# Applied Analytics Project: iFood Campaign

Omair Memon

## Project Overview

For this project, I have been tasked by a company named iFood to effectively analyze their data to predict customers who will respond to a new marketing campaign. If done correctly, this analysis can allow the company to maximize their profit by targeting customers who have a high likelihood of responding to the campaign, as well as not missing out on customers who would respond. Businesses commonly run into the problem of not targeting the correct customers during their campaigns, which causes them to miss out on retaining customers who demonstrate high probabilities of responding to said marketing campaign. iFood is willing to campaign towards more people, as long as it means that the company is capturing more people with a high probability of responding.

## Data Source and Preparation

The iFood dataset was sourced from Kaggle. My data cleaning process began on the data spreadsheet itself, where I changed binary variables to factors. After importing the new data set into my R studio, the bulk of my cleaning and preprocessing began. I started by looking at missing values in my data and I found that there were 24 missing values for the variable ‘income’. By looking at other variables, it is indicative that ‘Education’ would play a major role in the income amount that the customer generates. Therefore, instead of dropping these missing values, I filled them in by taking the median income of the specific education that the customer had. In addition, I created dummy variables for ‘education’ and ‘marital’ due to their categorical nature. For marital, the dummy variables introduced variables which had few observations such as ‘Alone’, ‘Absurd’, or ‘YOLO’, introducing instances of rarity. To reduce noise and the chance of our model overfitting, I grouped these variables based on the presence of kids or teens at home into more common marital occurrences such as ‘Single’ and ‘Married’.

With a dataset of this many variables, it was important to find respective variables that would not provide any insights into our analysis. Hence, a correlation matrix was created to find variables that had no actual correlation with ‘Response’. The correlation matrix below

shows us that the variable “NumWebPurchases’ has no correlation with ‘Results’, so I removed that variable from the data frame.

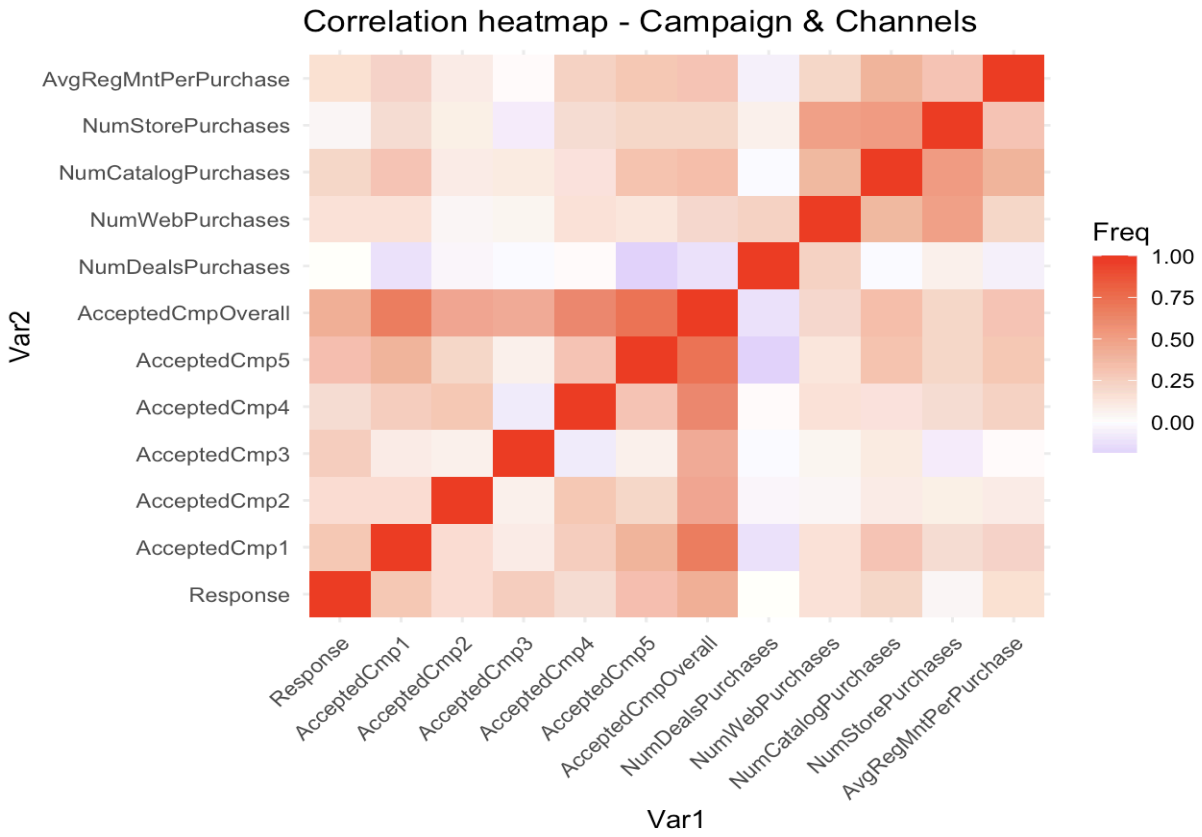


Figure 1: Correlation Matrix

## Model Selection and Rationale

After a thorough cleaning and preprocessing, I considered two primary machine learning models for the task of predicting whether a customer will respond to the campaign or not: logistic regression and ridge regression. Logistic regression was my initial choice due to its robustness in binary classification problems, such as the one at hand. Specifically, it allows for a probabilistic nature, or in other words, it provides the probability of each classification. This allows me to set different probability thresholds which will be a key factor later on. Furthermore, due to our extensive list of variables, I confidently felt that a model with regularization techniques would be necessary. Therefore, a ridge model was chosen for handling multicollinearity to penalize the size of coefficients and prevent overfitting.

## Feature Engineering and Selection

In addition to cleaning the data, there were numerous variables that were feature engineered into my model. Feature engineering was pivotal in enhancing my model's ability to predict positive customer response, as I will discuss further on. Overall, I feature engineered a total of 6 new variables. To start, I aggregated individual spending into a comprehensive 'Total Spending' metric, to help distinguish between everyday spending and luxury spending. To further define a split between luxury products and regular products, I engineered a variable that included the total amount of regular products. Another variable was created to display the total number of purchases to capture the frequency of customer interactions. I aggregated responses to various campaigns into one metric, providing a clear indicator of a customer's receptiveness to marketing initiatives. A higher score for this metric indicates a higher likelihood of a positive response to future campaigns, making it an invaluable predictor. Another variable was created to display averages per purchases to provide insights into spending patterns per transaction.

For the ridge regression model in particular, my thoughtful process of feature engineering reduced the risk of overfitting. In addition, it contributed to higher precision and recall in both models, as I will discuss in the next section. Finally, my feature engineered variables mitigated the effects of outliers and reduced dimensionality, which led to more accurate model predictions.

## Model Training and Validation

Overall, I created three different models, each of which had certain steps that differed it from the previous. It is important to note that for our models, the goal is to not simply just increase our accuracy. Remember, iFood is willing to dish out the campaign to more people, even if that increases the likelihood of false positives. The goal for the company is to capture as many true positives for responding campaigns so they can increase their sales. In other words, the goal of our models should be to maximize our recall, while attempting to maintain a relatively decent precision and accuracy score.

Furthermore, using the 'caret' package, all of our models allocated 80% of the data for the training model, with the remaining 20% reserved for testing. This validation technique allowed for an unbiased assessment of the model's capabilities at predicting on unseen data. The accuracy, precision, and recall scores enabled us to effectively understand our outcomes at having correct predictions for the results as a whole, and specifically true positives. There are more training methods for each distinct model that was used, however, I believe that it would

be more beneficial for the reader to describe these differences during the next section of the reading, to help them understand what is causing these models to perform differently.

## Results and Interpretation

My goal with this part of the report is to thoroughly explain the results of each model used, as well as why certain parametric decisions were made. The first model built, a simple logistic regression model, was constructed to observe the improvements in results from my comprehensive preprocessing steps, which were not used in Milestone 2's initial logistic regression model. My preprocessing and feature engineering proved to be successful, resulting in an accuracy of 87%, a precision of 45%, and a recall of 57%, indicating significant improvements from the Milestone 2 model. An 87% accuracy rate is impressive, as the model accurately predicts the response of each customer at a very high rate. However, our precision and recall is lower than I want, and iFood will miss out on tons of customers who would respond to campaigns. In any improvements that will be made to the model, we understand that precision and recall may have a tradeoff. In other words if we find a way to increase recall, then our precision will decrease because we will now have more false positives.

Nonetheless, iFood’s goal is to capture as many positive responding customers as possible. Therefore, the model should be tuned to force a higher recall rate, while attempting to maintain a solid accuracy, as we do not want the company to send the marketing campaign to all their customers as this will increase their costs substantially. My first idea to increase the recall rate was to tweak with the probability threshold of the logistic regression model. By lowering the probability threshold, theoretically, we should be able to capture more positive responses. However, this was not the case. After lowering the probability threshold to 0.25, our precision shot up to 90% indicating that we are accurate in predicting responses, but our recall shot down to 24%.

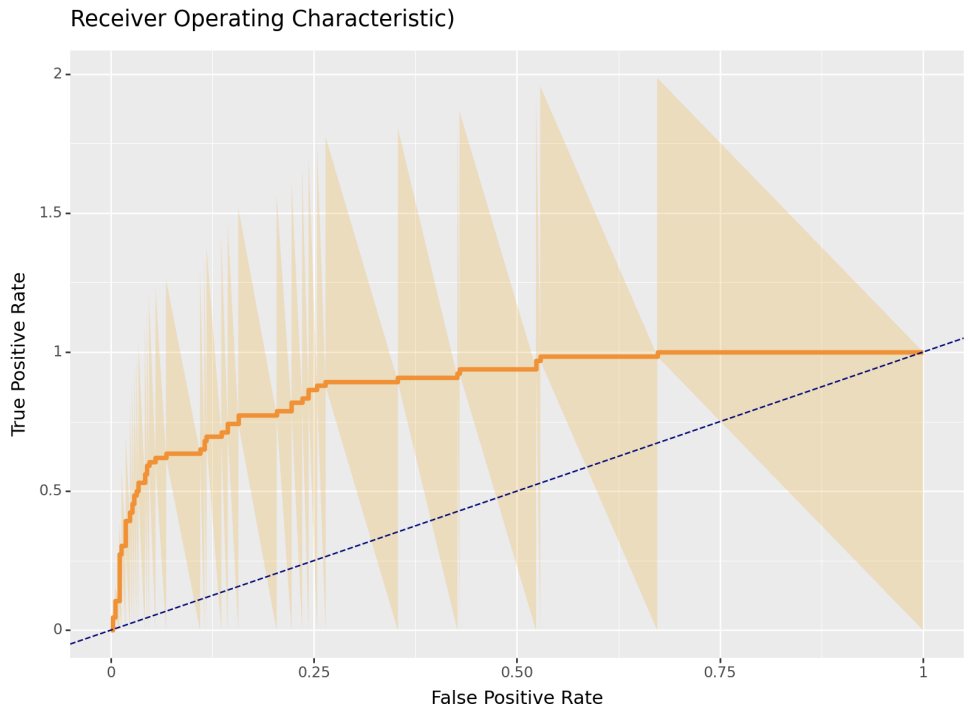


Figure 2: ROC-AUC Curve

My final model was a ridge model that incorporated certain techniques which enabled us to increase our recall to a satisfactory amount. Due to our abundance of variables, and newly feature engineered variables that may be causing multicollinearity, a ridge regression would prove to be beneficial in filtering out this dimensionality. Additionally, a deeper analysis of our data showed that the responses are imbalanced, as the majority of the data for responses is 'No'. To address this, an oversampling technique was used in our R code to make the class distribution more even. With an accuracy of 78%, a precision of 37%, and a recall of 70%, our ridge model proved to be more appropriate for iFood's situation. As you can see based on the ROC curve above, a high AUC score is shown indicating a high distinction between true positives and true negatives. Because of the resampling and dimensionality reduction of our model, we were able to provide a recall of 70% while maintaining a high accuracy. Although our precision was lowered, iFood will be able to capture the majority of campaign responders with this model. A confusion matrix with the model's predictions has been included below.

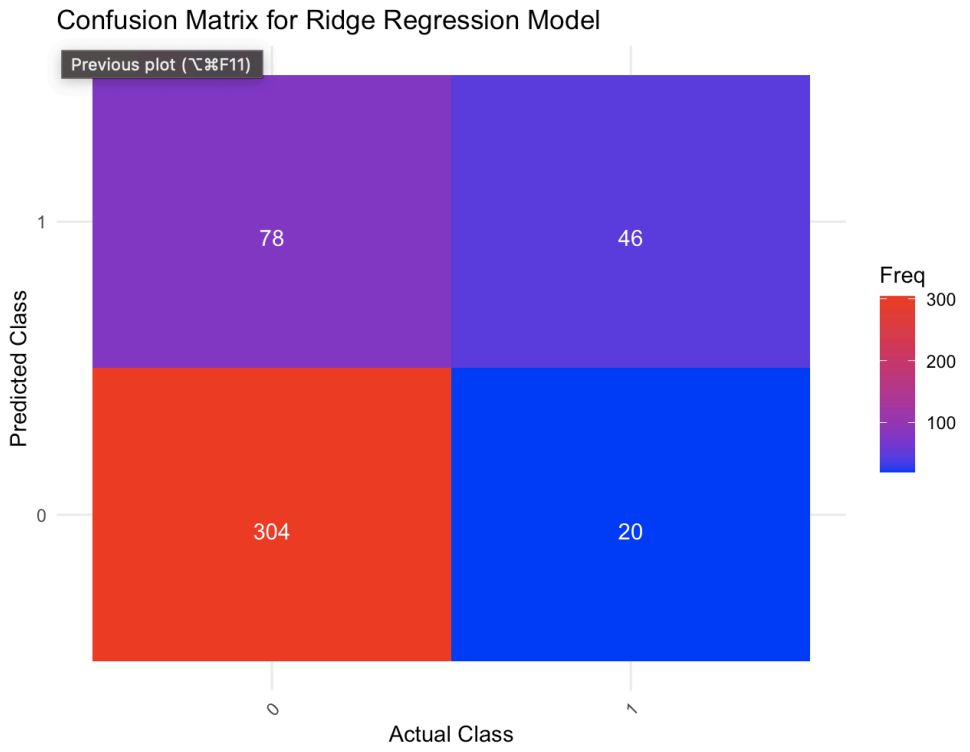


Figure 3: Confusion Matrix

## Lessons Learned and Challenges

There were many challenges that I faced during the preprocessing stage of this project. Specifically, thinking of variables to feature engineers as well as ways to fill in missing values and handle dimensionality required me to think outside of the box. I needed to create feature engineered variables that would play a role in improving the model, not just random variables that would play no role. Some of the methods that I employed are techniques that I had never done before this, and throughout this process, I learned a lot. Tuning my model to increase my recall posed many challenges. I firstly thought that simply tweaking with the probability threshold would significantly improve my recall, but this was not the case. This forced me to explore other methods such as resampling to achieve the desired result. Handling an imbalance data set was a significant challenge, but I was able to figure out the R script to do this. This project allowed me to incorporate most of what we learned throughout this course.

## Future Work and Improvement

Overall, I am extremely happy with the results of my model and the improvements that I made, however, in the future, I would like to employ advanced machine learning techniques such as neural networks when I have a better understanding of it conceptually. Also, I would like to use PCA to reduce the number of features while still retaining essential information. This could prove to be great for feature selection and dimensionality reduction, and may be an improvement from my ridge model. Finally, I would explore other resampling techniques such as stratified cross validation to address the imbalancement in our dataset. As said earlier, I am more than content with my model, but I plan to revisit these projects in the future and deploy these techniques that I have mentioned to find ways to improve my results.