International Journal of Research in Advent Technology, Special Issue, March 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

Hire Prediction System Using Machine learning

Abir Ojha, Prity Kumari, Jyoti Gupta and Nirmal Kumar Gupta
Department of Computer Science and Engineering
Jaypee University, Anoopshahr, U.P.

Abstract:-Machine Learning is defined as an application of artificial intelligence where available information is used by using algorithms to process or assists the processing of statistical data. While Machine Learning implicates concepts of automation and involves a high level of generalization to get a system that performs well on yet unseen data instances, it requires human guidance. Machine learning is a relatively new discipline in Computer Science that provides a collection of data analysis techniques. Some of these techniques are based on well-established statistical methods (e.g., logistic regression), while many others are not. The purpose of research is to predict the probability of a college student of getting hired or not getting hired by creating an application in Windows that will use few details as input and will give us an output of either being “hired” or “not hired”. The first phase of our paper involves the collection of data and creates a user-friendly GUI for end users. The final phase of the work is aimed at using algorithms in Machine Learning for the task of predicting the probability of a college student of getting either hired or not hired. We will present our findings in the Result section of our windows application.

Index Terms- Machine learning, Random forest, SVM, Hire prediction, Decision tree.

1. INTRODUCTION

Machines are increasingly doing “intelligent” things: Face book recognizes face in photos, Siri understands voices, and Google translates websites. The fundamental insight behind these breakthroughs is as much statistical as computational. Machine intelligence became possible once researchers stopped approaching intelligence tasks procedurally and began tackling them empirically. Face recognition algorithms, for example, do not consist of hard-wired rules to scan for certain pixel combinations, based on human understanding of what constitutes a face. Instead, these algorithms use a large dataset of photos labelled as having a face or not. To estimate a function f (x) that predicts the presence y of a face from Pixels x.
Machine learning (or rather “supervised” machine learning, which is the focus of this research paper) revolves around the problem of prediction: produce predictions of y from x. The appeal of machine learning is that it manages to uncover generalizable patterns. In fact, the success of machine learning at intelligence tasks is largely due to its ability to discover complex structure that was not specified in advance. It manages to fit complex and very flexible functional forms to the data without simply over fitting, it finds functions that work well out-of-sample.

We approach the problem from a nonverbal perspective where behavioural feature extraction and inference are automated. This paper presents a computational framework for the automatic prediction of hire ability. To this end, we collected a dataset of people who applied for a job where their marks in communication skills, of first round, their percentage in 12th class (grade), the number of internships they have done and the number of backlogs. We then evaluated several machine learning methods for the prediction of hire ability of the candidates based of the training data and showed the feasibility of conducting such a task.

2. RELATED WORK

Lodato MA et al. [1] proposed a predicting professional preferences for intuition –based hiring in which they presented a new improved predicting professional preferences for intuition –based hiring. This was the first study to examine the characteristics of HRM professionals that are associated with a preference for intuition-based hiring. The basic profile consisted of a general tendency to prefer feeling-based decision making in all life domains, less knowledge of HRM, and employment in a smaller organization. From a theoretical standpoint, their research suggests that scholars should begin developing propositions about the causal connections among these and related constructs. They need a better understanding of why people reject evidence-based practice, and what moderating variables are amenable to change. From a practical standpoint, their research suggests possible markers for predicting which HRM practitioners will be more receptive to science-based training. Kirimi J.M. et al [2] proposed an application of data mining classification in employee performance prediction in which they presents a new improved application of data mining classification in employee performance prediction. Their research paper focused on the possibility of building a classification model for predicting employee performance. Many performance attributes were tested using performance appraisal score for the year 2012 and 2013. Some of the attributes were found effective on the performance prediction. The Experience attribute had the maximum gain ratio, which made it the starting node and most effective attribute. Other attributes that appeared on the decision tree include Age, Qualification, Gender, Marital Status, Training and Performance Appraisal Score. For management of the School and HR Department, their model, and its subsequent enhancements, can be used in predicting the employee performance. Several actions can be taken in this case to avoid any risk related to hiring poorly performed employee.

3. RESEARCH METHODOLOGY

1. *Logistic Regression:*

In Logistic Regression modeling of an event occurring (Hired) versus event is not occurring (Not Hired). Using Logistic Regression, we identify probability of an event occurring (Hired) and event is not occurring (Not hired). Logistic Regression is one of the most used Machine Learning algorithms for binary classification. It is a simple Algorithm that you can use as a performance baseline, it is easy to implement and it will do well enough in many tasks. Therefore, every Machine Learning engineer should be familiar with its concepts. The building block concepts of Logistic Regression can also be helpful in deep learning while building neural networks. Like many other machine learning techniques, it is borrowed from the field of statistics and despite its name; it is not an algorithm for regression problems, where you want to predict a continuous outcome. Instead, Logistic Regression is the go-to method for binary classification. It gives you a discrete binary outcome between 0 and 1. To say it in simpler words, its outcome is either one thing or another. In this implementation we have use Binomial Distribution. When binary data and binomial distribution are mixed then it becomes Logit Function. Binomial distribution is a classification of binary data. Binomial distribution is used for find unknown probability.

The Logistic Equation:

Logistic regression achieves this by taking the log odds of the event ln(P/1-p) where, P is the probability of event. So, P always lies between 0 and 1.

Firstly, we will find Odds;

$$Odds = \frac{P(Hired)}{P(Not\ Hired)}$$

$$Odds = \frac{P}{1-P}$$

Where, P is the probability of event (Hired). So, P always lies between 0 and 1.

Logit function;

$$z_i = ln\left(\frac{P_i}{1-p_i}\right) = \alpha + \beta_1 x_1 + \cdots + \beta_n x_n$$

Taking exponent both side of above equation;

$$P_i = E(y = 1|x_i) = \frac{e^Z}{1+e^Z} = \frac{e^{\alpha+\beta_i x_i}}{1+e^{\alpha+\beta_i x_i}}$$

2. *Decision Tree:*

We have to use another algorithm in our project is Decision Tree for prediction of student which will be hired or not hired. Decision Tree Analysis is a general, predictive modeling tool that has applications spanning a number of different areas. In general, decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A decision tree is a tree-like graph with nodes representing the place where we pick an attribute and ask a question; edges represent the answers to the question; and the leaves represent the actual output or class label. They are used in non-linear decision making with simple linear decision surface.

Decision trees classify the examples by sorting them down the tree from the root to some leaf node, with the leaf node providing the classification to the example. Each node in the tree acts as a test case for some attribute, and each edge descending from that node corresponds to one of the possible answers to the test case. This process is recursive in nature and is repeated for every sub tree rooted at the new nodes.

A general algorithm for a decision tree can be described as follows:

1. Pick the best attribute. The best attribute is one which best splits or separates the data.
2. Ask the relevant question.
3. Follow the answer path.
4. Go to step 1 until you arrive to the answer.

The best split is one which separates two different labels into two sets.

Now that we know what a Decision Tree is, we'll see how it works internally. There are many algorithms out there which construct Decision Trees, but one of the best is called as **ID3 Algorithm**. ID3 Stands for **IterativeDichotomiser3**.Before discussing the ID3 algorithm, we'll go through few formulas.

Calculation:

Entropy:

$$H(S) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$$

Intuitively, it tells us about the predictability of a certain event (Hired or not hired).

Information Gain

Repeat until we run out of all attributes, or the decision tree has all leaf nodes. Information gain is also called as Kullback - Libeler divergence denoted by$IG(S,A)$ for a set S is the effective change in entropy after deciding on a particular attribute A. It measures the relative change in entropy with respect to the independent variables.

$$IG(S,A) = H(S) - H(S,A)$$

Alternatively,

$$IG(S,A) = H(S) - \sum_{i=0}^{n} P(x) * H(x)$$

Where$IG(S, A)$ is the information gain by applying feature A. H(S) is the Entropy of the entire set, while the second term calculates the Entropy after applying the feature A, where P(x) is the probability of event x.
ID3 Algorithm will perform following tasks recursively

1. Create root node for the tree
2. If all examples are positive, return leaf node 'positive'
3. Else if all examples are negative, return leaf node 'negative'
4. Calculate the entropy of current state H(S)
5. For each attribute, calculate the entropy with respect to the attribute 'x' denoted by H(S, x)
6. Select the attribute which has maximum value of IG(S, x)
7. Remove the attribute that offers highest IG from the set of attributes

3. *Random Forest:*

We have to use another algorithm in our project is Random Forest algorithm for prediction of student which will be hired or not hired. Random Forest is one of the most versatile machine learning algorithms available today. With its built-in assembling capacity, the task of building a decent generalized model (on any dataset) gets much easier.

In fact, the easiest part of machine learning is *coding*. If you are new to machine learning, the random forest algorithm should be on your tips. Its ability to solve—both regression and classification problems along with robustness to correlated features and variable importance plot gives us enough head start to solve various problems.
Random forest is a tree-based algorithm which involves building decision trees, then combining their output to improve generalization ability of the model. The method of combining trees is known as an ensemble method. Assembling is nothing but a combination of weak learners (individual trees) to produce a strong learner. Random Forest can be used to solve regression and classification problems. In regression problems, the dependent variable is continuous. In classification problems, the dependent variable is categorical. In simple words, Random forest builds multiple decision trees (called the forest) and glues them together to get a more accurate and stable prediction. The forest it builds is a collection of Decision Trees, trained with the bagging method.
Trivia: The random Forest algorithm was created by Leo Brie man and Adele Cutler in 2001 [3].
How does random forest work?
Creating a random forest
Step 1: Create a Bootstrapped Data Set
Step 2: Creating a Decision Trees
Step 3: Go back to step 1 and repeat
Step 4: Predicting the outcome of a new data point

4. *Support Vector Machine:*

We have to use another last algorithm in our project is Support Vector Machine algorithm for prediction of student which will be hired or not hired. Support vector machine is another simple algorithm that every machine learning expert should have in his/her arsenal. Support vector machine is highly preferred by many as it produces significant accuracy with less computation power. Support Vector Machine, abbreviated as SVM can be used for both regression and classification tasks. But it is widely used in classification objectives.
An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification, implicitly mapping their inputs into high-dimensional feature spaces.
HYPERPLANE
Hyper plane is a line that linearly separates and classifies a set of data.
SUPPORT VECTOR
Support vectors are the data points nearest to the hyper plane, the points of a data set that, if removed, would alter the position of the dividing hyper plane.
MARGIN
The distance between the hyper plane and the nearest data point from either set is known as the margin.
How to compute optimal hyper plane:
Let's introduce the notation used to define formally a hyper plane:
$f(x) = \beta_0 + \beta^T x$
Where β is known as the weight vector and β_0 as the bias.

4. RESEARCH WORK

In this project research paper the purpose is that we predict the probability of college student of getting hired or not getting hired by creating an application in Windows that will use few details as input and will give us an output of either being "hired" or "not hired". The first phase of our paper involves the collection of data and creates a user-friendly GUI for end users. The final phase of the work is aimed at using algorithms in Machine Learning for the task of predicting the probability of a college student of getting either hired or not hired. We will present our findings in the Result section of our windows application. In this project we have used 4 types of algorithms to calculate the majority of student to getting hired or not hired (1) Logistic regression (2) Decision tree (3) Random forest (4) SVM
Our first algorithms is logistic regression which we have implemented .In logistic regression; we are going to implement how the logistic regression model works in machine learning. The logistic regression model is one member of the supervised classification algorithm family. The building block concepts of logistic regression can be helpful in **deep learning** while building the neural networks. In this phase we measure the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function. Our second Algorithms is

Decision tree which we have implemented. Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. That will provide best predictor in our project. Our third Algorithms which we have implemented are Random forest .Random forest builds multiple decisions tree and merge them together to get a more accurate and stable prediction. And our last Algorithms which we have implemented is SVM which is support vector machine. The idea of SVM is simple: The algorithm creates a line or a hyper plane which separates the data into classes. In this phase I plan on offering a high- level of overviews of SVM.I will talk about the theory behind SVMs, its application for non-linearly separable datasets and a quick example of implementation of SVMs in Python as well.

5. CONCLUSION AND RESULT

This was a study to predict the hire ability of candidates out of which most of them are college students by the help of a dataset that had labels and marks in communication skills, of first round, their percentage in 12^{th}class (grade), the number of internships they have done and the number of backlogs. With the help of this data set we implement machine learning algorithms, we split the data into a training model and testing model and finally we tried to predict the chance of getting either hired or not hired based on the algorithms and the given data set. We finally conclude our findings in by creating a user friendly graphic user interface in form of a windows application and show the output of the each applied algorithm in a specific entry and then finally we take the majority of the results given by those algorithms and show it in the result section of our windows application which we placed at the top of our application, just below the title or heading of our application.

REFRENCES

[1] Lodato, M. A., Highhouse, S., & Brooks, M. E. (2011). Predicting professional preferences for intuition-based hiring. Journal of Managerial Psychology, 26(5), 352-365.

[2] Kirimi, J. M., & Moturi, C. A. (2016). Application of Data Mining Classification in Employee Performance Prediction. International Journal of Computer Applications, 146(7), 28-35.

[3] Cutler, A., & Breiman, L. (1994). Archetypal analysis. Technometrics, 36(4), 338-347.