February 2017

MacKenzie Seale
CLASSPASS

Predicting Subscriber Conversion

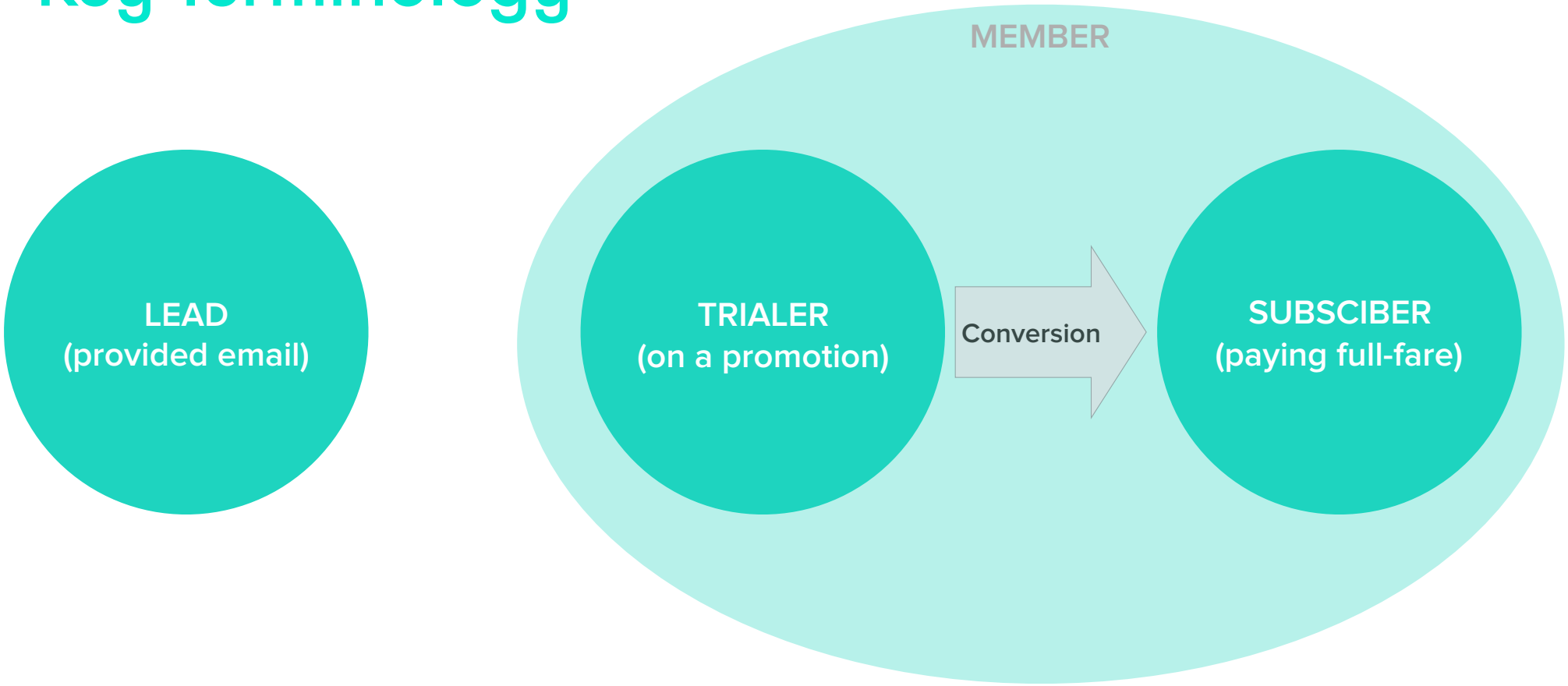
Agenda

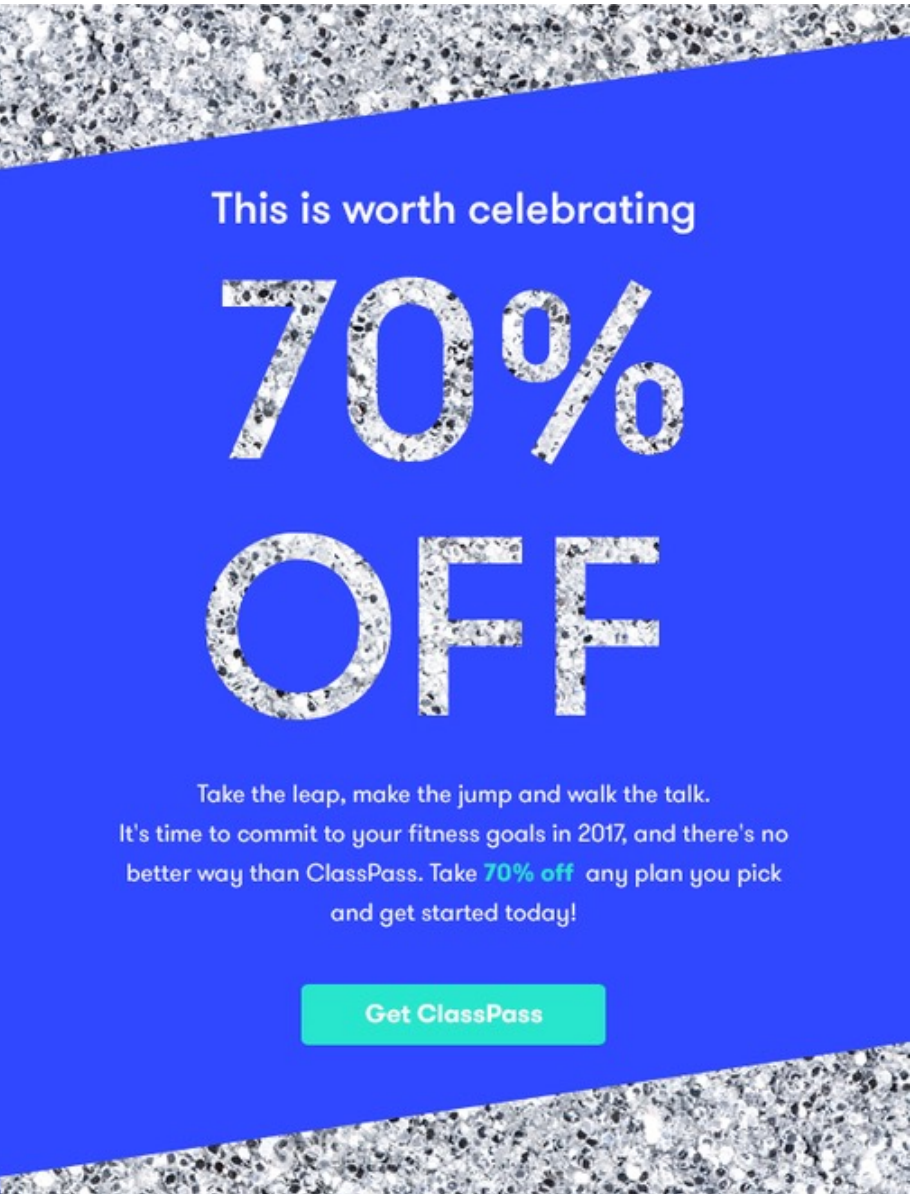
1. Background
2. Problem & Hypothesis
3. Dataset
4. Models
5. Conclusions & Next Steps

Background

ClassPass is a platform business

Key Terminology





Members fuel growth

Omni-channel multi-promotion marketing approach

Examples:

- 70% off email promotion to leads
- 3 month commitment advertised on Facebook
- Refer-a-friend banners on our homepage
- $19trial direct mail offer with Handy
- 50% off affiliate promotion with theSkimm

Problem & Hypothesis

PROBLEM STATEMENT:

Determine the association between the behavior of Trialers and the likelihood to convert to full-fare Subscribers.

HYPOTHESIS

Trialers that are more active on the platform are more likely to convert than those that do not use or rarely use the platform.

Let the sparring begin

- Using **supervised learning** for **classification** of binary outcome (conversion with yes = 1 and no = 0)

- Scoring models on
  - **Accuracy:** How often is it actually a conversion?
  - **Precision:** How similar are converters' behaviors?
  - **Recall:** Do we find everyone that will convert?
  - **F1 score:** How balanced are precision & recall?
  - **F1 score (on training data):** How balanced were precision & recall on the training data?

Dataset

Wrote customer query to select 42 variables

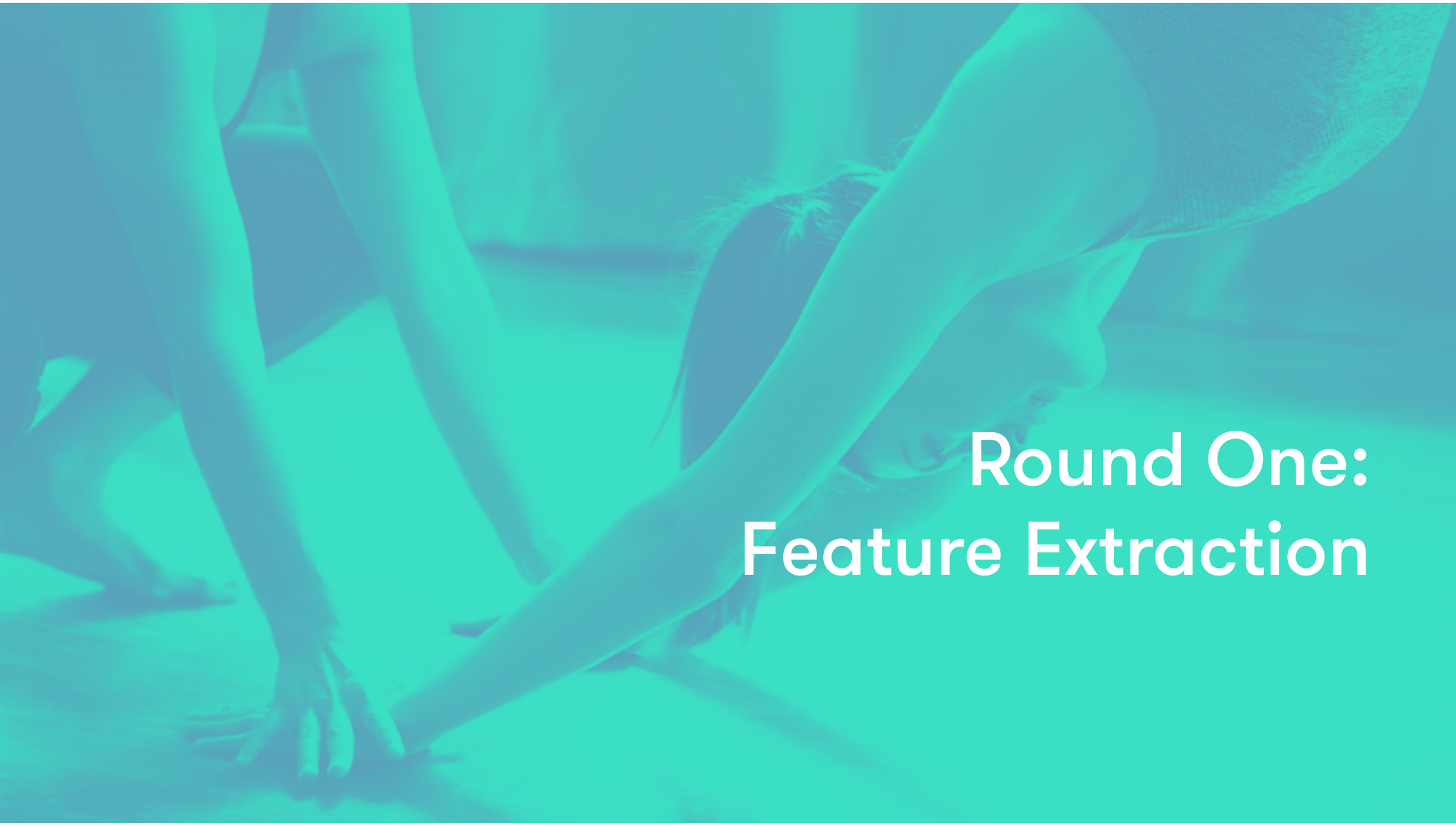
Selected the variables based off of known and predicted impact on subscriber conversion

- is_converted
- lead_created_month
- lead_created_day_of_week
- is_paidsocial_lead
- is_organic_lead
- is_seo_studios
- is_inviteafriend
- is_email
- is_referral
- is_paidsocial
- is_organic
- user_acquisition_month
- user_acquisition_day_of_week
- promo_days
- lead_to_promo_days
- is_three_studio_visits
- is_two_studio_visits
- is_four_studio_visits
- is_onboarded
- user_country
- user_msa_id
- avg_class_rating
- avg_days_booking_to_class
- avg_peak_classes
- reservations_barre_count
- reservations_boxing_count
- reservations_cycling_count
- reservations_dance_count
- reservations_gym_count
- reservations_martialarts_count
- reservations_pilates_count
- reservations_rowing_count
- reservations_strengthtrng_count
- reservations_yoga_count
- reservations_attended_t1_count
- reservations_attended_t2_count
- reservations_attended_t3_count
- reservations_attended_count
- reservations_missed_count
- reservations_late_cancel_count
- cost_of_all_reservations
- distinct venues count

Cleaning the data

- Almost 600k rows
- Performed initial cleaning while writing the query
  - Created booleans for categorical features
  - Grouped reservation counts by genre (barre, yoga, boxing, strength training etc)
  - Grouped reservation counts by tier (1, 2, 3)
  - Grouped reservation counts by type (attended, missed, late cancelled)
  - Aggregating data with averages
- Once loaded into Python continued cleaning
  - Replaced the NA's with 0's where appropriate

Models



Round One: Feature Extraction

Three classification models

	Decision Tree Classifier	Random Forest Classifier	Extra Trees Classifier
Accuracy	0.757	0.789	0.995
Precision	0.816	0.836	0.997
Recall	0.808	0.840	0.994
F1	0.812	0.838	0.995
F1 (train)	0.811	0.996	0.996
Inputs	41 features	41 features	41 features
Specifications	max depth = 5 min samples leaf = 10	trees = 30	none!
Most Important Features	1. Promo Days 2. User Acquisition Month 3. Distinct Venues Count 4. Is Onboarded? 5. User MSA ID	1. Promo Days 2. Cost of Reservations 3. User Acquisition Month 4. Distinct Venues Count 5. Res Attended Count	1. User Acquisition Month 2. Avg Peak Classes 3. Promo Days 4. Is Organic 5. Avg Class Rating

Reduced to 10 Features

1. User Acquisition Month
2. Avg Peak Classes (determined by peak time/day)
3. Promo Days (how long is the promotion)
4. Is Organic (a member we did not pay for)
5. Avg Class Rating
6. Cost of Reservations
7. Count of Distinct Venues
8. Is Onboarded (completed our sign-up flow)
9. Avg Days from Booking to Class
10. User MSA ID (city)

Round Two:
More Models Less Features

Cutting the variables by 75% had minimal impact on performance

Decision Tree Classifier	Score	New Score
Accuracy	0.757	0.753
Precision	0.816	0.824
Recall	0.808	0.790
F1	0.812	0.806
F1 (train)	0.811	0.806

Random Forest Classifier	Score	New Score
Accuracy	0.789	0.772
Precision	0.836	0.822
Recall	0.840	0.829
F1	0.838	0.826
F1 (train)	0.996	0.985

Extra Trees Classifier	Score	New Score
Accuracy	0.995	0.980
Precision	0.997	0.988
Recall	0.994	0.981
F1	0.995	0.985
F1 (train)	0.996	0.985

Logistic Regression

	Score
Accuracy	0.676
Precision	0.681
Recall	0.947
F1	0.792
F1 (train)	0.791

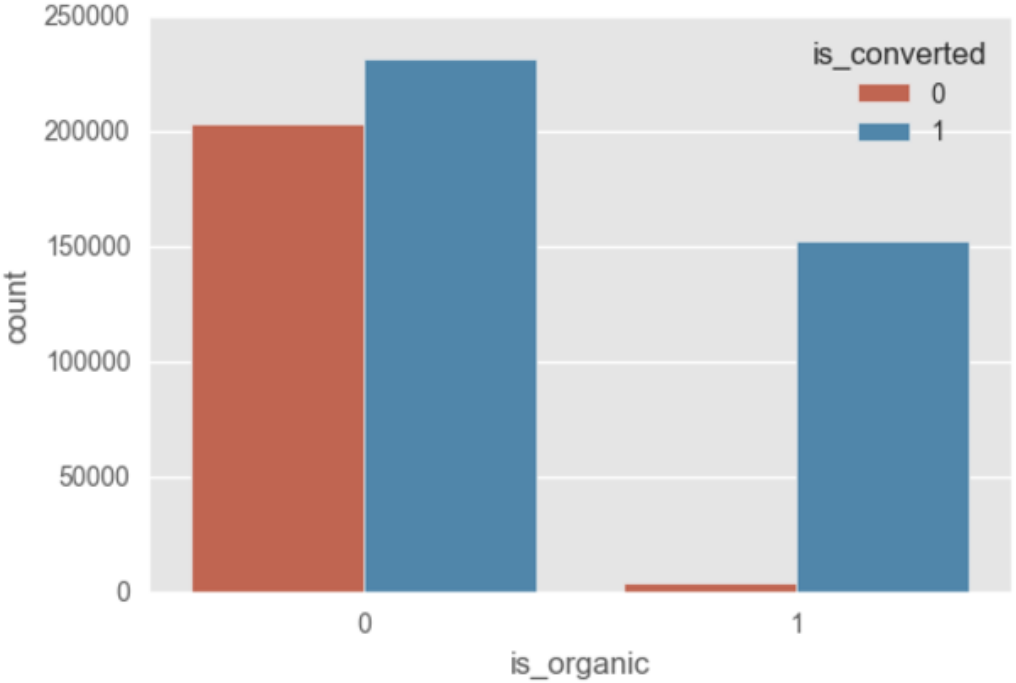
Feature	Coefficient
User Acquisition Month	0.365
Avg Peak Classes	-0.131
Promo Days	-0.191
Is Organic	1.294
Avg Class Rating	-0.305
Cost of Reservations	0.384
Avg Days Booking to Class	-0.116
Is Onboarded	0.355
Distinct Venues Count	0.264

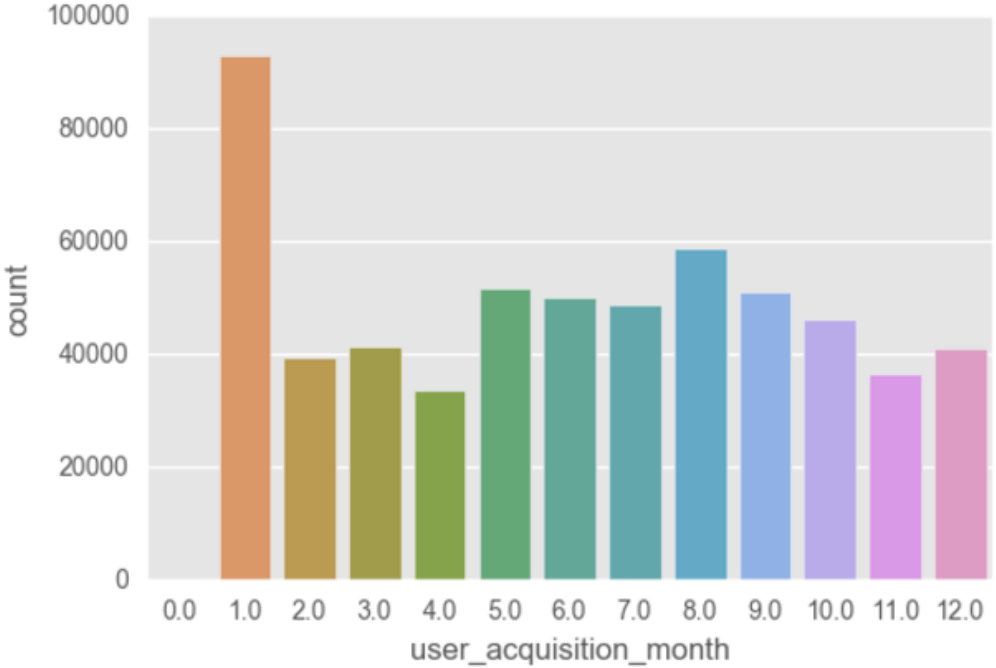
K Nearest Neighbors

	Score
Accuracy	0.653
Precision	0.656
Recall	0.981
F1	0.786
F1 (train)	0.785

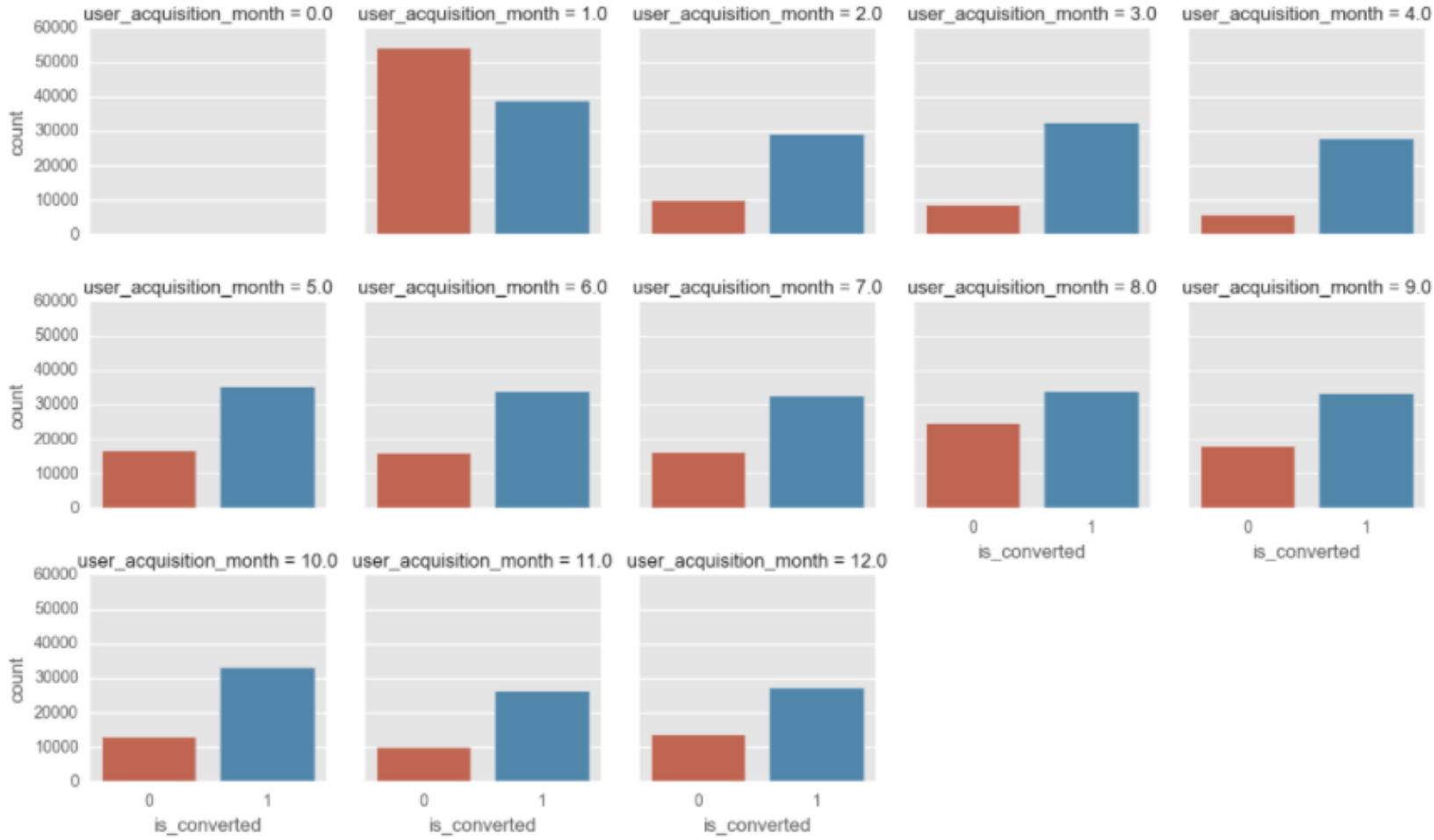
- Input 10 features
- 15 neighbors of uniform weight
- Discovered how expensive KNN is!

Studying the two most important features





New Year's Resolutions ≠ New Subscribers



Conclusions & Next Steps

Comparing the final models

- Decision Tree & Random Forest are better for **accuracy & precision**
- Logistic Regression & KNN are better for **recall**
- Extra Trees performed the best across all scoring methods

	Decision Tree	Random Forest	Extra Trees	Logistic Regression	K Nearest Neighbors
Accuracy	0.753	0.772	0.980	0.676	0.653
Precision	0.824	0.822	0.988	0.681	0.656
Recall	0.790	0.829	0.981	0.947	0.981
F1	0.806	0.826	0.985	0.792	0.786
F1 (train)	0.806	0.985	0.985	0.791	0.785

Feature Opportunity

Things to do

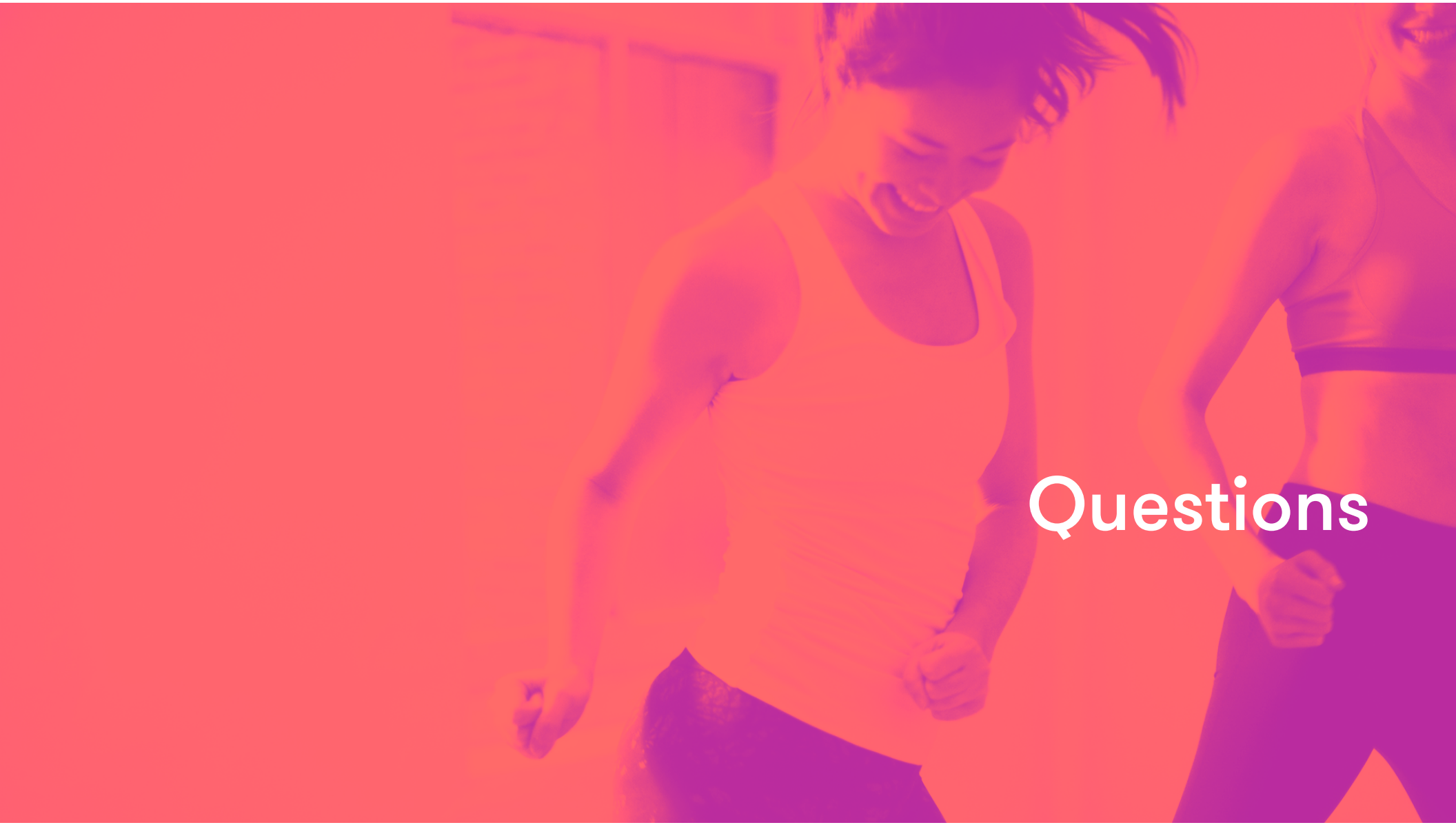
- Positive correlation with Avg Days Booking to Class suggests that we should encourage Trialers to book classes the day they signup
- Onboarding should be a requirement before booking first class
- Trialers should be encouraged to explore new studios as variety is an important product quality

Things not to do

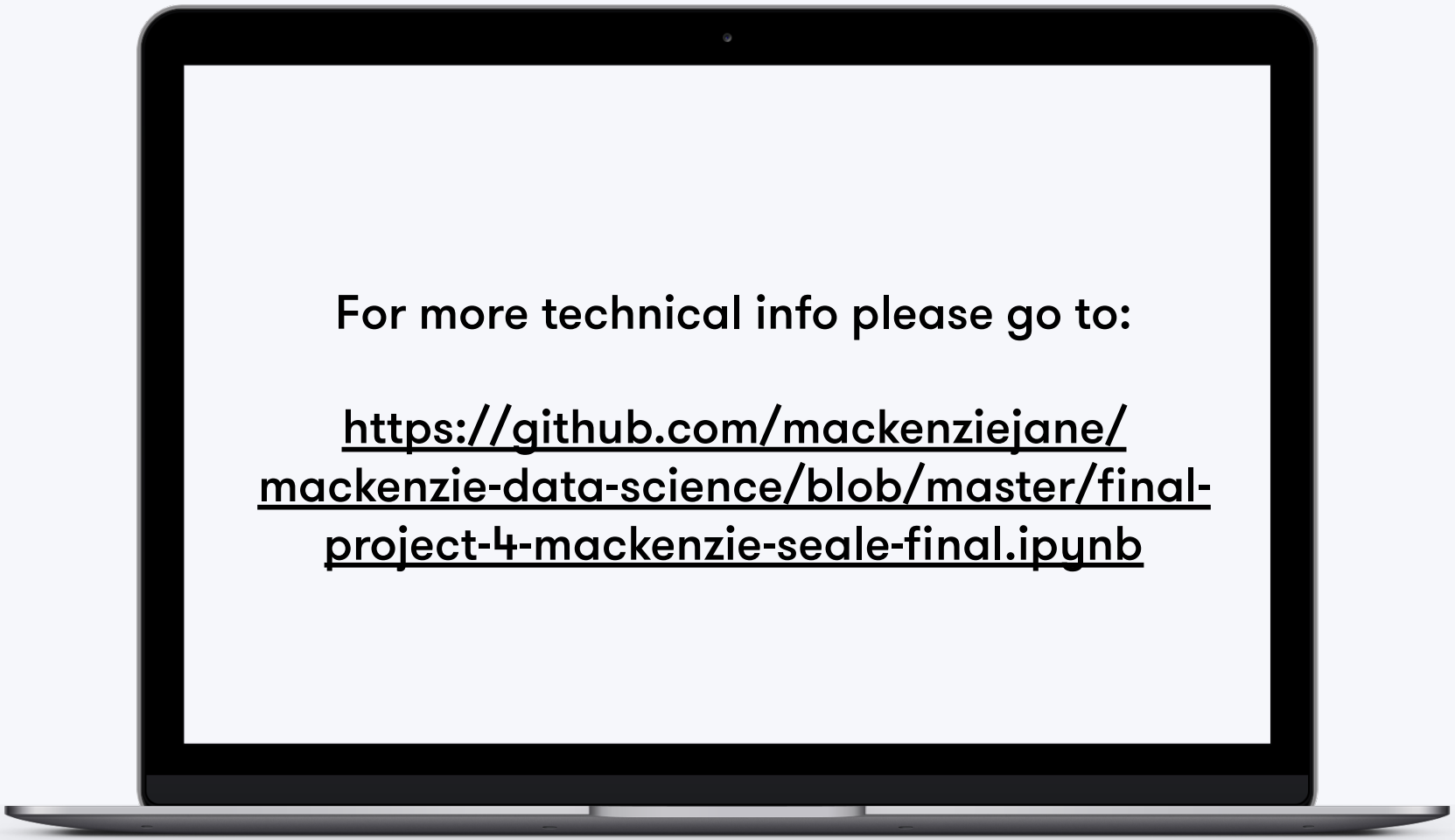
- Worry about getting Trialers peak class times
- Read into class ratings of Trialers

Next Steps

- Learn more about organic traffic
- Determine feature transformations needed to improve Accuracy & Precision of Logistic Regression & KNN
- Further examine Extra Trees for overfitting
- Reduce features to make models more robust


Questions

Appendix


For more technical info please go to:
https://github.com/mackenziejane/mackenzie-data-science/blob/master/final-project-4-mackenzie-seale-final.ipynb