Project Abstract:

STAT480 – Statistical Consulting

Description:

I was assigned a partner to work on an unstructured project for a real industry client. The data was massive which included 15 KPIs, and 5 economic indicators (including the client’s business revenue) with some dating back to January 1900. The goal was to go nuts with the data, and report back with any interesting findings, particularly which KPIs and corresponding lags did well for predicting the client’s revenue. We decided to use lags and leads up to two years out, and ideally at least one year out.

Throughout our time working on this project, we had three presentations; one for our analysis plan (to colleagues), one for our results (to colleagues), and one to the client with results simplified for business action. We also had a study report to document and back up our analyses. It should also be noted that there were also two concurrent projects, one other for this class, and the Waste Data Project for another class.

Analysis methodologies:

This project was largely time series analysis. I worked on a prescriptive regression with ARIMA errors model. This topic was barely touched on in a previous class, but this was the clear choice to use based on the extensive research I performed. My partner cleaned the data while I was still researching, and we worked on our separate sections.

Programming:

I used JMP and R for EDA, and programmed in R. The data was in a .xlsx file.

Transferable Skills:

Time Management, Teamwork, Leadership, Critical Thinking, Research, Programming

Post Project Debrief:

This was an intense project, and my first experience with an actual client. Halfway through the semester we had our first schedule meeting to kick off the project, of which we had plenty of rehearsal for asking the right questions and to fully understand the client’s problem. It turned out that our client was very prepared, as it was more of a presentation on their end; but I was able to ask pertinent questions at the end. We only contacted them to set their expectations on when we would present our findings, even though it might’ve benefited us to ask follow up questions about the project get feedback on our process throughout our time working on their problem.

As for the analysis itself, it was brutal. I didn’t know what direction to go at first, and I had issues reaching out to my partner and professor for direction, feedback, or help in general. I attribute this inaction to being a detriment from working remote, as the class met once a week. The sunk-cost fallacy also made it difficult to know when to ask for help instead of continuing to research approaches for the project.

Once I settled on the Regression with ARIMA errors direction, it was a bit easier, but following internet guides was difficult. There were so many different authors giving different suggestions, and some suggestions that just didn’t align with my direction. Things picked up speed once my professor realize how much I was struggling, and with me reviewing my past class that briefly mentioned my approach.

Then there was the issue with programming as well. Once I started picking up steam, I was also running short on time, which was adding stress and clouding my judgement. I went on many tangents and directions that were completely unnecessary. There was also a lot to brush up on considering the basis of my model was about Regression, and my project in that class was also rough (see Saratoga Homes Project). I’d surely benefit from having consistent hours to work on a team and bounce ideas back and forth.

When it came to fitting the model itself, it was such a gruelingly long and manual process. I’m still pretty sure it couldn’t be automated, but I’d love to figure out how. The process itself was to pretty much prewhiten the data to the response and take the suggested lags from this and fit the model. I would then use ARIMA errors to make sure the model assumptions were in check. It was surely an issue of scale, as there were many KPIs that had high collinearity and many lags across each one. I believe I was fitting, swapping, and removing from a group of 90 IVs. The end result was also questionably worth it. The model fit the error assumptions for ARIMA, but as my professor said: “it’s pretty fragile if including insignificant variables are necessary to meet the assumptions.” The client’s revenue was also already being modelled with a MAPE of about .95, so it was all an effort of excessive perfection. Keeping in mind the end goal though, the emphasis had to be placed on the KPIs, though, so this was really the only route.

When it came to the final weeks of the project, we were struggling with time, and I can definitely say I wasn’t procrastinating at the beginning, just struggling. We were able to get a presentation for class ready in time, but there was clearly need for polish. Fortunately, for the client presentation we had enough feedback to make a polish presentation, of which the client was happy with.

If I were to start this project over, I’d definitely adopt a more open conversation with my professor about what direction to go in, and use office hours as much as I could. There really is no room for pride when it all comes down to the client getting what they need. Analysis wise, I hope to learn better ways to do this, preferably with automation.

I did manage to contact my partner afterwards to ask for feedback, and I’m glad to say they didn’t have much to add, just to be more confident in myself. They also mentioned I was “friendly and approachable in and out of class”. I’m glad I came off that way, because they were also a pleasure to work with.