|  |  |  |
| --- | --- | --- |
|  | | |
|  | | |
| KIIT(sponsor).png | | |
|  | | |
|  | | |
| A major project report on | | |
|  | | |
| **FOREIGN EXCHANGE RATE PREDICTION USING ARTICLE DATA ( WEB SCRAPPING )** | | |
|  | | |
| submitted in partial fulfillment of the requirements for the degree of | | |
|  | | |
| B. Tech | | |
| In | | |
| Electronics and Telecommunication Engineering | | |
|  | | |
| By | | |
|  | | |
| **ANISH**  **HARSH BANSAL**  **SUMIT SAHA**  **TINA SASMAL**  **UTSAB GHOSH** | **1804430**  **1804442**  **1804485**  **1804486**  **1804488** |
|  | |
| under the guidance of | |
|  | |
| **Prof. V K Shrivastava** | |
|  | |
| |  | | --- | | School of Electronics Engineering | | **KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY**  (Deemed to be University)  BHUBANESWAR | |  | | APRIL 2021 | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CERTIFICATE**  This is to certify that the project report entitled **“FOREIGN EXCHANGE RATE PREDICTION USING ARTICLE DATA ( WEB SCRAPPING )”** submitted by   |  |  | | --- | --- | | **ANISH**  **HARSH BANSAL**  **SUMIT SAHA**  **TINA SASMAL**  **UTSAB GHOSH** | **1804430**  **1804442**  **1804485**  **1804486**  **1804488** |   in partial fulfilment of the requirements for the award of the **Degree of Bachelor of Technology** in **Electronics and Telecommunication Engineering**  is a bonafide record of the work carried out under my(our) guidance and supervision at School of Electronics Engineering, KIIT (Deemed to be University). | | | |
| se | Signature of Supervisor 1  Prof. V K Shrivastava  School of Electronics Engineering  KIIT (Deemed to be University) | |
|  | | | |
|  | | | |
| **The Project was evaluated by us on \_\_\_\_\_\_\_\_\_\_\_\_\_** | | | |
|  | | | |
| EXAMINER 1 | | EXAMINER 2 | |
| EXAMINER 3 | | EXAMINER 4 | |

**ACKNOWLEDGEMENTS**

We feel immense pleasure and feel privileged in expressing our deepest and most sincere gratitude to our supervisor **Professor V K Shrivastava** , for his excellent guidance throughout our project work. His kindness, dedication, hard work and attention to detail have been a great inspiration to us. Our heartfelt thanks to you sir for the unlimited support and patience shown to us. We would particularly like to thank him for all his help in patiently and carefully correcting all our manuscripts. We acknowledge the support received from Web Scrapping organization website for all its valuable information about the modules

**STUDENT SIGNATURE**

|  |  |  |
| --- | --- | --- |
| Roll Number | Name | Signature |
| **1804430** | **ANSIH** | WhatsApp Image 2021-11-12 at 4.53.42 PM |
| **1804442** | **HARSH BANSAL** | **WhatsApp Image 2021-11-12 at 6.08.18 PM** |
| **1804485** | **SUMIT SAHA** | WhatsApp Image 2020-10-20 at 10.06.12 PM |
| **1804486** | **TINA SASMAL** | **WhatsApp Image 2021-11-12 at 7.34.18 PM** |
| **1804488** | **UTSAB GHOSH** | **se2** |

Date:- **14.11.21**

**ABSTRACT**

We have implemented machine learning techniques in order to predict the foreign exchange rates of respective countries based on the data that were fetched from various sources . We have performed web scrapping using the API key generated from the NYT articles website.Then we have used that raw data remodeled it and converted it into a usable csv formatted file on which the rest of the algorithms were performed on.We have performed various regression models on the respective countries and selected the best model on the basis of Mean Squared Error(MSE) values. We used the articular data and the macro economical data to efficiently predict the subsequent years foreign exchange rates. The novelty of our project relies on utilizing the Latent Dirichlet Allocation to cluster news articles into topics and further understand the semantic meanings behind each topic and applying it to foreign exchange rates.

To give you a model, a HTML or XML document is changed over to DOM. What DOM does is that it communicates the construction of records and how a file can be gotten to. PHP gives DOM augmentation.The Document Object Model, or DOM, characterizes the style, structure and the substance inside XML records. DOM parsers are by and large utilized by scrubbers that need to get a top to bottom perspective on the construction of a page. Scrubbers can utilize a DOM parser to get the hubs containing data, and afterward utilize an instrument, for example, XPath to scratch site pages. Undeniable internet browsers like Internet Explorer or Firefox can be implanted to separate the whole page or simply aspects of it, regardless of whether the substance produced is dynamic in nature.

**Keywords** - Banks, Exchange rates , Tpot regressor, Web-scrapping, Machine learning.

**TABLE OF CONTENTS**

Abstract

Table of Contents

List of Figures

List of Tables

List of symbols/ abbreviations

CHAPTER I: INTRODUCTION 1

1.1 Overview 1

CHAPTER 2: METHODOLOGY FOR PREDICTING THE FER 2 2.1 Methodologies 2

2.2 Data Collection 3

2.3 Web Scrapping 4

2.4 Web Scrapping Methods 6

2.5 Topic Modelling 8

CHAPTER 3: IMPLEMENTATION OF OUR MODEL FOR

PREDICTING THE EXCHANGE RATES 10 3.1 Data Analysis 10

3.2 Data Exploration 14

3.3 Data Cleaning 15

3.4 Data Training 16

3.5 Designing of Model 17

CHAPTER 4: DESCRIPTIVE STATISTIC ALANALYSIS OF OUR

MODEL 4.1 GDP over years

4.1.1 Correlation Matrix 25

4.1.2 Plots 26

4.2 FER over years

4.2.1 Correlation Matrix 30

4.2.2 Plots 31

4.3 Modelling 34

CHAPTER 5: OPTIMIZATION OF MODEL

5.1: Optimized Techniques 36

5.2: Feature Importance 38

5.3: Future Prospects 40

5.4: Conclusion 42

5.5: Planning and Project management (Gantt chart) 44

REFERENCES 46

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Fig no.** | **Description** | **Page No.** |
| 2.4 | Topics of LDA Model | 4 |
| 3.5 | Tpot regressor Model | 6 |
| 3.5 | Tpot Pipeline Model | 8 |
| 3.5 | RidgeCV Model | 8 |
| 3.5 | Adaboost Regressor Model | 9 |
| 3.5 | LinearSVR Model | 9 |
| 4.1.1 | Correlation Matrix of GDP | 11 |
| 4.1.1 | Distribution of Countries FER | 22 |
| 4.1.2 | Plot of GDP | 14 |
| 4.2.1 | Correlation Matrix of FER | 11 |
| 4.3 | Error Plot of all topics | 32 |
| 5.2 | Feature Importance Plots | 14 |

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **Fig no.** | **Description** | **Page No.** |
| 2.4 | Topics of LDA Model | 4 |
| 3.1 | FERS of Different Countries | 6 |
| 4.1.1 | IQR range of FER of different countries | 8 |
| 4.3 | Error rates | 8 |
| 5.1 | Optimized techniques | 9 |
| 5.5 | Planning and Project Management | 9 |

**CHAPTER 1**

# INTRODUCTION

* 1. **OVERVIEW**

Machine learning (ML) is the study of computer [algorithms](https://en.wikipedia.org/wiki/Algorithm) that improve automatically through experience and by the use of data. It is seen as a part of [artificial intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence). Machine learning algorithms build a model based on sample data, known as "[training data](https://en.wikipedia.org/wiki/Training_data)", in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as [email filtering](https://en.wikipedia.org/wiki/Email_filtering) and [computer vision](https://en.wikipedia.org/wiki/Computer_vision), where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

A subset of machine learning is closely related to [computational statistics](https://en.wikipedia.org/wiki/Computational_statistics), which focuses on making predictions using computers; but not all machine learning is statistical learning. The study

of [mathematical optimization](https://en.wikipedia.org/wiki/Mathematical_optimization) delivers methods, theory and application domains to the field of machine learning. [Data mining](https://en.wikipedia.org/wiki/Data_mining) is a related field of study, focusing on [exploratory data](https://en.wikipedia.org/wiki/Exploratory_data_analysis) analysis through [unsupervised learning](https://en.wikipedia.org/wiki/Unsupervised_learning). In its application across business problems, machine learning is also referred to as [predictive analytics](https://en.wikipedia.org/wiki/Predictive_analytics).

Our project basically surrounds upon the idea of implementing machine learning techniques and algorithms in order to predict a desired result which serves our purpose both efficiently and accurately.In the case of our project we had the purpose or goal to predict the foreign currency exchange rates for subsequent years based on the previous years data.We basically focused on the macro economical data,the foreign exchange rates and the article data in order to predict the desired result.Starting from fetching the data till deployment our project had various steps and separate teams were made with capable members and knowledge to preform the task.Contributions were welcomed from all of the team members outside of that specific groups and necessary credits were given.

Our project workflow had the following steps/divisions:

* Data Collection
* Topic Modelling
* Data analysis
* Modelling
* Deployment

**CHAPTER 2**

## METHODOLOGY FOR PREDICTING THE FOREIGN EXCHANGE RATE

## 2.1. Data Collection

Our data collection was primarily focused upon utilizing the datasets : Foreign exchange rates,Macro economical indicator Data and Article Data.For the foreign exchange rates the yearly average foreign exchange data was downloaded from year 1981 to year 2019 for the following following: Dollar-Yuan (China), Dollar-Rupee (India), Dollar-Yen (Japan), Dollar-Pound (Great Britain), Dollar-Franc (Switzerland), and Dollar-Canadian Dollar (Canada).

For the news article data we utilized the New York Times Article Search API to collect metadata from around 2000 articles per year from 1981 to 2019.We focused on the topics which were most relevant to affect the exchange rates.A python web scraper code was built in order to harvest article data and meta data for the respective years in separate csv files which were later processed and implemented along with the other datasets.

For the macro economic indicators we fetched the Gross Domestic Product(GDP),Power Purchasing Parity(PPP),Inflation rate(INFL),Export rates and the import rates from the world bank website.Later after deciding among our team members we decided to move forward with the GDP and PPP data based on the feature importance and impact that they had towards the model accuracies.

### 2.2. Web Scrapping:

Collecting data from websites using an automated process is known as web scraping. Some websites explicitly forbid users from scraping their data with automated tools. Websites do this for two possible reasons:

The site has a good reason to protect its data. For instance, Google Maps doesn’t let you request too many results too quickly.

Making many repeated requests to a website’s server may use up bandwidth, slowing down the website for other users and potentially overloading the server such that the website stops responding entirely.

Web Scrapping can be done using HTML Parsing using libraries like Beautifulsoup, html5lib. Process involved using this method are –

1.Get HTML as String done with request module

2.Parsing Html with html5lib

3.Once html is fetched and parsed the next step is to manipulate the tree using beautifulsoup functions to get our job done.

Second way is to use the API Key provided by the respective websites like New York times.

#### 2.3. Research on methods of Web Scrapping:

Research on web scrapping was done in order to get the required technique to scrap the new York times data. Some of the techniques found to scrap the data are Human copy-and-paste. The simplest form of web scraping is manually copying and pasting data from a web page into a text file or spreadsheet.

1. Text pattern matching.
2. HTTP programming
3. HTML parsing.
4. DOM parsing.
5. Vertical aggregation.
6. Semantic annotation recognizing
7. Computer vision web-page analysis.
8. Using API Key

**Human copy-and-paste :** The simplest form of web scraping is physically reordering information from a website page into a text record or bookkeeping page. Here and there even the best web-scratching innovation can't supplant a human's manual assessment and reorder, and now and then this might be the main useful arrangement when the sites for scratching unequivocally set up obstructions to forestall machine mechanization.

**HTTP programming :** [Static](https://en.wikipedia.org/wiki/Static_web_page" \o "Static web page) and [dynamic web pages](https://en.wikipedia.org/wiki/Dynamic_web_page" \o "Dynamic web page) can be retrieved by representing HTTP demands on the remote web server utilizing attachment programming.

**HTML parsing :** Numerous sites have huge assortments of pages produced progressively from a hidden organized source like a data set. Information of a similar classification are ordinarily encoded into comparable pages by a typical content or format. In information mining, a program that recognizes such formats in a specific data source, extricates its substance and makes an interpretation of it into a social structure, is known as a covering. Covering age calculations expect that info pages of a covering acceptance framework adjust to a typical layout and that they can be effectively recognized as far as a URL normal plan. Besides, some semi-organized information inquiry dialects, like XQuery and the HTQL, can be utilized to parse HTML pages and to recover and change page content.

**DOM parsing :** The HTML Dom parser is acceptable decision, as it empowers us to access and utilize HTML effectively and easily. One can parse site pages as a DOM (Document Object Model) tree which is in a manner a portrayal of which undertakings can acquire induction to what portions of the pages.

To give you a model, a HTML or XML document is changed over to DOM. What DOM does is that it communicates the construction of records and how a file can be gotten to. PHP gives DOM augmentation.The Document Object Model, or DOM, characterizes the style, structure and the substance inside XML records. DOM parsers are by and large utilized by scrubbers that need to get a top to bottom perspective on the construction of a page. Scrubbers can utilize a DOM parser to get the hubs containing data, and afterward utilize an instrument, for example, XPath to scratch site pages. Undeniable internet browsers like Internet Explorer or Firefox can be implanted to separate the whole page or simply aspects of it, regardless of whether the substance produced is dynamic in nature.

**Vertical aggregation :** Vertical accumulation stages are made by organizations with admittance to enormous scope registering ability to target explicit verticals. A few organizations even run these information gathering stages on the cloud. Creation and checking of bots for explicit verticals is done by these stages with for all intents and purposes no human mediation. Since the bots are made naturally dependent on the information base for the particular vertical, their productivity is estimated by the nature of information separated.

**Semantic annotation recognizing :** It consists of two things - Xpath and Google sheets

**XPath:** XML Path Language, or XPath, is a query language that deals with XML records. Since XML records depend on a tree-like construction, XPath can be utilized to explore across the tree by choosing hubs dependent on an assortment of boundaries. XPath can be utilized related to DOM parsing to extricate a whole site page and distribute it at the objective site.

**Google Sheets :** Google Sheets can be utilized as a scratching apparatus, and it's very famous among scrubbers. From inside Sheets, a scrubber can utilize the IMPORTXML (,) capacity to scratch information from sites. This is helpful when the scrubber needs explicit information or examples to be separated from a site. You can utilize this order to check if your site is scratch evidence.

**Computer vision web-page analysis :** Computer vision additionally assists you with distinguishing site pages rapidly, permitting us to deliberately pull item data, pictures, recordings, articles and different information without figuring out pointless data.

Computer vision techniques empower you to precisely recognize key pieces of a site and concentrate those fields as organized information. This organized information then, at that point, empowers you to look for explicit picture types or text, or even explicit individuals.

Computer vision also allows us to (among other things):

* **Analyze images** – Using tagging, descriptions, and domain-specific models, it can identify content and label it accordingly, apply filters and settings, and separate images by type or even color scheme
* **Read text in images** – It can perceive words regardless of whether they are inserted inside pictures or in any case incapable to be separated, duplicated or stuck into a text record (called OCR, or Optical Character Recognition)
* **Read handwriting** – If information on a page is handwritten or an image of handwriting, it can also recognize and translate it into text (OCR)
* **Analyze video in real time** – Computer vision empowers us to extricate outlines from recordings from any gadget for examination

Certain e-commerce sites use computer vision to perform image analysis in their predictive analytics efforts to forecast what their customers will want next, for example. This can save an enormous amount of time when it comes to pulling, analyzing and using that data effectively.

Since it deals with organized information, Computer vision additionally gives us cleaner information that would help us to use to construct applications, illuminate our advertising choices. We can rapidly see designs in informational collections and recognize elements that you might have in any case missed.

**Text pattern matching :** A basic yet amazing way to deal with extract data from pages can be founded on the UNIX grep order or ordinary articulation coordinating with offices of programming dialects (for example Perl or Python). Customary Expressions(text design coordinating)

The fundamental thought of normal articulations is we characterize an example (the "ordinary articulation" or "regex") that we need to coordinate in a text string and afterward search in the string to return matches. Some of these patterns look pretty strange because they contain both the content we want to match and special characters that change how the pattern is interpreted. Regular expressions come up constantly when parsing string data and are a fundamental device to learn essentially at a fundamental level!

There are 3 pieces of info we need to extract from the text table :

* The names of the presidents
* The names of the colleges
* The salaries

First up is the name. In this regular expression, I make use of the fact that each name is at the start of a line and ends with a comma. The code below creates a regular expression pattern, and then searches through the string to find all occurrences of the pattern:

**# Create a pattern to match names**  
name\_pattern = re.compile(r'^([A-Z]{1}.+?)(?:,)', flags = re.M)

**# Find all occurrences of the pattern**  
names = name\_pattern.findall(content)

Like I said, the pattern is pretty complex, but it does exactly what we want! Don’t worry about the details of the pattern, but just think about the general process: first define a pattern, and then search a string to find the pattern.

**Scrapping Using HTML Parsing:**

We started by writing a code to scrap data using html parsing. With the help of request module, we called html module as string using the request module. Then Parsing was done with the help of htlm5lib and finally manipulated the tree using beautifulsoup. The final result was that the data was successfully fetched in the html form.

**Scrapping Using API Key:**

After understanding the complexity of the problem and matching it with our current knowledge, it was decided to go through the various codes available, try to understand them, debug them and try to find every possible way to use it for our purpose. So we went through the several codes, starting with the provided one on the problems research paper. After understanding this particular code, we tried to debug it and we were partially successful.After this we went through several other codes, and was able to use a web scrapper successfully to fetch the code.

We were successful in fetching the data through this method but the sudden change in timeline and then after analyzing the time taken through this method and the time complexity, other method was used. The Data was provided and another task of extracting and cleaning the data was provided.

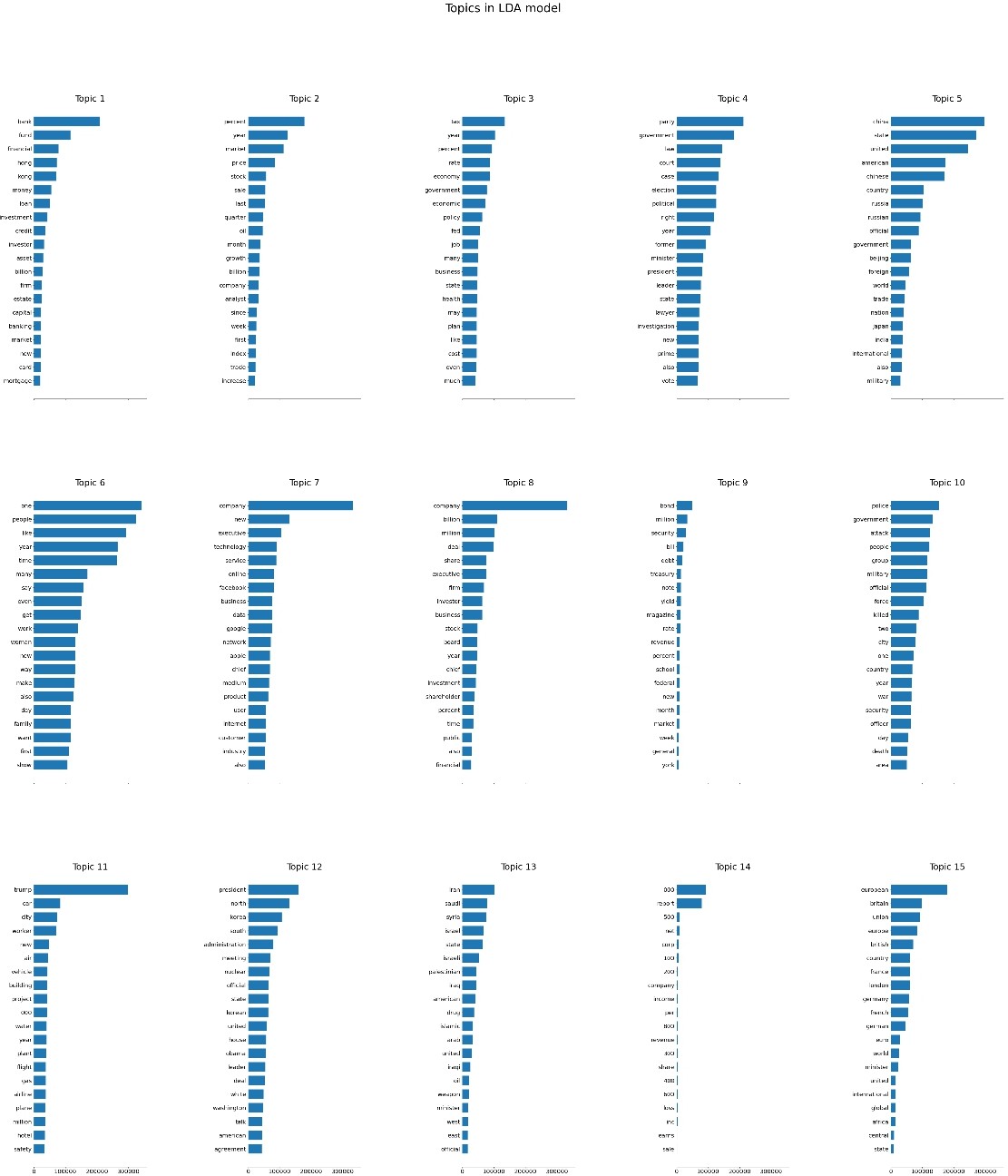
**2.4. Topic Modelling**

In [natural language processing](https://en.wikipedia.org/wiki/Natural_language_processing), the latent Dirichlet allocation (LDA) is a [generative statistical](https://en.wikipedia.org/wiki/Generative_model) model that allows sets of observations to be explained by [unobserved](https://en.wikipedia.org/wiki/Latent_variable) groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics. LDA is an example of a [topic model](https://en.wikipedia.org/wiki/Topic_model) and belongs to the [machine learning](https://en.wikipedia.org/wiki/Machine_learning) toolbox and in wider sense to the [artificial intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence) toolbox.

To generate features from news articles, we have used one of the NLP techniques called Latent Dirichlet Allocation (LDA). Ideally each articles is categorised in some generic topics but LDA clusters the articles as per the probability for each topic based on the count-vectorization of each words in all the articles. The articles of each year are combined and transformed into single dataframe consisting of probability of every article belonging to a particular topic which is named as distribution. We tried many values for number of topics (5, 10, 15, 20, 25, 30, 40, 50).

A distribution dataframe is saved for respective number of topics. The challenge was to transform a topic modelling dataframe to merge with economic factors which consist of single row for each year while there were more than 4000 rows of articles published every year. So to resolve this issue, we calculated the mean of probabilities of a each year’s topic and converted them into single row for a particular year. This dataframe is named as weights. There is no particular method to decide an optimised number of topic for a corpus, that’s why we had to figure out a way to do so.

We used a model optimization technique called GridSearchCV but it yielded a different topic value in each run. Hence, we decided to do trial and error method to decide optimal number of topic value. We found that dataframe with 15 topics gives best accuracy for prediction. The accuracies of 15 topics are listed below :



The plot shows the importance of particular word in every topic. Below are the words belonging to each topic.

|  |  |
| --- | --- |
| Topic 0 | bank, fund, financial, hong, kong, money, loan, investment, credit, investor, asset, billion, firm, estate, capital,banking, market, new, card, mortgage |
| Topic 1 | percent, year, market, price, stock, sale, last, quarter, oil, month, growth, billion, company, analyst, since, week, first, index, trade, increase |
| Topic 2 | tax, year, percent, rate, economy, government, economic, policy, fed, job, many, business, state, health, may, plan, like, cost, even, much |
| Topic 3 | party, government, law, court, case, election, political, right, year, former, minister, president, leader, state, lawyer, investigation, new, prime, also, vote |
| Topic 4 | china, state, united, american, chinese, country, russia, russian, official, government, beijing, foreign, world, trade, nation, japan, india, international, also, military |
| Topic 5 | one, people, like, year, time, many, say, even, get, work, woman, new, way, make, also, day, family, want, first, show |
| Topic 6 | company, new, executive, technology, service, online, facebook, business, data, google, network, apple, chief, medium, product, user, internet, customer, industry, also |
| Topic 7 | company, billion, million, deal, share, executive, firm, investor, business, stock, board, year, chief, investment, shareholder, percent, time, public, also, financial |
| Topic 8 | bond, million, security, bill, debt, treasury, note, yield, magazine, rate, revenue, percent, school, federal, new, month, market, week, general, york |
| Topic 9 | police, government, attack, people, group, military, official, force, killed, two, city, one, country, year, war, security, officer, day, death, area |
| Topic 10 | trump, car, city, worker, new, air, vehicle, building, project, 000, water, year, plant, flight, gas, airline, plane, million, hotel, safety |
| Topic 11 | president, north, korea, south, administration, meeting, nuclear, official, state, korean, united, house, obama, leader, deal, white, washington, talk, american, agreement |
| Topic 12 | iran, saudi, syria, israel, state, israeli, palestinian, iraq, american, drug, islamic, arab, united, iraqi, oil, weapon, minister, west, east, official |
| Topic 13 | 000, report, 500, net, corp, 100, 200, company, income, per, 800, revenue, 300, share, 400, 600, loss, inc, earns, sale |
| Topic 14 | european, britain, union, europe, british, country, france, london, germany, french, german, euro, world, minister, united, international, global, africa, central, state |

**CHAPTER 3**

**IMPLEMENTATION OF OUR MODEL FOR PREDICTING THE EXCHANGE RATES**

**3.1. Data Analysis**

In data mining, Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics, often with visual methods. EDA is used for seeing what the data can tell us before the modelling task. It is not easy to look at a column of numbers or a whole spreadsheet and determine important characteristics of the data. It may be tedious, boring, and/or overwhelming to derive insights by looking at plain numbers. Exploratory data analysis techniques have been devised as an aid in this situation.

Exploratory data analysis is generally cross-classified in two ways.

1. First, each method is either non-graphical or graphical.
2. And second, each method is either univariate or multivariate (usually just bivariate).

Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to

* maximize insight into a data set
* uncover underlying structure
* extract important variables
* detect outliers and anomalies
* test underlying assumptions
* develop parsimonious models

“Before Training, any model using any algorithm Data Preprocessing is that the most significant step and will be the primary step. The data Preprocessing contains several checkpoints (steps) such as:

**Step 1**: Import Libraries: The essential Libraries for Data preprocessing I used are Pandas for data manipulation and analysis, Numpy for numerical analysis, Matplotlib and Seaborn for better visuals and graphical stats of the data.

**Step 2:** Import the Dataset: This downloaded the dataset from Kaggle, and then downloaded the dataset using the pandas library.

**Step 3:** Taking care of Missing Data in Dataset: After evaluation of this dataset, I found no missing values in the dataset.

**Step 4:** Encoding categorical data: This dataset contains some Categorical values such as ‘countries name’, so we need to encode these categorical data into an encoded format to better train our model, to do this I used get\_Dummies() method of pandas and this converted the whole Categorical values in the dataset into binary values.

**Step 5:** Splitting the Dataset into the Training set and Test Set: To split this dataset into Test and Train dataset to train our machine learning model I used the capable machine learning library of python, scikit-learn or sklearn. Using its model selection method to create testing data by picking random values from the available dataset for model prediction, or we can say Supervised Learning.

**Step 6:** Feature Scaling: Since all the data, available in a standard format, so here I do not use any feature scaling techniques.

Just based on analyzing our input data set, we find out the minimum and maximum foreign exchange rate between countries and also visualize the distribution of FER. The maximum and minimum fer over countries is shown in the correlation matrix.

|  |  |  |
| --- | --- | --- |
| **Countries Name** | **Maximum FER** | **Minimum FER** |
| India | 1.97 | 0.67 |
| China | 1.98 | 0.66 |
| Japan | 1.85 | 0.54 |
| UK | 1.98 | 0.19 |
| Switzerland | 1.84 | 0.76 |
| Canada | 1.92 | 0.84 |

### **3.2. DATA EXPLORATION**

This helps us unravel specific patterns or characteristics to understand ideas and implement new policies. While data mining doesn't necessarily reveal every detail, it helps form a bigger picture of specific trends or areas to study. Using manual methods and automated tools, we explore the data to determine which model or algorithm is best for the later stages of data analysis. Using machine learning for exploratory data analysis, it helps us track data sources and explore data for big analytics. While manual data mining can be useful to focus on specific data sets of interest to you, machine learning provides a much broader focus, providing transformable useful insights. your business's understanding of patterns and trends.

窗体顶端

窗体底端

In our project, we mainly target two goals :

1. To highlight traits of single variables
2. Reveal patterns and relationships between variables.

To highlight traits of single variables , we represent each row as a person who was a survey respondent. The columns are the feature values corresponding to those people. We have 30,000 observations and 25 features. We have the same 30,000 observations, and one target variables that we have labels for.

**A. Feature Evaluation**

The evaluation of features was performed using various functions found in the scikit −learn library; a library for machine learning. It’s use in the project can be found towards the end of the data Cleaning And Preparing.py file.

A project specific improvement here would be to base the comparison of the three different selection methods used (correlation heat map, Pearson’s correlation coefficient and mutual information regression) on numerical methods. Instead, a visual evaluation was considered by investigating their plots. Implementing a function which would provide an overall score, cross-validating over all selection methods, would be a more robust and systematic path to take.

**B. Feature Selection**

Following the exploratory stage of feature evaluation, the final step of removing irrelevant features was taken in the last section of data Cleaning and Preparing.py. This step was made thoroughly easy due how the data frames had been prepared in the earlier section of the code. Finally, the training and testing sets are exported as CSV files and can be found in the folder prepared data Files under Train.csv and

Test.csv.

**3.3. DATA CLEANING**

* **Missing values** - We checked for missing values and removed them for better accuracy.
* **Extreme values or Outliers** - Outliers in data refer to observations that are divergent from a generalized pattern in a data sample. Outliers can skew data considerably, and should be highlighted and addressed before extracting insights.

**3.4. Data Training and Modelling**

To train and develop a model, first of all, we need to the dependent and independent variables. To find these variables, first I used to find the correlation between the variables of the output and then separates my variables into two different axes we call it x and y where the x-axis contains all the independent variable and y-axis having the dependent variable, in our model its rate price. Using sklearn.model\_selection library and its train\_test\_split function. GridSearchCV has been used for this model to find the best hyperparameters for our model prediction.

## 3.5. Design:

**ANALYTICAL METHOD AND TECHNOLOGY USED**

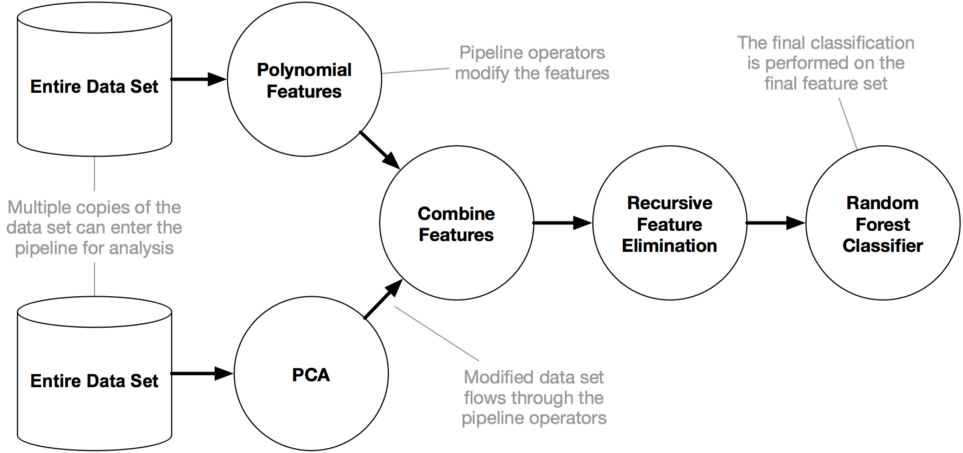
We will be using different approach to our problem for which we have libraries like ScikitLearn which can help us in predicting our outcomes.

**AutoML - Tpot Regressor**

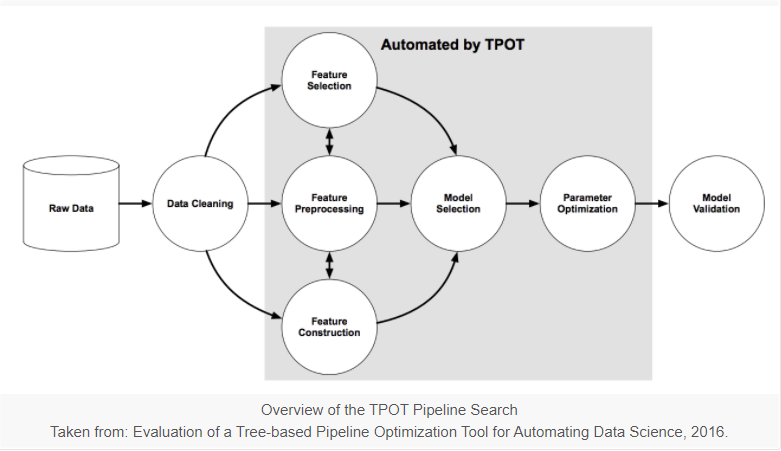
Automated machine learning (AutoML) refers to techniques for automatically discovering successful models for predictive modeling tasks with very little user involvement. TPOT is an open source library for running AutoML in Python.

It uses the popular machine learning library ScikitLearn for data transformations and machine learning algorithms, and uses a global stochastic genetic programming search to efficiently explore a system of patterns works best for a given data set.

Tree-Based Pipeline Optimization Tool, or TPOT for short, is a Python library for automated machine learning. TPOT uses a tree structure to represent a model pipeline for a predictive modeling problem, including data modeling and preparation algorithms and model hyperparameters. Then an optimization procedure is performed to find a tree structure that works best for a given data set. More precisely, a genetic programming algorithm, designed to perform global stochastic optimization on programs represented as trees.



The following figure from the TPOT paper shows the factors involved in pipeline research, including data cleaning, feature selection, feature processing, feature building, model selection, and optimization super parameter.



The initialization, tuning, and scoring of the TPOT classifier is similar to that of any other sklearn classifier. Here is the format:

**tpot = TPOTClassifier()  
tpot.fit(X\_train, y\_train)  
tpot.score(X\_test, y\_test)**

TPOT comes with its own variant of onehot encryption. Note that it may automatically add it to a path because it treats features with less than 10 unique values as a taxonomy. If you want to use your own encryption strategy, you can encrypt your data and then include it in TPOT.

You can choose the scoring criteria for tpot.score (although a bug with Jupyter and multiple processor cores prevents you from having a custom scoring criteria with multiple processor cores in a Jupyter notebook).

It seems that you cannot change the scoring criteria that TPOT uses internally when searching for the best path, only the scoring criteria to use in the test suite after TPOT selects the best algorithms. This is an area where some users may want more control. Maybe this option will be added in a future version.

TPOT writes information about the best performing algorithm and its exact score to a file with Tpot.export(). You can choose how much granularity you want to see while TPOT is running and have it write the pipeline to the output file on the fly in case it ends prematurely for some reason (e.g. your Kaggle Kernel crashes).

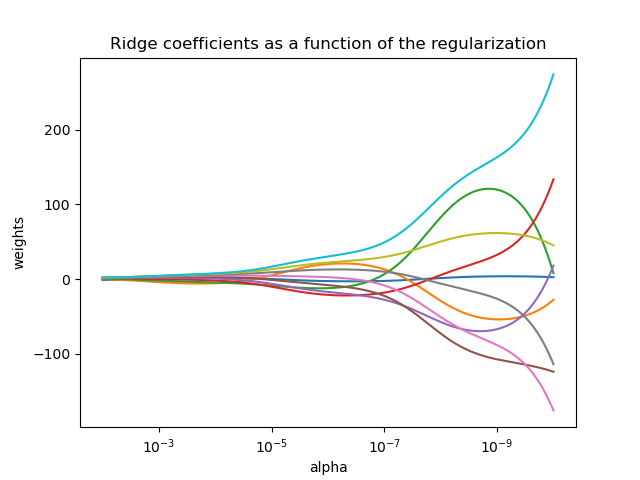
TPOT is designed to work over a period of time - hours or even a day. Although less complex problems with smaller data sets can see great results within minutes. You can tweak some parameters to make TPOT complete searches faster, but with less overhead when searching for an optimal path. It is not designed to be a complete search for preprocessing steps, feature selection, algorithms and settings, but it can be over if you set up its settings thoroughly. than.

**Here are the default TPOTClassifier parameters:**

generations=100,   
population\_size=100,   
offspring\_size=None # Jeff notes this gets set to population\_size  
mutation\_rate=0.9,   
crossover\_rate=0.1,   
scoring="Accuracy", # for Classification  
cv=5,   
subsample=1.0,   
n\_jobs=1,  
max\_time\_mins=None,   
max\_eval\_time\_mins=5,  
random\_state=None,   
config\_dict=None,  
warm\_start=False,   
memory=None,  
periodic\_checkpoint\_folder=None,   
early\_stop=None  
verbosity=0  
disable\_update\_check=False

After the application of Tpot regressor on our six different country respective datasets, we came across 4 different working models that gave us the best scores. The models are respectively :

1. **RidgeCV()**

Ridge regression is part of the L2 regularization family of regressions. This differs from the L1 regularity which limits the size of the coefficients by adding a fine equal to the absolute value of the magnitude of the coefficients. This results in sparse models, while in Ridge regression the penalty is equal to the square of the magnitude of the coefficients. 

#### Objective = RSS + α \* (sum of square of coefficients)

Here, α (alpha) is the parameter which balances the amount of emphasis given to minimizing RSS vs minimizing sum of square of coefficients. α can take various values:

1. ****α = 0:****

* The objective becomes same as simple linear regression.
* We’ll get the same coefficients as simple linear regression.

1. ****α = ∞:****

* The coefficients will be zero. Why? Because of infinite weightage on square of coefficients, anything less than zero will make the objective infinite.

1. ****0 < α < ∞:****

* The magnitude of α will decide the weightage given to different parts of objective.
* The coefficients will be somewhere between 0 and ones for simple linear regression.

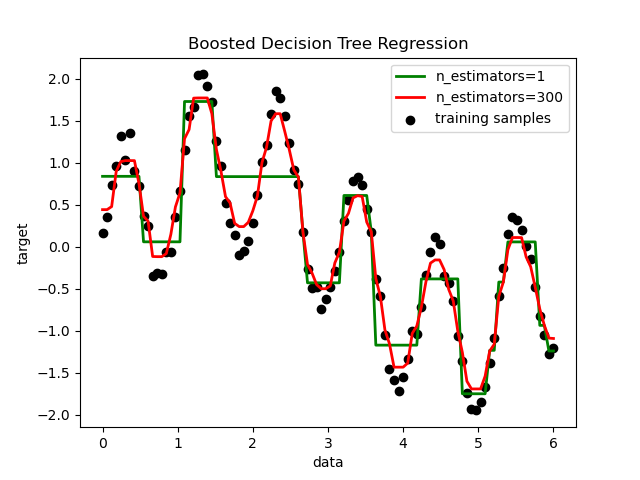
1. **AdaBoostRegressor(learning\_rate=0.1/0.001,loss="square"/”linear”,**

**n\_estimators=100)**

When nothing works, Boost works. Nowadays, many people use XGBoost or LightGBM or CatBoost to win competitions at Kaggle or Hackathons. AdaBoost is the first springboard in the Boost world.

AdaBoost was one of the first boost algorithms to adapt to solving practice. Adaboost helps you to combine several "weak classifiers" into one "strong classifier". Here are some (fun) facts about Adaboost :

* The weak learners in AdaBoost are decision trees with a single split, called decision stumps.
* AdaBoost works by putting more weight on difficult to classify instances and less on those already handled well.
* AdaBoost algorithms can be used for both classification and regression problem.



The power of the assembler is such that we can still build powerful assemblies even when the individual models in the assembly are extremely simple.

Decision sequences are the simplest model we can build, just guessing the same tag for every new example no matter what it looks like. The accuracy of such a model will be better if we guess any answer, 1 or 0, is the most common in the data. Let's say, if 60% of the examples are 1s, then we will get 60 accuracy just by guessing 1 each time.

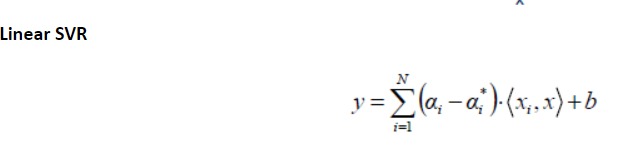
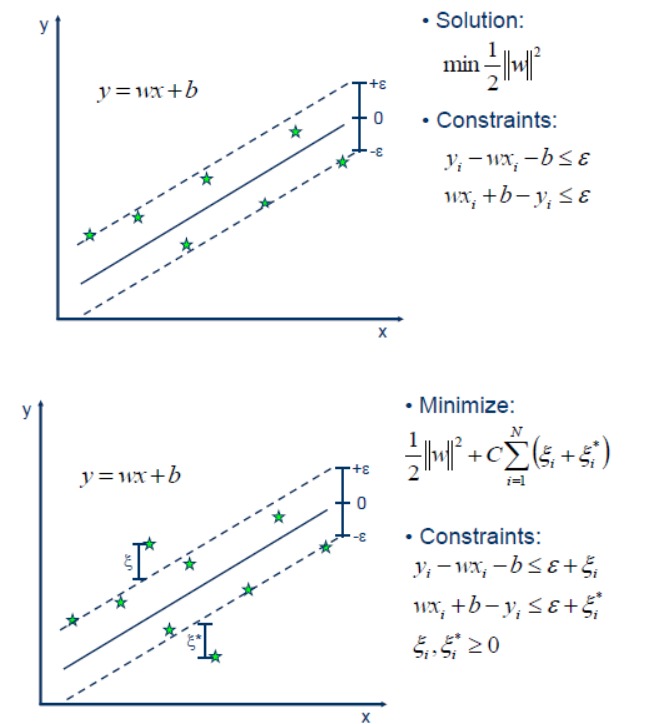
Decision distortions improve this by splitting the examples into two subsets based on the value of a property. Each strain selects a trait, such as X2, and a threshold, T, and then divides the examples into two groups on either side of the threshold.

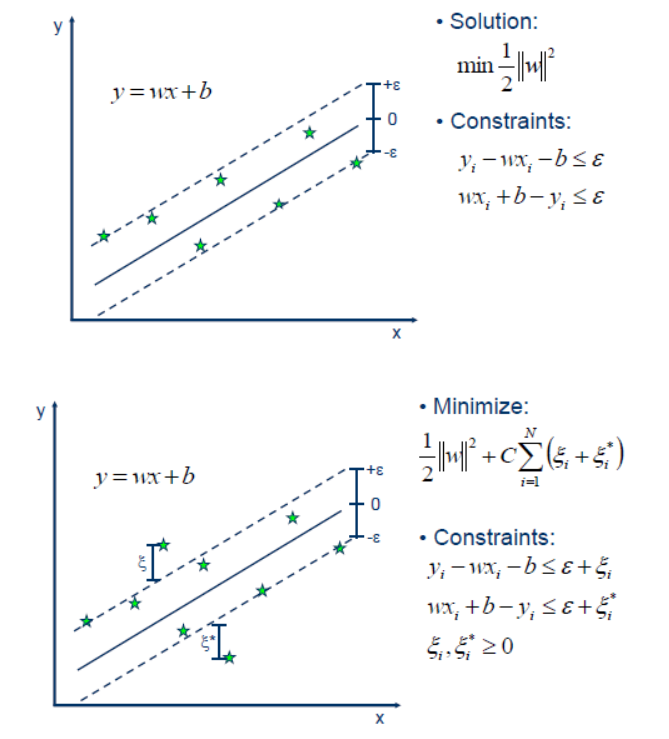
To find the decision scheme that best fits the examples, we can try each characteristic of the input as well as each possible threshold and see which gives the best accuracy. While it may seem naive that there are infinitely many choices for thresholding, the two different thresholds are only significantly different if they place examples from different sides of the split. Then, to try all possibilities, we can sort the examples by the characteristic in question and try a decreasing threshold between each pair of adjacent examples.

The algorithm just described could be further improved, but even this simple version is extremely fast compared to other ML algorithms (e.g. neural network training).

1. **LinearSVR(C=25.0,dual=False,epsilon=0.0001,loss="squared\_epsilon\_insensitive", tol=1e-05)**

Support Vector Machine can also be used as a regression method, keeping all the key features that characterize the algorithm (maximum profit). Support vector regression (SVR) uses the same principles as SVM for classification, with only some minor differences. First of all, because the output is a real number, it is difficult to predict the available information, which has infinitely many possibilities. In the regression case, a tolerance boundary (epsilon) is fixed as an approximation to the SVM that would have requested it from the problem. But besides this fact, there is also a more complicated reason, so the algorithm is also more complicated to consider. However, the main idea is always the same: to minimize error, by personalizing the profit-maximizing hyperplane, remember that partial error is accepted.





1. **LassoLarsCV(normalize=False)**

Lasso regression analysis is a method of removing and selecting variables for a linear regression model. The goal of lasso regression is to obtain a subset of predictors to minimize the prediction error for a quantitative response variable. To do this, lasso imposes a constraint on the model's parameters that causes the regression coefficients of certain variables to drop to zero. Variables with zero regression after the removal process will be dropped. removed from the model. Variables with non-zero regression coefficients are most closely associated with the response variable. The explanatory variables can be quantitative, categorical, or both.

**CHAPTER 4**

**DESCRIPTIVE STATISTIC ALANALYSIS OF OUR MODEL**

* 1. **GROSS DOMESTIC PRODUCT OF EACH COUNTRY OVER YEARS:**

**India :** Statistics show that India's GDP from 1986 to 2020, forecast to 2026. In 2019, India's GDP is about 2.87 trillion US dollars. Real GDP figures are an even more reliable tool for estimating which direction a country's economy might go, because they are adjusted for inflation and reflect changes in prices. reality.

**China :** China's gross domestic product in 2019 was $99.8651 billion, or $14.4 trillion (nominal). But in terms of purchasing power parity (PPP), China became the world's second-largest economy as early as 1999, overtaking Japan and toppling the United States to become the largest economy since 2014. From 1979 to 2010, China's average annual GDP growth was 9.91%, reaching an all-time high of 15.2% in 1984 and a record low of 3.8% in 2010. 1990. At current prices, the country's average annual GDP growth over those 32 years was 15.8%, reaching an all-time high of 36.41% in 1994 and a record 6.25%. in 1999.

**Japan :** Statistics show that the gross domestic product (GDP) in Japan from 1986 to 2019, forecasted to 2026. In 2019, Japan's gross domestic product was about 5.05 trillion US dollars. This puts Japan in third place in the world GDP ranking. By comparison, China's GDP was $8.390 billion in 2012, marking the world's second-largest economy.

**UK :** Britain's gross domestic product is starting to show signs of recovery since it plummeted after the financial crisis. However, some concerns have arisen about the represented forms of employment, the stability of employment and whether they are criticized by government officials simply as validating reforms pointed out by opponents. quoted as having “thought motives” or not. In 2019, GDP per capita in the UK was around $42,416.6. In the same year, the total population of Great Britain was about 64.6 million people. The UK is among the top countries in the global GDP rankings.

**Switzerland :**  In 2019, Switzerland's GDP reached about $732.19 billion. Switzerland's gross domestic product was largely unaffected by the 2008 global economic crisis. After briefly falling to negative territory in 2009, Switzerland's gross domestic product/GDP growth was positive every year, 2010 was the most successful year since the crisis. However, Switzerland's share of global GDP (adjusted for purchasing power parity) has steadily declined over the past decade. In addition, Switzerland leads the ranking of countries with the highest wealth per adult, with an average national wealth of over US$540 million per Swiss adult. Following that, Switzerland also has one of the main countries with the highest percentage of millionaires, this is probably due to the unparalleled tax rates in Switzerland.

**Canada :** In 2020, Canada's gross domestic product is about 1.64 trillion US dollars. Canada is the second largest country in the world after Russia and the largest country in North America. Despite its large size, Canada has a relatively small population of only about 35.9 million people. However, it is estimated that Canada's total population will reach around 37.5 million in 2019. Not only is Canada's population growing, but the economy is slowly recovering from the 2008 global financial crisis. . The unemployment rate in Canada in 2010 was around 8% (263,696). In addition, the gross domestic product per capita in 2014 was among the highest in the world, and in the same year the country's gross domestic product increased by more than 2.5% year-on-year. Canada's economic growth is the result of political stability and economic reforms following the global financial crisis.

* + 1. **Correlation:** This matrix represents the correlation between all possible pairs of values ​​in an array. The correlation coefficient represents the strength or weakness of the correlation between two exchange rate pairs. Correlation coefficients are expressed as values ​​and can range from 1 to 1, the decimal representing the coefficient. This is a statistical measure of how related currency pairs are in terms of value and will move together. Therefore, we compared the ratios and created a correlation matrix as shown below:



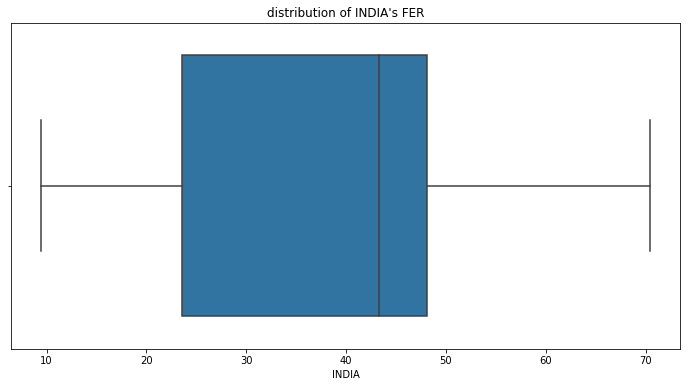
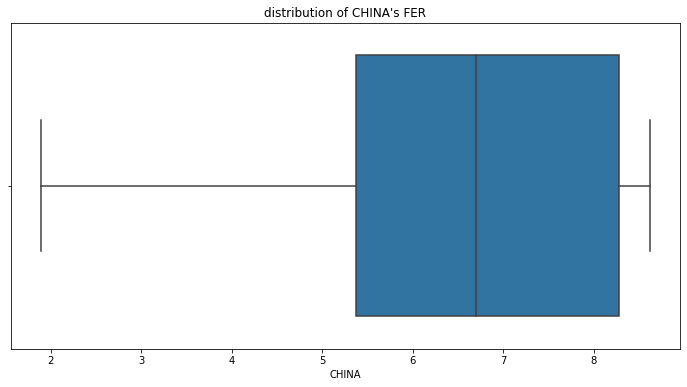
The following figure shows the correlation matrix on a heatmap. We can derive the following insights from this correlation matrix.

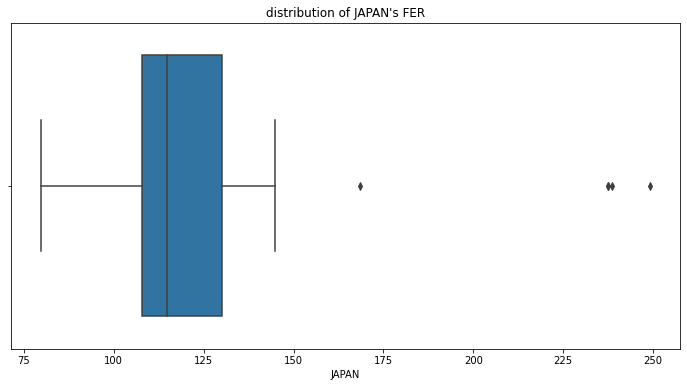
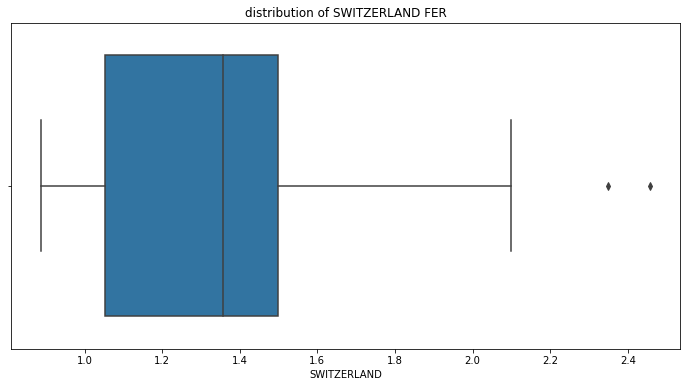
1. The dark colors shows highest relationship between GDPs of different countries and light color shows the weak relationship between countries.
2. India and China has the highest relationship over GDP over the years followed by Switzerland and Canada, Canada and UK.
3. Japan has the weak relationship with India and China over the years followed by Switzerland and Canada, Canada and UK.
4. Switzerland and Canada, Canada and UK has the neutral relationship with other countries over GDP over years.

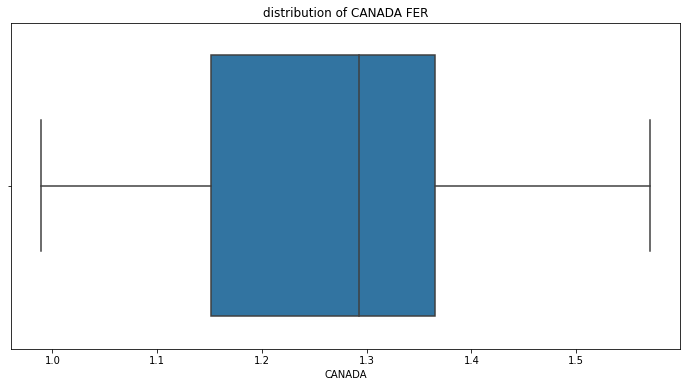
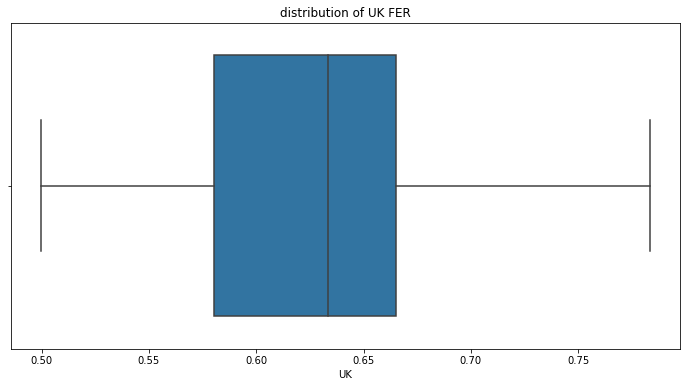
Based on this, we have analysed the weight-age of the words in their respective above topics (0-14) on the basis of distributions of FER of the countries from the meta-data. Through this, we get to know which words occurred the most in our meta data articles and going to effect our FER over years. We have done this with the help of box plot. A boxplot is a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”). It tells us about our outliers and what their values are. We also get to know if our data is symmetrical, how tightly our data is grouped, and if and how our data is skewed.

The Inter quartile range (IQR) of several countries is listed below along with box plots -:

|  |  |
| --- | --- |
| **Countries Name** | **IQR range** |
| India | 24.56 |
| China | 2.90 |
| Japan | 22.34 |
| UK | 0.085 |
| Switzerland | 0.44 |
| Canada | 0.21 |





#### Plot showing GDP of each country over years from 1981-2019:

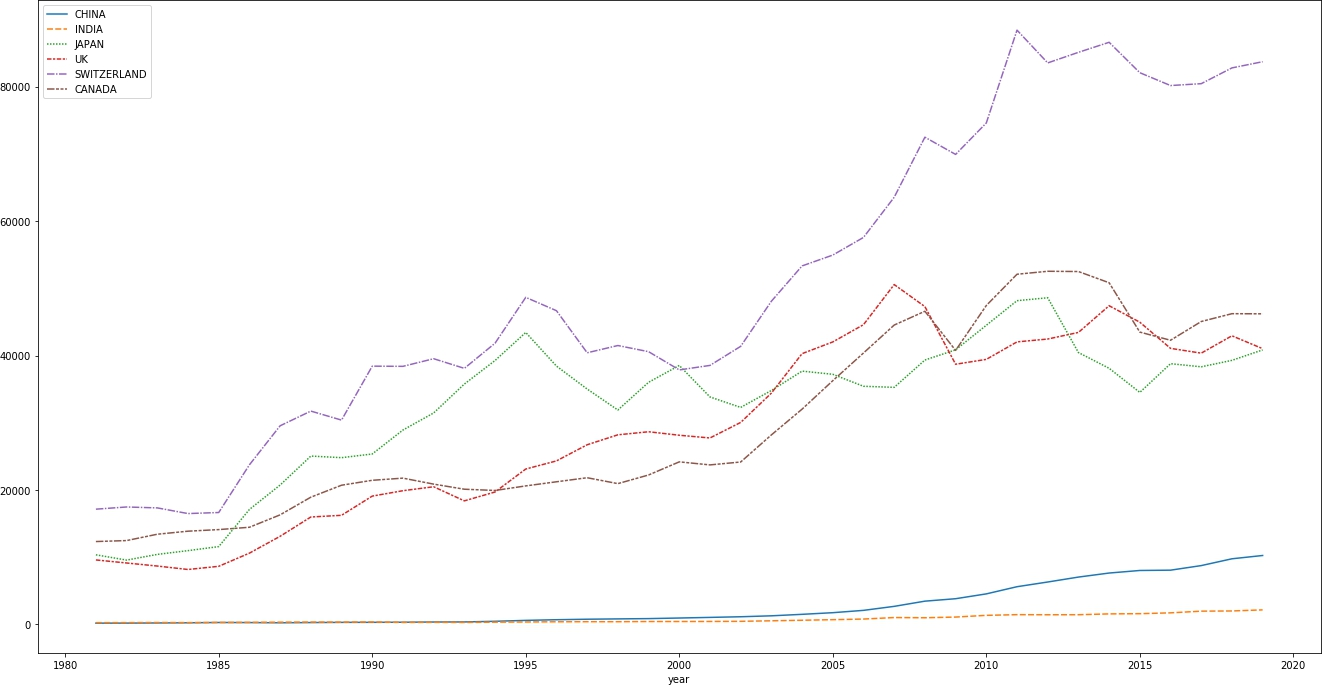
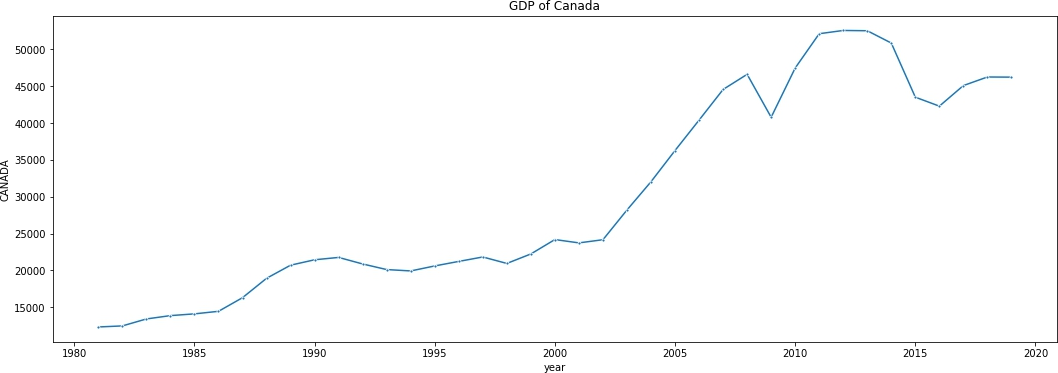


Fig.2.2

It is clearly seen that the GDP has increase exponentially from 2000 to 2019 and the major increase is shown in Switzerland. It indicates that the economic condition of Switzerland from 2000-2019 was far better from rest of the countries. Due to this, the FER should be more for Switzerland as compared with other countries.

**Plot showing GDP of Canada over years:**



We see that the GDP of Canada with respect of USA is increasing over time though it has decrease from the years 1992-1994, fluctuate between the years 1995-2003 and then sudden increase from 2003-2009 and again fluctuate.

**Plot showing GDP of Switzerland over years:**

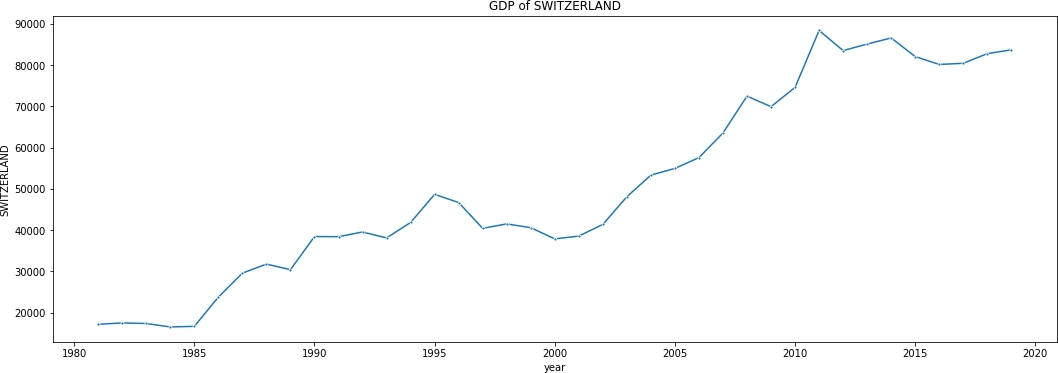


Fig.2.4

We see that the GDP of Switzerland with respect of USA keeps on fluctuating over years.

**Plot showing GDP of UK over years:**

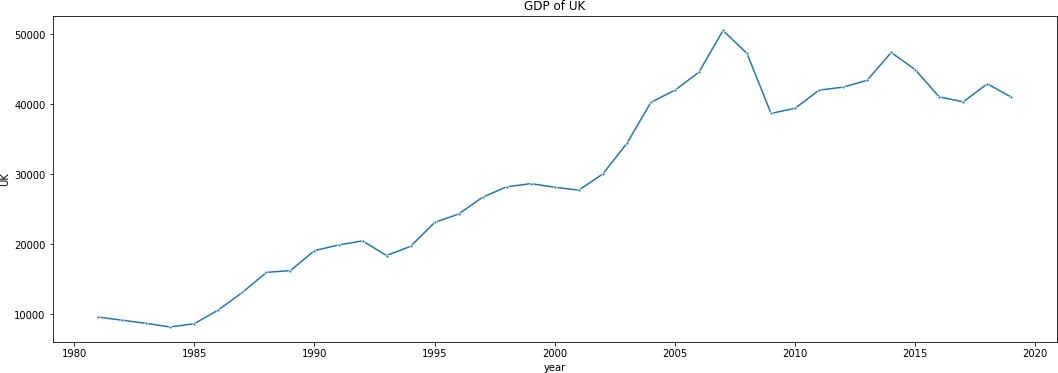
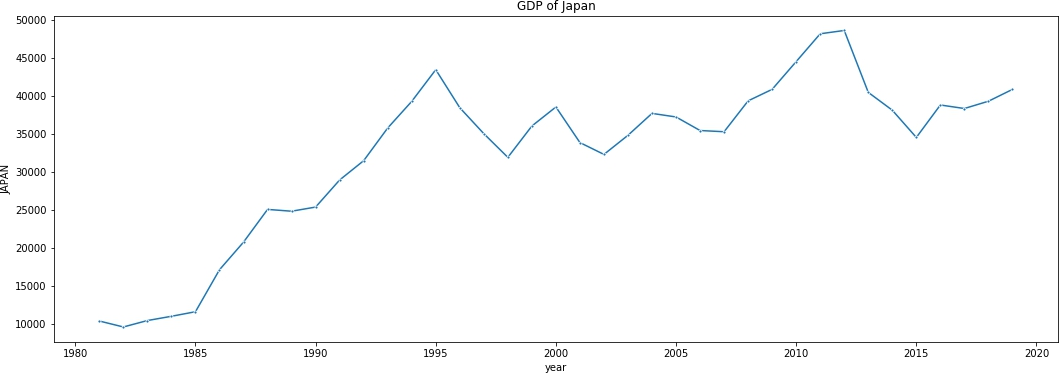


Fig.2.5

We see that the GDP of UK with respect of USA is increasing over time till 2003 but in the latter years its showing some downfall.

**Plot showing GDP of Japan over years:**

Fig.2.6

We see that the GDP of Japan with respect of USA keeps on fluctuating over years then also its shows some increase in the later part.

**Plot showing GDP of India over years:**

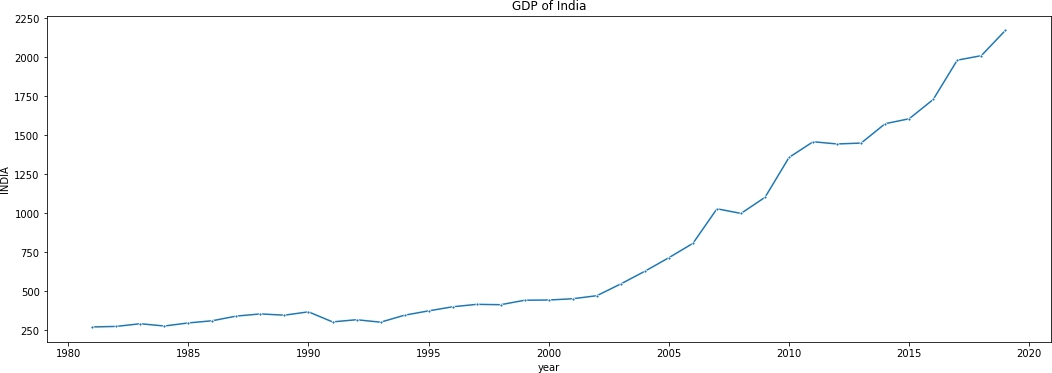


Fig.2.7

We see that the GDP of India with respect of USA is increasing over time though it has decrease from the years 1992-1994, fluctuate between the years 1995-2003 and then sudden increase from 2003-2009 and again fluctuate.

**Plot showing GDP of China over years:**

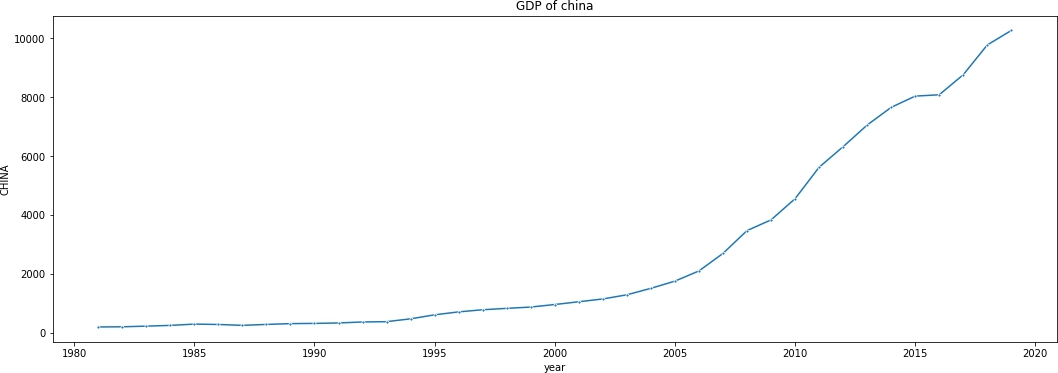
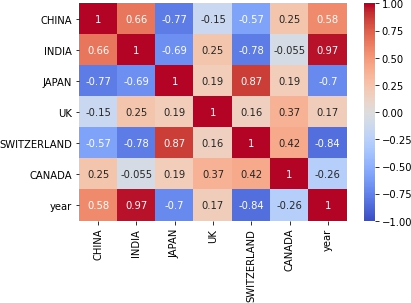


Fig.2.8

We see that the GDP of India with respect of USA is increasing exponentially over time from 2000-2020

* 1. **Foreign exchange rate dataset: 4.2.1. Correlation:**

The following figure shows the correlation matrix on a heatmap. We can derive the following insights from this correlation matrix.

1. **Heatmap with dark colors shows positive relationship between fers of different countries and light color shows the inverse relationship between countries.**
2. There is a high negative correlation between Switzerland and year.
3. China is neither correlated with japan nor with Switzerland
4. Japan has the lowest FER over years (negative correlated)
5. SWITZERLAND and JAPAN has the positive relationship over FER over the years followed by INDIA and CHINA similarly INDIA and SWITZERLAND has the inverse FER over years.

**Plot showing FER of each country over years from 1981-2019:**

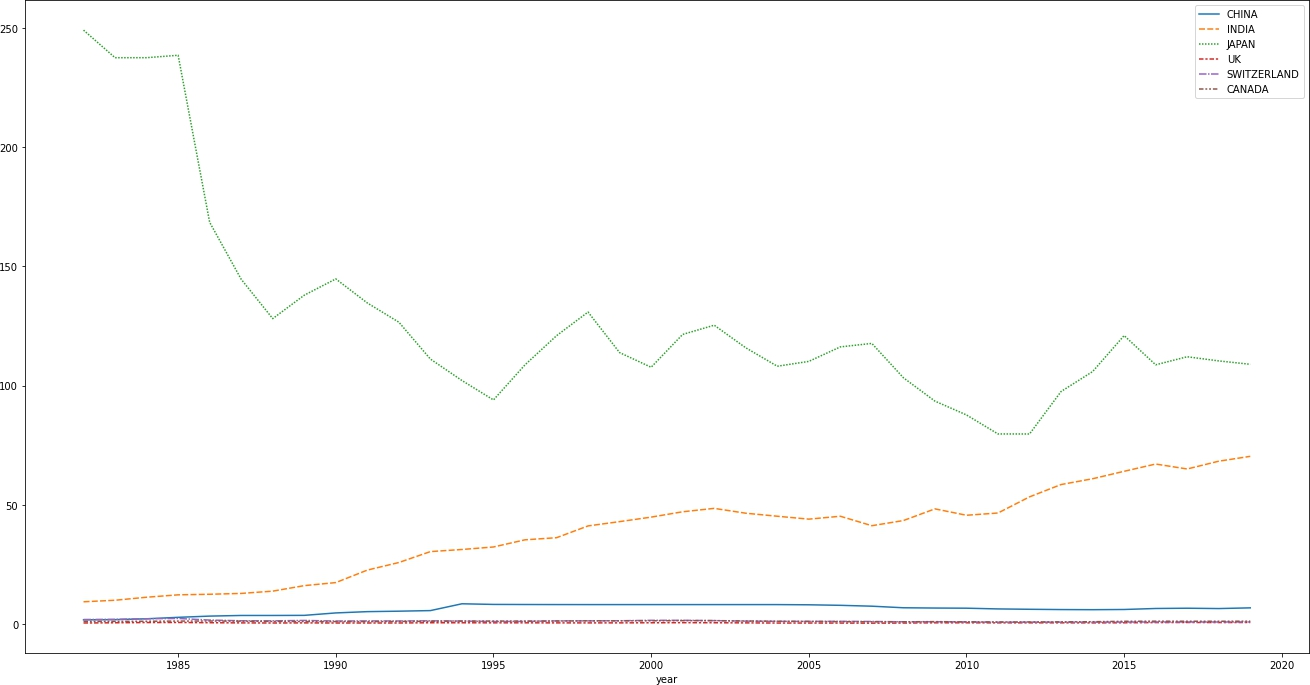


Fig.2.9

As we can see that the FER of japan has drastically fall from 1985 to 2000 and still it fluctuate over years from 2000 to 2020. FER Switzerland is constant through out the year. Also, India and china has increasing slope which it indicates that our model predicts the right output

**Plot showing FER of Canada over years:**

## 

Fig.2.10

**Plot showing FER of India over years:**

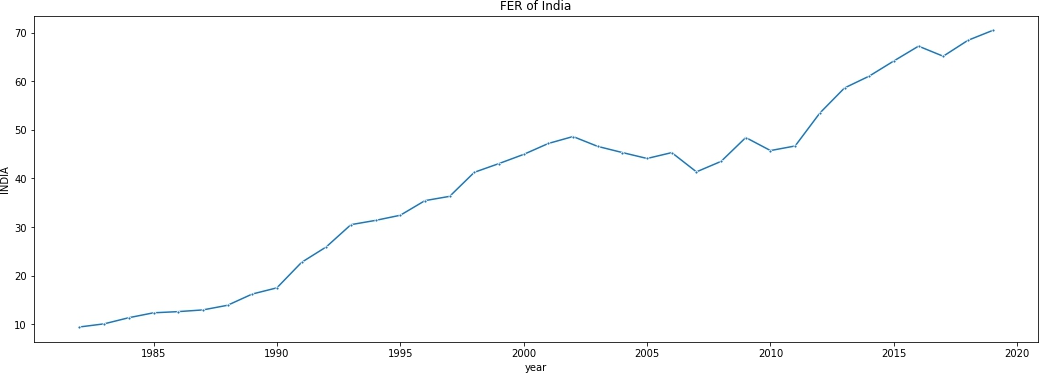


Fig.2.11

The line graph of FER of India shows that it keeps on fluctuating between increasing and decreasing of FER.

**Plot showing FER of Switzerland over years:**

## 

Fig.2.12

**Plot showing FER of Japan over years:**

## 

Fig.2.12

**Plot showing FER of UK over years:**

## 

Fig.2.14

**Plot showing FER of China over years:**

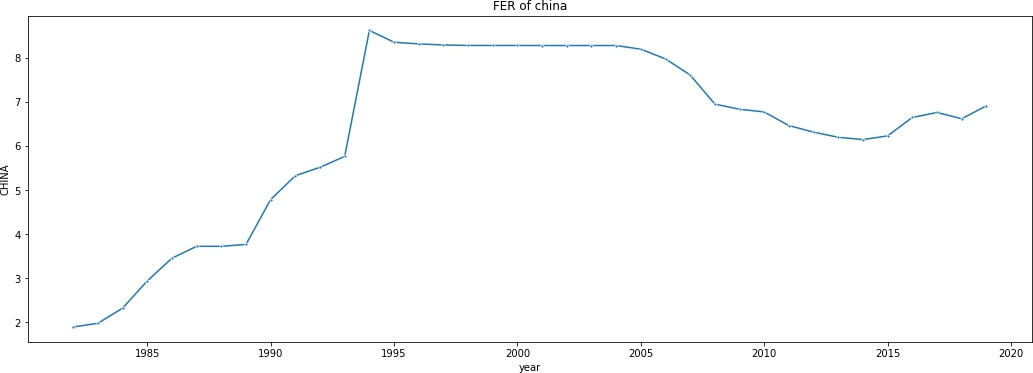


Fig.2.15

The line chart of GDP of China shows that it starts increasing from 1981- 1987 with a slight decrease in the year 1988 and again increase with much increase in 1989 and keeps on increasing till 1994 with a sudden decreases till year 1996 and remains stable till 2005. It fluctuate till 2015 and then it increases.

* 1. **Modelling**

We attempted to predict next year’s exchange rate using the popular models like Linear, Ridge, Lasso, Support Vector, Decision Tree and Random Forest regressors. We implemented these models using Scikit- Learn module. While selecting best accuracy model, we encountered with an issue of deflecting accuracy in every run even after fixing seed value. To solve this problem, we tried to select model using automated machine learning module called TPOT. TPOT is a tool that automatically creates and optimizes machine learning pipelines using genetic programming. We trained our data ( combination of economic factors and weights ) for every country and each topic number value and as a result we got optimal model pipelines with optimal parameters. We got best accuracy (mse) with weights having 15 topics.

Below is the mean error rates in % (only positive) for each country and every number of topics:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Count ry** | **Topics = 5** | **Topics = 10** | **Topics = 15** | **Topics = 20** | **Topics = 25** | **Topics = 30** | **Topics = 40** | **Topics**  **= 50** |
| **Canad a** | -0.002 | -0.002 | -0.004 | -0.002 | -0.002 | -0.003 | -0.002 | -0.002 |
| **China** | -0.045 | -0.213 | -0.094 | -0.227 | -0.223 | -0.247 | -0.021 | -0.193 |
| **India** | -3.849 | -2.780 | -4.773 | -3.005 | -4.446 | -2.014 | -4.67 | -3.325 |
| **Japan** | -139.321 | -228.697 | -213.008 | -185.098 | -178.335 | -58.269 | -69.139 | -129.13  3 |
| **Swiss** | -0.025 | -0.019 | -0.018 | -0.025 | -0.013 | -0.015 | -0.009 | -0.017 |
| **UK** | -0.00 | -0.002 | -0.002 | -0.001 | -0.002 | -0.001 | -0.001 | -0.001 |

The error plots are shown below :

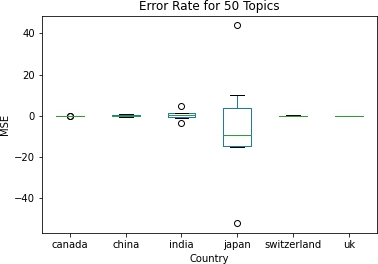
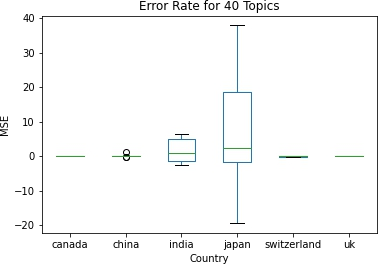
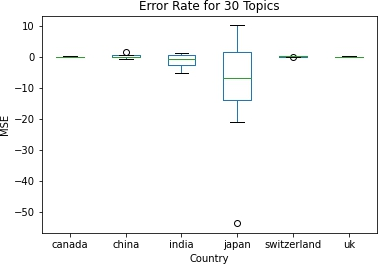
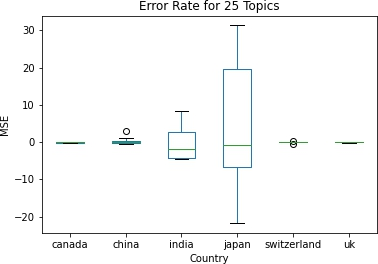
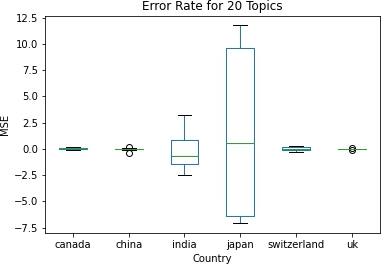
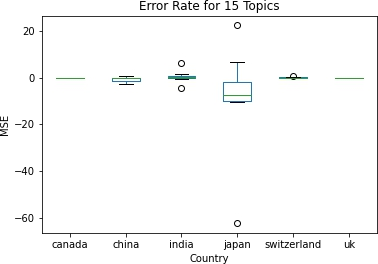
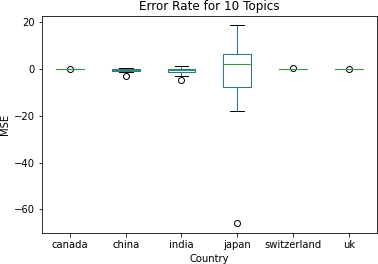
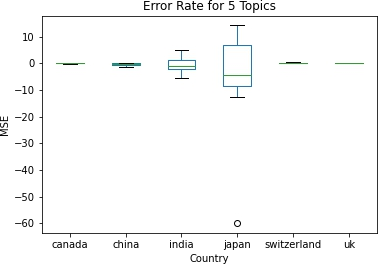


Fig.2.16. Error Plots for all topics

**CHAPTER 5**

**OPTIMIZATION OF MODEL**

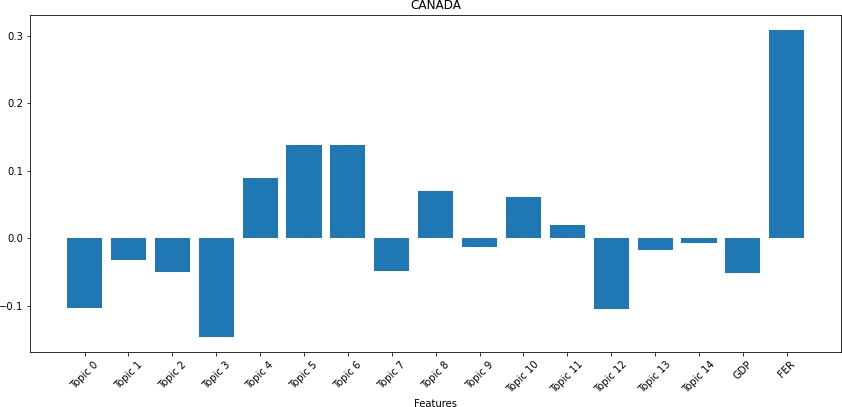
**5.1. Optimized Techniques**

The optimised models with number of topics as 15 with parameters for prediction are:

|  |  |  |
| --- | --- | --- |
| **Country** | **Models** | **Error Rate** |
| Canada | RidgeCV(), | 0.013% |
| China | AdaBoostRegressor(learning\_rate=0.1,  loss="square", n\_estimators=100) | 0.669% |
| India | LinearSVR(C=25.0, dual=False, epsilon=0.0001,  loss="squared\_epsilon\_insensitive", tol=1e-05) | 0.535% |
| Japan | LassoLarsCV(normalize=False), | 9.124% |
| Switzerland | AdaBoostRegressor(learning\_rate=0.001, loss="linear", n\_estimators=100) | 0.133% |

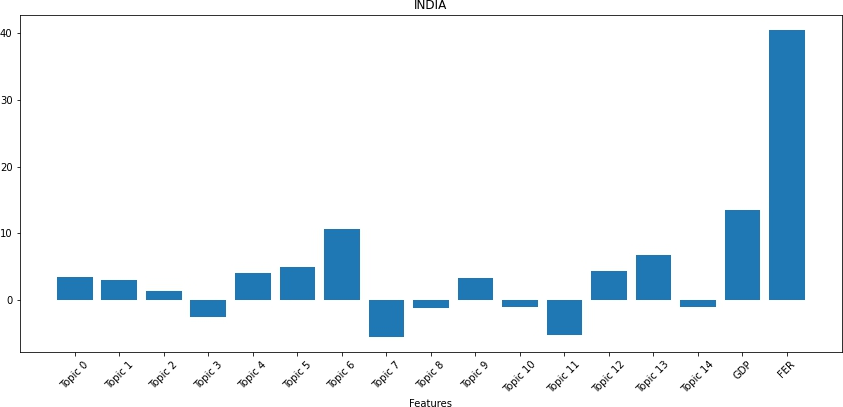
**5.2. Feature Importance**

* For canada, words come under Topic 5 & 6 has the more importance as these are the words which affect the FER most

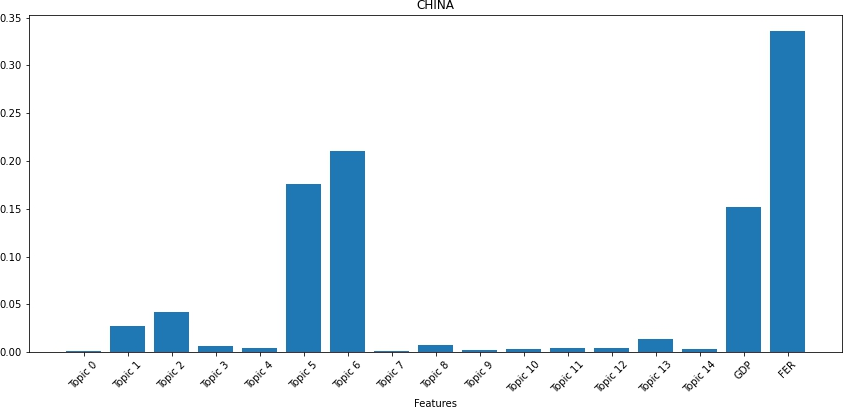


-

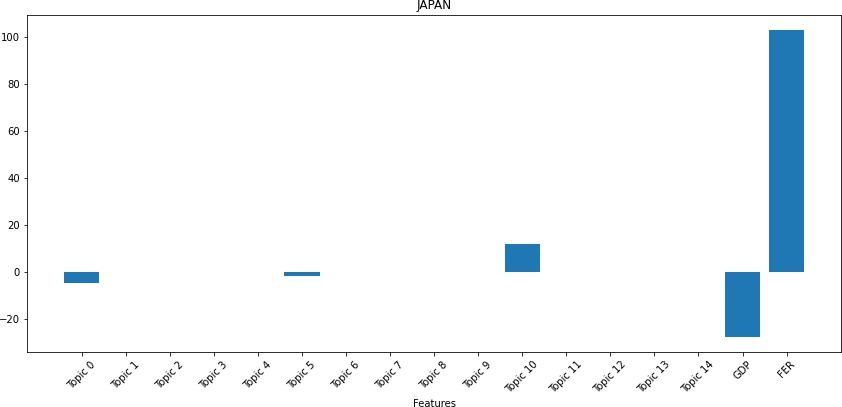
* For India, words come under Topic 6 has the more importance as these are the words which affect the FER most.



* For China, words come under Topic 5 & 6 has the more importance as these are the words which affect the FER most.



* For Japan, words come under Topic 10 has the more importance as these are the words which affect the FER most.



* For canada, words come under Topic 4 has the more importance as these are the words which affect the FER most

## 

## 5.3. Future Prospects

Considering the model which we have made by ML algorithms, there is very much scope of **Value Proposition of our model in future, since the idea which we have implemented is new and** no one else tried it till date.

In general terms, a value proposition comprises the value any organization promises to provide its customers, and based on that value the customers should prefer to purchase their products or services.

To Bring our project on web we have used HTML,CSS 3 for Frontend while backend is handled by Flask. We have used Heroku to deploy our project.

* The value proposition of our website is unique and consists of exchange rate metrics. It facilitates users with the foreign exchange rates over years, so that users can manage their trade.
* By holding the capital’s whole organization along with other features like inspection, currency rate, YOY comparison.
* Technology-driven features and strategies: Our website offers exact currency rate over years
* ECO Website: Our website facilitates users to get a currency inspection report which is objective-based and unbiased.
* Full circle trust score: Our website ensures the information regarding the currency exchange rate are trustworthy with whom they are dealing. It is very useful for critical trust factors i.e., verified exchange rates, auto inspection, capital’s dynamic flow. Traders can do more better objective assessment while purchasing a stock as it consists of 4 parameters to ease decision-making i.e., currency rates (country wise comparison), years, stock simulator and graphs
* This platform is useful for traders and consumers in calculating the fair market rates of goods in other countries. This is an unbiased and data-driven tool that provides a guaranteed 100% customer satisfaction. Using this tool, consumers can also check whether the vehicle is available in the market or not.
* This is a tool that benefits consumers with the pre-stocking research facility. It provides rate details, GDP, history of rates of respective countries. Also, using this tool, buyers can search any countries foreign exchange rates
* Findcars.com Assist: It helps buyers by guiding in buying stocks. It is basically an online assistance tool that helps users at every step of sale or purchase of stocks of other countries and other services.

We have used HTML,CSS 3 for Frontend while backend is handled by Flask. We have done this in following steps :

1. At first, we have pushed the code in Git hub. The link is attached below.
2. Following this, we have created an account on Heroku. Heroku is a cloud platform that lets companies build, deliver, monitor and scale apps — we're the fastest way to go from idea to URL, bypassing all those infrastructure headaches.
3. Linked the Github account account to Heroku
4. Deploy the model
5. Web app is ready

### Frontend

### HTML : The HyperText Markup Language, or HTML is the standard markup language for documents designed to be displayed in a web browser. Web browsers receive HTML documents from a web server or from local storage and render the documents into multimedia web pages. HTML describes the structure of a web page semantically and originally included cues for the appearance of the document. HTML can embed programs written in a scripting language such as JavaScript.

We have made a single HTML file i.e. index.html and used it showcase our project we are taking 3 inputs which are as follows:

1. From Country
2. To Country
3. Amount
4. Year

After entering these values we need to click the predict button which will take these values in backend and parse the predicted foreign exchange rates in front end.



**CSS :** Cascading Style Sheets (CSS) is a style sheet language used for describing the presentation of a document written in a markup language such as HTML. CSS is a cornerstone technology of the World Wide Web, alongside HTML and JavaScript. CSS is designed to enable the separation of presentation and content, including layout, colors, and fonts.This separation can improve content accessibility, provide more flexibility and control in the specification of presentation characteristics, enable multiple web pages to share formatting by specifying the relevant CSS in a separate.

We have used multiple CSS features like hover in our Predict button ,Animation in our output result .we have used background image ,google fonts ,multiple colours. We have used box model a box that wraps around every HTML element. It consists of: margins, borders, padding, and the actual content. The image below illustrates the box model:

**Final Deployment Link:** [http://currexc.herokuapp.com](http://currexc.herokuapp.com/)

**Raw Data drive link :** <https://drive.google.com/drive/folders/1p_yZYrH3SyXNqFYOf>D58VEuaxLCcjZJE

**GitHub link :** https://github.com/anisghosh10/Major-Group-7\_11

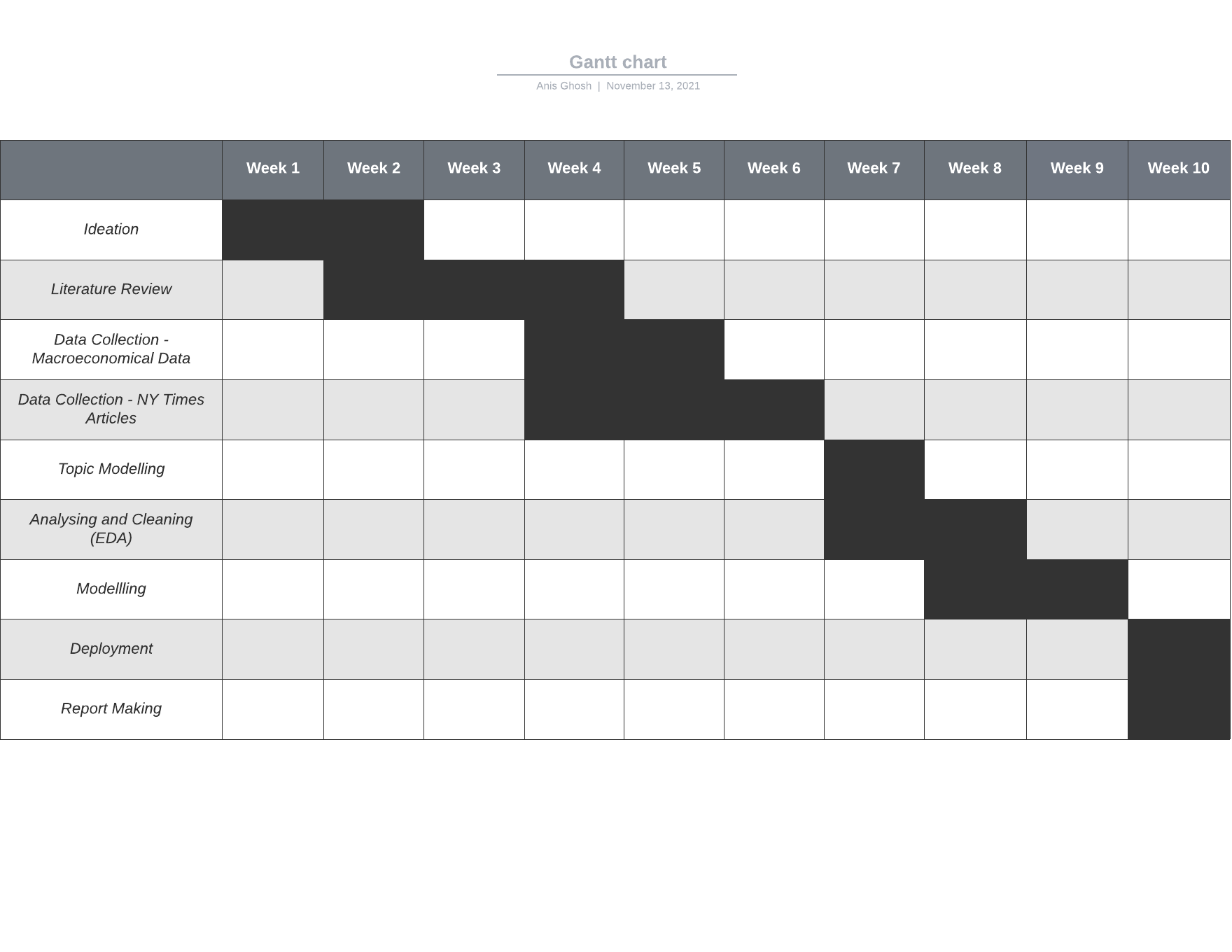
**5.4. Conclusion**

The machine learning (web scrapping) industry is considered one of the booming and evergreen sectors across the world especially in India where it holds a prominent position in the growth of the Indian economy. The industry provides useful insights into trading, automation, and majorly into data industry. This concept of ours is a growing and evolving concept that has resulted into many revolutionary ideas and methodologies and techniques that holds the potential to change our future and puts some light into futuristic automated machine driven societies. It already holds huge importance in the data driven industry and it will continue to increase its importance. A subset of machine learning is closely related to [computational statistics](https://en.wikipedia.org/wiki/Computational_statistics), which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of [mathematical optimization](https://en.wikipedia.org/wiki/Mathematical_optimization) delivers methods, theory and application domains to the field of machine learning. [Data mining](https://en.wikipedia.org/wiki/Data_mining) is a related field of study, focusing on [exploratory data](https://en.wikipedia.org/wiki/Exploratory_data_analysis) analysis through [unsupervised learning](https://en.wikipedia.org/wiki/Unsupervised_learning). In its application across business problems, machine learning is also referred to as [predictive analytics](https://en.wikipedia.org/wiki/Predictive_analytics)

**5.5. PLANNING AND PROJECT MANAGEMENT :**

|  |  |  |
| --- | --- | --- |
| **Activity** | **Starting week** | **Number of weeks** |
| Ideation | 1st week of August | 3 |
|
|
| Literature Review | Last week of August | 2 |
| Data Collection - Macro economical Data | 1st week of September | 2 |
| Data Collection - NY Times articles | 3rd week of September | 4 |
| Topic Modelling | 3rd week of October | 1 |
| EDA | 4th week of October | 1 |
| Modelling (Building) | 4th week of October | 1 |
| Evaluate the performance of our model | 1st week of November | 1 |
| Deployment | 1st week of November | 2 |
| Preparation of project report | 2nd week of November | 1 |
| Preparation of project presentation | 2nd week of November | 1 |

The Gantt chart is shown below:



## References:

<https://data.worldbank.org/indicator/pa.nus.fcrf>

[https://developer.nytimes.com](https://developer.nytimes.com/)

[https://developer.nytimes.com](https://developer.nytimes.com/)

[GitHub - HS189/FinancialArbitrage2: sadbois](https://github.com/HS189/FinancialArbitrage2)

[https://stackoverflow.com](https://developer.nytimes.com/)

<https://www.youtube.com/watch?v=mrExsjcvF4o>

[7] GONGGI, S., 2011. New model for residual value prediction of used cars based

on BP neural network and non-linear curve fit. In: Proceedings of the 3 rd IEEE International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), Vol 2. pp. 682-685, IEEE Computer Society, Washington DC, USA.

[8] LEXPRESS.MU ONLINE. 2014. Available from: http://www.lexpress.mu/ [Accessed 17 January 2014]

[9] LE DEFI MEDIA GROUP. 2014. Available from: http://www.defimedia.info/ [Accessed 17 January 2014]

1. GELMAN, A. AND HILL, J., 2006. Data Analysis Using Regression and Multilevel Hierarchical Models. Cambridge University Press, New York, USA.764 Sameerchand Pudaruth

[11] WEKA 4: DATA MINING SOFTWARE IN JAVA. 2014. Available from: http://www.cs.waikato.ac.nz/ml/weka/index.html [Accessed 17 January 2014].

1. LI, Y. H. AND JAIN, A. K., 1998. Classification of Text Documents. The Computer Journal, Vol. 41, pp. 537-546.
2. QUINLAN, J. R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann.

[14] MITCHELL, T. M., 1997. Machine Learning. McGraw-Hill, Inc. New York,

NY, USA.

[15] Strauss, Oliver Thomas, and Morgan Scott Hansen. "Advanced data science systems and methods useful for auction pricing optimization over network." U.S. Patent Application No. 15/213,941.

[16] Xinyuan Zhang , Zhiye Zhang and Changtong Qiu, “Model of Predicting the Price Range of Used Car”, 2017

[17] W.A. Awad and S.M. ELseuofi, “Machine Learning Method for SpamEmail Classification”, 2011

[18] Durgesh K. Srivastava, Lekha Bhambhu, “Data Classification Method using Support Vector Machine”, 2009

[19] Pudaruth, Sameerchand. "Predicting the price of used cars using machine learning techniques." Int. J. Inf. Comput. Technol 4.7 (2014): 753-764.

[20] Noor, Kanwal, and Sadaqat Jan. "Vehicle Price Prediction System using Machine Learning Techniques." International Journal of Computer Applications 167.9 (2017).

[21] Kuiper, Shonda. "Introduction to Multiple Regression: How Much Is Your Car Worth?." Journal of Statistics Education 16.3 (2008)

1. <http://ripublication.com/irph/ijict_spl/ijictv4n7spl_17.pdf>

[23] Gareth, J., Daniela, W., Trevor, H., & Tibshirani, R. (2013). An Introduction to Statistical

[24] Learning (Vol. 8). <https://doi.org/10.1016/j.peva.2007.06.006>

[25] Hurwitz, E., & Marwala, T. (2012). Common mistakes when applying computational intelligence and machine learning to stock market modelling. arXiv preprint arXiv:1208.4429.

[26] Pal, N., Arora, P., Kohli, P., Sundararaman, D., & Palakurthy, S. S. (2018, April). How Much Is My Car Worth? A Methodology for Predicting Used Cars’ Prices Using Random Forest. In Future of Information and Communication Conference (pp. 413–422). Springer, Cham.

[27] Raschka, S., & Mirjalili, V. (2017). Python machine learning. Packt Publishing Ltd.

[28] Agencija za statistiku BiH. (n.d.), retrieved from: http://www.bhas.ba . [accessed July 18, 2018.]

[29] Listiani, M. (2009). Support vector regression analysis for price prediction in a car leasing application (Doctoral dissertation, Master thesis, TU Hamburg-Harburg).

[30] Richardson, M. S. (2009). Determinants of used car resale value. Retrieved from: https://digitalcc.coloradocollege.edu/islandora/object /coccc%3A1346 [accessed: August 1, 2018.] [4] Wu, J. D., Hsu, C. C., & Chen, H. C. (2009). An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. Expert Systems with Applications, 36

[31] 7809-7817. [5] Du, J., Xie, L., & Schroeder, S. (2009). Practice Prize Paper—PIN Optimal Distribution of Auction Vehicles System: Applying Price Forecasting, Elasticity Estimation, and Genetic Algorithms to Used-Vehicle Distribution. Marketing Science, 28(4), 637-644.

[32] Gongqi, S., Yansong, W., & Qiang, Z. (2011, January). New Model for Residual Value Prediction of the Used Car Based on BP Neural Network and Nonlinear Curve Fit. In Measuring Technology and Mechatronics Automation (ICMTMA), 2011 Third International Conference on (Vol. 2, pp. 682-685). IEEE.

[33] Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. Int. J. Inf. Comput. Technol, 4(7), 753-764.

[34] Noor, K., & Jan, S. (2017). Vehicle Price Prediction System using Machine Learning Techniques. International Journal of Computer Applications, 167

[35] , 27-31. [9] Auto pijaca BiH. (n.d.), Retrieved from: https://www.autopijaca.ba. [accessed August 10, 2018].

[36] Weka 3 - Data Mining with Open Source Machine Learning Software in Java. (n.d.), Retrieved from: https://www.cs.waikato.ac.nz/ml/weka/. [August 04, 2018].