Project-LTV 내부 공유용

Table of Contents

During the (2021.04.05) Presentation

- 우선, 전략팀/이매진의 궁금증은?
- Motivating Example
    - LTV 기여 what?

    ▼ list

```
address
completed_lectures_1day
completed_lectures_3days
completed_lectures_10days
completed_lectures_30days
completed_lectures_90days
completed_lectures_150days
nonmission_posts_1day
nonmission_posts_3days
nonmission_posts_10days
nonmission_posts_30days
nonmission_posts_90days
nonmission_posts_150days
posts_1day
posts_3days
posts_10days
posts_30days
posts_90days
posts_150days
comments_1day
comments_3days
comments_10days
comments_30days
comments_90days
comments_150days
cheers_before
cheers_10days
cheers_30days
cheers_90days
cheers_150days
completed_class
brand
boundness
internal_category
package_name
amount
klass_state
order_ticket_diff
lv_count_1day
lv_count_3days
lv_count_10days
lv_count_30days
lv_count_90days
lv_count_150days
lv_web_ratio_150days
lv_ios_ratio_150days
lv_android_ratio_150days
lv_app_ratio_150days
lv_distinct_hours_150days
count_distinctdays_3days
count_distinctdays_10days
count_distinctdays_30days
count_distinctdays_90days
count_distinctdays_150days
rebuy_365days
```

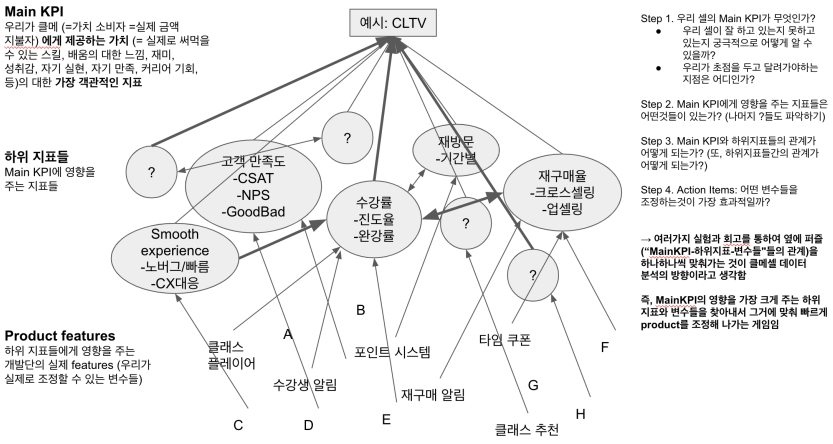
1. Research Questions

1. (The Big Question) 유저들의 LTV에 가장 크게 기여하는 변수 (특징, 행동, 등)들은 어떤것들이 있을까?
2. 해당 변수들이 상대적으로/절대적으로 LTV에 얼만큼 기여하는가?

2. Contexts

▼ Context 1: Project-LTV를 진행하게된 계기

1. 작년 11월 (전략팀→클메셀)로 옮기면서 전사적인 방향성에 대해서 생각하게 되었는데, **결국에 우리가 집중해야할 부분이 LTV가 아닌가, 그리고 LTV에 기여하는 변수들이 무엇인가 찾는것이 우선순위가 아닐까 제안한적이 있었음**
    - 클메셀 데이터 분석의 진행 방향에 관한 생각 1차 버전 (2020.11.10)
        - https://docs.google.com/presentation/d/1xtqqt97adPuMhLeGmg3nvB25_AgIBCoUEFV91mLFGuk/edit#slide=id.ga1



2. 이러한 프로젝트는 각 관계에 대한 개별 분석을 하나하나씩 진행하기엔 할게 넘 많고, 통제 변수들을 control할 수 없는 환경 속에서의 분석은 의미가 없다고 판단되어 Data Science 프로젝트로 가져갈 수 밖에 없었음
3. **클메셀 & 윌리랑 논의한 결과 아래 3개 이유에 의해 Project-LTV는 우선순위 3군 (1위 - 클메셀 수강시간, 2위 - 콘셀 PD 대시보드)으로 낮춰 side로 진행하게 되었음**
    a. 그때 당시 클메셀에 더 직접적인 지표는 새롭게 떠오르고 있던 "수강시간"이 더 적합하다고 판단되어 이 쪽을 먼저 더 파보기로함
    b. Data Science Project는 크림슨 1인이서 진행하기에 스펙이 너무 컸음
    c. 중간에 크림슨이 훈련소로 가버림 ㅎ
4. 훈련소 갔다와서 Project-LTV를 아에 Project-수강시간으로 진행하기로함
    a. Goal Variable이 LTV가 아닌 수강시간으로 정해서 프로젝트 진행 → 여기서 얻은 방법론/인사이트를 추후에 LTV-Project에 적용해보는 전략임
5. **최근 들어 (21년 3월 말) Project-수강시간에서 유의미한 결과가 나오기 시작하여 최근 일주일 가량 (2021.03.31 ~ 2021.04.05) Project-LTV로 분석을 진행해보았음**
    a. 그리고 때마침 이매진과 전략팀 쪽에서도 재구매 & 관련 변수들에 관한 질문들을 재시해주었음
6. **아래 분석 내용이 최근에 따끈따근하게 진행하게된 Project-LTV의 결과임**
    a. (클래스 리워드-얼버 전환율 예측 모델 프로젝트를 예시로) 보통 2-3명의 인원의 collaboration으로 1-2달 정도 소요될것으로 예상되는 프로젝트를 1인이 LEAN하고 쿽하게 진행하게 됨에 따라 결과들이 완전하지 않을 수 있음을 주의
    b. 더 큰 사이즈의 data processing, 다른 data scientist 친구들의 cross checking, prediction score 높이기, 등 과정 추가로 진행 후 결과 공유할 예정

▼ Context 2: 일반 데이터 분석과 데이터 과학 기반 분석은 장단점이 무엇인가?

- Data Science Main Pros
    1. 아래 예시 개별 분석 과제들을 **한꺼번에** 분석 가능
        - i) 게시글 수와 재구매의 관계
        - ii) web vs app 사용과 재구매의 관계
        - iii) 첫구매한 제품의 패키지 타입과 재구매의 관계
        - iv) 기타 이론적으론 무한대의 가정을 분석하는 과제가 있을 수 있음
    2. 한꺼번에 분석함으로써 **다른 변수들의 영향 최소화**
    3. 개별 변수와 재구매의 관계에 대한 **상대적 기여도 순위를** 메길 수 있음
- Data Science Main Cons

- 1. Data Collection & Processing에 노력/시간/노가다가 오래 소요
- 2. Data Science 인사이트 필요 (통계, 코딩, Polishing, 등)

3. Modelling Process (proj_ltv_ver5)

▼ User Group (Dataset Row Count ~= 30000)

- New Users (First Ever Ticket Start Date between the two dates below)
  - 2020.01.01 ≤ Ticket Start Date < 2020.04.01
- 이상한 데이터 잘라내고 총 3만명 가량 있음

▼ X: User Features (Dataset Column Count = 68)

- 각종 유저 변수들 (특징 & 행동 데이터) 20-30가지를 모음
- 몇 가지 변수들은 아래 기준들로 추가 정제 시킴:
  - Ticket Start Date 이후 1일간 / 3일간 / 10일간 / 30일간 / 90일간 / 150일간 각각의 기간별로
    - 예시로 "completed_lectures"라는 변수 하나는 각각의 기간별로 총 6개 변수로 쪼개짐
  - 정성 데이터는 0과 1 변수로 나오게금 변환 시킴:
    - 예시로 "brand"라는 변수 하나는 brand_career, brand_creative, brand_money, brand_others 이런식으로 변수 4개로 쪼개짐

▼ Y: Goal Variable LTV/재구매 기준

- **기준 1. rebuy_365_days → Main으로 사용한 Y (이유: 재구매율이 어짜피 낮아서 재구매 했는가 여부를 파악하는게 우선 과제라고 판단)**
  - 위 User Group의 각각 유저별로 Ticket Start Date 이후 365일간 재구매를 하였는가? (1 or 0)
- 기준 2. sum_revenue_365days
  - 위 User Group의 각각 유저별로 Ticket Start Date 이후 365일간의 총 구매 amount (순거래아님)

▼ Fitted Models

Model
Random Forest
Decision Tree
Logistic Regression
Naive Bayes
Support Vector Machines
Perceptron
Stochastic Gradient Decent
KNN
XGBoost

- 이중 Random Forest가 test_accuracy_score가 가장 높게 나와 Random Forest를 메인으로 사용함

4. Results (proj_ltv_ver5)

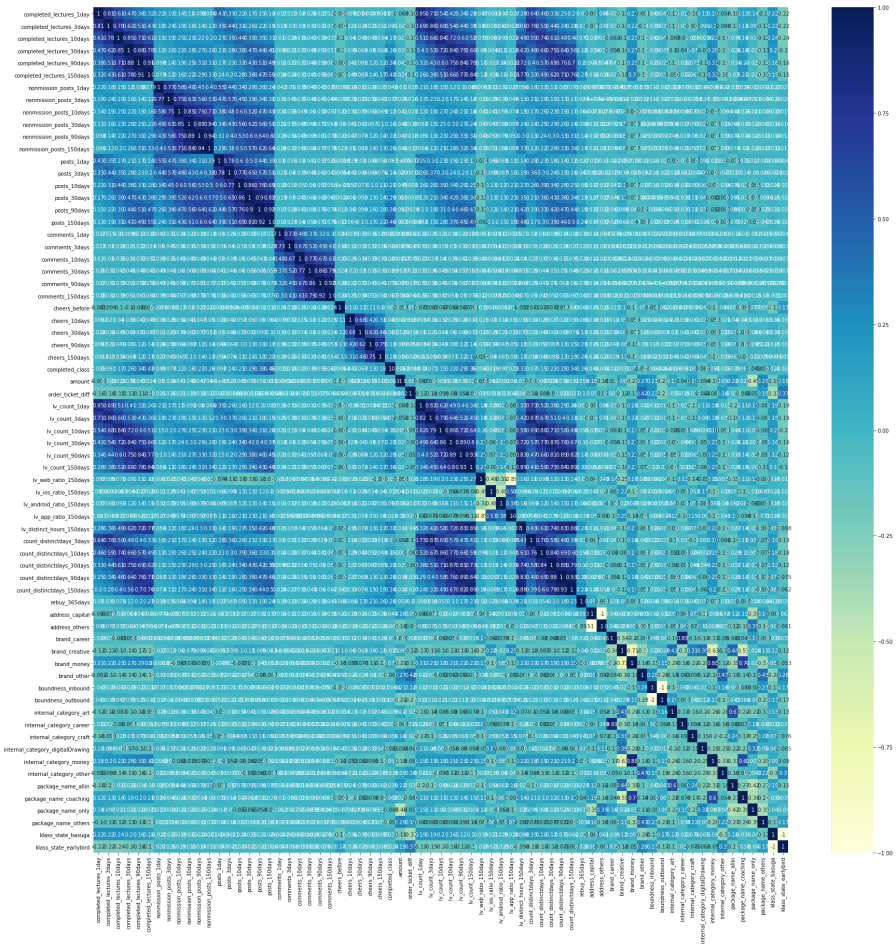
Simple Correlations

▼ Y와의 상관계수가 가장 높았던 X들 20개

```
completed_lectures_150days    26.76
completed_class               26.16
cheers_150days                26.08
posts_150days                 23.70
count_distinctdays_150days    23.39
lv_count_150days              22.51
cheers_90days                 21.39
lv_distinct_hours_150days     21.32
completed_lectures_90days     20.34
posts_90days                  18.96
count_distinctdays_90days     17.51
nonmission_posts_150days      17.44
lv_count_90days               17.03
nonmission_posts_90days       13.99
cheers_30days                 13.48
posts_30days                  12.93
completed_lectures_30days     12.46
count_distinctdays_30days     10.40
order_ticket_diff             10.39
lv_count_30days               10.24
Name: rebuy_365days, dtype: float64
```

- 현재 변수들의 개별 상관계수가 매우 높은 편이 아니여서 X들을 더 모아야함 (다른 친구들의 creativity/questions 필요, data processing/engineering 도움 필요)

▼ (참고) Correlation Matrix: 모든 변수들의 서로간의 상관관계 (feature가 넘 많아서 숫자가 안보임 ㅋㄷ;;)



Random Forest Model Scores

▼ Accuracy Scores

```
Train Score 99.94 %
Test Scores Mean: 76.54
Test Scores: [0.76606811 0.76024768 0.76975477]
Test Scores Standard Deviation: 0.003913703164490253
```

▼ Confusion Matrix

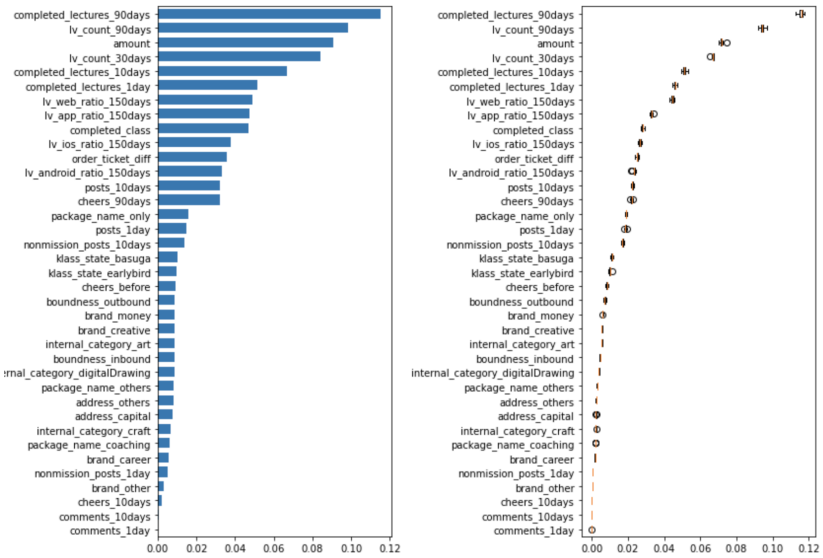
```
[[3579  365]
 [1071 1042]]
              precision    recall  f1-score   support

           0       0.77      0.91      0.83      3944
           1       0.74      0.49      0.59      2113

    accuracy                           0.76      6057
   macro avg       0.76      0.70      0.71      6057
weighted avg       0.76      0.76      0.75      6057

F1 Score 0.5920454545454544
```

▼ 개선의 여지 있음

- X row를 늘린다
- 좋은 X feature들을 더 찾는다
- 파생 X 생성
- F-1 높일 수 있는 방법 생각
- Outliers들 추가 제거
- User Segmentation

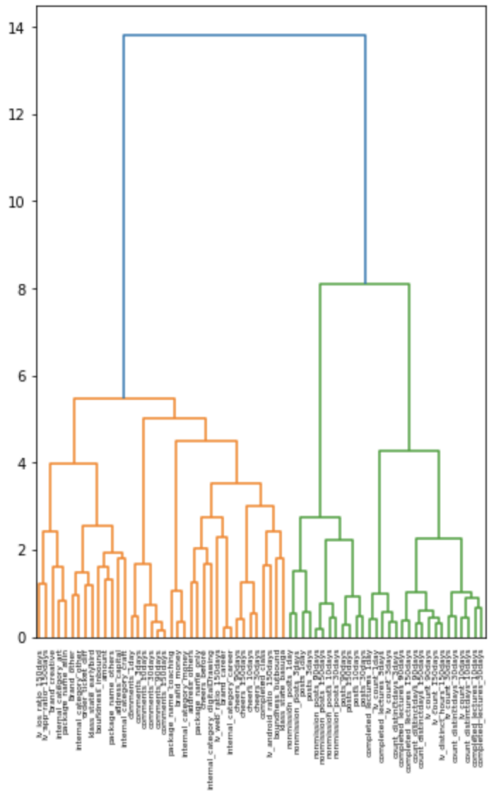
Random Forest Permutation Feature Importance Scores

▼ Permutation Feature Importance



▼ ^이게 무슨 뜻인가?

- Feature Importance를 통해서 Random Forest 예측 모델에 가장 크게 기여한 변수들이 무엇인가?를 알 수 있음
  - 하지만 X들이 Numerical 변수들일 경우 일반적인 Feature Importance를 사용하면 안되고 Permutation Feature Importance를 사용해야함
  - 이것을 진행하려면 X간의 상관관계를 최소화해서 변수들을 새로 선택해야함 (dendrogram 그려서 correlation hierarchy 파악 후, 이것을 통해서 변수들을 추려서 진행함, hierarchy level=1)



▼ 이 중에 t-test 패스한 친구들 (p-value < 0.05)을 다시 추리면:

1. completed_lectures_10days
2. amount
3. lv_web_ratio_150days
4. completed_class
5. order_ticket_diff
6. posts_10days
7. cheers_90days
8. package_name_only

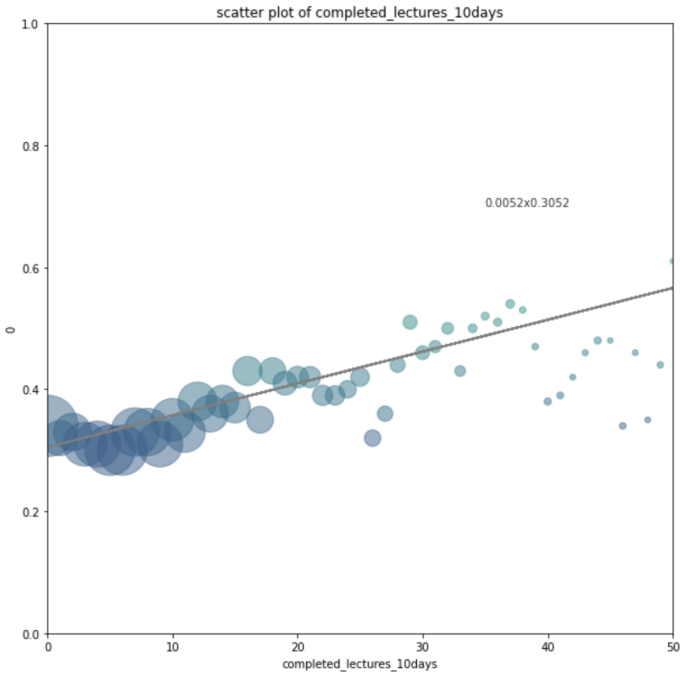
Scatter Plots

▼ "High Contribution Xs" vs Y

6a) Completed Lectures Within 10 days

•••

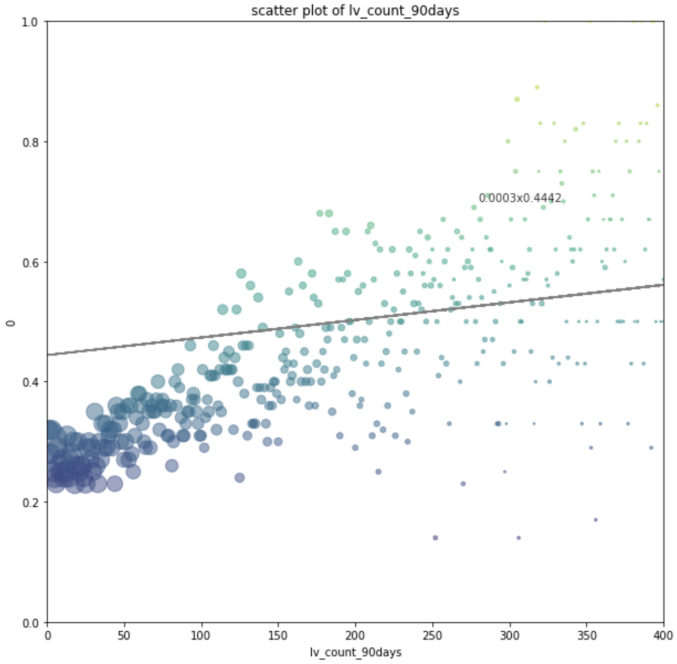
```
0.0052x0.3052
```



6b) Lecture View Count Within 90 days (p-value test didnt pass)

•••

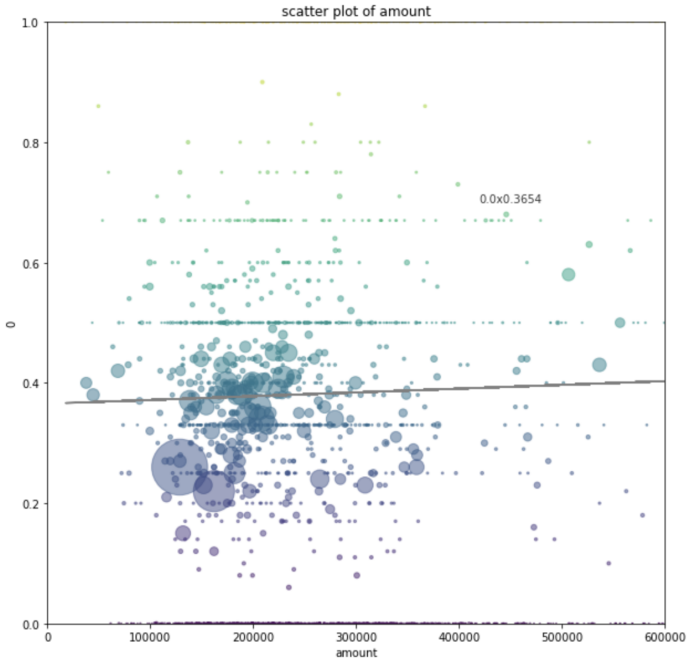
```
0.0003x0.4442
```



6c) Class Price

• • •

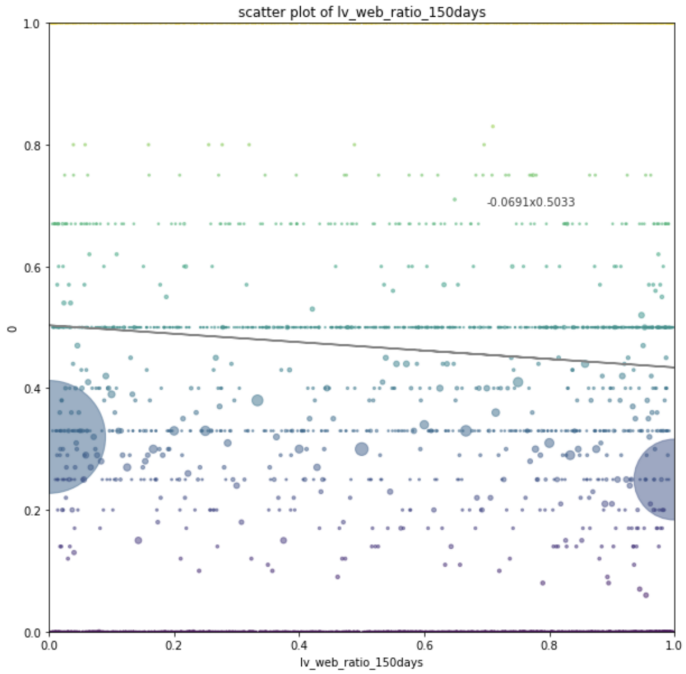
```
0.0x0.3654
```



6d) Web Lecture View Ratio Within 150 days

• • •

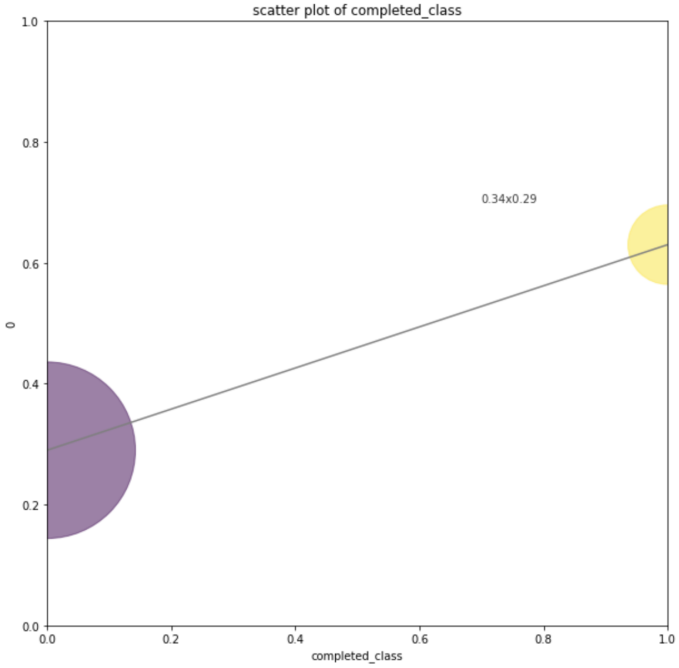
```
−0.0691x0.5033
```



6e) Completed Class or Not

•••

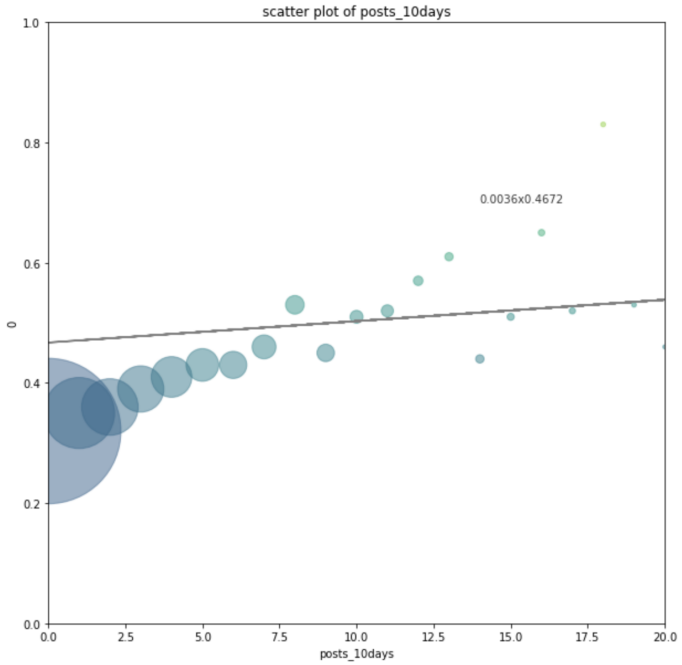
```
0.34x0.29
```



6f) Number of Posts Within 10 Days

•••

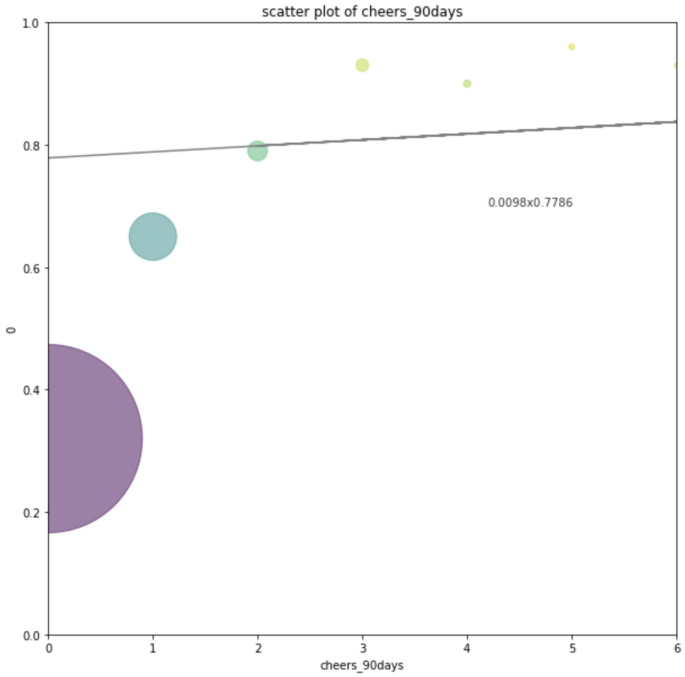
```
0.0036x0.4672
```



6g) Number of Cheers Within 90 Days

•••

```
0.0098x0.7786
```



▼ Other Xs vs Y (번외편)

- 카테고리는 유의미하다가 나오진 않았지만 일단 avg들을 보긴했음

6h) By Categories

	internal_category	sum_revenue_365days	count_users	avg
0	career	63771813	47	1356847.09
1	founded	1138648329	1348	844694.61
2	dataAndDevelopment	157549281	249	632728.04
3	music	1052003368	1811	580896.39
4	signature	49943681	99	504481.63
5	craft	1537729022	3199	480690.54
6	stock	252614133	588	429615.87
7	cooking	324293092	787	412062.38
8	lifestyle	219265836	557	393655.00
9	art	2034645991	5211	390452.12
10	digitalDrawing	1857098400	4771	389247.20
11	sns	579971950	1506	385107.54
12	careerVideoAndDesign	596776586	1628	366570.38
13	oa	162767905	459	354614.17
14	photograph	599027076	1786	335401.50
15	workout	172852196	528	327371.58
16	writeContent	59706674	203	294121.55
17	onlineShop	1494405852	5504	271512.69

- age/gender 데이터 있는 User Group 3000명을 토대로 다시 분석해봄

5. Future Action Items

- Score 개선 (위에 Random Forest Model Scores 섹션에 "개선의 여지 있음" 부분 참고)
- 더 많은 User Group Data Processing (With other Data Scientists', Engineers' help)