**Executive Summary**

**Problem Statement**

To achieve the goal of getting better understanding of our customers and identifying who are most likely to respond positively to the next marketing campaign, this project will be mainly focused on customer profiling by using k-means cluster and predictive analysis by building logistic regression model.

**Exploratory Analysis**

Chart, bar chart

Description automatically generated

The graph above shows the percentage of customers shop via the different channels. The most popular channels are store and web. In the data dictionary, number of visits to company’s website in the last month is also included in the purchase channels (we didn’t include it in the graph because it is not actually a channel). In the last month, a total of 11,909 visits have been made, and 9,150 of orders were made through company’s website.

Chart, bar chart

Description automatically generated

The graph above indicates the proportion of product categories that are purchased from the retailer. As we can see, in terms of amount spent, wines and meat are the most purchased product categories.

**Demographic Analysis**

Chart, histogram

Description automatically generatedChart, bar chart, histogram

Description automatically generated

Graphs shown above demonstrate the age and marriage status of our consumers. Based on the left graph, most of our customers are between 40 and 60 years old. According to further calculations, the average age of our customers is 53, and the median age of our customers is 52. There are also some outliers over 110 years old and this might be caused by inaccurate data inputs during information collection. The distribution of responses to the last marketing campaign is similar to that of age, from which the relationship between age and response cannot be inferred.

As we can see in the right graph, most of our customers are married or have partners. In total of 2,240 records, 77 of them are widows and 480 of them are single, and very few of them belong to other categories such as alone, YOLO, or absurd. Proportionally, it seems that single people have more responses to marketing campaigns, so in the process of building the model we need to consider the impact of marital status on response.

In conclusion, our main consumers are Quinquagenarian who are married or have partners. We might want to consider launching some family size products to satisfy their needs.

Chart, bar chart

Description automatically generatedChart, bar chart

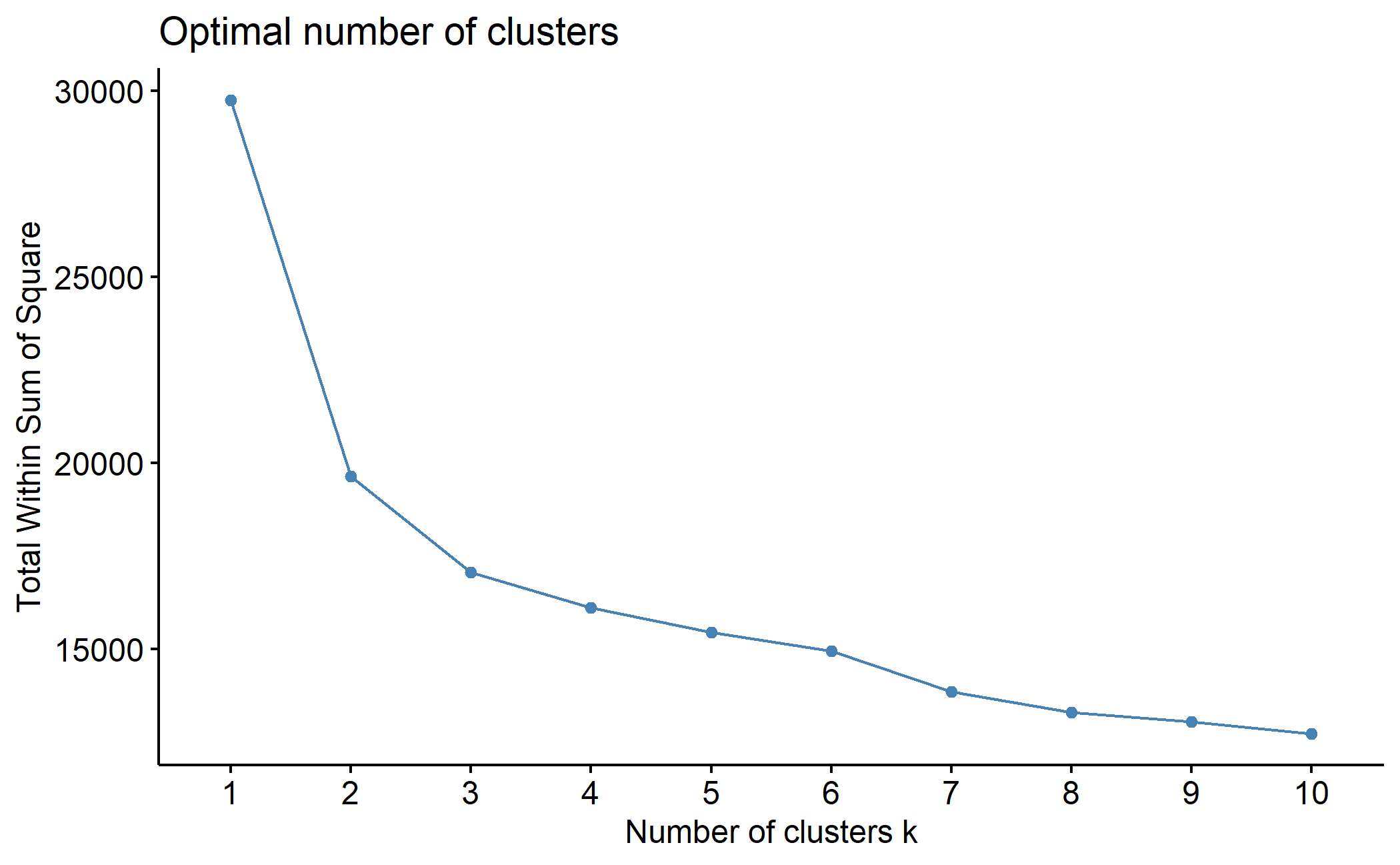
Description automatically generated

These two graphs showing above indicate the kids at home and education level of our customers. Since our main customers are in their 50s, most of them have no kids at home, and 899 of customers have 1 kid at home. In terms of education level, most of customers are at graduation level, and 856 of total have either a master’s degree or a PhD, which means our customers are mainly well-educated or highly educated people.

Proportionally, the above graphs do not imply the relationship between kids at home and education level and the response, which we will consider later in the model building session.

**Cluster Analysis**

By using fviz\_nbclust(method = “wss”)[[1]](#footnote-1) function in R, it could generate an elbow plot and let us visually choose the optimal number of clusters. With respect to the graph below, we choose k = 7 as the optimal number of clusters.



Since the cluster sizes are not uniform, we standardized the data before clustering and generated 7 clusters as following.

K-means clustering with 7 clusters of sizes 138, 710, 362, 249, 286, 332, 163.

Diagram

Description automatically generated

Text

Description automatically generated with medium confidence

Text

Description automatically generated with low confidence

Text

Description automatically generatedText

Description automatically generated

Cluster Profile

Cluster 1 – Gold Investors (average age of 56)

* Invest gold the most (average amount of spend in last 2 years is 167.38)
* Love online shopping
* Actively involved in the 3rd campaign
* Never complain in the last 2 years

Channel: E-mail

Offer: Online Coupon – for order over $300, get 5% off on gold

The average household income for cluster 1 is about $61,838.84. Although they are not highest income group, they spent a lot on gold in the last 2 years. They are also the most preferred cluster for online shopping – the average number of purchases made through the company’s website in this cluster is almost 7. They showed an unusual level of engagement in the 5th market campaign, we assume that the 5th campaign might involve what they are interested in – gold. If the company wants to attract this group to spend more, then we should make more marketing activities about gold.

Cluster 2 – Younger & Frugal (average age of 46)

* Youngest group with lowest income (yearly household income of $32,584.30)
* Have most kids at home (almost every family has a child at home)
* Don’t buy groceries or gold often (the lowest among 7 clusters)
* Don’t consume much at the company
* Not active in marketing campaigns (especially for 1st, 2nd, 5th)

We do not suggest offering discount or do market campaign to this group.

For cluster 2, they are not our main or target customers. People in this group are experiencing their 40s and might have a high school or college kid who hasn’t moved out yet. They don’t have much household income and they probably need to pay the tuition for their kids. Thus, they are frugal and don’t spend much on anything. They are not interested in deals or marketing campaigns (these activities are not attractive to them, and they might only buy things they need). We suggest the company do not take this group as the target consumers.

Cluster 3 – Not Significant in Any Categories (average age of 58)

* Not active membership for the company
* The second old group
* Weak spending/purchasing power
* Not interested in many things

Channel: Email

Offer: Reward e-coupon with 10% off for orders over $300

Cluster 3 is similar to the cluster 2, the difference is that they are older and earn slightly more. Since they are older, they don’t have many kids at home, and with an average household income of $61030.34, they are not even average spenders at the company. This group doesn’t seem to be attracted to marketing campaigns, but they do care a little bit more about deals. We don’t think this is our target consumers either, but if the company would give it a try, reach out them via email and offer some deals might motivate them to spend.

Cluster 4 – Wealthy & Big Spenders (average age of 54)

* Highest income with least kids at home
* Wines & meat lovers
* Don’t care about deals or discounts
* Catalogue buyers
* Acted positively in the 1st, 4th and 5th marketing campaigns

Channel: Mailer/Catalogue

Offer: Exclusive 15% off on wines and meat

With the highest household income of $81,490.18, cluster 4 is one of the groups with the strongest purchasing power. With the fewest kids at home, they are the middle class who enjoy their 50s. They are big spenders on wines and meat, and they don’t care about deals or discount. As our target consumers, they acted positively to almost every marketing campaign and love to spend money on the company’s products. The company should consider offering more marketing campaigns to motivate them to purchase.

Cluster 5 – Healthy Main Consumers (average age of 51)

* Buy the most fruit, meat, fish, sweets
* Shopping preferences span all channels (in-store the most)
* The second young group
* Made some complains in the last 2 years

Channel: SMS/Mailer

Offer: Catalogue coupon or use their membership card to get 15% off on fruit, meat, fish and sweets

Cluster 5 includes our loyal customers who buy most of stuff with the company (they are our target customers). Although they complained a lot, they continue to shop at us. We suggest the company to take their feedback and give some benefits back to them. The company could also send surveys to them and offer discounts after they finished.

Cluster 6 – Senior Citizens (average age of 61)

* Oldest among the groups
* Not main consumers (not into anything, buy everything little)
* Not interested in marketing campaigns
* Complained the most in the last 2 years

Channel: Mailer

Offer: Finish the customer satisfaction survey and get 10% off discount

Cluster 6 complained the most in the last 2 years, since they spend very little at the company, they might be customers that the company has lost. To win them back, we recommend the company to conduct a survey on them and collect the reasons why they don’t shop at us anymore. The company should make some changes to attract them to spend again.

Cluster 7 – Deal Chasers (average age of 53)

* Love to buy things online
* Discount lovers (average number of purchases made with a discount = 6.84)
* Have many kids at home
* Buy everything little

Channel: Email

Offer: E-coupon – 10% off with orders over $150

Cluster 7 seems to purchase only with a deal. In order to motivate them to spend, we recommend the company to give them some deals with a consumption threshold. Since they are not actively involved in the past marketing campaigns, the company should try to give them some discounts during marketing campaigns to see if this will attract them to purchase more.

Timeline

Description automatically generated

The graph above shows that, overall, cluster 4 and cluster 5 include a high percentage of customers who responded to positively to the previous campaign based on the variable response. As we have discussed before, these two groups are our loyal customers, and they are the target groups for the next marketing campaign. Although cluster 2 seems to have relatively larger amounts of people responded positively to the last campaign, due to the large sample base of cluster 2, the overall response of this group was not positive in a percentage term. The company should follow the above recommendations for each cluster, give back to loyal customers and attract potential consumers to spend more at their place.

**Model Report**

We built two logistic regression models for predicting customers’ response. We analyzed the overall performance of these models and find the most optimized one as the final model.

Full Model Analysis

Text

Description automatically generated

In the full model, we used all the variables except customer id, birth year and the register date of membership. To build a general model, we want to avoid unique attributes such as customer id and date of membership, and since we’ve created a new variable called age - simply using 2022 deduct birth year, we removed the birth year.

Table

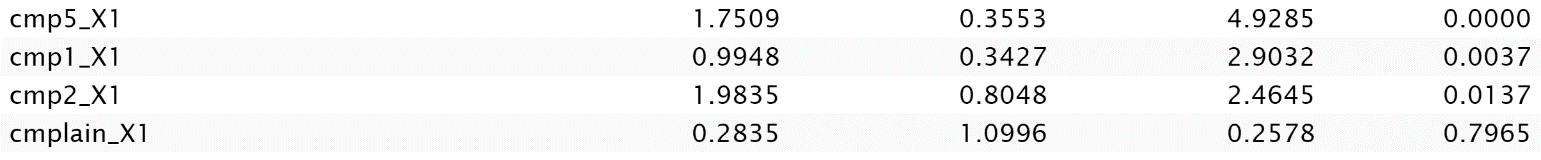
Description automatically generated

Table

Description automatically generated

Table

Description automatically generated



As graph above shows, many variables of the full model are not statistically significant with a p-value over 0.05, which means they are not useful estimators for predicting customer response.

Full Model ROC Chart[[2]](#footnote-2)

Chart

Description automatically generated

Full Model Confusion Matrix

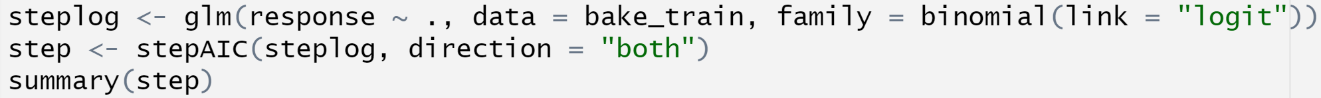
Chart

Description automatically generatedGraphical user interface

Description automatically generated with medium confidence

Reduced model

We used stepwise[[3]](#footnote-3) function to refine the model recipe and select useful determinants for predicting customer response.



After stepwise selection, we got our final recipe with all effective predictors and then apply them to our final model. The following images contain all the variables used in our final model, as we can see here, most of variables are statistically significant with a p-value less than 0.05. Although of them may have higher p-value than the threshold, since they are important to the model performance, we decided to keep them.

Table

Description automatically generated

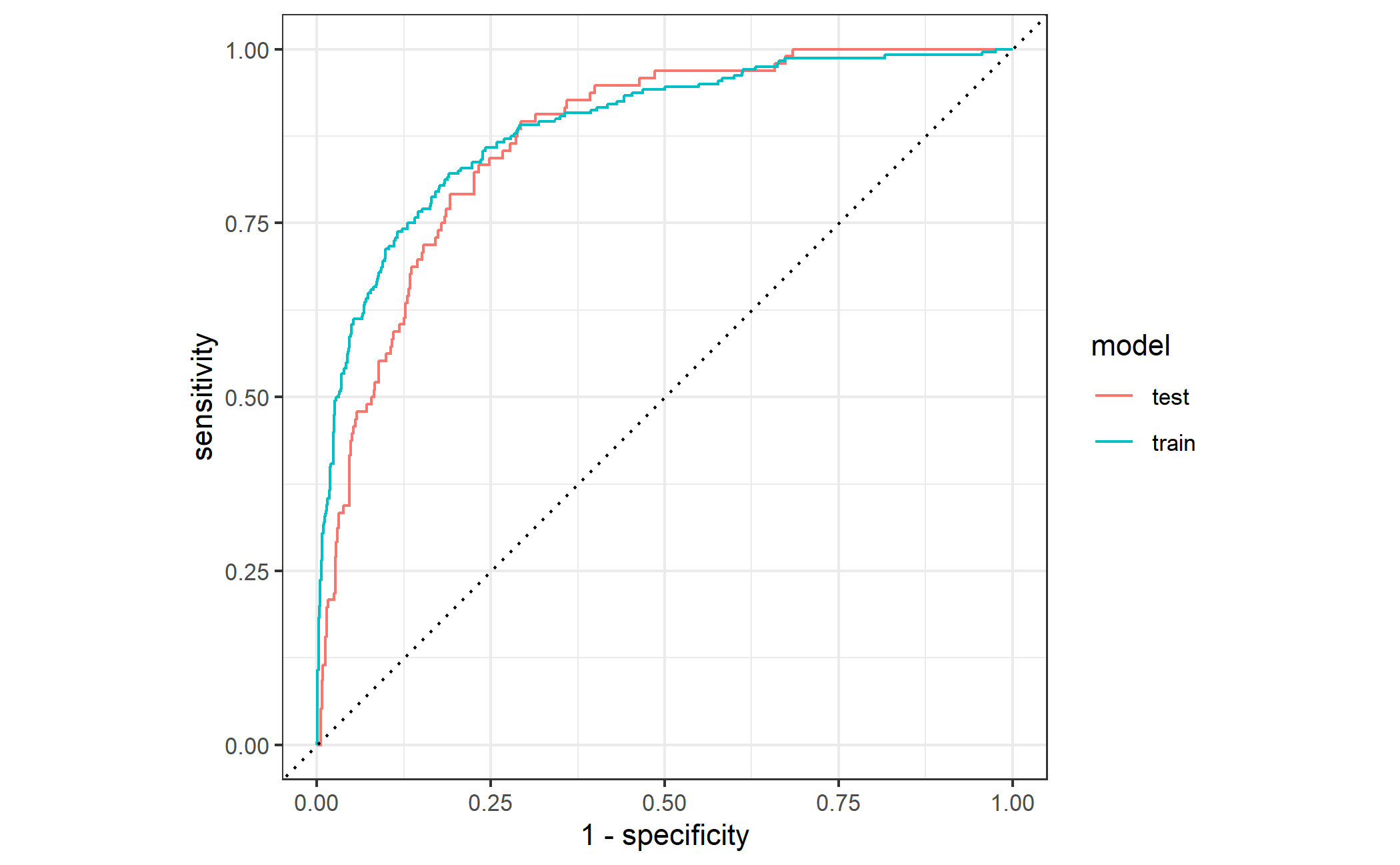
Table

Description automatically generated

A picture containing table

Description automatically generated

Final Model ROC Chart



Final Model Confusion Matrix

Chart

Description automatically generatedChart, bubble chart

Description automatically generated

Model Comparison

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Partition | AUC | Accuracy[[4]](#footnote-4) | Precision[[5]](#footnote-5) | Recall[[6]](#footnote-6) |
| Logistic Full Model | train | 0.8894798 | 0.8985969 | 0.7612903 | 0.4916667 |
| test | 0.8652344 | 0.8690476 | 0.5540541 | 0.4270833 |
|  |  |  |  |  |  |
| Logistic Final Model | train | 0.8890939 | 0.8998724 | 0.7677419 | 0.4958333 |
| test | 0.8703704 | 0.8779762 | 0.5972222 | 0.4479167 |

Based on the chart above, the final model performs the best as it has the highest test scores for AUC, accuracy, precision and recall. An AUC = 0.8704 means that the model is 87.04% accurate in distinguishing positive and negative responses. An accuracy of 0.8780 demonstrates that we predict correctly for 87.80% of customer responses. A precision = 0.5972 tells that for all positive responses cases we predicted, 59.72% of them are correctly identified. A recall = 0.4479 indicates that we can correctly identify 44.79% of all positive responses that actually happened.

Besides, the final model is better because it is much simpler and the gap between its training and test dataset is smaller than the full model – which means it is less overfitting than the full model.

But there are also some issues with the final model since it’s precision and recall are not considered as decent. This might be due to a skewed dataset where most of its records are negative responses to marketing campaigns. In order to further develop and refine the model, we highly recommend the company to collect more data on positive responses so that we could learn more about the attributes of those customers who are actively involved in marketing campaigns.

**Prediction Analysis**

After applying the final model to the new customer market file to predict potential responses of customers towards the next marketing campaign, in a total of 40 people, 35 of them respond passively to the marketing campaign, while 12.5% of customers act positively. Here is the recommendation for each customer for the next campaign (predicted response 0 – no, 1 – yes):

|  |  |  |
| --- | --- | --- |
| Predicted Responses | Customer ID | Recommended Classification |
| 0 | 111 | Do not contact |
| 0 | 112 | Do not contact |
| 0 | 113 | Do not contact |
| 0 | 114 | Do not contact |
| 1 | 115 | Contact |
| 0 | 116 | Do not contact |
| 0 | 117 | Do not contact |
| 0 | 118 | Do not contact |
| 0 | 119 | Do not contact |
| 0 | 120 | Do not contact |
| 0 | 121 | Do not contact |
| 0 | 122 | Do not contact |
| 0 | 123 | Do not contact |
| 0 | 124 | Do not contact |
| 0 | 125 | Do not contact |
| 0 | 126 | Do not contact |
| 0 | 127 | Do not contact |
| 0 | 128 | Do not contact |
| 0 | 129 | Do not contact |
| 0 | 130 | Do not contact |
| 1 | 131 | Contact |
| 0 | 132 | Do not contact |
| 1 | 133 | Contact |
| 1 | 134 | Contact |
| 0 | 135 | Do not contact |
| 0 | 136 | Do not contact |
| 0 | 136 | Do not contact |
| 0 | 138 | Do not contact |
| 0 | 139 | Do not contact |
| 0 | 140 | Do not contact |
| 0 | 141 | Do not contact |
| 0 | 142 | Do not contact |
| 0 | 143 | Do not contact |
| 0 | 144 | Do not contact |
| 0 | 145 | Do not contact |
| 0 | 146 | Do not contact |
| 1 | 147 | Contact |
| 0 | 148 | Do not contact |
| 0 | 149 | Do not contact |
| 0 | 150 | Do not contact |

Based on our predictions, the company should contact customers with ID of 115, 131, 133, 134, 147 for the next marketing campaign as they are more likely to response actively.

1. Nbclust() is a function that compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters. For each k, calculate the total within-cluster sum of square (wss). Plot the curve of wss according to the number of clusters k. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters. [↑](#footnote-ref-1)
2. An ROC curve entails how capable the model is at distinguishing between results, being true negative, false negative, true positive, or false positive. As ROC\_AUC approaches 1, the model is more likely to predict a positive as positive, and a negative as negative. [↑](#footnote-ref-2)
3. Stepwise selection reduced the complexity of the model without compromising its accuracy. It performs model selection by AIC. It has an option called direction, which can have the following values: “both”, “forward”, “backward”. Here we used “both” as direction and it will help us remove useless variables/estimators that do not improve the model performance. [↑](#footnote-ref-3)
4. Accuracy is one metric for evaluating models, it measures what percentage of the total sample did we predict accurately. [↑](#footnote-ref-4)
5. Precision is a percentage of true positive among all predicted positive cases. [↑](#footnote-ref-5)
6. Recall is a percentage of true positive among all actual positive cases. [↑](#footnote-ref-6)