Social Media Analytics

Kashish Mandani, TE EXTC, Batch C, UID: 2017120038
Rohit Lambade, TE EXTC, Batch C, UID: 2018220069
Rahul Mane, TE EXTC, Batch C, UID: 2018220070

***Abstract*—To predict whether the customers purchase through the social media advertisements or not . The target variable is purchase of people through social network ad . For this dataset our trainset is “Gender”,”Age” and our testset is “Salary”.**

The classification goal is to predict if the customer will purchase anything from social media on influence of any advertisement. We have used Random forest classifier which creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

I. Introduction

MACHINElearning is the science of getting computers to act without being explicitly programmed. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome. Machine learning is so pervasive today that you probably use it dozens of times a day without knowing it. Many researchers also think it is the best way to make progress towards human-level AI. In this class, you will learn about the most effective machine learning techniques, and gain practice implementing them and getting them to work for yourself. More importantly, you’ll learn about not only the theoretical underpinnings of learning, but also gain the practical know-how needed to quickly and powerfully apply these techniques to new problems. Finally, you’ll learn about some of Silicon Valley’s best practices in innovation as it pertains to machine learning and AI.

This course provides a broad introduction to machine learning, datamining, and statistical pattern recognition. Topics include: (i) Supervised learning (parametric/non-parametric algorithms, support vector machines, kernels, neural networks). (ii) Unsupervised learning (clustering, dimensionality reduction, recommender systems, deep learning). (iii) Best practices in machine learning (bias/variance theory; innovation process in machine learning and AI). The course will also draw from numerous case studies and applications, so that you’ll also learn how to apply learning algorithms to building smart robots (perception, control), text understanding (web search, anti-spam), computer vision, medical informatics, audio, database mining, and other areas. Here numpy and scipy are the libraries Pandas — to load the data file as a Pandas data frame and analyze the data. If you want to read more on Pandas, feel free to check out my post! From Sklearn, I’ve imported the datasets module, so I can load a sample dataset, and the linear$_{m}odel, soIcanrunalinearregressionFromSklearn, sub-librarymodel_{s}election, I'veimportedthetrain_{t}est_{s}plitsoIcan, well, splittotrainingandtestsetsFromMatplotlibI'veimportedpyplot$ $Thesampleofdatausedtofitthemodel.TestDataset : Thesampleofdatausedtoprovideanunbiasedevaluationofafinalmodelfitonthetrainingdataset.$

The code for using Random Forest Classifier follows this sequence Import library Create model Train Predict

Classification accuracy alone can be misleading if you have an unequal number of observations in each class or if you have more than two classes in your dataset. Calculating a confusion matrix can give you a better idea of what your classification model is getting right and what types of errors it is making. Visualizations help in identifying patterns, relationships, and outliers in data.. It helps to build a compelling story based on visuals. Insights gathered from the visuals help in building strategies for businesses.

======================

II. Methodology

In the random forest, we grow multiple trees in a model. To classify a new object based on new attributes each tree gives a classification and we say that tree votes for that class. The forest chooses the classifications having the most votes of all the other trees in the forest and takes the average difference from the output of different trees. In general, Random Forest built multiple trees and combines them together to get a more accurate result.

While creating random trees it split into different nodes or subsets. Then it searches for the best outcome from the random subsets. This results in the better model of the algorithm. Thus, in a random forest, only the random subset is taken into consideration.

A very simple Random Forest Classifier implemented in python. The sklearn.ensemble library was used to import the RandomForestClassifier class. The object of the class was created. The following arguments was passed initally to the object: n$_{e}stimators = 10criterion =' entropy'$

The inital model was only given 10 decision tree, which resulted in a total of 10 incorrect prediction. Once the model was fitted with more the decision trees the number of incorrect prediction grew less.

It was found that a the optimal number of decision trees for this models to predict the answers was 200 decision trees. Hence the n$_{e}stimatorargumentwasgivenafinalvalueof200.$

Anything more that 200 will result in over-fitting and will lead further incorrect prediction. Within the social media interface, ads have proven to be the most effective source for brand discovery.

III. Result Analysis

We have your dataset of Social Network Ads. We subdivide this dataset into several subset. Per subset we train a Decision

Tree algorithm. In the end, we have trained several decision trees each returning their prediction for each observation in the dataset i.e. on basis of the gender and age and on observations of their salary we can easily predict the count of people purchasing products on influence of social media advertisements. We are now able to choose, per observation, what is the most probable answer to be expected. Individually, the predictions made by each model may not be accurate but combined together those predictions will be closer to the mark on average.

A. Applications

There are several applications where the random forest can be applied. We will discuss some of the sectors where random forest can be applied. We will also look closer when the random forest analysis comes into the role.

Banking Sector: The banking sector consists of most users. There are many loyal customers and also fraud customers. To determine whether the customer is a loyal or fraud, Random forest analysis comes in. With the help of a random forest algorithm in machine learning, we can easily determine whether the customer is fraud or loyal. A system uses a set of a random algorithm which identifies the fraud transactions by a series of the pattern.

Medicines: Medicines needs a complex combination of specific chemicals. Thus, to identify the great combination in the medicines, Random forest can be used. With the help of machine learning algorithm, it has become easier to detect and predict the drug sensitivity of a medicine. Also, it helps to identify the patient's disease by analyzing the patient's medical record.

Stock Market: Machine learning also plays role in the stock market analysis. When you want to know the behavior of the stock market, with the help of Random forest algorithm, the behavior of the stock market can be analyzed. Also, it can show the expected loss or profit which can be produced while purchasing a particular stock.

E-Commerce: When you will find it difficult to recommend or suggest what type of products your customer should see. This is where you can use a random forest algorithm. Using a machine learning system, you can suggest the products which will be more likely for a customer. Using a certain pattern and following the product's interest of a customer, you can suggest similar products to your customers.

IV. Conclusion

Random Forest Classifier being ensembled algorithm tends to give more accurate result. This is because it works on principle, Number of weak estimators when combined forms strong estimator. Even if one or few decision trees are prone to a noise, overall result would tend to be correct. Even with small number of estimators = 30 it gave us high accuracy as 97The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model's prediction on this subset. The data is split it into a training and testing sets, made predictions based on this data and tested the predictions on the test data. Classification accuracy alone can be misleading if you have an unequal number of observations in each class or if you have more than two classes in your dataset. Calculating a confusion matrix can give you a better idea of what your classification model is getting right and what types of errors it is making. Visualizations help in identifying patterns, relationships, and outliers in data.. It helps to build a compelling story based on visuals. Insights gathered from the visuals help in building strategies for businesses.

===

V. References

1. D. H. Fisher, ”Knowledge Acquisition via Incremental Conceptual Clustering”, Machine Learning, vol. 2, no. 2, pp. 139-172, 1987. CrossRef Google Scholar 2. L. Dent, ”A Personal Learning Apprentice”, Proc. 10th Nat'l Conf. Artificial Intelligence (AAAI '92), pp. 96-103, 1992. Google Scholar 3. J. Fisher, S. Hinds, D. D'Amato, ”A Rule-Based System for Document Image Segmentation”, Proc. 10th Int'l Conf. Pattern Recognition, pp. 567-572, 1990. View Article Full Text: PDF (663KB) Google Scholar 4. J. C. Schlimmer, D. Fisher, ”A Case Study of Incremental Concept Induction”, Proc. Fifth Nat'l Conf. Artificial Intelligence (AAAI '87), pp. 496-501, 1987. Google Scholar 5. M. Compton, Intelligent Purchase Request System, pp. 56-57, 1992. Google Scholar