**Rapportage Data Verwerking, Beschrijvende Statistiek en Correlatie Analyses**

In deze rapportage geven we een korte omschrijving van de uitgevoerde data verwerking, benoemen we de missing waardes per variabele en de meest belangrijke en opvallende beschrijvende statistieken en gevonden correlaties. In een verdere bijlage(s) zullen de volledige beschrijvende statistieken en correlaties gegeven worden. Verder zullen we advies uitbrengen over mogelijke vervolgstappen.

**Data verwerking**

*Verwijderde variabelen*

Een aantal variabelen zullen niet meegenomen worden omdat ze niet bruikbaar zijn voor de voorgestelde modelering. In sommige gevallen gebeurt dit omdat de variabelen niet meer in gebruik zijn of omdat ze geen of nauwelijks informatie bevatten of dezelfde informatie als andere variabelen. In andere gevallen is dit omdat de data ongestructureerd is in de vorm van open vragen. Dit betekent niet dat deze variabelen geen relevante informatie bevatten. Echter gezien de tijdsdruk en gebrek aan beschikbare domeinkennis is het niet efficiënt hier tijd aan te besteden. Een overzicht van verwijderde variabelen en de reden van verwijdering is terug te vinden in bijlage 1.

*Categorische variabelen*

Veel categorische variabelen bevatten naast duidelijk categorieën ook vervuilde data. Hierbij is gekeken of deze data kon worden omgezet naar bruikbare categorieën. Indien mogelijk is dit gedaan indien niet mogelijk werden deze vervangen door “other” of als missing ingevuld afhankelijk van de situatie. Categorieën die minder dan 10 keer voorkwamen zijn tevens omgezet naar “other”. Daarnaast zijn vanuit sommige variabelen nieuwe categorische variabelen gecreëerd. Zo is postcode omgezet naar de variabelen Gemeente, Provincie en Randstad en is er voor utm\_campaign en utm\_adgroup locatie juist afgesplitst. Zo zijn bijvoorbeeld `adgroup` genaamd `elektromonteur\_amsterdam` omgezet in ` elektromonteur ` en is er een variabele `adgroup\_location` waarin `amsterdam` wordt aangegeven.

Bij de variabele leeftijd bleek dat deze deels als cijfer is ingevuld maar ook enige tijd als categorische variabele met categorieën 18-30 jaar, 30-40 jaar, 50 of ouder. Er is hier voor gekozen om een variabele te maken leeftijd\_cat en alle cijfermatige leeftijden ook om te zetten naar deze drie categorieën.

*Continue variabelen*

Een aantal datum variabelen zijn omgezet naar data. De leeftijd variabelen bevatte een aantal categorische antwoorden, deze zijn omgezet naar cijfers waarbij een willekeurig getal binnen de categorie is toegekend. Dus bijvoorbeeld “18-30 jaar” wordt een willekeurig getal tussen 18 en 30. Verder is de variabele “Hoe lang in dienst/werkloos” omgezet naar “jaar\_ervaring” hierbij is uit de tekst zo goed mogelijk het aantal jaar ervaring onttrokken. Het wordt aangeraden om die in de toekomst niet als open invulveld te geven maar als cijfer in aantal jaren te vragen.

*Selectie van kandidaten*

Op basis van gesprekken met Timo is ervoor gekozen om enkel kandidaten met de status van prioriteit 1 mee te nemen. Daarnaast is tevens besloten om kandidaten die een ingevulde waarde hadden voor de variabele `afwijsBasisGegevens` ook uit te sluiten. Oorspronkelijk waren er 9943 kandidaten waarvan 173 starter. Na selectie waren er nog 4129 kandidaten over waarvan 136 starter. Het kan mogelijk waardevol zijn om te kijken of de kandidaten waarbij de prioriteit missing was alsnog een prioriteit toe te kennen om zo mogelijk meer kandidaten mee te kunnen nemen. Alle beschrijvende statistieken zijn uitgevoerd over de selectie van kandidaten.

**Missing waardes**

De dataset bevat helaas veel missende waardes. Deze zijn redelijk willekeurig verdeeld over de data waardoor er bijna geen kandidaten zijn die complete data hebben.

|  |  |
| --- | --- |
| **Column** | **Missings** |
| cdate | 0 |
| recruitercode | 471 |
| belafspraak | 198 |
| utm\_source | 173 |
| utm\_medium | 80 |
| utm\_campaign | 2046 |
| utm\_adgroup | 2596 |
| conversiepunt | 241 |
| pagina | 718 |
| uitkomstTelefonischDeal | 309 |
| leeftijd | 216 |
| geboortedatum | 2340 |
| Ben je in het bezit van rijbewijs? | 2338 |
| beschikking tot eigen vervoer? | 2339 |
| score 1 | 2754 |
| score 2 | 2766 |
| score 3 | 2768 |
| Voorkeursbranche | 2392 |
| Werksituatie | 2326 |
| Strevon startsalaris | 2475 |
| Strevon werktijden | 2474 |
| groupid | 0 |
| stage | 0 |
| status | 0 |
| starter | 0 |
| utm\_campaign\_location | 3568 |
| utm\_campaign\_no\_loc | 2046 |
| utm\_adgroup\_location | 3586 |
| utm\_adgroup\_no\_loc | 2596 |
| referrer\_clean | 697 |
| leeftijd\_cat | 217 |
| beschikking tot eigen vervoer?\_clean | 2339 |
| jaar\_ervaring | 2817 |
| Plaats | 235 |

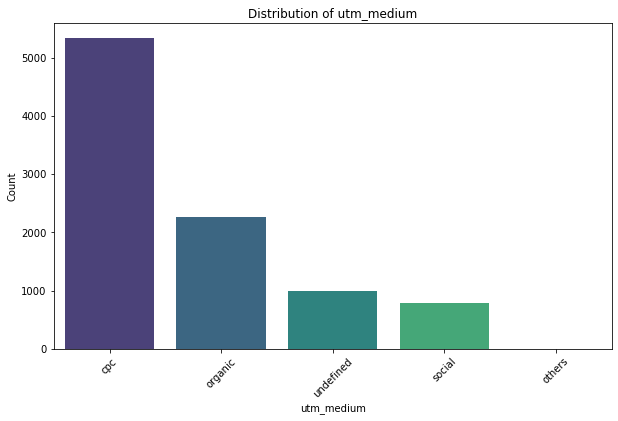
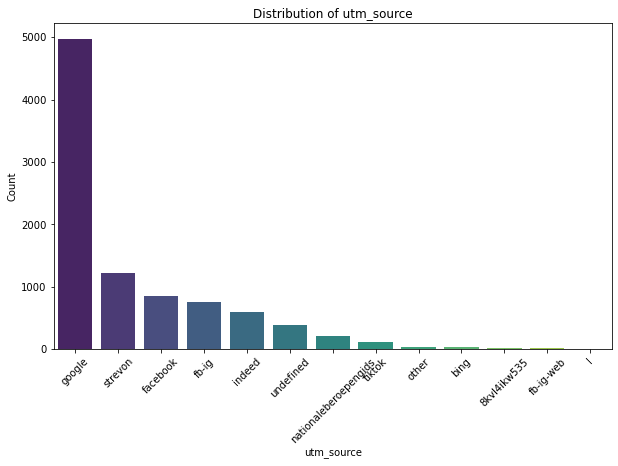
**Beschrijvende Statistiek**

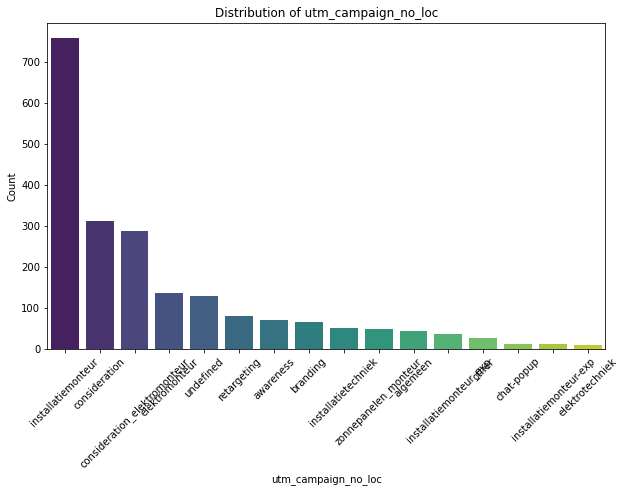
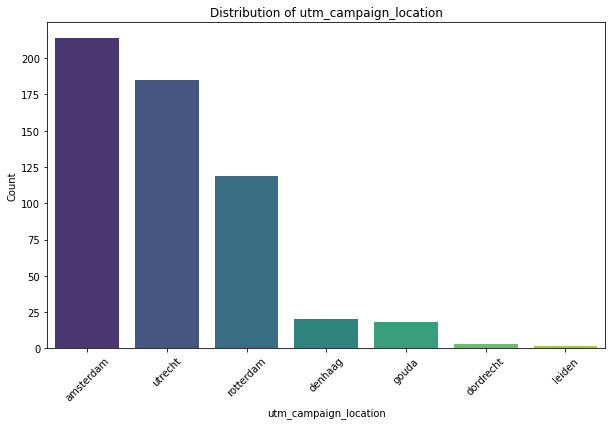
*Starter*

In totaal bleken er van de 4129 kandidaten dus 136 starter te zijn en 3993 geen starter.

*Utm\_source, utm\_medium, utm\_campaign, utm\_adgroup*

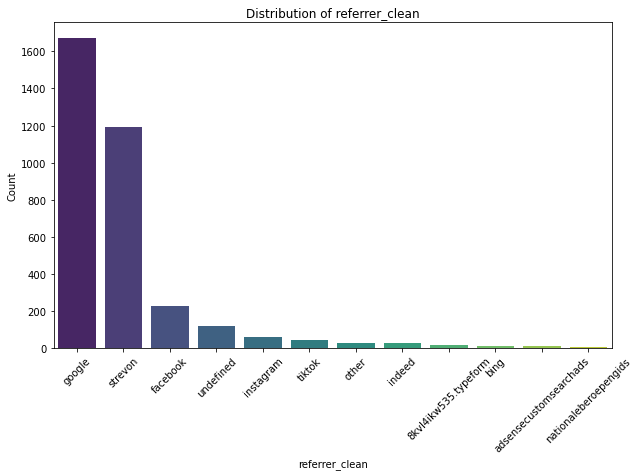
De meeste kandidaten kwamen via google binnen. Daarnaast kwamen deze meestal door middel van cpc (paid advertising). Hiervan kwamen veruit de meeste kandidaten via campagnes gefocused op ‘installatiemonteur’, gevolgd door campagnes ‘consideration’ en ‘consideration\_elektromonteur’. Van de localized campagnes, kwamen de meeste kandidaten binnen via Amsterdam, Utrecht en Rotterdam. De waardes voor adgroup waren redelijk gelijkwaardig verdeeld.





*Referrer*

De meeste mensen kwamen binnen via referrer google of strevon.

**

*Conversiepunt*

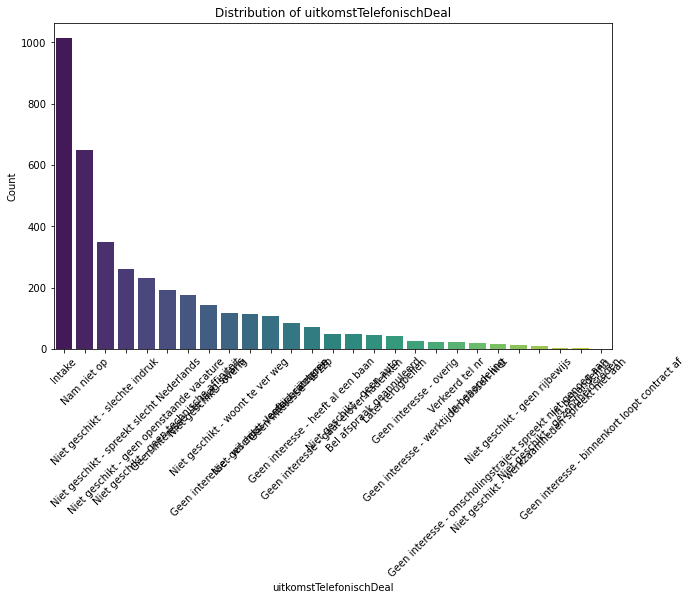
Bijna alle kandidaten converteerden op de website, slechts een klein deel op het leadform.

*Pagina*

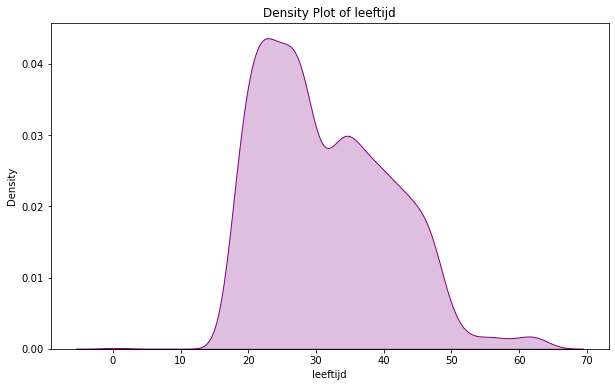
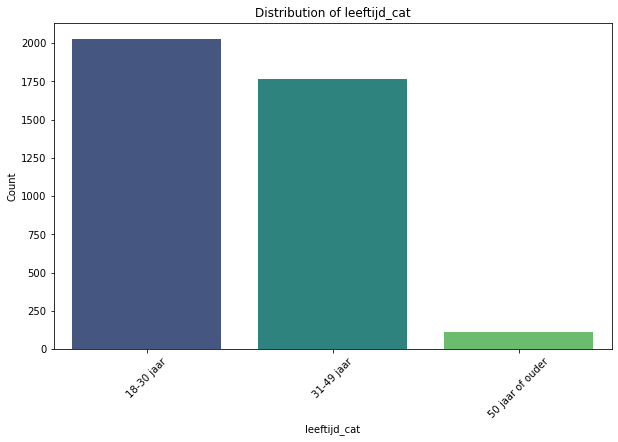
Het aantal pagina’s vanwaar conversie kwam was groot en zeer verspreid. Hier lijkt weinig waardevolle informatie in te zitten. Wel lijkt het alsof het gros van de kandidaten via pagina’s die beginnen met ‘omscholing-‘ converteerden.

*UitkomstTelefonischDeal*

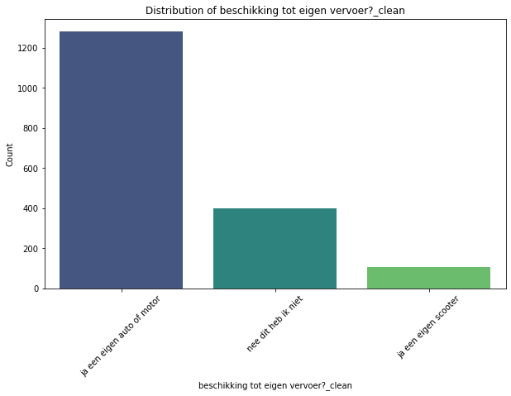
Uit telefonisch contact hadden de meeste kandidaten een “intake” of “nam niet op”.

**

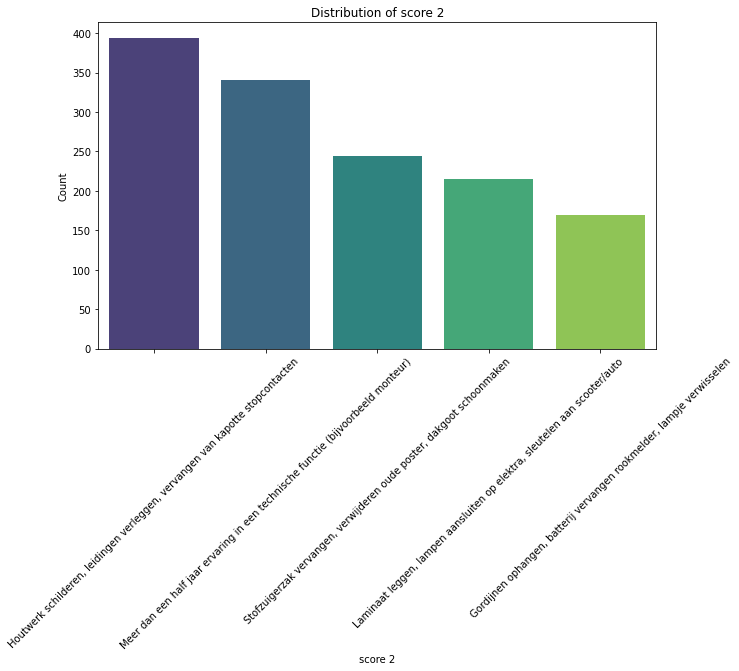
*Leeftijd*

De meeste kandidaten waren jong en zaten in de categorie 18-30 jaar maar kort gevolgd door de group 31-49 jaar. Er waren nauwelijks kandidaten van 50 jaar of ouder.   


*Ben je in bezit van rijbewijs*

**Hier antwoorden 1790 kandidaten met ja, 1 kandidaat met nee en de rest was missing. Hiervan hadden 1280 beschikking over een scooter of motor en 109 een scooter. De rest had geen beschikking over eigen vervoer.

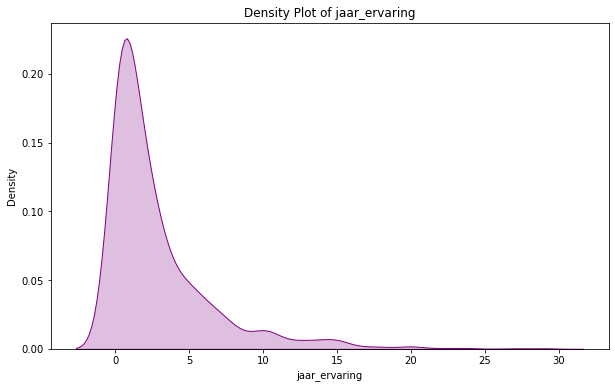
*Score 1, score 2 en score 3*

**Score 1 lijkt eerder een andere vraag te zijn geweest. Veel vreemde antwoorden zijn verwijderd en enkel de vaste categorieën zijn behouden. Op score 2 en score 3 zien we wel consequent dezelfde antwoord categorieën. Omdat onduidelijk is wat de achterliggende vraag was voor deze scores is het moeilijk de gevonden waardes te interpreteren.



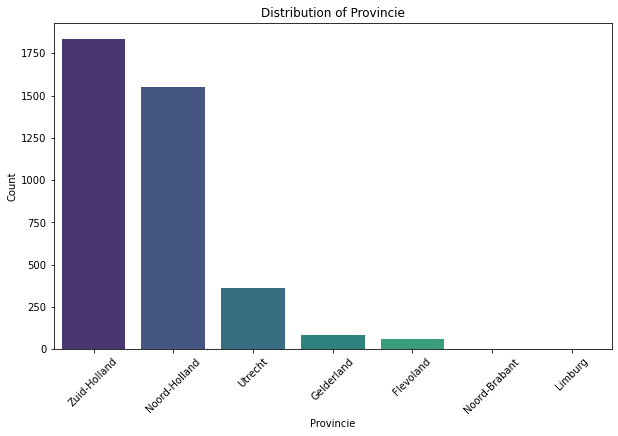
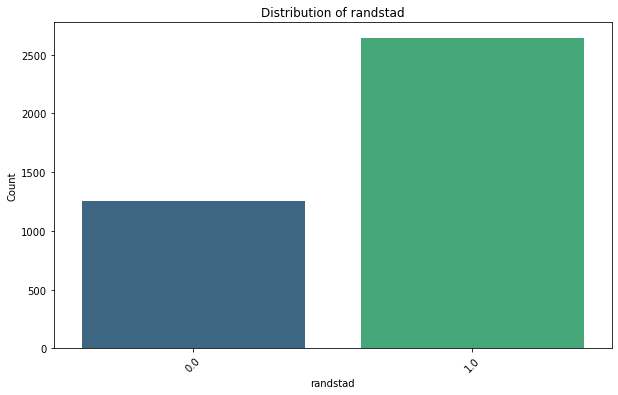
*Jaar ervaring*

De meeste kandidaten hadden geen ervaring of weinig jaren ervaring.

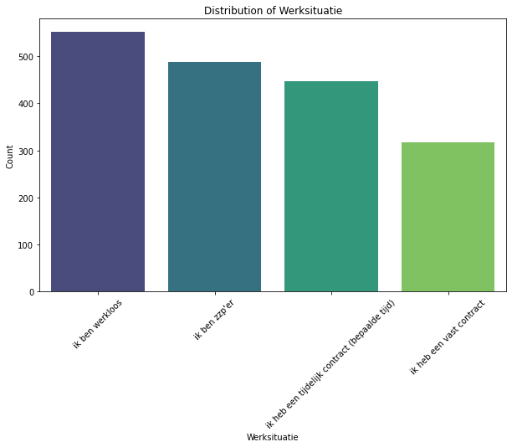
**

*Provincie*

Meeste kandidaten kwamen uit Zuid-Holland en Noord-Holland. Twee van de drie kandidaten woonde in de randstad.

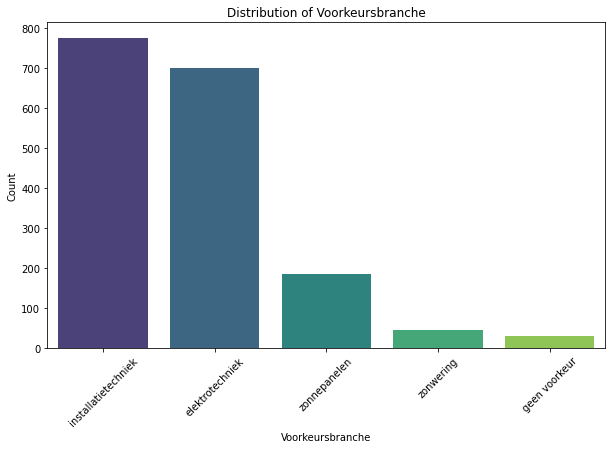
**

*Werksituatie*

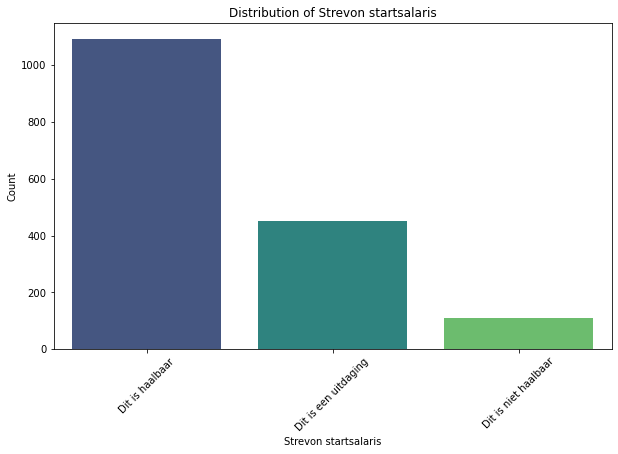
**De meeste kandidaten waren werkloos of zzp’er of hadden een tijdelijk contract.

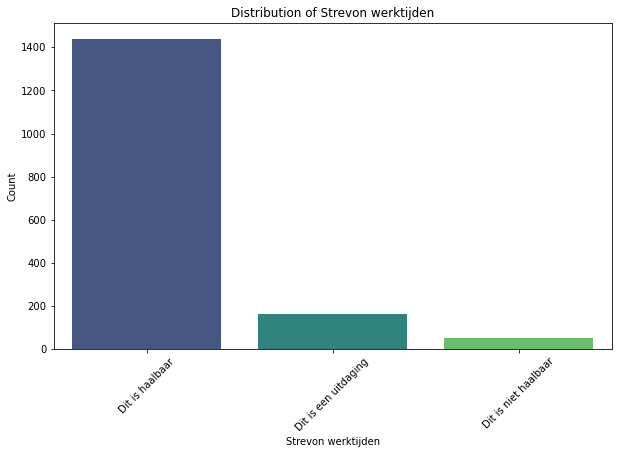
*Voorkeursbranche*

De meeste kandidaten gaven de voorkeur aan werken in installatietechniek of elektrotechniek.

**

*Strevon Startsalaris, Strevon werktijden*

**Strevon startsalaris en werktijden waren in de meeste gevallen haalbaar.

**

**Correlatie analyse**

Om een correlatie analyse uit te kunnen voeren zijn de verschillende categorische variabelen omgezet naar dummy variabelen. Dit houdt in dat voor elke categorie een variabele werd gemaakt die een 1 kreeg als de kandidaat deze categorie had en een 0 wanneer dat niet zo was. De correlatie analyse liet zien dat er vijf variabelen waren met correlatie hoger dan 0.1 met wel of niet starter worden zijn. Omdat we nog niet weten wat groupid, status en stage zijn is het mogelijk dat we deze variabelen niet mee kunnen nemen. Ook uitkomstTelefonischDeal intake is twijfelachtig.

| **variable** | **correlation** |
| --- | --- |
| groupid | 0.849236 |
| status | -0.696374 |
| stage | 0.323257 |
| uitkomstTelefonischDeal\_Intake | 0.320525 |
| Voorkeursbranche\_installatietechniek | 0.105904 |

**Advies voor vervolgstappen**

Vanaf nu kunnen we verschillende modellen gaan testen en vergelijken. Ervan uitgaande dat we alle variabelen mogen gebruiken. We zullen eerst een simpel basismodel opstellen als baseline. Daarna zullen we meer complexe modelstructuren toepassen en vergelijken met de baseline. Uiteindelijk hopen we een bruikbaar model te vinden.

Als dit niet tot een bruikbaar model leidt dan kunnen we kijken of we alsnog de open vragen kunnen gebruiken. Dit kunnen we doen door deze te categoriseren of door een tekst categorisatie model te bouwen. Dit laatste lijkt echter niet ideaal doordat we relatief weinig data hebben voor dergelijke modellen.

Er is nog veel onduidelijkheid over welke variabelen wel of niet gebruikt kunnen worden. Belangrijk is dit zo snel mogelijk helder te krijgen zodat de juiste keuzes gemaakt kunnen worden in het vervolg. Het is belangrijk dat data in de toekomst dan ook op dezelfde manier wordt verzameld om te garanderen dat het model goed blijft werken.

*Overige opmerkingen*

Veel data is niet goed bruikbaar omdat vragen in loop van tijd zijn veranderd of gewijzigd. Daarnaast zijn er veel open vragen waarvan de data tevens niet goed bruikbaar is. Om in de toekomst betere modellen te kunnen bouwen is het belangrijk dat er geen vragen zomaar worden aangepast. Er moet een nieuwe vraag gemaakt worden. En de oude dient te worden verwijderd. Let hierbij op dat als de oude vraag werd gebruikt in het model dit tot problemen kan leiden en het model mogelijk onbruikbaar wordt. Tevens is het belangrijk op te letten op antwoord opties. Leeftijd heeft in het verleden categorische antwoord opties gehad en is nu continu.

Sommige vragen zijn open of deels open. Dit soort vragen zijn niet goed bruikbaar om modellen te ontwikkelen. Tekst categorisatie en interpretatie van open vragen kan alleen gebeuren met grote hoeveelheden data.

Er zijn relatief weinig starters. Dit maakt het moeilijk om modellen te trainen hierop.

*Bijlage 1: Verwijderde variabelen*

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Variabele** | **Reden voor verwijdering** |  |
| 1 | id | Geen info |  |
| 2 | firstName | Privacy |  |
| 3 | lastName | Privacy |  |
| 4 | afwijsRedenContact | Niet meer in gebruik |  |
| 5 | uitkomstIntake | Niet meer in gebruik |  |
| 6 | datumIntake | Niet meer in gebruik |  |
| 7 | belafspraakId | Geen info |  |
| 8 | platformId | Geen info |  |
| 9 | keyword | Geen info |  |
| 10 | gclid | Geen info |  |
| 11 | functienaam | Niet meer in gebruik |  |
| 12 | branche | Niet meer in gebruik |  |
| 13 | welkeFunctieZoekJe | Niet meer in gebruik |  |
| 14 | Salarisindicatie | Niet meer in gebruik |  |
| 15 | datum en tijd gesprek | Niet meer in gebruik |  |
| 16 | contractgesprek | Niet meer in gebruik |  |
| 17 | Source | samengevoegd met `utm\_source` |  |
| 18 | Medium | samengevoegd met `utm\_medium` |  |
| 19 | campagneNaam | samengevoegd met `utm\_campaign` |  |
| 20 | Adgroup | samengevoegd met `utm\_adgroup` |  |
| 21 | UitkomstTelefonischContact | Zelfde informatie als UitkomstTelefonschDeal -> Intake |  |
| 22 | Motivatie | Open veld onbruikbaar |  |
| 23 | Huidige/Laatste functie | Open veld onbruikbaar |  |
| 24 | Meest trotste project | Open veld onbruikbaar |  |

*Bijlage 2: top 100 correlaties*

|  | **variable** | **correlation** |
| --- | --- | --- |
| 0 | groupid | 0.849236 |
| 1 | status | -0.696374 |
| 2 | stage | 0.323257 |
| 3 | uitkomstTelefonischDeal\_Intake | 0.320525 |
| 4 | Voorkeursbranche\_installatietechniek | 0.105904 |
| 5 | Voorkeursbranche\_elektrotechniek | 0.104854 |
| 6 | Voorkeursbranche\_zonwering | 0.085188 |
| 7 | uitkomstTelefonischDeal\_Nam niet op | -0.079772 |
| 8 | uitkomstTelefonischDeal\_Niet geschikt - slecht... | -0.056077 |
| 9 | Ben je in het bezit van rijbewijs?\_ja | -0.054653 |
| 10 | Plaats\_Haarlem | 0.053815 |
| 11 | Gemeente\_Haarlem | 0.053815 |
| 12 | utm\_adgroup\_no\_loc\_elektrotechniek\_omscholing | 0.051710 |
| 13 | uitkomstTelefonischDeal\_Niet geschikt - spreek... | -0.047842 |
| 14 | uitkomstTelefonischDeal\_Niet geschikt - geen o... | -0.045030 |
| 15 | recruitercode | -0.044149 |
| 16 | score 3\_Vervangen batterij afstandsbediening, ... | -0.043572 |
| 17 | Strevon startsalaris\_Dit is een uitdaging | -0.042873 |
| 18 | leeftijd | -0.042650 |
| 19 | beschikking tot eigen vervoer?\_clean\_ja een ei... | -0.041548 |
| 20 | uitkomstTelefonischDeal\_Niet geschikt - geen t... | -0.040644 |
| 21 | beschikking tot eigen vervoer?\_Ja een eigen au... | -0.040141 |
| 22 | Provincie\_Noord-Holland | 0.039181 |
| 23 | leeftijd\_cat\_31-49 jaar | -0.039036 |
| 24 | uitkomstTelefonischDeal\_Geen interesse - te la... | -0.038942 |
| 25 | leeftijd\_cat\_18-30 jaar | 0.038372 |
| 26 | pagina\_opleidingstrajecten/elektrotechniek/avo... | 0.037164 |
| 27 | utm\_adgroup\_no\_loc\_elektromonteur\_bbl | 0.036440 |
| 28 | utm\_campaign\_elektromonteur\_amsterdam | 0.036316 |
| 29 | score 2\_Meer dan een half jaar ervaring in een... | -0.035652 |
| 30 | pagina\_omscholing-zonwering-optin | 0.035536 |
| 31 | uitkomstTelefonischDeal\_Niet geschikt - overig | -0.034956 |
| 32 | Strevon werktijden\_Dit is haalbaar | -0.032547 |
| 33 | utm\_medium\_social | -0.032164 |
| 34 | uitkomstTelefonischDeal\_Niet geschikt - woont ... | -0.031377 |
| 35 | uitkomstTelefonischDeal\_Geen interesse - wil e... | -0.030957 |
| 36 | Strevon werktijden\_Dit is een uitdaging | -0.030585 |
| 37 | Plaats\_Hoogvliet Rotterdam | 0.030518 |
| 38 | Strevon startsalaris\_Dit is niet haalbaar | -0.030389 |
| 39 | uitkomstTelefonischDeal\_Niet geschikt - leefti... | -0.030246 |
| 40 | randstad\_0.0 | -0.030135 |
| 41 | pagina\_opleidingstrajecten/installatietechniek... | -0.029812 |
| 42 | utm\_source\_fb-ig | -0.029514 |
| 43 | conversiepunt\_website | 0.029133 |
| 44 | score 2\_Houtwerk schilderen, leidingen verlegg... | -0.027609 |
| 45 | utm\_adgroup\_no\_loc\_installatietechniek\_bbl | 0.027553 |
| 46 | score 1\_Meer dan een half jaar ervaring in een... | -0.027433 |
| 47 | utm\_adgroup\_location\_amsterdam | 0.027074 |
| 48 | uitkomstTelefonischDeal\_Geen interesse - is zzp | -0.026756 |
| 49 | Provincie\_Gelderland | -0.026433 |
| 50 | utm\_campaign\_no\_loc\_elektromonteur | 0.026423 |
| 51 | Gemeente\_Amersfoort | -0.026142 |
| 52 | Plaats\_Amersfoort | -0.025824 |
| 53 | Werksituatie\_ik heb een tijdelijk contract (be... | -0.024876 |
| 54 | Werksituatie\_ik ben werkloos | -0.024649 |
| 55 | utm\_source\_indeed | -0.024637 |
| 56 | uitkomstTelefonischDeal\_Geen interesse - heeft... | -0.024236 |
| 57 | utm\_campaign\_location\_amsterdam | 0.024188 |
| 58 | score 1\_Laminaat leggen, lampen aansluiten op ... | -0.023726 |
| 59 | utm\_medium\_organic | 0.023478 |
| 60 | Provincie\_Utrecht | -0.023366 |
| 61 | score 3\_Meer dan een half jaar ervaring in een... | -0.023309 |
| 62 | conversiepunt\_leadform | -0.023289 |
| 63 | leeftijd\_cat\_50 jaar of ouder | -0.022816 |
| 64 | Provincie\_Flevoland | -0.022599 |
| 65 | randstad\_1.0 | 0.022464 |
| 66 | utm\_adgroup\_undefined | -0.022354 |
| 67 | utm\_adgroup\_no\_loc\_undefined | -0.022354 |
| 68 | utm\_campaign\_elektromonteur\_utrecht | 0.021863 |
| 69 | Plaats\_Almere | -0.021835 |
| 70 | Gemeente\_Almere | -0.021835 |
| 71 | utm\_adgroup\_elektromonteur\_bbl | 0.021310 |
| 72 | pagina\_opleidingstrajecten/elektrotechniek/bbl | 0.021169 |
| 73 | Plaats\_Rotterdam | -0.021157 |
| 74 | utm\_adgroup\_no\_loc\_installateur\_opleiding | 0.021122 |
| 75 | utm\_adgroup\_no\_loc\_omscholingstraject | -0.020639 |
| 76 | Strevon werktijden\_Dit is niet haalbaar | -0.020639 |
| 77 | referrer\_clean\_facebook | -0.020546 |
| 78 | utm\_campaign\_installatiemonteur | -0.020312 |
| 79 | uitkomstTelefonischDeal\_Geen interesse - gaat ... | -0.020225 |
| 80 | utm\_campaign\_no\_loc\_zonnepanelen\_monteur | -0.020015 |
| 81 | score 1\_Gordijnen ophangen, batterij vervangen... | -0.020008 |
| 82 | uitkomstTelefonischDeal\_Niet geschikt - geen auto | -0.019803 |
| 83 | utm\_adgroup\_loodgieter\_opleiding\_utrecht | 0.019619 |
| 84 | referrer\_clean\_tiktok | -0.019589 |
| 85 | utm\_adgroup\_no\_loc\_werkleertraject | -0.019469 |
| 86 | utm\_campaign\_no\_loc\_algemeen | -0.019372 |
| 87 | uitkomstTelefonischDeal\_Bel afspraak geannuleerd | -0.019372 |
| 88 | beschikking tot eigen vervoer?\_clean\_nee dit h... | -0.019284 |
| 89 | utm\_campaign\_branding | 0.019259 |
| 90 | utm\_campaign\_no\_loc\_branding | 0.019259 |
| 91 | beschikking tot eigen vervoer?\_Nee dit heb ik ... | -0.019024 |
| 92 | pagina\_opleidingstrajecten-bouw-techniek | -0.018669 |
| 93 | uitkomstTelefonischDeal\_Later terugbellen | -0.018482 |
| 94 | utm\_adgroup\_omscholingstraject | -0.018482 |
| 95 | utm\_adgroup\_no\_loc\_elektromonteur\_opleiding | -0.018021 |
| 96 | utm\_campaign\_installatiemonteur\_exp | -0.017787 |
| 97 | utm\_campaign\_no\_loc\_installatiemonteur\_exp | -0.017787 |
| 98 | Provincie\_Zuid-Holland | -0.017767 |
| 99 | Plaats\_Wijchen | -0.017308 |

*Bijlage 3: Beschrijvende statistieken continue variabelen*

|  | **recruitercode** | **leeftijd** | **groupid** | **stage** | **status** | **starter** | **jaar\_ervaring** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| count | 3658.000000 | 3911.000000 | 4129.000000 | 4129.000000 | 4129.000000 | 4129.000000 | 1312.000000 |
| mean | 7.134773 | 31.561238 | 1.455074 | 22.014774 | 1.870671 | 0.032938 | 2.873336 |
| std | 2.232737 | 9.506050 | 1.253089 | 38.768113 | 0.491926 | 0.178495 | 3.713893 |
| min | 1.000000 | 17.000000 | 1.000000 | 2.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 7.000000 | 24.000000 | 1.000000 | 5.000000 | 2.000000 | 0.000000 | 0.500000 |
| 50% | 7.000000 | 29.000000 | 1.000000 | 5.000000 | 2.000000 | 0.000000 | 1.500000 |
| 75% | 9.000000 | 39.000000 | 2.000000 | 15.000000 | 2.000000 | 0.000000 | 4.000000 |
| max | 12.000000 | 64.000000 | 10.000000 | 148.000000 | 2.000000 | 1.000000 | 29.000000 |

**Beschrijvende statistiek DateTime variabelen**

|  | **cdate** | **geboortedatum** | **belafspraak** |
| --- | --- | --- | --- |
| count | 4129 | 1789 | 3931 |
| mean | 2023-06-15 07:20:00.085250560+00:00 | 1993-08-25 06:03:01.106763520+00:00 | 2023-07-05 19:35:14.500127232+00:00 |
| min | 2021-07-30 11:46:23+00:00 | 1972-11-24 00:00:00+00:00 | 2022-03-31 14:30:00+00:00 |
| 25% | 2022-12-16 14:56:43+00:00 | 1987-10-12 00:00:00+00:00 | 2023-01-10 08:35:00+00:00 |
| 50% | 2023-07-13 16:52:47+00:00 | 1995-03-04 00:00:00+00:00 | 2023-08-02 13:30:00+00:00 |
| 75% | 2023-12-13 21:40:13+00:00 | 2000-05-17 00:00:00+00:00 | 2024-01-12 09:45:00+00:00 |
| max | 2024-05-29 14:53:57+00:00 | 2023-05-25 00:00:00+00:00 | 2024-06-05 08:00:00+00:00 |