# Ramp Technical Assignment

Michail C. Melonas

25 October 2020

## 1) Calculate a generic retention curve, weighted by cohort size over the last 180 days of data. (This is done to capture temporal recency)

Retention rates measure the proportion of users whom continue product-use over time. These rates are useful to understanding specific product features and user sub-population behaviours.

Figure 1 below presents the retention observations and curve fitted to the provided cohort data. Using the 180 most recent groups (starting and ending on 2020/03/14 and 2020/09/09, respectively), the observations were weighted according to size (i.e., daily new users). As can be seen, it was found that a power function,

$$r(x) = 0.428x^{-0.432} \qquad \text{with } x > 0,$$

delivered a good fit to the retention data.

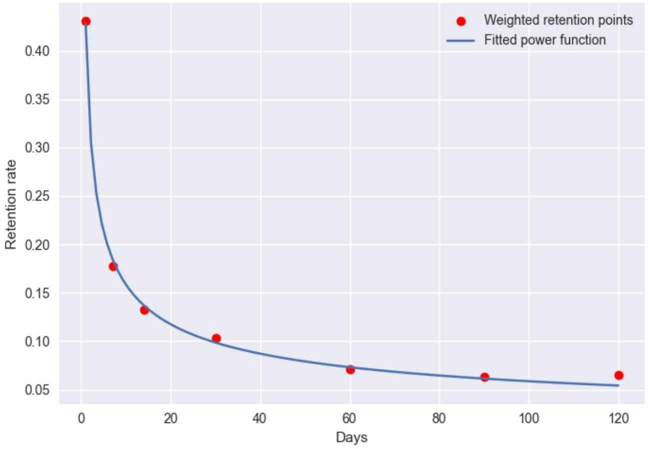

Figure 1: Weighted average retention values and fitted retention curve.

## 2) Formulate an appropriate model that describes the relationship:

$$\boldsymbol{DAU(t) \sim DNU(t) \cdot \{D_x | x\}}$$

We assume that daily active users are based on two factors: new users and returning users. That is, $DAU(j)$ consists of those users who were obtained in the most recent period, say $t = j$, and of those users retained from previous periods, $t = 0, 1, \ldots, j - 1$. Mathematically, we can deconstruct $DAU(j)$:

$$\begin{aligned} DAU(j) &= DNU(j) + DNU(j - \mathit{1})\ D_1(j-1) + \ldots + DNU(\mathit{0})\ D_j(0) \\ &= \sum_{x=0}^{j} DNU(j - x)\ D_x(j - x) \end{aligned}$$

where $DNU(t)$ is the number of new users gained at time $t$, and $D_x(t)$ (with $D_0(j) = 1$) is the proportion of users having joined at time $t$ whom are still active $x$ periods later.

Of course, looking forwards, we won't have access to $DNU(t)$ observations. Also, the future retention rates, $\{D_x(t)|x\}$, are both unknown and are expected to depart from the historic rates over time (e.g., due to changes to the underlying product, or network effects encouraging lost users to return). Therefore, if we wish to make claims about future user statistics, we are left with having to settle for forecasts. Consider the estimator:

$$\begin{aligned} \widehat{DAU}(j + n|t = j) = &\widehat{DNU}(j + n) + \widehat{DNU}(j + n - \mathit{1})\ \widehat{D}_1 + \ldots + \widehat{DNU}(j + \mathit{1})\ \widehat{D}_{n-1} \\ &+ DNU(j)\ \widehat{D}_n + \ldots + DNU(\mathit{0})\ \widehat{D}_{n+j} \end{aligned}$$

where we assume that we are currently at time $t = j$ and wish to predict daily active users $n$ periods ahead. Here, $\widehat{DNU}(t)$ denotes predicted daily active users at time $t$, and $\widehat{D}_x$ is the estimated proportion of users still active $x$ periods after signing up.

Looking at the above estimator, we note that new users and retention rates require distinct estimation strategies (the former being related to acquisition efforts, and the latter being product driven). We propose to approach the problem of future new users as a univariate time series. Retention rates can be obtained via a parametric retention curve fitted to historic $D_x(t)$ values.

## 3) Train / fit the model described in 2, over the 180 day range described in 1 and forecast DAU ahead two years.

Figure 2 tracks the 180 most recent user acquisitions. There appears to be (some) regularity in this time series. However, the task of making claims up to 2 years into the future meant that the fitting of a traditional time series model (e.g., of ARIMA-type) was not feasible. For this reason we have opted for a very simple approach to the modelling of future daily new users: assume this value to be a constant which is equal to the average DNU figure over the most recent 180 day period.

Figure 3 displays the result of our proposed model.

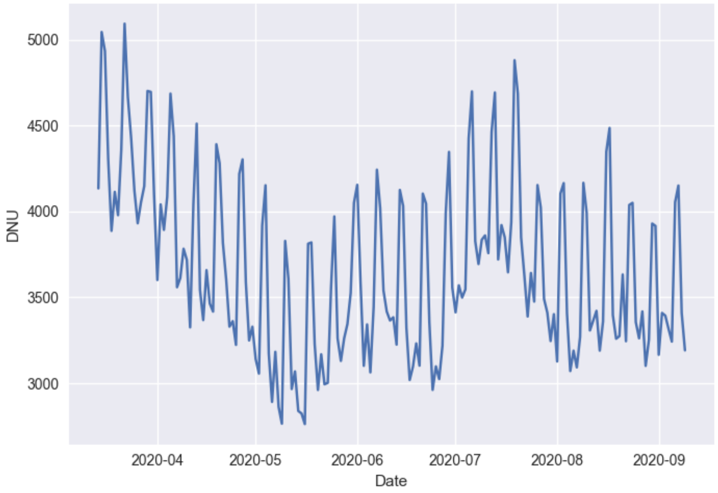

Figure 2: Daily new users for the period 2020/03/14 to 2020/09/09.

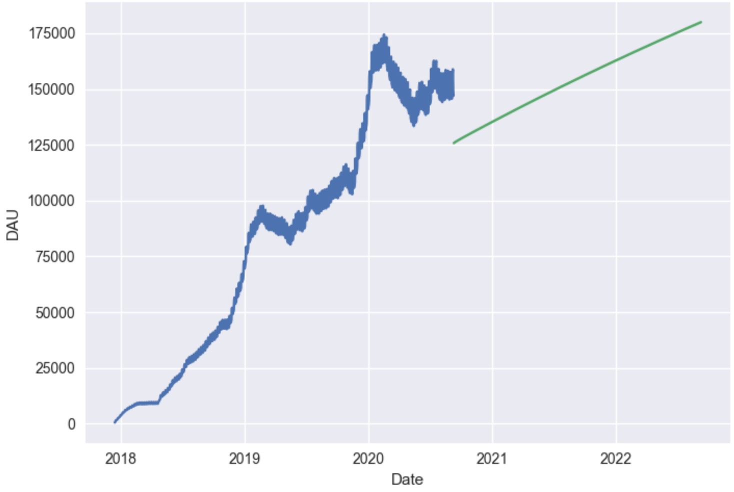

Figure 3: Actual daily active users for the period 2017/12/15 to 2020/09/09, and predicted usage for 2020/09/10 to 2022/09/09.

## 4) Using whatever means deemed necessary, validate your findings.

The proposed approach is faulted to a degree (this can be seen from the discontinuity in Figure 3). However, its partial success can be validated by applying the model the 100 most recent observations. That is, by comparing the predicted daily active usage to the (actual) observed values. This is demonstrated in Figure 4 below. Although we appear to be underestimating the observed values, the prediction are not far off, and track the correct direction. This suggests that our approach might be useful for the task of predicting daily active usage.

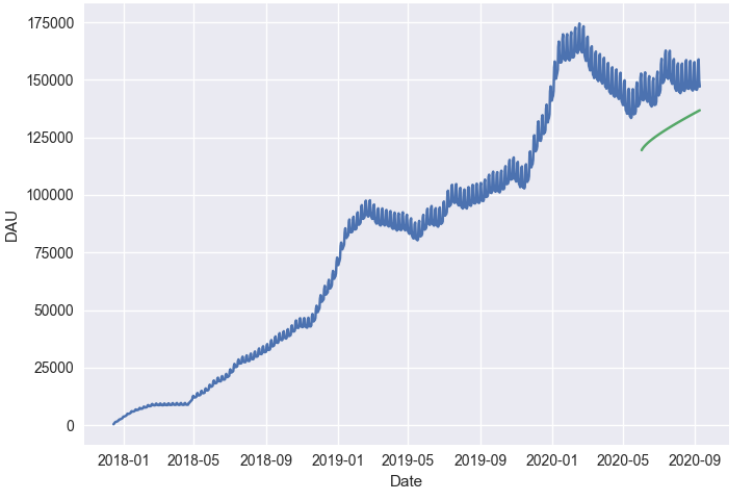

Figure 4: Actual daily active users for the period 2017/12/15 to 2020/09/09, and predicted usage for 2020/06/02 to 2020/09/09.

## 5) Using what you have learned from the data, recommend areas of improvement and further study.

The most obvious short-coming of the above proposed approach is the modelling of daily new users. An improved time series modelling (i.e., one that is more advanced than simply averaging past values and assuming the number of future acquisitions to remain constant over time) would undoubtedly improve accuracy. Perhaps access to the underlying drivers of new users (e.g., marketing efforts) would allow a covariate-based approach to the time series. In addition, access to said drivers would allow for scenario planning (i.e., comparing future daily active users or revenue under varying sets of circumstances).

Another possible area of improvement is the retention modelling component. Although the power function was found to be superior when compared to linear, quadratic and exponential functions, other approaches (e.g., a survival- or non-parametric modelling) are yet to be tested.