Mobile Operator Users Churn Data Analysis

Alvydas Vitkauskas

Initial Data Overview

- We were given usage and churn data for 66,469 active users:

Month	Number of users
June	57,656
July	58,218
August	66,469

- Judging by the user_intake tag, the number of new users that were still active in the month of August was:

New in month	Still active in August
June	1,150
July	1,384
August	9,379

- All 562 additional users in July and 8,251 users in August were considered being new users for that months and are included in the new user numbers above, even if for some of them user_intake tag was not set.

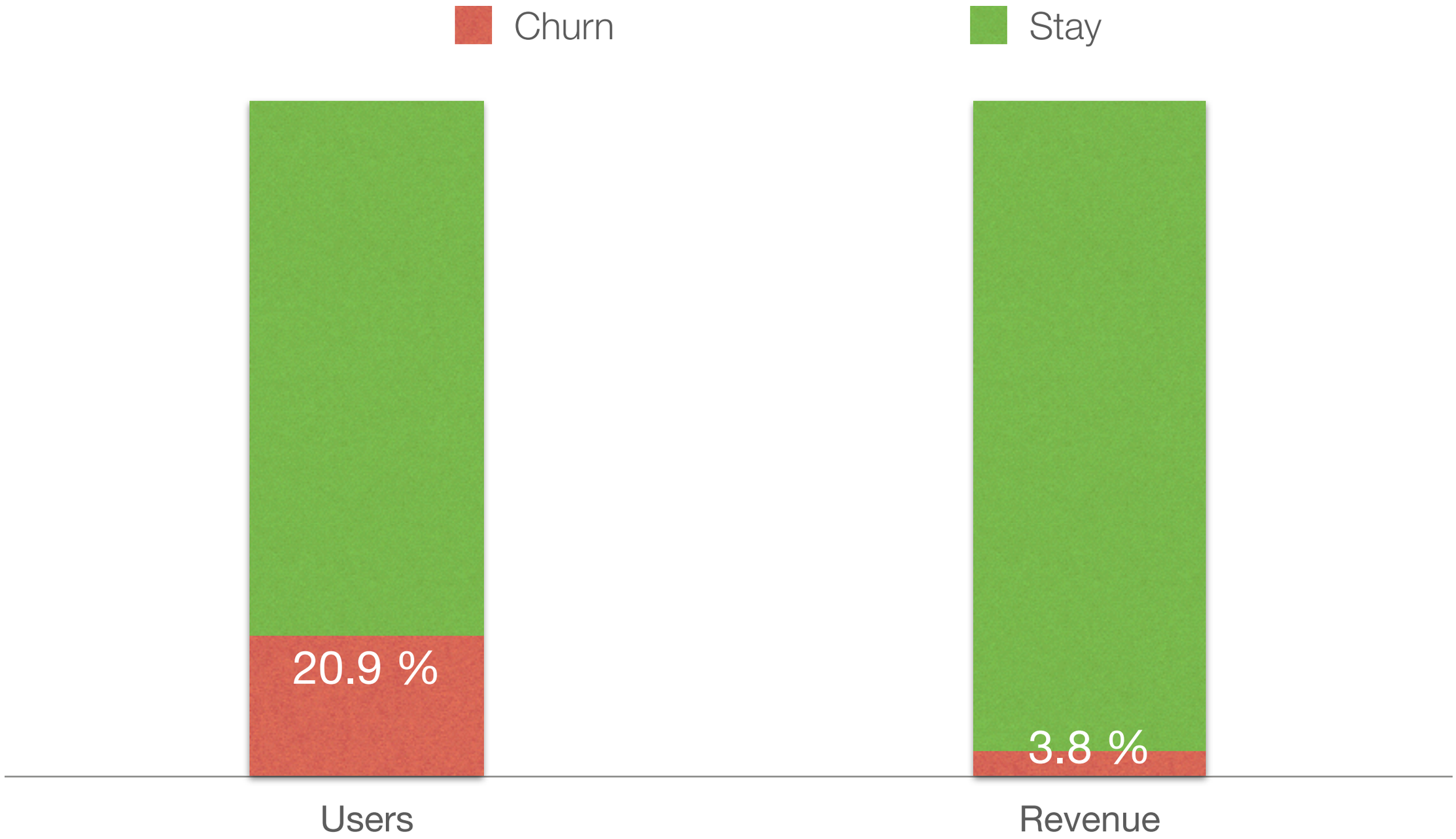
Data Cleaning

- The following data cleaning operations were performed:
    - all 562 additional users in July and 8,251 users in August were considered being new users for that months and their user_intake tag was set to 1
    - if user_intake = 1 & user_lifetime > 31,
      user_lifetime was set to the median lifetime of correct new users
    - if user_intake = 0 & user_lifetime > 15,000,
      user_lifetime was set to the maximum lifetime of correct old users
    - all the month-to-month sequences of user_lifetime were updated to follow the +31 days pattern
    - all inactivity counters for June that were bigger than the lifetime of the user were adjusted and were set to the value of the lifetime
    - month-to-month inactivity change cannot be bigger than 31 days and therefore were adjusted accordingly

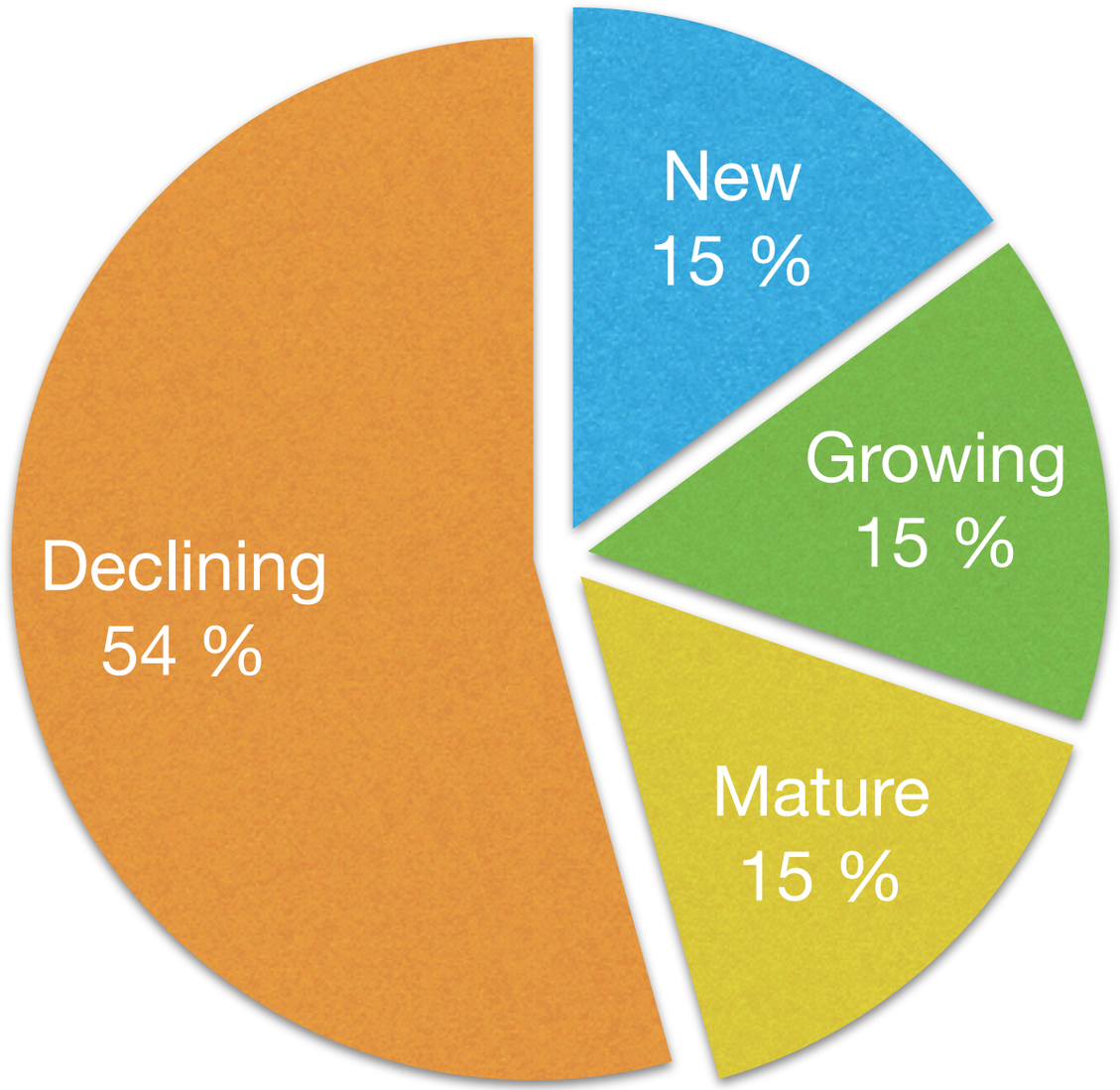
Overall User and Revenue Churn Rate

- Overall user churn rate in September was 20.9 %.
- Though overall revenue churn was only 3.8 %.
- 16,6 % of users who started using services in June and 17,6 % of users who started using services in July were churning in September.
- But September churn rate for users who started in August was very high at 67,1 %.
- As we are given data about the active users only, we could suspect that similar high churn rate of new users was possible in July and August, and that additional 16-17 % of new users are churning on the second and third month of their lifecycle.

Overall Users and Revenue Churn Rate

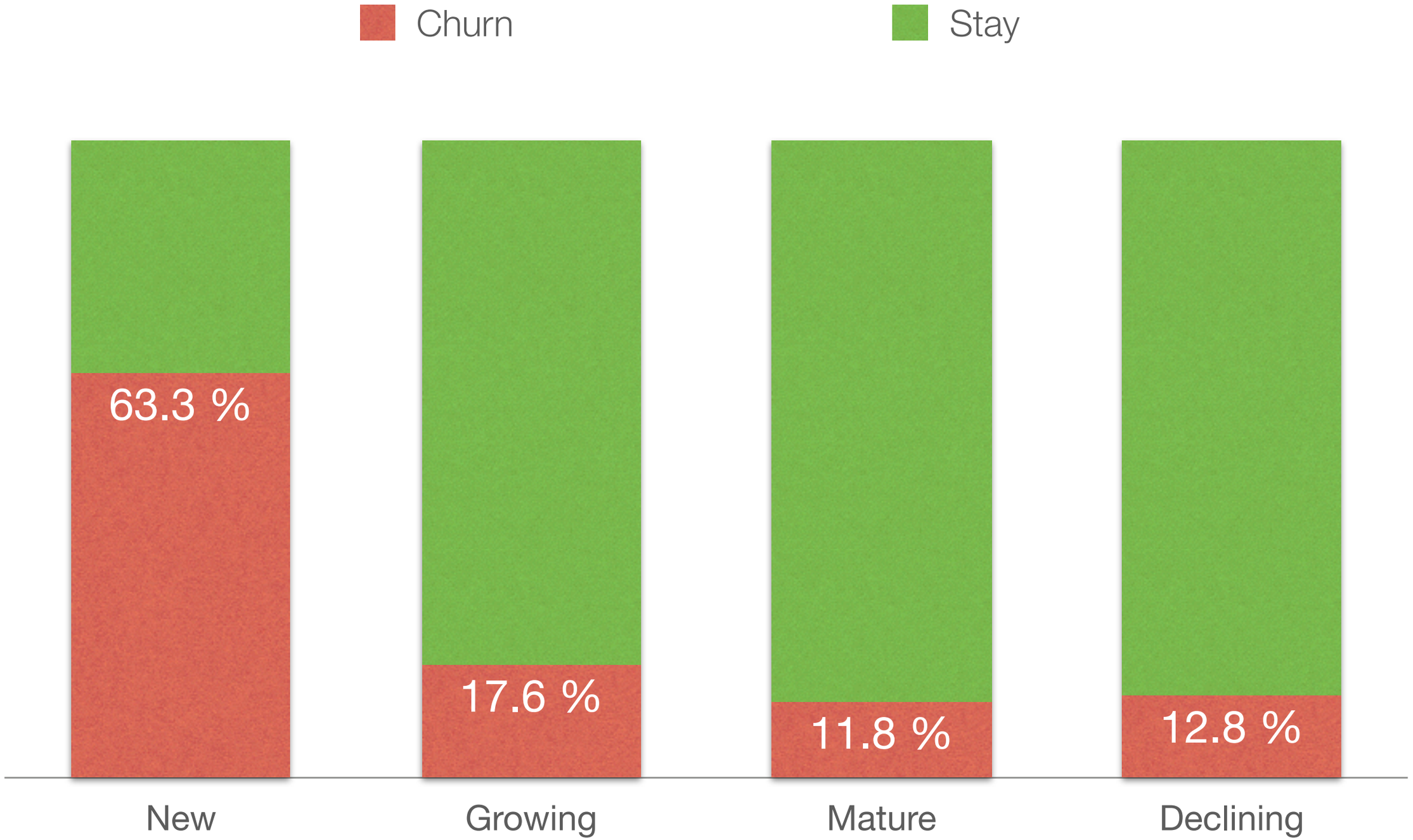


User Base Segmentation by Lifecycle

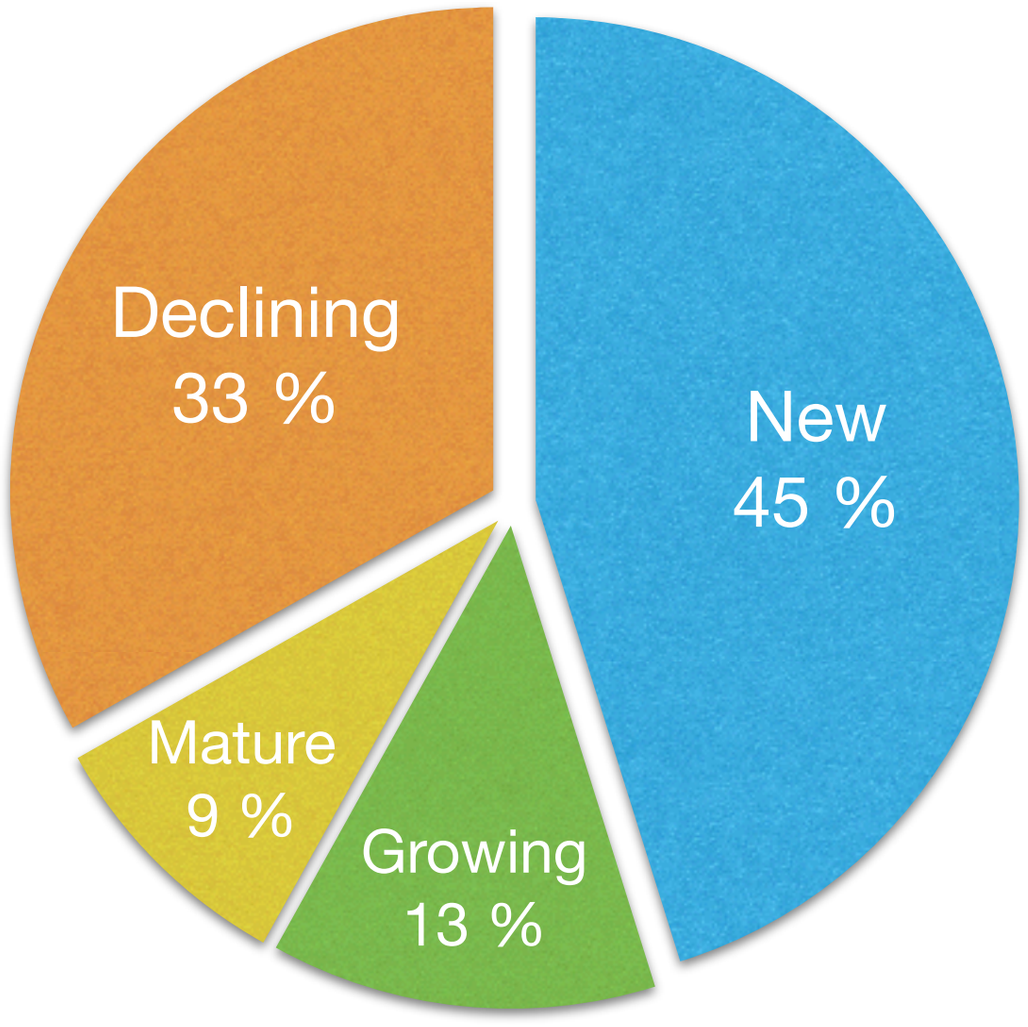


New (0-3 months), Growing (3-12 months), Mature (12-24 months), Declining (24+ months)

User Churn Rate by Lifecycle Segment

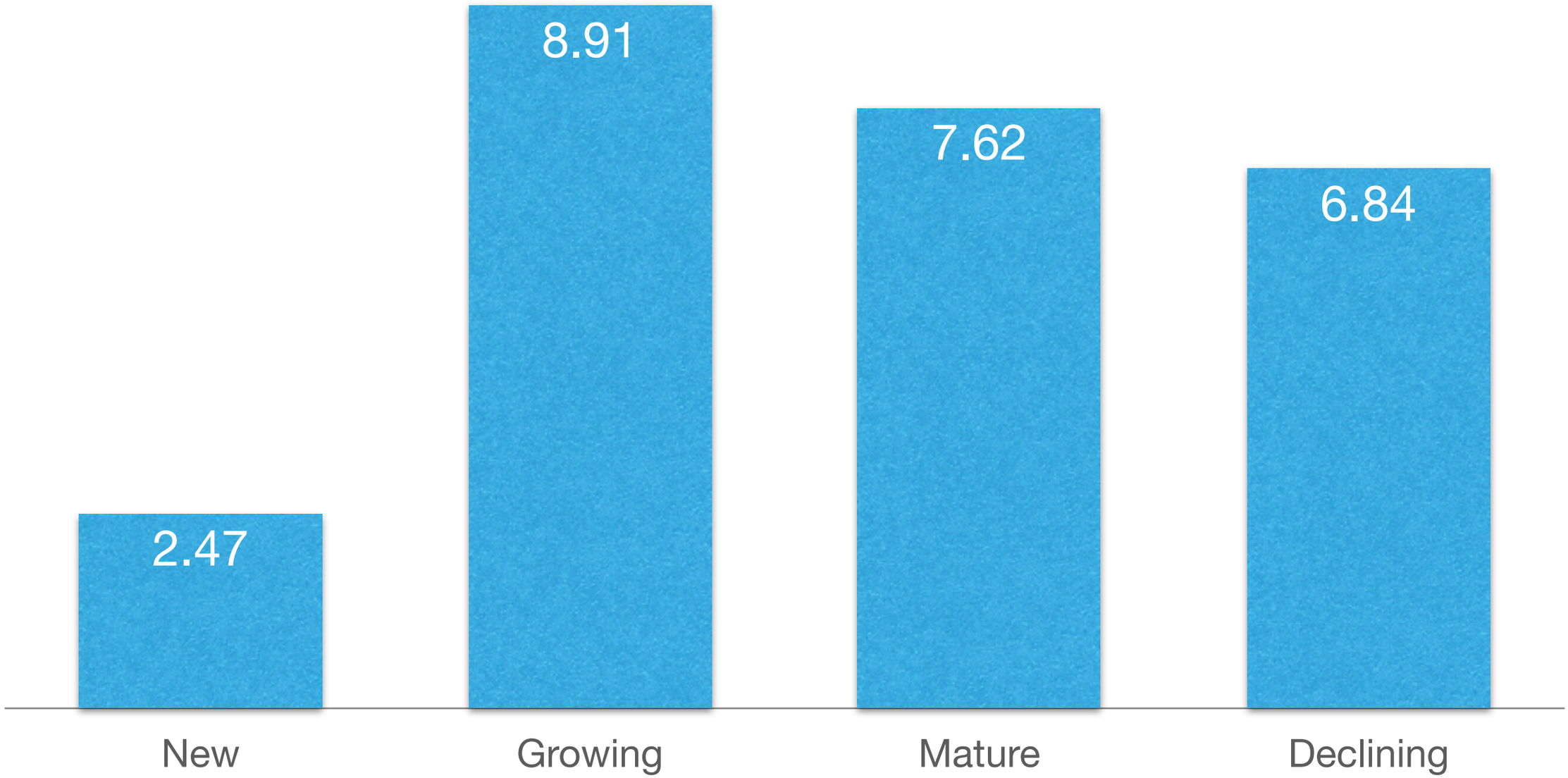


Composition of Churning Users by Lifecycle

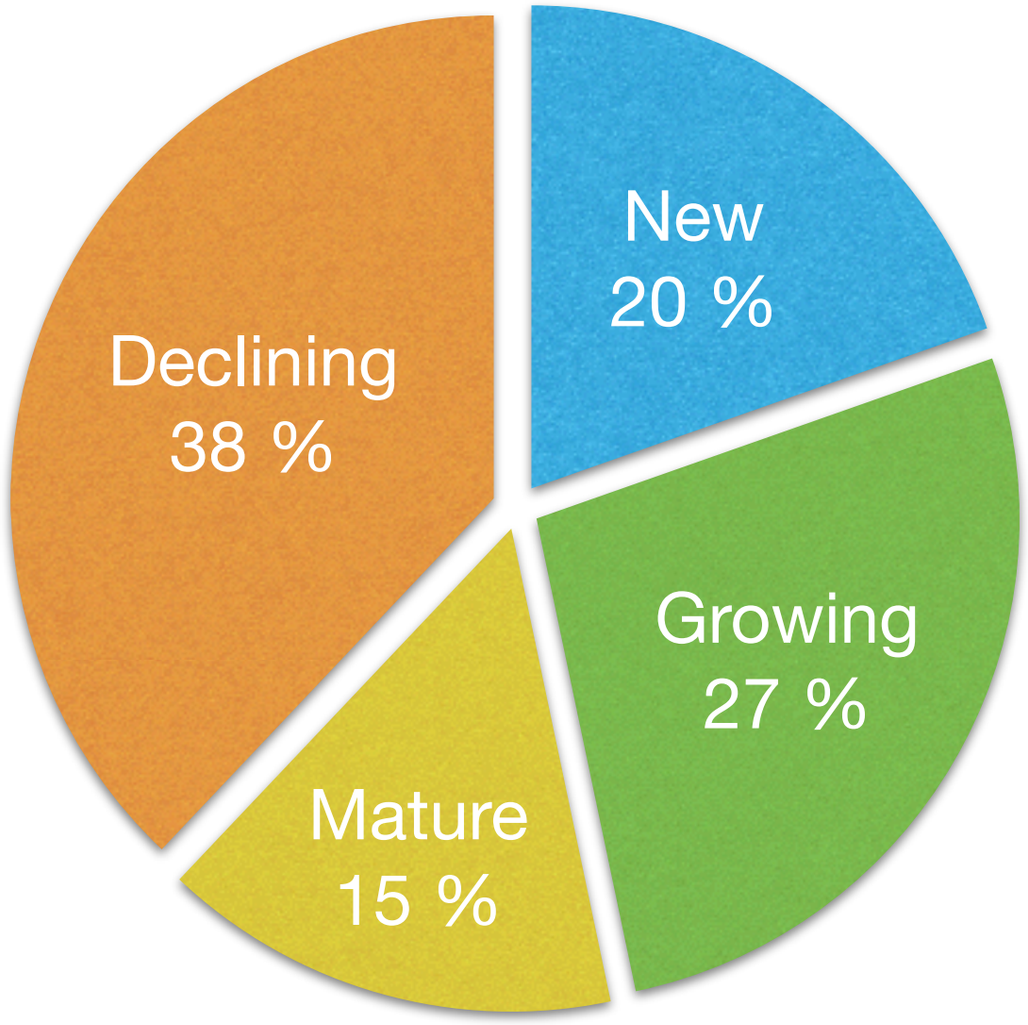


New users constitute the biggest part (45%) of all churners because of their high propensity to churn, and Declining users also make a big part (33%) because more than half of all users belong to that category.

ARPU by Lifecycle Segment

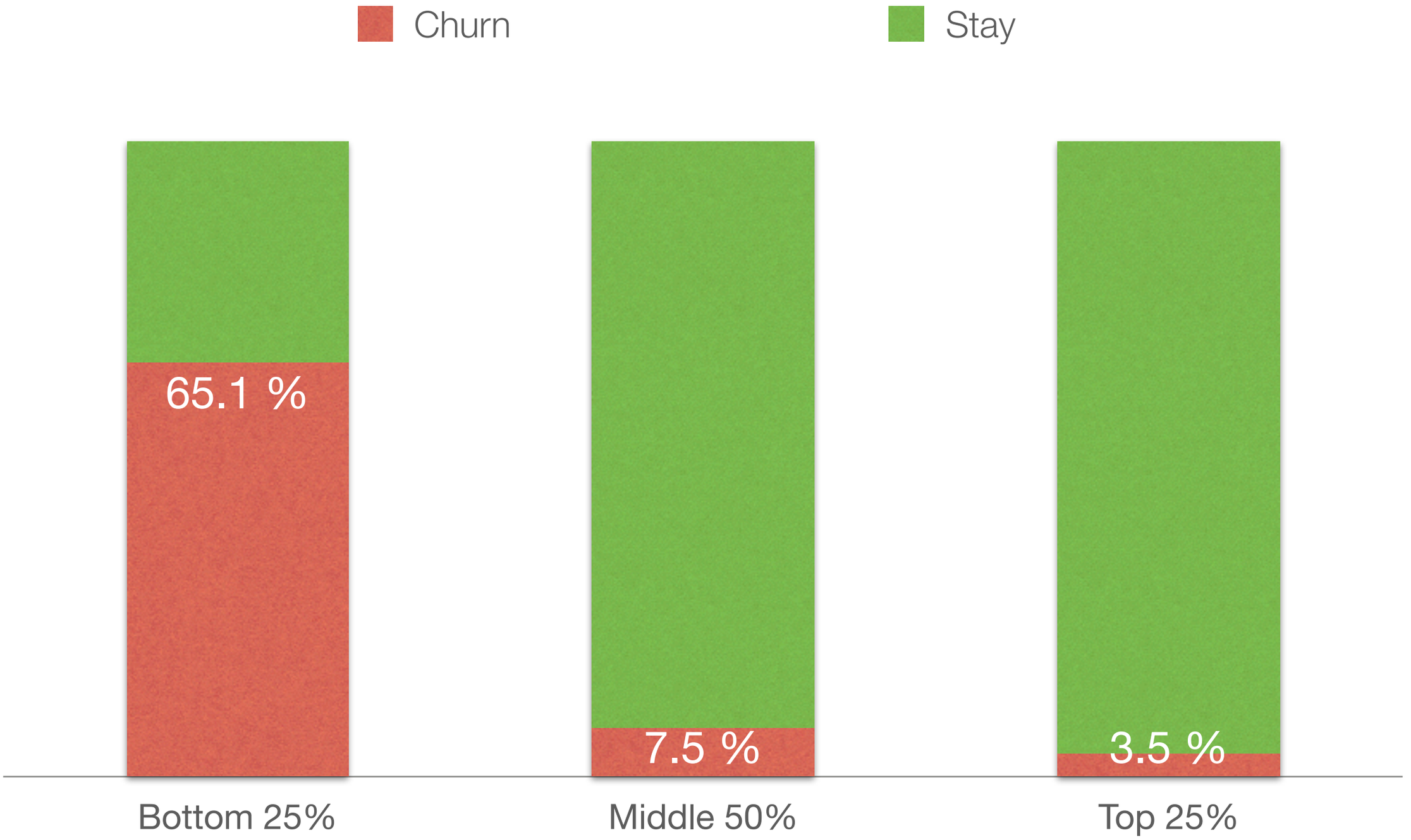


Composition of Churning Revenue by Lifecycle

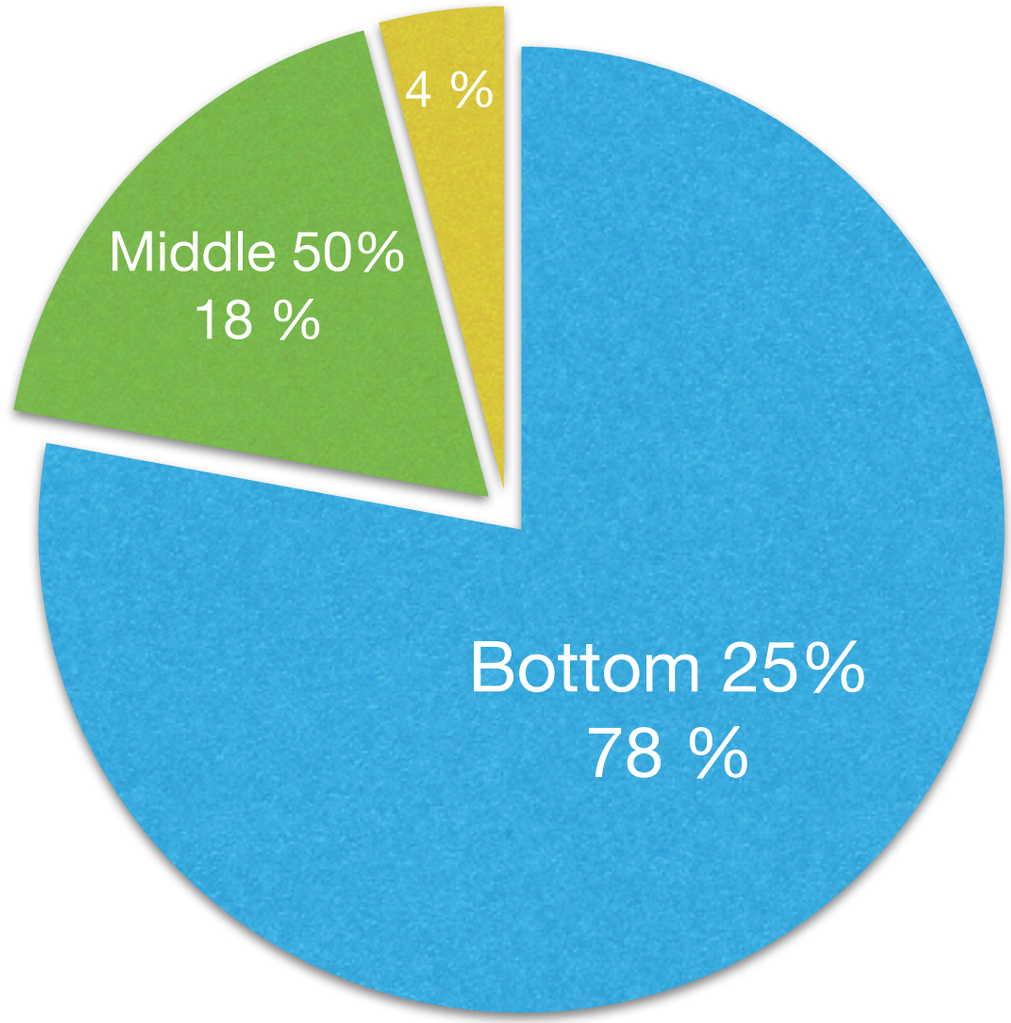


Even when New users have the biggest propensity to churn, still the biggest part of churning revenue comes from leaving Declining users because of their bigger ARPU and big overall proportion in the user base.

User Churn Rate by Customer Value

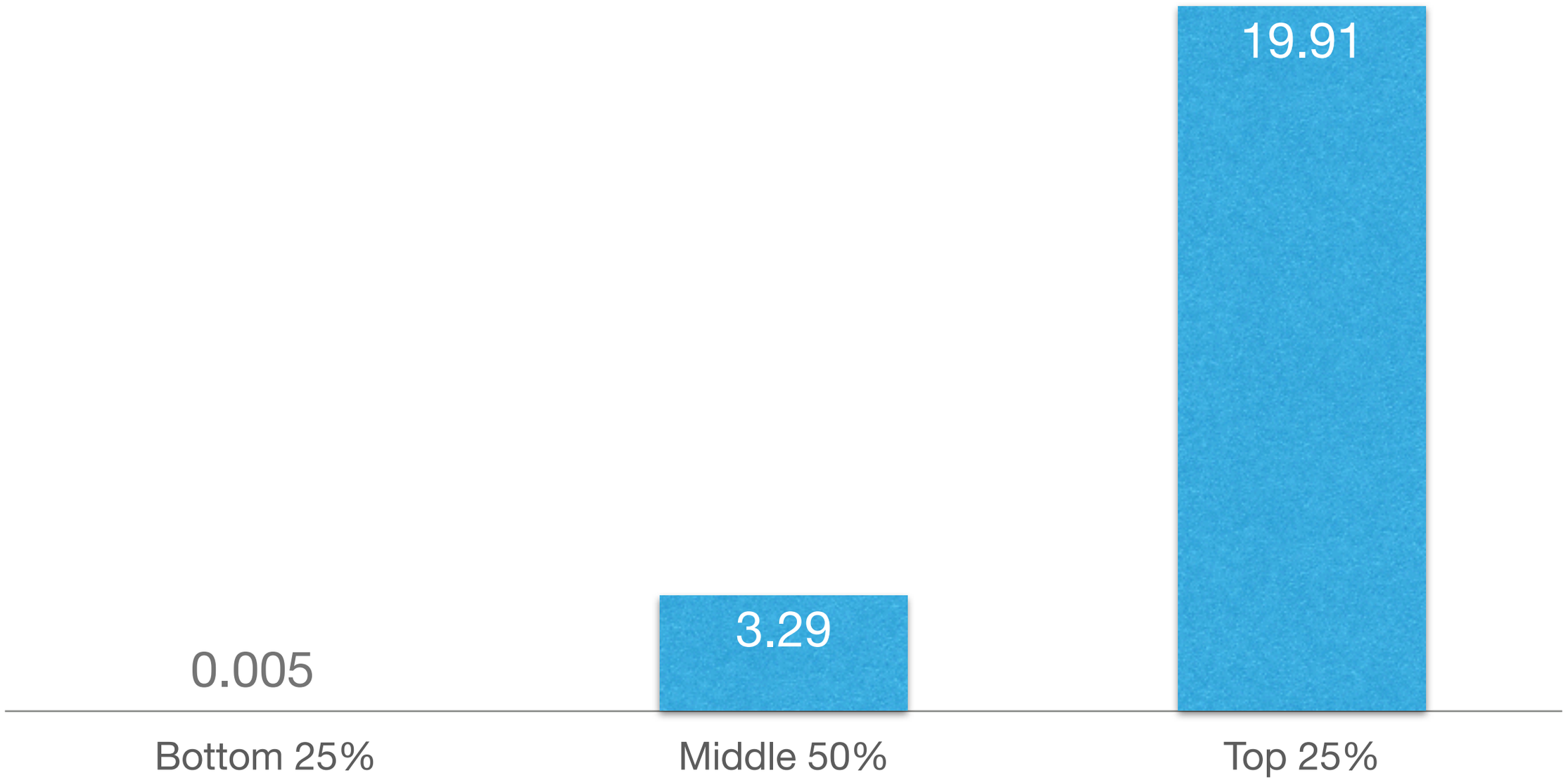


Composition of Churning Users by Value Groups

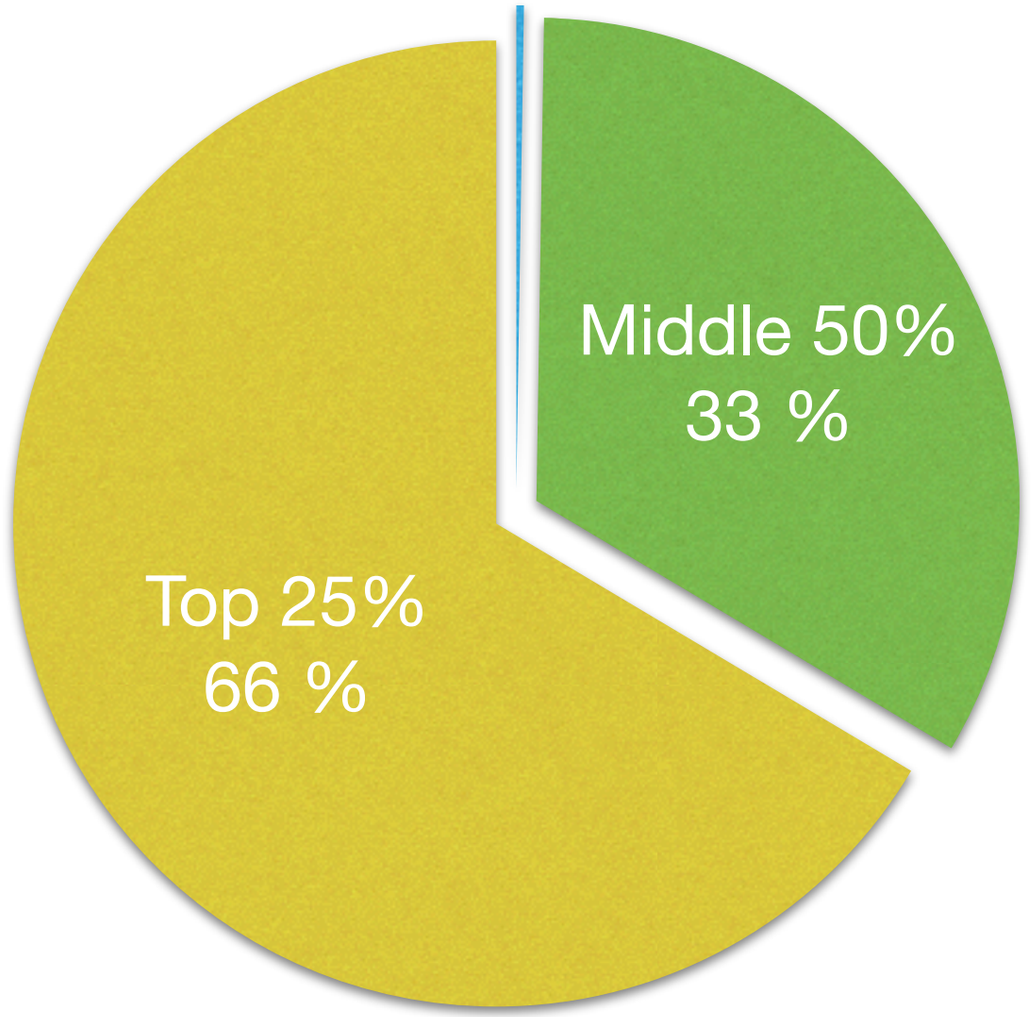


Over 3/4 of the churners are from the Bottom 25% value group.

ARPU by Customer Value Group



Composition of Churning Revenue by Value Groups



But 99% of revenue churn comes from Top 25% and Middle 50% users.

Zero ARPU Users

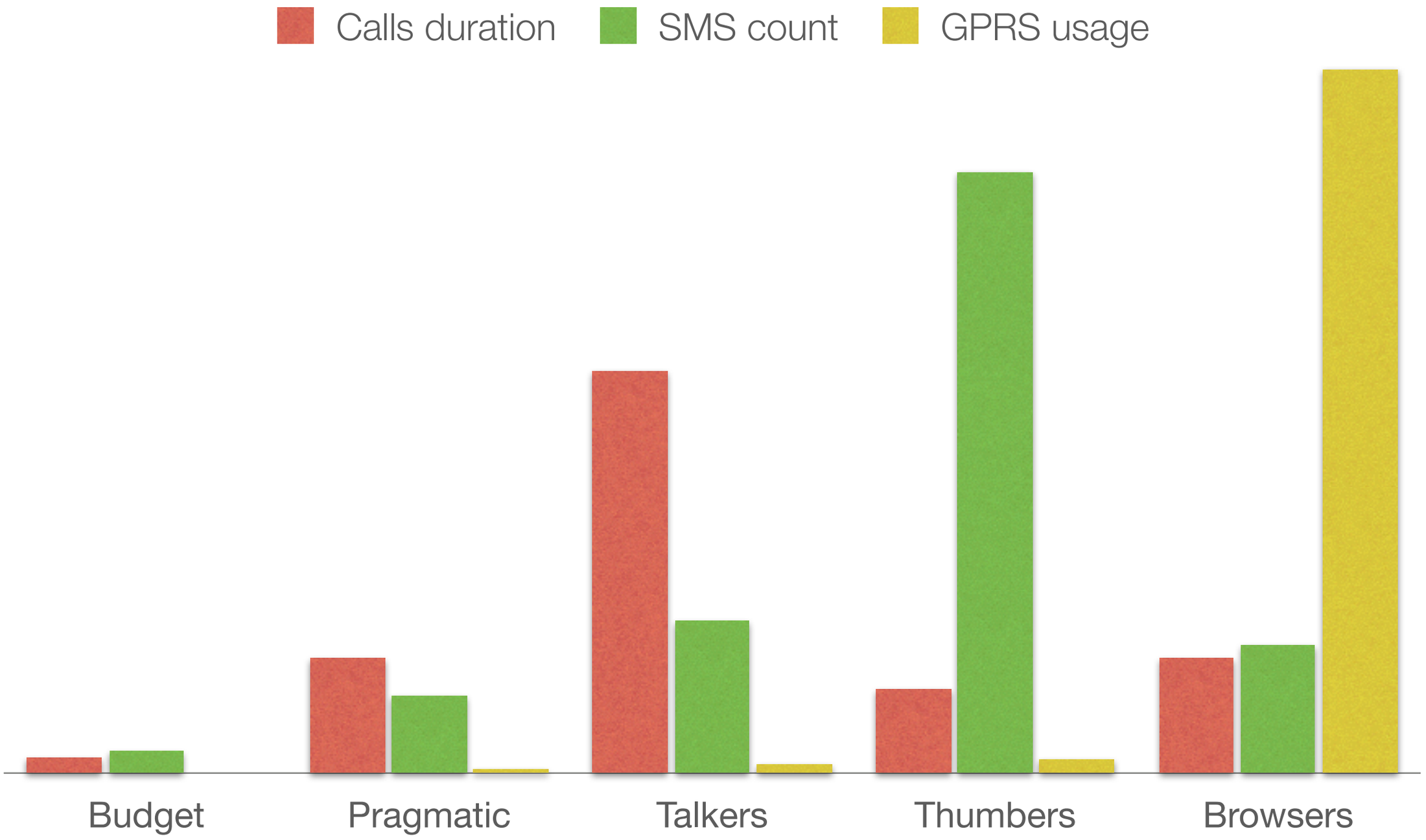
- 14,946 users (22.5 % of total user base) did not spend anything during these three months

- 50 % of them are New users and 30 % are Declining

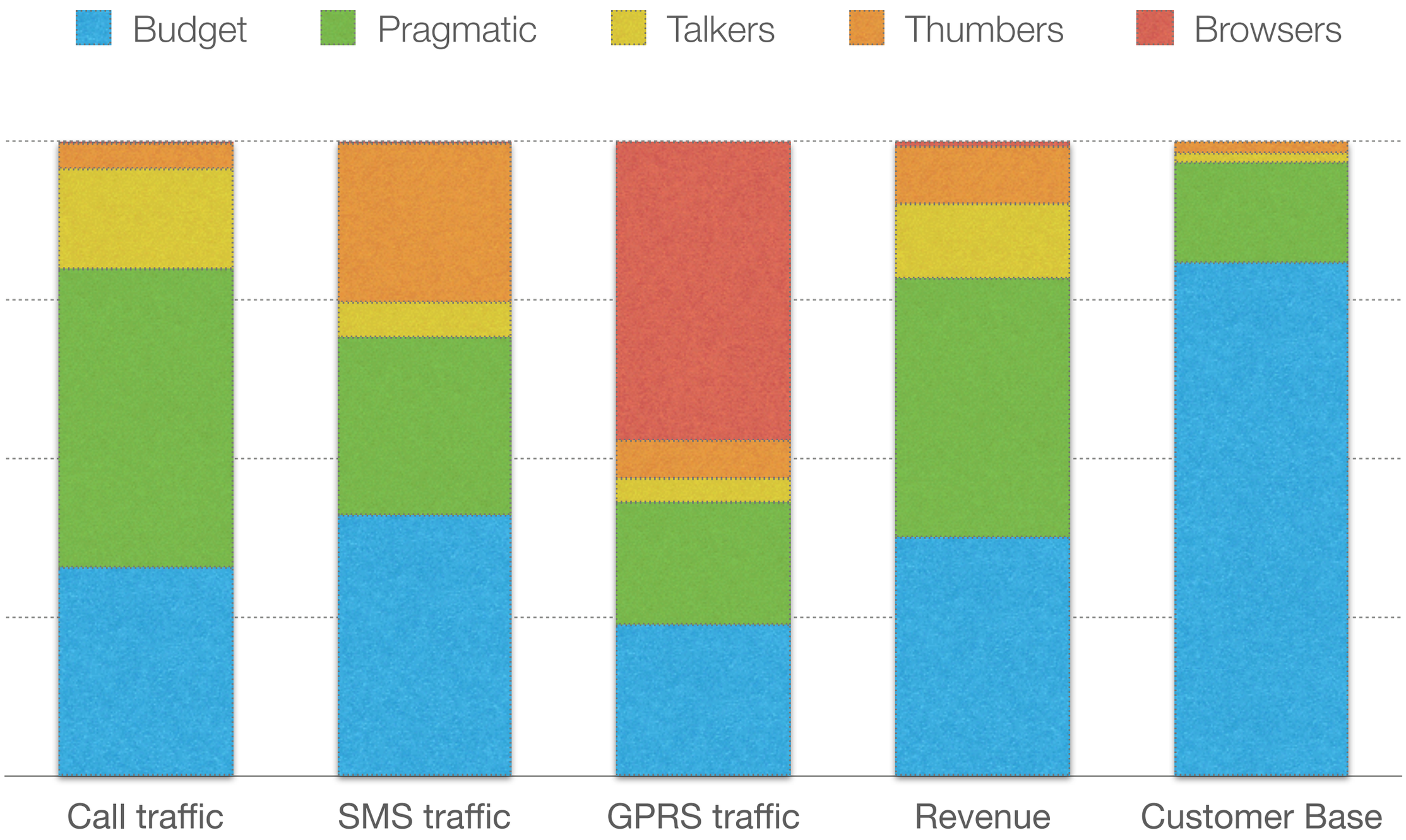
- 69 % of these Zero ARPU users churned in September

- As Zero ARPU users make 90 % of the Bottom 25% value segment, that is why 78 % of churning users belong to Bottom 25% segment, but they only make 1 % of the churning revenue.

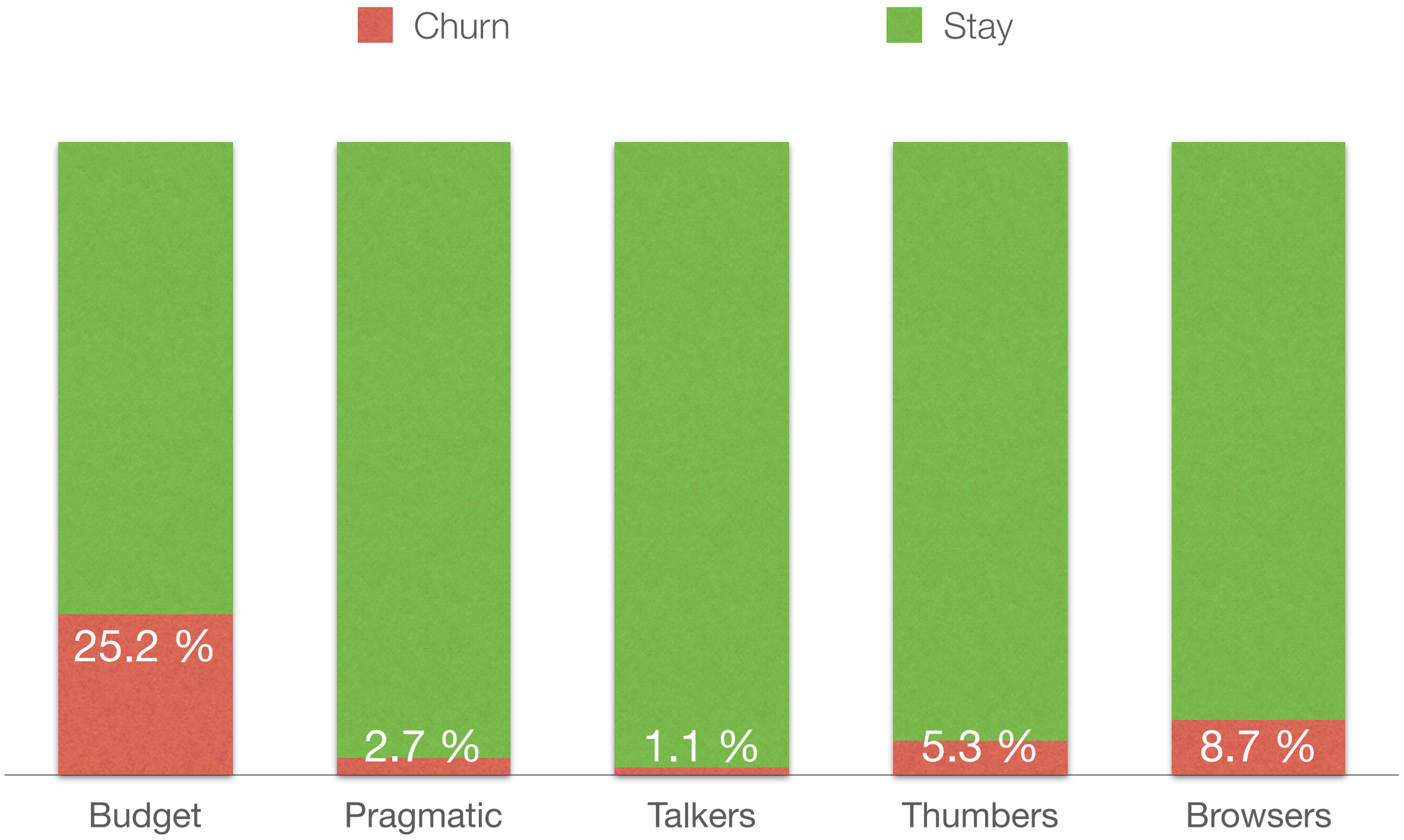
User Segmentation by Usage Profile



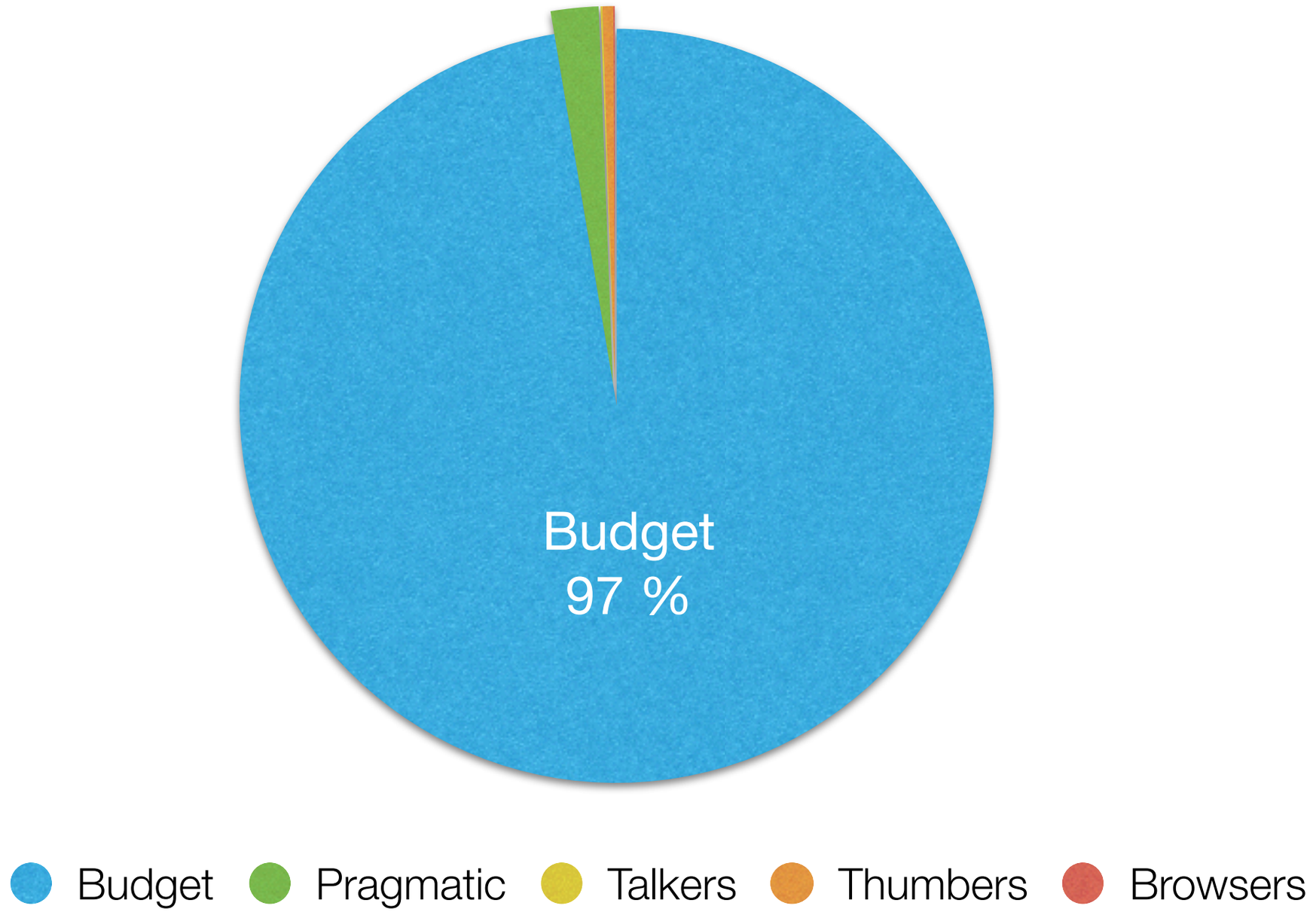
Overall View on Customer Base Composition



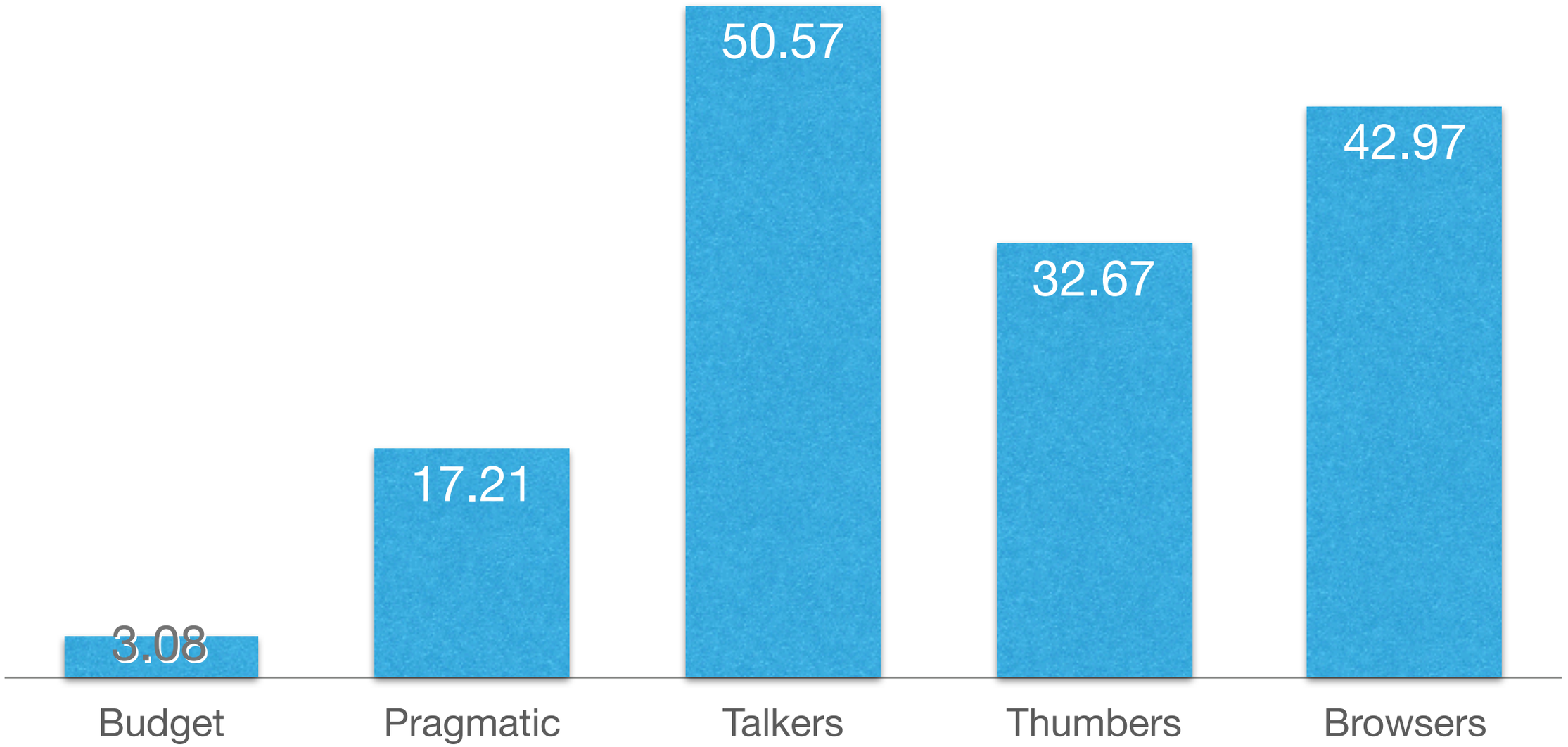
User Churn Rate by Usage Profile Segment



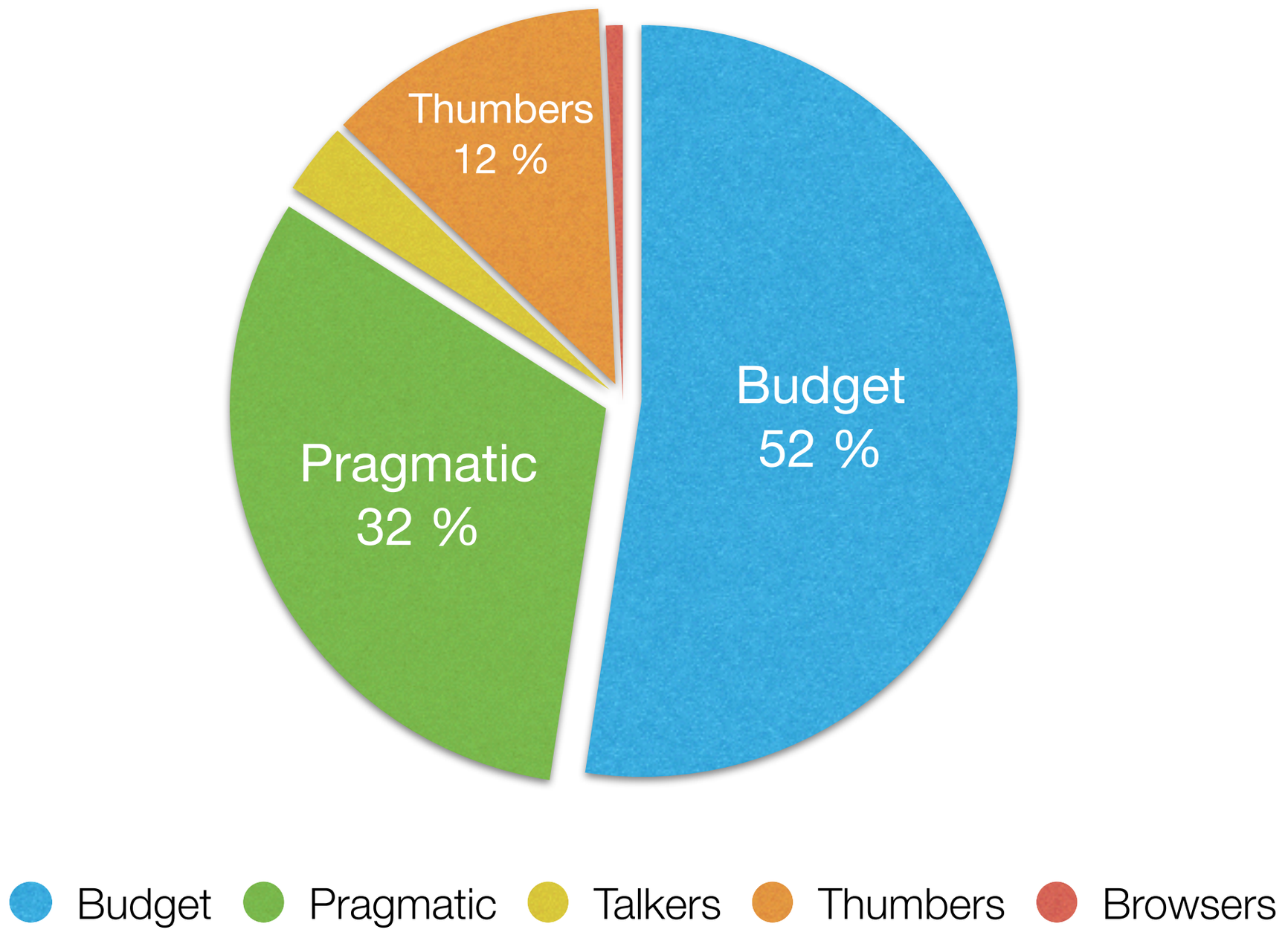
Composition of Churning Users by Usage Profile



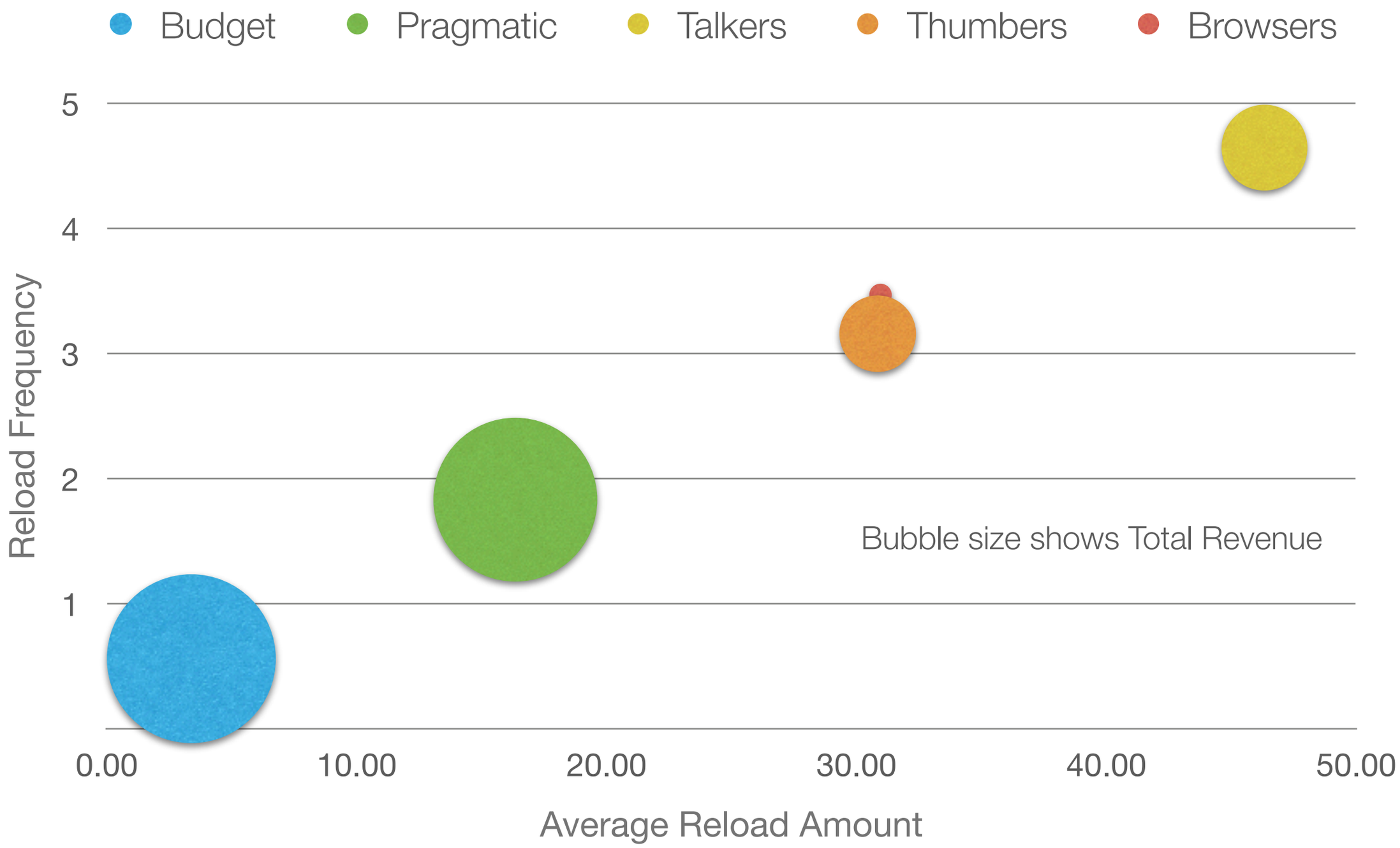
ARPU by Usage Profile Segment



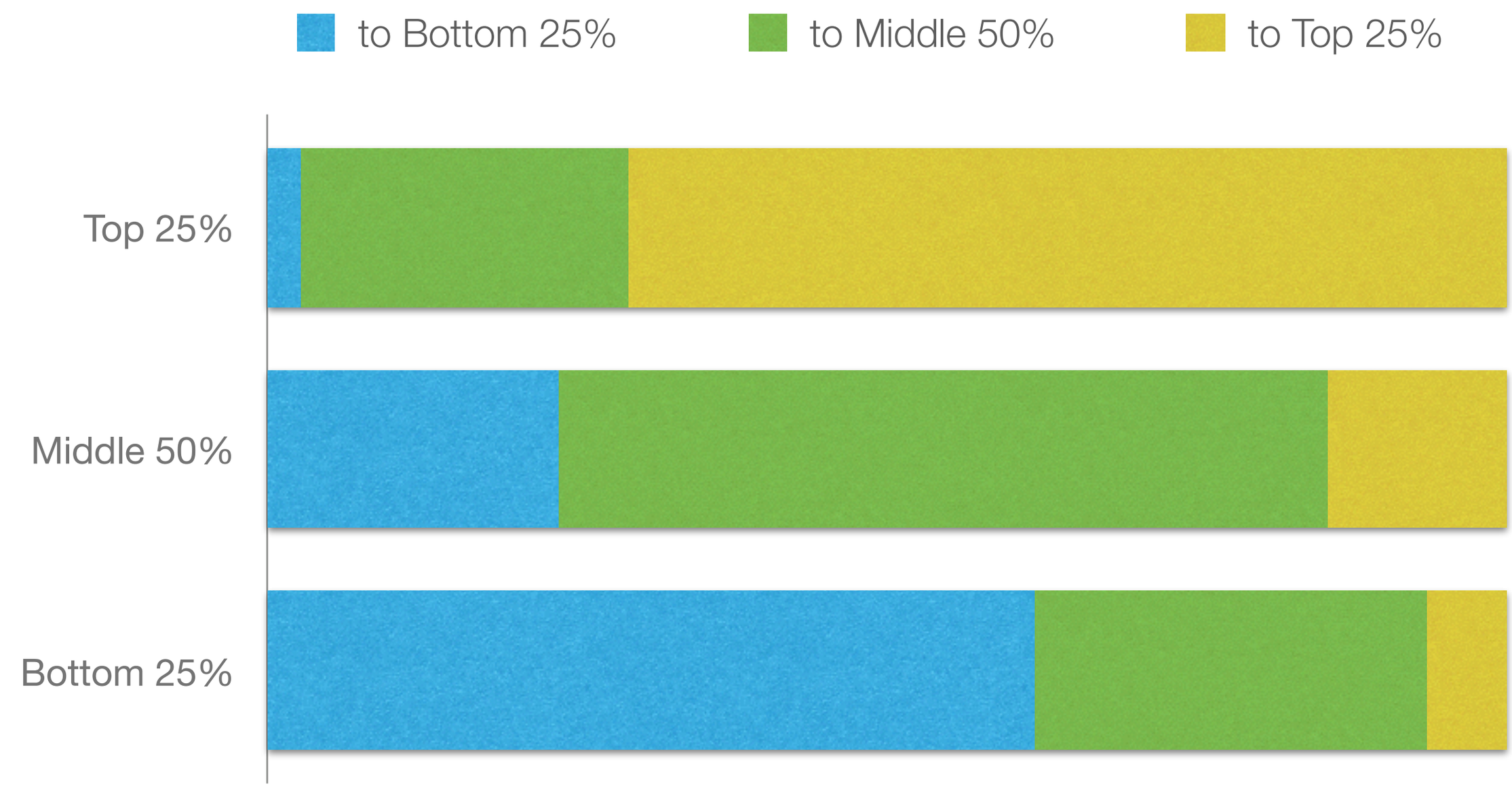
Composition of Churning Revenue by Usage Profile



RFM Segmentation by Usage Profile



Migration Trend between Value Segments



The Main Churn Issues to Target with Modelling

- From all segmentation and data analysis shown above, we could say that company has a big issue with the new customers as 63 % of them churn just the next month after starting to use the services.

- And even if majority of these users are either paying very little or even paying nothing in this first month (so when churning they do not take away the existing revenue from the company), but these lost new users is the lost potential revenue of the future.

- So the first target for churn modelling could be the New Users segment.

- Another important issue with the churn is lost revenue. And in this regard I decided to model the propensity to churn of the „Pragmatic“ users according to usage profile segmentation, as they constitute the second biggest part of the churning revenue (first biggest being „Budget“ users, and big part of them would be targeted through the campaign addressed to the New users anyway, as these segments partially overlap).

Modelling the New Users Churn

- As majority of New users churn the next month from intake, so we should act fast and therefore we should model the propensity to churn based on the data just from one month.

- We have 9,924 new users in August with 6,282 of them churning in September. This is quite well balanced segment in regard of our target variable, and I decided to try two most popular techniques on it: logistic regression and random forest.

- I did try different techniques to find the best explanatory variables for logistic regression, but no trials gave me results superior to the simple approach of leaving all the variables of 8th month in regression, so finally I took this straightforward way.

- Random forests by their design choose different sets of predictors when building individual trees, so they are well suited for the problems with big number of explanatory variables. Also, I used R implementation of conditional inference random forests (`cforest` in `party` package) as they deal better with predictors varying in their scale of measurement or number of categories.

Comparison of Two Prediction Models

- To better compare the expected accuracy of two models, I chose to run a 5-fold cross-validation for both of them, averaging the results of all 5 runs.

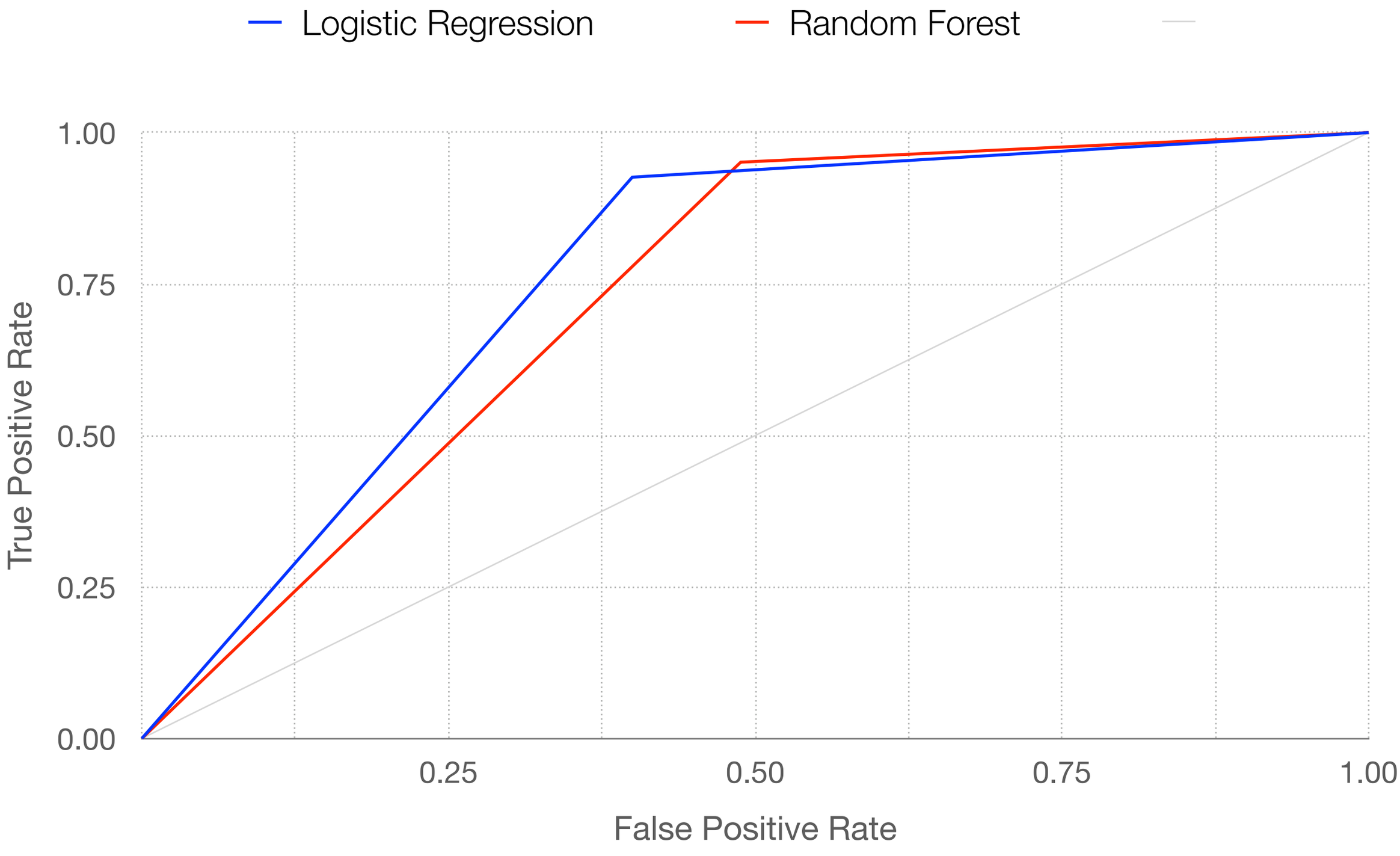
- Both models did well with quite similar results, random forests giving a bit better true positive rate (95,1 % vs. 92,6 %) at the expense of the bigger false positive rate (48,8 % vs. 40,0 %).

- I think it's important that true positive rates of both models is high, as it's important not to miss the churners in our incentives campaign. Also it should be mentioned that logistic regression had it's benefit in easier computational requirements.

- On the other hand, users constituting false positive results for the models are in some way similar to the churners, so even if they did not really churn in September, they could churn in the coming months, so it probably does no harm to also target them with incentives.

- ROC curves of the models are provided for comparison on the next slide.

ROC Curves of the Models



Modelling the „Pragmatic“ Users Churn

- We have 10,456 users in Pragmatic segment, but only 284 of them churned in September, so this segment is very much unbalanced with regard to our target variable: the class of interest (churners) is much smaller than the majority class of non-churners.

- A classification algorithm would under-perform when trained on this set (actually it would predict all users as non-churners in our case). Therefore we have to use some technique to deal with this problem.

- I decided to user a simple under-sampling technique and prepared the training set composed of all the minority class members (churners) and the same number of randomly selected majority class members.

- Then the random forest model was fit to the training set and evaluated on the full set of Pragmatic users. The number of trees in the forest was increased in this model due to the bigger number of predictor variables, as all the variables of 3 months were used for modelling.

Evaluation of the „Pragmatic“ Users Churn Model

- The resulting model gave us 81,7 % true positive rate with 22,0 % false positive rate (the confusion matrix provided below).

		Actual	
		Churn	Non-Churn
Predicted	Churn	232	2244
	Non-Churn	52	7928

- From practical point of view this would mean that we would be addressing 2,476 users (23.7 % of „pragmatic“ users or 3.7 % of all user base) to reach and possibly stop from churning 232 churners.

- But if the incentives campaign is successful, this would let us to potentially retain 25,7 % (4,327.71) of the churning revenue. Depending on the cost of incentives campaign per user, this could be well worth pursuing.