CAPSTONE OBJECTIVE: OPTIMIZING TARGETED TELEMARKETING

PROCESS OVERVEIW:

Problem Statement:

The current business state of a Portuguese banking institution, is one of fiscal instability, lost revenue and lowered long-term client engagement, in comparison to the last fiscal year (Moro, Cortez, & Rita, 2014). To maintain corporate competitiveness, the retail bank initiated a telemarketing campaign based on the notion that their clientele weren't depositing as frequently as they had been in the past (Moro, Cortez, & Rita, 2014). The marketing campaign was deployed via direct phone calls to promote the subscription of long-term deposits. With the desire to optimize output and reduce human and financial resources for the campaign, a study was established to determine which targeted clients would most likely respond positively to the telemarketing scheme and subscribe to the long-term deposits.

Distinguishing the success rate of the telemarketing phone calls is of high value as it highlights whether the retail bank in question is able to successfully reach consumers and sell them the campaign. More specifically, which clients are most financially viable to target (in terms of reduced opportunity cost and avoiding financing marketing ploys aimed at the wrong clientele). Marketing campaigns are a typical strategy implemented to enhance business engagement. By using this targeted marketing ploy, the retail banking firm is aiming to meet a desired business state of increased client long-term deposits and consequential financial stability.

Domain of Interest:

The banking industry is that in question, with strong focus on marketing operations. With this in mind, the retail banking sector is facing a significant amount of disruption as the digital age is exponentially changes business processes and client behaviour. Specifically, the shift towards mobile and online banking, away from on-site banking to reap benefits of trade (Lee & Lee, 2020). High competition is prevalent from disruptive start-ups and "neobanks" (Hopkinson & Klarova, 2019) that are changing service expectations - which is only further highlighting that efficiency gains from technology are of elevated value to the consumers.

These factors all motivate the need to adaptively strategize and implement new business processes and market various revenue-stimulating sectors, as is seen in this particular study (Moro, Cortez, & Rita, 2014). As multiple marketing schemes are implemented to adjust for this operational shift, the ability to successfully target clients is critical in achieving a competitive advantage.

Although the industry value chain in question centres around fiscal activities (wealth management, financial guidance and asset/liability administration ect.), this particular project is focused on the success rate of a telemarketing campaign. Due to the generalizability of such marketing schemes, this project is henceforth highly relevant to multiple industries facing the need to instigate competitive marketing operations.

Stakeholders:

The stakeholders in questions are in direct relation to retail banking. This covers the official Board of Directors, Board of Auditors, Advisory Board and shareholders of such institutions. Each stakeholder is impacted by the success rate of the marketing campaign, promoting long-term deposits. The higher the success rate, the higher the financial investment actions

of the banking entity (on products or services with a higher rate of return). Increasing the bank's investment capabilities, increases the services available to clients. For obvious reasons, shareholders will only benefit from increased revenue and stability of long-term deposits, and so forth. From a client perspective, long-term benefits include increased saving capabilities, decreased investment volatility, and increased interest rates.
In line with the project, it can be assumed that clientele wish to gain these competitive rates on their long-term deposits and shareholders will expect to benefit from increased financial flow and business activities to promote the economic success of their institutional foundations.

Business Question:
The main business question that requires attention is: *Can customers be targeted more effectively to increase the ratio of successful telemarketing phone calls?*
The aim is to deliver valuable customer insight to the stakeholders that will increase the chance of campaign success.
The required rate of accuracy should be very high as the consequences of incorrect predictability would result in the investment of unsuccessful marketing campaign, increased opportunity cost (missing out on better opportunities/client engagement), reduced revenue and investment capabilities, lowered financial certainty/ increased volatility.

Data Question:
From a data science perspective, the question we are attempting to answer is: *Can a Machine Learning model predict which clients are most-likely to successfully respond to telemarketing calls aimed at selling long-term deposits?*
The data required to answer this question involves:

- Survey data of the marketing campaign
- Attributes of the targeted clients from pre-existing data from the Portuguese retail bank in question
- Resulting response of the customer (successful or unsuccessful subscription to the long-term deposits)

Data:
The relational dataset was sourced from the UCI Machine Learning Repository (UCI, 2014), based on the aforementioned 2014 study (Moro, Cortez, & Rita, 2014). The data is related with direct marketing campaigns of the Portuguese banking institution, of which were stemmed from direct phone calls. Often, more than one contact to the same client was required, in order to access if the product in question (long-term deposit) would be subscribed ('yes') or not ('no'). There were originally four datasets in question, as follows:

1. bank-additional-full.csv with all examples (41,188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analysed in [Moro et al., 2014]
2. bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.
3. bank-full.csv with all examples (45,211) and 17 inputs, ordered by date (older version of this dataset with less inputs).
4. bank.csv with 10% of the examples (4521) and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).

The smallest datasets are provided to test more computationally demanding machine learning algorithms (e.g. SVM).

In regard to the reliability of the data, it was established within a highly cited and publicised research paper (Moro, Cortez, & Rita, 2014). The original collection was released publicly for future research and educational purposes, as well as an adaption by the Banco de Portugal that contained revised attributes to accommodate for socio economic factors. Coinciding with the fact that that the collection and study of the data acquired spanned from 2008 to 2014, I can contest that the source is very reliable. With respect to the raw data, the completeness and quality of the figures is varied. There are significant incidents that have unknown variables, as would be expected in the case of telemarketing data. In hindsight, a lot of wrangling will have to be done to get the dataset to a stage that a predictive model can be successfully built upon.

The original data was generated for the founding journal (Moro, Cortez, & Rita, 2014) in the following stages:

1. A Portuguese retail bank was addressed, with data collected from 2008 to 2013 (note the inclusion of the financial crisis).
2. A significant set of 150 features was analysed related with bank client, product and social-economic attributes.
3. A semi-automatic feature selection was initiated in the modelling phase within the study, resulting in a reduced set of 21 attributes.

Although this targeted data was publicised for future studies, the exact figures are not available on an ongoing basis. In the context of the specialized curation of the data with reference to the retail banking case study, this may present as a limitation as future comparisons will not be easily made. However, due to the relatability of the data in correspondence to a multitude of industries, the potential for replication to some degree is implied.

DATA SCIENCE PROCESS

Process Flow

The data science process flow encompasses core concepts of data collection from the UCI, storing it in an appropriate manner (SQLite in this case due to the structured, relational nature of the data obtained) pre-processing and exploration, unsupervised and unsupervised modelling, evaluation across appropriate metrics and, finally, suggested implementation of the insightful findings by stakeholders. The flow amongst these processes is not linear and in practice will re-iterate until the satisfactory condition (goal, desired output or meaningful insight) is met.

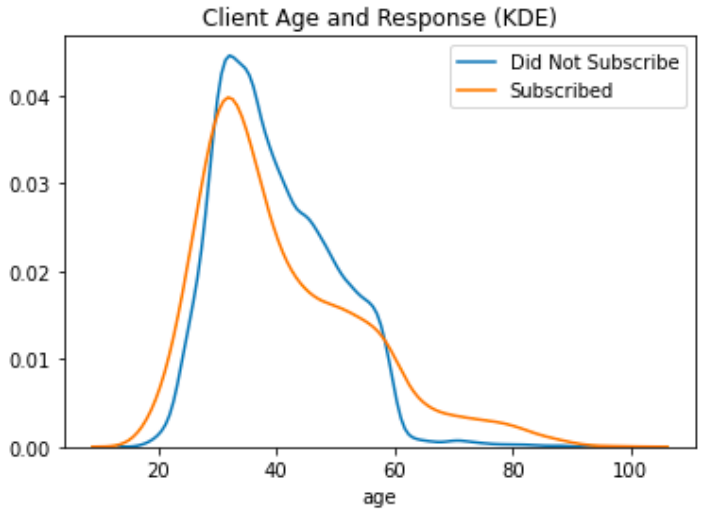
The data pipeline that was used to initially wrangle the raw data consisted of the following:

- Removing nulls: substituting “unknown” incidents to NaNs, removing or replacing them with the mode (categorical incidences) and/or dropping columns with significant missing data.
- Grouping of data for clearer categorisation and visualisation (attributes such as customer age and educative background)

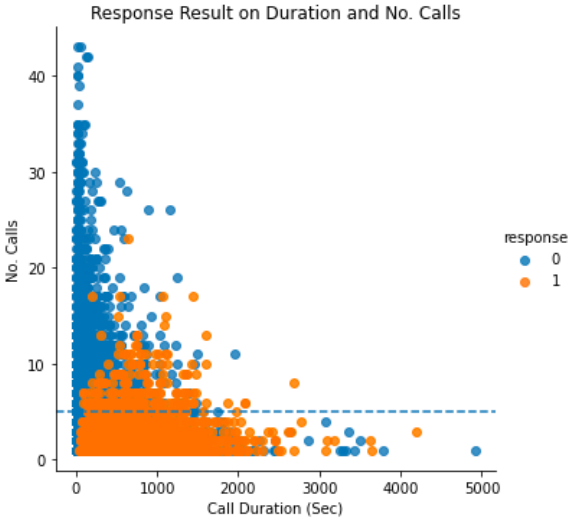
Data Analysis:

Three noteworthy insights:

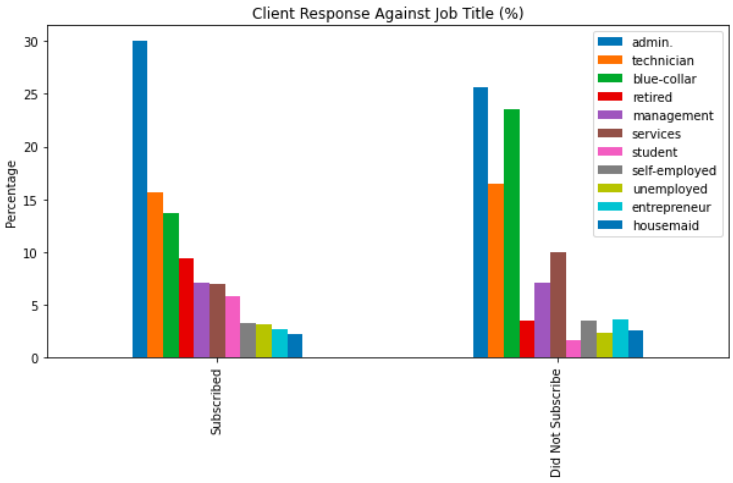
1. After categorising client ages into bins, the group with the highest number (count) of customers that responded positively to the telemarketing campaign were in their "30s". However, once these figures were normalized, the finding actually showed that "20s" were more likely to subscribe, as well as those in their "60s" and "70s+". This was also efficiently shown in the KDE plot, distinguishing the response across the age groups in smooth manner.



2. Response result on duration and number of calls: Higher subscription rate when the client received less than 5 telemarketing calls. In other words, clients were most likely to reject the campaign if they received more than 5 calls. The plot significantly distinguishes the clients which subscribed to the long-term deposit from the clients who did not. Clients who subscribed to the deposit where contacted fewer times and had longer call duration



3. Response against client occupation, as a normalized proportion: clients that worked in admin had the highest percentage of subscription rates. Retired clients were also more likely to subscribe. Blues collar workers, on the other hand, saw a noticeably bigger segment that did not subscribe. Although not as extreme, this same observation can be made for those working in the service industry.



Modelling:

Feature Selection:

Using the client response as the target variable was self-explanatory as it is the key component to the dataset problem. The basis was to predict whether the response outcome of clients (whether they subscribed to the long-term deposits or not) could be modelled using various predictor variables. It is also binary, which prompts the classification analysis.

As correlation wasn't as significant as expected in regard to informative variables, predictive features were considered all-inclusive unless there were high unknowns or misleading values. Such attributes were deemed incomplete and removed. When considering an unsupervised form of machine learning, K-means, feature selection was more targeted, and a more precise predictor matrix was used. In this case, only attributes that were directly related to the client were obtained to produce potentially insightful clusters.

Feature Engineering:

A few key stages were implemented in the feature engineering process. As the data was predominately categorical, the relevant attributes had to be transformed to numerical data for modelling. This was initially achieved with a combination of mapping (for binary attributes) as well as one-hot encoding using *pandas.get_dummies* for nominal variables. Following this the predictor matrix had to be standardized using *StandardScaler* to ensure the original data is scaled alongside the newly encoded attributes. If the data was not to be standardized, features with greater numerical values would misleadingly have greater influence on the predictive models than those that are numerically smaller.

Another feature engineering method that was implemented throughout the modelling process was *SMOTE*, or Synthetic Minority Over-Sampling Technique (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). It was apparent that the data is extremely biased with regards to the resultant client subscribers (target variable) - only 11.3% of clients subscribed to the long-term deposit. To account for the class imbalance, it is necessary to implement a resampling technique on the training dataset. When considering imbalanced classification data, most ML models will ignore or behave poorly on the minority class. This is a critical point as in this case, the minority class (client subscription) is the most desired outcome. SMOTE (Synthetic Minority Oversampling Technique) is such a form of data augmentation that addresses this issue by oversampling the minority class.

Models:

The base supervised machine learning model used was logistic regression (Cramer, 2002). As the target variable of interest is binary and categorical in nature, linear regression would not be a feasible model as predictions can fall external to the discrete output (1 or 0)**.** As a model, logistic regression adapts so that the predicted target variable remains discrete. Logistic regression was run and compared twice, once with SMOTE and once without.

Support Vector Machine, SVM, was implemented as a more modern machine learning model for binary classification problems (Cortes & Vapnik, 1995). The concept of SVM is based upon the mapping of a hyper-plane that differentiates two classifiers, such as the customer response in this case.

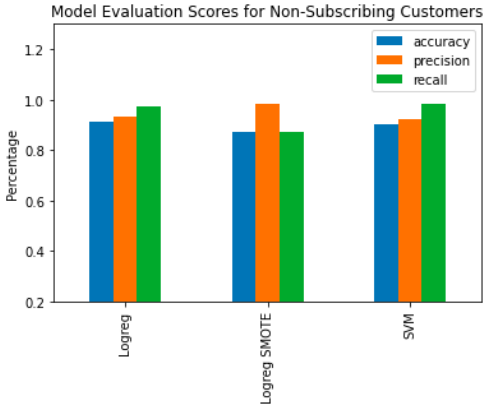
With regards to an unsupervised ML challenge, I wanted to see whether one cluster of client attributes be likely to respond more positively to the telemarketing campaign than another. To solve this, an iterative algorithm by the name of K-Means (MacQueen, 1967) was used to

partition the targeted data into distinct sub-groups and find potentially insightful customer attributes.

Outcomes:

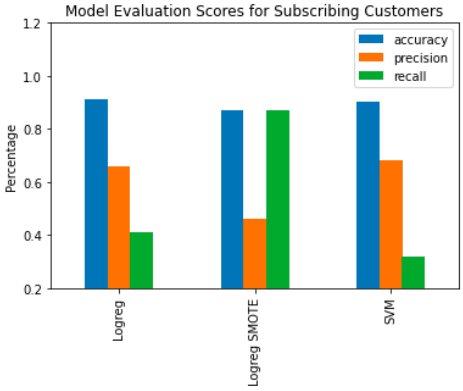
Supervised ML:

The comparative evaluation metrics used for the supervised learning models involved an accuracy score (informally, the fraction of predictions the model got right), recall (how sensitive the classifier model is at detecting positive instances), precision (the proportion of correct positive identifications) as well as a Receiver Operating Characteristic (ROC) curve (specificity vs sensitivity).



The numerical results were as follows:

- Logreg:
  - Accuracy: train 91.19%, test 90.86%
  - Recall: didn't subscribe (0) 97%, subscribed (1) 41%
  - Precision: didn't subscribe (0) 93%, subscribed (1) 66%
  - ROC curve area: train 94%, test 93%
- Logreg with SMOTE:
  - Accuracy: train test 86.75%
  - Recall: didn't subscribe (0) 87%, subscribed (1) 87%
  - Precision: didn't subscribe (0) 98%, subscribed (1) 46%
  - ROC curve area: train 94%, test 94%
- SVM:
  - Accuracy: train 90.51%, test 90.47%
  - Recall: didn't subscribe (0) 98%, subscribed (1) 32%
  - Precision: didn't subscribe (0) 92%, subscribed (1) 68%



Whilst the evaluation metrics saw little divergence across the model for non-subscribing clients, there were significant differences across the models for subscribing customers. Accuracy Score maintained consistency but deemed potentially unreliable on an unbalanced classifier dataset. However, accuracy of below 90% (as detected in the resampled logistic regression model) implies less risk of overfitting the data (increasing generalizability). Logistic regression was comparatively insensitive in its detection of classes until the data was synthetically resampled (SMOTE) - from 41% recall to 87%! Although the precision score is low for the resampled logistic regression model (relatively speaking) for the subscribing clients, it was the highest in the non-subscribing clients. SVM, on the other hand, had stronger precision but significantly low recall in detecting positive responses (subscribing clients). These figures suggest that the model to most aptly predict the responses of clients would be logistic regression model on a synthetically resampled dataset to rectify the imbalanced classifier.

Unsupervised ML:
The K-Means unsupervised clustering challenge brought to light some insightful findings. The aim was to cluster the data to determine whether a group is more likely to respond to the telemarketing calls in a certain way. To find the optimal number of clusters, 'k', the elbow method (Thorndike, 1953) was used, but provided too subtle to use as concrete evidence. As the target is binary, two clusters were initially established and then increased from there until meaningful insights were drawn. When k = 2, of which was the most insightful, it was evident that Cluster 0 was more likely to subscribe to the long-term deposits. The customer attributes of this cluster were younger than 40 years of age, had a tertiary level of education, accounted for the largest proportion of singles and had administration/ technician/ service roles. Cluster 1 contrastingly was less likely to subscribe, above the age of 40, married, in a "blue-collar" job role and a base level of education.

Implementation:
The main objective of this project was to provide insights to retail banking stakeholders with the aim to optimize marketing strategies and improve effectiveness. This project should enable retail banks to develop a more granular understanding of their client base and predict their customers response to their telemarketing campaign. This is done by analysing client features to identify which group is more likely to respond positively to the campaign and subscribe to long-term deposits. Understanding customer clustering leads to more effective campaigns, informed product design and greater overall client satisfaction.
As a response to stakeholders, implementing targeted campaigns based off predictive modelling, clustered client attributes with increased positive response potential and refined methods of contact (e.g. ensuring no more than five calls are made to each client to increase success rate), will ensure operational effectiveness can be optimised.

With direct reference to the business question and, consequently, the data question, they were both successfully answered. By implementing supervised and unsupervised machine learning models, it was established that with a combination of logistic regression (with the aid of a synthetic resampling technique) and K-means clustering, I was able to predict whether clients were going to respond positively to a telemarketing campaign with 86.75% accuracy along with the corresponding attributes of those more likely to subscribe (less than 40 years old, higher education, admin roles). In response, stakeholders can more effectively target their telemarketing attempts to increase the ratio of successful responses.

Limitations and Future Adaptations:
The confidence of the predictive modelling can be improved by adjusting the following limitations:

- A significantly imbalanced classifier. Although this was synthetically corrected, a bias far less than 11.3% in favour to an unsuccessful subscription is not ideal for analysis.
- Handling of multiple categorical variables was significantly challenging when visualisation clustering and decision boundaries post-modelling due to one-hot encoding. Ideally, a clustering visualisation to observe the coordinates of the responding centroids would have been constructed to aid in client attribute hypothesising.
- Incompleteness of data was an issue. Significant amount of unknowns resulted in a lot of cleaning, removing and/or replacing with mode values of interest.

As a continuation, further exploration of client attributes with reference to business engagement (which retail banking services do they usually engage in, how long have they been loyal to the business, how many services do they use ect.) would've been insightful. More informative data from previous campaigns could have also been sourced to enhance predictions and business implementations. Such valuable data could include reasons why clients responded a certain way in the past (ratings, reasoning, contact experience). Implementing more models of interest to further compare or investigate the business problem could include classification models such as neural networks, Bayesian networks and random forest. Improvement to quantifying stakeholder gain could also be achieved by constructing a predictive monetary cost-benefit analysis, comparing cases before and after data-informed targeted telemarketing campaigns. Profit gained or lost with response to the success of predictive modelling would aptly concrete which model would yield the most operational interest for stakeholders.

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