ML IN RECRUITMENT: PREDICTING JOB POST ENGAGEMENT THROUGH CLASSIFICATION

Tony Xu 261173499
Vinay Govias 261143063
Xinran Yu 260922576
Yichen Yu 261136583
Oyundari Batbayar 260713572

LIST OF CONTENTS

EXECUTIVE SUMMARY

OBJECTIVE AND METHODOLOGY

JOB POSTING POPULARITY?

- **Leveraged supervised learning** to predict the likelihood of a job posting receiving an above-average application rate
- **Analyzed a comprehensive dataset** from LinkedIn, focusing on diverse job-related features.
- **Implemented data cleaning, feature engineering, and extraction** of key insights from **textual job descriptions.**
- **Alignment with Recruitment Dynamics** Ensured that models align with current recruitment trends and business objectives, including aspects like work types.

MODELING APPROACH

PRECISION ACCURACY RECALL?

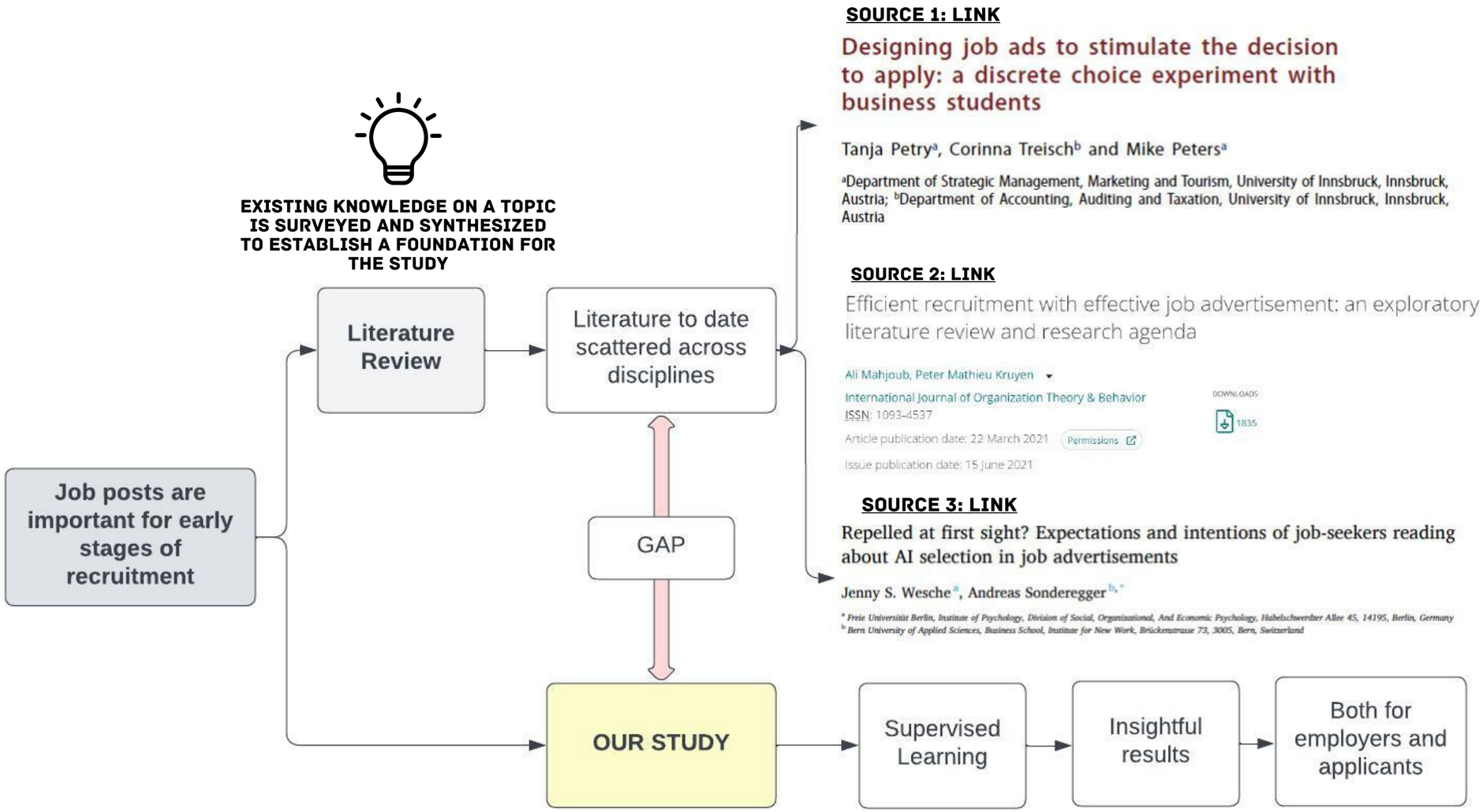
- **Diverse Model Selection -** Employed a range of models (MLP, Logistic Regression, Gradient Boost, KNN, Random Forest) to ensure robustness and versatility in predictions.
- **Feature Engineering -** Rigorously processed and transformed data to create meaningful, predictive features from complex job-related information.
- **Optimization and Tuning -** Employed GridSearchCV for hyperparameter tuning, ensuring optimal model performance.
- **Incorporation of Feature Importance Scoring -** Integrated feature importance scores to identify and prioritize the most impactful variables driving recruitment success.

BUSINESS IMPLICATIONS

MONEY AND/OR WELFARE?

- **Identifying Key Attractors -** Determine specific job features that significantly attract applicants.
- **Optimized Job Descriptions -** Tailor job descriptions to emphasize elements most valued by candidates, increasing application rates.
- **Resource Allocation -** Direct recruitment resources and efforts towards aspects with the highest impact on application rates.
- **Market Alignment -** Align job offerings with current market trends and candidate preferences, enhancing competitive positioning.

DESCRIPTION OF THE ANALYTICS PROBLEM



SOURCE 1: LINK

Designing job ads to stimulate the decision to apply: a discrete choice experiment with business students

Tanja Petry[a], Corinna Treisch[b] and Mike Peters[a]

[a]Department of Strategic Management, Marketing and Tourism, University of Innsbruck, Innsbruck, Austria; [b]Department of Accounting, Auditing and Taxation, University of Innsbruck, Innsbruck, Austria

SOURCE 2: LINK

Efficient recruitment with effective job advertisement: an exploratory literature review and research agenda

Ali Mahjoub, Peter Mathieu Kruyen

International Journal of Organization Theory & Behavior

ISSN: 1093-4537

Article publication date: 22 March 2021 Permissions

Issue publication date: 15 June 2021

DOWNLOADS 1835

SOURCE 3: LINK

Repelled at first sight? Expectations and intentions of job-seekers reading about AI selection in job advertisements

Jenny S. Wesche[a], Andreas Sonderegger[b,*]

[a] Freie Universität Berlin, Institute of Psychology, Division of Social, Organizational, And Economic Psychology, Habelschwerdter Allee 45, 14195, Berlin, Germany
[b] Bern University of Applied Sciences, Business School, Institute for New Work, Brückenstrasse 73, 3005, Bern, Switzerland

DESCRIPTION OF CONCEPTUAL MAP

- Our literature review underscores the importance of job posts in **attracting suitable candidates at the initial stages of recruitment.**
- By identifying a gap in the literature, where the specifics at the job posting level are not sufficiently explored, the study aims to offer a **detailed examination of how discrete elements within job advertisements** influence the decision-making process of prospective applicants.
- The research is poised to use **supervised learning** methodologies, suggesting an analytical approach that could decipher complex patterns and **provide nuanced insights**.
- The outcomes of the study are anticipated to be **valuable for both employers**, in crafting effective job listings, and **applicants**, in discerning job features that align with their expectations and career aspirations.

DATA DESCRIPTION

Source:
https://www.kaggle.com/datasets/shashankshukla123123/linkedin-job-data

Scraped from LinkedIn India. Raw data contains 7927 rows and 15 columns

Includes various tech job roles located in India

Contains information on job location, description, and other related parameters

Contains duplicated entries; messy data; and other issues that needs to be resolved.

TABLE 1. DATA DESCRIPTION

Field	Description
job_ID	Unique identifier for each job posting.
job	The title of the job posting.
location	The location of the job posting.
company_id	The unique identifier for the company offering the job.
company_name	The name of the company offering the job.
work_type	The type of work offered (e.g. full-time, part-time, etc.).
full_time_remote	Indicates if the job is a full-time remote position.
no_of_employ	The number of employees at the company offering the job.
no_of_application	The number of applications received for the job.
posted_day_ago	The number of days ago the job was posted.
alumni	Indicates if the job posting is for alumni of a certain organization.
Hiring_person	The name of the person responsible for hiring for the job.
linkedin_followers	The number of LinkedIn followers of the hiring person.
hiring_person_link	A link to the LinkedIn profile of the hiring person.
job_details	Detailed information about the job, including responsibilities and requirements.

DATA PREPROCESSING - OVERVIEW

Skills Extract

Use tf–idf score to filter the top 2000 words, then select a list of technical skills (~200) based on prior text and build dummy variables.

Target Variable

Convert application number and job posted time into the speed of application number per hour and convert it to above/below average.

Dummy variable

Build dummy variables based on job location, company size, education level, and experience level (some features extracted from job descriptions).

Company benefits

Extract benefits related words such as growth potential, company culture and benefits related from a prebuilt list. Uses wordnet synsets to look for synonyms as well of the words on the list, to make the extraction more robust.

Industry and # of employees

Extract industry information, assign macro industries, and create ranks based on the number of employees.

Hiring Person Info

1 or 0 based on if Hiring persons details are present in the job post

Remote Work

Based on if the job allows for hybrid or remote work.

LinkedIn follower column

Removed text and commas and extracted the number of followers for each company on LinkedIn.

Data Cleaning

Performed data cleaning procedure, such as standardize format of the features, removing duplicates, etc.

Class Imbalance Correction

Used SMOTE to improve the distribution of average and above average application rate instances in the dataset.

Feature Selection

Performed LASSO and Random Forest Feature Selection and compared the results, to remove redundant or irrelevant variables, keeping only the important predictors.

DATA PREPROCESSING

TEXT ANALYTICS COMPANY BENEFITS EXTRACTION FROM JOB POST

STEPS	NATURAL LANGUAGE TOOLKIT (NLTK) LIBRARY	WORDNET DATA DOWNLOAD	SYNONYM EXTRACTION	CATEGORIES AND KEYWORDS	BUILDING A SYNONYM DICTIONARY	CREATING DUMMY VARIABLES FOR CATEGORIES
DESCRIPTION	The code used the **NLTK** library, a popular Python package for natural language processing .	The function **get_synonyms** extracted synonyms for a given word using WordNet. For each 'synset' that a word belonged to, the function found all 'lemmas' and added their base forms to a set of synonyms.	Downloaded **WordNet** data, which was a large lexical database of English. Grouped English words into sets of synonyms called synsets.	For each category, expanded **the list of keywords by adding their synonyms.**	Defined several categories relevant to job descriptions (like **'Benefits-Related', 'Company-Culture',** etc.) and **associated a list of keywords with each category.**	**Created a new column in the DataFrame and populated it with 1 or 0,** indicating whether any of the keywords or their synonyms were present in the job description.

DATA PREPROCESSING

FEATURE SELECTION USING RANDOM FOREST AND LASSO

RANDOM FOREST FOR FEATURE SELECTION

- Utilizes an ensemble of decision trees.
- Features are ranked by importance for selection.

LASSO REGRESSION FOR FEATURE SELECTION

- Applies linear regression with a regularization term.
- Shrinks coefficients of less important features to zero. This resulted in sparse models, focusing on key features.

Using these two methods, we built a list of variables (~100) that we decided to keep in the final Models.

DATA PREPROCESSING
SKILLS EXTRACTION

TF-IDF

Term Frequency-Inverse Document Frequency is a numerical statistic used to reflect **how important a word is to a document** in a collection or corpus.

It's a product of two statistics:

Term Frequency (TF), which measures how frequently a term occurs in a document

Inverse Document Frequency (IDF), which gauges the rarity of the term across the corpus.

TF-IDF increases with the number of times a word appears in the document but is offset by the frequency of the word in the corpus, helping to control for the fact that **some words were generally more common than others.**

TF-IDF was used to extract the most commonly occuring technical skills from the dataset and these words were made into dummy variables.

DATA PREPROCESSING - SMOTE (SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE)

ADDRESSES CLASS IMBALANCE

ENHANCES THE MINORITY CLASS BY GENERATING SYNTHETIC DATA POINTS.

PROCESS

- **Identifies the minority class** in the dataset.
- For each minority class instance, finds its **k-nearest neighbors in the feature space.**
- **Generates new synthetic instances** along the line joining each instance and its neighbors.

ADVANTAGES

- **Creates new, synthetic samples** rather than duplicating, enhancing data diversity.
- **Helps reduce model bias** towards the majority class, improving classification performance.
- **Can prevent overfitting**, a common issue with simple oversampling techniques.

CONSIDERATIONS

- **Effectiveness can vary** with the choice of 'k' (number of neighbors).
- **Best applied only to training data** to prevent data leakage.
- **May require feature selection** or reduction in high-dimensional datasets.

EXPLORATORY DATA ANALYSIS (EDA)

FIGURE 1. JOB DISTRIBUTION BY LOCATION

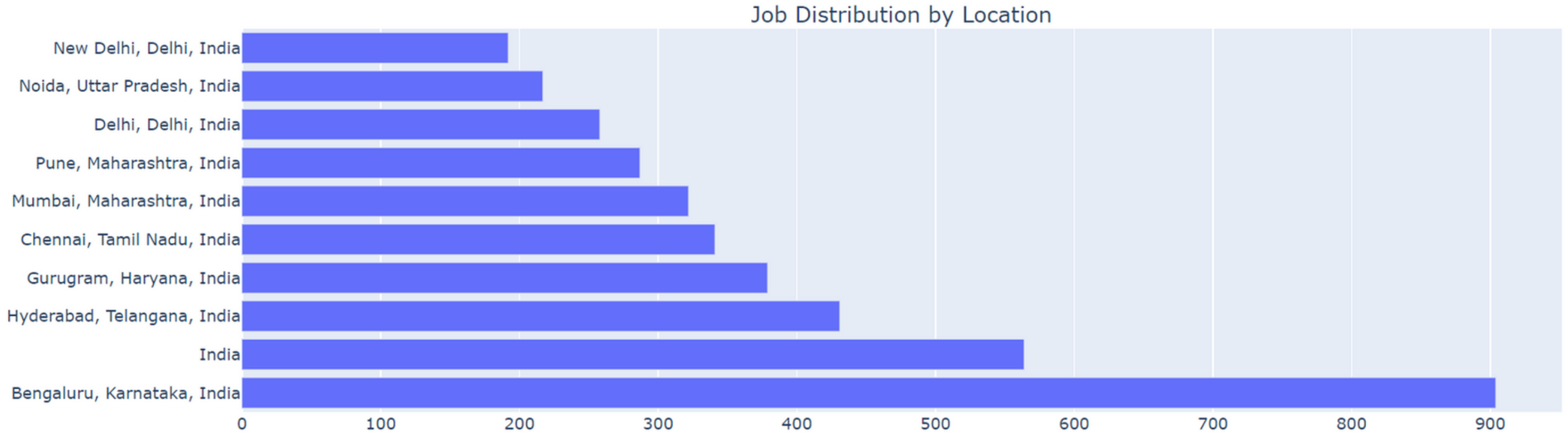
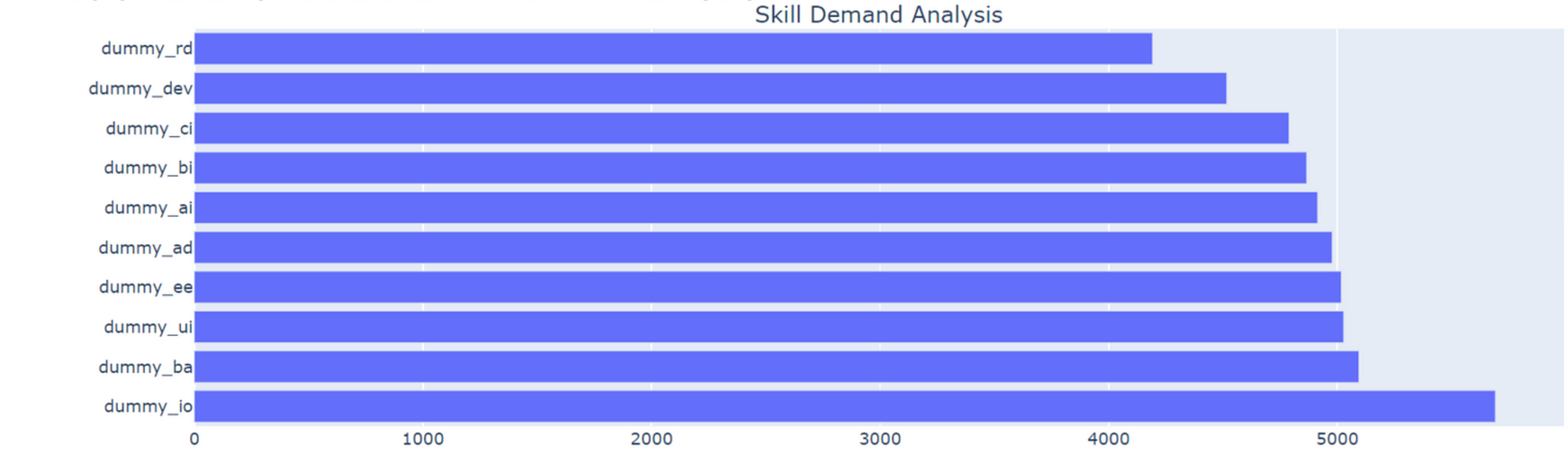


FIGURE 2. SKILL DEMAND ANALYSIS

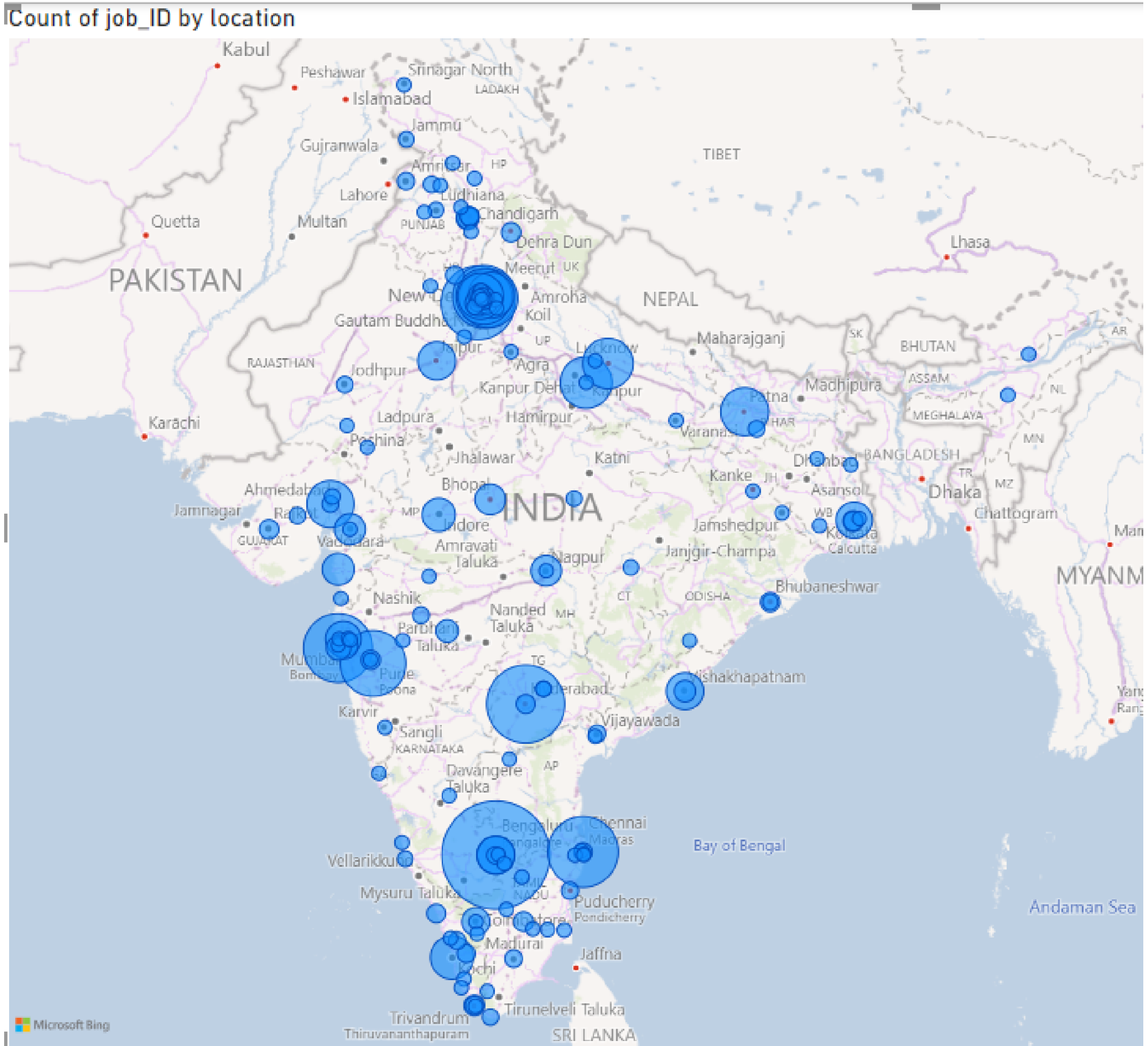


KEY FINDINGS

- **VARIETY OF JOB SKILLS** ARE DEMANDED IN THE DATASET.
- **JOBS LOCATED MAINLY IN THE TECH CLUSTER** OF INDIA (NEXT MAP).

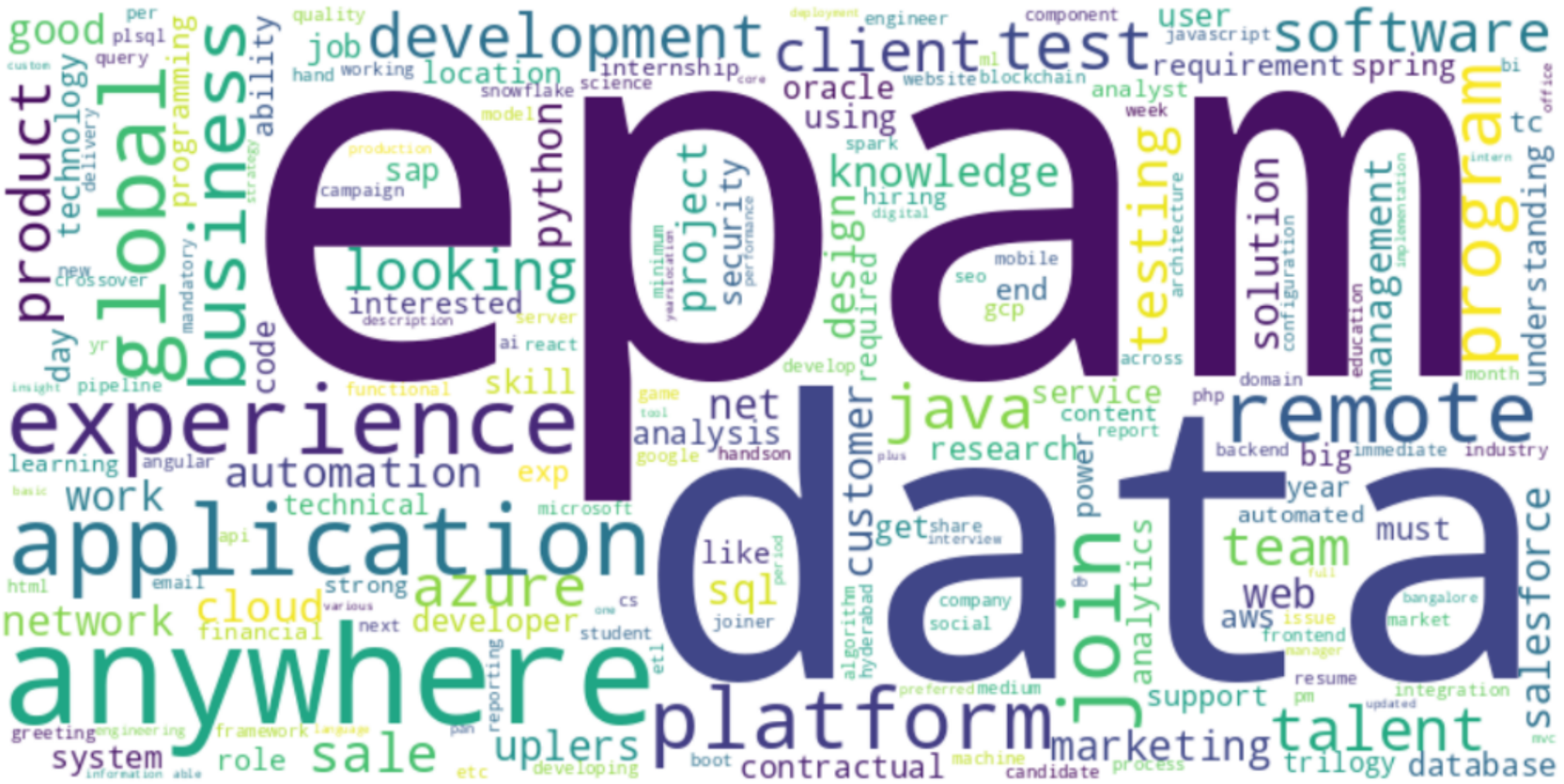
EDA

FIGURE 3. NUMBER OF JOB POSTS BY LOCATION



EDA

FIGURE 4.



- **KEY WORDS INCLUDE DATA, EPAM (**A COMPANY NAME), AND ANYWHERE, HIGHLIGHTING THE CURRENT TREND IN INDIA’S JOB MARKET.

EDA FINDINGS

- Majority of job opportunities concentrated in certain urban areas. Located primarily in India’s tech cluster.
- A significant distribution between remote and on-site jobs, a clear preference of one type over the other.
- Certain skills are in high demand, indicated by high frequency in job description
- High diversity in job titles, indicating the range of specialties in demand.

FIGURE 5.

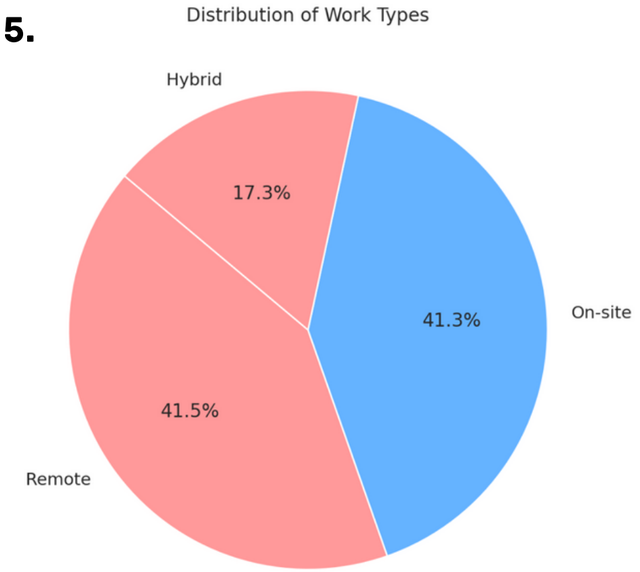
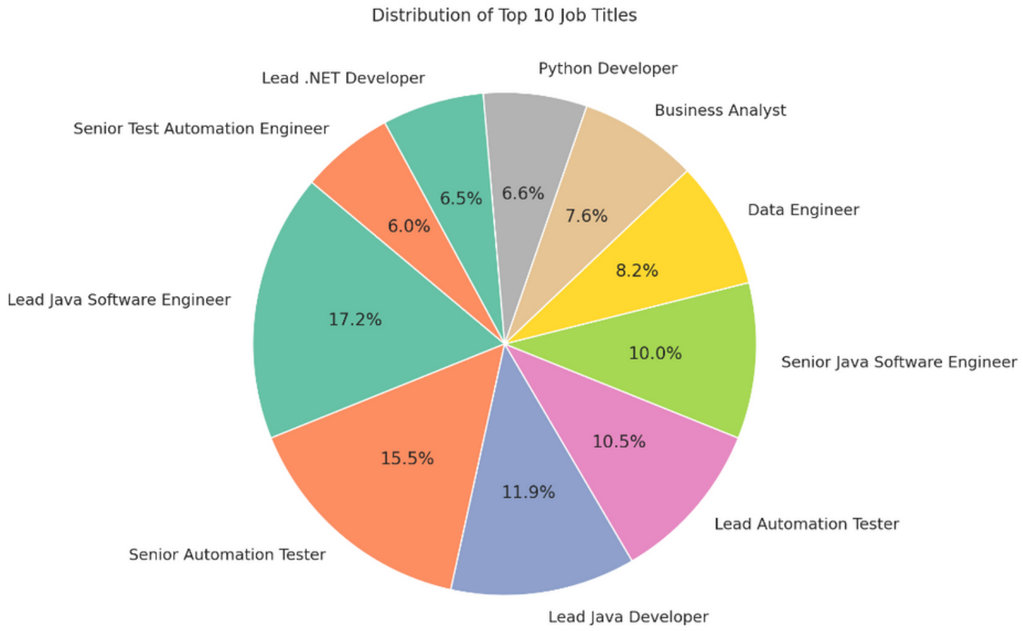
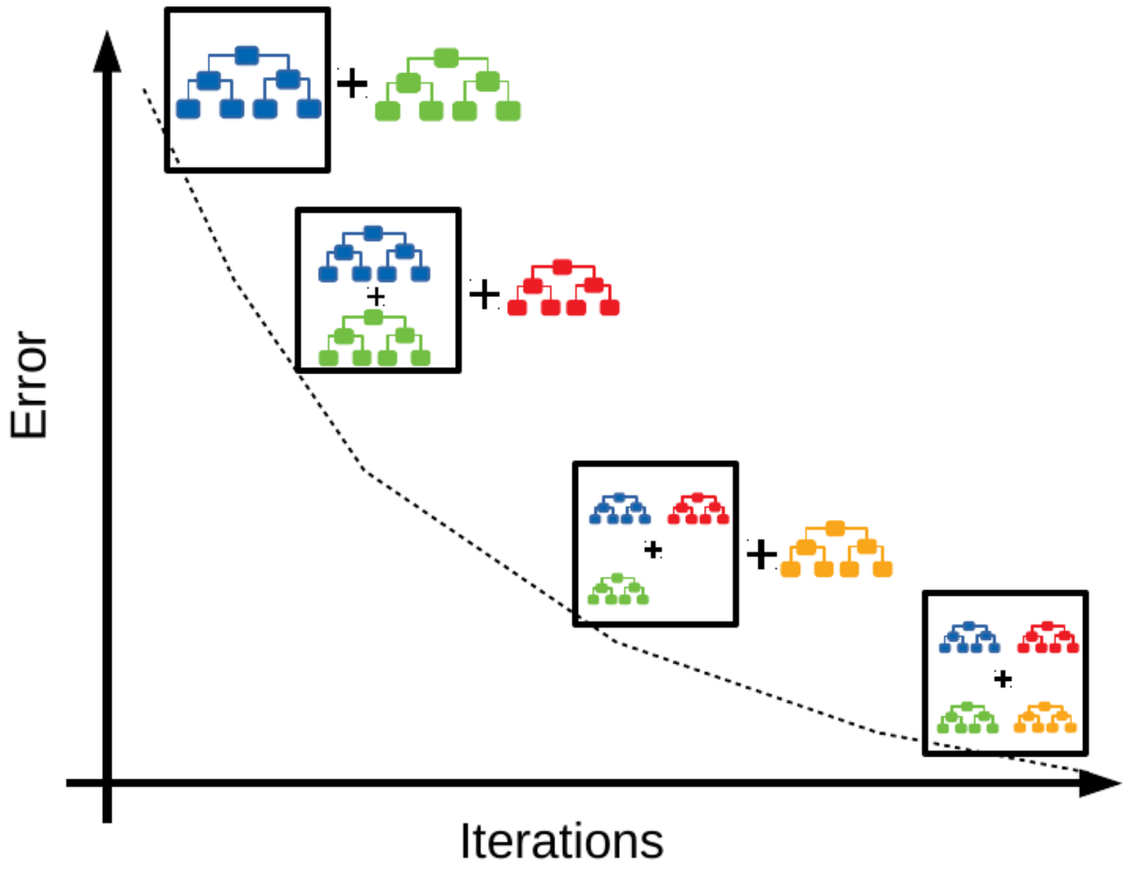


FIGURE 6.



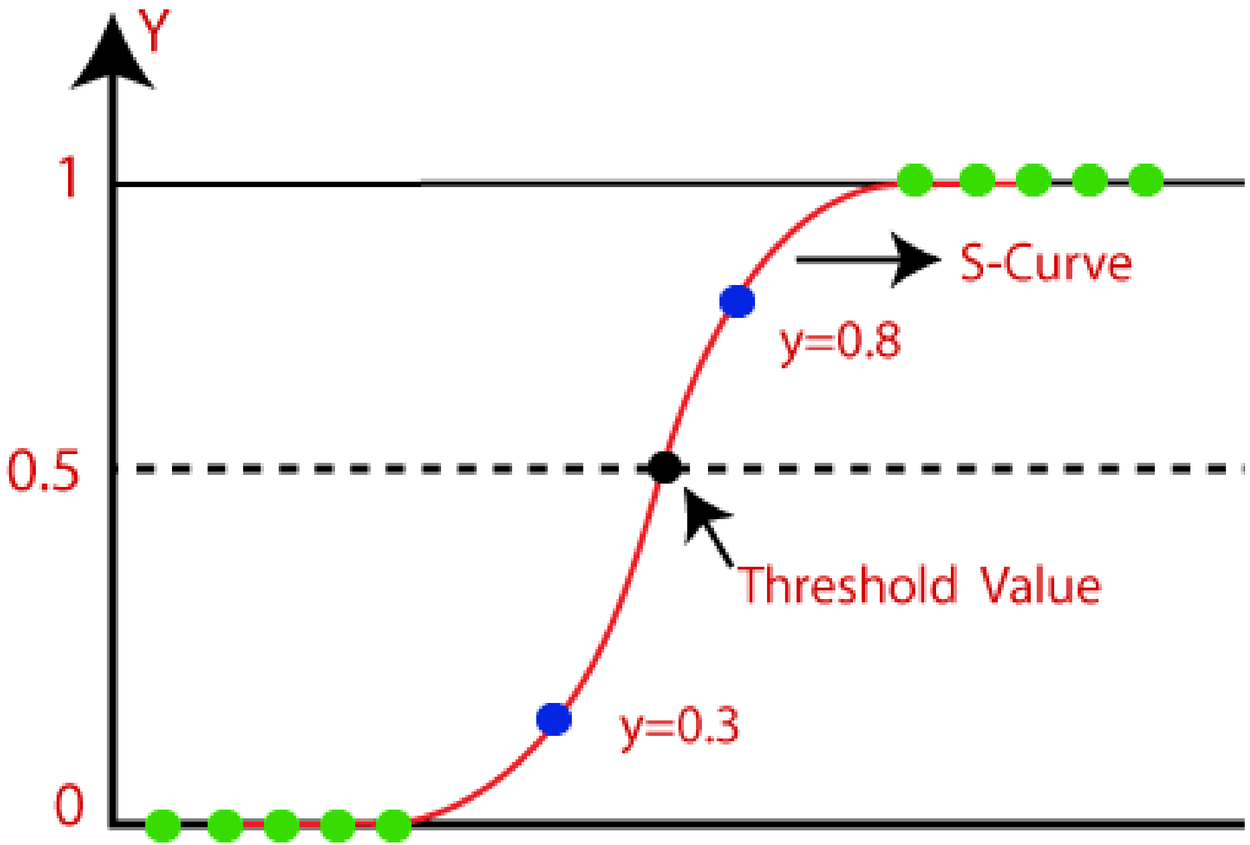
MODEL SELECTION



GRADIENT BOOSTING

- **A MACHINE LEARNING TECHNIQUE** FOR CLASSIFICATION THAT BUILDS AN ENSEMBLE OF WEAK PREDICTION MODELS,TYPICALLY DECISION TREES, IN A STAGE-WISE FASHION.
- **IT OPTIMIZES A COST FUNCTION OVER FUNCTION SPACE** BY SEQUENTIALLY ADDING WEAK LEARNERS TO MINIMIZE THE MODEL'S PREDICTIVE ERROR.

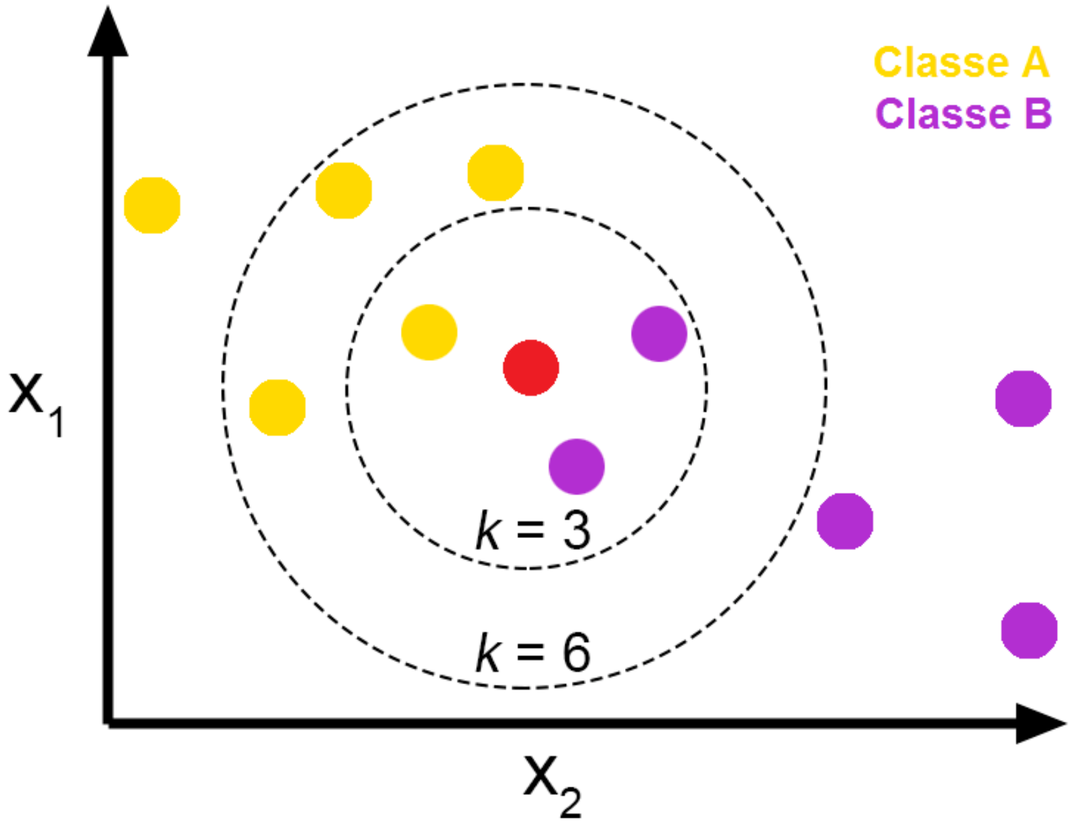
MODEL SELECTION



LOGISTIC REGRESSION

- **A STATISTICAL MODEL THAT PREDICTS THE PROBABILITY OF A BINARY OUTCOME** BY APPLYING THE LOGISTIC FUNCTION TO A LINEAR COMBINATION OF FEATURES.
- **IT IS OFTEN USED FOR BINARY CLASSIFICATION TASKS** AND PROVIDES A PROBABILISTIC FRAMEWORK THAT QUANTIFIES THE UNCERTAINTY OF PREDICTIONS.

MODEL SELECTION

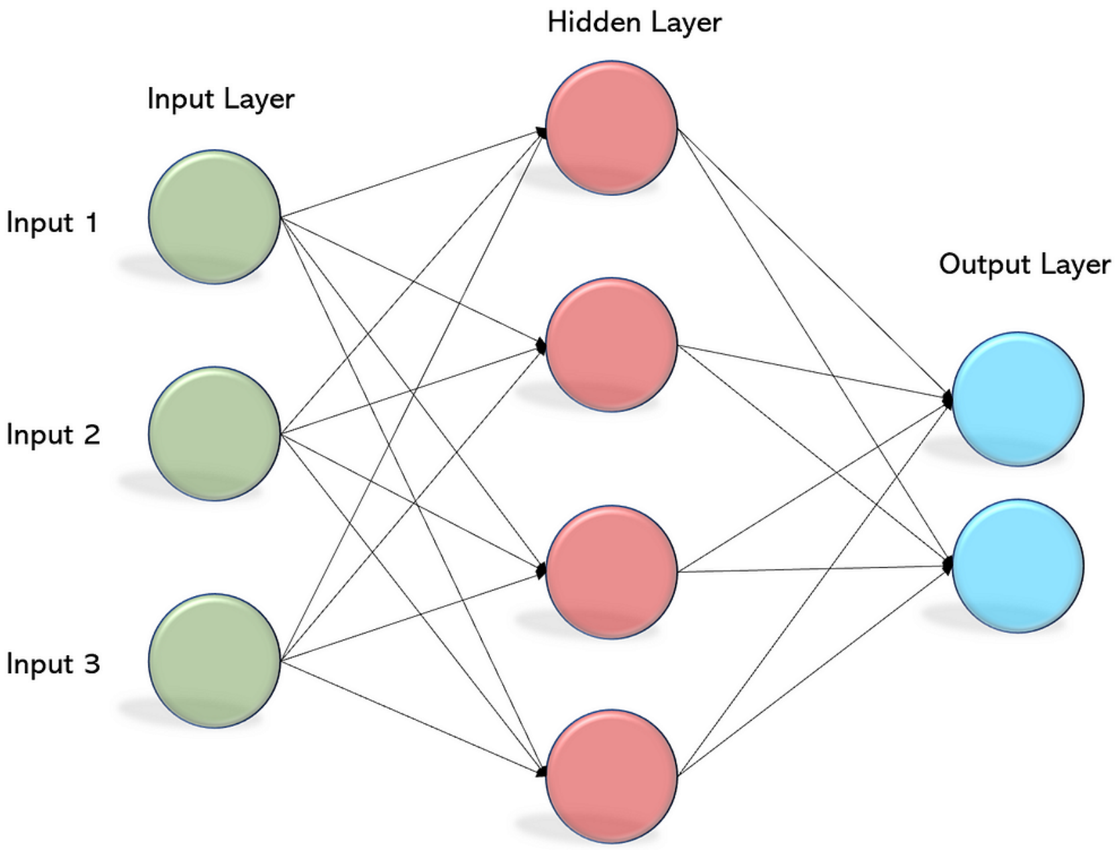


K NEAREST NEIGHBOURS

- **KNN IS A NON-PARAMETRIC, LAZY LEARNING ALGORITHM** THAT CLASSIFIES A DATA POINT BASED ON HOW ITS NEIGHBORS ARE CLASSIFIED.

- **IT ASSIGNS THE CLASS WHICH IS MOST FREQUENT** AMONG ITS K NEAREST NEIGHBORS, WHERE K IS A USER-DEFINED CONSTANT.

MODEL SELECTION



ARTIFICIAL NEURAL NETWORKS

- **ANN-MLP IS A CLASS OF FEEDFORWARD ARTIFICIAL NEURAL NETWORKS** THAT CONSISTS OF AT LEAST THREE LAYERS OF NODES: AN INPUT LAYER, HIDDEN LAYERS, AND AN OUTPUT LAYER. IT USES BACKPROPAGATION FOR TRAINING.

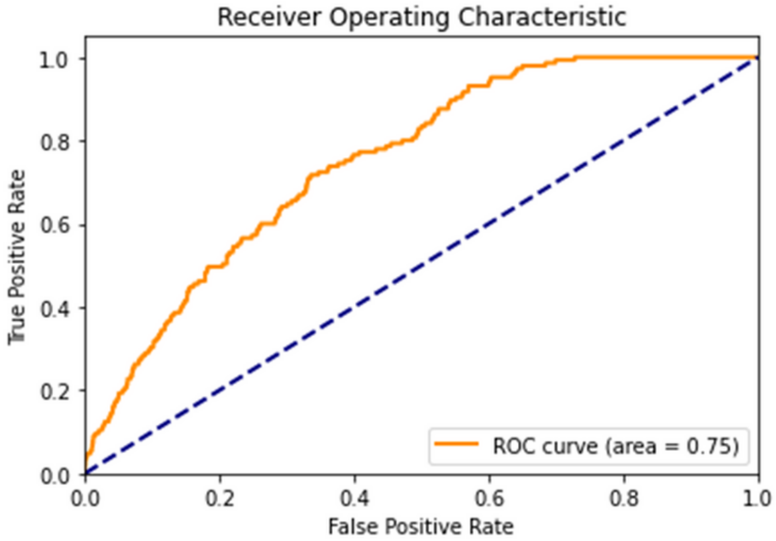
- **CAN MODEL COMPLEX NON-LINEAR RELATIONSHIPS** FOR CLASSIFICATION TASKS.

MODEL SELECTION

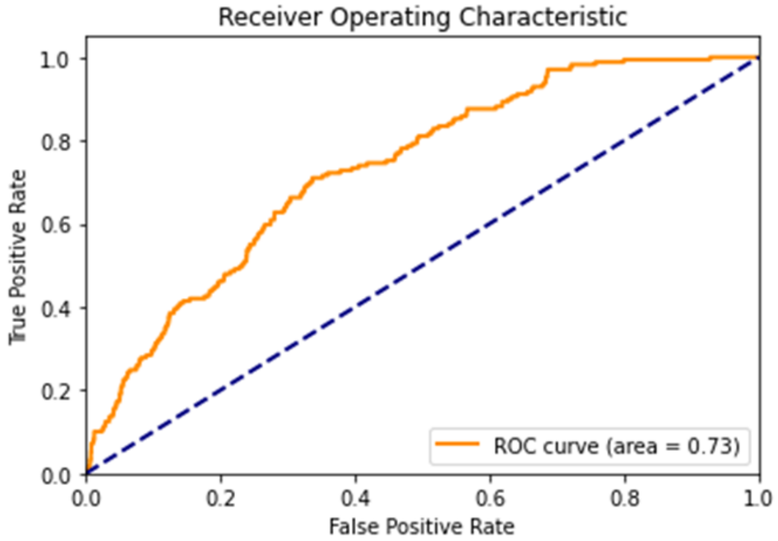
	Logistic Regression	KNN	ANN - MLP	Gradient Boosting (RF Feature Selection)	Random Forest
RECALL	0.20	0.05	0.22	0.24	0.41
PRECISION	0.28	0.1	0.22	0.29	0.20
ACCURACY	0.89	0.86	0.87	0.83	0.86

AFTER BUILDING AND COMPARING ALL THE MODELS, THE FOLLOWING SCORES ARE FOR CLASS 1 (GETTING ABOVE AVERAGE APPLICATION RATE) OF THE BEST MODEL OF EACH TYPE

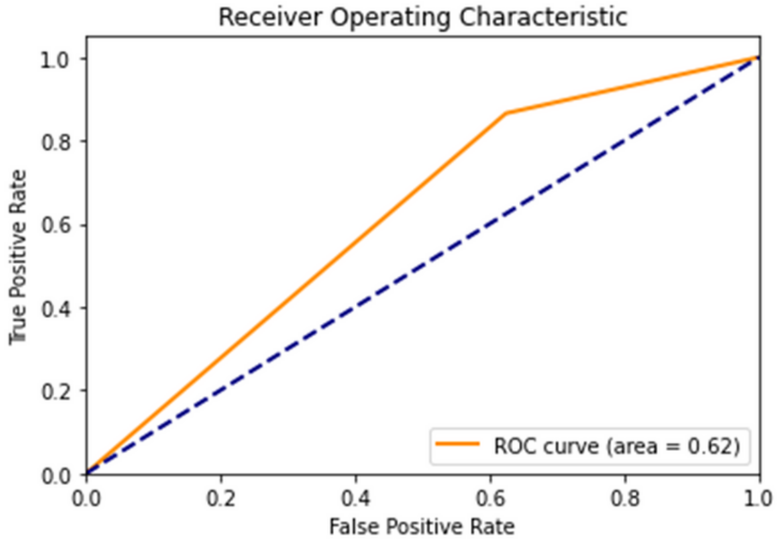
MODEL SELECTION - CONTD.



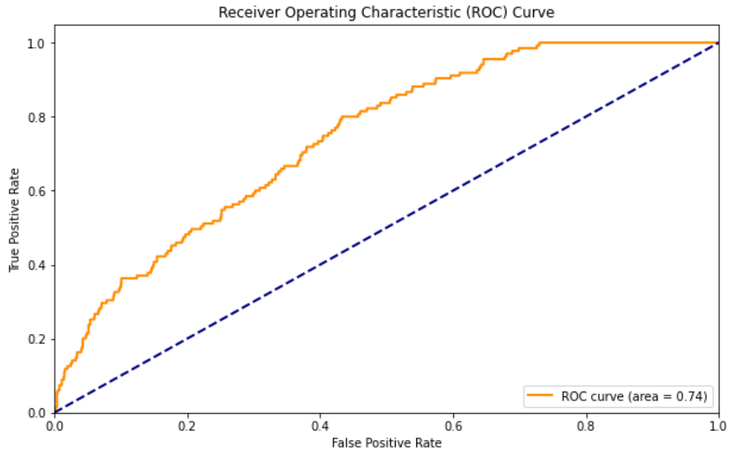
Random Forest AUC = 0.75



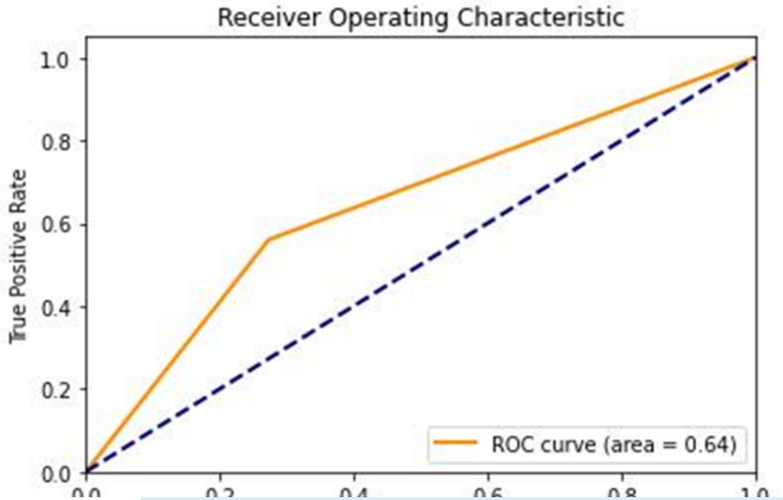
Gradient Boost AUC=0.73



KNN AUC=0.62

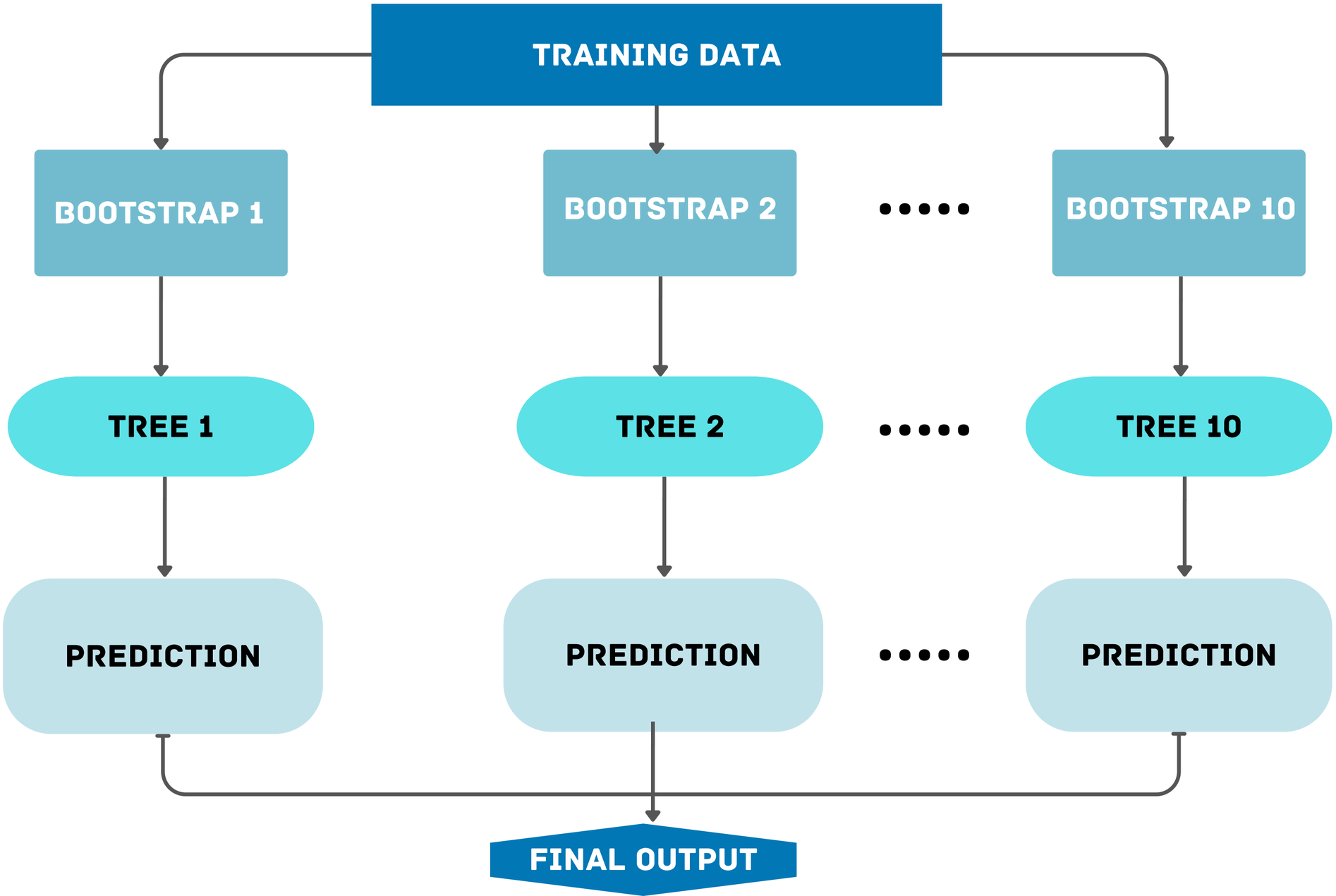


Logistic Regression AUC=0.74



ANN AUC=0.64

RANDOM FOREST - BAGGING



RANDOM FOREST - BAGGING

- **Random Forest is a non-linear model** made up of many decision trees. It's known for high accuracy, robustness, and interpretability compared to GB , ANN etc.
- **It does not have coefficients** like a regression model. Instead, it has feature importance.
- **Hyperparameter tuning is crucial** for optimizing the model's performance.
- **Random Forest inherently uses bagging**, creating trees on different samples and averaging their predictions.

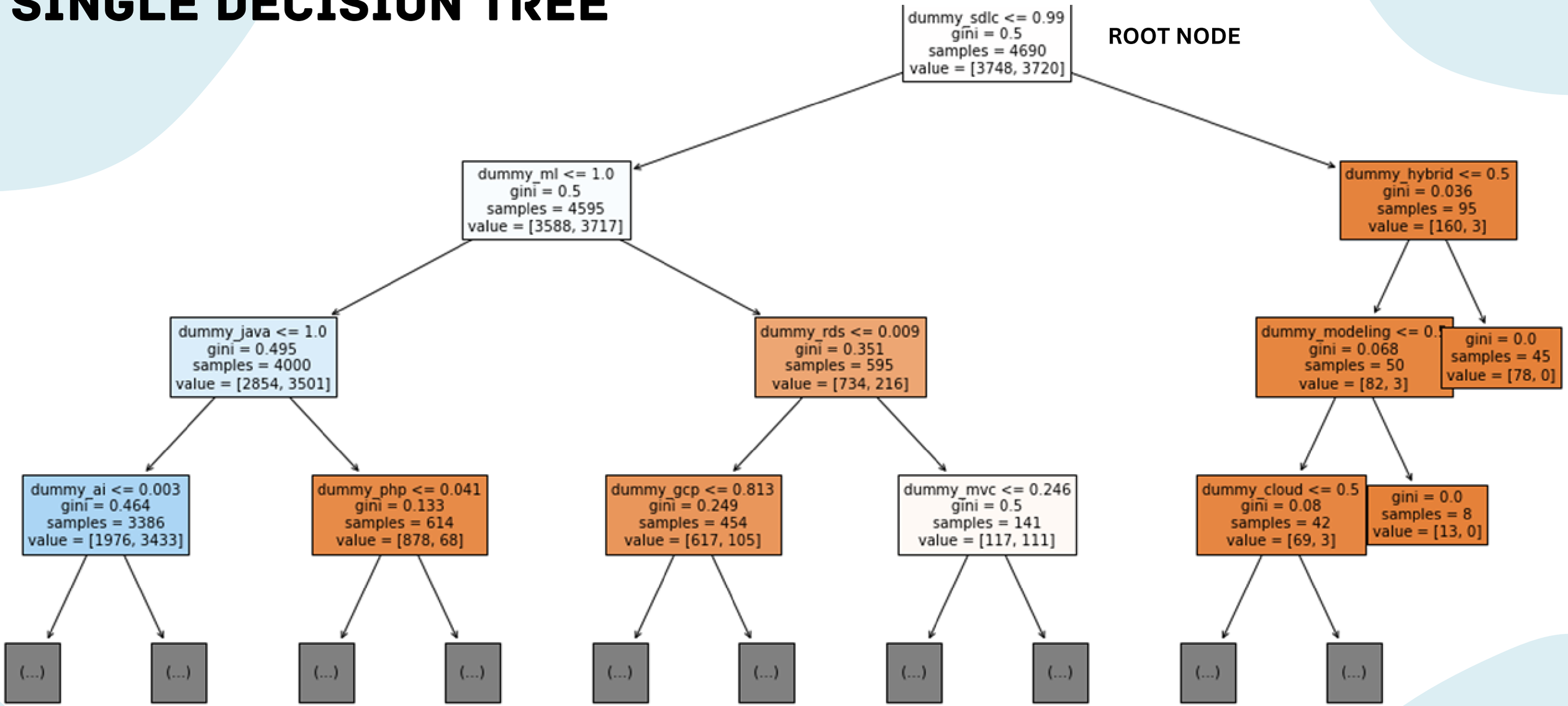
- **ROC curve is a plot** that shows the performance of a classification model at all classification thresholds.
- **AUC, or Area Under the Curve,** measures the entire two-dimensional area underneath the entire ROC curve, providing an aggregate measure of performance across all possible classification thresholds.
- **The closer the AUC is to 1,** the better the model's prediction capabilities.

RANDOM FOREST - HYPER PARAMETER TUNING

Tuning applied using GridSearchCV

- **n_estimators:** The number of trees
- **max_depth:** The maximum depth of the trees
- **min_samples_split:** The minimum number of samples required to split an internal node (decision node) ,
- **min_samples_leaf:** The minimum number of samples required to be at a leaf node(terminal node)
- **max_features:** The number of features to consider when looking for the best split

RANDOM FOREST - ANALYZING A SINGLE DECISION TREE



RANDOM FOREST - ANALYZING A SINGLE DECISION TREE

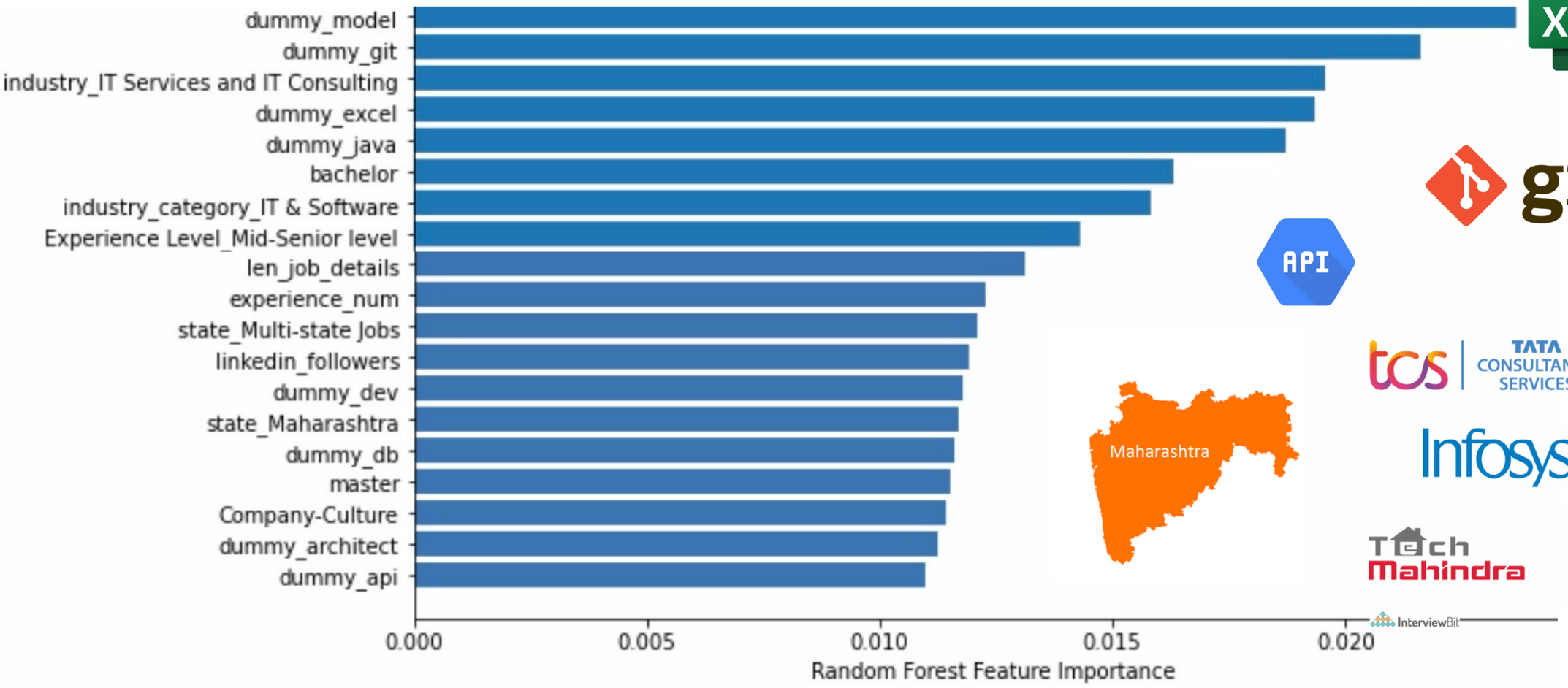
DECISION TREE VISUALIZATION

- **A single tree from the Random Forest** gives us insight into the model's decision-making process.
- **Nodes represent features that split the data,** with the Gini index measuring each split's purity.
- **The branches represent the binary outcomes** of each decision, culminating in leaves that make the final prediction.

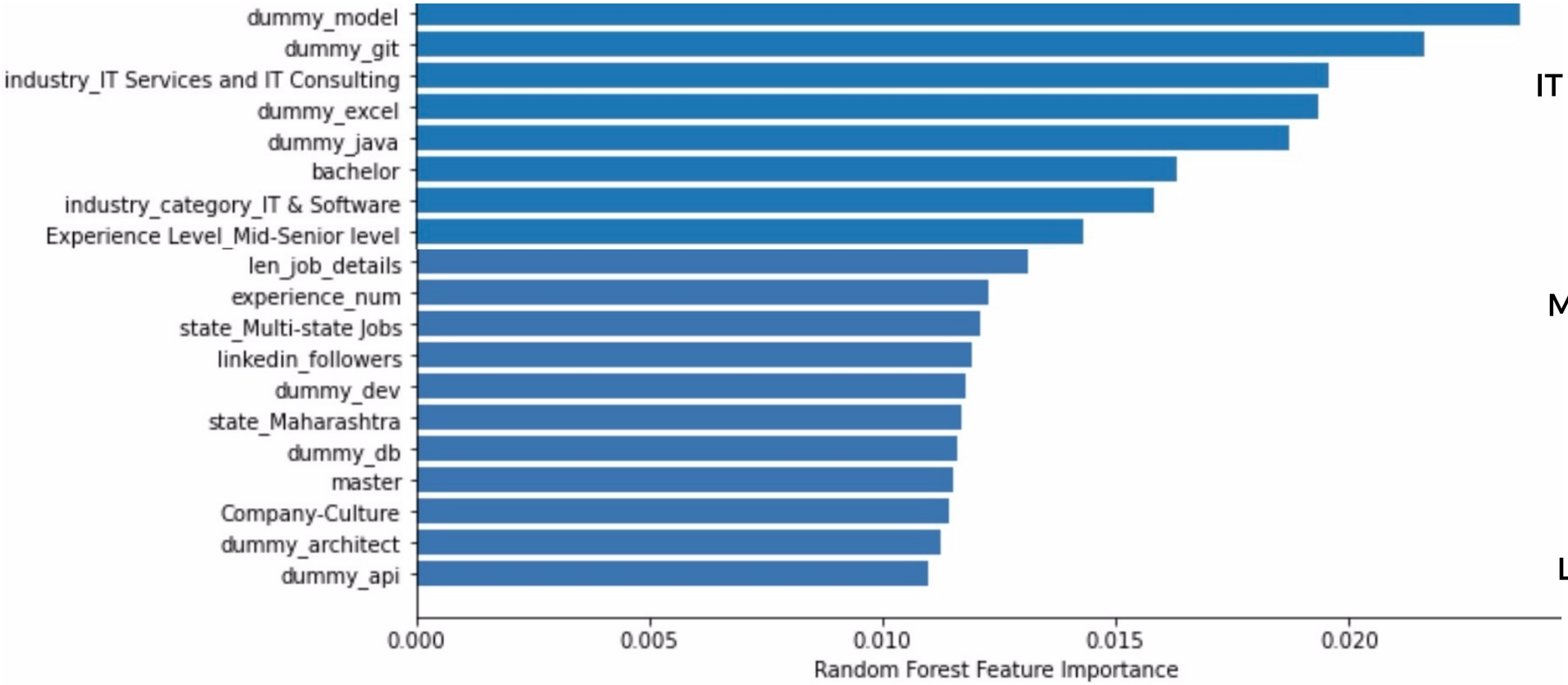
THE GINI INDEX

- **The Gini index is a measure of impurity or purity** used while creating the decision tree.
- **A lower Gini index indicates a better split;** a Gini of 0 means all elements belong to one class, indicating perfect purity.

IMPORTANT FEATURES



INTERPRETATION



COMPANY/INDUSTRY LEVEL

INDUSTRY

IT services and Consulting

IT & Software

COMPANY

10001+ Employees

More LinkedIn Followers

Company Culture

Multi-state Jobs

Maharashtra

JOB POSTING

Longer Job Description

INTERPRETATION

COMPANY/INDUSTRY LEVEL

INDUSTRY
IT services and Consulting
IT & Software

COMPANY
10001+ Employees
More LinkedIn Followers
Company Culture
Multi-state Jobs
Maharashtra

JOB POSTING
Longer Job Description

Industry

- High attraction observed in IT services and consulting industry.
- IT & Software sector equally popular.
- Both sectors known for rapid innovation, advanced work environments, and exposure to the latest technology.

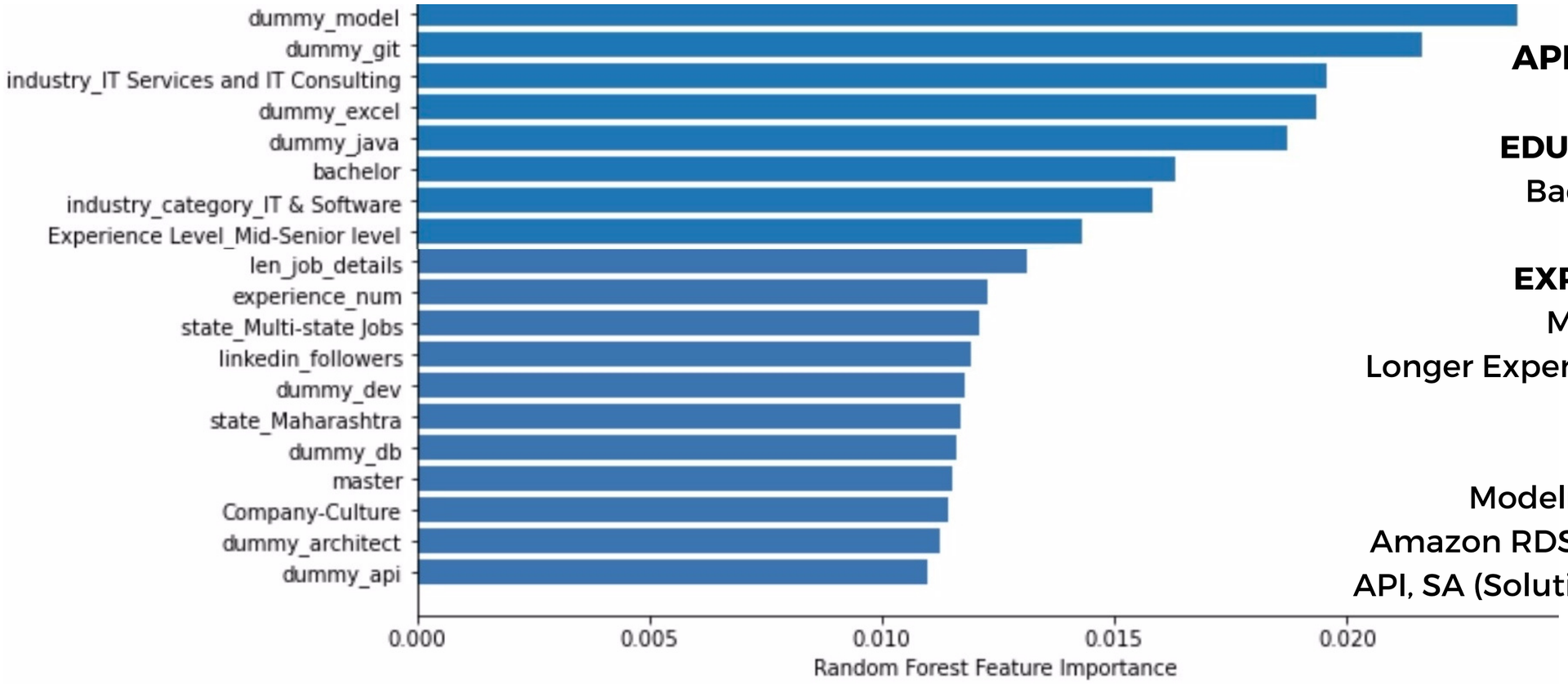
Company

- Larger companies (over 10,001 employees) with substantial LinkedIn following more appealing to job seekers.
- Reasons:
  - Comprehensive learning experiences.
  - Diverse career paths.
  - Opportunities to work on large-scale projects.
  - Access to extensive networks and resources.

Job Description

- Multi-state companies, particularly those in Maharashtra, more attractive to applicants.- Indicates significant geographical influence on job preferences.
- Longer job descriptions correlate with a higher number of applicants. - Detailed information helps candidates assess job suitability and interest.
- Inclusion of a positive company culture in job descriptions (such as inclusivity and balance) significantly attracts applicants.

INTERPRETATION



APPLICANT LEVEL

EDUCATIONAL LEVEL

Bachelor → Master

EXPERIENCE LEVEL

Mid-Senior Level

Longer Experience Requirement Better

TECH SKILLS

Modeling, Git, Excel, Java,
Amazon RDS, Dev Ops, DB (Database),
API, SA (Solution Architect), Architecture

INTERPRETATION

APPLICANT LEVEL

EDUCATIONAL LEVEL
Bachelor → Master

EXPERIENCE LEVEL
Mid-Senior Level
Longer Experience Requirement Better

TECH SKILLS
Modeling, Git, Excel, Java,
Amazon RDS, Dev Ops, DB (Database),
API, SA (Solution Architect), Architecture

Education

- Bachelor's degrees more popular among job seekers for initial industry entry. - Offers wide range of opportunities.
- Master's degrees valued for enhancing competitiveness and specialization.

Experience

- Job postings for mid-senior level positions with longer experience requirements draw more interest.
- Entry-level positions often filled through campus recruitment.
- Mid-senior level professionals more likely to apply online.

Tech Skills

- Modeling, Git, Excel, Java, Amazon RDS, Dev Ops, DB (Database), API, SA (Solution Architect), Architecture
- These skills are either popularly learned or in high demand in the industry.

INFLUENCERS OF JOB APPLICATION RATES ON LINKEDIN IN INDIA - EXPAND

TALENT ACQUISITION IN INDIA'S IT SECTOR

- India's IT industry: a significant contributor to GDP and employment (Bundhun, 2022)
  - The gap between the availability of skilled professionals and the industry's growing needs.
- The widening gap between talent demand and supply (Bundhun, 2022)
- The necessity for skilled workers in critical and niche areas (Bundhun, 2022)
  - A surge in demand for skilled workers in critical areas is outpacing supply, creating a landscape where talent acquisition becomes not just an operational task but a strategic imperative.

TARGETING EDUCATED AND EXPERIENCED PROFESSIONALS

- Attract professionals with Bachelor and higher degrees.
- Focus recruitment on mid-senior level experience.
- Align with market demand for seasoned expertise.
  - Ensuring that the workforce is equipped to handle complex, evolving technological landscapes.

INFLUENCERS OF JOB APPLICATION RATES ON LINKEDIN IN INDIA - EXPAND

DETAILED JOB DESCRIPTIONS

- Emphasize the role of job description detail in attracting candidates (Bhardwaj, 2023)
  - Detailed descriptions act as a first point of contact, informing candidates about the role and the company culture.
- Present data or a quick case study showing increased engagement with detailed postings.
- Recommendation: Urge businesses to invest in crafting comprehensive job descriptions.

RETENTION IN LARGE-SCALE IT ENTERPRISES

- Competitive compensation and flexible work environments to combat high attrition rates (Bundhun, 2022).
- Continuous learning opportunities as a retention strategy.
- Addressing the challenges of cross-offers and job market competition.

INFLUENCERS OF JOB APPLICATION RATES ON LINKEDIN IN INDIA - EXPAND

UPSKILLING FOR INDUSTRY-RELEVANT TECH SKILLS

- High demand for skills in Modeling, Git, Excel, Java, databases, systems analysis, and architecture.
- Implement reskilling and upskilling initiatives.
- Training programs to update the workforce skill set.
    - These training programs are designed to update the workforce's skill set, making them more adaptable and responsive to the fast-paced changes in technology.

BRIDGING THE GRADUATE SKILL GAP

- Addressing the mismatch between graduate skills and industry needs.
- Industry-academia partnerships for curriculum alignment.
- Reducing onboarding time through pre-employment training.
    - This collaboration is vital to ensure that the curriculum is aligned with on-the-ground industry needs, reducing the time and resources spent on training new hires to be job-ready.

INFLUENCERS OF JOB APPLICATION RATES ON LINKEDIN IN INDIA - EXPAND

CONCLUSION

In conclusion, the IT and Software industry in India is on a transformative journey. Through strategic talent acquisition, retention, and continuous skill development, the sector will be able to overcome its current challenges. This holistic approach will not only fill the immediate talent needs but will also lay the foundation for sustainable growth and innovation.

REFERENCE

Bhardwaj, N. (2023a, July 14). Talent trends shaping the India job market in 2023. India Briefing News. https://www.india-briefing.com/news/emerging-talent-trends-shaping-india-job-market-in-2023-28900.html/

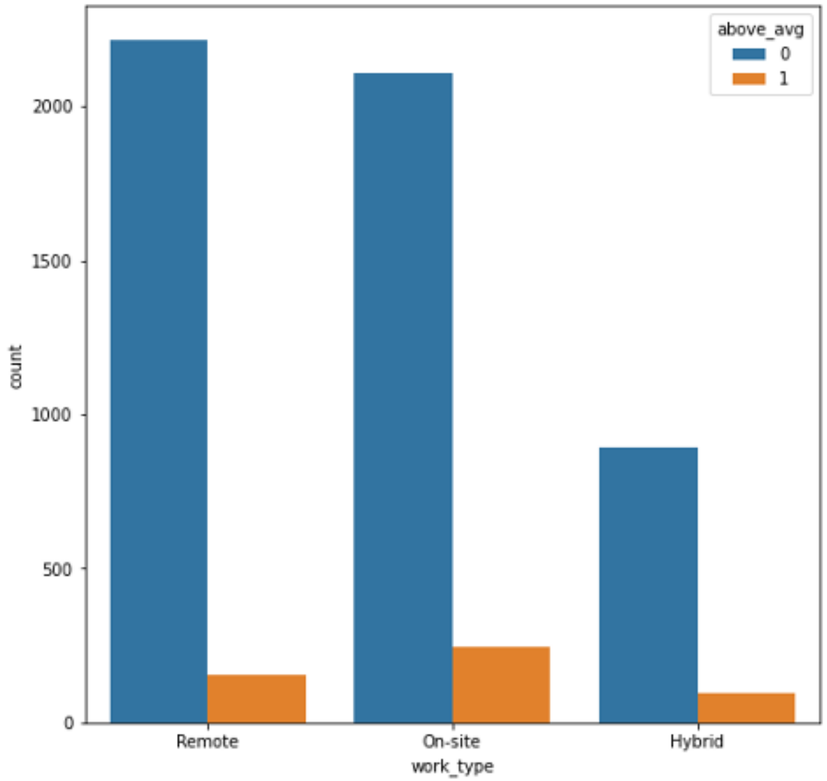
Bundhun, R. (2022, March 14). Here’s why India’s it sector is facing a talent war. The National. https://www.thenationalnews.com/business/technology/2022/03/14/heres-why-indias-it-sector-is-facing-a-talent-war/#:~:text=India%27s%20IT%20industry%20employs%204,Reuters%20%0A%0ARebecca%20Bundhun

THANKS FOR WATCHING

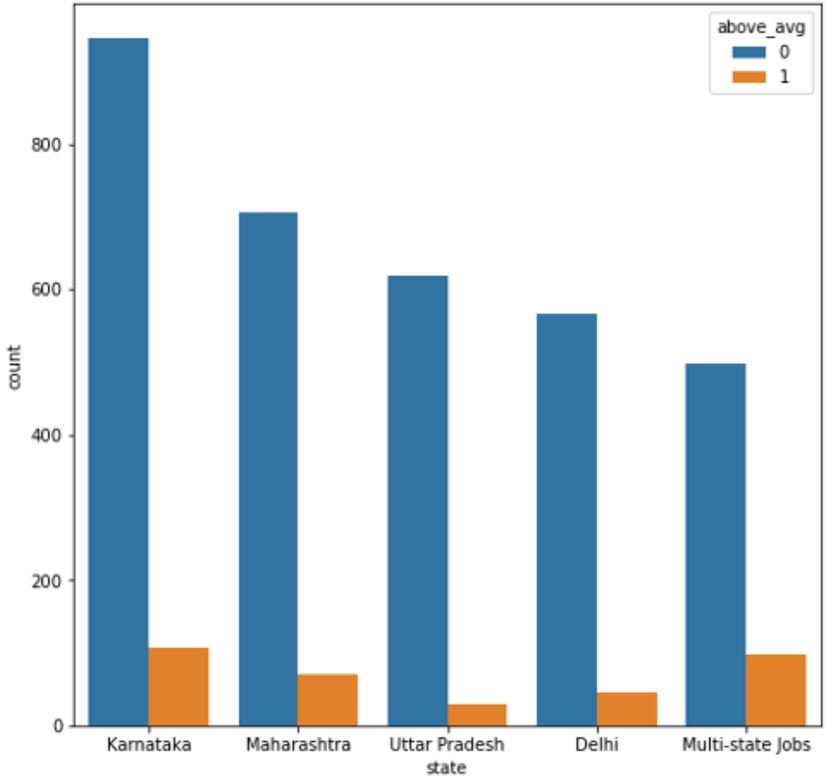
Any questions?

APPENDIX

CHART PAGE

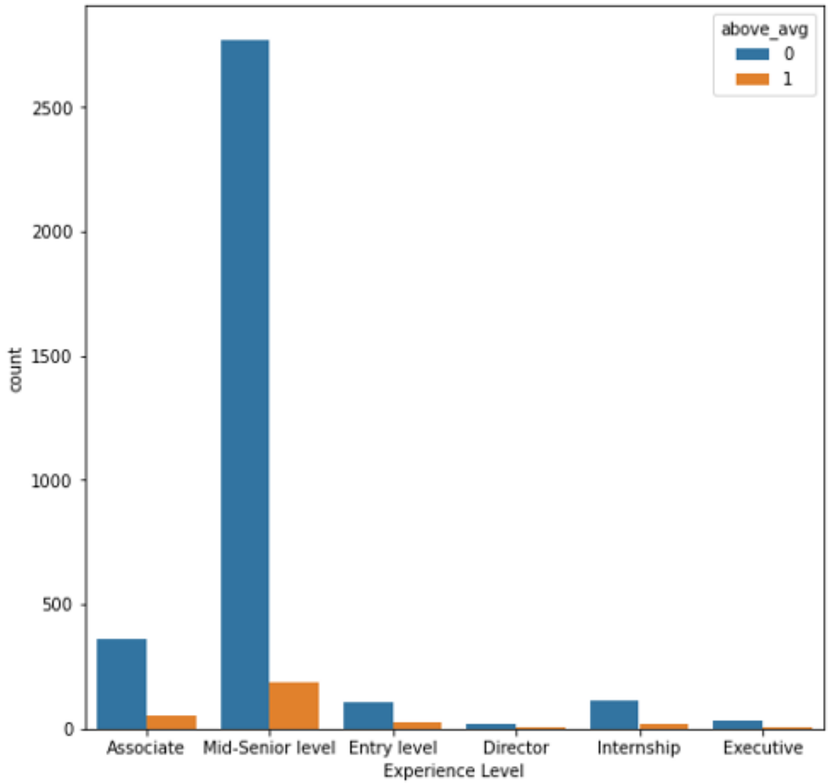


HISTOGRAM #1: APPLICATION RATE THAT IS ABOVE AVERAGE VS. WORK TYPE

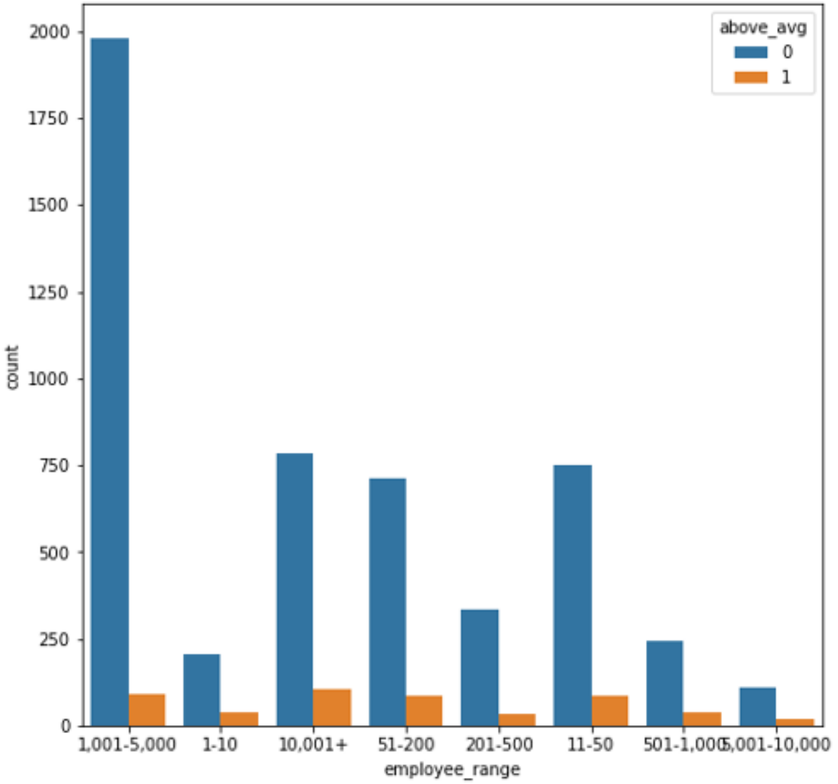


HISTOGRAM #2: APPLICATION RATE THAT IS ABOVE AVERAGE VS. TOP 5 STATES BY NUMBER OF OCCURRENCES

CHART PAGE



HISTOGRAM #3: APPLICATION RATE THAT IS ABOVE AVERAGE VS. EXPERIENCE LEVEL



HISTOGRAM #4: APPLICATION RATE THAT IS ABOVE AVERAGE VS. EMPLOYEE RANGE