Predicting Telemarketing Call Duration for Banking Product Subscriptions

Predictive Modelling (2024 P3A)

Koen van Esterik
kd.vanesterik@student.han.nl

29th of June 2025

0. Abstract

1. Introduction

Voor elk commercieel bedrijf bestaat de behoefte om winst te maken. Dit om op basis daarvan investeringen te doen, groei te realiseren en uiteindelijk bestaanszekerheid te garanderen. Naast deze behoefte bestaat ook de verplichting richting aandeelhouders, die een belang hebben bij waardecreatie om daarmee de waarde van hun aandelen te vergroten.

Dit geldt ook voor investeringsmaatschappijen die opereren op het gebied van investeren in andere bedrijven en daarbij winst te genereren. Het succes van een investeringsmaatschappij wordt bepaalt door het vermogen om de juiste investeringsbeslissingen te nemen en daarmee de waarde van aandelen van hun aandeelhouders te maximaliseren.

Met dit als context richten we onze aandacht op de investeringsmaatschappij Blackrock, die een investering overweegt - om een Portugese bank over te nemen. Deze Portugese bank voert naast reguliere financiële activeitien onder andere ook telemarketing campagnes uit, om financiële producten te verkopen.

De meest robuuste telemarketing operatie bestaan uit de campagnes om abonementen voor bank depositos te werven. Ze gebruiken hiervoor gegevens van bestaande klanten, die willekeurig gekozen worden om vervolgens uit te bellen. Dit met wisselend succes, omdat het moeilijk te voorspellen is - hoe een klant gaat reageren op een mogelijk storend telefoongesprek. Veel gesprekken zijn opgenomen in een dataset, waarin is aangegeven of de prospect een bank deposito abonnement wil afnemen.

Hier ligt voor Blackrock een mogelijkheid om dieper strategisch inzicht te krijgen met betrekking tot de overname van de bank. Dit door te onderzoeken of winst

gemaximaliseerd kan worden door de willekeurige keuze van het telemarketing process te optimaliseren.

De bijgehouden dataset kan dienen als input voor het gesuggereerde onderzoek van de optimalisaite van het telemarketing process. Dit onderzoek zal gebaseerd zijn op het ontwikkelen van een machine learning voorspellingsmodel. Dit voorspellingsmodel zal als inzicht bieden, die Blackrock kan gebruiken om mee te laten wegen tot overname van de Portugese bank.

Dit onderzoek laat zich verwoorden als: - Reduceer willekeurig gekozen telemarketing campagne prospects, - Door de ontwikkeling van een voorspellingsmodel, - Die beter presteert dan willekeurig kiezen, - Om winstmaximalisatie te realiseren

Het type onderzoek dat gesuggereerd wordt, is gebaseerd op Design Science Research. Dit omdat DSR de nadruk legt op het voldoen van stakeholder requirements, in plaats van traditioneel wetenschappelijk onderzoek - wat de nadruk legt op het genereren van algemene kennis.

De taak die onderzocht moet worden is binaire classificatie, want het onderzoek is gebaseerd om te voorspellen of een prospect een bank deposito abonnement wil afnemen - ja of nee. Deze informatie is aanwezig in de eerder genoemde dataset en zullen als labels dienen. De overige klantinformatie aanwezig in de dataset zullen als predictors dienen.

De dataset in kwestie heeft overigens een mix aan categorische en numerieke waarden binnen een aantal predictors, die voor complicaties zorgen met het ontwikkelen van een voorspellingsmodel. Deze waarden beslaan een impliciete splitsing tussen prospects die wel en niet benaderd zijn binnen een telemarketing campagne. Om dit op te lossen zal de splitsing expliciet gemaakt worden door twee verschillende datasets aan te maken, waarmee respectievelijke twee verschillende voorspellingsmodellen ontwikkeld worden.

Deze splitsing kan mogelijk twee verschillende additionele behoeften bevredigen: 1. Verschillende datasets voor respectievelijk wel en niet benaderd: betekent een ruwe vorm van klantprofilering, wat volgens eerder onderzoek een positief effect heeft in telemarketing campagnes. 2. Verschillende modellen voor respectievelijk wel en niet benaderd: betekent mogelijk betere prestaties, omdat deze ontwikkeld worden op basis van gespecificeerde datasets.

De onderzoekstaak om voorspellingsmodellen te ontwikkelen is gebaseerd op het trainen van modellen met verschillende algoritmen - om vervolgens het best presterende model te kiezen. Primair zal de metric Receiver Operating Characteristics (ROC) voor de modelselectie gebruikt worden. De ROC metric is namelijk gefocust op het detecteren van `false negatives`. Deze `false negatives` treden op wanneer het voorspellingsmodel een potentiële prospect niet identificeert. Dit kan leiden tot gemiste kansen met betrekking tot de werving van deposito abonnementen en daarmee passend om als modelselectie metric voor dit onderzoek te gebruiken.

Het uitendelijk gekozen voorspellingsmodel zal getest worden op data die het

model nog niet heeft gezien. De uitkomsten hiervan zullen gebruikt worden om een drietal Return On Investment (ROI) berekeningen te maken: 1. Een berekening van de huidige werkwijze waar prospects willekeurig worden gekozen. 2. Een berekening van de voorgestelde werkwijze waar benaderde prospects door het voorspellingsmodel worden gekozen. 3. Een berekening van de voorgestelde werkwijze waar niet benaderde prospects door het voorspellingsmodel woden gekozen.

Voor de ROI berekening stellen we de volgende formule voor:

$$ROI = \frac{Profit}{Cost} = \frac{CR * R - C}{C}$$

waarbij

- N, Number of people called;
- C, Kosten per call;
- R, Revenue per succesful subscription;
- CR, Conversion Rate (as a decimal, e.g., 0.02 for 2

De uitkomsten van deze berekeningen zal Blackrock, naast andere overwegingen, dienen om een gewogen besluit te nemen met betrekking tot de overname van de Portugese bank.

2. Methodology

2.1. Dataset

2.2. Exploratory Data Analysis

Om lekkage van data te voorkomen is de feature `duration` niet meegenomen in het train proces. Dit omdat de waarde hiervan niet gebruikt kan worden - wanneer er een voorspelling gedaan wordt, want de eindgebruiker (call agent) weet pas na afloop hoe lang het gesprek heeft geduurd.

2.3. Models

! Beschrijf alle modellen en waarom ze gekozen zijn:

- Neural Networks: als referentiepunt van de originele studie
- Random Forest: simpeler van aard, maar performance van neurale netwerken kan bijhouden
- AdaBoost: een iteratie op random forest met als doel om zwak ingeschatte resultaten te boosten
- XGBoost: gebaseerd op generative model, waardoor sampling mogelijk is

2.4. Preprocess

Een aantal features in de dataset zijn kwalitatieve waarden en moeten omgevormd worden naar kwantitatieve waarden om als input voor een model te kunnen dienen. De features in kwestie zijn:

- job
- marital
- education
- contact
- poutcome

3. Results

3.1. Metrics

- ROC
- Calibration
- Recall

Metric	AdaBoost	Neural Net	Random Forest
AUC	0.7465	0.7626	0.7739
ALIFT	0.2465	0.2626	0.2739

3.1. Model Evaluation

! Beschrijf alle evaluaties en waarom ze gekozen zijn:

- Confusion matrix
- Probalility calibration
- Cost vs threshold analysis

4. Discussion

5. Conclusion

References