R Notebook

Take Home Final

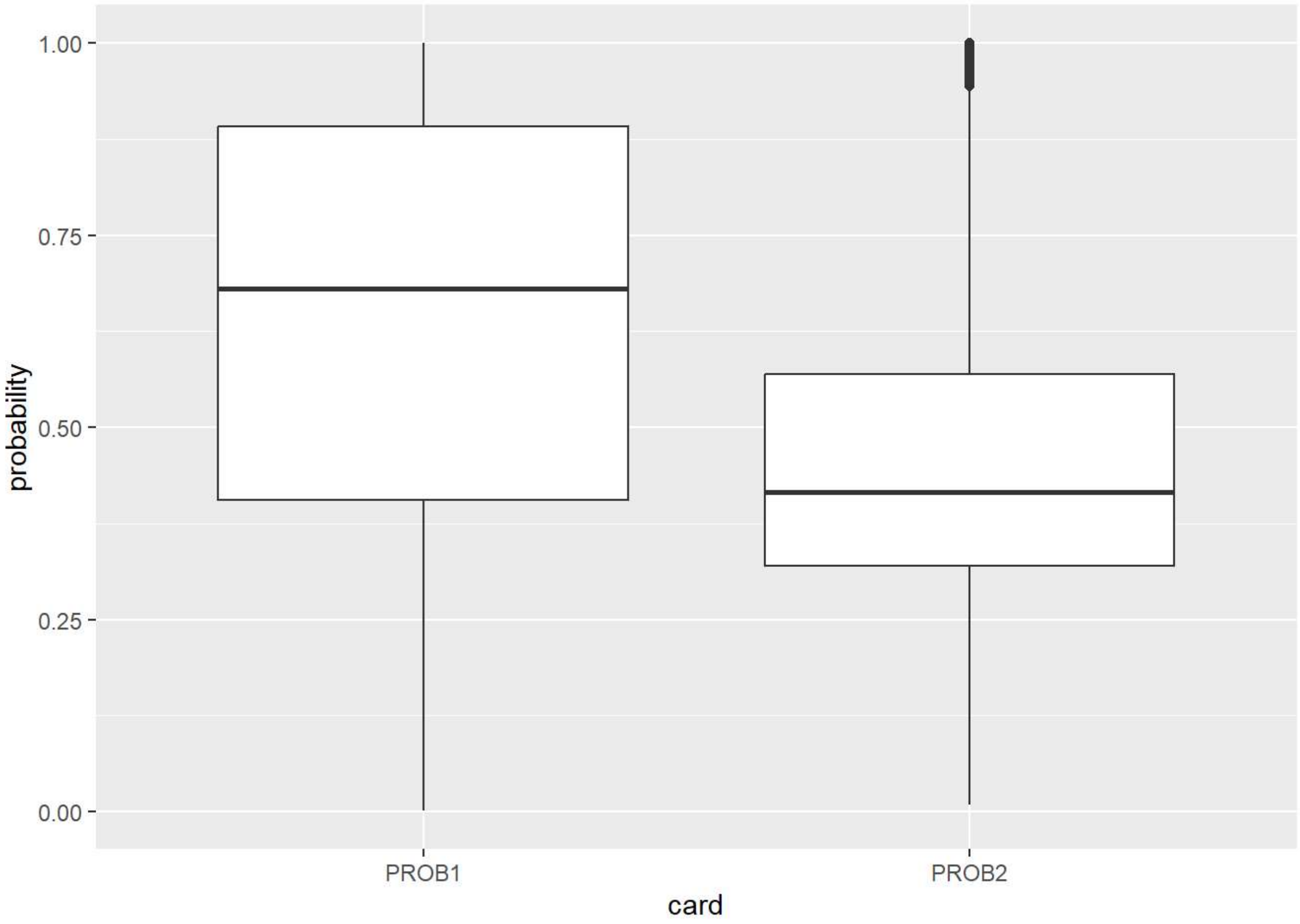
Load Libraries and Data

```
library(tidyverse)
library(readxl)
library(ggplot2)
library(caret)
library(pROC)

bancaja_cc <- read_excel("Bancaja-Developing Customer Intelligence Spreadsheet.xlsx", sheet = 2)
```

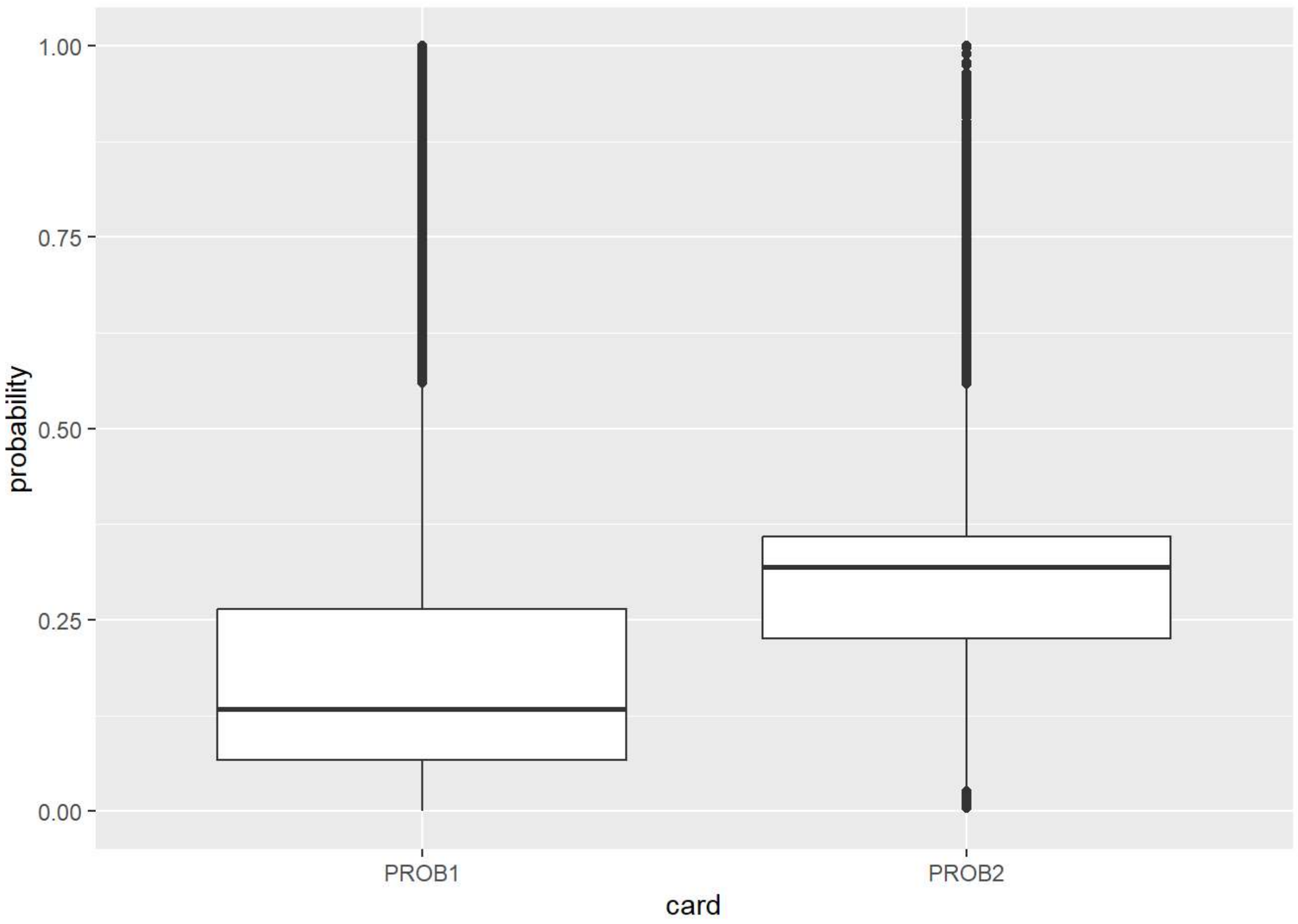
Question 1

In the Bancja case, you are presented with predicted probabilities of two credit card ownership models: PROB1 and PROB2. Which of these two models do you think is better? Justify your choice.

```
bancaja_cc %>%
  filter(HASCARD == 1) %>%
  pivot_longer(cols = PROB1:PROB2, names_to = "card", values_to = 'probability') %>%
  ggplot(aes(x = card, y = probability))+
  geom_boxplot()
```



```
bancaja_cc %>%
  filter(HASCARD == 0) %>%
  pivot_longer(cols = PROB1:PROB2, names_to = "card", values_to = 'probability') %>%
  ggplot(aes(x = card, y = probability))+
  geom_boxplot()
```



```
prob1_roc <- roc(bancaja_cc$HASCARD, bancaja_cc$PROB1)
prob2_roc <- roc(bancaja_cc$HASCARD, bancaja_cc$PROB2)

roc.test(prob1_roc, prob2_roc)
```

```
##
##  DeLong's test for two correlated ROC curves
##
## data:  prob1_roc and prob2_roc
## Z = 59.967, p-value < 2.2e-16
## alternative hypothesis: true difference in AUC is not equal to 0
## 95 percent confidence interval:
##  0.1592162 0.1699756
## sample estimates:
## AUC of roc1 AUC of roc2
##   0.8855892   0.7209933
```

Question 2

Assuming that Bancja can only send 5,000 mailings, which model would generate the most credit card sales? How would you know?

```
set.seed(123)
data_part <- createDataPartition(bancaja_cc$HASCARD, p = .8,
                                 times = 1, list = FALSE)

ban_train <- bancaja_cc[data_part, ]
ban_test <- bancaja_cc[-data_part, ]

q2_model_1 <- glm(HASCARD ~ PROB1, data = ban_train, family = 'binomial')

q2_model_2 <- glm(HASCARD ~ PROB2, data = ban_train, family = 'binomial')

q2_pred_1 <- predict.glm(q2_model_1, newdata = ban_test, type = 'response')
q2_pred_2 <- predict.glm(q2_model_2, newdata = ban_test, type = 'response')

summary(q2_model_1)
```

```
## 
## Call:
## glm(formula = HASCARD ~ PROB1, family = "binomial", data = ban_train)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.4239  -0.5515  -0.4067   0.4884   2.3896  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -2.80080    0.02980  -94.00   <2e-16 ***
## PROB1        5.68451    0.06364   89.32   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 36485  on 27999  degrees of freedom
## Residual deviance: 23220  on 27998  degrees of freedom
## AIC: 23224
## 
## Number of Fisher Scoring iterations: 4
```

```
summary(q2_model_2)
```

```
##
## Call:
## glm(formula = HASCARD ~ PROB2, family = "binomial", data = ban_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2750  -0.8530  -0.6967   1.0941   2.2075
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.39790    0.03474  -69.02   <2e-16 ***
## PROB2        4.90977    0.08632   56.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 36485  on 27999  degrees of freedom
## Residual deviance: 32468  on 27998  degrees of freedom
## AIC: 32472
##
## Number of Fisher Scoring iterations: 4
```

```
q2_conf_1 <- table(ban_test$HASCARD, q2_pred_1 > .5)
q2_conf_1
```

```
##
##     FALSE TRUE
##   0  4066  426
##   1   813 1695
```

```
q2_conf_2 <- table(ban_test$HASCARD, q2_pred_2 > .5)
q2_conf_2
```

```
##
##     FALSE TRUE
##   0  4108  384
##   1  1600  908
```

```
q2_accuracy_1 <- sum(diag(q2_conf_1)) / sum(q2_conf_1)
q2_accuracy_1
```

```
## [1] 0.823
```

```
q2_accuracy_2 <- sum(diag(q2_conf_2))/ sum(q2_conf_2)
q2_accuracy_2
```

```
## [1] 0.7165714
```

```
ban_samp <- bancaja_cc[sample(nrow(bancaja_cc), 5000), ]

samp_pred_1 <- predict.glm(q2_model_1, newdata = ban_samp, type = 'response')
samp_pred_2 <- predict.glm(q2_model_2, newdata = ban_samp, type = 'response')

samp_roc_1 <- roc(ban_samp$HASCARD, ban_samp$PROB1)
samp_roc_1_1 <- roc(ban_samp$HASCARD, samp_pred_1)
samp_roc_2 <- roc(ban_samp$HASCARD, ban_samp$PROB2)
samp_roc_2_1 <- roc(ban_samp$HASCARD, samp_pred_2)

roc.test(samp_roc_1, samp_roc_2)
```

```
## 
##  DeLong's test for two correlated ROC curves
## 
## data:  samp_roc_1 and samp_roc_2
## Z = 22.988, p-value < 2.2e-16
## alternative hypothesis: true difference in AUC is not equal to 0
## 95 percent confidence interval:
##  0.1498508 0.1777856
## sample estimates:
## AUC of roc1 AUC of roc2 
##   0.8921299   0.7283116
```

```
roc.test(samp_roc_1_1, samp_roc_2_1)
```

```
## 
##  DeLong's test for two correlated ROC curves
## 
## data:  samp_roc_1_1 and samp_roc_2_1
## Z = 22.988, p-value < 2.2e-16
## alternative hypothesis: true difference in AUC is not equal to 0
## 95 percent confidence interval:
##  0.1498508 0.1777856
## sample estimates:
## AUC of roc1 AUC of roc2 
##   0.8921299   0.7283116
```

```
sum(samp_pred_1 >= .5)
```

```
## [1] 1554
```

```
sum(samp_pred_2 >= .5)
```

```
## [1] 953
```

```
conf_1 <- table(ban_samp$HASCARD, samp_pred_1 > .5)
conf_1
```

```
## 
##      FALSE TRUE
##   0  2876  330
##   1   570 1224
```

```
conf_2 <- table(ban_samp$HASCARD, samp_pred_2 > .5)
conf_2
```

```
## 
##      FALSE TRUE
##   0  2905  301
##   1  1142  652
```

```
accuracy_1 <- sum(diag(conf_1)) / sum(conf_1)
accuracy_1
```

```
## [1] 0.82
```

```
accuracy_2 <- sum(diag(conf_2))/ sum(conf_2)
accuracy_2
```

```
## [1] 0.7114
```

Question 3

When you divide Bancja customers younger than 40 compared to 40 and over, does your perception of model performance for PROB1 and PROB2 differ? Why or why not?

```
over40 <- bancaja_cc %>%
  filter(AGE >= 40)

under40 <- bancaja_cc %>%
  filter(AGE < 40)

over_roc_1 <- roc(over40$HASCARD, over40$PROB1)
over_roc_2 <- roc(over40$HASCARD, over40$PROB2)

under_roc_1 <- roc(under40$HASCARD, under40$PROB1)
under_roc_2 <- roc(under40$HASCARD, under40$PROB2)

roc.test(over_roc_1, over_roc_2)
```

```
## 
##  DeLong's test for two correlated ROC curves
## 
## data:  over_roc_1 and over_roc_2
## Z = 49.374, p-value < 2.2e-16
## alternative hypothesis: true difference in AUC is not equal to 0
## 95 percent confidence interval:
##  0.1345362 0.1456590
## sample estimates:
## AUC of roc1 AUC of roc2 
##   0.8773641   0.7372666
```

```
roc.test(under_roc_1, under_roc_2)
```

```
## 
##  DeLong's test for two correlated ROC curves
## 
## data:  under_roc_1 and under_roc_2
## Z = 26.519, p-value < 2.2e-16
## alternative hypothesis: true difference in AUC is not equal to 0
## 95 percent confidence interval:
##  0.2116975 0.2454868
## sample estimates:
## AUC of roc1 AUC of roc2 
##   0.8857103   0.6571181
```

```
over_pred_1 <- predict.glm(q2_model_1, newdata = over40, type = 'response')
over_pred_2 <- predict.glm(q2_model_2, newdata = over40, type = 'response')

under_pred_1 <- predict.glm(q2_model_1, newdata = under40, type = 'response')
under_pred_2 <- predict.glm(q2_model_2, newdata = under40, type = 'response')

summary(over_pred_1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.05729 0.08964 0.15873 0.32244 0.52020 0.94703
```

```
summary(over_pred_2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.08478 0.22466 0.30497 0.35356 0.43306 0.92482
```

```
summary(under_pred_1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.05729 0.39726 0.77807 0.65516 0.91505 0.94699
```

```
summary(under_pred_2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.08761 0.21658 0.31844 0.38542 0.50773 0.92496
```

```
sum(over_pred_1 > .5)
```

```
## [1] 8079
```

```
sum(over_pred_2 > .5)
```

```
## [1] 5568
```

```
sum(under_pred_1 > .5)
```

```
## [1] 2502
```

```
sum(under_pred_2 > .5)
```

```
## [1] 925
```

Question 4

In the Bancja case, is there value in segmentation beyond age at 40? Why or why not?

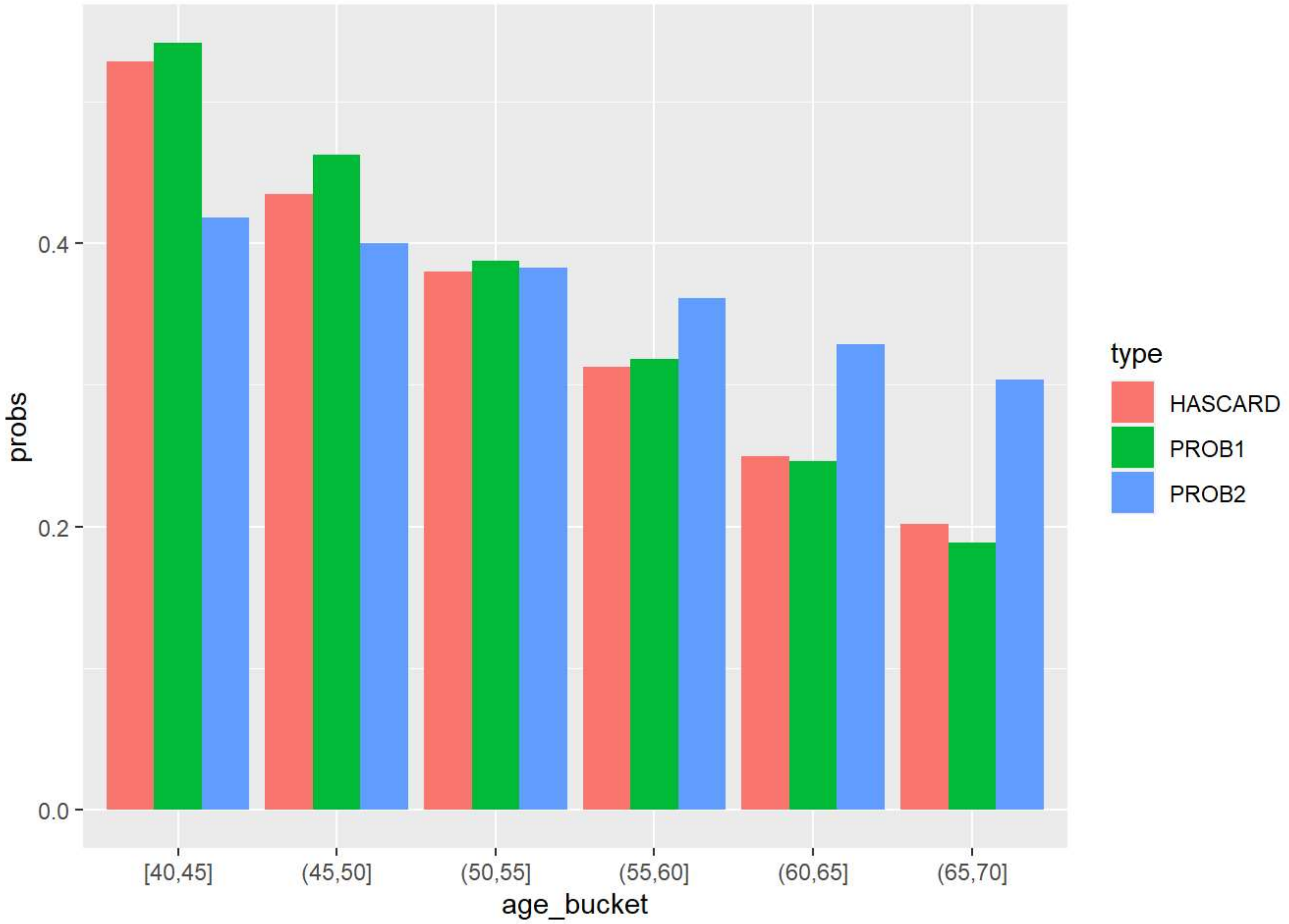
Split ages by 5 year buckets

```
q4 <- over40[over40$AGE > 40, ]

q4$age_bucket <- cut(q4$AGE[q4$AGE > 40], breaks = seq(40, max(q4$AGE), by = 5), include.lowest
= TRUE)

avg_probs <- aggregate(cbind(PROB1, PROB2) ~ age_bucket, data = q4, FUN = mean)

HASCARD_prop <- aggregate(HASCARD ~ age_bucket, data = q4, FUN = mean)

seg_results <- merge(avg_probs, HASCARD_prop, by = "age_bucket")

seg_results
```

age_bucket	PROB1	PROB2	HASCARD
<fct>	<dbl>	<dbl>	<dbl>
(45,50]	0.4622538	0.3998946	0.4342934
(50,55]	0.3876557	0.3823305	0.3800637
(55,60]	0.3179203	0.3610584	0.3126214
(60,65]	0.2464726	0.3283140	0.2494216
(65,70]	0.1887890	0.3039484	0.2018810
[40,45]	0.5411233	0.4179431	0.5281662

6 rows

```
seg_results %>%
  pivot_longer(PROB1:HASCARD, names_to = 'type', values_to = 'probs') %>%
  ggplot(aes(x = age_bucket, y = probs, fill = type)) +
    geom_bar(stat = "identity", position = 'dodge')
```


0.4
0.2
0.0
probs
[40,45]
(45,50]
(50,55]
(55,60]
(60,65]
(65,70]
age_bucket
type
HASCARD
PROB1
PROB2