DOI 10.29042/2017-1988-1993

Helix Vol. 7(5): 1988-1993

Predictive Analytics as a Service on Tax Evasion using Gaussian Regression Process

[1] S.Kishore Babu, [*2] S.Vasavi

[1]Andhra Loyola Institute of Engineering & Technology College, India, [2] VR Siddhartha Engineering College, India

*Email: [1] skbjpj@gmail.com, [*2] vasavi.movva@gmail.com*

Received:8th May 2017, Accepted: 15th June 2017, Published: 1st September 2017

Abstract

Predictive analytics combines the capabilities of statistical analysis, machine learning and data mining. Vast amount of unstructured data produced by various public and private sectors such as government, health insurance, social media and academics gave the way for text analytics to make an insight into finding risk. Predictive analytics can forecast trends, determines statistical probabilities and to act upon fraud and security threats for big data applications such as business trading, fraud detection, crime investigation, banking, insurance, enterprise security, government, healthcare, e-commerce, and telecommunications**.** Predictive analytics as a service (PAaaS) framework is proposed in our earlier works. This paper gives solution to one of the application fraud detection in income tax data. The solution is based upon ensemble model that uses Gaussian process with varying hyper parameters. Performance measures NRMSE and COD are used to analyse the model. Test results proved that the third hyper-parameter values yielded a good result with less error rate and more variance which is reliable for a predictive model.

Keywords—Predictive Analytics, Gaussian process, Data Transformation , Statistical Computing, Tax Administration

I. INTRODUCTION

Data analytics uncover hidden patterns and correlations from large volumes of data by using techniques from statistics, machine learning, artificial intelligence and data mining. Predictive analytics (PA) refers to predictions such as customer relationship management, cross sell, healthcare insurance, risk management in banking, telecommunications, fraud detection about the future through analysis of data. Data analytics can be categorized into descriptive, predictive and prescriptive models. Descriptive models uses data aggregation to conclude upon what has happened, like mentioning the relationship between the data and describes past, predictive analytics that uses statistical models to forecast the future like what may happen and prescriptive analytics uses optimization techniques to suggest the ways of outcomes and their possible effects. Further predictive models can be categorized into classification models and regression models. Classification models determine class labels (categorical) where as regression models help in predicting a numeric value. Many techniques have been developed for predictive modeling such as SVM, Bayesian methods, neural networks, regression models, k-NN, uplift models and decision trees. But ensemble models proved to be achieving good accuracy when compared to others, reason being, they train several similar models and combines results so that a best model can be derived to predict new data. As explained in [1], predictive models can find relationship between outcome and dependent variables.Similarly descriptive models are used to form clusters of objects with similar characteristics. There are six phases for predictive analytic process. In the initial phase project is defined with outcomes, objectives, scope and the deliverables from the project. In the next phase data is collected from various sources and is analyzed. This analysis requires strategies for preprocessing such as data cleaning, transformation and data modeling so that useful data is extracted for further processing. Subsequently validate the initial hypothesis using statistical models. The next phase is predictive modeling for forecasting the future. Results after implementation can be deployed for using it in the day to day decision making. The last phase is monitoring the model in order to ensure that it is providing the expected results.

Fraud detection is necessary for any financial system. Nowadays, some people are deceiving government by not paying the taxes correctly. A huge loss is being reported by the government. Government has no proper estimate of how much tax is to be collected from the people. If it has a proper observation of the taxes received from previous years, government can make decisions regarding, amount of taxes that should be collected. A proper estimate of these amounts can make the task of fraud detection easier. If we can reduce the fraud that is associated with the tax collection, there will be an increase in the income for the government, which can be used for development activities. So the fraud can be detected by extracting the insights from the data which is discussed in this paper. This paper focus on the computing layer of our framework described in [2], that applies predictive analytic algorithm Gaussian process on income tax data set to identify fraud in projected tax values. The paper is organized as follows: section 2 presents literature survey on various predictive analytic algorithms and various existing web services. Section 3 outlines the proposed framework. Conclusions and future work is given in section 4.

II. MATERIALS & METHODS

Work reported in [3] explores the compliance and revenue consequences of the use of predictive analytics in an agent-based model that draws upon the behavioral approach to tax compliance. The belief and social custom feed into the occupational choice between employment and two forms of self-employment. It is shown that the use of predictive analytics yields a significant increase in revenue over a random audit strategy by affecting the subjective belief and enhancing the social custom. Agent-based modeling for tax evasion is discussed in [4]. Credit scoring and econometric analysis methods from predictive analytics are used. Their analysis compared the outcome of predictive analytics based on tax return data with that of random audits. Decision Trees are used for identifying behavior patterns and census data allows for the classification of those likely to commit fraud. Forecasting analysis such as Predict tax collection revenues to detect anomalies or errors during a particular time period is made. Predictive analytics and predictive risk modeling (PRM) models are reviewed such as Deloitte Consulting, the Public Consulting Group, the Case Commons Casebook model, and cutting edge work in New Zealand as reported in [5]. In an article published by economic times stated that Indian Government wants to use big data analytics tools in income tax department so as to differentiate black money holders from genuine tax-payers[6]. In another article [7], reported that predictive analytics can search vast amount of tax data for detecting tax evasion. Authors of [8] explained how Predictive data analytics saves lives and taxpayer dollars in New York City. In the work reported by [9] concluded that predictive analytics as a service should be dynamic and self improving so that it will provide right data to the right person at the right time and place. Authors of [10] discussed about feature engineering strategies for credit card fraud detection. Periodic behavior of the time of a transaction is analyzed using the Von Mises distribution. For the experiments they used three cost-insensitive classification algorithms: decision tree (DT), logistic regression (LR) and a random forest (RF), with and without the Bayes minimum risk threshold (BMR), and the cost-sensitive algorithms - cost-sensitive logistic regression (CSLR) and cost-sensitive decision tree (CSDT). Authors did not mention about features creation and their individual impact. A method is proposed in [11] to minimize the investment risk in stock market by predicting the returns of a stock using ensemble learning random forest algorithm. A case study on tax collection per month of the Federal Patrimony Department (SPU) is presented in [12]. In this authors analyzed Rule-Based Systems and Neural Networks classifiers for fraud detection and finally Neural Networks predictors for detecting fraud in time series data of the SPU. But their method works for small samples only. ANN along with Genetic algorithm is used to find credit card fraud detection in [13]. Deep learning, advanced machine learning algorithms are used for fraudulent detection in [14], Authors of [15] has used regression for predicting behavior of stock market analysis.

PROPOSED FRAMEWORK:

2.1 **Pre-Processing Using Z-Score Normalization**

Fig.1. presents flowchart of the proposed system. Data collected from variety of sources may have incomplete, noisy and in-consistent data. Preprocessing helps for data smoothing and normalization. Our framework uses Z-Score normalization for normalizing the data. When there are outliers, Z-Score proves to be better when compared to Min-Max Normalization. Z-score normalization is calculated using Eq (1) and Eq (2)

V' = (V - Mean) / StDev (1)

$$Z=(x-\mu)/\sigma \qquad (2)$$

Where Z is Z-Score, *X* is the data element, *μ* is the mean and σ is the standard deviation.

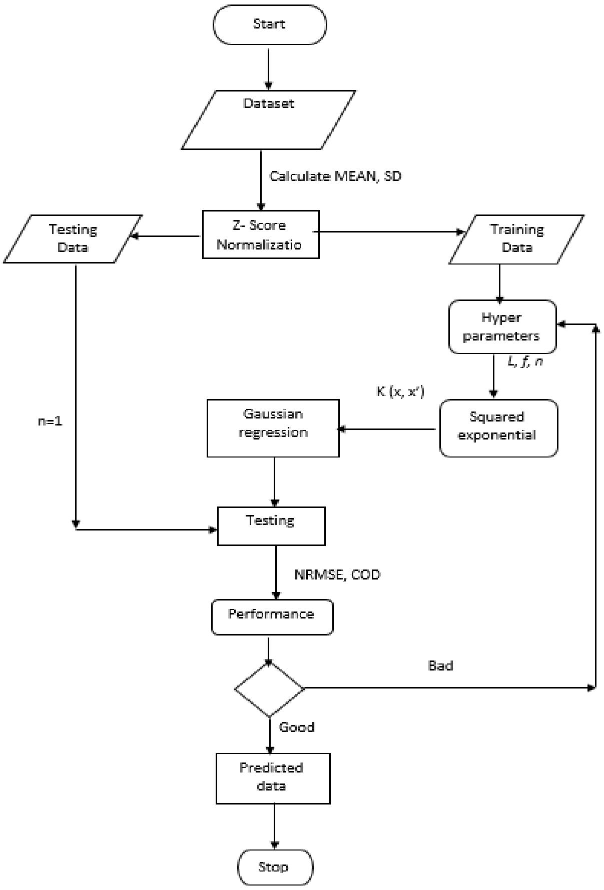


Fig.1.Flow chart of the proposed system

2.2 **Gaussian Process (GP)**

Gaussian process can be defined as given in Eq (3):

$$f(x) \sim GP(m(x),k(x,x_0)) \qquad (3)$$

Where m(x) is the mean function which is calculated using Eq (4) and $k(x,x_0)$ is the covariance function (kernel function of the Gaussian process).

$$m(x)=E[f(x)] \qquad (4)$$

Generally mean function is assumed to be zero, but authors of [16] states that a Gaussian process can have a zero mean function. More information about the non-zero mean functions could be found in [16]. For our implementation, we used zero value. Covariance between the two random variables f(x) and f(x₀) is calculated using Eq (5).

$$k(x,x_0)=E[(f(x)-m(x))(f(x_0)-m(x_0))] \quad (5)$$

2.3 Gaussian Process Regression

Covariance function should be selected so as to produce a distribution of random functions. For efficient forecasting, GP requires good number of training samples and targets. Squared Exponential covariance function is used to calculate covariance function as given in Eq(6).

$$y=f(x)+\varepsilon \quad (6)$$

Where Gaussian noise Ɛ is defined as $\varepsilon \sim \mathrm{N}(0,{\sigma_n}^2)$. Sometimes Squared Exponential covariance function as given in Eq(7) and Eq(8) is used for covariance function

$$f(x) \sim GP(0,k(x,x_0)) \quad (7)$$

where

$$k(x,x_0)=\sigma_f^2 \exp[(x-x_0)^2/2l^2] \quad (8)$$

Covariance function related to the target values y, denoted as $\mathrm{cov}(x,x_0)$,

And is defined using Eq(9):

$$\mathrm{cov}(x,x_0)=k(x,x_0)+{\sigma_n}^2\delta(x,x_0) \quad (9)$$

Where $\delta(x,x_0)$ is the Kronecker delta function which is equal to 1 if $x=x_0$ and 0 otherwise.

2.4 Selection Of Hyper-Parameters

The noise variance and the parameters in the kernel function are taken as the free hyper-parameters in Gaussian process [16]. This is represented by the vector *θ in Eq (10)*:

$$\theta=\{l,{\sigma_f}^2,{\sigma_n}^2\} \quad (10)$$

Where l is the characteristic length scale, ${\sigma_f}^2$ is the signal variance and ${\sigma_n}^2$ is the noise variance. In Gaussian Process Regression (GPR), these three parameters are obtained by learning the data. Bayesian model is used to infer these parameters (marginal likelihood maximization method). According to the Bayes rule, we can represent the posterior probability of the parameters as given in Eq (11):

$$p(\theta \mid x,y)=p(y \mid x,\theta)p(\theta)/p(y \mid x) \quad (11)$$

$p(y|x,\theta)$ is the marginal likelihood which is to be maximized, and *p(θ)* is the prior probability of the parameters.

The marginal likelihood can be calculated using Eq (12).

$$p(y|x,\theta)=\int p(y|f,x)p(f|x)df \quad (12)$$

The log value of the marginal likelihood gives the parameters which has been given in Eq (13) [16]:

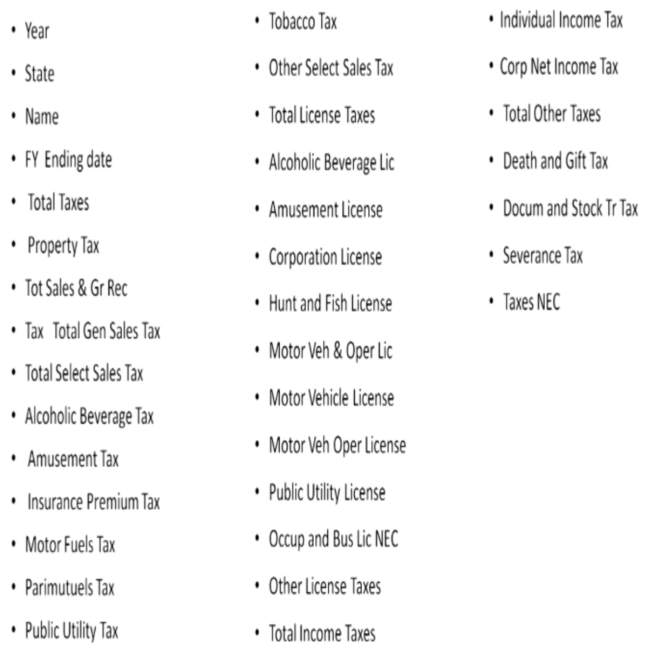
$$L(\theta)=\log p(y|x,\theta)=(-1/2)y^Tk^{-1}y^{-1}-(1/2)\log|k|-(n/2)\log 2\pi \quad (13)$$

Final step is to find best set of features for creating a predictive model because some features are better for predicting the target than others. Also, some features have a strong correlation with other features, so they will not add much new information to the model, and thus can be removed.

III Results and Discussion

3.1 Dataset used

Input dataset is State Government Tax Collections historical data available at https://www.census.gov/govs/statetax/historical_data.html. Fig.2. presents attribute list. Data from the Annual Survey of State Government Tax Collections are available online for each year from 1992 to the present. A historical dataset for years 1951 to present is available in Excel format. This paper presents the results of the analysis for the subset of dataset that comprises data of all the states and for the years 1994 to 2014. Also we considered property tax as the target variable for all the states over the last 20 years. R Studio software that provides a platform for statistical computing and handling large data sets is used for implementing the algorithm.



- Year
- State
- Name
- FY Ending date
- Total Taxes
- Property Tax
- Tot Sales & Gr Rec
- Tax Total Gen Sales Tax
- Total Select Sales Tax
- Alcoholic Beverage Tax
- Amusement Tax
- Insurance Premium Tax
- Motor Fuels Tax
- Parimutuels Tax
- Public Utility Tax
- Tobacco Tax
- Other Select Sales Tax
- Total License Taxes
- Alcoholic Beverage Lic
- Amusement License
- Corporation License
- Hunt and Fish License
- Motor Veh & Oper Lic
- Motor Vehicle License
- Motor Veh Oper License
- Public Utility License
- Occup and Bus Lic NEC
- Other License Taxes
- Total Income Taxes
- Individual Income Tax
- Corp Net Income Tax
- Total Other Taxes
- Death and Gift Tax
- Docum and Stock Tr Tax
- Severance Tax
- Taxes NEC

Fig.2. Attribute list

3.2 Perforance Measures

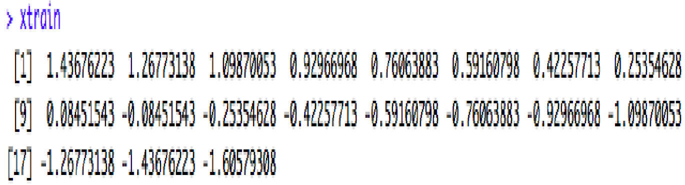
Normalized Root Mean Square Error (NRMSE can be calculated using Eq (14).

$$NRMSE=\sqrt{\frac{1}{N}\frac{\Sigma_i(d_i-y_i)^2}{\Sigma_i(d_i)^2}} \quad (14)$$

Coefficient of determination (COD) can be calculated using Eq (15)

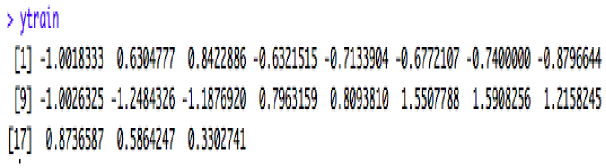
$$COD = R^2 = 1 - \frac{\Sigma_i (x_i - y_i)}{\Sigma_i \left(x_i - \bar{x}_i \right)^2} \qquad (15)$$

The training data sets are taken into xtrain and ytrain vectors as shown in Fig 3 and Fig 4. xtrain vector contains the year column values of training data set and ytrainvector contains the property tax data of training data set. The testing data set also have been into xtest and ytest vectors. A sequence of points are generated, at which Gaussian functions will be applied. v1, v2, v3 are the lines generated in the graph by applying Gaussian function at the points defined by sequence as shown in the Fig.5.



```
> xtrain
 [1]  1.43676223  1.26773138  1.09870053  0.92966968  0.76063883  0.59160798  0.42257713  0.25354628
 [9]  0.08451543 -0.08451543 -0.25354628 -0.42257713 -0.59160798 -0.76063883 -0.92966968 -1.09870053
[17] -1.26773138 -1.43676223 -1.60579308
```

Fig.3. Training Vector: xtrain



```
> ytrain
 [1] -1.0018333  0.6304777  0.8422886 -0.6321515 -0.7133904 -0.6772107 -0.7400000 -0.8796644
 [9] -1.0026325 -1.2484326 -1.1876920  0.7963159  0.8093810  1.5507788  1.5908256  1.2158245
[17]  0.8736587  0.5864247  0.3302741
```

Fig.4. Training Vector: ytrain

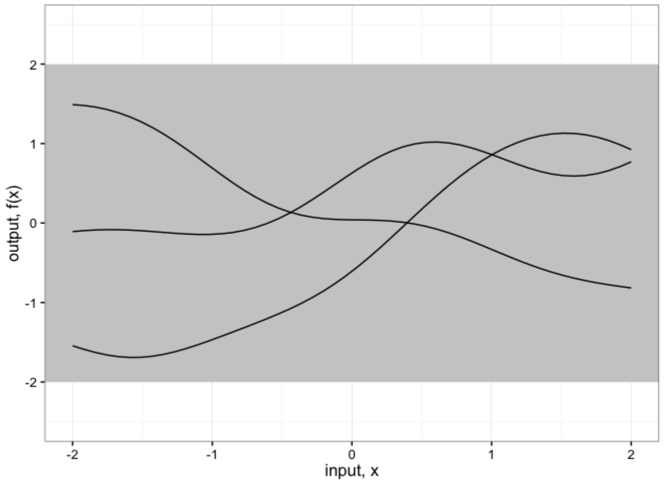


Fig.5. Sample Gaussian Functions plot

Now, after selecting the hyper-parameters values $\theta = \{1,1,0.004225\}$ and passing these values to Gaussian function, the covariance matrices are calculated. These covariance matrices are used to determine Gaussian function values for the corresponding input sequence of random numbers that are equally distributed using sequence function. The predicted value for the testing value will be the actual mean of the posterior Gaussian distribution. So, to find the prediction value for the testing value, the testing values is passed to x.star vector in the implementation code. After calculating the prediction values for all the states, the performance of model is calculated using NRMSE. The parameter y containing the predicted values and x containing the observed values are passed to NRMSE function. This function gives the error rate in the prediction process. Coefficient of determination describes how much of the variance between the two variables is described by the linear fit. Considering the prediction value, the calculated COD is :

COD = 0.90, for hyper parameters *l*=0.5, σ_f =1, σ_n =0.080

COD= 0.99, for hyper parameters *l*=1, σ_f =1, σ_n =0.065

The graph generated by taking the hyper-parameters *l*=0.5, σ_f =1, σ_n =0.080is as shown in Fig.6. The three variables v1, v2, v3 represent the Gaussian Functions.

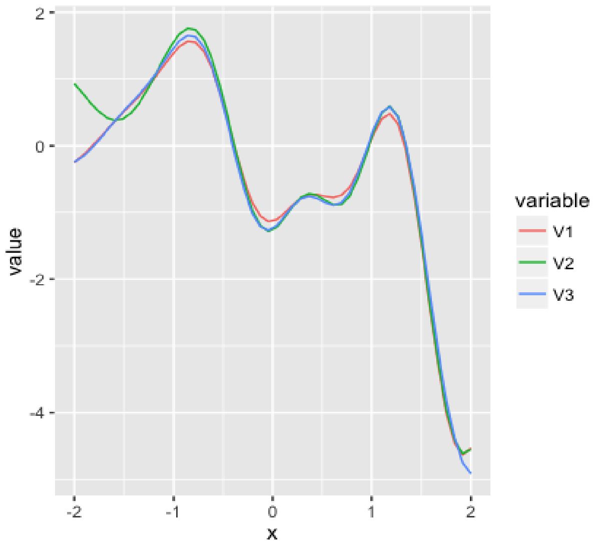


Fig.6. Training data graph plot with more error rate

On calculating the performance measure, the error rate that is obtained is high. Hence on changing the hyper-parameters to *l*=1, σ_f =1, σ_n =0.065, the graph plot generated after actual prediction is shown in Fig.7.

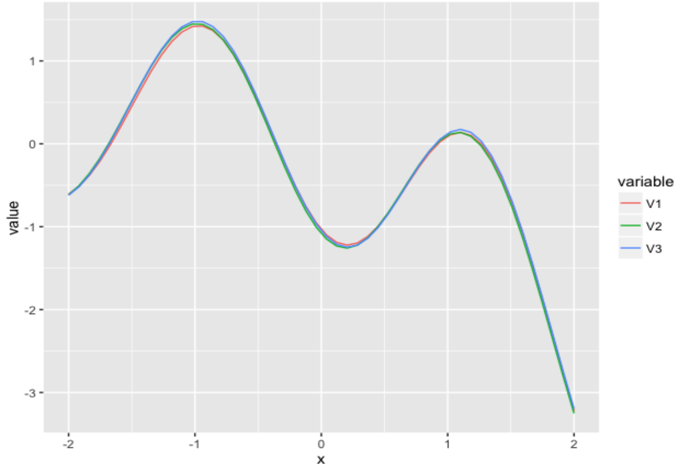


Fig.7. Training data graph plot

3.3 Analysis of performance measures:

Table 1 compares the performance of the model by varying the Hyper-parameters values.

Table 1: NRMSE and COD values with varying Hyper-parameters

S.NO	Hyper-parameters(l,σ_f,σ_n)	NRMSE	COD
1	0.5, 1, 0.80	54.4	0.90
2	1, 1, 0.1	14.3	0.98
3	1, 1, 0.065	3.8	0.99

By observing the table 1, it is obvious that different hyper-parameters produced different NRMSE and COD values. Considering the first Hyper-parameter values, the length-scale

is taken as 0.5, it means the function values change quickly and the noise variance is taken as 0.080 which means more noise is added to functions. So this generated more error rate and less variance. Now considering the second hyper-parameter values, the length-scale and noise variance are changed in such a way that it decreased the error rate and increased variance when compared to previous values. As more noise variance is added to the function in the previous two hyper-parameter values, the noise variance is decreased, that is less than the previous values. So the third hyper-parameter values yielded a good result with less error rate and more variance which is reliable for a predictive model.

IV Conclusion

This paper presented one of the case studies, where we build a predictive model that forecast the tax values. Regression has been used successfully in building the predictive model by selecting suitable hyper-parameters that are used in defining Gaussian functions. The performance of the predictive model has been measured by using Normalized Root Mean Square Error (NRMSE) and Coefficient of Determination (COD) performance measures of different hyper-parameter values such as $\theta = \{1,1,0.004225\}$, $\theta = \{0.5,1,0.0064\}$ and $\theta = \{1,1,0.001\}$. It was observed that when the value is $\theta = \{1,1,0.004225\}$ performance is good when compared to $\theta = \{0.5,1,0.0064\}$.

The field of Artificial Neural Networks, Gaussian process, Linear Regression model has diverse opportunities for future research in the predictive analytics. This paper presented results of Gaussian Process prediction technique to build the predictive model. In future, work can be extended by implementing various other techniques such as Artificial Neural Networks and Linear Regression Model techniques. By implementing all these techniques, we can determine which technique is more reliable by using performance measures such as NRMSE and COD.

V References

1. Frank Buytendijk and Lucie Trepanier, "Predictive Analytics: Bringing the Tools to the Data," *An Oracle White Paper*, Sep. 2010
2. S.Kishore Babu, S.Vasavi, K.Nagarjuna, “Framework for Predictive Analytics as a Service using ensemble model”, IEEE IACC 2017 [Inpress]
3. Nigar Hashimzade, Gareth D. Myles, Matthew D. Rablen, “Predictive Analytics and the Targeting of Audits”, 2014 https://tarc.exeter.ac.uk/media/universityofexeter/businessschool/documents/centres/tarc/publications/discussionpapers/Predictive_28-10-14.pdf
4. Gareth Myles with Nigar Hashimzade, Frank Page, Matt Rablen,” Targeting audits using predictive analytics”, 2013 https://tarc.exeter.ac.uk/media/universityofexeter/business school/documents/centres/tarc/publications/Targeting_Audits_Using_Predictive_Analytics.pdf
5. Dr.Thomas Packard, “*SACHS Literature Review: Predictive Analytics in Human Services*”, Sandiego state university school of social work, February 2016 https://theacademy.sdsu.edu/wp-content/uploads/2016/03/sachs-predictive-analytics-report-feb-2016.pdf
6. Sachin Dave, http://cio.economictimes.indiatimes.com/news/business-analytics/income-tax-department-to-use-analytics-to-look-for-discrepancies-in-bank-accounts/55849827 Last accessed on December 10th 2016
7. Learning Lab #2: Big Data and Predictive Analytics, https://www.vertexinc.com/blog/tax-matters/learning-lab-2-big-data-and-predictive-analytics Last accessed on May 4th 2017
8. Alex Howard, Predictive data analytics is saving lives and taxpayer dollars in New York City”, https://www.oreilly.com/ideas/predictive-data-analytics-big-data-nyc June 26th 2012 Last accessed on May 4th 2017
9. Ajit, Predictive Analytics as a service for IoT http://www.opengardensblog.futuretext.com/archives/2014/10/predictive-analytics-as-a-service-for-iot.html October 21st 2014 Last accessed on May 4th 2017
10. Alejandro Correa Bahnsen, Djamila Aouada, Aleksandar Stojanovic, Björn Ottersten ‘Feature engineering strategies for credit card fraud detection’, *Expert Systems With Applications*, , pp. 134–142 2016
11. Luckyson Khaidem, Snehanshu Saha, Sudeepa Roy Dey, ‘Predicting the direction of stock market prices using random forest’, **arXiv:1605.00003**,2016.
12. Antonio Manuel Rubio Serrano, João Paulo Carvalho Lustosa da Costa, Carlos Henrique Cardonha, Araripleno Almeida Fernandes,d Rafael Timóteo de Sousa Júnior, Neural Network Predictor for Fraud Detection: A Study Case for the Federal Patrimony Department, 2012 pp:61-66
13. Raghavendra Patidar, Lokesh Sharma,” Credit Card Fraud Detection Using Neural Network”, International

Journal of Soft Computing and Engineering (IJSCE) Volume-1, Issue-NCAI2011, June 2011 pp:32- 38

14. K. Anupriya, C. Kanimozhi, "Predicting Eshopping Data Using Deep Learning", Middle-East Journal of Scientific Research 24 (S1): 250-256, 2016
15. Farhad Soleimanian Gharehchopogh., Tahmineh Haddadi Bonab., Seyyed Reza Khaze., "A Linear Regression Approach To Prediction Of Stock Market Trading Volume: A Case Study," IJMVSC, vol. 4, no. 3, Sep. 2013, pp:25-31
16. Rasmussen, C.E., Williams, C. K. I, "*Gaussian processes for Machine Learning", 2006, chapter 4,* The MIT Press, Cambridge.pp:79-102