Analysis of Number of Child Births and Infant Deaths in India

- Guided by Arnab Bhattacharya

Contributors

Group Number: 5	
Name	**Roll No. (Email-ID)**
Lavlesh Mishra	19111048 (lavleshm@iitk.ac.in)
Kuldeep Kumar Solanki	19111045 (kuldeeps@iitk.ac.in)
Jaydeep Meda	19111039 (jaydeepm@iitk.ac.in)
Aditya Jain	20111418 (kun20@iitk.ac.in)
Rohit Singh	20111004 (adityaj20@iitk.ac.in)

Broad Aims of the Project

We aim to create a model which uses HMIS(Health Management Information System) and CENSUS data to predict the following for a district in a particular year.

- Number of births
- Number of still births
- Number of infant deaths

Datasets Used

Performance of Key Health Management Indicators for each district in India (HMIS)

- Released by Ministry of Health under National Health Mission flagship program
- Seeks to provide effective healthcare to the rural population throughout the country
- Includes the key indicators which affects health of mother and child during pregnancy and at the time of delivery
- Data for financial years 2008 to 2019

Census Data 2001 and 2011

- Available at district level for total population, age, disability, education, migration, religion, and various other features
- Generated projections of census data for the years 2008 to 2018

Unstructured Data

HMIS Data:

- Data we found was in unstructured excel files

Census Data:

- Census data is semi-structured
- We converted them into structured comma separated files (csv).

Financial Year: 2018-19

Provisional Figures for the Period April to March

These indicators are under compilation (83,84,85,86,197,198)

	Indicators	1		2		3		4		5		6		7		8		9		10	
		Total number of pregnant women Registered for ANC		Number of Pregnant women registered within first trimester		% 1st Trimester registration to Total ANC Registrations		Number of pregnant women received 4 or more ANC check ups		TT2 given to Pregnant women (numbers)		TT Booster given to Pregnant women (numbers)		% Pregnant Woman received 4 ANC check ups to Total ANC Registrations		% Pregnant women received TT2+ TT Booster to Total ANC Registration		Number of Pregnant women given 180 IFA tablets		% Pregnant women given 180 IFA to Total ANC Registration	
		2018-19	2017-18	2018-19	2017-18	2018-19	2017-18	2018-19	2017-18	2018-19	2017-18	2018-19	2017-18	2018-19	2017-18	2018-19	2017-18	2018-19	2017-18	2018-19	2017-18
	_Jammu & Kashmir	3,95,092	3,99,307	2,66,690	2,58,865	67.5	64.8	2,88,808	2,47,859	1,76,547	1,74,937	55,651	56,517	73.1	62.1	58.8	58	1,40,277	2,14,965	35.5	53.8
1	Anantnag	28,918	27,773	24,691	25,353	85.4	91.3	16,029	19,526	20,516	21,546	6,025	6,002	55.4	70.3	91.8	99.2	8,009	14,662	27.7	52.8
2	Badgam	12,379	12,592	12,289	12,104	99.3	96.1	9,632	9,707	12,060	11,465	27	1,366	77.8	77.1	97.6	101.9	8,592	10,347	69.4	82.2
3	Bandipora	7,140	7,641	5,954	6,279	83.4	82.2	5,155	5,167	5,337	5,436	1,883	2,030	72.2	67.6	101.1	97.7	5,602	4,891	78.5	64
4	Baramula	22,241	25,103	18,756	20,940	84.3	83.4	20,931	22,510	15,479	14,386	7,796	9,418	94.1	89.7	104.6	94.8	21,892	23,461	98.4	93.5

STATE	DISTRICT	SUB-DIST	TOWN_VII	WARD	EB	LEVEL	NAME	TRU	No_HH	TOT_P	TOT_M	TOT_F	P_06
24	01	0000	00000000	0	0	DISTRICT	Kachchh	Urban	96009	474892	247682	227210	65094
24	01	0002	40101000	0	0	TOWN	Rapar (M)	Urban	4327	23057	11857	11200	3934
24	01	0003	40102000	0	0	TOWN	Bhachau (I	Urban	5703	25389	13310	12079	4284
24	01	0004	40103000	0	0	TOWN	Anjar (M)	Urban	14411	68343	35341	33002	9859
24	01	0005	40104000	0	0	TOWN	Bhuj (M)	Urban	19483	98528	51768	46760	11651
24	01	0008	40105000	0	0	TOWN	Mandvi (M)	Urban	8045	42355	21620	20735	5485
24	01	0009	40106000	0	0	TOWN	Mundra (C	Urban	2680	12931	6650	6281	1775
24	01	0010	40107000	0	0	TOWN	Gandhidha	Urban	29872	151693	79379	72314	21151
24	01	0010	40108000	0	0	TOWN	Kandla (C	Urban	2979	14695	8469	6226	2115
24	02	0000	00000000	0	0	DISTRICT	Banas Kar	Urban	52072	275501	144831	130670	41710

Data Preprocessing - HMIS

Characteristics

- 385 csv files for 28 States and 7 Union Territories
- Data set has 163 attributes
- Each files stores the data for two years and has 325 columns

Cleaning

- Removed attributes having missing values
- Removed redundant attributes

Data Preprocessing - CENSUS

Characteristics

- Available district wise data for year 2001 and 2011
- Data set has 64 and 94 attributes for the year 2001 and 2011 respectively

Inconsistency in data

- The attribute 'All ages' has data for the people whose age are not known
- 'Literate' attribute includes figures for 'literates without educational level' and 'educational levels not classifiable'
- District IDs are outdated. Since 2011 many districts are newly created which led to the change of district IDs
- Ever married women includes currently married, widowed, divorced and separated
- There are missing values in the dataset, represented by NaN

Generation of Census Data

- The Growth Rate(r) is calculated using the 2001 data as Principal(p) and 2011 data as the final amount(p+i). Here i is the Interest.

$$(1+r)^{10} = \frac{Census - 2011 \ \ stats}{Census - 2001 \ \ stats}$$

- Growth rate r is constant over here, which means the interpolation is geometric and not linear.
- Census data for year $Y = (1+r)^{(Y-2001)} * P$

Feature Selection

HMIS

- Out of 163 attributes we filtered out 42 attributes for our project
- There are many attributes which represent the percentage of the other attributes, consequently, have been dropped due to multi collinearity

CENSUS

- 6 Attributes have been selected out of 64
  - Area name
  - Population
  - Literate persons
  - Main workers
  - Marginal workers
  - Non-workers

Data Integration

- Merged HMIS and Census data. Both the datasets has one same attribute i.e., District Name
- This attribute is ‘Area Name’ in Census data, and it is ‘Indicator’ in HMIS data
- Performed natural join on this common attribute
- Conflicts in district names are resolved using edit distance.

Data Exploration

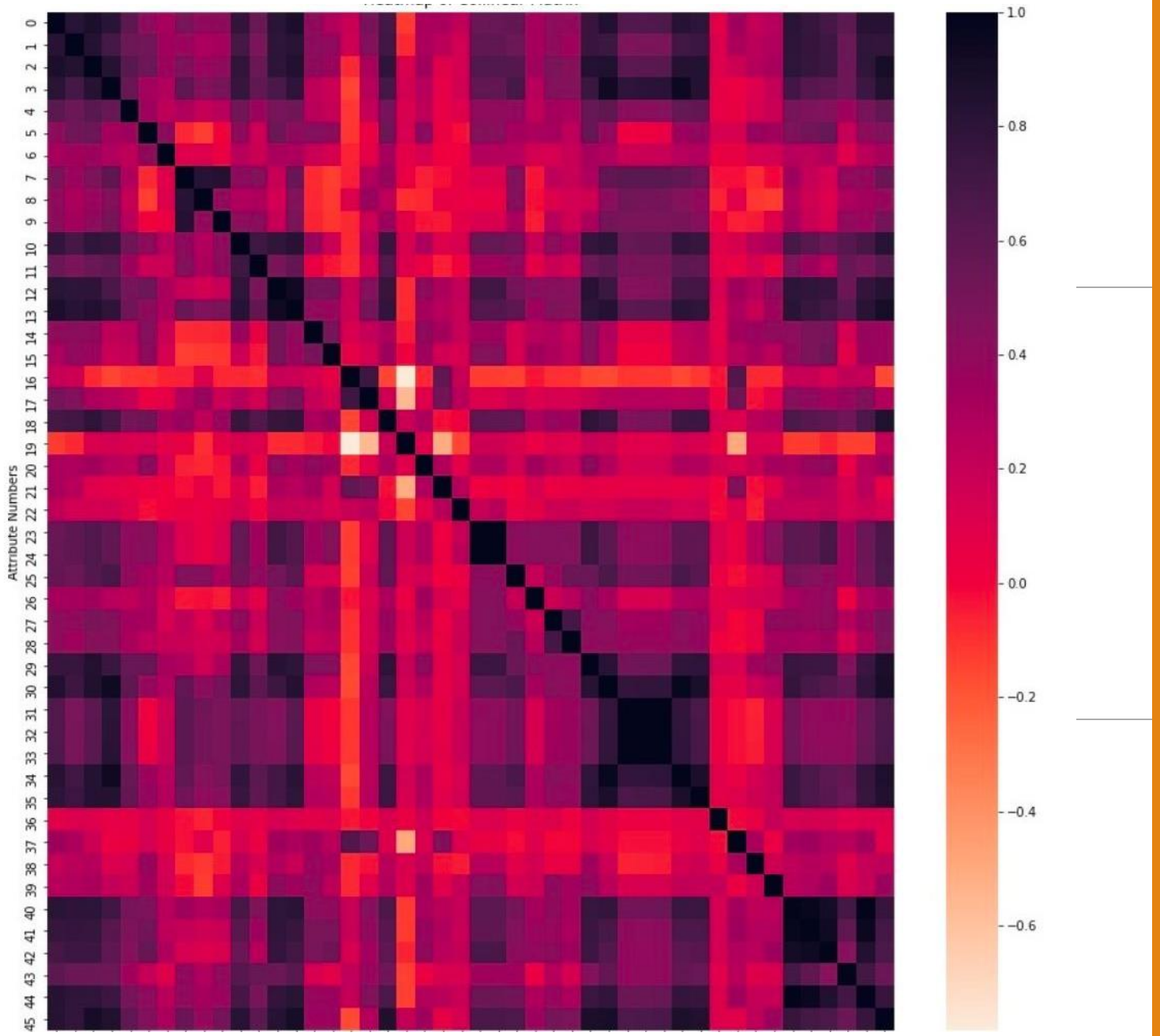
- Univariate Analysis
- Independent and Dependent variable identification
- Multivariate Analysis
- Missing Value Treatment
- Outlier Analysis

Univariate Analysis

- Exploring variables/attributes one by one
- Out of 47 attributes only one is categorical and all the others are numerical
- Since we are solving regression problem, categorical data "Indicator" is of no use in our prediction. So, we removed it from the data set.

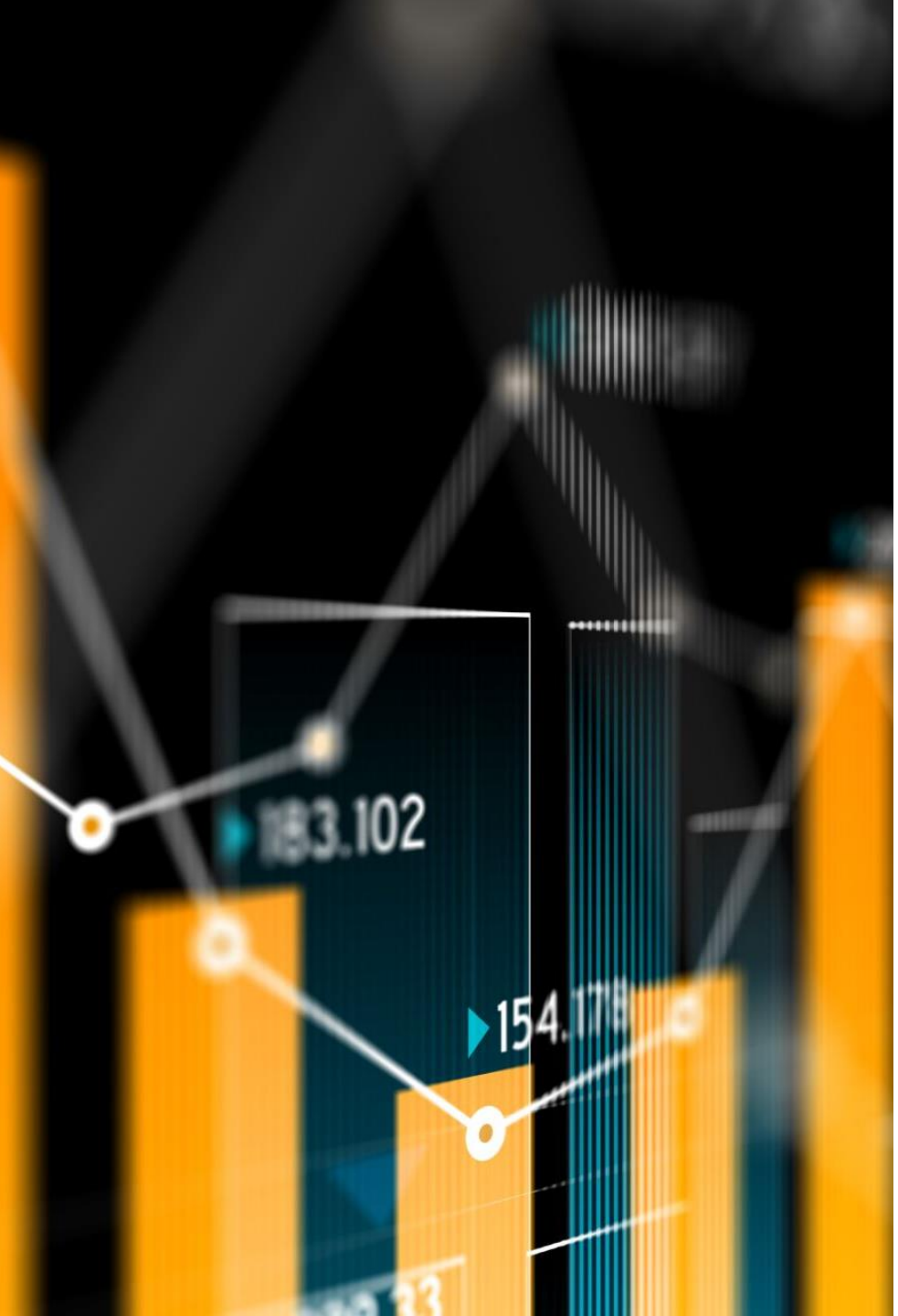
Independent and Dependent variable identification

- Identify Predictor (Input) and Target (output) variables.
- Case 1: Predicting number of Births in a year
  - Independent\Target attribute: 'Total Number of reported live births'
  - Dependent attributes: Rest 45 attributes of the dataset.
- Case 2: Predicting number of Still-Births in a year
  - Independent\Target attribute: 'Total Number of reported Still Births'
  - Dependent attributes: Rest 45 attributes of the dataset.
- Case 3: Predicting number of Infant-Deaths in a year
  - Independent\Target attribute: 'Total Number of reported Infant Deaths'
  - Dependent attributes: Rest 45 attributes of the dataset.



Multivariate Analysis

1. HeatMap of Collinear matrix of dataset
2. Dark squares in the HeatMap shows high correlation between the attributes



Missing Value Treatment

- Most tricky part of this data exploration as there are no NaN values in our dataset at this stage.
- All NaN values in Census data were taken care at the individual pre-processing of the census files
- In HMIS data there are no NaN values but there are values which are equal to zero.
- The tricky part is whether these zero values are missing values or actual value of the data example is zero.

Outlier Analysis

IQR Test(Interquartile Range)

- The data is sorted in ascending order and split into 4 equal parts
- The data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR are outliers

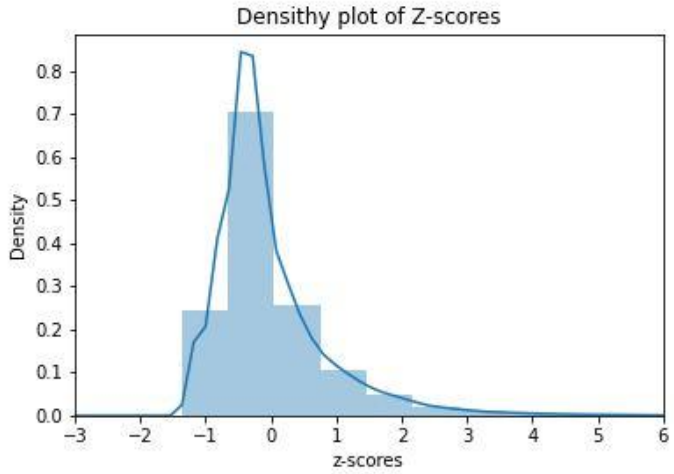
Attributes (Top 10) having most number of Outliers

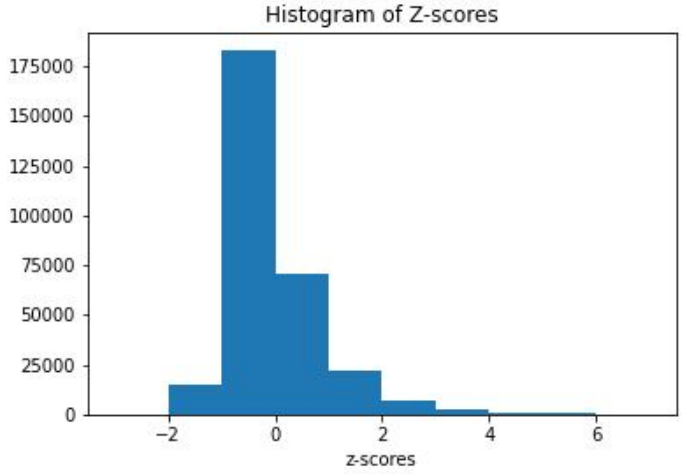
Attributes Name	No. of Outliers
Adverse Events Following Imunisation (Others)	610
IUCD insertions done (pvt. facilities)	532
Number of home deliveries attended by SBA trained (Doctor/Nurse/ANM)	414
Number of Vasectomies Conducted (Public + Pvt.)	406
Number of Minor Operations	392
Total Number of MTPs (Public) reported	384
Number of C-section deliveries conducted at private facilities	357
Number having severe anaemia (Hb<7) treated at institution	327
Number of Major Operations	319
Number of home deliveries attended by Non SBA trained (trained TB/Dai)	295
Number of C-section deliveries conducted at public facilities	278
Condom pieces distributed	272
Total Number of Abortions (Spontaneous/ Induced) Reported	267
Number of Home deliveries	256

Outlier Analysis

Z-Score

- Z score helps to understand if a data value is greater or smaller than mean and how far away it is from the mean
- Z score tells how many standard deviations away a data point is from the mean
- If the z score of a data point is more than 3, it indicates that the data point is quite different from the other data points





Prediction

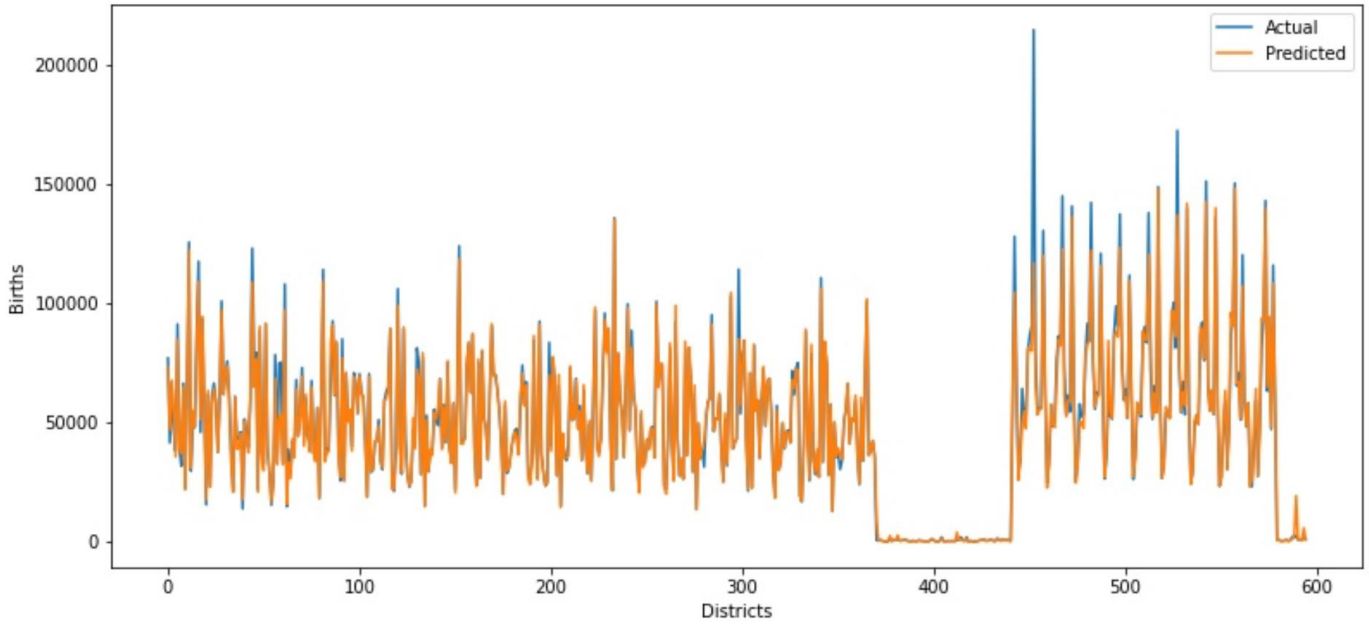
CASE 1: PREDICTION OF NUMBER OF BIRTHS

Measure	Linear Regression	Random Forest
Training Mean Absolute Error	2828.6132176902	574.652419301165
Test Mean Absolute Error	5109.2537316244	2229.61431932773
Training Root Mean Square Error	5362.25745574993	1564.56984390149
Testing Root Mean Square Error	8402.35904407914	5916.33023171844
Training Accuracy(R-square)	95.7276759961367	99.6362870467858
Testing Accuracy(R-square)	93.3642138572999	96.7100088929496

Prediction

CASE 1: PREDICTION OF NUMBER OF BIRTHS

- Random forest result for births(Actual vs Predicted)



Prediction

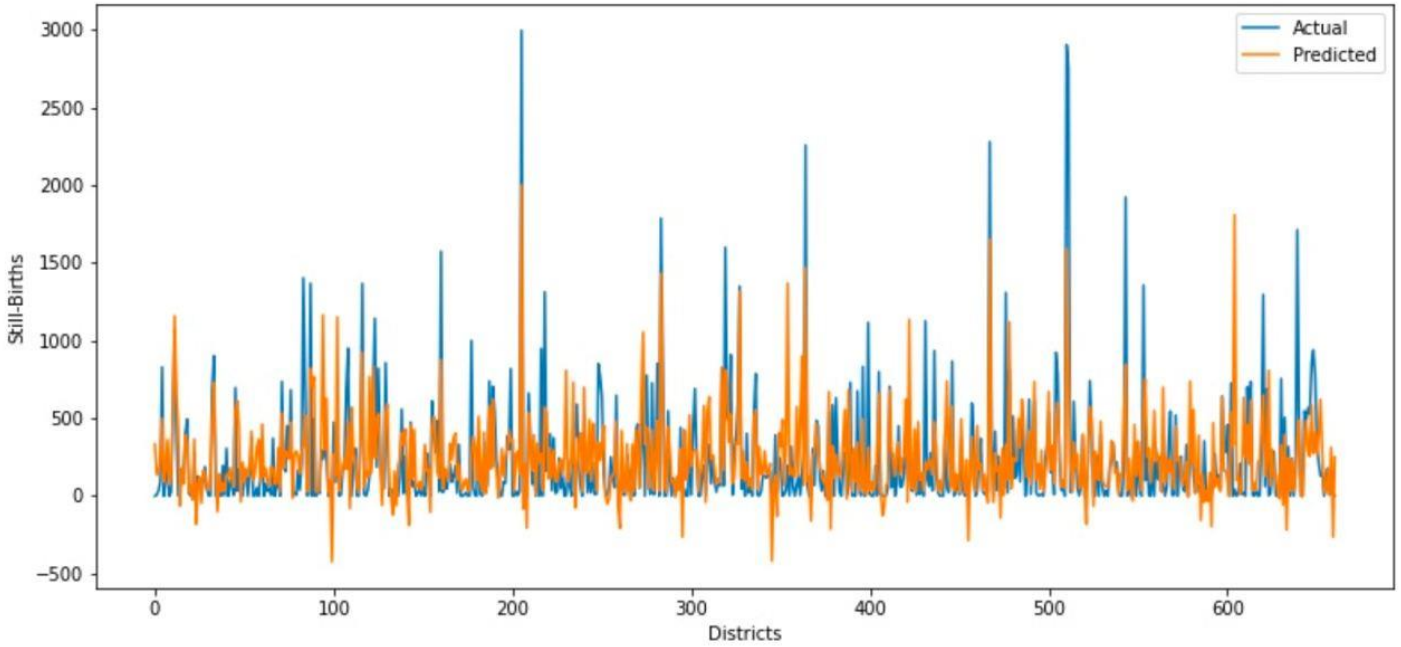
CASE 2: PREDICTION OF NUMBER OF STILL-BIRTHS

Measure	Linear Regression	Random Forest
Training Mean Absolute Error	1023.99261949344	89.88106998654105
Test Mean Absolute Error	2390.04152966319	541.9993948562783
Training Root Mean Square Error	2235.31247540752	396.0021534095236
Testing Root Mean Square Error	4821.80489352115	2139.573840681855
Training Accuracy(R-square)	88.124395801068	99.62728718459579
Testing Accuracy(R-square)	92.4555099954537	98.51452557349563

Prediction

CASE 2: PREDICTION OF NUMBER OF STILL-BIRTHS

- Random forest result for still births(Actual vs Predicted)



Prediction

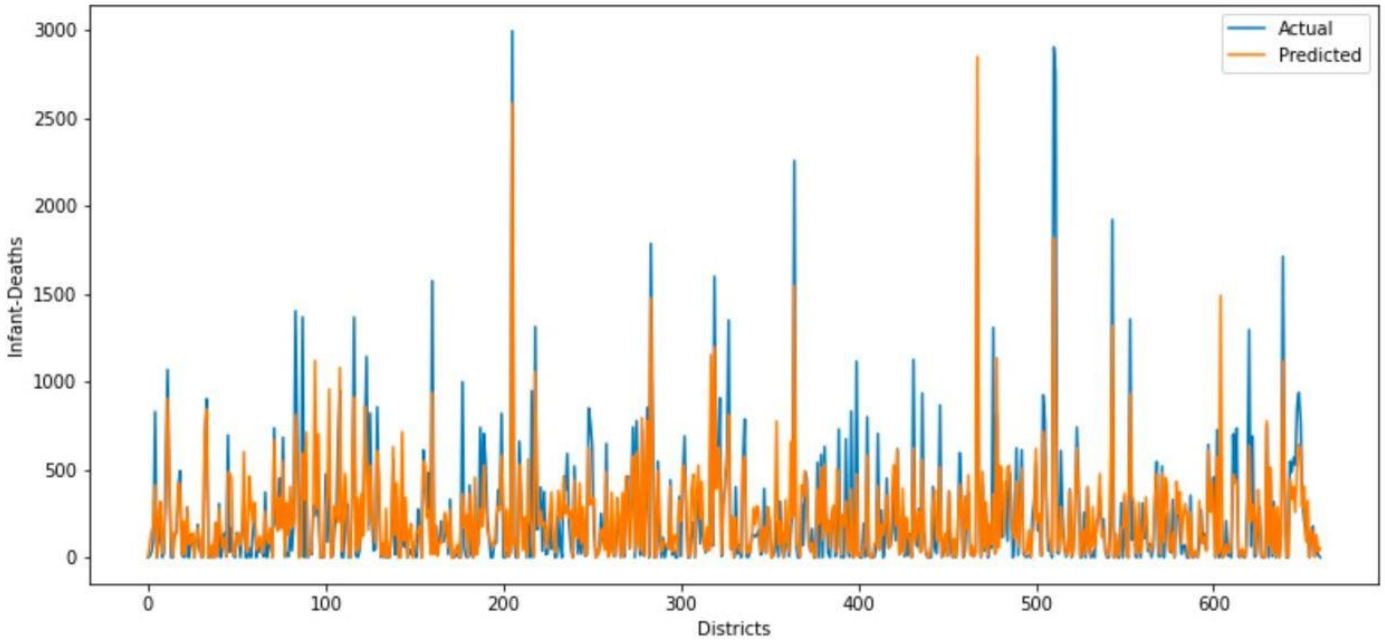
CASE 3: PREDICTION OF NUMBER OF INFANT DEATHS

Measure	Linear Regression	Random Forest
Training Mean Absolute Error	176.98444922111506	44.47679676985195
Test Mean Absolute Error	171.44992601028855	113.2001815431165
Training Root Mean Square Error	284.300609512833	87.23965230976026
Testing Root Mean Square Error	265.06288848278155	201.14927401387413
Training Accuracy(R-square)	48.09697806858291	95.12455070018713
Testing Accuracy(R-square)	49.51743097935387	70.56988968985102

Prediction

CASE 3: PREDICTION OF NUMBER OF INFANT DEATHS

- Random forest result for still infant deaths(Actual vs Predicted)



Results

Model Name	Problem Statement	Training Accuracy	Testing Accuracy
Linear Regression	Child Births	0.957267	0.933956
Random Forest	Child Births	0.993489	0.960923
Linear Regression	Still Births	0.881243	0.924555
Random Forest	Still Births	0.996272	0.985145
Linear Regression	Infant Deaths	0.480969	0.495174
Random Forest	Infant Deaths	0.9512345	0.705698

Discussions

- For the prediction of total number of births, ‘Total number of pregnant women Registered for ANC’, ‘Number of pregnant women received 3 ANC check ups’, ‘Deliveries Conducted at Public Institutions’ have come out as the most influential factors.
- For the prediction of total number of still births, ‘Total reported deliveries’, ‘TT2 or Booster given to Pregnant women’, ‘Number of Minor Operations’ have come out as the most influential factors.
- For the prediction of total number of infant deaths, ‘Number of Major Operations’, ‘Number of C-section deliveries conducted at public facilities’, ‘Number of home deliveries attended by Non-SBA trained’ have come out as the most influential factors.

Discussions

- Used Linear regression and Random Forest regression for prediction
- Random Forest Regressor Model performs the best with training accuracy of 99.15% and testing accuracy of 94.41%
- Decision Trees are great for obtaining non-linear relationships between input features and the target variable
- Random forest is an ensemble of decision trees. This is to say that many trees, constructed in a certain “random” way form a Random Forest
- The averaging makes a Random Forest better than a single Decision Tree hence improves its accuracy and reduces overfitting
- A prediction from the Random Forest Regressor is an average of the predictions produced by the trees in the forest

THANK YOU