Optimizing Targeted Telemarketing Strategies

Institute of Data Capstone Project

Anna-Maria Schreiner, Aspiring Data Scientist

Today’s Agenda

1. Who am I and what can I do for YOU?
2. Case Context
3. Defining the Problem
4. Exploring the Data
   - Process Flow
   - Initial Findings
5. Delivering the Solution
   - Feature Engineering
   - Machine Learning Models
6. Overall Findings
   - Supervised ML
   - Unsupervised ML
7. Next Steps

Anna-Maria Schreiner

- **Academic Background:**
  - Bachelor of International Business Management at University of Queensland (Australia) and Universität St Gallen (Switzerland)
  - RMIT/ Udacity Business Analytics Nanodegree
  - Institute of Data Graduate Certificate of Data Science and Artificial Intelligence (in completion)

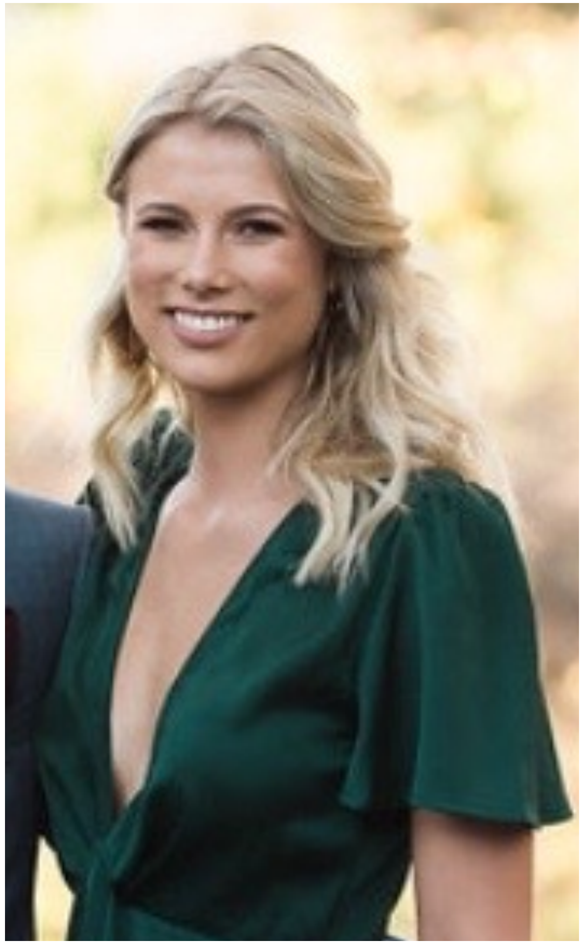
THE UNIVERSITY OF QUEENSLAND AUSTRALIA

RMIT University

Universität St.Gallen

UDACITY

Institute of Data



Case Context: Declining Revenue in Retail Banking

- Retail bank problem area:
  - Problem → Declining revenue
  - Strategized Solution → deploying direct telemarketing campaigns to promote long-term deposits
- How can I help?
  - Modeling the success of the marketing campaigns will be indispensable knowledge gained and add significant value:
    - ✓ Success rate of the campaign (present and future)
    - ✓ Which clients are most financially viable to target
    - ✓ Optimize marketing strategies and improve effectiveness

Time to Solve Some Problems!

Business Question

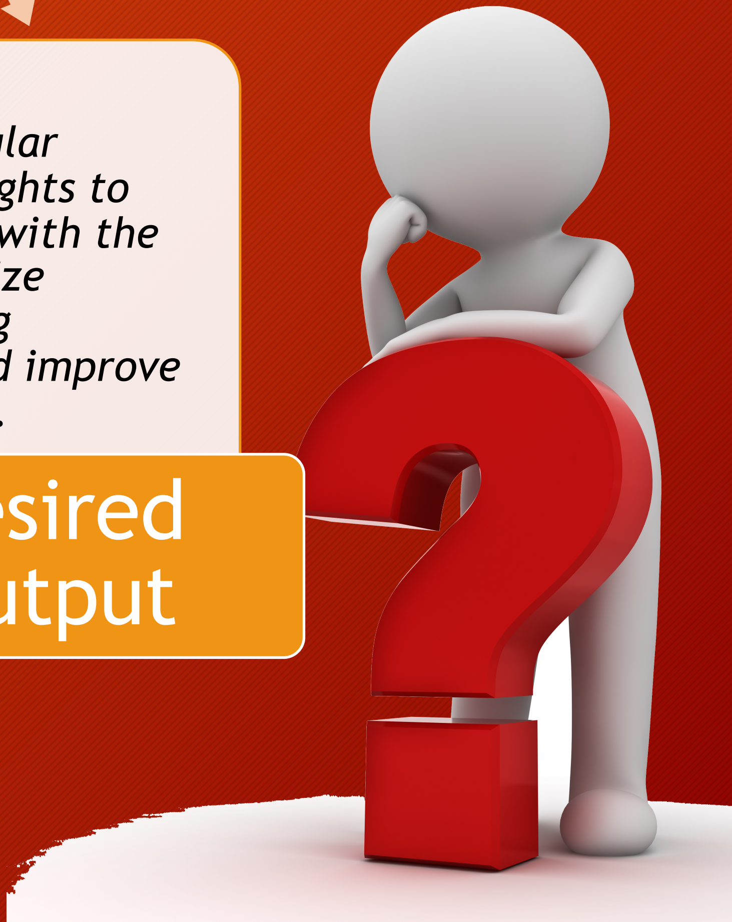
- *Can customers be targeted more effectively to increase the ratio of successful telemarketing phone calls?*

Data Question

- *Can a Machine Learning model predict which clients are most likely to successfully respond to telemarketing calls?*

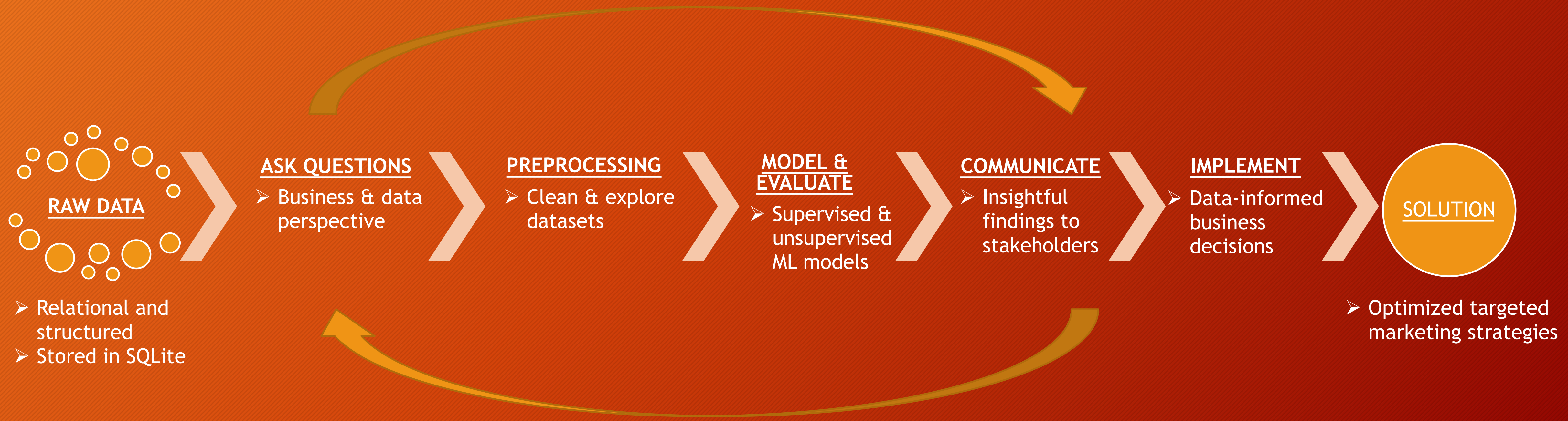
Desired Output

- *Provide granular customer insights to stakeholders with the aim to optimize telemarketing strategies and improve effectiveness.*

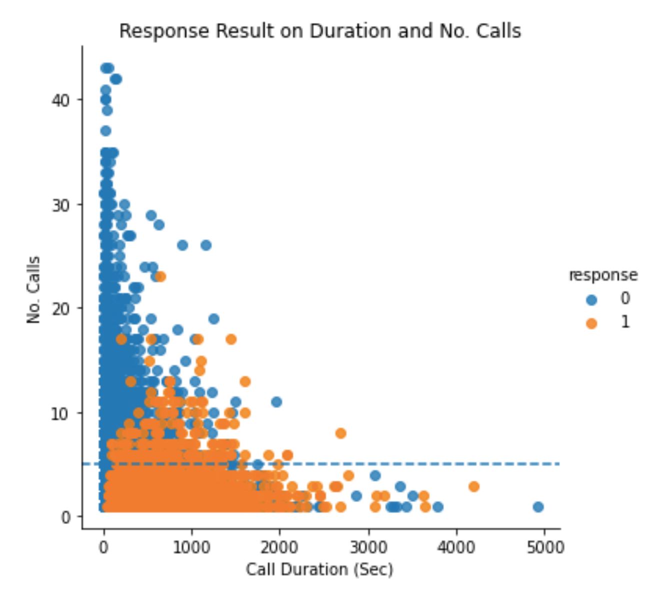


Exploring the Data: Data Science Process Flow

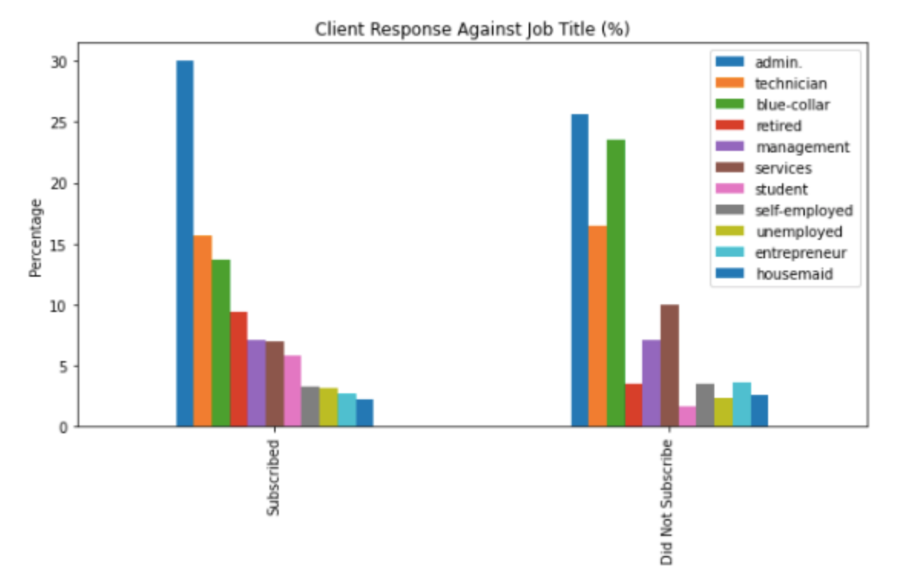
To be iterated in line with the business and data questions, via stakeholders



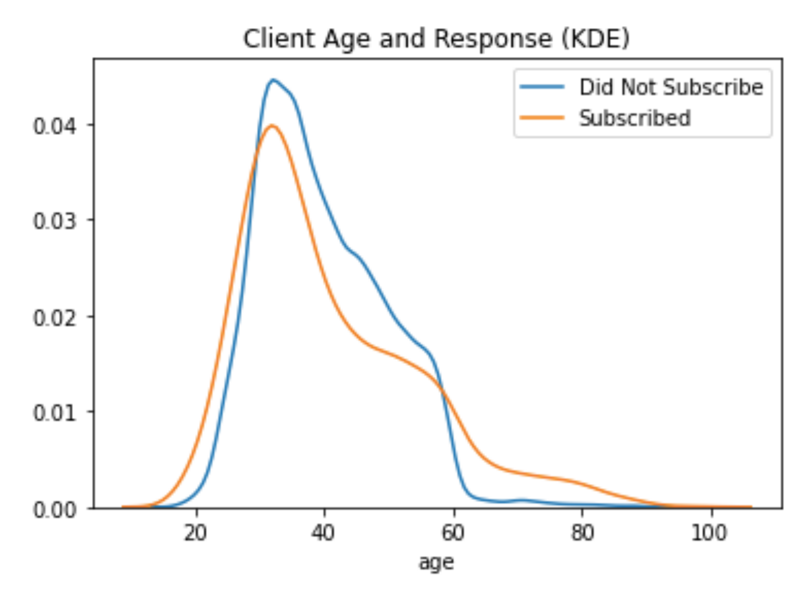
1.



2.



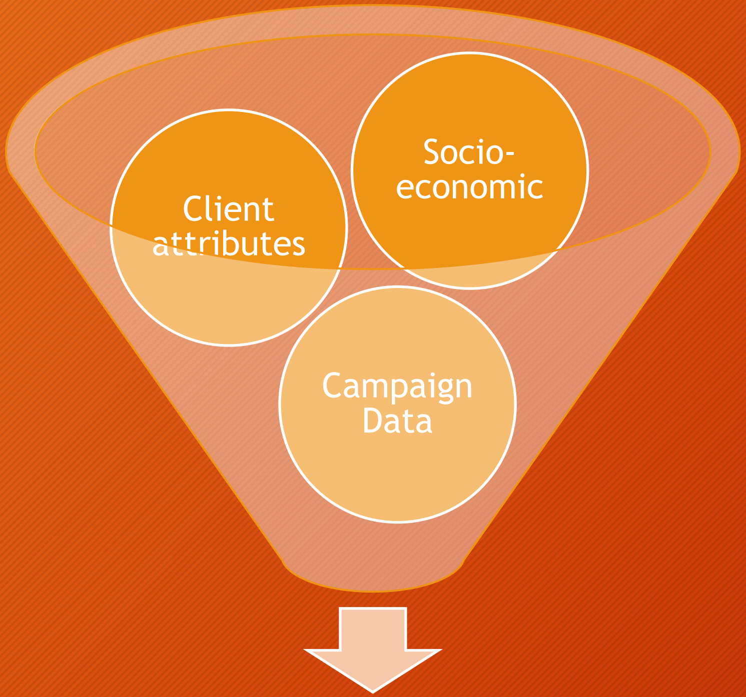
3.



Exploring the Data: Initial Findings

Delivering the Solution: Feature Engineering

Inputs: Predictor Variables



Output: Client Response (Binary Classification)

Preparing the Feature Vector:

1. To maintain a strong model, most features were used - only those that deemed inaccurate or incomplete were removed.
2. Categorical attributes had to be converted numerical by mapping binary variables or one-hot encoding nominal variables
3. Normalized scaler to standardize dummy variables with original numerical data
4. SMOTE: Synthetic resampling technique for class imbalance

Delivering the Solution: ML Models

Binary Classification Modelling (Supervised):

1. Logistic regression (With and without synthetic resampling)
   - Adapts so that binary target variable remains discrete
   - Logit function: $\boldsymbol{log}\left(\frac{\boldsymbol{p}}{\boldsymbol{1-p}}\right) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 X}$
   - Calculate predicted y: $\hat{\boldsymbol{y}} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 X}$
   - Calculate p (probability of True, where 0 < p > 1):

$$\boldsymbol{p} = \frac{\boldsymbol{e^{\hat{y}}}}{\boldsymbol{e^{\hat{y}} + 1}} = \frac{\boldsymbol{1}}{\boldsymbol{1 + e^{-\hat{y}}}}$$

2. Support Vector Machine
   - Concept based upon the mapping of an optimal hyper-plane for linearly separable classifiers
   - Function in Lagrangian form:

   Minimize: $J(w, b, a) = \frac{1}{2}w^T w - \sum_{i=1}^{N} a_i d_i (w^T x_i + b) + \sum_{i=1}^{N} a_i$

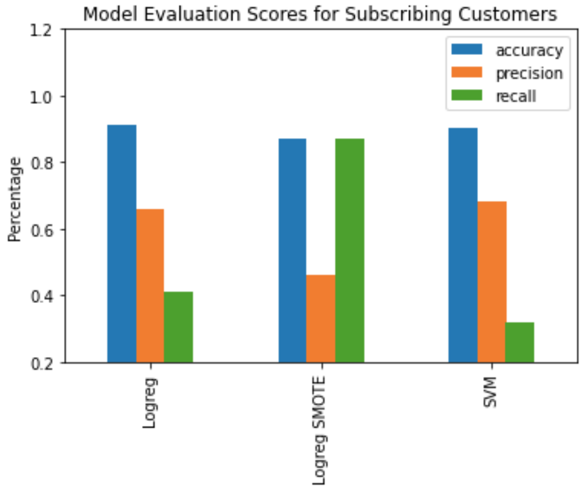
Clustering Challenge! (Unsupervised):

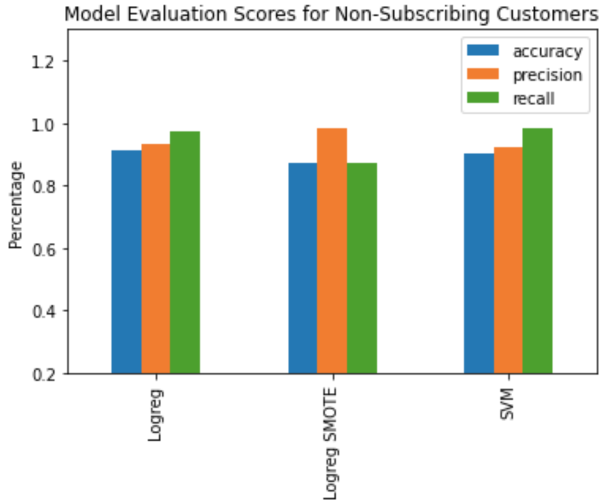
1. K-means
   - An iterative algorithm that attempts to partition the data into distinct subgroups
   - Objective function: $J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \parallel x^i - \mu k \parallel^2$
   - Will one cluster of client attributes be likely to respond more positively to the telemarketing campaign than another?

(1) Findings: Supervised ML

- Model Evaluation
  - Accuracy Score maintained consistency but deemed potentially unreliable on an unbalanced classifier dataset. However, accuracy of below 90% implies less risk of overfitting the data.
  - Logistic regression was comparatively insensitive in its detection of classes UNTIL the data was synthetically resampled (SMOTE) (From 41% recall to 87%!)
  - SVM had stronger precision but significantly low recall in detecting positive responses (subscribing clients)
- Conclusion:
  - With 87% accuracy (room for generalizability) and 87% recall (sensitivity in detecting positive instances) the model to most effectively predict the response of customers would be logistic regression on a synthetically resampled dataset.

	accuracy	precision	recall
Logreg	0.91	0.66	0.41
Logreg SMOTE	0.87	0.46	0.87
SVM	0.90	0.68	0.32





(2) Findings: Unsupervised ML

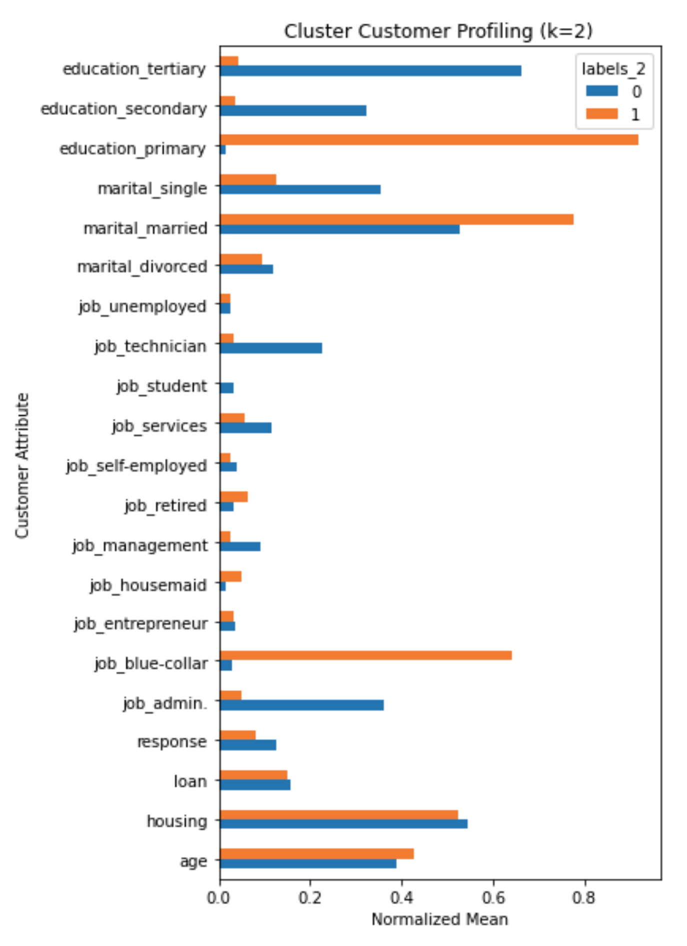
K-Means Customer Profiling (when optimal k=2)

CLUSTER 0

- More likely to subscribe
- Age: Less than 40
- Education: Tertiary
- Marital: Highest representation of singles
- Job: administration and technician roles

CLUSTER 1

- Less likely to subscribe
- Age: 40+
- Education: Base-level education
- Marital: Married
- Job: Blue-collar roles



Next Steps

Implementation of insights:

- Targeted campaigns to ensure optimized success rate based of:
  - Proficient predictive modelling
  - Clustered client attributes with increased successful response potential
  - Refined methods of contact (e.g. no more than 5 calls per client)
- Stakeholder value:
  - Optimized marketing strategies = more efficient targeted approach
  - Understanding customer needs = more effective campaigns, smarter product design and greater customer satisfaction

Future adaptations:

- Exploration of client attributes to business engagement
- Modelling monetary cost-benefit analysis across multiple models of interest
- Adjusting for limitations caused by a significantly imbalanced classifier, incompleteness of data and handling of multiple categorical variables

Any Questions?

References

- Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002, June). SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321-357.
- Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning, 20*, 273-297.
- Cramer, J. (2002). *The Origins of Logistic Regression.* Technical Report, Tinbergen Institute.
- Hopkinson, G., & Klarova, D. (2019, July). How Neobanks' Business Models Challenge Traditional Banks. *Young Graduate News*.
- Kim, K.-H., Lee, C.-S., Jo, S.-M., & Cho, S.-B. (2015, November). Predicting the success of bank telemarketing using deep convolutional neural network. *7th International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, 314-317.
- Landreth, O. (1992). Banco Comercial Português: Putting Marketing at the Forefront. In *European Corporate Strategy* (pp. 84-105). London, United Kingdom: Palgrave.
- Landreth, O. (1992). Banco Comercial Português: Putting Marketing at the Forefront. In *European Corporate Strategy.* London: Palgrave.
- Lee, S., & Lee, D. (2020). "Untact": a new customer service strategy in the digital age. *Service Business, 14*, 1-22.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 281-297.
- Moro, S., Cortez, P., & Rita, P. (2014, June). A data-driven approach to predict the success of bank telemarketing. *Science Direct, 62*, 22-31.
- Tang, H. (2014, December). A Comparison of Two Modeling Techniques in Customer Targeting For Bank Telemarketing. *Scholar Works*.
- Thorndike, R. (1953, December). Who belongs in the family? *Psychometrika, 18*, 267-276.
- UCI. (2014, June). *Bank Marketing Data Set.* Retrieved November 2020, from Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing