**Citadel European Regional Datathon – Personal Summary**

**Summary:**

Undertaking this project was a fantastic experience for a collection of reasons. This was my first experience conducting analysis on real-world ‘Big Data’. This threw up many challenges for me which needed to be solved in a short time-window (detailed in a separate section below). It was exciting and rewarding to search for solutions to these problems. Another rewarding aspect of the project was collaborating with my three teammates: Ben, Wilfried, and Armando. I learned a huge amount from each teammate, particularly Armando, who worked with me on the time series modelling section. The subject matter was very topical; user privacy is of major concern in the digital age. click through rate (CTR) provides an interesting metric related to this issue, since if CTR is rising, perhaps it is because websites have a better understanding of their users.

**Skills Developed:**

* Modelling variance clustering of data.
* Google Colab for collaborative coding projects.
* Presentation skills. I was very impressed by the standard of work done by my teammates in presenting the report and infographic in a neat and informative way.
* Natural Language Processing. I had no previous experience in this area and learned a lot from the methods Wilfried and Ben employed to use headline sentiment as a predictor of CTR.

**Challenges Encountered and Solutions Used:**

* Blocking the data such that comparisons were not distorted by external sources of variance. We developed a process to group distinct headlines by image. One added complexity was the variety of test structures in the dataset. Some tests included multiple images with one headline, others included multiple headlines with one image, as well as tests with multiple images and headlines. We had to locate tests which had more than one distinct headline for a given image, so that a comparison could be made. Where tests had multiple images, we split the tests by image for the purposes of analysis.
* Coefficient estimation of ARIMA + GARCH model. This was perhaps to most difficult problem to solve, for two reasons. Python provides modules which allow GARCH coefficient estimation, but not ARIMA + GARCH estimation (to the best of my knowledge). Although the model is additive, the coefficients must be estimated simultaneously to get consistent results. A possible solution would be to calculate the likelihood function by hand and then use a maximisation algorithm. The difficulty was the time scale I had to produce the results. The coefficient estimation was one of the last pieces of analysis to be conducted and I didn’t think I would be able to execute the solution above. Instead, I decided to switched programming languages. I imported the cleaned data set into R-studio, which has a package capable of performing the required coefficient estimation.
* Technical issues. I had difficulty installing the requisite packages to use for parameter estimation of the time series model in R-Studio. After completing the project (and a painstaking search through stack exchange) I identified that the issue was caused by duplicated files in my r directory. Again, the section conducted in r towards the end of the project meant time was of the essence. I quickly identified that the issue was not straightforward and would take too long to solve. A simple and quick solution was to conduct the analysis on R cloud, since it did not interact with my own file system.

**Shortcomings and Further Study:**

* There was potential to examine and partition the variance of headlines and packages as a nested factorial design. I had not studied this method at the time of the competition.
* I would like to investigate the predictive power of the model chosen for the CTR time series. In addition, further diagnosis of residuals could be conducted to determine is more information could be extracted from the right skewed white noise.
* A mistake was made in the model statement of an ARIMA(0,1,1) + GARCH(1,1) in the executive summary. I added the conditional volatility to the ARIMA part of the model, rather than subbing it in for the white noise terms. The actual process of parameter choice and estimation was unaffected by this mistake.
* In general, the subject matter was highly interesting. In our project, we only scratched the surface of what the Upworthy research archive has to offer. I would love conduct further analysis of the data, and am particularly interested in incorporating google analytics data (such as session length) into the analysis, as I believe this may add a further level of insight to the patterns we saw.