Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

A Regression Approach to Distribution and Trend Analysis of Quarterly Foreign Tourist Arrivals in India

Navoneel Chakrabarty
Jalpaiguri Government Engineering College
Jalpaiguri, West Bengal, India
E mail: nc2012@cse.jgec.ac.in

Abstract: International Tourism has been a very important contributor to a country's economic development. In developing countries like Argentina, Brazil, India etc., the tourism industry plays an important role in the Gross Domestic Product and Foreign Exchange Earnings. Now a days, India has been welcoming a highly impressive number of foreign tourists from all round the globe annually. This study aims at analysing the distribution and trend of foreign tourists visiting India, among the four quarters of a year, given different configurations of Gross Domestic Product and Foreign Exchange Earnings using Machine Learning and Regression Analysis. A 4 headed Machine Learning Model has been constructed and trained independently for the purpose. Finally, the results of the four individually trained sub models are collected together for Trend Anal ysis and Distribution Analysis. This final evaluation is done for the year 2012 post to Model Construction, Training, Tuning and Individual Validations of the 4 sub models. It has been found that the Distribution and Trend Analysis have been almost similar to the Original Distribution and Trends of Foreign Tourists among the four quarters of 2012. This similarity in Distribution Analysis has been shown using visualizations like Pie Chart and that in Trend Analysis has been shown using Line Plots.

Keywords: International Tourism • Gross Domestic Product • Foreign Exchange Earnings • Machine Learning • Regression Analysis • Trend Analysis Distribution Analysis

Introduction

In recent years, India has seen the growth of Tourism Industry by leaps and bounds. Its rich flora and fauna and 25 world famous destinations are the eye-catching elements for International Travellers. Tourists from countries like Bangladesh, United States, United Kingdom, Canada, Malaysia and Sri Lanka are the maximum visitors in India. As the Indian Tourism Industry has flourished significantly, India has been ranked 7th in Asia Pacific in World Tourism [1] and 24th in the world for its cultural resources and World Heritage sites. The quarter wise trend and distribution analysis of Foreign Tourist Arrivals also gives an idea of the favourable sea sons and weather conditions preferred by overseas visitors for their visit to India on an average. For example, if according to the FTA Distribution for a particular year, it is found that there has been more fraction of tourists visiting India in 1st Quarter i.e., January to March, then it can be

ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

interpreted that in that year, foreign tourists preferred Winter Season as ideal weather for sightseeing in India. As India is a developing nation, it has a strong inter dependency between Gross Domestic Product and Foreign Tourism Demand. In other words, GDP serves as an important determinant of Foreign Tourism Demand and hence, Foreign Tourist Arrivals can also be determined almost accurately. The GDP is categorized sector wise, each having a strong inter dependency with International Tourism: Gross domestic product at market prices output approach Gross value added at basic prices, total activity Agriculture, and forestry and fishing Industry, including energy manufacturing. Construction Services Distributed trade, repairs, transportation, accommodation, food services and activities Real estate activities Public admin, education, human health. So, with the distribution and trend analysis, it can also be shown that how much inter dependent the GDP and Foreign Tourism are, in India. Likewise, analysis of Quarterly Foreign Tourist Distribution is also helpful for the Tourism Industry for proper maintenance of the Heritage Sites of India and hence preserving the rich flora and fauna to behold them in front of foreigners. This paper has been structured as follows: Introduction, Literature Review, Proposed Methodology, Implementation Details, Sub Model Performance Analysis, Distribution and Trend Analysis Evaluation and Conclusion.

Literature Review

Several Approaches have been applied by Researchers and Statisticians for predicting Tourist Arrival Patterns.Sun et al. [2] applied Machine Learning and used Internet Search Index for forecasting Tourist Arrivals and compared the forecasting performance of 2 different search engines, Baidu and Google.Hugo David dos Reis Barbosa Ricardo [3] proposed a Data Mining Approach to forecast tourism de mand for Lisbon's Region Claveria et al. [4] focussed on forecast horizon and its influence on forecasting performance of Machine Learning Algorithms, Support Vector Regression and Neural Networks for predicting Tourism Demand in Spain. Yi Chung Hu [5] developed a Predictive Model for determining Foreign Tourists for the Tourism Industry in China and Taiwan using Soft Computing based Grey Markov Models. Oscar Garc ıa Rodr´ıguez [6] used Online Search Engine Data in the form of queries for forecasting tourism arrivals to Balearic Islands. Biljana Petrevska [7] performed a Time Series Modelling, applying Autoregressive Integrated Moving Average (ARIMA) for Tourism Demand Prediction in Former Yugoslav Republic of Macedonia. Yu et al. [8] applied a Deep Learning Approach for Statistical Modelling and Prediction for Tourism Economy using Dendritic Neural Network. Ali et al. [9] implemented Support Vector Machinesand Artificial Neural Networks for modelling Singapore Tourist Arrivals to Malaysia.Noersasongko et al. [10] proposed a Deep Learning Model for Tourism Arrival Forecasting using Genetic Algorithm based Neural Network. Cankurt et al. [11] used Machine Learning Techniques in developing models for tourism demand forecasting in Turkey with trend analysis, seasonal and cyclic components. Neupane et al. [12] proposed an empirical study for modelling Monthly Foreign Tourist Arrivals and its risk in Nepal.

Proposed Methodology

ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

The data for the analysis was extracted from Quarterly Foreign Tourist Arrivals in India Dataset or FTA Dataset, publicly available at Kaggle [13]. The dataset includes 4 sub data sets, each representing the determinants of Foreign Tourist Arrivals and Foreign Tourist Arrivals for a quarter of every year (3month tire) from 2005 to 2016. Every sub dataset consists of 41 features, which are the factors affecting Foreign Tourist Influx and the numeric value of Foreign Tourist Arrivals in each quarter i.e., from January to March, April to June, July to September and October to December. All the 41 features are continuous variables containing information on different categories of Gross Domestic Product and their subcategories (in Indian Rupee Billions), along with Foreign Exchange Earnings (in Indian Ru pee Crores) as shown in Table 1.Here, each of these configurations of GDP is called Category of GDP, having their share in multiple sectors, which are called Sub Categories of GDP. A Regression Approach to Distribution and Trend Analysis of Quarterly Foreign Tourist Arrivals in India 3

ID	Feature Name		Feature Description
1	GDP Configurations	CQRSA	National currency, current prices, quarterly levels, seasonally adjusted
2		CQR	National currency, current prices, quarterly level
3		VNBQRSA	National currency, constant prices, national base year, quarterly levels, seasonally adjusted
4		VNBQR	National currency, constant prices, national base year, quarterly levels
5	Foreign Exchange Earnings		Earnings made from the profits incurred by selling goods and services to foreign visitors in the economy out of the foreign currency brought by them

Table 1. Feature Headers

3.1 Analysis of 1st Quarter i.e., January to March

3.1.1 Trend Analysis of Foreign Tourist Arrivals in the 1st Quarter from 2005 to 2016

ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

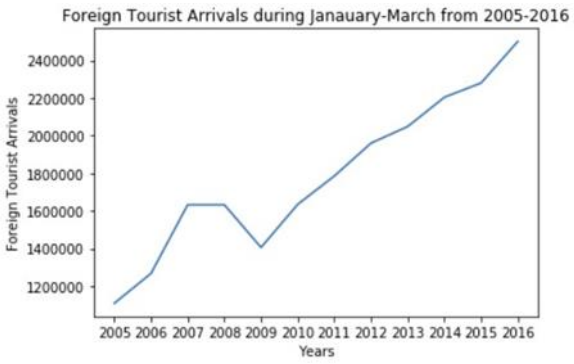


Fig 1. Yearly Trends of Foreign Tourist Arrivals in India during the 1st Quarter

From Fig 1, it is evident that from 2005 to 2007, there has been a gradual increase in FTA in India which became constant for a year, up to 2008. Then, there was a drop in FTA from 2008 to 2009 and since then FTA has been strictly increasing till 2016.

3.1.2 Feature Selection and Verification

As there are 41 features for only 12 instances (2005 to 2016), feature selection is very essential. The Random Forest Regressor is used for feature selection as per feature importances returned by it. The features are assigned importance's as per their selection as best split in the ensemble of decision trees.

The Algorithm for Random Forest Regressor is given below:
Algorithm 1. Random Forest Regressor
-Input: training set Z = {(x1,y1),...,(xn,yn)}
-B: number of iterations
-Tb(x): bth Decision Tree
1. for b=1...B:
2. A bootstrap sample of size n is drawn from Z.
3. A decorrelated decision tree, Tb(x) is grown.
4. f(x) =PB
b=1 Tb(x)

Out of 41 features in the dataset, for the 1st Quarter, 4 most important features are selected as per Random Forest Regressor (with 10 estimators) Feature Importance Scores shown in Table 2.

ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

ID	Feature Name		Random Forest Regressor Score
1	CQRSA	market prices - output approach	0.08874022
		basic prices and total activity	0
		agriculture, forestry and fishing	0
		industry, including energy	0.00789719
		manufacturing	0.02209012
		construction	0.00098083
		services	0.01514922
		distributed trade, repairs, transportation, accommodation, food services and activities	0.0126678
		real estate activities	0.0094151
		public admin, education, human health.	0.00294061
2	CQR	market prices - output approach	0.16452942
		basic prices and total activity	0
		agriculture, forestry and fishing	0.00173814
		industry, including energy	0.01782935
		manufacturing	0
		construction	0.00312828
		services	0.07594435
		distributed trade, repairs, transportation, accommodation, food services and activities	0.00606703
		real estate activities	0.09953914
		public admin, education, human health.	0
3	VNBQRSA	market prices - output approach	0
		basic prices and total activity	0.00210095
		agriculture, forestry and fishing	0.00036553
		industry, including energy	0.08014486
		manufacturing	0.00205321
		construction	0
		services	0.00205321
		distributed trade, repairs, transportation, accommodation, food services and activities	0.00242961
		real estate activities	0.01210568
		public admin, education, human health.	0.02733832
4	VNBQR	market prices - output approach	0.07288322
		basic prices and total activity	0.08058026
		agriculture, forestry and fishing	0
		industry, including energy	0.00073515
		manufacturing	0
		construction	0.00181453
		services	0.00453129
		distributed trade, repairs, transportation, accommodation, food services and activities	0
		real estate activities	0
		public admin, education, human health.	0.08262048
5	Foreign Exchange Earnings		0.10107216

Table 2. Feature Study for 1st Quarter

From Table 2, 4 most important features are found to be:

- National currency, current prices, quarterly levels' Gross domestic product at market prices output approach
- Foreign Exchange Earnings
- National currency, current prices, quarterly levels' Real Estate Activities
- National currency current prices, quarterly levels' Gross domestic product at market prices output approach (seasonally adjusted)

Correlation Matrix is shown in Fig 2, in the form of a Heat-Map displaying Feature-to-Feature and Feature-
to-Label Pearson Correlations where the features are the selected features from Random Forest Regressor's
Feature Importance Scores.

ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

Fig 2. Heat-Map showing Feature-to-Feature and Feature-to-Label's Pearson Correlation Coefficients in the Analysis of 1st Quarter

From the heat-map in Fig 2, it is verified by Pearson Correlation Coefficients, that the selected features have high correlation values with the Target (FTA) values. A Scatter-Plot of FTA is shown to depict its dependency on the Most Important Feature (National currency, current prices, quarterly levels' Gross domestic product at market prices - output approach) in Fig 3.

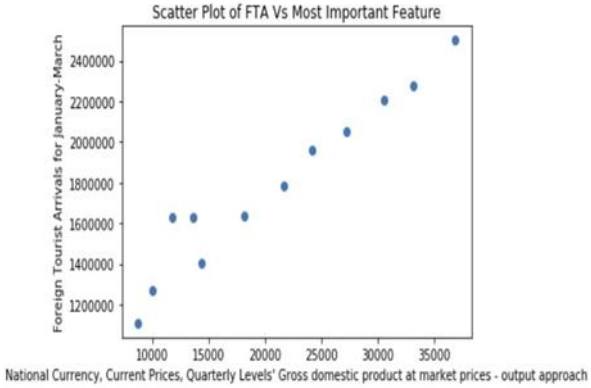


Fig 3. Scatter-Plot showing the dependency of FTA in the 1st Quarter with CQR's Gross Domestic Product at market prices – output approach

3.1.3 Data Pre-processing

1. Shuffling: The whole dataset is shuffled in a consistent way such that information regarding the features and FTA values of different years remain presenting the Training Set and Validation Set.
2. Splitting: Now, the dataset is split into Training and Validation Sets, with 80% data (here 9 instances), made available for Training Purposes and 20% data (here 3 instances) in the Validation Set.
3. Feature and Target-Value Scaling: All the 4 selected features and FTA values are Standard Scaled by the formula:

ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

x = (x - u)/sigma

where,
- u is the Mean of all the data-points in the training set belonging to a particular attribute.
  u = (x1 + x2 + x3….xn)=n
- sigma is the Standard Deviation of all the data-points in the training set belonging to a particular attribute.

$$sigma = \sqrt{\sum_{i-1}^{n} (x_i - u)^2 / n} \quad (1)$$

x is the considered data-point belonging to a particular attribute. The scaling procedure is done for the data-points in the Validation Set also by taking the same Mean and Standard Deviation from the Training Set for different attributes

3.1.4 Learning Algorithm

The learning algorithm used to build the Regression Model is an Ensemble Learning and Boosting Algorithm known as Gradient Boosting Regressor. It combines the outputs of many weak regressors (decision tree regressors) to produce a powerful ensemble (this process is boosting) using Gradient Descent Minimization of the Target Function in the Functional Space. The Algorithm for Gradient Boosting Regressor is given below:

- Input: training set Z = {(x1,y1),...,(xn,yn)}
- M: number of iterations
- v: learning rate

1. $f_0(x) = (1/n) * \sum_{i=1}^{n} y_i$
2. for m=1...M:
3. $y^p = y - f_{m-1}(x)$ (residual)
4. A decision tree $h_m(x)$ is fitted to the targets, y^p
5. $f_m(x) = f_{m-1}(x) + v * h_m(x)$
6. return $f_M(x)$

Algorithm 2. Gradient Boosting Regressor

3.1.5 Training the Model

The hyper-parameters of the Gradient Boosting Regressor, number of estimators (regression trees), maximum depth and learning rate are tuned using Grid-Search. The best set of hyper-parameters so

ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

found, were 8 regression trees, maximum depth of 3 and learning rate of 0.3.The resulting model has been shown in Fig 4 for the 8th Regression Tree which is the most correct one among the 8 regression trees as at each step, the error in the previous regression tree gets corrected by the following regression tree.

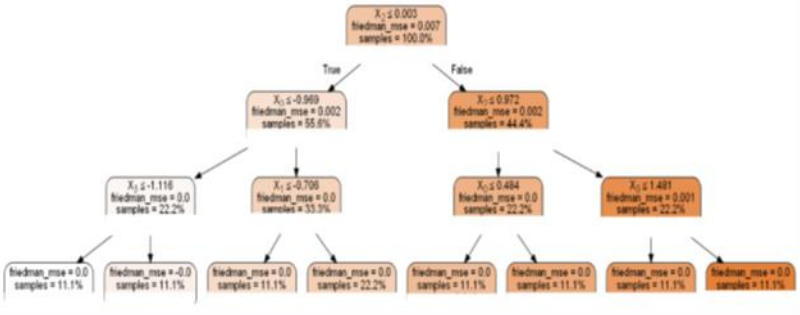


Fig 4. Structure of the 8th Regression Tree in GBR for the 1st Quarter Analysis

The summary of the Grid-Search hyper-parameter tuning is shown in Fig 5. It depicts the variation of Grid Search Mean Score with Max-Depth for different learning rates for 8 estimators.

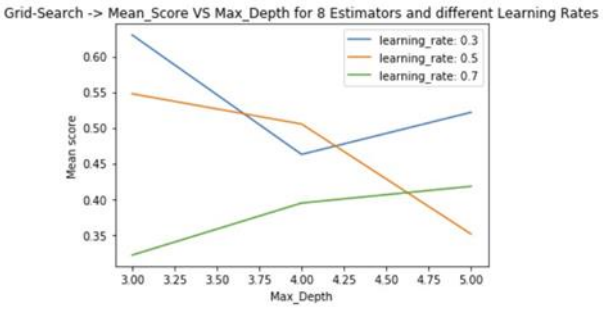


Fig 5. Grid Search Summary on Mean Score for the 1st Quarter Analysis

ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

3.2 Analysis of 2nd Quarter i.e., April to June

3.2.1 Trend Analysis of Foreign Tourist Arrivals in the 2nd Quarter from 2005 to 2016

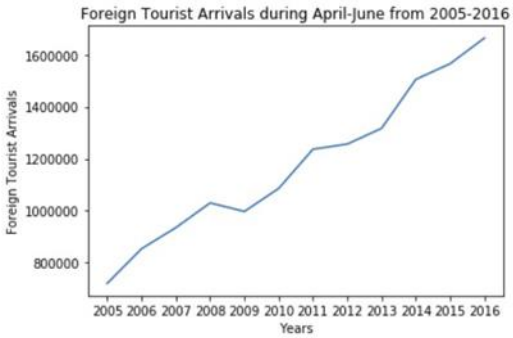


Fig 6. Yearly Trends of Foreign Tourist Arrivals in India during the 2nd Quarter

From Fig 6, it is evident that from 2005 to 2008, there has been a gradual increase in FTA in India. It had dropped for a year till 2009. Since then FTA has been strictly increasing at varying rates till 2016

3.2.2 Feature Selection and Verification.

Out of 41 features in the dataset, for the 2nd Quarter, 6 most important features are selected as per Random Forest Regressor (with 10 estimators) Feature Importance Scores shown in Table 3.
From Table 3, 6 most important features are found to be:

– National currency, constant prices, national base year, quarterly levels seasonally adjusted's Construction.
– National currency, current prices, quarterly levels' Public admin; education; human health
– National currency, constant prices, national base year, quarterly levels, seasonally adjusted's Manufacturing
– Foreign Exchange Earnings.
– National currency, constant prices, national base year, quarterly levels, seasonally adjusted's Distrib. trade, repairs; transp.; accommod., food serv. Activ
– National currency, current prices, quarterly levels, seasonally adjusted's Manufacturing.

ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

ID	Feature Name		Random Forest Regressor Score
1	CQRSA	market prices - output approach	0.0019547
		basic prices and total activity	0
		agriculture, forestry and fishing	0.01432205
		industry, including energy	0.07232376
		manufacturing	0.07584391
		construction	0
		services	0.00869051
		distributed trade, repairs, transportation, accommodation, food services and activities	0.0014967
		real estate activities	0.02185489
		public admin, education, human health.	0.00683048
2	CQR	market prices - output approach	0
		basic prices and total activity	0.00032274
		agriculture, forestry and fishing	0
		industry, including energy	0.00080167
		manufacturing	0.00082824
		construction	0
		services	0
		distributed trade, repairs, transportation, accommodation, food services and activities	0.00073219
		real estate activities	0.00934275
		public admin, education, human health.	0.16694698
3	VNBQRSA	market prices - output approach	0
		basic prices and total activity	0.00927072
		agriculture, forestry and fishing	0.00297263
		industry, including energy	0.00687398
		manufacturing	0.09932957
		construction	0.24066903
		services	0
		distributed trade, repairs, transportation, accommodation, food services and activities	0.09378982
		real estate activities	0
		public admin, education, human health.	0.00118568
4	VNBQR	market prices - output approach	0.01404005
		basic prices and total activity	0.01155724
		agriculture, forestry and fishing	0.01077851
		industry, including energy	0
		manufacturing	0
		construction	0.00283957
		services	0.01292117
		distributed trade, repairs, transportation, accommodation, food services and activities	0.01411048
		real estate activities	0
		public admin, education, human health.	0
5	Foreign Exchange Earnings		0.09736998

Table 3. Feature Study for 2nd Quarter

A Correlation Matrix is shown in Fig 7, in the form of a Heat-Map displaying Feature-to-Feature and Feature to-Label Pearson Correlations where the features are the selected features from Random Forest Regressor's Feature Importance Scores.

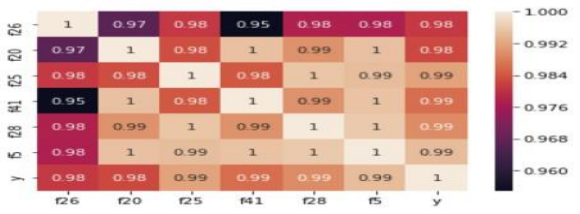


Fig 7. Heat-Map showing Feature-to-Feature and Feature-to-Label's Pearson Correlation Coefficients in the Analysis of 2nd Quarter

ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

From the heat-map in Fig 7, it is verified by Pearson Correlation Coefficients, that the selected features have high correlation values with the Target (FTA) values. A Scatter-Plot of FTA is shown to depict its dependency on the Most Important Feature (National currency, constant prices, national base year, quarterly levels seasonally adjusted's Construction) in Fig 8.

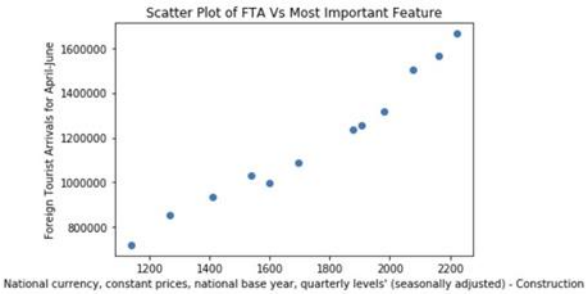


Fig 8. Scatter-Plot showing the dependency of FTA in the 2nd Quarter with VNBQRSA's Construction

3.2.3 Data Pre-processing

The Data Pre-Processing is done by following the same methodology steps as done in the Analysis of 1st Quarter.

3.2.4 Learning Algorithm

The learning algorithm used to build the Regression Model is an Ensemble Learning and Boosting Algorithm known as Gradient Boosting Regressor. The Algorithm of the Gradient Boosting Regressor is shown as Algorithm 2.

3.2.5 Training the Model

The hyper-parameters of the Gradient Boosting Regressor, number of estimators (regression trees), maximum depth and learning rate are tuned using Grid-Search. The best set of hyper-parameters so found, were 29 regression trees, maximum depth of 3 and learning rate of 0.1.

The resulting model has been shown in Fig 9 for the 29th Regression Tree which is the most correct one among the 29 regression trees as at each step, the error in the previous regression tree gets corrected by the following regression tree.

ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

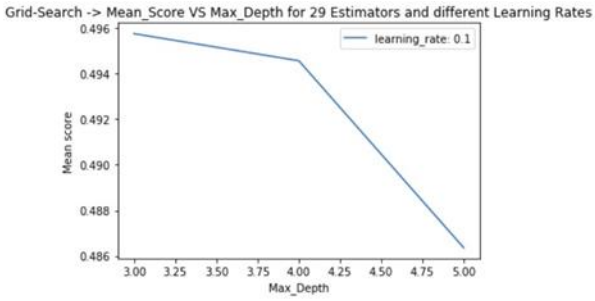


Fig 9. Structure of the 29th Regression Tree in GBR for the 2nd Quarter Analysis

The summary of the Grid-Search hyper-parameter tuning is shown in Fig 10. It depicts the variation of Grid Search Mean Score with Max-Depth for the learning rate of 0.1 and for 29 estimators.

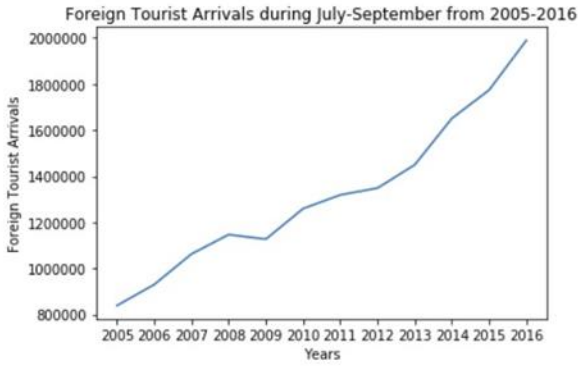


Fig 10. Grid Search Summary on Mean Score for the 2nd Quarter Analysis

The summary of the Grid-Search hyper-parameter tuning is shown in Fig 10. It depicts the variation of Grid Search Mean Score with Max-Depth for the learning rate of 0.1 and for 29 estimators.

3.3 Analysis of 3rd Quarter i.e., July to September

3.3.1 Trend Analysis of Foreign Tourist Arrivals in the 3rd Quarter from 2005 to 2016

3.3.2 Feature Selection and Verification

Out of 41 features in the dataset, for the 3rd Quarter, 6 most important features are selected as per Random Forest Regressor (with 10 estimators) Feature Importance Scores shown in Table 4.
From Table 4., 6 most important features are found to be:
– National currency, current prices, quarterly levels' Distrib. trade, repairs; transp.; accommod., food serv. Activ.
– National currency, constant prices, national base year, quarterly levels' Agriculture, forestry and fishing
– National currency, constant prices, national base year, quarterly levels, seasonally adjusted's Agriculture, forestry and fishing

ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

– National currency, current prices, quarterly levels, seasonally adjusted's Services
– National currency, constant prices, national base year, quarterly levels, seasonally adjusted's Gross domestic product at market prices - output approach
– National currency, current prices, quarterly levels' Industry, including energy
A Correlation Matrix is shown in Fig 12, in the form of a Heat-Map displaying Feature-to-Feature and Feature-to-Label Pearson Correlations where the features are the selected features from Random Forest Regressor's Feature Importance Scores.

ID	Feature Name		Random Forest Regressor Score
1	CQRSA	market prices - output approach	1.04205535e-02
		basic prices and total activity	2.79253121e-02
		agriculture, forestry and fishing	3.42270352e-04
		industry, including energy	7.78076778e-02
		manufacturing	1.53341523e-02
		construction	1.63291797e-02
		services	8.35133434e-02
		distributed trade, repairs, transportation, accommodation, food services and activities	7.80066904e-02
		real estate activities	5.91625531e-04
		public admin, education, human health.	3.07478747e-03
2	CQR	market prices - output approach	0
		basic prices and total activity	1.52085816e-02
		agriculture, forestry and fishing	1.81977516e-05
		industry, including energy	7.97976205e-02
		manufacturing	5.80398042e-04
		construction	0
		services	3.34796872e-03
		distributed trade, repairs, transportation, accommodation, food services and activities	9.98740822e-02
		real estate activities	0
		public admin, education, human health.	2.50481558e-03
3	VNBQRSA	market prices - output approach	2.35772512e-04
		basic prices and total activity	4.90342508e-04
		agriculture, forestry and fishing	8.49871221e-02
		industry, including energy	7.13619132e-02
		manufacturing	3.53793850e-02
		construction	5.31355706e-04
		services	6.26931422e-04
		distributed trade, repairs, transportation, accommodation, food services and activities	1.26234968e-02
		real estate activities	3.60874137e-03
		public admin, education, human health.	6.84714626e-02
4	VNBQR	market prices - output approach	8.01600203e-02
		basic prices and total activity	1.62267626e-03
		agriculture, forestry and fishing	8.53151107e-02
		industry, including energy	1.27514193e-02
		manufacturing	2.11950014e-04
		construction	1.18586385e-02
		services	3.89689090e-04
		distributed trade, repairs, transportation, accommodation, food services and activities	1.72608067e-03
		real estate activities	1.29466472e-02
		public admin, education, human health.	0
5	Foreign Exchange Earnings		2.39876577e-05

Table 4. Feature Study for 3rd Quarter

ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

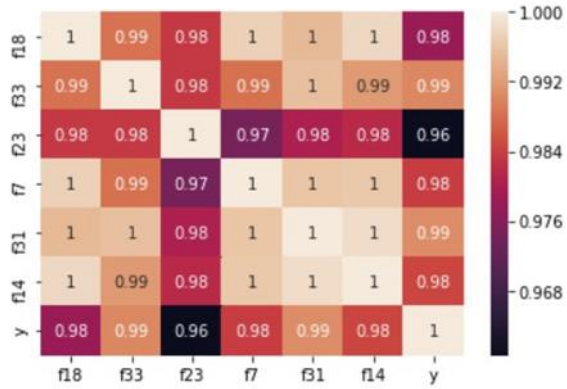


Fig 12. Heat-Map showing Feature-to-Feature and Feature-to-Label's Pearson Correlation Coefficients in the Analysis of 3rd Quarter

From the heat-map in Fig 12 , it is verified by Pearson Correlation Coefficients, that the selected features have high correlation values with the Target (FTA) values. A Scatter-Plot of FTA is shown to depict its dependency on the Most Important Feature (National currency, current prices, quarterly levels'-Distrib. trade, repairs; transp.; accommod., food serv. Activ.) in Fig 13

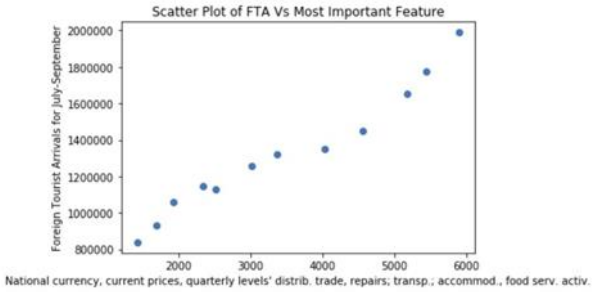


Fig 13. Scatter-Plot showing the dependency of FTA in the 3rd Quarter with CQR's Distrib. trade, repairs; transp.; accommod., food serv. Activ.

3.3.3 Data Pre-processing

The Data Pre-Processing is done by following the same methodology steps as done in the Analysis of 1st Quarter.

3.3.4 Learning Algorithm

The learning algorithm used to build the Regression Model is an Ensemble Learning and Boosting Algorithm known as Gradient Boosting Regressor. The Algorithm of the Gradient Boosting Regressor is shown as Algorithm 2.

ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

3.3.5 Training the Model

The hyper-parameters of the Gradient Boosting Regressor, number of estimators (regression trees), maximum depth and learning rate are tuned using Grid-Search. The best set of hyper-parameters so found, were 28 regression trees, maximum depth of 3 and learning rate of 0.1.

The resulting model has been shown in Fig 14 for the 28th Regression Tree which is the most correct one among the 28 regression trees as at each step, the error in the previous regression tree gets corrected by the following regression tree.

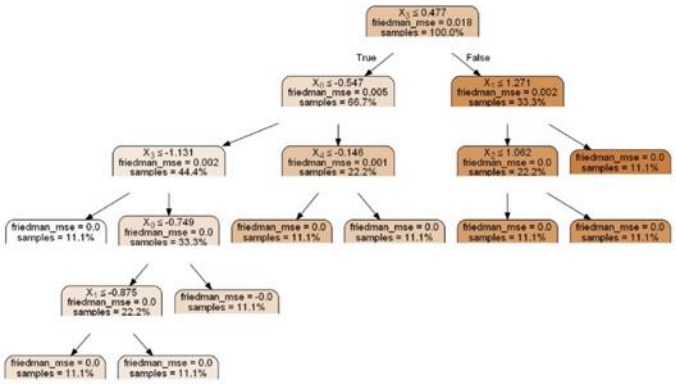


Fig 14. Structure of the 28th Regression Tree in GBR for the 3rd Quarter Analysis

The summary of the Grid-Search hyper-parameter tuning is shown in Fig 15. It depicts the variation of Grid Search Mean Score with Max-Depth for the learning rate of 0.1 and for 28 estimators.

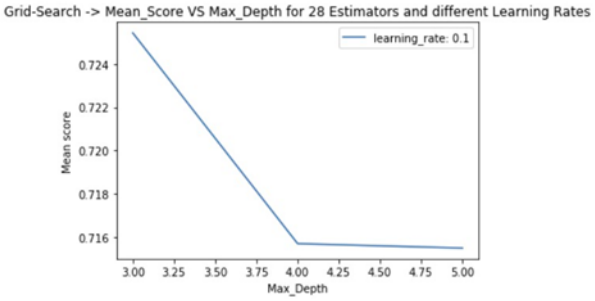


Fig 15. Grid Search Summary on Mean Score for the 3rd Quarter Analysis

ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

3.4 Analysis of 4th Quarter i.e., October to December

3.4.1 Trend Analysis of Foreign Tourist Arrivals in the 4th Quarter from 2005 to 2016

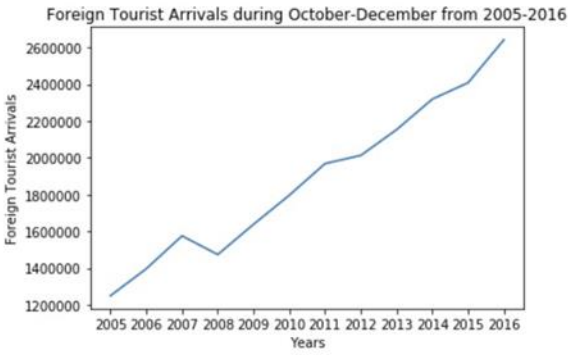


Fig 16. Yearly Trends of Foreign Tourist Arrivals in India during the 4th Quarter

From Fig 16, it is evident that from 2005 to 2007, there has been a gradual increase in FTA in India. It had dropped for a year till 2008. Since then FTA has been strictly increasing at varying rates till 2016.

3.4.2 Feature Selection and Verification

Out of 41 features in the dataset, for the 4th Quarter, 5 most important features are selected as per Random Forest Regressor (with 10 estimators) Feature Importance Scores shown in Table 5.

From Table 5, 5 most important features are found to be:
– Foreign Exchange Earnings.
– National currency, constant prices, national base year, quarterly levels, seasonally adjusted's Construction.
– National currency, constant prices, national base year, quarterly levels' Gross value added at basic prices, total activity.
– National currency, current prices, quarterly levels' Distrib. trade, repairs; transp.; accommod., food serv. Activ.
– National currency, constant prices, national base year, quarterly levels, seasonally adjusted's Real estate activities.
A Correlation Matrix is shown in Fig 17, in the form of a Heat-Map displaying Feature-to-Feature and Feature-to-Label Pearson Correlations where the features are the selected features from Random Forest Regressor's Feature Importance Scores.

ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

ID	Feature Name		Random Forest Regressor Score
1	CQRSA	market prices - output approach	7.54456088e-02
		basic prices and total activity	1.87425241e-04
		agriculture, forestry and fishing	0
		industry, including energy	9.61326913e-03
		manufacturing	1.18216688e-03
		construction	8.23513556e-05
		services	2.26920878e-03
		distributed trade, repairs, transportation, accommodation, food services and activities	4.84330126e-05
		real estate activities	4.23609478e-03
		public admin, education, human health.	7.52294809e-04
2	CQR	market prices - output approach	3.87272956e-04
		basic prices and total activity	5.80240084e-03
		agriculture, forestry and fishing	0
		industry, including energy	1.66524723e-02
		manufacturing	3.12873946e-02
		construction	1.18113915e-02
		services	7.82441512e-02
		distributed trade, repairs, transportation, accommodation, food services and activities	9.29181414e-02
		real estate activities	9.60630085e-03
		public admin, education, human health.	6.84637513e-03
3	VNBQRSA	market prices - output approach	2.52674102e-04
		basic prices and total activity	1.12281116e-02
		agriculture, forestry and fishing	1.65547016e-02
		industry, including energy	1.22489313e-02
		manufacturing	2.45763470e-04
		construction	1.36504450e-01
		services	1.04605275e-02
		distributed trade, repairs, transportation, accommodation, food services and activities	1.21443072e-04
		real estate activities	8.75553063e-02
		public admin, education, human health.	1.35652030e-02
4	VNBQR	market prices - output approach	1.08934038e-03
		basic prices and total activity	9.60191534e-02
		agriculture, forestry and fishing	7.63299415e-02
		industry, including energy	5.72852484e-03
		manufacturing	1.45505027e-02
		construction	1.87375772e-03
		services	9.89750748e-03
		distributed trade, repairs, transportation, accommodation, food services and activities	1.11364149e-03
		real estate activities	6.34478276e-03
		public admin, education, human health.	2.09086862e-03
5	Foreign Exchange Earnings		1.48852113e-01

Table 5. Feature Study for 4th Quarter

ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

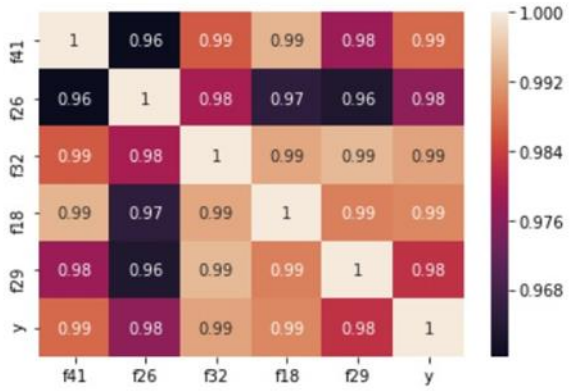


Fig 17. Heat-Map showing Feature-to-Feature and Feature-to-Label's Pearson Correlation Coefficients in the Analysis of 4th Quarter

From the heat-map in Fig 17, it is verified by Pearson Correlation Coefficients, that the selected features have high correlation values with the Target (FTA) values. A Scatter-Plot of FTA is shown to depict its dependency on the Most Important Feature (Foreign Exchange Earnings) in Fig 18

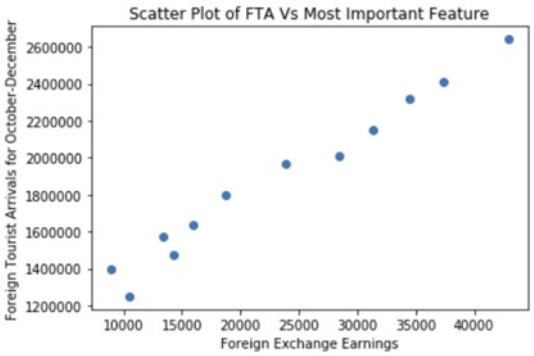


Fig 18. Scatter-Plot showing the dependency of FTA in the 4th Quarter with Foreign Exchange Earnings

3.4.3 Data Pre-processing

The Data Pre-Processing is done by following the same methodology steps as done in the Analysis of 1st Quarter.

3.4.4 Learning Algorithm

ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

The learning algorithm used to build the Regression Model is an Ensemble Learning and Boosting Algorithm known as Gradient Boosting Regressor. The Algorithm of the Gradient Boosting Regressor is shown as Algorithm 2.

3.4.5 Training the Model

The hyper-parameters of the Gradient Boosting Regressor, number of estimators (regression trees), maximum depth and learning rate are tuned using Grid-Search. The best set of hyper-parameters so found, were 10 regression trees, maximum depth of 5 and learning rate of 0.2. The resulting model has been shown in Fig 19 for the 10th Regression Tree which is the most correct one among the 10 regression trees as at each step, the error in the previous regression tree gets corrected by the following regression tree.

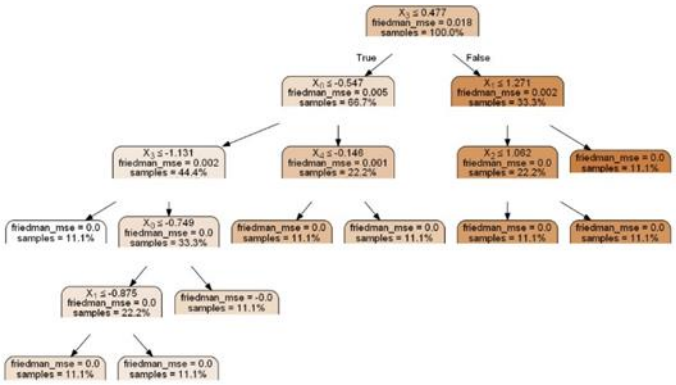


Fig 19. Structure of the 10th Regression Tree in GBR for the 4th Quarter Analysis

The summary of the Grid-Search hyper-parameter tuning is shown in Fig 20. It depicts the variation of Grid Search Mean Score with Max-Depth for different learning rates and for 10 estimators.

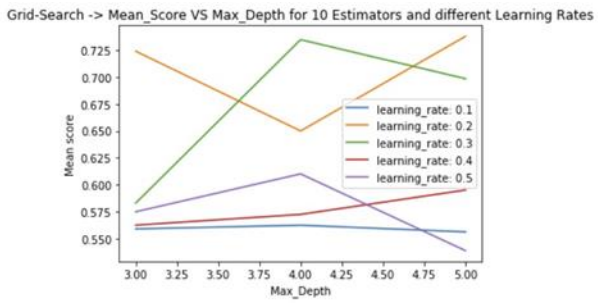


Fig 20. Grid Search Summary on Mean Score for the 4th Quarter Analysis

ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

4 Implementation Details

The overall development of the 4 Sub-Models is done using Python's Scikit-Learn Machine Learning Toolbox on a machine with Intel(R) Core(TM) i5-8250U processor, CPU @ 1.60 GHz 1.80 GHz and 8 GB RAM. The Visualizations are done using Python's Plotting Libraries, Matplotlib & Seaborn and the Tree Visualizations are done using Python Libraries, Graphviz and PyDotPlus.

5 Sub-Model Performance Analysis

5.1 Statistical Results

The Performance Analysis of the Sub-Models is done on the following metrics:
1. Mean Absolute Error
2. Mean Square Error
3. Root Mean Square Error
4. Coefficient of Determination or R2

After training, the trained sub-models are validated on their respective Validation Sets. The Performance Analysis for all the sub-models are tabulated in Table 6.

Quarter	Training MAE	Validation MAE	Training MSE	Validation MSE	Training RMSE	Validation RMSE	Coefficient of Determination
1st Quarter/January to March	0.05	0.163	0.003	0.037	0.058	0.193	0.902
2nd Quarter/April to June	0.041	0.047	0.002	0.005	0.048	0.072	0.988
3rd Quarter/July to September	0.045	0.033	0.003	0.001	0.055	0.033	0.998
4th Quarter/October to December	0.075	0.094	0.012	0.006	0.107	0.081	0.988

Table 6. Sub-Model Performance Analysis

5.2 Visualization Results

The nature of Regression Fits for the 4 Sub-Models are visualized in the form of Fitting Diagrams and Training and Validation Loss Convergence Diagrams:

5.2.1 Nature of Regression Fit of the 1st Sub-Model

The Fitting Diagram of the 1st Sub-Model for the prediction of Foreign Tourist Arrivals in the 1st Quarter is shown in Fig 21.

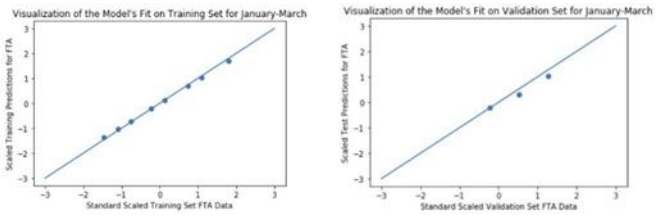


Fig 21. Fitting Diagram for the 1st Quarter FTA Prediction

ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

From Fig 21, it can be interpreted that this Regression Fit is close to a Perfect-Fit. The Training and Validation Loss Convergence Diagram of the 1st Sub-Model for the prediction of Foreign Tourist Arrivals in the 1st Quarter is shown in Fig 22. Here, the Loss is referred as the Mean Square Error.

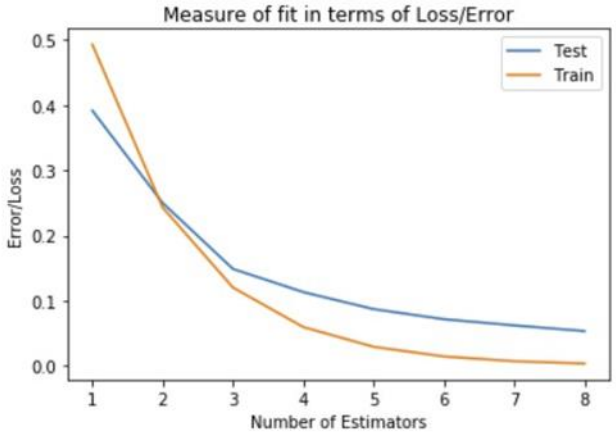


Fig 22. Training and Validation Loss Convergence Diagram for 1st Quarter FTA Prediction

From Fig 22, it can be concluded that finally (at the 8th estimator), the Training and Validation MSE/Loss has a difference of 0.034 between them, making it very close to a perfect fit.

5.2.2 Nature of Regression Fit of the 2nd Sub-Model

The Fitting Diagram of the 2nd Sub-Model for the prediction of Foreign Tourist Arrivals in the 2nd Quarter is shown in Fig 23.

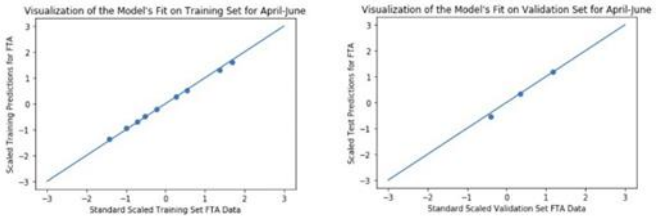


Fig 23. Fitting Diagram for the 2nd Quarter FTA Prediction

From Fig 23, it can be interpreted that this Regression Fit is very close to a Perfect-Fit. The Training and Validation Loss Convergence Diagram of the 2nd Sub-Model for the prediction of Foreign Tourist Arrivals in the 2nd Quarter is shown in Fig 24. Here, the Loss is referred as the Mean Square Error.

ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

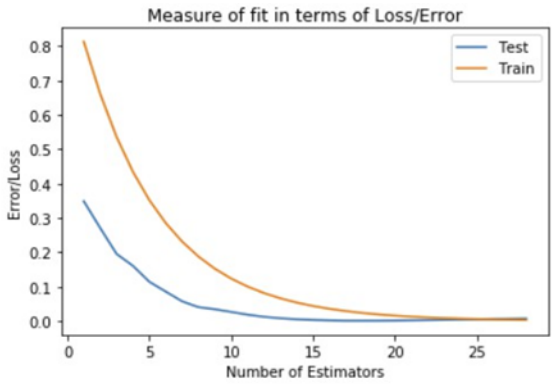


Fig 24. Training and Validation Loss Convergence Diagram for 2nd Quarter FTA Prediction

From Fig 24, it can be concluded that finally (at the 29th estimator), the Training and Validation MSE/Loss has almost converged with a slight difference of 0.003. So, it can be called almost a perfect-fit.

5.2.3 Nature of Regression Fit of the 3rd Sub-Model

The Fitting Diagram of the 3rd Sub-Model for the prediction of Foreign Tourist Arrivals in the 3rd Quarter is shown in Fig 25.

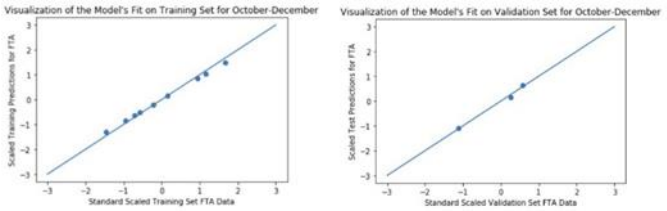
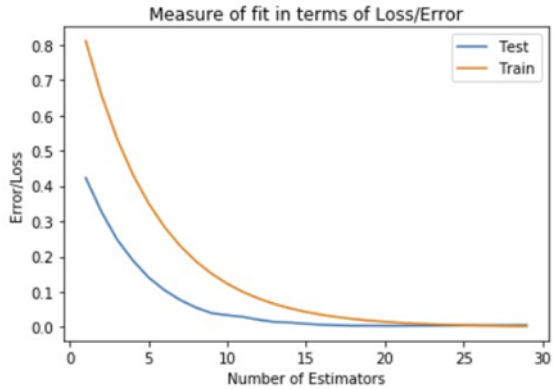


Fig 25. Fitting Diagram for the 3rd Quarter FTA Prediction

From Fig 25, it can be concluded that this Regression Fit is roughly, a Perfect-Fit.The Training and Validation Loss Convergence Diagram of the 3rd Sub-Model for the prediction of Foreign Tourist Arrivals in the 3rd Quarter is shown in Fig 26. Here, the Loss is referred as the Mean Square Error.



ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

Fig 26. Training and Validation Loss Convergence Diagram for 2nd Quarter FTA Prediction

From Fig 26, it can be concluded that finally (at the 28th estimator), the Training and Validation MSE/Loss has a difference of 0.002 between them which is, quite acceptable to be called as a Perfect fit.

5.2.4 Nature of Regression Fit of the 4th Sub-Model

The Fitting Diagram of the 4th Sub-Model for the prediction of Foreign Tourist Arrivals in the 4th Quarter is shown in Fig 27.

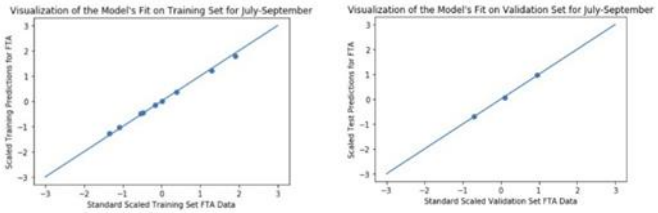
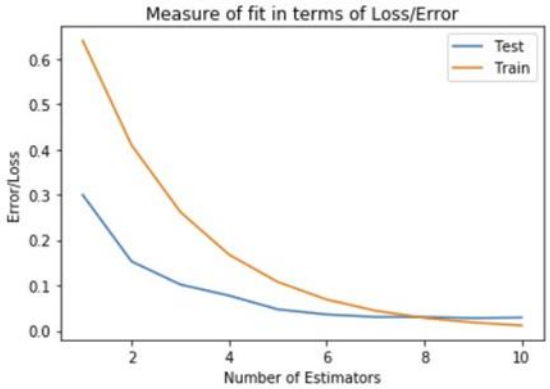


Fig 27. Fitting Diagram for the 4th Quarter FTA Prediction

From Fig 27, it can be concluded that this Regression Fit is nearly an approximate Perfect-Fit. The Training and Validation Loss Convergence Diagram of the 4th Sub-Model for the prediction of Foreign Tourist Arrivals in the 4th Quarter is shown in Fig 28. Here, the Loss is referred as the Mean Square Error



ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

Fig 28. Training and Validation Loss Convergence Diagram for 4th Quarter FTA Prediction

From Fig 28, it can be concluded that finally (at the 10th estimator), the Training and Validation MSE/Loss has a difference of 0.006 between them which is, quite acceptable to be called as a Perfect fit.

6 Evaluation of Distribution Analysis and Trend Analysis

It has been found that, the 4 Validation Sets of the 4 Sub-Models, have the year, 2012 as a common instance which is validated, after they are trained independently. So, 2012 can be used for our Distribution Analysis and Trend Analysis of Foreign Tourist Arrivals over the 4 quarters. This is achieved in the following steps:

1. The scaled predictions given by the 4 sub-models are obtained for FTA of 2012 and are re-scaled back to their unscaled values.
2. The unscaled values, so obtained are then rounded off to their nearest integers.
3. Finally, to show the FTA (Foreign Tourist Arrival) Distribution among the 4 quarters and comparing with that of the Original Distribution, Pie Charts are constructed both for the Distribution Analysis, obtained from the 4-headed Machine Learning Model and the Original FTA Distribution. Pie-Charts are shown in Fig 29.
4. For the Trend Analysis of FTA in the 4 quarters of 2012 and comparing with the Original Trend, Line Plots are constructed and shown in Fig 30.

Fig 29. FTA Distribution Analysis among 4 quarters of 2012

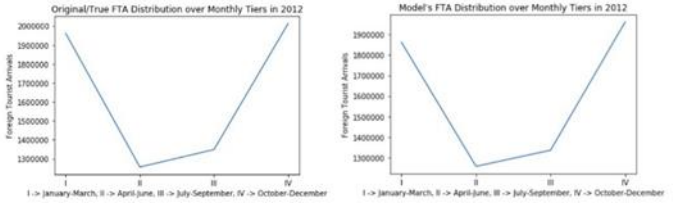


Fig 30. FTA Trend Analysis in the form of Line Plots in the 4 quarters of 2012

ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

7 Conclusion

This paper proposed a Machine Learning with Regression Analysis Approach for Foreign Tourist Arrival Distribution Analysis and Trend Analysis over the 4 quarters of a year from categorical with multi-sector (subcategorical) GDP values and Foreign Exchange Earnings. From the end results shown in the form of Pie Charts and Line Plots, it is evident that the 4-Headed Machine Learning Model performed well, predicting almost accurately the Distribution and Trends of International Tourists visiting India at different time frames in a year. Also, it is therefore verified that, there is a strong inter-dependency between Gross Domestic Product and Foreign Tourism Demand in India. The research done in this paper can be performed for any other developing country, after relevant data collection which will help enhance and flourish the Tourism Industry and hence making it, a paradise for foreign tourists. Future Scope of this work involves the application of more advanced Regression Techniques and Statistical Methods for Trend and Distribution Analysis. Also preparation of Time Series Tourist Data can help in Time Series Analysis by identifying patterns in Time Series Data for more effective and accurate trend analysis. Moreover, several modelling and forecasting procedures can be employed on time series data namely, Exponential Smoothing, Seasonal Decomposition Models and Spectrum Analysis.

8 Compliance with Ethical Standards

Funding: The study is not funded by any organization. Conflict of Interest: The corresponding author states that there is no conflict of interest.

References

[1] https://en.wikipedia.org/wiki/World_Tourism_ rankings
[2] Sun, Shaolong, et al. ”Forecasting tourist arrivals with machine learning and internet search index.” Tourism Management 70 (2019): 1-10.
[3] Ricardo, Hugo David dos Reis Barbosa. Forecasting tourism demand for Lisbon´ s region: a data mining approach. Diss. 2018.
[4] Claveria, Oscar, Enric Monte, and Salvador Torra. ”Modelling tourism demand to Spain with machine learning techniques. The impact of forecast horizon on model selection.” arXiv preprint arXiv:1805.00878 (2018).
[5] Hu, Yi-Chung. ”Predicting foreign tourists for the tourism industry using soft computing-based Grey–Markov models.” Sustainability 9.7 (2017): 1228.
[6] Garc´ıa Rodr´ıguez, Oscar. ”Forecasting tourism arrivals with an online search engine data: A study of the Balearic Islands.” (2017).

ISSN: 2582-2640 (online)

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No. 01
Pages: 57-82
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.1.006

[7] Petrevska, Biljana. ”Predicting tourism demand by ARIMA models.” Economic research-Ekonomska istraˇzivanja 30.1 (2017): 939-950.

[8] Yu, Ying, et al. ”Statistical modeling and prediction for tourism economy using dendritic neural network.” Computational intelligence and neuroscience 2017 (2017).

[9] Ali, Rafidah, and Ani Shabri. ”Modelling Singapore Tourist Arrivals to Malaysia by Using SVM and ANN.” (2017).

[10] Noersasongko, Edi, et al. ”A tourism arrival forecasting using genetic algorithm based neural network.” Indian Journal of Science and Technology 9.4 (2016).

[11] Cankurt, S., and A. Subasi. ”Developing tourism demand forecasting models using machine learning techniques with trend, seasonal, and cyclic components.” Balkan Journal of Electrical and Computer Engineering 3.1 (2015).

[12] Neupane, Hari Sharma, Chandra Lal Shrestha, and Tara Prasad Upadhyaya. ”Modelling monthly international tourist arrivals and its risk in Nepal.” NRB Economic Review 24.1 (2012): 28-47. 13. https://www.kaggle.com/navoneel/fta-data nces

ISSN: 2582-2640 (online)