Machine Learning For Profit Based Investing

Predictive Modelling (2024 P3A)

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

7th of July 2025

1. Introduction

Voor elk commercieel bedrijf bestaat de plicht om winst te maken. Dit om op basis daarvan investeringen te doen, groei te realiseren en continuiteit te handhaven. [1] Deze verantwoordelijkheid is essentieel voor het bedrijf richting werknemers maar ook richting aandeelhouders. Voor aandeelhouders bestaat dit uit waardecreatie, om daarmee het rendement van de investeringen van aandeelhouders te vergroten.

Dit geldt ook voor investeringsmaatschappijen, die voornamelijk investeren in andere bedrijven om daarbij waardecreatie en rendement te realiseren. Het succes van een investeringsmaatschappij wordt daarbij bepaalt door het vermogen om de juiste investeringsbeslissingen te nemen en daarmee winst voor aandeelhouder te maximaliseren.

Met dit als context heeft de investeringsmaatschappij Blackrock een onderzoek ingesteld, om mogelijk een Portugese bank over te nemen. Deze investering zou in theorie een goed rendement op kunnen leveren voor de aandeelhouders van Blackrock. De Portugese bank voert, naast reguliere financiële activeitien, telemarketing campagnes uit om financiële producten te verkopen. Blackrock wil onderzoeken of zij waardecreatie voor aandeelhouders kan realiseren met haar kennis en expertise met betrekking tot optimalisatie van deze telemarketing campagnes.

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

De bijgehouden dataset dient als input voor het onderzoek. Het onderzoek richt zich op een selectieprocedure van uit te bellen telemarketing prospects. Deze

proceducure wordt gebasseerd op een machine learning voorspellingsmodel. Dit voorspellingsmodel gebruikt de gegevens van bestaande klanten en voorspelt vervolgens of deze een deposito abonnement willen afnemen. Deze voorspellingen zuillen antwoord geven op de vraag - hoeveel kan de effectiviteit verbeterd worden mbt telemarketing campagnes.

Het onderzoek laat zich verwoorden als:

- Vergroot de conversie ratio van telemarketing campagnes,
- Door de ontwikkeling van een selectieprocedure op basis van een voorspellingsmodel,
- Die beter presteert dan uit te bellen naar alle telemarketing prospects,
- Om winstmaximalisatie voor de aandeelhouders van Blackrock te realiseren.

Het voorspellingsmodel is gebaseerd op binaire classificatie, omdat we willen voorspellen of een prospect een bankdeposito-abonnement zal afsluiten: ja of nee. Deze informatie is aanwezig in de eerder genoemde dataset en zal als trainings- en test-data dienen.

Binaire classificatie bestaat uit een begeleide leermethode om gegevens te categoriseren in één van twee mogelijke resultaten. Dit om op basis hiervan voorspellingen te doen op nieuwe, ongeziene gegevens. Bij het evalueren van de prestaties van het binaire classificatiemodel worden deze termen gebruikt:

- True Positive (TP): Het model voorspelt correct een positieve uitkomst.
- False Negative (FN): Het model voorspelt incorrect een negatieve uitkomst.
- False Positive (FP): Het model voorspelt incorrect een positieve uitkomst.
- True Negative (TN): Het model voorspelt correct een negatieve uitkomst.
- Thresholds: De bepaling waar voorspellingen van het model worden ingedeeld.

Vverschillende standaard metrics worden berekend door middel van deze termen - bijv. accuracy, precision, recall, etc. Deze metrics dienen om een classificatiemodel te evalueren. In dit onderzoek gebruiken we deze termen, om te beoordelen welke de meeste winst oplevert. Daarbij stellen we een eigen metric voor die de meeste winst berekent op basis van de eerder genoemde termen (TP, FN, FP, TN en thresholds). We noemen deze metric de **Maximum Profit** (MP) metric.

Eerst bepalen we de voorspellingen voor elke threshold op basis van de kansberekeningen van het classificatiemodel:

$$\vec{y_{pred}} = \sum_{i=1}^{\vec{t}} \begin{cases} 1 & \text{if} \vec{y_{probs}} \geq t_i \\ 0 & \text{otherwise} \end{cases}$$

waarbij:

- $\vec{y_{pred}}$, vector met voorspellingen voor alle drempelwaarden;
- $\vec{y_{probs}}$, vector met probabilities;

- $\vec{t}$, vector met alle drempelweaarden;

Daarna voeren we de voorspellingen in een confusion matrix om de TP's, FN's, FP's en TN's voor alle thresholds te bepalen:

$$\vec{tps}, \vec{fns}, \vec{fps}, \vec{tns} = \sum_{i=1}^{\vec{t}} \text{confusion\_matrix}(y_{\vec{true}}, y_{\vec{pred}})$$

waarbij:

- $y_{\vec{pred}}$, vector met voorspellingen voor alle drempelwaarden;
- $y_{\vec{true}}$, vector met werkelijke waarden;
- $\vec{t}$, vector met alle drempelweaarden;
- $\vec{tps}$, vector met TP's voor alle drempelwaarden;
- $\vec{fps}$, vector met FN's voor alle drempelwaarden;
- $\vec{fps}$, vector met FP's voor alle drempelwaarden;
- $\vec{fps}$, vector met TN's voor alle drempelwaarden;

Tot slot calculeren en bepalen we de maximale winst voor alle drempelwaarden:

$$p = \max\left(r * \vec{tps} - c * (\vec{tps} + \vec{fps})\right)$$

waarbij:

- p, scalar met de maximale winst;
- r, scalar met de opbrengst per succesvol gesprek;
- c, scalar met de kosten per gesprek;
- $\vec{tps}$, vector met TP's voor alle drempelwaarden;
- $\vec{fps}$, vector met FP's voor alle drempelwaarden;

De MP metric introduceren we, om een verbinding te leggen tussen het technische gedeelte van het onderzoek en de business case. Dit met de gedachte dat standaard metrics vaak onvoldoende inzicht bieden met betrekking tot de strategische beslissingen [2] die Blackrock doorgaans moet maken.

De MP metric zal uitkomst bieden bij het evalueren van alle mogelijke classificaitemodellen en zal een uiteindelijke model selecteren op basis van de maximale winst. Vervolgens zal het geselecteerde classificatiemodel gebruikt worden om berekeningen en daarmee een vergelijking te maken met:

1. De huidige werkwijze waar alle prospects gebeld worden.
2. De voorgestelde werkzijze waar het classificatiemodel een voorselectie aan prospects maakt.

Deze vergelijking zal Blackrock antwoord geven op de vraag of het verbeterpotentieel van de telemarketing campagnes dusdanig is, dat de investering van Blackrock in de Portugese bank gerechtvaardigd is vanuit aandeelhouders perspectief.

2. Methodology

Dit onderzoek kan uitgevoerd worden met elke willekeurige predictive data analysis tool setup, maar wij hebben onder andere de volgende systeemconfiguratie gebruikt:

- Python 3.11
- PDM
- Jupyter
- Pandas
- Numpy
- Scikit-Learn

De beschrijving in de repository [3] voor dit onderzoek geeft aan, hoe je dit onderzoek moet opzetten. Dit zodat het onderzoek zelf en de resultaten gevalideerd kunnen worden.

2.1. Dataset

De dataset voor het onderzoek laat zich als volgt beschrijven:

- tijdsreeks data
- 41,000+ instances
- 20 features
  - 10 numeriek
  - 10 categoriaal
- Target met binaire waarden: yes/no
- Geen missende waarden

De data is verzameld in de periode mei 2008 t/m november 2010. Onderstaande opsomming laat de structuur van de dataset zien:

```
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             41188 non-null  int64
 1   job             41188 non-null  object
 2   marital         41188 non-null  object
 3   education       41188 non-null  object
 4   default         41188 non-null  object
 5   housing         41188 non-null  object
 6   loan            41188 non-null  object
 7   contact         41188 non-null  object
 8   month           41188 non-null  object
 9   day_of_week     41188 non-null  object
 10  duration        41188 non-null  int64
 11  campaign        41188 non-null  int64
 12  pdays           41188 non-null  int64
 13  previous        41188 non-null  int64
```

```
14  poutcome        41188 non-null  object
15  emp.var.rate    41188 non-null  float64
16  cons.price.idx  41188 non-null  float64
17  cons.conf.idx   41188 non-null  float64
18  euribor3m       41188 non-null  float64
19  nr.employed     41188 non-null  float64
20  y               41188 non-null  object
```

Dit overzicht dient als referentie voor meerdere beschrijvingen in dit document.

2.2. Data Cleaning

Afgezien dat er geen waarden missen in de dataset, is er wel een andere taak qua data cleaning vereist. De beschrijving van de dataset geeft aan dat de feature *duration* verwijderd moet worden. Dit om opmerkelijke voorsellingen te voorkomen.

Naar onze mening heeft dat te maken met mogelijke data leakage op het moment van trainen van een voorspellingsmodel. Dit met de gedachte dat een voorspelling van een gesprek vooraf niet gemaakt kan worden, als de *duration* van datzelde gesprek nog niet vastgelegd is.

Deze *duration* feature moet daarom verwijderd worden.

2.3. Feature Engineering

De dataset is opgebouwd als een tijdsreeks. Echter mist er een concrete timestamp in de data. Door de periodebepaling van de databeschrijving te gebruiken - samen met de waarden in de *month* feature, kan het jaar per instance bepaald worden. Deze aangemaakte *year* feature zal meerdere inzichten qua data analyse mogelijk maken.

2.4. Preprocessing

Een aantal taken aan preprocessing dienen uitgevoerd te worden, voordat modeltraining kan plaatsvinden.

1. De target feature omvormen van *ja/nee* naar numerieke waarden.
2. Alle categorische waarden omvormen naar numerieke waarden.
3. Alle numerieke waarden omvormen naar een range van nul naar één.
4. Eventuele imbalance verbeteren door extra voorbeelden van ondervertegenwoordigde gegevens te creëeren.

Deze stappen zijn van belang om een goed presterend model te ontwikkelen.

2.5 Train Test Split

De volgende split aan train- en test-date dient uitgevoerd te worden:

- shuffle
- no stratify
- 20% test size

Het uiteindelijke classificatiemodel moet patronen aan prospect profielen ontdekken, die onafhankelijk van elkaar te bepalen zijn. Daarom stellen we een reguliere split voor, ookal is de dataset als tijdsreeks opgebouwd.

2.5. Model Shortlist

De modellen die in aanmerking komen om geevalueerd te worden, voldoen aan de volgende condities:

- Het model moet voorzien zijn van een probability functie (om gebruik te kunnen maken van de MP metric).
- Het model mag niet langer dan maximaal een minuut trainen per data batch.

De modellen die voldoen aan deze condities zijn:

- AdaBoost
- Gradient Boosting
- K-Nearest Neighbors
- Logistic Regression
- Random Forest
- XGBoost

Wellicht bestaan er meer modellen die aan de bovenstaande condities voldoen, maar voor dit onderzoek beperken we ons tot deze shortlist.

2.6. Procedure

De procedure om het onderzoek uit te voeren bestaat uit de volgende stappen:

1. Laad de dataset in.
2. Pas de beschreven data cleaning toe.
3. Pas de beschreven transformaties toe.
4. Split de dataset in een train- en test-set.
5. Hypertune alle modellen op basis van de MP metric als score.
6. Selecteer de best presterende parameters voor elk model.
7. Modelleer alle modellen door middel van cross-validatie.
8. Voorspel de probabilites van alle model door middel van cross-validatie.
9. Bereken de maximale winst voor alle modellen.
10. Evalueer de maximale winst voor alle modellen.
11. Selecteer het model met de hoogst maximale winst.
12. Gebruik de voorspellingen van het geselecteerde model om de huidige werkwijze met de voorgestelde werkwijze te vergelijken.

Wellicht heb je nu een kopje koffie verdiend.

3. Exploratory Data Analysis

Voordat we de beschreven procedure van het onderzoek hebben uitgevoerd, hebben we tevens de dataset zelf onderzocht. Hierbij hebben we een aantal opmerkelijkheden ontdekt, die mogelijk van belang zijn voor de prestaties van het uiteindelijke voorspellingsmodel.

3.1. Data Imbalance

De aangemaakte *year* feature geeft de mogelijkheid om de data per jaar te categorisen. Deze categorisatie laat de volgende twee opmerkelijkheden zien.

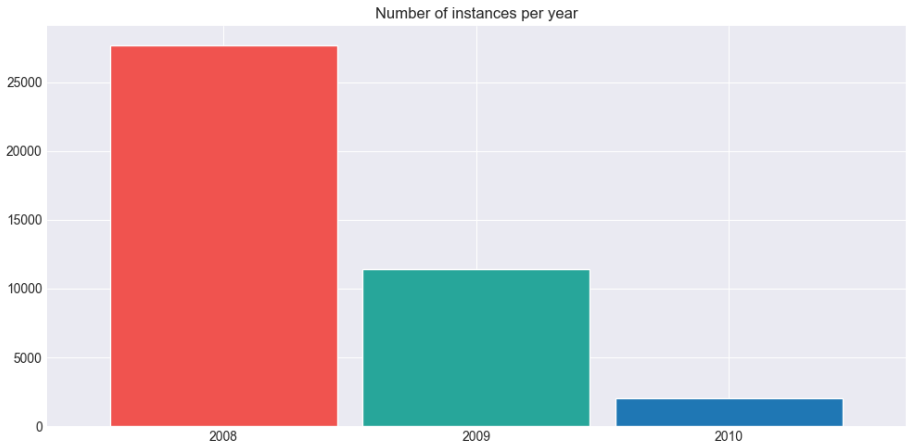

Figure 1: Number of instances per year

De bulk van de data is geconcentreerd in het jaar 2008, zoals figuur 1 aantoont.

Daarnaast verschilt de verdeling van de target per jaar aanzienlijk, zoals figuur 2 aantoont.

We weten niet waarom deze imbalance in de data aanwezig is. Een speculatie is dat er in 2008 een financiële crisis woedde en dat er daarom in dat jaar meer negatief is geantwoord dan andere jaren, op de vraag van de telemarketing campagne. Maar dit kunnen we niet verifiëren, omdat we geen toegang hebben tot de eigenaren van de dataset.

3.2. Approached vs Not-Approached

De *pdays* feature geeft volgens de databeschrijving het aantal dagen aan sinds het laatste contact binnen de lopende telemarketing campagne. Met de waarde *999* als uitzondering op die regel, want dit geeft aan dat de prospect nog niet is benaderd.

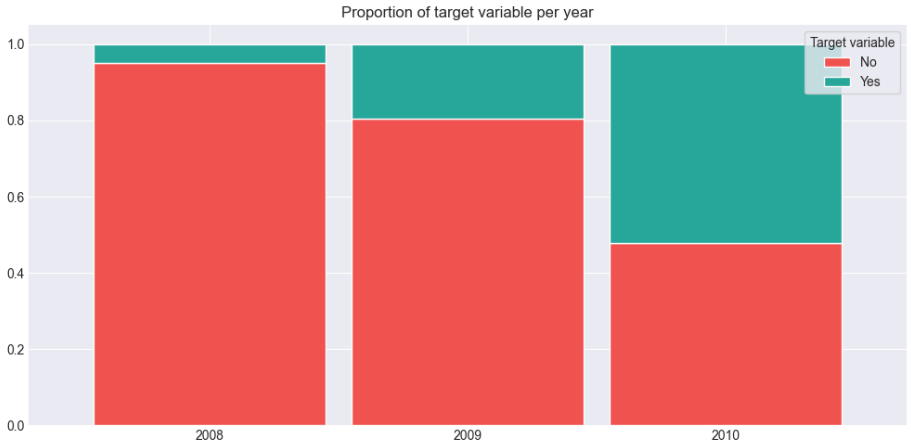

Figure 2: Proportion of target variable per year

Deze mix van numerieke en categorische waarden binnen één feature kan een probleem opleveren bij het trainen van een voorspellingsmodel. Omdat een model dit contextuele onderscheid niet kan maken.

Figuur 3 toont aan dat de verdeling van benaderde ten opzichte van niet-benaderde prospects in het jaar 2008 marginaal is.

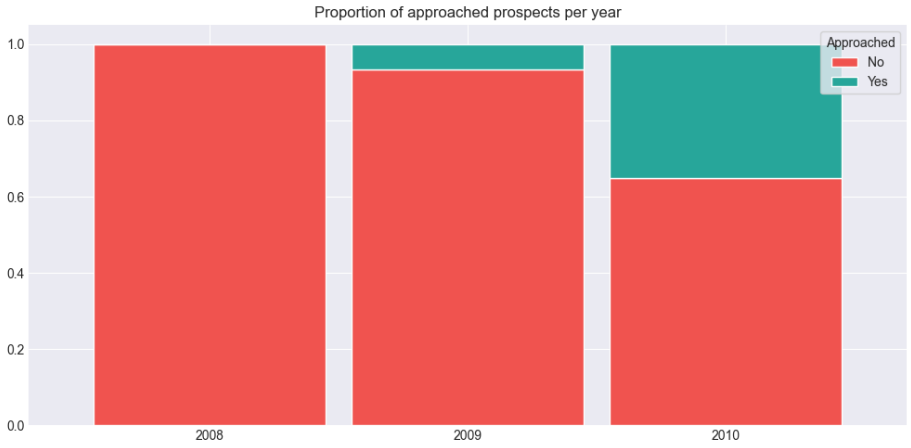

Figure 3: Proportion of approached prospects per year

Daarnaast is het zo dat de instances die geclassificeerd kunnen worden als `not-approached`, wel degelijk informatie bevatten die suggeren dat de prospect al wel benaderd is. Dit geeft het vermoeden dat er fouten in dataset aanwezig zijn. Helaas kan dit niet geverifiëerd worden.

4. Results

We gebruiken cross-validatie om in eerste instantie de shortlist aan classificaite-modellen met elkaar te vergelijken. Deze cross-validatie maakt gebruik van de train-set en levert tevens de input voor de voorgestelde MP metric. De metric heeft een ratio aan revenue en cost nodig en gebruikt de train-set om die uit te rekenen. Dit omdat we cross-validatie willen doen op basis van een realistische verhouding die in de train-set aanwezig is.

Setting	Value
Hourly Wage	50.00
Cost Per Call	100.00
Revenue Per Success	400.00

De bovenstaande verhoudingen geven de volgende resultaten:

Model	Optimal Threshold	Profit	Profit Margin
AdaBoost	0.25	328,300	43.02%
Gradient Boosting	0.24	354,100	45.89%
K-Nearest Neighbors	0.21	248,300	42.06%
Logistic Regression	0.23	351,500	45.16%
Random Forest	0.2	**362,500**	**42.73%**
XGBoost	0.24	203,800	22.53%

De MP-plots in figuur 4 dienen als volgt gelezen te worden: bij een threshold van 0 worden alle prospects gebeld, terwijl er bij een threshold van 1 geen enkele prospect gebeld wordt.

Met deze resultaten is het duidelijk dat het Random Forest model de meeste winst oplevert. Dit met de volgende hyperparameters:

```
{
  'max_depth': 10,
  'min_samples_leaf': 3,
  'min_samples_split': 10,
  'n_estimators': 100
}
```

Deze hyperparameters leveren de profit curve plot in figuur 5 op.

Omdat we gebruik maken van de door het model voorspelde probabilities is het van belang dat deze correct gecalibreerd zijn. Dit omdat de MP metric deze gebruikt om de winst per threshold te berekenen. Zoals in figuur 6 aangetoond wordt, is het Random Forest model redelijk gecalibreerd. Dit betekent dat de

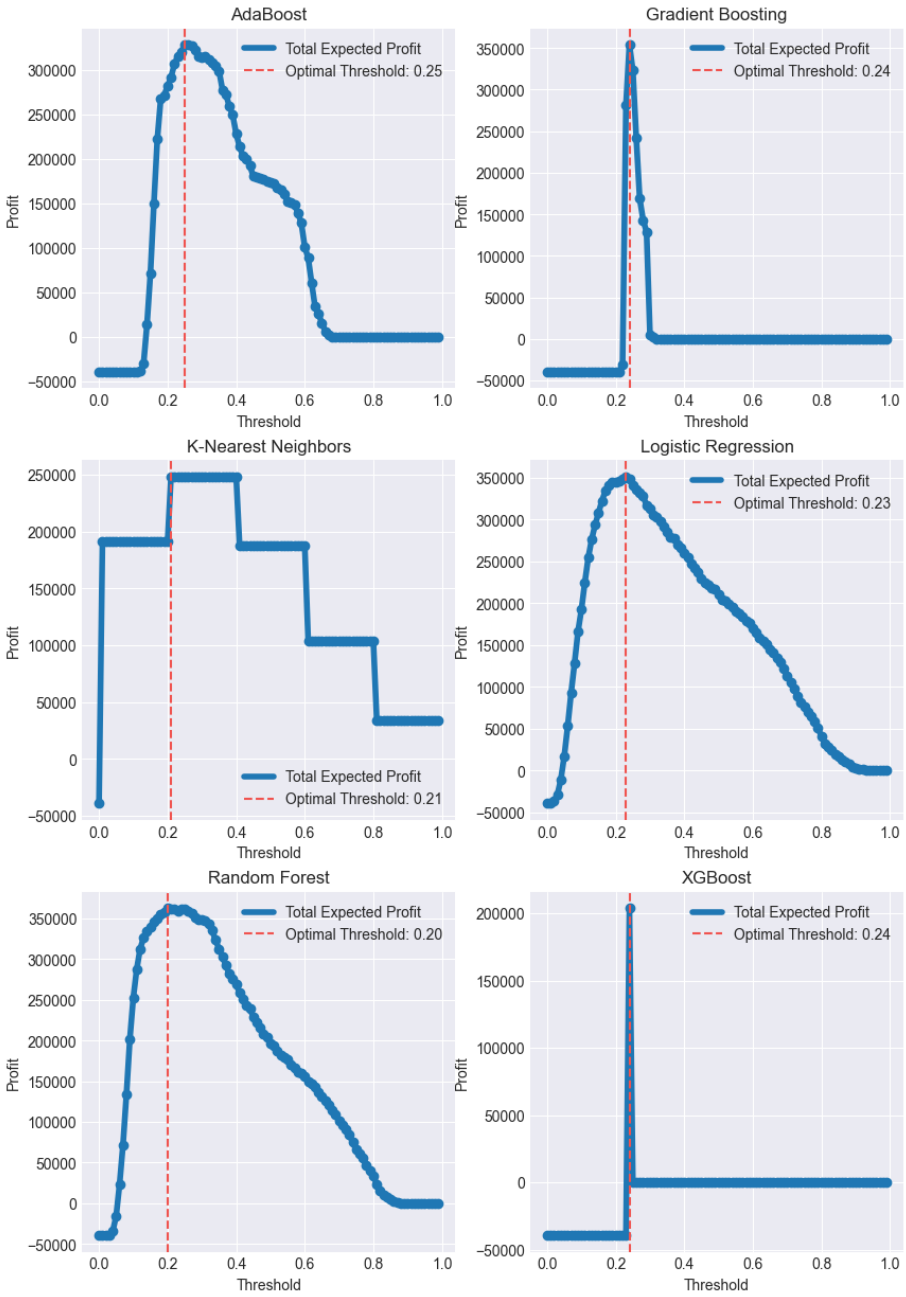

Figure 4: Model Selection

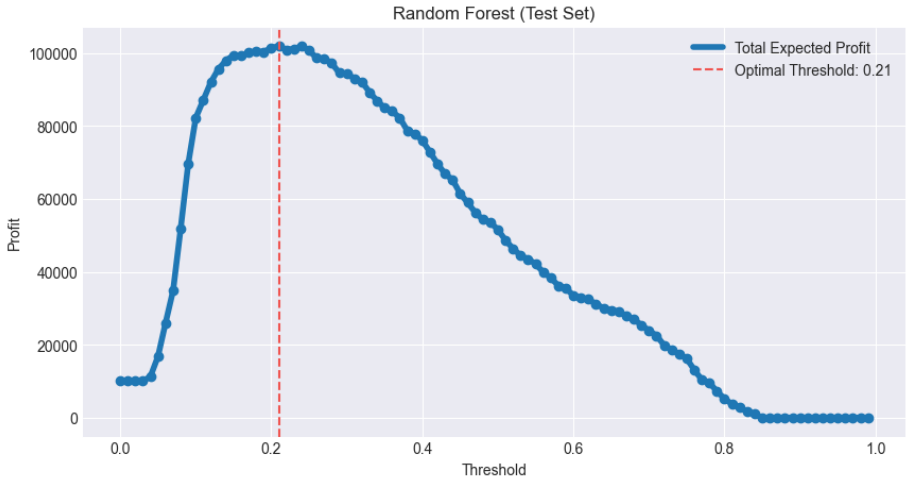

Figure 5: Model Evaluation

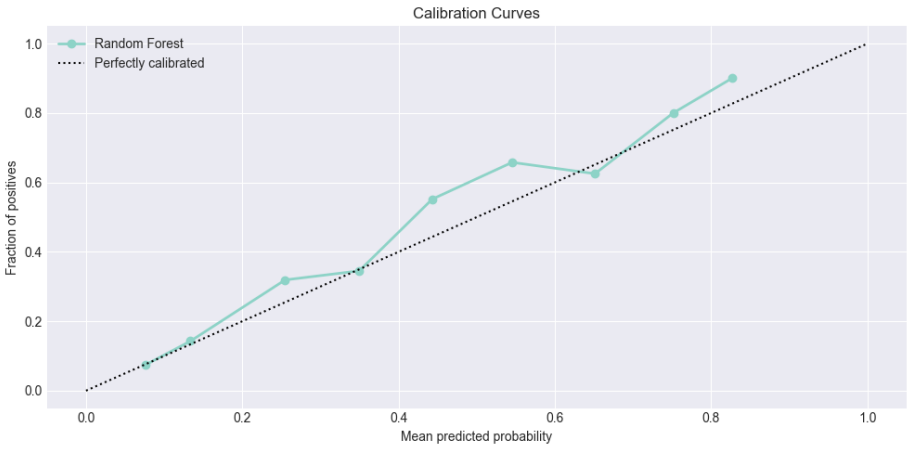

Figure 6: Model Calibration

voorspelde probabilities overeenkomen met de werkelijke kansen van de positieve klasse.

Tevens willen we nog evalueren of het geselecteerde model niet under- of over-fit. Dit doen we met dezelfde MP metric, maar dan op basis van genormaliseerde waarden die de metric berekent. Dit omdat absolute winst getallen niet geschikt zijn om een learning curve te berekenen, doordat deze waarden niet in dezelfde eenheid of schaal zijn - waardoor ze geen eerlijke vergelijking mogelijk maken. De learning-curves in figuur 5 illustreren de evaluatie van het geselecteerde model.

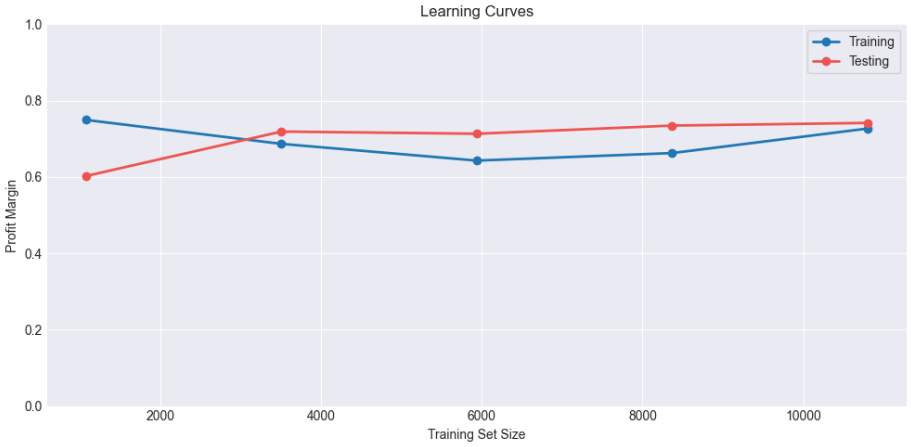

Figure 7: Learning Curves

Veel gedaan ... misschien nu tijd voor een dansje.

5. Discussion

Tijdens het onderzoek zijn er een aantal bevindingen gedaan, die mogelijk om vervolgonderzoek vragen.

- De dataset is qua instances relatief sterk geconcentreerd in het jaar 2008.
- De dataset is uit balans met de betrekking tot de target feature voor vooral het jaar 2008.
- De *pdays* feature heeft een mix van numerieke en categorische waarden.
- De dataset is relatief oud - gezien dit onderzoek in 2025 plaatsvindt.

Om deze issues op te lossen in een mogelijke vervolgonderzoek, is er contact nodig met de eigenaren van de dataset. Dit om meer domeinkennis op te doen en op basis daarvan meer gefundeerde besluiten te nemen met betrekking tot de uitvoering ervan.

6. Conclusion

Als conclusie willen we de effectiviteit van de huidige en voorgestelde telemarketingprocessen analyseren en hun impact op de winstgevendheid evalueren.

De revenue-cost ratio voor de berekeningen zijn als volgt:

Setting	Value
Cost Per Call	100.00
Revenue Per Success	400.00

En geven de volgende resultaten:

Procedure	Profit
Call All Prospects	10,000
Call Preselected Propspects	102,000

Met een evaluatie van deze resultaten kunnen we concluderen dat er aanzienlijke optimalisatiemogelijkheden bestaan binnen het telemarketingproces. Dit leidt ons tot de conclusie dat de investering van Blackrock in de Portugese bank gerechtvaardigd is vanuit het perspectief van de aandeelhouders.

Een oud Nederlands gezegde is nu goed van toepassing: “Gas op die lolly!”

References

[1] “*Principles of Corporate Finance*,” *Wikipedia*. Feb. 04, 2024. Accessed: Jun. 23, 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Principles_of_Corporate_Finance&oldid=1203323000

[2] “Data Science for Business: What you need to know about Data Mining and Data Analytic Thinking.” Accessed: Jul. 05, 2025. [Online]. Available: https://fosterprovost.com/publication/data-science-for-business-what-you-need-to-know-about-data-mining-and-data-analytic-thinking/

[3] K. van Esterik, “Vanesterik/mads-telemarketing-assignment.” Jul. 04, 2025. Accessed: Jul. 05, 2025. [Online]. Available: https://github.com/vanesterik/mads-telemarketing-assignment