**University of Columbia Master of Data Science Capstone Project**

**(Addendum)**

**PROJECT TITLE\***

Evaluating the Attractiveness of a Country for Business Investment using Worlds Bank Indicators

***Note:*** *This document is for reference to the main proposal. To get a comprehensive understanding, the students first need to be introduced to the main proposal.*

*During the course of the project, while discovering the minute details, some modification to the original project proposal may be necessary.*

**Background material:**

As a background study, student can refer manufacturing and service industry related literature from journal articles, and arxiv articles. These articles can be found/downloaded from Google Scholar or university library online portal. Three ready references are given below.

Rocha, I. L. (2018). Manufacturing as driver of economic growth. *PSL Quarterly Review*, *71*(285), 103-138.

Heshmati, A. (2003). Productivity growth, efficiency and outsourcing in manufacturing and service industries. *Journal of economic surveys*, *17*(1), 79-112.

Buera, F. J., & Kaboski, J. P. (2012). The rise of the service economy. *American Economic Review*, *102*(6), 2540-69.

**Snippets of sample data:**

Data can be downloaded from <https://data.worldbank.org/>

Full list of URLs is mentioned in the main document. A snippet of data as excel file is attached in the email. Please note the World Bank maintains data in isolation in many separate files. Students need to download the data sets per indicator/variable wise as mentioned in table 1 and table 2 of the main document and to prepare a consolidated one. Some data may be found missing. In such cases, the students need to impute the missing data programmatically.

**Students' desirable skills/background:**

Familiarity to Statistics, Programming language Python/R are required.

**Clarifications on the goal:**

Goal of the Project:

1. To formulate a Machine Learning model for assessing the best investment destination.
2. To understand the data cleansing process (e.g., missing values, normalization etc.)
3. To evaluate the error metrics of multiple models and select the best performing model

This project will utilize Multivariate Time Series (MTS) models**.** MTS can be done using Vector Auto Regressive (VAR) technique or Long Short-Term Memory (LSTM) method. The students also need to do data cleansing (standard Python libraries are available for missing value imputation).

This project will assess the “Manufacturing value added as % of GDP” to measure a country’s attractiveness for business investment in Manufacturing, and “Service value added as % of GDP” to measure a country’s attractiveness for business investment in Service.

**Methods:**

Time series analysis can be done by below mentioned two methods.

* Multi Variate Time Series Model: One file attached in the email
* Long Short-Term Memory (LSTM): Two files attached in the email

***Note: Plethora of sources available in internet to understand the above concepts.***

**Deliverables:**

1. The visualization report of “Manufacturing value added as % of GDP” and “Service value added as % of GDP” trend.
2. Comparative assessment report detailing the predictor variables’ behavior in different countries, influence on the target variables, and thereby indicating the best investment destination with reasons.