische methode gebruikt kan worden om de samenhang tussen twee categorische variabelen te berekenen.

Variabele 1 / **Variabele 2**	Categorisch	Continue
Categorisch	**Cramérs V-statistiek**	**R-kwadraat**
Continue	**R-kwadraat**	**Pearson correlatiecoëfficiënt**

Tabel 4: Mogelijke statistische methoden om de samenhang tussen 2 variabelen te berekenen

De samenhang tussen twee categorische variabelen in een kruistabel wordt uitgedrukt door Cramér's V-statistiek. Voor de samenhang tussen twee continue variabele wordt gebruikt gemaakt van de Pearson correlatiecoëfficiënt en om de samenhang tussen een categorische en een continue variabele aan te geven, wordt de R-kwadraat gebruikt.
Omdat de onafhankelijke variabelen van de dataset van het onderzoek categorische zijn, kan de sterkte van de samenhang in de steekproef worden gemeten met Cramér's V-statistiek. Deze samenhangsmaat kan rechtstreeks door SPSS berekend worden.

Cramér's V Statistiek:

Cramér's V is een door de Zweedse wiskundige en statisticus Harold Cramér ontwikkelde associatiemaat voor twee categorische variabelen, dus variabelen die slechts op nominale schaal gemeten zijn.

Cramér's V wordt in het algemeen als de meest aantrekkelijke maat gezien voor grotere kruistabellen, omdat de bovengrens hiervan 1 is (bij volledige associatie).

Steekproef:

In een steekproef wordt Cramér's V gedefinieerd aan de hand van de kruistabel met r rijen en k kolommen en waargenomen frequenties O_{ij} in de steekproef.

Kolomtotaal = som van de aantallen in de cellen van een kolom van een kruistabel. Regeltotaal = som van de aantallen in de cellen van een rij uit een kruistabel.

Bereken op basis hiervan de verwachte frequentie:

$$E_{ij} = \frac{\text{regeltotaal\_in\_rij\_i} * \text{regeltotaal\_in\_kolom\_j}}{\text{steekproefomvang\_N}}$$

De χ^2-grootheid is:

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

De formule voor de samenhang V is:

$$V = \sqrt{\frac{\chi^2}{n * \min(r-1, k-1)}},$$

waarbij:
n: de totale steekproef grootte.
Min (r-1, k-1): betekent welke het minst is, de aantal kolomen of aantal rijen. Bijvoorbeeld: als we een 2 x 3 tabel hadden dan zouden we (2-1) kiezen.

In de onderstaande tabel zijn de waarden opgenomen die inzicht geeft in de onderlinge samenhang tussen twee variabelen.

V=0.00	geen samenhang
V=0.10	zwakke samenhang
V=0.25	redelijk sterke samenhang
V=0.50	sterke samenhang
V=0.75	zeer sterke samenhang
V=1.00	volledig samenhang

Tabel 5: Mogelijke uitkomsten van Cramér's V statistiek

4 Logistische regressie

In dit hoofdstuk wordt beschreven waarom logistische regressie gekozen wordt en de theorie hierover. Ook de onafhankelijke variabelen worden geselecteerd en vervolgens de binomiaal-verdeling beschreven. Daarna worden de onbekende parameters geschat en eventueel de kwaliteit van het model wordt bepaald.

4.1 Logistische regressiemodel versus lineaire regressie

Af te lezen uit de onderstaande tabel zien we dat er logistische regressie gebruikt wordt wanneer de afhankelijk variabele Y binair is. Als de afhankelijk variabele Y continu is, wordt de lineaire regressie gebruikt. Voor het onderzoek is gekozen voor logistische regressie. Aangezien de dataset meer dan één onafhankelijk variabele heeft en de afhankelijke variabele binaire is, is het mogelijke model dat gebruikt wordt voor het onderzoek de logistische regressie (zie onderstaande tabel).

Onafhankelijke variabele X	Afhankelijke variabeleY continu	Afhankelijk variabeleY binair
1 X	Lineaire regressie	Logistische regressie
> 1X	Lineaire regressie	Logistische regressie

Tabel 6: Mogelijke modellen

Waarom is de logistische regressie gekozen en niet bijvoorbeeld lineaire regressie?

- Het grootste probleem van lineaire regressie is dat de door lineaire regressie voorspelde afhankelijke variabele een waarde heeft die groter kan zijn dan 1 en kleiner dan 0. Het is daarom aan te raden om logistische regressie bij de analyse te gebruiken wanneer men te maken heeft met een dichotome afhankelijke variabele.

- Logistische regressie is verwant aan lineaire regressie. Een afhankelijke variabele wordt verklaard aan de hand van één of meerdere onafhankelijke variabelen. Bij lineaire regressie is het noodzakelijk dat de afhankelijke variabele minstens intervalgeschaald is. Bij logistische regressie valt deze beperking weg. Logistische regressie heeft tot doel een categorische variabele te verklaren, voorzien van twee groepen, aan de hand van interval-, ratiogeschaalde en/of categorische variabelen. Deze specifieke combinatie van meetniveaus van de afhankelijke en de onafhankelijke variabelen maakt logistische regressie de aangewezen techniek om te gebruiken.

- Het resultaat van een klassieke R-kwadraat (R^2)[1] zal bij lineaire regressie lager zijn en minder goed interpreteerbaar. De reden waarom de R^2 lager is juist omdat de afhankelijke variabele enkel de waarde 0 of 1 bevat.

4.2 Logistische regressiemodel

Door logistische regressie toe te passen kan aan de hand van een aantal onafhankelijke variabelen de kans voorspeld worden dat een kandidaat geschikt is voor een bepaalde functie. Logistische regressie wordt namelijk gebruikt om een binaire afhankelijke variabele (twee groepen, in dit geval geschikt of niet geschikt) te voorspellen. Binaire logistische regressie is een lineaire methode. Alle informatie omtrent de verklarende variabelen wordt samengevat in één enkele lineaire combinatie van deze variabelen. De ruimte van de verklarende variabelen wordt opgesplitst in twee half-ruimten, die elk corresponderen met een waarde van de te voorspellen binaire variabele.
De afhankelijke variabele dient bij logistische regressie een zogenaamde 0 – 1 variabele te zijn. In dit onderzoek is dat de keuze tussen geschikt (waarde = 1) en niet geschikt (waarde = 0).
De onafhankelijke variabelen bestaan uit n kenmerken. De logistische regressieanalyse berekent voor elk van de onafhankelijke variabelen een coëfficiënt die aangeeft of de desbetreffende variabele een positief dan wel een negatief effect heeft op de kans voor een geschikte of niet geschikte kandidaat voor een bepaalde functie.
Op elk kandidaat kan een verzameling van i onafhankelijke variabelen worden gemeten die kan worden voorgesteld als de vector $x = (x_1, x_2, \ldots, x_n)$.
Hier wordt er gebruikt gemaakt van een statistisch model dat de kans weergeeft dat iemand tot groep 1 behoort met gegeven de karakteristieken $x_1, x_2, x_3, \ldots, x_n$. Dit wordt gemodelleerd als:

$$P_{geschikt} = P(y = 1 | x_1, x_2, x_3, \ldots, x_n) = \frac{1}{(1 + \exp^{(-(b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots\ldots + b_nx_n))})} \qquad (1)$$

$$P_{niet\ geschikt} = P(y = 0 | x_1, x_2, x_3, \ldots, x_n) = 1 - P_{geschikt},$$

waarbij:

Y=1: aangeeft dat de kandidaat ‘geschikt’ is,
Y=0: aangeeft dat de kandidaat ‘niet geschikt’ is, en
$x_1, x_2, \ldots x_n$: de verklarende variabelen aangeeft.
Op basis van de trainingsampel worden de onbekende parameters $b_1, b_2, b_3, \ldots$ van dit model geschat. Voor het schatten van de parameters $b_0, b_1, b_2, b_3, \ldots$ wordt hieronder beschreven welke techniek hiervoor kan worden gebruikt.

[1] Een andere manier om aan te geven hoe goed het regressie-model is, is met R-kwadraat. R-kwadraat wordt bepaald door het kwadraat te nemen van de proportie verklarende variantie R. De verklarende variantie R wordt gegeven door:

$$R = \frac{\text{verklarendevariantie}}{\text{totaalvariantie}} = \frac{\Sigma(Y_{est} - \overline{Y})^2}{\Sigma(Y - \overline{Y})^2}$$

Y_{est} : voorspelde uitkomst
Y: waargenomen uitkomst
$\overline{Y}$: gemiddelde waargenomen uitkomsten

Een schatter voor $(b_1x_1 + b_2x_2 + b_3x_3 + ... + b_nx_n)$ moet berekend worden voor elke kandidaat afzonderlijk, waarna de kans op het voorkomen van dat geschikt-type berekend wordt aan de hand van vergelijking (1). Vervolgens worden de berekende kansen van vóórkomen voor verschillende kandidaten tegen elkaar afgewogen, en is de kandidaat met de hoogste kans van vóórkomen de voorspelde kandidaat.
In de bijlage B ziet u de dataset die de vijfenzeventig onafhankelijke variabelen $x_1, x_2, \ldots x_{75}$ en 1 afhankelijk variabele Y staan. Deze dataset wordt later in drie groepen opgesplitst namelijk in een dataset voor de functie Tester, een dataset voor de functie Java en een dataset voor de functie ECM. De reden voor het opsplitsen is dat voor elke kandidaat de mogelijkheid bestaat om geschikt te zijn voor de functie Tester, Java of ECM. Maar het kan ook zijn dat de kandidaat niet alleen geschikt zal zijn voor de functie Tester maar ook voor de functie Java. Bovendien ziet de mogelijke skills variabele voor elke functie geheel anders uit en qua aantallen varianten (mogelijke skills per functie is heel anders). Nog een andere belangrijke reden is dat het niet efficiënt en effectief is als de overbodige onafhankelijke variabelen in het opgezette model worden meegenomen. Bijvoorbeeld: De functie Tester heeft in totaal 18 skills. De skills zijn aangeduid voor de functie Tester met waar een T achter staat (zie bijlage C). De functie Java heeft in totaal 23 skills. De skills zijn aangeduid voor de functie Java met waar een J achter staat (zie bijlage C). De functie ECM heeft in totaal 15 skills. De skills zijn aangeduid voor de functie ECM met waar een S achter staat (zie bijlage C).

4.3 Binomiaal verdeling

In paragraaf 4.2 is de formule voor het berekenen van de geschiktheid van de kandidaat al bekend. Om het duidelijker te maken wordt de formule van de verdeling en van de link functie in deze paragraaf beschreven.

De keuze van de binomiale verdeling:
In dit onderzoek hebben we te maken met de binomiale verdeling. De aannames worden:

(i) $nY \sim Bin(n,\mu)$,
(ii) $\eta = x^T\beta$,
(iii) $\eta = g(\mu) = \log[\mu/(\mu-1)]$

In dit model is de relatie tussen component (i) en (iii) via de link functie g te vinden. Deze link functie moet differentieerbaar en monotoon zijn.

De tabel hieronder laat zien wat de natuurlijke parameter van de binomiale verdeling is.

Verdeling	Natuurlijke parameter	b	c
Binomiaal	$\log(\frac{p}{1-p})$	nlog(1-p)	$\log\begin{bmatrix} n \\ y \end{bmatrix}$

Tabel 7: Parameters van de Binomiale verdeling

Neem een stochastische grootheid Y waarvan de kansverdeling afhangt van een parameter θ. De verdeling kan worden geschreven als:

$$f_i(y) = f_i(y, \theta) = \exp\left(\frac{y\theta_i - b(\theta_i)}{\phi / A_i} + c(y, \phi / A_i) \right), \quad (2)$$

$$EY_i = \mu_i = b'(\theta_i), \; i = 1,\ldots,n$$

$$VarY_i = b''(\theta_i)\phi / A_i, \; i = 1,..,n \quad (3)$$

Het symbool A_i in (2) staat voor een vóóraf bekende constante. Dus wordt de vorm van f_i bepaald door de functies b en c. De b en c moeten zó gekozen zijn dat f_i een kansdichtheid is, en zodat vergelijking (3) van hierboven geldt. Een kansdichtheid heeft de volgende eigenschappen:

1) De verdeling is vastgelegd door verwachting en variantie.
2) De variantie van Y_i is een functie van zijn gemiddelde.

De waarden van b en c worden weergegeven in de onderstaande tabel.

Verdeling	$b(\theta_i)$	ϕ	A_i	$c(y_i, \phi / A_i)$
Binomiaal	$\log(1+e^{\theta_i})$	1	n_i	$\log\begin{pmatrix} n_i \\ n_i y_i \end{pmatrix}$

Tabel 8: Waarden van $b(\theta_i)$, ϕ, A_i, $c(y_i, \phi / A_i)$ van de binomiale verdeling

De keuze van de link functie:

	predictor			
respons		continu	categorisch	mixed
	continu	regressie	ANOVA	ANCOVA
	categorisch	logistische regressie	logit model	logistische regressie of logit model

Tabel 9: Mogelijkheden van de link functie

Bovenstaande tabel laat ons zien dat de Logit link functie gebruikt wordt als de "predictor" en de "respons" allebei categorische zijn. De Logit en Probit link functies worden vaak gebruikt in combinatie met een error-functie die binomiaal verdeeld is, omdat deze linkfuncties in het interval (0,1) op de reële as transformeren. Als de logit link functie wordt gebruikt in combinatie met een binomiale verdeling spreekt men ook wel van een logistische regressie model.
Logistische regressie werkt met kansverhoudingen in plaats van fracties. De kansverhouding, ook wel odds genoemd, is de verhouding tussen de fracties van de twee mogelijke uitkomsten. Als p de fractie bij de ene uitkomst is (in dit onderzoek is bijvoorbeeld de waarde "geschikt"), dan is 1 – p de fractie bij de andere uitkomst (in dit onderzoek is bijvoorbeeld de waarde "niet geschikt"). De ODDS is:

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

De kans p wordt bij logistische regressie uitgedrukt in termen van verklarende variabelen. De logaritme van de kansverhouding (p/(1-p)) wordt gemodelleerd als een lineaire functie van verklarende variabelen. Deze transformatie van de kansverhouding wordt de log odds of de logit genoemd:

$$Logit(p) = \log(\frac{p}{(1-p)})$$

Hieronder zijn de waarden van g(x) en $g^{-1}(x)$ die bij de Logit link functie horen.

g(x)	$g^{-1}(x)$
$\log(\frac{x}{1-x})$	$(\frac{e^x}{1+e^x})$

Tabel 10: De g(x) en $g^{-1}(x)$ van Logit link functie

4.4 Parameterschatting

Voor het schatten van de parameters gaan we de maximum likelihood methode gebruiken. Aangezien deze methode de meest betrouwbare resultaten oplevert in vergelijking tot de kleinste kwadratenmethode. Methode “Fisher scoring” is een bekende iteratieve methode om de schatting van de parameters te benaderen.

4.4.1 Maximum likelihood

De regressiecoëfficiënten β worden bepaald met behulp van de zogenaamde ‘maximum likelihood method’ (Wilks, 1995), het maximaliseren van de logaritme van de likelihood functie. De essentie van deze methode is dat de parameters gevonden worden die de grootste kans hebben dat de geobserveerde data voortbrengen. De likelihood is gedefinieerd als het product van de kansen dat de geobserveerde data is waargenomen. Gewoonlijk wordt de log van de likelihood gebruikt, omdat het makkelijker te werken is met een sommatie dan met een product.
Maximum likelihood schatting zoekt de parameters die de log likelihoodfunctie maximaliseren. De maximum likelihood schatter van β is $\hat{\beta} = (\hat{\beta}_0,....,\hat{\beta}_P)^T$. Het is de waarde die wordt verkregen door de log likelihoodfunctie van de exponentiële familie te maximaliseren naar β.

$$l(\theta) = l(\beta) = \sum_{i=1}^{n}\left(\frac{Y_i\theta_i - b(\theta_i)}{\phi / A_i} + c(Y_i, \phi / A_i)\right),$$

waarbij:
β: de vector is van constanten die via de linkfunctie wordt gelinkt aan de parameter θ.

4.4.2 Fisher scoring

$\hat{\beta}$ moet numeriek worden bepaald. Een bekende methode hiervoor is “Fisher scoring” die door Nelder en Wedderburn (1972) was beschreven. Deze methode begint met een startoplossing β^0 en update deze oplossing tot β^1 op de volgende manier:

$$\beta^1 = \beta^0 + \left\{E_{\beta^o}\left(-\frac{\partial^2 l}{\partial\beta\partial\beta^T}\right)\right\}^{-1}\frac{\partial l}{\partial\beta},$$

waarbij de beide afgeleiden worden geëvalueerd in het punt β^0 en de verwachting wordt berekend alsof β^0 de ‘ware parameter’ is. De startoplossing β^0 kan worden verkregen door alle parameterwaarden op nul te zetten of de resultaten van een vorig gefitte Gegeneraliseerde Lineaire Models (GLM) te gebruiken.

De β^0 wordt dan vervangen door β^1 en het updaten wordt weer herhaald. De methode stopt als $\beta^m - \beta^{m-1}$ klein genoeg is. Op dat moment wordt $\hat{\beta}$ gegeven door de laatste β^m.

De nauwkeurigheid van een schatting van een parameter bepalen

Om de nauwkeurigheid van een schatting van een parameter te bepalen, kan de asymptotische variantiematrix van de schatter $\hat{\beta}$ worden gebruikt. Deze wordt gegeven door de inverse van de Fisher informatiematrix:

$$\left\{ E_{\beta^o}\left(-\frac{\partial^2 l}{\partial \beta \partial \beta^T} \right) \right\}^{-1} = \phi (X^T W X)^{-1}$$

waarbij:
W: de n x n diagonaalmatrix is met als i-de element.

$$w_i^0 = A_i \{ g'(\mu_i^0)^2 b''(\theta_i^0) \}^{-1}$$

Als je deze variantie wilt gebruiken heb je wel schattingen nodig voor β, W en ϕ. Als schatter voor ϕ wordt de volgende formule gebruikt:

$$\hat{\phi} = \frac{\chi^2}{n - p - 1}$$

waarbij:
χ^2 : de gegeneraliseerde Pearson chikwadraat statistiek is (zie de subparagraaf 4.5.2 voor de uitgebreide uitleg).
Een andere manier om ϕ te schatten is door de totale deviantie schatter te gebruiken:

$$\hat{\phi} = \frac{D}{n - p}$$

waarbij:
D: de deviantie (zie de subparagraaf 4.5.1 voor de uitgebreide uitleg). Met behulp van de schattingen voor de variantie, kan een (asymptotische), $(1-\alpha/2)100\%$ betrouwbaarheidsinterval voor β_j worden gemaakt (zie de formule hieronder).

$$\left[\beta_j - t_{n-p-1;(1-\frac{\alpha}{2})} \sqrt{(\hat{\phi}((X^T \hat{W} X)^{-1})_{ij})}; \beta_j + t_{n-p-1;(1-\frac{\alpha}{2})} \sqrt{(\hat{\phi}((X^T \hat{W} X)^{-1})_{ij})} \right],$$

waarbij:
$\hat{\phi}$: is de schatter voor ϕ
n: het aantal waarnemingen
p: het aantal variabelen (exclusief constante)

Om te controleren welke variabelen wel of niet in het model opgenomen moeten worden, zal de corresponderende t-ratio gebruikt kunnen worden. De t-ratio is de geschatte waarde gedeeld door de geschatte standaarddeviatie, en vergelijkt deze ratio met plus of min het (1-$\frac{\alpha}{2}$) kwantiel van de t-verdeling met (n-p-1) vrijheidsgraden (zie tabel kritieke waarden van t-verdeling, bijlage I). Als de t-ratio niet significant is, is dit een teken dat de bijbehorende onafhankelijke variabele uit het model verwijderd moet worden.

4.5 Kwaliteit van het model

Voor het meten van de kwaliteit van een logistische regressiemodel kunnen de deviantie ("Deviance") of de "Pearson chisquared statistic" of de ROC curve worden gebruikt.

4.5.1 De deviantie

De likelihood ratio is een goede maatstaf om de goodness-of-fit van het model te bepalen.

$$\lambda = \frac{L(\hat{\beta}_{\max}; y)}{L(\hat{\beta}; y)},$$

waarbij:

$L(\hat{\beta}; y)$: de maximum likelihood waarde is van de likelihood functie van het onderzochte model.

In de praktijk wordt vaak de logaritme van de likelihood ratio gebruikt, omdat hiermee makkelijker te rekenen is. Dit is het verschil tussen de likelihoodfuncties:

$$\ln \lambda = l(\hat{\beta}_{\max}; y) - l(\hat{\beta}; y)$$

In de praktijk wordt vaker de statistiek $\ln \lambda$ gebruikt die vermenigvuldigd wordt met -2. De reden voor dit is dat -2 $\ln \lambda$ chikwadraat verdeeld is. Omdat de log likelihood negatief is, de -2 log likelihood is positief, en de hogere waarde van dit aangeeft dat het model dat onderzocht wordt een slechte beschrijving van de data geeft, als je het vergelijkt met het verzadigde model (verzadigde model is een algemener model dat het maximaal mogelijke aantal parameters heeft, met andere woorden een verzadigd model is een volledig model). De deviantie D wordt gegeven door:

$$D = 2\left[l\left(\hat{\beta}_{\max}; y\right) - l(\hat{\beta}; y)\right]$$

waarbij:

$\hat{\beta}^{\max}$: de parameter is vector van het verzadigde model

$\hat{\beta}$: de maximum likelihood schatter is van het te onderzoeken model

$l\left(\hat{\beta}_{\max}; y\right)$: de likelihoodfunctie is voor het verzadigde model geëvalueerd in het punt $\hat{\beta}^{\max}$

$l(\hat{\beta}; y)$: de maximum likelihood waarde is van de likelihood functie van het onderzochte model

4.5.2 Pearson chisquared statistic

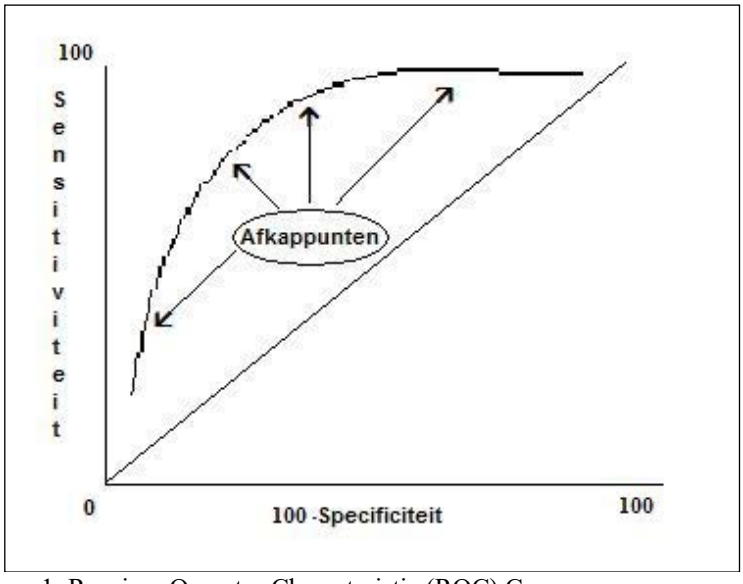
De Pearson chikwadraat statistiek wordt gegeven door:

$$\chi^2 = \phi \sum_{i=1}^{n} \frac{\{Y_i - E_\beta Y\}^2}{Var_\beta(Y_i)} = \sum_{i=1}^{n} \frac{\{Y_i - b'(\hat{\theta}_i)\}^2}{b''(\hat{\theta}_i)/A_i}$$

Het is handig om de deviantie en Pearson chikwadraat statistiek te gebruiken om twee geneste modellen met elkaar te vergelijken. Twee modellen zijn genest als het ene model kan worden verkregen door een parameter aan het andere model toe te voegen. Of, andersom, het andere model kan worden verkregen door een parameter weg te halen uit het “grotere” model.

4.5.3 ROC curve

Een van de technieken om de kwaliteit van het model te valideren is de Receiver Operator Characteristic (ROC) curve. ROC - statistiek is gebaseerd op de simpele gedachte dat wanneer we twee kandidaten nemen van wie er een geschikt en de ander niet geschikt is, een goed model voor de geschikte kandidaat een grotere kans om te schikken dient te voorspellen dan voor de niet geschikte kandidaat. De ROC - statistiek geeft voor een willekeurig paar van een geschikte en een niet geschikte de kans aan dat dat paar in de correcte volgorde door het model voorspeld wordt. Deze statistiek geeft een indruk van het discriminerend vermogen van het model tussen geschikte en niet geschikte. Deze ROC - statistiek wordt berekend als de proportie paren van een niet geschikte en een geschikte kandidaat waarin voor de geschikte kandidaat de voorspelde kans groter is dan voor de niet geschikte kandidaat.



Figuur 1: Receiver Operator Characteristic (ROC) Curve

De ROC curve bevindt zich in een vierkant: x-as van 0 – 100% en y-as van 0 – 100%. Bevindt men zich op de diagonale lijn dan heeft de test geen discriminerend vermogen, met andere woorden de diagonaal is de ROC curve behorende bij de meest waardeloze test (bij elk afkappunt is het percentage terecht-positieven gelijk is aan het percentage fout-positieven). Het afkappunt is de waarde die toelaat een onderscheid te maken tussen negatief (lager dan het afkappunt) en positief (hoger dan het afkappunt). Hoe meer de curve de links bovenste hoek benadert hoe beter het discriminerend vermogen tussen de geschikte en niet geschikte kandidaten. Een ROC curve is een grafiek waarbij de ratio van de echt positieven (100- specificiteit) vergeleken wordt met een reeks afkappunt waarden om een positief resultaat te definiëren. Met andere woorden: de ROC curve is een grafische weergave van de sensitiviteit (% terecht-positieven) op de y-as en 100-specificiteit (% fout-positieven) op de x-as, voor elk afkappunt in de range van testuitslagen.

Oppervlakte onder de ROC-curve (area under curve):
De oppervlakte onder die diagonaal is gelijk aan de helft van het vierkant en is gelijk aan 0.5. Hoe groter het ROC oppervlakte van een test, des te meer de curve ervan in de linker-bovenhoek ligt, des te beter de test is. De interpretatie van het ROC oppervlakte is min of meer het percentage kandidaten dat door de test correct geclassificeerd kan worden. Hosmer en Lemeshow (2000) geven als algemene regel dat een model niet discrimineert bij $\mathrm{ROC(area)} = 0.5$, acceptabel is bij $0.7 \leq \mathrm{ROC(area)} \leq 0.8$, excellent bij $0.8 < \mathrm{ROC(area)} \leq 0.9$ en 'outstanding' bij $\mathrm{ROC(area)} > 0.9$.

4.6 Selectie van de onafhankelijk variabelen

In deze paragraaf 4.6 wordt beschreven hoe de waarden voor de β's geschat zullen worden. Subparagraaf 4.6.1 gaat over de methoden die gebruikt zullen worden om de onafhankelijke variabelen te selecteren. In subparagraaf 4.6.2 wordt het algemene stappenplan voor het selecteren uiteengezet.

4.6.1 De mogelijke methoden

Er is voor logistische regressieanalyse ook een aantal methoden (Enter, Forward en Backward) beschikbaar. De methode Enter is een volledige regressieprocedure die in een gezamenlijke opname van alle variabelen voorziet. De meest gebruikte methode van deze is de achterwaartse (Backward) selectieprocedure. Hieronder wordt de definitie van elke methode en eventueel de vergelijking tussen de voorwaartse (Forward) en de achterwaartse selectieprocedure in het kort weergegeven.

	Voorwaartse selectieprocedure	Achterwaartse selectieprocedure
vergelijking	De voorwaartse selectieprocedure start met een model zonder predictoren. De onafhankelijke variabelen worden één voor één aan het model toegevoegd, waarbij telkens gekeken wordt of het model hierdoor verbeterd. Indien dit zo is, dan wordt de variabele opgenomen en wordt gezocht naar de volgende significante variabele. Indien dit niet zo is, dan wordt de opgenomen variabele verwijderd en wordt gezocht naar de volgende significante variabele. Kortom: tijdens het bepalen van de variabelen wordt bovendien gekeken of het nieuwe model beter is (in vergelijking met het model dat tot op dat moment bepaald was). Indien bij het toevoegen van een nieuwe variabele, het verwijderen van een reeds opgenomen variabele leidt tot betere resultaten dan wordt wanneer dit zo is, de nieuwe variabele toegevoegd en wordt de variabele, die al opgenomen was, verwijderd.	De achterwaartse regressieprocedure vertrekt dus van de situatie waarin alle variabelen worden opgenomen en verwijdert systematisch de variabelen met de laagste voorspelkracht. Zo wordt elke variabele geëvalueerd door de verklaringskracht van het gereduceerde model (na verwijdering van de variabele) telkens opnieuw te vergelijken met de verklaringskracht van het volledige model (inclusief de beschouwde variabele).
zwakke punten	Algemeen wordt aanvaard dat deze methode niet geschikt is om significante variabelen op te sporen. De reden hiervoor is dat de variabelen die op deze manier gevonden worden in veel gevallen enkel significant zijn in de dataset waarop gewerkt wordt.	
log likelihood	Deze waarde is afhankelijk van de steekproefgrootte en het aantal parameters in het model. Daarom wordt de vergelijking van de log likelihood gemaakt tussen het initiële/voorgaande model en het model met de parameter die overwogen wordt om opgenomen te worden. In het eerste model met een constante en een parameter, dan wordt vergeleken met een model waarbij enkel de constante werd berekend. De verbetering van het nieuwe model waarmee we de afhankelijke variabele kunnen schatten op basis van de geselecteerde variabelen, wordt weergegeven door de log likelihood statistiek. De log likelihood statistiek geeft de reductie weer in de fout bij het voorspellen van de afhankelijke variabele in het model waarbij de nieuwe onafhankelijke variabele is toegevoegd aan het model dat tot die stap gebruikt werd. De significant geeft aan hoe groot de kans is dat de verbetering in het model die gevonden werd in het uiteindelijke model per toeval ontdekt werd. Een waarde kleiner dan 0.05 is een algemeen aanvaarde norm. Een waarde 0.000 (dus kleiner dan 0.001) is dan ook zeker ruim voldoende.	Deze procedure schat het model door iteratief elke variabele te verwijderen en de verandering in de log-likelihood te evalueren. Onder de nulhypothese zijn de regressiecoëfficiënten van de verwijderde variabelen nul. De likelihood-ratio test meet dan de verhouding tussen de likelihood van het gereduceerde model ten opzichte van de likelihood van het volledige model. Bij een significantniveau ($p < 0.05$) wordt de variabele in het model opgenomen.

Tabel 11: Vergelijking tussen de voorwaartse en de achterwaartse selectieprocedure

4.6.2 Stappenplan (Backward)

In dit onderzoek wordt het onderstaande stappenplan gebruikt voor de drie typen functies (Tester, Java ontwikkelaar en ECM.). Per functie afzonderlijk is de uitwerking van het gekozen model in de bijlage van respectievelijk D, E en F en telkens in onderdeel (4) opgenomen.

Stap 1: Kies de selectiemethode Backward Wald bij het binaire logistische regressie model in SPSS. Alleen de geselecteerde onafhankelijke variabelen voor deze selectiemethode Backward worden gebruikt voor het verwijderen en eventueel toevoegen. Het huidige model is een volledig model.

Stap 2: Gebaseerd op de Maximum likelihood Estimates (MLE) van het huidige model worden de parameters bèta's, de betrouwbaarheidsintervallen, Wald statistiek en de significantie in deze stap uitgerekend voor elke onafhankelijke variabele van het model.

Stap 3: Kies de variabele met de hoogste significantie. Als deze significantie kleiner is dan 0.05, ga dan naar stap 5; anders, als het huidige model zonder deze variabele met de hoogste significantie is hetzelfde als het vorige model, dan is het model af; anders, ga naar de volgende stap.

Stap 4: Wijzig het huidige model bij het verwijderen de variabele met de hoogste significantie uit het model. Alle parameters voor dit gewijzigde model worden uitegerekend en ga terug naar stap 2.

Stap 5: Controleer of er nog te kiezen variabele is die nog niet in het model zit. Wanneer er geen meer is, dan is het model af; anders, ga naar volgende stap.

Stap 6: Gebaseerd op MLEs van het huidige model, wordt de score statistiek voor alle variabelen berekend die nog niet in het model zijn en zoek naar hun significantie.

Stap 7: Kies de variabele met de laagste significantie. Als deze significantie kleiner is dan 0.05, ga dan naar de volgende stap; anders, is het model af.

Stap 8: De variabele met de laagste significantie is de beste variabele om te worden opgenomen in het huidige model. Als het model niet hetzelfde is als een van de vorige modellen, dan worden de parameters voor het nieuwe model geschat en ga terug naar stap 2; anders, is het model af.

5 Model geschikt prospect Tester

<< VERTROUWELIJK >>

6 Model geschikt prospect Java

<< VERTROUWELIJK >>

7 Model geschikt prospect ECM

<< VERTROUWELIJK >>

8 Conclusies en aanbevelingen

<< VERTROUWELIJK >>

9 Literatuurlijst

- Alfred DeMaris (1995). *A Tutorial in Logistic Regression*. Journal of Marriage and the Family, 57: 956-968.
- David W. Hosmer and Stanley Lemeshow (1989). *Applied Logistic Regression*. New York: Wiley.
- Hosmer, D. W., and S. Lemeshow. 2000. *Applied Logistic Regression*, 2nd ed. New York: John Wiley and Sons.
- Jennings, D. E. 1986. *Outliers and residual distributions in logistic regression*. *Journal of the American Statistical Association,* 81: 987-990
- Kleinbaum, D. G. 1994. *Logistic Regression: A Self-Learning Text*. New York: Springer-Verlag.
- Lammers, Ben Pelzer & John Hendrickx (1998). *Inleiding Loglineaire Analysen*. Nijmegen: Vakgroep Methoden.
- Marija J. Norušis (1997). SPSS Professional Statistics 7.5. Chicago, SPSS Inc.
- MILES, J.; SHEVLIN, M. (2000).

Verschillende internet pagina's:

- http://www.leeuwendaal.nl/leeuwendaal/adviesgroepen/werving-en-search/arbeidsmarktonderzoek.732.lynkx
- http://www.let.rug.nl/~nerbonne/teach/stats/Moore-McCabe-H15.pdf
- http://www.ru.nl/socialewetenschappen/rtog/naslagwerk/onderdelen/logistische/
- http://www2.chass.ncsu.edu/garson/PA765/logistic.htm
- Binary logistische regressie
- http://www.let.leidenuniv.nl/history/RES/VStat/html/overzicht_spss.html
- http://books.google.nl/books?id=rDTCnK_3RrgC&pg=PA286&lpg=PA285&ots=53x7DLYKjC&dq=meervoudig+logistische+regressie&sig=EsgbAiWjNV2JZkL8W8DbUhZTJ0w#PPA282,M1
- http://www.mcl.fh-osnabrueck.de/~articles/Zinke/SPSS/Tutorial/sample_files/
- http://www.smartdrill.com/CaseStudies/Bank_loan_credit_risk_logistic_regression_model.html
- http://dissertations.ub.rug.nl/FILES/faculties/ppsw/2004/t.a.leonida/samenvat.pdf
- http://courses.ncssm.edu/math/Stat_Inst/PDFS/REG3_LOG.pdf
- http://habe.hogent.be/stat/statistiek/thesis/thesis2.html
- http://www.let.leidenuniv.nl/history/RES/stat/html/les9.html

10 Begrippenlijst

Binomiaal verdeling
The binomial verdeling (De term wordt voor het eerst gebruikt door Yule in 1911) wordt gedefinieerd als:

$$f(x) = \left(\frac{n!}{x!(n-x)!} \right) p^x q^{n-x}$$

for x = 0, 1, 2, ..., n
waarbij:
p :is de kans op succes voor elke test
q :is gelijk aan 1-p
n :is het aantal onafhankelijke test.

Chikwadraattoets
Met chikwadraat (χ^2) kan worden getoetst of er een statistisch significant verband is tussen twee categorale (nominale of geclassificeerde) variabelen. Indien beide variabelen onafhankelijk van elkaar zijn dan zal de verdeling van de waarnemingen volledig op toeval berusten. De variabelen hebben dan immers geen invloed op elkaar (De Vocht, 2002).

Confouder
Als we geen rekening houden met vooropleiding, schatten we het effect van studie-uren op het cijfer totaal verkeerd in. Vooropleiding wordt een confounder genoemd.

Dichotomisch
De wijze van determineren waarbij telkens uit twee (soms meer) kenmerken een keus gedaan moet worden.

F-waarde
De F-waarde wordt berekend door het aandeel verklaarde variantie (Regression) te delen door het aandeel onverklarende variantie (Residual).

Frequentietabel
Een van de meest voorkomende statistische procedures van de beschrijvende statistiek is het maken van een frequentietabel. Bij de tabel worden statistische maten berekend (zoals rekenkundig gemiddelde en standaarddeviatie).
Frequentietabel is nuttig voor een eerste overzichtsblik op je gegevens.

Hosmer en Lemeshow Goodness-of-Fit Test (Hosmer & Lemeshow, 1989) Er wordt nagegaan of er significante verschillen zijn tussen de frequenties zoals die in de data waargenomen worden en de frequenties zoals die door het model voorspeld worden (zie Sig. Bij Hosmer en Lemeshow Goodness test).
Als deze verschillen niet significant zijn, dan kunnen we concluderen dat het model goed bij data past.
Als significantie van de Goodness of Fit test groter is dan 0.05, dan kunnen we concluderen dat het model goed bij de data past.
In deze test moeten we voorzichtig zijn met heel grote steekproeven en met kleinere steekproeven. Bij kleine steekproeven zal de test aangegeven dat het model past, terwijl dit niet het geval is. En bij heel grote steekproeven geeft de test vaak aan dat er significante verschillen zijn, terwijl het model toch goed bij de data past.

Indicatorvariabele
Om categorische variabelen zoals bijvoorbeeld geschikt in een regressie model te gebruiken, worden indicator variabelen gebruikt. Een voorbeeld van een indicatorvariabele is:

X = 1 als de kandidaat geschikt is
X = 0 als de kandidaat niet geschikt is

Constante
De constante β_0. De constante in de regressievergelijking geeft aan welke waarde we op de afhankelijke variabele verwachten als de onafhankelijke variabele gelijk is aan 0.

Normale verdeling of Gauss-verdeling (genoemd naar de Duitse wiskundige Carl Friedrich Gauss) is een begrip uit de kansrekening. Deze verdeling vindt onder meer toepassing in de statistiek. Het is een continue kansverdeling met een asymptotisch gedrag. De bijbehorende kansdichtheid is hoog in het midden, en wordt naar lage en hoge waarden steeds kleiner zonder ooit echt nul te worden. Door de vorm wordt deze kansdichtheid ook wel klokkromme of Gausscurve genoemd. Ze wordt gegeven door de formule:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},$$

waarin twee parameters, μ en σ, voorkomen. De normale verdeling wordt wel genoteerd als $N(\mu,\sigma^2)$-verdeling, wat wil zeggen dat het een normale verdeling is met verwachtingswaarde μ en standaardafwijking σ.
De integraal van deze functie, voor x lopend van $-\infty$ tot $+\infty$, is precies 1.
De normale verdeling is symmetrisch om het centrum: de verwachtingswaarde μ van de verdeling is het 'middelpunt' van de grafiek van de verdelingsfunctie. De 'breedte' van de grafiek van de kansdichtheid wordt gekarakteriseerd door de standaarddeviatie σ (of de variantie σ^2).

Multicollineariteit
Is een probleem dat zich enkel in binaire regressieanalyse voordoet. Het betreft het optreden van een hoge correlatie tussen sommige verklarende variabelen, waardoor de schatting van de coëfficiënten minder betrouwbaar wordt.

Missing values
De term “missing values” wordt waarschijnlijk bij sommige tabel gebruikt, met daarachter een aantal. Dit aantal is het aantal eenheden waarvan bij de sommige onderzochte variabele geen waarde bekend is. In SPSS worden deze eenheden totaal uitgesloten voor de analyse.

Nulhypothese
Een test begint bij het stellen van een nulhypothese. Hierin wordt gesteld dat de resultaten gelijk zijn onder alle condities en dat er geen verschil bestaat tussen bijvoorbeeld de gemiddelden van twee steekproeven. Men hoopt echter dat deze nulhypothese niet waar is, d.w.z. dat men deze kan verwerpen. De p-waarde is de waarschijnlijkheid dat deze nulhypothese waar is. Indien de p-waarde < 0.05 dan weet men dat er een kans bestaat van minder dan 1 op 20 dat deze nulhypothese waar is. Men kan dan op goede gronden veronderstellen dat er bepaalde (nog niet bekende) oorzaken zijn om de nulhypothese te verwerpen.
Hoe kleiner een steekproef is, des te groter moet men het gebied van "betrouwbaarheid" nemen. Dit, omdat bij kleine steekproeven (N < 200), BIAS en andere "sample errors" een veel belangrijker rol spelen dan bij grote steekproeven.

Een nulhypothese stelt altijd: Er is geen verschil of geen effect of geen associatie / correlatie tussen de situatie in de steekproef ten opzichte van de situatie in de populatie (of een andere steekproef of een bepaalde norm). De nulhypothese wordt onderzocht onder een bepaald maximaal betrouwbaar gebied (CI = Confidence Interval). Pas indien na toetsing de p-waarde buiten dit maximale betrouwbare gebied blijkt te vallen en terecht komt in een zogenaamd overschrijdingsgebied: [100% - CI% = a(alpha)] is er sprake van een statistisch significant (verschillend) resultaat. De waarde van alpha wordt dus van tevoren vastgesteld.

P-waarde

Terwijl de Nulhypothese stelt dat er geen verschil bestaat, hoopt of tracht men met statistische testen aan te tonen dat er weldegelijk een statistisch "bewijs" aanwezig is voor een verschil (altijd bij een van te voren vastgestelde waarde van alpha en een overeenkomstig maximaal betrouwbaarheidsgebied, CI).

Dit "bewijs" wordt beoordeeld aan de hand van de p-waarde. Valt de p-waarde buiten het van tevoren gestelde betrouwbaarheidsinterval dan stelt men dat de Nulhypothese NIET WAAR is onder de geteste omstandigheden. De p-waarde is derhalve de kleinste waarde van alpha waarvoor een nulhypothese verworpen wordt. Een p-waarde kan worden berekend op 3 manieren: Eenzijdig rechts van het gemiddelde, eenzijdig links van het gemiddelde en tweezijdig.

Veel statistische computer programma's geven p-waarden bij gebruik van onderzoeksresultaten waarbij sprake is van een nulhypothese. Tevens worden deze p-waarden in veel publicaties vermeld.

Een p-waarde is in feite de kans op het optreden van het waargenomen steekproef-gemiddelde (of de correlatie) volgens het toeval. Hoe kleiner deze p-waarde, des te minder sprake is van een toevallig resultaat. Een p-waarde gaat echter uit van de aanname dat de gestelde nulhypothese WAAR is. Daarom zegt de p-waarde niet alles en moet, indien mogelijk, ook een Power-analyse worden uitgevoerd.

Regressiecoëfficiënt

De regressiecoëfficiënt geeft aan hoeveel verschil in y we kunnen verwachten voor iedere eenheid toename (positieve coëfficiënt) of afname (negatieve coëfficiënt) in x.

Residuen

Verschil tussen werkelijke en geschatte waarden van de afhankelijke variabele.

ROC curve

Receiver Operating Characteristics curve.

Samenhang tussen twee variabelen

Twee variabelen hangen samen als sommige waarden van één variabele vaker voorkomen bij bepaalde waarden van de tweede variabele dan met andere waarden van de tweede variabele.

Significant

Indien grens-waarde van de van tevoren vastgestelde alpha is overschreden dan moet men de nulhypothese verwerpen. Er is dan namelijk een statistisch significant effect ("bewijs") geconstateerd. Alpha is de significantie-drempel (drempelgebied voor de overgang van wel => niet meer betrouwbaar voor de nulhypothese). Indien de p-waarde groter of gelijk is aan deze alpha dan moet men de nulhypothese accepteren.

Dus indien een verkregen p-waarde = 0.0356 bij gekozen alpha □=0, dan is er sprake van ongeveer 3,5 % toeval, maar moet de nulhypothese worden verworpen (een significant resultaat) omdat men bij alpha □=0.05 nog (ten hoogste) 5% accepteert als een "toevallig resultaat".

Bij een gekozen waarde van alpha = 0.01 (1%) moet dezelfde nulhypothese echter worden aanvaard. Het hangt dus volstrekt af van de waarde van alpha die men kiest.
Bij verwerping van de nulhypothese spreekt men van: een statistisch significant resultaat.

Stapsgewijze meervoudige regressie

Bij stapsgewijze meervoudige regressie wordt per stap een onafhankelijke variabele in het regressiemodel opgenomen op basis van de F-waarde. De onafhankelijke variabele met de laagste significantie (hoogste F-waarde) wordt steeds aan het model toegevoegd. Daarbij wordt gecontroleerd voor de invloed van de variabelen die al in het model zijn opgenomen. Hierdoor veranderen de waarden van de constante en de partiele regressiecoëfficiënten bij iedere stap opnieuw.
De variabelen worden toegevoegd op volgorde van hun relatieve invloed op de afhankelijke variabele Y. Het model voltooid als de significanties van alle nog niet opgenomen variabelen groter zijn dan 0.05. Alleen de significante variabelen worden dus in het model opgenomen.

Te verklarende variabele, verklarende variabele

Een te verklaren variabele meet de uitkomst van een onderzoek. Een verklarende variabele poogt de waargenomen uitkomsten te verklaren.

Vrijheidsgraden

Het aantal vrijheidsgraden (df) van Regression is gelijk aan het aantal onafhankelijke variabelen.
Het aantal vrijheidsgraden (df) van Residual is gelijk aan het aantal cases minus het aantal onafhankelijke variabelen minus 1.

Wald statistiek

De Wald-statistiek geeft aan hoe sterk de bijdrage van de parameter is. Deze Wald waarden zijn gelijk aan het kwadraat van (B/S.E.).

$$\text{Wald} = \left(\frac{B}{S.E.}\right)^2$$

Bijlage A Beschrijving van de dataset

<< VERTROUWELIJK >>

Bijlage B Samenstelling groep kandidaten

De onderstaande frequentietabel biedt inzicht in de frequentie van waarden van elk verklarende variabele. Dit betekent dat er geturfd wordt hoe vaak waarden voorkomen en wat de absolute en relatieve frequenties zijn van die waarden.

<< VERTROUWELIJK >>

		Frequency	Percent	Valid Percent	Cumulative Percent
nationaliteit	geen Nederlands nationaliteit	5	1.7	1.7	1.7
	Nederlands nationaliteit	123	41.4	41.4	43.1
	onbekend	169	56.9	56.9	100.0
	Total	297	100.0	100.0	
werkvergunning	heeft Nederlands werkvergunning	295	99.3	99.3	99.3
	onbekend	2	.7	.7	100.0
	Total	297	100.0	100.0	
Nederlands taal	spreekt geen Nederlands	13	4.4	4.4	4.4
	spreekt Nederlands	78	26.3	26.3	30.6
	onbekend	206	69.4	69.4	100.0
	Total	297	100.0	100.0	
werkervaring	werkervaring junior	75	25.3	25.3	25.3
	werkervaring medior	163	54.9	54.9	80.1
	werkervaring senior	28	9.4	9.4	89.6
	onbekend	31	10.4	10.4	100.0
	Total	297	100.0	100.0	
managementervaring	heeft geen managementervaring	211	71.0	71.0	71.0
	heeft managementervaring	27	9.1	9.1	80.1
	onbekend	59	19.9	19.9	100.0
	Total	297	100.0	100.0	
hbo opleiding	heeft geen hbo opleiding	71	23.9	23.9	23.9
	heeft hbo opleiding	90	30.3	30.3	54.2
	onbekend	136	45.8	45.8	100.0
	Total	297	100.0	100.0	
wo opleiding	heeft geen wo opleiding	108	36.4	36.4	36.4
	wo opleiding	53	17.8	17.8	54.2
	onbekend	136	45.8	45.8	100.0
	Total	297	100.0	100.0	
gewenst functie Tester	geen gewenst functie tester	231	77.8	77.8	77.8
	gewenst functie tester	66	22.2	22.2	100.0
	Total	297	100.0	100.0	
gewenst functie Java ontwikkelaar	geen gewenst functie java	179	60.3	60.3	60.3
	gewenst functie java	118	39.7	39.7	100.0
	Total	297	100.0	100.0	
gewenst functie ECM	geen gewenst functie ecm	278	93.6	93.6	93.6
	gewenst functie ecm	19	6.4	6.4	100.0
	Total	297	100.0	100.0	
dienstverband vast	geen dienstverband vast	91	30.6	30.6	30.6
	dienstverband vast	206	69.4	69.4	100.0
	Total	297	100.0	100.0	
dienstverband tijdelijk	geen dienstverband tijdelijk	135	45.5	45.5	45.5
	dienstverband tijdelijk	162	54.5	54.5	100.0
	Total	297	100.0	100.0	
dienstverband projectbasisTijdelijk	geen dienstverband projectbasisTijdelijk	143	48.1	48.1	48.1
	dienstverband projectbasisTijdelijk	154	51.9	51.9	100.0
	Total	297	100.0	100.0	
dienstverband contract	geen dienstverband contract	135	45.5	45.5	45.5
	dienstverband contract	162	54.5	54.5	100.0
	Total	297	100.0	100.0	
gewenst salaris	onbekend salaris	130	43.8	43.8	43.8
	gewenst salaris minder dan 2000 euro per maand	5	1.7	1.7	45.5
	gewenst salaris tussen 2000 en 2500	19	6.4	6.4	51.9
	gewenst salaris tussen 2500 en 3000	28	9.4	9.4	61.3
	gewenst salaris tussen 3000 en 3500	21	7.1	7.1	68.4
	gewenst salaris meer dan 3500 minder dan 9000	51	17.2	17.2	85.5
	gewenst salaris tussen 9000 en 14000	43	14.5	14.5	100.0
	Total	297	100.0	100.0	
dienstverband stagiair	geen dienstverband stagiair	284	95.6	95.6	95.6
	dienstverband stagiair	13	4.4	4.4	100.0
	Total	297	100.0	100.0	
dienstverband projectbasis	geen dienstverband projectbasis	135	45.5	45.5	45.5
	dienstverband projectbasis	162	54.5	54.5	100.0
	Total	297	100.0	100.0	
gewenste status fulltime	geen gewenste status fulltime	19	6.4	6.4	6.4
	gewenste status fulltime	278	93.6	93.6	100.0
	Total	297	100.0	100.0	
gewenste status parttime	geen gewenste status parttime	197	66.3	66.3	66.3
	gewenste status parttime	100	33.7	33.7	100.0
	Total	297	100.0	100.0	
gewenste locatie Groningen	geen gewenste locatie Groningen	268	90.2	90.2	90.2
	gewenste locatie Groningen	22	7.4	7.4	97.6
	onbekend	7	2.4	2.4	100.0
	Total	297	100.0	100.0	
gewenste locatie Friesland	geen gewenste locatie Friesland	266	89.6	89.6	89.6
	gewenste locatie Friesland	24	8.1	8.1	97.6
	onbekend	7	2.4	2.4	100.0
	Total	297	100.0	100.0	
gewenste locatie Drenthe	geen gewenste locatie Drenthe	263	88.6	88.6	88.6
	gewenste locatie Drenthe	27	9.1	9.1	97.6
	onbekend	7	2.4	2.4	100.0
	Total	297	100.0	100.0	
gewenste locatie Overijssel	geen gewenste locatie Overijssel	251	84.5	84.5	84.5
	gewenste locatie Overijssel	39	13.1	13.1	97.6
	onbekend	7	2.4	2.4	100.0
	Total	297	100.0	100.0	
gewenste locatie Gelderland	geen gewenste locatie Gelderland	229	77.1	77.1	77.1
	gewenste locatie Gelderland	61	20.5	20.5	97.6
	onbekend	7	2.4	2.4	100.0
	Total	297	100.0	100.0	
gewenste locatie Utrecht	geen gewenste locatie Utrecht	156	52.5	52.5	52.5
	gewenste locatie Utrecht	134	45.1	45.1	97.6
	onbekend	7	2.4	2.4	100.0
	Total	297	100.0	100.0	
gewenste locatie Noord Holland	geen gewenste locatie Noord Holland	131	44.1	44.1	44.1
	gewenste locatie Noord Holland	159	53.5	53.5	97.6
	onbekend	7	2.4	2.4	100.0
	Total	297	100.0	100.0	
gewenste locatie Zuid Holland	geen gewenste locatie Zuid Holland	156	52.5	52.5	52.5
	gewenste locatie Zuid Holland	134	45.1	45.1	97.6
	onbekend	7	2.4	2.4	100.0
	Total	297	100.0	100.0	
gewenste locatie Zeeland	geen gewenste locatie Zeeland	290	97.6	97.6	97.6
	onbekend	7	2.4	2.4	100.0
	Total	297	100.0	100.0	
gewenste locatie Noord-Brabant	geen gewenste locatie Noord Brabant	241	81.1	81.1	81.1
	gewenste locatie Noord Brabant	49	16.5	16.5	97.6
	onbekend	7	2.4	2.4	100.0
	Total	297	100.0	100.0	
gewenste locatie Limburg	geen gewenste locatie Limburg	275	92.6	92.6	92.6
	gewenste locatie Limburg	15	5.1	5.1	97.6
	onbekend	7	2.4	2.4	100.0
	Total	297	100.0	100.0	
gewenste locatie Flevoland	geen gewenste locatie Flevoland	230	77.4	77.4	77.4
	gewenste locatie Flevoland	60	20.2	20.2	97.6
	onbekend	7	2.4	2.4	100.0
	Total	297	100.0	100.0	
beschikbaar direct	geen beschikbaar direct	80	26.9	26.9	26.9
	beschikbaar direct	41	13.8	13.8	40.7
	onbekend	176	59.3	59.3	100.0
	Total	297	100.0	100.0	
skills Java	heeft geen skills Java	260	87.5	87.5	87.5
	heeft wel skills Java	37	12.5	12.5	100.0
	Total	297	100.0	100.0	
skills Linux	heeft geen skills Linux	289	97.3	97.3	97.3
	heeft wel skills Linux	8	2.7	2.7	100.0
	Total	297	100.0	100.0	
skills Itil	heeft geen skills Itil	281	94.6	94.6	94.6
	heeft wel skills Itil	16	5.4	5.4	100.0
	Total	297	100.0	100.0	
skills Prince twee	heeft geen skills Prince	292	98.3	98.3	98.3
	heeft wel skills Prince	5	1.7	1.7	100.0
	Total	297	100.0	100.0	
skills Html	heeft geen skills Html	275	92.6	92.6	92.6
	heeft wel skills Html	22	7.4	7.4	100.0
	Total	297	100.0	100.0	
skills Javascript	heeft geen skills Javascript	282	94.9	94.9	94.9
	heeft wel skills Javascript	15	5.1	5.1	100.0
	Total	297	100.0	100.0	
skills Oracle	heeft geen skills Oracle	292	98.3	98.3	98.3
	heeft wel skills Oracle	5	1.7	1.7	100.0
	Total	297	100.0	100.0	
skills MySQL	heeft geen skills Mysql	276	92.9	92.9	92.9
	heeft wel skills Mysql	21	7.1	7.1	100.0
	Total	297	100.0	100.0	
skills SQL	heeft geen skills SQL	271	91.2	91.2	91.2
	heeft wel skills SQL	26	8.8	8.8	100.0
	Total	297	100.0	100.0	
skills Tmap	heeft geen skills Tmap	263	88.6	88.6	88.6
	heeft wel skills Tmap	34	11.4	11.4	100.0
	Total	297	100.0	100.0	
skills Testframe	heeft geen skills Testframe	288	97.0	97.0	97.0
	heeft wel skills Testframe	9	3.0	3.0	100.0
	Total	297	100.0	100.0	
skills Information Systems Examination Board	heeft geen skills Iseb	290	97.6	97.6	97.6
	heeft wel skills Iseb	7	2.4	2.4	100.0
	Total	297	100.0	100.0	
skills International Software Testing Qualification Board	heeft geen skills Istqb	290	97.6	97.6	97.6
	heeft wel skills Istqb	7	2.4	2.4	100.0
	Total	297	100.0	100.0	
skills Functional Acceptance Test	heeft geen skills Fat	297	100.0	100.0	100.0
skills Gebruikersacceptatie Test	heeft geen skills Gat	295	99.3	99.3	99.3
	heeft wel skills Gat	2	.7	.7	100.0
	Total	297	100.0	100.0	
skills Quicktestpro	heeft geen skills Quicktestpro	296	99.7	99.7	99.7
	heeft wel skills Quicktestpro	1	.3	.3	100.0
	Total	297	100.0	100.0	
skills Testdirector	heeft geen skills Testdirector	286	96.3	96.3	96.3
	heeft wel skills Testdirector	11	3.7	3.7	100.0
	Total	297	100.0	100.0	
skills Geautomatiseerd testen	heeft geen skills GeautomatiseerdTesten	297	100.0	100.0	100.0

Bijlage C Verklarende variabelen

Hieronder staan de vijfenzeventig verklarende variabelen. De skills variabelen voor de functie Tester, de functie Java en de functie ECM zijn gemarkeerd achtereenvolgens met de letter T, J en S.

<< VERTROUWELIJK >>

Bijlage D voor de functie Tester

Deze bijlage staat uit:
1. Een overzicht van de 50 verklarende variabelen x_1, x_2, …. x_{50};
2. Grafieken van geschiktheidresultaten;
3. Correlatiematrix;
4. Tabel achterwaartse selectieprocedure;
5. Ondernomen stappen om het eindresultaat te krijgen.

<< VERTROUWELIJK >>

Bijlage E voor de functie Java

In deze bijlage staat:
1. Een overzicht van de 55 verklarende variabelen x_1, x_2, …. x_{55};
2. Grafieken van geschiktheidresultaten;
3. Correlatiematrix;
4. Tabel achterwaartse selectieprocedure;
5. Ondernomen stappen om het eindresultaat te krijgen.

<< VERTROUWELIJK >>

Bijlage F voor de functie ECM

In deze bijlage staat:

1. Een overzicht van de 47 verklarende variabelen x_1, x_2, …. x_{47};
2. Grafieken van geschiktheidresultaten;
3. Correlatiematrix;
4. Tabel achterwaartse selectieprocedure;
5. Ondernomen stappen om het eindresultaat te krijgen.

<< VERTROUWELIJK >>

Bijlage G Excelsheet

<< VERTROUWELIJK >>

Bijlage H Chikwadraatverdeling

df	P = 0.05	P = 0.01	P = 0.001
1	3.84	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
6	12.59	16.81	22.46
7	14.07	18.48	24.32
8	15.51	20.09	26.13
9	16.92	21.67	27.88
10	18.31	23.21	29.59
11	19.68	24.73	31.26
12	21.03	26.22	32.91
13	22.36	27.69	34.53
14	23.69	29.14	36.12
15	25.00	30.58	37.70
16	26.30	32.00	39.25
17	27.59	33.41	40.79
18	28.87	34.81	42.31
19	30.14	36.19	43.82
20	31.41	37.57	45.32
21	32.67	38.93	46.80
22	33.92	40.29	48.27
23	35.17	41.64	49.73
24	36.42	42.98	51.18
25	37.65	44.31	52.62
26	38.89	45.64	54.05
27	40.11	46.96	55.48
28	41.34	48.28	56.89
29	42.56	49.59	58.30
30	43.77	50.89	59.70

df	P = 0.05	P = 0.01	P = 0.001
31	44.99	52.19	61.10
32	46.19	53.49	62.49
33	47.40	54.78	63.87
34	48.60	56.06	65.25
35	49.80	57.34	66.62
36	51.00	58.62	67.99
37	52.19	59.89	69.35
38	53.38	61.16	70.71
39	54.57	62.43	72.06
40	55.76	63.69	73.41
41	56.94	64.95	74.75
42	58.12	66.21	76.09
43	59.30	67.46	77.42
44	60.48	68.71	78.75
45	61.66	69.96	80.08
46	62.83	71.20	81.40
47	64.00	72.44	82.72
48	65.17	73.68	84.03
49	66.34	74.92	85.35
50	67.51	76.15	86.66
51	68.67	77.39	87.97
52	69.83	78.62	89.27
53	70.99	79.84	90.57
54	72.15	81.07	91.88
55	73.31	82.29	93.17
56	74.47	83.52	94.47
57	75.62	84.73	95.75
58	76.78	85.95	97.03
59	77.93	87.17	98.34
60	79.08	88.38	99.62

df	P = 0.05	P = 0.01	P = 0.001
61	80.23	89.59	100.88
62	81.38	90.80	102.15
63	82.53	92.01	103.46
64	83.68	93.22	104.72
65	84.82	94.42	105.97
66	85.97	95.63	107.26
67	87.11	96.83	108.54
68	88.25	98.03	109.79
69	89.39	99.23	111.06
70	90.53	100.42	112.31
71	91.67	101.62	113.56
72	92.81	102.82	114.84
73	93.95	104.01	116.08
74	95.08	105.20	117.35
75	96.22	106.39	118.60
76	97.35	107.58	119.85
77	98.49	108.77	121.11
78	99.62	109.96	122.36
79	100.75	111.15	123.60
80	101.88	112.33	124.84
81	103.01	113.51	126.09
82	104.14	114.70	127.33
83	105.27	115.88	128.57
84	106.40	117.06	129.80
85	107.52	118.24	131.04
86	108.65	119.41	132.28
87	109.77	120.59	133.51
88	110.90	121.77	134.74
89	112.02	122.94	135.96
90	113.15	124.12	137.19
91	114.27	125.29	138.45
92	115.39	126.46	139.66
93	116.51	127.63	140.90
94	117.63	128.80	142.12
95	118.75	129.97	143.32
96	119.87	131.14	144.55
97	120.99	132.31	145.78
98	122.11	133.47	146.99
99	123.23	134.64	148.21
100	124.34	135.81	149.48

Bijlage I Kritieke waarden voor de t-verdeling

degrees of freedom (df)	alpha = (1-tailed)	0.25	0.1	0.05	0.025	0.01	0.005	0.001
	alpha ( ) = (2-tailed)	0.5	0.2	0.1	0.05	0.02	0.01	0.002
1		1.000	3.078	6.314	12.706	31.821	63.657	318.309
2		0.816	1.886	2.920	4.303	6.965	9.925	22.327
3		0.765	1.638	2.353	3.182	4.541	5.841	10.215
4		0.741	1.533	2.132	2.776	3.747	4.604	7.173
5		0.727	1.476	2.015	2.571	3.365	4.032	5.893
6		0.718	1.440	1.943	2.447	3.143	3.707	5.208
7		0.711	1.415	1.895	2.365	2.998	3.499	4.785
8		0.706	1.397	1.860	2.306	2.896	3.355	4.501
9		0.703	1.383	1.833	2.262	2.821	3.250	4.297
10		0.700	1.372	1.812	2.228	2.764	3.169	4.144
11		0.697	1.363	1.796	2.201	2.718	3.106	4.025
12		0.695	1.356	1.782	2.179	2.681	3.055	3.930
13		0.694	1.350	1.771	2.160	2.650	3.012	3.852
14		0.692	1.345	1.761	2.145	2.624	2.977	3.787
15		0.691	1.341	1.753	2.131	2.602	2.947	3.733
16		0.690	1.337	1.746	2.120	2.583	2.921	3.686
17		0.689	1.333	1.740	2.110	2.567	2.898	3.646
18		0.688	1.330	1.734	2.101	2.552	2.878	3.610
19		0.688	1.328	1.729	2.093	2.539	2.861	3.579
20		0.687	1.325	1.725	2.086	2.528	2.845	3.552
21		0.686	1.323	1.721	2.080	2.518	2.831	3.527
22		0.686	1.321	1.717	2.074	2.508	2.819	3.505
23		0.685	1.319	1.714	2.069	2.500	2.807	3.485
24		0.685	1.318	1.711	2.064	2.492	2.797	3.467
25		0.684	1.316	1.708	2.060	2.485	2.787	3.450
26		0.684	1.315	1.706	2.056	2.479	2.779	3.435
27		0.684	1.314	1.703	2.052	2.473	2.771	3.421
28		0.683	1.313	1.701	2.048	2.467	2.763	3.408
29		0.683	1.311	1.699	2.045	2.462	2.756	3.396
30		0.683	1.310	1.697	2.042	2.457	2.750	3.385
31		0.682	1.309	1.696	2.040	2.453	2.744	3.375
32		0.682	1.309	1.694	2.037	2.449	2.738	3.365
33		0.682	1.308	1.692	2.035	2.445	2.733	3.356
34		0.682	1.307	1.691	2.032	2.441	2.728	3.348
35		0.682	1.306	1.690	2.030	2.438	2.724	3.340
36		0.681	1.306	1.688	2.028	2.434	2.719	3.333
37		0.681	1.305	1.687	2.026	2.431	2.715	3.326
38		0.681	1.304	1.686	2.024	2.429	2.712	3.319
39		0.681	1.304	1.685	2.023	2.426	2.708	3.313
40		0.681	1.303	1.684	2.021	2.423	2.704	3.307
41		0.681	1.303	1.683	2.020	2.421	2.701	3.301
42		0.680	1.302	1.682	2.018	2.418	2.698	3.296
43		0.680	1.302	1.681	2.017	2.416	2.695	3.291
44		0.680	1.301	1.680	2.015	2.414	2.692	3.286

45		0.680	1.301	1.679	2.014	2.412	2.690	3.281
46		0.680	1.300	1.679	2.013	2.410	2.687	3.277
47		0.680	1.300	1.678	2.012	2.408	2.685	3.273
48		0.680	1.299	1.677	2.011	2.407	2.682	3.269
49		0.680	1.299	1.677	2.010	2.405	2.680	3.265
50		0.679	1.299	1.676	2.009	2.403	2.678	3.261

Bijlage J Code in R voor correlatiematrix

<< VERTROUWELIJK >>