

Group 27 – Final Report

Introduction/Background

The dataset we used for our project is provided by iFood, a leading food delivery app in Brazil similar to Door Dash in the United States. The data provided by iFood is a list of customers that iFood tried to sell to through previous marketing campaigns. The attributes cover several key areas ranging from customer profile information, product preferences, historical campaign successes/failures and channel performance for these tactics.

Objective

Our objective was to predict the success of a future marketing campaign and use that information to better target marketing results that will lead to higher overall sales based on campaigns. We looked at multiple probabilistic classification and supervised learning techniques to analyze this dataset. We used three core models: Logistic Regression, Random Forest, and Naïve Bayes. Some of the questions we explored are:

1. What is the relationship between customer attributes and response to product-specific campaigns?
2. Does the response to marketing campaigns vary between customers buying products online vs. store?

Problem Statement

Can we predict the success of the last marketing campaign based on customer attributes utilizing probabilistic models?

Hypothesis

If we can develop a probabilistic machine learning model with high accuracy, we can better target campaigns to people who have a high likelihood of response, which will increase overall sales conversion rates and total sales by campaign.

Literature Survey

We surveyed multiple articles related to our problem statement. These articles helped determine which attributes will be important for our analysis. [Nike Just Do It\[1\]](#) is the case study of Nike's "Just do it" campaign. The campaign focused on tying the brand to human capability to achieve. It is a potential research article on customer-centric marketing campaigns. [International Marketing Fails](#) focused on how campaign strategies varied across international markets. In [The Difference in Marketing for a retail store vs. e-commerce,\[2\]](#) we found the difference in marketing campaigns for online and brick-and-mortar stores to attract customers. In comparison, [Amazon Framework Analysis STP Case Study\[3\]](#) is a case study on the relevance of customer attributes in creating models for customer segmentation, positioning, and targeting.

Description of datasets

The dataset, [Marketing Analytics](#), was used in a Kaggle competition and is based on data from [iFood](#), a popular food delivery app in Brazil. The dataset contains meta-information about customers and

interactions with those customers from previous iFood marketing campaigns. The response variable included in the dataset indicates (y/n) if the customer responded to a recent marketing campaign, and it includes response variables for five previous campaigns for comparison.

The dataset includes 28 variables for 2,240 customers. These variables provide insights into:

- Customer profiles
- Product preferences
- Campaign success/failures
- Channel performance

These are the attributes we believe will be most important to our analysis:

- Age
- Income
- Country
- Product purchase-related variables
- Education

Note: *Between the project proposal and progress report, the dataset has been changed on Kaggle. Though most of the columns are the same, the key differences are:*

- *Year_Birth was changed to Age*
- *Country was removed*
- *There are 35 fewer observations*

Approach and Methodology

Methodology

We set out to build a model that could predict the likelihood of a prospect/current customer responding to a future marketing campaign with high accuracy. We focused on three probabilistic machine learning methods: Logistic Regression, Naïve Bayes Classifier, and Random Forest. These models have been founded on probabilities, and Random Forest uses majority voting among all its trees to predict the class of any data point provided. Having a model that outputs probabilities will allow us to identify and prioritize whom we may want to market-specific campaigns to in the future.

To assess model performance, we split the original dataset into train, validation, and test sets with 75%, 12.5%, and 12.5% split, respectively. This will allow us to fit our model to the dataset, select a model based on performance versus the validation set and then see how the model would perform in a real-world scenario holding back one last set of data for testing.

Approach

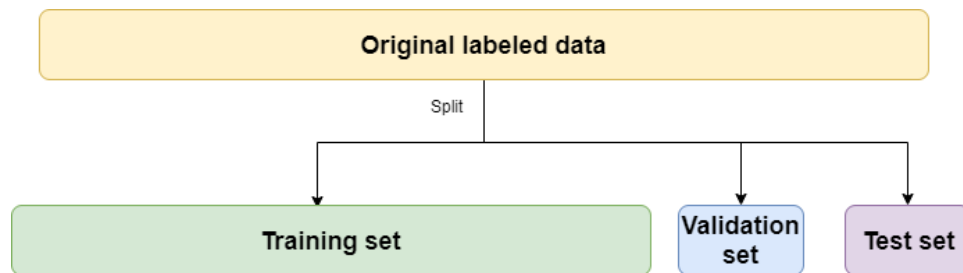
The various models were tested using the following metrics:

3. Accuracy – correctly predicted responses / total responses
4. Question answered: Out of all of the responses, how many did we correctly predict?
5. $(TP + TN) / (TP + FP + FN + TN)$
6. Precision – true positive values / total predicted true values

7. Question answered: Out of all the responses we predicted would respond, how many responded?
8. $TP / (TP + FP)$
9. Recall – also known as sensitivity, correctly predicted true positive responses / actual true positives
10. Question answered: Out of all responses, how many did we correctly capture?
11. $TP / (TP + FN)$

Steps:

12. Assess model performance through the following process flow

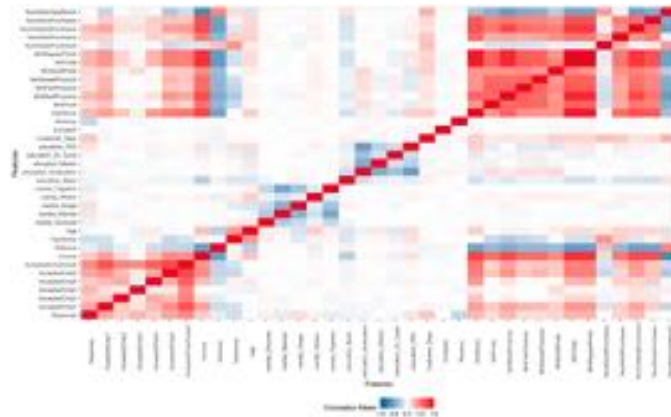


- a. Train
 - i. Optimize for accuracy
 - ii. Test with and without cross-validation and hyperparameter tuning to reduce overfitting and ideally return higher performance
 - b. Validation
 - i. Accuracy
 - ii. Precision
 - iii. Recall
 - c. Test
2. Prioritize prospective/current customers whom we think will respond to a future marketing campaign

Data Cleaning and Prepping:

The data set is already well-prepared for analysis, most variables are numeric/binary, and there is one record per customer. We tried different combinations of variables in the initial experimentation phase depending on the model.

We ran a basic correlation plot to help narrow down important features. This analysis showed that factors like education and marital status might have little impact on our model, and we should focus more on purchase history and response to other campaigns.



Assumptions

There are a few education-related columns for which the data dictionary is unavailable. These column names are as below. We made the following assumptions while interpreting the following columns.

- **education_2n Cycle:** Graduate level course work
- **education_Basic:** High-school level graduation
- **education_Graduation:** Graduation from a 4-year degree program
- **education_Master:** Graduation from a 2-year postgraduate program
- **education_PhD:** Doctorate level education

Synthetic Variables

After a preliminary analysis of the data, we think that creating a few additional explanatory variables might simplify the analysis and reduce the dimensionality of the models. We will create these variables in the manner specified below and test them out in our models to see if it yields better results.

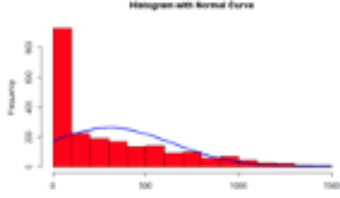
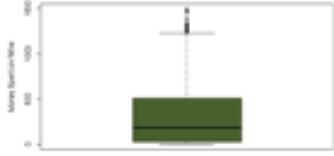
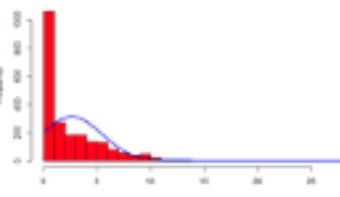
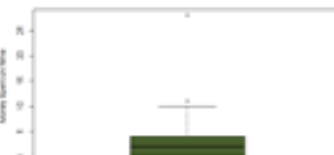
1. **TotalSpend:** Total amount spent on wines, fruits, meat, fish, and gold products
2. **TotalPurchases:** Sum of deal, catalog, store, and web purchases.
3. **TotalDependents:** Sum of columns Kidhome and Teenhome, which indicates the total number of dependent kids in the household

Variables that need to be excluded

Preliminary analysis also revealed that columns Z_ConstContact and Z_Revenue are of no use in analysis as they contain constant values of 3 and 11, respectively, across all rows.

Outlier Detection

To understand the data better, we ran the Shapiro-Wilks test and Grubbs test* to understand the normality of the data and if there are any outliers. We noticed several columns that do not have normally distributed data and/or have few outliers. A few examples are given below.

<p>Shapiro-Wilk normality test</p> <p>data: df\$MntWines W = 0.8399, p-value < 2.2e-16</p> <p>Grubbs test for one outlier</p> <p>data: df\$MntWines G = 3.51661, U = 0.99439, p-value = 0.4737 alternative hypothesis: highest value 1493 is an outlier</p>		
<p>Shapiro-Wilk normality test</p> <p>data: df\$NumCatalogPurchases W = 0.84235, p-value < 2.2e-16</p> <p>Grubbs test for one outlier</p> <p>data: df\$NumCatalogPurchases G = 9.05961, U = 0.96274, p-value < 2.2e-16 alternative hypothesis: highest value 28 is an outlier</p>		

** We understand that for Grubbs test to yield correctly, the assumption of normality must be first met.
This exercise was done to get a general sense of distribution across all columns*

Exploratory Analysis

Principal Component Analysis

Given the high number of variables, we wanted to see if Principal Component Analysis could help reduce the dimensionality of the problem statement. Hence, we built a PCA model that regressed on Response using other explanatory variables. We noticed some reduction in the number of columns as the first 22 columns obtained in PCA analysis were able to explain ~95% variance in the data. Since the reduction in the number of variables was not significantly helpful, plans to develop further on PCA and to integrate it with other models were dropped.

Multiple Regression Model

This model will help understand the associative relationship between predictor variables/customer attributes and response to marketing campaigns/total sales. After data cleaning and adding new variables, we regressed total sales on customer attributes like Education, Marital Status, Income, Kids at home, Teens at Home, and Age to find out the statistical significance of these predictor variables. VIF of these predictor variables was also calculated to detect multicollinearity. The model gave an R-squared of 0.73, which means these customer attributes can explain a 73% variation in total sales. These results gave us a preliminary idea about the predictors that might be statistically significant in predicting customers' responses to campaigns.

Logistic Regression

We built a logistic regression model to ensure that we are headed in the right direction to answer our research questions. We created a binary variable based on response to the first five marketing campaigns. Then iteratively, we built logistic regression models starting with predictors at a significance level of 0.05 and ending at a significance level of 0.01. For each of these models, we predicted a response to the next campaign with a classifier threshold of 0.5 and compared our results against the actual response to the sixth campaign. The model gave 90% accuracy in predicting the response to the sixth campaign. The model might be overfitting because of the many predictor variables in play and might have higher variance with the net new data set. To address this issue, we decided to use stepwise variable selection to reduce the number of predictor variables and achieve a tradeoff between bias and variance.

Baseline Models

Logistic regression from our exploratory analysis and random forest were natural choices for baseline models. Random Forest was the best choice among decision tree models, given the substantial number of variables, which was a suggestion we received from TAs on our proposal. The idea behind Logistic Regression was that even though Response is a binary variable, having a probabilistic model will give us the control to adjust the threshold and thus gives us a realistic way of incorporating risk tolerance of executives of the company. Instead of solely relying on one classification model, we also wanted to try out another classification model which would provide us with probabilities, and as such, we went with Naïve Bayes because it is a simple model to construct. It also provides hyper tuning capabilities, which is relevant for our dataset since we have a relatively large number of predictors and a smaller number of rows.

Random Forest

Given the substantial number of variables and the possibility of highly correlated variables in clusters, we wanted to see if the Random Forest model could perform better than other models. We built a simple RF model with the same training dataset consistent with other models. The Random Forest approach was undoubtedly convenient. There was no need for cross-validation or a separate test set to get an unbiased estimate of the test set error since it is estimated internally during the run.

Naïve Bayes

The purpose of testing a baseline version of the NB versus a hyper-tune version is to see if accuracy can improve model performance. We selected the caret package specifically because it allows for k-fold cross-validation of our dataset that will keep our model from overfitting against the data. K-fold cross-validation is essential in this analysis since the number of observations is small, and the k-fold technique will address it.

Additionally, the baseline models have cross-validation, so if hyperparameter tuning leads to no additional accuracy, we can be confident that the baseline model will still not overfit.

Model Used:

- Naïve Bayse – nb.cv() function

Cross Validation:

- R's Caret Package
- 5 K-fold CV
- Repeated 2 times

Optimized Models

Naïve Bayes: Hyperparameter Tuning

We wanted to tune our Naïve Bayes model to attempt to optimize it for improved performance.

We used cv as a starting point from our baseline model for our parameter tuning and then added gridsearch functionality of R's Caret package.

Features included in our hyperparameter tuning:

- Userkernel = if true use a kernel density estimate for continuous variables versus a Gaussian density estimate if false
- laplace = provides an additive smoothing effect which is to help with the zero-probability issue that is common with naive Bayes
- adjust = selecting the bandwidth for kernel density
- small bw leads to under smoothing
- large bw leads to over smoothing

After gridsearch was successfully run, it output the optimal parameters to maximize accuracy, which were the following:

Accuracy was used to select the optimal model using the largest value.

The final values used for the model were

- laplace = 1
- usekernel = TRUE
- adjust = 0.25

Stepwise Regression

The underlying goal of Stepwise Logistic Regression is to find a set of independent variables that significantly influence the dependent variable (Response). We used both forward stepwise and backward stepwise regression to identify the variables that best explain our response variable. After identifying the variables, we created two models using those variables and predicted the accuracy of each model.

K-fold(K=10) cross validation along with Stepwise

Further, we explored K-fold(K=10) cross-validation along with Stepwise. We created two k-fold models from each set of variables obtained from forward and backward stepwise logistic regression and predicted the accuracy of each model.

Assumptions: We have kept a threshold value of 0.5 for predicting probability. Any prob higher than 0.5 is considered as 1 and lower than 0.5 is considered as 0.

All the models were trained on the same training dataset (75%), consistent with other models. It should be noted that in all the models, we are treating Response as a factor.

Results

We observed that adding synthetic variables to the mix did not help improve the accuracy in any significant way. Hence after experimenting independently on the models, the team decided to drop the use of synthetic variables in the final version of our models used in the model selection process.

Below is the table of results we obtained on the test dataset with all the models we built.

Model	Precision	Recall	Accuracy
Basic Naives Bayes	90.5%	85.9%	80.3%
Optimal Naive Bayes	90.0%	89.3%	82.5%
Stepwise Forward Logistic Regression	89.5%	98.3%	88.72%
Random Forest	88.93%	96.15%	86.54%

Below is the table of results we obtained on validation dataset with the final selected model.

Model	Precision	Recall	Accuracy
Stepwise Forward Logistic Regression	92.27%	98.76%	91.67%

Naïve Bayes: With Naïve Bayes we achieved 80.3% accuracy.

Random Forest: Looking at the importance plot of model_rf we can say that the most significant predictor variables are AcceptedCmpOverall, Recency, Customer_Days, MntMeatProducts, Income, MntTotal. Our intuition before the model was that positive response to previous campaigns and income would be one of the top predictors along with education. From the graph, it looks like barring education the intuition was true. The random forest model achieved 86.54% accuracy on the test dataset.

Naïve Bayes: Hyperparameter Tuning: With this model we achieved 82.5% accuracy. Interestingly we didn't obtain any significant improvement in results with hyper-tuning the parameters.

Logistic Regression:

Stepwise Regression

- marketing_data_fullmodel – We created this model for performing simple logistic regression on all the variables in our dataset. We achieved an accuracy of 89.45% for this model.

- **stepwise_backward**- We created this model from the variables obtained from backward stepwise regression. We achieved an accuracy of 87.27% on the test data set.
- **stepwise_forward** - We created this model from the variables obtained from forward stepwise regression. We achieved an accuracy of 88.72% on the test data set.

K-fold(K=10) cross validation along with Stepwise:

- **kfold_backward**- We applied 10-fold cross validation to the variables obtained from backward stepwise regression and created this model. We achieved an accuracy of 87.27% on the test data set.
- **kfold_forward** - We applied 10-fold cross validation to the variables obtained from forward stepwise regression and created this model. We achieved an accuracy of 89.09% on the test data set.

Logistic regression helped us fine tune our models systematically and gave us a set of statistically significant variables that do the best job of predicting the Response variable.

We achieved the highest accuracy of 91.67% from forward stepwise regression model tested on our test data set. The set of variables which are statistically significant as per this model is AcceptedCmpOverall, Recency, NumDealsPurchases, AcceptedCmp4, NumStorePurchases, Customer_Days, marital_Married, marital_Together, education_PhD.

Final Model Selection: Based on the results obtained we selected logistic regression with forward stepwise variable selection as our final model. We further tested the final model on test dataset, and we obtained 91.76% accuracy.

Interpretation of results

After iterating through multiple models, logistic regression gave the highest accuracy especially the one with forward variable selection. The interpretation of the coefficients of this model is as following:

- **AcceptedCmpOverall** – This predictor shows the total number of responses to campaigns in the past. The log odds of responding to a new campaign increase by 1.882 with one unit increase in overall campaign acceptance in the past, holding all other predictors constant OR the odds of responding to a new campaign increase by 556% with one unit increase in overall campaign acceptance in the past, holding all other predictors constant. ***Customers who have frequently responded to campaigns in the past will have a higher probability of responding to future campaigns because of their affinity towards the app and to get food delivered at their doorstep.***
- **Customer_Days** – This predictor reflects the total number of days a customer has been buying from the company. The log odds of responding to a new campaign increase by 0.004442 when the customer has been with the company for one more day, holding all other predictors constant OR the odds of responding to a new campaign increase by 0.44% when the customer has been with the company for one more day, holding all other predictors constant. ***Customers***

with longer tenure usually stick to the tried and tested delivery options and campaigns will decrease retention turnover.

- **Recency** – This predictor tells the number of days since the last purchase. The log odds of responding to a new campaign decreases by 0.02891 with one day increase in the days since last purchase was done, holding all other predictors constant OR the odds of responding to a new campaign decreases by 2.84% with one day increase in the days since last purchase was done, holding all other predictors constant. ***Customers who buy frequently through the app are more likely to respond to a campaign as they might be inclined towards ordering food from outside and campaigns will increase retention.***
- **Teenhome** – This predictor indicates the number of teens in the household. The log odds of responding to a new campaign decrease by 1.043 with one additional teen, holding all other predictors constant OR the odds of responding to a new campaign decreases by 64.76% with one additional teen, holding all other predictors constant. ***Teens are tech savvy and will be trying out multiple apps providing the same service and as such the probability of responding to campaigns decreases.***
- **marital_Married1** – This categorical predictor states whether customer is married or not. The log odds of responding to a new campaign decrease by 1.508 for customer who is married compared to customer who is not married, holding all other predictors constant OR the odds of responding to a new campaign decreases by 77.86% for customer who is married compared to customer who is not married, holding all other predictors constant. ***Customers who are married tend to eat out less and as such it makes sense why the response to campaign will be on the lower side.***
- **marital_Together1** - This categorical predictor states whether customer is with a partner or not. The log odds of responding to a new campaign decrease by 1.255 for customer who has a partner compared to customer who does not have a partner, holding all other predictors constant OR the odds of responding to a new campaign decreases by 71.49% for customer who has a partner compared to customer who does not have a partner, holding all other predictors constant. ***Customers with partners tend to eat out less and as such it makes sense why the response to the campaign will be on the lower side.***
- **education_PhD1** - This categorical predictor states whether a customer has a PhD or not. The log odds of responding to a new campaign increase by 1.009 for customer who has a PhD compared to customer who does not, holding all other predictors constant OR the odds of responding to a new campaign decreases by 174% for customer who has a PhD compared to customer who does not, holding all other predictors constant. ***Customers with advanced***

degrees might be busier in their lives and as such the likelihood of ordering food from outside is high. Campaigns will attract customers in this profile.

- **NumDealsPurchases** - This predictor indicates the number of deals purchased by the customer. The log odds of responding to a new campaign increase by 0.2448 with one additional unit of deal purchased, holding all other predictors constant OR the odds of responding to a new campaign increase by 27.73% with one additional unit of deal purchased, holding all other predictors constant. ***Customers who have frequently responded to campaigns in the past will have a higher probability of responding to future campaigns because of their affinity towards the app and to get food delivered at their doorstep.***
- **AcceptedCmp41** - This categorical predictor states whether a customer responded to campaign# 4 or not. The log odds of responding to a new campaign decrease by 1.018 for customer who responded to campaign# 4 compared to customer who did not, holding all other predictors constant OR the odds of responding to a new campaign decreases by 63.86% for customer who responded to campaign# 4 compared to customer who did not, holding all other predictors constant. ***This is an interesting predictor. Without knowing all the details, one thing we can infer is that the terms of campaign# 4 were not favored upon by customers and hence should be avoided in the future.***
- **MntMeatProducts** - This predictor indicates the amount of meat products purchased by the customer. The log odds of responding to a new campaign increase by 0.002363 with one additional unit of meat purchased, holding all other predictors constant OR the odds of responding to a new campaign increase by 0.24% with one additional unit of meat purchased, holding all other predictors constant. ***As compared to other products, customers who buy meat products frequently have a higher probability to respond to campaigns. This makes sense as meat is a staple food item.***
- **NumStorePurchases** - This predictor indicates the number of purchases made in store by the customer. The log odds of responding to a new campaign decrease by 0.1522 with one additional purchase done in store, holding all other predictors constant OR the odds of responding to a new campaign decrease by 14.11% with one additional purchase done in store, holding all other predictors constant. ***Customers who prefer to shop in brick-and-mortar stores will not be inclined to make orders through a food delivery app and will be less likely to respond to campaigns.***

Based on our analysis:

- *The company should be spending most of the campaign budget on customers who have a longer tenure with the company, are not married and are without teenage dependents.*
- *Customers with advanced education should be part of the target population.*

- *The campaigns should be focused on offering discounts for ordering meat products through the app.*
- *The campaigns should be targeted at customers who do not visit a physical store frequently.*
- *The company does not need to create opportunity cost by creating test and control groups for customers who fall in the above segment and instead should switch offer tactics for every campaign.*
- *Marketing should focus on doing an A/B test for customers who are outside of the above segmentation to validate the kind of offers that will attract them towards their app and to opt for food delivery. Multiple variations of the content, design, and placement of the campaigns can be tested to see the variation in responses.*

References

^[1] <https://www.linkedin.com/pulse/how-important-social-impact-brands-marketing-strategy-jay-sung/>

^[2] <https://www.thebalancesmb.com/compare-brick-and-mortar-stores-vs-online-retail-sites-4571050>

^[3] <https://github.com/nailson/ifood-data-business-analyst-test/blob/master/iFood%20Data%20Analyst%20Case.pdf>