Voorspelling van Boodschappenlijstjes

Petra Tol

Stageverslag

Voorspelling van Boodschappenlijstjes

Petra Tol

Stageverslag

Universiteit:
Vrije Universiteit Amsterdam
Faculteit der Exacte Wetenschappen
Boelelaan 1081
1081 HV Amsterdam

Begeleiders:
Marianne Jonker
Rene Bekker

Stagebedrijf:
Albert Heijn
Afdeling Business Analytics
Provinciale weg 11
1506 MA Zaandam

Begeleider:
John-Paul van Doorn

Juli 2009

Voorwoord

Dit stageverslag vormt de afsluiting van mijn masteropleiding Mathematics aan de Vrije Universiteit te Amsterdam. Mijn zes maanden durende stage vond plaats bij Albert Heijn in Zaandam, op de afdeling Business Analytics. Hier heb ik mij beziggehouden met het opstellen van een wiskundig model waarmee boodschappenlijstjes van individuele klanten kunnen worden voorspeld, op basis van de aankoopgegevens in het verleden.

Ik heb deze stage als een leerzame, maar zeker ook leuke afsluiting van mijn studie ervaren, dankzij de verschillende personen die mijn stage mede mogelijk hebben gemaakt. Ik wil Egbert Dijkstra bedanken voor het aanbieden van de stageplaats en John-Paul van Doorn voor de begeleiding vanuit Albert Heijn. Marianne Jonker en Rene Bekker wil ik bedanken voor de begeleiding vanuit de Vrije Universiteit. Ze wisten zeer regelmatig tijd te maken om mijn vorderingen met me door te spreken en nieuwe ideëen aan te dragen.

Tot slot wil ik mijn collega's bij Albert Heijn bedanken voor de plezierige tijd, in het bijzonder de eerder genoemde personen en Maryam Miradi en Jeanine Schoonemann. Tevens wil ik Maryam bedanken voor de hulp bij het extra onderzoek aan het einde van mijn stage.

Petra Tol
Zaandam, juli 2009

Samenvatting

[vertrouwelijk]

Inhoudsopgave

Hoofdstuk 1

Inleiding

In deze inleiding beschrijven we ten eerste de geschiedenis van het bedrijf Albert Heijn en gaan we in op het concept bonuskaart en zijn toepassingen (zie ook de website van Albert Heijn, [1]). Vervolgens wordt de probleemstelling van deze stage beschreven, waarvan de oplossing een basis moet zijn voor een nieuwe toepassing van de bonuskaart. Ook de toegevoegde waarde van deze toepassing voor Albert Heijn wordt besproken. Ten slotte geven we een beschrijving van de structuur van dit stageverslag.

1.1 Albert Heijn

De basis voor de supermarktketen Albert Heijn werd gelegd op 27 mei 1887 toen de 21-jarige Albert Heijn een klein kruidenierswinkeltje in Oostzaan overnam van zijn vader. Na zeven jaar breidt hij uit met een filiaal in Purmerend en zowel het assortiment als het aantal filialen groeit verder. Op 29 april 1920 draagt Albert Heijn de leiding van het bedrijf over aan zijn zoons Gerrit en Jan en schoonzoon Johan Hille. Op dat moment telt het bedrijf, dat altijd de naam van zijn oprichter is blijven dragen, zo'n 75 filialen.

In 1952 opent Albert Heijn de eerste zelfbedieningszaak en het bedrijf groeit uit tot de grootste supermarktketen van Nederland. Tegenwoordig is Albert Heijn een dochterbedrijf van Koninklijke Ahold en telt meer dan 750 winkels. Deze filialen zijn onderverdeeld in verschillende winkeltypes: wijkwinkels, AH XL, AH to go en Albert. Bij deze laatste bestel je je boodschappen via internet en laat ze tot in de keuken bezorgen.

De huidige missie van het bedrijf is 'Het alledaagse betaalbaar, het bijzondere bereikbaar'. Hiermee heeft het bedrijf eigenlijk dezelfde doelstelling voor ogen als de oprichter eens had: 'Prima Kwaliteit, Groote Omzet, Kleine Winst. Arm en rijk kunnen bij mij hun inkopen doen'. Het 'bijzondere' in de huidige missie doelt bijvoorbeeld op de bijzondere producten die Albert Heijn in het assortiment heeft, maar er zijn ook andere zaken waarin het bedrijf zich onderscheidt van de concurrentie. Zo zijn de mogelijkheid tot zelfscannen en het terugzien van je aangekochte producten op internet zaken waarmee Albert Heijn voorop loopt.

Daarnaast hecht Albert Heijn er veel waarde aan om een bijdrage te leveren aan een gezonde en duurzame samenleving. Het uitgangspunt hierbij zijn

de vier belangrijke pijlers: gezondheid, duurzame handel, klimaat en lokale betrokkenheid.

1.2 Bonuskaart

In 1997 werd de bonuskaart geïntroduceerd bij Albert Heijn. In totaal zijn er 10 miljoen kaarten uitgegeven, waarvan er 5,3 miljoen actief gebruikt worden. De bonuskaart is een persoonlijke kaart die de klant korting geeft op bonusaanbiedingen in alle winkels van Albert Heijn, inclusief de webwinkel. Met de kaart kunnen ook Air Miles worden gespaard en met behulp van de bonussleutelhanger kan een gevonden sleutelbos gratis worden opgestuurd naar het bij Albert Heijn opgegeven huisadres.

Regelmatig worden groepen klanten met een bonuskaart, die aangegeven hebben daar interesse in te hebben, geselecteerd voor speciale aanbiedingen, uitnodigingen, proeverijen, enzovoorts. Het zelfscannen en de mogelijkheid tot het terugzien van je aankopen op internet gebeurt ook aan de hand van je bonuskaartnummer. Zo heeft de bonuskaart vele toepassingen en deze stage is het begin van een poging deze mogelijkheden nog verder uit te breiden.

1.3 Probleemstelling

De centrale vraag tijdens deze stage luidt als volgt:

> Hoe kunnen de boodschappenlijstjes van Albert Heijns beste klanten op intelligente wijze worden voorspeld?

Aan de hand van de aankoopgegevens per bonuskaartnummer zoeken we naar een model om voor een klant de aankopen voor de komende week te kunnen voorspellen. Daarbij zoeken we naar een balans tussen het aantal artikelen dat we op de boodschappenlijst plaatsen en de betrouwbaarheid van die lijst. We willen namelijk aan de ene kant zo veel mogelijk aankopen voorspellen, maar aan de andere kant willen we ook dat de lijst niet te veel producten bevat die de klant niet zal kopen.

Verder moeten we er rekening mee houden dat producten specifieke artikelnamen ('ah biologische halfvolle melk 500 ml') en groepsnamen ('melk') hebben. Deze zogenaamde MIAC-groepen (Management Information And Control), zijn waarschijnlijk beter voorspelbaar, maar de omschrijving is voor de klant 'onherkenbaar'. We zullen moeten beslissen hoeveel van deze groepsnamen we op de lijst willen zetten.

1.4 Toegevoegde waarde voor Albert Heijn

Het voorspelde boodschappenlijstje dient in eerste instantie als service aan de klant en moet daarmee de klantloyaliteit verhogen. Verder wordt de toegevoegde waarde voor Albert Heijn gezocht in een hogere omzet per klant, doordat klanten herinnerd worden aan het meenemen van bepaalde producten. Als ze die producten vergeten, zullen ze die wellicht bij een andere supermarkt halen.

Andere voorbeelden van mogelijke toepassingen van de boodschappenlijstjes zijn het geven van een optimale route door de winkel, het doen van persoonlijke

aanbiedingen en waarschuwen als een product op de boodschappenlijst niet voorradig is en de klant zo mogelijk een alternatief bieden. Het uitwerken van deze toepassingen is geen onderdeel van deze stage, het uitzoeken of het überhaupt mogelijk is om het aankoopgedrag van klanten te voorspellen is dat wel.

1.5 Structuur van het verslag

We beginnen het onderzoek met verschillende inleidende onderzoeken naar het aankoopgedrag van klanten, beschreven in hoofdstuk 2. Hier zoeken we bijvoorbeeld naar een verband tussen de aankoophoeveelheden en de tussenaankooptijden. Ook kijken we hoe vaak de verschillende producten terugkeren in de lijst met aankopen van de klant. Om de opgestelde modellen uiteindelijk te kunnen beoordelen beschrijven we een aantal criteria welke we zullen gebruiken voor het vergelijken van de modellen.

In hoofdstuk 3 testen we een eerder opgesteld model voor het voorspellen van boodschappenlijstjes. Vervolgens wordt in hoofdstuk 4 de ontwikkeling van een nieuw model beschreven en wordt deze vergeleken met het eerdere model. In hoofdstuk 5 beschrijven we enkele aanvullende onderzoeken naar het aankoopgedrag van klanten. Ten slotte zijn in hoofdstuk 6 de conclusies en aanbevelingen te vinden ten aanzien van het voorspellen van de boodschappenlijst met dit nieuwe model.

Om het verslag ook leesbaar te houden voor de lezers zonder wiskundige achtergrond, zijn de uitgebreidere wiskundige beschrijvingen geplaatst in bijlage A. In bijlage B zijn de algoritmen om de gevonden modellen toe te passen op de data beschreven in pseudo-code, zodat deze eenvoudig te implementeren zijn in de gewenste programmeertaal. De overige bijlagen bevatten schermvoorbeelden van een mogelijke applicatie en resultaten van het aanvullende onderzoek.

Hoofdstuk 2

Inleidend onderzoek

[vertrouwelijk]

Hoofdstuk 3

Regressieanalyse

[vertrouwelijk]

Hoofdstuk 4

Nieuw model

[vertrouwelijk]

Hoofdstuk 5

Aanvullend onderzoek

[vertrouwelijk]

Hoofdstuk 6

Conclusies en aanbevelingen

[vertrouwelijk]

Bijlage A

Wiskundige beschrijvingen

In deze bijlage geven we eerst een beschrijving van de bekende en minder bekende gebruikte wiskundige begrippen in dit verslag. Vervolgens geven we een wiskundige beschrijving van de gebruikte modellen.

A.1 Begrippen

De hieronder beschreven begrippen zijn in diverse naslagwerken terug te vinden, bijvoorbeeld [5]. Bij de beschrijvingen van de begrippen zijn $x = (x_1, x_2, \ldots, x_n)$ en $y = (y_1, y_2, \ldots, y_n)$ twee steekproeven van n datapunten voor de grootheden X en Y.

Verwachting

Het verwachting van X wordt aangeduid met μ_X en wordt geschat met het steekproefgemiddelde $\bar{x} = \sum_{i=1}^{n} x_i$.

Variantie

De variantie van X wordt aangeduid met $\sigma^2(X)$ en wordt geschat met de steekproefvariantie $s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$.

Standaardafwijking

De standaardafwijking van X is gelijk aan de wortel van de variantie en daarom aangeduid met $\sigma(X)$. De waarde wordt dan ook geschat met de wortel van de steekproefvariantie, s_x.

Relatieve standaardafwijking

De relatieve standaardafwijking van X, oftewel de variatiecoëfficiënt, is gelijk aan de standaardafwijking gedeeld door de verwachting. Deze wordt dus geschat met $\frac{s_x}{|\bar{x}|}$. Deze waarde maakt het mogelijk om voor verschillende grootheden de afwijking van het gemiddelde te vergelijken, wat met alleen de standaardafwijking niet zinvol is als de gemiddelden verschillen.

Covariantie

De covariantie $\text{cov}(X, Y)$ is een maat voor het lineaire verband tussen twee variabelen X en Y. De covariantie geeft aan hoeveel de waarde van X zal toenemen als de waarde van Y toeneemt. $\text{Cov}(X, Y)$ wordt geschat door de steekproefcovariantie van x en y te bepalen, die wordt gegeven door $c_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$.

Correlatiecoëfficiënt

Ook de correlatiecoëfficiënt ρ is een maat voor de samenhang tussen twee variabelen, echter in dit geval is de waarde geschaald zodat verschillende waarden onderling goed te vergelijken zijn. De correlatiecoëfficiënt van X en Y wordt gegeven door:

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sigma(X)\sigma(Y)}.$$

De schatter voor deze waarde, de steekproefcorrelatiecoëfficiënt, wordt dus gegeven door $r_{xy} = \frac{c_{xy}}{s(x)s(y)}$

De waarde van r_{xy} ligt altijd in het interval $[-1, 1]$, waarbij een waarde van -1 of 1 duidt op een sterk negatief, respectievelijk positief lineair verband tussen x en y. Een waarde van 0 betekent dat er geen lineair verband is. Er kan dan nog wel een ander, bijvoorbeeld kwadratisch, verband bestaan.

Met behulp van een correlatietoets kunnen we toetsen of de correlatiecoëfficiënt ρ verschilt van 0. De nulhypothese en de alternatieve hypothese definiëren we daarbij als volgt:

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

We meten hoe groot de kans is onder H_0 dat de afwijking van r ten opzichte 0 gebaseerd is op toeval. Is deze kans kleiner dan 0.05, dan verwerpen we de nulhypothese en noemen we de waarde van r significant verschillend van 0. Kunnen we de nulhypothese niet verwerpen, dan hebben we geen reden om aan te nemen dat er een lineair verband bestaat tussen x en y. Bij weinig waarnemingen is de waarde van r vaak niet significant.

A.2 Regressieanalyse

Regressieanalyse wordt gebruikt om te proberen het gedrag van een responsvariabele te beschrijven aan de hand van één of meer verklarende variabelen. De meest simpele vorm van regressieanalyse is enkelvoudige lineaire regressie (zie ook [6]), waarbij we verwachten dat er een rechtlijnig verband te vinden is tussen de responsvariabele y en één verklarende variabele x. De lijn die dit verband beschrijft wordt dan gegeven door de volgende vergelijking:

$$y = \alpha + \beta x.$$

Hier is α de de zogenaamde intercept, oftewel de waarde van y als x gelijk is aan 0. De helling β geeft aan hoe de waarde van y verandert als x een eenheid

verandert. Indien we de waarde van de parameters α en β weten, kunnen we voor elke waarneming van de verklarende variabele x, de waarde van de responsvariabele y bepalen.

Stel nu dat we n waarnemingen hebben van de variabelen x en y, die we aangeven met (x_i, y_i) voor $i = 1, \ldots, n$. Deze datapunten zullen in werkelijkheid bijna nooit op een rechte lijn liggen. Het verschil tussen de waargenomen waarde voor y en de waarde die voorspeld werd door de regressielijn noemen we een statistische fout. Deze fouten kunnen onstaan door bijvoorbeeld meetfouten of effecten van variabelen die niet in het model zijn opgenomen. We nemen aan dat deze statistische fouten onderling ongecorreleerd zijn, met verwachting 0 en variantie σ^2.

Het enkelvoudige lineaire regressiemodel kunnen we in het kort als volgt beschrijven:

$$\begin{aligned} y_i = \alpha + \beta x_i + e_i \qquad & i = 1, 2, \ldots, n \\ \text{met } E(e_i) &= 0 \\ \text{var}(e_i) &= \sigma^2 \\ \text{cov}(e_i, e_j) &= 0 \qquad i \neq j \end{aligned}$$

Het residu $\hat{e}_i$ is een schatter voor e_i en geeft de verticale afstand tussen de gevonden regressielijn en de waargenomen waarde y_i. We willen natuurlijk graag dat deze afstand zo klein mogelijk is. Daarom zoeken we schattingen voor α en β die de som van de gekwadrateerde residuen minimaliseert. Die zogenaamde kleinste kwadratenschatters $\hat{\alpha}$ en $\hat{\beta}$ worden dus gegeven door die waarden van α en β die de volgende functie minimaliseren:

$$\sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

Deze schatters worden dan gegeven door

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \text{en} \\ \hat{\alpha} &= \bar{y} - \beta \bar{x}. \end{aligned}$$

Hierbij zijn $\bar{x}$ en $\bar{y}$ de gemiddelden van alle x_i, respectievelijk y_i $(i = 1, \ldots, n)$.

Om te bepalen hoe goed het gedrag van de responsvariabele wordt beschreven door de regressielijn, kunnen we de waarde van de determinatiecoëfficiënt gebruiken, welke meestal R^2 genoemd wordt. R^2 is het deel van de variabiliteit in y dat wordt verklaard door regressie naar x en wordt gegeven door

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

met $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$.

De adjusted R^2 houdt rekening met het aantal variabelen p dat in het model wordt opgenomen en het aantal waarnemingen dat we gebruiken voor de regressie. In het geval van enkelvoudige regressie hebben we dus $p = 1$. De adjusted R^2 wordt gegeven door

$$\text{adjusted } R^2 = 1 - \frac{n-1}{n-p-1}(1 - R^2).$$

Deze waarde is beter geschikt om te gebruiken bij een klein aantal waarnemingen en gebruiken we daarom in dit onderzoek.

A.3 Exponential Smoothing

Exponential smoothing is een voorspelmethode voor tijdreeksen, waarbij we elke voorspelling baseren op de voorgaande voorspelling en de werkelijke waarneming (zie ook [4]). Bij single exponential smoothing hebben we een voorspelling $\hat{y}_t$ voor de waarde van de tijdreeks op tijdstip t. Zodra we de werkelijke waarneming y_t voor dat tijdstip hebben, weten we ook de grootte van de fout $y_t - \hat{y}_t$. We passen de voorspelling dan aan met behulp van deze fout. De voorspelling voor het volgende tijdstip wordt dan gegeven door

$$\hat{y}_{t+1} = \hat{y}_t + \alpha(y_t - \hat{y}_t) = \alpha y_t + (1-\alpha)\hat{y}_t, \qquad \alpha \in [0,1].$$

Hoe dichter α bij 1 ligt, hoe groter de invloed van de fout op de volgende voorspelling. Bij het voorspellen van de boodschappenlijstjes staat y_t voor de verstreken tijd tussen aankoop t en $t+1$ van een bepaald artikel en is $\hat{y}_t$ dus de voorspelde waarde van die tussentijd.

De eerste voorspelling die we doen moeten we op een andere wijze dan hierboven bepalen, aangezien we deze niet kunnen baseren op de voorgaande voorspelling. Een algemene methode om deze waarde te bepalen is het gemiddelde nemen van de eerste k waarnemingen en dit als eerste voorspelling te gebruiken. In ons geval hebben we soms weinig waarnemingen, dus nemen we $k = 1$, oftewel $\hat{y}_2 = y_1$. We kunnen $\hat{y}_t$ nu schrijven als een functie van $y_1, \ldots, y_{t-1}$:

$$\hat{y}_t = \alpha \sum_{i=1}^{t-2} (1-\alpha)^{i-1} y_{t-i} + (1-\alpha)^{t-2} y_1$$

We zouden graag de kwaliteit van verschillende voorspellingen willen vergelijken, zoals we dat bij lineaire regressie kunnen doen met behulp van de adjusted R^2. Bij exponential smoothing zijn er verschillende opties voor een dergelijke maat, bijvoorbeeld de sum of squared errors (SSE), mean squared error (MSE), de mean absolute percentage error (MAPE) of de mean absolute scaled error (MASE). Men is het er niet over eens welke van deze maten het beste geschikt is (zie ook [3]). Als we waarnemingen $y_1, \ldots, y_n$ hebben, worden de verschillende maten als volgt bepaald:

$$SSE = \sum_{i=2}^{n} (\hat{y}_i - y_i)^2$$

$$MSE = \frac{1}{n-1} \sum_{i=2}^{n} (\hat{y}_i - y_i)^2$$

$$MAPE = \frac{1}{n-1} \sum_{i=2}^{n} \left| 100 \cdot \frac{\hat{y}_i - y_i}{y_i} \right|$$

$$MASE = \frac{1}{n-1} \sum_{i=2}^{n} \left| \frac{\hat{y}_i - y_i}{\frac{1}{n-1} \sum_{t=1}^{n} |y_t - y_{t-1}|} \right|$$

De grootte van de SSE en de MSE is afhankelijk van de grootte van de y-waardes en daarom niet geschikt om de resultaten van verschillende datasets te vergelijken. De MAPE en de MASE zijn wel geschaald naar de grootte van de waarneming en daarom beter geschikt voor dit doeleinde.

De waarde van α wordt vaak van tevoren vastgesteld. Gewoonlijk ligt die waarde ergens tussen de 0.05 en 0.30 ([2], p.111). Als er voldoende observaties zijn kan α ook zo worden gekozen dat deze de MSE minimaliseert.

Bijlage B

Pseudo Code

[vertrouwelijk]

Bijlage C

Schermvoorbeelden applicatie

[vertrouwelijk]

Bijlage D

Resultaten aanvullend onderzoek

[vertrouwelijk]

Bibliografie

[1] Albert Heijn website, april 2009. `http://www.ah.nl`.

[2] A.C. Harvey. *Time series models.* The MIT Press, 1993.

[3] R.J. Hyndman and A.B. Koehler. Another look at measures of forecast accuracy. *Monash University Australia*, Mei 2005.

[4] R.J. Hyndman, A.B. Koehler, J.K. Ord, and R.D. Snyder. *Forecasting with exponential smoothing.* Springer, 2008.

[5] B.B. van der Genugten. *Inleiding tot de waarschijnlijkheidsrekening en mathematische statistiek.* Stenfert Kroese, 1986.

[6] S. Weisberg. *Applied linear regression.* John Wiley & Sons, 1985.