Revenue Forecasting Model methodology

Juan Barragan
Financial Engineering

Contents

1. Introduction

The digital economy is thriving, yet financial funding is harsh for pure digital players. Is somehow easy to understand why, namely, traditional banks rely on balance sheets and tangible assets for assessing the creditworthiness of any business. But as it happens, digital enterprises often, don't have balance sheets spanning several years of activity and owe mostly intangible assets like the ones of websites and brands. A turn point is the current ease of access to almost real time data on banking transactions and advertisement spend. This drive us to study another way of assessing risk and creditworthiness: namely, a healthy digital business has healthy transactions and spends.

In this note we explain how to use these transactions, concretely aggregated weekly revenue and aggregated weekly spend on advertisement for projecting the future revenue of a company and henceforth having a better view on the risk of a lending position on that company.

2. The data set

At Acme Corp, we have a hundred of clients but nonetheless for our study we concentrate the dataset to about 50. These 50 are the companies where we dispose of enough transactions data spanning for at least one year. This is so as, we want to forecast about six months of revenue, we need at least two times that amount. The methodology will consist on training a model on the fist 6 months (Or more if available) and then test on the last 6 months. Using the daily revenue, we aggregate it into weekly revenue concentrating the value on each Monday.

3. Tools

3.1. Time Series

The main tool for modeling the revenue will be the Time Series, a simple *Random Variable* X_t indexed by a discrete ordered parameter t which is equally spaced among adjacent instances.

3.2. Stationarity

We suppose implicitement that the time series we are analyzing are *stationary*. Let $\{X_t, t \in \mathbb{N}$ be our time series, suppose $Var(X_t) = \mathbb{E}((X_t - \mu)^2) < \infty$, where $\mu = \mathbb{E}(X_t)$. The autocovariance function, $\gamma_X(\cdot, \cdot)$ of X_t is defined as:

$$\gamma_X(r, s) = \mathbb{E}[(X_r - \mathbb{E}X_r)(X_s - \mathbb{E}X_s)]$$

We say that the series is stationary if:

$$\mathbb{E}|X_t|^2 < \infty, t \in \mathbb{N}$$
$$\mathbb{E}X_t = \mu, t \in \mathbb{N}$$

and

$$\gamma_X(r, s) = \gamma_X(r + t, s + t) \quad r, s, t \in \mathbb{N}$$

3.3. ARIMA

Let L be the *lag* or *shift* operator, $LX_t = X_{t-1}$, let ϵ_t be a suite of gaussian noises (With normal distribution, mean 0 variance σ) then an ARIMA(p, q, d) time series is

$$\Big(1 - \sum_{i=1}^{p} \alpha_i L^i\Big)(1 - L)^d X_t = \Big(1 + \sum_{i=1}^{q} \beta_i L^i\Big)\epsilon_t$$

ARIMA is composed by three parts. First the Auto-Regressive term:

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i}$$

this supposes that the revenue of the considered company behaves just like a linear combination of past values. Let $\nabla X_t = X_t - X_{t-1} = (1 - L)X_t$, then we have the differentiating therm,

$$(1 - L)^d X_t$$

which is helpful for eliminating seasonality and polynomial trend, for example if X_t is such that

$$X_t = Y_t + \sum_{j=0}^{k} a_j t^j$$

With Y_t stationary then we can differentiate k times, and we obtain:

$$\nabla X_t = k! a_k + \nabla^k Y_y \tag{1}$$

Which is now a stationary process with mean $k!a_k$. The last term

$$\Big(1 + \sum_{i=1}^{q} \beta_i L^i\Big)\epsilon_t$$

amount to a linear combination of current and past error terms. So we suppose the revenue of an enterprise weekly aggregated follows an ARIMA(p, d, q) process. We have tested a number of methods for calibrating ARIMA for each enterprise including Autocorrelations and hyperparameter tuning.

3.4. Latent ROAS

We also tried a custom model:

$$\begin{pmatrix} X_t \\ S_t \end{pmatrix} = \sum_{i=1}^{p} \begin{pmatrix} \alpha_i X_{t-i} \\ \beta_i S_{t-i} \end{pmatrix}$$

Where X_t is the revenue ans S_t is the amount of money the company has spent on advertisement. We suppose that only $q < p$, β coefficients are non-zero, i.e. $\beta_k = 0$ for $k \leq q$. More specifically, we take 12 weeks of revenue et 4 weeks of spend.

Roughly then, we suppose that the revenue of an enterprise is linearly dependent on past revenues and that the only way of changing that dynamics is to spend on advertisement. ROAS means Return On Advertisement Spend and measures the effectiveness of an ad campaign. As sometimes companies like Facebook and Google cannot sometimes fully track the client (Specially by confidentiality and privacy reasons) we don't know what ROAS is. So our approach is to consider it as a latent variable depending on ad spend. We calibrated the latent ROAS using simple Linear Regression on past observed values.

4. Results

We have separated the available data by enterprise on two sets *train* and *test*. The test one corresponding to the last 6 months (Actually 24 weeks) these values will remain unknown to the model. The training set is the remaining data (at least 24 + 2 weeks) We run exhaustive calculations' for hyperparameter tuning for getting the ARIMA parameters (p, d, q), up to 12 weeks of lagged revenue ($1 \leq p \leq 12$) up to 2 differentiation ($0 \leq d \leq 2$), meaning that we make the hypothesis of no trend, linear trend or quadratic trend) As a reference benchmark, we have compared performance with the *Moving Average* of ten weeks revenue. A word of caution here, the moving average is not actually projecting unknown values but rather are calculated on observed ones, that is, we allowed to the moving average to *know the future* via the test set. Error measure is a tweaked version of MAPE (Mean Absolute Percentage Error). We calculate the mean

absolute value of observed X_t against forecasted one F_t over the mean observed values:

$$E = \frac{\sum_{i=1}^{24} |X_i - F_i|}{\sum_{i=1}^{24} X_i}$$

The next figures (figures 1 - 4) shows how ARIMA performs on very different situations.

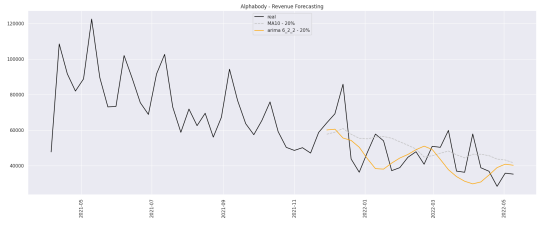

Figure 1: *Toto corp ARIMA forecast*

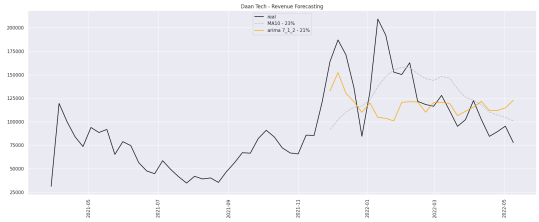

Figure 2: *Foe Tech ARIMA forecast*

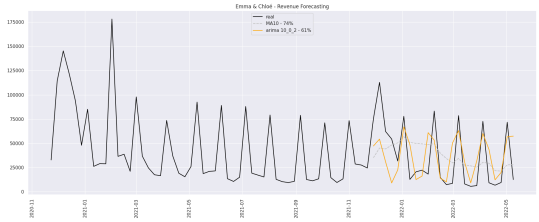

Figure 3: *You and Me corp ARIMA forecast*

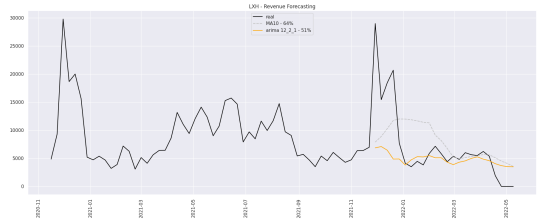

Figure 4: *YHX ARIMA forecast*

So we remark that ARIMA is robust: it outperforms almost every time the moving average. Knowing that the moving average uses observed values, this is indeed a pretty good performance. On the other hand, performance of our homemade latent

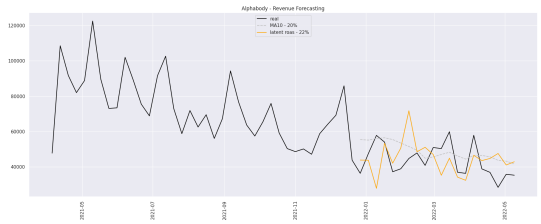

Figure 5: *Toto corp Latent ROAS*

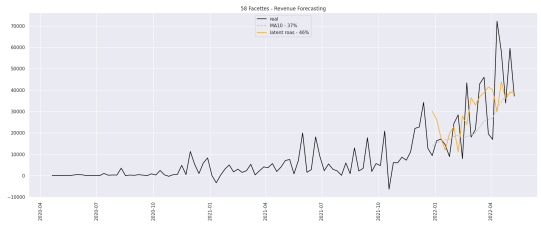

Figure 6: *58 Grades of Shade corp Latent ROAS*

ROAS is more contrasted as is shown on the next figures (fig 5 - 8)

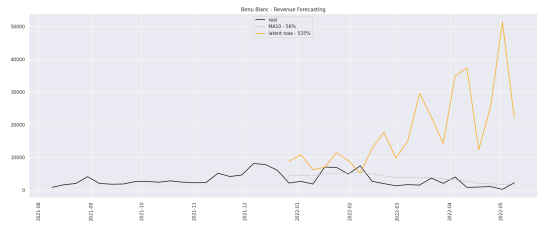

Figure 7: *Monsieur and Madame Blanc Latent ROAS*

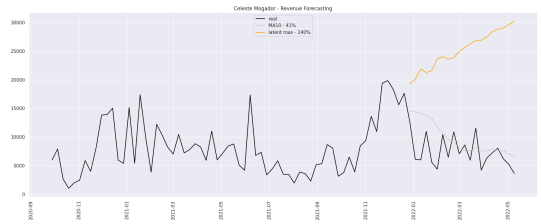

Figure 8: *Tacos Mogador Latent ROAS*

We also remark that although latent ROAS performs well on some situations it fails to eliminate some false trends.

5. Next Steps

We would like to continue to investigate the latent ROAS model as it gives us more freedom for studying other impacting variables. We can for example use the differentiating formula (1) for correcting false trends and forcing the time series to be stationary.

Another idea is to take into account that most e-commerce companies experience big sales on specific periods of the year. Namely, around the end of the year. To this end, we would like to introduce growth revenue correction by adding bump functions around these specific dates. On figures 9 we see how revenue (In percentage of annual revenue) is concentrated by month.

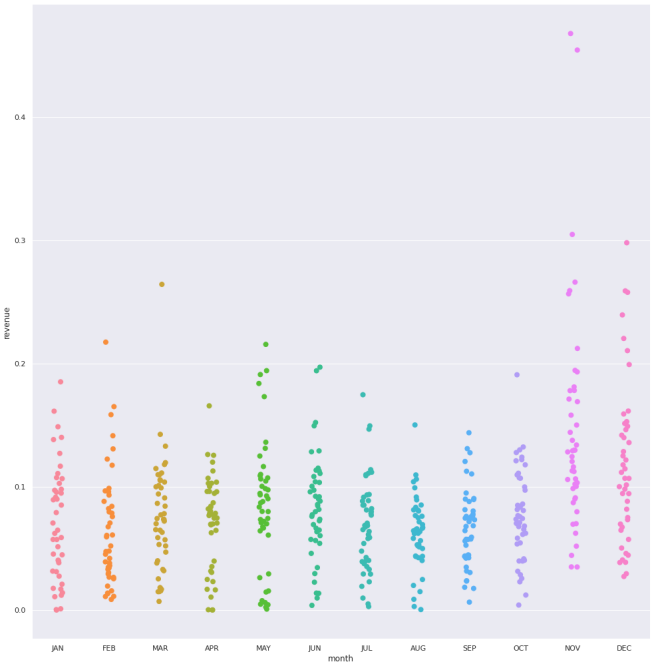

Figure 9: *Revenue percentage by month*

The figure 10 shows how growth behaves on percentage from months November to December and January to December.

Figure 10: *Revenue growth on key months*

The way to correct bump in sales can take the form of *Radial functions* which are functions with a bump concentrated around a point:

$$R(t) = \kappa e^{-\gamma|t-t_0|^2}$$

So as to have the revenue X_t modified as for example,

$$X_t(1 + r_X e^{-\gamma|t-t_0|^2})$$

r_X being the growth on revenue for that period. In order to calibrate these new parameters, r_X and γ we need to collect extensive data, cluster companies along verticals and infer their growth (and further decrease) on holidays.

Another interesting technique to explore could be the *Kalmann Filter*.

A. Summary of results

Here's the summary of errors obtained by ARIMA, by company for all the studied enterprises.

company name	MA10	ARIMA error
A	74%	61%
B	56%	49%
C	46%	41%
D	70%	79%
E	15%	14%
F	71%	81%
G	105%	216%
H	64%	60%
I	63%	60%
J	15%	16%
K	10%	10%
I	41%	72%
J	63%	45%
K	57%	42%
L	34%	30%
M	34%	85%
N	57%	55%
O 360	52%	59%
P	60%	89%
Q	44%	45%
R	30%	27%
S	26%	26%
T	39%	39%
U	20%	27%
V	38%	34%
W	87%	58%
X	18%	24%
Y	108%	54%
Tacos Mogador	38%	31%
Z	50%	50%
AA	38%	38%
VV	14%	16%
BB	54%	45%
CC	77%	79%
DD	16%	14%
EE	34%	39%
FF	43%	47%
GG	94%	72%
HH	65%	58%
II	16%	15%
JJ	50%	46%
KK	72%	62%
LL	22%	40%
MM	49%	53%
NN	20%	20%
OO	23%	21%
PP	64%	51%
QQ	17%	16%
RR	56%	48%
SS	77%	41%
WW	16%	14%

References

[1] P. J. Brockwell, R. A. Davis, Time Series: Theory and Methods. *Springer*

[2] W. H. Greene, Econometric Analysis *Prentice Hall*

[3] J. D. Hamilton, Time Series Analysis *Princeton University Press*

[4] L. Wasserman, All of Statistics *Springer texts in Statistics*