Отчет: стоит ли мало-активных клиентов банка мотивировать использовать больше продуктов?

Представим такую ситуацию: иновационный менеджер банка разработал идею нового продукта, нацеленного на использовани менее активными клиентами чтобы в дальнейшем они перешли в разряд активных. Внедрением такого продукта заинтересовалось руководство.

Продукт планируется быть глубоко интегрированным в экосистему работы банка, а значит будет иметь необходимость в использовании сразу большим числом клиентов из категории менее активных.

Нам же предстоит понять, нужно ли удерживать менее активных пользователей (большая ли часть из них покинула банк) и стоит ли приступать к практической разработке продукта вдобавок к основным, что уже имеются у банка, и может ли это оказаться плохой идеей и в перспективе повлиять на отток клиентов.

Вопрос, на который я постараюсь ответить звучит следующим образом: следует ли добавлять дополнительный продукт пользователем с небольшой активностью?

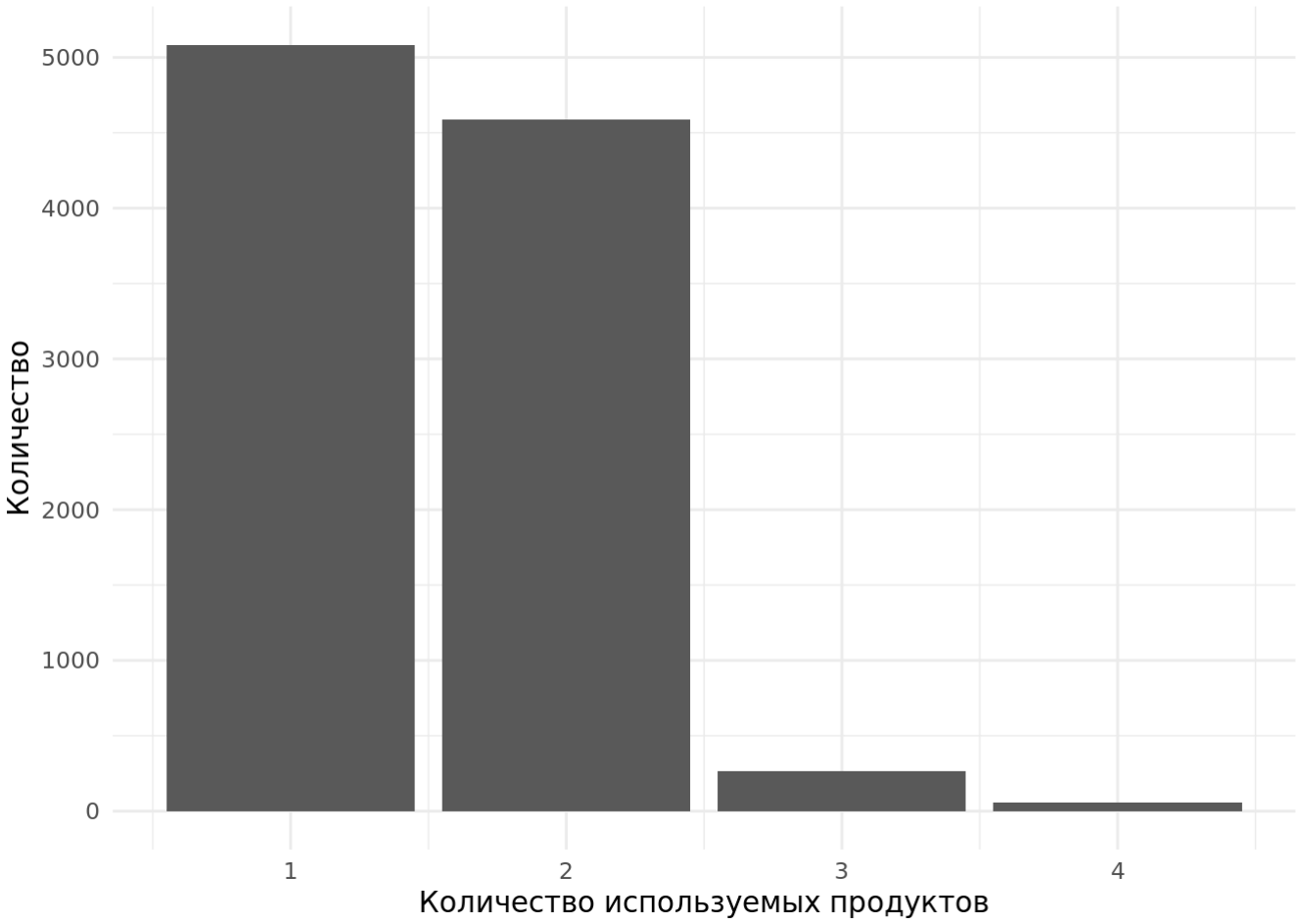
Содержание таблицы

В нашем случае интересны количество продуктов, активность клиента, и, конечно, покинул ли он банк

```
##  [1] "CustomerId"      "Surname"         "CreditScore"     "CountryId"
##  [5] "Gender"          "Age"             "Tenure"          "Balance"
##  [9] "NumOfProducts"   "HasCrCard"       "IsActiveMember"  "EstimatedSalary"
## [13] "Exited"
```

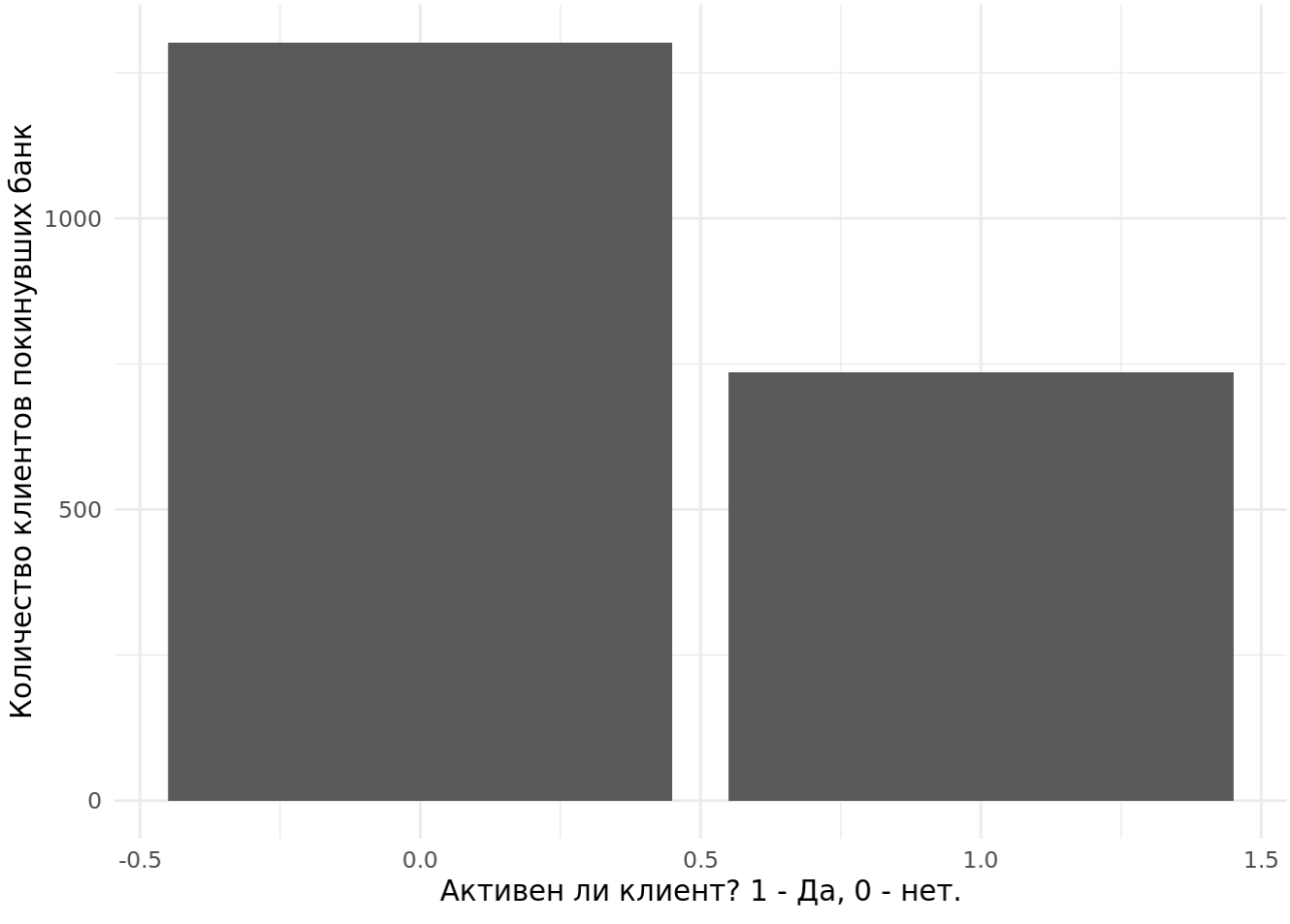
Посмотрим на распределение интересующей характеристики - количество продуктов

Можно заметить, что все клиенты банка используют минимум один продукт, значит, для удобства можно поделить данные в бинарный формат 1 (больше 1 продукта) и 0 (1 продукт), так как при внедрении нового продукта большое количество мало-активных клиентов станут пользоваться более чем одним продуктом



Посмотрим на менее активных клиентов

Видно, что отказались от услуг банка в наибольшей степени менее активные клиенты, значит можно посмотреть, влияет ли на их отказ количество продуктов, как и было запланированно ранее



Также лучше узнать более конкретные значения по характеристике покинувших банк клиентов среди мало-активных.

```
##   Exited Quantity
## 1      0     3547
## 2      1     1302
```

Теперь мы знаем, что клиентов покинувших банк в 3 раза меньше, чем клиентов пользующихся его услугами. Это может пригодиться для применения весов для характеристики, так как выборка несбалансированна.

Модели для анализа ситуации до внедрения нового продукта

Для начала построим модель дерева принятия решений с кросс валидацией и посмотрим на результат предсказания

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 709 260
##          1   0   0
##
##                Accuracy : 0.7317
##                  95% CI : (0.7026, 0.7594)
##     No Information Rate : 0.7317
##     P-Value [Acc > NIR] : 0.5167
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 1.0000
##             Specificity : 0.0000
##          Pos Pred Value : 0.7317
##          Neg Pred Value :    NaN
##              Prevalence : 0.7317
##          Detection Rate : 0.7317
##    Detection Prevalence : 1.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : 0
##
```

Модель вышла крайне плохой, все предсказания в категории ‘клиент не покинет банк’. Такая модель нам не подойдет. Вероятно, проблема кроется в несбалансированости выборки. Попробуем решить эту проблему.

Модель с применением весов

Попробуем применить веса к значению, количество которого гораздо меньше противоположного, в нашем случае это категория клиентов покинувших банк. Тут нам пригодится один из предыдущих пунктов. Мы помним, что клиентов покинувших банк в 3 раза меньше, чем оставшихся клиентов. Попробуем увеличить вес переменной в 3 раза.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 401  81
##          1 308 179
##
##                Accuracy : 0.5986
##                  95% CI : (0.5669, 0.6296)
##     No Information Rate : 0.7317
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.199
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.5656
##             Specificity : 0.6885
##          Pos Pred Value : 0.8320
##          Neg Pred Value : 0.3676
##              Prevalence : 0.7317
##          Detection Rate : 0.4138
##    Detection Prevalence : 0.4974
##       Balanced Accuracy : 0.6270
##
##        'Positive' Class : 0
##
```

Результат плохой, аккураси близко к 50%, что равно угадыванию значений случайным образом, сенситивити и спесифисити тоже не впечатляют

Модель с даун-сэмплингом

Попробуем еще один метод выравнивания несбалансированной выборки, может, выйдет получить результат лучше. На этот раз попробуем down-sampling.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 401  81
##          1 308 179
##
##                Accuracy : 0.5986
##                  95% CI : (0.5669, 0.6296)
##     No Information Rate : 0.7317
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.199
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.5656
##             Specificity : 0.6885
##          Pos Pred Value : 0.8320
##          Neg Pred Value : 0.3676
##              Prevalence : 0.7317
##          Detection Rate : 0.4138
##    Detection Prevalence : 0.4974
##       Balanced Accuracy : 0.6270
##
##        'Positive' Class : 0
##
```

Результат остался таким же

В итоге, лучший результат, который нам удалось получить это аккураси и сенситивити в 60% и спесифисити в 70%. В ситуации, что есть сейчас, можно посудить, что менее активные клиенты, пользующиеся одним продуктом, в большинстве своем не покидают банк. Вероятно, не стоит менять эту ситуацию.

Результат довольно плохой, но все же попробуем предсказать результат в том случае, если новый продукт будет разработан и будет использоваться менее активными пользователями.

Модель для предсказания оттока после внедрения нового продукта

Для реализации данной модели добавим каждому мало-активному клиенту дополнительный продукт.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1587  369
##          1    5   38
##
##                Accuracy : 0.8129
##                  95% CI : (0.7951, 0.8298)
##     No Information Rate : 0.7964
##     P-Value [Acc > NIR] : 0.03452
##
##                   Kappa : 0.1352
##
##  Mcnemar's Test P-Value : < 2e-16
##
##             Sensitivity : 0.99686
##             Specificity : 0.09337
##          Pos Pred Value : 0.81135
##          Neg Pred Value : 0.88372
##              Prevalence : 0.79640
##          Detection Rate : 0.79390
##    Detection Prevalence : 0.97849
##       Balanced Accuracy : 0.54511
##
##        'Positive' Class : 0
##
```

Модель вышла плохой. Вероятно, проблема кроется в несбалансированости выборки. Попробуем решить эту проблему.

Модель для предсказания с применением весов

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 790 167
##          1 802 240
##
##                Accuracy : 0.5153
##                  95% CI : (0.4931, 0.5374)
##     No Information Rate : 0.7964
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0544
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.4962
##             Specificity : 0.5897
##          Pos Pred Value : 0.8255
##          Neg Pred Value : 0.2303
##              Prevalence : 0.7964
##          Detection Rate : 0.3952
##    Detection Prevalence : 0.4787
##       Balanced Accuracy : 0.5430
##
##        'Positive' Class : 0
##
```

Результат плохой, аккураси близко к 50%, что равно угадыванию значений случайным образом, сенситивити и спесифисити также не впечатляют

Выводы

Даже если закрыть глаза на то, что результат сопоставим со случайным распределением (50%), можно увидеть, что доля клиентов, покинувших банк увеличилась. Поэтому, можно заключить, что внедрять подобный продукт для мало-активных пользователей не стоит.

Вероятно, подобные пользователи не заинтересованы в более продвинутом использовании банковских услуг, либо просто не нуждаются в них. Поэтому, попытка перевода их в активную категорию может сделать только хуже.

Однако, качество полученных моделей довольно низкое, а значит ответ на вопрос нельзя дать с максимальной уверенностью.