International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 4 451 – 455

Comparative Analysis of Classification Models on Income Prediction

[1]Bhavin Patel, [2]V. Kakulapati, [3]VVSSS Balaram
[1.2,3] Sreenidhi Institute of Science & Technology,
Yamnampet, Ghatkesar, Hyderabad, India
[1] *bhavinpatel.rns@hotmail.com*, [2]*vldms@yahoo.com*, [3]*vbalaram@sreenidhi.edu.in*

Abstract: Predictive Analytics is the underlying technology that can simply be described as an approach to scientifically utilize the past to predict the future to help coveted results. It is the branch of cutting edge analytics which is utilized to make predictions about unfamiliar events. Predictive analytics utilizes different procedures from information mining, insights, modeling, machine learning and artificial Intelligence. It includes extraction of data from information and is utilized to predict patterns and behavior patterns. It can be connected to an unfamiliar event or interest whether past, present or future. It helps being used of statistical algorithms information and machine learning strategies to distinguish the probability of future results in light of chronicled information. Income Determination is an important application of predictive analytics where customer segmentation takes place based on different demographical data. In this paper, we attempt to identify this purpose with a novel approach using different classification techniques to minimize the risk and cost involved to predict certain income levels. Here we have demonstrated the performance of each algorithm particularly on identification of customers using classification techniques. In addition, we provide an investigation analysis on true positives, false negatives, scored labels and scored probabilities.

Keywords: *Predictive Analytics, Statistics, Machine Learning, Data Mining, Classification*

*****

I. INTRODUCTION

With the power to predict income shifts based on past transactions, the Income Prediction Model can enable you to achieve a greater understanding of consumer and market behavior. With the intelligence you gain, you can better target programs or apps to specific income demographics and promote the right offers to the right consumers.

Sales and Marketing campaigns often require buyers with certain levels of disposable incomes for their products and solutions. Most of them are limited in their ability to provide accurate results and information. Almost, all companies use some sort of statistical modeling and regression techniques to predict potential customers of value to them. They use various tools such as the open source R programming language with features such as ggplots to visualize the trends or powerful scripting languages such as Python to create models. However, Retailers generally require more informed decisions about the type of products they need to stock and in order to accomplish this task and they seek commercial tools that are expensive.

The absence of good information is the greatest obstruction to associations trying to utilize predictive analytics. To make more accurate predictions [1] they need attributes of those products that buyers have bought in the past and a demographic attributes of customer such as Age, Gender, and Location.

The primary tool regression analysis [2] is used by associations for projecting analytics. For instance, an analyst may hypothesize a place of independent factors that are statistically associated with the purchase of the product. Using regression, how much every variable effects the behavior may be known. In this paper, we attempt to understand an emerging yet popular ML tool such as the Azure Machine Learning Studio that can assist us in understanding such similarities and differences.

II. RELATED WORK:

There are two major machine learning techniques, one is classification which is used to assign each dataset to predefined sets and another one is the prediction which is used to predict continuous valued function. The intention of categorization is to precisely calculate the objective class for every access in a certain data set.

SVM and PCA [3] are utilized to create and assess income prediction data in light of the Current Population Survey given by the Census Bureau of U.S. An itemized factual review focused for relevant element determination is found to increase the productivity and even enhance classification precision.

Vrushali [4] investigated and application of dissimilar algorithms like J48, Naïve Bayes and Random Forest Classifiers evaluates by fertility index to increase the algorithm performance.

According to Y. Bengio [5] networks are suitable in information-rich situations and are usually utilized for retrieving implanted facts in the structure of regulations,

quantitative assessment of these regulations, clustering, self-organization, categorization and regression. They have an improvement, over other kinds of machine learning algorithms for scaling.

S.Archana et al [6] emphasis about the major category of categorization algorithm includes C4.5, k-nearest neighbor classifier, Naive Bayes, SVM, and IB3. They explained a common survey of dissimilar categorization algorithms and their benefits and drawbacks.

III. OUR APPROACH

To develop a predictive for our experimentation may utilize the sample information set from the UCI repository where thousands of datasets belonging to different categories are available for free use and download. The dataset contains various attributes that are tagged as uninterrupted as they stand for numerical values whereas some factors have a predefined list of probable values. A sample 'Census Income Dataset' looks like the following:

Information set uniqueness:	Multivariate	The count of Instances:	48842
Feature uniqueness:	Categorical, Integer	The count of Attributes:	14
Related Tasks:	categorization	omitted Values:	Yes

Upon the successful selection of the dataset the next major step is to create an environment where one can model and test data. But in order to accomplish this feat, the missing values and inconsistencies (noise and outliers) must be removed, so that there is no duplication or unnecessary repetition and only unique values must exist. The omitted values require be updating with accurate values or dropping for accurate results. Clean Missing Data an inherent feature with the ML Studio might be used where the appropriate cleaning mode may be selected such as:

Custom Substitution Value, Replace with Mean/Median/Mode, Remove Entire Row, Remove Entire Column.

The next step is to split the dataset or to partition it for Training Data (Utilized for generating a analytical representation based on inherent prototypes originate in chronological information and Validation Data (Used for creating a new predictive model against known outcomes). Here again we can use the splitting mode such as: divide tuples, Recommender divisions, normal appearance or virtual appearance.

IV. PREDICTIVE ANALYSIS

Now we insert an ML design so that can train the model which is utilized to evaluate the data. The question isn't whether you can find the answers. The question is how. So, depending on what you find out you can choose the appropriate algorithm. For instance:

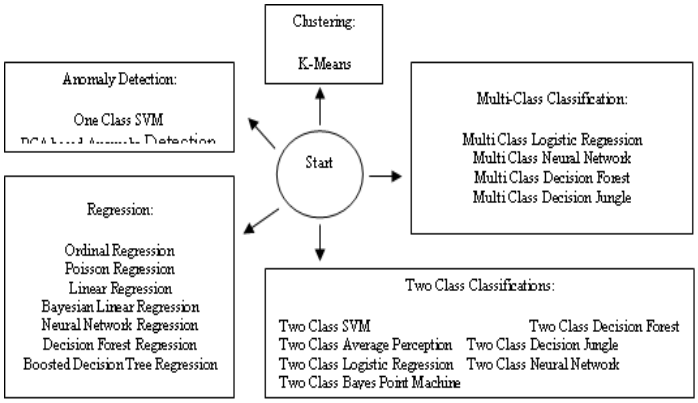

Fig 4.1. Framework of predictive analytics of income Prediction

4. i. Neural Network

This is one of the useful classification algorithms that utilize the perception of neurons that logically represents the working of the human brain. In this classification process, data values are representing by the neurons and the connectivity is representing by synapses. It is basically the layered approach in which there are two mai n layers called two end points represented by input and output layer. Other then these two layers, the model also have intermediate layers called hidden layers. On each layer some weight age is assigned. The middle layer of the network defines weights to different input values so that effective classification will be done. This allows the dataset as the input layer, and characterizes it as the network nodes. The predictor weights are applied to these nodes in the hidden layer. This layer actually defines the degree of connectivity between the nodes. After adjusting the weightage, output layer is derived as the final result [7, 8].

4. ii. Support Vector Machine

SVM is another robust and successful classification algorithm.SVM basically works as the linear separator between two data points to identify two different classes in the multidimensional environment. The main aim of SVM is to exploit the boundary between classes and to reduce the space between points. SVM basically defines the dealing of interaction respective to the features and the repetitive features. This divides the dataset in two vector sets beneath n dimensional space vector. This algorithm basically builds a hyper-
plane environment so that each element is compared with

respective to the separated linear line. The hyper-plane concept is presented to perform the data separation based on the largest distance analysis to identify the classes. To reduce the error ratio, the largest margin classifier is defined. The work also includes the analysis based on margin vector along with support vector analysis [9].

4. iii. Regression: It generates a condition to depict the factual connection between at least one of the indicators and the reaction variable and to anticipate new perception.

a. Ordinal Regression

This is utilized when the tag or objective column contains numbers, yet the numbers represent to a positioning or order instead of a numeric estimation. Anticipating ordinal numbers requires an different design than predicting the estimations of numbers on a continuous scale, due to the numbers allocated to represent to rank order don't have essential scale.

b. Poisson Regression

This is a special kind of regression that is typically utilized to model calculations. Approximation the quantity of emergency service calls during an event, or estimating customer inquiries relating to a promotion. Creating contingency tables since the response variable has Poisson dispersion; hidden suspicions about likelihood dissemination are unique in relation to minimum square relapses and Poisson models must be translated uniquely in contrast to other regression models.

c. Linear Regression

For predictive task, linear regression is a better choice. This type of regression is likely to effort well on high-dimensional, meager information sets not having complication. This is in Azure Machine Learning Studio utilizes to solve regression crisis: The typical regression predicament occupies a particular independent factor and a dependent factor. This is called simple regression.

d. Bayesian Linear Regression

In insights, the Bayesian way to deal with regression is frequently distinguished diversely in relation to frequent approach. The Bayesian approach utilizes linear relapse supplemented by extra data as an earlier likelihood appropriation. Earlier the information regarding the factors is connected with possibility capacity to produce approximations for the factors. Interestingly, the incessant advance, stand for regular least-square linear regression, except that the information includes adequate estimation to create a significant mode

e. Neural Network Regression

A neural framework can be considered as a weighted coordinated non-cyclic graph. The nodes of the graph reorganizing in layers and are relating with weight edges to nodes in the accompanying layer. The main layer is called as the information layer. The final layer is called as the yield layer. The output layer contains a single node on the account of the regression model.

The left over layers are called unseen layers. To process the yield of the network on given information case, esteem is calculated for every node in the unseen layers and in the yield layer. For every hub, the esteem is set by figuring the weighted total of the estimation of the nodes in the past layer and applied an initiation capacity to that weighted aggregate.

The structure of neural network model graph includes the following attributes:

- The count of unseen layers
- The count of nodes in every unseen layer
- relationships between the layers
- selection of creation functions
- loads on the grid edges

The organization of the graph and activation capacity are resolved by the consumer. The weights on the edges are originated while setting up the neural framework on information data. These networks can be computationally costly; due to various hyper-parameters and the arrangement of routine structure topologies. In spite of the fact that much of the time neural systems create preferable outcomes over different calculations, acquiring such outcomes may include a considerable lot of clearing (cycles) over hyper-parameters.

f. Decision Forest Regression

These trees are non-parametric models that perform a succession of simple tests for every instance, navigating a binary tree information structure until a leaf node attained.

These trees are effective in both computation and memory utilization amid preparing and prediction. They can describe non-straight decision limits. They perform integrated feature determination and categorizations are versatile within the view of noisy elements. Regression contains a group of decision trees. Each tree in a regression decision forest yields a Gaussian distribution by method for prediction. An amassing is performed over the group of trees to find a Gaussian distribution nearest to the joined appropriation for all trees in the model.

g. Boosted Decision Tree Regression

Boosting is solitary of several classic methods for creating ensemble models, along with bagging, random forests, and so forth. These trees in AML Studio utilize a proficient

execution of the MART GBS. To solve regression problems, gradient boosting is a machine learning technique. This constructs every failure tree in a stage insightful form, utilizing a predefined misfortune capacity to quantify the mistake in every progression and accurate for it in the subsequent. In this manner the prediction model is really a grouping of weaker forecast models. In regression issues, boosting constructs a progression of trees in a step-wise form, and afterward chooses the optimal tree utilizing a subjective differentiable loss work

. In this paper, since we are concerned with predicting values and estimating the relationship between variable we will use Regression and in order to predict categories and identify what categories the new information belongs to we use Classification.

Here we need to select the column to be expected stand on further columns. We start training the model and determine its suitability for the solution. We later visualize the newly trained data.

V. EVALUATING EXPERIMENT RESULTS.

In order to calculate the consequences, we establish the representation's accuracy. Here we notice the set of curves and metrics that are useful in evaluating the model. For instance, we have two additional parameters:

Scored-Labels: It predicts 'True' if the probability is greater than .5 and 'False' if it isn't.

Scored-Probabilities*:* It is the probability that algorithm has decided, that the above record belongs to 'True-Category'

Accuracy=Correctly Predicted Observations= 0.94
(Indicates 94% accuracy)
Total Number of Events

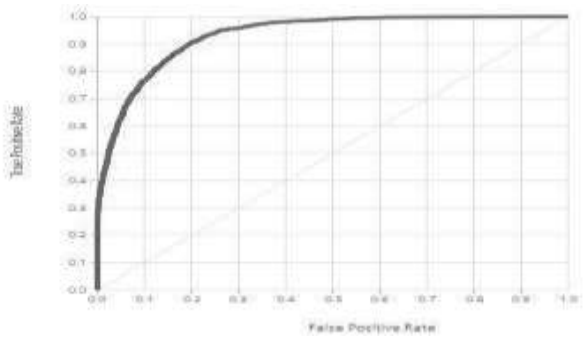

Fig 3.2. Recipient Operating Characteristic (ROC) Curve:

The above arc is known as 'Receiver Operating Characteristic' design. The Horizontal alignment stand for the ratio of Counterfeit Positives *(*event is 'Negative') and the Vertical alignment stand for the ratio of Exact Positives (event is 'Positive')

Exact- +ve: 1630
Counterfeit- -ve: 722
Counterfeit- +ve: 498
Exact- -ve: 6918
Positive-Label: >50K
Negative-Label: <=50K
Accuracy: 0.90

Confusion Matrix: The following matrix is known as "*Confusion Matrix"*. It predicts the Scored Labels against the actual class. For instance, this matrix indicates the accuracy of the *Multi-Classification Decision Forest* Algorithm against the actual classes i.e. High, Medium and Low.

In short, it predicts the Scored Probabilities of the predicted class.

The '*Average-Accuracy'* of the Multi-Classification Decision Forest using the sample dataset turned out to be 0.7.

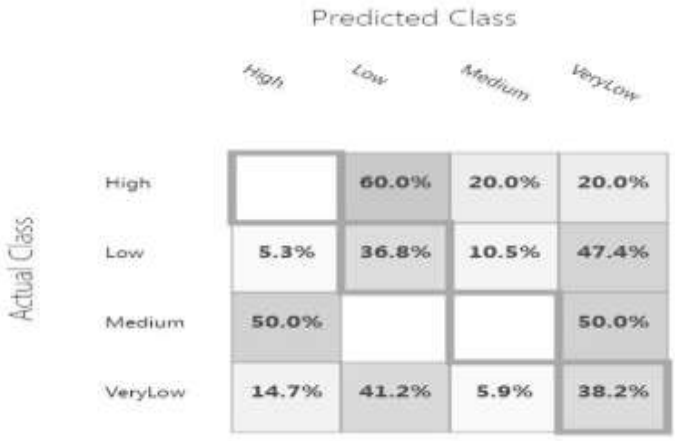

Fig 3.3. Multi-Classification Decision Forest

VI. CONCLUSION AND FUTURE WORK

With the proposed model, we measure the accuracy of various classification models directly by comparing the whole dataset. This new comparison gives the positive and improved result using the given metrics. The algorithm retrains itself each time an input variable is passed and compares itself to the previous scored label. However, we also observe that at times it gives us negative results. In order to overcome such negative impact, we enhance the model by carefully training it.

REFERENCES

[1] Anton Antonov et al., "Classification and Association Rules for Census Data" Mathematical for prediction algorithms, March 30, 2014 .

[2] Azamat Kibekbaev et al., "Benchmarking Regression Algorithms for Income Prediction Modeling", 2015 International Conference on Computational Science and

Computational Intelligence, 978-1-4673-9795-7/15 $31.00 © 2015 IEEE,DOI 10.1109/CSCI.2015.162, pp: 180-185.

[3] A Lazar. “Income Prediction via Support Vector Machine”, IEEE conference on Machine Learning and applications,16-18 Dec. 2004 DOI: 10.1109/ICMLA.2004.1383506.

[4] Vrushali “Comparative Analysis of Classification Techniques on Soil Data to Predict Fertility Rate for Aurangabad District” IJETTCS, Volume 3, Issue 2, March – April 2014, ISSN 2278-6856, PP:200-203.

[5] Y. Bengio et al., "Introduction to the special issue on neural networks for data mining and knowledge discovery," IEEE Trans. Neural Networks, vol. 11, pp. 545-549, 2000.

[6] S.Archana et al., “Survey of Classification Techniques in Data Mining”, International Journal of Computer Science and Mobile Applications, Vol.2 Issue. 2, February- 2014.

[7] Kumari et al , “Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction”, International Journal of Computer Science and Technology Vol. 2, Issue 2, pp. 304-308, 2011.

[8] Ture, M et al,” Comparing classification techniques for predicting essential hypertension”, Expert Systems with Applications 29, pp. 583–588, 2011.

[9] Burges, C., “A Tutorial on Support Vector Machines for Pattern Recognition” Data Mining and Knowledge Discovery, Vol. 2, pp. 121-167, 1998