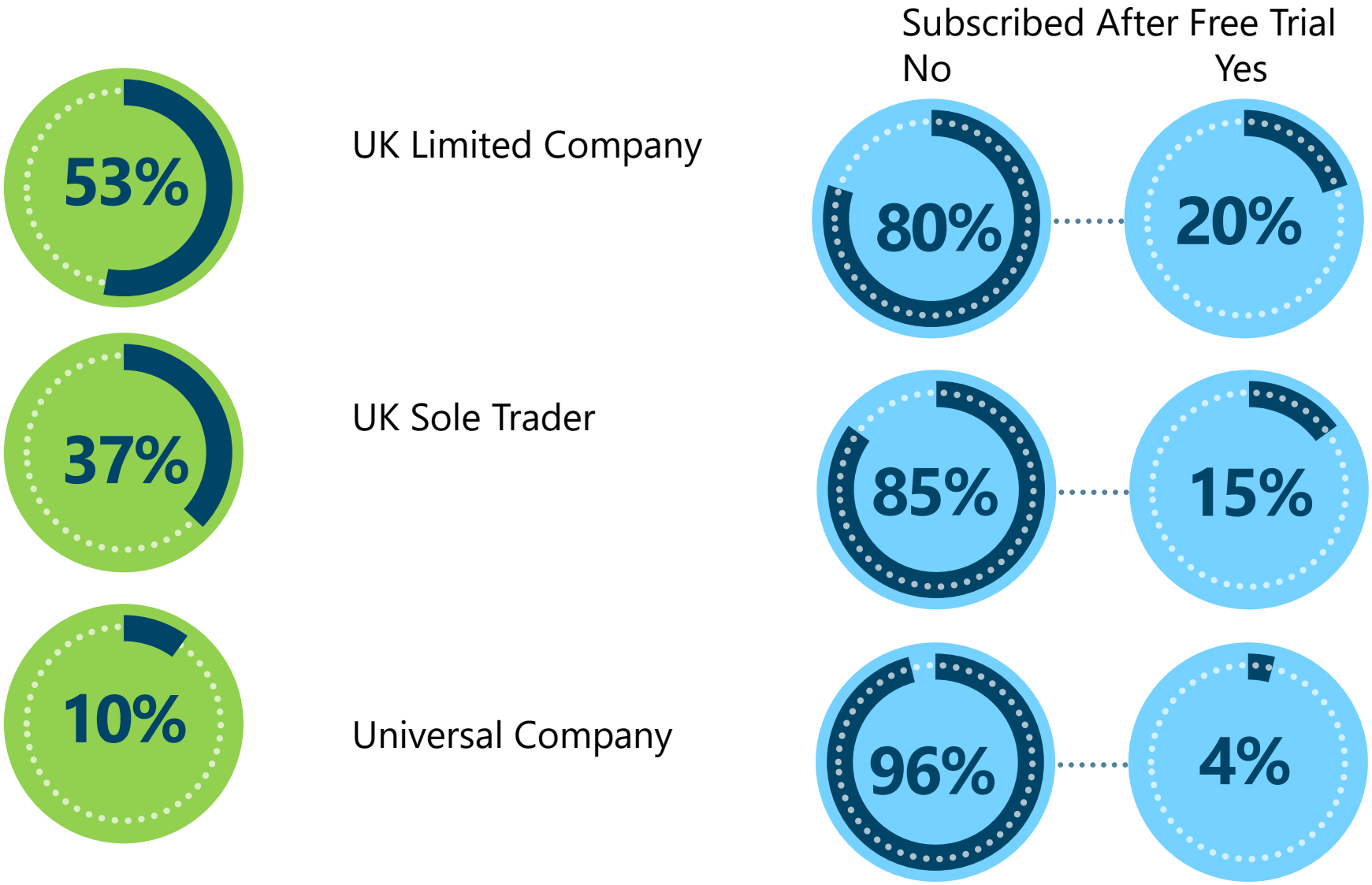
FreeAgent Data Analysis

Zheng Zhao

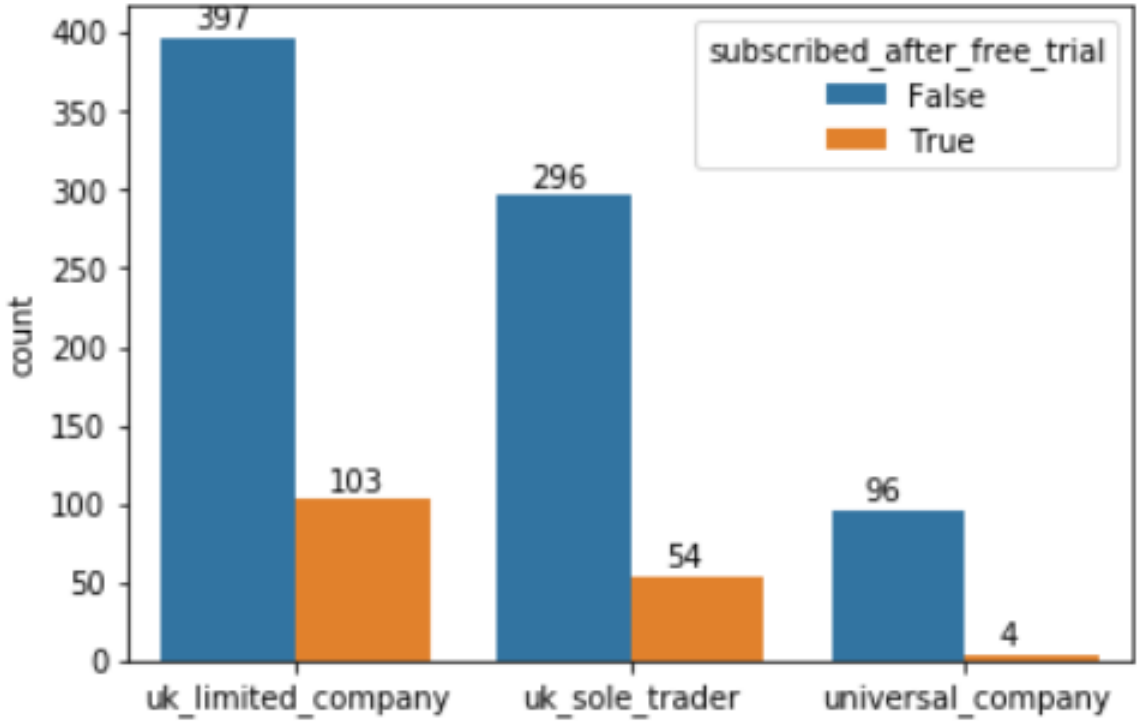
OVERALL STATS

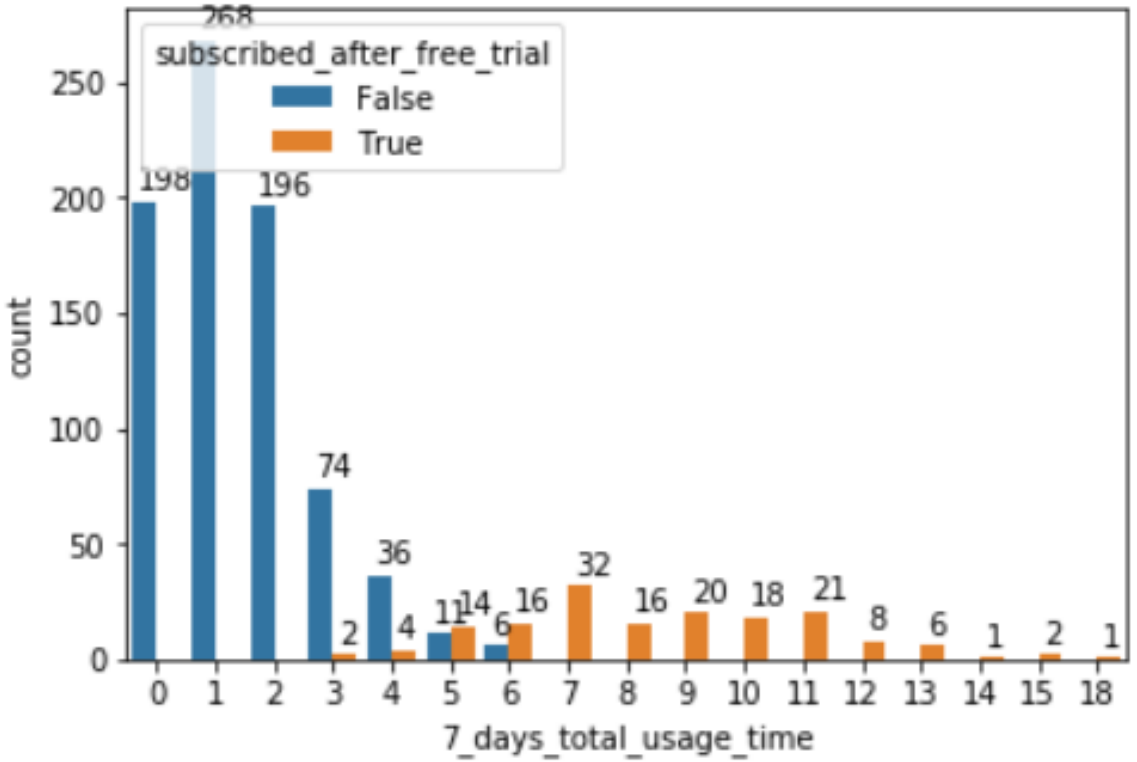
- There are 950 data entries in provided datasets.
- There are three type of companies:
- UK Limited Company, UK Sole Trader, Universal Company
- The total number of times each customer used the system in the first seven days of the free trial ranges from 0 to 18 times.
- Majority customers used the system only once or did not used at all during the seven day period.

DATA BREAKDOWN



GRAPHS


subscribed_after_free_trial
False
True
count
400
350
300
250
200
150
100
50
0
397
103
296
54
96
4
uk_limited_company
uk_sole_trader
universal_company


subscribed_after_free_trial
False
True
count
250
200
150
100
50
0
198
268
196
74
36
2
4
11
14
6
16
32
16
20
18
21
8
6
1
2
1
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 18
7_days_total_usage_time

ANALYSING PROCESS

- First I tried to visualize the data to see how the data is distributed in terms of company type and subscription after free trial
- Then I wrote a python script that processed *engagement_report.log* and stored data and combined with *attributes_report.csv* to get activities of all customers and for training the prediction model in next stage.
- I first get each customer's system usage on each of the first seven days of the free trial.
- Then I added the sum of 7 day's total usage for each customer as an attribute to the data because that's more meaningful and better for visualization purposes.
- After put all data together, I visualized the relationship between total_system_usage_time for each customer and status of subscription to get answer for first question.

QUESTION 1

Examine the provided data. Are there any differences in properties of behaviour between those customers that subscribe and those that do not?

- In general, UK limited companies and UK sole traders tends to subscribe more than Universal companies does.
- Customers that subscribed used the system a lot more frequent during the first 7 days of the free trial than those who did not subscribe.

MODEL BUILDING

- All the data now is in *attributes_report.csv* , I used it as training dataset for the prediction model
- For baseline model, I used a dummy model that will just predict a customer will not subscribe after the trial given this has higher probability in the training dataset.
- Final model uses SVC with kernel being Radial Basis Function
- Given this is a relatively small dataset, this model's performance should not be treated as golden.
- In the future, a better model can be trained with larger amount of data.
- Outlier detection and removal could improve future performance as well.

MODEL SELECTION

	classification accuracy
• Baseline model	0.8305
• Naive Bayes model	0.9474
• Logistic Regression model	0.9758
• LR with regularization parameter C=0.152	0.9779
• Linear SVC model	0.9758
• Radial Basis Function SVC	0.9821
• Polynomial SVC	0.9811
• Random Forest model	0.8595