Marketing

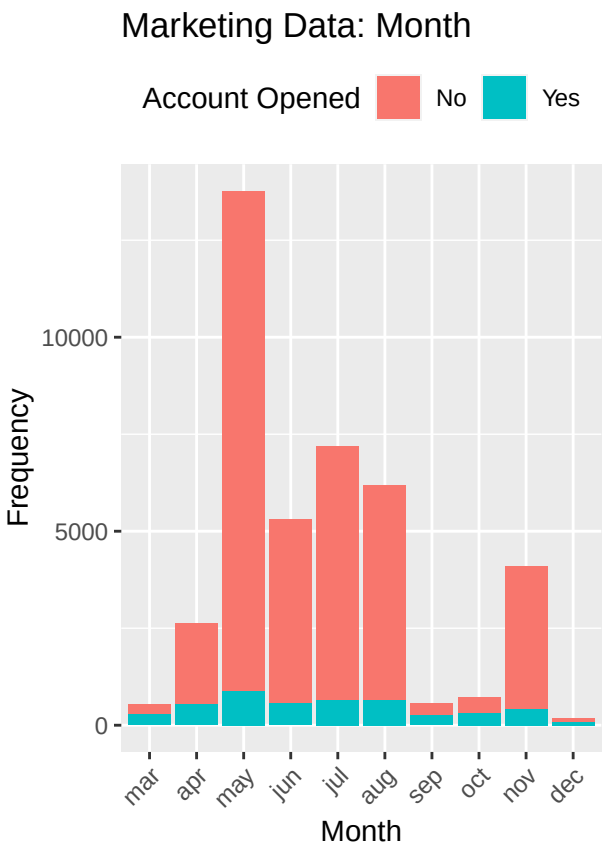
Jillian Maw

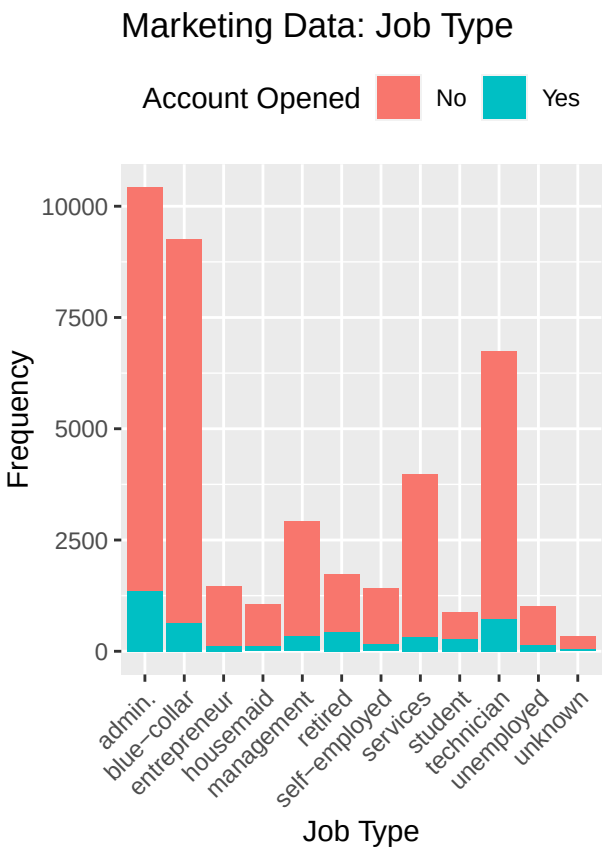
4/8/2022

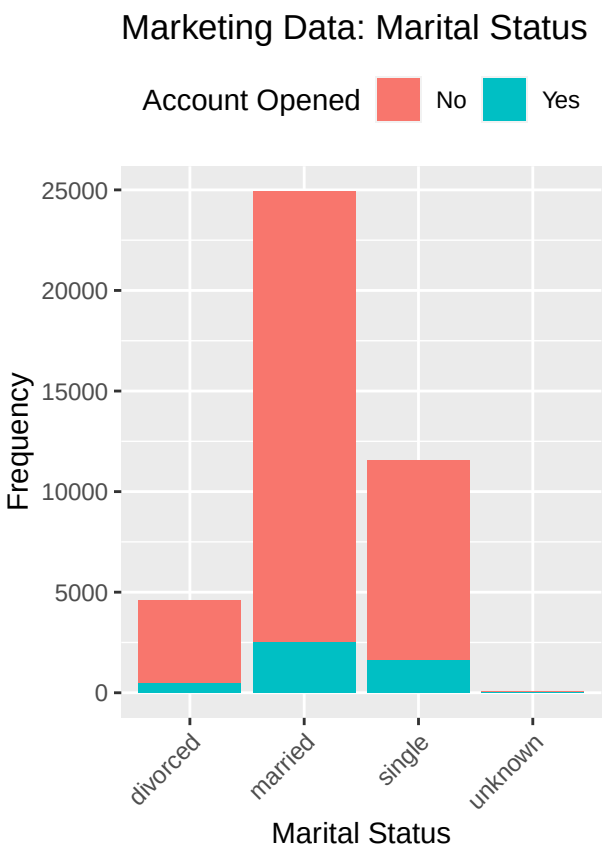
Section 1: Introduction and Problem Background

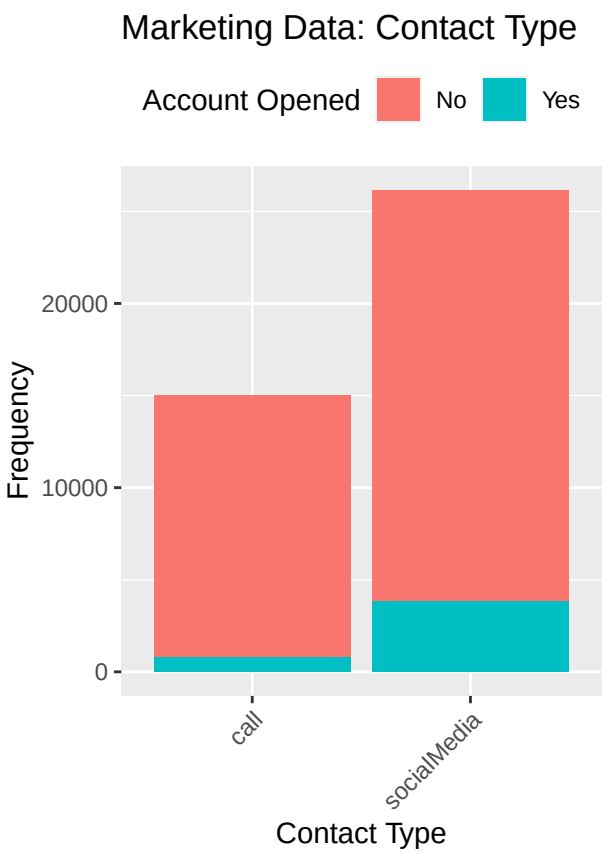
I have been asked to help your bank understand who your potential customers for credit cards are, so as to better help you all target the interested individuals in future marketing campaigns. The effectiveness of your bank's campaign depends on its ability to reach interested individuals; by analyzing the data you have collected, I will be able to help determine which customers are likely to purchase a credit card based on their characteristics, the method of contact, and the number of times to contact. Determining these things will also allow your bank to prioritize contacts. Since the response variable of interest is whether or not a customer opened a new credit card account, I will be creating a logistic regression model to determine the probability a potential customer opens an account or not.

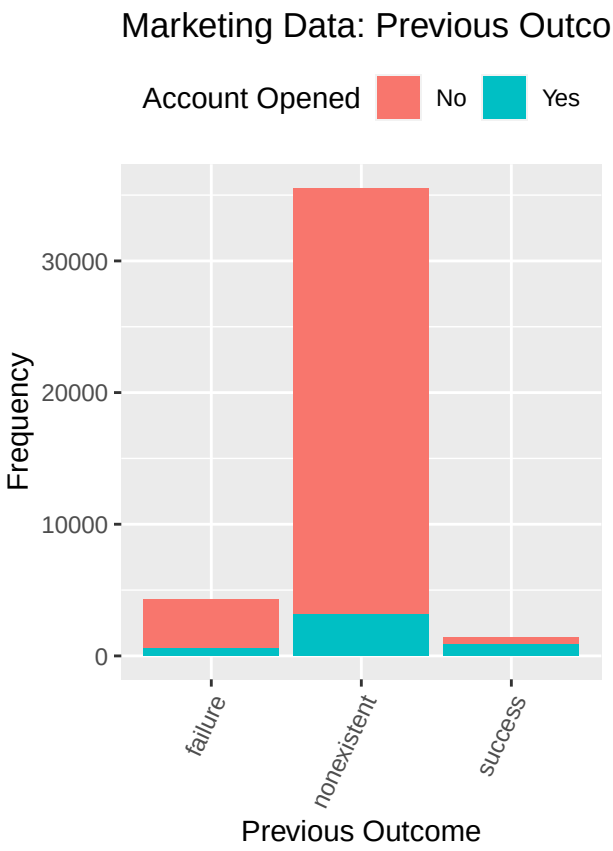
The 10 different variables from the data set your bank had for a current marketing campaign, which was designed to get customers to open a new bank credit card account. Those variables are age, type of job, marital status, type of communication contact, the last month of contact, the number of contacts performed during this campaign and for this client, if this client has been previously contacted in a different campaign, the number of contacts performed before this campaign and for this client, the outcome of the previous marketing campaign, and if the client opened a new account. As can be seen in the graphs below, most people did not choose to open new accounts, but a few trends can be seen. More accounts were opened in the summer months, when more people were presumably contacted, judging by the frequency of people who must have been asked. People with administrative jobs were both asked more and more likely to say yes to a new account. Married people were more likely to open new accounts. People contacted via social media were more likely to open new accounts. Not many previous contacts were contacted, but those who were tended to open new accounts, by the appearance of the graph. The correlation plot has no strong relationships between variables except for a slight one. 0.59 out of 1 between the number of contacts performed before this campaign and for this client and if this client been previously contacted in a different campaign. The histograms show right skews for all three variables, but as I am using the Bernoulli distribution instead of the Normal distribution to show the results of this data set later, this should not be an issue.

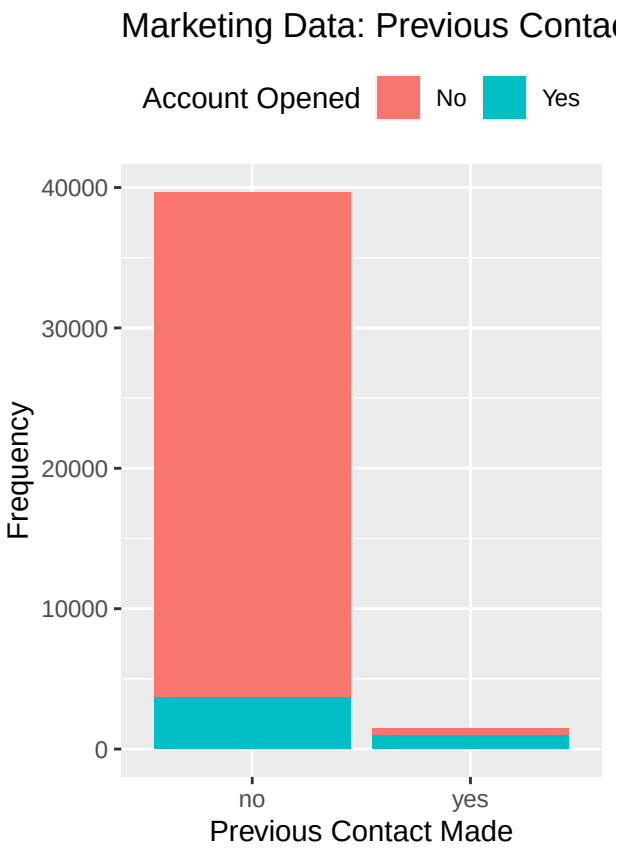

Marketing Data: Month
Account Opened
No
Yes
Frequency
10000
5000
0
mar
apr
may
jun
jul
aug
sep
oct
nov
dec
Month

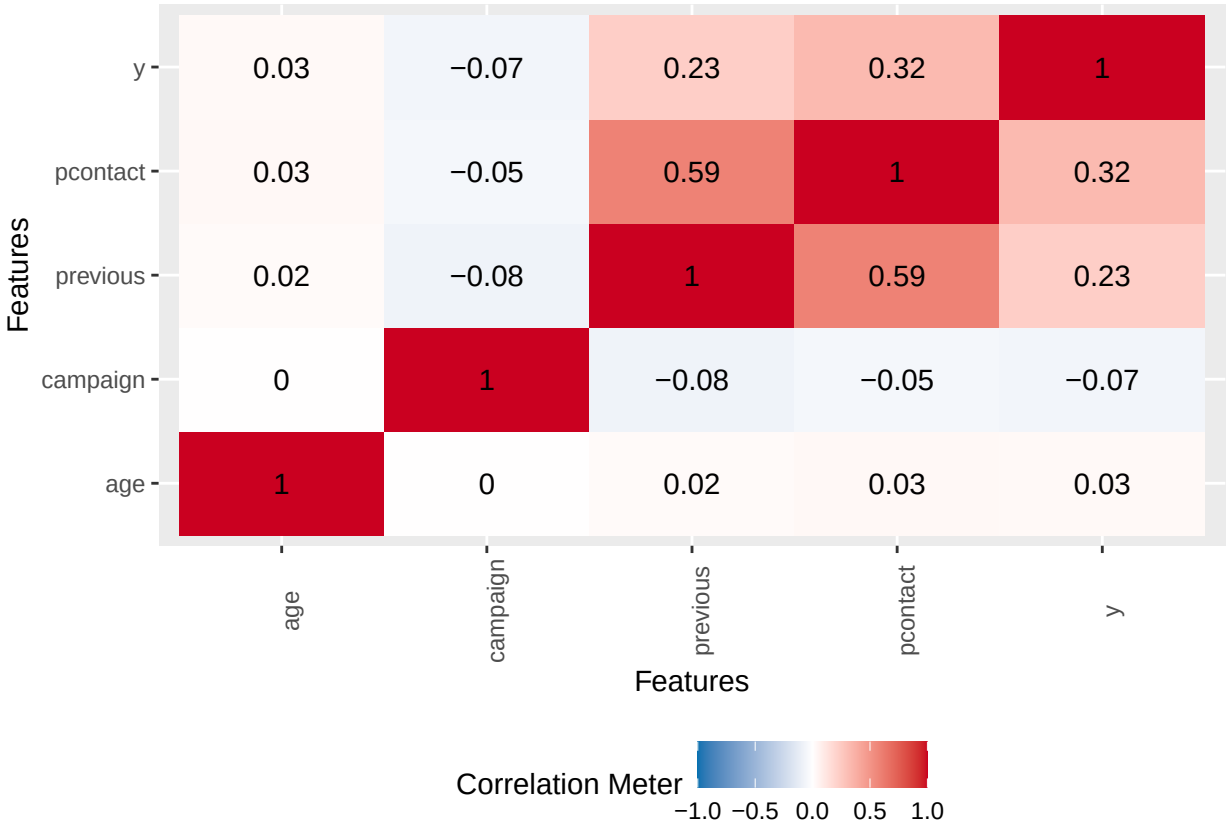

Marketing Data: Job Type
Account Opened
No
Yes
Frequency
10000
7500
5000
2500
0
admin.
blue-collar
entrepreneur
housemaid
management
retired
self-employed
services
student
technician
unemployed
unknown
Job Type

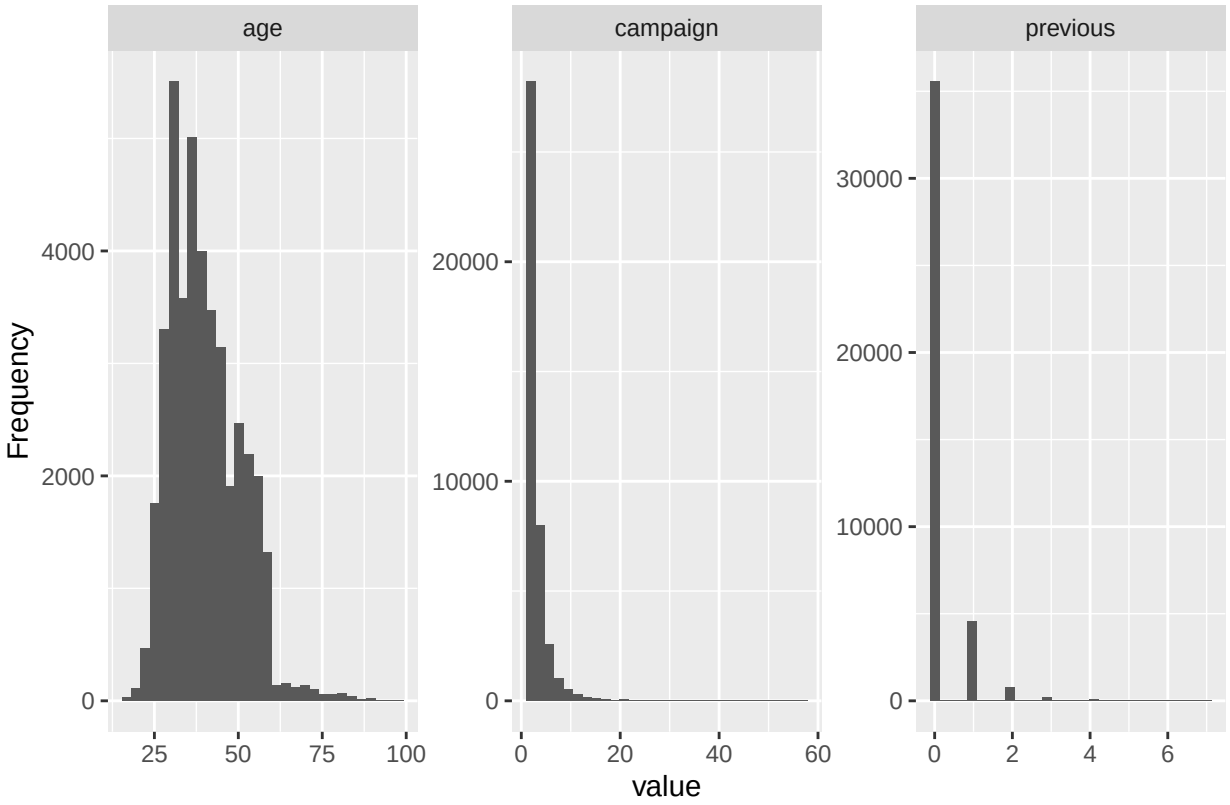

Marketing Data: Marital Status
Account Opened
No
Yes
25000
20000
15000
10000
5000
0
Frequency
divorced
married
single
unknown
Marital Status


Marketing Data: Contact Type
Account Opened
No
Yes
20000
10000
0
Frequency
call
socialMedia
Contact Type


Marketing Data: Previous Outcor
Account Opened
No
Yes
30000
20000
10000
0
Frequency
failure
nonexistent
success
Previous Outcome


Marketing Data: Previous Conta
Account Opened
No
Yes
40000
30000
20000
10000
0
Frequency
no
yes
Previous Contact Made


Features
y
pcontact
previous
campaign
age
0.03 −0.07 0.23 0.32 1
0.03 −0.05 0.59 1 0.32
0.02 −0.08 1 0.59 0.23
0 1 −0.08 −0.05 −0.07
1 0 0.02 0.03 0.03
age
campaign
previous
pcontact
y
Features
Correlation Meter
−1.0 −0.5 0.0 0.5 1.0



Section 2: Statistical Modeling

Due to the number of explanatory variables in the data being less than 40, I can use the "best subset selection" variable selection procedure (also referred to as the exhaustive variable selection procedure) because it is the best method for minimizing the Aikake Information Criteria, the Bayesian Information Criteria, or the Predictive Error, and maximizes the Adjusted R^2 for a logistic regression model. These criteria are the best indicators for determining the fit of the model. In addition, the "best subset selection" procedure tests all possible variables with each other, allowing us to know with certainty that I have the best model. I decided to use the Aikake Information Criteria, or AIC, model comparison criterion, because I am looking to predict if a woman will have a positive diabetes diagnosis, and the AIC model comparison criterion is optimized for making predictions, compared to the Bayesian Information Criteria, or BIC, which is optimized for making inferences. After using the "best subset selection" variable selection method and the AIC model comparison criteria, I decided the most important variables to include were the type of job, marital status, type of communication contact, the last month of contact, the number of contacts performed during this campaign and for this client, if this client has been previously contacted in a different campaign, the number of contacts performed before this campaign and for this client, and the outcome of the previous marketing campaign. In short, every variable except the age variable.

Below is the logistic regression model to determine if a client will open a new account:

$$y_i \overset{ind}{\sim} \text{Bern}(p_i)$$

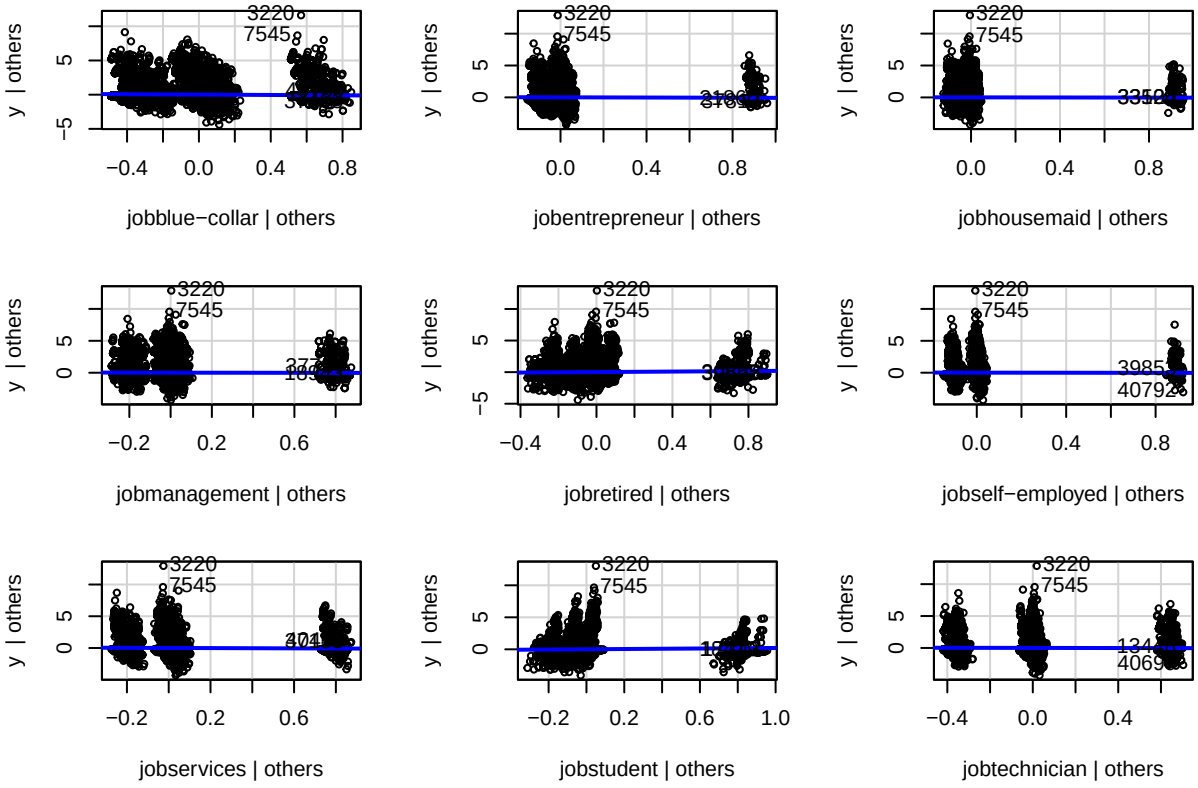
$$log(\frac{p_i}{1-p_i}) = \beta_0 + \sum_{j=1}^{P} x_{ij}\beta_j, \text{ where } \beta_j \text{ represents the following variables:}$$

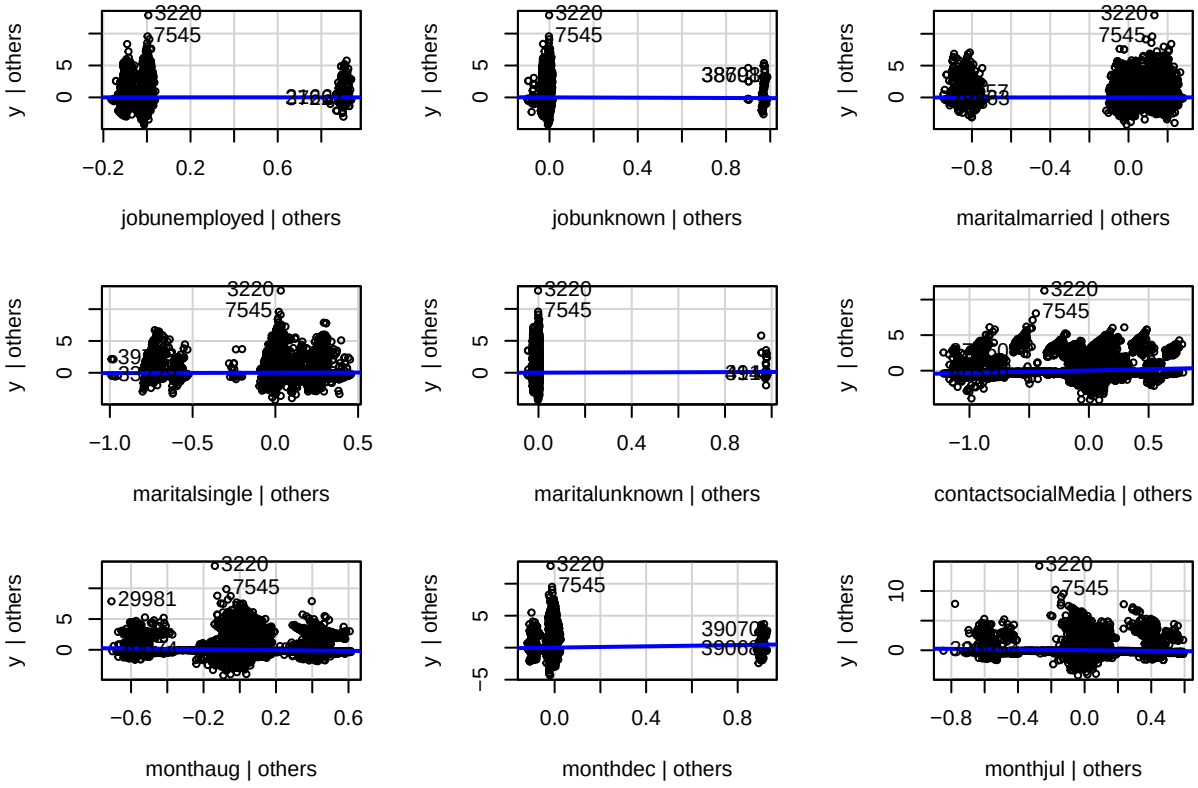
- β_1 for a job type of blue-collar, also written as $\beta_{\text{I(Job=Blue-Collar)}}$
- β_2 for a job type of entrepreneur, also written as $\beta_{\text{I(Job=Entrepreneur)}}$
- β_3 for a job type of housemaid, also written as $\beta_{\text{I(Job=Housemaid)}}$
- β_4 for a job type of management, also written as $\beta_{\text{I(Job=Management)}}$
- β_5 for a job type of retired, also written as $\beta_{\text{I(Job=Retired)}}$
- β_6 for a job type of self-employed, also written as $\beta_{\text{I(Job=Self-Employed)}}$
- β_7 for a job type of services, also written as $\beta_{\text{I(Job=Services)}}$
- β_8 for a job type of student, also written as $\beta_{\text{I(Job=Student)}}$
- β_9 for a job type of technician, also written as $\beta_{\text{I(Technician)}}$
- β_{10} for a job type of unemployed, also written as $\beta_{\text{I(Job=Unemployed)}}$
- β_{11} for a job type of unknown, also written as $\beta_{\text{I(Job=Unknown)}}$
- β_{12} for a marital status of married, also written as $\beta_{\text{I(Marital=Married)}}$
- β_{13} for a marital status of single, also written as $\beta_{\text{I(Marital=Single)}}$
- β_{14} for a marital status of unknown, also written as $\beta_{\text{I(Marital=Unknown)}}$
- β_{15} for a contact communication type of social media, also written as $\beta_{\text{I(Contact=Social Media)}}$
- β_{16} for the last contact month of the year being March, also written as $\beta_{\text{I(Month=March)}}$
- β_{17} for the last contact month of the year being May, also written as $\beta_{\text{I(Month=May)}}$
- β_{18} for the last contact month of the year being June, also written as $\beta_{\text{I(Month=June)}}$
- β_{19} for the last contact month of the year being July, also written as $\beta_{\text{I(Month=July)}}$
- β_{20} for the last contact month of the year being August, also written as $\beta_{\text{I(Month=August)}}$
- β_{21} for the last contact month of the year being September, also written as $\beta_{\text{I(Month=September)}}$
- β_{22} for the last contact month of the year being October, also written as $\beta_{\text{I(Month=Ocotober)}}$
- β_{23} for the last contact month of the year being November, also written as $\beta_{\text{I(Month=November)}}$
- β_{24} for the last contact month of the year being December, also written as $\beta_{\text{I(Month=December)}}$
- β_{25} for the number of contacts performed during this campaign and for this client, also written as $\beta_{\text{(Campaign)}}$
- β_{26} for if this client has been previously contacted in a different campaign, also written as $\beta_{\text{(Previous)}}$
- β_{27} for the outcome of the previous marketing campaign being nonexistent, also written as $\beta_{\text{I(Previous Outcome=Nonexistent)}}$
- β_{28} for the outcome of the previous marketing campaign being success, also written as $\beta_{\text{I(Previous Outcome = Success)}}$
- β_{29} for the number of contacts performed before this campaign and for this client, also written as $\beta_{\text{(Previous Contact)}}$

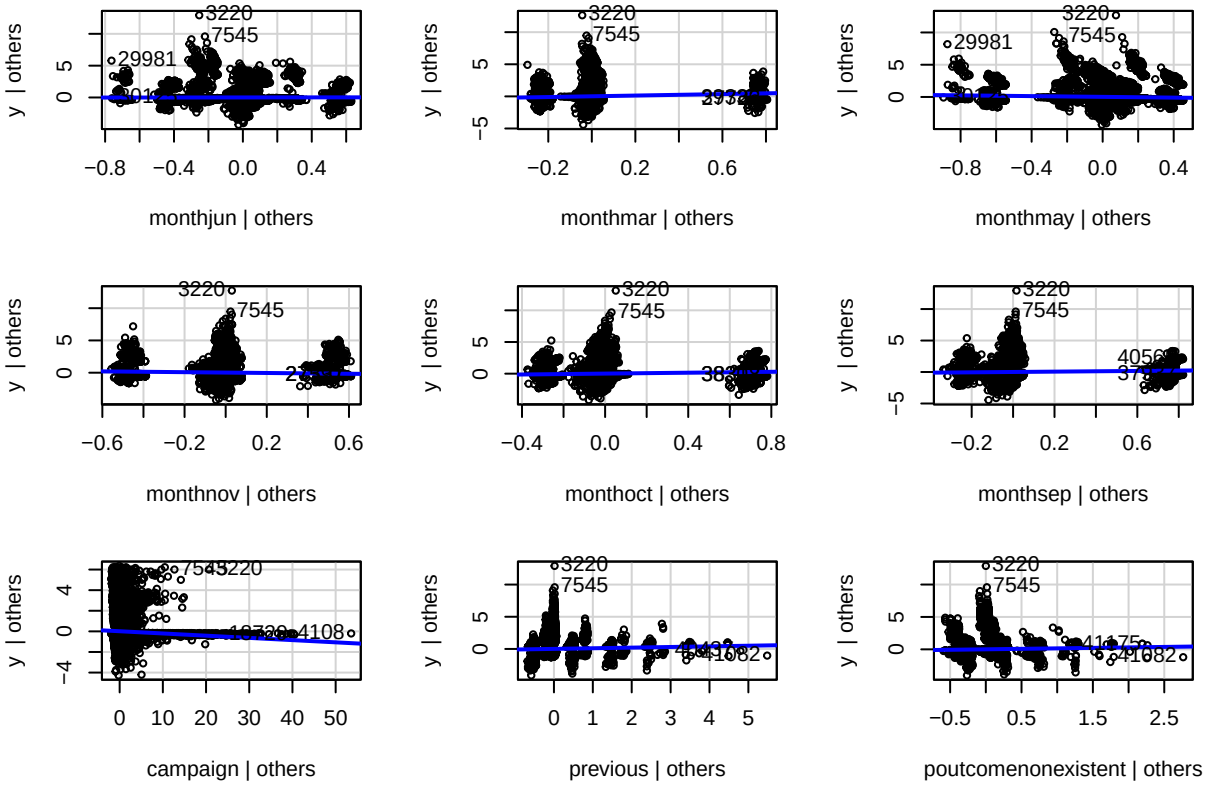
To demonstrate how to interpret this long list of variables in the model, I provide two examples of coefficient interpretation, one categorical effect and one quantitative effect. One categorical effect is that for the category job, if the

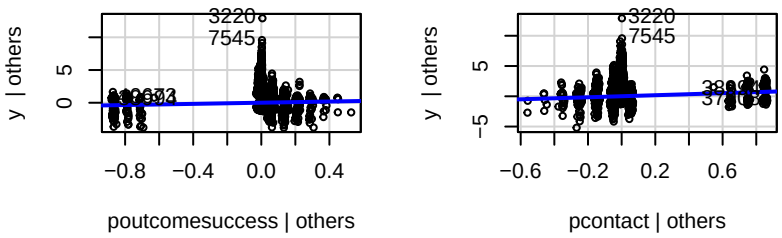
Section 3: Model Validation

The assumptions made for this model include that the data set contains independent data and, when the dataset is graphed by log-odds, it is linear. We can assume the data is independent because When we check the linearity of the log-odds of the data set, we can create scatterplots of the explanatory variables (β_1 through β_{29}, as explained above) against the response variable, y, whether they opened a new account or not, and check that the regression lines modeling those relationships are monotonic, or always increasing or always decreasing. Below are the scatterplots proving these relationships.

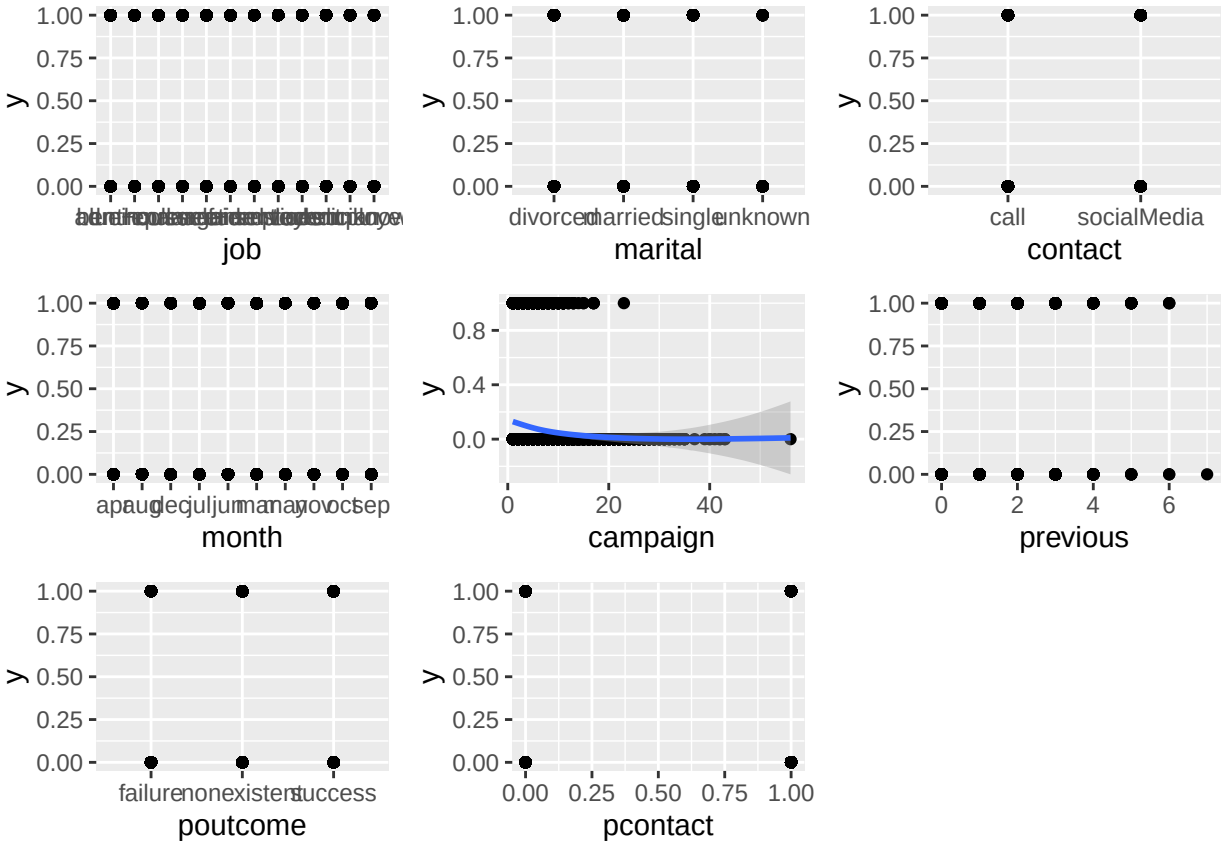

y | others
jobblue−collar | others
jobentrepreneur | others
jobhousemaid | others
jobmanagement | others
jobretired | others
jobself−employed | others
jobservices | others
jobstudent | others
jobtechnician | others
3220
7545


y | others
jobunemployed | others
jobunknown | others
maritalmarried | others
maritalsingle | others
maritalunknown | others
contactsocialMedia | others
monthaug | others
monthdec | others
monthjul | others
3220
7545
29981


y | others
monthjun | others
monthmar | others
monthmay | others
monthnov | others
monthoct | others
monthsep | others
campaign | others
previous | others
poutcomenonexistent | others
3220
7545
29981

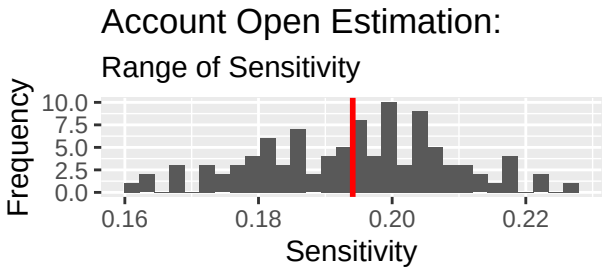


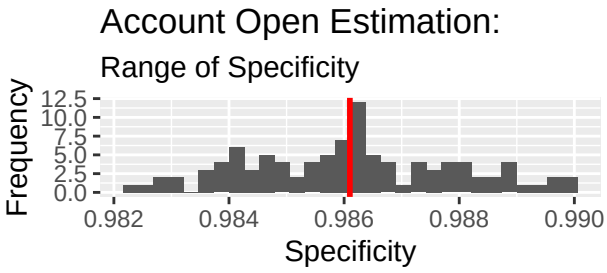

```
## Warning: Computation failed in ‘stat_smooth()‘:
## x has insufficient unique values to support 10 knots: reduce k.
## Computation failed in ‘stat_smooth()‘:
## x has insufficient unique values to support 10 knots: reduce k.
```

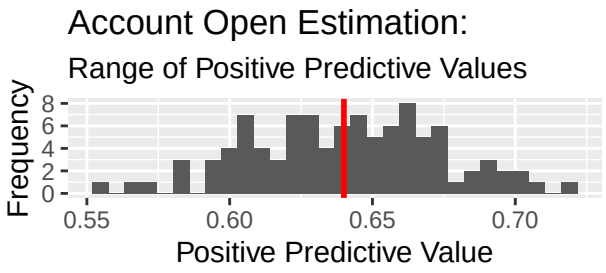


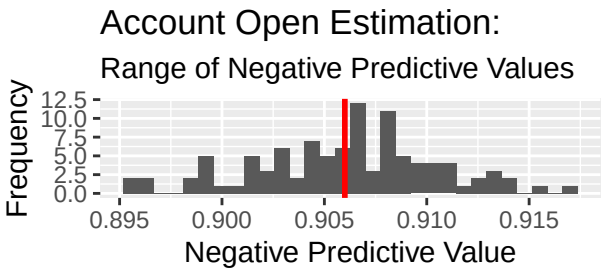
As can be seen in the graph above, the threshold for classification that minimizes the misclassification rate is 0.4839, as this is where the relationship between the threshold and the error rate was lowest recorded.

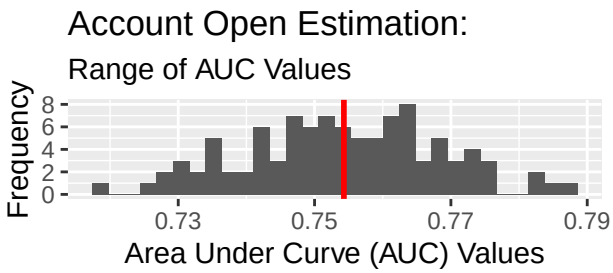
To test the predictive ability of the logistic regression model I created, I ran a cross-validation study classifying clients as either will open or will not open an account, using the previously found threshold of 0.4839. I found the cross validation generally produced average results of sensitivity, specificity, positive predicted value, and negative predicted value of 0.1941, 0.9861, 0.64, and 0.906, respectively. To help with proving the cross validation, I included graphics below that illustrate the 100 tests done to prove the predictive ability of the new logistic regression model.











Section 4: Results

Below is the fitted model for the data set with my best estimates of the coefficients for β_j:

$\hat{y}_i \overset{ind}{\sim} \text{Bern}(p_i)$,$log(\frac{p_i}{1-p_i})$ = -2.5613 + -0.3914 x_i I(Job=Blue-Collar)+ -0.2016 x_i I(Job=Entrepreneur)+ -0.1595 x_i I(Job=Housemaid)+ -0.0933 x_i I(Job=Management)+ 0.531 x_i I(Job=Retired)+ -0.1218 x_i I(Job=Self-Employed)+ -0.2992 x_i I(Job=Services)+ 0.5013 x_i I(Job=Student)+ -0.157 x_i I(Job=Technician)+ -0.0004 x_i I(Job=Unemployed)+ -0.1826 x_i I(Job=Unknown)+ 0.0451 x_i I(Marital=Married)+ NA x_i I(Marital=Single)+ 0.3119 x_i I(Marital=Unknown)+ 0.9821 x_i I(Contact=Social Media)+ 1.1673 x_i I(Month=March)+ -0.7889 x_i I(Month=May)+ -0.0298 x_i I(Month=June)+ -0.7631 x_i I(Month=July)+ -0.8401 x_i I(Month=August)+ 0.5968 x_i I(Month=September)+ 0.7531 x_i I(Month=October)+ -0.8744 x_i I(Month=November)+ 0.9179 x_i I(Month=December)+ -0.0744 x_i (Campaign)+ 0.2081 x_i (Previous)+ 0.2976 x_i I(Previous Oucome=Nonexistent)+ 0.8432 x_i I(Previous Outcome=Success)+ -0.0933 x_i I(Job=Management)+ 1.4338 x_i (Previous Contact)·

Based on the fitted model above, I think are the effects of the above, selected variables on whether a client opens an account or not to be divisible in three parts: a positive effect, a negative effect, and a negligible effect. I will list each group below in those lists, with a level of uncertainty.

The variables that have a **positive** effect on the likelihood of opening an account:
- I am 95% confident that if a client has the job type “retired”, all other variables remaining the same, that the likelihood of the client opening an account will increase between the range of 47.1807% and 96.2056%.
- I am 95% confident that if a client has the job type “student”, all other variables remaining the same, that the likelihood of the client opening an account will increase between the range of 37.3827% and 97.9273%.
- I am 95% confident that if a client has the marital status “single”, all other variables remaining the same,

that the likelihood of the client opening an account will increase between the range of 7.8024% and 37.7142%.
- I am 95% confident that if a client is contacted via social media, all other variables remaining the same, that the likelihood of the client opening an account will increase between the range of 141.3584% and 195.6131%.
- I am 95% confident that if a client's last contact month of the year is September, all other variables remaining the same, that the likelihood of the client opening an account will increase between the range of 46.0922% and 125.4872%.
- I am 95% confident that if a client's last contact month of the year is October, all other variables remaining the same, that the likelihood of the client opening an account will increase between the range of 74.5147% and 158.2149%.
- I am 95% confident that if a client's last contact month of the year is December, all other variables remaining the same, that the likelihood of the client opening an account will increase between the range of 76.9144% and 253.3853%. - I am 95% confident that if the number of contacts performed before this campaign and for this client increase by 1, all other variables remaining the same, that the likelihood of the client opening an account will increase between the range of 10.3436% and 37.586%.
- I am 95% confident that if the outcome of contacting the client in the previous marketing campaign was nonexistent, all other variables remaining the same, that the likelihood of the client opening an account will increase between the range of 13.8128% and 59.7061%.
- I am 95% confident that if the outcome of contacting the client in the previous marketing campaign was a success, all other variables remaining the same, that the likelihood of the client opening an account will increase between the range of 57.242% and 243.6725%.
- I am 95% confident that if the client having been previously contacted in a different campaign is true, all other variables remaining the same, that the likelihood of the client opening an account will increase between the range of 181.9219% and 523.9078%.

The variables that have a **negative** effect on the likelihood of opening an account:
- I am 95% confident that if a client has the job type "blue-collar", all other variables remaining the same, that the likelihood of the client opening an account will decrease between the range of -39.2904% and -24.763%.
- I am 95% confident that if a client has the job type "entrepreneur", all other variables remaining the same, that the likelihood of the client opening an account will decrease between the range of -33.5871% and -0.2066%.
- I am 95% confident that if a client has the job type "services", all other variables remaining the same, that the likelihood of the client opening an account will decrease between the range of -35.4046% and -15.1249%.
- I am 95% confident that if a client has the job type "technician", all other variables remaining the same, that the likelihood of the client opening an account will decrease between the range of -23.0188% and -5.1775%.
- I am 95% confident that if a client's last contact month of the year is May, all other variables remaining the same, that the likelihood of the client opening an account will decrease between the range of -59.9503% and -48.4239%.
- I am 95% confident that if a client's last contact month of the year is July, all other variables remaining the same, that the likelihood of the client opening an account will decrease between the range of -59.1331% and -46.7956%.
- I am 95% confident that if a client's last contact month of the year is August, all other variables remaining the same, that the likelihood of the client opening an account will decrease between the range of -62.2297% and -50.6476%.
- I am 95% confident that if a client's last contact month of the year is November, all other variables remaining the same, that the likelihood of the client opening an account will decrease between the range of -63.984% and -51.7232%.
- I am 95% confident that if the number of contacts performed during this campaign and for this client increase by 1, all other variables remaining the same, that the likelihood of the client opening an account will decrease between the range of -8.8655% and -5.5031%.

The variables that have a **negligible** effect on the likelihood of opening an account:
- I am 95% confident that if a client has the job type "housemaid", all other variables remaining the same, that the likelihood of the client opening an account will be between the range of -32.5407% and 6.6816%.
- I am 95% confident that if a client has the job type "management", all other variables remaining the same,

that the likelihood of the client opening an account will be changed between the range of -20.941% and 4.7067%.
- I am 95% confident that if a client has the job type "self-employed", all other variables remaining the same, that the likelihood of the client opening an account will be changed between the range of -27.1722% and 6.9261%.
- I am 95% confident that if a client has the job type "unemployed", all other variables remaining the same, that the likelihood of the client opening an account will be changed between the range of -19.2427% and 22.8832%.
- I am 95% confident that if a client has the job type "unknown", all other variables remaining the same, that the likelihood of the client opening an account will be changed between the range of -44.826% and 22.0677%.
- I am 95% confident that if a client has the marital status "married", all other variables remaining the same, that the likelihood of the client opening an account will be changed between the range of -6.5281% and 17.3062%.
- I am 95% confident that if a client has the marital status "unknown", all other variables remaining the same, that the likelihood of the client opening an account will be changed between the range of -36.4393% and 166.5326%.
- I am 95% confident that if a client's last contact month of the year is June, all other variables remaining the same, that the likelihood of the client opening an account will be changed between the range of -16.5444% and 12.9172%.

From the above ranges, it can be seen that social media contact is effective in marketing by anywhere between 141.3584% and 195.6131%, so there is strong evidence that social media contact is more effective than calling. Additionally, for repeated contacting, if the number of contacts performed before this campaign and for this client increase by 1, all other variables remaining the same, that the likelihood of the client opening an account will increase between the range of 10.3436% and 37.586%, so there is strong evidence that repeated contacting will increase the likelihood of a person opening a new account.

The file `ShouldWeContact.csv` contains information on a few people that the bank is considering contacting. My recommendation is to not contact them, as the likelihood of them opening a new account is well below my found threshold of 48.39%. If the bank still wishes to try in order to increase the previous contact number for future campaigns, however, I would contact them, in order from most likely to least likely: the single administrator at 11.9272%, the married 38-year-old administrator at 10.4233%, the single technician at 10.3736%, the married technician at 8.7671%, the divorced management worker at 8.0504%, the married 27-year-old administrator at 7.3416%, the married blue-collar worker (who was contacted on social media first) at 6.6605%, the married 48-year-old administrator at 4.0843%, the divorced technician at 4.0366%, and the married 48-year-old blue-collar worker (who was called first) at 2.7984%.

Section 5: Conclusions

It can be seen that the logistic regression model can predict if a client will open a new account. All the variables your bank collected, except age, helped increase the model's accuracy. Contacting via social media and with repeated contacts bothe increased those odds as well.

The bank executives could consider testing different packages of new account offers on clients to see if particularly low APR or cash back offers would make clients more likely to open a new account.

Appendix of Code

```
knitr::opts_chunk$set(echo = FALSE, include = FALSE)
library(forcats)
library(tidyverse)
library(vroom) #faster data reading
library(dplyr)
```

```r
library(DataExplorer) #for plot explorations
library(ggplot2) #for professional looking graphics
library(GGally) #for ggpairs
library(car) #for variance inflation factors & added-variable plots
library(bestglm) #for variable selection procedures
library(knitr) #for pretty tables with kable
# library(kableExtra) #for if kable needs to be landscape
library(gridExtra) #for making grids on same row
library(pROC) #for making a ROC curve on step 7
options(scipen = 999) #for preventing scientific notation
bank <- vroom("~/R programming/STAT_330/Marketing.csv", show_col_types = FALSE)
bank_contact <- vroom("~/R programming/STAT_330/ShouldWeContact.csv", show_col_types = FALSE)
bank <- bank %>% mutate(job=as.factor(job), marital=as.factor(marital),
                        contact=as.factor(contact), month=as.factor(month),
                        poutcome=as.factor(poutcome))
bank_contact <- bank_contact %>% mutate(job=as.factor(job), marital=as.factor(marital),
                        contact=as.factor(contact), month=as.factor(month),
                        poutcome=as.factor(poutcome))
bank_contact$pcontact <- (bank_contact$pcontact == 'yes') * 1
set.seed(86)
# (a)  In your own words, describe the background of the problem and the goals of the study. (What char
# (b)  Summarize what data you are going to use to fulfill the goals mentioned above.  Explore the data
job.bar <- ggplot(data = bank, aes(factor(job), ..count..)) +
  geom_bar(aes(fill = factor(y)), position = "stack")+
  xlab('Job Type') +
  ylab('Frequency') +
  ggtitle('Marketing Data: Job Type') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "top") +
  guides(fill = guide_legend(title = "Account Opened")) +
  scale_fill_hue(labels = c("No", "Yes"))

marital.bar <- ggplot(data = bank, aes(factor(marital), ..count..)) +
  geom_bar(aes(fill = factor(y)), position = "stack")+
  xlab('Marital Status') +
  ylab('Frequency') +
  ggtitle('Marketing Data: Marital Status') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "top") +
  guides(fill = guide_legend(title = "Account Opened")) +
  scale_fill_hue(labels = c("No", "Yes"))

contact.bar <- ggplot(data = bank, aes(factor(contact), ..count..)) +
  geom_bar(aes(fill = factor(y)), position = "stack")+
  xlab('Contact Type') +
  ylab('Frequency') +
  ggtitle('Marketing Data: Contact Type') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "top") +
  guides(fill = guide_legend(title = "Account Opened")) +
  scale_fill_hue(labels = c("No", "Yes"))

month.bar <- ggplot(data = bank, aes(factor(month), ..count..)) +
  geom_bar(aes(fill = factor(y)), position = "stack")+
  xlab('Month') +
  ylab('Frequency') +
```

```
  ggtitle('Marketing Data: Month') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "top") +
  guides(fill = guide_legend(title = "Account Opened")) +
  scale_fill_hue(labels = c("No", "Yes")) +
  scale_x_discrete(limits = c("mar", "apr", "may", "jun", "jul", "aug", "sep",
                              "oct", "nov", "dec"))

poutcome.bar <- ggplot(data = bank, aes(poutcome, ..count..)) +
  geom_bar(aes(fill = factor(y)), position = "stack") +
  guides(fill = guide_legend(title = "Account Opened")) +
  scale_fill_hue(labels = c("No", "Yes")) +
  theme(axis.text.x = element_text(angle = 65, hjust = 1), legend.position = "top") +
  xlab('Previous Outcome') +
  ylab('Frequency') +
  ggtitle('Marketing Data: Previous Outcome')

pcontact.bar <- ggplot(data = bank, aes(pcontact, ..count..)) +
  geom_bar(aes(fill = factor(y)), position = "stack") +
  theme(legend.position = "top") +
  guides(fill = guide_legend(title = "Account Opened")) +
  scale_fill_hue(labels = c("No", "Yes")) +
  xlab('Previous Contact Made') +
  ylab('Frequency') +
  ggtitle('Marketing Data: Previous Contact')

grid.arrange(month.bar, job.bar, nrow = 1)
grid.arrange(marital.bar, contact.bar, nrow = 1)
grid.arrange(poutcome.bar, pcontact.bar, nrow = 1)
bank$pcontact <- (bank$pcontact == 'yes') * 1
bank$y <- (bank$y == 'yes') * 1
plot_correlation(bank, type = "continuous")
plot_histogram(bank)
# (a)  Using justifiable techniques, identify which variables (if any) are important to include in your
# (b)  Mathematically write out your statistical model that you will use to achieve the goals of the an
vs.res <- bestglm(as.data.frame(bank), IC="AIC", method="exhaustive",
                  TopModels=1, family=binomial)
#vs.res$BestModel
pred.logreg <- glm(y~job+marital+contact+month+campaign+previous+poutcome+pcontact,
                   bank, family = binomial)
# (a)  Describe and justify each assumption you are making in your statistical model.
# (b)  How well does your model fit all of the data? Justify your answer by selecting an appropriate cu
# (c)  How well does your model do at predicting new customers who will open an account? Justify your a
avPlots(pred.logreg, ask = FALSE)

job.scatter <- ggplot(data = bank, mapping=aes(y=y, x=job)) +
  geom_point() +
  #geom_jitter(width=0.15,height=0.15) +
  # xlab('Glucose') +
  # ylab('Diabetes Diagnosis') +
  # ggtitle('Linearity Assumption Scatterplot', subtitle = 'Gluscose Level') +
  geom_smooth()

marital.scatter <- ggplot(data = bank, mapping=aes(y=y, x=marital)) +
```

```
  geom_point() +
  #geom_jitter(width=0.15,height=0.15) +
  # xlab('Body Mass Index') +
  # ylab('Diabetes Diagnosis') +
  # ggtitle('Linearity Assumption Scatterplot', subtitle = 'Body Mass Index') +
  geom_smooth()

contact.scatter <- ggplot(data = bank, mapping=aes(y=y, x=contact)) +
  geom_point() +
  #geom_jitter(width=0.15,height=0.15) +
  # xlab('Pedigree Strength of Diabetes') +
  # ylab('Diabetes Diagnosis') +
  # ggtitle('Linearity Assumption Scatterplot',
  #         subtitle = 'Pedigree Strength of Diabetes') +
  geom_smooth()

month.scatter <- ggplot(data = bank, mapping=aes(y=y, x=month)) +
  geom_point() +
  #geom_jitter(width=0.15,height=0.15) +
  # xlab('Age') +
  # ylab('Diabetes Diagnosis') +
  # ggtitle('Linearity Assumption Scatterplot', subtitle = 'Age') +
  geom_smooth()

campaign.scatter <- ggplot(data = bank, mapping=aes(y=y, x=campaign)) +
  geom_point() +
  #geom_jitter(width=0.15,height=0.15) +
  # xlab('Age') +
  # ylab('Diabetes Diagnosis') +
  # ggtitle('Linearity Assumption Scatterplot', subtitle = 'Age') +
  geom_smooth()

previous.scatter <- ggplot(data = bank, mapping=aes(y=y, x=previous)) +
  geom_point() +
  #geom_jitter(width=0.15,height=0.15) +
  # xlab('Age') +
  # ylab('Diabetes Diagnosis') +
  # ggtitle('Linearity Assumption Scatterplot', subtitle = 'Age') +
  geom_smooth()

poutcome.scatter <- ggplot(data = bank, mapping=aes(y=y, x=poutcome)) +
  geom_point() +
  #geom_jitter(width=0.15,height=0.15) +
  # xlab('Age') +
  # ylab('Diabetes Diagnosis') +
  # ggtitle('Linearity Assumption Scatterplot', subtitle = 'Age') +
  geom_smooth()

pcontact.scatter <- ggplot(data = bank, mapping=aes(y=y, x=pcontact)) +
  geom_point() +
  #geom_jitter(width=0.15,height=0.15) +
  # xlab('Age') +
  # ylab('Diabetes Diagnosis') +
```

```
  # ggtitle('Linearity Assumption Scatterplot', subtitle = 'Age') +
  geom_smooth() #method='glm'

suppressMessages(grid.arrange(job.scatter, marital.scatter, contact.scatter,
                              month.scatter, campaign.scatter, previous.scatter,
                              poutcome.scatter, pcontact.scatter, nrow = 3))

bank.probs <- predict.glm(pred.logreg, type="response")
thresh <- seq(from=0, to=1, length=10000)
#Empty vector to hold misclassification rates
misclass <- rep(NA,length=length(thresh))
for(i in 1:length(thresh)) {
  #If probability greater than threshold then 1 else 0
  my.classification <- ifelse(bank.probs>thresh[i], 1, 0)
  # calculate the pct where my classification not eq truth
  misclass[i] <- mean(my.classification!=bank$y)
}

#Find threshold which minimizes misclassification
threshold <- thresh[which.min(misclass)]
threshold <- round(threshold, digits = 4)
thresh.plot <- ggplot() +
  geom_line(aes(x=thresh, y=misclass)) +
  xlab("Threshold") +
  ylab("Error Rate") +
  ggtitle("Threshold vs. Misclassification Rate") +
  geom_vline(xintercept = threshold, col = "red", lwd = 1) +
  annotate("text", x = threshold + 0.2, y = 0.3,
           label = paste("Minimum Threshold =", threshold), col = "red", size = 3.5)
print(thresh.plot)

#ROC (Receiver Operating Characteristic) Curve
my.roc <- roc(bank$y, bank.probs, levels = c(0,1), direction = "<")
thresh.check <- round(auc(my.roc)[1], digits = 4)
roc.plot <- ggplot() +
  geom_line(aes(x=1-my.roc[["specificities"]], y=my.roc[["sensitivities"]])) +
  geom_abline(intercept=0, slope=1, color ='red') +
  xlab("Percent of True Negatives") +
  ylab("Percent of True Positives") +
  ggtitle("Receiver Operating Characteristic Curve")
print(roc.plot)

#confusion matrix
best.class <- ifelse(bank.probs>threshold, 1, 0)
#table(diagnosis$diabetes, best.class)
confusion.mat <- addmargins(table("Predicted" = best.class, "Actual" = bank$y))
confusion.mat
# Choose number of CV studies to run in a loop & test set size
n.cv <- 100
n.test <- round(.1*nrow(bank))

## Set my threshold for classifying
cutoff <- threshold
```

```r
## Initialize matrices to hold CV results
sens <- rep(NA, n.cv)
spec <- rep(NA, n.cv)
ppv <- rep(NA, n.cv)
npv <- rep(NA, n.cv)
auc <- rep(NA, n.cv)

## Begin for loop
for(cv in 1:n.cv){
  ## Separate into test and training sets
  test.obs <- sample(1:nrow(bank), n.test)
  test.set <- bank[test.obs,]
  train.set <- bank[-test.obs,]

  ## Fit best model to training set
  train.model <- glm(y~job+marital+contact+month+campaign+previous+poutcome+pcontact,
                     data=train.set, family=binomial)

  ## Use fitted model to predict test set
  pred.probs <- predict.glm(train.model,newdata=test.set, type="response")
  #response gives probabilities

  ## Classify according to threshold
  test.class <- ifelse(pred.probs>cutoff, 1, 0)

  ## Create a confusion matrix
  conf.mat <- as.data.frame.matrix(
    addmargins(table("Actual"=factor(
      test.set$y, levels=c(0,1)),
      "pred" = factor(test.class, levels=c(0,1)))))

  ## Pull of sensitivity, specificity, PPV and NPV using bracket notation
  sens[cv] <- conf.mat[2,2]/conf.mat[2,3]
  spec[cv] <- conf.mat[1,1]/conf.mat[1,3]
  ppv[cv] <- conf.mat[2,2]/conf.mat[3,2]
  npv[cv] <- conf.mat[1,1]/conf.mat[3,1]

  ## Calculate AUC
  auc[cv] <- auc(roc(test.set$y, pred.probs, levels = c(0,1), direction = "<"))
} #End for-loop

mean.sens <- round(mean(sens), digits = 4)
mean.spec <- round(mean(spec), digits = 4)
mean.ppv <- round(mean(ppv), digits = 4)
mean.npv <- round(mean(npv), digits = 4)
mean.auc <- round(mean(auc), digits = 4)
CV.sens <- ggplot() +
  geom_histogram(mapping=aes(x=sens)) +
  xlab('Sensitivity') +
  ylab('Frequency') +
  ggtitle('Account Open Estimation:',
          subtitle = 'Range of Sensitivity') +
  geom_vline(xintercept = mean.sens, col = "red", lwd = 1)
```

```
CV.spec <- ggplot() +
  geom_histogram(mapping=aes(x=spec)) +
  xlab('Specificity') +
  ylab('Frequency') +
  ggtitle('Account Open Estimation:',
          subtitle = 'Range of Specificity') +
  geom_vline(xintercept = mean.spec, col = "red", lwd = 1)

CV.ppv <- ggplot() +
  geom_histogram(mapping=aes(x=ppv)) +
  xlab('Positive Predictive Value') +
  ylab('Frequency') +
  ggtitle('Account Open Estimation:',
          subtitle = 'Range of Positive Predictive Values') +
  geom_vline(xintercept = mean.ppv, col = "red", lwd = 1)

CV.npv <- ggplot() +
  geom_histogram(mapping=aes(x=npv)) +
  xlab('Negative Predictive Value') +
  ylab('Frequency') +
  ggtitle('Account Open Estimation:',
          subtitle = 'Range of Negative Predictive Values') +
  geom_vline(xintercept = mean.npv, col = "red", lwd = 1)

CV.auc <- ggplot() +
  geom_histogram(mapping=aes(x=auc)) +
  xlab('Area Under Curve (AUC) Values') +
  ylab('Frequency') +
  ggtitle('Account Open Estimation:',
          subtitle = 'Range of AUC Values') +
  geom_vline(xintercept = mean.auc, col = "red", lwd = 1)

suppressMessages(grid.arrange(CV.sens, CV.spec, CV.ppv, CV.npv, CV.auc, nrow=3))
# (a)  Based on your fitted model, what do you think are the effect(s) (if any) of the selected variabl
# (b)  Is there evidence that social media vs. personal contact is more effective in marketing?
# (c)  Does repeated contacting seem to increase the likelihood of a person taking out an account?
# (d)  The file `ShouldWeContact.csv` contains information on a few people that the company is consider
bank.beta.0 <- round(as.numeric(coef(pred.logreg)["(Intercept)"]), digits = 4)
bank.beta.1 <- round(as.numeric(coef(pred.logreg)["jobblue-collar"]), digits = 4)
bank.beta.2 <- round(as.numeric(coef(pred.logreg)["jobentrepreneur"]), digits = 4)
bank.beta.3 <- round(as.numeric(coef(pred.logreg)["jobhousemaid"]), digits = 4)
bank.beta.4 <- round(as.numeric(coef(pred.logreg)["jobmanagement"]), digits = 4)
bank.beta.5 <- round(as.numeric(coef(pred.logreg)["jobretired"]), digits = 4)
bank.beta.6 <- round(as.numeric(coef(pred.logreg)["jobself-employed"]), digits = 4)
bank.beta.7 <- round(as.numeric(coef(pred.logreg)["jobservices"]), digits = 4)
bank.beta.8 <- round(as.numeric(coef(pred.logreg)["jobstudent"]), digits = 4)
bank.beta.9 <- round(as.numeric(coef(pred.logreg)["jobtechnician"]), digits = 4)
bank.beta.10 <- round(as.numeric(coef(pred.logreg)["jobunemployed"]), digits = 4)
bank.beta.11 <- round(as.numeric(coef(pred.logreg)["jobunknown"]), digits = 4)
bank.beta.12 <- round(as.numeric(coef(pred.logreg)["maritalmarried"]), digits = 4)
bank.beta.13 <- round(as.numeric(coef(pred.logreg)["martalsingle"]), digits = 4)
bank.beta.14 <- round(as.numeric(coef(pred.logreg)["maritalunknown"]), digits = 4)
bank.beta.15 <- round(as.numeric(coef(pred.logreg)["contactsocialMedia"]), digits = 4)
```

```r
bank.beta.16 <- round(as.numeric(coef(pred.logreg)["monthmar"]), digits = 4)
bank.beta.17 <- round(as.numeric(coef(pred.logreg)["monthmay"]), digits = 4)
bank.beta.18 <- round(as.numeric(coef(pred.logreg)["monthjun"]), digits = 4)
bank.beta.19 <- round(as.numeric(coef(pred.logreg)["monthjul"]), digits = 4)
bank.beta.20 <- round(as.numeric(coef(pred.logreg)["monthaug"]), digits = 4)
bank.beta.21 <- round(as.numeric(coef(pred.logreg)["monthsep"]), digits = 4)
bank.beta.22 <- round(as.numeric(coef(pred.logreg)["monthoct"]), digits = 4)
bank.beta.23 <- round(as.numeric(coef(pred.logreg)["monthnov"]), digits = 4)
bank.beta.24 <- round(as.numeric(coef(pred.logreg)["monthdec"]), digits = 4)
bank.beta.25 <- round(as.numeric(coef(pred.logreg)["campaign"]), digits = 4)
bank.beta.26 <- round(as.numeric(coef(pred.logreg)["previous"]), digits = 4)
bank.beta.27 <- round(as.numeric(coef(pred.logreg)["poutcomenonexistent"]), digits = 4)
bank.beta.28 <- round(as.numeric(coef(pred.logreg)["poutcomesuccess"]), digits = 4)
bank.beta.29 <- round(as.numeric(coef(pred.logreg)["pcontact"]), digits = 4)
bank.prob <- predict.glm(pred.logreg, newdata = bank_contact, type = "response")
pred.perc <- round(100 * as.numeric(bank.prob), digits = 4)
# (a)  Briefly summarize the main findings of your analysis in 1 paragraph and without using statistica
# (b)  Identify 1-2 "next steps" that the executives should consider to better understand and predict c
#End of exam's code
```