TP Enrollment Forecasting:

Does a Straight Edge Make a Good Data Scientist?

Since the dawn of time, humankind has wondered things about the future, and then failed to predict them. Our quest to predict future monthly enrollments at the TP level has made sizeable improvements over the oracles and charlatans of the days of yore, but just like them we still face the fundamental problem of using past information to predict the future. Our project began with a simple linear model, taking the previous 26 months’ enrollments to project future enrollments as a straight line struck across the graph. If we go back to this first file now, and evaluate its predictions of August – October, we see that with an RMSE of 28.8, it actually does a reasonably good job (Chart 1). However, by utilizing the tools of modern data science, surely this prediction must be akin to child’s play.

Our next step was to write an uncomfortably complicated query that added into the historical data set the ratio of active TruEx and TPP hospitals to all active hospitals, as well as the presence of both radio and TV DTC advertisement campaigns. From this initial exploration, we saw some pretty promising sea-born pair plots that suggested a few interesting things (Chart 2):

* Most of the TPs’ highest enrolling months appear to be during TV DTC months.
* Radio likely did not have as much impact on enrollments as TV.
* Excluding Karen Belanger outliers, there doesn’t appear to be a strong correlation with TruEx or TPP active hospital ratios.
* Excluding Karen Belanger, there is a correlation between enrollments and literal numbers of active TPP and TruEx hospitals, but this is confounded by our definition of “active”.

Undeterred, we set out to build a multivariate model that would tell us the impact of each of these factors on enrollments, so that we could potentially assign dollar values to them from a modelling perspective. The multivariate OLS model was short lived, because during a lunchroom conversation with Emily Wright, I was reminded that I had violated the OLS assumption of IID response variables (Please don’t tell Prof. Evgeniya Duzhak). In fact, total enrollments plotted against frequency of occurrence don’t look anything like an Independently Identically Distributed random variable. They look like Chart 3, which more or less matches a gamma distribution.

Great, so we threw away the multivariate model and moved to a Poisson regression, but immediately threw that out as well because the average of our predicted variable was not equal to the variance. The average monthly enrollment value was 116.3, but the variance was about 12,312. So we used a negative binomial regression, which finally matched our prerequisite assumptions. This was great, because we finally had a model that would take each one of our carefully queried variables and use it to find the slope of enrollments with respect to that variable, all else equal. But this was bad because now we had a model that couldn’t be used to generate predictions for each TP in each month for the next year or so, because we would have to feed them in one at a time and collect the results. And while yes, we were in the process of figuring out how to take this model and feed it an array of variables so we could collect an array of results, and then iterate over all the months and TPs, we never quite got that far, because enter: Dave Jaw and DataRobot.

Dave suggested that this would be a great use case for DataRobot, and showed me how to use the tool. It’s a pretty amazing tool, because it takes in your data set, and then trains up to about 50 models on it, and then goes back and validates the top performers against a holdout set. Most of the time, we got model BP 32 XGBoost back as the top performer (or very near the top). The next week or so however, in a phone meeting with Data Robot, we were told of their new time series feature, and they offered to let me experiment with it during a trial. Utilizing the times series model I think we learned a few things that merit repeating. First, everything we had already done in the Poisson regression and the regular XGBoost model was clearly wrong, because we had been training a model using past data to evaluate past data. If we want to correctly train a model to predict the future, we need to make sure we are recreating the naivety of past epochs by training it on older periods than the one we test it on. And if we want to evaluate models correctly, then we need to withhold the most recent data and test them on it by hand. Secondly, we learned from this that nothing is as predictive of future enrollments as past enrollments.

If you take a look at the top 20 or so variables the time series created for itself, they are all lags on or aggregates of enrollments. This implies that DTC campaign flags and high TPP ratio territory months are less predictive of future enrollments than the 20th lagged enrollment variable. That’s pretty amazing, because even after adding in almost 20 new variables from the Regional Market Report – enrollments by source, certs by source, issued certs, web leads, cancellations, etc. – out of the 20 top derived variables in the time series model, fully 18 of them were still just enrollment lags. The two that took slots 19 and 20 were both lagged web leads for pets under 5, and further examination reveals that there is a pretty good relationship between future enrollments and past web leads, but the relationship is not strong enough to inform near-term events; ie, it’s predictive in a general sense, but Data Robot is still grabbing that feature and deriving a 6 month and 12 month median or mean before using it as a variable in its model[[1]](#footnote-1). By contrast, a simple 6 month mean on enrollments proves to be an extremely accurate predictor up to about the 200-250 enrollments per TP per month range (Chart 5)[[2]](#footnote-2). This makes intuitive sense, and further supports Monica’s original method of quarterly forecasting for TPs by taking the average of the 6 months prior. However, this one measurement can be improved by adding in additional derived enrollment features and lags.

Since the best Data Robot model we used was a time series using Territory Partner as a series identifier, we can conclude that the most relevant data is just historical enrollments at the TP level. However, this bakes seasonality into itself by lagging enrollments and weighting those lagged variables by different amounts. We talked to an engineer at Data Robot why the model doesn’t use any of the other features, and after making some initial improvements after our first phone call, it became fairly evident that put simply, it’s because our data set isn’t very predictive of enrollments, and really the best we can do at the TP level is to look at historical performance. These results are disappointing to anyone who wanted the final model to attach a definitive dollar amount to programs like DTC and TPP, but I think the final model successfully accomplishes what it set out to do: find the best relationship between data that will be known at time of prediction, and the target variable. This shows that these models simply don’t find TPP or DTC to be a good predictor of future enrollments. Trupanion Express is almost as bad, and web leads are only marginally better. An improved quarterly model could potentially be built in Data Robot that utilizes web leads and certs, but likely it wouldn’t be much better than what we currently have[[3]](#footnote-3).

Compare these to our seasonally-adjusted linear model, where we simply take 26 months’ worth of preceding data to build the equation below:

yij = εj(β0i + β1iX)

Where yij represents enrollments for TP i in month j

And εj represents the seasonality term in month j

And β0i represents the y intercept of a linear model of Y regressed on X for TP i

And β1i represents the derivative of enrollments with respect to month for TP i

And X represents months since the beginning of 2016

(Chart 6)

This linear model does well in predicting August – October enrollments by TP, with an RMSE only slightly above that of the XGBoost time series model. That’s a very good RMSE, and corresponds to an MAE of 17.4 if you exclude the NULL TPs, for which it did a particularly poor job predicting[[4]](#footnote-4). The time series model does about the same, which immediately brings us back to the opening question: does this only mildly adjusted line on a page do as good a job as the best model DataRobot can train? Not likely, no.

Why did we build the β1i coefficients off of 26 months of data rather than 24? Or 30? Or any other number? I think it’s a combination of we cheated and it got lucky. We chose 26 months because we found a stretch of data over which most TPs exhibited the most linear growth, and that oddly linear trend happened to continue past the end of the training set. What happens, for example, if we were to take column D of this file[[5]](#footnote-5) on the [Summaries] tab and ctrl + H for $M to $H? Then the naïve model’s RMSE jumps to 39.9, a fully 53.3% higher RMSE than the XGBoost time series model. Experimenting with different time spans for that β1i coefficient returns RMSEs that are anywhere between 5% better than the XGBoost to well over 50% worse. The take away is that it is unstable. The XGBoost time series model maintains a near 25 RMSE no matter what subset of the data it is trained on. So no, a straight edge does not make a good data scientist. However, we no longer own the Time Series feature of Data Robot, so what should we do next quarter when it comes time to add in our own projections for TPs? I think we have two main options: use the linear model or use a non-time series version of the XGBoost BP 32 model trained on months as categorical dummy variables. From my perspective, we should try both for some time, and keep score of what actually does a better job when the future is truly unknown, both to the model and the people building the model. I would bet that the XGBoost model maintains it’s relatively low RMSE, while the seasonally-adjusted linear model proves more volatile. Only the future will tell.

References

Chart 1

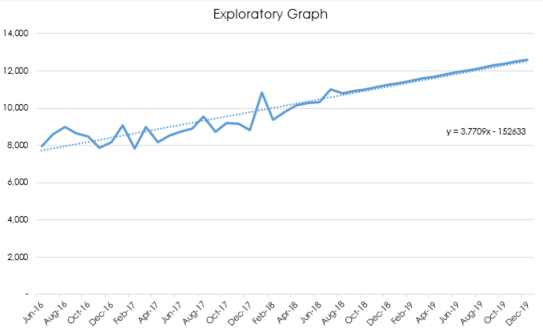


Chart 1 displays the pure linear forecast we first built back in Q3.

Chart 2

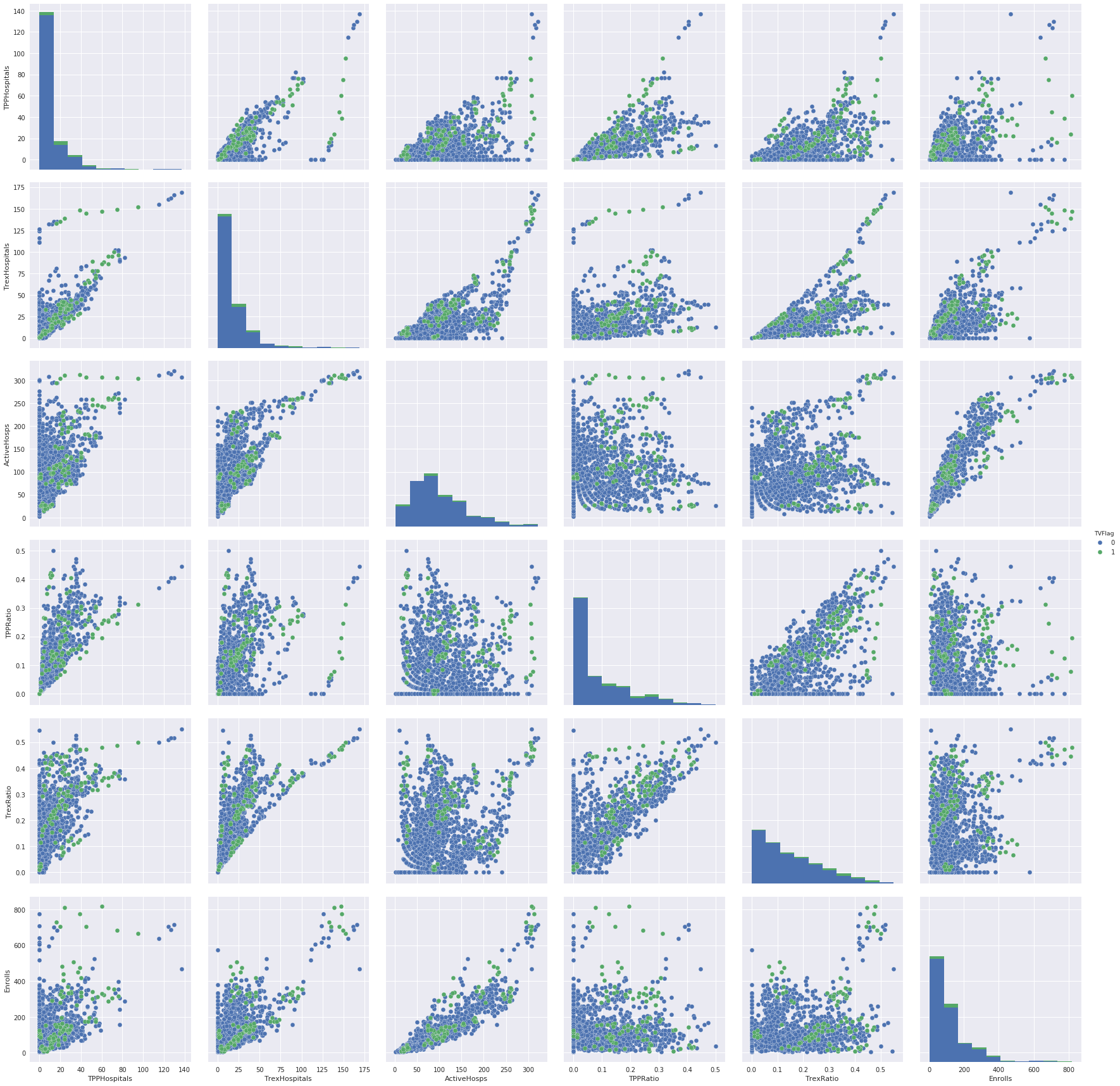


Chart 2 shows a bunch of sea-born pair plots with all TV month-TP combinations colored green. Enrollments are on the y-axis on the bottom most row.

Chart 3

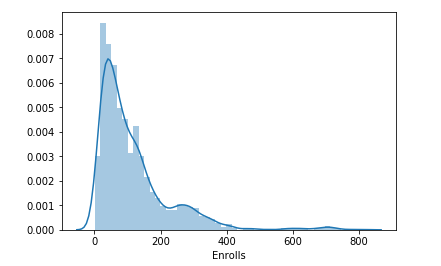


Chart 3 shows the frequency of enrollment counts for month-TP combinations over the last 60 or so months.

Chart 4

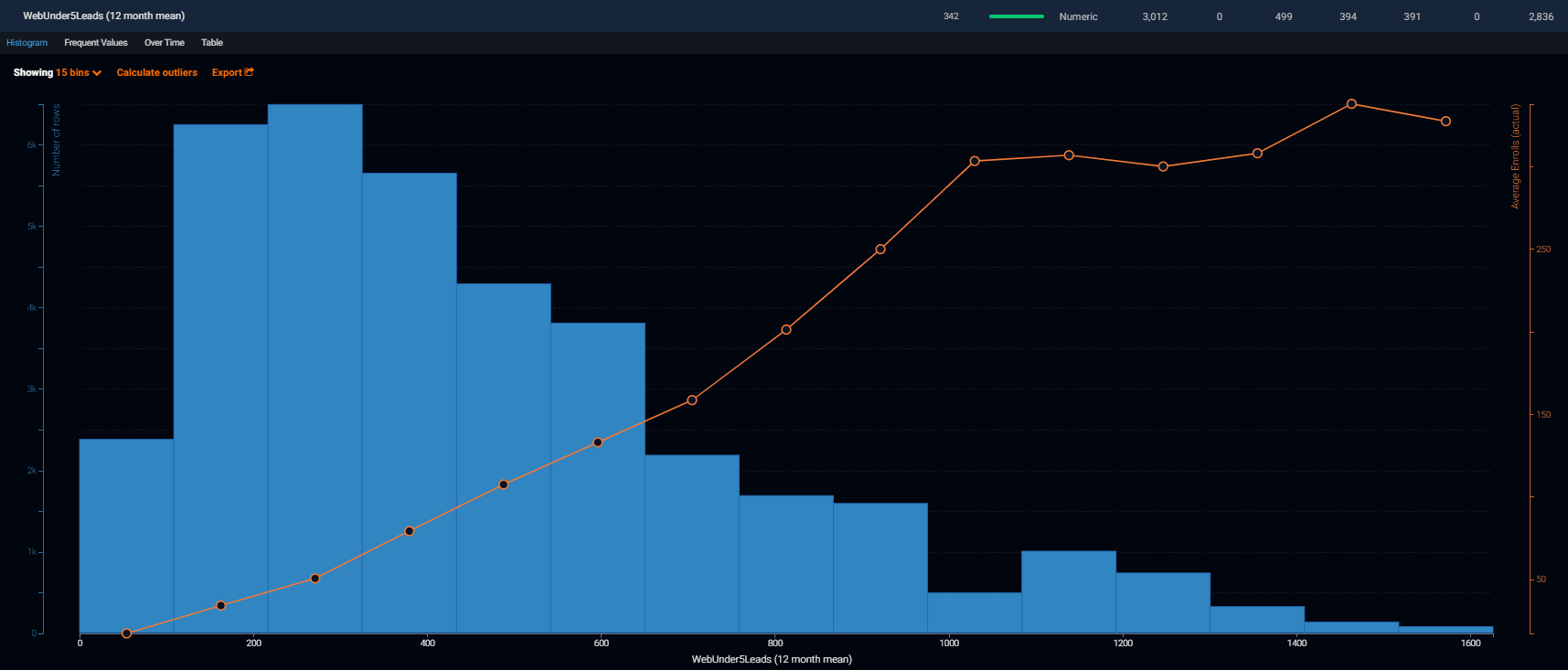


Chart 4 shows the frequency of web leads under 5 in month-TP combinations, but also includes on the second y-axis the actual average enrollments. A perfect orange line represents a 1:1 predictor variable.

Chart 5

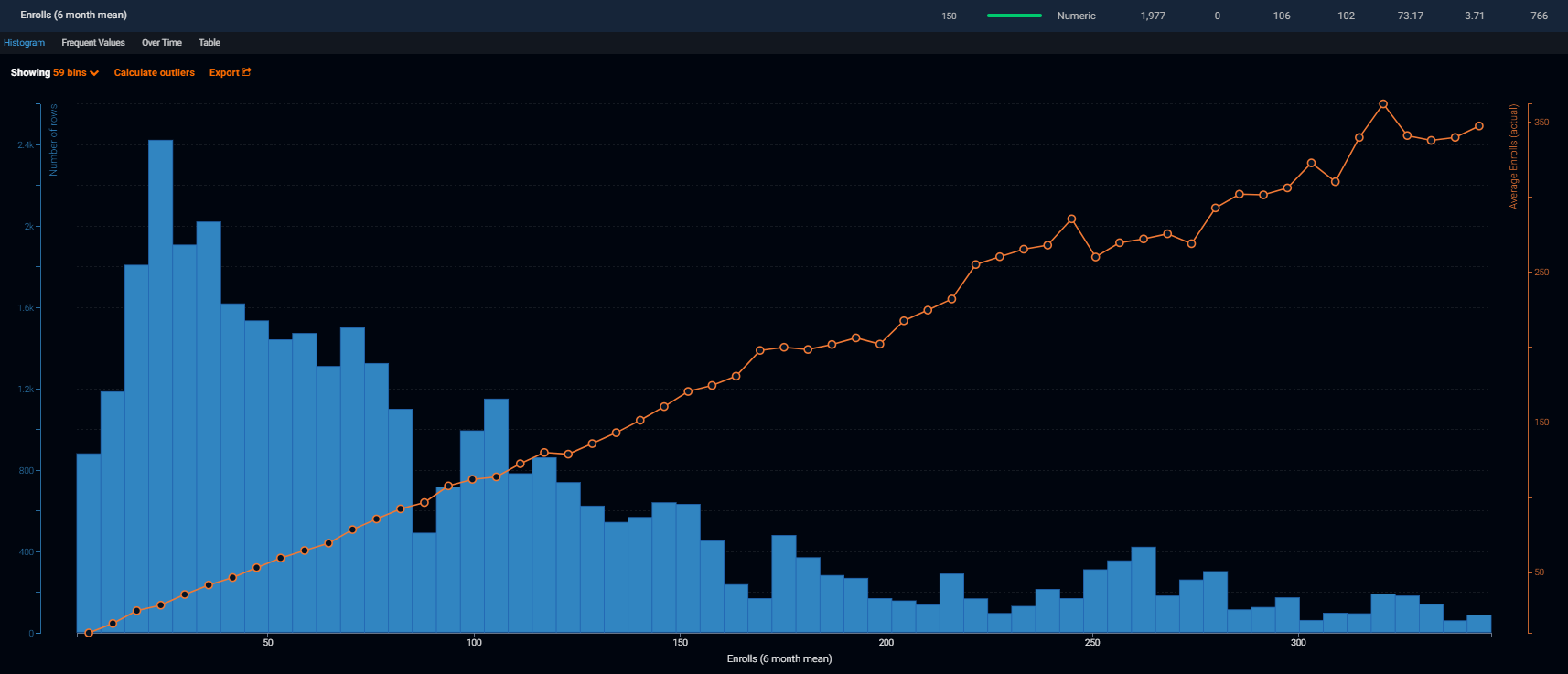


Chart 5 is similar to 4, but shows a lagged 6 month mean enrollment as the predictor rather than lagged 12 month mean web leads under 5. You can see the orange line is nearly linear, especially for lower enrollment count month-TP combinations.

Chart 6

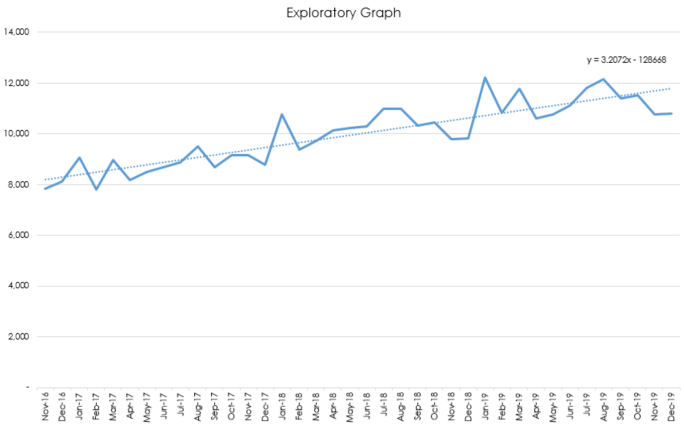


Chart 6 displays the Chart 1 aggregate data with our seasonal adjustment factor multiplied into all the predictions as εj in the equation on page 3.

1. You’ll notice in Chart 4 that this variable has a much more predictive effect before it reaches about 1,100 web leads under 5. After that point, it tends to level out. Could be worth exploring the relationship between web leads and enrollments, because these charts suggest a diminishing marginal return after about 1,300 total web leads, and almost no relationship after 1,600 per territory per month. Of course, this is sparse data, because there are likely just a few territories in that range. [↑](#footnote-ref-1)
2. The more linear the orange line in Charts 4 and 5, the more accurate the predictor. This one for example, is extremely reliable until the data gets sparser. You’ll notice that TPs like Karen Belanger and Michelle Rosen are not even included in this chart. [↑](#footnote-ref-2)
3. In addition, we don’t own this feature in Data Robot anymore, as our free trial expired towards the end of our experimenting, so it’s even less likely to outperform now. [↑](#footnote-ref-3)
4. The XGBoost model does a comparatively good job at outlier prediction, so the errors on NULL TP predictions are much lower for the XGBoost time series model. [↑](#footnote-ref-4)
5. [A:\Projects\SalesMarketing\Enrollment Forecasting\Enrollment TP Forecast (2018-11-28) Seasonality Backtest.xlsx](file:///A:\Projects\SalesMarketing\Enrollment%20Forecasting\Enrollment%20TP%20Forecast%20(2018-11-28)%20Seasonality%20Backtest.xlsx) [↑](#footnote-ref-5)