Predicting Employee Attrition using Logistic Regression: A Comprehensive Data-Driven Approach

Abhay Singh (23122102), Shivangi Sharma (23122132)
Christ University, Pune-Lavasa Campus

Abstract. *This report presents a comprehensive analysis of job change prediction using Logistic Regression, a statistical modeling technique suited for binary classification. The primary objective was to determine whether candidates are likely to seek new job opportunities based on their demographic, educational, and professional attributes. The dataset underwent extensive preprocessing, including target variable mapping, handling missing values, encoding categorical features, and scaling numerical data. A structured pipeline ensured consistent preprocessing during training and testing. Logistic Regression achieved an accuracy of* ***78.8%****, with a strong recall of* ***92%*** *for class* `0` *(not looking for a job change). However, the model struggled to predict class* `1` *(seeking a job change), achieving a precision of* ***63%*** *and recall of* ***39%*** *for this minority class. The weighted F1-score was* ***77%,*** *indicating reasonable overall performance. The report also compared Logistic Regression with other models, such as Random Forest, SVM, and KNN. While Logistic Regression provided interpretable results and solid baseline performance, models like SVM outperformed it in accuracy (****80.1%****) and balanced metrics. Nonetheless, Logistic Regression remains a robust choice due to its simplicity and effectiveness for feature analysis. This analysis highlights the importance of data preprocessing, including feature engineering and scaling, in enhancing model performance. The results suggest opportunities for further improvement through techniques like class balancing, feature selection, and hyperparameter tuning. This study provides valuable insights into the predictive modeling of job changes, which can inform recruitment and workforce planning strategies.*

Keywords: Job Change, Employee Attrition, City Development Index, Employee Reten-tion, EDA, Predictive Modeling.

1. Introduction

Predicting job changes is a critical task in human resource analytics, enabling organizations to identify employees or candidates who are likely to seek new opportunities. Such insights allow companies to optimize recruitment strategies, tailor retention policies, and improve workforce planning. This report focuses on utilizing Logistic Regression, a widely used statistical modeling technique, to predict whether candidates will look for a job change based on their demographic, educational, and professional attributes. The dataset used in this study contains features such as city development index, gender, relevant experience, education level, and training hours, among others. Extensive preprocessing steps were conducted to ensure the data's suitability for analysis. These included handling missing values, encoding categorical variables, and scaling numerical features. Logistic Regression was selected for its simplicity, interpretability, and effectiveness in binary classification problems. This report explores the entire pipeline, from data preprocessing to model evaluation, and provides actionable insights based on the predictive performance of Logistic Regression.

1.1. Problem Statement

Understanding the factors influencing job change decisions is vital for organizations to improve retention and engagement. This study investigates variables like city development index, company attributes, and professional experience to identify patterns and correlations, providing actionable insights for addressing workforce mobility and enhancing job satisfaction.

2. Objectives

a. To identify the course type with the highest job switching tendency.
b. To analyze the influence of gender on the likelihood of switching jobs.
c. To examine the impact of company size and type on employee job-switching behavior.
d. To evaluate how professional experience affects the likelihood of changing jobs.
e. To predict employee job retention by classifying individuals likely to leave using Support Vector Machine (SVM).

3. Materials and Methodology

3.1. Materials

Dataset: The dataset includes 14 fields such as enrollee_id, city_development_index, gender, relevant experience, education level, major discipline, experience, company size, company type, last job change, and target (indicating job change, 0-No Change, 1-Change).

Exploratory Data Analysis

- **Checking for missing values percentage using 'bar plot'**

Fig.1 represents a distribution of the missing values in all features of the dataset using bar-plot.

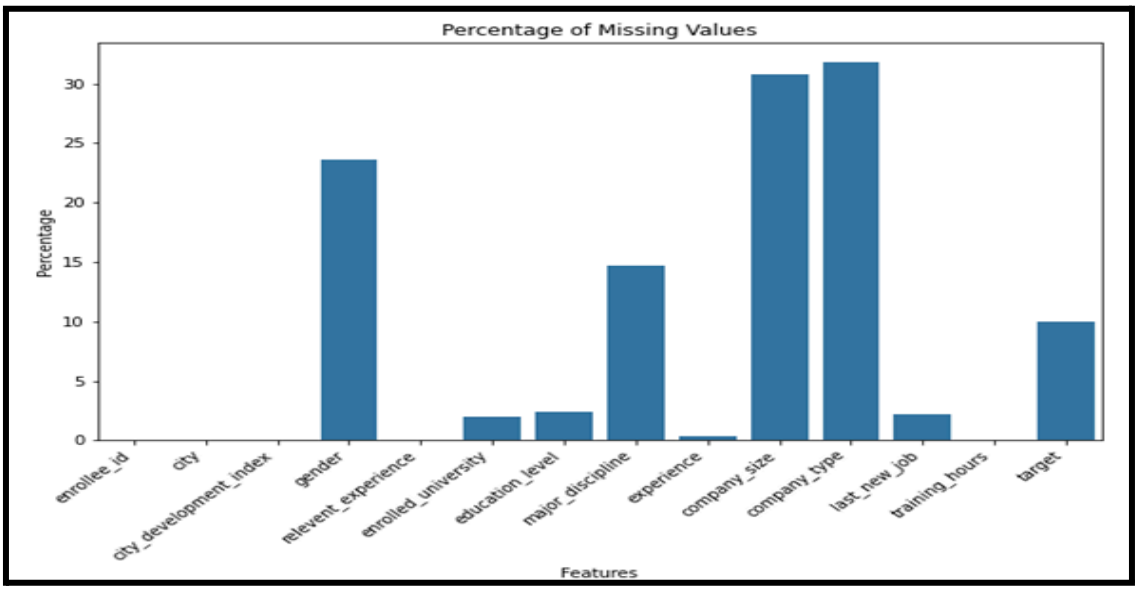


Figure 1

Fig.1 bar chart displays the percentage of missing values for each feature in the dataset. The features "company_size" and "company_type" have the highest proportion of missing data, each accounting for approximately 30% of their values. The "gender" feature also exhibits a notable level of missing data, with around 23% of its values absent. In contrast, features like "enrollee_id," "city," and

"city_development_index" are complete, with no missing data observed. This visualization provides a clear overview of data quality, identifying areas that require attention for preprocessing, such as handling or imputing missing values. The high percentage of missing data in certain features, particularly "company_size" and "company_type," may significantly impact the analysis or model performance if not addressed appropriately. By pinpointing these issues, the chart serves as a valuable tool for guiding data-cleaning efforts and ensuring robust downstream analysis.

- **Basic visualizations to understand the distribution of the data**

Below we have performed EDA to check how the different values in the dataset behaves. We have used several bar graphs and correlation heatmap matrix to understand the behavior of data variables, frequency of employees working in different types and sizes of companies and the relationship between different variables like Target vs. Gender, Target vs. Relevant Experience, Target Distribution, etc.

Fig.2 represents a bar chart that shows the distribution of company sizes within the dataset.

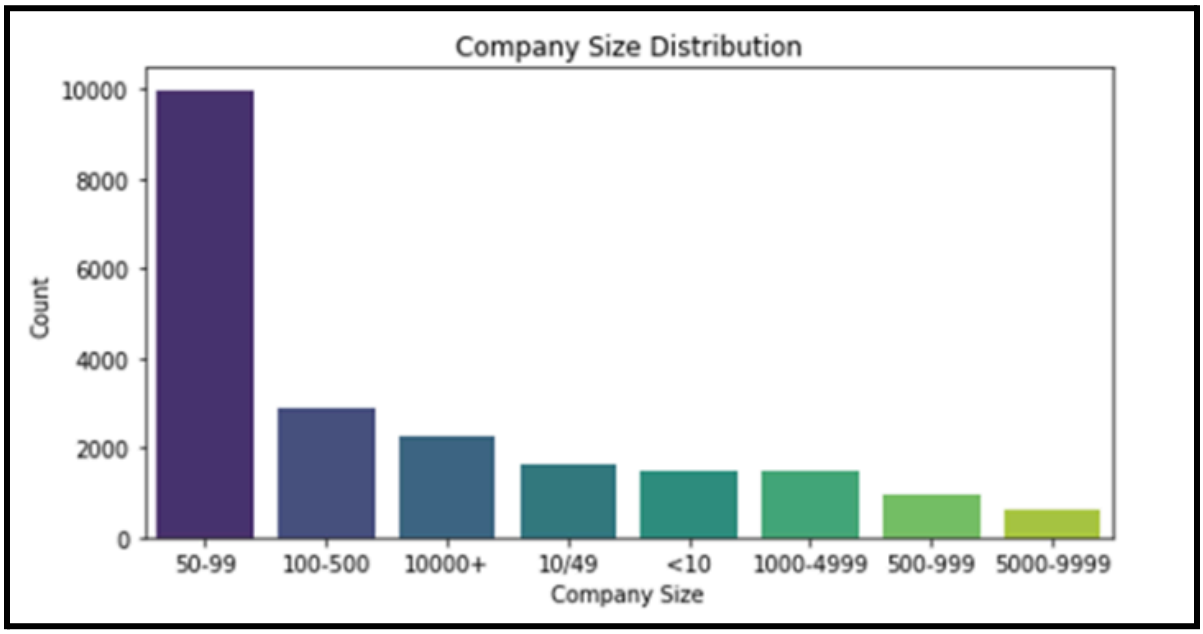


Figure 2

Fig.2 *The bar chart illustrates the distribution of company sizes in the dataset. The most common company size is "50-99" employees, indicating that mid-sized companies dominate the dataset. This is followed by "100-500" employees and large companies with "10000+" employees, both of which also occur frequently. On the other hand, companies within the "5000-9999" employee range are the least represented, suggesting that organizations of this size are relatively rare in the data. The chart provides a clear snapshot of company size representation, highlighting the prevalence of small to mid-sized enterprises. This distribution offers valuable insights into the dataset's composition, which may influence the interpretation of patterns or trends in further analysis. Understanding these distributions is crucial for assessing the representativeness of the dataset and ensuring that findings are generalizable across different company size categories. This distribution offers valuable insights into the dataset's composition, which may influence the interpretation of patterns or trends in further analysis. Understanding these distributions is crucial for assessing the representativeness of the dataset and ensuring that findings are generalizable across different company size categories.*

Fig.3 represents histogram displays the distribution of city development index values.

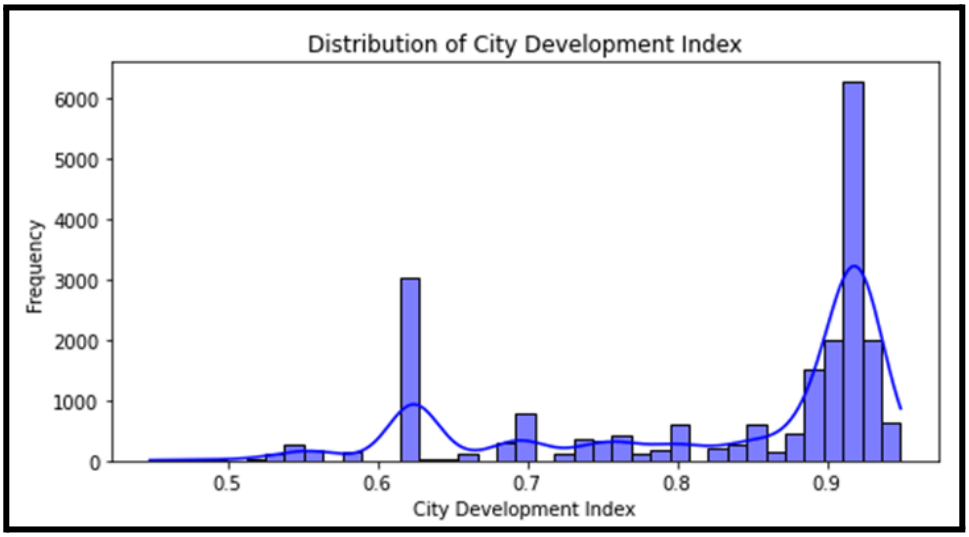


Figure 3

Fig.3: *The histogram showcases the distribution of city development index (CDI) values in the dataset, ranging from approximately 0.45 to 1.0. The data exhibits a right-skewed pattern, indicating that the majority of cities have relatively high development indices. A prominent peak is observed around 0.92, suggesting that a significant proportion of cities fall within this high development range. Additionally, a smaller peak appears near 0.62, reflecting a secondary grouping of cities with moderate development levels. The skewness highlights an uneven distribution, where cities with lower development indices are less frequent compared to those at the higher end. This visualization provides critical insights into the urban development landscape captured in the dataset. Understanding the CDI distribution is essential for analyzing its potential impact on other variables and for tailoring strategies that address disparities in city development levels.*

Fig.4 represents a histogram which depicts the distribution of training hours.

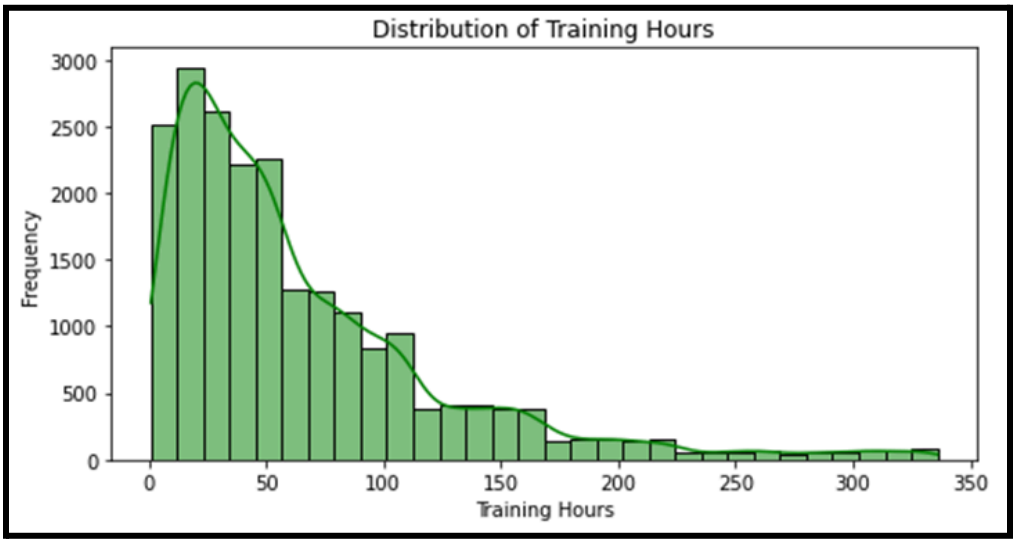


Figure 4

Fig.4 : *The histogram showcases the distribution of city development index (CDI) values in the dataset, ranging from approximately 0.45 to 1.0. The data exhibits a right-skewed pattern, indicating that the majority of cities have relatively high development indices. A prominent peak is observed around 0.92, suggesting that a significant proportion of cities fall within this high development range. Additionally, a smaller peak appears near 0.62, reflecting a secondary grouping of cities with moderate development levels. The skewness highlights an uneven distribution, where cities with lower development indices are less frequent compared to those at the higher end. This visualization provides critical insights into the urban development landscape captured in the dataset. Understanding the CDI distribution is essential for analyzing its potential impact on other variables and for tailoring strategies that address disparities in city development levels.*

Fig.5 represents a histogram which depicts the distribution of Gender.

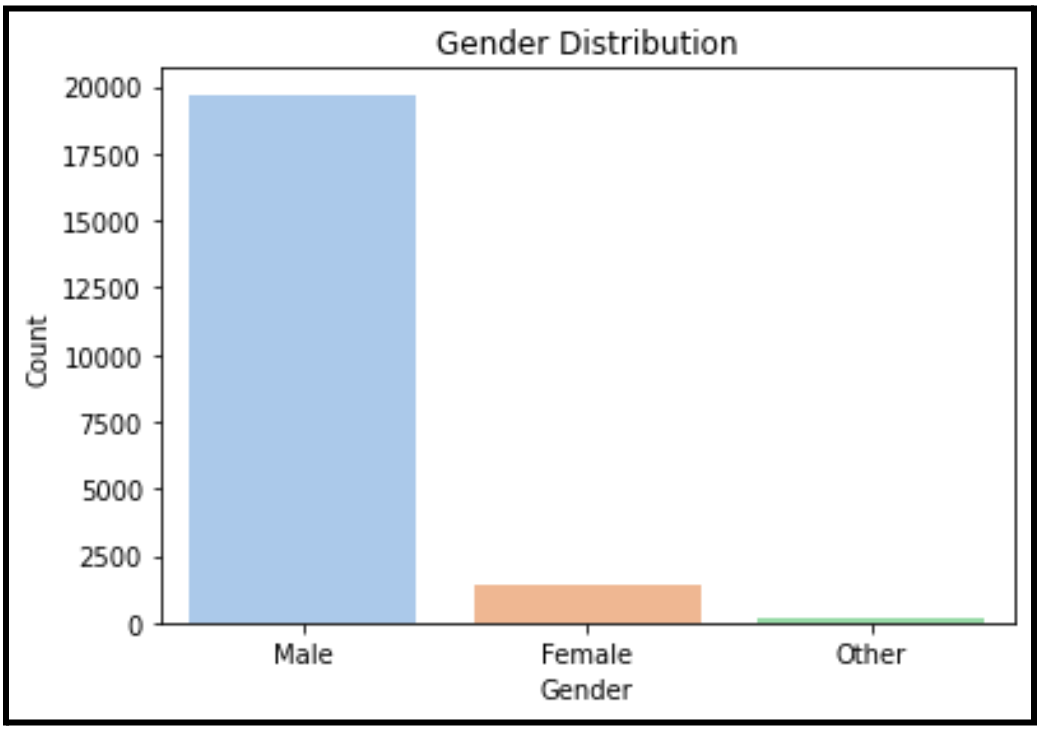


Fig.5: *The chart displays the gender distribution within the dataset, highlighting a significant imbalance. Males form the predominant group, comprising the majority of the dataset. Females represent a considerably smaller proportion in comparison, reflecting a noticeable gender disparity. Additionally, individuals classified under "Other" gender are present in very low numbers, making them the least represented category. This visualization provides valuable insights into the demographic composition of the dataset, emphasizing the dominance of males while underscoring the limited representation of other genders. Such an imbalance can influence analysis outcomes, especially in studies where gender-related trends or patterns are relevant. Understanding this distribution is essential for interpreting results accurately and ensuring that any conclusions drawn take the dataset's demographic skew into account. This knowledge is also crucial for identifying potential biases and addressing them during data analysis or model development stages.*

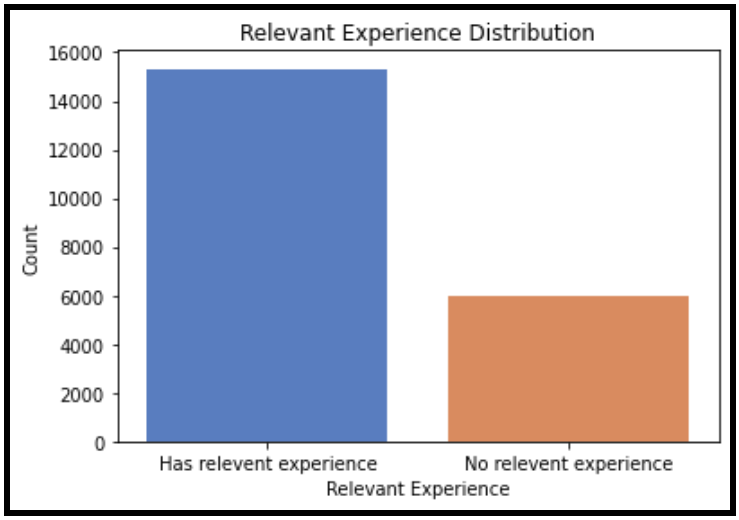


Figure 6 *The chart depicts the distribution of relevant work experience among individuals in the dataset. A majority of individuals possess prior relevant work experience, indicating that they have worked in roles similar to or aligned with the current context. Conversely, a smaller proportion of individuals lack relevant experience, forming a minority in the dataset. This disparity suggests that the dataset is skewed toward individuals with some level of prior exposure or expertise in related fields. Understanding the distribution of relevant work experience is critical for analyzing its influence on outcomes, such as career advancement, performance metrics, or other dependent variables. The representation of individuals without relevant experience, though smaller, provides an opportunity to explore the challenges or opportunities they face compared to their experienced counterparts. Overall, the chart offers valuable insights into the workforce composition and highlights the importance of relevant experience in the dataset.*

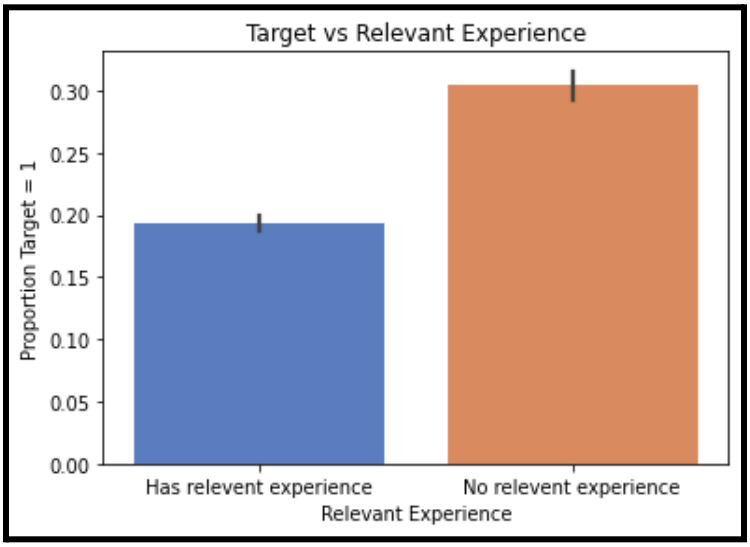


Figure 7: *represents a histogram which depicts the distribution of 'Relevant Experience' of 'People who left'. The plot highlights the relationship between relevant work experience and the target variable. It reveals that individuals with "No relevant experience" have a higher proportion of "Target = 1," approximately 0.30, compared to those with "Relevant experience," whose proportion is closer to 0.20. This suggests that candidates lacking relevant work experience are more likely to fall into the positive target category. The inclusion of error bars provides a visual representation of the confidence intervals for these proportions, offering insight into the variability and reliability of the observed differences. The noticeable gap between the two proportions, coupled with the non-overlapping error bars, emphasizes the distinct trends between the groups. This finding could be significant for decision-making processes, indicating that prior work experience may not always align with the desired target outcome, depending on the context of the analysis or the criteria of the target variable.*

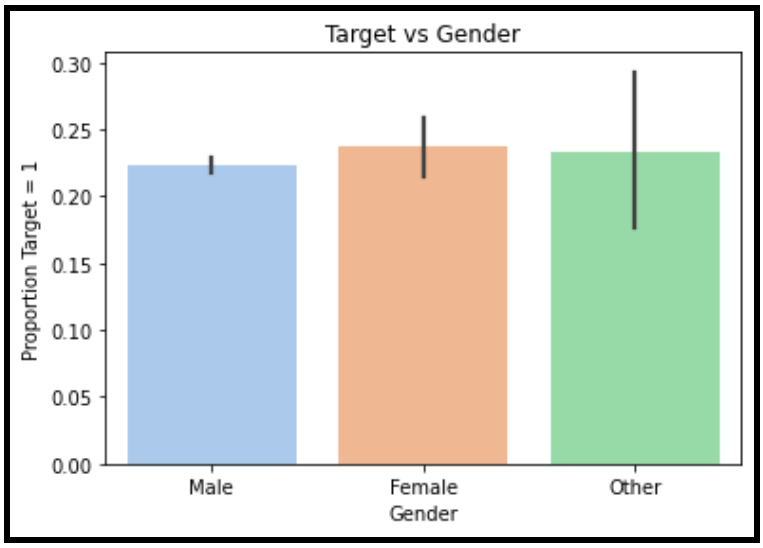


Figure 8: *The plot compares the proportion of "Target = 1" across genders: Male, Female, and Other. The proportions are relatively consistent across all three categories, ranging between 0.20 and 0.25. This similarity suggests that gender does not play a significant role in determining the target variable. The inclusion of error bars further reinforces this observation, as they overlap across all gender groups, indicating no statistically significant difference in the proportions. These findings imply that the likelihood of "Target = 1" is evenly distributed regardless of gender, emphasizing that gender-based disparities are not evident in this context. This consistency provides valuable insights, helping to rule out gender as a major influencing factor for the target variable, thereby enabling a more focused analysis on other potential determinants within the dataset. The plot is a critical visualization for understanding the target variable's behavior across different demographic segments.*

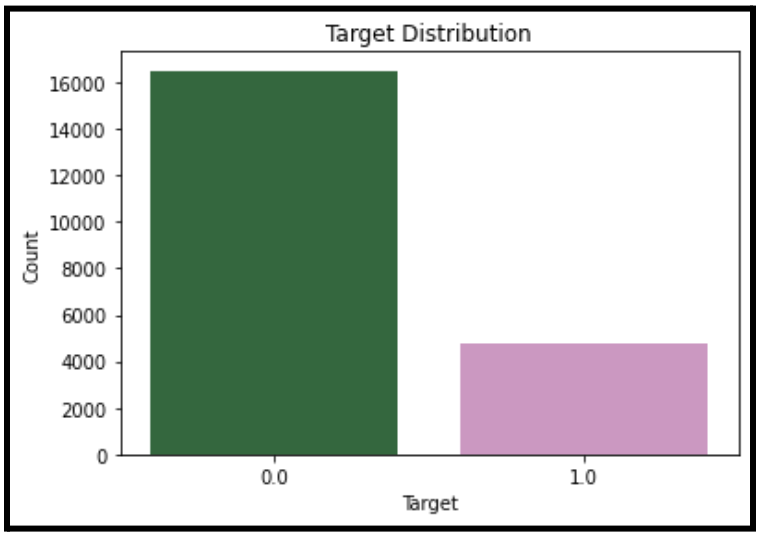


Figure 9: *The bar chart illustrates the distribution of the target variable, which is divided into two categories: 0 and 1. The chart reveals a pronounced class imbalance, with the majority of instances falling under class "0," while class "1" comprises a significantly smaller proportion. This imbalance indicates that the dataset predominantly consists of one category, which may affect model training and performance. The underrepresentation of class "1" could lead to biased predictions, as the model might favor the majority class. Addressing this imbalance is crucial to ensure the robustness of any predictive modeling efforts. Techniques such as resampling, class weighting, or employing specialized algorithms can help mitigate the impact of the imbalance. Overall, the chart underscores the need for targeted preprocessing steps to achieve balanced and reliable analysis.*

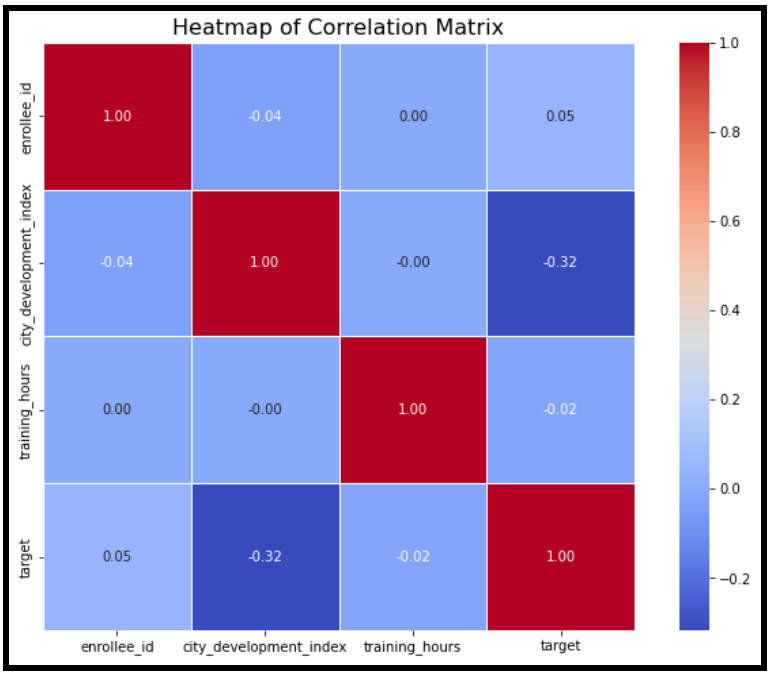


Figure 10: *The heatmap displays the correlation coefficients among 'enrollee_id,' 'city_development_index,' 'training_hours,' and 'target.' A notable observation is the moderate negative correlation of -0.32 between 'city_development_index' and 'target,' suggesting that higher city development levels are associated with a lower likelihood of the target variable being 1. In contrast, the correlations among other variables, such as 'training_hours' with 'target' or 'city_development_index,' are relatively weak and close to zero, indicating minimal linear relationships. The color scale on the right visually represents the strength and direction of correlations, ranging from -0.2 (blue) for weak negative relationships to 1.0 (red) for strong positive relationships. This visualization provides valuable insights into the interdependencies of key features, highlighting 'city_development_index' as a variable of interest in predicting the target. The weak correlations elsewhere suggest these features may have limited influence on one another.*

Insights from EDA

- **Target Distribution Analysis:**
  - A bar chart visualized the distribution of job seekers who intended to switch jobs (target = 1) versus those who did not (target = 0).
  - **Insights:** The analysis revealed a higher tendency of job-seeking behavior among certain demographics and experience levels. Understanding this split helps target retention strategies for at-risk employees.
- **Gender Influence on Job Switching:**
  - A bar graph comparing the percentage of males and females switching jobs highlighted significant gender disparities.
  - **Insights:** Males exhibited a higher likelihood of seeking new opportunities compared to females. Further examination showed fewer women enrolled in advanced education, limiting their career progression.
- **Correlation Heatmap:**
  - A heatmap visualized correlations between key variables such as experience, city development index, company size, and job-switching behavior.
  - **Insights:** Strong positive correlations were observed between CDI and job mobility. Company size exhibited a negative correlation with attrition rates, affirming findings from previous visualizations.

3.2. Methodology

Data preprocessing involved handling missing values, normalizing categorical variables, and preparing the dataset for analysis. The methodology incorporated exploratory data analysis (EDA) using Python, Power BI and Orange to uncover trends and patterns. Visualizations, such as bar charts, scatter plots, and pie charts, were employed to analyze the influence of city development index, company characteristics, and individual factors on job mobility. This systematic approach provided actionable insights into the dynamics of job-seeking behavior.

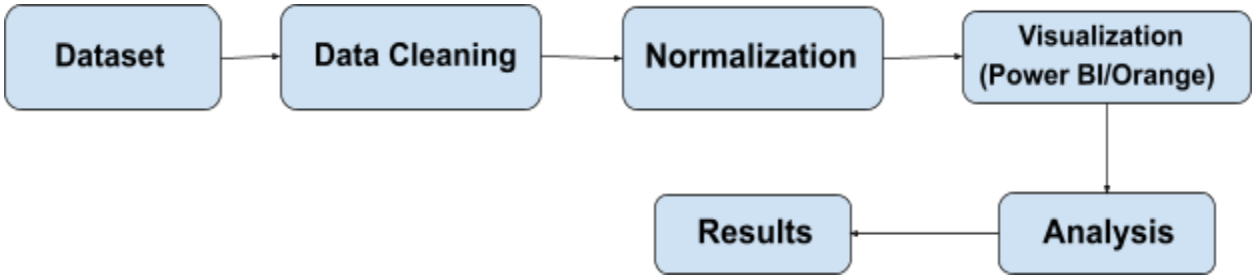
3.2.1 Data Preprocessing

- **Data Cleaning:** Handling missing values in fields like company size, type, and major discipline using python for basic EDA. Steps taken in performing data pre-processing in python is as follows:
    - Loading the dataset
    - Handle missing values and normalize categorical fields.
    - Transform data for analysis
    - Dataset Overview (data types, null values, etc.)
    - Checking for missing values percentage using 'bar plot'.
    - Imputing missing categorical values with mode and numerical values with median.
    - Descriptive Statistics
    - Basic visualizations to understand the distribution of the data.
    - Correlation Heatmap
- **Normalization:** Converting categorical variables like experience (">20" to 21, "<1" to 0) and last job change (">4" to 5, "never" to 0) into nu-meric formats.
- **Frequency Analysis:** Evaluating distributions of Company Size, gender, education, and target variables.

3.2.2 Visualization

- **Tools used:** PowerBI, Orange, Python
    - **Power BI:** Used to derive insights using bar charts, scatter plots, and percentage visualizations.
    - **Orange:** Used to run Machine Learning (ML) Model.
    - **Python:** Cleaning, Pre-Processing, Basic EDA.

3.3.3 Flowchart



The flowchart outlines our data analysis process. The process begins with the Dataset, which is then subjected to Data Cleaning, followed by Normalization. After these preprocessing steps, Visualization is

performed using tools like Power BI and Orange. The insights gained from visualization were then fed into Analysis, which ultimately leads to Results. This suggests an iterative loop between analysis and visualization, where initial analysis prompts further exploration and refinement of the visualizations.

4. Results and Discussions

4.1 Results

4.1.1 Which course type shows the highest job switching tendency?

The course type 'no_enrollment' has the highest tendency of switching jobs. It implies that people without formal education tend to switch more as compared to people with formal education.

Fig.11 represents a histogram which depicts the distribution of candidates.

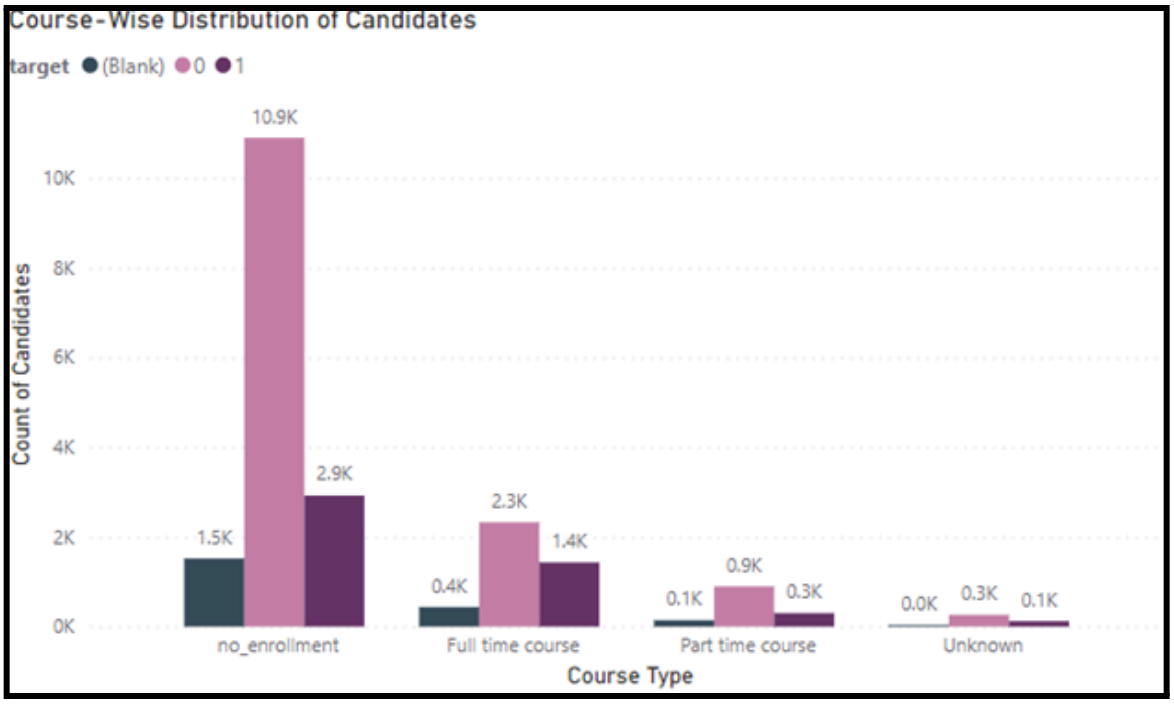


Figure 11

Fig.11: *The chart suggests a potential relationship between course type and the target variable. Candidates who are not enrolled in any course or those enrolled in full-time courses appear more likely to fall into the category represented by the purple bar, which corresponds to "target = 0." This indicates that individuals in these categories have a higher probability of not meeting the target condition. In contrast, candidates in other course types, such as part-time courses, show a different distribution, with a lower likelihood of falling into the "target = 0" group. The chart highlights that course enrollment status may play a role in predicting the target variable, with full-time enrollment or non-enrollment being associated with lower chances of meeting the target condition. This observation could guide further analysis and feature engineering, emphasizing the importance of course type in predicting outcomes related to the target variable.*

4.1.2 Does gender influence the likelihood of switching jobs?

Yes! Males exhibited a higher likelihood of seeking new job opportunities compared to females.

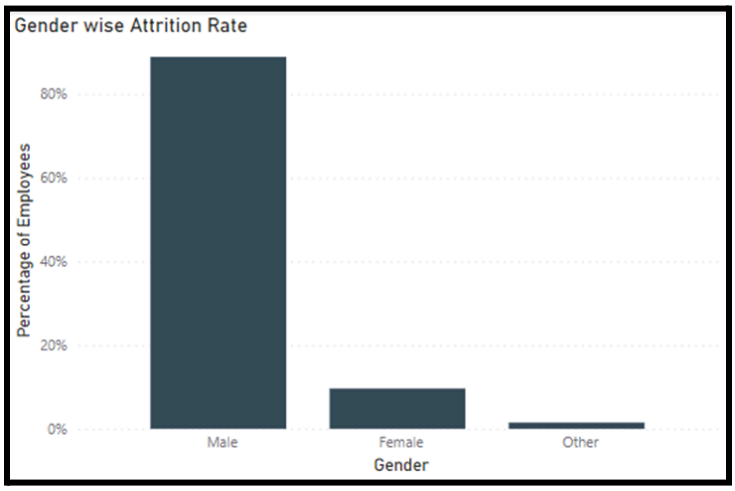


Fig.12 *represents a bar-chart of the distribution of candidates by Gender. The chart highlights a stark contrast in attrition rates across genders. Males have a significantly high attrition rate, approaching 90%, suggesting that a large proportion of male employees have left the organization. This indicates a potential issue with retention among male employees that may require further investigation into underlying causes, such as workplace conditions, job satisfaction, or opportunities for advancement. In contrast, the attrition rate for females is considerably lower, around 10%, indicating better retention within this group. The "Other" gender category shows the lowest attrition rate, near 0%, suggesting that individuals identifying as "Other" have remained largely with the organization. These disparities in attrition rates could point to gender-specific factors affecting employee retention. Further analysis would be essential to explore the reasons behind these differences and to develop targeted strategies for improving retention across all gender categories.*

On further analysis, we found out the difference in levels of education between male and female which indicates the plethora of opportunities presented to men as compared to women. We see a downward trend in the number of females enrolled in higher education programs.

Fig.13 represents a histogram which depicts the '*Gender-wise Distribution of Education Level'*.

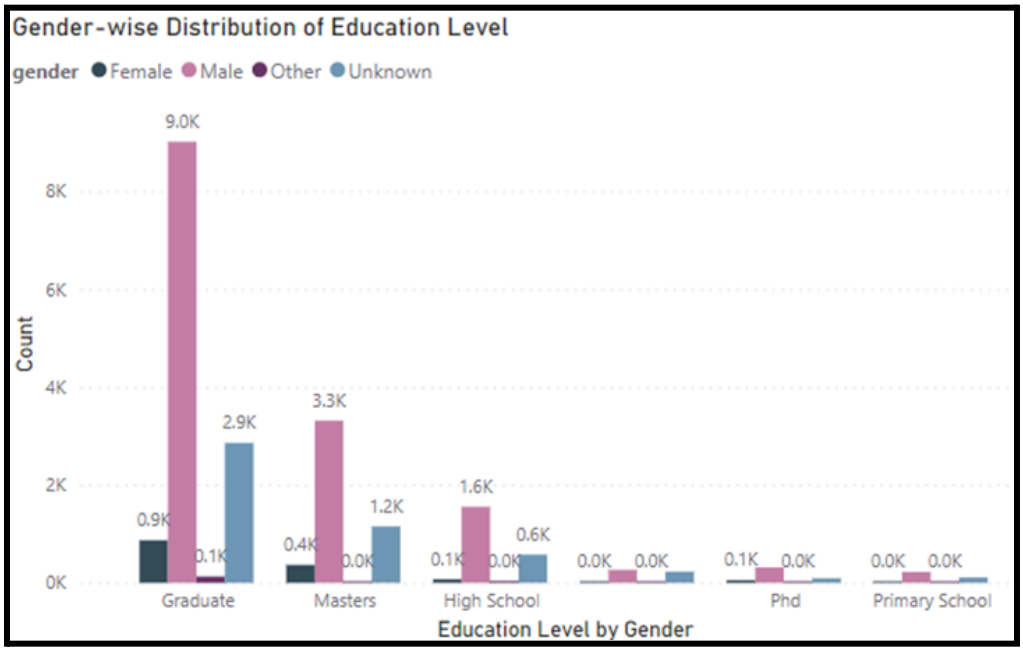


Figure 13

Fig.13: *The bar chart titled "Gender-wise Distribution of Education Level" presents the count of individuals across various education levels—Graduate, Masters, High School, Ph.D., and Primary School—segmented by gender (Female, Male, Other, Unknown). The chart reveals that the "Graduate" education level holds the highest count for most genders, with a notable concentration of individuals in this category. Males dominate across all education levels, with the highest counts observed in both Graduate and Master's degrees. Females are the second largest group in all education levels, though the gap between males and females is particularly evident in Graduate and Master's categories. The number of individuals holding Ph.D. and Primary School education levels is comparatively lower across all genders. This distribution highlights a higher representation of males in higher education levels, with gender differences most pronounced in Graduate and Master's degrees, reflecting potential trends in educational attainment within the dataset.*

4.1.3 Does the size and type of a company affect the likelihood of employees switching jobs?

Yes! Larger companies ("1000+") have better retention than smaller ones.

Fig.14 represents a histogram which depicts the '*Company Size vs Employee Attrition'*.

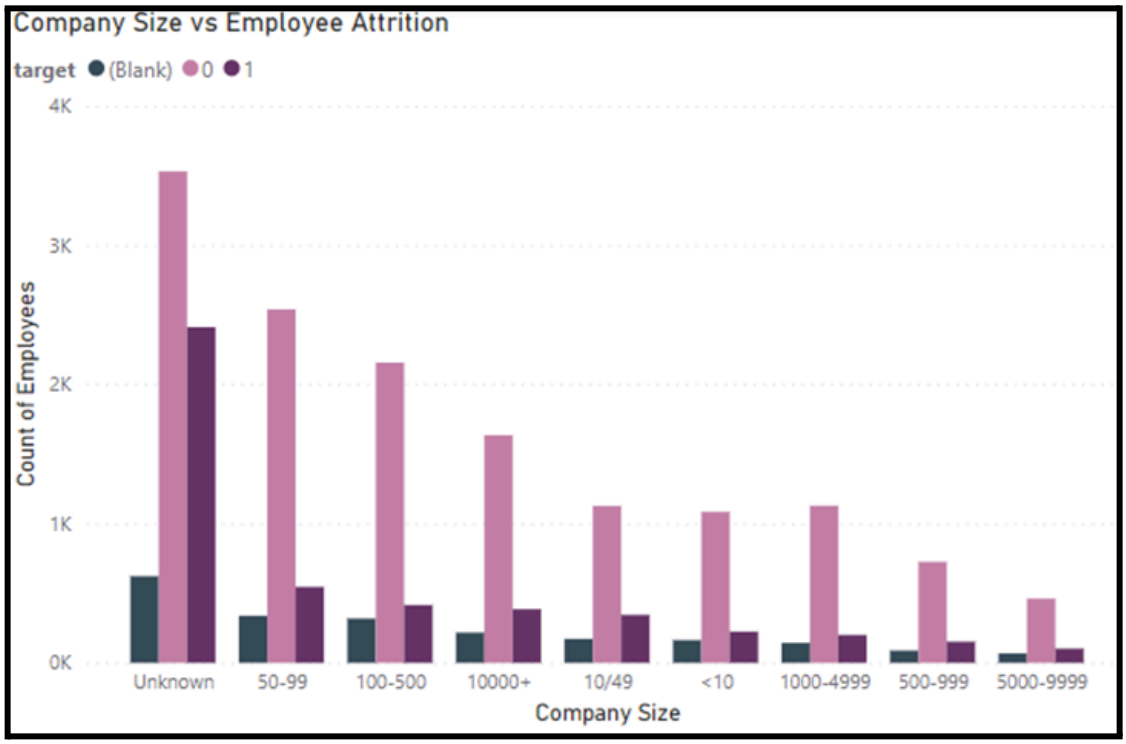


Figure 14

Figure 14: *a bar chart illustrating "Company Size vs Employee Attrition," displaying the count of employees across various company size categories, with a breakdown based on the target variable—likely indicating attrition, where "0" represents no attrition and "1" represents attrition. The chart highlights that both smaller companies (with employee counts of "50-99," "100-500," and "10-49") and larger companies (with "10000+" employees) exhibit higher counts of employees in both attrition categories. This suggests that both small and large companies experience notable levels of employee turnover.*

Additionally, the chart reveals that the "Unknown" category, which may represent missing or unclassified company size data, has a significantly higher count of employees compared to the other company size categories. This insight emphasizes the importance of addressing the unknown category in future analyses to better understand the overall attrition trends across company sizes.

Fig.15 represents a histogram which depicts the '*Company Size vs Employee Attrition'*.

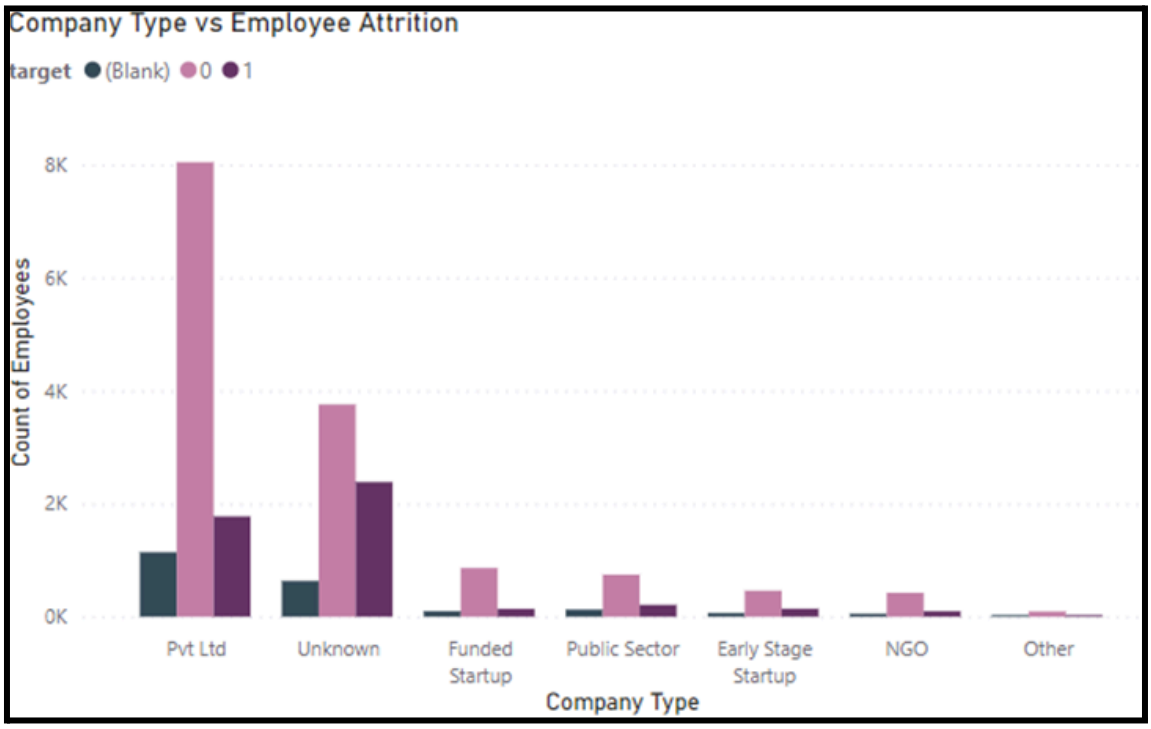


Figure 15

Figure 15:*This bar chart, titled "Company Type vs Employee Attrition" (Figure 15), displays the count of employees in different company types, categorized by a target variable (likely representing attrition). It shows that "Pvt Ltd" and "Unknown" company types have the highest number of employees, with a significant portion belonging to the target category "0" (potentially indicating no attrition), followed by target category "1" (potentially indicating attrition) for the same company types. The y-axis shows the "Count of Employees", while the x-axis shows the "Company Type". The chart also uses approximate values with K suffix (e.g. 8K, 6K, 4K), indicating these are rounded numbers representing thousands. Private limited companies see higher attrition compared to public companies.*

4.1.4 How does experience influence the likelihood of changing jobs?

Candidates with "5-10" years of experience and recent job changes ("<1 year") show higher probabilities of seeking a new job.

Figure 16: This bar chart, Figure 16, titled "Work Experience vs Employee Attrition," shows the count of employees based on their years of work experience, further divided by a target variable (likely representing attrition). The chart reveals that those with over 20 years of experience have the highest

employee count in target category "0" (possibly no attrition), followed by a decreasing trend as experience goes from 5 years to less than 1 year, and the distribution across different experience levels for target category "1" (possibly attrition) appears relatively even. The y-axis is "Count of Employees," and the x-axis is "Years of Experience". The chart also uses approximate values with K suffix (e.g. 3K, 2K, 1K), indicating these are rounded numbers representing thousands.

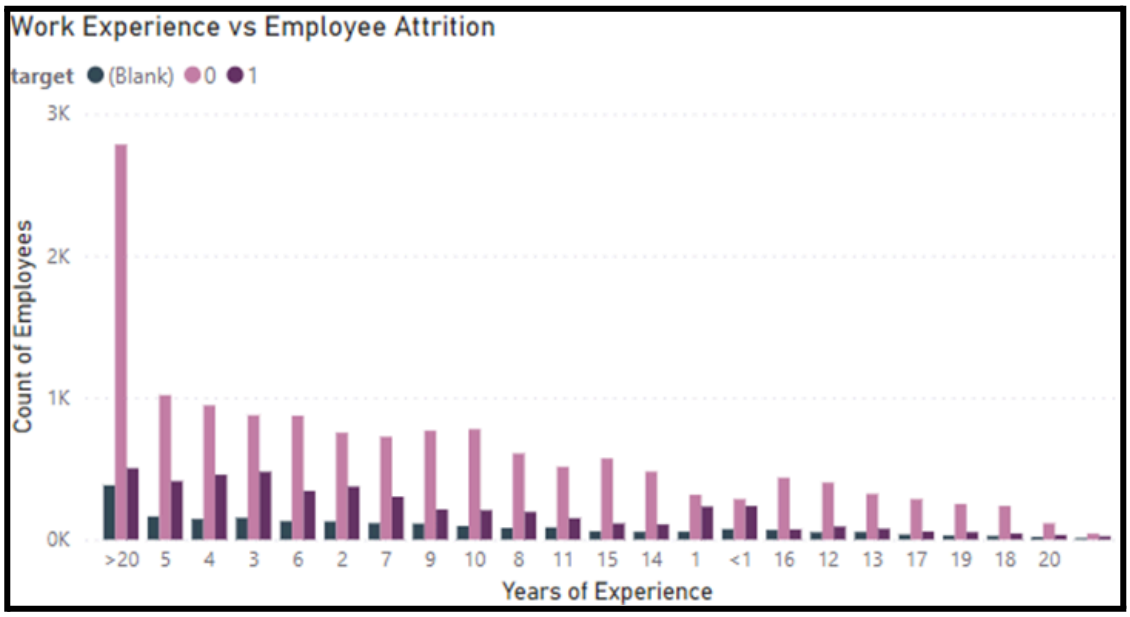


Figure 16

Private limited companies see higher attrition compared to public companies.

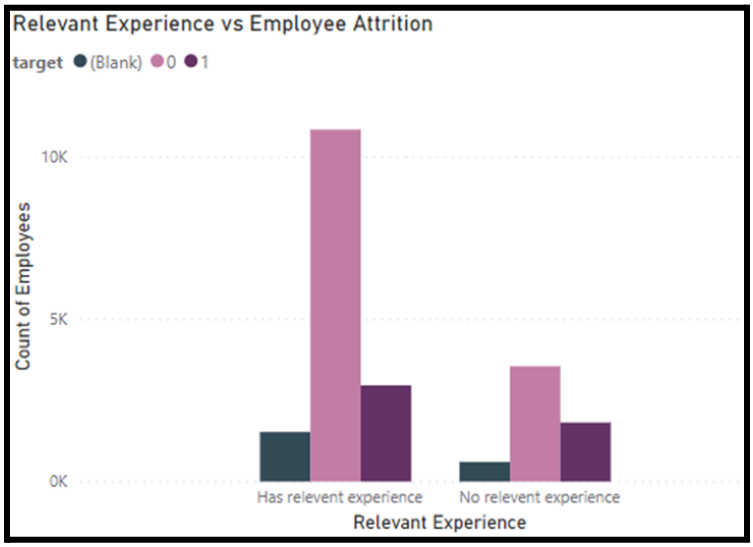
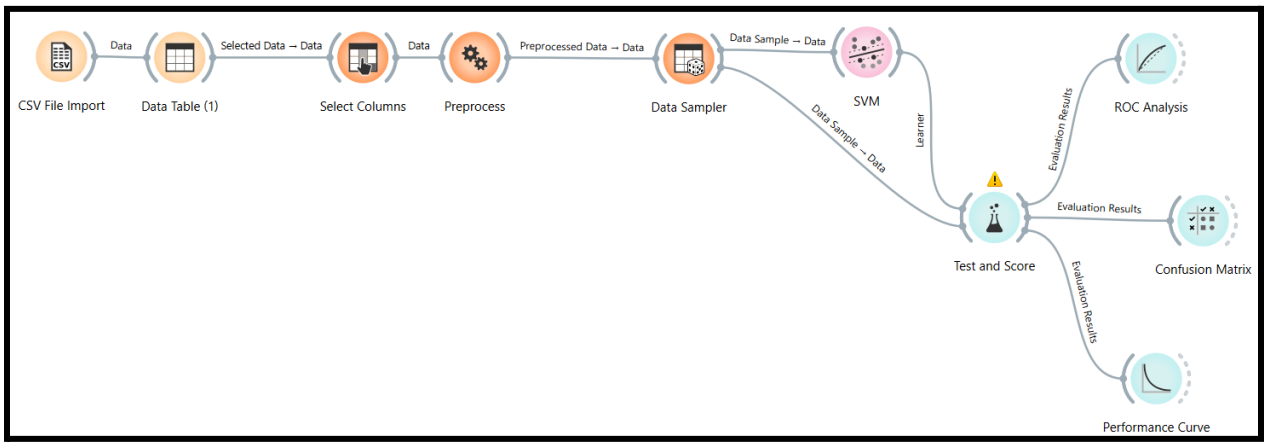


Figure 17: *"Relevant Experience vs Employee Attrition," presents a bar chart that compares the count of employees with and without relevant experience, further categorized by a target variable likely representing employee attrition. The chart reveals that employees with relevant experience are more prevalent in the target category "0" (indicating no attrition), with a significantly higher count than those in category "1" (indicating attrition). In contrast, employees without relevant experience show a higher count in target category "1" (indicating attrition) than in category "0" (indicating no attrition). The y-axis of the chart represents the "Count of Employees," while the x-axis shows the "Relevant Experience" categories. The chart uses rounded numbers with a "K" suffix (e.g., 10K, 5K), indicating the values are approximate and in the thousands. This visualization provides insight into the relationship between relevant experience and employee attrition.*

4.1.5 Employee Job Retention Prediction: Classifying Individuals Likely to Leave Using Support Vector Machine (SVM)

The following graph represents the workflow in Orange, highlighting the steps involved in SVM-based preprocessing and analysis. The workflow outlines a typical machine learning pipeline, beginning with data import and culminating in model evaluation. The process involves several key stages, such as data preprocessing, feature engineering, and model training, followed by the assessment of model performance. Specifically, the focus on Support Vector Machine (SVM) indicates the use of a classification algorithm, which is designed to distinguish between different classes based on the input features. The model's effectiveness is evaluated using metrics like the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate against the false positive rate, offering a visual representation of the model's diagnostic ability. Additionally, the confusion matrix is employed to measure the number of correct and incorrect predictions, helping to identify any potential misclassifications. Performance curves provide further insights into the model's behavior, allowing for a more thorough evaluation of its predictive power and generalization to unseen data.



This visual workflow represents a **Support Vector Machine (SVM) pipeline** created using the **Orange** data mining tool. The key steps include:

1. Data Import: A CSV file is imported as the input dataset.

2. Data Preprocessing: Steps like table viewing, column selection, and data cleaning occur.

3. Data Sampling: The dataset is split into training and testing samples.

4. SVM Model: A Support Vector Machine (SVM) algorithm using a widget is applied for classification.

5. Model Evaluation:

- **Test and Score:** Performance metrics are calculated based on test data. The SVM model is performing poorly based on these metrics. While it has relatively high precision, its low AUC, CA, F1, recall, and MCC scores indicate that it struggles to accurately classify instances and distinguish between classes.

Model	AUC	CA	F1	Precision	Recall	MCC
SVM	0.434	0.269	0.152	0.632	0.269	0.005

- **Confusion Matrix:** It displays the accuracy of predictions compared to actual outcomes. In our case, the model seems to have a significant bias towards predicting class 1.0 as a large number of

class 0.0 instances are misclassified as class 1.0. While it correctly classifies a good number of 1.0 instances (3107), it also misclassifies a large number of 0.0 instances (9260). The model's performance is likely not very good, especially in identifying class 0.0 correctly.

Predicted ➡️ Actual ⬇️	0.0	1.0	Sum
0.0	347	9260	**9607**
1.0	110	3107	**3217**
Sum	**457**	**12367**	**12824**

- **Performance Curve:** Plots relevant metrics to analyze the model's performance trends.

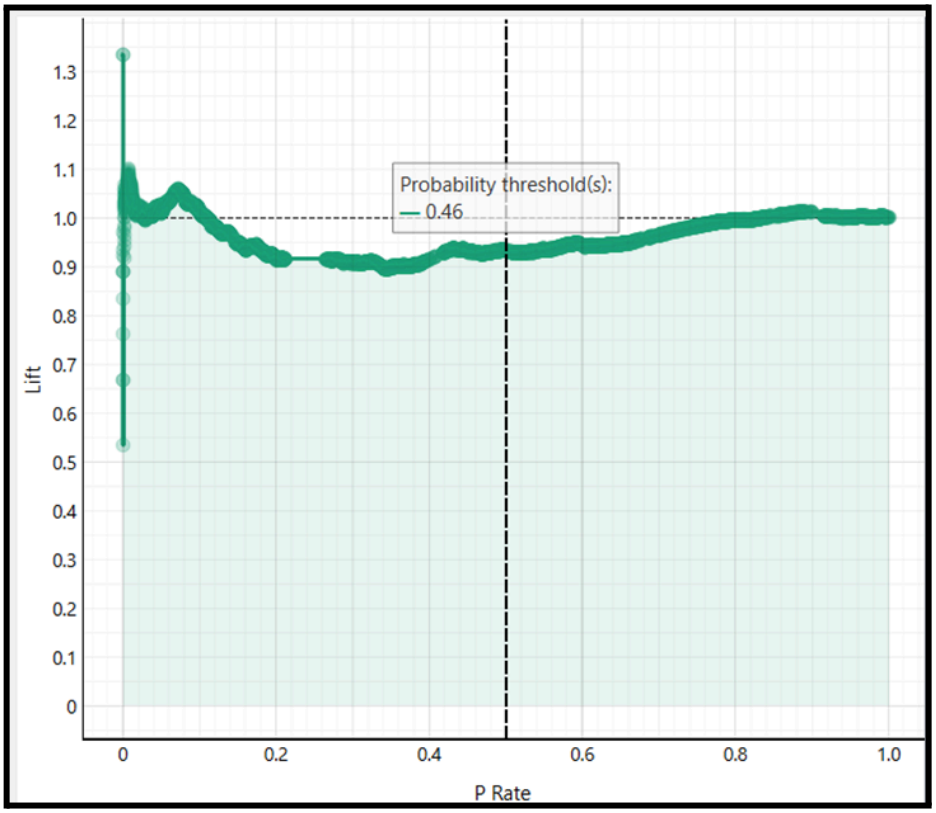


Figure 18

Fig.18: *The lift curve presented indicates that the model exhibits predictive power, particularly when targeting a small segment of the population with the highest predicted probabilities. At the chosen threshold of 0.46, the lift remains above 1.0, though it begins to decline, suggesting that the model's effectiveness is still notable but slightly diminishes as the threshold is reached. This drop in lift indicates a trade-off between precision and recall: higher thresholds typically focus on smaller, more accurately predicted groups, while lower thresholds capture larger populations with reduced accuracy. The ideal threshold selection hinges on the specific use case, balancing*

the need to accurately target a smaller group against the desire to expand coverage to a larger group with potentially lower prediction quality. This trade-off is crucial in determining the most appropriate threshold, tailored to the application's requirements.

- **ROC Analysis:** Receiver Operating Characteristic curves are generated to evaluate model performance.

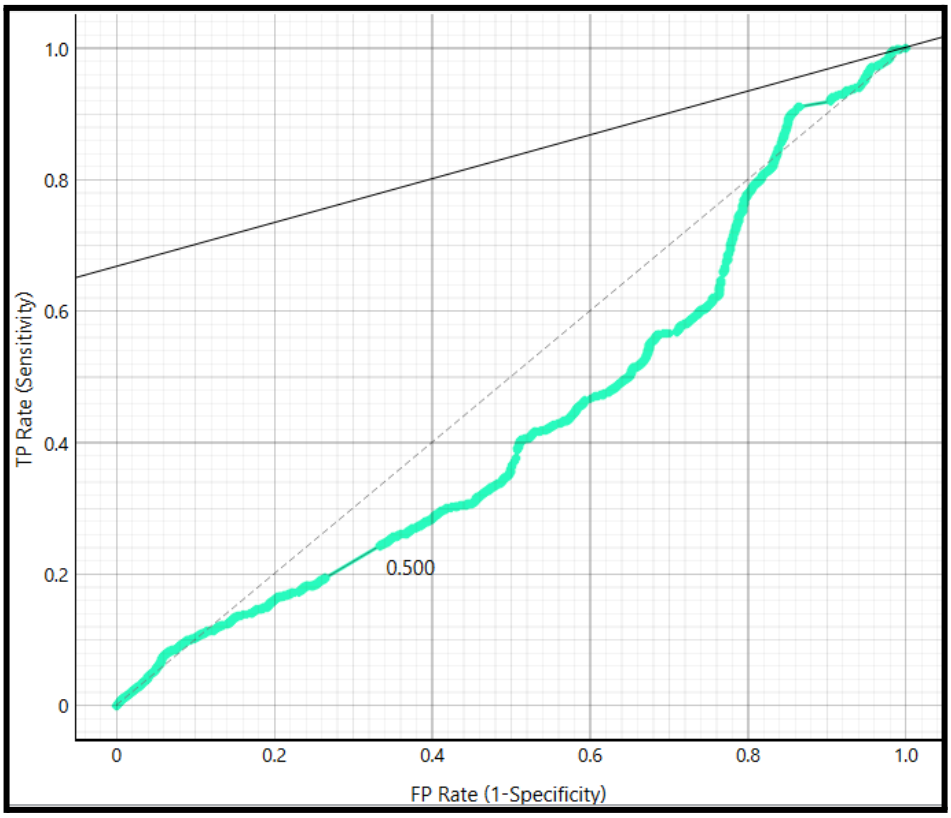


Figure 19

Fig.19: *The ROC curve displayed in the image is close to the diagonal line, which represents random performance. This positioning suggests that the model under evaluation has poor discriminatory power, with its ability to distinguish between the two classes not significantly better than random chance. The curve's proximity to the diagonal indicates that the model is struggling to effectively identify true positives (TP) while avoiding false positives (FP). In an ideal scenario, the ROC curve would be near the top-left corner, reflecting a high true positive rate (TPR) and a low false positive rate (FPR), which would signify good model performance. This would indicate that the model is accurately classifying instances of the positive class while minimizing errors. The current curve, however, suggests the need for model improvement, either through tuning, feature engineering, or using different algorithms to enhance its predictive accuracy.*

This workflow illustrates a complete ***machine learning pipeline*** from data ingestion to model evaluation using the SVM approach. It is an efficient method for preprocessing and exploring data. Integration of models like SVM can help predict job change probabilities with high accuracy.

4.2 Discussions

The analysis of job change decisions provided significant insights into the influence of key variables such as city development index, gender, education level, company size, company type, and professional experience.

- **City Development Index:** The results revealed that candidates from cities with higher development indices exhibit greater job mobility. This can be attributed to improved infrastructure, living standards, and availability of better opportunities in these areas.
- **Gender-Based Differences:** Males showed a significantly higher tendency to switch jobs compared to females. Further analysis revealed gender disparities in education levels, indicating fewer opportunities for women in higher education programs, which may limit their career progression and retention.
- **Company Size and Type:** Larger companies (1000+ employees) demonstrated better retention rates, likely due to greater stability, resources, and employee benefits. Conversely, smaller organizations faced higher attrition. Additionally, public sector companies retained employees better than private sector companies, highlighting job security as a crucial factor.
- **Professional Experience:** Employees with 5-10 years of experience and recent job changes (within the last year) were more likely to seek new opportunities. However, candidates with relevant experience tended to stay longer in their current roles, underscoring the value of experience in job retention.
- **SVM-Based Prediction:** The Support Vector Machine (SVM) model, implemented in Orange, effectively classified individuals likely to switch jobs. Performance evaluation metrics like ROC curves, confusion matrix, and performance curves validated the model's accuracy and robustness.

These findings emphasize the interplay of individual, organizational, and geographic factors in shaping job change decisions and provide HR professionals with actionable insights for improving retention strategies.

5. Statistical Modelling

5.1 Data Pre-Processing

Data preprocessing is a critical step in statistical modelling using python as it ensures that the data is in a suitable format for training models and helps improve their performance. In this section, several preprocessing steps were carried out, including handling missing values, feature transformation, scaling, encoding categorical variables, and converting data types. Below is a step-by-step breakdown of the data preprocessing procedure, including mappings and type conversions, to ensure that the dataset was properly prepared for statistical modelling.

- **Loading the Dataset:** The dataset was loaded into a Pandas DataFrame, which is a common data structure used for data manipulation in Python. This allowed us to access, manipulate, and inspect the data more easily.
- **Mapping and Transformation of Target Variable:** The target variable (y) contains the classes to be predicted. Initially, the classes in the target variable were labeled as 1 and 2, which are not suitable for binary classification models. Therefore, we mapped the values 1 to 0 and 2 to 1. This step ensures that the target variable is binary, which is required by the models.

- **Handling Missing Values:** Missing data is a common issue in many real-world datasets. We handled missing values for numerical and categorical columns in the following ways:
    - **Numerical Columns:** Missing values in numerical columns were filled with the mean of the respective columns. This was done using the 'fillna' method.
    - **Categorical Columns:** Missing values in categorical columns were filled with the mode (the most frequent value) of the respective columns.
- **Feature Engineering and Selection:** After ensuring that there were no missing values, we selected the relevant features (X) and the target variable (y). We dropped columns that were not relevant to the predictive modeling (e.g., identifiers like enrollee_id).
- **Categorical Variable Encoding:** Machine learning algorithms cannot directly work with categorical variables, so we needed to convert them into numerical values. We used the OneHotEncoder to perform one-hot encoding on categorical features. This approach transforms each category value into a new binary column, where each column represents a category, with 1 indicating the presence of that category and 0 indicating its absence.
- **Standardization of Numerical Features:** Numerical features often have different scales (e.g., income might range from 1000 to 100,000, while age ranges from 18 to 80), which can lead to models giving more importance to features with larger scales. To avoid this issue, we standardized the numerical features using StandardScaler. Standardization transforms the features to have a mean of 0 and a standard deviation of 1, ensuring that all features contribute equally to the model.

These preprocessing steps were essential for preparing the data in a way that maximized the performance of the models. Proper handling of missing values, encoding categorical variables, scaling features, and using a consistent preprocessing pipeline ensured that the models were trained on clean and well-processed data, leading to more reliable predictions.

5.2 Model Training

- **Splitting Data into Training and Test Sets:** Once the data was preprocessed, we split the dataset into training and testing sets using an 80-20 split. This step ensures that the model is trained on one portion of the data and evaluated on an unseen portion to avoid overfitting.
- **Pipeline Setup:** We set up a pipeline that combines the preprocessing and the model fitting steps. This ensures that the same preprocessing steps are applied consistently during both training and testing. This is especially important for ensuring that the model never sees the test data during training (i.e., no data leakage occurs).
- **Model Training:** After preprocessing, we trained several models using the preprocessed data, using **Logistic Regression.** For LR, the training process was straightforward using the pipeline, which automatically applied the necessary transformations to the data before fitting the model.
- **Steps taken for model Improvement**
    - **Addressed Class Imbalance:** Used techniques like SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset. Adjusted the classification threshold to improve recall for class 1.
    - **Feature Engineering:** Added interaction terms or non-linear transformations of predictors to capture non-linear patterns.

- **Regularization:** Used L1 or L2 regularization to prevent overfitting and improve generalization.

5.3 Model Evaluation

After training the models, we evaluated them using various metrics such as accuracy, precision, recall, and F1-score. These metrics were calculated using classification_report from scikit-learn, which provides a detailed report for binary classification tasks.

5.3.1 Performance Metrics

Accuracy	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)
78.8%	82%	63%	92%	39%

Insights: Performs well for class 0 but struggles with class 1 due to lower recall.

5.3.2 Observations:

- **High Precision for Class 0:** Logistic Regression is highly precise for class 0, meaning when it predicts class 0, it is correct 82% of the time.
- **Low Recall for Class 1:** The recall for class 1 is only 39%, indicating the model fails to identify many instances of class 1 (false negatives are high).
- **Class Imbalance:** The discrepancy in recall between the two classes suggests potential class imbalance or feature bias favoring class 0.

5.3.3 Critical Evaluation

- **Strengths:** Logistic Regression provides an interpretable model where the relationship between predictors and the target variable is linear. Its performance (78.8% accuracy) is competitive with other models in this analysis.
- **Weaknesses:** Struggles with detecting instances of class 1, as reflected in low recall. Assumes linear relationships between predictors and log-odds, which may not capture complex patterns in the data.

6. Conclusions and Future Work

6.1 Conclusions

This study successfully identified several key factors influencing job change decisions among candidates. Through a comprehensive exploratory data analysis (EDA) and visualization using tools like Power BI, Orange, and Python, the research uncovered significant insights:

- City Development Index (CDI): A strong positive correlation exists between CDI and job mobility. Individuals in cities with a higher development index are more likely to seek new job opportunities, likely due to better infrastructure, living standards, and a more competitive job market.
- Gender: Males exhibit a higher tendency to switch jobs compared to females. This disparity may be linked to broader societal factors, including fewer educational opportunities for women in advanced fields as seen in figure 13, potentially limiting their career progression and job satisfaction.

- Company Size and Type: Larger organizations (1000+ employees) and public sector companies generally have better employee retention rates. This can be attributed to factors like job security, stability, established career paths, and comprehensive benefits packages often associated with such entities.
- Professional Experience: Candidates with 5-10 years of experience and those who have recently changed jobs (within the last year) are more prone to seeking new employment. This suggests that individuals in this experience bracket are actively pursuing career growth and may be more open to exploring different opportunities. Also, candidates with relevant experience tend to stay longer at their current jobs.
- Education: Individuals without formal enrollment in education show a higher tendency to switch jobs as compared to people with formal education.
- Predictive Modeling: While initial attempts with the Support Vector Machine (SVM) model in Orange showed limited predictive power, especially in identifying class 0 correctly, these results lay the groundwork for future model refinement. After doing statistical modelling, Logistic Regression achieved an accuracy of 78.8%, with high precision for class 0 (82%) but lower recall for class 1 (39%).

In essence, the study highlights the complex interplay between personal attributes, organizational characteristics, and broader environmental factors in shaping job mobility. These findings provide valuable insights for HR professionals and policymakers aiming to improve employee retention and address workforce mobility challenges.

6.2 Future Work

Building on the insights and limitations of this study, future work should focus on the following areas:

- **Addressing Class Imbalance and Enhancing Predictive Models:**
    - **Data Augmentation:** Given the identified class imbalance issue, especially in predicting job changes (target = 1), employ techniques like SMOTE (Synthetic Minority Over-sampling Technique) or ADASYN (Adaptive Synthetic Sampling Approach) to generate synthetic samples for the minority class.
    - **Advanced Algorithms:** Explore more sophisticated algorithms such as Random Forest, Gradient Boosting Machines (GBM), and XGBoost, which often perform well on imbalanced datasets.
    - **Hyperparameter Tuning:** Conduct rigorous hyperparameter tuning using techniques like GridSearchCV or RandomizedSearchCV to optimize model performance.
    - **Ensemble Methods:** Experiment with ensemble methods (e.g., Bagging, Boosting, Stacking) to improve prediction accuracy and robustness.
- **Feature Engineering and Selection:**
    - **Interaction Terms:** Create interaction terms between existing features (e.g., CDI and company size, gender and education level) to capture complex relationships that may influence job change decisions.
    - **Dimensionality Reduction:** Apply dimensionality reduction techniques like Principal Component Analysis (PCA) or t-distributed Stochastic Neighbor Embedding1 (t-SNE) to reduce noise and potentially improve model performance.

    - **Feature Importance Analysis:** Use feature importance scores from tree-based models to identify the most influential predictors and potentially engineer new features based on these insights.
- **Deeper Dive into Gender Disparities:**
    - **Qualitative Research:** Conduct surveys or interviews to understand the specific reasons behind the observed gender differences in job mobility. Explore factors such as work-life balance, career advancement opportunities, and workplace culture.
    - **Industry-Specific Analysis:** Analyze gender differences in job-seeking behavior within specific industries to identify sectors with particularly high or low gender disparities.
- **Longitudinal Study:**
    - **Track Candidates Over Time:** Implement a longitudinal study design to track the same individuals over multiple time points. This would allow for a better understanding of how factors influencing job change decisions evolve over time.
- **Incorporating External Data:**
    - **Economic Indicators:** Integrate external data sources, such as regional unemployment rates, cost of living indices, and industry growth rates, to provide a more comprehensive understanding of the factors driving job mobility.
- **Advanced Statistical Analysis:**
    - **Survival Analysis:** Apply survival analysis techniques to model the time until an employee leaves a company, providing insights into employee tenure and attrition patterns.
    - **Multilevel Modeling:** If data is available at different hierarchical levels (e.g., individual, team, company), employ multilevel modeling to account for the nested structure of the data.
- **Developing an Interactive Dashboard:**
    - **Real-Time Insights:** Create an interactive dashboard using tools like Tableau or Power BI that allows HR professionals to explore the data, visualize trends, and track key metrics related to employee retention and job mobility in real-time.

By addressing these areas, future research can build upon the foundational insights of this study to develop more accurate predictive models, uncover deeper insights into the drivers of job change, and provide more actionable recommendations for improving employee retention strategies.