**Problem:**

Predicting revenue for 3, 6, 9, 12, and 15 months is an important topic for the company I’m employed by. At present, the machine learning (ML) models built address this topic from a univariate perspective, using only revenue as a feature to predict future revenue.

The impact of this approach is that the financial analysts that create predictions in their respective markets across the globe produce forecasts that most of the time prove to be more accurate forecasts than that of the models.

# Benefits:

While there is no expectation that a machine learning (ML) model will replace the need for financial analysts, there is a desire for the ML models to be accurate enough to serve as a reliable benchmark that analysts reference when they create their forecasts.

# Objective:

This project proposes to research a number of features (around 50) that could potentially be used to create a multi-variate model that predicts future revenue better than the current univariate models do.

These features were identified by talking with a number of financial analysts across my company, in which they were asked what they take into account when they derive their forecasts.

**Data:**

The data available for this study is from 2012 to 2021 and covers over 50 countries that my company has operations in. The label to predict (revenue) is tracked at a monthly level so total instances available for each country is around 100 (8.5 years \* 12 months = 100 months of data).

*Data Sensitivity:*

To avoid the need for a potential Non-Disclosure Agreement (NDA), I propose the following…

1. Will not mention the name of my company in the study

2. Will reference the countries as Country1… Country50

3. Will reference the features as Feature1… Feature50

**Approach:**

This is a time series problem, in which the objective is to create a multi-variate model with a better MAPE than the current univariate models…

*Data Wrangling:*

1. This data has some missing values and will require limited transformations to bring the data to a state where multi-variate models can be created.

*Data Storytelling:*

1. Analyze the features...
   1. look for correlation between each feature (or groups of features) and the label
   2. look for correlation between features
   3. evaluate if the data has a linear, gaussian, or other distribution type
2. Employ machine learning algorithms, such as Random Forest and Granger Causality, to identify which features are best to carry forward into modeling. Forward stepwise regression will most likely be employed as well.
3. Various visualizations of the data (boxplots, residplots, histograms, etc...)

*Machine Learning:*

1. Algorithms to employ will be determined based on the Data Storytelling. ARIMA and Exponential Smoothing are anticipated to be the algorithms, but others might be identified as well.
2. Requirements stated in the course syllabus will be applied to meet the expectations for this course.