International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

ELSEVIER SSRN

IJIEMR Transactions, online available on 15th Apr 2022. Link :http://www.ijiemr.org/downloads.php?vol=Volume-11&issue=ISSUE-04

DOI: 10.48047/IJIEMR/V11/I04/34

Title **A MACHINE LEARNING MODEL FOR PREDICTING THE PROFESSION AS SOFTWARE ENGINEER IN ASTROLOGY**

Volume 11, Issue 04, Pages: 239-247

Paper Authors
Udaya Sri Kompalli, Dr. Rudresh M Shastri

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

A MACHINE LEARNING MODEL FOR PREDICTING THE PROFESSION AS SOFTWARE ENGINEER IN ASTROLOGY

Udaya Sri Kompalli,
Research Scholar,
Maharshi College of Vedic Astrology
Udaipur, Rajasthan

Dr. Rudresh M Shastri,
Director of R&D , Head,
Dept.of Vedic Astrology,
Maharshi College of Vedic Astrology
Udaipur, Rajasthan

ABSTRACT

The Jyothish sastra is prominent and primitive in all sciences. It is developing from the beginning of the creation and is evergreen.

"यदा सुखा मयूरनं नागनं मनयो याद

तद्वत वेदांत शास्त्रं ज्योतिषम् मूर्हनिस्थितम्"

Astrology is supreme among vedic sciences like a peacock's tail and a gem on the snake's head. Beautiful peacock tail is known for material affairs and glorious rays of gem for knowledge. Planets are the main sources for predicting astrology. In the Solar System planets revolve around the sun in their orbits. Sages have defined the birth of a person as purely the game played by God. He is the blessing given by God. When a person is born we look into the time and date of birth. As they resemble the life of the native. We need to draw the birth chart by seeing the planetary positions in the solar system at the time of birth. The screenshot of the planetary positions at the time of birth is nothing but the "Rashi Chart" or "Birth Chart".

The human nature itself shows the interest about knowing the future. The present study is to predict the possibility of a person to select his profession as Software Engineer. In this research work, the data is collected from the working Software Engineers in different software companies and designed a model which will predict the profession by the training given to the mode.

Total 700 records were collected for the study from working professional Software Engineers and 300 records were collected who are not software engineers. To connect Scientific evaluation with the classic Vedic Astrology. Artificial Intelligence is the branch of science which has been evolving fastly and influencing more. In this present study, the data were collected to predict the profession by implementing various supervised learning classification techniques such as Logistic Regression, Support Vector Machine, Decision Trees, Random Forest, Naïve Bayes and KNN algorithms and the results were compared for accuracy.

Keywords

Artificial Intelligence, Machine Learning, Astrological Prediction, Classification Techniques, Predicting the profession as Software Engineer, Python.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

1. INTRODUCTION

Astrology is the science from old ages. It is the different branch which is related to the study of nature and behavior of a human being basing on the planetary positions of the native in birth chart. Now, in this trending era of technical innovations we try to implement the technical theories and implement the relation with astrology. In artificial Intelligence we are trying to develop a model which will learn the facts from the data provided to that and prepare different decision rules by its own. In general, the vedic astrology shows that the planet positions in different bhavas will yield the defined results. Basing on the nature and features of planets and stars an astrologer will predict the native nature and behavior. In this study we are trying to develop the model which will define the rules by the class defined as "Software Engineer", and "Other". The decision tree is used to take decisions which are built by the model.

Initially, for prediction we need to convert the available data from horoscope into a tabular format and pass it to the classification model. Then it will generate different rules and show the results of accuracy.

2. LITERATURE REVIEW

The birth chart of the native shows the planetary positions at the time of his birth. Basing on that one can analyze the results of each bhava. In total there are 12 bhavas which will show different aspects of the life of native. An astrologer can predict about the different levels of his life basing on the stars and planets. A Birth Chart is the screenshot of the planets in the universe. A lot of literature is available on astrology but the scientific validity could not be proved. In recent times there was as great change in the techniques applied and a lot of research has been done to establish a scientific validity for astrology.

Some researchers strongly believe that it is a part of hidden science which is not disclosed to a normal man. The ancient astrologers have made a belief that it is only possible to predict by devotional influence of God.

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans: *The* ability to learn. Machine learning is actively being used today, perhaps in many more places than one would expect. It is the learning and building of algorithms that can learn from and make predictions on data sets that are provided to the model to learn

There are two branches of machine learning.

1. Supervised Learning
2. Unsupervised Learning

Supervised machine Learning on a predefined set of training samples, which then facilitate its ability to reach an accurate conclusion when given new data. When the data is divided or classified based on the class manually and let the model learn from the given data which is known.

Unsupervised machine learning bunch of data and must find path there in. When the data is unclassified the model itself learns the path and pattern from the given data and predicts the results.

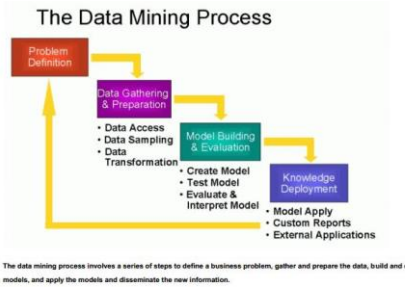
3. SUPERVISED LEARNING

In this paper we implemented the supervised machine learning method. For this we first need to collect the data and identify the columns. We have collected the data related to date of birth, time of birth, place of birth and generated a Birth chart. Next converting the whole data into dataset and given to the model to training.

In this paper we studied different supervised learning algorithms as Supervised learning

can be divided into 1) classification and ii)Regression. When the class attribute is discrete it is called Classification, when the class attribute is continuous, it is regression.

3.1 Decision Tree Learning (Random Forest, Decision Tree) –



The data mining process involves a series of steps to define a business problem, gather and prepare the data, build and evaluate mining models, and apply the models and disseminate the new information.

Decision is learnt from training the data set -Each non-leaf node in a tree represent a feature and each branch represent a value that the feature can take - Instances are classified by following a path that starts at the root node and ends at a leaf by following branches based on instance feature values -Construction of decision trees is based on heuristics.

3.2. Naive Bayes Classifiers

Bayesian network is a model that encodes probabilistic relationships among variables of interest. In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Naïve Bayesian technique is generally used for intrusion detection in combination with statistical schemes, a procedure that yields several advantages, including the capability of encoding interdependencies between variables and of predicting events, as well as the ability to incorporate both prior knowledge and data.

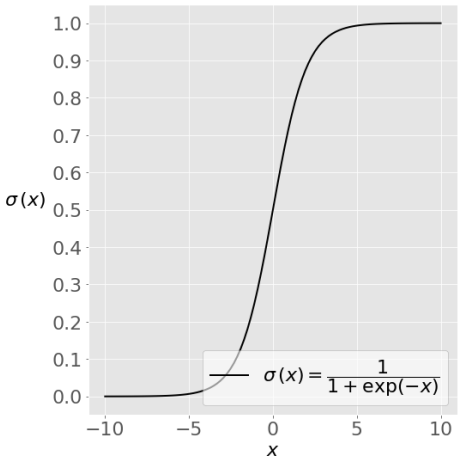
3.3. Support Vector Machine

An SVM maps input (real valued) feature vectors into a higher dimensional feature space through some nonlinear mapping. SVMs are developed on the principle of structural risk minimization [16]. -Structural risk minimization seeks to find a hypothesis (h) for which one can find lowest probability of error whereas the traditional learning techniques for pattern recognition are based on the minimization of the empirical risk, which attempt to optimize the performance of the learning set

3.4. Logistic Regression:

Logistic Regression is a statistical procedure for predicting the value of a dependent variable from an independent variable when the relationship between the variables can be described with a discrete data. Logistic regression is a fundamental classification technique that belongs to the group of linear classifiers and is similar to polynomial and linear regression. Logistic regression is fast and relatively uncomplicated, to interpret the results. As it's a method for binary classification, it can also be applied to multiclass problems.

To implement Logistic Regression one must have an idea about sigmoid curve. Where linear regression is a straight line passing through x and y axis Sigmoid curve shows the data as either 0 or 1.



The logistic regression of some dependent variable x on the set of independent variables x = (x_1 , …, x_r), where r is the number of predictors (or inputs), start with the known values of the predictors x_i and the corresponding actual response (or output) y_i for each observation i = 1, …, n.

Logistic regression is a linear classifier, so we use linear function $f(x)=b_0+b_1x_1+\ldots b_nx_n$ also called the logit. The variables $b_0, b_1, \ldots b_n$ are the estimators of the regression coefficients, which are also called the predicted weights or just coefficients. The logistic regression function (**x**) is the sigmoid function of (**x**): (**x**) = 1 / (1 + exp(−f(**x**)). As such, it's often close to either 0 or 1. The function (**x**) is often interpreted as the predicted probability that the output for a given **x** is equal to 1. Therefore, 1 − (x) is the probability that the output is 0.

Logistic regression determines the best predicted weights b_0 , b_1 , …, b_r such that the function p(x) is as close as possible to all actual responses y_i, i = 1, …, n, where n is the number of observations. The process of calculating the best weights using available observations is called **model raining** or **fitting**. There's one more important relationship between (**x**) and (**x**), which is that log(p(**x**) / (1 − p(**x**))) = f(**x**). This equality explains why (**x**) is the logit. It implies that (**x**) = 0.5 when (**x**) = 0 and that the predicted output is 1 if (**x**) > 0 and 0 otherwise.

3.5 UNSUPERVISED LEARNING

The unsupervised learning is the process of grouping the instances of similar objects. The label for each instance is not known to the clustering algorithm. This is the main difference between supervised and unsupervised learning. Any clustering algorithm requires a distance measure. Instances are put into different clusters based on their distance to other instances. The most popular distance measure for continuous features is the Euclidean distance: $d(X,Y) = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 + \ldots + (x_n-y_n)^2}$.

3.6. KNN:

KNN methodology is a commonly used clustering technique. In this analysis the user starts with a collection of samples and attempts to group them into 'k' Number of Clusters based on certain specific distance measurements. -K-Nearest Neighbor clustering generates a specific number of disjoint, flat (non-hierarchical) clusters.

Classification performance can be studied using the following format

Binary classification has four possible types of results:

1. **True negatives:** correctly predicted negatives (zeros)
2. **True positives:** correctly predicted positives (ones)
3. **False negatives:** incorrectly predicted negatives (zeros)
4. **False positives:** incorrectly predicted positives (ones)

The most straightforward indicator of **classification accuracy** is the ratio of the number of correct predictions to the total number of predictions (or observations).

Precision for Binary Classification

In an imbalanced classification problem with two classes, precision is calculated as the number of true positives divided by the total number of true positives and false positives.

True Positives (TP) False Positives(FP)

Precision = TP / (TP + FP)

The result is a value between 0.0 for no precision and 1.0 for full or perfect precision. If it is nearer to 0.9 we can say that the model is perfectly working.

Recall for Binary Classification

In an imbalanced classification problem with two classes, recall is calculated as the number of true positives divided by the total number of true positives and false negatives.

Recall = TP / (TP + FN)

The result is a value between 0.0 for no recall and 1.0 for full or perfect recall.

F-Measure

Once precision and recall have been calculated for a binary or multiclass classification problem, the two scores can be combined into the calculation of the F-Measure.

The traditional F measure is calculated as follows:

F-Measure = (2 * Precision * Recall) / (Precision + Recall)

METHODOLOGY

As we have selected Predictive Astrology, one branch of Astrology trying to implement these in Machine Learning Techniques and comparing with the different classification Techniques. For this the sample data has been collected from the working Software Engineers for 700 members and non Software Engineers for 300 members. We have collected the Birth Charts basing on the data given by them and converted into the Excel format by specifying the existence of planet in the house. Total instances taken for the analysis were 700 Software Engineers and 300 non Software Engineers. Noting the planetary positions in the following format. In this project we are trying to evaluate the causes for the native selecting the career as a Software Engineer. We have collected 700 actual details of the working Software Engineers in different Software Companies and 300 are the normal people who are not related to software field. We now examine different machine learning techniques with the available data and check for the accuracy and different rules generated by the Machine that it has learnt from the data.

Birth Chart is the main tool used for predictions of Astrology, that is based on the planetary positions in various Rasi. We now convert the existing birth charts of all the natives into an Excel format by noting the planets in each Rasi. The house number is noted down in the Planet column. Such as if Sun is located in Aries represent the value as 1, if in Sagittarius the value will be 9. Prepare a Excel Sheet for all the Data from the birth chart.

The following table shows the format. The columns are known as attributes and the rows are known as instances.

Sl.No.	Attribute	Type	Value
1	LAGNA	NUMERIC	1 to 12
2	SUN	NUMERIC	1 to 12
3	MOON	NUMERIC	1 to 12
4	MARS	NUMERIC	1 to 12
5	MERCURY	NUMERIC	1 to 12
6	JUPITER	NUMERIC	1 to 12
7	VENUS	NUMERIC	1 to 12
8	SATURN	NUMERIC	1 to 12
9	RAHU	NUMERIC	1 to 12
10	KETU	NUMERIC	1 to 12
11	ARIES	NUMERIC	1 to 12
12	TAURUS	NUMERIC	1 to 12
13	GEMINI	NUMERIC	1 to 12
14	CANCER	NUMERIC	1 to 12
15	LEO	NUMERIC	1 to 12
16	VIRGO	NUMERIC	1 to 12
17	LIBRA	NUMERIC	1 to 12
18	ACQUARIUS	NUMERIC	1 to 12
19	SCORPIO	NUMERIC	1 to 12
20	SAGITTARIUS	NUMERIC	1 to 12
21	CAPRICON	NUMERIC	1 to 12
22	ACQUARIUS	NUMERIC	1 to 12
23	PISCES	NUMERIC	1 to 12
24	CLASS	STRING	SE,OTH

Table : Format to store the data in Excel

7. EXPERIMENTS AND EVALUATION

We have designed a model by selecting all the 1000 records. When the data is supplied

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

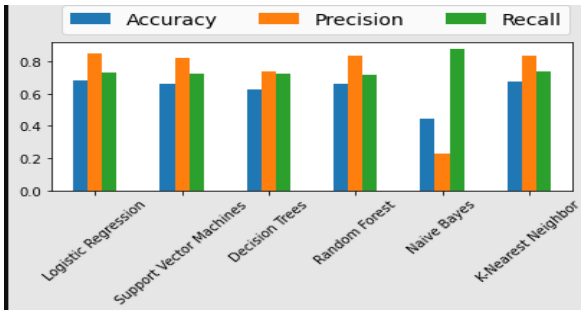
to the model it gets trained by the data. We have taken the data as 60% for training and 40% of data for Testing the results on the model we have developed.
Initially we have taken the data for 50 cases of software engineers and 30 cases of non software engineers. Then it is observed that

```
classifier = LogisticRegression
(solver='lbfgs', max_iter=100)
```

Logistic Regression for 50+30
True Positive(TP) = 10
False Positive(FP) = 6
True Negative(TN) = 4
False Negative(FN) = 4
Accuracy of the binary classification = 0.583

Classification Report TRAIN :		
	precision recall	f1-score support
0	0.77 0.50	0.61 20
1	0.77 0.92	0.84 36
accuracy	0.77	56
macro avg	0.77 0.71	0.72 56
weighted avg	0.77 0.77	0.75 56
76.785		
Classification Report TEST :		
	Precision Recall	F1-score Support
0	0.50 0.40	0.44 10
1	0.62 0.74	0.67 14
accuracy	0.58	24
macro avg	0.56 0.56	0.56 24
weighted avg	0.57 0.58	0.57 24
58.333		
Accuracy:	58.333	

	Accuracy	Precision	Recall
Logistic Regression	0.6800	0.8473	0.7318
Support Vector Machines	0.6618	0.8210	0.7255
Decision Trees	0.6254	0.7368	0.7253
Random Forest	0.6581	0.8368	0.7162
Naive Bayes	0.4436	0.2263	0.8775
K-Nearest Neighbor	0.6763	0.8315	0.7348



In the same way the sample size has been increased to check the change in accuracy. The has been a better improvement in the accuracy when the size of the data set is 250 each category.

Logistic Regression for 250+250
True Positive(TP) = 37
False Positive(FP) = 29
True Negative(TN) = 45
False Negative(FN) = 39
Accuracy of the binary classification = 0.547
Accuracy: 54.667

	Accuracy	Precision	Recall
Logistic Regression	0.54666	0.48684	0.56060
Support Vector Machines	0.56666	0.48684	0.58730
Decision Trees	0.53333	0.50000	0.54285
Random Forest	0.52000	0.47368	0.52941
Naive Bayes	0.52000	0.35526	0.54000

K-Nearest Neighbor	0.44666	0.40789	0.44927

In the same way the sample size has been increased to check the change in accuracy. The has been a better improvement in the accuracy when the size of the data set is 700 working and 300 not working. The training set was taken as 70% and 30% we get the following accuracy.

```
True Positive(TP)  =  195
False Positive(FP) =  82
True Negative(TN)  =  14
False Negative(FN) =  9
Accuracy of the binary
classification = 0.697
```

To improve the accuracy we have tried to implement the above with 60% and 20% training and testing set.

```
True Positive(TP)  =  130
False Positive(FP) =  54
True Negative(TN)  =  10
False Negative(FN) =  6
Accuracy of the binary
classification = 0.700
```

Classification Report TRAIN :			
	Pre Rec	F1-s	support
0	0.49 0.17	0.26	236
1	0.73 0.93	0.81	563
accuracy		0.70	799
70.33792240300374			
Classification Report TEST :			
0	0.62 0.16	0.25	64
1	0.71 0.96	0.81	136
accuracy		0.70	200
Accuracy: 70.000			

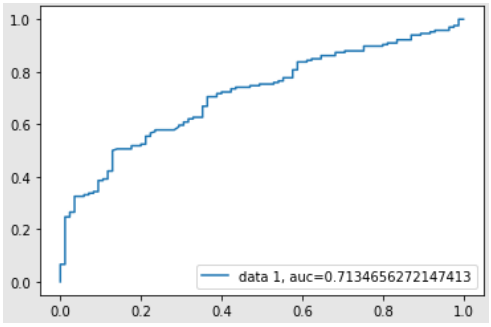
Model	Accuracy	Precision	Recall
Logistic Regression	0.700	0.955882	0.706522
Support Vector Machines	0.320	0.000000	0.000000
Decision Trees	0.620	0.735294	0.714286
Random Forest	0.680	0.838235	0.730769
Naive Bayes	0.595	0.595588	0.757009
K-Nearest Neighbor	0.660	0.838235	0.712500

By the above classifications it is evident that the Logistic Regression is best giving the results as 70% of accuracy. Now we select the above model and try to predict by providing the new data and evaluation is observed

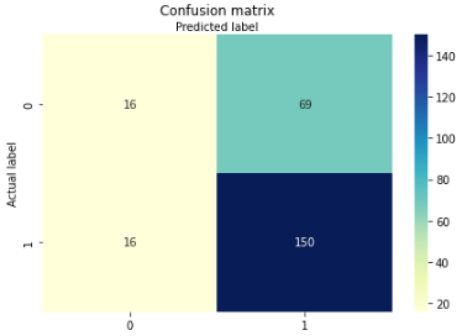
When the number of instances decrease, the performance of the algorithm will be less. For this we are collecting more number of test cases and observe the performance of algorithms. Machine Learning indicates high performance when the data is large. The ability to learn by example makes artificial neural networks very flexible and powerful. Some algorithms show high performance on the data and some show poor performance on the same data. So we select the best evaluated algorithm and compare the results with new data. For this we have selected the data of 50 working and 30 non working professionals and a prediction.csv file is generated.
Predicted values for 50+30 cases

```
True Positive(TP)  =  35
False Positive(FP) =  15
True Negative(TN)  =  7
False Negative(FN) =  25
Accuracy of the binary
classification = 0.700
```

ROC Curve



Confusion Matrix:



8. CONCLUSION AND FUTURE WORK

Large numbers of technologies are being developed for the extraction of meaningful data from huge collections of data using different Machine Learning techniques. Different tools, algorithms and methods which are being used to mine and analyze the data, perform differently on the data collections. Choosing the best algorithm to use for a specific analytical task can be a challenge. While you can use different algorithms to perform the same business task, each algorithm produces a different result, and some algorithms can produce more than one type of result.

8. REFERENCES

- Brihat Jatakam By Varahamihira
- Saravali
- Uttara Kalamrita By Kalidas, Translated By PS Sastri
- Chaplot N, Dhyani P, P. Rishi O. Astrological Prediction For Profession Doctor Using Classification Techniques Of Artificial Intelligence. *Int J Comput Appl*. 2015;122(15):28-31. Doi:10.5120/21778-5052
- Chaplot, N., Dhyani, P., & Rishi, O. P. (2013). *A Review On Machine Learning Concepts For Prediction Based Application "*. *I*(Ii), 12–18.
- Ganesh S, D. H. (2014). Comparative Study Of Data Mining Approaches For Prediction Heart Diseases. *IOSR Journal Of Engineering*, *4*(7), 36–39. Https://Doi.Org/10.9790/3021-04733639
- Jagtap, S. B., & G, K. B. (2013). *Census Data Mining And Data Analysis Using WEKA*. 35–40. Http://Arxiv.Org/Abs/1310.4647
- Kelly, I. W. (1997). Modern Astrology: A Critique. *Psychological Reports*, *81*(3), 1035–1066. Https://Doi.Org/10.2466/Pr0.1997.81.3.1035
- Kulkarni, P. S., Belokar, V. C., Sane, S. S., Bhale, N. L., Dept, I., W Poly, K. K., & Dept, C. (2012). Heart Disease Prediction From Horoscope Of A Person Using Data Mining. *International Journal Of Scientific And Research Publications*, *2*(7), 1–6. Www.Ijsrp.Org
- Safwat Jamil, L. (2016). *IJESRT International Journal Of Engineering Sciences &Amp; Research Technology Data Analysis Based On Data Mining Algorithms Using Weka WORKBENCH*. *5*(8), 262–267. Http://Awww.Ijesrt.Com
- Shajan, R., & Gladston Raj, S. (2019). Horoscope Analysis And Astrological Prediction Using Biased Logistic Regression (BLR). *International Journal Of Innovative Technology And Exploring Engineering*, *8*(12), 2187–2193.

About Authors:

Dr. Rudresh M Shastri:

An academician working as Director of R&D, HOD, in Department of Astrology at Shree Maharshi College of Vedic Astrology, Affiliated by ICPEM Regd., Under Niti Aayog P.C., Government of India, having 10+ years of experience in Vedic Astrology and Co founder of Sri Jyotisham. Graduated in MBA and MSc Psychology from Bharathiar University. Master of Arts in Vedic Astrology and winner of Gold Medal in PhD in Vedic Astrology from Maharshi College of Vedic astrology. Certified as life time Associate Research Astrologer by Maharshi College of Vedic Astrology.

UdayaSri Kompalli

A Research Scholar and an Academician, working as Associate Professor, in Department of CSE, NRI Institute of Technology, Pothavarappadu, Krishna Dist, Andhra Pradesh, India. Having 25 years of experience in Teaching Field with a Qualification of M.B.A., M.Sc., IT, M.Tech(CSE), MA(Astrology) and pursuing Ph.D. at Shree Maharshi College of Vedic Astrology, Udaipur. Won Total of 7 Awards out of which 3 are International Awards, and A World Book of Record Holder in 2022. Interested in developing Software for Examinations in Degree College and published 30 different papers in International Journals, Scopus and UGC Indexed Journals. Completed 40 Online Certification Courses in the time of Lockdown.