Analytics Exercise: Predicting Job Slot Product Retention

PA 446 Midterm

Alexis Kwan

We would like your help to better understand factors that affect Job Slots product retention. Please use the data provided in Exhibit 1 and 2 to answer the following questions.

We anticipate that it will take ~2-5 hours. You can analyze the data using whatever tools you feel will help you best extract insights.
Please present your responses to each question in a single PowerPoint deck (or PDF). In your write-up, include your assumptions and provide explanations for your answers.
Additionally, please feel free to include any supporting materials you produced while addressing these question (scripts, model outputs, visualizations etc.).

PART A: Data wrangling in R (for reference, the orignial assignment for Part A asked for SQL. You are required to use R)

1) Total Contract Value is defined as the total amount that customers committed to spend. Write R tidyverse pipe that returns the Total Contract Value ('total_contract_value' field) for each state (name) by month (using 'start_date') in the provided datasets.

```
joined = perf %>%
  left_join(locs, on=city_id) %>%
  mutate(start_date = mdy(start_date),
         start_month = month(start_date))
```

```
joined %>%
  group_by(state_name, start_month) %>%
  summarise(sum(total_contract_value))
```

```
## # A tibble: 120 x 3
## # Groups:   state_name [10]
##    state_name start_month `sum(total_contract_value)`
##    <chr>            <dbl>                       <dbl>
##  1 California           1                      426490
##  2 California           2                      425800
##  3 California           3                      540365
##  4 California           4                      214100
##  5 California           5                      316720
##  6 California           6                      330920
##  7 California           7                      423805
##  8 California           8                      433150
##  9 California           9                      411900
## 10 California          10                      508865
## # ... with 110 more rows
```

2) For all employers who purchased >1 product with [REDACTED], write a R tidyverse pipe to return the 'job_slots' and 'click_marketplace_value' values for the second transaction by employer.

*Note: You only have to provide the R code needed to accomplish the questions above; output results are not required

```
joined %>%
  arrange(employer_id, start_date) %>%
  group_by(employer_id) %>%
  mutate(trans_n = row_number(employer_id)) %>%
  arrange(employer_id) %>%
  filter(trans_n == 2)
```

```
## # A tibble: 7,039 x 17
## # Groups:   employer_id [7,039]
##    employer_id city_id contract_id start_date end_date   renewed_flag job_slots
##          <dbl>   <dbl>       <dbl> <date>     <chr>             <dbl>     <dbl>
##  1       95253    2407     2575140 2017-12-19 1/18/2018             1        15
##  2      231360    7807     1544592 2016-11-29 12/28/2016            1        15
##  3      266960    9299     1981272 2017-05-26 6/25/2017             1        15
##  4      268747    5386     1775124 2017-03-05 4/4/2017              0        15
##  5      279307    6116     1544544 2016-11-29 12/28/2016            0        15
##  6      279707    5201     1247448 2016-07-17 8/16/2016             1        15
##  7      304187    6606     1720188 2017-02-14 3/13/2017             1        15
##  8      333840    1580     1556052 2016-12-02 1/1/2017              1        15
##  9      337413    9716     2622360 2018-01-03 2/2/2018              0        15
## 10      339760   11373     1168128 2016-06-12 7/11/2016             1        15
## # ... with 7,029 more rows, and 10 more variables: total_contract_value <dbl>,
## #   applications <dbl>, apply_start_clicks <dbl>,
## #   click_marketplace_value <dbl>, job_listings <dbl>, city_name <chr>,
## #   state_id <chr>, state_name <chr>, start_month <dbl>, trans_n <int>
```

PART B: Metric Design
3) What metrics would you propose to measure the quality of services [REDACTED] provided to our clients? How does performance vary in terms of:
- job_slots?
- total_contract_value?
- click_marketplace_value?

Quality of service ought relate directly in some way to the deliverable or outcome that client cares about. In this case they care about the number of applications they get vs. the cost-per-application. It would be hard to directly influence number of applications that actually get through (due to external factors) we can at least influence traffic and clicks to some degree. Therefore we should look to a function of apply_start_clicks for our measure of quality.

```
perf %>% summarise(median(apply_start_clicks/job_slots, na.rm = T))
```

```
## # A tibble: 1 x 1
##   `median(apply_start_clicks/job_slots, na.rm = T)`
##                                               <dbl>
## 1                                              2.33
```

Based on the table above we can see that the client gets about 2 application start clicks for every job slot in a contract.

```
perf %>% summarise(median(apply_start_clicks/total_contract_value, na.rm = T))
```

```
## # A tibble: 1 x 1
##   `median(apply_start_clicks/total_contract_value, na.rm = T)`
##                                                          <dbl>
## 1                                                       0.0404
```

For total contract value, they get about median value of 0.04 clicks for every dollar spent on a contract.

```
perf %>% summarise(median(apply_start_clicks/click_marketplace_value, na.rm = T))
```

```
## # A tibble: 1 x 1
##   `median(apply_start_clicks/click_marketplace_value, na.rm = T)`
##                                                             <dbl>
## 1                                                          0.0327
```

In terms of the marketplace value, clients receive about 0.03 clicks per dollar of marketplace value.

PART C: Retention Modeling and Analysis
4) Which factors or combination of factors best correlate with an employer's likelihood to retain (i.e. renewed_flag = 1)? And how well does your chosen method correlate with retention? Please list any assumptions you made and explain why you chose your methodology.

```
joined = joined %>%
  mutate(market_diff = click_marketplace_value - total_contract_value)

glm1 = glm(renewed_flag ~ market_diff + apply_start_clicks, data = joined, family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(glm1)
```

```
##
## Call:
## glm(formula = renewed_flag ~ market_diff + apply_start_clicks,
##     family = binomial, data = joined)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.4559  -1.4080   0.7828   0.8953   2.1359
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          7.889e-01  1.298e-02   60.79   <2e-16 ***
## market_diff          2.101e-04  7.494e-06   28.03   <2e-16 ***
## apply_start_clicks -1.602e-03  1.251e-04  -12.81   <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: 45910  on 37756  degrees of freedom
## Residual deviance: 44526  on 37754  degrees of freedom
## AIC: 44532
##
## Number of Fisher Scoring iterations: 5
```

Market_diff represents how much value over or under the market rate that we are delivered to the client relative to the competition. apply_start_clicks shows how much value we deliver at the beginning of the application process. These are factors that directly relate to our quality of service and are levers related to different aspects of an application journey. I hypothesize that these two factors affect how likely a client is to renew a contract, and using logistic regression we can predict if they will renew or not.

5) Based on your analysis, what modifications would you recommend we make to our ad platform algorithm to improve retention?

Both of the factors decribed above appear to have a statistically significant relationship to the renewal status. Since the market differential has a positive affect on how likely a client is to renew, we ought to change the bidding algorithm to allow for a larger differential between the click market value versus the contract value. This could mean boosting market value for underperforming applications multiple times higher than before. Additionally since start clicks negatively affect the probability we should modify the search algorithm to direct fewer applicants to applications, perhaps focusing on applicants who have a higher probablity to finish an application, like those with higher relevance to the job, versus a larger pool of applicants.