

Predicting Employability from User Personality using Ensemble Modelling

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering

in

Information Security

Submitted By

Sahil Sharma

(Roll No. 801333023)

Under the supervision of:

Dr. Deepak Garg

Associate Professor

Dr. Prashant Rana

Assistant Professor



COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

July 2015

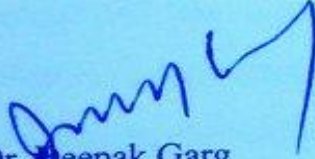
CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*Predicting Employability from User Personality using Ensemble Modelling*", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Information Security* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. Deepak Garg and Dr. Prashant Singh Rana*. Also it refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.


Sahil Sharma

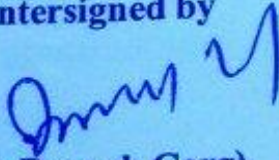
This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.


Dr. Deepak Garg

Head of Department

C.S.E Department

Countersigned by

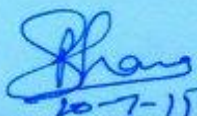

(Dr. Deepak Garg)

Head

Computer Science and Engineering Department*

Thapar University

Patiala


20-7-15
Dr. Prashant Singh Rana

Assistant Professor

C.S.E Department


(Dr. S. S. Bhatia)

Dean (Academic Affairs)

Thapar University

Patiala

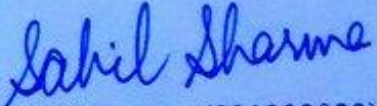
Acknowledgement

I would begin by saying that I am amazingly appreciative to my aide **Dr. Deepak Garg** for tolerating me as his understudy for working in theory and giving me opportunity to pick my point. He generally propelled me for doing the research at a quick pace and diving profound into exploration papers on the grounds that we may discover a path in them about what we need to do precisely in the early phases of the exploration work.

I additionally recognize my co-guide **Dr. Prashant Rana**, who has helped me, roused and guided me for the execution of the thought which I had been bearing for 6 months by then, since the first week of his joining in Thapar University. Prashant Sir taught me R programming and gave me genuine bits of knowledge into machine learning and information mining. Deepak Sir and Prashant Sir, both assumed their extraordinary part in my exploration work from begin to end. Both have their qualities and exceptional method for taking care of the understudy like me who knew nothing aside from having a thought.

I am thankful to **Dr. Jhiliak Bhattacharya**, P.G. Organizer, Information Security, Thapar University for the inspiration and motivation for the fruition of the proposal. I likewise express my appreciation to **Dr. S. S. Bhatia**, Dean of Academic Affairs in the University, for making procurements of framework, for example, library offices, PC labs outfitted with web office, tremendously helpful for the learners to furnish themselves with the most recent in the field.

Last yet not the slightest I need to express my ardent on account of my guardians, grandparents and my companions who with their provocative perspectives, veracity and entire hearted collaboration helped me in doing my proposal.


Sahil Sharma (801333023)

Abstract

In field of machine learning numerous down to earth executions have been done for diverse purposes. Be that as it may, less work has been done in field of psychology research taking help of machine learning and data mining strategies. Big five identity model is the most broadly acknowledged psychology research identity display that is acknowledged in the field of applied psychology. Five variables can characterize one's identity as indicated by Big five model and they are Openness, Conscientiousness, Extrovertness, Agreeableness, and Neuroticism, subsequently the acronym OCEAN. In the data that has been prepared and tried for the outcomes that are given in this paper, aggregate of 19719 perceptions have been recorded where each individual has given their reactions to aggregate of 50 inquiries. 10 inquiries for every identity attribute were inquired. Theory has been taken to choose the order of individual, in the event that he/she is employable taking into account a few scopes of diverse identity components. At last, 6 machine learning models have been checked for their exhibitions to foresee the employability of 70-30 model. Models with 90% above precision have been checked with their changes and blends for checking which of the mix is anticipating the best result i.e. with most astounding exactness and ensemble model for this assignment is chosen. Group procedure verifies that the proposed model will perform on mixture of datasets with high precision. On the off chance that base models don't give adequate result, ensemble model will beat any base model with best result conceivable.

Table of Contents

Sr. No.	Title	Page No.
i.	Certificate	i
ii.	Acknowledgement	ii
iii.	Abstract	iii
iv.	Table of contents	iv-v
v.	List of Figures	vi
vi.	List of Tables	vii
1.	Introduction	1-2
2.	Literature Review	3-10
	2.1 The Big Five Personality Test	3-4
	2.1.1 Openness	3
	2.1.2 Conscientiousness	3
	2.1.3 Extrovertness	4
	2.1.4 Agreeableness	4
	2.1.5 Neuroticism	4
	2.2 Machine Learning	4-6
	2.2.1 Supervised Machine Learning	5
	2.2.2 Unsupervised Machine Learning	5
	2.2.3 Ensemble Learning	6
	2.3 Data Mining	6-10
	2.3.1 Random Forest Algorithm	7
	2.3.2 Decision Tree Algorithm	7
	2.3.3 ADA Boost	8
	2.3.4 Support Vector Machine	8
	2.3.5 Linear Model	9
	2.3.6 Neural Network Model	10
3.	Problem Formulation and Objectives	11-15
	3.1 Understanding practicality of Machine Learning and Data Mining Techniques	11

	3.2 To study details of data mining techniques	12
	3.3 Realizing the value of data cleansing	12
	3.4 Efficient use of R programming and Rattle tool	12
	3.5 Advantages of Rattle over traditional R studio	13
	3.6 Detailed analysis through graphical representation	13
	3.7 K-Fold Validation	13
	3.8 Significance of Ensemble Modelling	13
	3.9 Analysis of Ensemble Models	14
	3.10 Value of MS-Excel for Data Scientist	14
4.	Tools and Methodology	16
5.	Implementation and Results	17-32
	5.1 Raw Data Collection	17
	5.2 Data Cleansing	17
	5.3 Hypothesis Applied	19
	5.4 Data Sampling: Training and Testing	20
	5.5 Supervised Learning and K-Fold Validation	23-30
	5.5.1 Supervised Learning	23
	5.5.2 K-Fold Validation	23
	5.6 Predictions of Models Applied	30
	5.7 Ensembling OR Ensemble Modelling	30
	5.7.1 How to do Ensemble Modelling using Microsoft Excel	31
	5.8 Result: Final Prediction	32
6.	Conclusion and Future Scope	33
	Video Presentation	34
	References	35-36
	List of Publications	37

List of Figures

Figure No.	Description	Page No.
1	Decision Tree	08
2	SVM – Linear and Kernel Classification	09
3	Methodology	16
4	K-Fold Graph	29

List of Tables

Table No.	Description	Page No.
1	OCEAN Personality Traits	01
2	Raw Data	18
3	Raw Data After Cleansing	18
4	Hypothesis for Deciding Employability	19
5	Employability calculation using Hypothesis	20
6	Prediction results in Testing File	21
7	Results of Classification Models – seed 42	22
8	K-Fold Validation – seed 42	24
9	K-Fold Validation – seed 857279	24
10	K-Fold Validation – seed 429822	25
11	K-Fold Validation – seed 620316	25
12	K-Fold Validation – seed 373955	26
13	K-Fold Validation – seed 798731	26
14	K-Fold Validation – seed 754410	27
15	K-Fold Validation – seed 1111	27
16	K-Fold Validation – seed 645946	28
17	K-Fold Validation – seed 549403	28
18	Comparison of Ensemble Models	30
19	Ensemble Generation in Excel	32

In present time, employment is the essential need of each passing college alumni and post graduate in every nation. Rising populace as of late and rising number of school graduates passing their school consistently has increased current standards of employability in the matter of choice technique at any firm in any part going from data innovation, educating, exploration, deals, advertising and so forth. Organizations have begun considering the personality component of the individual being enlisted. Identity is the center of the individual that will choose the amount of good or terrible the individual will perform in the organization. His/her disposition towards work will choose if the undertaking will be conveyed on time. There is an exceptionally famous personality test that is accessible for anybody to take to check what the key components of their identity are, and the name of the model is Big five personality test. In beginning of Big five identity show, the five attributes that characterized the model were Surgency, Agreeableness, Conscientiousness, Emotional Stability and Intellect [1]. Be that as it may, later more famous portrayal of this model came as O.C.E.A.N [2] and the elements are Openness, Conscientiousness, Extrovertness, Agreeableness and Neuroticism. Fundamental qualities of both the forms of the Big five models are same. Table I indicates mixed bag of feelings in connection to every identity quality.

TABLE I. OCEAN PERSONALITY TRAITS

Big five feature	Emotions related to high score	Emotions related to low score
Openness	Fantasy, aesthetics, feeling deeply, curious, analytical, creative, imaginative.	Conventional, change-resistant, unimaginative, fear of risk taking.
Conscientiousness	Thorough, confident, cautious, ordered, dutifulness, logical, full of aspiration, self-disciplined.	Non-disciplined, easy going, unordered, spontaneous.
Extrovertness	Warmth, assertiveness, outgoing, gregariousness, excitement seekers.	Emotionally less expressive, feeling discomfort around

		people, prefer to stay alone, introspective-thinkers
Agreeableness	Trust others, direct, generous, compliance, humble.	No trusting others, cunning, greedy, rude, rebellious.
Neuroticism	Anxious, angeriness, depressed, inferiority complex, impulsive, vulnerable.	Optimistic, calm, composed, happy.

We first find the employability of the person based on 50 big five personality questionnaire and also study the impact of personality in deciding whether the person is employable or not. Also, required things to be employable will be suggested because personality decides the attitude of the person including the dedication, discipline and learning attitudes. So, that is how this model works. Users will enter their answers based on whatever they want to rate as, 1 – very inaccurate, 2 – moderately inaccurate, 3 – neither accurate nor inaccurate, 4 – moderately accurate, 5 – very accurate with the statement.

Bagging and Boosting are the two techniques that are used in different machine learning techniques. Bagging can be defined as Train-Test-Repeat method, using the K-Fold Validation Technique. K-Fold Validation is discussed in section 3.7 and section 5.5.2. Whereas, Boosting applies to the technique of feeding the data slowly and making the machine learn by increasing the weight gradually and making machine more intelligent in terms of artificial intelligence. [16]

Ensemble models combine more than one machine learning model making itself more effective than one single base model results [4, 5, 6]. Classification machine learning models have been taken into account because predicting the employability is the binary class prediction i.e. either the person is employable (class 1) or the person is unemployable (class 0). General analysis of the models will show what percentage of people were employable out of the complete dataset of 19719 observations. Inference will be made about what characteristics should be developed to be employable. This ensemble model is checked for its efficiency by using R programming, so it has great advantage over other ways for building ensembles in terms of handling huge databases. Using server side programming for relatively large databases which can't be handled on personal computers and laptops, this ensemble model will still work.

The relation between personality traits is studied under “The Big Five”. In our case study also data set is built on the test using the Big Five factor markers from international personality item pool, developed by Lewis R. Goldberg in 1992. [1]

2.1. The Big Five Personality Test

This is a five factor model, popularly known as the big five model built on characteristics of the personality in the field of psychology. The big five characteristics consists of five elements, described in Table I, OCEAN Personality Traits. Different types of analyses has been done based on personality traits. Prediction of personality using twitter data [2], Emotions and personality has been monitored from Facebook profiles for full one year and then analyzing when the people are happiest and not happy month-wise and the emotions analysis through the statuses of people in those periods [8]. Personality has also been studied in a different way on social platform such as analyzing the profiles of famous and not-famous people on Facebook [9].

Detailed discussion of the five characteristics of the personality is as follow:

2.1.1. Openness

Openness is closely related to open mind concept. Curiosity, new ideas, balanced discussions possible too because they are open to other person’s ideas. New experiences are welcomed too. High scorers with this characteristics are highly on emotions too, because they feel more, their aura is affected, hence they can see things which others can’t see. Openness is related to overall human body, they like to learn new things more than others.

2.1.2 Conscientiousness

Conscientious is related to the people who have developed nature of discipline and like to complete work in time in a planned manner. People scoring high in this aspect are more organized, hence more effective. They are perfect to be a leader. As, they have art of planning... they have an edge over others. Only clear mind can have the clarity of thoughts

and vice-versa. Conscientious people have characteristics of punctuality. They are ordered and does thorough work always.

2.1.3 Extrovertness

Extrovertness is a characteristic we have heard always in contrast to introvertness. Extrovert is the person who has the ability to connect with others better and are livelier in the outer world. They think less and act more. Introverts are the ones who are known to be the thinkers and most often all the life changing inventions have been done by the introverts. Introvertness makes you think and streamline thoughts are more common in an introvert if they have the knowledge of value of thoughts. Extroverts often act on their instincts.

2.1.4 Agreeableness

This characteristic relates to direct attitude person, being humble, generous, and full of compliance and trust others. People with high score in this criteria are good team players. Though they are low on leadership quality but they make a very good team work.

2.1.5 Neuroticism

This characteristic relates to the negative attitude such as anger, depression, anxious, inferiority complex. People high on this area are not best fit to be employable or work under pressure.

2.2. Machine Learning

Machine Learning is relatively a new field when compared to psychology. Broadly machine learning can be classified in to two main categories:

- i. Supervised Machine Learning.
- ii. Unsupervised Machine Learning.
- iii. Ensemble Learning

2.2.1. Supervised Machine Learning

Supervised machine learning is known to be supervised for a reason and that reason being machine is provided with the basic data from the user end. Data may be in different forms such as excel sheet, csv (comma separated values) sheet, MATLAB file or some other format. Concept behind supervised learning being machine will learn the cases from the data input and do predictions based on that. Naïve Bayes method is the best example for showing how this work is done. Interested candidates can study Naïve bayes implementation from a book named “Data Smart” written by a data scientist himself and all the concepts are implemented easily in Microsoft Excel.

There are two cases in supervised machine learning and those are:

- Regression Technique
- Classification Technique

Regression algorithms work on real values and Classification algorithms work on whole numbers. This statement is the sole explanatory statement for practical implementation of the supervised machine learning area using R programming. I have worked in R, so all claims are made in relation with R programming only.

Though raw data is never purely consisting of real values or whole number values. Raw data has a lot of noise i.e. unwanted data or unnecessary data. We always have to make the raw data take form, from which we can help our problem statement to be solved. In my case also raw data obtained was very huge and complex, but as a student and researcher in field of machine learning and data mining, I learnt the data cleansing process and applied efficiently on that raw data supporting my practical and problem solving.

2.2.2. Unsupervised Machine Learning

Unsupervised learning, as the name suggests is related to machine learning by itself with some algorithms without interference of the user giving the input of data. Most popular unsupervised machine learning is Clustering.

Clustering is the technique in which the distance formula plays the main role in deciding the mean of the points falling under one region. Different regions formation takes place with time when the points come in contact with each other in reference to the distance calculated.

2.2.3. Ensemble Learning

Ensemble modelling is the combination of two or more machine learning and data mining algorithms which work as combined unit to give out the best result possible from either algorithm. This technique is widely used these days in research in different fields.

Usually this technique increases the overall efficiency of the results because of the unity factor of the two models coming in to play. But this technique might also decrease the efficiency in cases where the data is vast and one model is better than other with great margin.

This technique also plays an important role in increasing the area of datasets for one topic of research. Wide variety of data sets might be thrown at the model and this will still give a decent result where one single model might not perform well at all.

This technique has been used in our implementation and this will be discussed in detail in chapter 5.

2.3 Data Mining

Data mining and Machine learning are very closely related terms. Data mining algorithms are the techniques that are used to get the results out of dataset provided to the machine. Machine learning is the technique which is used as a superset while applying data mining algorithms. R programming has a list of around 200 data mining algorithms available in library which can be implemented on the dataset. Though with every data mining algorithm, machine learning category is provided. And based on that knowledge different data mining algorithms are applied on the data set.

In our work, there is usage of 6 data mining algorithms which have taken place and they are as follows:

1. Decision Tree
2. Ada Boost
3. Random Forest
4. Support Vector Machine
5. Linear Model
6. Neural Network

2.3.1 Random Forest Algorithm

Random forests are the ensemble learning method in itself majorly used in regression and classification. When the training of the machine is done, it generates the multitude of the decision trees. Classification is finding the mode of classes and Regression is the mean prediction of individual trees. Random forests corrects the over-fitting habit of the decision trees to their train set.

Random forest shows us the variable importance too, which tells the contributing percentage of the attribute to the final prediction. So, because of this characteristic of random forest model, it becomes very easy for the feature selection. Random forest technique uses multiple decision trees for the improvement of classification rate.

2.3.2 Decision Tree Algorithm

Generally, a decision tree is a decision support tool using tree like graph of decisions and the possible outcomes. And, decision tree learning uses the above explained decision tree as the predictive model, which uses the observations as the attributes for predicting the target value. The tree models, those can take only a finite set of values are called classification trees. And, where the tree variable can take value that is continuous in nature i.e. the real numbers, are called the regression trees.

In the following figure, Report Submission Example has been taken to understand the decision tree. In this instance, if report is not submitted on time then the result will be that person will be failing. And if the person is submitting on time, further classification is done on the basis of plagiarism for grading. Following is the figure 1.

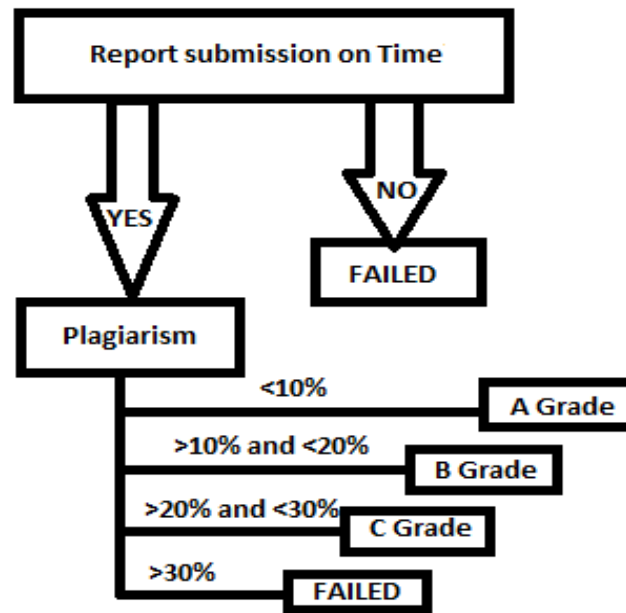


Figure 1. Decision Tree

2.3.3 ADA Boost

In classification methods, boosting plays a high significance role. Boosting signifies application of the classification algorithm in the gradual manner, by changing the weightage of partition of training data in comparison to testing dataset, and then having classifiers taken in those weighted majority. This boosting technique is used in combination with many classification algorithms for improving the model's performance.

This boosting technique can be very easily explained in terms of statistics as additive modelling.

2.3.4 Support Vector Machine [SVM]

SVM or Support Vector Machine is the supervised learning model which is widely used in pattern recognitions, used in regression as well as classification modelling.

Suppose there are some examples for training of the model, SVM will try to classify it accordingly into as much as high precision of classification.

SVM is of two types:

- Linear SVM Classifier.
- Kernel Based SVM Classifier. [Non - Linear]

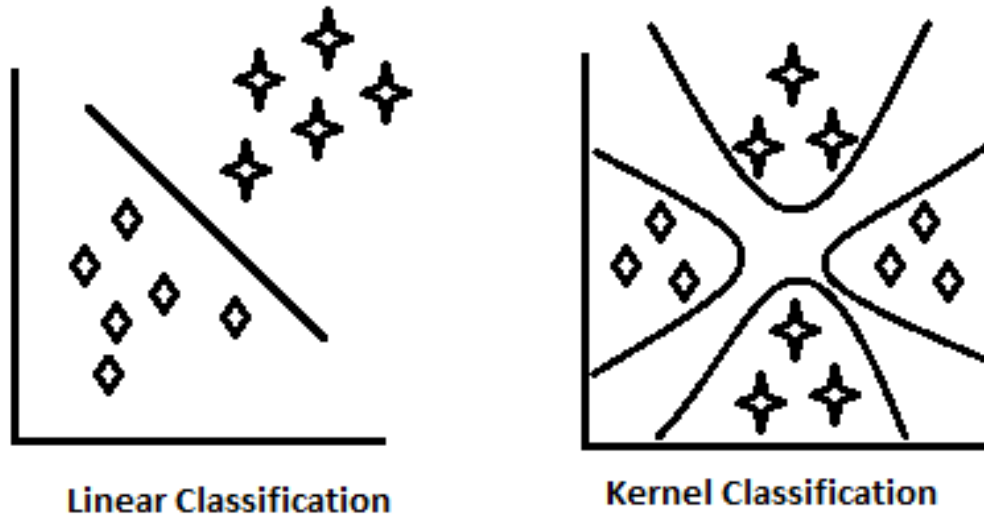


Figure 2. SVM – Linear and Kernel Classification

2.3.5 Linear Model

Linear Model is used here in a very close context to linear regression model. Linear regression can be explained as the modelling of relationship of scalar dependent variable and independent variable. Where there is only one independent variable, that case is called the simple linear regression. And for more than one independent variable, it is called multivariate linear regression.

So, if we analyze our dataset in which employability is predicted based on the response of an individual for the 50 questions of the Big five Personality questionnaire, all 50 responses act as independent attributes for one prediction variable i.e. target value having 0 or 1 value for non-employable and employable respectively. Hence, Target value is dependent variable on 50 responses.

2.3.6 Neural Network Model

This model is based on neurons working in brain. Signal is passed from one neuron to another till the output is given by body. Same logic is applied in neural networking in machine learning. Different weights are assigned to the inputs that are given to the machine and working like neurons the inputs make sense with each other and output is given.

This is used in supervised learning, unsupervised learning and reinforcement learning. These days DNA reprogramming is also being done in the latest research. It is based on neural networking only.

The main problem statement is to predict whether the person is employable or not on the basis of a questionnaire developed in corresponding to psychological characteristics.

But, there are many sub-objectives of this research such as:

1. To understand the practicality of Machine Learning and Data Mining Techniques.
2. To study details of data mining techniques.
3. Realizing the value of data cleansing before final data feeding to machine.
4. Efficient use of R programming and Rattle tool.
5. Advantages of Rattle over traditional R Studio.
6. Doing detailed analysis through graphical representation.
7. To check the consistency of the models being tested by changing the seed value, hence changing the training and testing dataset which is being randomly selected while the machine learning and analysis is done by the models. This procedure is known as K-Fold Validation, where K is the number of tests run by changing the seed value every time.
8. To show the significance of ensemble modelling.
9. To be able to see the detailed analysis of combinations of different ensemble mixtures by permutations of different data mining models.
10. How MS-Excel can be a handy-tool for doing the Data Science Business

3.1 Understanding practicality of Machine Learning and Data Mining Techniques

Data mining plays a great role in finding the relation and statistical inferences from the data provided in the dataset. In our work, we have analysed the data using different data mining techniques in excel [3]. Machine learning models have also been tested for the different values required in classification analysis, such as Precision, Confusion Matrix, ROC value, Accuracy of models, RMSE value [7]. Different data mining methods such as

Random Forest [11], SVM [10, 11], Linear Model [11], Neural Network [11], Ada Boost [12] and Decision Tree [11] has been referred for the purpose of comparison of their results and their implementation in R.

3.2 To study details of data mining techniques

This has been studied in section 2.3 of chapter 2.

3.3 Realizing the value of data cleansing

Data cleansing is a term related to getting the valuable information from the raw data and removing the noisy data (data not valuable to us). This can be done easily in Microsoft Excel and is widely useful technique for every data scientist.

3.4 Efficient use of R programming and Rattle tool

R Programming is the standard statistical programming which is used in machine learning field through-out the world. This has its own advantages over other languages and Python language is the closest to what R is capable of. R has a library defined of list of data mining algorithms available for machine learning at goo.gl/3DWT2s. There are over 200 models available for different kind of data sets, considering regression, classification, multivariate, clustering etc. kinds of analyses required.

Rattle is a tool just like weka, but Rattle being more powerful as it is completely built through R programming only. Rattle is very helpful for beginners in field of machine learning. Graph charts can be built in it. Procedure for machine learning and data mining is also very easy through rattle. Graphical user interface is always easy to access and easily understandable to the beginners. But this tool has its own limitations, not more than 4-5 models can work together by relying only on this tool. Maximum advantage of R programming can only be taken for accessing up to 15-30 models through R programming using the libraries defined.

3.5 Advantages of Rattle over traditional R Studio

Rattle has following advantages:

1. Its graphical user interface is better than that of R Studio.
2. It's easier for beginners to work on Rattle than starting directly on R Studio, because its prerequisite is to at least know one programming language, preferably python.

3.6 Detailed analysis through graphical representation

Graphical representations are much more explanatory than theoretical of tabular results. So, time and again in this report, graphs will be plotted for more easy representation of whatever field result is to be displayed.

3.7 K-Fold Validation

K – Fold Validation is the most important part of overall analysis. Explaining it in very crisp form, this is the concept of checking the consistency of the model when predicting the target value with the use of machine learning algorithm. In our case, we have done 10-fold validation, meaning, dataset and different data mining and machine learning models will be checked with different seed value for 10 different times, ensuring every time different training and testing data set portions are picked and then checking the accuracy for consistency. If the accuracy is consistent and results are acceptable, then we conclude that the model is indeed a good model for predicting the target value.

3.8 Significance of Ensemble Modelling

Ensemble modelling is the most crucial part of our research. Combining two or more machine learning models and then comparing those combinations to find the best possible combination which will be recommended as the best ensemble possible for problem to be solved.

3.9 Analysis of Ensemble Models

Detailed analysis of ensemble models will be done in the Implementation and Results chapter, section 5.7. Combinations of top 5 machine learning algorithms will be checked for which one is the best.

3.10 Value of MS-Excel for Data Scientist

The Machine Learning models always provide the values as results developed by different algorithms based on their technical working and the training-testing data set provided to them. It's already discussed earlier that data cleaning is a very essential part of the machine learning experience. There are more than one way to clean data for final dataset to be ready to be fed to machine for the prediction. Fastest way to edit data by professionals is by using python scripting or in Linux kernel by applying AWK statements. But, for convenience work is done in excel. Microsoft Excel is widely used by data scientists for cleaning the data.

How about we take an illustration of sentiment analysis raw data. In the event that we gather tweets from twitter site for say, cricket world cup. At that point the crude information when removed by python scripting would comprise of distinctive tweets, spot, time, creator and so forth with the content proclamation that we oblige that is the 140 character tweet for opinion characterization as positive, negative or unbiased tweet. Sentiment Analysis is a splendid field to do research on. Assessments characterization is done generally nowadays to know the reaction or input of individuals in today's field of long range interpersonal communication and e-trade business. We should return to our subject.

Exceed expectations can be utilized to straightforwardly take the segments needed for our machine learning knowledge. Find and Replace choice can be utilized helpful while making the information the way we need, as needed in tweets segmented. When we get the crude information of tweet, it is in twofold quotes, it can be effectively evacuated utilizing discover and supplant.

Subsequently, Microsoft Excel is useful for the information researcher for information purging and preparing the information comes about that are given by the machine learning calculations.

Tools used for implementation of the problem solution are as follows:

- ✓ R Studio: Version 0.98.1102 - © 2009-2014 RStudio, Inc.
- ✓ Microsoft Excel 2013

Methodology used can be seen in the following figure:

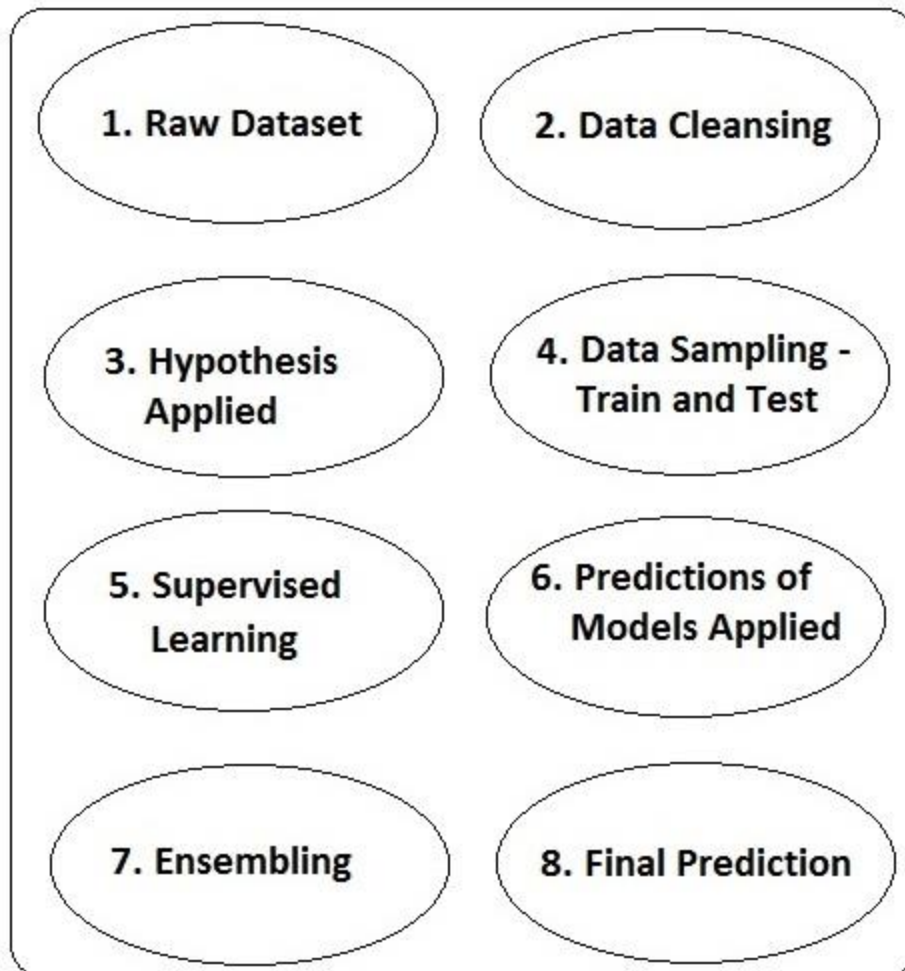


Figure 3. Methodology

Implementation technique and Results can be illustrated in following flow of data:

1. Raw Data Collection
2. Data Cleansing
3. Hypothesis Generation
4. Data Sampling: Training and Testing
5. Supervised Learning and K-Fold Validation
6. Predictions of Models Applied
7. Ensembling
8. Result: Final Prediction

5.1 Raw Data Collection

The dataset that has been used in the analysis is secondary data and was collected through an interactive online personality test [13]. Participants deliberately decided to get their information stored for edification and research purposes, those who did not allow, their information was not included in dataset. Participants came to test through different sources including another page on test website, Facebook, Google, from URLs with edu in their domain (e.g. xxx.edu.au) etc and some other sources. Total 50 questions (rating on scale of 1-5) were answered by every candidate, so there are 50 features in the data set helping in predicting the 51st feature which is binary value of the classifier indicating employability. Moreover, race (13 type of races such as mixed race, arctic, Caucasian etc.), age (range 13-80), engnat (English native language or not), hand (left-handed, right-handed or both), gender (male, female or other) and country was also recorded in raw dataset.

5.2 Data Cleansing

Data cleansing is spinal cord of data science. A true data scientist always finds something out of the noisy data by the art of data cleansing. There are many techniques of doing data cleansing from which we opted to clean the noise out of raw data using Microsoft Excel.

Our data was in csv (comma separated values) format. So, there was a need of shunting out of the unwanted columns out of the raw data. Below is the screenshot table of our raw data:

Table 2. Raw Data

1	race	age	engnat	gender	hand	source	country	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10
2	3	53	1	1	1	1	US	4	2	5	2	5	1	4	3	5	1
3	13	46	1	2	1	1	US	2	2	3	3	3	3	1	5	1	5
4	1	14	2	2	1	1	PK	5	1	1	4	5	1	1	5	5	1
5	3	19	2	2	1	1	RO	2	5	2	4	3	4	3	4	4	5
6	11	25	2	2	1	2	US	3	1	3	3	3	1	3	1	3	5
7	13	31	1	2	1	2	US	1	5	2	4	1	3	2	4	1	5
8	5	20	1	2	1	5	US	5	1	5	1	5	1	5	4	4	1
9	4	23	2	1	1	2	IN	4	3	5	3	5	1	4	3	4	3
10	5	39	1	2	3	4	US	3	1	5	1	5	1	5	2	5	3
11	3	18	1	2	1	5	US	1	4	2	5	2	4	1	4	1	5
12	3	17	2	2	1	1	IT	1	5	2	5	1	4	1	4	1	5
13	13	15	2	1	1	1	IN	3	3	5	3	3	3	2	4	3	3
14	13	22	1	2	1	2	US	3	3	4	2	4	2	2	3	4	3
15	3	21	1	2	1	5	US	1	3	2	5	1	1	1	5	1	5

There was a need of selecting only the important columns who would contribute positively in our research. So, according to the arisen need, first 7 columns were shunted out of our raw data in process of data cleansing and the data on which hypothesis was applied looked like in the below table:

Table 3: Raw Data after Cleansing

1	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	A1	A2	A3	A4	A5	A6
2	4	2	5	2	5	1	4	3	5	1	1	5	2	5	1	1	1	1	1	1	1	5	1	5	2	3
3	2	2	3	3	3	3	1	5	1	5	2	3	4	2	3	4	3	2	2	4	1	3	3	4	4	4
4	5	1	1	4	5	1	1	5	5	1	5	1	5	5	5	5	5	5	5	5	5	1	5	5	1	5
5	2	5	2	4	3	4	3	4	4	5	5	4	4	2	4	5	5	5	4	5	2	5	4	4	3	5
6	3	1	3	3	3	1	3	1	3	5	3	3	3	4	3	3	3	3	3	4	5	5	3	5	1	5
7	1	5	2	4	1	3	2	4	1	5	1	5	4	5	1	4	4	1	5	2	2	2	3	4	3	4
8	5	1	5	1	5	1	5	4	4	1	2	4	2	4	2	2	3	2	2	2	5	5	1	5	1	5
9	4	3	5	3	5	1	4	3	4	3	1	4	4	4	1	1	1	1	1	1	2	5	1	4	3	3
10	3	1	5	1	5	1	5	2	5	3	2	4	5	3	3	5	5	4	3	3	1	5	1	5	1	5
11	1	4	2	5	2	4	1	4	1	5	5	2	5	2	3	4	3	2	3	4	2	3	1	4	2	4
12	1	5	2	5	1	4	1	4	1	5	5	3	5	3	2	5	3	3	4	3	2	4	2	4	1	5
13	3	3	5	3	3	3	2	4	3	3	1	5	3	3	2	3	2	3	2	4	4	4	2	2	5	5
14	3	3	4	2	4	2	2	3	4	3	3	3	3	3	2	2	4	4	2	3	1	4	1	5	1	5
15	1	3	2	5	1	1	1	5	1	5	5	3	5	2	5	5	3	2	5	3	1	1	1	4	2	3

As we can see the above screenshot which is showing columns A to T representing ratings of 23 users for 10 questions each for Extrovertness and Neuroticism on scale of 1-5. Similarly, 19719 cases of 50 questions rating for O.C.E.A.N Personality.

5.3 Hypothesis Applied

For our purpose, we have taken only the rating of all the questionnaire as the attributes as the input and the hypothesis case has been generated by taking care of the qualities required by the employee [14]. There are certain qualities that are directly linked to the personality traits. Hence the hypothesis in the Table II – Hypothesis for deciding employability has been generated including the neutral emotions.

Table 4. Hypothesis for Deciding Employability

Big Five Characteristic	Range of values acceptable (on scale of 5)	Number of people in acceptable range (out of 19719)
Openness [O]	$2 < O \leq 5$	19681
Conscientiousness [C]	$2 < C \leq 5$	19644
Extrovertness [E]	$2 < E \leq 5$	19658
Agreeableness [A]	$2 < A \leq 5$	19675
Neuroticism [N]	$0 < N \leq 3$	9385

After making the data the way we needed for our experiment to go on and applying the hypothesis on it, we got something like in below table:

Table 5: Employability calculation using Hypothesis

1	Openness	Conscientiousness	Extrovertness	Agreeableness	Neuroticism	Employable
2	3.1	3.1	3.2	3.2	1.9	1
3	2.6	2.8	2.8	3.1	2.9	1
4	4.1	3.3	2.9	3.8	4.6	0
5	3.7	3.4	3.6	3.7	4.3	0
6	2.2	2.6	2.6	4	3.2	0
7	3.5	3.7	2.8	3.4	3.2	0
8	2.9	3	3.2	3.7	2.5	1
9	3	3	3.5	3.1	1.9	1
10	3.7	3.5	3.1	3.3	3.7	0
11	3.3	3.2	2.9	2.7	3.3	0
12	3.8	3.2	2.9	2.9	3.6	0
13	3.4	3.4	3.2	3.7	2.8	1
14	3.1	3.3	3	3.2	2.9	1
15	3.1	3.3	2.5	2.7	3.8	0
16	3.2	3.2	3.2	3	3.1	0
17	3.5	3	2.6	2.9	2.3	1
18	3.5	3.2	2.8	2.5	2.3	1
19	3.1	2.8	3.2	3.4	3.2	0
20	2.9	3.2	2.8	2.6	2.6	1
21	3.2	2.3	2.6	2.9	2.9	1
22	3.5	3	3.1	3.4	4	0
23	3.3	3.1	2.8	2.6	1.8	1

5.4 Data Sampling: Training and Testing

This is the part of implementation where we choose to do the partitioning of our data in to training and testing data. In our experiment, [70, 30] was the partition we used for our practical purposes because that's the standard partitioning ratio.

[70, 30] means 70% of the data set is dedicated to training and 30% of data is dedicated to testing the algorithms if they predict the data being tested as accurately as possible, compared to their original number. This prediction is done on the basis of training data we have fed to the machine algorithms.

For instance 17919 observations are being observed in our dataset. And, this means on the basis of [70, 30] partition, 12543 observations (70% training set) are being fed to the

machine for the learning of the pattern of employability and non-employability we have earlier studied through hypothesis generated values. Now, those 12543 observations will help in the prediction of the rest of the 5376 (30% testing set), using the 6 models we have implemented in R. And that output sheet will be saved as csv file too for the verification of the accuracy that we opted to find in MS-Excel again, for the simplicity purposes. Part of output can be seen in below table:

Table 6. Prediction results in Testing File

1	Employable	rpart	ada	rf	ksvm	glm	nnet
2	1	1	1	1	1	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1
6	1	1	0	1	1	0	0
7	0	0	0	0	0	0	0
8	1	1	1	1	1	1	1
9	0	0	0	0	0	0	0
10	1	1	1	1	1	0	1
11	0	1	0	0	0	0	0
12	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0
14	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0
17	1	1	1	1	1	1	1
18	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0
20	1	1	1	1	1	1	1
21	0	0	0	0	0	0	0
22	1	1	1	1	1	1	1
23	1	0	1	1	1	1	1

Results for seed value 42 and training-testing sampling can be seen in below table:

Table 7: Results of Classification Models-seed 42

Model Name	Method, Package	Accuracy	Confusion Matrix	ROC	Precision/ Recall
Decision Tree	rpart, rpart	0.883874	0.116	0.9325	0.950
Random Forest	rf, randomForest	0.941346	0.058	0.9881	0.991
Ada Boost	ada, ada	0.953685	0.046	0.9914	0.994
Support Vector Machine	svm, ksvm	0.985970	0.014	0.9974	0.998
Linear Model	lm, glm	0.986139	0.013	0.9957	0.997
Neural Network	neuralnet, neuralnet	0.943374	0.0566	0.9882	0.991

There are 6 columns in the table. All of these columns have been very carefully selected for the purpose of comparing classification models. For e.g. If the values of true positive and true negative are to be compared then in the regression model, it is known as Error Matrix but in the classification model, same thing is referred to as Confusion Matrix.

First column in table 7 tells us the name of the model of which different of which different characteristics and results are been compared for the classification modelling. Second column describes the Method and Package name used in the R scripting for the desired models to be used. Third column describes the accuracy of the models at this seed value, more the accuracy, better the model. Fourth column belongs to Confusion Matrix value, this tells us the total error of values that did not match when matrix of original values versus predicted values was made. Fifth column ROC value, tells us ratio of true positive rate versus false positive rate [15]. Column six signifies Precision and Recall value, precision is the value for ratio of number of true positives to the total number of positives predicted.

And Recall value is ratio of how much actual positives a model can identify to the total number of positives. Here, both are equal. Also, from meaning of both precision and recall, both are one and same thing.

5.5 Supervised Learning and K-Fold Validation

5.5.1 Supervised Learning

Supervised learning term has been used because data has been fed to machine through us. And algorithms are being applied too, saving the results in the tabular form.

Finding the accuracy has been the main motive for selecting the models that could be further used in ensemble modelling. But, we can't be sure if the model is consistent or not in giving the results and that accuracy being shown in one practical is apt reason to select it for ensemble modelling.

Therefore, need arose for K-Fold Validation Testing.

5.5.2 K-Fold Validation

K-Fold Validation is the most important step for verifying if the results given by the models are consistent and that is different sample of data is picked up randomly in the data set, accuracy won't be affected much.

Procedure for K-Fold Validation was simple enough:

1. Set the random seed value.
2. Set the partition as [70, 30], as kept in the first time practical.
3. Run the algorithms and generating the test result file.
4. Save the result file as csv file.
5. Find the accuracy generated by different machine learning models in R.
6. Repeat steps 1 to 5 for 10 times.

Finding and storing the results in tabular form for every training-testing sampling with 10 different seed values is the procedure required for 10-fold validation testing. We have taken K=10 for our validation testing.

In following 10 tables, we elaborately show the results that we have got during the K-Fold validation testing. After those, there is a proof with graph for showing the consistency for top 5 models.

Table 8. K-Fold Validation - Seed 42

Model Name	Model Accuracy	Confusion Matrix Value	ROC Value	Predicted vs Observed	Recall
Decision Tree	88.3874	0.116	0.9325	0.6208	0.950
Random Forest	94.1346	0.058	0.9881	0.7954	0.991
Ada Boost	95.3685	0.046	0.9914	0.8571	0.994
Support Vector Machine	98.597	0.014	0.9974	0.9455	0.998
Linear Model	98.6139	0.013	0.9957	0.9317	0.997
Neural Network	94.3374	0.0566	0.9882	0.8291	0.991

Table 9. K-Fold Validation - Seed 857279

Model Name	Model Accuracy	Confusion Matrix Value	ROC Value	Predicted vs Observed	Recall
Decision Tree	88.33277	0.1166723	0.9340	0.6218	0.951
Random Forest	94.55529	0.0541089	0.9897	0.8022	0.992

Ada Boost	95.70511	0.0429489	0.9935	0.8692	0.995
Support Vector Machine	98.85019	0.0114981	0.9983	0.9544	0.999
Linear Model	98.6811	0.0131890	0.9964	0.9330	0.997
Neural Network	94.42002	0.0557998	0.9886	0.8308	0.992

Table 10. K-Fold Validation - Seed 429822

Model Name	Model Accuracy	Confusion Matrix Value	ROC Value	Predicted vs Observed	Recall
Decision Tree	88.10007	0.1189993	0.9279	0.6097	0.947
Random Forest	93.91481	0.0608519	0.9874	0.7922	0.991
Ada Boost	94.9121	0.0508789	0.9899	0.8475	0.993
Support Vector Machine	98.24206	0.0175794	0.9966	0.9380	0.997
Linear Model	98.79986	0.0120013	0.9942	0.9299	0.996
Neural Network	95.63895	0.0436105	0.9923	0.8716	0.994

Table 11. K-Fold Validation - Seed 620316

Model Name	Model Accuracy	Confusion Matrix Value	ROC Value	Predicted vs Observed	Recall
Decision Tree	88.43813	0.1156187	0.9308	0.6213	0.949
Random Forest	93.79648	0.0628803	0.9874	0.7966	0.991

Ada Boost	95.09804	0.0490196	0.9910	0.8558	0.993
Support Vector Machine	98.25896	0.0174104	0.9969	0.9425	0.998
Linear Model	98.61393	0.0138607	0.9950	0.9289	0.996
Neural Network	95.38540	0.0461460	0.9912	0.8629	0.994

Table 12. K-Fold Validation - Seed 373955

Model Name	Model Accuracy	Confusion Matrix Value	ROC Value	Predicted vs Observed	Recall
Decision Tree	87.89723	0.1210277	0.9265	0.6036	0.946
Random Forest	93.83029	0.0615280	0.9880	0.7941	0.991
Ada Boost	95.33469	0.0466531	0.9914	0.8560	0.994
Support Vector Machine	98.47870	0.0152129	0.9974	0.9413	0.998
Linear Model	98.27586	0.0172413	0.9958	0.9268	0.997
Neural Network	94.67546	0.0532454	0.9893	0.8444	0.992

Table 13. K-Fold Validation - Seed 798731

Model Name	Model Accuracy	Confusion Matrix Value	ROC Value	Predicted vs Observed	Recall
Decision Tree	88.10007	0.1189993	0.9295	0.6119	0.948
Random Forest	93.69506	0.0632183	0.9876	0.7928	0.991
Ada Boost	95.08114	0.0491886	0.9914	0.8541	0.994

Support Vector Machine	98.54632	0.0145368	0.9974	0.9412	0.998
Linear Model	98.71535	0.0128465	0.9953	0.9316	0.997
Neural Network	94.35429	0.0564570	0.9884	0.8382	0.991

Table 14. K-Fold Validation - Seed 754410

Model Name	Model Accuracy	Confusion Matrix Value	ROC Value	Predicted vs Observed	Recall
Decision Tree	88.67478	0.1132522	0.9312	0.6235	0.949
Random Forest	93.64435	0.0633874	0.9867	0.7896	0.990
Ada Boost	94.91210	0.0508789	0.9900	0.8471	0.993
Support Vector Machine	98.10683	0.0189317	0.9963	0.9347	0.997
Linear Model	98.52941	0.0147058	0.9951	0.9298	0.996
Neural Network	93.69506	0.0630493	0.9868	0.8191	0.990

Table 15. K-Fold Validation - Seed 1111

Model Name	Model Accuracy	Confusion Matrix Value	ROC Value	Predicted vs Observed	Recall
Decision Tree	87.81271	0.1218729	0.9156	0.5967	0.938
Random Forest	94.11765	0.0588235	0.9882	0.7972	0.991
Ada Boost	95.52062	0.0447937	0.9913	0.8582	0.994

Support Vector Machine	98.27586	0.0172413	0.9967	0.9401	0.998
Linear Model	98.91819	0.0108181	0.9945	0.9367	0.996
Neural Network	94.97972	0.0502028	0.9898	0.8494	0.993

Table 16. K-Fold Validation - Seed 645946

Model Name	Model Accuracy	Confusion Matrix Value	ROC Value	Predicted vs Observed	Recall
Decision Tree	87.96484	0.1203516	0.9274	0.6093	0.946
Random Forest	94.01623	0.0596687	0.9881	0.7944	0.991
Ada Boost	95.55443	0.0444557	0.9921	0.8611	0.994
Support Vector Machine	98.52941	0.0147058	0.9973	0.9410	0.998
Linear Model	98.20825	0.0179175	0.9958	0.9283	0.997
Neural Network	94.72617	0.0527383	0.9910	0.8457	0.993

Table 17. K-Fold Validation - Seed 549403

Model Name	Model Accuracy	Confusion Matrix Value	ROC Value	Predicted vs Observed	Recall
Decision Tree	87.77890	0.122211	0.9241	0.5972	0.944
Random Forest	93.47532	0.065077	0.9869	0.7881	0.990
Ada Boost	95.11494	0.048850	0.9903	0.8493	0.993

Support Vector Machine	98.32657	0.016734	0.9971	0.9407	0.998
Linear Model	98.81677	0.011832	0.9950	0.9326	0.996
Neural Network	95.16565	0.048343	0.9911	0.8589	0.993

Now we plot the cross-validation graph based on the accuracies of best 5 models:

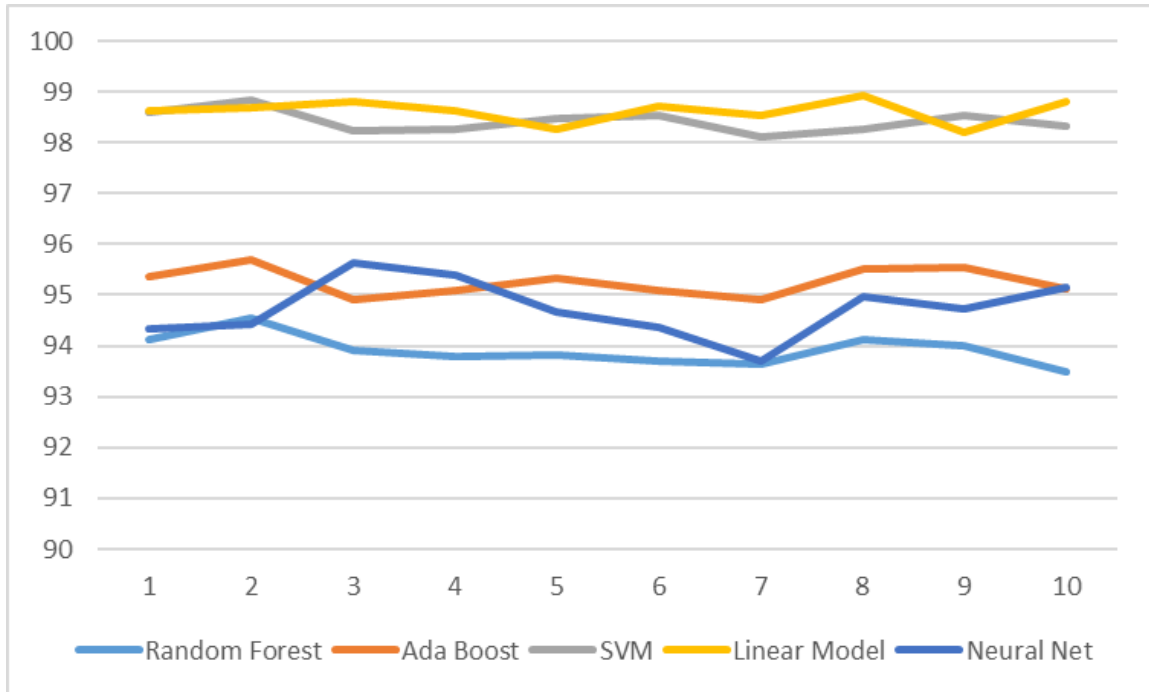


Figure 4. K-Fold Graph

The purpose behind such a splendid precision of these models can be clarified effortlessly. At the point when model is prepared with vast number of cases and related results, precision shoots up. Here, there are 19719 perceptions. In this way, when we do training testing in 70-30 model, we are preparing our models with 13803 cases, which is really high number of cases to prepare a model as the expectation is paired in nature i.e. employable or not employable. For multivalued results, this high precision would not so much come. Additionally, the procedure utilized for discovering the estimations of O.C.E.A.N qualities was straight forward in light of inputs in scope of 1-5 as appraisals. Along these lines, above clarifies high precision. In addition, this can be said that when we diminish the training/testing proportion, precision will diminish. Along these lines, precision can be put

in expanding request of [Training, Testing] proportion blends as [40, 60] < [50, 50] < [60, 40] < [70, 30] etc.

5.6 Predictions of Models Applied

Till now, we know that our models are consistent in their performance. Now, the next step was allowing these models to be tested in ensemble learning. So, these predictions were applied for ensemble modelling.

5.7 Ensembling OR Ensemble Modelling

As we can see the performance of all the five models is very consistent, so the next step is finding the best ensemble combination out of these five models. It is important to compare all the possible combinations for final ensemble model because we don't know which combination might come out as the best, serving the greater range of data sets.

Total of 20 combinations were designed to be checked for their accuracy at the values we achieved at seed 42 with 70-30 model of training-testing. And the table for accuracy comparison is as follows:

Table 18. Comparison of Ensemble Models

Sr. No.	Possible Ensembles	Accuracy
1.	Random Forest + Ada Boost	94.99662000
2.	Random Forest + Support Vector Machine	96.90669371
3.	Random Forest + Linear Model	96.97430696
4.	Random Forest + Neural Network Model	95.26707235
5.	Ada Boost + Support Vector Machine	97.58282623
6.	Ada Boost + Linear Model	97.70114943
7.	Ada Boost + Neural Network Model	96.04462475
8.	Support Vector Machine + Linear Model	98.79986477
9.	Support Vector Machine + Neural Network Model	97.46450304
10.	Linear Model + Neural Network Model	97.48140636
11.	Random Forest + Ada Boost + Support Vector Machine	95.90939824

12.	Random Forest + Support Vector Machine + Linear Model	98.56321839
13.	Random Forest + Linear Model + Neural Network	97.49830967
14.	Ada Boost + Support Vector Machine + Linear Model	98.54631508
15.	Ada Boost + Support Vector Machine + Neural Network	97.71805274
16.	Ada Boost + Linear Model + Neural Network	97.71805274
17.	Support Vector Machine + Linear Model + Neural Network	98.51250845
18.	Random Forest + Ada Boost + Support Vector Machine + Linear Model	97.65043949
19.	Random Forest + Support Vector Machine + Linear Model + Neural Network	98.22515213
20.	Random Forest + Ada Boost + Support Vector Machine + Linear Model + Neural Network	97.81947262

From table 18, blend of Support Vector Machine and Linear Model (at number 8) is ended up being the best conceivable gathering with a precision of 98.7998%. Thus, for discovering the employability of a man taking into account Big five Personality Test, troupe model of Support Vector Machine and Linear Model together ends up being the best amid usage.

5.7.1 How to do ensemble modelling using Microsoft Excel?

Suppose we have the value of prediction of 2 machine learning models names A and B as 0 and 1 respectively, then to find the result for ensemble i.e. combining the result of the two models need to be combined using union mathematical concept.

Formula for Ensembling n number of models accuracies in excel is:

Ensemble combination value = **UNION (Model A, Model B,... , Model N)**

Above formula can be made to work in different forms.

Below is the screenshot of ensemble modelling done in MS-Excel:

Table 19. Ensemble Generation in Excel

1							1st	1st Accuracy	2nd	2nd Accuracy	3rd	3rd Accuracy
2	Employable	rf	ada	ksvm	glm	nnet	rf+ada	94.9966193	rf+svm	96.90669371	rf+lm	96.97430696
3		1	1	1	0	0	0	1	1	1	1	1
4		0	0	0	0	0	0	0	1	0	1	0
5		0	0	0	0	0	0	0	1	0	1	0
6		1	1	1	1	1	1	1	1	1	1	1
7		1	1	0	1	0	0	1	1	1	1	1
8		0	0	0	0	0	0	0	1	0	1	0
9		1	1	1	1	1	1	1	1	1	1	1
10		0	0	0	0	0	0	0	1	0	1	0
11		1	1	1	1	0	1	1	1	1	1	1
12		0	0	0	0	0	0	0	1	0	1	0
13		0	0	0	0	0	0	0	1	0	1	0
14		0	0	0	0	0	0	0	1	0	1	0
15		1	1	1	1	1	1	1	1	1	1	1

5.8 Result: Final Prediction

Combination of SVM and Linear model is best for this kind of application where based on machine learning and psychological data of big five personality or data like these ratings of 0-5, prediction of employability is needed to be done.

This report goes into point by point usage and the systems needed for them. This can be extremely useful for the intrigued people for figuring out how to do profound investigation and learn different strategies in data science. Case in point, how to bode well for us has been examined. How information purifying is an imperative piece of usage of machine learning models has likewise been talked about.

The consequences of our usage have demonstrated that troupe model of SVM and Linear Model are as one the best blend to get the coveted result productively. Organizations can utilize this model to anticipate the individual's capacity to perform well at their association in light of the fact that identity is the most critical element that assumes an imperative part to choose whether the execution of the representative will be at standard with the necessities of the firm or not. In future, examination should be possible by diving more profound into this theme by discovering what sort of occupation is suitable for distinctive scope of identities utilizing multivalued arrangement of models.

Video Presentation

Link to Video Presentation:

<https://www.youtube.com/watch?v=9fhZt6PND9c>

References

- [1] Goldberg Lewis R., “The development of markers for the Big-Five factor Structure.”, *Psychological Assessment*, Vol. 4, No. 1, Pages 26-42, 1992
- [2] Quercia D., Kosinski M., Stillwell D., Crowcroft J., “Our Twitter Profiles, Our Selves: Predicting Personality with Twitter”, *IEEE International Conference on Privacy, Security, Risk and Trust, and IEEE International conference on Social Computing*, Pages 180-185, 2011.
- [3] Tang H., “A Simple Approach of Data Mining in Excel”, *4th International conference on Wireless Communications, Network and Mobile Computing*, Pages 1-4, 12-14 Oct, 2008.
- [4] E. Alpaydin, “Introduction to Machine Learning”, The MIT Press, February 2010.
- [5] R. Polikar, “Ensemble Based Systems in Decision Making”, *IEEE Circuits and systems Magazine*, Third Quarter, 2006.
- [6] Lei Xu, Adam K, Ching Y. Suen, “Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition”, *IEEE Transactions on Systems*, Vol. 22, No. 3, May/June 1992.
- [7] Rana P.S, Sharma H, Bhattacharya M, Shukla A, “Quality assessment of modelled protein structure using physicochemical properties”, *Journal of Bioinformatics and Computational Biology*, Imperial College Press, 2014.
- [8] Farnadi G., Sitaraman G., Rohani M., Kosinski M., Stillwell D., Moens M.F., Davalos S., Cock M.D., “How are you doing? Emotions and Personality in Facebook”, *EMPIRE Workshop of the 22nd International Conference on User Modeling, Adaptation and Personalization (UMAP 2014)*.
- [9] Quercia D., Lambiotte R., Stillwell D., Kosinski M., Crowcroft J., “The Personality of Popular Facebook Users”, *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 2012.
- [10] Burges J.C., “A Tutorial on Support Vector Machines for Pattern Recognition”, *Bell Laboratories, Lucent Technologies, Data Mining and Knowledge Discovery 2*, Pages 121-167, 1998.
- [11] Maroco J., Silva D., Rodrigues A., Guerreiro M., Santana I., Mendona A., “Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests”, *BMC Research Notes*, 2011.
- [12] Tae-Ki An, Moon-Hyun Kim, “A New Diverse Adaboost Classifier”, *International Conference on Artificial Intelligence and Computational Intelligence*, Pages 359-363, 2010.
- [13] <http://personality-testing.info/tests/BIG5.php>
- [14] Rehman M., Mahmood A.K., Salleh R., Amin A., “Mapping Job Requirements of Software Engineers to Big Five Personality Traits”, *International Conference on Computer & Information Science (ICCIS)*, Vol. 2, Pages 1115-1122, 12-14 June, 2012.

- [15] Williams G., "Data Mining with Rattle and R – The Art of Excavating Data for Knowledge Discovery", Springer
- [16] Kaur A., Kaur K., "Performance Analysis of Ensemble Learning for Predicting Defects in Open Source Software", IEEE International Conference on Advances in Communications and Informatics (ICACCI-2014), New Delhi, Pages 219-225, 24-27 Sept, 2014.

List of Publications

[1] Sharma S., Garg D., Rana P.S, “Predicting Employability from User Personality using Ensemble Modelling”, in proceedings of IEEE, Fourth International Conference on Advances in Computing, Communications and Informatics (ICACCI-2015), SCMS, Kerala, August 10-13, 2015.

[Accepted]