Census Revenue Dataset

Shivam Tiwari[1], Prateek Parmar[2]
Kangaroo Software Pvt Ltd, Indore , India

Abstract- *This project aims to analyse census data provided by Census bureau database, while considering different factors such as net worth,structure at work,clients, market , product reputation, contribution in share market and risks with product etc. Also details about the companies are taken into account before signing them into a contract like product details,number of clients,stock market registered or not,yearly income. These companies can be big multinational companies to small startups and even small shops. Using random forest model to analyse the stock behaviour and Support vector machine model(SVM) as classification algorithm. The dataset encourages to draw valuable insights and conclusions. The conclusions drawn might help in delivering wiser decisions and also invest wisely in stock market. A website is developed as a platform for this project which consists of details about the companies after the analysis and the website can be used by the companies to reach out to us.*

Keywords- Analysis, Algorithm, Prediction, Stock market ,Random forest model,Support Vector Machine.

I. INTRODUCTION

The for-profit companies basically deliver products and services to generate revenues and provide profit for the benefit of the ones who own the company. Revenues, total costs of manufacturing and profit are the basic financial components of a business. A company must generate enough revenue to cover for its costs and earn a profit to continue as an ongoing enterprise,. Our aim is to find out meaningful insights that can be the basis for many clever decisions. Our dataset contains records with various attributes such as net worth,structure at work,clients, market , product reputation, contribution in share market and risks with product etc. Exploratory analysis will be done between dependent and independent variables. Not all the attributes are relevant for our analysis. The selection of the useful attributes will be based on the outcomes of the various algorithms. The variables are numeric as well as multiple factors. This analysis helps us to make smarter decisions while accepting the requests from companies that are willing to collaborate in our project. The platform we use here is a website which can be used to reach out to us and it also consists data about different companies, the results of analysis after analysing the trends in stock market for better decision making. We basically reach out to all the companies from giant MNCs to small scale companies that are willing to invest a small sum of money in our scheme. Then according to our policies we form a bond with them. The small amount is collected from these companies and a part of this money is invested in stock market. Our analysis helps us in making proper decisions and invest wisely. Hence if the company is in need for money, they can ask us for a certain amount.

II. RELATED WORKS

In this section we intend to provide an insight on different works done in the field of analytics to understand their significance in the relation with revenue. Sharath R et al, mentioned financial gain as their major domain of concentration and appropriate inferences were derived on that so as to extend the efficiencies,algorithms like neural nets, support vector machines(SVM) are often targeted for prediction and classification. The result from the model is biased because all the individual algorithms have additional or less the similar accuracy. [1] Bricker J examined the importance of standing in unit consumption[2].Financial choices victimization unit knowledge from the Survey of client Finances (SCF) were joined to neighborhood knowledge within the yankee Community Survey(ACS). The observations counsel that increasing inequality may need broader economic considerations. It includes a discount within the rate of savings and increase in the debt. the share of status cars within the given space is related with unit financial gain.Koskinen et al[3], outlined modeling of variation in individual wage in European nation by manipulating a superior dataset. A linear mixed-effects model was calculable based on age, the period of employment, and Gross Domestic product. The model provides predictors for individual wages. the information was divided into eight subgroups - each gender and financial gain quartiles. The proposed model was Associate in Nursing extension of the fundamental linear regression. It permits some model parameters to be drawn from a likelihood distribution. Meyer et al[4], evaluated consumption and financial gain measures of the fabric well-being of the poor.Consumption is most well-liked over financial gain. one in every of the long term goals of this analysis is to boost financial gain and consumption. Kennickell et al[5], place an attempt tounderstand the link between financial gain and wealth. Few made individuals seem within the lowest financial gain group. it's been witnessed that older operating individuals have higher properties and financial gain than their younger colleagues. On the opposite hand, folks that were retired from their job consumemore wealth and lesser financial gain than the children. so the link between income and wealth isn't robust.John R. Baldwin[6] evaluated the extent of labour productivity in Canada relative thereto of the u. s. in 1999. In doing therefore, he addresses 2 main problems. the primary is that the equivalence of the measures of gross domestic product and labour inputs that the applied mathematics agency of every country produces. Second, it investigates however a index number will be created to reconcile estimates of Canadian and U.S. gross domestic product per hour worked that square measure calculated in Canadian and U.S. bucks severally. Aleksander JARZĘBOWICZ[7] did comparative analysis of techniques enclosed in IIBA, REQB and IREB standards and also the list of issues rumored by practitioners related to techniques recommended as effective solutions.

.

III. INFERENCE

From the paper, “Data Analytics to predict the financial gain and Economic Hierarchy on Census Data”[1], we discovered that Benford’s law and Naive Thomas Bayes algorithms were accustomed predict higher pension schemes of the retired folks. however execution time ought to be reduced. In the paper, “Modeling and Predicting Individual Salaries: A Finnish Case Study”[3], An extension of linear regression model was used. They didn't record the part-time workers’ knowledge. They enclosed solely the working hours knowledge of regular staff on a daily basis.The analysis thereby resulted in restricted analysis on people that had stable jobs within the non-public sector.Here in, “A survey on identifying and addressing business analysis problems ”[7], a comparative analysis of techniques enclosed in IIBA, REQB and IREB standards was done and also the list of issues rumored by practitioners related to techniques were recommended as effective solutions to the given paper,“Factors predicting the labour productivity in the Us and Canada ”[6], used the equivalence of the measures of GDP and labour inputs that the applied mathematics agency in every country produces.However the indicators are often made to reconcile estimates of Canadian and U.S. GDP per hour worked that square measure calculated in Canadian and U.S. bucks severally. once doing so, and taking into consideration different assumptions concerning Canada/U.S. prices, the paper provides point estimates of Canada’s relative labour productivity of the whole economy in 1999 of around 94% that of the u. s.. The paper points out that a minimum of a ten decimal point confidence interval ought to be applied to those estimates. the scale of the vary is especially sensitive to assumptions that square measure created concerning import and export costs.

IV. PROPOSED METHODOLOGY

After referring the previous literature and also the inferences drawn from them, for the convenience of implementation, we will perform analysis between numerous attributes to analyse the financial condition of each and every company that we contact and use a website as our platform. Python language is used for our dataset. Analytics offer numerous techniques to handle all the data about the conmpanies and the factors that we will consider while making contact with the company. We will use various algorithms to carry out our analysis and the outcome will be displayed on our website. The website will consist data about each company.The analysis will be done by considering the following factors:

- Yearly Income
- Net worth and profit
- Structure at work
- Clients
- Market Network
- Product value
- Future value and uses of company product
- Contribution in the share market
- Risk with the product

The analysis will be done on the basis of these factors and then displayed on the website. Our database will contain all the data and the output will be evaluated using data analysis algorithms and then the output will be displayed on the website.Then there will be a section in our website to siaplay the trends in stock market. The page will show all the graphs related to the trends and how the company is doing in stock market if its registered. This will help us keep a record of the trends in market for proper investment..

V. ALGORITHMS USED

We use two algorithms for our analysis of the data based such as random forest model to analyse the stock behaviourand Support vector machine model(SVM) as classification algorithm.

A. Random Forest Model

Random Forest is a combination of multiple trees. Each tree is supplied with some random sample of knowledge with replacement and provides completely different classifications. It takes votes from the results of all the trees and chooses the classification having most votes and once the dependent variable is continuous, it takes the mean from the outputs given by completely different trees. The number of attributes given to the trees for classification is taken into account by taking the root of the total range of attributes. though every tree works on completely different attributes. therefore every tree can have a different root node and split. therefore for the ultimate result, the output of all the trees square measure thought-about. the amount of tress to be taken is tuned. One vital feature of random forest is that it shows the foremost vital as well as least vital variables within the dataset The random forest algorithmic program is wont to solve each classification and regression issues. It will handle large dataset with n range of input variables. It automatically takes care of the missing values. Since it is combination of multiple trees, the accuracy is expected to be a lot of more than the one call tree. Random forest isn't nearly as good for regression problems because it is for the classification issues.

B. Support Vector Machine Model

A Support Vector Machine (SVM) may be a discriminativeclassifier formally outlined by a separatinghyperplane. In alternative words, given labeled coaching information (supervised learning),the algorithmic program outputs associate best hyperplane that categorizes new examples. In 2 dimentional area this hyperplane may be a line dividing a plane in 2 elements wherever in every category lay in either aspect.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.When data is unlabelled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups. The support vector clustering algorithm, created by Hava Siegelmann and Vladimir Vapnik, applies the statistics of support vectors, developed in the support vector machines algorithm, to categorize unlabeled data, and is one of the most widely used clustering algorithms in industrial applications. The Regularization parameter (often termed as C parameter in python’s sklearn library) tells the SVM optimisation what quantity you wish to avoid misclassifying every coaching example. for giant values of C, the optimisation can opt for a smaller-margin hyperplane if that hyperplane will a higher job of obtaining all the coaching points classified properly. Conversely, a really tiny worth of C can cause the optimizer to appear for a larger-margin separating hyperplane, albeit that hyperplane misclassifies additional points.

VI. CONCLUSION

With the assistance of our work, we'll be analysing the financial gain of the company and analyzing the factors that powerfully have an effect on the financial gain. This data will help us keep a record of the financial records of all the companies and according to the factors that are considered by the company that we analyse using the dataset that we collect from various sources , we form a contract with the companies and lay down the policies. The website is a platform that we use for keeping the data after analysing it and the various factors that are mentioned. Then we have another page on the website that will consist of trends in stock market which will help us in making smarter decisions while investing in stock market. We will display the output of our data on the website about all the companies that are willing to sign a contract with us.

REFERENCES

[1] Sharath R , Krishna Chaitanya S, Nirupam K N, Sowmya B J and Dr K G Srinivasa(2016), Data Analytics to predict the Income and Economic Hierarchy on Census Data. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Bricker, J., Ramcharan, R. and Krimmel, J. (2014), Signaling status: the impact of relative income on household consumption and financial decisions, Federal Reserve Board, Finance and Economics Discussion Series (FEDS) Working Paper no. 2014-76. K. Elissa, “Title of paper if known,” unpublished.

[3] Koskinen, Lasse, TapioNummi, and JanneSalonen. Modelling and Predicting Individual Salaries: A Study of Finland's Unique Dataset. Finnish Centre for Pensions,2005.Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[4] Meyer, Bruce D., and James X. Sullivan. Measuring the wellbeing of the poor using income and consumption. No. w9760. National Bureau of Economic Research, 2003.

[5] Kennickell, Arthur B. "Using income data to predict wealth." Federal Reserve Board. Retrieved on April 22 (1999): 2005.

[6] R.Baldwin, J.Pierre Maynard, J. Tanguaye, M., Wong, F. AND Yan, B. "A Comparison of Canadian and U.S. Productivity Levels: An Exploration of Measurement Issues",2018

[7] C. Jayavarthini, 2 Ishu Todi, 3Kshitij Kumar Agarwal,” Analysis and Prediction of adultT income”

[8] Support vector machine. (2018). .http://dbpedia.org/page/Support_vector_machine

[9] ("Support vector machine", 2018) https://en.wikipedia.org/wiki/Support_vector_machine

[10] Xiang Yu, Zhihong Tian, Jing Qiu, and Feng Jiang (“A Data Leakage Prevention Method Based on the Reduction of Confidential and Context Terms for SmartMobileDevices”)https://www.hindawi.com/journals/wcmc/2018/5823439/