Improving Influenza Prediction Rate by Combining Traditional, Search and Social Media Data sources

Shrinivas V Shanbhag (16IT244), Gaurav U D (16IT113), Abhilash V (16IT201)
Information Technology
National Institute of Technology Karnataka
Surathkal, India 575025

***Abstract*—Influenza, which is also termed as the flu, is one of the infectious disease caused by the virus known as influenza virus. Influenza outbreak poses a brilliant task for clinical professionals as every year they claim around half of a million lives international. Predicting when a virulent disease may occur is vital in order that the preventive measures can be taken on the right time. In the past examination done over this theme, different statistical models were used by different organisations to forecast flu rate in USA. This project considers the forecast data given by different sources namely CDC, Twitter, Athena health, Google Trends and Google Flu trends, and strengthens the forecast using various ensemble models. The project has total twelve models, out of which five are forecasting models based on five different data, and latter ones are ensemble models. The proposed model gives considerably good predictions when compared to other existing systems.**

***Index Terms*—Influenza, CDC, ILI, Regression, Linear Regression, SVM, LASSO, Ensemble Model.**

I. Introduction

Influenza is a viral contamination that attacks the respiratory machine especially nostril, throat and lungs. Influenza is commonly referred to flu, but it's not the same as belly "flu" viruses that cause diarrhea and vomiting. Predictions of the dynamics of non-seasonal and seasonal flu outbreaks stays a wonderful undertaking. On world wide basis they cause up to five hundered thousand deaths a year and in the United States of America (US)[1] an estimation of three thousand to fifty thousand deaths a year is obtained. The Centers for Disease Control and Prevention (CDC) of United States constantly monitors the level of Influenza Like Illness (ILI) move within the US populace by way of gathering facts from physicians' reports that file the share of sufferers visible in clinics who show off Influenza like illnesses (ILI) symptoms. To estimate the Influenza Like Illness (ILI) activity in the United States, many attempts are been taken ahead to the release of CDC reports, some using a combination of machine learning and mathematical models and Others the usage of non conventional Internet based totally statistics structures together with Google Trends, Google Flu trends, Yahoo, Baidu Internet searches, Twitter posts, Wikipedia article views, Athena Health statistics base, Flu Near You, and clinicians' databases queries. This project specifically makes a speciality of non-traditional Internet-primarily based procedures, in which statistical and system getting to know algorithms able to presenting actual-time ("now forged") and forecast estimates of Influenza like ailments (ILI) by means of leveraging data from more than one resources which includes: Google Trends, Google Flu Trends, almost real-time health center visit facts supplied by way of Athena health facts base, Twitter posts, and statistics from CDC. Simple and effective models were built to forecast flu rates using these database, and to strengthen the forecasts of these weak predictors, ensemble models is used to provide real-time ("now cast") and forecast estimates .The proposed methodology exploits the statistics from each facts source and achieves to as it should be are expecting weekly ILI predictions for one week, two week, 3 week and 4 week ahead of the release of CDC's ILI reviews, successfully generating forecasts 3 weeks into the destiny. We trained our ensemble approach during the 2011–2012and 2012–2013 flu seasons and is evaluated on 2013-2014 and 2014-2015 flu seasons.

II. Literature Survey

A. Literature Review

In [2], Lamb et al. developed two phase approach to find flu infection rates from twitter data, firstly they used supervised learning based classification model, In which labeling of the set of tweets as non related or associated, and then categorised the associated tweets as flu related tweets or other viral disease related tweet or normal tweet. It was successful in classifying between reports of normal and flu related tweets. But had the disadvantage of inherent deficiency of twitter data under representation of elder, children and internet deprived areas. N Generous et al.[3] uses Wikipedia article get entry to logs and reputable sickness prevalence reviews, were used to build linear models to analyze and forecast on 14 diseases context, and finds the forecast value to be significant up to 28 days, but does not perform well with fast changing diseases pattern like influenza/flu. Ginsberg, J., Mohebbi, M., Patel, R. et al. [4] uses the Google search engine volume of specific terms for the prediction of Influenza Like Illness through linear models, and will be able to continuously provide real time estimates of ILI based on geographical boundaries, but overestimated the flu over certain time periods due to extreme media coverage and attention i.e. was not able to separate hype from actual infection rate. Mark S. et al. [5] developed the participatory flu surveillance system (FNY) used to gather data, segregate based on symptoms and make estimation through sampling, this uses clean collection of data, but is limited by samples, geography locations, ability to forecast for extended period of days.

B. Outcome of Literature Survey

Social media based surveillance go away at the peak of flu season by making higher estimation than ground truth , whereas government and voluntary based estimation has delayed estimation, and lesser coverage. Combination of the data sources may complement the advantages of one method over others disadvantage, thereby giving accurate forecast prediction.

C. Issues and Challenges

1) Proper clean data won't be available easily, manual cleaning is the main challenge.
2) Proper synchronisation of data is not there among data from five sources.
3) It is not possible to use present data, the project is done on the past data available.

D. Motivation

The recent influenza prediction systems do not give accurate results of flu rate for two to four weeks ahead. These already existing flu rate prediction systems do not extract any information from other data sources as discussed earlier. In order to extract information from all the data sources and use them wisely to improve flu prediction rate, this project uses ensemble models which uses predictions from single predicting model trained using each data source is used to provide real-time predictions (coming week prediction) and forecast estimates.

E. Problem Statement

Statistical and machine learning based methodology to combine information from search, social media and other statistics assets and data source to enhance Influenza Surveillance.

F. Objectives

1) Collection of data from five sources (Google Flu Trends, Google Trends, CDC data, Athena data, Twitter data)
2) Data cleaning and pre processing.
3) Machine-learning model for each type of data
4) Applying ensemble models
5) Visualisation of results

III. PROPOSED METHODOLOGY

The diagram 1 gives the abstract view of the developed model. Linear Regression model and Linear SVR model are used to predict each data source, and Linear Regression, LASSO, Linear SVR, AdaBoost regression with decision-trees, Kernel Ridge, Random-Forest Regressor, Gaussian Regressor are the regression model used as an ensemble models in this work.

A. Linear Regression Model

Linear regression is one of the classic regression model mainly used in time series problem. Here default weights were assigned to each feature of the data set. this statistical model will try to find linear relation among all the featured and with the output variable. This model tries to develop linear relation ship by fitting all the data points to linear line so that regression can be achieved. Here linear relation ship is found between input feature and output flu rate, hence it is termed as Linear Regression model. When features are plotted on x-axis and flu rate is plotted on y-axis, Linear Regression model gives the straight line which fits all the data points, according to the following formula,

$$y = b_0 + m_1 b_1 + m_2 b_2 + m_3 b_3 + \ldots\ldots + m_n b_n \quad (1)$$

Here b_0 is called as intercept point, and $m_1, m_2, ..., m_n$ are input variables.

B. Linear SVR Model

This model is based on the core principle of support vector machine, it is more flexible when compared with other forecasting models, because it can easily learn dependencies between features and output variable, even though the dimensions of the data used is very high. when compared with the Linear Regressor model, which strive to minimise the error rate, the SVR model strive to suit the error within a certain threshold. That means two boundary lines are defined on both sides of the regression line, and it is made sure that maximum of all the data points will fit within the boundary lines, and near to the in-between regression line, according to the following equation.

$$e \leqslant y - m_b - b_0 \leqslant +e \quad (2)$$

Here b and b_0 are the linear regression parameters, and e is nothing but distance of boundary lines from central regression line.

C. LASSO Regression Model

LASSO regression model is also a linear regression model, even this assigns some weights to each features like Linear Regression model, but it works on the shrinkage policy. It mainly try to regularise the weights used to learn dependencies among features to predict output. LASSO exhibits simple and sparse nature. LASSO works effectively when the data used is having high levels of multi co linearity or when automation of certain part of the model like feature selection or elimination is required. The equation used by this model is,

$$\sum_{i=1}^{n} \left(y_i - \sum_j x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (3)$$

This is the loss function, which should be minimised, where y is output, and x is input, and B are parameters, which were trying to shrink to zero.

D. AdaBoost Regression with Decision-Trees

In this model we use Decision trees as sub model, and Adaboost is used to stabilise the Decision Trees. Decision Trees is one of the machine learning model which is mainly used for classification and regression that means supervised learning. decision trees will split the data into different branches such that information gain from each branch is high, this means particular split will be effective in making forcast. Here whole data is represented in a tree form, where internal nodes are splits, and leafs are actual data points. While predicting it will find a average of output variables of the particular subtree. But this tree formation will be highly sensitive for the data, minute change in data point will change the tree structure. In order to overcome this issue we are using boosting methodology called AdaBoost layer. Adaptive Boosting (AdaBoost) regression model fits data by assigning weights to each version of data in order to generalise the model. At each iteration weights for wrongly predicted tree is improved, and final prediction will be the median of all tree values.

E. Kernel Ridge Regressor

Kernel ridge regression (KRR) is one of the regression model which uses ridge regression (linear least squares with l2-norm regularization) where kernel functions were used to improve existing ridge regression. It for this reason learns a linear function inside the space prompted through the respective kernel and the facts. It uses non-linear functions to execute non-linear kernel strategy in the original space. The way model learns the parameter is more over like support Vector Regression model. KRR uses squared error loss as loss function, and uses l2 regularization. KRR model works well for moderate dimension data when compared with other regression models like SVR. But the learned model is non sparse and not efficient when compared to SVR in speed.

F. Random-Forest Regressor

Random forest is one of the classic supervised machine learning model, and is used as an ensemble model using bootstrap aggregation as an ensemble methodology and decision tree were used as sub models. Here random subsets of training data is used to train one decision tree, and similarly n-decision trees were trained. During testing all trees will predicts some value and voting/mean is used to give final prediction.

G. Gaussian Regressor

Gaussian process is one among the classic regression models which is not limited by any functional form, So in preference to calculating the possibility distribution of parameters of a particular function, GPR calculates the chance distribution over all admissible capabilities that suit the information. Here the model calculates the posterior using training data, and predictive posterior distribution is computed on the specified points of interest.

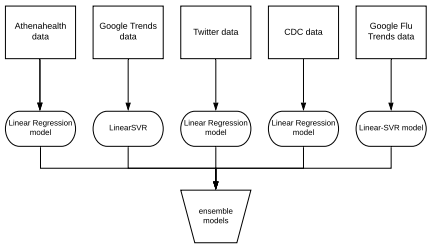

Fig. 1. Proposed Architecture

IV. WORK DONE

A. Data Collection

The proposed model requires data from five sources, called CDC database, Google Trends, Google Flu Trends, Athena database, Twitter Data set.

B. Data Source

1) CDC data: The CDC (Center for Disease Control) data coprises of ILI data for every week, where flu rate of every week is provided from 1907 till now. Every data point has eleven features including flu rate. This data gives total number of people suffered with ILI in that week in USA. This data is obtained freely from ILInet. via the online tool, here both historical and present data can be obtained [6].

2) Athena data: This is the data collected and managed by Athena health. It collected data from the patients who were seeking medical help from Athena health, which describes how many people are having flu type disease and how many are having other viral disease. We collected data from July 2009 to February 2015. Athena fitness information is commonly available at the least one week beforehand of CDC ILI reports.

3) Google Trends Data: We used Google Trends API, and searched Google trend values for 100 search words which are flu related words, and is done for every week, from 2004 till 2016. This data is directly obtained in the tabular format from Google Data sets.

4) Google Flu Trends: Google Flu Trends' provides the predicted value of flu rate in North America for every week[7]. The result of Google flu trend model is used to build prediction model. Data from 2003 till 2016 is collected from Google flu Trends web page.

5) Twitter data set: First twitter data is collected from twitter API, and is labeled as ILI related or not, then classification model is developed and number of tweets related to ILI and how many are not related is calculated from 2015 till 2019, the whole procedure is explained in [8,9].

Characteristics of the above mentioned data, is described in TABLE I.

C. Data Preprocessing

We used the above mentioned data from 2011 till 2015, where training is done on the 2011, 2012 year data, and tested on the remaining data. We used GA as one of optimisation

TABLE I
CHARACTERISTICS OF DATA

Data set	No. of Entries	No. of Features	Starting Year	Ending Year
CDC data	1147	11	1997	2019
Twitter data	314	5	2015	2009
Athena data	366	4	2009	2016
G flu trends	620	160	2003	2015
G trends	653	131	2004	2016

techniques to extract good features from Google Trends and Google Flu trends, application of GA to Google Trends data and Google Flu Trends data, helped in extracting informative feature and reduced it to nearly 50 features out of 162. We used manual row elimination technique to remove whole row if any of the variable of data point is not available. In this project we synchronised the data from five data sources, for the developed ensemble model.

D. Model Building

The proposed model predicts flu rate of the coming, next, two and three weeks ahead. For this we used data generated by five organisations as mentioned earlier, we used this data and built five estimators, where Linear Regressor model is used for CDC, Athena and Twitter data, along with Athena and Twitter data we used last three weeks flu rate to make it more accurate and efficient. For Google trend data and Google Flu trend data we used Genetic Algorithm with Linear SVR model. The output of these models is given as input to seven ensemble models called Linear Regression, LASSO, Linear SVR, AdaBoost regression with decision-trees, Kernel Ridge, Random-Forest Regressor, Gaussian Process Regressor as shown in Fig. 1.

V. RESULTS AND ANALYSIS

Evaluation Metrics used in this work are Root Mean Squared Error (RMSE) and Accuracy. Let yt_i and yp_i denote ith true value and predicted value respectively and n denote the number of datapoints.

1. RMSE: A measure of the difference between predicted and true values is defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (yt_i - yp_i)^2}{n}} \tag{4}$$

2. Accuracy: A measure of correctness to total number of predictions. In regression problems, accuracy is measured as the coefficient R^2. It is defined by following equation

$$R^2 = (1 - u/v) \tag{5}$$

where u is the residual sum of squares defined as

$$u = \sum_{i=1}^{n} (yt_i - yp_i)^2 \tag{6}$$

and v is the total sum of squares defined as

$$v = \sum_{i=1}^{N} \left(yt_i - \overline{yt}\right)^2 \tag{7}$$

TABLE II
ACCURACY OF ENSEMBLE MODELS

Ensemble Models	Accuracy			
	0-lag	1-lag	2-lag	3-lag
Linear Regressor	0.725	0.763	0.602	0.581
LASSO Regressor	0.735	0.764	0.602	0.581
LinearSVR Regressor	0.757	**0.773**	0.621	0.545
AdaBoost regression with DT	0.709	0.599	0.411	0.552
Kernel-Ridge	0.719	0.766	0.598	0.582
Random Forest Regressor	**0.759**	0.636	0.464	**0.632**
Gaussian Process Regressor	0.725	0.763	0.602	0.581
LASSO [8]	0.753	0.718	0.612	0.6
SVM(RBF) [8]	0.694	0.753	**0.659**	0.541
AdaBoost [8]	0.635	0.624	0.518	0.529
CDC Baseline	0.682	0.706	0.624	0.624

The best possible score is 1.0 and it can be also negative in worst case scenario.

In this work we are comparing results of proposed five single predictors and seven ensemble model with CDC's Baseline prediction and other existing systems. Results are found for zero, one, two and three lag of weeks, that means predicting coming weeks flu rate, next weeks flu rate, next to next weeks flu rate and next three weeks flu rate respectively, which is described in TABLE II and TABLE III, and visualised by plotting graph for each models.

A. Tabulation of Ensemble Models

Tabulation of results of all the proposed ensemble models and already existing model is done, and results of all the five weak predictors and well performed ensemble models is visualised in Fig 2. and accuracy is summarised in TABLE II, and RMSE is summarised TABLE III.

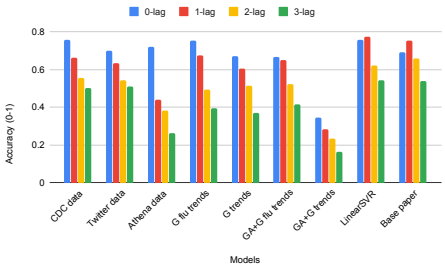

Fig. 2. Comparison of weak predictors

B. Plot of CDC Flu model

Linear Regression Model is trained on CDC data set for 2011-2013 year, and tested for 2013-2015 year, and results of test data is plotted for zero, one, two and three lag of weeks, and visualised in Fig. 3.

C. Plot of Twitter Flu model

Linear Regression Model is trained on CDC data set for 2011-2013 year, and tested for 2013-2015 year, and results of test data is plotted for zero, one, two and three lag of weeks, and visualised in Fig. 4.

TABLE III
RMSE OF ENSEMBLE MODELS

Ensemble Models	RMSE			
	0-lag	1-lag	2-lag	3-lag
Linear Regressor	0.608	0.565	0.737	0.757
LASSO Regressor	0.602	0.565	0.737	0.757
LinearSVR Regressor	0.572	0.554	0.719	0.789
AdaBoost regression with DT	0.626	0.735	0.897	0.782
Kernel-Ridge	0.615	0.561	0.74	0.755
Random Forest Regressor	0.569	0.705	0.856	0.71
Gaussian Process Regressor	0.608	0.565	0.737	0.757
LASSO [8]	0.213	0.477	0.766	0.857
SVM(RBF) [8]	**0.176**	0.352	0.527	0.507
AdaBoost [8]	0.251	**0.334**	**0.446**	**0.503**
CDC Baseline	0.501	0.71	0.863	0.977

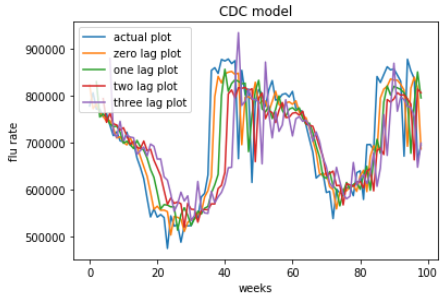

Fig. 3. Plot of CDC data

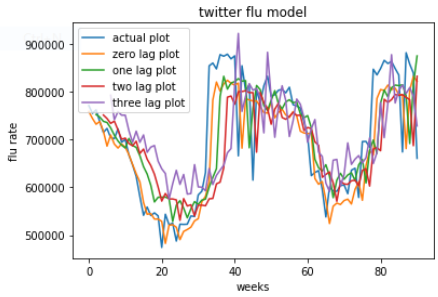

Fig. 4. Plot of Twitter Flu data

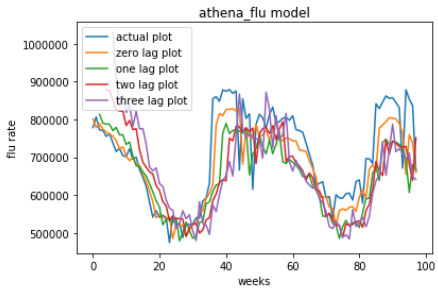

Fig. 5. Plot of Athena Flu data

D. Plot of Athena Flu model

Linear Regression Model is trained on CDC data set for 2011-2013 year, and tested for 2013-2015 year, and results of test data is plotted for zero, one, two and three lag of weeks, and visualised in Fig. 5.

E. Plot of Google Trends Flu model

Fig. 6. Plot of Google Trends Flu data

Linear SVR Model is trained on CDC data set for 2011-2013 year, and tested for 2013-2015 year, and results of test data is plotted for zero, one, two and three lag of weeks, and visualised in Fig. 6.

F. Plot of Google Flu Trends model

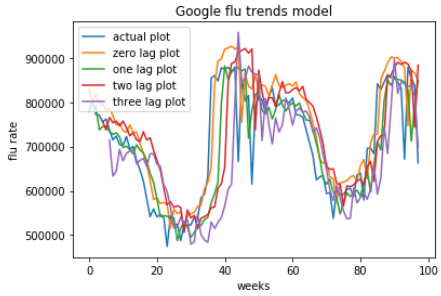

Fig. 7. Plot of Google Flu Trends data

Linear SVR Model is trained on CDC data set for 2011-2013 year, and tested for 2013-2015 year, and results of test data is plotted for zero, one, two and three lag of weeks, and visualised in Fig. 7.

G. Plot of Linear Regression ensemble model

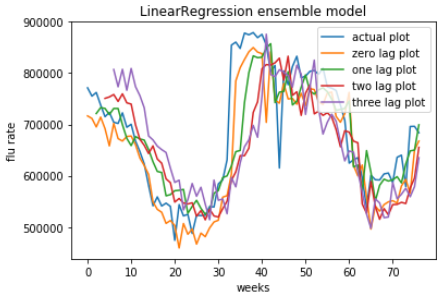

Fig. 8. Plot of Linear Regression ensemble model

Linear Regressor Model is used as an ensemble model, which is trained on predicted values of five single predictors

for 2011-2013 year, and tested for 2013-2015 year, and results of test data is plotted for zero, one, two and three lag of weeks, and visualised in Fig. 8.

H. Plot of Linear SVR ensemble model

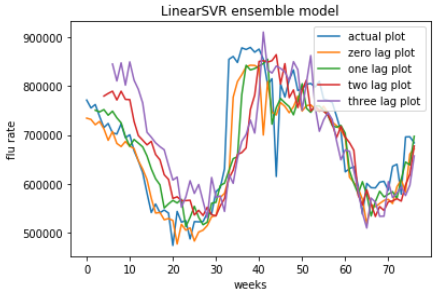

Fig. 9. Plot of Linear SVR ensemble model

Linear SVR model is used as an ensemble model, which is trained on predicted values of five single predictors for 2011-2013 year, and tested for 2013-2015 year, and results of test data is plotted for zero, one, two and three lag of weeks, and visualised in Fig. 9.

I. Plot of LASSO Regression ensemble model

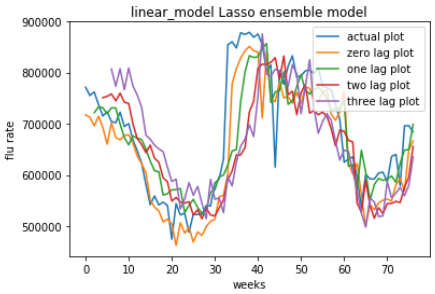

Fig. 10. Plot of LASSO Regression ensemble model

LASSO Regression model is used as an ensemble model, which is trained on predicted values of five single predictors for 2011-2013 year, and tested for 2013-2015 year, and results of test data is plotted for zero, one, two and three lag of weeks, and visualised in Fig. 10.

J. Plot of AdaBoost Regression + DT ensemble model

AdaBoost Regression with Decision Trees model is used as an ensemble model, which is trained on predicted values of five single predictors for 2011-2013 year, and tested for 2013-2015 year, and results of test data is plotted for zero, one, two and three lag of weeks, and visualised in Fig. 11.

K. Plot of KernelRidge Regressor ensemble model

KernelRidge Regressor is used as an ensemble model, which is trained on predicted values of five single predictors for 2011-2013 year, and tested for 2013-2015 year, and results of test data is plotted for zero, one, two and three lag of weeks, and visualised in Fig. 12.

Fig. 11. Plot of AdaBoost Regression + DT ensemble model

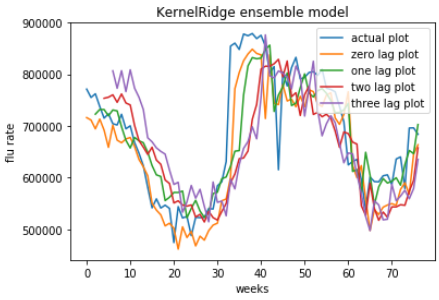

Fig. 12. Plot of KernelRidge Regressor ensemble model

L. Plot of Random-Forest Regressor ensemble model

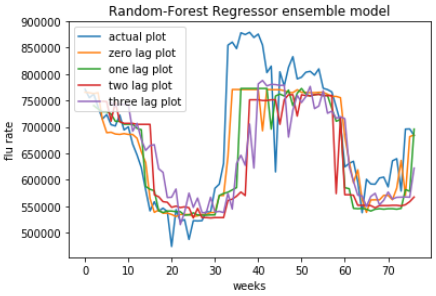

Fig. 13. Plot of Random-Forest ensemble model

Random-Forest Regressor is used as an ensemble model, which is trained on predicted values of five single predictors for 2011-2013 year, and tested for 2013-2015 year, and results of test data is plotted for zero, one, two and three lag of weeks, and visualised in Fig. 13.

M. Plot of Gaussian Regressor ensemble model

Gaussian Regressor is used as an ensemble model, which is trained on predicted values of five single predictors for 2011-2013 year, and tested for 2013-2015 year, and results of test data is plotted for zero, one, two and three lag of weeks, and visualised in Fig. 14.

For zero week lag Random Forest Regressor worked better than other proposed systems, and outperformed other existing systems. For one week lag Gaussian Process Regressor and Linear Regressor has given more accurate results when

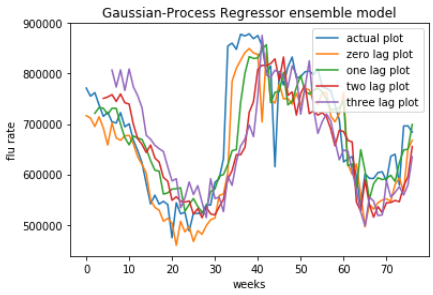

Fig. 14. Plot of Gaussian Regressor ensemble model

compared to other systems. For two and three weeks lag our model RMSE got decreased when compared to other existing systems.

VI. Conclusion

In this paper, we have presented five prediction models and used ensemble models to overcome the weak-ness of each model and make efficient model by using information from five data sources. Our ensemble approach uses real-time and historical information to accurately forecast flu estimates one, two, and three weeks into the future. The result obtained in proposed model of the paper has improved when compared to other existing systems. For future work, we would like to extend this study with deep learning time series models, and would like to use data from other sources to improve the existing model.

References

1 Who (2015) influenza(seasonal).
URL http://www.who.int/mediacentre/factsheets/fs211/en/index.html.

2 A. Lamb, M. J. Paul, M. Dredze, Separating fact from fear: Tracking flu infections on twitter, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Atlanta, Georgia, 2013, pp. 789–795 (Jun. 2013).
URL https://www.aclweb.org/anthology/N13-1097

3 N. Generous, G. Fairchild, A. Deshpande, S. Y. Del Valle, R. Priedhorsky, Global disease monitoring and forecasting with wikipedia, PLOS Computational Biology 10 (11) (2014) 1–16 (11 2014). doi:10.1371/journal.pcbi.1003892.
URL https://doi.org/10.1371/journal.pcbi.1003892

4 M. M. H. P. R. S. B. L. S. M. S. B. L. Ginsberg, Jeremy, Detecting influenza epidemics using search engine query data, Nature (2009/02/01). doi:10.1038/nature07634.
URL https://doi.org/10.1038/nature07634

5 M. S. Smolinski, A. W. Crawley, K. Baltrusaitis, R. Chunara, J. M. Olsen, O. Wójcik, M. Santillana, A. Nguyen, J. S. Brownstein, Flu near you: Crowdsourced symptom reporting spanning 2 influenza seasons, American Journal of Public Health 105 (10) (2015) 2124–2130, pMID: 26270299 (2015). arXiv:https://doi.org/10.2105/AJPH.2015.302696, doi:10.2105/AJPH.2015.302696.
URL https://doi.org/10.2105/AJPH.2015.302696

6 National, regional, and state level outpatient illness and viral surveillance.
URL https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html

7 Google flu trends.
URL https://www.google.org/flutrends/about/

8 M. Santillana, A. T. Nguyen, M. Dredze, M. J. Paul, E. O. Nsoesie, J. S. Brownstein, Combining search, social media, and traditional data sources to improve influenza surveillance, PLOS Computational Biology 11 (10) (2015) 1–15 (10 2015). doi:10.1371/journal.pcbi.1004513.
URL https://doi.org/10.1371/journal.pcbi.1004513