**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 can build a logistic regression model with stepwise algorithm to select most important factors and predict whether the customer will stick to the three-meal-per-week plans after two-weeks promotion period.

-Assumptions of potential subscribers to three-meal-per-week plans:

* People who frequently order food delivery
* People who are wealthy to purchase meal delivery service
* People who have more people in the household

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)

Hypothesis 1: People who frequently order food delivery will likely upgrade

* People who order frequently indicate a need of delivery services

#LastOrder – if recent, may need upgrade

#NumDeliv – if many, may need upgrade

# NumMeals – if many, may need upgrade

Hypothesis 2: People who are wealthy would like to stick to the three meal per week plans

* Peope who have high income would afford the upgraded service

# DA\_Income – if high, may upgrade

Hypothesis 3: People who need to take care of others living in the household may suit to the three-day meal plan

* People who need to take care of other family members may not have time to cook

# DA\_Under20 – if many, may need upgrade

# DA\_Over60 – if many, may need upgrade

# DA\_Single – if many, may not need upgrade

Hypothesis 4: People order many meals per delivery may want to do three meal per week

* People order many meals per delivery indicate maybe they keep some meals to the next day

# MealsPerDeliv – if many, may need upgrade

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 data:

83.5% of missing data in Weeks3Meals

* We will convert them to 0 since the missing data is due to no subscription

73.6% of missing data in Disc

* We will impute them with NoDiscount since the missing value indicates the customer is not student nor senior.

35.4% of missing data in Title

* We will remove them since it’s non-reporting value and we can’t impute them with other indicators

3.43% of missing data in DA\_Income, DA\_Over60 and DA\_Single

* We will remove the cases with missing value since 3.43% are small enough to ignore

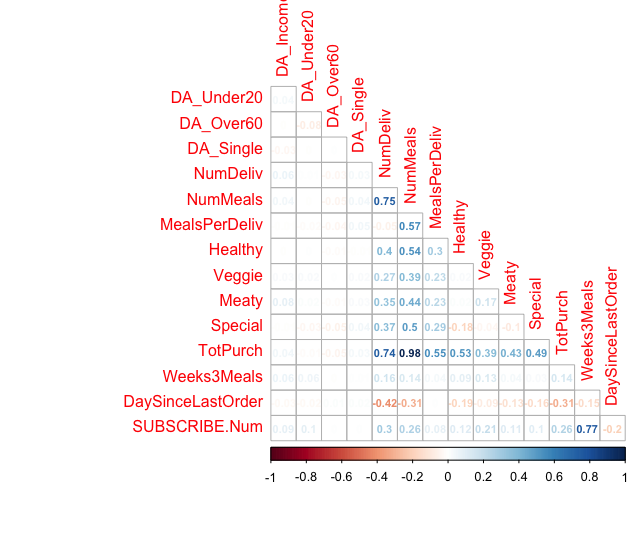
-Correlation:

TotPurch is highly correlated with NumMeals at 0.98

NumMeals is highly correlated with NumDeliv at 0.75

TotPurch is highly correlated with NumDeliv at 0.74

SUBSCRIBE.NUM is highly correlated with Weeks3Meals at 0.77



-We have reduced the dataset from Observations: 1,400 and Variables: 21 to Observations: 1,352 and Variables: 20

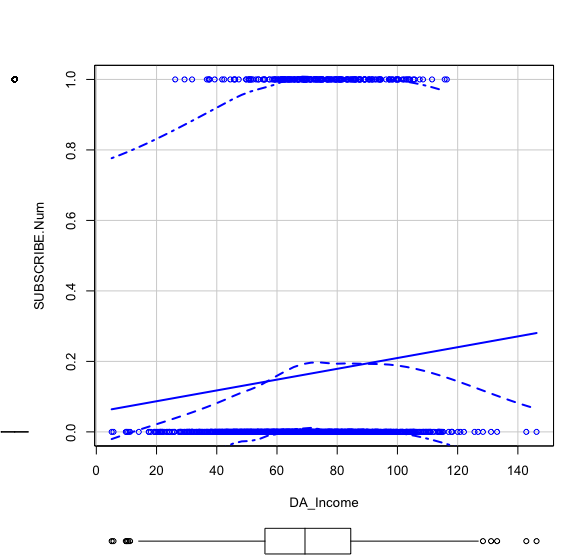
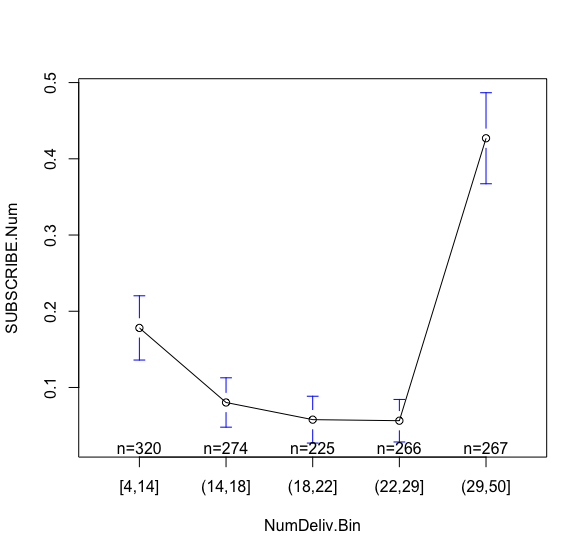
1. Document your model development steps (i.e., identify the important predictor variables, variable selection, variable transformation, and model assessment and final selection). Provide the necessary visual graphs or charts to support your decisions along the steps (15 points)

-Variable transformation:

We created DaysSinceLastOrder to track the days since last orders for each customer.

We log transferred DA\_Income since it is non-linear and concave with SUBSCRIBE.

We squared NumDeliv since it is non-linear and U-shape with SUBSCRIBE.

-Variable selection:

We realize the Total purchase is a trivially related variable to NumMeals and NumDeliv since the amount is added up with the number of meals delivered. Therefore, we will remove TotPurch.

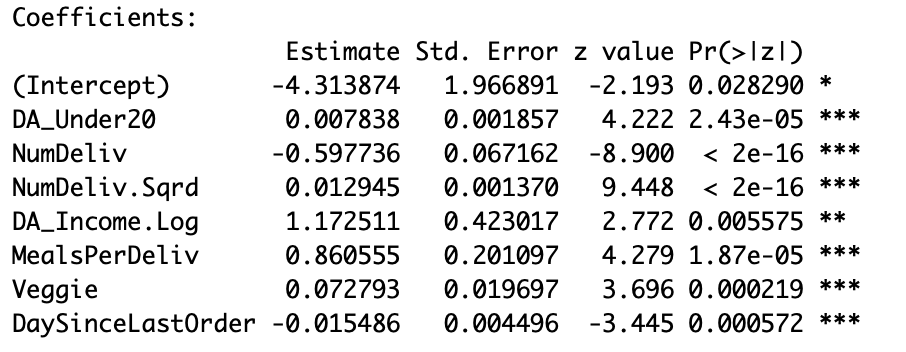
We remove customer ID and postal code because they have no numerical value to the model.

We remove Weeks3Meals since it only has values for people who subscribed. So it does not provide value for our model.

-Most important predictor variables:

We have the maximal model with AIC: 474.73 and McFadden Rsquared: 0.359

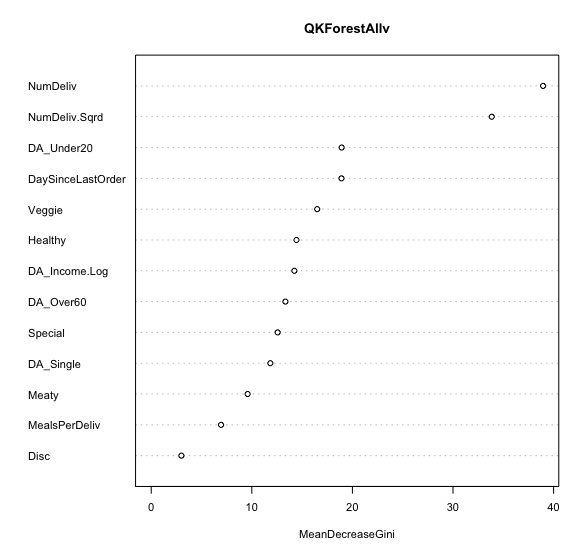
We use stepwise algorithm to select the best predictors with minimum AIC:



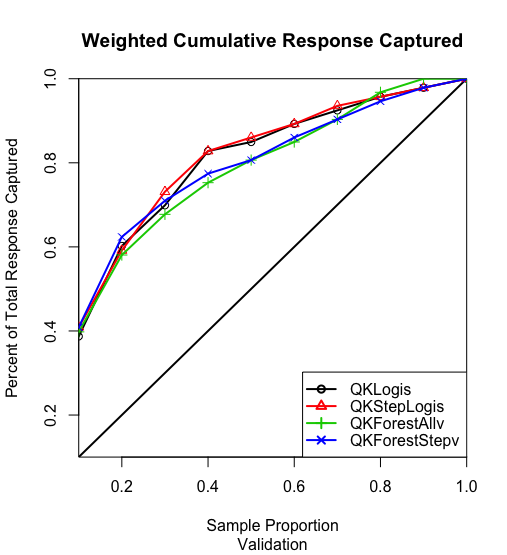
with AIC: 463.6 and McFadden Rsquared: 0.355

We run Random Forest and the most important predictors are:

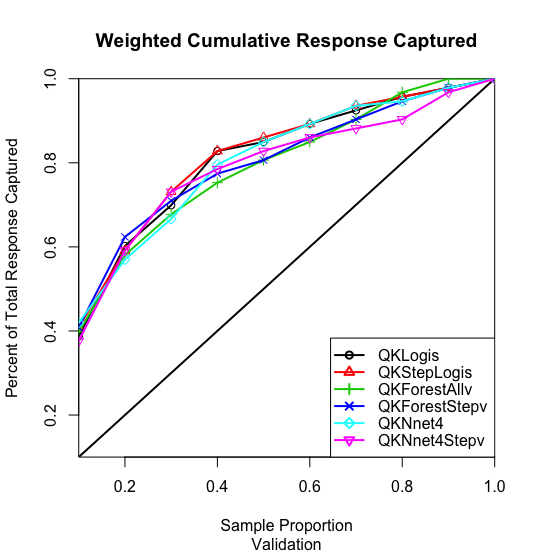
Number



We run random forest with stepwise selected variables and it is slightly better capturing given top 20% of the population but overfitting afterwards. Random forest underperforms logistic regression with selected variables and it means that logistic regression with proper variable transformation does pretty good job to predict the validation set.



Finally, we run neural network to test more on nonlinearity in the variables. We find that neural network with all variables underperforms logistic regression and neural network with stepwise selected variables is overfitting the validation set. Therefore, we can see that the best model here is logistic regression with stepwise selection.

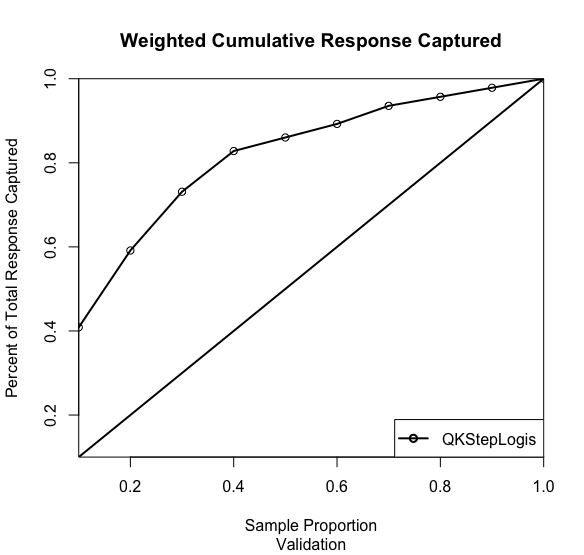
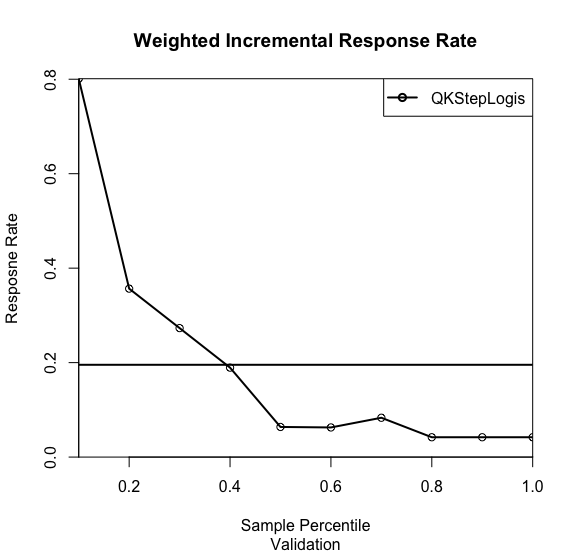


1. Interpret your final model and final model’s lift-chart on validation sample (include your final model’s lift chart in your document). (10 points)

-The lift chart shows that the Logistic regression with the variables that are selected by stepwise algorithm predicts the best on the validation set.

-The best model contains

* Last date of the order
* Average income of households in the customer’s Census Dissemination Area
* Number of households in customer’s Census Dissemination Area with individuals under 20 years of age
* Number of deliveries ordered in last 6 months
* Number of meals ordered in last 6 months
* Average number of meals per delivery
* Number of vegetarian meals ordered last 6 months



Calculating break-even response rate:

ER\*a = cost

Cost: 40%x100($)

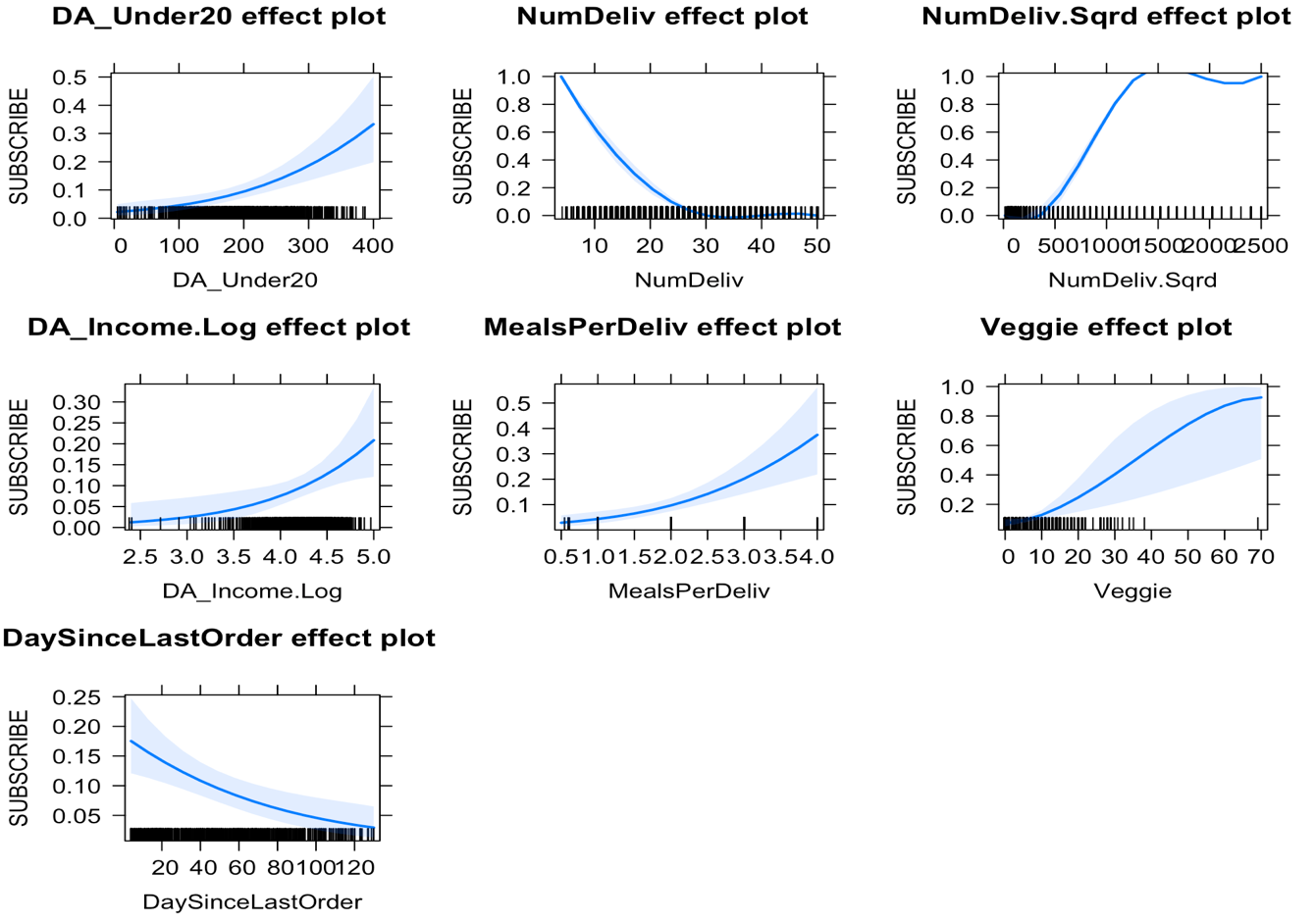
ER: 120%x100($)

a = 40/120 = 0.33

1. Attach your R script file for verification purpose (10 points).

**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:

* What are the characteristics of those customers who are most likely subscribe to the promotional offer?
* Higher number of individuals under 20 years of age living in the customer’s neighborhood (~0.3 effect size)
* High number of deliveries ordered in last 6 months (1 effect size)
* High income of households in customer’s neighborhood (~0.2 effect size)
* High average number of meals per delivery (~0.4 effect size)
* High number of vegetarian meals ordered last 6 months (~0.8 effect size)
* Low days since last order (~0.2 effect size)



* Are there any implications for communication content?
* What are the characteristics of those who subscribe and do not opt-out after two weeks?
* Can you identify customers who were not offered the promotion that you would now offer it to?
* Yes, we can identify those potential customers who have high chance of continuing the offer after two weeks period.
* Is this upselling promotion profitable? Why or why not?
* What should the next steps be for the company or your analysis? What information will be useful for your future steps?

For the business understanding, you can incorporate some point format but the rationale for your recommendation should be fully supported and well explained. The evaluation of this part will be based on the clarity of your communications, and evidence and logic of your recommendations.

(Data Dictionary on next page)

**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.