Background:

Marketing to a specific audience is what drives customers to choose one bank over another. With the evolution of technology, competition is exponentially increasing, making it more challenging to attract customers to subscribe to new products and services. The bank's marketing team has orchestrated campaigns to encourage customers to subscribe to their term deposit programs. However, attracting and retaining customer subscriptions can be costly and labour-intensive. One way to reduce these expenses is by focusing promotions on potential customers who are likely to open a deposit account. These target customers will be referred to in this proposal as high-value customers. This raises the question: how can the marketing team accurately identify the demographic to market to for the most cost-effective outcome?

Therefore, this proposal aims to discuss using machine learning techniques to analyse the bank's data and classify high-value customers.

The potential benefits of conducting this project include:

- Increased profitability and stable cash flows by targeting high-value customers with tailored offers and promotions, the bank can achieve larger and more stable cash flows. Attracting customers who are more likely to make significant deposits and maintain long-term relationships will drive higher revenues and contribute to a more predictable financial outlook.
- Increased customer retention rate by classifying high-engagement customers based on behaviours and preferences, the bank can develop strategies to increase customer loyalty and reduce churn. Proactively addressing the needs of identified high-value customers and offering incentives that align with their preferences will strengthen customer relationships and increase the lifetime value of each customer.
- Attract more cost-effective customers acquisition by marketing to individuals who share similar characteristics with current subscribers, the bank can reduce acquisition costs while maximising the impact of its marketing efforts. This data-driven approach ensures that marketing resources are focused on the most promising prospects.
- Gain a competitive advantage by leveraging data-driven insights to continuously improve and adapt products and services ensures that the bank stays ahead of its competitors. By being more responsive to market, social, and economic trends, as well as evolving customer needs, the bank can differentiate itself in the marketplace, attract more customers, and secure a stronger industry position.
- Enhanced operational efficiency by targeting a specific demographic based on data analysis can streamline operations, reduce costs associated with customer acquisition, and improve product and service delivery. This focus on high-value customers allows the bank to optimise resources and ensure that marketing and service efforts are both effective and efficient.

The insights expected from this project include:

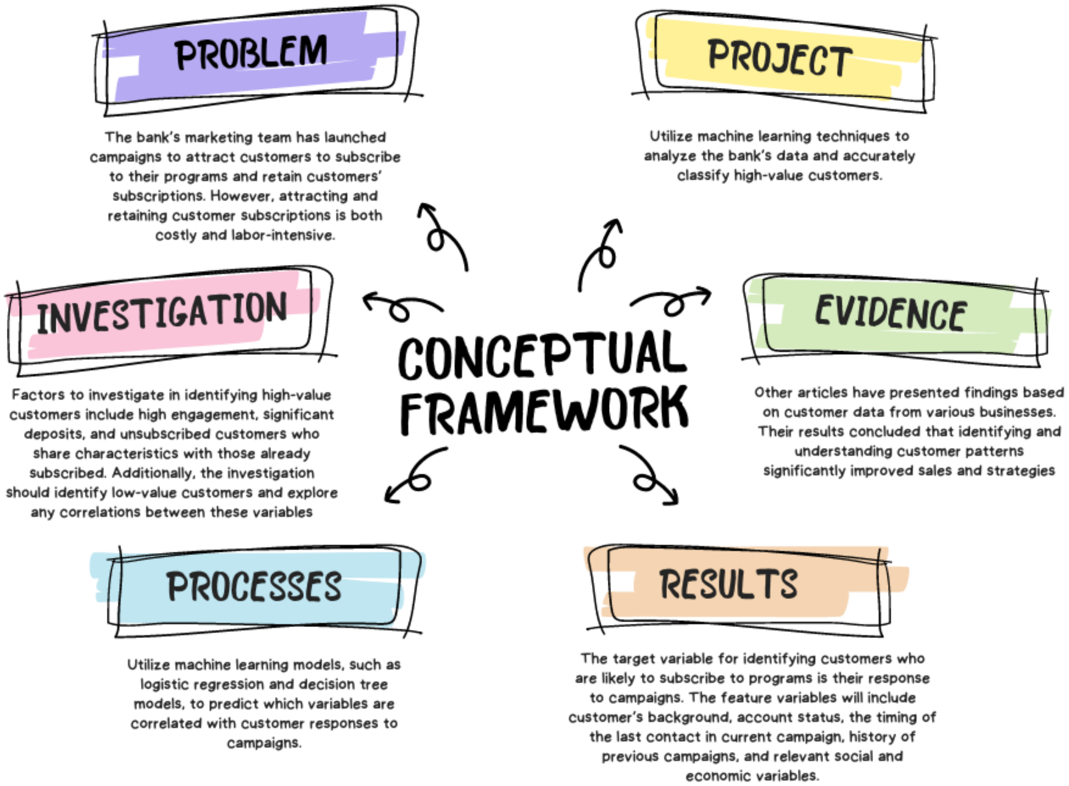
- Identify the characteristics of customers who have high engagement with campaigns. Key features to help identify include the number of contacts during a campaign, the number of contacts from previous campaigns, the duration of contact with the bank, and the timing of the last contact with the customer.
- Identify unsubscribed customers who share similar characteristics with those who have subscribed. Relevant features may include age, occupation, marital status, education level, credit history, and whether they have a housing or personal loan.
- Identify the customers with significant deposits by analysing social and economic factors such as the employment variation rate, consumer price index, consumer confidence index, and the Euribor 3-month rate, the bank can identify customers with significant deposits.
- Differentiate between high-value and low-value customers. High-value customers are those who significantly contribute to the bank's revenue, either through larger deposits or sustained long-term engagement. By identifying the characteristics and behaviours of these customers, the bank can allocate resources more efficiently, focusing on retaining and nurturing high-value customers while also developing strategies to convert low-value customers into high-value ones.

In addition to the benefits and insights gained from this project, the knowledge acquired will empower the bank to make more informed and strategic decisions, including:

- Refined its marketing strategies to create cost effective marketing by understanding who the potential customers are allows the bank to create cost-effective marketing strategies. This precise targeting minimises wasted resources on broad, ineffective campaigns and ensures that marketing efforts are focused on the most promising leads, leading to a more efficient allocation of the marketing budget.
- Insights into customer characteristics and behaviours will guide the development of new banking products and services that better meet the needs of high-value customers. This could involve introducing new term deposit options, personalised financial advice, or loyalty programs that enhance customer satisfaction and retention.
- By addressing inefficiencies in current marketing operations and customer engagement processes, the bank can allocate resources more effectively. Understanding customer behaviours, such as response rates and number of contacts, ensures that staff time and marketing budgets are spent on activities that deliver the highest return on investment.

Conceptual Framework

Breakdown of the project's concepts:



CONCEPTUAL FRAMEWORK

PROBLEM

The bank's marketing team has launched campaigns to attract customers to subscribe to their programs and retain customers' subscriptions. However, attracting and retaining customer subscriptions is both costly and labor-intensive.

PROJECT

Utilize machine learning techniques to analyze the bank's data and accurately classify high-value customers.

INVESTIGATION

Factors to investigate in identifying high-value customers include high engagement, significant deposits, and unsubscribed customers who share characteristics with those already subscribed. Additionally, the investigation should identify low-value customers and explore any correlations between these variables

EVIDENCE

Other articles have presented findings based on customer data from various businesses. Their results concluded that identifying and understanding customer patterns significantly improved sales and strategies

PROCESSES

Utilize machine learning models, such as logistic regression and decision tree models, to predict which variables are correlated with customer responses to campaigns.

RESULTS

The target variable for identifying customers who are likely to subscribe to programs is their response to campaigns. The feature variables will include customer's background, account status, the timing of the last contact in current campaign, history of previous campaigns, and relevant social and economic variables.

Based on this framework, three key goals have been established as criteria for completing this project:

- Accurately segment customers into high-value and low-value groups based on their characteristics and behaviours.
- Determine which variables significantly influence a customer's likelihood to respond to marketing campaigns.

Similar investigations have been conducted using machine learning to analyse and predict the effectiveness of various business services.

A study by Samer Nofal from the German Jordanian University applied machine learning predictive models to identify high-value bank customers with current accounts based on their transactions (Nofal, 2023). This study utilised clustering algorithms, neural networks, support vector machines, and decision trees (Nofal, 2023). The models achieved an average accuracy of 97%, leading researchers to conclude that these methods effectively segmented high-value customers (Nofal, 2023).

The Vellore Institute of Technology conducted an analysis to identify customer churn using machine learning algorithms (Prabadevi, Shalini, & Kavitha, 2022). This study tested algorithms such as stochastic gradient boosting, random forest, logistic regression, and k-nearest neighbours, achieving respective accuracies of 83.9%, 82.6%, and 78.1% (Prabadevi, Shalini, & Kavitha, 2022). The study concluded that machine learning effectively classifies customer churn in the banking sector (Prabadevi, Shalini, & Kavitha, 2022).

Researchers from the University of Granada published an article on leveraging machine learning to identify changes in customer behaviour by classifying bank customer data (Santiago, Pedro, & Francisco, 2020). The study employed random forest and causal forest models to segment the data (Santiago, Pedro, & Francisco, 2020). The random forest model achieved an accuracy of 88.41%, suggesting that banks should address changes in customer behaviour by offering personalised digital services (Santiago, Pedro, & Francisco, 2020).

The findings from these studies indicate that machine learning can lead to insightful and accurate analyses. Therefore, this proposal should be considered, as it suggests that the application of machine learning would provide valuable insights and accurate customer segmentation.

Variable Selection

The target variable is a categorical binary variable representing customer responses to the latest campaign—whether they opened a new account or not. This variable is chosen because it reflects the likelihood of customer engagement with campaigns. The variables "Contact" and "Nr_employed" will not be examined as they are deemed irrelevant to the project's objectives. The 20 feature variables include a mix of categorical and numeric data, allowing for a comprehensive analysis of factors influencing customer behaviour.

BANK CLIENT DATA:

VARIABLE	TYPE	IMPORTANCE
AGE	Numeric	Age of the customers is important to keep track of because it would help differentiate the high value customers and the low value customers.
JOB	Categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown'	Job is the customer's current occupation. This is important to investigate because it would help differentiate the high value customers and the low value customers.
MARTIAL	Categorical: 'divorced', 'married', 'single', 'unknown' ; note: 'divorced' means divorced or widowed	Customer's marital status is important to investigate because it would help differentiate the high value customers and the low value customers.
EDUCATION	Categorical: : 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown'	Customer's education is important to investigate because it would help differentiate the high value customers and the low value customers.

THE NATURE AND STATUS OF THEIR EXISTING ACCOUNTS WITH THE BANK:

VARIABLE	TYPE	IMPORTANCE
DEFAULT	Categorical: 'no', 'yes', 'unknown'	Defaults indicates whether the customer has credit in default. Considering the customer's account status is important to this investigation because it would differentiate the high value customers and the low value customers.
HOUSING	Categorical: 'no', 'yes', 'unknown'	Housing refers to the customers having a housing loan. Considering the customer's account status is important to this investigation because it would differentiate the high value customers and the low value customers.
LOAN	Categorical: 'no', 'yes', 'unknown'	Loan defines whether the customers have a personal loan. Considering the customer's account status is important to this investigation because it would differentiate the high value customers and the low value customers.

RELATED WITH THE LAST CONTACT OF THE CURRENT CAMPAIGN:

VARIABLE	TYPE	IMPORTANCE
CONTACT	Categorical: 'cellular', 'telephone'	The type of communication the customers use to contact bank to subscribe to programs. This variable is not important to consider because it shows the form of communication use to communicate with banks.
MONTH	Categorical: 'jan', 'feb', 'mar', ‘may’, ‘apr’, ‘jun’, ‘jul’, ‘aug’, ‘sep’, ‘oct’, 'nov', 'dec'	The last contact month from the customer to engaged with the campaign. This variable is important because it shows which month customers are more active.
DAY_OF_WEEK	categorical: 'mon', 'tue', 'wed', 'thu', 'fri'	The last contact day of the week from the customer to engage with campaign. This variable is important because it shows which month customers are more active.
DURATION	numeric in seconds	The duration of the last contact with the customer. This variable is important because it shows the amount of time the customer engaged with the campaign. This can identify the potential customers who are more likely to subscribed to programs.

OTHER ATTRIBUTES:

VARIABLE	TYPE	IMPORTANCE
CAMPAIGN	numeric, includes last contact	Campaign is the number of contacts performed during this campaign and for this customer. This variable is important because it shows how frequent the customer engaged with this campaign. This can identify the potential customers who are more likely to subscribed to programs.
PDAYS	numeric; 999 means client was not previously contacted	Pdays is the number of days that passed by after the client was last contacted from a previous campaign. This variable is important because it shows the activeness of the customers engaging with future campaigns. This can identify the potential customers who are more likely to subscribed to programs.
PREVIOUS	numeric	Previous is the number of contacts performed before this campaign and for this client. This variable is important because it shows the activeness of the customers engaging with past campaigns. This can identify the potential customers who are more likely to subscribed to programs.
POUTCOME	categorical: 'failure', 'nonexistent', 'success'	Outcome of the previous marketing campaign. This variable is important because it shows the likelihood of this customer subscribing to programs from engaging with promotions. This can identify the potential customers who are more likely to subscribed to programs from being influenced by campaigns.

SOCIAL AND ECONOMIC CONTEXT ATTRIBUTES:

VARIABLE	TYPE	IMPORTANCE
EMP_VAR_RATE	numeric	The customer’s employment variation rate (quarterly indicator). This is important to investigate because it would help differentiate the high value customers and the low value customers.
CONS_PRICE_IDX	numeric	The customer’s consumer price index (monthly indicator). This is important to investigate because it would help differentiate the high value customers and the low value customers.
CONS_CONF_IDX	numeric	The customer’s consumer confidence index (monthly indicator). This is important to investigate because it would help differentiate the high value customers and the low value customers.
EURIBOR3M	numeric	Customer’s euribor 3 month rate (daily indicator). This is important to investigate because it would help differentiate the high value customers and the low value customers.
NR_EMPLOYED	numeric	The number of employees (quarterly indicator). This is not important to investigate because this shows show the current number of employees working at the bank.

Methods of analysis /Analysis Plan

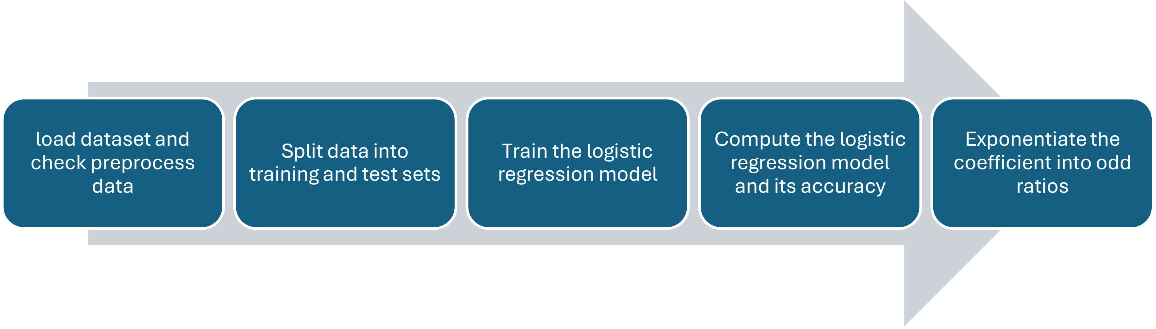
The models used to analyse our dataset include logistic regression and the decision tree model, specifically Classification and Regression Trees (CART). These models are appropriate for our analysis because logistic regression effectively constrains predictions to the interval [0, 1], making it suitable for binary classification, while decision trees are adept at predicting the most likely outcome based on feature variables for each observation in the binary categorical target variable. Both models share similar data requirements: a categorical binary target variable and a mix of categorical and numerical feature variables.

Logistic Regression

To validate the accuracy of the logistic regression model, these assumptions were established:

- Each observation is independent of the others, implying no correlation between any feature variables or residuals across observations.
- The dependent variable is binary or dichotomous, meaning it can take only two possible outcomes.
- A linear relationship is assumed between the features and the log odds of the target variable. This assumption allows logistic regression to model the probability of the outcome effectively.
- It is assumed that the independent variables are not highly correlated with each other. High multicollinearity can lead to unreliable estimates of regression coefficients and affect the model's performance.
- The errors (residuals) are assumed to be independent of each other, indicating no autocorrelation in the residuals.
- The model assumes that there are no significant outliers that could disproportionately influence the estimates.
- The dataset is sufficiently large since the model generally requires a large sample size to provide reliable estimates and ensure the stability of the model.

The key steps to perform a logistic regression:



load dataset and check preprocess data

Split data into training and test sets

Train the logistic regression model

Compute the logistic regression model and its accuracy

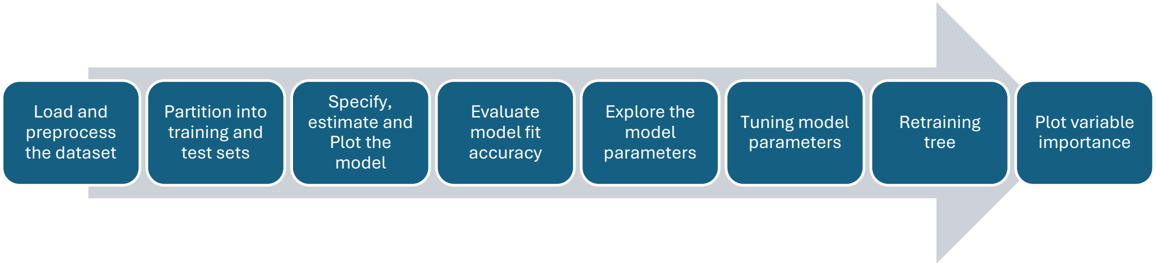
Exponentiate the coefficient into odd ratios

Decision Trees

The following assumptions were established to validate the decision trees' accuracy:

- The model assumes that each observation in the dataset is independent, implying there is no correlation between observations.
- Partitioning the feature variables can capture nonlinear relationships between the features and the target variable. This flexibility allows the decision tree to model complex interactions between variables.
- The dataset is a large sample size because it is crucial to create meaningful splits and avoid overfitting. Small datasets may lead to overfitting, resulting in less reliable predictions.

The key steps to perform a decision trees model:



Load and preprocess the dataset

Partition into training and test sets

Specify, estimate and Plot the model

Evaluate model fit accuracy

Explore the model parameters

Tuning model parameters

Retraining tree

Plot variable importance

Form of results

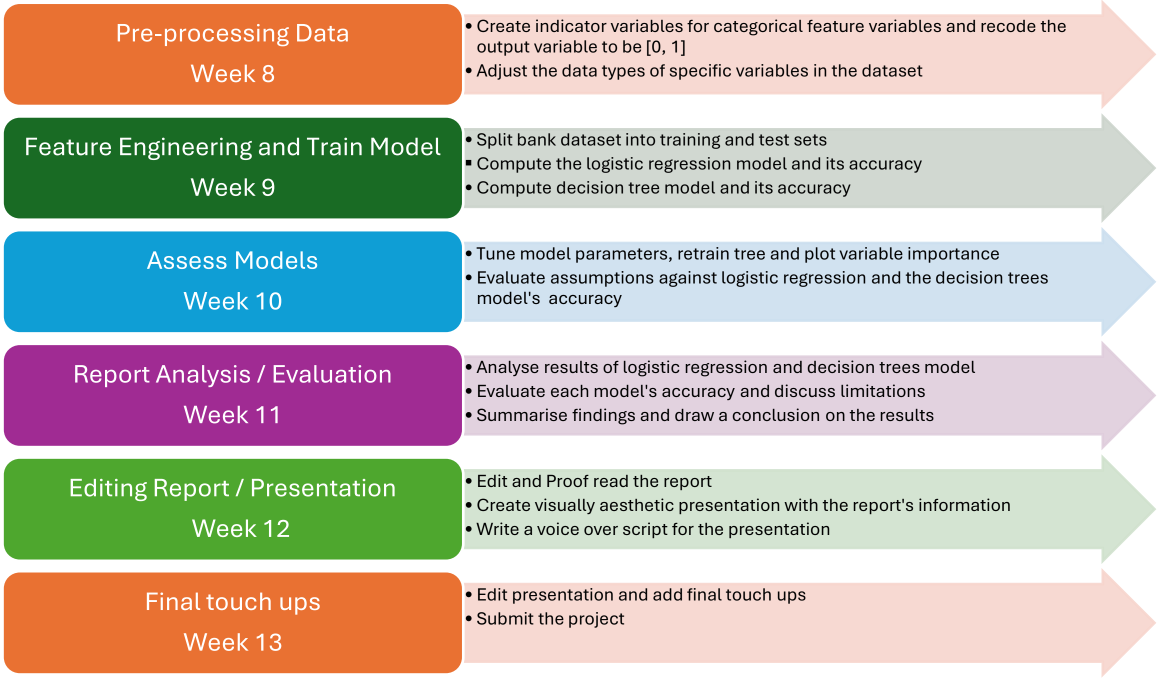
Based on the variables, the logistic regression and decision trees will potentially be the following outcomes (variables for decision trees were rank from most correlated to least):

Variable	Significance with relation to Responses	Coefficients
Age	p<0.05	Low positive
Job	p<0.05	Low positive
Martial	p>0.05	Low negative
Education	p>0.05	Low negative
Default	p>0.05	Low negative
Housing	p>0.05	Low negative
Loan	p>0.05	Low negative
Month	p>0.05	Low negative
Day of the week	p>0.05	Low negative
Duration	p<0.05	High positive
Campaign	p<0.05	High positive
Pdays	p<0.05	High positive
Previous	p<0.05	High positive
Poutcomes	p<0.05	High positive
Emp_var_rate	p<0.05	High positive
Cons_price_idx	p<0.05	High positive
Cons_conf_idx	p<0.05	High positive
Euribor3m	p<0.05	High positive

Variable	**Yes**	**No**
Duration	Greater than 200	Less than 200
Euribor3m	Less than 1	Greater than 1
Cons_conf_idx	Less than 30	Greater than 30
Emp_var_rate	Less than 0	Greater than 0
Cons_price_idx	Less than 93	Greater than 93
Poutcomes	Success	Failure
Pdays	Less than 30 days	Not contacted
Campaign	Greater than 2	Less than 2
Previous	No difference	No difference
Month	No difference	No difference
Age	No difference	No difference
Job	No difference	No difference
Martial	No difference	No difference
Education	No difference	No difference
Default	No difference	No difference
Housing	No difference	No difference
Loan	No difference	No difference
Month	No difference	No difference
Day of the week	No difference	No difference

This proposal outlines the rationale and process for using machine learning techniques to investigate the bank's data. Completing this project will offer several benefits, including more cost-effective customer acquisition, a competitive advantage, enhanced operational efficiency, increased profitability, and improved customer retention. These advantages collectively contribute to enhancing the bank’s overall value. By successfully identifying customers who are most likely to engage, the bank can maximise profits while minimising costs.

Below is a timeline summarising the key activities and actions involved in this project:



Pre-processing Data
Week 8

- Create indicator variables for categorical feature variables and recode the output variable to be [0, 1]
- Adjust the data types of specific variables in the dataset

Feature Engineering and Train Model
Week 9

- Split bank dataset into training and test sets
- Compute the logistic regression model and its accuracy
- Compute decision tree model and its accuracy

Assess Models
Week 10

- Tune model parameters, retrain tree and plot variable importance
- Evaluate assumptions against logistic regression and the decision trees model's accuracy

Report Analysis / Evaluation
Week 11

- Analyse results of logistic regression and decision trees model
- Evaluate each model's accuracy and discuss limitations
- Summarise findings and draw a conclusion on the results

Editing Report / Presentation
Week 12

- Edit and Proof read the report
- Create visually aesthetic presentation with the report's information
- Write a voice over script for the presentation

Final touch ups
Week 13

- Edit presentation and add final touch ups
- Submit the project

Next Steps

Integrating machine learning to better understand the target audience will significantly reduce costs and labour efforts in developing low-cost, effective future campaigns. The next step, following this proposal, is to apply machine learning techniques to the dataset. Without this approach, it will be challenging to maintain competitiveness while minimising expenses and labour. Therefore, this project should be prioritised and completed without delay.

References

Nofal, S. (2023, September 25). *Identifying highly-valued bank customers with current accounts based on the frequency and amount of transactions*. Retrieved from PubMed Central: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11255440/

Prabadevi, Shalini, & Kavitha. (2022, September 8). *Customer churning analysis using machine learning algorithms*. Retrieved from ScienceDirect: https://www.sciencedirect.com/science/article/pii/S2666603023000143

Santiago, Pedro, & Francisco. (2020, October 28). *A machine learning approach to the digitalization of bank customers: Evidence from random and causal forests*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7593085/