A System to Predict the Subscription Likelihood of a Bank Customer

Babatunde Racheal

3/04/23

INTRODUCTION

The banking sector in today's world has been considered one of the most lucrative, productive and competitive sectors amongst other industries in the economy and like every other sector, it also looks for avenues to increase revenue, customer satisfaction and overall growth every year (Pakurár et al., 2019). Organizations or businesses resort to marketing initiatives as a form of outsourcing to boost their bottom lines and differentiate themselves from competitors. Direct marketing is used by businesses when they want to reach a specific group of their customers to accomplish a goal. Clients can be reached from afar via call centers to streamline the management of campaigns and these call centers facilitate interactions with customers via a variety of channels, including telephones. Telemarketing, or sales promotions through a call centre, and this is still so due to the remoteness effect (Asare-Frempong and Jayabalan 2017).

The goal of a study by Alsolami, Saleem and AL-Ghamdi (2020) utilized bank data to predict the efficiency of bank telemarketing campaigns. An important hindrance to this is that banks are reluctant to provide room to outsiders due to restrictions and privacy issues as banks possess highly confidential data and have a strict policy in sharing the customer's data; this research addressed these and other pressing issues. The primary question that this study aimed to answer is "how do data mining algorithms enhance the procedure of banks' telemarketing campaigns" The proposed model assisted banking institutions to better comprehend their ideal customers. The next step of the study focused on the difficulty banks face when sifting through customer information and attempting to get in touch with them about deposits or other matters. Increased focus on customer relationship management is another development in banking telemarketing. Banks no longer aim to simply make a profit from their customers, but rather to encourage meaningful relationships with them and this

means that telemarketers learn to empathize with their clients and offer tailored solutions to their unique problems (Osifo, 2020a).

Telemarketing is still an integral part of the banking industry's marketing strategy, but it has developed significantly in recent years to adapt to changing consumer preferences and recent technological possibilities (Xie et al., 2023).

DATA SUMMARY

The data provided for this study includes the information gathered from phone calls to the bank's customers. The data included 17 columns and the last column is the subscription output this study is interested in predicting. Performing exploratory data analysis on the data, one issue the data have is the presence of outliers because some variables like age and customer balance widely differ in value. In an attempt to reduce the effect of outliers in the data, the scale transformation was first used but later opted for the log transform option because the scaling procedure gave our data a negative to a positive range which will likely not be suitable for the model.

METHODS

The Simple Vector Machine and the Naive Bayes classification models are the models this study will be considering. Both classification and regression issues can be solved with a support vector machine. Finding a maximum marginal hyper-plane by categorizing datasets is the central objective of SVM, it is an efficient strategy for forecasting which marketing and sales leads are most likely to sign up for a service (Sunil, 2019).

Often remember that the hyper-parameters you pick and the quality of the data you feed SVMs can greatly affect the accuracy of the model. Therefore, it is essential to perform thorough analyses of the model's outputs and to pay close attention to its development (Zhen & Wenjuan 2016).

The major benefit stated in Stecanella's(2018) study of using SVM is that it performs well even when dealing with a large number of attributes. In high dimensions, it performs reasonably well and since SVM is kernel-based, it can model complicated and practical issues. With SVM, the possibility of over fitting is reduced. A couple of downsides to choosing the method is the selection of kernel function and when we have a reasonably large dataset involved it takes a long time.

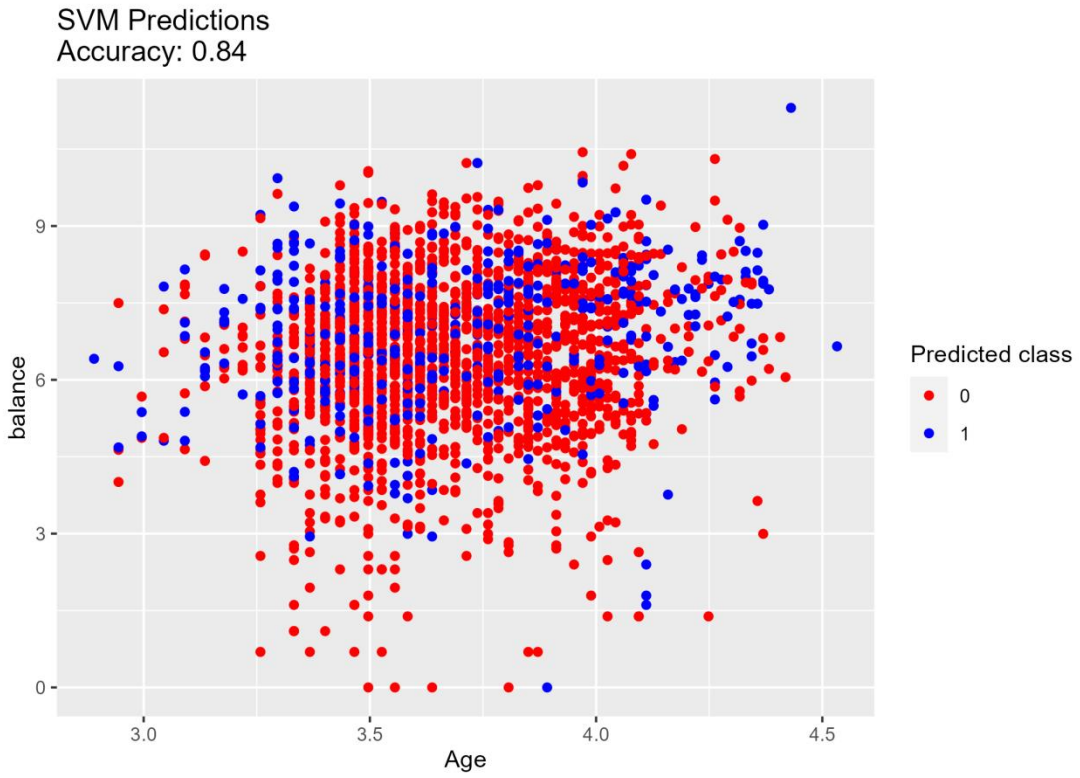
Naive Bayes is a probabilistic algorithm that uses the joint probabilities of a data point's features to determine the likelihood that it belongs to a given class. It is centred on conditional probability theory, specifically Bayes' theorem. The system is based on the assumption that individual characteristics are unrelated to one another. Despite this disadvantage, the classifier is straightforward and produces respectable results (www.turing.com, 2022).

The following will be measures used to evaluate the performance of both models, a confusion matrix table details the percentage of correct classifications, incorrect classifications, and total classifications for each class in the validation set. A way to also measure the accuracy of this table is by investigating the accuracy, sensitivity, specificity and precision and ROC metrics.

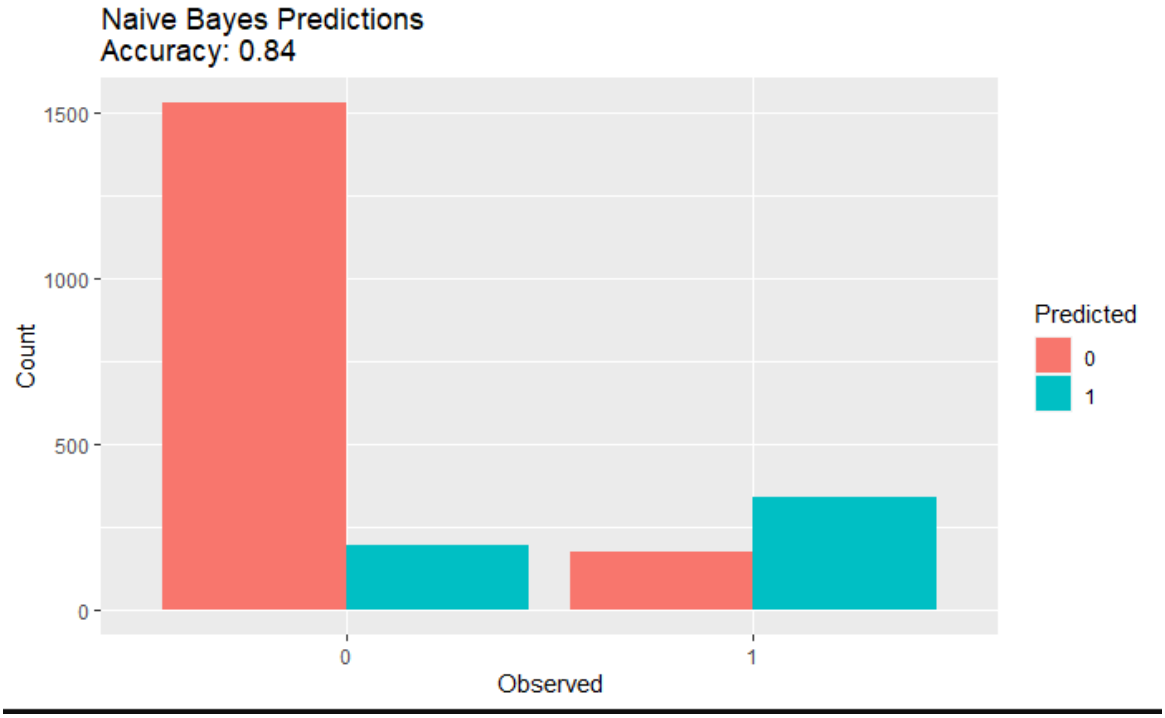
RESULTS

The performance indicators for SVM are an accuracy of 84% which gives us the proportion of how many instances out of all instances in the test set were correctly classified, with a precision of 87.74%, this is the percentage of true positive predictions out of all positive predictions made by the model and recall of 91.81% which can also be called sensitivity (also known as true positive rate) is the proportion of actual positive instances that are correctly identified as positive by the model.

Sensitivity = True Positives / (True Positives + False Negatives). Specificity for this model is 57.09%, this implies that the proportion at which the model predicts true negative values (subscribed 1) as a negative value is 57.09%.



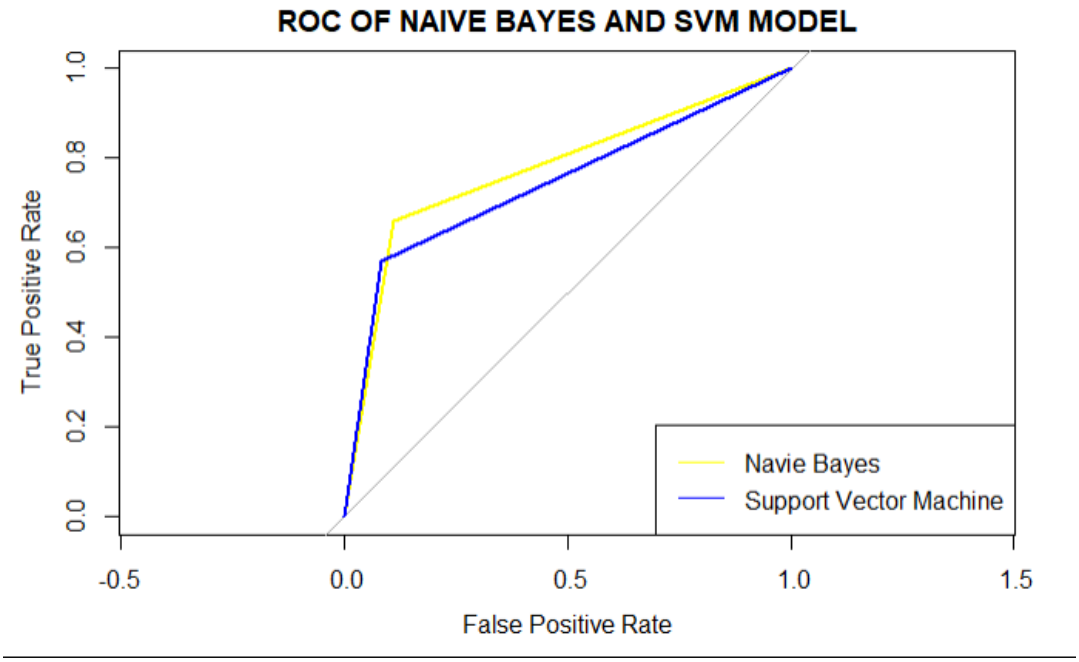
The plot shows the points predicted as subscribed i.e. 1 and not subscribed 0 based on two variables age and balance from the data. One quick visual access point that can be derived from this visualization is that the predicted class for not subscribed is more than subscribed and this is accurate because the data are originally given as more non-subscribed customers than subscribed meaning the precision for this model is accurate.



Evaluating similar metrics for the naïve Bayes in order to compare model performance, this bar plot visualizes the confusion matrix table i.e. the proportion of positives predicted as positives, negatives predicted as negatives and it can also be deducted that Support Vector Machine has slightly higher accuracy measure than that of the Naïve Bayes despite the percentage of specificity being low (ability to predict true negative), I will recommend the support vector machine because the cost of making false negative can be low compared to the cost of making false positive i.e. if we say a customer will subscribe to the bank's service and the customer does not subscribe, a substantial amount of resource would have been directed to this particular customer, meanwhile this study aims to minimize the marketing cost of the bank while also increasing the revenue.

COMPARING THE METRICS OF BOTH MODELS

	SUPPORT VECTOR MACHINE	NAÏVE BAYES
ACCURACY	83.82%	83.55%
RECALL	91.81%	88.85%
SPECIFICITY	57.09%	65.83%
PRESCION	87.74%	89.68%



If you plot the ROC curves for SVM and Naive Bayes, you'll notice that they intersect at a point where the FPR and TPR are identical for both methods. Therefore there is no longer a definitive answer to the question of which ROC curve performs better; instead, the answer depends on whether the study values high precision and low recall or low precision and high recall.

PURPOSE AND LIMITATION OF THE STUDY

The study can be implemented by learning which benefits customers who use the bank's service value the most. Increased customer satisfaction and retention, as well as revenue growth, can result from using this information to develop and modify products or services to better meet the needs of customers. The goal may also be to foresee which subscribers are most likely to terminate their subscriptions. By being proactive with measures like targeted promotions, personalized communications, and enhanced customer service, the bank can reduce turnovers and maintain revenue.

Customers (and their intended communications) are traditionally segmented using basic factors like life stages and demographic. However, this is no longer adequate and a way

banks can boost their marketing return on investment and make better use of customer segmentation strategies is if they take a fresh look at their data and use it creatively (Buglewicz, 2023).

Examples of constraints include the data does not take into account the potential influence of exogenous variables, such as market conditions at a given period of the campaign and incorporating a variable that tracks a customer's spending pattern, for example, customers might want a service that keeps them grounded especially during the festive period. This will help them stay current on their preferences and is essential because customers' preferences and behavior are dynamic and can shift over time.

CONCLUSION

The banking industry is under increasing pressure to increase profits and reduce costs, making optimization aiming at telemarketing a critical issue. European banking has evolved significantly since the 2008 financial crisis. In particular, pressure was applied to the Portuguese banking sector to raise capital requirements (e.g., by capturing more long-term deposits). Moro, Cortez and Rita (2014).

The predictive model can be used alongside other data sources and analytics tools to learn more about customer behavior and preferences and by incorporating data that cannot be gotten from just calls alone, the bank can also use this to make better decisions and boost its overall business performance.

The model is helpful for financial institutions that want to boost their subscription rates and their bottom line. Caution should be exercised when acting on the model's predictions, and their validity and utility should be evaluated on an ongoing basis. The model also helps the bank learn more about its customers' habits, which improves its ability to meet their specific

requirements through targeted advertising and product development. Customer happiness, loyalty, and spending will all rise as a result.

REFERENCE

Alsolami, F.J., Saleem, F. and AL-Ghamdi, A.A.-M. (2020). مجلةجامعةالملكعبدالعزيز - عالومالحاسباتوتقنيةالمعلومات. *PREDICTING THE ACCURACY OF TELEMARKETING PROCESS IN BANKS USING DATA MINING*, vol 9. doi:https://doi.org/10.4197/comp.

Asare-Frempong, J. and Jayabalan, M. (2017). Predicting customer response to bank direct telemarketing campaign. *2017 International Conference on Engineering Technology and Technopreneurship (ICE2T)*. [online] doi:https://doi.org/10.1109/ice2t.2017.8215961.

Bansala, R., Singh, J. and Kaur, R. (2022). *Machine learning and its applications: A Review*. [online] Available at: https://www.researchgate.net/publication/338685514_Machine_learning_and_its_applications_A_Review [Accessed 25 Feb. 2023].

Buglewicz, S. (2023). *Advanced Strategies for Customer Segmentation in Banking*. [online] CCG. Available at: https://www.customer.com/blog/financial-marketing/advanced-strategies-for-customer-segmentation-in-banking/ [Accessed 29 Mar. 2023].

Moro, S., Cortez, P. and Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62, pp.22–31. doihttps://doi.org/10.1016/j.dss.2014.03.001.

Osifo, S. (2020a). CUSTOMER RELATIONSHIP MANAGEMENT AS A TOOL FOR IMPROVING BANK PERFORMANCE AND NATION-BUILDING. *The Academy Management Journal*, [online] 15. Available at: https://www.researchgate.net/publication/350994166_CUSTOMER_RELATIONSHIP_MANAGEMENT_AS_A_TOOL_FOR_IMPROVING_BANK_PERFORMANCE_AND_NATION_BUILDING [Accessed 29 Mar. 2023].

Pakurár, M., Haddad, H., Nagy, J., Popp, J. and Oláh, J. (2019). The Service Quality Dimensions that Affect Customer Satisfaction in the Jordanian Banking Sector. *Sustainability*, [online] 11(4), p.1113. doi:https://doi.org/10.3390/su11041113.

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.

Stecanella, B. (2018). *An introduction to Support Vector Machines (SVM)*. [online] MonkeyLearn Blog. Available at: https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/ [Accessed 30 Mar. 2023].

Sunil, R. (2019). *Understanding Support Vector Machine algorithm from examples (along with code)*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/ [Accessed 30 Mar. 2023].

www.turing.com. (2022). *Naive Bayes Algorithm in ML: Simplifying Classification Problems*. [online] Available at: https://www.turing.com/kb/an-introduction-to-naive-bayes-algorithm-for-beginners [Accessed 30 Mar. 2023].

Xie, C., Zhang, J.-L., Zhu, Y., Xiong, B. and Wang, G.-J. (2023). How to improve the success of bank telemarketing? Prediction and interpretability analysis based on machine learning. *Computers & Industrial Engineering*, 175, p.108874. doihttps://doi.org/10.1016/j.cie.2022.108874.