**Predictive Models for Telemarketing**

**INTRODUCTION**

According to an article by Boada(2022) discussed telemarketing as a tool used by organizations to reach out to prospective customers by phone and sell them their products.

Advertising, in which a business makes an unsolicited first contact with a potential customer to sell them a product or service, is the most well-known form of telemarketing (Queensland Government 2022).

Having an understanding of your customer’s experience can also benefit other departments of the business, benefits such as customer acquisitions, event attendance confirmation, post-sale follow-up, and customer satisfaction surveys (Schwager and Meyer, 2007)

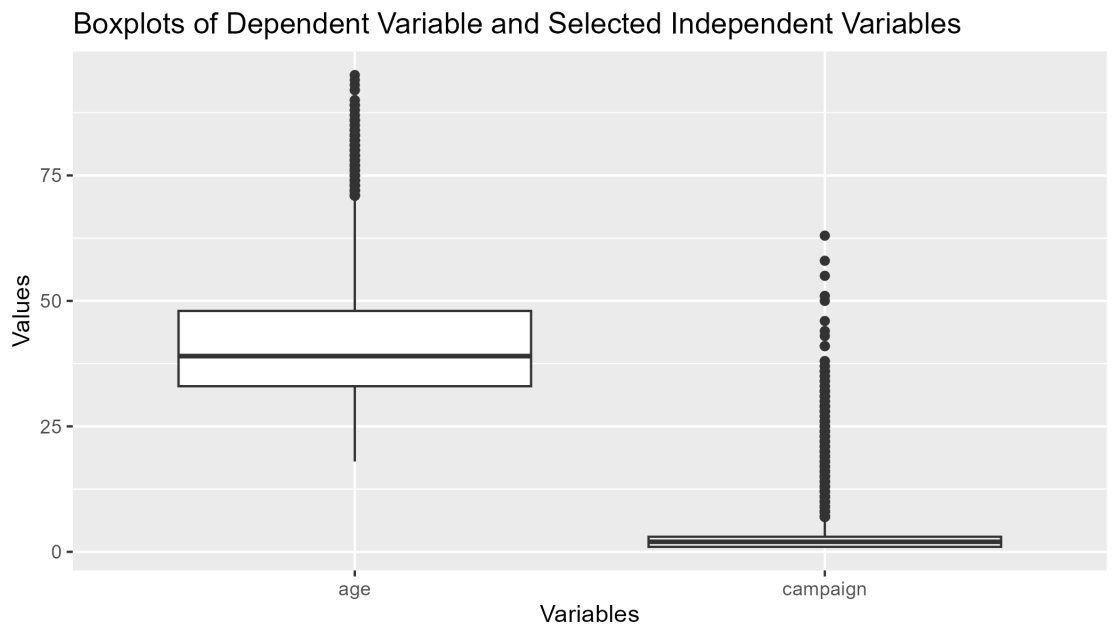
A major benefit is that organizations now understand their customers by having an insight into their thought process when buying or subscribing to a product or service in today's hyper-competitive business environment. This brings organizations to focus their marketing efforts to maximize their return on investment by knowing which customers are likely to subscribe to a product or service (Marotta 2021).

Commercial banks now readily look for avenues to gain insights into the characteristics of their target audience and develop a better understanding of what motivates customers (Garg et al. 2017).

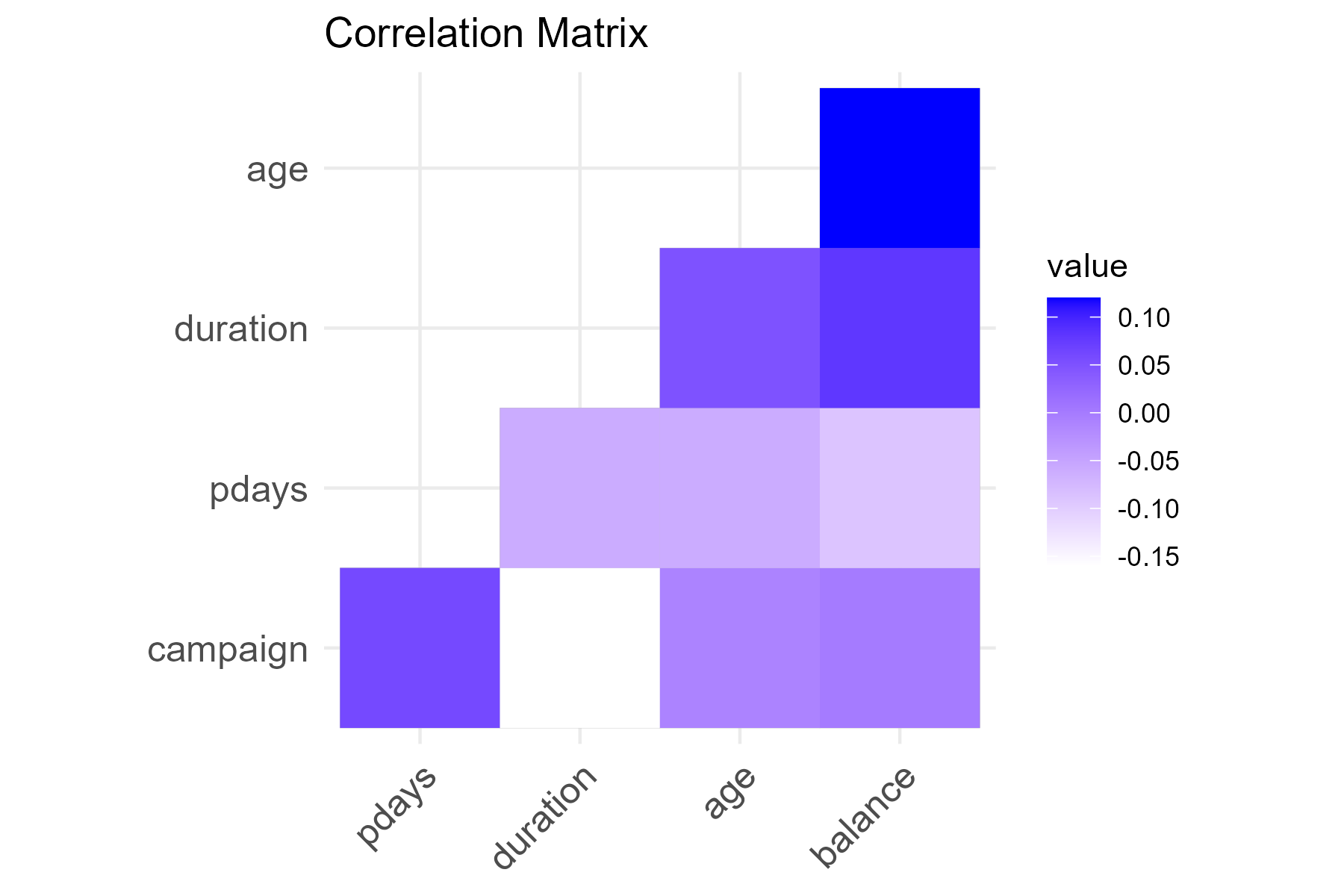
Based on a study on Miner KKBox provides streaming services to millions of users, with over 30 million soundtracks available for free and premium (paid) streaming on a variety of devices. This study follows a process whereby paid subscribers are classified into meaningful categories based on transactional and listening behaviors. Segmentation insights aided in the development of customized strategies to improve relationships with customers and profitability. It also provided a summary of the predictive model developed to identify customer turnovers for the KKBox music subscription service.

Because acquiring new customers is usually expensive in any business or organization, a subscription service like KKBox would benefit financially from investing in customer retention (Junior and Dinis, 2017).

**DATA OVERVIEW**

I will start by providing an overview of the dataset and the variables used in the model. The data provided for this study contains data on phone calls made to customers asking them to subscribe to a service; this dataset has 17 columns which were further categorized into four groups. During the exploratory phase of this analysis, the dataset indicated a large presence of outliers among numerical variables. 

Also during this phase, a correlation matrix was derived to get more insights into the relationships between the variables present in the data before model creation. Looking at some of the variables, a high correlation between its customer's current account balance and the age of the customer can be observed from the plot.



Following this, we will describe the methodology used to create each model and evaluate its performance. I will also be comparing the results of the two models used in this study and the best one that predicts whether or not a customer subscribed for the product.

**METHODS**

This study looks at logistic regression as one of the methods to use in classifying and predicting customers that subscribed or not. Logistic regression is a straightforward but effective model for predicting a binary outcome that is, whether or not an occurrence took place. It is a classification model used in supervised machine learning which is used in almost every industry, for example in marketing, healthcare, and others and is an essential piece of tool for any data scientist (Mudalige et al., 2021c).

In the case of this study, we want to predict whether a customer will subscribe based on the details gathered from calling the customer. Access to important predictor variables that are likely to be linked to the desired outcome is required by the model. In this case, after selecting some of the variables suspected to be influential, the logistic regression model is seen to be more accurate when all the variables in the data are used as predictor variables but If input variables are highly correlated with one another (known as multicollinearity), then the effect of each on the regression model becomes less precise (Ranganathan, Pramesh and Aggarwal, 2017).

A report on the machine learning classification model by Jiang(2018) states that a logistic regression model is a common tool for statistical classification and discriminant analysis. It finds extensive use in several situations, such as marketing management. This is also generally considered to be easily interpretable which means that the coefficients of the model will be used to interpret the relationship between the dependent and independent variables easily for the bank's marketing team.

The measure of forecast this model will also be using will be accuracy, sensitivity and specificity. We will be looking at how well the models predict which customers will subscribe to the service and which customers will not. A high accuracy score, for example, indicates that the model is making overall accurate predictions, whereas a high sensitivity score indicates that the model is correctly identifying the majority of customers who will not subscribe to the service.

In comparison with the logistic regression model, we will look at the Support Vector Machine as a tool for predicting a customer's likelihood of subscription. This algorithm has a good generalization performance, which means that it can make accurate predictions on new or unseen data. In interpreting the forecast measure of an SVM model in predicting the outcome of a customer subscription status, an appropriate metric to consider identifying customers who are likely to subscribe is specificity (Sarker, 2021).

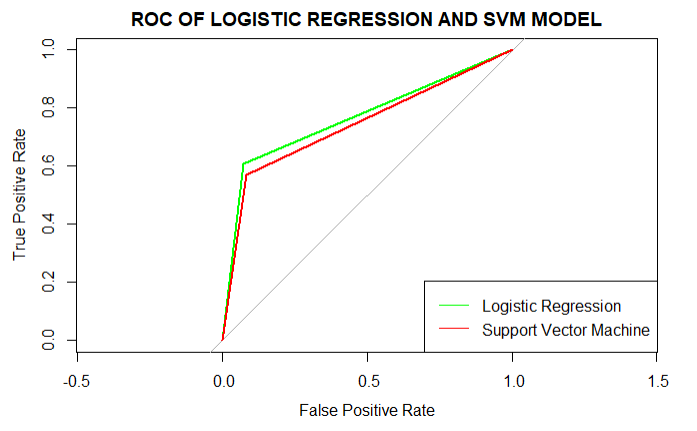
**RESULTS**

Looking at the target or dependent variable y and the p-value of all the independent variables, the variable with a higher p-value and looking at the variable of importance output some variables are being exempted from the second logistic regression model to determine whether or not these set of variables are more influential than using all the variables provided in the dataset. These set of variables are month, contact, housing, loan, duration, previous, day, campaign and previous outcome. Accessing the summary of both models, the AIC including all the variables in the dataset is 3602.7 and that of the selected variables is 3617 The basic idea of AIC is to penalize the inclusion of additional variables to a model. It adds a penalty that increases the error when including additional terms. The AIC for the two modifications are closely related so to avoid the issue of multicollinearity, the second Logistic regression model with fewer variables should be used.

Another modification that was made to the Logistic regression model was tuning the hyper-parameters to get an even more accurate result of the model. Looking at the three modifications of the logistic regression model, the one with the modification of a slightly higher accuracy value and the highest specificity value is the second model, which implies that the percentage at which the model generally identifies true values (predicting correctly that a customer will or will not subscribe to the bank's service) among all true predictions is 85.33%. Looking at the specificity percentage (the ability of the model to detect true negatives) of the third model, the value is 53.79% which is low compared to other modifications of the logistic model which are 67.54% and 71.2%.

Since the cost of choosing a false positive (i.e., predicting a positive outcome when the true outcome is negative which means that the model will identify that a particular customer will not subscribe when the true outcome is subscribed) is high. The logistic regression model with the best modification is the second logistic regression.

The metrics for SVM are accuracy 84%, precision 88.8% and recall which is also considered as the sensitivity 91.58%. Comparing the specificity metrics i.e. the percentage that the negative outcome 1 will be predicted as positive outcomes 0 for both the SVM and logistic model, the percentage specificity for the logistic regression model is 71.2% which is higher than that of SVM 61.3%. The second logistic regression model is still considered the best model to predict if a customer will subscribe to the bank's service and should be recommended to t he marketing team of the bank.



The FPR and TPR for SVM and Logistic Regression models are almost equal at the point where their respective ROC curves intersect. The question of which ROC curve performs better is thus no longer answerable categorically; rather, it now depends on whether the study prioritizes high precision and low recall or low precision and high recall.

**CORPORATE PURPOSE & LIMITATIONS**

The model can be used to determine which benefits are most important to customers who use the bank's service. This result can be used to create new products or improve existing ones to better meet the needs of customers, resulting in increased customer satisfaction and retention, as well as revenue growth. The purpose can also be to identify customers who are at risk of cancelling their service. The bank can reduce turnovers and preserve revenue by taking proactive measures such as targeted promotions, personalized communications, or improved customer service.

**LIMITATIONS**

The preference and behavior of a customer can change over time and the model should be able to keep up i.e. include a variable that will monitor the spending pattern of the customer to be updated on the preferences of the customer.

Subscription behavior may be influenced by factors such as a customer's life events or external market conditions that are not captured in this data.

**CONCLUSION**

The model provides useful insights into customer behavior, enabling the bank to better tailor its marketing and product offerings to the needs of its customers. This will lead to increased customer satisfaction, retention, and revenue growth. It is also important to pay attention to the limitations of the model and also continually evaluate and update the model as new data becomes available or as customer behavior changes over time.

In a view to gaining a more comprehensive understanding of customer behavior and preferences, the predictive model can be used in conjunction with other data sources and analytics tools (IBM, 2022).

The bank can make more informed decisions and improve its overall business performance by integrating data from multiple sources and not just calls alone.

Ultimately, the model is a valuable tool for banks looking to increase customer subscriptions and revenue growth. However, it is critical to proceed with caution when interpreting the model's predictions and to constantly assess their accuracy and usefulness in making business decisions.

**REFERENCES**

Boada, N. (2022). *What Is Telemarketing? Definition, Advantages, and Examples*. [online] www.cyberclick.net. Available at: <https://www.cyberclick.net/numericalblogen/what-is-telemarketing-definition-advantages-and-examples> [Accessed 18 Mar. 2023].

Garg, A., Grande, D., Miranda, G.M.-L., Sporledar, C. and Windhagen, E. (2017). *Analytics in banking: Time to realize the value*. [online] McKinsey & Company. Available at: <https://www.mckinsey.com/industries/financial-services/our-insights/analytics-in-banking-time-to-realize-the-value> [Accessed Mar. 2023].

IBM (2022). *What Is Predictive Analytics? | IBM*. [online] www.ibm.com. Available at: <https://www.ibm.com/topics/predictive-analytics> [Accessed 23 Mar. 2023].

Jiang, Y. (2018). Using Logistic Regression Model to Predict the Success of Bank Telemarketing. *International Journal on Data Science and Technology*, [online] Vol. 4(ISSN: 2472-2200 (Print); ISSN: 2472-2235 (Online)), p.7. doi:<https://doi.org/10.11648/j.ijdst.20180401.15>.

Junior, C. and Dinis, G. (2017). *Churn Analysis in a Music Streaming Service : Predicting and understanding retention*. [online] *kth.diva-portal.org*. Available at: <http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-215709> [Accessed 30 Mar. 2023].

Marotta, D. (2021). *Customer Insights: What They Are & Why They Matter*. [online] Hitachi Solutions. Available at: <https://global.hitachi-solutions.com/blog/what-are-customer-insights/> [Accessed 27 Mar. 2023].

Mudalige, S., Ganegoda, G.U., Jayatilake, C. and Alexis, D. (2021a). Involvement of Machine Learning Tools in Healthcare Decision Making. *Journal of Healthcare Engineering*, [online] 2021, pp.1–20. doi:<https://doi.org/10.1155/2021/6679512>.

Queensland Government (2022). *Using direct marketing*. [online] www.business.qld.gov.au. Available at: <https://www.business.qld.gov.au/running-business/marketing-sales/marketing/activities/direct-marketing> [Accessed 23 Mar. 2023].

Ranganathan, P., Pramesh, C.S. and Aggarwal, R. (2017). Common pitfalls in statistical analysis: Logistic regression. *Perspectives in Clinical Research*, [online] 8(3), pp.148–151. doi:<https://doi.org/10.4103/picr.PICR_87_17>.

Sarker, I.H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, [online] 2(3). doi:<https://doi.org/10.1007/s42979-021-00592-x>.

Schwager, A. and Meyer, C. (2007). *Understanding Customer Experience*. [online] Harvard Business Review. Available at: <https://hbr.org/2007/02/understanding-customer-experience> [Accessed 20 Mar. 2023].