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

**PROJECT TITLE\***

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

**MOTIVATION, BACKGROUND AND OVERVIEW: Please state briefly what is the problem that the project tackles. The projects need to be focused on a data science problem that is engaging, relevant, clearly defined and of the right scope for a semester. When assessing the proposals we will be looking for a diverse set of problems that address different topics and technical requirements that our students can address. The evaluation criteria will include: Is this a data-science project? Can our students learn about a data science application in the real world? Is the proposed research problem important and can potentially have a big impact? Will our students be excited about it? Please provide your project description having these criteria in mind.**

Business investment is a major pillar of economic growth. New business investment decision is fraught with many risks and other factors. New investment depends on the factors pertinent to policies and regulations, technology, and risk factors (policy risk, economic risk etc.). New investment decision also depends on the prospect and growth potential of the respective industry sector/vertical. In this capstone project we will focus on two major sectors: Manufacturing and service.

This capstone project will evaluate a country’s attractiveness in business investment in:

1. Manufacturing industry
2. Service industry

Here the attractiveness of a country for business investment in Manufacturing and service is represented as *“Manufacturing value added as % of GDP”* and *“Services value added % of GDP”* respectively.

**Evaluation of a country’s attractiveness in Manufacturing**

A country's attractiveness for business investment in Manufacturing depends on many factors. The World Bank maintains a large repository of relevant data that influence Manufacturing. Table 1 lists the possible factors that influence Manufacturing. In this table first, 9 predictor variables are listed and then the target variable is mentioned. For each factor, index as per variable naming practice in computer programming environment, short description, unit of measurement are mentioned.

**Table 1.** Factors for attractiveness in Manufacturing

Predictor variables

| **Sl. No** | **Index** | **Short Description** | **Unit** |
| --- | --- | --- | --- |
| 1 | Air\_freight\_million\_ton\_km | Air freight in million ton-km measured in metric tons times kilometer traveled. | Million ton-km |
| 2 | Container\_port\_traffic\_TEU | Container port traffic (TEU:20-foot equivalent units) is the shipment of containers from port; inbound and outbound. | Number of containers |
| 3 | Railways\_goods\_trans\_million\_ton-km | Goods transport by railways in million ton measured in metric tons times kilometer traveled. | Million ton-km |
| 4 | Logistic\_performance | It is the quality of the logistic related infrastructure (1 being low and 5 being high). In this index, trade, and transport related infrastructure are evaluated. | Numeric index (1=lowest to 5=highest) |
| 5 | Industry\_value\_added\_current\_USD | Total output in manufacturing, industry, construction, power, water, mining etc. in USD. It is the net output of a sector after adding up all outputs minus intermediate inputs and assets depreciation. | USD |
| 6 | GDP\_per\_capita\_current\_USD | It is the GDP divided by total midyear population of a particular country. | USD |
| 7 | Final\_consumption | Final consumption or total consumption expenditure is the expenditure for household final consumption plus the government final consumption. | USD |
| 8 | Access\_finance | Access to finance is the percentage of firms those use banks to finance investments. | Percentage of firms using bank |
| 9 | New\_business\_density | The number of new limited liability companies registered in a year per 1000 people (ages between 15-64). | Number of business entity per 1000 people per year |
|  | Target Variable |  |  |
| 1 | Manufacturing\_value added\_%\_of\_GDP | Manufacturing value added is the net output of a sector after adding up all outputs minus intermediate inputs. It also ignores the depreciation of goods. | Percentage of GDP in USD |

Please note that “Manufacturing\_value added\_%\_of\_GDP” reflects a country’s attractiveness for business investment in Manufacturing.

**Evaluation of a country’s attractiveness in Service**

A country's attractiveness for business investment in Service depends on many factors. The World Bank maintains a large repository of relevant data that influence Service. Table 2 lists the possible factors that influence Service. In this table first, 11 predictor variables are listed and then the target variable is mentioned.

**Table 2.** Factors for attractiveness in Service

Predictor variables

| **Sl. No** | **Index** | **Short Description** | **Unit** |
| --- | --- | --- | --- |
| 1 | Access\_electricity | Percentage of population with access to electricity. | Percentage of a country's population |
| 2 | Industry\_value\_added\_current\_USD | Total output in manufacturing, industry, construction, power, water, mining etc. in USD. | USD |
| 3 | Employment\_in\_industry\_%\_of\_total\_employment | It is the person’s occupation in an industry to produce goods or services for an earning. The industry sector includes mining and quarrying, manufacturing, construction, and public utilities. | Percentage of total employment |
| 4 | GDP\_per\_capita\_current\_USD | It is the GDP divided by total midyear population of a particular country. | USD |
| 5 | Final\_consumption | Final consumption or total consumption expenditure is the expenditure for household final consumption plus the government final consumption. | USD |
| 6 | Primary\_completion\_rate | The number of enrollments minus the number of people who are repeating in the last grade of the primary school, divided by the population at the last grade. | Percentage |
| 7 | Educational\_attainment | The percentage of people ages 25 and above who completed education in post-secondary non-tertiary level. | Percentage |
| 8 | Individuals\_using\_the\_Internet\_%\_of\_population | Internet users are individuals who have accessed the internet recently using a computer, cell phone, or by any device. | Percentage of population |
| 9 | Poverty\_headcount\_ratio | It is the percentage of the population who are earning less than $1.90 per day at 2011 international price. | Percentage of population |
| 10 | Access\_finance | Access to finance is the percentage of firms those use banks to finance investments. | Percentage of firms using bank |
| 11 | New\_business\_density | The number of new limited liability companies registered in a year per 1000 people (ages between 15-64). | Number of business entity per 1000 people per year |

|  | Target variable |  |  |
| --- | --- | --- | --- |
| 1 | Services\_value\_added\_%\_of\_GDP | It is the value added in wholesale and retail service. Service value added is the net output of a sector after adding up all outputs minus intermediate inputs. | Percentage of GDP in USD | |

Please note that “Service\_value added\_%\_of\_GDP” reflects a country’s attractiveness for business investment in Service.

A multivariate time series forecasting model can be built to find the manufacturing and service trend in a particular country.

**Multivariate Time Series (MTS)**

A Multivariate time series has more than one time-dependent variable. Each variable depends not only on its past values but also has some dependency on other variables. This dependency is used for forecasting future values.

MTS can be done using Variable Auto Regressive (VAR) or Long Short-Term Memory (LSTM).

A country wise comparative assessment can be done; which country is performing better than the others. For example, a comparison of investment environment for next 5 years among 3 countries in Manufacturing and Service sector.

| Multivariate Time series Prediction for next 5 years | | | |
| --- | --- | --- | --- |
| Country Name | Target variable: Industry type (Y) | No of Predictor variables(X) | Remarks |
| Country 1 | Manufacturing\_value added\_%\_of\_GDP | 9 | Refer Table 1 for details about no of predictors |
| Services\_value\_added\_%\_of\_GDP | 11 | Refer Table 2 for details about no of predictor |

| Multivariate Time series Prediction for next 5 years | | | |
| --- | --- | --- | --- |
| Country Name | Target variable: Industry type (Y) | No of Predictor variables(X) | Remarks |
| Country 2 | Manufacturing\_value added\_%\_of\_GDP | 9 | Refer Table 1 for details about no of predictors |
| Services\_value\_added\_%\_of\_GDP | 11 | Refer Table 2 for details about no of predictor |

| Multivariate Time series Prediction for next 5 years | | | |
| --- | --- | --- | --- |
| Country Name | Target variable: Industry type (Y) | No of Predictor variables(X) | Remarks |
| Country 3 | Manufacturing\_value added\_%\_of\_GDP | 9 | Refer Table 1 for details about no of predictors |
| Services\_value\_added\_%\_of\_GDP | 11 | Refer Table 2 for details about no of predictor |

Country 1, Country 2, and Country 3 can be compared against each other for prudent investment decision. For example, in this capstone project, three countries can be:

* USA
* China
* India

We can compare the attractiveness of these three countries for business investment in Manufacturing, and Service. We can also increase the number of countries for more holistic comparative assessment. We can also increase/decrease the number of predictors after framing the actual business problem.

**Performance Evaluation Metrices**

To evaluate the performance of forecasting modeling, mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), Coefficient of determination (COD). These metrices can be stated as follows:

where, and represent the predicted and actual figures; N denotes the total sample size, and implies an average of actual observations. In addition to the above error metrices, Mean Absolute Percentage Error(MAPE), and Symmetric Mean Absolute Percentage Error(sMAPE) can also be calculated.

We suggest carrying out the entire exercise using Python based various software libraries in interactive Jupyter notebook. However, R can also be used.

All the data are available with the World Bank.

**The dataset(s) can be public or private. Please keep in mind that the students will need to list the project on their CV and the report will be public. All datasets must be submitted by Friday, July 23, 2021 for Fall 2021.**

**DATASET: Please provide a detailed description of the type of data that is required to address the problem. For example, is this social media data, medical data, financial data, etc? What is the size of the data. Will the organization provide the majority of the data or is the data accessible via other avenues/ sources? How much of the data is available? Do the students need to gather data? In assessing the projects, the availability and type of data will play an important role. Please consider these evaluation criteria for data requirements when submitting the proposal: Is the data set clearly defined? Is the data set complex and big enough for creating learning opportunities? Is the data set ready? (availability, need for processing) Does the data require extensive computing resources (if yes, can the affiliates provide resource/funding?)**

Dataset is public and can be downloaded from the below URLs.

[World Bank Open Data | Data](https://data.worldbank.org/)

<https://data.worldbank.org/>

Specific source/URL are given below

Dataset regarding **Evaluation of a country’s attractiveness in Manufacturing**

Predictor variables (please refer Table 1)

| **Sl. No** | **Index** | **Source/URL** |
| --- | --- | --- |
| 1 | Air\_freight\_million\_ton\_km | https://data.worldbank.org/indicator/IS.AIR.GOOD.MT.K1 |
| 2 | Container\_port\_traffic\_TEU | https://data.worldbank.org/indicator/IS.SHP.GOOD.TU |
| 3 | Railways\_goods\_trans\_million\_ton-km | https://data.worldbank.org/indicator/IS.RRS.GOOD.MT.K6 |
| 4 | Logistic\_performance | https://data.worldbank.org/indicator/LP.LPI.OVRL.XQ |
| 5 | Industry\_value\_added\_current\_USD | https://data.worldbank.org/indicator/NV.IND.TOTL.CD |
| 6 | GDP\_per\_capita\_current\_USD | https://data.worldbank.org/indicator/NY.GDP.PCAP.CD |
| 7 | Final\_consumption | https://data.worldbank.org/indicator/NE.CON.TOTL.CD |
| 8 | Access\_finance | https://data.worldbank.org/indicator/IC.FRM.BNKS.ZS |
| 9 | New\_business\_density | https://data.worldbank.org/indicator/IC.BUS.NDNS.ZS |
|  | Target Variable |  |
| 1 | Manufacturing\_value added\_%\_of\_GDP | https://data.worldbank.org/indicator/NV.IND.MANF.ZS |

Dataset regarding **Evaluation of a country’s attractiveness in Service**

Predictor variables (please refer Table 2)

| **Sl. No** | **Index** | **Short Description** |
| --- | --- | --- |
| 1 | Access\_electricity | https://data.worldbank.org/indicator/EG.ELC.ACCS.ZS |
| 2 | Industry\_value\_added\_current\_USD | https://data.worldbank.org/indicator/NV.IND.TOTL.CD |
| 3 | Employment\_in\_industry\_%\_of\_total\_employment | https://data.worldbank.org/indicator/SL.IND.EMPL.ZS |
| 4 | GDP\_per\_capita\_current\_USD | https://data.worldbank.org/indicator/NY.GDP.PCAP.CD |
| 5 | Final\_consumption | https://data.worldbank.org/indicator/NE.CON.TOTL.CD |
| 6 | Primary\_completion\_rate | https://data.worldbank.org/indicator/SE.PRM.CMPT.ZS |
| 7 | Educational\_attainment | https://data.worldbank.org/indicator/SE.SEC.CUAT.LO.ZS |
| 8 | Individuals\_using\_the\_Internet\_%\_of\_population | https://data.worldbank.org/indicator/IT.NET.USER.ZS |
| 9 | Poverty\_headcount\_ratio | https://data.worldbank.org/indicator/SI.POV.DDAY |
| 10 | Access\_finance | https://data.worldbank.org/indicator/IC.FRM.BNKS.ZS |
| 11 | New\_business\_density | https://data.worldbank.org/indicator/IC.BUS.NDNS.ZS |

|  | Target variable |  |
| --- | --- | --- |
| 1 | Services\_value\_added\_%\_of\_GDP | https://data.worldbank.org/indicator/NV.SRV.TOTL.ZS |

**DATA TYPE: Public data is data made available by a third party and is available to the general public. Novel data is data that has been recently published by the proposer or will be made public as part of this project. Private data is data that cannot be made available after the project ended. Please check all that apply.\***

[x] Uses Public Data

[ ] Uses Private Data

[ ] Uses Novel data

**HOW WILL THE DATASET BE MADE AVAILABLE? For example: CSV/XLS file, remote database, raw images or documents, REST endpoint, etc.**

**Type of Data\***

[] Graphs, Networks

[] Text Data

[] Audio/Image/Video

[] Geospatial

[x] Time Series

[x] Numerical data

**Work Requirements (Check all that apply)\***

[ x ] R

[ ] Scraping (including API)

[x] Database (e.g. SQL)

[x] Preprocessing

[x] Visualization

[ ] Option 6

[ x ] App/tool building

[x ] Other: Machine Learning/AI model ( Programming Language Python)

**RESEARCH GOALS**

**Project Topic\***

[ x] Social Good

[ ] Biomedical

[ ] Physical Sciences (chemistry, climate, etc.)

[ ] Consumer

[ ] Social Media

[ x ] Finance and Economics

**Data Science Areas in this project?\***

[ x ] Statistics

[ x] Causal inference

[ ] Deep Learning

[ ] Reinforcement Learning

[x] Algorithms

**OUTCOME\***

[x] Model

[x] Report

[x] Paper

[ ] Software

[ ] Other

**SKILLS: What skills should students expect to learn through their project? Check all that apply.\***

[x] Project planning and scoping

[x] Data acquisition and scraping

[ ] Data versioning and management

[x] Data cleaning

[x] Combining data sources

[x] Exploratory data analysis and visualization

[x] Supervised modeling

[ ] Unsupervised modeling

[ ] Establishing evaluation metrics

[ ] Working with text data

[ ] Working with image data

[x] Working with time series data

[x] Working with tabular data

[ ] Working with geospatial data

**What is the goal of this project? What questions do you want answered? What has been done already to achieve this goal?\***

Prudent business investment decisions are essential for a corporations’ growth. Business investment has many implications on economic growth, and socio-economic development for a country. With this project driven by data science, we can find out what factors drive the business investment decision in a country. A comparative assessment among multiple countries can be essential for a corporation to decide where to invest more prudently as a next move.

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 metrices of multiple models and select the best performing model

Outcome/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.

Some of the literature which can be helpful are listed below.

Berghout, E., & Renkema, T. J. (2001). Methodologies for investment evaluation: a review and assessment. *Information Technology evaluation methods and management*, 78-97.

Bondarenko, S., Shlafman, N., Kuprina, N., Kalaman, O., Moravska, O., & Tsurkan, N. (2021). Planning, Accounting and Control as Risk Management Tools for Small Business Investment Projects. *Emerging Science Journal*, *5*(5), 650-666.

De Clercq, D., Meuleman, M., & Wright, M. (2012). A cross-country investigation of micro-angel investment activity: The roles of new business opportunities and institutions. *International Business Review*, *21*(2), 117-129.

Oliner, S., Rudebusch, G., & Sichel, D. (1995). New and old models of business investment: a comparison of forecasting performance. *Journal of Money, Credit and Banking*, *27*(3), 806-826.

World Bank. (2020). *Global Investment Competitiveness Report 2019/2020: Rebuilding Investor Confidence in Times of Uncertainty*. The World Bank.

**What are the ethical considerations?\***

The project will use open-source public dataset.

**Are there any ethical concerns about the proposed project such as privacy, transparency, and bias that we should pay special attention to?\***

No.

**What is the relevant background needed for the project? In order to make sure we build the right team of students for each project, please provide information on the relevant background information that someone working on the project should have. What technical skills they should have and/or relevant literature (please provide citations) or tools (please provide links) they will need to know or be able to learn.\***

The student should have skills or willing to quickly learn time series modeling, predictive models, forecasting and decision science, coding skills in Python/R, database knowledge and paper writing and good presentation skills.

**What are the quantitative and/or qualitative metrics that can be used to judge the successful completion of the capstone project?\***

Some of the metrics would be - understanding and clear definition of problem, approach to solve the problem, steps undertaken with intermediate results like accurate forecasts, algorithms used, impact of the solution, report and presentation, quality of the paper, quality of the software/code.

**Are international students on a F1 or J1 student visa eligible to work on this project?\***

[x] Yes

[ ] No

**Are you willing and/or able to work with students who are currently physically in another country (if time zone is not an issue?)\***

[x] Yes

[ ] No

[ ] Maybe

**Are you willing to work with two teams of students?\***

[x ] Yes

[ ] No

[] Maybe

**If Yes, please indicate how the project may be appropriate to engage two teams.**

[ x ] Option 1

**An important aspect of the capstone project is the opportunity for students to work with professionals across different industries or academic research labs. Thus each organization must provide mentorship to the students so that they can receive constant feedback and guidance (while each team will also have a faculty advisor the organization mentor will play a crucial role in guiding the team). Please specify who will work with the students and what are their qualifications or training? What amount of time per week do they intend to devote to working with the project team? Each mentor will also ideally help the DSI faculty advisor assess the success of the project at the end which will translate to a grade for the students.   
  
Each capstone project will be mentored by at least one industry mentor and one faculty mentor, with the industry project proposer(s)/mentor(s) as the primary mentor(s). Industry mentor(s), in addition to monitoring the project progress and provide timely guidance to the capstone team, are expected to  
  
1. Meet with the team on a bi-weekly basis (teleconference is fine)  
2. Review the midterm progress report and provide comments to both the team and the course instructor  
3. Attend the final poster presentation session  
4. Review and evaluate the team's final report  
5. Provide comments on each team member's participation**

**Mentor Name\***

Sourav Ghosh

**Mentor Email Address\***

[sourav.g.ghosh@accenture.com](mailto:sourav.g.ghosh@accenture.com)

**Mentor Title\***

Data Science Senior Manager

**Mentor Department/Division\***

Accenture, Data & AI

**Mentor Phone Number\***

+919051037776

**Mentor 2 Phone Number\***

+919903856697

**Mentor LinkedIn Profile**

<https://www.linkedin.com/in/sourav-ghosh-b75a1354/>

**Current Resume**

Add file

**What amount of time per week do you intend to devote to working with the project team?**

10%

**Mentor 2 Name**

Paritosh Pramanik

**Mentor 2 Email Address**

[paritosh.pramanik@accenture.com](mailto:paritosh.pramanik@accenture.com)

**Mentor 2 Title**

Data Science Associate Manager

**Mentor 2 Department/Division**

Accenture, Data & AI

**Mentor 2 LinkedIn Profile**

<https://www.linkedin.com/in/paritosh-pramanik-b11b53105/>

**Current 2 Resume**

Your answer

**What amount of time per week do you intend to devote to working with the project team?**

10%