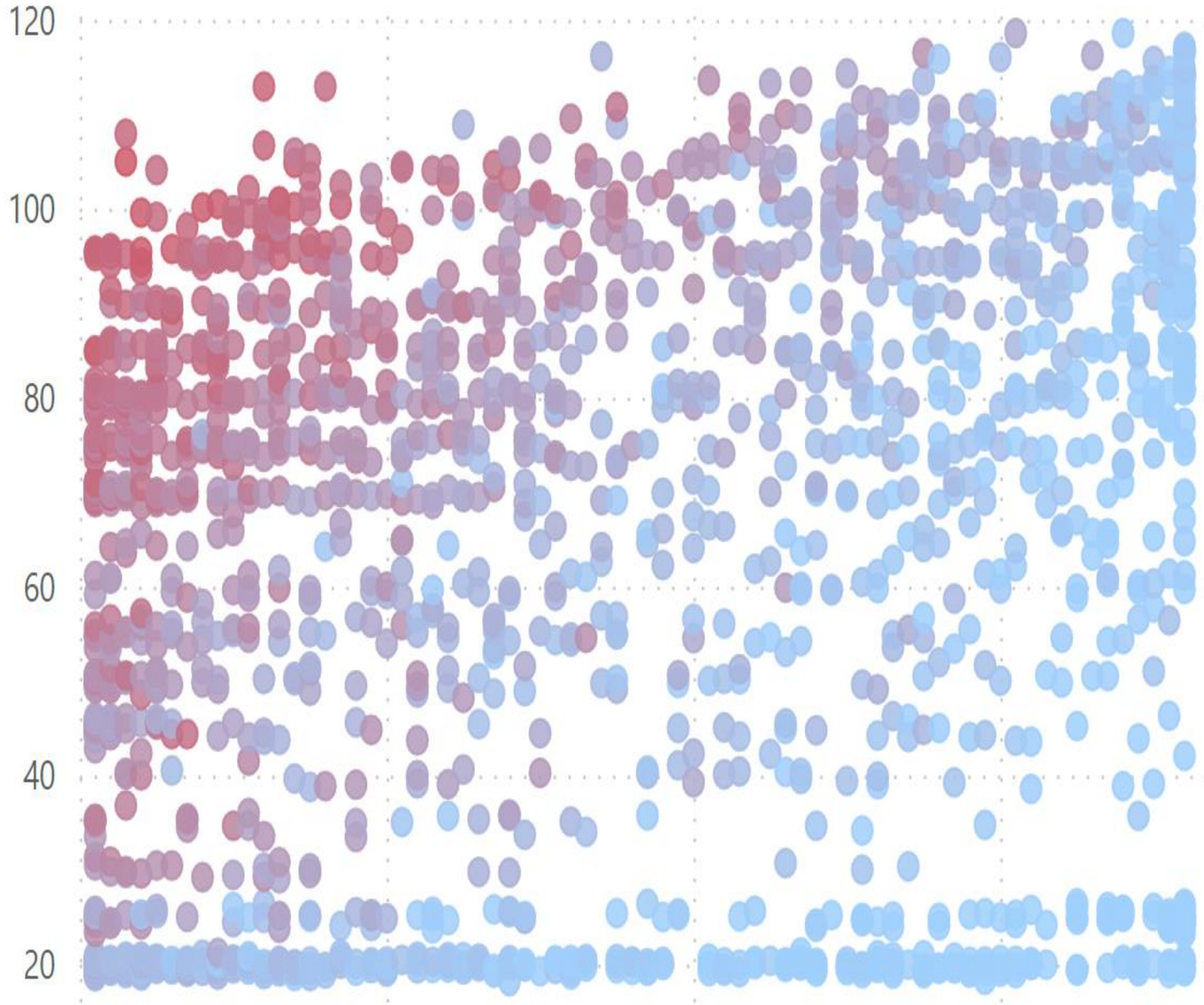


LOGISTIC REGRESSION

ASSESSMENT 3

ZZBU6514 MANAGING CUSTOMER ANALYTICS (STEPHAN TSENG)

DAVID OTTO, UNSW Z5379919

CONTENTS

Executive Summary 3
Research Background and Objectives 4
Description of Sample 4
Methodological Approach 5
  Logistic Regression 5
  Methodology 6
  Assumptions & Proposed Scenarios 6
  Model Evaluation and Selection 7
Results 8
  Influencing Factors 8
  Churn Rates 9
  Profitability 10
Implications 11
  Recommendations 11
  Limitations 12
Conclusion 12
References 14

Figure 1: Variables used in Datasets 5
Figure 2: Analysis Methodology 6
Figure 3: Evaluation of Regression Models 7
Figure 4: Variables Used in Each Model and their Significance (p-value; less than 0.05 is significant) 8
Figure 5: Standardised Coefficients of Influencing Factors 8
Figure 6: Churn Rates in Datasets 9
Figure 7: Number of Customers by Likelihood of Churn 9
Figure 8: Maximum Profitability by Churn Threshold 10
Figure 9: Chart of Scenarios for Campaign (recommended scenarios highlighted) 11
Figure 10: Table of Scenarios for Campaign (recommended scenarios in bold) 11

EXECUTIVE SUMMARY

Questions have arisen at SY regarding the cost and benefit of its cold calling campaign to prevent the churn of its customers. This report employs a logistic regression model to determine the probability of churn for a set of customers, and based on this it performs calculations to determine the profitability of targeting customers as opposed to simply contacting all customers. With the data provided, targeting all customers provides the highest profitability, however a targeted approach, such as targeting only those who are less likely to churn than to remain, could bring a better return on investment per dollar.

Word Count of body, excluding tables and figures: 1,313 words

RESEARCH BACKGROUND AND OBJECTIVES

The SY telecommunications firm has traditionally called all customers before their contract renewal to encourage them to renew with new special plans. However, questions about the effectiveness of the strategy of cold-calling every customer have arisen. Many customers decide not to renew their contract (i.e., churn), potentially wasting the costs involved. Indeed, the industry has seen more freedom of choice for customers and consumer protection in recent years (ITU, 2018, p. 53) (Muneiah & Rao, 2019, pp. 8167-8168), thus churn rates remain a constant challenge for SY.

Customer relationship management plays a pivotal role for retention in the industry (Sharma, et al., 2018, p. 316) (Bell, et al., 2005), but the effectiveness of cold calling is mixed (Schultz, 2023). However, data platform vendors also provide statistics that suggest that proper data can make cold calling strategies more effective (Cannon, 2023), so we could use data analytics in cold calling strategies to bring the costs of marketing under control.

This report will employ a logistic regression model to determine the factors of churn and the profitability of the cold calling strategy. It will show results to answer whether a targeted approach is more profitable than targeting everyone, and provide suggestions as to where the threshold should be set.

DESCRIPTION OF SAMPLE

Two datasets are provided: a **Model Development** dataset containing data on the previous customers with results for the previous campaign on whether they churned or not, and a **Prediction** dataset containing data for customers that are up for renewal and thus subject to a future campaign.

A summary of the variables for each customer are given below in Figure 1. Variables describing categories were modified to binary values so that they can be processed in a regression. No outliers were found in the data.

Variable	Description	Treatment
customerID	Customer ID number	Removed
gender	Gender of customer (“Male” or “Female”)	Changed to “**MaleGender**”; 1 = Male, 0 = Female
SeniorCitizen	Whether customer is a senior citizen (1 = yes, 2 = no)	Used as is
Partner	Existence of partner (“Yes” or “No”)	Converted variables to 1 = yes, 0 = no
Dependents	Existence of dependents (“Yes” or “No”)	Converted variables to 1 = yes, 0 = no
tenure	Number of months customer has been with SY	Used as is
PhoneService	Connection of landline phone service (“Yes” or “No”)	Converted variables to 1 = yes, 0 = no
Contract	Whether customer’s contract is “Short term” or “Long term”	Changed to “**LongContract**”; 1 = Long term, 0 = Short term
PaperlessBilling	Whether customer uses paperless billing (“Yes” or “No”)	Converted variables to 1 = yes, 0 = no
MonthlyCharges	Dollar value of the customer’s current monthly payment	Used as is
TotalCharges	Total dollar value of money the customer has paid for services	Used as is
Churn (Model Development only)	Whether customer churned (“Yes”) or renewed their contract (“No”)	Converted variables to 1 = yes (churned), 0 = no (did not churn)

Figure 1: Variables used in Datasets

METHODOLOGICAL APPROACH

Logistic Regression

For this analysis, a logistic regression is performed on the Model Development dataset. Logistic regression is a popular method that has been used extensively in telecom churn prediction (Jain, et al., 2020) (Mustafa, et al., 2021). Logistic regression is advantageous as it is a simple yet effective method that gives inference on how important each feature is in predicting the outcome, and it gives precise probabilities of the predictions (Grover, 2023).

Methodology

Figure 2: Analysis Methodology

For this analysis, several iterations will be performed to find the optimal regression model. This is done to remove variables with multicollinearity issues, which can reduce the precision of models (Frost, 2017), and remove variables which are not significant in predicting churn. The optimal model will be selected based on evaluations of accuracy and goodness-of-fit.

The optimal logistic regression model will then be applied to the Prediction dataset. Based on the probability of churn for all customers in the dataset, calculations will be performed to determine the potential cost, revenue, profit, and advertising-to-sales ratio (A-S ratio) for various scenarios.

Assumptions & Proposed Scenarios

The calculations are performed on the following assumptions:

1. **Cost of $7** per customer contact.
2. Potential **revenue of $50** for successful retention.
3. **Weighted revenue** of the $50 potential revenue multiplied by the probability of not churning.

The calculations will be performed for the following scenarios:

1. Calling **all customers**.
2. Calling the customers with a **churn rate below the threshold of 86%** (given $50 \times (1 - 0.86) = 7$, meaning a higher churn rate would give a negative weighted profitability).
3. Calling customers who are **less likely to churn** than to renew (churn rate of less than 50%).

Additional scenarios will also be suggested based on the results of the analysis.

Model Evaluation and Selection

The following models were evaluated, and the optimal model was selected. Whilst this model has one factor with low significance, its higher goodness-of-fit results using the Hosmer-Lemeshow test and accuracy means that it is optimal for this exercise.

No.	Accuracy (percent correct on training data)	Goodness-of-fit (Nagelkerke Pseudo-R^2) (Higher is better)	Goodness-of-fit (Hosmer-Lemeshow test) (Higher is better)	Remarks
1	81.4%	0.355	0.002	Multicollinearity issues with Tenure and TotalCharges. Tenure was chosen due to its stronger significance.
2	79.4%	0.353	0.049	
3	79.4%	0.352	0.018	
4	78.8%	0.345	0.055	
5	79.6%	0.346	0.003	
6	79.0%	0.338	0.001	
7	79.4%	0.353	0.136	Selected model
8	78.9%	0.329	0.081	

Figure 3: Evaluation of Regression Models

Model	1	2	3	4	5	6	7	8
SeniorCitizen	<0.01	<0.001	<0.001	--	<0.001	--	<0.001	<0.001
Tenure	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	
MonthlyCharges	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
TotalCharges	0.003	--	--	--	--	--	--	<0.001
MaleGender	0.867	0.857	--	--	--	--	--	--
Partner	0.958	0.956	--	--	--	--	--	--
Dependents	0.125	0.099	--	--	--	--	0.073	--
PhoneService	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
LongContract	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
PaperlessBilling	<0.001	<0.001	<0.001	<0.001	--	--	<0.001	<0.001

Figure 4: Variables Used in Each Model and their Significance (p-value; less than 0.05 is significant)

RESULTS

Influencing Factors

The logistic regression gives information on influencing factors. It appears that tenure, the monthly charge amount, and contract term type have the greatest influence on whether a customer churns.

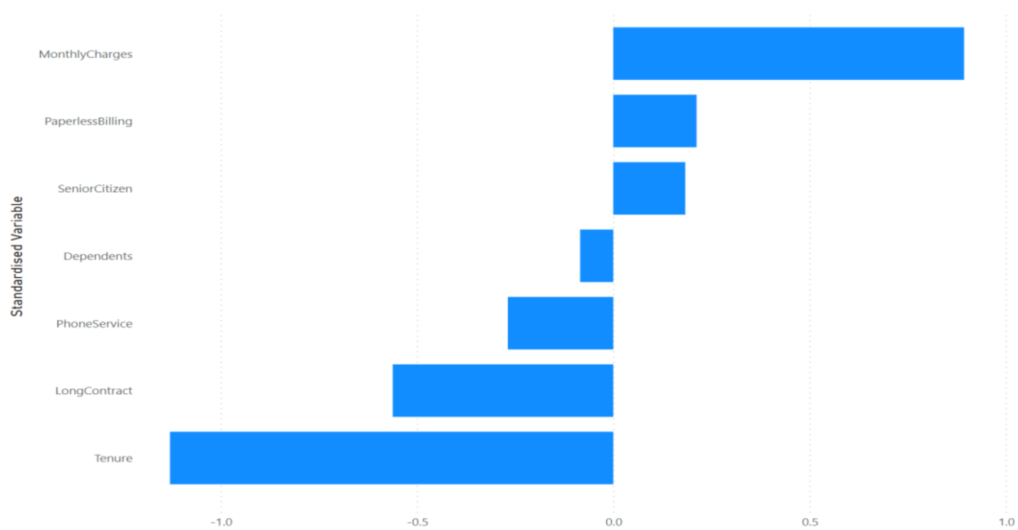


Figure 5: Standardised Coefficients of Influencing Factors

Churn Rates

	Model Development (actual)	Model Development (model)	Prediction (model)
Churn	1086 (26.2%)	785 (19.0%)	327 (17.8%)
No churn	3054 (73.8%)	3355 (81.0%)	1508 (82.2%)

Figure 6: Churn Rates in Datasets

Figure 6 shows that churn rates range from 17 to 26 percent of the cohort, making high probabilities of churn unlikely to be predicted. The highest churn rate for Prediction is 0.858, below the profitability threshold of 86%. This means that, for this dataset, setting the threshold of 86% will be ineffective, as no customers are above that threshold.

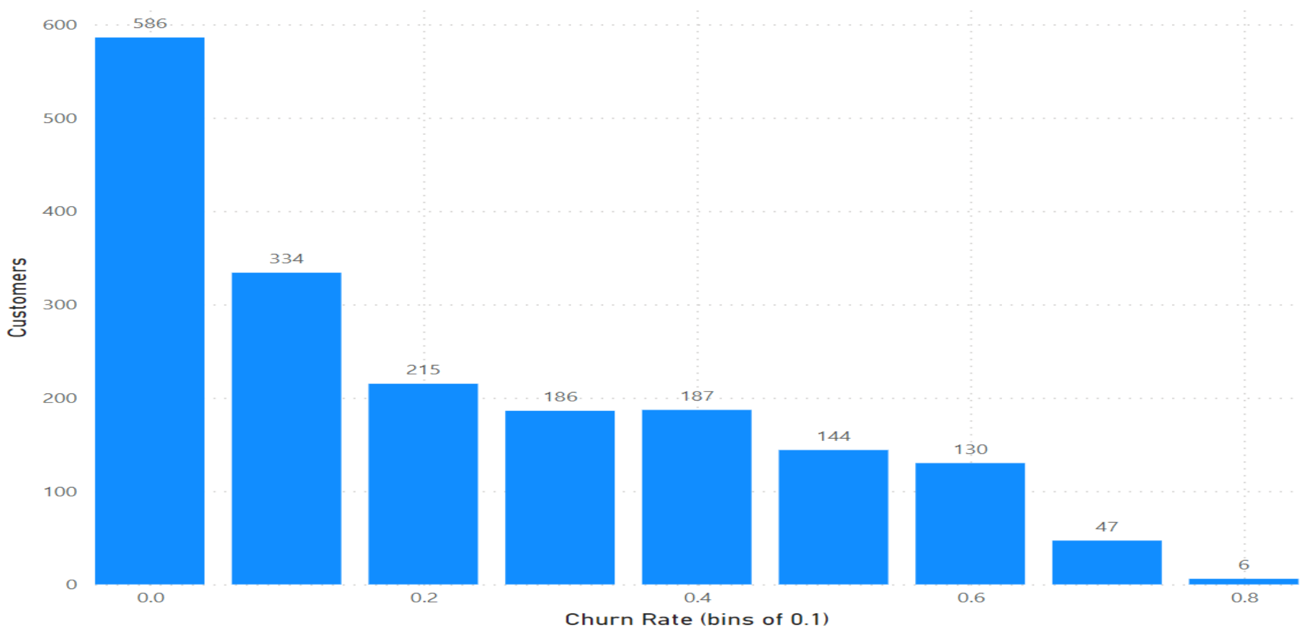


Figure 7 :Number of Customers by Likelihood of Churn

Profitability

Figure 8 shows the maximum profitability at thresholds in intervals of 0.1.

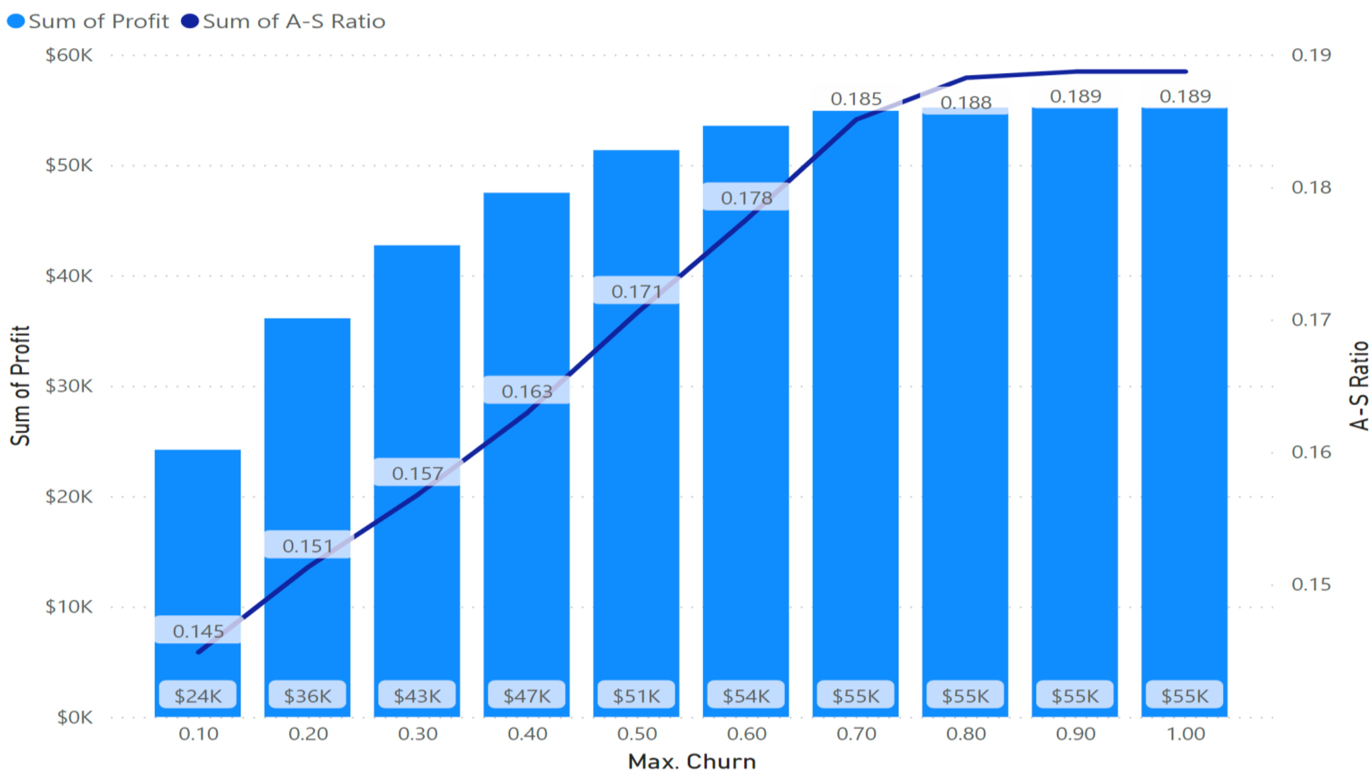
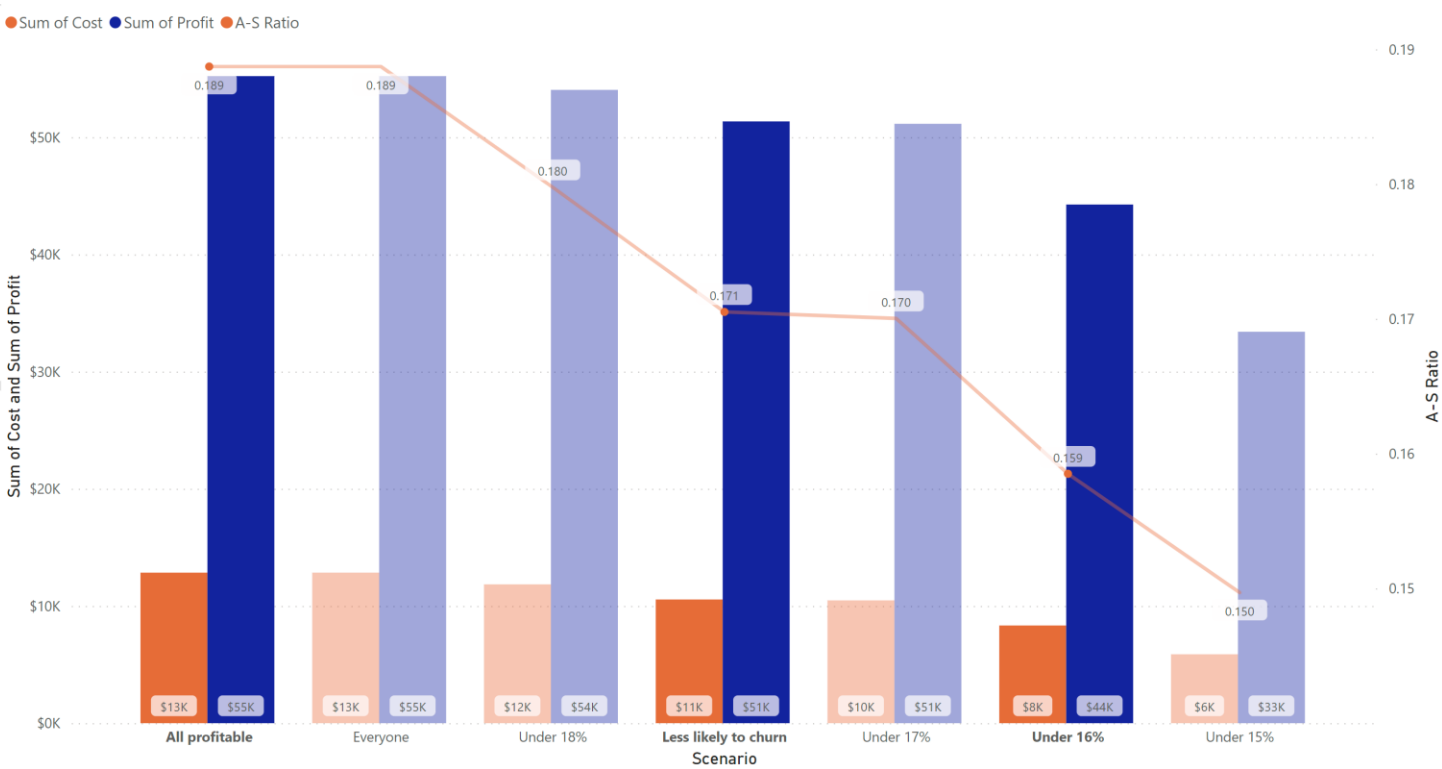


Figure 8: Maximum Profitability by Churn Threshold

In terms of profit alone, it is advantageous to include all customers in this dataset. However, whilst including all customers in the dataset would deliver the highest profit, we can see that the returns diminish the higher the churn threshold is set. SY might consider a lower threshold to control costs and deliver greater returns for lower expenditure.

Only contacting those who are less likely to churn (threshold of 50%) is an ideal solution, delivering an A-S ratio of just over 0.17 and returning a profit of over $51,000. Although these ratios could be seen as high (Media Group Online, 2022), due to the higher amounts of profit to be gained it is recommended that SY pursue at least 0.16 to deliver decent returns. Potential scenarios are detailed below.



Q

Figure 9: Chart of Scenarios for Campaign (recommended scenarios highlighted)

Scenario	Max. Churn	Revenue	Cost	Profit	A-S ratio	Contacts	Percentage
Under 15%	17%	$ 39,282.14	$ 5,880.00	$ 33,402.14	0.150	840	46%
Under 16%	**33%**	**$ 52,597.53**	**$ 8,337.00**	**$ 44,260.53**	**0.159**	**1191**	**65%**
Under 17%	49%	$ 61,628.82	$ 10,479.00	$ 51,149.82	0.170	1497	82%
Under 18%	63%	$ 65,885.75	$ 11,844.00	$ 54,041.75	0.180	1692	92%
Less likely to churn	**50%**	**$ 61,906.76**	**$ 10,556.00**	**$ 51,350.76**	**0.171**	**1508**	**82%**
All profitable	**86%**	**$ 68,063.58**	**$ 12,845.00**	**$ 55,218.58**	**0.189**	**1835**	**100%**
Everyone	100%	$ 68,063.58	$ 12,845.00	$ 55,218.58	0.189	1835	100%

Figure 10: Table of Scenarios for Campaign (recommended scenarios in bold)

IMPLICATIONS

Recommendations

This report recommends the following actions.

- For pure profitability**, all customers up to the given threshold of 86% churn should be contacted** (**100%** of customers).

    - Given the model and the lower likelihood of churn, the given threshold would have minimal effect. However, keeping the threshold would ensure profitability does not go down in the event many customers are found likely to churn.
- However, returns diminish and the higher the threshold, the lower the returns on investment.
    - **Contacting customers more likely to stay than churn is effective** (contact **82%** of customers).
    - To minimise costs, the threshold could be set to where the **A-S ratio of the campaign is under 16%** (**max. churn of 33%;** contact **65%** of customers).

Limitations

In addition to general caveats of using logistic regression (possibility of being too close to the model development set and rigidity due to the linear nature) (Grover, 2023), it is important to note that there is not enough data to evaluate whether the cold calling itself is effective, as we do not have data on persons who were not called.

For the purposes of obtaining data on the effectiveness of the cold calling, SY could conduct a randomised controlled trial in the future where customers are separated into random groups, then one group is cold-called and one group is not (Dattani, 2022).

CONCLUSION

This report has identified potential scenarios for SY to employ to deliver the most profitable outcomes in its cold calling campaign for renewing customers. It has found that the most profitable method is to contact as many customers as possible up to the given churn rate of 86%, above which customers would make a loss, although this would be realistically similar to contacting all customers. SY could, however, get a greater return on investment through limiting the customers they contact by limiting the number of contacts.

Using logistic regression analysis in this way can be effective in reducing expenditures for SY. The model in this report can target customers by their likelihood of churning. With additional data on whether the customer was called or not, it can potentially be expanded to measure the effectiveness of the cold calling itself.

REFERENCES

Bell, S. J., Auh, S. & Smalley, K., 2005. Customer Relationship Dynamics: Service Quality and Customer Loyalty in the Context of Varying Levels of Customer Expertise and Switching Costs. *Journal of the Academy of Marketing Science,* 33(2), pp. 169-183.

Cannon, W., 2023. 50 Cold Calling Statistics to Learn From in 2023. *UpLead.* [Online] Available at: https://www.uplead.com/cold-calling-statistics/ [Accessed 18 Jun 2023].

Dallimore, E. J., Hertenstein, J. H. & Platt, M. B., 2013. Impact of Cold-Calling on Student Voluntary Participation. *Journal of Management Education,* 37(3), pp. 305-341.

Dattani, S., 2022. Why randomized controlled trials matter and the procedures that strengthen them. *Our World In Data.* [Online] Available at: https://ourworldindata.org/randomized-controlled-trials [Accessed 19 Jun 2023].

Frost, J., 2017. Multicollinearity in Regression Analysis: Problems, Detection, and Solutions. *Statistics By Jim* [Online] Available at: https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/ [Accessed 18 Jun 2023].

Grover, K., 2023. Advantages and Disadvantages of Logistic Regression. *Open Genus.* [Online] Available at: https://iq.opengenus.org/advantages-and-disadvantages-of-logistic-regression/ [Accessed 18 Jun 2023].

ITU, 2018. *Global ICT Regulatory Outlook 2018,* Geneva: ITU Publications. Jain, H., Khunteta, A. & Srivastava, S., 2020. Churn Prediction in Telecommunication using Logistic Regression and Logit Boost. *Procedia Computer Science,* Volume 167, pp. 101-112.

Media Group Online, 2022. *2022 Advertising-to-Sales Ratios.* [Online] Available at: https://mediagrouponlineinc.com/wp-content/uploads/2022/07/AdToSalesRatios22b.pdf [Accessed 19 Jun 2023].

Muneiah, J. N. & Rao, C. D. V. S., 2019. Customer's class transformation for profit maximization in multi-class setting of Telecom industry using probability estimation decision trees. *Journal of Intelligent & Fuzzy Systems,* Volume 37, pp. 8167-8197.

Mustafa, N., Ling, L. S. & Razak, S. F. A., 2021. Customer churn prediction for telecommunication industry: A Malaysian Case Study. *F1000 Research,* 10(1274).

Schultz, E., 2023. Infographic: 30 Must-Know Sales Prospecting Stats and What They Mean for Sellers*. Rain Group.* [Online] Available at: https://www.rainsalestraining.com/blog/infographic-30-sales-prospecting-stats-and-what-they-mean-for-sellers [Accessed 18 Jun 2023].

Sharma, V., Joseph, S. & Poulose, J., 2018. Determinants of consumer retention strategies for telecom service industry in Central India. *Problems and Perspectives in Management,* 16(2), pp. 306-320.