Bonterra Take-Home Assessment – Overview Write-up

David Li / [davidjli@umich.edu](mailto:davidjli@umich.edu) / 248-918-8782

Opening Remarks / Problem Statement Identification

Our Likely Model Type / Use Case would be Time Series Forecasting, due to the rolling temporal nature of the dataset among Organizations and their gift amounts and specific dates. Our Target Column of Interest is “gift” indicating the gift amount, and the variable we are interested in developing a model to predict future gift sizes for current donors.

Exploratory Data Analysis:

- The majority of OrgIDs all come from one Org, 10109128 at ~40000 rows. The next most frequent org is 17863152, at only ~1500 rows. So the data is heavily representative of a lot of data for this frequent Org.

- The DonorID frequency graph shows multiple ~12 DonorIDs all with high frequencies. Beyond these 12, the frequencies are spotted and varies across many DonorIDs

- The gift column shows high frequencies of specific amounts, namely the usual threshold generic donation amounts (i.e. 10, 20, 50, 100, 25, etc.) Since these values have much more frequency than the many other values, they are probably pre-select / default options but also allowing for custom amounts.

- The statistical summary of the gift column shows the majority of results (i.e most percentile thresholds) in the 1-100 range, indicating the majority of the spread of the results are reasonably located here. However, the mean and standard deviation are exceptionally high, likely influenced by the maximum value of 350,000.

- With giftdate, we prior mentioned the large range of a very early and a future date as the minimum and maximum date ranges of the data. Additionally, generating frequencies for the dates shows that there is one date that has an astounding high frequency seen, 2020-04-07 at ~18000 times. Again, this is exceptionally high given the frequency of many other dates is at most 60 times.

Modelling / Future Extensions & Steps:

Our Preliminary Model of Choice will be ARIMA, which stands for AutoRegressive Integrated Moving Averages – it is a common “first-choice” model for Time Series applications.

RMSE is Root Mean Squared Error, which is a relatively common metric used to assess the overall distance (i.e. "accuracy") between our predicted data points and actual data points. The lower the number this is, the better performant the model is.

In order to quickly determine an ideal set of parameters (with the limited time for this assessment) to use for our ARIMA model, we can implement a simple grid search to try a few different combinations of the required parameters for the model and choose the parameter set that yields the best model performance.

Our chosen Model Parameter set of (p = 0, d = 1, q = 1) showed an RMSE of 661.789, which is not amazing but is a good baseline for future extensions and improvements to the model. An in-depth explanation of these parameters can be found on the web with ARIMA documentation.

Areas of Improvement:

- Our grid search is a simplified, "brute force" approach to determining the p,d,q hyperparameters needed for the ARIMA model. There are statistical processes and methods to assist in getting a narrower sense of what the hyperparameters should be or approximately be.

- We had only implemented the model for one medium-sized data of OrgID. Not having models for the other various OrgIDs could miss important gaps or anomalous patterns that need to be corrected for.

- There are many other Time Series models that may outperform our ARIMA model if we do more extensive exploratory analysis (i.e. if seasonality is a strong factor, the Seasonal ARIMA model may be a more obvious choice).

- There may be more sensible groupings that are sufficient to apply here for multiple benefits (i.e. aggregating gift amounts in a way to reduce the data points and improve model speed and possibly performance).

- We noticed there were some outlier gift amounts. We may want to resolve/justify these data points and decide if they should be handled/removed for better model performance.

Potential Next Steps:

- Do deeper data plotting / analysis to investigate and extract more data patterns that can further inform our modelling setup (i.e. if seasonality is a significant component)

- Explore/Try other Time Series Models, such as SARIMAX, LSTM, Prophet, etc.

- Retry the Modelling with different aggregations/groupings of the data.

- Extend to other OrgIDs and extract learnings/insights to help us better model in future iterations.