BENNETT UNIVERSITY
TIMES OF INDIA GROUP

Foreign, Commonwealth & Development Office
UK Government

N·S·D·C
National Skill Development Corporation
Transforming the skill landscape

Study on ‘External factors affecting success rates in skilling programs’

April 2021

Table of Content

List of Tables

List of Figures

Introduction to Bennett University

Extending the core journalistic principles of Trust, Knowledge and Public Service, Bennett, Coleman and Co. Ltd. (BCCL) established over 178 years ago, has continually undertaken initiatives for the betterment of Indian society. The group started expanding into the education sector with the launch of TimesPro - which focuses on short-term courses that enhance employability. Bennett University, a state private university in Uttar Pradesh with the aim of providing Ivy League quality of education to undergraduate and postgraduate students making them, ‘life and career ready’. Bennett University located at Greater Noida, commenced its operations on 08 Aug 2016. The University fosters a proactive environment of Innovation and Entrepreneurship, while enhancing skills in all areas of higher education through the internationally acclaimed Centers of Excellence such as Centre for Innovation & Entrepreneurship (CIE) and Centre of Executive Education (CEE).

Vipul Kumar Mishra
Assistant Professor
Computer Science and Engineering Department
Bennett University
Email- Vipul.mishra@bennett.edu.in

Acknowledgements

This report has been prepared under the aegis of Skills for Jobs programme of the Foreign, Commonwealth & Development Office (FCDO)[1]. The report would not have been possible without the constant support, hard work and encouragement of the colleagues from Bennett University, National Skill Development Corporation (NSDC) and FCDO. We are particularly grateful to FCDO for introducing us to the varied stakeholders of the vocational education and training landscape and on-boarding us as experts for the analysis. The Skills for Jobs programme has allowed us to derive insights from the data sets which has been collected across different variables which can potentially assist in creating more informed policy decisions.

We would also like to express our gratitude towards NSDC for providing us the problem statement and the data samples to explore interlinkages of the different variables to determine the placement possibilities of the potential and trained candidates in the short-term skill development domain. NSDC also provided us inputs and feedback during the course of the project.

In addition, we convey our gratitude to all those who have in some way or the other, contributed towards the successful completion of this report.

[1] The Foreign & Commonwealth Office (FCO) and the Department for International Development (DFID) merged on 1 September 2020 to form the Foreign, Commonwealth & Development Office (FCDO)

Note to Readers and Disclaimer

1. The study was commissioned by FCDO in partnership with the Bennett University. The purpose of this study is to help NSDC and other relevant stakeholders in the skilling ecosystem to get latest relevant evidence through analytics and machine learning techniques. It is not intended to be a comprehensive summary of evidence. The contents do not constitute professional advice on behalf of the UK's FCDO or Bennett University

2. The information contained herein is of a general nature and is not intended to address the circumstances of any particular individual or entity. The report shall not be a substitute for any due diligence to be carried out by any party. No one should act on such information without appropriate professional advice after a thorough examination of the particular situation.

3. While information obtained from the public domain or external sources has not been verified for authenticity, accuracy or completeness, Bennett University have obtained information, as far as possible, from sources generally considered to be reliable. Bennett University assume no responsibility for such information.

4. In connection with the report or any part thereof, FCDO or Bennett University does not owe duty of care (whether in contract or in tort or under statute or otherwise) to any person or party to whom the Report is circulated to and FCDO or Bennett University shall not be liable to any party who uses or relies on this Report. FCDO and Bennett University thus disclaim all responsibility or liability for any costs, damages, losses, liabilities, expenses incurred by such third party arising out of or in connection with the Report or any part thereof.

5. The key variables (high correlation or high attribution variables) identified in the study are the outcome of carefully chosen and implemented AI/ML methodologies. Though they are able to explain the target variable i.e. placement (total / individual placement), they may not necessarily constitute a cause-effect relationship with the target. A detailed cause-effect map may only be derived after a careful assessment of impact of individual variables on placement. Bennett University, as well as FCDO, in no way recommends the use of any variable as a lever for the management of placement performance before proper professional due-diligence.

6. By reading/ viewing the report, the reader of the report shall be deemed to have accepted the terms mentioned hereinabove.

7. The study was completed in March 2021 and is based on the publicly available data from various ministry in India and data shared by NSDC and has not considered the developments subsequent to the completion of the study.

Glossary

AI	Artificial Intelligence
CART	Classification & Regression Trees
FCDO	Foreign, Commonwealth & Development Office
MIS	Management Information System
ML	Machine Learning
MSDE	Ministry of Skill Development and Entrepreneurship
NSDC	National Skill Development Corporation
PMKVY	Pradhan Mantri Kaushal Vikas Yojana
R&D	Research and Development
TC	Training Centre
TP	Training Partner
STT	Short Term Training
UK	United Kingdom
VET	Vocational Education and Training
ITI	Industrial Training Institute

Executive summary

With a focused skill development policy spanning across both fresh skilling and certification of prior knowledge together with a structured vocational education and skill development set up, India has made some commendable strides. The world's largest skill development scheme-Pradhan Mantri Kaushal Vikas Yojana (PMKVY), implemented under the aegis of the Ministry of Skill Development and Entrepreneurship (MSDE) and National Skill Development Corporation (NSDC), has amplified the reach and acceptance of skill development across divergent beneficiary groups. With close to five years of its operations, PMKVY has generated multiple data points about the varied stakeholders in the training value chain.

One of the key focus areas of NSDC is to foster private sector participation in the training and placements of the trained candidates. Given the importance of placements and industry tie ups in the skill development space, it is imperative to understand the linkages that placement may have with the environmental context in which training has been conducted for example local industry demand. Through the data provided by NSDC of 1.5 lakh candidates and open source data available from various ministries of India, Bennett University has conducted analysis on impact of external factor such as population, industrial development, living standard, basic education, literacy, labour force etc. on placement outcome of skilling initiatives using Artificial Intelligence (AI) and Machine Learning (ML) techniques and Statistical analysis. Moreover, Bennett University has developed predictive modelling using Artificial Intelligence (AI) and Machine Learning (ML) techniques to forecast the probability of a trainee getting placed.

The analysis on impact of external factors on placement rates in PMKVT STT has been done using two methodologies: first is to find Pearson and Spearman correlation between the placement and the external variable and second is to find the relative importance of factors using tree-based method. Person and spearman correlation of external factor with placement percent clearly indicate that *living standard and industry presence* in the district have very high correlation with placement percentage. Whereas spearman correlation with placement count indicates that if a district has more industries such as *business service companies, real state and renting* companies then total placement in that district is likely to be high. Whereas, the Pearson correlation analysis, finds that some of the attributes of a district *'No. of colleges', 'Primary with Upper Primary sec/higher sec. Enrolment in school', 'Number of ITI Trade Count', 'No of ITI', 'Population Age(15-21)', and 'Person with Under Graduate'* are most correlated with total number of placement at the district.

The predictive model, which predicts the probability of placement of a candidate at the end of training, using the tree-based methodology, seeks to understand the contribution of external factors which are influencing placements. The model achieves close to 78 percent accuracy in the placement prediction which is 3% better than that from previous study wherein external attribute such as population, industrial development, living standard, basic education, literacy, employment were not utilized (Previous study used only training center data - TC location, partner name, batch size, sector name, job role etc.). An attribute analysis of variables and their influence on placements is also calculated. The placement is not only a marker of a training centre's quality of training but also helps us understand industry demand and relevance of a job role. The analysis revealed the importance of external factors such as "Upper Primary with sec./higher sec. Enrolment in district", "Primary with Upper Primary Enrolment in district" and "Overall Literacy in district" along with internal variables such as age group, name of particular training center (TC), training partner (TP), job role details with regard to placements.

Moreover, the subjective analysis of the NSDC data sets pointed out that sectors like "Apparel", "Electronic and Hardware", and "Healthcare" are best performing sectors in terms of placement with "self-employed tailor", "retail sales associates", "documentation assistant" and "customer care executive" as the preferred job roles.

Context setting

Over the years, the technical training and vocational education landscape in India has been evolving with the interfacing of the new age technologies and skills being introduced in the ecosystem. Cusped with a demographic dividend and with a fast-paced economy, the country has a colossal responsibility of channelizing the youth potential in economically productive ways. Hence a trained workforce is a quintessential force for fuelling the industrial growth engine.

MSDE through the National Skill Development Corporation, a public-private partnership (PPP) entity, has been catalysing short-term trainings through a plethora of schemes, the largest being the Pradhan Mantri Kaushal Vikas Yojana (PMKVY). Launched in 2015 with an aim to skill close to one crore youth by 2020, the focus of PMKVY is now on Industry 4.0 and new age digital skills. Over the two phases of PMKVY, NSDC and other stakeholders of the VET ecosystem have amassed an enormous data of trainees and trainings.

With the proliferation of Big Data and other relevant technologies, a bevy of insights can be garnered through large data sets. Vast amounts of digital data have been captured through the training cycle especially for the short-term training. As on August 2020, close to 34 lakh candidates have received fresh skill trainings, 33 lakh candidates have enrolled for being certified for their prior experience and nearly 1.5 lakh candidates have been enrolled under special projects[2]. Data related to their demographics, job profiles and trainings have been recorded by the private training entities on the centralized portal namely the Skill Development Management System (SDMS) created by NSDC.

Going forward, the volume, velocity and variety of the data will continue to increase. Therefore, the endeavour now is to appropriately visualize and interpret the data for creating insights and opportunities in further embellishing the scheme and the larger VET ecosystem. One of the major goals of Big Data and related technologies is to create knowledge. The sources, methods and incumbent analysis from Big Data methods supplement and enrich established statistics. These analytics allow a detailed view at a granular level with space-related insights in real time, as well as predictive analysis.

Machine Learning and Artificial Intelligence methods are employed for undertaking an analysis of the multivariate data sets generated through the training lifecycle. This data has been leveraged to understand how the impact of various internal and external attributes are on the placement prospects of a potential trainees. Using a range of AI/ML based methodologies, a model was created for gathering insights from the data and their underlying meaning for drawing the relevant information which can potentially not only inform and strengthen the policy and budgetary prerogatives of the government but also assist in understanding the aptitude and guide the career aspirations of the VET trainees.

In first phase of study, we developed an AI model to predict the placement of the candidate. In addition, we also ranked the attributes based on the impact on placement. The study used only internal attributes such as candidate age group, education, location and training center location, batch size etc. In continuation to first phase, in this phase (second) phase, we want to analyse impact of external factors such as population of district, education status of district, industries in district and living standard of the district into the placement of a candidate.

[2] https://pmkvyofficial.org/Dashboard.aspx

Overview of problem statement

The objective of this study was to analyse whether external factors are correlated with success rates in the PMKVY 2.0 – Short Term Training (STT) program. The study aimed to identify which external variables affect placement outcomes of the STT program. External factors considered in the study are district population (age wise, gender wise), industry in district, employment of the district, education status of the district, sectoral presence, number of colleges for higher education in district, Industrial Training Institute (ITI) presence in district, Schools in the district and living standard of the district. In addition, sector-specific drivers of employment can also be expected to affect availability of work (employment / entrepreneurship) opportunities. The study aims to assess whether such factors have an impact on the employment outcomes of individuals who have undergone skill training under PMKVY 2.0-STT, and how these external factors interplay with factors intrinsic to program implementation, such as candidate profiles (gender, age, education).

To conduct this study, the data for population, living standard, industry, education, education and training, Higher education, and employment are collected from various external sources such as census, Social Economy and Cast Census (SECC), Unified District Information System for Education (U-DISE), All India Survey on Higher Education (AISHE) etc. NSDC also provided information of 1,50,000 enrolled candidates on 40+ specific attributes including age, gender, education, location, training partner, job role, placements, results etc.

Approach and detailed methodology

Given the problem statement of finding correlation of placement with the external factor such as population of district, literacy rate, industry situated in district, education status, to name a few, a three-step process was identified. A detailed overview of each of the steps have been delineated below:

1. Data Collection and Pre-processing

Data is collected across five broad categories namely – Population data, Education and training data, Industry data, Living standard data, Labour force data. Detailed data points are:

Table 1: District wise Population Data[3]

Attribute name	Description
Total population	Total population
Population Age (15-24) Absolute	Number of People between Age (15-24)
Population Age (5-14) Absolute	Number of People between Age (5-14)
Population Total male	Total male Population
Population Total female	Total female Population
Sex Ratio	Number of females in 1000 males
decadal growth rate	decadal growth rate gives an overview of the total population growth in a particular decade

Table 2: District wise primary Education and training data[4]

Attribute name	Description
Total Number of Schools	School counts in district
Primary Enrolment	Number of students enrolled in primary
Primary with Upper Primary Enrolment	Number of students enrolled in Primary with Upper Primary
Primary with Upper Primary sec/higher sec. Enrolment	Number of students enrolled in Primary with Upper Primary sec/higher sec.
Upper Primary Only Enrolment	Number of students enrolled in Upper Primary Only
Upper Primary with sec./higher sec. Enrolment	Number of students enrolled in Upper Primary with sec./higher sec

Table 3: District wise Education level (highest education)[5]

Attribute name	Description
Population Having Highest Education Level of Illiterate	These attributes describe total number of persons with highest education level.
Population Having Highest Education Level of Literate but Below Primary	
Population Having Highest Education Level of Primary	
Population Having Highest Education Level of Middle	

[3] Population data: https://censusindia.gov.in/2011-Common/CensusData2011.html Date of access- 15 February 2021
[4] School data http://udise.in/drc.htm Date of access- 18 February 2021
[5] School data http://udise.in/drc.htm Date of access- 18 February 2021; College data https://aishe.gov.in/aishe/collegeDirectoryIndex?hasReportLink=index Date of access- 20 February 2021; ITI data: https://www.ncvtmis.gov.in/Pages/ITI/Count.aspx Date of access- 21 February 2021

Attribute name	Description
Population Having Highest Education Level of Secondary	
Population Having Highest Education Level of Higher Secondary	
Population Having Highest Education Level of Undergraduate	
Population Having Highest Education Level of Postgraduate	
Population Having Highest Education Level of Diploma Certificate	
Population Having Highest Education Level of PG Diploma	
Population Having Highest Education Level of M.Phil.	
Population Having Highest Education Level of Ph.D.	
Overall Literacy	Literacy percentage in district
Female Literacy	Female Literacy percent in district
College count (Higher Education)	Total number of colleges for higher education in district
Trade Count in ITI	Total trade offered by all ITI in district
ITI count	Number of ITI in district

Table 4: District wise Industry presence (number of companies registered and active in a particular sector in a district)[6]

Attribute name	Description
Business Services	Number of companies in a particular sector in a district. For example, first attribute represents number of companies registered and active in business services category similarly other attributes also represent total company in other sectors. District information taken from registered address.
Manufacturing (Food stuffs)	
Real Estate and Renting	
Trading Community,	
personal & Social Services	
Agriculture and Allied Activities	
Construction	
Manufacturing (Textiles)	
Transport, Storage and Communications	
Electricity, Gas & Water companies	
Manufacturing (Paper & Paper products, Publishing, printing and reproduction of recorded media)	
Manufacturing (Machinery & Equipment)	
Manufacturing (Others)	
Finance	
Manufacturing (Metals & Chemicals, and products thereof)	
Manufacturing (Leather & products thereof)	
Mining & Quarrying	
Manufacturing (Wood Products)	
Insurance	

[6] Industry data: http://www.mca.gov.in/MinistryV2/archiveofmasterdatadetails.html Date of access- 16 February 2021

Table 5: Living standard[7]

Attribute name	Description
Total Households	These attributes describe living standard in a particular district, for rural and urban households. Broadly, household types are Normal household, institutional household, houseless and houseless with shelter in district. Attribute indicates number of households in each category in district. For example, “Number Normal Rural” indicate number of households in normal category in rural area in a particular district. Institutional Households- A group of unrelated persons who live in an institution and take their meals from a common kitchen is called an Institutional Household. Examples of Institutional Households are boarding houses, messes, hostels, hotels, rescue homes, jails, ashrams, orphanages, etc. Houseless Households- Households which do not live in buildings or census houses but in the open on roadside, pavements, in Hume pipes, under flyovers and staircases, or in the open in places of worship, mandaps, railway platforms, etc. are treated as Houseless households.
Total Rural Household	
Total Urban Household	
Number Normal Rural	
Number Normal Urban	
Institutional Rural	
Institutional Urban	
Houseless Rural	
Houseless Urban	
Houseless with Shelter Urban	
Houseless without Shelter Urban	

Table 6: Labour Force Data[8]

Attribute name	Description
Not in Labour Force	Number of people who are not in the labour force
Employed	Number of people who are employed
Unemployed	Number of people who are unemployed

Table 7: NSDC Data[9]

Attribute name	Description
TC count	Number of Training centres in district
Total number of students placed in district	Number of students placed in district placed from any training center in that district
Total number of students enrolled for training in district	
Placement percentage in district	Placement percentage = Total placed (in district) *100 /Total enrolled (in district)
Candidate information	Age group, district, state, education level
Training Center (TC) information	TC name, TC district, TC state,
Enrolment information	Batch size, sector name, job role

District level data has been collected for the above data points from various sources[10].

[7] Living standard data: https://secc.gov.in/statePopulationLivingStatusUrban Date of access- 16 February 2021

[8] Labor bureau data : http://labourbureaunew.gov.in/ Date of access- 25 February 2021

[9] https://skillsip.nsdcindia.org/knowledge-products/predictive-analytics-skill-development-leveraging-artificial-intelligence-and date of access- 15 March 2021

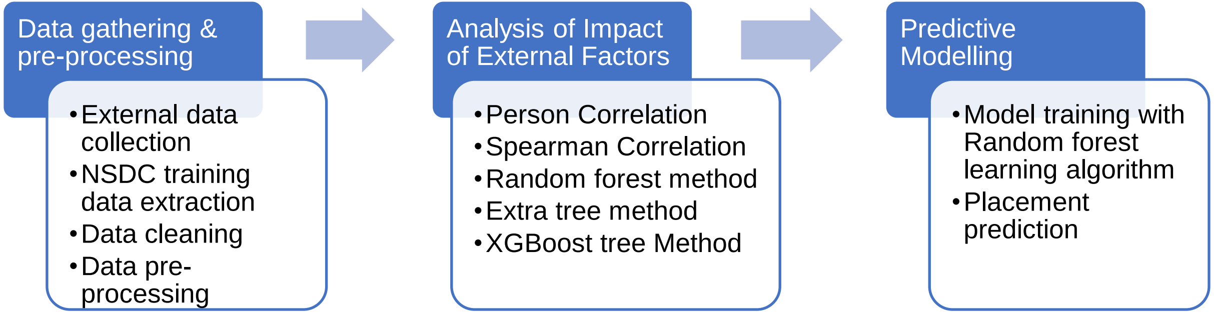
[10] Population data: https://censusindia.gov.in/2011-Common/CensusData2011.html Date of access- 15 February 2021; Living standard data: https://secc.gov.in/statePopulationLivingStatusUrban Date of access- 16 February 2021; Industry data: http://www.mca.gov.in/MinistryV2/archiveofmasterdatadetails.html Date of access- 16 February 2021; School data http://udise.in/drc.htm Date of access- 18 February 2021; College data https://aishe.gov.in/aishe/collegeDirectoryIndex?hasReportLink=index Date of access- 20 February 2021; ITI data: https://www.ncvtmis.gov.in/Pages/ITI/Count.aspx Date of access- 21 February 2021; Labor bureau data : http://labourbureaunew.gov.in/ Date of access- 25 February 2021

1.1. Data cleaning

To streamline and standardize the multivariate dataset provided by NSDC, a data cleaning endeavour was undertaken. The process consists of two steps; first, the attributes serving only as identifiers (each data point a unique distinct value, therefore not useful in analysis, as one cannot get any pattern from such data) such as training centre (Center), candidate id (Cand ID), smart centre id (Smart Centre), training batch id (Batch ID), name of training batch (Batch Name), date of beginning the training (Batch Start Date), date of closing the training (Batch End Date), date of birth of candidates (DoB), pin code of the training centre (Pin code) have been removed. Second, all the attributes that could be derived from other attributes represent duplicate information. It includes Training center name and Training center ID etc. both of which represent the same information. We converted text to numeric value as provided in next section. Similarly, candidate's birth year (Year of Birth), domicile details (Candidate Constituency, Sub District) id for a particular job role and sector, details about centre's constituency (Partner ID, TC constituency) have also been excluded from further analysis.

1.2. Data pre-processing

In the previous step, it is observed that all the attributes being considered for the analysis excluding 'Total Candidates in Batch' are categorical with data being in the text format. Therefore, this step is necessary to convert the categorical data from text to numerical format.



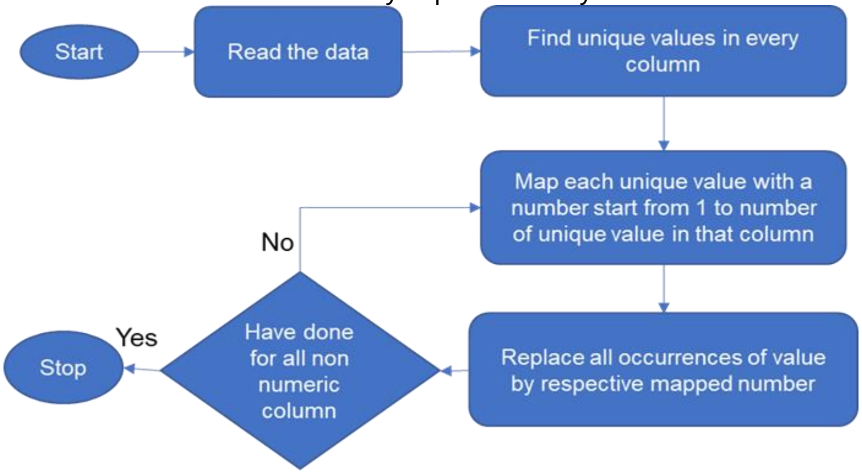
Data gathering & pre-processing
- External data collection
- NSDC training data extraction
- Data cleaning
- Data pre-processing

Analysis of Impact of External Factors
- Person Correlation
- Spearman Correlation
- Random forest method
- Extra tree method
- XGBoost tree Method

Predictive Modelling
- Model training with Random forest learning algorithm
- Placement prediction

Figure 1: Overview of the analysis

This is done by assigning a unique numerical indicator to each of unique entry of an attribute. For example, the attribute Education Level" has 6 unique data categories - "No Education", "10th std & below", "11th-12th std", "Diploma/ITI/Polytechnic/Other", "Undergraduate/ Graduate", "Postgraduate & above". Each of unique entry of an attribute is then replaced by 1,2,3,4,5,6 respectively. Similar conversion is applied to all the text attributes. For analysis, we are using tree-based methods therefore ordinal values which are being used as replacement to text data do not make any impact on analysis.



Start
Read the data
Find unique values in every column
Map each unique value with a number start from 1 to number of unique value in that column
No
Yes
Stop
Have done for all non numeric column
Replace all occurrences of value by respective mapped number

Figure 2: Pre-processing of non-numeric (text) data

1.3. Dataset preparation for analysis

The data have been collected from various sources in different formats. Therefore, pre-processing of data and preparation of data in suitable format is the first and crucial step in successful analysis. Data points such as TC count, number of students placed in district have been extracted from the data shared by NSDC. A small data sample is shown in Table 8.

Table 8: Collected data sample for analysis of external factor on placement

District	Population Age (15-21)	No. of colleges	ITI Trade Count	No of ITI	TC count	Total Enrolled for training	Overall Literacy	…	Number of students Placed
Pulwama	78828	17	21	3	11	93	65	…	55
Poonch	64910	6	16	3	8	98	68.69	…	29
Rajouri	87624	13	20	2	4	127	68.54	…	57
Reasi	43079	10	7	2	2	31	59.42	…	10
Srinagar	168934	48	30	2	13	319	71.21	…	118
Udhampur	74939	10	5	1	12	215	69.9	…	107
Chamba	73468	15	63	17	4	67	73.19	…	28
Kangra	190665	65	230	53	24	278	86.49	…	96

This data has been utilized for finding impact of external attributes on placement of enrolled student. For development of better placement prediction model another dataset has been prepared where, district level external data such as population, no. of colleges, overall literacy etc. have been clubbed with NSDC internal data, that have been used in our previous study[9] as shown in Table 9. Data cleaning and pre-processing of NSDC data is similar as our previous study for more detail of data cleaning and pre-processing please refer to [9].

Table 9: Sample data for predictive modelling

Gender	Job Role	Sector Name	Partner Name	TC District	Age group	No. of colleges	Trade Count	No of ITI	Total Number of Schools	Total Population	Overall Literacy	...	Placed
Male	CCTV Installation Technician	Electronics and Hardware	A.S. Education and Welfare Society	Jhajjar	19-21	69	127	15	885	956907	80.83	...	Yes

Gender	Job Role	Sector Name	Partner Name	TC District	Age group	No. of colleges	Trade Count	No of ITI	Total Number of Schools	Total Population	Overall Literacy	...	Placed
Female	Trainee Associate	Retail	Orchid Skills	Jhajjar	22-25	69	127	15	885	956907	80.83	...	Yes
Female	Sewing Machine Operator	Apparel	Orion Edutech Private Limited	Jhajjar	22-25	69	127	15	885	956907	80.83	...	No
Female	Assistant Beauty Therapist	Beauty and Wellness	Orion Edutech Private Limited	Jhajjar	22-25	69	127	15	885	956907	80.83	...	Yes
Male	Showroom Hostess - Customer Relationship Executive	Automotive	Webtech Universal Learning Pvt. Ltd.	Jhajjar	22-25	69	127	15	885	956907	80.83	...	No
Male	Retail Sales Associate	Retail	Orion Edutech Private Limited	Jhajjar	22-25	69	127	15	885	956907	80.83	...	Yes

Data collection and data preparation for analysis is very crucial part of the study. Data has been collected from various sources and converted into desired format. Moreover, non-useful data is discarded and categorical data is transformed into numeric data for conducting the analysis.

2. Analysis for impact of external factor on the placement:

After the step entailing the pre-processing of the data, an ‘Impact analysis’ is conducted to understand the impact of external factor on the placement of the candidates.

In the preliminary stage of the attribute analysis, a correlation graph was created to examine the linear relation among the various external attributes.

2.1. Correlation graph

To determine correlation between the placement and the external factors, two different correlation graphs have been plotted. One using the Pearson correlation and another using Spearman correlation. Where Pearson Correlation shows the linear relationship between two sets of data. In simple terms, it answers the question, can one draw a line graph to represent the data? However, Spearman's correlation coefficient, measures the strength and direction of association between two ranked variables. More technical detail of Person correlation and Spearman correlation are given in the appendix.

A correlation graph provides value in range of [-1, 1]. Both in Pearson and Spearman correlation, an absolute value of 1 indicates a perfect linear relationship between the variables. A value close to 0 indicates no linear relationship between the variables. Higher absolute value indicates strong correlation and lower value represents a weak relation. In general, absolute correlation value (only numerical value without positive or negative sign) should be greater than 0.3 for variables to be considered as strongly related. The sign of value represents the type of proportionality. Positive value indicates a directly correlation and a negative value indicates inverse correlation. In direct correlation, value of one variable increases with the other and in inverse correlation, value of the variable decreases with increase in the other.

From the Figure 3, it is clearly visible that living standard such as number of person “Houseless with shelter Urban”, Number of person living “normal Urban”, “total household in urban”, have very high correlation with placement percentage. Moreover, industry presence in the district such as number of finance companies in district also has high correlation with placement percentage.

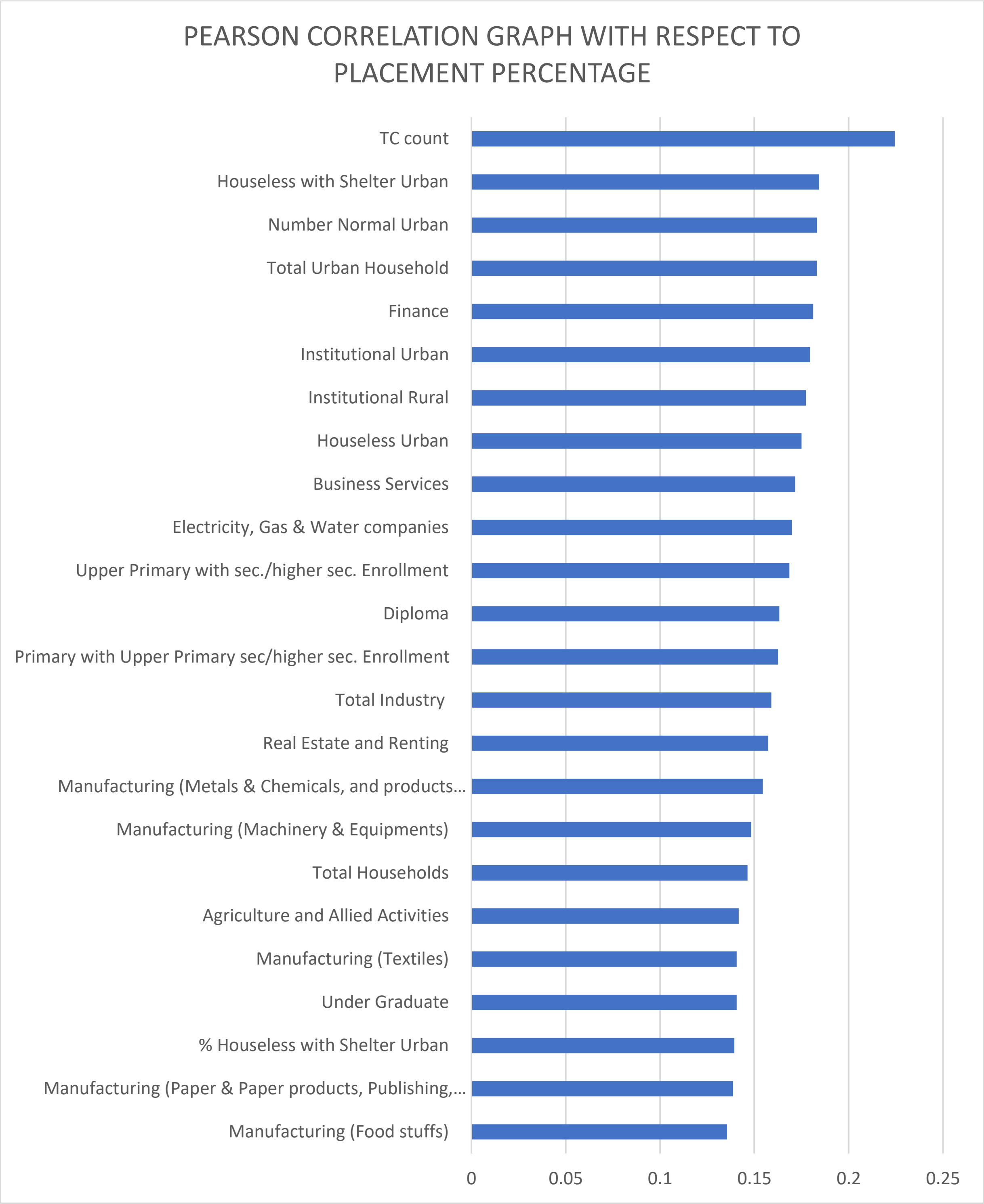

PEARSON CORRELATION GRAPH WITH RESPECT TO PLACEMENT PERCENTAGE

Figure 3: Pearson correlation of external attributes with placements percentage in district

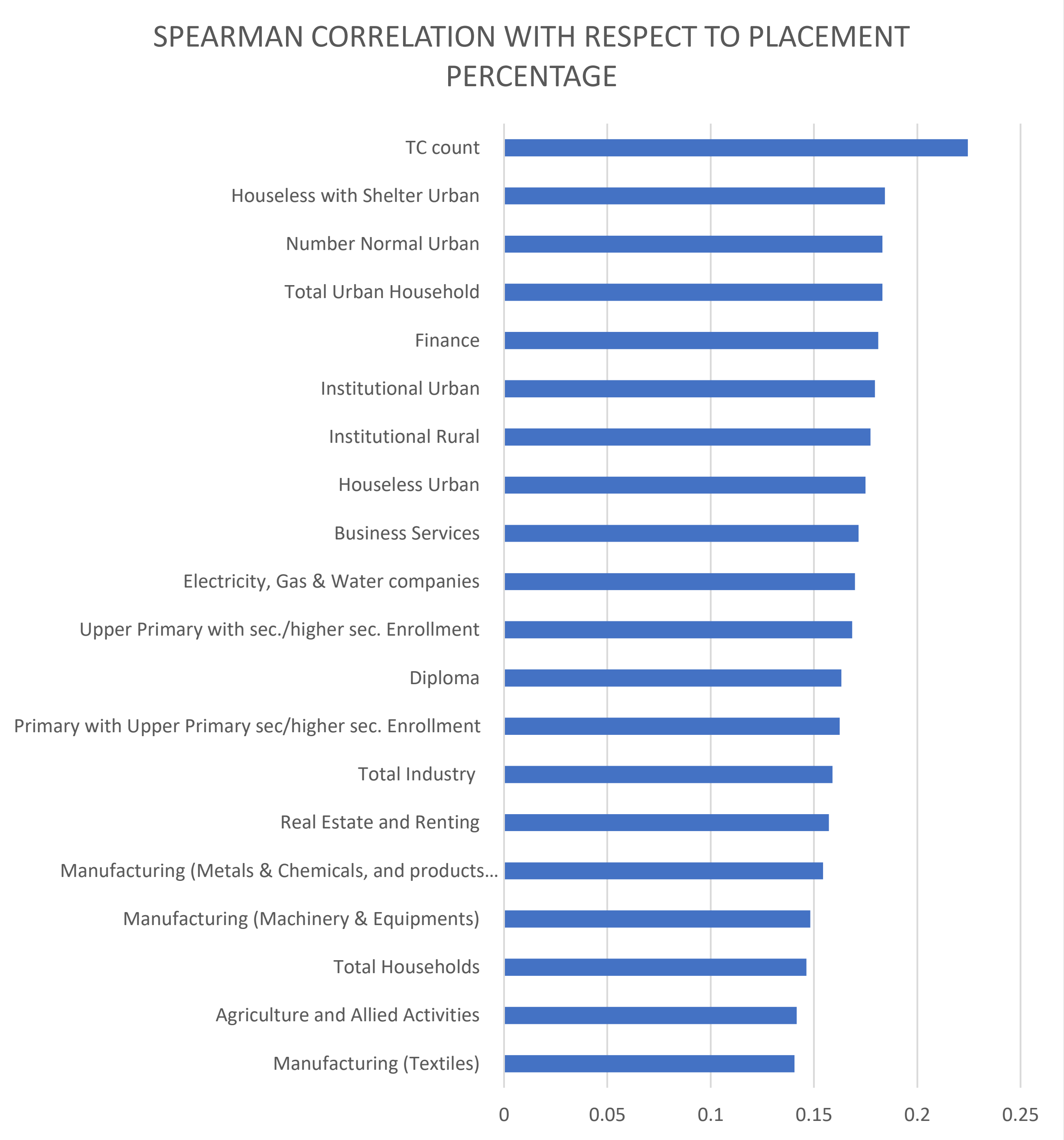


Figure 4: Spearman correlation of external attributes with placements percentage in district

From the Figure 4 also, it is clearly visible that living standard such as number of persons “Houseless with shelter Urban”, Number of person living “normal Urban”, “total household in urban”, have very high correlation with placement percentage. Moreover, industry presence in the district such as number of finance companies in district also has high correlation with placement percentage.

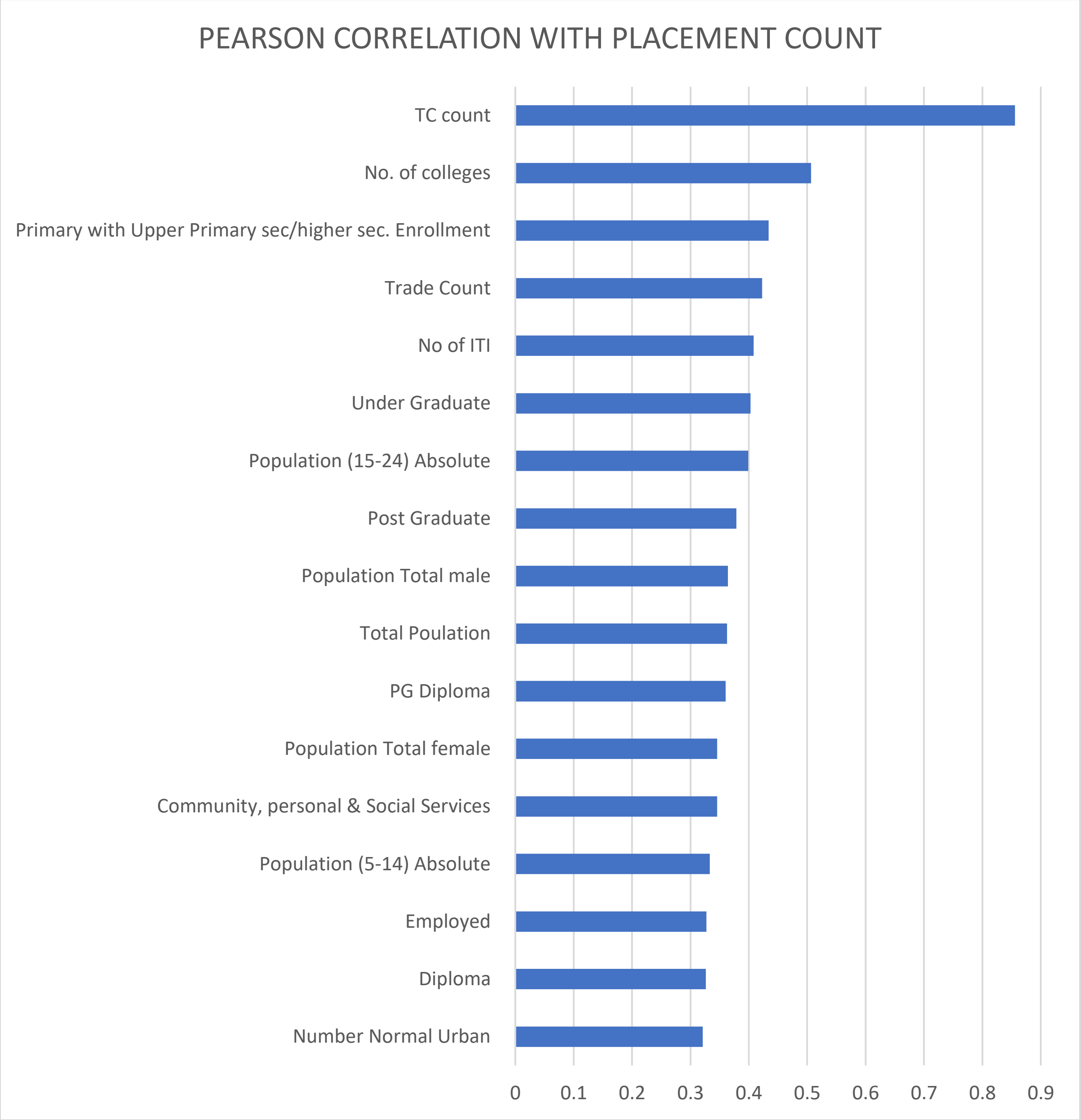


Figure 5: Pearson correlation of external attributes with number of placements in district

From the Figure 5, a clear inference is drawn that placement has very high correlation with TC count which is obvious. If a district has large number of TC, then there will be more total placement. Moreover, interesting point is placement also has very high relation with number of colleges in district which implies, if the district has high focus on higher education and has large number of colleges in district then placement of the student in that district is also high. Similarly, if a district has high number a primary and upper primary sec/higher sec enrolments, then placement is also high. As indicated in graph, if a district has a large young population, then placement is also high.

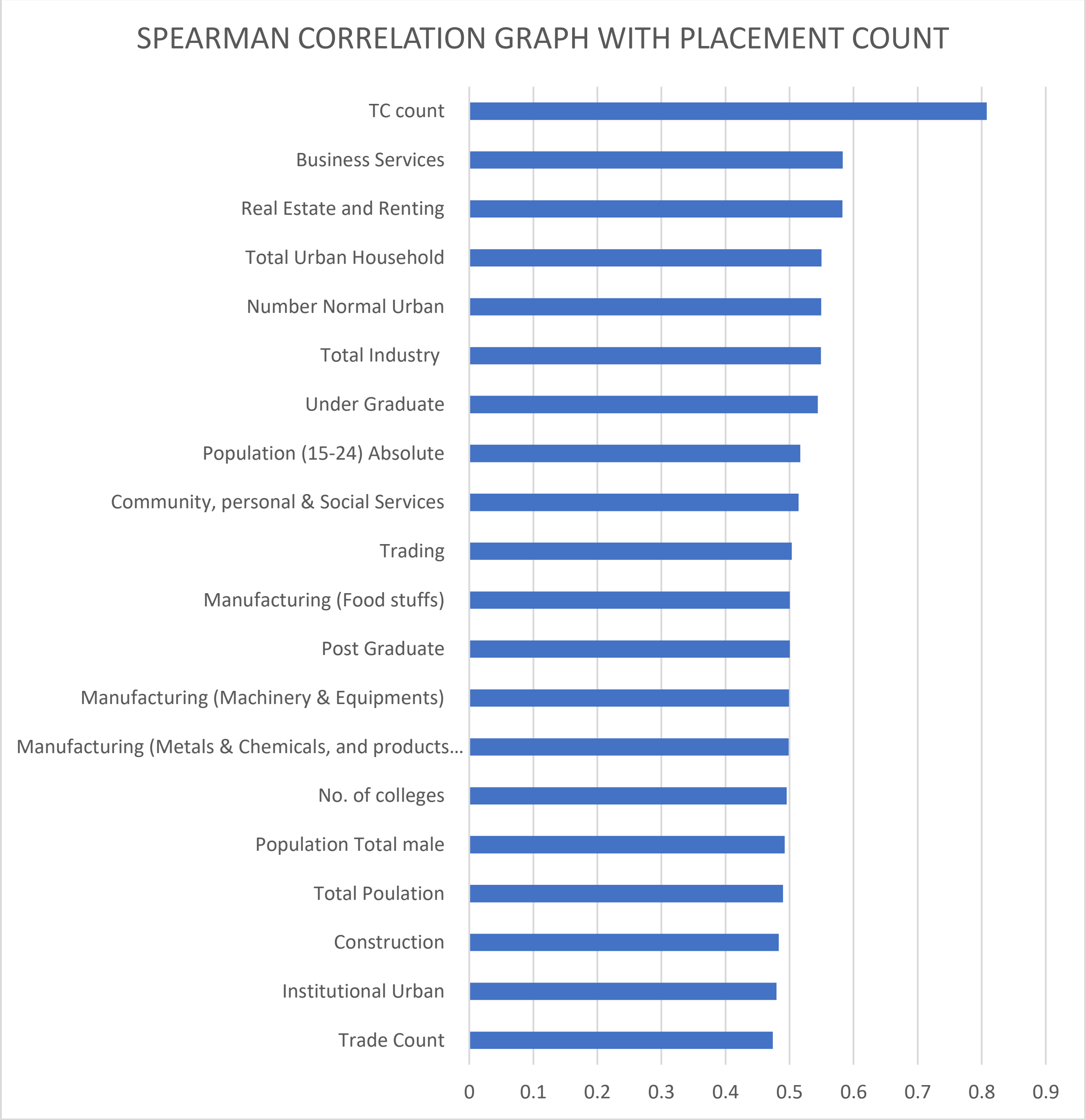


Figure 6: Spearman correlation of external attributes with number of placements in district

From the Figures 6, a clear inference is drawn that placement have very high correlation with TC count which is obvious. On the top of it, Spearman correlation indicates that if a district has more business service company more real state and renting company or more industries then placement also high of a student who trained in that district. It is clearly visible here also more young population aged between 15-21/24 have higher correlation with placement.

As shown Figure 7 and 8 placement of a candidate heavily depend upon the certification, result, and assessment status. Moreover, placement also depend upon geography and sector name and Job role. Another point to notice that lower age group of candidates have high possibility of placement.

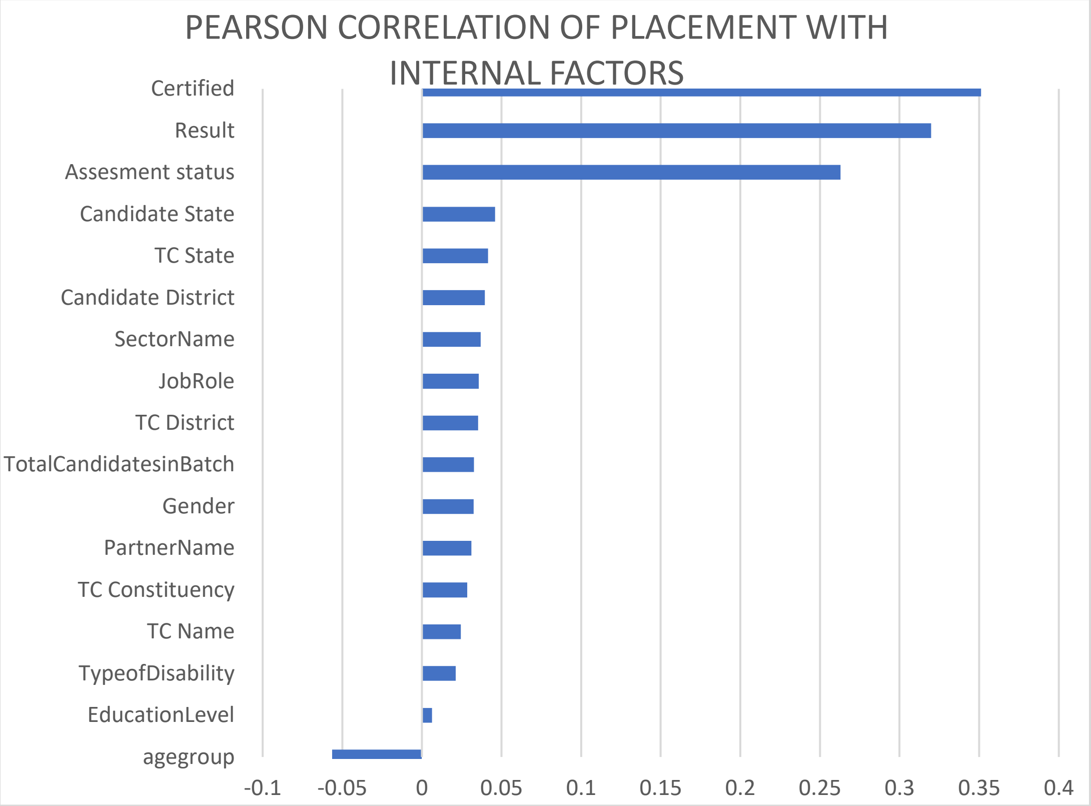


Figure 7: Pearson correlation of placement with internal factors

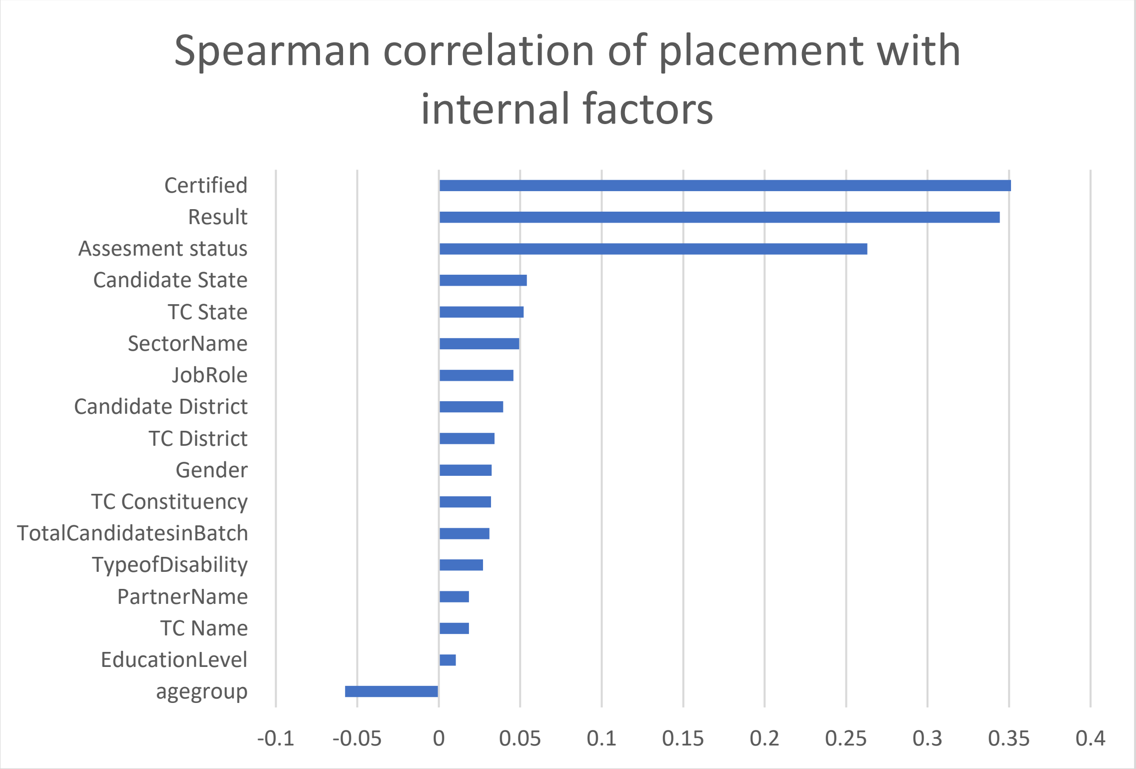


Figure 8: Spearman correlation of placement with internal factors

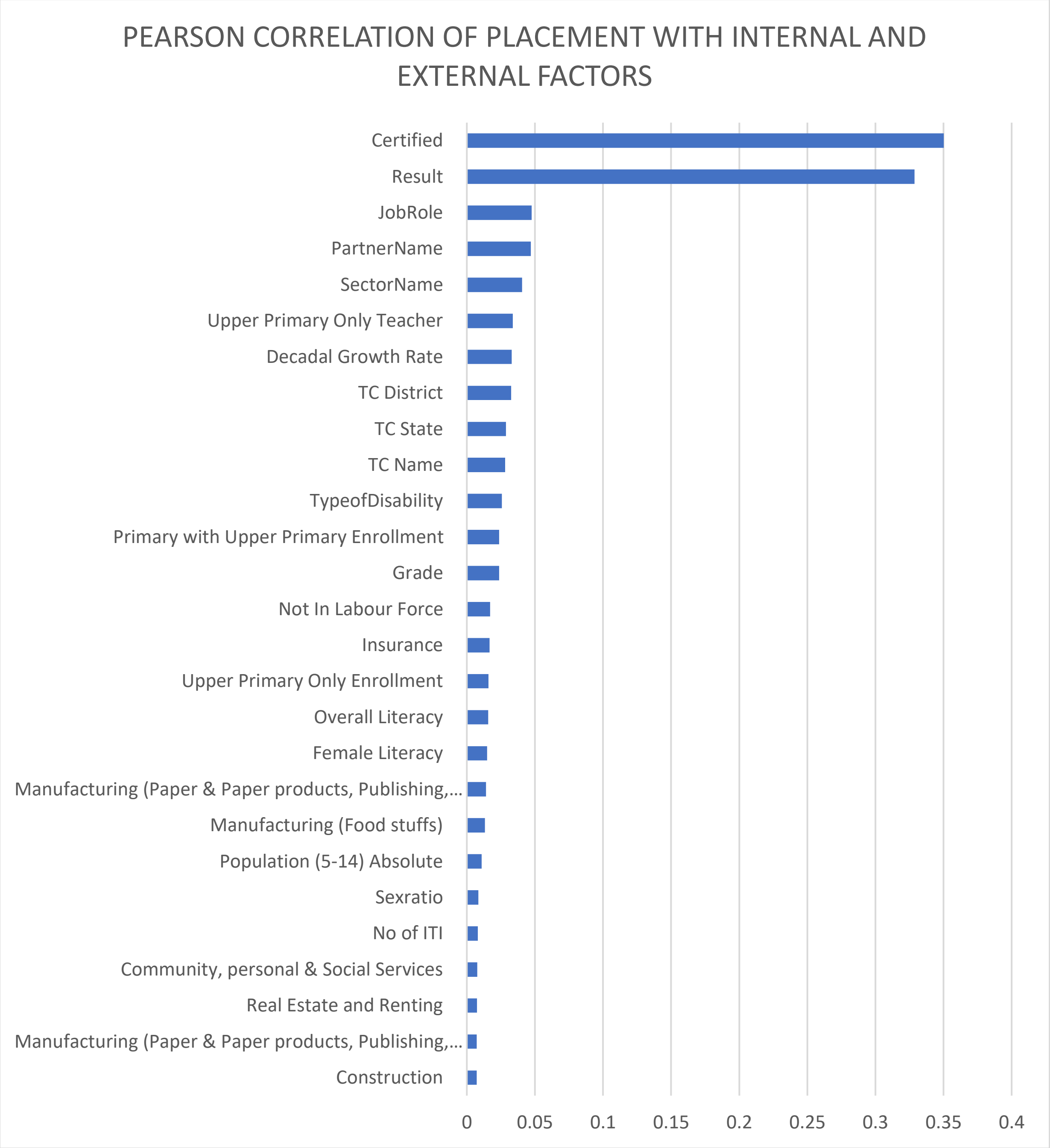


Figure 9: Pearson correlation of placement with internal factors and external factors

The correlation of external factors with placement has been measured using Person correlation and Spearman correlation. Person and spearman correlation of external factor with placement percent clearly indicate that living standard and industry presence in the district have very high correlation with placement percentage. Whereas spearman correlation with placement count indicates that if a district has more business service companies, more real state and renting companies or more industries then total placement in that district is likely to be high

3. Attribute Relevance Analysis

3.1. Tree based Methodology for attribute analysis

A tree-based model was adopted to explore the relation between placement and other attributes. Given the problem statement and scenarios at hand (where most of the data is categorical data), tree-based models are considered more appropriate and perform better than other models on this data. The four tree-based models considered for this analysis are Decision Tree, Random Forest, Extra Tree Classifier, XGBoost classifier. The explanations of these models have been given in the Annexures.

The attribute analysis and predictive modelling was based on the above approaches. After applying all the four models on the data, it was observed that Random Forest Model was best suited in providing results with a relatively higher accuracy than the other methods. In view of the same, the attribute relevance analysis in the further sections is based on Random Forest Model.

Table 10: Top 24 Attributes from relevance analysis for placement prediction

Attributes	Attribute Importance	Attribute ranking
Age group	0.111464	1
Job Role	0.089938	2
TC Name	0.08875	3
Partner Name	0.085398	4
Sector Name	0.069905	5
Certified	0.062504	6
Result	0.059081	7
Education Level	0.058685	8
Grade	0.050815	9
Gender	0.035845	10
Total Candidates in Batch	0.033634	11
TC State	0.002401	12
Decadal Growth Rate* of District	0.002289	13
TC District	0.002165	14
Houseless Rural (living standard)	0.002161	15
Upper Primary with sec./higher sec. Enrolment (education)	0.00213	16
Institutional Urban (living standard)	0.002116	17
Upper Primary Only Teacher	0.002082	18
Primary with Upper Primary Enrolment (education)	0.00208	19
Overall Literacy	0.002035	20
Institutional Rural (living standard)	0.002034	21
Female Literacy	0.002005	22
Sex Ratio	0.001983	23
Upper Primary Only Enrolment (education)	0.001964	24

Attribute relevance analysis shows that internal factors namely age of the candidate, job role, and name of the training center in which the candidate is enrolled have large contribution in successful placement of the candidate. However, among external factors - decadal growth rate of the district, total number of upper primary with sec./higher sec. enrolment, and overall literacy are key contributing factors in determining individual candidate's placement.

Attribute Ranking: The importance of the attributes and their associated rankings with respect to placement are presented in the Table 10. It is evident from the table that the attributes such as the age of the candidates, job role, name of the training center, sector name, and education level a have maximum impact

on placement (to predict probability of placement at enrolment stage). As we already know internal factor heavily impacting the success of training program in terms of placement, but as evident from table along with internal factor external factor such as decadal growth rate*, Upper Primary with sec./higher sec. Enrolment, overall literacy also impacting a lot on the placement of a candidate.
Note: *decadal growth rate gives an overview of the total population **growth** in a particular decade

4. Predictive Modelling

4.1. Model Training

Based on the attribute ranking shown in Table 10, a predictive modelling of placement was done. This modelling allowed us to explore the possibilities of placement of a skilled candidate. Going forward this analysis can help the training center's (TCs) to direct the potential candidates to specific sectors and job roles which are most suitable for them, especially in terms of placement.

For the predictive modelling the linear Regression, Decision Tree, Random Forest, Extra Tree methods, Artificial Neural Network, and XGBoost were used. Post explorations of the various models, the Random Forest method worked most efficiently. This method achieved a 78% accuracy while predicting if a candidate will be place or not after the training. On the other hand, in previous study[9] when only internal data was present, algorithm achieved 75% of accuracy on predicting placement status of a candidate. So, it evident from study that use of external data along with internal data can give 3% improvement in prediction of placement of a candidate.

For predictive modelling, the trained model was created using given data along with "Random Forest learning algorithm" is shown in Figure 10.

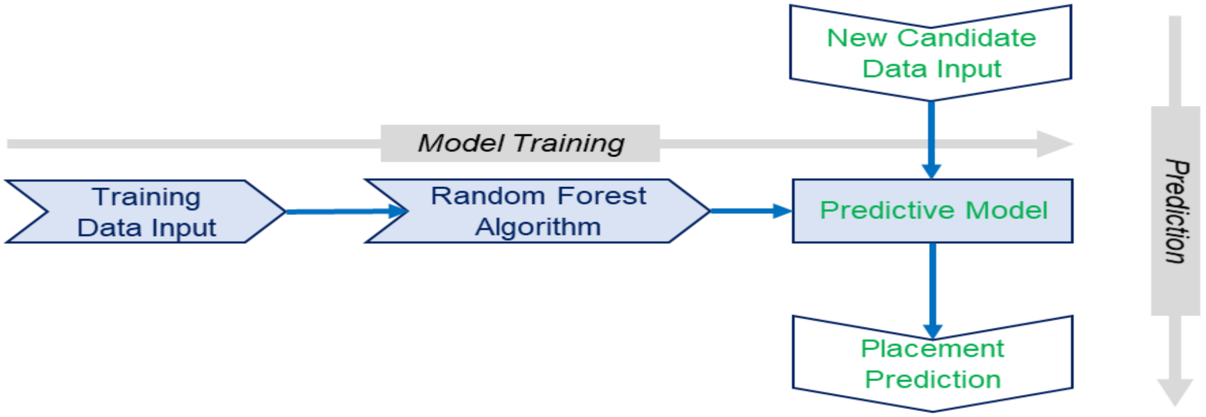


Figure 10: Predictive Modelling of placement

4.2. Prediction of placement

Post the creation and training of the model, one has to provide the new data as input and the model will be able to predict the chance of placement of a candidate. This model can be leveraged at the end of training to forecast that candidate will place or not with 78% confidence using his profile, training centre data and external district level data such as: *Age Group, Candidate District, Education Level, Gender, Candidate State, Type of Disability, TC Name, Partner Name, TC Constituency, TC District, TC State, Total Candidates in Batch, Job Role, Sector Name, and external data such district population, education status, industries in district, living standard.* An example of utilization of predictive modelling is shown in Figure 11.

Table 11: Input data Sample

Sno.	1	2	3	4	5	6
Gender	Male	Female	Female	Female	Male	Male
JobRole	CCTV Installation Technician	Trainee Associate	Sewing Machine Operator	Assistant Beauty Therapist	Showroom Hostess - Customer Relationship Executive	Retail Sales Associate
SectorName	Electronics and Hardware	Retail	Apparel	Beauty and Wellness	Automotive	Retail
PartnerName	A.S. Education and Welfare Society	Orchid Skills	Orion Edutech Private Limited	Orion Edutech Private Limited	Webtech Universal Learning Pvt. Ltd.	Orion Edutech Private Limited
TC District	Jhajjar	Jhajjar	Jhajjar	Jhajjar	Jhajjar	Jhajjar
Agegroup	19-21	22-25	22-25	22-25	22-25	22-25
Colleges count	69	69	69	69	69	69
Trade Count	127	127	127	127	127	127
ITI count	15	15	15	15	15	15
Total Number of Schools	885	885	885	885	885	885
Total Population	956907	956907	956907	956907	956907	956907
Overall Literacy	80.83	80.83	80.83	80.83	80.83	80.83
...	**...**	**...**	**...**	**...**	**...**	**...**

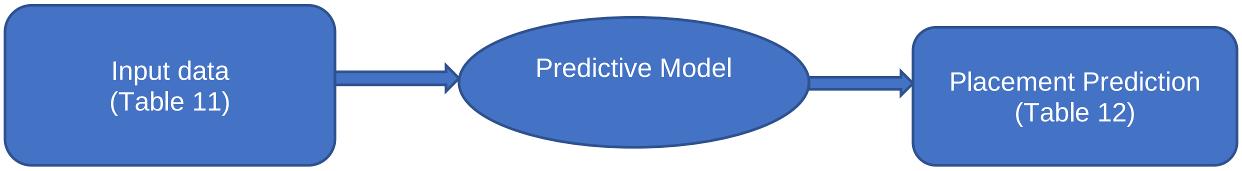


Figure 11:An execution example of predictive modelling

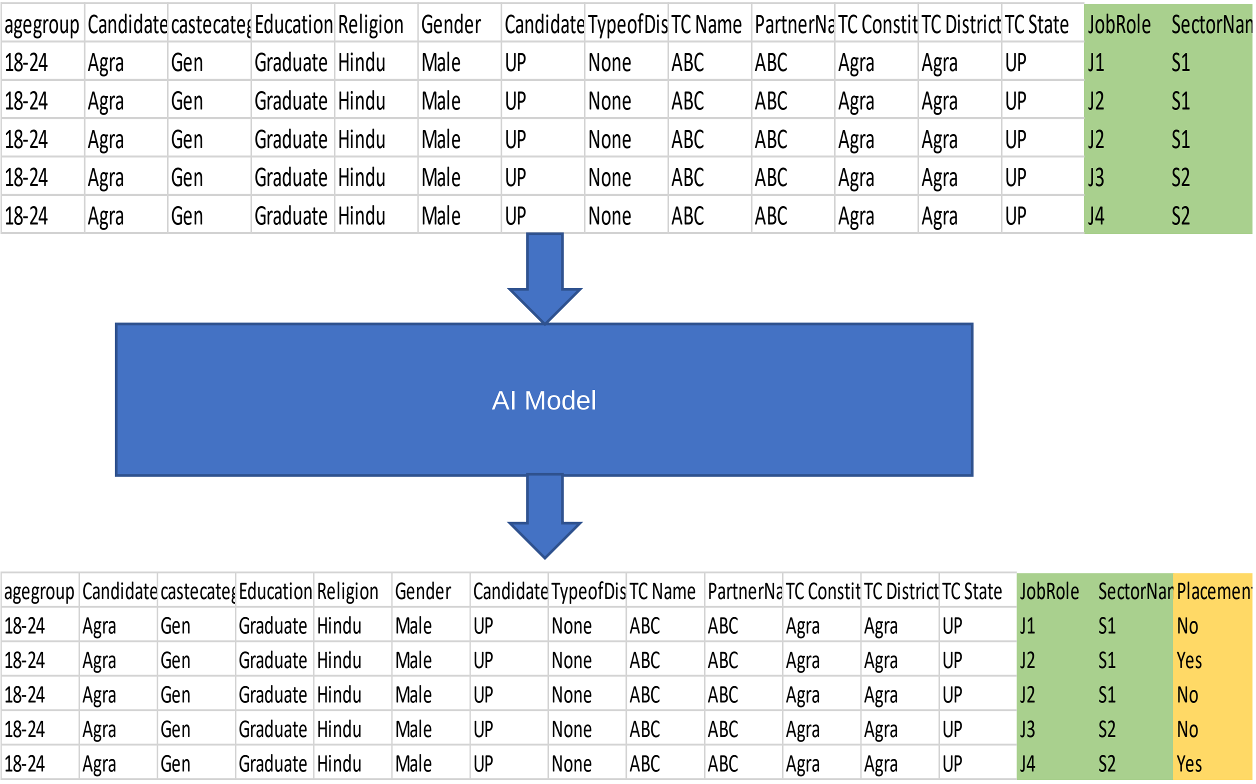
Table 12: Placement prediction output

Sno.	1	2	3	4	5	6
Gender	Male	Female	Female	Female	Male	Male
JobRole	CCTV Installation Technician	Trainee Associate	Sewing Machine Operator	Assistant Beauty Therapist	Showroom Hostess - Customer Relationship Executive	Retail Sales Associate
SectorName	Electronics and Hardware	Retail	Apparel	Beauty and Wellness	Automotive	Retail
PartnerName	A.S. Education and Welfare Society	Orchid Skills	Orion Edutech Private Limited	Orion Edutech Private Limited	Webtech Universal Learning Pvt. Ltd.	Orion Edutech Private Limited
TC District	Jhajjar	Jhajjar	Jhajjar	Jhajjar	Jhajjar	Jhajjar
Agegroup	19-21	22-25	22-25	22-25	22-25	22-25
Colleges count	69	69	69	69	69	69
Trade Count	127	127	127	127	127	127
ITI Count	15	15	15	15	15	15
Total Number of Schools	885	885	885	885	885	885
Total Population	956907	956907	956907	956907	956907	956907
Overall Literacy	80.83	80.83	80.83	80.83	80.83	80.83
...	...	...	...	...	...	...
Placed	**Yes**	**Yes**	**No**	**Yes**	**No**	**Yes**

4.3. Suggestive Modelling

The trained model which is described in figure 5 can also be used as suggestive modelling. For an example if a candidate in Agra went to a training centre and requested suggestion for a placement-oriented job role, the model can generate the placement probabilities of the candidate basis some generic data as mentioned previously. Hence this model can aid candidates wishing to pursue vocational education and skill development to select industry relevant courses. Similarly, if a candidate lives halfway in between Agra and Mathura, it provides him/her the option to undertake training in any of the district. Using the created model one can check the details of all training centers, job roles being taught in both the districts and exactly suggest candidates the right mix of job role, district and training centre for maximizing their placement potential. The flow diagram of the example is given in figure 12:

Using the Random Forest method, the predictive model created allowed one to gauge the placement of the candidate with 78% accuracy when external data also used for placement prediction along with the internal NSDC data. The use of such model can be leveraged at the enlistment and counselling phases to support the candidates to pass through towards industry focused skills.



agegroup	Candidate	castecateg	Education	Religion	Gender	Candidate	TypeofDis	TC Name	PartnerNa	TC Constit	TC District	TC State	JobRole	SectorNan
18-24	Agra	Gen	Graduate	Hindu	Male	UP	None	ABC	ABC	Agra	Agra	UP	J1	S1
18-24	Agra	Gen	Graduate	Hindu	Male	UP	None	ABC	ABC	Agra	Agra	UP	J2	S1
18-24	Agra	Gen	Graduate	Hindu	Male	UP	None	ABC	ABC	Agra	Agra	UP	J2	S1
18-24	Agra	Gen	Graduate	Hindu	Male	UP	None	ABC	ABC	Agra	Agra	UP	J3	S2
18-24	Agra	Gen	Graduate	Hindu	Male	UP	None	ABC	ABC	Agra	Agra	UP	J4	S2

agegroup	Candidate	castecateg	Education	Religion	Gender	Candidate	TypeofDis	TC Name	PartnerNa	TC Constit	TC District	TC State	JobRole	SectorNar	Placemen
18-24	Agra	Gen	Graduate	Hindu	Male	UP	None	ABC	ABC	Agra	Agra	UP	J1	S1	No
18-24	Agra	Gen	Graduate	Hindu	Male	UP	None	ABC	ABC	Agra	Agra	UP	J2	S1	Yes
18-24	Agra	Gen	Graduate	Hindu	Male	UP	None	ABC	ABC	Agra	Agra	UP	J2	S1	No
18-24	Agra	Gen	Graduate	Hindu	Male	UP	None	ABC	ABC	Agra	Agra	UP	J3	S2	No
18-24	Agra	Gen	Graduate	Hindu	Male	UP	None	ABC	ABC	Agra	Agra	UP	J4	S2	Yes

Figure 12: An execution example of suggestive modelling

As per the figure mentioned above, the model predicted that the candidate in question should choose job role J2 and sector S1 or job role J4 and sector S2 as these entail the highest probability of placement. Similarly, the predictions can be based on district or training partner or training centre name etc.

.

Risks and possible mitigation

The model was created based on the data shared by NSDC wherein 1.5 lakh random data points was shared out of a very large data set. Moreover, external data collected from various places are also not up to date. For example, population related data is 10 years old because Census 2011 data is only available for the analysis. Similarly, industry related data is 6 years old the latest available data was from year 2015. Similarly, Education and training data is from 2017. Therefore, the reported model performance and confidence on it is based on the shared data sets. A training of the created model on the entire available data set and the latest external data would create more accurate and useful results.

Moreover, the prediction from the model is a possibility based on the previous records and currently the model does not capture very uncertain events as well as the aptitude of a potential candidate. A more in-depth analysis would be required of ascertaining the best fitted job roles wherein their interests are documented as a variable for the model which is currently not there.

Recommendations

The AI/ML predictive model as created can be leveraged to garner deeper insights into the skill development ecosystem and the linkages with the placement structures. Based on the data sets shared by NSDC and public data set of various factors a lot of details could be understood, to ensure greater feasibility and enhanced use of the model, the following recommendations can be looked into:

1. **Data inconsistency:** During the process of data collection and preparation lots of inconsistency found in different agency data such as inconsistency in district name, spelling mistake etc. which make hard to prepare the data. Govt. should have a common naming convention for standard data points

2. **Incorporation of salary details**: If the salary details are well captured by the Government agencies and the private training partners then the model may also be able to predict the salary scales and range of the job roles when interlinked with the district and other relevant details of external environment. While in schemes such as PMKVY, the salary details are captured, it is important that it is made mandatory and even updated during the placement tracking process

3. **Integration of latest data sets in the MIS or data templates:** The external environment data used in the model is slight dated. Given that the skill development ecosystem and success of the system in terms of placements works in consonance with the industrial, demographic, and educational variables of the district. Therefore, NSDC may institute an internal mechanism to collect latest data or buy the relevant dataset from paid vendors. Latest integrated data (internal + external) will be quite useful in the current study as well as other data-driven studies

4. **Frequency of running the model:** It is imperative that the model runs at regular intervals to be able to add a dynamism to the model while also being able to take account of the constantly evolving economic and social indicators.

Conclusion

This report documented how the AI/ML and data science methods can be leveraged in getting a deep dive into the nuances of data collected from various external sources and the data generated during the training lifecycle. The analytics has the potential to map skills by requirement, identify potential impact of external factor that can be beyond control of NSDC such as number of young populations, total industry in a district etc. on the placement of the candidates and success of a training center, predictive analysis of demand for new occupations and skills – in quasi real time.

The insights from this exercise showcased the impact of external factors such as number of college in a district, number on primary and upper primary sec/higher sec enrolment, population aged between 15-24, and number of industries in district have high impact on placement along with the internal factor such as age groups, training center name, training partner name and. The predictive models which have been created using the external and internal data sets has the potential of capitalizing on key insights of the training value chain. Some of the ways in which the model can be leveraged are the following:

- **Creation of market relevant business models:** Given the stake of the private entities in skill development, an access to these insights would be able to help training providers in investing in infrastructure for skills which are market driven and would allow them to place candidates post the training. Hence better-informed investment opportunities can be created not just for the national but also for the international entities which would want to enter the Indian VET domain in varied capacities

- **Finding potential location to develop new training center:** Based on the impact of external factor determined in the analysis we can suggest location for new training centre which can have high probability of placement after the successfully completion of training

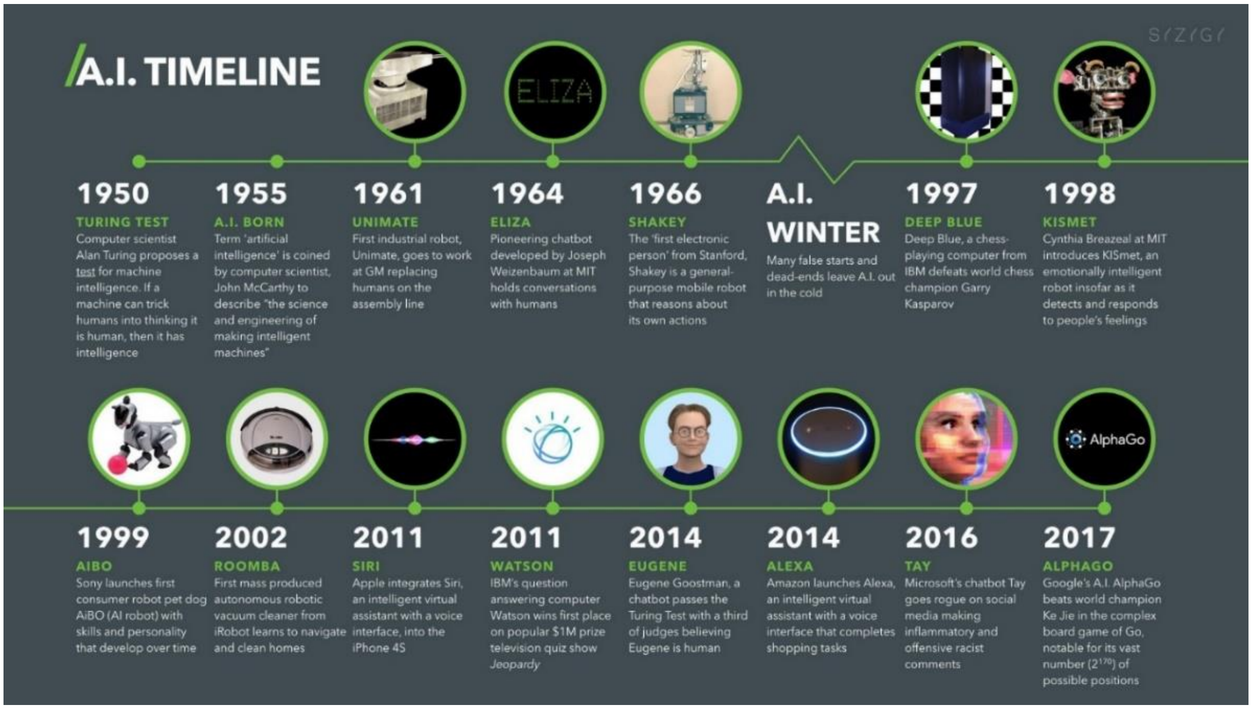
- **Course offerings:** The model through its analysis is able to understand the market relevance of skills. Since it is imperative that for any placement linked course, its industrial worth needs to be adjudged, NSDC can leverage the model to study the placement scenario at regular interval of times and create the necessary amendments in their course offerings

- **Policy making:** The agility of the model to map labour market intelligence would allow government officials to take more data driven decisions. Given the veracity of the results of mathematical models, the insights provided would be beneficial in analysing micro trends which cumulatively can be affect the scheme design and implementation. Moreover, budgetary allocations for trainings would be informed by the labour market intelligence allowing an emphasis of funding support to courses with higher placement probabilities and thereby have more trained candidates to satiate the industry demands. As mentioned in the current model, variables such as age no of colleges in district, total number of industry and number on primary and upper primary sec/higher sec enrolment have very strong correlation with placements, hence when scheme designs and funding is decided, it would be appropriate to take account of these variables in aligning the decisions.

Annexure

1. Introduction Artificial Intelligence: Past Present and Future

Artificial intelligence (AI) is not new. The term was coined in 1956 by John McCarthy, a computer science professor from Stanford University who organized an academic conference on the topic at Dartmouth College in the summer of that year. The field of AI has gone through a series of boom-bust cycles since then, characterized by technological breakthroughs that stirred activity and excitement about the topic. As you can see in Figure 13, today we in the era of an 'AI'. Artificial intelligence can be defined as human intelligence exhibited by machines; systems that approximate, mimic, replicate, automate, and eventually improve on human thinking. Throughout the past half-century a few key components of AI were established as essential: the ability to perceive, understand, learn, problem solve, and reason.



/A.I. TIMELINE

1950 TURING TEST
Computer scientist Alan Turing proposes a test for machine intelligence. If a machine can trick humans into thinking it is human, then it has intelligence

1955 A.I. BORN
Term 'artificial intelligence' is coined by computer scientist, John McCarthy to describe "the science and engineering of making intelligent machines"

1961 UNIMATE
First industrial robot, Unimate, goes to work at GM replacing humans on the assembly line

1964 ELIZA
Pioneering chatbot developed by Joseph Weizenbaum at MIT holds conversations with humans

1966 SHAKEY
The 'first electronic person' from Stanford, Shakey is a general-purpose mobile robot that reasons about its own actions

A.I. WINTER
Many false starts and dead-ends leave A.I. out in the cold

1997 DEEP BLUE
Deep Blue, a chess-playing computer from IBM defeats world chess champion Garry Kasparov

1998 KISMET
Cynthia Breazeal at MIT introduces KISmet, an emotionally intelligent robot insofar as it detects and responds to people's feelings

1999 AIBO
Sony launches first consumer robot pet dog AiBO (AI robot) with skills and personality that develop over time

2002 ROOMBA
First mass produced autonomous robotic vacuum cleaner from iRobot learns to navigate and clean homes

2011 SIRI
Apple integrates Siri, an intelligent virtual assistant with a voice interface, into the iPhone 4S

2011 WATSON
IBM's question answering computer Watson wins first place on popular $1M prize television quiz show *Jeopardy*

2014 EUGENE
Eugene Goostman, a chatbot passes the Turing Test with a third of judges believing Eugene is human

2014 ALEXA
Amazon launches Alexa, an intelligent virtual assistant with a voice interface that completes shopping tasks

2016 TAY
Microsoft's chatbot Tay goes rogue on social media making inflammatory and offensive racist comments

2017 ALPHAGO
Google's A.I. AlphaGo beats world champion Ke Jie in the complex board game of Go, notable for its vast number (2^{170}) of possible positions

Figure 13: AI timeline Source: Digital wellbeing, Paul Marsden

Despite the oversimplification that tends to define AI in the popular press, AI is not one single, unified technology. AI is actually a set of interrelated technology components that can be used in a wide variety of combinations depending on the problem it addresses. Generally, AI technology consists of sensing components, processing components, and learning components as shown in Figure 14.

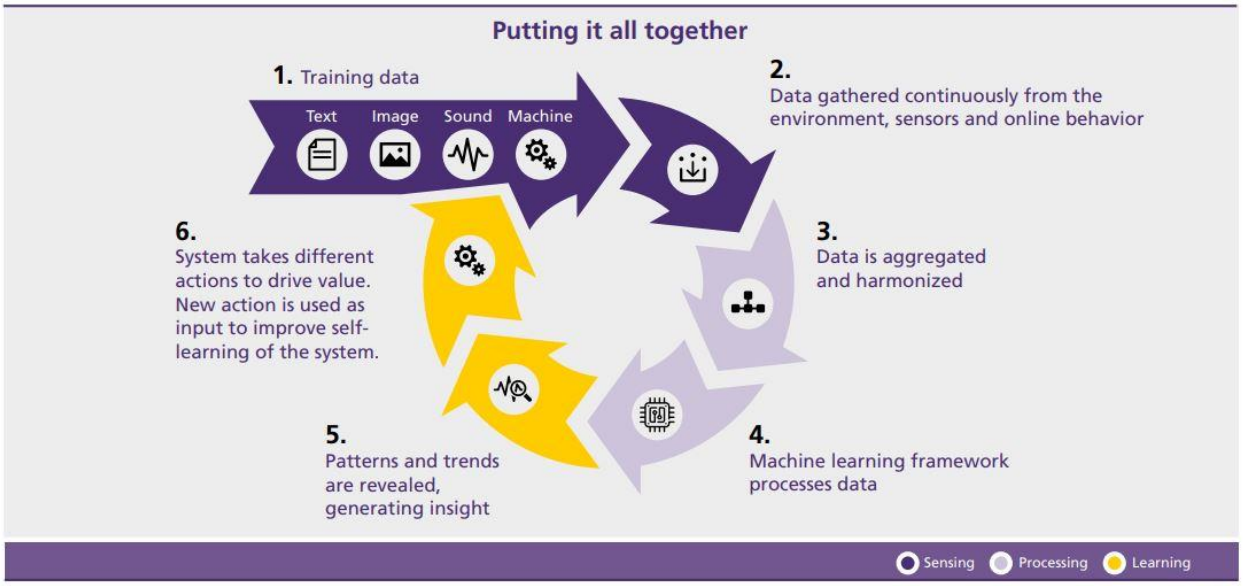

Putting it all together
1. Training data
Text
Image
Sound
Machine
2. Data gathered continuously from the environment, sensors and online behavior
3. Data is aggregated and harmonized
4. Machine learning framework processes data
5. Patterns and trends are revealed, generating insight
6. System takes different actions to drive value. New action is used as input to improve self-learning of the system.
Sensing
Processing
Learning

Figure 14: An AI learning cycle Source: IBM

2. Investment & Funding for AI

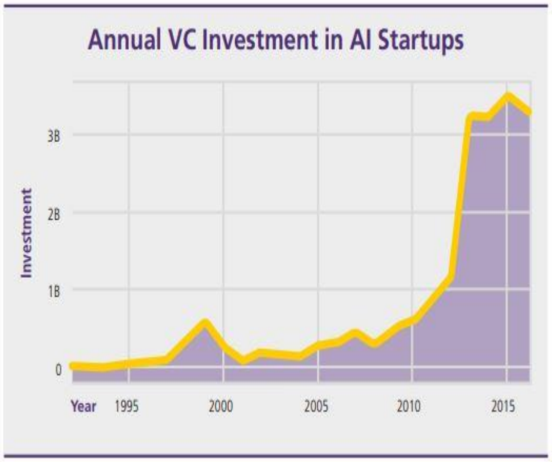


Figure 15: Investment in in US start-ups developing AI systems; Source: AIindex.org

Other contributing factors to the recent surge in progress and interest in AI are the precipitous spikes in venture capital investment in AI start-ups and corporate funding for AI R&D and acquisitions. In 2017 alone, a group of 100 AI start-ups raised $11.7 billion in aggregated funding across 367 deals, contributing to a sixfold increase in investment since 2000 as depicted in Figure 14. Among technology corporations, Baidu and Google specifically spent between $20-$30 billion on AI in 2016, where 90% was allocated for R&D and deployment, and 10% for acquisitions. Finally, 9,043 U.S. patents were issued to IBM in 2017, more than 3,300 of which were related to AI or cloud technologies.

The NSDC has an extensive network of skilling partners across the country. The partners receive funding and payments based on skilling targets they achieve during a point in time. One key challenge that NSDC faces is to analyse and utilize the existing data and make new plans for funding support to geographic area and sectors which are good performer in terms of placement. For this they are required to develop a system that can predict the placement. NSDC has over 20 lakhs trained individuals and over 538 training partners. To top it all, there are more than 37 Sector Skill Councils and over 10373 training centres scattered across the length and breadth of India[11]. To analyse such enormous data and develop a model that can predict the placement of a candidate is a complex task. In addition, a suggestive model, that can guide to a new candidate for choosing a sector and job role so that the placement probability will be more after completing the training.

[11] https://nsdcindia.org/about-us date of access- 08 September 2020

3. Algorithms and Methodology

3.1. Decision Tree Classifier[12]

Decision Tree is a type of supervised learning algorithm (having a predefined target variable) that is mostly used in classification problems. It works for both categorical and continuous, input and output variables. In this technique, the population or sample is split into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables.

Consider an example; There is a sample of 30 students with three variables gender (boy / girl), class (ix / x) and height (5 to 6 ft). 15 of students play cricket in leisure time. If a model is to be created for predicting who will play cricket during their leisure time, students playing cricket in their leisure time needs to be segregated based on highly significant input variable among all three as shown in Figure 16.

This is where Decision Tree helps, wherein it will segregate the students based on all values of three variable and identify the variable which creates the most apt homogeneous sets of students (which are heterogeneous to each other). For this example, as shown in the image below, it can be observed that the variable 'gender' is able to identify best homogeneous sets compared to the other two variables.

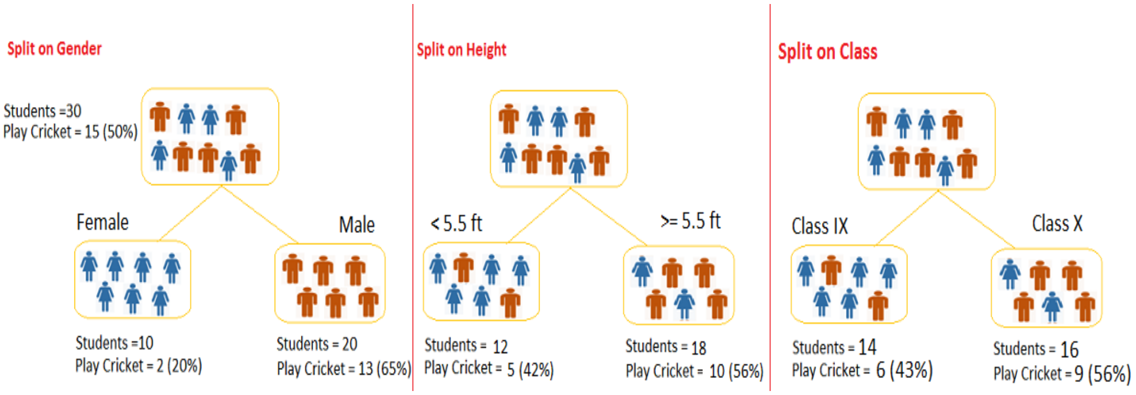


Figure 16: Example of decision tree

As mentioned above, decision tree identifies the most significant variable and its value lies in the method being able to estimate most appropriate homogeneous sets of population. For identifying the most significant variable and its value, Decision Tree employs various algorithms including Categorical Variable Decision Tree, which has categorical target variable (which is the interest of this analysis).

For the problem statement identified by NSDC, Categorical Variable Decision Tree was used. A Decision Tree which has a categorical target variable is referred to as a Categorical Variable Decision Tree. For understanding the placement scenario, the target variable was "candidate will be placed or not" i.e. YES or NO.

3.2. Random Forest method[13]

Random Forest is a versatile machine learning method capable of performing both regression and classification tasks. It also undertakes dimensional reduction methods, treats missing values, outlier values and other essential steps of data exploration. It is a type of ensemble learning method, where a group of weak models combine to form a powerful model. In this method, multiple trees are grown as opposed to a

[12] Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics, 21(3), 660-674.

[13] Liaw, A., & Wiener, M. (2002). Classification and regression by Random Forest. R news, 2(3), 18-22.; Parvin, H., MirnabiBaboli, M., & Alinejad-Rokny, H. (2015). Proposing a classifier ensemble framework based on classifier selection and decision tree. Engineering Applications of Artificial Intelligence, 37, 34-42.; Decision Tree vs. Random Forest – Which Algorithm Should you Use: https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/

single tree in CART model[14]. To classify a new object based on attributes, each tree gives a classification and the tree "votes" for that class. The Forest chooses the classification having the most votes (over all the trees in the Forest) and in case of regression, it takes the average of outputs by different trees. It works in the following manner wherein each tree is planted and grown as follows:

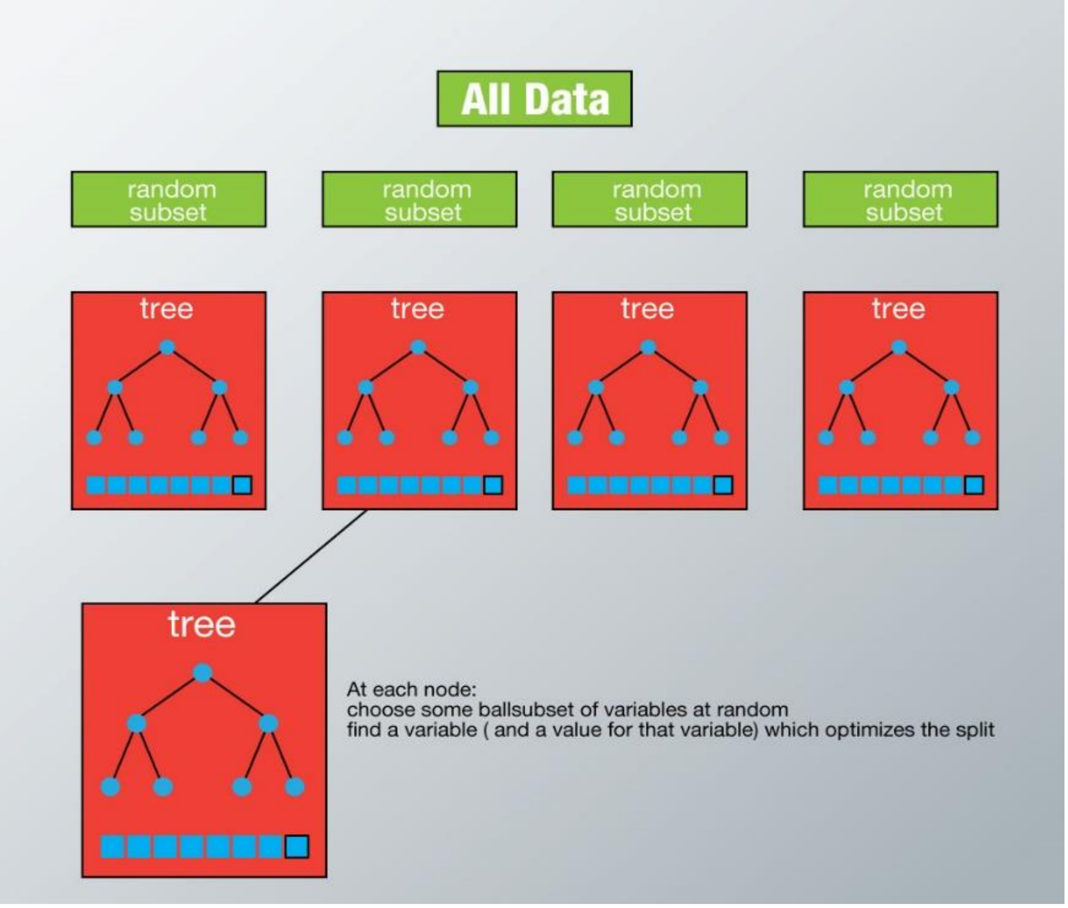

All Data
random subset
random subset
random subset
random subset
tree
tree
tree
tree
tree
At each node:
choose some ballsubset of variables at random
find a variable (and a value for that variable) which optimizes the split

Figure 17: Example of Random forest

Let us assume the number of cases in the training set is N. Then, sample of these N cases is taken at random but with replacement. This sample will be the training set for growing the tree.

If there are M input variables, a number m<M is specified such that at each node, m variables are selected at random out of the M. The best split on these m is used to split the node. The value of m is held constant while we grow the forest.

Each tree is grown to the largest possible extent and there is no pruning. Prediction of new data is done by aggregating the predictions of the n trees (i.e., majority votes for classification)

[14] The CART or Classification & Regression Trees algorithm is structured as a sequence of questions, the answers to which determine what the next question if any should be. The result of these questions is a tree-like structure where the ends are terminal nodes at which point there are no more questions. The main elements of CART are:

- Rules for splitting data at a node based on the value of one variable;
- Stopping rules for deciding when a branch is terminal and can be split no more;
- Finally, a prediction for the target variable in each terminal node.

3.3. Extra Tree Classifier[15]

Extra Trees Classifier is similar to Random Forest but with 2 key differences.
Considering a scenario wherein multiple decision trees are being built in the process, which would entail the requirement for multiple datasets. A best practice is not to train the decision trees on the complete dataset only on fraction of data (around 80 percent) for each tree. In a Random Forest, we draw observations with replacement wherein we can have repetition of observations in a random forest. While in an Extra Tree Classifier, observations are drawn without replacement, so there will not be any repetition of observations unlike in Random Forest model. The difference lies is the process of converting a non-homogeneous parent node into 2 homogeneous child nodes (best possible cases). In Random Forest, it selects the best split to convert the parent into the two most homogeneous child nodes. In an Extra Tree Classifier, it selects a random split to divide the parent node into two random child nodes. Summary of algorithms is depicted in Table 13.

Table 13: Summary of tree-based algorithms

	Decision Tree	Random Forest	Extra Trees
Number of trees	1	many	many
Number of features considered for split at each decision node	All Feature	Random subset of feature	Random subset of feature
Bootstrapping (drawing sample without replacement)	Not applied	No	Yes/No
How split is made	Best Split	Best Split	Random Split

3.4. XGBoost Classifier[16]

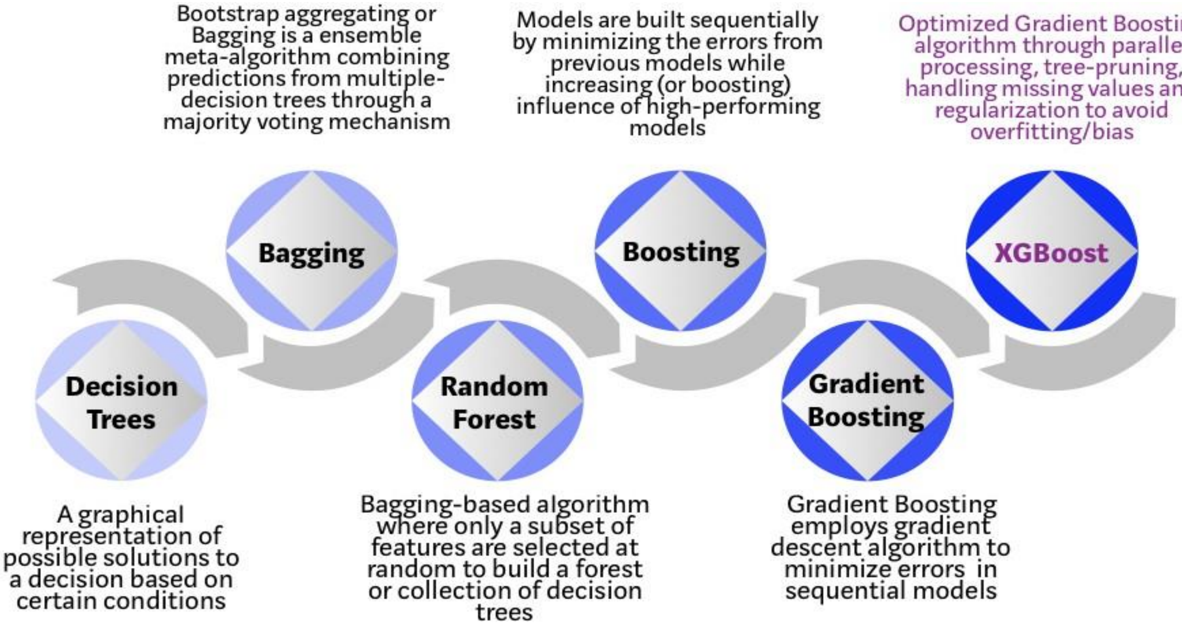
Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made. A popular example is the AdaBoost algorithm that weights data points that are hard to predict. Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models. This approach supports both regression and classification predictive modelling problems. XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now. Please see the figure 18 for the evolution of tree-based algorithms over the years.

Algorithmic Enhancements in XGBoost:

- *Regularization*: It penalizes more complex models through both LASSO (L1) and Ridge (L2) regularization to prevent overfitting.
- *Sparsity Awareness*: XGBoost naturally admits sparse features for inputs by automatically 'learning' best missing value depending on training loss and handles different types of sparsity patterns in the data more efficiently.
- *Weighted Quantile Sketch***:** XGBoost employs the distributed weighted Quantile Sketch algorithm to effectively find the optimal split points among weighted datasets.
- *Cross-validation*: The algorithm comes with built-in cross-validation method at each iteration, taking away the need to explicitly program this search and to specify the exact number of boosting iterations required in a single run.

[15] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine learning, 63(1), 3-42.; An Intuitive Explanation of Random Forest and Extra Trees Classifiers: https://towardsdatascience.com/an-intuitive-explanation-of-random-forest-and-extra-trees-classifiers-8507ac21d54b

[16] https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d Date of access- 20 February 2021



A graphical representation of possible solutions to a decision based on certain conditions

Bootstrap aggregating or Bagging is a ensemble meta-algorithm combining predictions from multiple-decision trees through a majority voting mechanism

Bagging-based algorithm where only a subset of features are selected at random to build a forest or collection of decision trees

Models are built sequentially by minimizing the errors from previous models while increasing (or boosting) influence of high-performing models

Gradient Boosting employs gradient descent algorithm to minimize errors in sequential models

Optimized Gradient Boosting algorithm through parallel processing, tree-pruning, handling missing values and regularization to avoid overfitting/bias

Figure 18: Evolution of XGBoost Algorithm from Decision Tree

3.5. Correlation coefficient[17]

A correlation coefficient measures the extent to which two variables tend to change together. The coefficient describes both the strength and the direction of the relationship. Minitab offers two different correlation analyses:

<u>Pearson product moment correlation</u>

The Pearson correlation evaluates the linear relationship between two continuous variables. A relationship is linear when a change in one variable is associated with a proportional change in the other variable. For example, you might use a Pearson correlation to evaluate whether increases in temperature at your production facility are associated with decreasing thickness of your chocolate coating.

<u>Spearman rank-order correlation</u>

The Spearman correlation evaluates the monotonic relationship between two continuous or ordinal variables. In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate. The Spearman correlation coefficient is based on the ranked values for each variable rather than the raw data.

Spearman correlation is often used to evaluate relationships involving ordinal variables. For example, you might use a Spearman correlation to evaluate whether the order in which employees complete a test exercise is related to the number of months they have been employed.

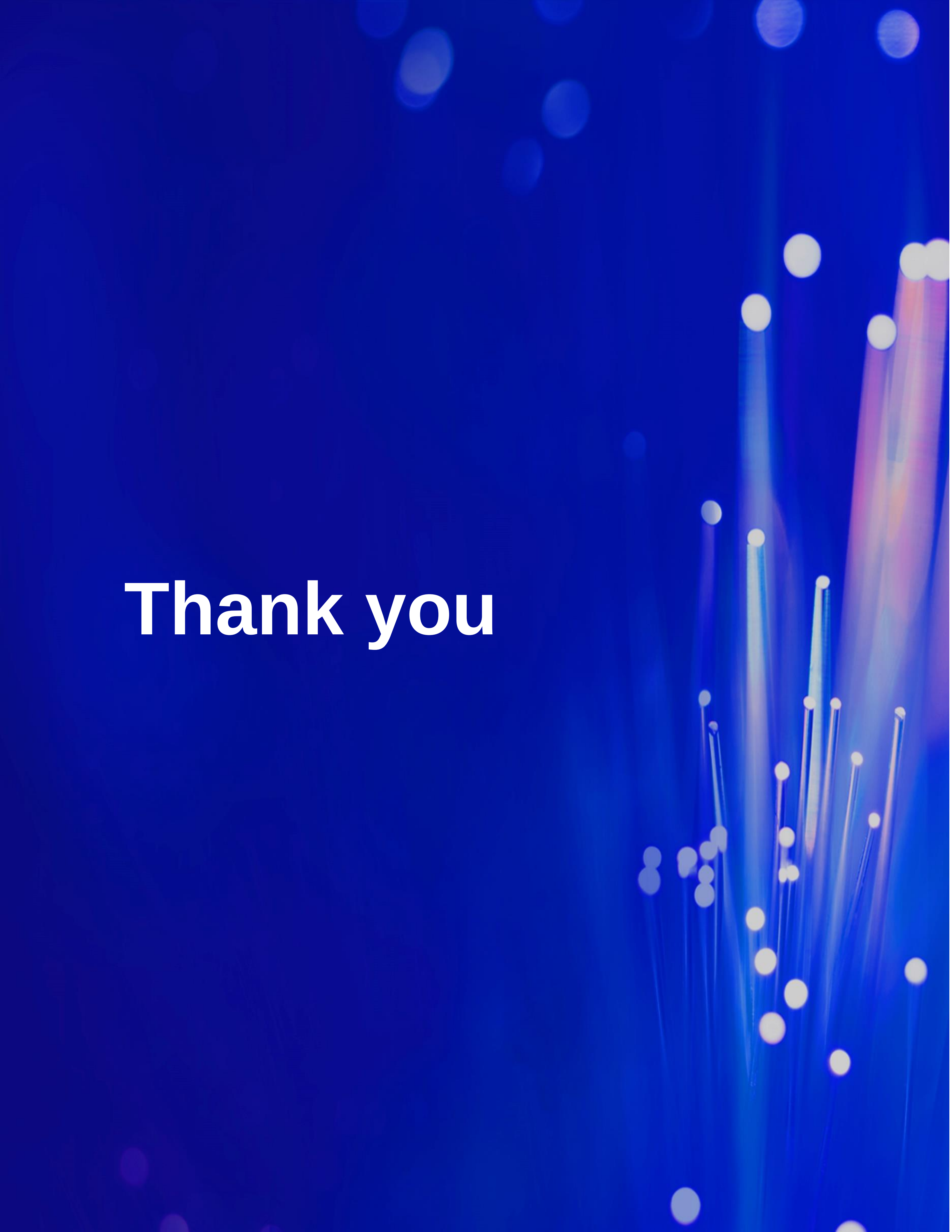
Correlation coefficients only measure linear (Pearson) or monotonic (Spearman) relationships.

[17] https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/supporting-topics/basics/a-comparison-of-the-pearson-and-spearman-correlation-methods/ Date of access- 20 February 2021

Accuracy of predictive modelling

Accuracy is the ratio of number of correct predictions to the total number of input samples.
The accuracy of a machine learning classification algorithm is one way to measure how often the algorithm classifies a data point correctly. Accuracy is the number of correctly predicted data points out of all the data points. More formally, it is defined as the number of true positives(TP) and true negatives(TN) divided by the number of true positives(TP), true negatives(TN), false positives(FP), and false negatives(FN). A true positive or true negative is a data point that the algorithm correctly classified as true or false, respectively. A false positive or false negative, on the other hand, is a data point that the algorithm incorrectly classified.

Accuracy = (TN + TP)/(TN+TP+FN+FP) = (Number of correct assessments)/Number of all assessments)



Thank you