**Rapportage Modelering 1**

In deze rapportage geven we een korte omschrijving van de laatste aanpassingen aan de data verwerking en rapporteren we de resultaten van de eerste ronde modelering. We laten zien welke modellen er getest zijn en welke modellen als beste presteerden. Vervolgens bespreken we de vervolgstappen om het model te trainen en toe te passen op verschillende momenten in de funnel.

**Laatste aanpassingen data verwerking**

*Verschil in dagen tussen invullen vragenlijst en eerste belmoment*

Op basis van gesprekken met Timo is besloten om ook te kijken naar het toevoegen van verschil in dagen tussen het moment dat een kandidaat de vragenlijst invult en wanneer de belafspraak staat ingepland. Op basis van de correlatie analyses leek dit echter nauwelijks een relatie te hebben tot het converteren tot starter.

**Modelering**

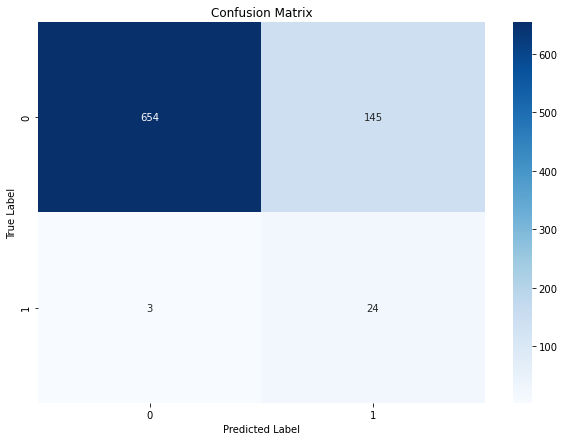
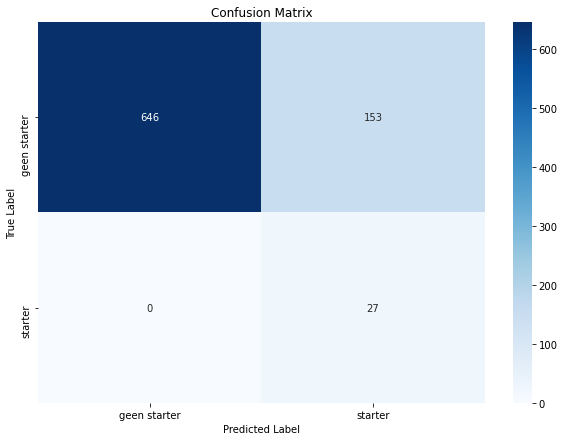
De uiteindelijk dataset bevatte in totaal 4129 bruikbare kandidaten. In totaal waren en 308 voorspellende variabelen en 1 afhankelijke variabele. Bij het testen zijn de variabelen stage, status en groupid niet meegenomen omdat onduidelijk was wat deze variabelen waren ten tijde van het modeleren.

Allereerst moest bepaald worden wat de meest relevante meetcriteria zou zijn. Omdat het belangrijkste was om zoveel mogelijk starters te signaleren en zo min mogelijk te missen werd er gekozen om modellen in eerste instantie te selecteren op basis van Recall en daarna te kijken naar Precision op voorspelling van starter. Dit houdt in dat we in eerste instantie kijken dat we zoveel mogelijk van de starters ook als starter voorspellen en ten tweede proberen om het aantal fout voorspelde starters zo klein mogelijk te maken. In totaal zijn er 10 experimenten uitgevoerd met verschillende modellen en methodes van data resampling. Twee modellen staken hierbij duidelijk boven de rest uit. Dit waren het model BalancedRandomForest model (BRF) en BalancedRandomForest model met RFE feature selectie (RFE-BRF).

Uit de classification plots en confusion matrixes kwam naar voren dat beide modelen sterk waren in het detecteren van starters. De RFE-BRF presteerde net iets beter en vond alle starters maar had iets meer false positives (mensen die als starter werden aangeduid maar het niet waren). Indien dit model gebruikt zou worden dan zou gefocussed kunnen worden op 20% van de deelnemers waarvan uiteindelijk 15% tot starter zou converteren. Let op dit is gebaseerd op de originele selectie van kandidaten met prioriteit 1. Deze waardes kunnen in praktijk anders uitvallen als meerdere selectie criteria van toepassing zijn.

**Advies**

Vervolgstappen zijn om RFE-BRF modellen te bouwen voor specifieke momenten in de funnel. Voorstel is om dit vlak na het invullen van de vragenlijst te doen, na het toekennen van de intake en bij deelnemers met in de aller laatste fase van de funnel op basis van groupid. Hiermee kan dan op verschillende momenten in het proces worden geoptimaliseerd om zo min mogelijk tijd te verspillen aan niet converterende kandidaten en daardoor meer focus te houden op die groep die wel waarschijnlijk zal converteren.

Model1: BRF Model2: RFE-BRF

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classification Report BLR: | | | | |
|  | precision | recall | f1-score | support |
| Geen Starter | 1.00 | 0.82 | 0.90 | 799 |
| Starter | 0.14 | 0.89 | 0.24 | 27 |
| accuracy |  |  | 0.82 | 826 |
| macro avg | 0.57 | 0.85 | 0.57 | 826 |
| weighted avg | 0.97 | 0.82 | 0.88 | 826 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classification Report RFE-BLR: | | | | |
|  | precision | recall | f1-score | support |
| Geen Starter | 1.00 | 0.81 | 0.89 | 799 |
| Starter | 0.15 | 1.00 | 0.26 | 27 |
| accuracy |  |  | 0.81 | 826 |
| macro avg | 0.57 | 0.90 | 0.58 | 826 |
| weighted avg | 0.97 | 0.81 | 0.87 | 826 |