Group 6: Megan Marini, David Ling, Amy Xu

**Assignment 2: Quick Kits Case and Data Dictionary**

QuickKits is a meal preparation and delivery service with nation-wide franchises. From the QuickKits website, customers choose a recipe and place the order for all the ingredients, sourced and chopped. QuickKits has four main categories of recipes, namely Healthy, Vegetarian, Meaty, and Specialty. Weekly meal plans are offered, with customers choosing from plans with two, three, four, or five meal deliveries per week. A delivery can consist of one or more meals. QuickKits recently decided to test an upselling promotional discount. On March 5, 2018, (2018/03/05), 10% percent of their customers on two-meal-per-week plans were offered a promotional discount of 40% for a three-meal-per-week plan, for two weeks. As the contribution margin for this service is approximately 20%, the promotion will run at a loss. At the end of the two-week period, subscribers to the promotion would remain on the three-week plan at regular price unless they opted out.

The first challenge is to build a predictive model of subscribers to the promotion, based on the 10% of customers that received the offer. Second, utilizing your model and other information in the case to build a business recommendation plan to the managers of QuickKits on how to improve their business.

Grading:

**Part1:** 70% on documentation of the model development, using the rapid model development framework.

1. Based on the background information, build and describe your mental model for subscription to the upselling promotion among the customers on two-meal-per-week plans. (10 points)

The objective is to maximize the profit by targeting highly potential two-meal-per-week plans subscribers who would likely subscribe to three-meal-per-week plans. We predict the three most important relationships will be customer loyalty, income, and family size.

1. **Customer Loyalty: The longer a customer has been subscribed to the QuickKits two-meal-per-week meal plan and the more frequently the more likely they will be to subscribe to the promotion.**

* If a customer has been subscribed to the meal plan for a longer time, it indicates they value and trust the brand. This trust would increase their willingness to try new things and promotions from the brand, or their willingness to support it by upping their number of meals.
* A long subscription time also indicates the customer enjoys the product enough to stick with it for a long time. This enjoyment would positively relate to wanting to increase the number of meals received per week.

1. **Income: People with higher income would be more likely to subscribe to the promotion than people with lower incomes.**

* People with higher incomes have more discretionary income and will be able to absorb the cost of an additional meal without much pain whereas lower income people may not be able to fit the extra meal in their budget, even with the promotion.

1. **Household Size: People with larger household sizes (3+) will be more likely to subscribe to the promotion.**

* For a person living by themselves or with one other person, buying more than 2 meals a week may be excessive. Depending on the meal size they may not be able to finish the meals in one week before the next set of deliveries come.
* On the other hand, people with large families/house sizes may be more receptive to receiving more meals as they will be spread over a larger number of people and they can see the monetary benefits from doing so.

1. Connect your mental model and the variables provided in the dataset; form a list (3~5 hypotheses) of hypotheses that can be tested using the data later; and briefly explain the rationale for each hypothesis. (10 points)
2. **Customer Loyalty**

* NumDeliv will be positively related to customer loyalty and positively related to the probability of subscribing to the promotion. The relationship will be linear.
  + If their deliveries are higher, it indicates they have been ordering consistently and thus are enjoying the company’s services.
* LastOrder will be positively related to the probability of subscribing to the promotion
  + If people ordered recently, they may want the upgrade compared to people who haven’t ordered for a while.
* MealsPerDeliv: People order many meals per delivery may have a higher probability of wanting to do three meals per week
  + People ordering many meals per delivery indicates they may keep some meals to the next day, and want to get the most out of their one delivery per week

1. **Income**

* DA\_Income will be positively related to the individuals income and positively related to probability of subscribing to the promotion non-linearly.
  + People with higher incomes have more spending money and flexibility to add another meal to their weekly delivery without feeling it financially.
  + People with lower incomes may not be able to afford the extra meal or prefer buying the rest of their weekly meals from the grocery store.
  + This relationship will increase at a decreasing rate as once an individual reaches a high enough level of income, they will not be as incentivized by discount promotions.

1. **Household Size**

* DA\_Under20 will be positively related to family size and positively related to the probability of subscribing to the promotion. The relationship will be linear.
  + It is likely that the “individual under 20 years of age” indicates there is a child in the household, which means the household would generally be 3+ people (mom, dad, children). Larger families, i.e. those with children, would be more likely to subscribe as they have a greater need for more meals.
* DA\_Single will be negatively related to household size and negatively related to the probability of subscribing to the promotion. The relationship will be linear.
  + People living by themselves have less need for a larger amount of meals than families. They may not be able to finish the 3 meals in the one-week window and it would be wasteful for them to subscribe.

1. Document your data cleaning steps (i.e., missing information, low frequencies in the factor variables, correlation, trivially related variables, case identifiers). (15 points)

**Missing information:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variables** | **NA %** | **Explanation** | **Solution** |
| DA\_Income, DA\_Over60, DA\_Single | 3.42% | These variables have a low NA rate, likely due to missing data on those Census Dissemination areas. | Delete cases with NA. This will not have a significant impact on the model since the NA rate is ~ 3%. |
| Title | 35% | The high NA rate is likely due to nonreporting of the member when they signed up. | Delete variable as title would not have an impact on probability of subscribing |
| Disc | 73% | This is meaningful missing data. NA means the customer is not senior or a student. | Create a new level turning all NAs into “NoDiscount” |
| Weeks3Meals | 83% | This relationship is trivially related to the target variable.Missing values indicate the customer did not subscribe to the meal plan. | Delete variable.  The variable should not be included in the model since it is trivially related to SUBSCRIBE. |

**Low frequencies in factor variables:**

* Instead of *LastOrder*, which has 121 levels that is not meaningful, we should use this factor to measure the time since the customer’s last order.
  + Solution: create a new variable, DaysSinceLastOrder, which is equal to Today’s date (2018-03-05) subtracted by LastOrder.
* *PostalCode* has 120 levels. With so many levels, interpreting each alone is likely not meaningful. Grouping them by region may be useful, but we do not have enough information or time in our rapid model development to do so.
  + Solution: exclude PostalCode from the model and use the DA variables.

**Correlation:**

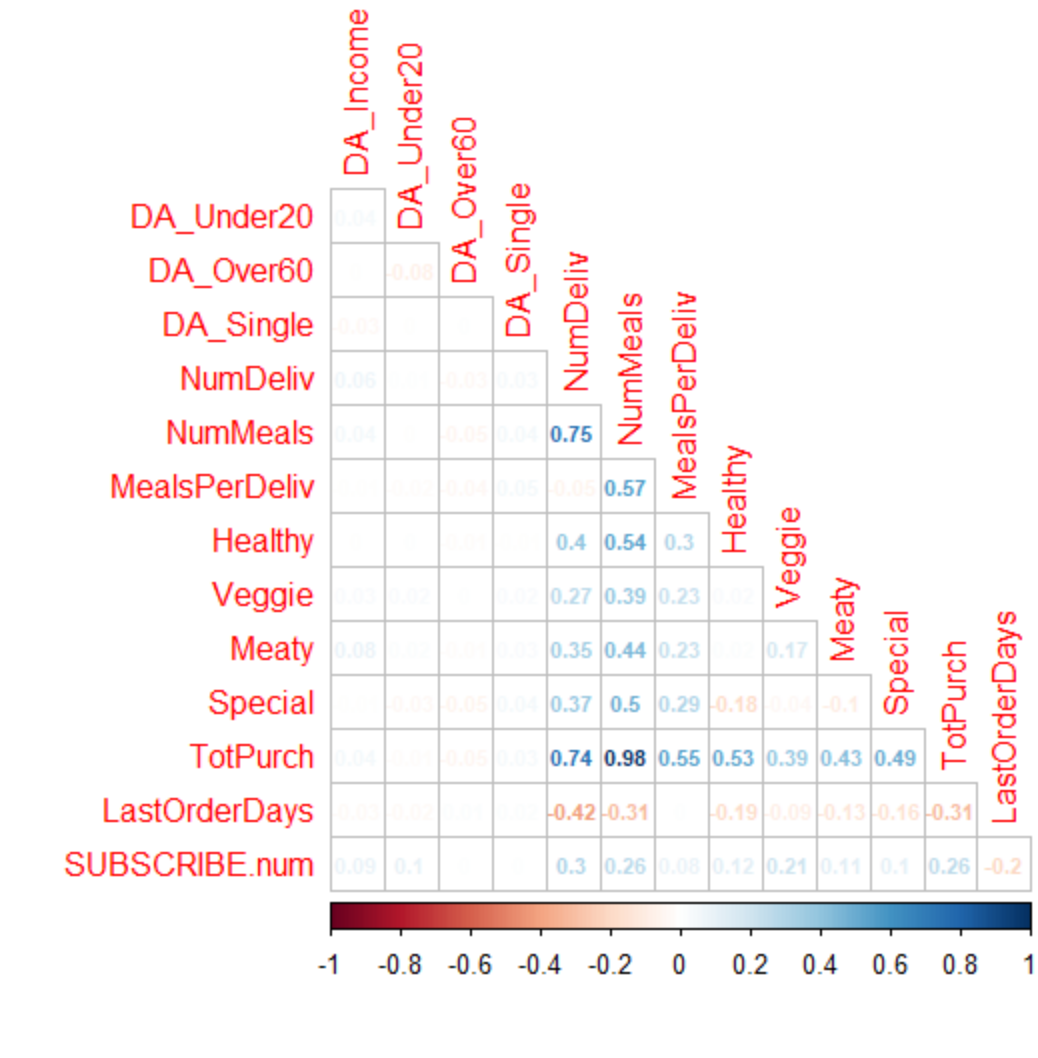
|  |  |  |
| --- | --- | --- |
| **Variables** | **Correlation** | **Explanation** |
| NumMeals & TotPurch | 0.98 | All of these variables are related to the frequency at which one uses QuickKits. Likely, NumDeliv/NumMeals cannot be used in the same model as TotPurch, although we will just keep it in mind for now and explore the data further when making the model. |
| NumDeliv & TotPurch | 0.74 |
| NumMeals & NumDeliv | 0.75 |
| SUBSCRIBE.NUM & Weeks3Meals | 0.77 | The variables are trivially related (see below) |

**Trivially related variables:**

* There is a case of reverse causality with *Weeks3Meals*. Weeks3Meals indicates if a member who signed up for the offer stayed after the promotion was over; however, staying with the promotion occurs *after* actually subscribing to the promotion, so it would have a very high correlation with SUBSCRIBE. We will not delete this variable, just leave it out of our analysis for now.

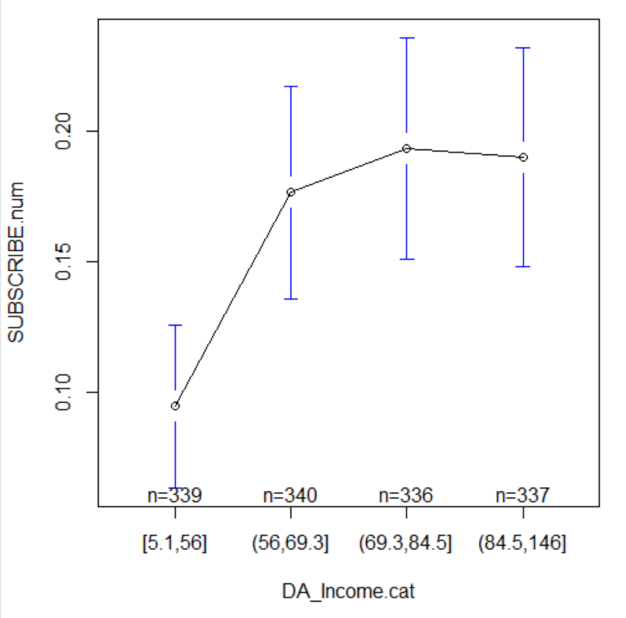
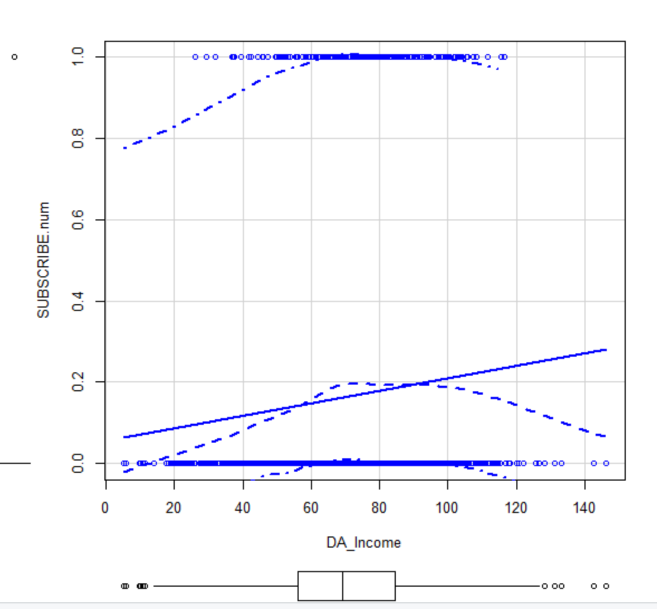
**Case identifiers:**

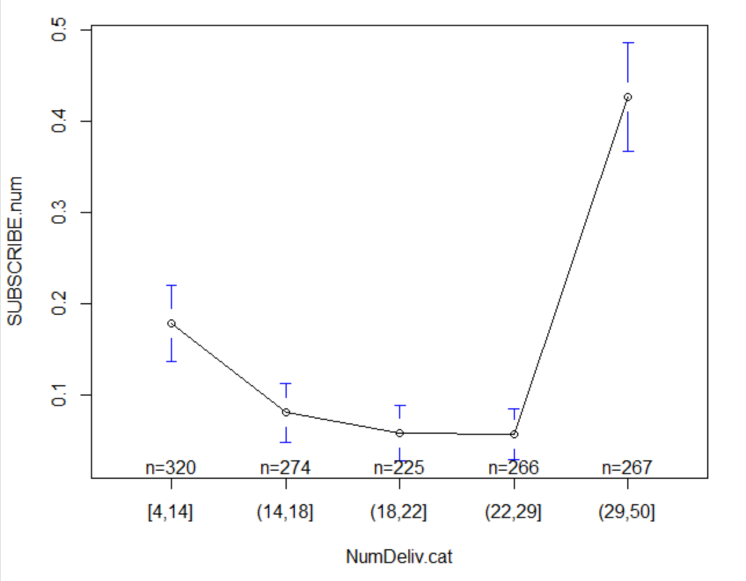
* *Custid* is a case identifier. To fix this, used rownames and deleted the values.
* *X* refers to the customer’s order in the database; custid is a better identifier. X should be deleted.



1. Document your model development steps (i.e., identify the important predictor variables, variable selection, variable transformation, and model assessment and final selection). Provide the necessary visual graphs or charts to support your decisions along the steps (15 points)

**Variable Transformation**

* We created LastOrderDays to track the days since last orders for each customer in replacement of LastOrder. To do this, we subtracted today’s date (2018-03-05) from the LastOrder variable.
* Since DA\_Income is non-linear and concave with SUBSCRIBE, we chose a log transformation.



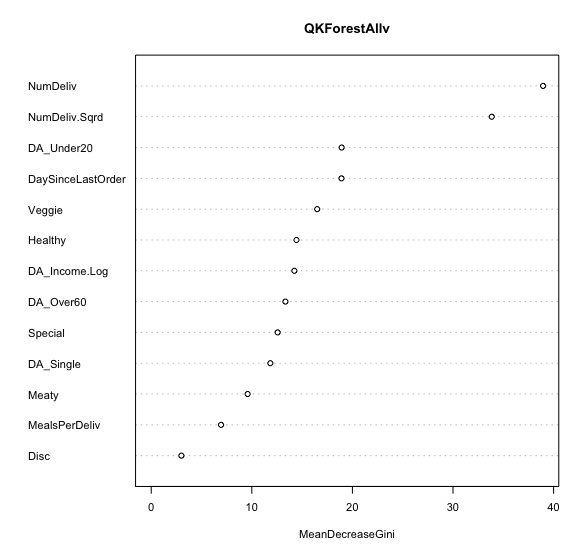
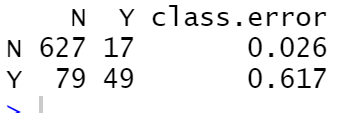
* Since NumDeliv is non-linear and u-shaped with SUBSCRIBE, we chose a square transformation.

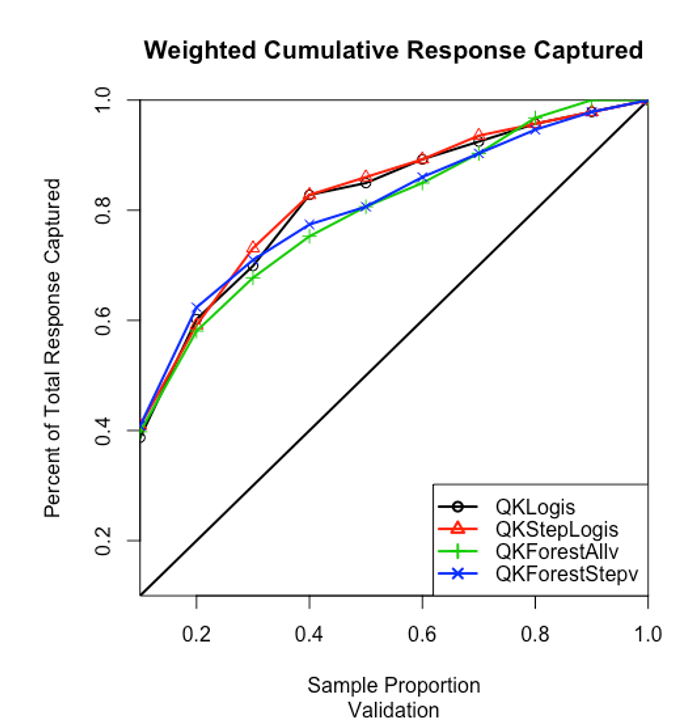
**Variable Selection**

* TotPurch is a trivially related variable to NumMeals and NumDeliv since the amount is added up with the number of meals delivered. The two will always move linearly together. This is supported by the correlation of 0.98. Since the random forest variable importance indicates that NumMeals and NumDeliv are more significant than TotPurch, we will remove TotPurch.
* Weeks3Meals does not provide value for our current model since it only has values for people who subscribe, therefore it is not included in the model.
* CustID and postalcode are not included in the model as they have no meaningful value or relationship to the probability of subscribing

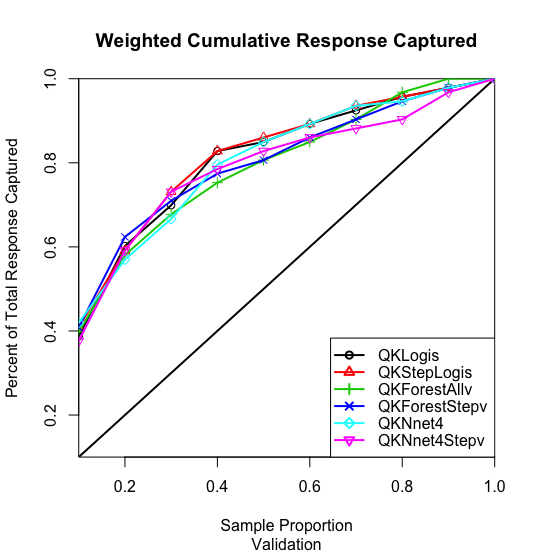
**Most Important Predictor Variables**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Most Important Variables** | **AIC** | **McFadden R-Squared** |
| Random Forest | 1. NumDeliv 2. DA\_Under20 3. DaysSinceLastOrder 4. Veggie 5. Healthy 6. DA\_Income.log 7. DA\_Over60 | N/A | |
| Stepwise Logistic | * DA\_Under20 * NumDeliv * DA\_Income.log * MealsPerDeliv * Veggie * LastOrderDays * NumDeliv.sq | 463.6 | 0.355 |
| Logistic | No variable selection method yet | 474.73 | 0.359 |

* As you can see in the above table, the Stepwise Logistic model has the lowest AIC at 463.6. Although we could not compute these measures for the Random Forest, a confusion matrix tells us that the Random Forest has a low error rate for predicting “N” (2.6%) but a high error rate for predicting “Y” (61.7%). There are many false positives.
* ****

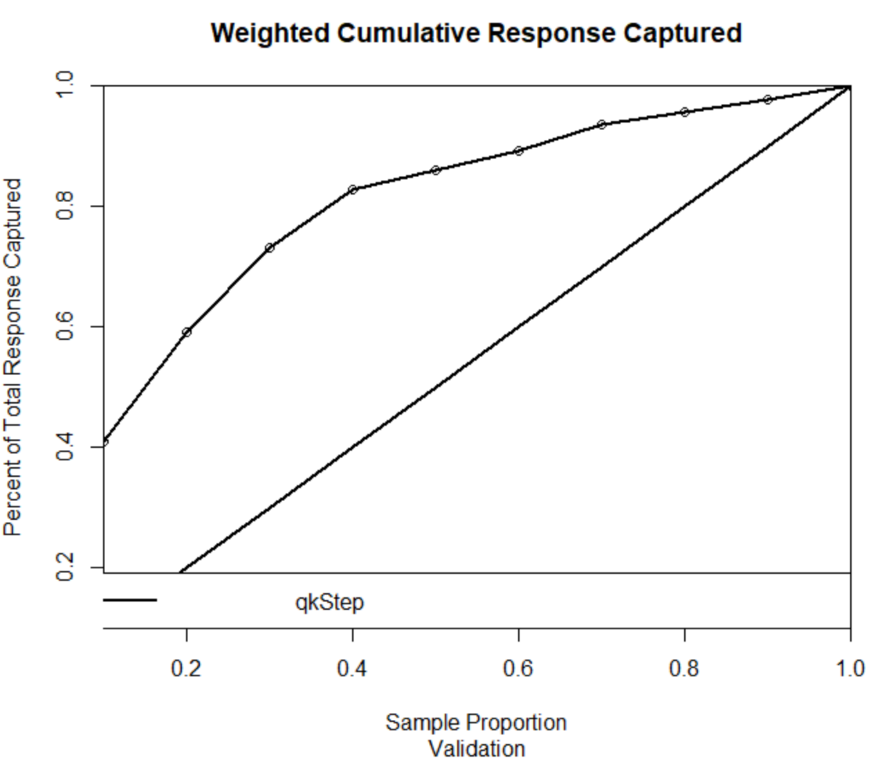
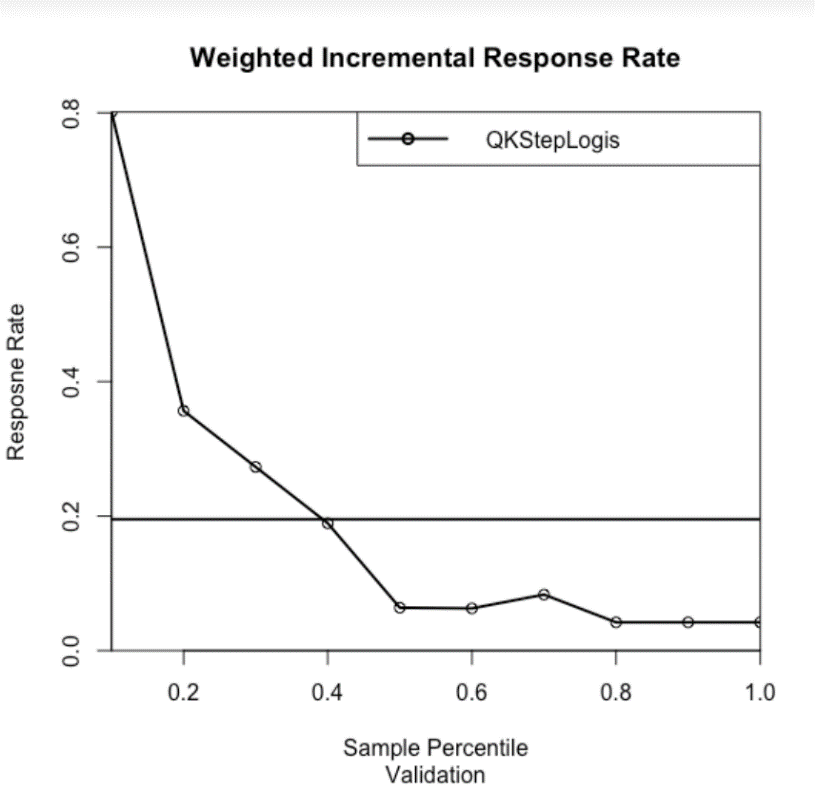


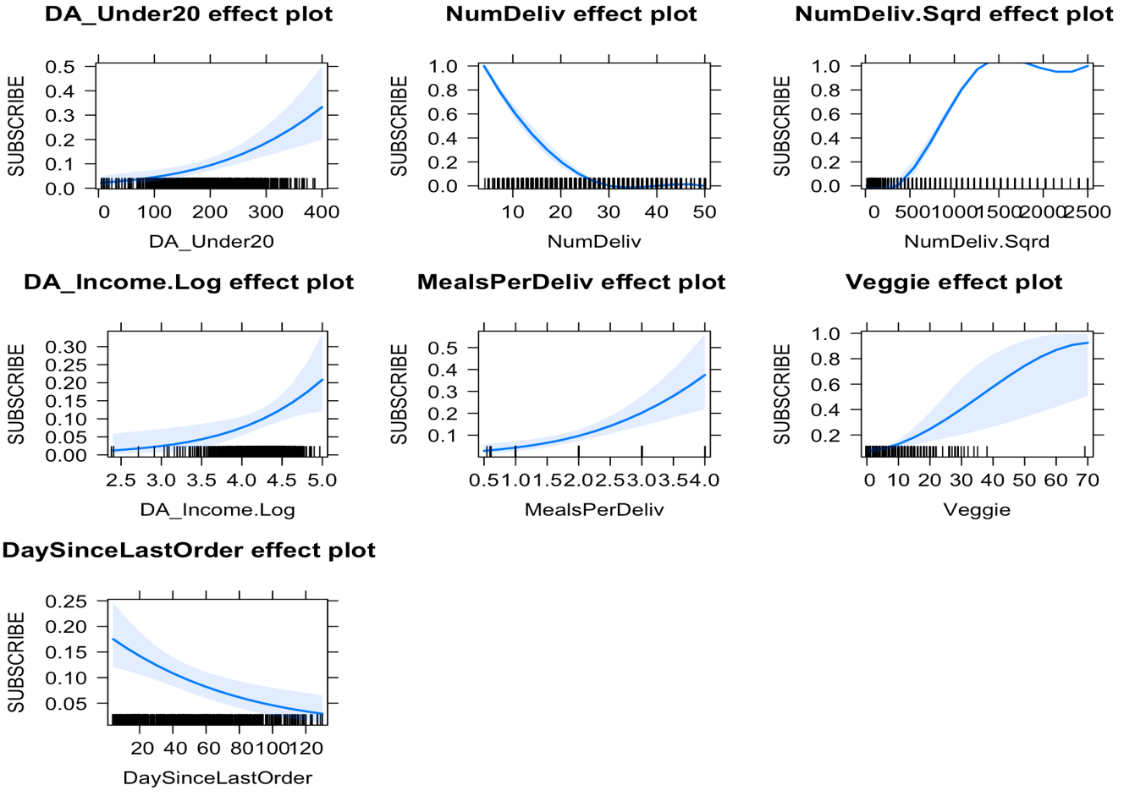
These results are confirmed by this lift chart. The true response rate is 19.54%. Random forest is slightly better at capturing the top 20% of the population but overfits the model afterward. Random forest underperforms logistic regression with selected variables, meaning that logistic regression with proper variable transformations does a pretty good job of predicting the validation set.

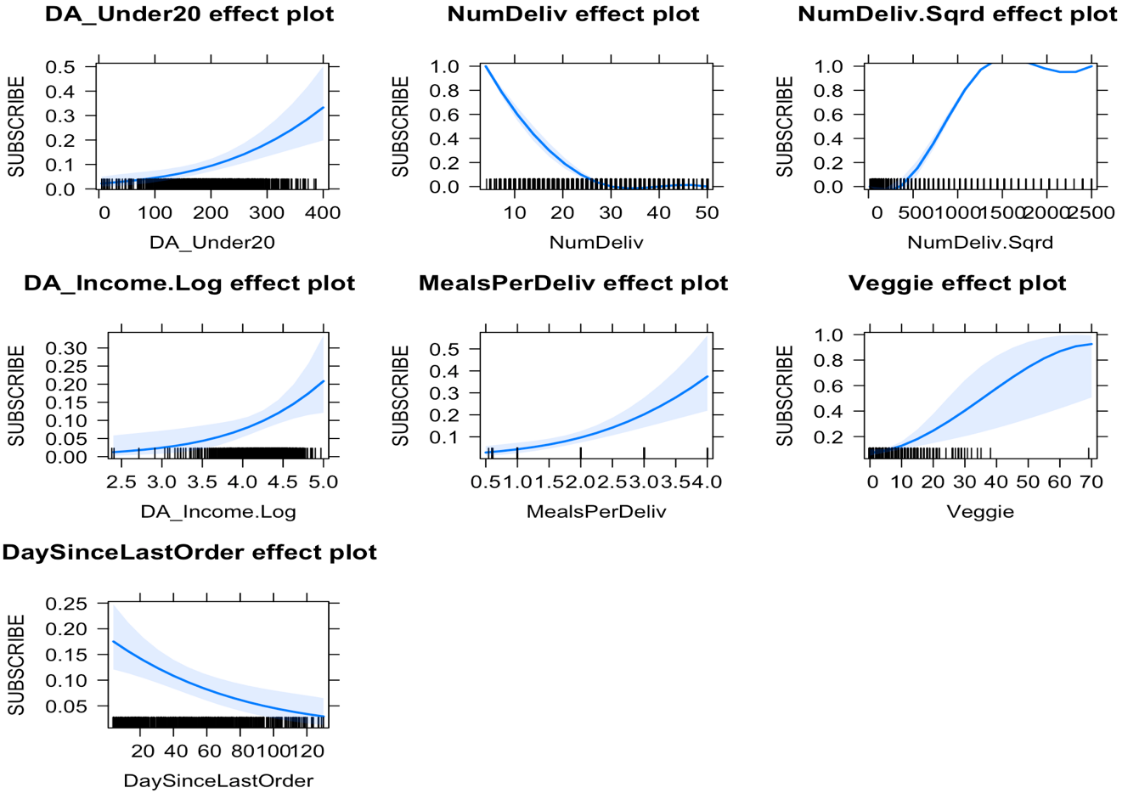


Finally, we ran neural network to test for more non-linearity in the variables. We find that neural network underperforms logistic regression and neural network with stepwise selected variables is overfitting the validation set. Therefore, the best model is logistic regression with stepwise selection and there are not any significant nonlinear relationships that are not being captured by the logistic regression.

1. Interpret your final model and final model’s lift-chart on validation sample (include your final model’s lift chart in your document). (10 points)



Our final model contains:



* DaysSinceLastOrder: The last date the customer ordered from QuickKits
  + Negatively related: The longer it’s been since their last order, the less likely a customer is to subscribe to the promotion
* DA\_Income: The average income of households in the customer’s Census Dissemination Area
  + Positively related: The higher one’s income of their area, the more likely they are to subscribe
* DA\_Under20: The number of households in the customer’s Census Dissemination Area with individuals under 20 years of age
  + Positively related: the more households with individual’s under 20 in there area, the more likely they are to subscribe
* NumDeliv: The number of deliveries in the last 6 months
  + Negatively related: the more deliveries one has in the past 6 months, the less likely they are to subscribe to the promotion
* The average number of meals per delivery
  + Positively related: the higher the number of meals per delivery, the more likely they are to subscribe to the promotion
* The number of vegetarian meals ordered in the last 6 months
  + Positively related: The more veggie meals one orders, the more likely they are to subscribe to the promotion.

Overall, these results confirmed most of our original hypotheses. The one outlier is NumDeliv; if someone has had more deliveries in the past 6 months, they are less likely to subscribe to the promotion. A possible reason for this is that the individual is already receiving lots of deliveries and does not see the value of adding an additional meal. They could not have the capacity to consume that extra meal.

A surprising variable not included in our hypothesis is Veggie. Possibly, there could be a correlation with Vegetarian Meal consumption and Income since it is generally seen as more expensive to be a vegetarian. Another reason could be that vegetarians have limited food options so they get more enjoyment and value from meal kits that provide variety to their meals.

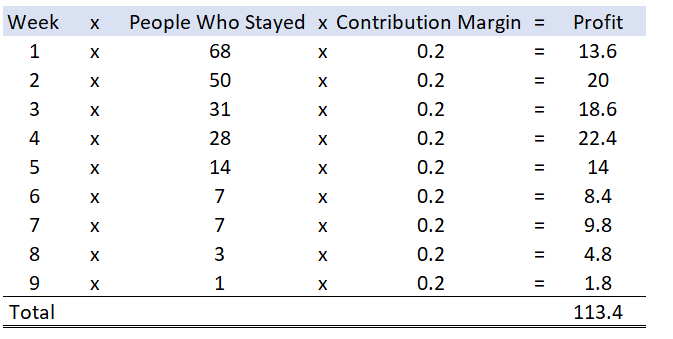
Based on the final incremental lift chart, our logistic model does better at capturing subscribers for the first 40% of the sample percentile validation. It captures ~ 80% of subscribers while contacting the first 40% of the sample, compared to the baseline which would be 40% captured for the first 40% of the sample.

**Part 2:** 30% on business understanding and commutations on the recommendations on the company’s promotion plan. Here is the list of potential questions (but not limited to these questions) that might (or might not) be addressed:

According to our final model, customers who are most likely to subscribe to the promotional offer live in areas where households contain many people under 20 and where average income is high. In terms of ordering behavior, they have a comparably low number of deliveries but they often order a higher number of meals per delivery than average customers. They also tend to be Vegetarian or enjoy Veggie meals and have recently ordered from QuickKits.

First, we will assess the profitability of the campaign. Each campaign runs at a 20% loss equal to Contribution Margin – Discount Rate (20%-40%). This loss should be multiplied by 2 since the promotion runs for a 2-week period. In the data, there are 231 respondents who subscribed (SUBSCRIBE = “Y”). These numbers must be multiplied by the price of a 3-meal kit which is unknown but will be represented by “x”. Thus, the total loss = 20%\*2\*231x = 92.4x.

The total revenues from the promotion equal to the number of people who stayed on the promotion after a meal kit, multiplied by the number of weeks they stayed with the 3-week meal kit and the contribution margin. Again, this all must be multiplied by the price of a meal kit (x).

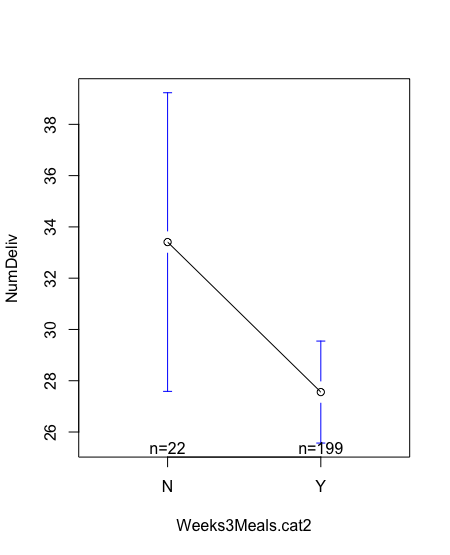


Thus, we can see the total profit equals to 113.4x-92.4x=21x.

The promotion ran at a profit.

For example, if the price of a 3-meal delivery kit is $100, the profit would be $2,100 (21\*100).

Knowing that this promotion is profitable, we would recommend a similar promotion but more focused on targeting individuals with characteristics that correspond with the Weeks3Meals variable, or probability of staying with the meal-plan after the promotion is over. We created a new dataset containing only people who subscribed (SUBSCRIBE =”Y”) and re-factored the Weeks3Meals variable to be “N” if Weeks3Meals = 0 (did not stay with 3 meals after promotion) or “Y” if Weeks3Meals = 1-9 (stayed with 3 meals after promotions). We then assessed the relationship of predictors with this new target variable and found that NumDeliv had the smallest range when compared to the target variable. See below:



As you can see, people are much more likely to stay with the subscription if they have had less deliveries in the past 6 months. Income also had some effect, with people being more likely to stay with the promotion if their income is higher, however the range was much larger for “N”.

In terms of future communication content, we can see that DA variables are very significant in the final model, making up 2 of 6 of the most important variables. This means that certain neighborhoods are more or less likely to subscribe than other neighbourhoods. QK could look into geographic targeting for direct mail campaigns to increase revenues while saving costs by being geographically focused. Also, we learned that people who order Veggie Kits are more receptive to marketing communications, which could lead to a future promotion focused on Vegetarian meals only. Lastly, we see that the probability of subscribing decreases steeply from recent order to a high number of days since last order, meaning marketing should only be directed to people who have ordered recently (~ 2 months).

Overall, knowing the characteristics of those who stay after the promotion period allows us to more effectively target profitable customers. In the future, we could look at the cost of marketing campaigns, profit per customer, and number of people reached to find the profit maximizing quantity. We could also use a survey to collect income and household data on an individual level rather than by DA. This would prove if the DA variables are truly related to probability of subscribing to a promotion or if the geographic area that one lives in is more important.

**DATA DICTIONARY:**  Variables in the Data Set **QK.csv**

Target Variable to predict probability of response to the promotion and upgrading to three-meal plan, **for the model performance competition**:

**SUBSCRIBE**: Customer signed up for the promotion or not: “Y” if signed up, “N” if not.

Customer Characteristics:

**Custid:** 7 digit customer identification number

**Disc:** Class of customer for standard discounts: “Student” or “Senior”

**Title:** “Mr” “Ms” “Mrs” or “Dr”

**LastOrder:** Date of last order: year/month/day

**Pcode:** Postal Code\*

**DA\_Income:** Mean Income of households in customer’s Census Dissemination Area, thousands of dollars

**DA\_Under20**: Number of households in customer’s Census Dissemination Area with individuals under 20 years of age

**DA\_Over60:** Number of households in customer’s Census Dissemination Area with individuals over 60 years of age

**DA\_Single:** Number of households in customer’s Census Dissemination Area with only 1 person

**NumDeliv:** Number of deliveries ordered in last 6 months

**NumMeals:** Number of meals ordered in last 6 months (each delivery can have several meals)

**MealsPerDeliv:** Average number of meals per delivery (NumMeals/NumDeliv)

**Healthy:** Number of healthy meals ordered last 6 months

**Veggie:** Number of vegetarian meals ordered last 6 months

**Meaty:** Number of meaty meals ordered last 6 months

**Special:** Number of specialty meals ordered last 6 months

**TotPurch:** Amount purchased last 6 months in dollars

**Weeks3Meals:** Number of weeks that a customer that signed up for the promotional offer stayed with the three-meal-plan, after the promotion was over.

\*Canada has 850,000 postal codes, which are combined into 54000 Dissemination Areas. A Dissemination Area typically has between 400 and 700 households.