Avanika Narayan
11/3/2018

A Statistical Approach to Modeling and Forecasting the Life Cycle of Various SaaS Business Models

Overview

The goal of this project is to build a robust model for projecting and understanding the revenue life cycle of different categories of SaaS companies and Saas business models. The growth of a SaaS business from both a revenue and an account perspective is the result of several contributing factors. Growth can be broken down into two parts: market penetration and drivers of growth. Drivers of growth are the sources of revenue such as new, churn, contraction, resurrection and expansion revenue that contribute to the overall revenue growth. Putting both concepts together: at different stages of growth in a business, different drivers of growth will contribute to the aggregate revenue gained or lost at varying magnitudes. The contribution of these sources of revenue to the top-line is strongly correlated with the underlying dynamics of the company question. For example, companies that sell services to small to medium sized business are subject to significantly more customer churn than companies that sell software to enterprise business. Deriving insights on how these factors affect the top-line revenue and the ultimate life cycle of a business will be incredibly insightful for forming *investment decisions*, f*orecasting stock prices* and fundamentally *understanding the health of a business.* The idea of modeling the revenue growth of companies at such a granular level has not yet been done and and doing so could provide profound insights on the underlying mechanics of SaaS business models that have gone otherwise unnoticed.

Model Description

1) DEFINE VARIABLES:

Let Q be a variable that represents a list of financial quarters in the history of a company where q_i is some arbitrary financial quarter.

1. **ARR(q_j) = ARR(q_{j-1}) + Net New ARR($q_j - q_{j-1}$)** : the total revenue recorded in a given quarter is revenue from the previous quarter summed with all new revenue earned over the subsequent quarter.
2. **Net New ARR($q_j - q_{j-1}$) = New ARR($q_j - q_{j-1}$) + Expansion ARR($q_j - q_{j-1}$) - Churned ARR($q_j - q_{j-1}$) - Contraction ARR($q_j - q_{j-1}$)** : the net new arr earned at the end of a given quarter is the sum of the revenue accumulated from the acquisition of new accounts, with new revenue from expansion of existing accounts minus any revenue lost from churned accounts or downgrades in accounts from the prior quarter.
3. Breaking down Net New ARR:
    a. **Total New ARR($q_j - q_{j-1}$) = New ARR($q_j - q_{j-1}$) + Expansion ARR($q_j - q_{j-1}$)**
    b. **Annual Gross Dollar Churn($q_j - q_{j-1}$) = Churned ARR($q_j - q_{j-1}$) + Contraction ARR($q_j - q_{j-1}$)**

Let us understand the components of growth and the variables they are dependent upon.

- Total New ARR (T):
  - **$T(q_j) = ARR(q_{j-1}) * Gr(q_j)$** where Gr = new arr growth rate. Let us denote Gr as some random variable where P(Gr = g) is dependent on some variable X. Based on observed data, I make that claim X is average contract value. Intuition? SMB-focused companies, which have smaller account values, grow significantly faster than enterprise-oriented businesses. Why? Companies that provide enterprise software have longer sales cycles and thus grow at slower rate. For SMB-focused companies, acquiring new customers takes a shorter sales cycles that requires less sales and marketing spend. Moreover, we note that the Gr in quarter j, $Gr(q_j)$, is dependent on the new arr growth rate of the prior quarter. Why is this the case? For the purposes of building financial models, the principle of a **time series** is incredibly handy. Rather than randomly sampling a growth rate for a given quarter, conditioning on the prior growth rates enables for a more accurate forecast of the growth.
    - Conclusion: **$T(q_j) = ARR(q_{j-1}) * Gr(q_j)$** where **$P(Gr(q_j) = g \mid Av = v, Gr(q_{j-1}) = g_{j-1})$** where Av is ACV
- % of Total New ARR from Expansions (E)
  - **$E(q_j) = T(q_j) * Er(q_j)$** where Er is a RV that represents the % of total new ARR that is comes from account expansion. Let us denote Er as a r.v. that is dependent on some variable X. Based on observed data, I make that claim that X is average recurring revenue. *Intuition?* The percent of new ARR from expansions and upsells is significantly higher for enterprise software providers in comparison to SMB-Focused companies. This is likely the effect of the "land-and-expand" philosophy of such business models. Often, great SMB businesses will grow with their customers and "move up market." This means that over time while a company may begin by focusing on SMB sales, they may expand and upsell into Enterprise size sales even with existing customers (as well as new customers). As a result, modeling E as a function of ARR made the most logical sense. Moreover, we note that when computing Er in quarter j, $Er(q_j)$, we condition upon our knowledge of the expansion percentage of the prior quarter.
    - Conclusion: **$E(q_j) = T(q_j) * Er(q_j)$** where **$P(Er(q_j) = e \mid A = a, Er(q_{j-1}) = e_{j-1})$** where A is ARR
- New ARR (Nr) - Revenue from new accounts:
  - **$Nr(q_j) = T(q_j) - E(q_j)$**
- Annual Gross Dollar Churn (Ch):
  - **$Ch(q_j) = ARR(q_{j-1}) * Cr(q_j)$** where Cr = gross annual churn rate**.** Let us denote Cr as some random variable where P(Cr = c) is dependent on some variable X**.** Based on observed data, I make that claim X is average contract value. *Intuition?* Smaller accounts churn more than larger accounts as the stakes to cancel a $100 subscription are far less than canceling a subscription worth $>250K. As a result, we model the Cr as an r.v. that is dependent on average contract value. Moreover, we note that when computing Cr for quarter j, $Cr(q_j)$, we condition upon our knowledge of the churn rate of the prior quarter.
    - Conclusion: **$Ch(q_j) = ARR(q_{j-1}) * Cr(q_j)$** where **$P(Cr(q_j) = c \mid Av = v, Cr(q_{j-1}) = c_{j-1})$** where Av is ACV

- Logo churn (L) - # of churned accounts
    - **$L(q_j)$ = Total customers(q_{j-1}) * $Lr(q_j)$** where $Lr(q_j)$ = Logo churn rate. Let us denote Lr as some random variable where P(Lr = l) is dependent on some variable **X.** Based on observed data, I make that claim X is average contract value. Intuition: SMB-focused companies face significantly higher customer churn rates than enterprise companies. Why? The stakes are higher for a big company to cancel a subscription than an small 50 person team! Moreover, we note that when computing Lr for quarter j, $Lr(q_j)$, we condition upon our knowledge of the logo churn rate of the prior quarter.
        - Conclusion: **$Ch(q_j)$ = Total customers(q_{j-1}) * $Lr(q_j)$** where **$P(Lr(q_j) = l \mid Av = v, Lr(q_{j-1}) = l_{j-1})$** where Av is ACV
- Churned ARR (Tc) - Revenue lost from churned accounts
    - **$Tc(q_j)$ = Total Customers(q_{j-1}) * $L(q_j)$ * ACV**
- Contraction ARR (Ca) - Revenue lost from downgrades and downsizes in accounts
    - **$Ca(q_j) = Ch(q_j) - Tc(q_j)$**

2) CHOOSE DISTRIBUTIONS

Having defined all our random variables, we must now choose the distributions that describe these random variables. To summarize, we need to determine distributions for the following variables:

- **Gr** : new arr growth rate
- **Er** : % of total new ARR that is comes from account expansion.
- **Cr** : gross annual churn rate
- **Lr:** Logo churn rate

Deciding what distribution best fit the semantic meaning of the r.v's was based on the idea that all of the growth rates must be values greater than 0. In other words, the variable Er (% of total new ARR that is comes from account expansion) cannot take on a negative value as this would be impossible. As a result, I decided that the best distributions to describe the variables would either be a gamma distribution, beta distribution or uniform distribution.

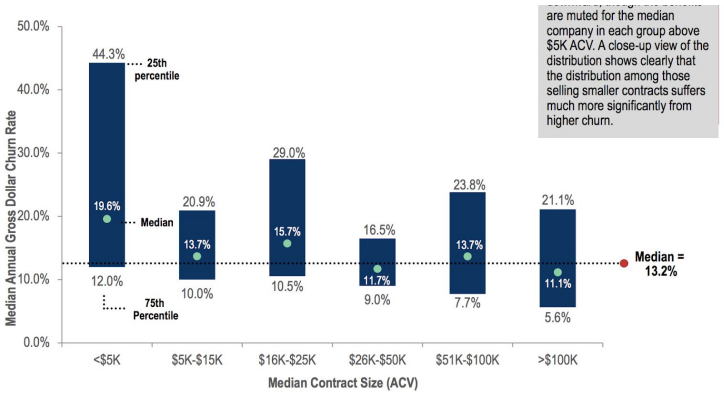
- $\boldsymbol{Gr} \sim Beta(a_g, b_g)$
- $\boldsymbol{Er} \sim Beta(a_e, b_e)$
- $\boldsymbol{Cr} \sim Beta(a_c, b_c)$
- $\boldsymbol{Lr} \sim Uni(x_l, y_l)$

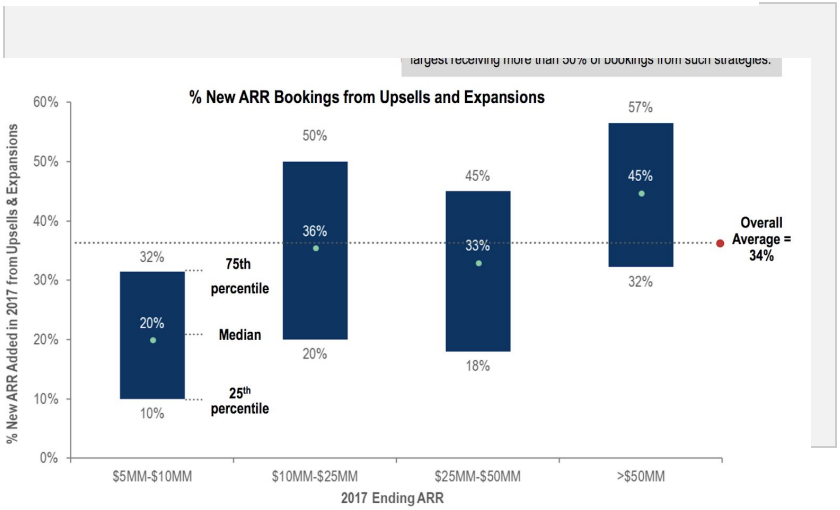
Having decided what distributions would be best suited for modeling the data, I embarked on a journey of estimating the parameters for each of these distributions.

Avanika Narayan
11/3/2018

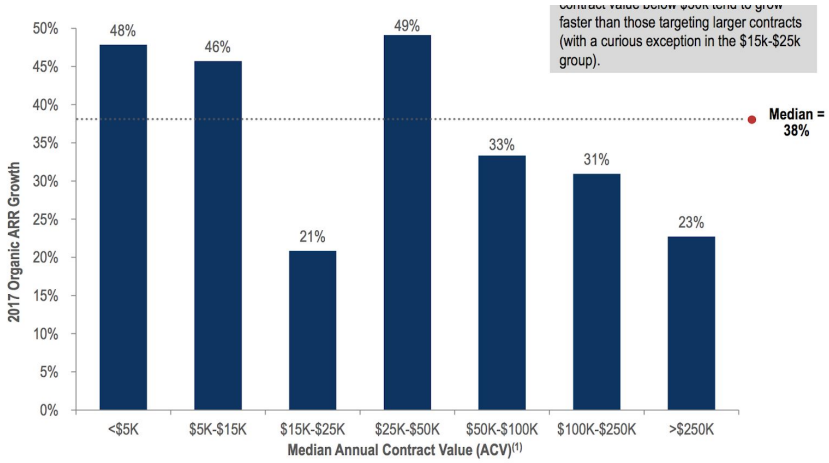
PARAMETER ESTIMATION

In order to determine the parameters for the given distributions I used data from 2018 KeyBank survey of SaaS companies. The data is as follows:





Avanika Narayan
11/3/2018



As seen above, I was needed to determine the parameters of the distributions of the variables (Er and Cr) based on the knowledge of three statistics (median, Q1, and Q3), and the distribution of Gr based on one statistic (median). To determine the parameters of the distributions for Er and Cr I used the principle of ***gradient descent*** to estimate the parameters *a* and *b* of the corresponding Beta distributions. I employed the use of gradient descent to minimize the following cost function:

$$(F(Q_1) - .25)^2 + (F(Q_2) - .5)^2 + (F(Q_3) - .75)^2 \text{ where } F(X) \text{ is the CDF of a Beta}$$

* To compute the partial derivatives of the CDF of the Beta with respect to *a* and *b*, I used the incomplete beta function

In other words, I was looking to find parameters (*a*,*b*) that maximized the likelihood that the resulting distribution had a Q1, Q3 and median that closely resembled the observed statistics in the KeyBank survey.

Choosing initial parameters:

I observed that in order for the algorithm to converge on the parameters which best minimized the loss function, I needed to set the initial values of *a* and *b* in the following way. I set the initial value of *a* to 1. For the initial setting of *b,* I used the formula for calculating the median of a Beta distribution, along with the observed median, in order to determine a ratio between *a* and *b:*

$$\frac{a - \frac{1}{3}}{a + b - \frac{2}{3}} = Median$$

$$b = \frac{a - \frac{1}{3} - aQ_2 + \frac{2}{3}Q_2}{Q_2} \text{ where } Q_2 = \text{median}$$

Having set *a* to 1, I used the formula to compute the initial value of *b.*

Parameters for Er and Cr:

Using this method, I calculated all the parameters for the distributions of Er and Cr as a function of ACV and ARR respectively (See below for the computed parameters). It is important to note, that prior data indicated that at different ranges of ACV, companies exhibit different distributions of churn rates and % of new ARR from expansions. Thus, for different ranges of ARR and ACV, the underlying distribution vary and so do the parameters of the Beta distributions that model the data.

```
Er = {
        '>250K': [1, 1.90],
        '50K-100K': [1, 1.02],
        '25K-50K': [1, 0.36],
        '<5K': [1, 0.39],
        '100K-250K': [1, 1.15],
        '15K-25K': [1, 2.17],
         '5K-15K': [1, 0.45]
}

Cr = {
        '26K-50K': [0.97, 4.697],
        '>100K': [0.94, 5.01],
        '<5K': [0.94, 2.40],
        '5K-15K': [0.964, 3.866],
        '51K-100K': [0.935, 3.866],
        '16K-25K': [0.971, 3.246]
}
```

Parameters for Gr:

Given that for the variable Gr, I was only provided with the median growth rates as a function of ACV, I simply determined the parameters values using the equation $b = \frac{a - \frac{1}{3} - aQ_2 + \frac{2}{3}Q_2}{Q_2}$ where Q_2 is the median. I set $a = 1$ and solved for *b.*

Parameters for Lr:

In determining the parameters for the the uniform distribution of the quarterly logo churn rate, I used the following metrics from Redpoint venture capitalist Tomasz Tungz:

	Monthly Churn Rate	Annual Churn Rate
SMB	3-7%	31-58%
Mid-Market	1-2%	11-22%
Enterprise	0.5-1%	6-10%

The values provided enabled me to compute the minimum and maximum values of the uniform distributions:

```
Lr  =  {
        'SMB': [.31, .58],
        'MID-MARKET': [.11, .22],
        'ENTERPRISE': [.06, .1],
}
```

3) BUILD THE MODEL

To understand how the model works, I will demonstrate by indicating how revenue for a given quarter q_i is computed.

From section 1 (variables) we know the following equation:
Net New ARR(q_j - q_{j-1}) = Total New ARR(q_j - q_{j-1}) - Ann. Gross Dollar Churn(q_j - q_{j-1})

Recall that T = Total New ARR and Ch = Annual Gross Dollar Churn are both variables that are computed as follows:

$T(q_j) = ARR(q_{j-1}) * Gr(q_j)$; $Ch(q_j) = ARR(q_{j-1}) * Cr(q_j)$

In order to compute $T(q_j)$ and $Ch(q_j)$, we must determine the value of $Gr(q_j)$ and $Cr(q_j)$.

Step 1: Determine Growth Rates

Recall that the following is an assumption made by the model:

- $P(Gr(q_j) = g \mid Av = v, Gr(q_{j-1}) = g_{j-1})$
- $P(Cr(q_j) = c \mid Av = v, Cr(q_{j-1}) = c_{j-1})$

- $P(Er(q_j) = e \mid A = a, Er(q_{j-1}) = e_{j-1})$
- $P(Lr(q_j) = l \mid Av = v, Lr(q_{j-1}) = l_{j-1})$

Here, $Gr(q_{j-1})$, $Cr(q_{j-1})$, $Er(q_{j-1})$, and $Lr(q_{j-1})$ are the growth rates of the prior quarter. Av is an r.v. that represents average account value and A is an rv that represents ARR: Let us begin by computing the rates:

$$Gr(q_j) = g_{j-1} * randomGrowthRate() + alpha*(g_{j-1} - convergence\_target)$$
$$Cr(q_j) = c_{j-1} * randomGrowthRate() + alpha*(c_{j-1} - convergence\_target)$$
$$Er(q_j) = e_{j-1} * randomGrowthRate() + alpha*(e_{j-1} - convergence\_target)$$
$$Lr(q_j) = l_{j-1} * randomGrowthRate() + alpha*(l_{j-1} - convergence\_target)$$

What do these variables means:

convergence_target: The convergence target is a function of the current ACV. Say, for example, the current ACV for a given company X is <5K. As a result, the current growth rate is sampled from the distribution which corresponds to the observed growth rates associated with companies that have an ACV < 5K. We know, however, that overtime accounts expand and new customers are acquired. As a result, the ACV may increases such that it enters a range of 5-10K. Based on the data we have seen so far, it is clear that companies in different ranges of ACV's tend to have different growth rates. As such, for different ranges of ACV's the distribution of growth rates also differs. This being said, when a company transitions from having an ACV <5K to an ACV between 5 - 10K, we must adjust the growth rate accordingly. So, for computation purposes this convergence target when company X is in the ACV range of <5K would be the mean growth rate of the distribution of growth rates for companies which have an ACV between 5-10K. Having this convergence model enables us to not have to randomly sample from the distributions every quarter based on current ACV. Rather, it allows for the model to consider the growth rate as a time series where the current growth rate is both a function of ACV and prior growth rates.

randomGrowthRate(): The function randomGrowthRate() helps model the inherent randomness in growth from quarter to quarter. The function simply samples from a uniform distribution with a range of .99 - 1.02.

alpha: constant which determines how quickly the growth rates converges to the convergence target. The current alpha values were chosen such that they maximized the likelihood that the model generated revenue outputs similar to those of data found for public companies.

Step 2: Compute Revenue for q_j

Hypothetically:

- ARR(q_{j-1}) = 1M,
- Total Number of Accounts in q_{j-1} = *1000*

After computing the rates above we have that:

- Gr(q_j) = .05
- Cr(q_j) = .02
- Er(q_j) = .07
- Lr(q_j) = .01

What is ARR in q_j *?*

ARR(q_j) = ARR(q_{j-1}) + T(q_j) - Ch(q_j)
T(q_j) = ARR(q_{j-1}) * Gr(q_j) = (1M * .05) = 50,000
Ch(q_j) = ARR(q_{j-1}) * Cr(q_j) = (1M * .02) = 20,000

So we have that:

ARR(q_j) = 1M + 50K - 20K = **1.03M**

What % of New ARR came from expansions - E(q_j)?!

E(q_j) = T(q_j) * Er(q_j) = 50000 * .07 = **3.5K**

Following this same logic, we can compute total number of churned accounts, revenue from new accounts, churned revenue from canceled accounts etc.

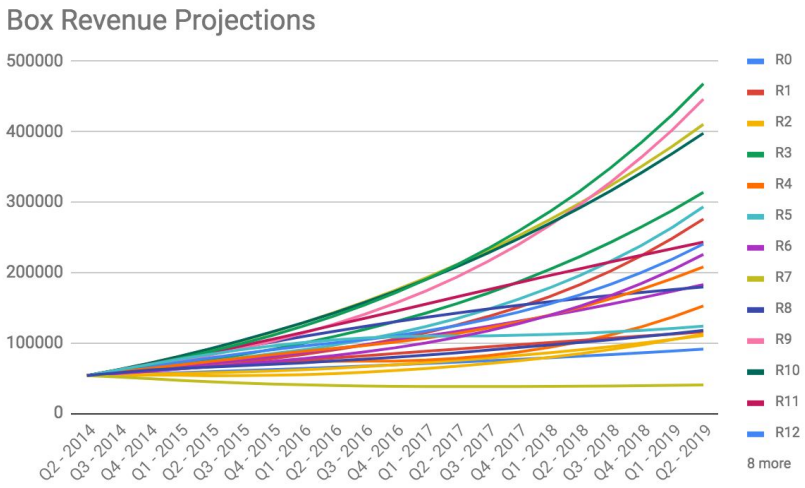
Step 3: Repeats Steps 1 and 2!!!!

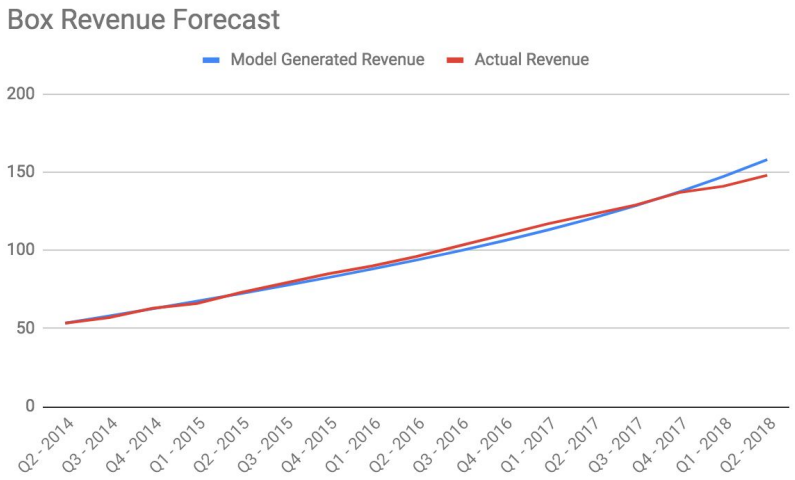
In order to generate a complete revenue projections for several quarters, compute steps 1 and 2 repeatedly!!!

Applications:

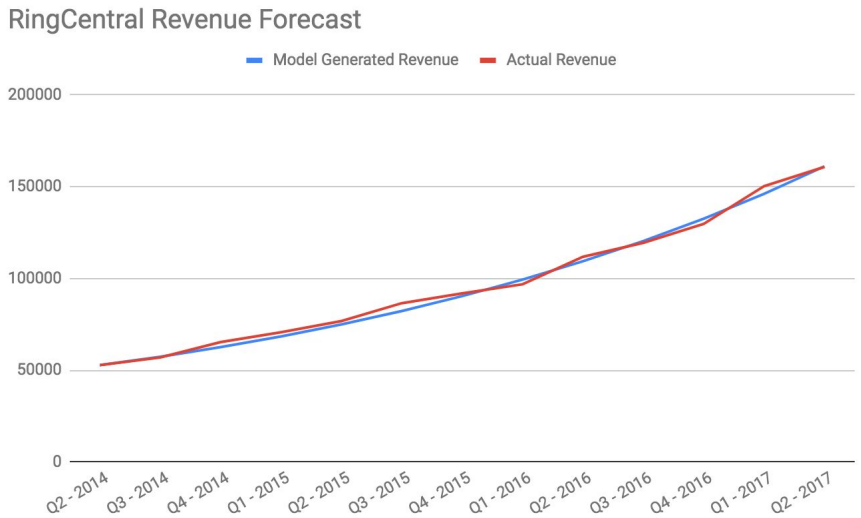
1. **Forecast revenue growth**
   a. Given an input of ARR @ IPO and number of customers: the model can make an incredibly accurate revenue projection! (See the below example of Box and RingCentral)

Ex. #1 Box:



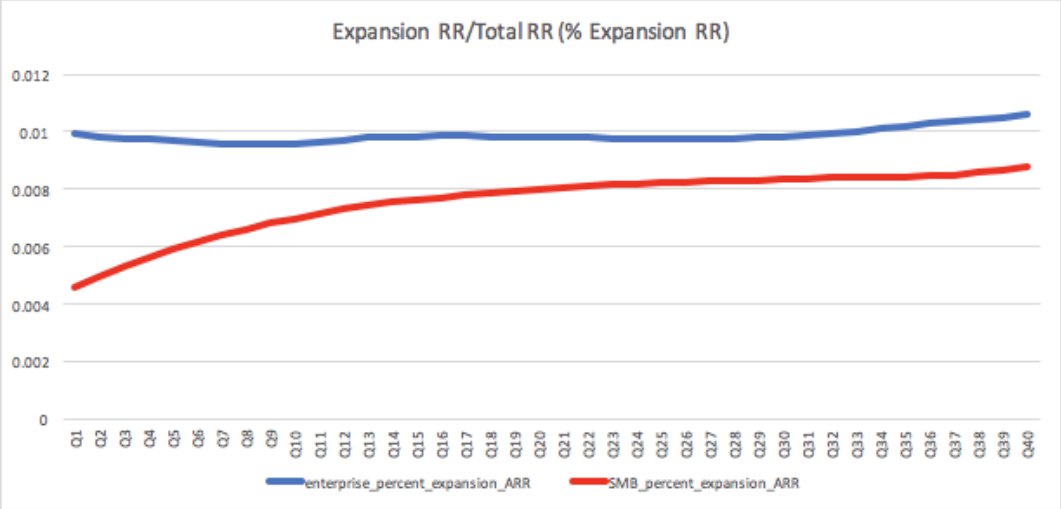


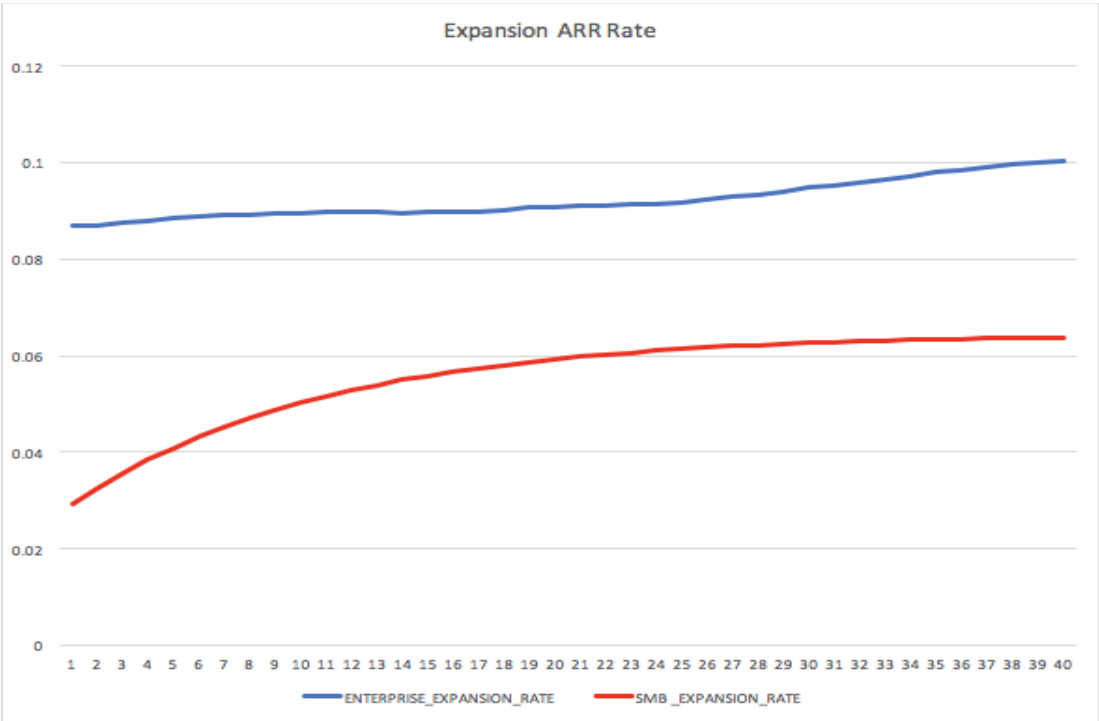
Ex. #2 RingCentral:

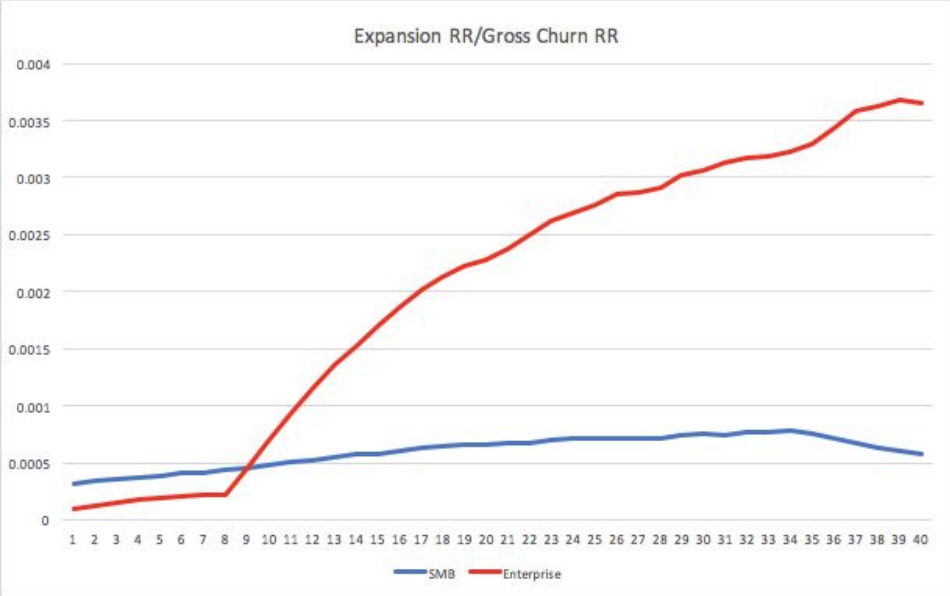


Avanika Narayan
11/3/2018

2. **Compare and Contrast Metrics Between SMB-Focused and Enterprise-Focused Companies**
    a. Let's take the example of an SMB company and Enterprise focused company both with 100 customers. The ACV for the SMB-focused company is $1000 and the ACV for the Enterprise-oriented company is $100,000. This implies an ARR of 100K and 10M respectively. We compare the % expansion RR, % churn RR and ratio of expansion RR/ churn RR for a total of 20 quarters:







Why is this interesting? Say I am a venture capitalist and I am presented a company that has 100 customers with an ACV of 100K. This company would be categorized as an SMB-focused company. I could use the model to derive insights on say, the average % expansion ARR for a company at this stage or the expected ratio of expansion RR/Churned RR.. Comparing against the expected value provides insights on whether a company is special. Clearly, special companies are ones that should be invested in!

3. **Making predictions about revenue forecasts for publicly traded companies**
   As shown in the screen capture, the model can help determine the probability that some revenue forecast for a given company is met. This could be incredibly useful if you are investing in stocks.

4. **And hopefully much more that I am unaware of :)**

Challenges:

1. Modeling distributions using only three statistics: This was one of the biggest challenges for me. I initially planned on determine the parameters of the distributions using bootstrapping but upon guidance from Professor Piech, I decided that parameter estimation was the best approach

2. Sparsity of data: Finding data to model this problem was one of the most difficult problems I faced. A large part of the project was piecing together analysis that I found online that provided insights on how the different drivers of revenue growth varied between different SaaS companies. I was also incredibly fortunate to have a friend who works at a venture capital company provide me with anonymized data
3. Understanding and modeling the problem: One of the hardest problems was understanding how the different drivers of growth played a role in determining the top-line revenue. I initially believed that expansion rates, churn rates etc. where all a function of penetration of the TAM. This, however, proved to be quite wrong and I had to rebuild the model under the assumption that the rates were a product of ACV rather than market penetration.

Sources:

- What Average Contract Value Is Best For A SaaS Company
- SMB or Enterprise - Which is the Better Go To Market in SaaS?
- The Maximum Viable Churn Rate for a StartupThe Maximum Viable Churn Rate for a Startup
- SaaS solutions: moving upmarket from smb to enterprise customers
- KeyBank 2018 National SaaS Survey
- ARR Growth for SaaS IPOs
- Yahoo Finance Workday Analysis
- Ideal Customer Target Size for a New SaaS Company
- Hubspot SaaS Benchmarking Study
- Macrotrends.net for consolidated financial earnings reports
- SEC.org

Avanika Narayan
11/3/2018