

BA706 Analytic Modeling

HR Analytics: Job Change Prediction of Data Scientists

For David Parent

Submitted by:

Deo Dominic Fajardo (301243074)

Marian Grace Ilagan (301236806)

Jeriel Madamba (301219390)

Justine Tinio (301241837)

Due date: December 16, 2022

| | | |
|----------|--|-----------|
| 1 | <i>Introduction</i> | 3 |
| 1.1 | Problem Statement | 3 |
| 1.2 | Objectives and Measurement | 4 |
| 2 | <i>Data Source</i> | 4 |
| 2.1 | Dataset Introduction | 4 |
| 2.2 | Initial Data Preparation | 5 |
| 3 | <i>Data Exploration</i> | 6 |
| 3.1 | Preliminary Exploration | 6 |
| 3.2 | Data Exploration and Visualization | 8 |
| 3.3 | EDA Summary | 19 |
| 4 | <i>Data Preparation</i> | 19 |
| 4.1 | Feature selection | 20 |
| 4.2 | Missing values – Imputation | 21 |
| 4.3 | Non-numeric values – Recode | 22 |
| 4.4 | Extreme values | 24 |
| 4.5 | Data Partition | 25 |
| 5 | <i>Model Exploration</i> | 26 |
| 5.1 | Decision Trees | 26 |
| 5.2 | Logistic Regression | 32 |
| 5.3 | Neural Network | 38 |
| 6 | <i>Results and Analysis</i> | 41 |
| 6.1 | Performance Measure / Model Assessment (ROC/ASE) | 41 |
| 6.2 | Feature Importance | 43 |
| 6.3 | Best model results | 44 |
| 7 | <i>Conclusion and Recommendations</i> | 45 |
| 7.1 | Conclusion | 45 |
| 7.2 | Recommendation | 46 |
| 7.3 | Further Research | 47 |
| | <i>References</i> | 48 |

Table of Contents

1 Introduction

1.1 Problem Statement

Harvard Business Review reported that data scientist is one of the most in demand and lucrative jobs in the 21st century with a 256% increase in Indeed job postings last 2019 as well as an average salary of almost \$200,00 in US for experienced professionals (2022). However, despite of being well compensated, data science professionals were found to have high turnover rate with an average retention of only 1.7 years (Globe News Wire, 2022). According to Omdena, there are three main reasons why data scientists quit their jobs: a) the lack of employee engagement which can be due to stressful work environment, lack of professional and personal growth opportunities or inability to provide business value with their machine learning models; b) a mismatch in employer expectations where the actual responsibilities given to the data scientists do not meet the job description causing them to become unsatisfied with the role; and, c) the lack of development opportunities for data scientists who want to keep up with the new technologies and challenges for professional career development (2022).

Managing employee attrition and retention has been a major activity in Human Resource where predictive analytics can be applied. Predictive models can help on HR decisions and strategies such as hiring employees who has a higher chance of staying in the company, forecasting drop-off candidate rates during the recruitment process, and predicting attrition caused by poor management and unmet expectations. In turn, these data-driven decisions will help the whole organization with their cost-saving targets and achieve more streamlined process.

To study the employee attrition and retention of data scientists, we will be using a public dataset from Kaggle and identify the factors affecting their decision to quit or stay.

1.2 Objectives and Measurement

The main objective of this project is to generate a model that will best predict the likelihood of whether a data scientist plans to move to another company or not. It also aims to identify which important features affect the candidate's decision.

Several models will be created using three different machine learning techniques specifically Decision tree, Regression and Neural Networks. Model assessment will be conducted to compare the models' performance based on average squared error (ASE) and receiver operator characteristic (ROC) index.

2 Data Source

2.1 Dataset Introduction

Our dataset from Kaggle called "HR Analytics: Job Change of Data Scientists" contains data scientists' information with 19,158 observations and 14 features (including the target variable). This is a binary classification problem with the target feature consists of either 0 (candidate is not looking to change job) or 1 (candidate is looking for a job change). Sample data from the file is shown in Figure 1 below as well the description for each feature in Table 1.

Figure 1. Sample data from the dataset

| enrollee_id | city | city_development_index | gender | relevent_experience | enrolled_university | education_level | major_discipline | experience | company_size | company_type | last_new_job | training_hours | target |
|-------------|----------|------------------------|--------|-------------------------|---------------------|-----------------|------------------|------------|--------------|----------------|--------------|----------------|--------|
| 8949 | city_103 | 0.92 | Male | Has relevent experience | no_enrollment | Graduate | STEM | >20 | | | 1 | 36 | 1 |
| 29725 | city_40 | 0.776 | Male | No relevent experience | no_enrollment | Graduate | STEM | 15 | 50-99 | Pvt Ltd | >4 | 47 | 0 |
| 11561 | city_21 | 0.624 | | No relevent experience | Full time course | Graduate | STEM | 5 | | | never | 83 | 0 |
| 33241 | city_115 | 0.789 | | No relevent experience | | Graduate | Business Degree | <1 | | Pvt Ltd | never | 52 | 1 |
| 666 | city_162 | 0.767 | Male | Has relevent experience | no_enrollment | Masters | STEM | >20 | 50-99 | Funded Startup | 4 | 8 | 0 |
| 21651 | city_176 | 0.764 | | Has relevent experience | Part time course | Graduate | STEM | 11 | | | 1 | 24 | 1 |
| 28806 | city_160 | 0.92 | Male | Has relevent experience | no_enrollment | High School | | 5 | 50-99 | Funded Startup | 1 | 24 | 0 |
| 402 | city_46 | 0.762 | Male | Has relevent experience | no_enrollment | Graduate | STEM | 13 | <10 | Pvt Ltd | >4 | 18 | 1 |
| 27107 | city_103 | 0.92 | Male | Has relevent experience | no_enrollment | Graduate | STEM | 7 | 50-99 | Pvt Ltd | 1 | 46 | 1 |
| 699 | city_103 | 0.92 | | Has relevent experience | no_enrollment | Graduate | STEM | 17 | 10000+ | Pvt Ltd | >4 | 123 | 0 |
| 29452 | city_21 | 0.624 | | No relevent experience | Full time course | High School | | 2 | | | never | 32 | 1 |
| 23853 | city_103 | 0.92 | Male | Has relevent experience | no_enrollment | Graduate | STEM | 5 | 5000-9999 | Pvt Ltd | 1 | 108 | 0 |
| 25619 | city_61 | 0.913 | Male | Has relevent experience | no_enrollment | Graduate | STEM | >20 | 1000-4999 | Pvt Ltd | 3 | 23 | 0 |
| 5826 | city_21 | 0.624 | Male | No relevent experience | | | | 2 | | | never | 24 | 0 |

Table 1. Data dictionary - details for each feature

| Column | Description | Range of possible values |
|------------------------|--|--|
| enrollee_id | Unique ID for each candidate | |
| city_code | City code | |
| city_development_index | Development index of the city (scaled) | Between 0 - 1 |
| gender | Gender of the candidate | Male, Female, Other |
| relevent_experience | Relevant experience to data science | Has relevant experience, No relevant experience |
| enrolled_university | Type of university course enrolled to (if any) | Full time course, no enrollment, part time course |
| education_level | Highest educational attainment of the candidate | Primary School, High School, Graduate, Masters, Phd |
| major_discipline | Education major discipline of candidate | STEM, Business Degree, Arts, Humanities, No Major, Other |
| experience | Candidate's total experience in years | <1, 1 to 20, >20 |
| company_size | No of employees in current employer's company. | Below 10, 10/49, 50-99, 100-500, 500-999, 1000-4999, 5000-9999, 10000+ |
| company_type | Type of current employer | Early Stage Startup, Funded Startup, NGO, Other, Public Sector, Pvt Ltd. |
| last_new_job | Difference in years between previous job and current job | 1,2,3,4,>4, never |
| training_hours | Number of training hours completed | Between 1 to 336 |
| target | 0 – Not looking for a job change 1 – Looking for a job change | 0, 1 |

2.2 Initial Data Preparation

Preliminary checking of the data file showed that it needs to be cleaned first before uploading in SAS. The special characters in the file have caused the variables that we want to consider as interval to be nominal. To preserve the original information, additional columns were created

which includes the corrected data. Table 2 summarizes the new columns added to the file, the reason for correction and changes that were done.

Table 2. Corrections done in the file before upload

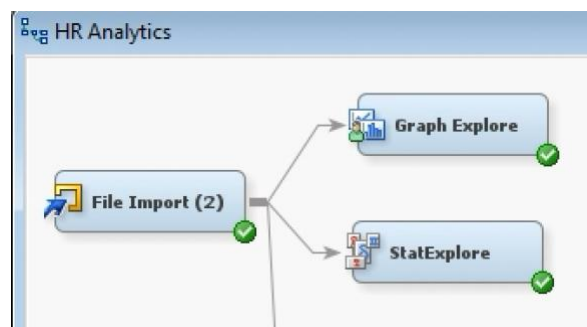
| New columns added | Data issue | Correction done |
|----------------------|---|--|
| experience_cleaned | Cannot be changed as interval in SAS due to < and > characters | Changed <1 to 0.5 and >20 to 21 to make this column an interval data. |
| company_size_cleaned | Cannot be changed as interval in SAS because the values are in range. One range is inputted as 10/49 instead of 10-49. | Used the midpoint of each range and rounding up to nearest whole number since this is count of employees. e.g. "5000-9999" = $(5000+9999)/2 = 7499.5 = 7,500$ |
| last_new_job_cleaned | Cannot be changed as interval in SAS due to > character and string "never". | Changed >4 to 5. "never" means the candidate didn't change their job. Considered never as blank instead and will impute in SAS. |

3 Data Exploration

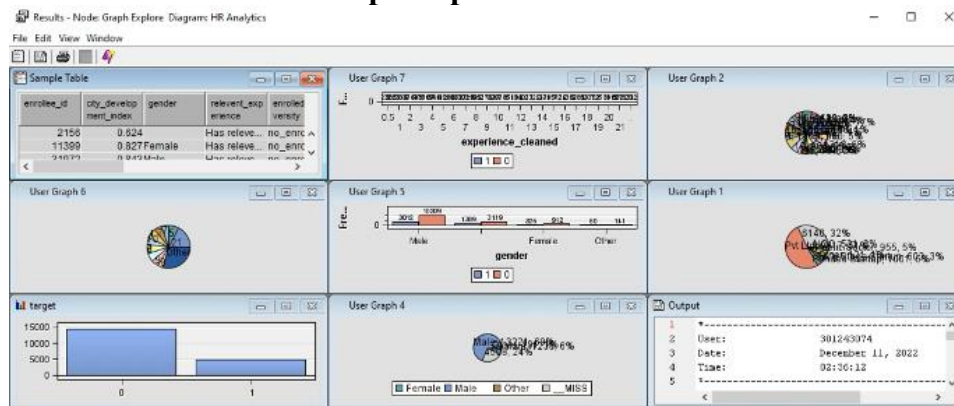
3.1 Preliminary Exploration

Using SAS Enterprise Miner, we can plot the data to find missing values, skewness, and other notable characteristics in each set of variables. We will analyze the data through bar graphs, pie charts, histograms, and summary statistics to explore the data through Graph Explore and Stat Explore node.

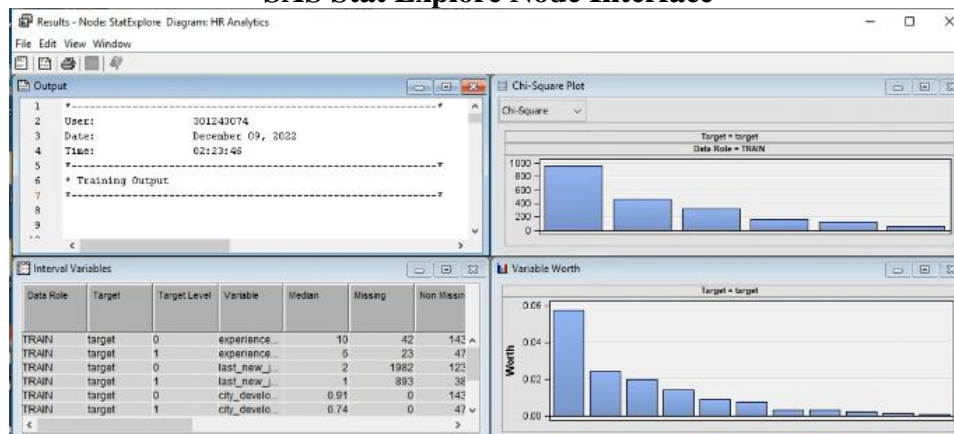
SAS Enterprise Miner Diagram with nodes, Graph Explore and StatExplore



SAS Graph Explore Node Interface



SAS Stat Explore Node Interface



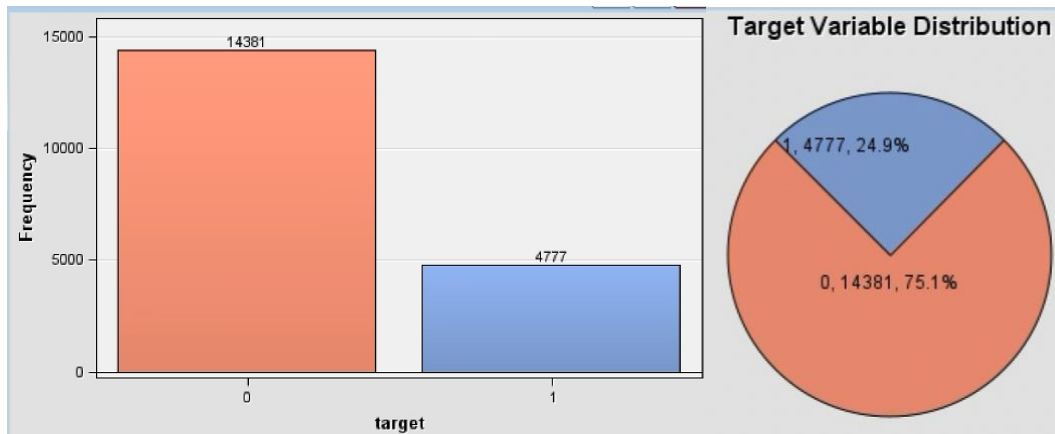
The findings in these preliminary explorations will be our basis in determining the features of our model, including which variables are to be imputed, recoded, and transformed (if necessary).

3.2 Data Exploration and Visualization

The following consists of plots and key findings generated for each variable.

TARGET

Figure 2. Distribution of target variable

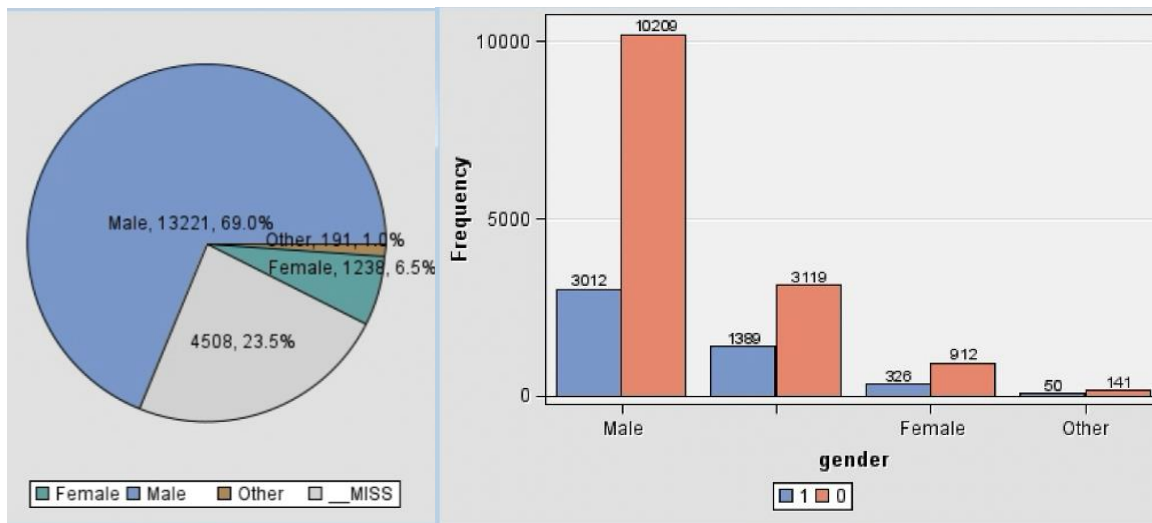


Key Findings:

- Data set is relatively imbalanced between candidates looking for a job change (1) versus those who are not (0).
- Those candidates who are not looking for a job change make up most of the data at 75.1%.

GENDER

Figure 3. Percentage of Gender & Distribution of Target Variable by Gender

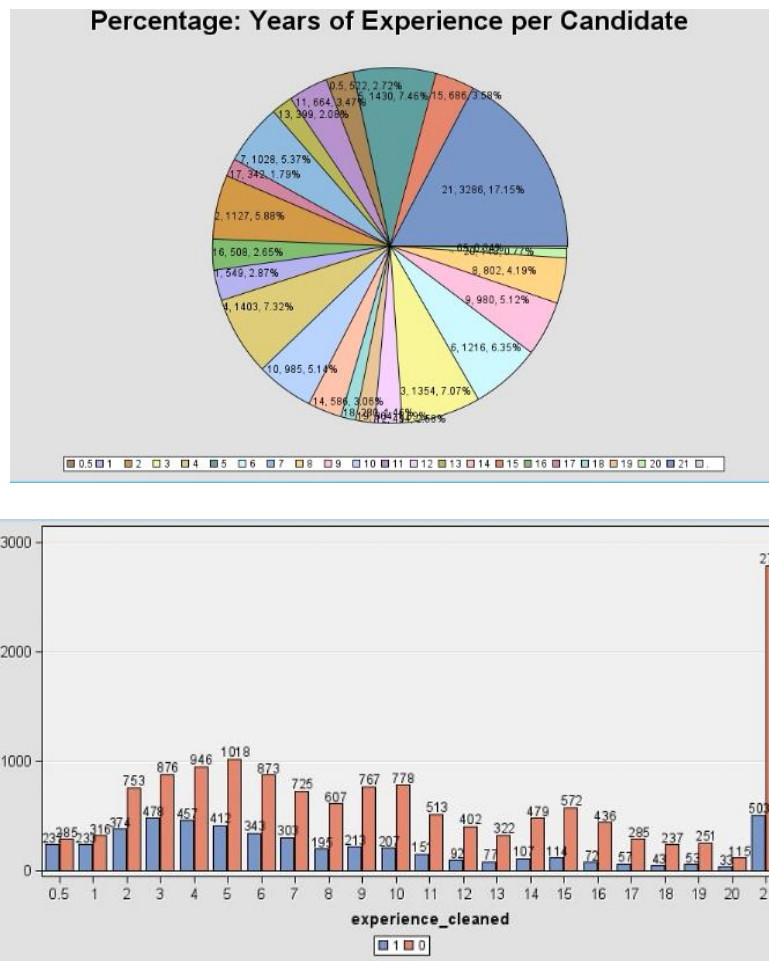


Key findings:

- Males compose most of the dataset at 69%. Female is at 6.5% and 'Other' is at 1.0%
- The remaining 23.5% are missing values which is still a noticeably large number.
- With the distribution of target variables based on Gender, Male are still most likely to both change jobs and not change jobs. On the other hand, the least likely for both targets are labelled as "Other".

EXPERIENCE

Figure 4. Percentage of Experience & Distribution of Target Variable by Experience

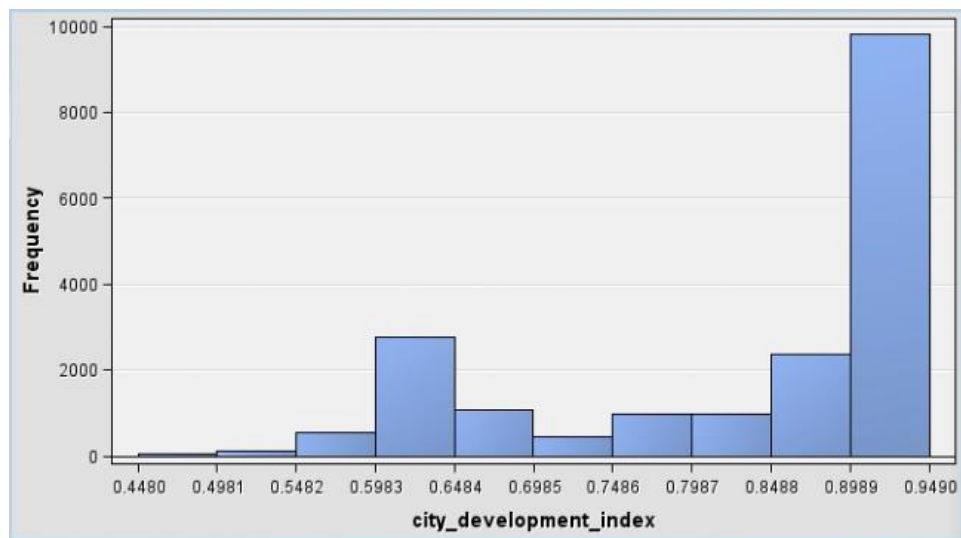


Key Findings:

- According to the pie chart, the mode of this variable is 21 years.
- There are only 65 records that are missing for this variable.
- Those with 21 years of experience lead the data at not looking for a job change, followed by 5 years.
- Those with 21 years of experience also make the most of those looking for a job change, followed by those with 3 years of experience.

CITY DEVELOPMENT INDEX

Figure 5. Histogram: City Development Index

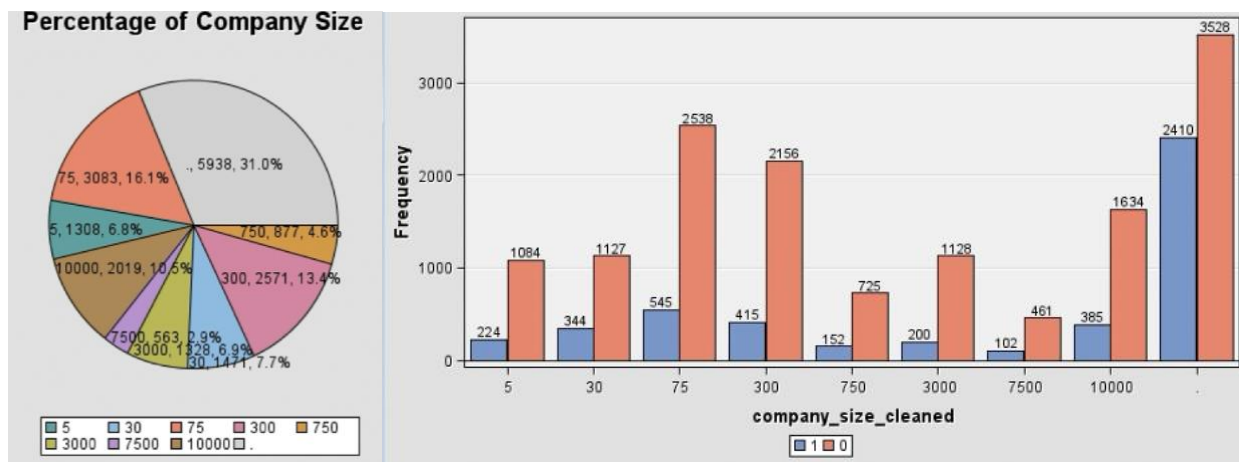


Key Findings:

- There are no missing values for the City Development Index variable.
- For this variable, the values are scaled between zero and 1.

COMPANY SIZE

Figure 6. Percentage of Company Size & Distribution of Target Variable by Company size

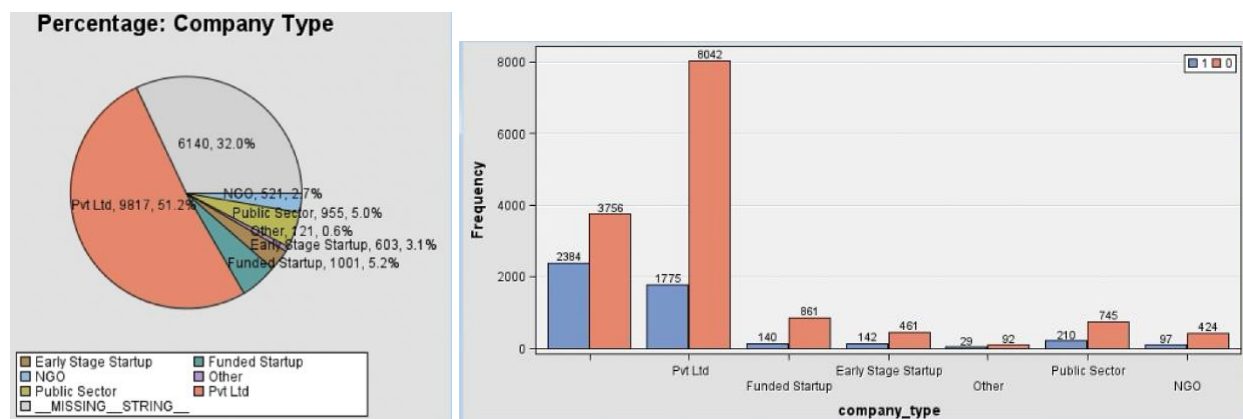


Key Findings:

- There are 5938 missing values for Company Size.
- Apart from the missing values, candidates coming from companies with 75 employees lead at changing and not changing jobs.
- Those coming from companies with 7500 employees fall behind at both mentioned target values.

COMPANY TYPE

Figure 7. Percentage of Company Type & Distribution of Target Variable by company type

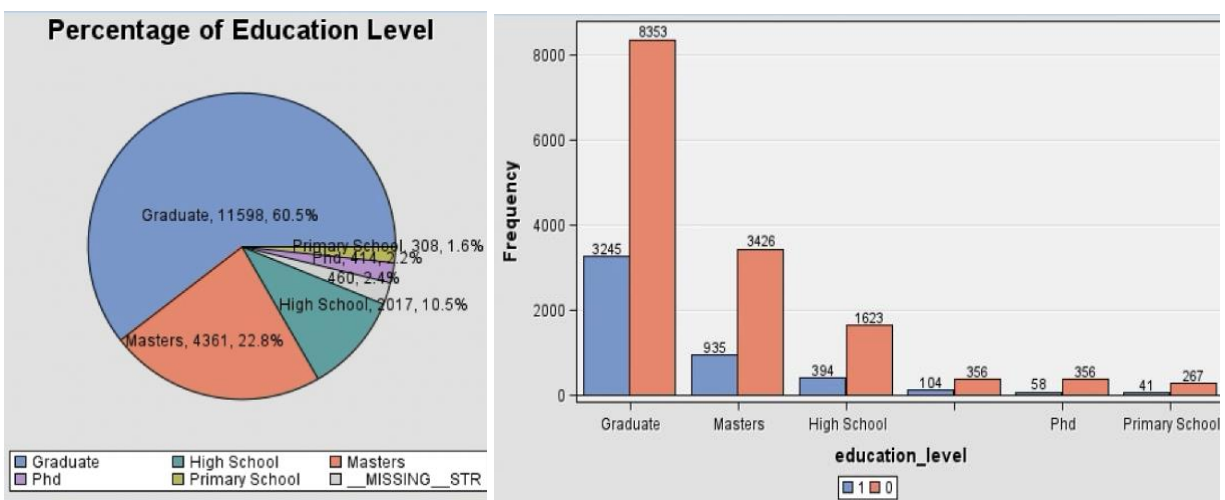


Key Findings:

- There are 6140 missing values within the Company Type Variable. This makes up 32% of the data.
- More than half of the data set are Pvt Ltd (51.2%).
- When the missing values are disregarded, Pvt Ltd employees leads the target variable at both job change seekers and non-seekers.

EDUCATION LEVEL

Figure 8. Percentage of Education Level & Distribution by Target Variable by Education level



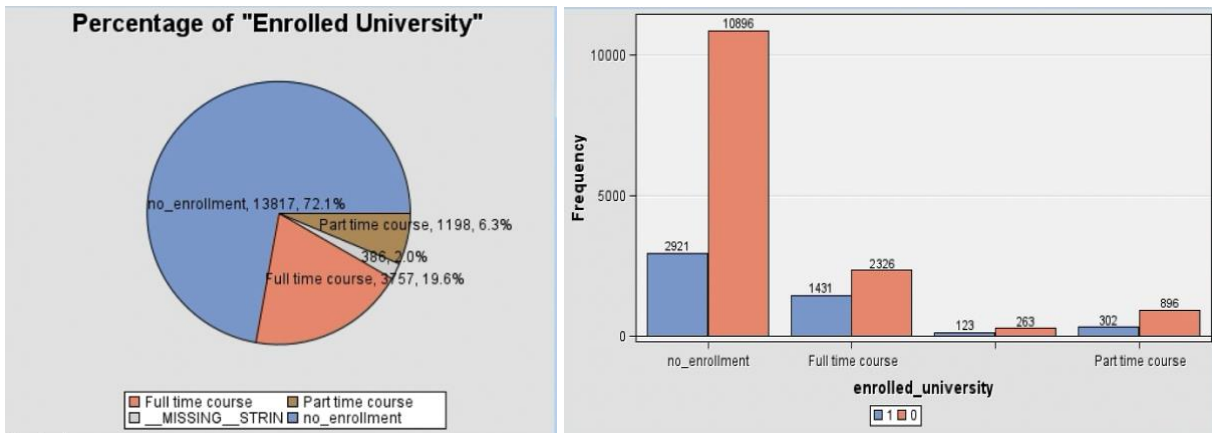
Key Findings:

- The missing values for this variable amount to 460, which is only 2.4% of the entire Education Level data set.
- “Graduate” consists the majority of the Education Level variable data at 60.5%.
- When distributed against the target variable, “Graduate” has the highest amount of both job seekers & non-seekers.

- We can choose to group the variables Graduate, Masters and PHD to reduce dimensionality.

ENROLLED UNIVERSITY

Figure 9. Percentage of Enrolled University & Distribution of Target Variable by type of University Course

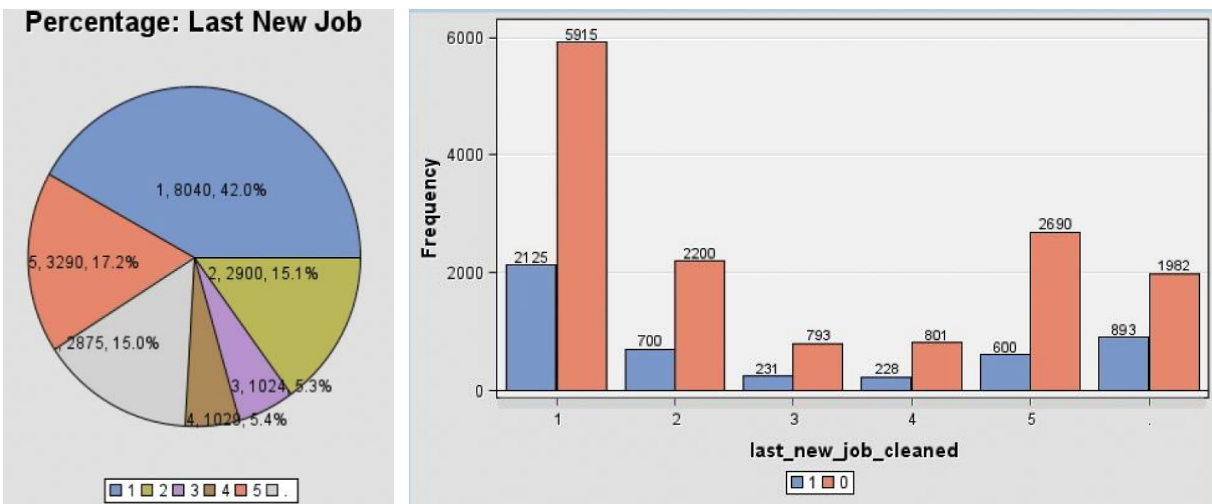


Key Findings:

- 2.0% of this variable are missing values, amounting to 386 only.
- The “no_enrollment” value dominates the variable data set with 72.1%, or 13,817 records.
- For both looking and not looking for job change, “no-enrollment” has the highest frequency when matched with the target variable.

LAST NEW JOB

Figure 10. Percentage of Last New Job & Distribution by Target Variable

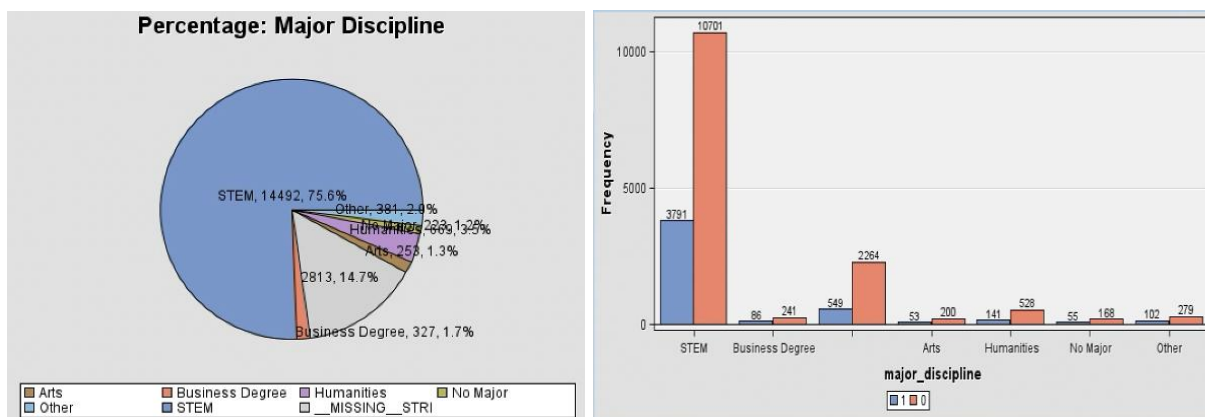


Key Findings:

- There are 2875 (15.0%) missing values for the Last New Job variable.
- The mode of the last_new_job variable is 1 year (difference between previous job and current job).
- When the variable Last New Job is distributed against the target variable, 1 year is still the highest count on both job changers and non-changers.

MAJOR DISCIPLINE

Figure 11. Percentage & Distribution of Target Variable by Major Discipline

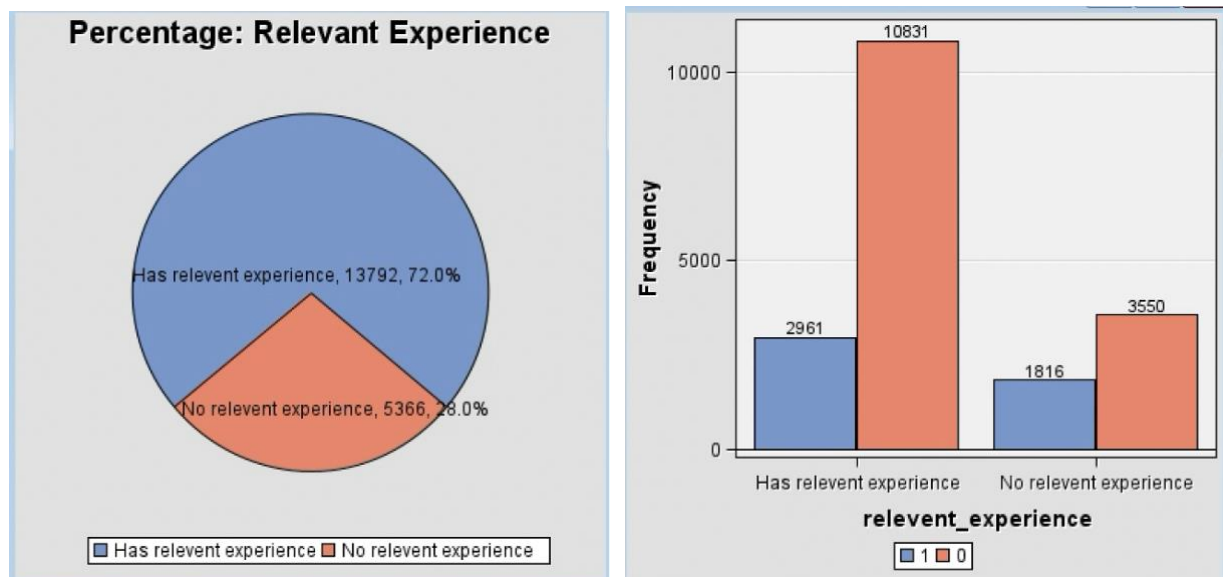


Key Findings:

- 14.7% of the Major Discipline variable data are missing, amounting to 2813.
- At 75.6% (14,492), the majority of this variable is “STEM”.
- There are 7 levels in the Major Discipline variable. 6 of these share the 24.4 percent minority.
- For the distribution of target variable by Major Discipline, STEM is found to be the most for both looking and not looking for a job change.

RELEVANT EXPERIENCE

Figure 12. Percentage of Relevant Experience & Distribution of Target Variable by Relevant experience

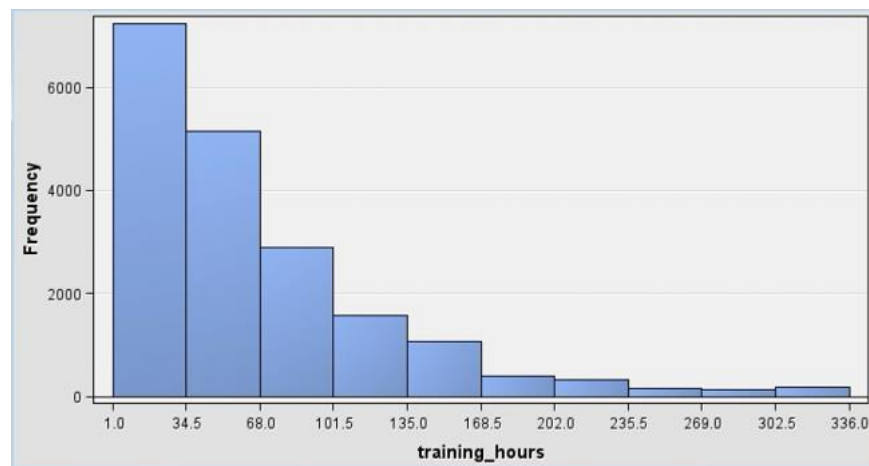


Key Findings:

- With only two levels for “Relevant experience”, there are no missing values for this variable.
- Those candidates that have relevant experience have the highest number of looking and not looking for job change (target variable).

TRAINING HOURS

Figure 13. Histogram: Training Hours



Key Findings:

- There are no missing values for the “Training hours” variable.
- The values are scaled between 1 and 336.

CITY

Figure 14. Sample Statistics of City Variable

| Sample Statistics | | |
|-------------------|--------------|------------------|
| Obs # | Variable ... | Number of Levels |
| 1 | city | 123 |
| 2 | enrollee_id | . |
| 3 | target | . |

Key Findings:

- The variable city has shown to have 123 number of levels, which is notably high and too specific to consider for the model.

SKEWNESS

Figure 15. Summary Statistics of Interval Variables



| Data Role | Target | Target Level | Variable | Skewness | Me |
|-----------|--------|--------------|------------------------|----------|----|
| TRAIN | target | 1 | training_hours | 1.856487 | |
| TRAIN | target | 0 | training_hours | 1.804095 | |
| TRAIN | target | 0 | company_size_cleaned | 1.427699 | |
| TRAIN | target | 1 | company_size_cleaned | 1.366018 | |
| TRAIN | target | 1 | last_new_job_cleaned | 1.065592 | |
| TRAIN | target | 1 | experience_cleaned | 0.888693 | |
| TRAIN | target | 0 | last_new_job_cleaned | 0.715807 | |
| TRAIN | target | 0 | experience_cleaned | 0.286495 | |
| TRAIN | target | 1 | city_development_index | 0.008548 | |
| TRAIN | target | 0 | city_development_index | -1.45179 | |

Key Findings:

- To observe the skewness, the summary statistics for the interval variables was generated.
- Training hours, company size, and last new job has shown the highest level of skewness in the list.

INTERVAL VARIABLES

Figure 16. Minimums & Maximums



| Data Role | Target | Target Level | Variable ▲ | Minimum | Maximum | Median | Missing |
|-----------|--------|--------------|------------------------|---------|---------|--------|---------|
| TRAIN | target | 0 | city_development_index | 0.448 | 0.949 | 0.91 | 0 |
| TRAIN | target | 1 | city_development_index | 0.448 | 0.949 | 0.74 | 0 |
| TRAIN | target | 0 | company_size_cleaned | 5 | 10000 | 300 | 3528 |
| TRAIN | target | 1 | company_size_cleaned | 5 | 10000 | 300 | 2410 |
| TRAIN | target | 0 | experience_cleaned | 0.5 | 21 | 10 | 42 |
| TRAIN | target | 1 | experience_cleaned | 0.5 | 21 | 6 | 23 |
| TRAIN | target | 0 | last_new_job_cleaned | 1 | 5 | 2 | 1982 |
| TRAIN | target | 1 | last_new_job_cleaned | 1 | 5 | 1 | 893 |
| TRAIN | target | 0 | training_hours | 1 | 336 | 48 | 0 |
| TRAIN | target | 1 | training_hours | 1 | 336 | 46 | 0 |

3.3 EDA Summary

Through the exploratory data analysis, it is observed that some variables have missing values. Those with missing values are Gender, Experience, Company Size, Company Type, Education Level, Enrolled University, Last New Job, and Major Discipline. Each missing data can be fixed through imputation later in the model.

When the interval variables' minimum and maximum values are examined, no unusual values are observed.

Additionally, some variables have been found to be too specific. City variable has 123 number of levels. Major discipline may be reduced through recoding. For example, we can apply recoding to this variable by using STEM and NON_STEM as our recoded values to prevent any unnecessary data dimensions. The same can be done for Education level. The variables Graduate, Masters, and PHD can be recoded into one as "Degree holder".

For skewness, training_hours, company_size, and last_new_job were observed to be the most skewed amongst all the interval variables. We can use data transformation for these if the values exceed the desired skewness threshold.

4 Data Preparation

In data preparation, we will define the variables that will be used in the models as well as the split of train and validation data. The imputation, replacement and transformation nodes that will be used by some models will be discussed here as well.

4.1 Feature selection

The file import node was used to import our dataset to SAS. Then the ‘target’ variable was set to binary, to flag if the observation is looking for a career change. Then, the following variables were rejected from our import:

Table 3. Rejected columns in SAS

| Variable Name | Reason for Rejection |
|---------------|---|
| city | The variable has 123 unique |
| company_size | This variable was cleaned using excel, the observations were changed from categorical to interval. The variable was replaced by “company_size_cleaned”. |
| experience | Cleaned using Excel, as the observations contains character (“<”, “>”) which cannot be used for interval. The following observations were change: “>20” = 21 “<1” = 0.5 The variable was replaced by “experience_cleaned”. |
| Last_new_job | Cleaned using Excel, as the observations contains character (“<”, “>”) which cannot be used for interval. The following observations were change: “>4” = 5 The variable was replaced by “last_new_job”. |

A total of thirteen variables will be employed in our models in its final form.

Figure 17. Feature selection

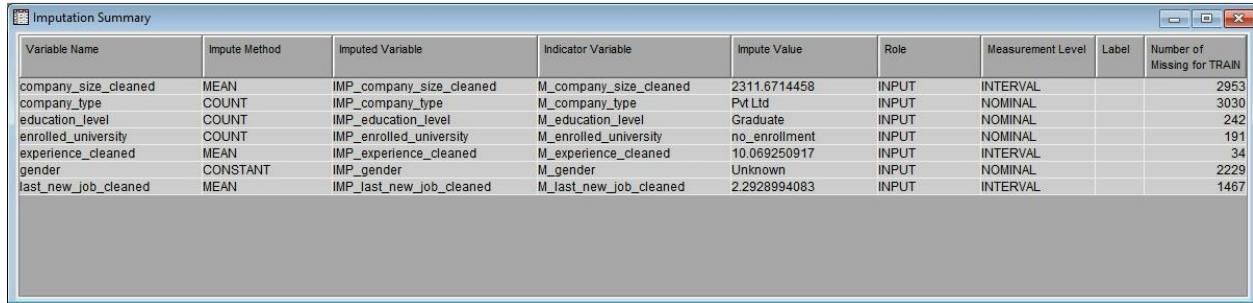
| Variables - FIMPORT2 | | | | | | | |
|--|----------|----------|--------|-------|------|-------------|-------------|
| <div> <div>(none) ▾</div> <div> <input type="checkbox"/> not Equal to ▾ <input type="text"/> </div> </div> <div> Columns: <input type="checkbox"/> Label <input type="checkbox"/> Mining <input type="checkbox"/> Basic </div> | | | | | | | |
| Name | Role | Level | Report | Order | Drop | Lower Limit | Upper Limit |
| city | Rejected | Nominal | No | | No | . | . |
| city_development_index | Input | Interval | No | | No | . | . |
| company_size | Rejected | Nominal | No | | Yes | . | . |
| company_size_cleaned | Input | Interval | No | | No | . | . |
| company_type | Input | Nominal | No | | No | . | . |
| education_level | Input | Nominal | No | | No | . | . |
| enrolled_university | Input | Nominal | No | | No | . | . |
| enrollee_id | ID | Nominal | No | | No | . | . |
| experience | Rejected | Nominal | No | | Yes | . | . |
| experience_cleaned | Input | Interval | No | | No | . | . |
| gender | Input | Nominal | No | | No | . | . |
| last_new_job | Rejected | Nominal | No | | Yes | . | . |
| last_new_job_cleaned | Input | Interval | No | | No | . | . |
| major_discipline | Input | Nominal | No | | No | . | . |
| relevant_experience | Input | Nominal | No | | No | . | . |
| target | Target | Binary | No | | No | . | . |
| training_hours | Input | Interval | No | | No | . | . |

4.2 Missing values – Imputation

To find correlation between variables, all observations should have no missing or null values. In order to use all observations, imputation can be implemented using descriptive statistics. For categorical variables, we will set the imputation to the most frequent observation, while for numerical variables we will use the mean of the variable. The new imputed variable will have a corresponding indicator variable (e.g. M_company_size_cleaned) which is a flag that denotes if the value of a specific variable was imputed, therefore missing. These imputation flags will help us interpret if the missing data is statistically significant to the model.

The missing observations for each variable were imputed as shown in Figure 18 below.

Figure 18. Imputation of missing values



| Variable Name | Impute Method | Imputed Variable | Indicator Variable | Impute Value | Role | Measurement Level | Label | Number of Missing for TRAIN |
|----------------------|---------------|--------------------------|------------------------|---------------|-------|-------------------|-------|-----------------------------|
| company_size_cleaned | MEAN | IMP_company_size_cleaned | M_company_size_cleaned | 2311.6714458 | INPUT | INTERVAL | | 2953 |
| company_type | COUNT | IMP_company_type | M_company_type | Pvt Ltd | INPUT | NOMINAL | | 3030 |
| education_level | COUNT | IMP_education_level | M_education_level | Graduate | INPUT | NOMINAL | | 242 |
| enrolled_university | COUNT | IMP_enrolled_university | M_enrolled_university | no_enrollment | INPUT | NOMINAL | | 191 |
| experience_cleaned | MEAN | IMP_experience_cleaned | M_experience_cleaned | 10.069250917 | INPUT | INTERVAL | | 34 |
| gender | CONSTANT | IMP_gender | M_gender | Unknown | INPUT | NOMINAL | | 2229 |
| last_new_job_cleaned | MEAN | IMP_last_new_job_cleaned | M_last_new_job_cleaned | 2.2928994083 | INPUT | INTERVAL | | 1467 |

Note that the impute node will be used by regression and neural networks only.

4.3 Non-numeric values – Recode

To reduce the curse of dimensionality, some input levels of the following categorical variables were consolidated using the replacement node.

- Recode Major discipline node** - The observations from the ‘major_discipline’ variable was compressed from seven categories to (STEM, Business Degree, Arts, Humanities, No Major, Other, Missing) to three categories (STEM, NO_COLLEGE, NON_STEM) using the replacement node. For the reason that we will not take any business decisions related to the other combination for NON_STEM disciplines. This recode node will be used by all models including decision tree.

Figure 19. Recode results for major discipline

The screenshot shows a software window titled 'Results - Node: Recode Major discipline Diagram: HR_Analytics 2'. It contains two main sections:

Total Replacement Counts

| Variable | Role | Label | Train |
|------------------|-------|-------|-------|
| major_discipline | INPUT | | 4666 |

Output

Replacement Values for Class Variables

| Variable | Formatted Value | Type | Character Unformatted Value | Numeric Value | Replacement Value | Label |
|------------------|-----------------|------|-----------------------------|---------------|-------------------|-------|
| major_discipline | | C | | . | NO_COLLEGE | |
| major_discipline | Humanities | C | Humanities | . | NON_STEM | |
| major_discipline | Other | C | Other | . | NON_STEM | |
| major_discipline | Business Degree | C | Business Degree | . | NON_STEM | |
| major_discipline | Arts | C | Arts | . | NON_STEM | |
| major_discipline | No Major | C | No Major | . | NON_STEM | |

b. Recode_gender_educationallevel_companytype node – This node combines the input levels of gender, educational level and company type. This node will be used by regression and neural network models.

- The gender variable contains four input levels (Male, Female, Unknown, and Other). To mitigate the decisions for the unknown, we recoded the observation by combining it with the other using the replacement node, then we assigned the observation name to U_O. Doing this can help our decision focus on three observations.
- For educational level feature, Graduate, Masters and Phd was recoded to Degree Holder because the educational attainment after secondary education did not make a big difference in the results. Input levels are reduced from five (Primary School, High School, Graduate, Masters and Phd) to three input levels (Primary School, High School and Degree Holder).
- For company type, input level is reduced from six to four: Startup (Early Startup and Funded Startup), Public_NGO (Public Sector and NGO), Pvt Ltd., and Other.

Figure 20. Recode results for company type, gender, and education level

The screenshot shows the 'Results - Node: Recode_gender_educationallevel_companytype' window. It contains two main sections: 'Total Replacement Counts' and 'Output'.

Total Replacement Counts

| Variable | Role | Label | Train | Validation |
|---------------------|-------|-------------------------|-------|------------|
| IMP_company_type | INPUT | Imputed company_type | 1556 | 1524 |
| IMP_education_level | INPUT | Imputed education_level | 8403 | 8430 |
| IMP_gender | INPUT | Imputed gender | 2333 | 2366 |

Output

31 Replacement Values for Class Variables

| Variable | Formatted Value | Type | Character Unformatted Value | Numeric Value | Replacement Value | Label |
|---------------------|---------------------|------|-----------------------------|---------------|-------------------|-------------------------|
| IMP_company_type | Funded Startup | C | Funded Startup | . | Startup | Imputed company_type |
| IMP_company_type | Public Sector | C | Public Sector | . | Public_NGO | Imputed company_type |
| IMP_company_type | Early Stage Startup | C | Early Stage Startup | . | Startup | Imputed company_type |
| IMP_company_type | NGO | C | NGO | . | Public_NGO | Imputed company_type |
| IMP_education_level | Graduate | C | Graduate | . | DEGREE HOLDER | Imputed education_level |
| IMP_education_level | Masters | C | Masters | . | DEGREE HOLDER | Imputed education_level |
| IMP_education_level | Phd | C | Phd | . | DEGREE HOLDER | Imputed education_level |
| IMP_gender | Unknow | C | Unknown | . | 0_U | Imputed gender |
| IMP_gender | Other | C | Other | . | 0_U | Imputed gender |

4.4 Extreme values

Extreme values or outliers can skew the data and make the model biased. Variables with skewed data distribution have skewness more than absolute value of 1 or 2. In our data, interval variables training_hours and IMP_company_size_cleaned are a bit skewed (skewness of 1.8 and 1.9) as shown in Figure 15 in EDA section. Transform nodes using cap and floor, and log transform will be prepared for regression and neural network models.

Cap and Floor

After cap and floor, training hours' skewness lowered down to 1.4 but IMP_company_size_cleaned did not changed.

The screenshot shows the 'Interval Variables' window with the following data:

| Variable | Replace Variable | Lower limit | Upper Limit | Label | Limits Method | Replacement Method | Lower Replacement Value | Upper Replacement Value |
|-----------------------|---------------------|-------------|-------------|------------------------------|---------------|--------------------|-------------------------|-------------------------|
| IMP_company_size... | REP_IMP_compan... | -6920.52 | 11543.86 | Imputed company_size_cleaned | STDDEV | COMPUTED | -6920.52 | 11543.86 |
| IMP_experience_cl... | REP_IMP_experie... | -10.1036 | 30.24211 | Imputed experience_cleaned | STDDEV | COMPUTED | -10.1036 | 30.24211 |
| IMP_last_new_job... | REP_IMP_last_ne... | -2.08169 | 6.667491 | Imputed last_new_job_cleaned | STDDEV | COMPUTED | -2.08169 | 6.667491 |
| city_development_i... | REP_city_develop... | 0.457216 | 1.199688 | city_development_index | STDDEV | COMPUTED | 0.457216 | 1.199688 |
| training_hours | REP_training_hours | -112.156 | 241.339 | training_hours | STDDEV | COMPUTED | -112.156 | 241.339 |

Results - Node: StatExplore (5) Diagram: HR_Analytics 2

File Edit View Window

Interval Variables

| Data Role | Target | Target Level | Variable | Median | Skewness | Missing | Non Missing | Minimum | Maximum | Mean |
|-----------|--------|--------------|------------------------------|----------|----------|---------|-------------|----------|---------|-------|
| TRAIN | target | 0 | REP_IMP_experience_cleaned | 9 | 0.307206 | 0 | 7190 | 0.5 | 21 | 10.72 |
| TRAIN | target | 1 | REP_IMP_experience_cleaned | 6 | 0.898531 | 0 | 2389 | 0.5 | 21 | 8.099 |
| TRAIN | target | 0 | REP_city_development_index | 0.91 | -1.4521 | 0 | 7190 | 0.457216 | 0.949 | 0.853 |
| TRAIN | target | 1 | REP_city_development_index | 0.738 | 0.02904 | 0 | 2389 | 0.457216 | 0.949 | 0.753 |
| TRAIN | target | 0 | REP_IMP_last_new_job_cleaned | 2 | 0.800238 | 0 | 7190 | 1 | 5 | 2.348 |
| TRAIN | target | 1 | REP_IMP_last_new_job_cleaned | 2 | 1.103426 | 0 | 2389 | 1 | 5 | 2.12 |
| TRAIN | target | 0 | REP_training_hours | 47 | 1.426707 | 0 | 7190 | 1 | 241.339 | 64.11 |
| TRAIN | target | 1 | REP_training_hours | 46 | 1.468203 | 0 | 2389 | 1 | 241.339 | 61.67 |
| TRAIN | target | 0 | REP_IMP_company_size_cleaned | 750 | 1.608674 | 0 | 7190 | 5 | 10000 | 2307. |
| TRAIN | target | 1 | REP_IMP_company_size_cleaned | 2311.671 | 1.942886 | 0 | 2389 | 5 | 10000 | 2323. |

Log Transform after cap and floor

Results - Node: Transform Variables Diagram: HR_Analytics 2

File Edit View Window

Transformations Statistics

| Source | Method | Variable Name | Skewness | Formula | Number of Levels | Non Missing | Missing | Minimum | Maximum | Mean | Standard Deviation | Kurtosis | Label |
|--------|----------|------------------------------|----------|----------------|------------------|-------------|---------|----------|----------|----------|--------------------|----------|-------------|
| Input | Original | IMP_company_size_cleaned | 1.674517 | | | 9579 | 0 | 5 | 10000 | 2311.671 | 3077.396 | 1.602315 | Imputed... |
| Input | Original | training_hours | 1.81574 | | | 9579 | 0 | 1 | 336 | 64.5914 | 58.91588 | 3.854468 | |
| Output | Computed | LOG_IMP_company_size_cleaned | -0.54498 | log(IMP_c... | | 9579 | 0 | 1.791759 | 9.21044 | 6.346903 | 2.156035 | -0.79101 | Transfor... |
| Output | Computed | LOG_training_hours | -0.35423 | log(trainin... | | 9579 | 0 | 0.693147 | 5.820083 | 3.793817 | 0.938802 | -0.1967 | Transfor... |

Log Transform without cap and floor

Results - Node: Transform Variables Diagram: HR_Analytics 2

File Edit View Window

Transformations Statistics

| Source | Method | Variable Name | Skewness | Formula | Number of Levels | Non Missing | Missing | Minimum | Maximum | Mean | Standard Deviation | Kurtosis | Label |
|--------|----------|------------------------------|----------|-----------------|------------------|-------------|---------|----------|----------|----------|--------------------|----------|-------------|
| Input | Original | IMP_company_size_cleaned | 1.674517 | | | 9579 | 0 | 5 | 10000 | 2311.671 | 3077.396 | 1.602315 | Imputed... |
| Input | Original | training_hours | 1.81574 | | | 9579 | 0 | 1 | 336 | 64.5914 | 58.91588 | 3.854468 | |
| Output | Computed | LOG_IMP_company_size_cleaned | -0.54498 | log(IMP_co... | | 9579 | 0 | 1.791759 | 9.21044 | 6.346903 | 2.156035 | -0.79101 | Transfor... |
| Output | Computed | LOG_training_hours | -0.35423 | log(training... | | 9579 | 0 | 0.693147 | 5.820083 | 3.793817 | 0.938802 | -0.1967 | Transfor... |

4.5 Data Partition

In supervised machine learning, the dataset is needed to be trained in order to predict other observation. In this project the dataset is then split into 50% train and 50% validation using the data partition node. Train dataset will be used to build the model while validation data will be used to check the model performance and avoid overfitting or underfitting. This dataset is now ready for modeling.

5 Model Exploration

5.1 Decision Trees

For creating classification systems based on several variables or for creating prediction algorithms for a target variable, decision tree methodology is a frequently used data mining technique. Using branch-like segments that form an inverted tree with a root node, internal nodes, and leaf nodes, this technique divides a population into subgroups (Song & Lu, 2015).

Advantage:

- Decision trees are easy to understand and interpret since they can be seen.
- In contrast to other models, decision trees are capable of handling both categorical and numerical observations.
- It requires less data preparation than other models that call for data normalization.

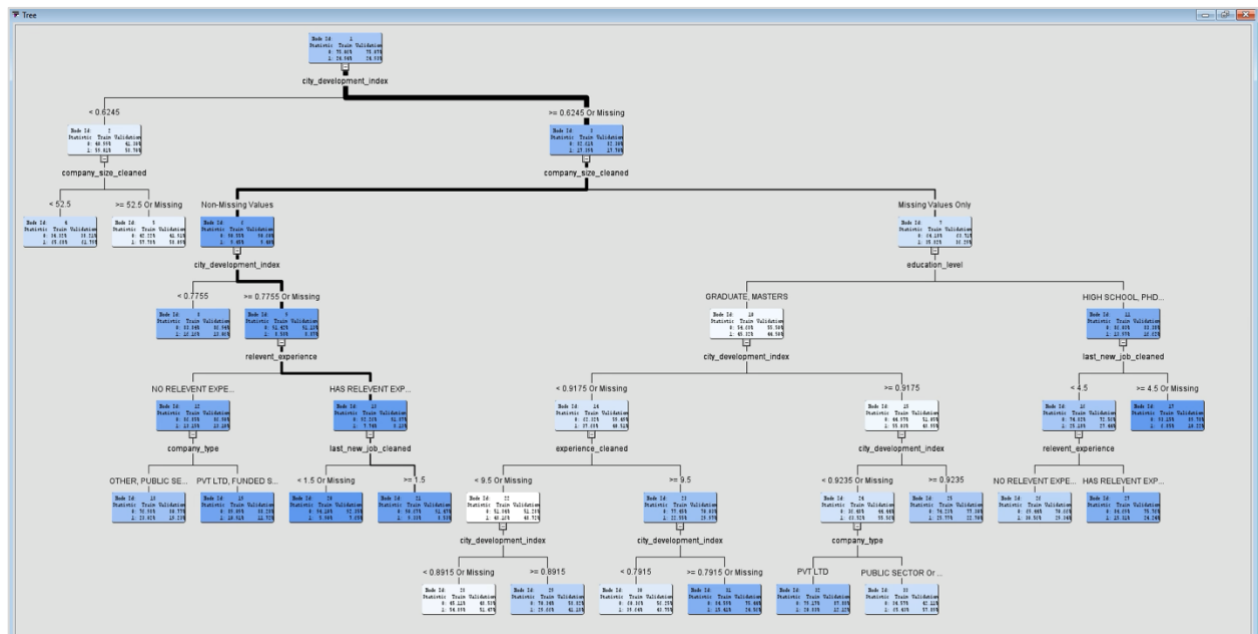
Disadvantage:

- does a poor job of handling data that is unbalanced and produces bias.
- Little changes in the data might result in different decision trees.

Maximal Decision Tree

It is a decision tree that has reached its full potential and has not yet undergone any pruning. To run a Maximal decision tree, we set the sub tree method to largest and the assessment measure to average square error. The result showed an ASE of 0.142565 with 17 leaves.

Figure 21. Maximal decision tree



| Competing Rules For Node 1 | | | | |
|----------------------------|------------------------|----------|--------------------|--|
| Split Variable | Variable Description | -Log(p) | Number of Branches | |
| city_development_index | city_development_index | 285.6961 | 2 | |
| company_size_cleaned | company_size_cleaned | 118.5231 | 2 | |
| company_type | company_type | 95.7444 | 2 | |
| experience_cleaned | experience_cleaned | 54.6879 | 2 | |
| enrolled_university | enrolled_university | 42.4411 | 2 | |

Variable Importance

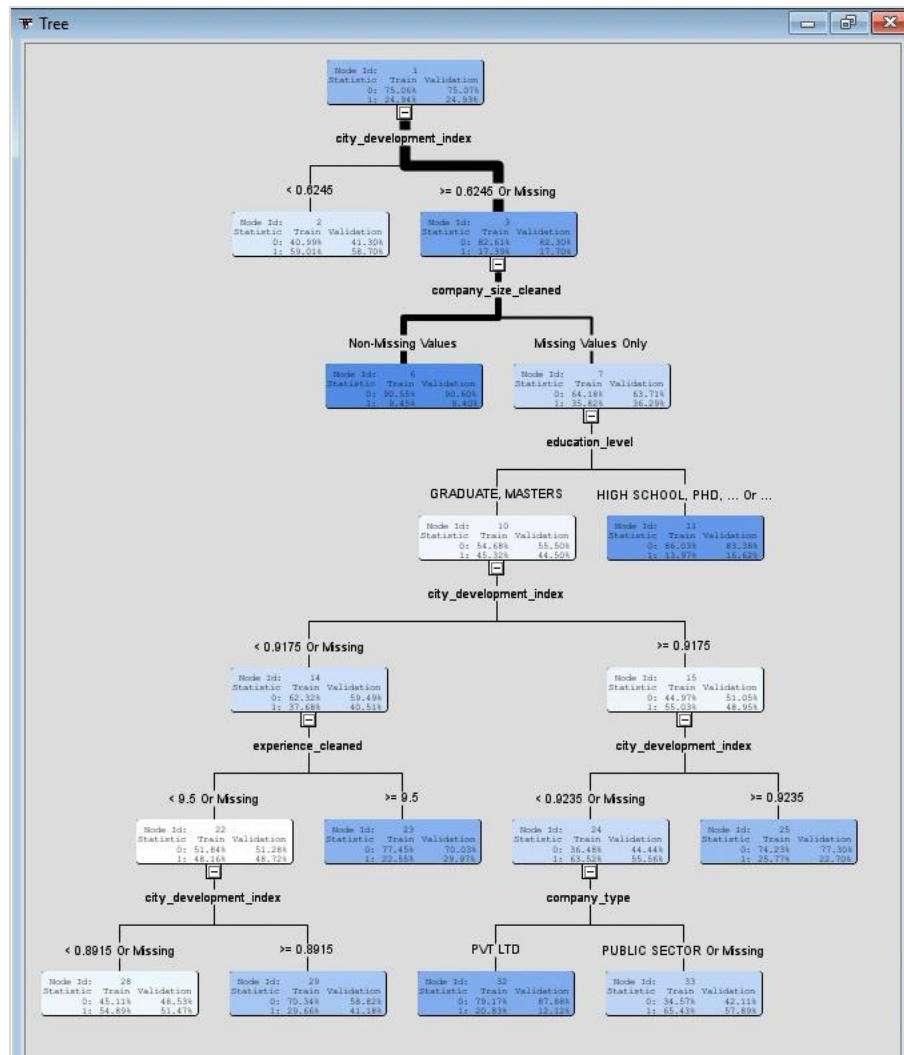
| Variable Name | Label | Number of Splitting Rules | Importance | Validation Importance | Ratio of Validation to Training Importance |
|------------------------|-------|---------------------------|------------|-----------------------|--|
| city_development_index | | 6 | 1.0000 | 1.0000 | 1.0000 |
| company_size_cleaned | | 2 | 0.6317 | 0.6982 | 1.1052 |
| education_level | | 1 | 0.4105 | 0.3937 | 0.9591 |
| experience_cleaned | | 1 | 0.2240 | 0.1596 | 0.7124 |
| last_new_job_cleaned | | 2 | 0.1541 | 0.1389 | 0.9014 |
| company_type | | 2 | 0.1463 | 0.1710 | 1.1687 |
| relevent_experience | | 2 | 0.1071 | 0.0767 | 0.7161 |

Based on logworth the competing variables for the first split are `city_development_index` and `company_size_cleaned`. For the variable importance the `company_size_cleaned` explains 63.17%, while `education_level` explains 41.05% of the variability explained by `city_development_index`.

Misclassification Tree

To find the optimal tree, pruning can be performed to the maximal tree using misclassification which is the percentage of results that differ from the predictions. The result showed an ASE 0.143443 with 9 leaves.

Figure 22. Misclassification tree



| Competing Rules For Node 1 | | | | |
|----------------------------|------------------------|----------|--------------------|--|
| Split Variable | Variable Description | -Log(p) | Number of Branches | |
| city_development_index | city_development_index | 285.6961 | 2 | |
| company_size_cleaned | company_size_cleaned | 118.5231 | 2 | |
| company_type | company_type | 95.7444 | 2 | |
| experience_cleaned | experience_cleaned | 54.6879 | 2 | |
| enrolled_university | enrolled_university | 42.4411 | 2 | |

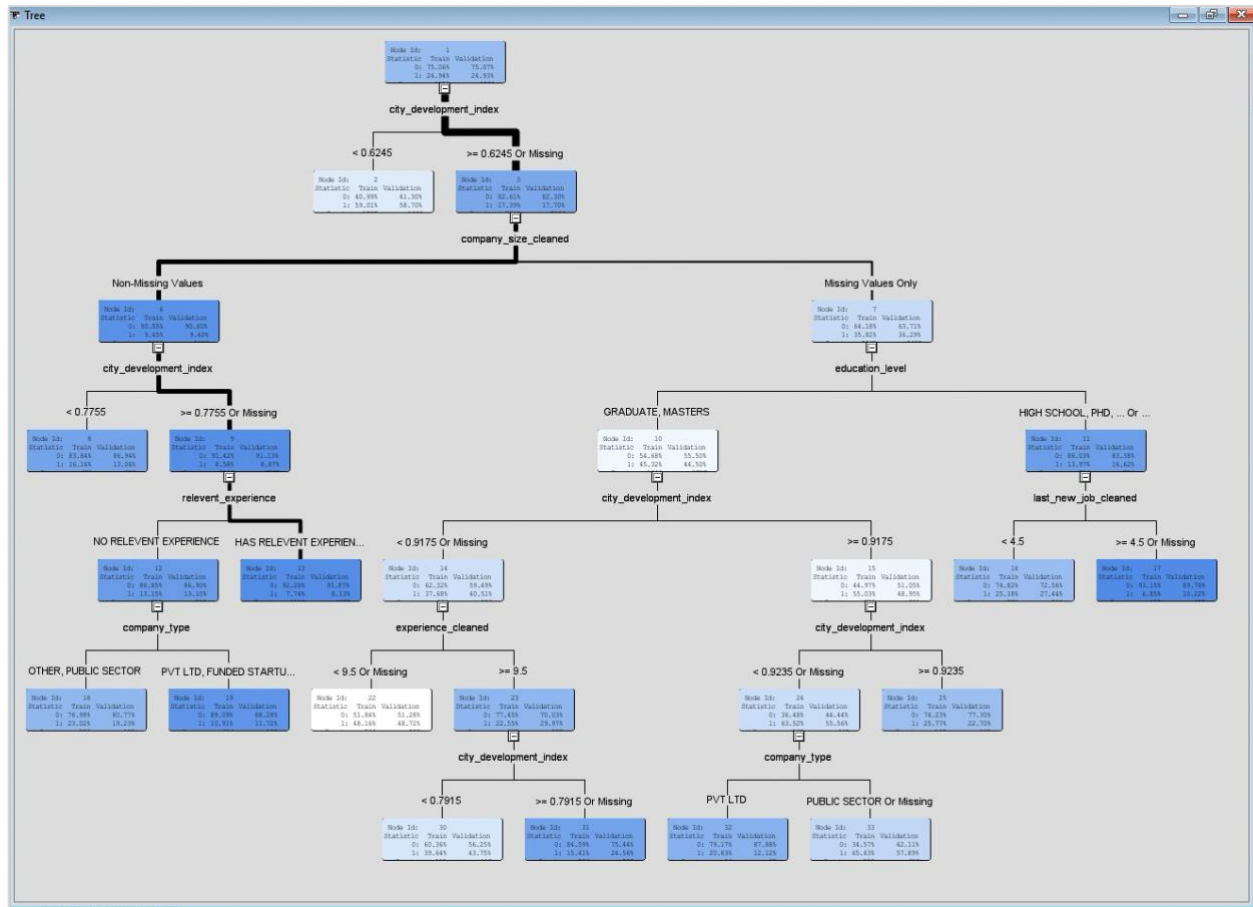
| Variable Importance | | | | | |
|------------------------|-------|---------------------------------|------------|--------------------------|---|
| Variable Name | Label | Number of Splitting Rules | Importance | Validation Importance | Ratio of Validation to Training Importance |
| city_development_index | | 4 | 1.0000 | 1.0000 | 1.0000 |
| company_size_cleaned | | 1 | 0.6364 | 0.7023 | 1.1036 |
| education_level | | 1 | 0.4161 | 0.3961 | 0.9518 |
| experience_cleaned | | 1 | 0.2271 | 0.1606 | 0.7069 |
| company_type | | 1 | 0.1270 | 0.1675 | 1.3193 |

The competing variables for the first split are city_development_index and company_size_cleaned. For the variable importance the company_size_cleaned explains 63.64%, while education_level explains 41.61% of the variability explained by city_development_index.

Probability Tree

Probability tree is where the leaves and branches were prune, moving the population to leaves that has the best split. This will optimize the performance of the decision tree by letting the tree grow to its optimal size. The result showed an ASE of 0.14232 with 33 leaves.

Figure 23. Probability Tree



| Split Variable | Variable Description | -Log(p) | Number of Branches |
|------------------------|------------------------|----------|--------------------|
| city_development_index | city_development_index | 285.6961 | 2 |
| company_size_cleaned | company_size_cleaned | 118.5231 | 2 |
| company_type | company_type | 95.7444 | 2 |
| experience_cleaned | experience_cleaned | 54.6879 | 2 |
| enrolled_university | enrolled_university | 42.4411 | 2 |

Variable Importance

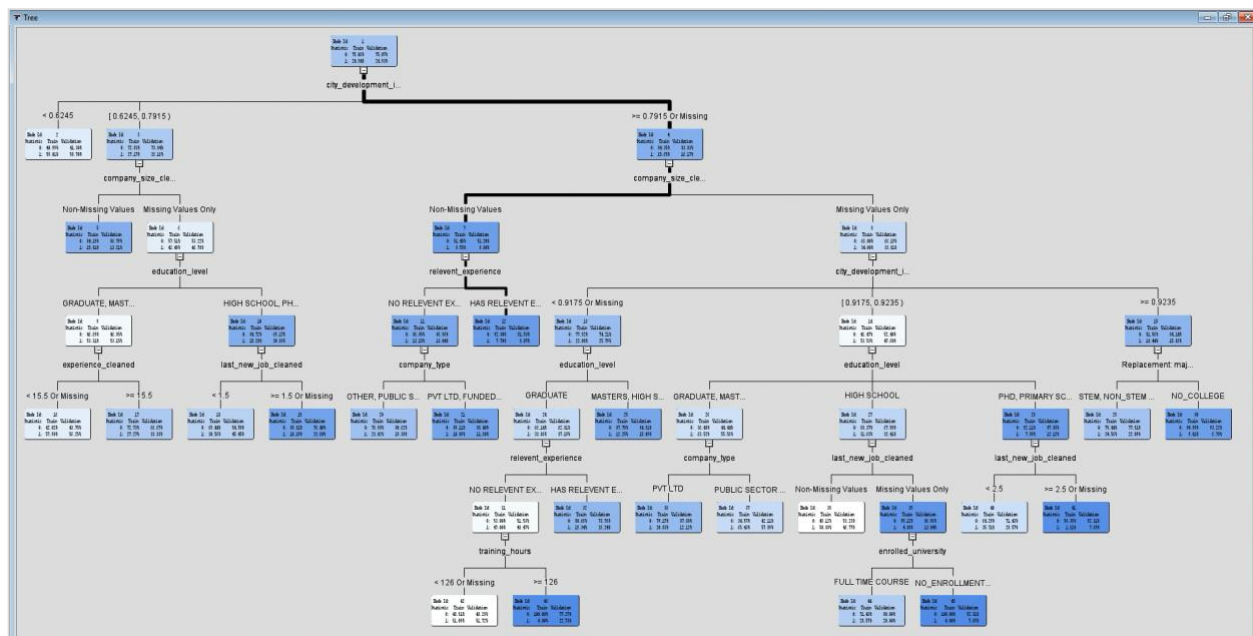
| Variable Name | Label | Number of Splitting Rules | Importance | Validation Importance | Ratio of Validation to Training Importance |
|------------------------|-------|---------------------------|------------|-----------------------|--|
| city_development_index | | 5 | 1.0000 | 1.0000 | 1.0000 |
| company_size_cleaned | | 1 | 0.6352 | 0.6982 | 1.0991 |
| education_level | | 1 | 0.4154 | 0.3937 | 0.9479 |
| experience_cleaned | | 1 | 0.2267 | 0.1596 | 0.7041 |
| company_type | | 2 | 0.1480 | 0.1710 | 1.1551 |
| last_new_job_cleaned | | 1 | 0.1418 | 0.1389 | 0.9796 |
| relevant_experience | | 1 | 0.0809 | 0.0767 | 0.9478 |

The competing variables for the first split are city_development_index and company_size_cleaned. For the variable importance the company_size_cleaned explains 63.52%, while education_level explains 41.54% of the variability explained by city_development_index.

Probability Tree – 3 Branches

Since the probability has the lowest ASE among the decisions tree, we then tried running it again by changing the maximum branch to three splits instead of just two. The result showed an ASE of 0.141066 with 22 leaves.

Figure 24. Decision Tree - Three way split tree



| Competing Rules For Node 1 | | | |
|----------------------------|------------------------|----------|--------------------|
| Split Variable | Variable Description | -Log(p) | Number of Branches |
| city_development_index | city_development_index | 298.2304 | 3 |
| company_size_cleaned | company_size_cleaned | 118.5231 | 2 |
| company_type | company_type | 96.2808 | 3 |
| experience_cleaned | experience_cleaned | 60.0066 | 3 |
| enrolled_university | enrolled_university | 43.3556 | 3 |

| Variable Importance | | | | | |
|------------------------|-------------------------------|---------------------------|------------|-----------------------|--|
| Variable Name | Label | Number of Splitting Rules | Importance | Validation Importance | Ratio of Validation to Training Importance |
| city_development_index | | 2 | 1.0000 | 1.0000 | 1.0000 |
| company_size_cleaned | | 2 | 0.5919 | 0.6603 | 1.1157 |
| education_level | | 3 | 0.4069 | 0.3359 | 0.8255 |
| relevant_experience | | 2 | 0.1727 | 0.1464 | 0.8477 |
| last_new_job_cleaned | | 3 | 0.1564 | 0.1380 | 0.8823 |
| company_type | | 2 | 0.1426 | 0.1650 | 1.1570 |
| experience_cleaned | | 1 | 0.1198 | 0.0814 | 0.6797 |
| training_hours | | 1 | 0.1120 | 0.0447 | 0.3995 |
| REP_major_discipline | Replacement: major_discipline | 1 | 0.0729 | 0.0786 | 1.0783 |
| enrolled_university | | 1 | 0.0393 | 0.0256 | 0.6509 |

The competing variables for the first split are city_development_index and company_size_cleaned. For the variable importance the company_size_cleaned explains 59.19%, while education_level explains 40.69% of the variability explained by city_development_index.

By comparing the ASE from all decision trees model, the probability tree with three branches has the best result (optimal tree):

Table 4. Decision tree - model summary

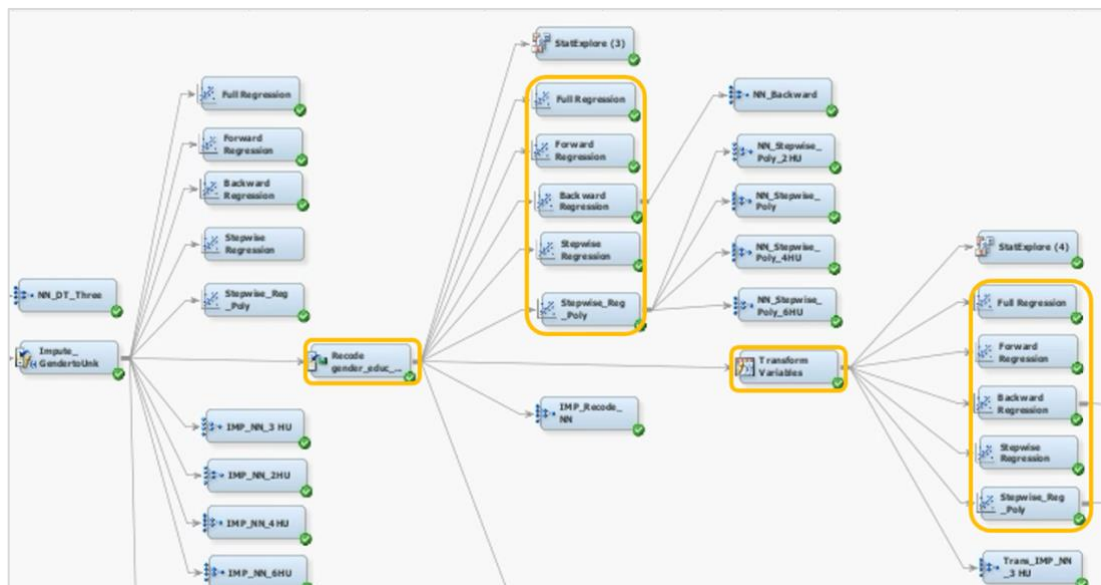
| Model | Average squared error (ASE) |
|-------------------------------|-----------------------------|
| Maximal Decision Tree | 0.142565 |
| Classification Tree | 0.143443 |
| Probability Tree | 0.14232 |
| Probability Tree – 3 Branches | 0.141066 |

5.2 Logistic Regression

Logistic regression (also known as the logit model), was formulated by David Cox in 1958 and can be used for both classification and class probability estimation with its range bounded between 0 and 1. A classification model rather than a regression model, logistic regression is used for dichotomous outcome variables and relies on the “odds” of the event as it applies a nonlinear log transformation to the odds ratio, thereby not requiring a linear relationship between variables.

In this study, logistic regression was used to predict whether the candidate will be looking for a job change or will stay with the company after training. During the model development process, missing input values were included using data imputation since one cannot run nor do math with missing data on logistic regression. Moreover, data transformations were also executed in preparation for the regression to ensure linearity, achieve normality, and stabilize the variance. Although cap and floor were initially added instead of the transform node, these were showing unsatisfactory average squared error results on the models compared to the outcome by applying log transformations, hence, the method was not applied. Transforming the data via a log transformation allows us to select the best mathematical transformation for variables such as categorical or binary variables which can help reduce the skewness, leading to a more accurate model. Outliers or extreme values were also handled, and variables were reduced to avoid overfitting. Model complexity was also increased by using the Polynomial Regression aside from the standard regression methods. This model together with the Full, Forward, Backward, and Stepwise selection models were utilized, and the results are shown below.

Figure 25. Diagram: Regression model



Sequential Selection - Full

With full regression, all inputs are used to fit the model. Results show that the lowest average squared error which is at 0.158557 belongs to the model connected to the transform node.

| Description of the Full Regression node | Average squared error (ASE) |
|---|-----------------------------|
| Connected to the recode node | 0.159674 |
| Connected to the transform node | 0.158557 |

Sequential Selection - Forward

Starting off with an empty model, the forward selection model adds significant variables one at a time and stops until the criterion has been met. The model connected to the transform node has the lower ASE at 0.15873.

| Description of the Forward Selection model | Average squared error (ASE) |
|--|-----------------------------|
| Connected to the recode node | 0.159724 |
| Connected to the transform node | 0.15873 |

Sequential Selection - Backward

In backward regression, it begins with all the predictor variables in consideration then gradually removes the least significant variable one after the other until no variable is left in the model. Similar to the full selection method, the model connected to the transform node has the least ASE at 0.158545.

| Description of the Backward elimination model | Average squared error (ASE) |
|---|-----------------------------|
| Connected to the recode node | 0.159665 |
| Connected to the transform node | 0.158545 |

Sequential Selection - Stepwise

The stepwise selection model combines both the forward and backward selection methods to test which significant variables best optimizes the model. It adds multiple variables while simultaneously removing those that are not significant. The same figures were generated similar

to the full selection model, in which the stepwise model connected to the transform node has the lower ASE at 0.15873.

| Description of the Stepwise model | Average squared error (ASE) |
|-----------------------------------|-----------------------------|
| Connected to the recode node | 0.159724 |
| Connected to the transform node | 0.15873 |

Polynomial Regression

Polynomial regression enables predictions to match the input or target association and minimize prediction bias. This regression node adds additional interaction terms and can be done either selectively or autonomously. With stepwise polynomial regression model, the lowest ASE was coming from the model connected to the recode node.

| Description of the Stepwise polynomial model | Average squared error (ASE) |
|--|-----------------------------|
| Connected to the recode node | 0.155537 |
| Connected to the transform node | 0.157175 |

The valuation data generated from the regression nodes allows us to choose the best model which is the Stepwise Polynomial Regression having the lowest ASE at 0.155537. Although the two polynomial regressions have the lowest ASE among the regression models created, the Backward Selection connected to the transform node will be used for a better interpretation of the results. It will allow us to distinctly measure the relationship between the variables and clearly define the constant effect of the predictor on the likelihood or probability that an outcome will occur.

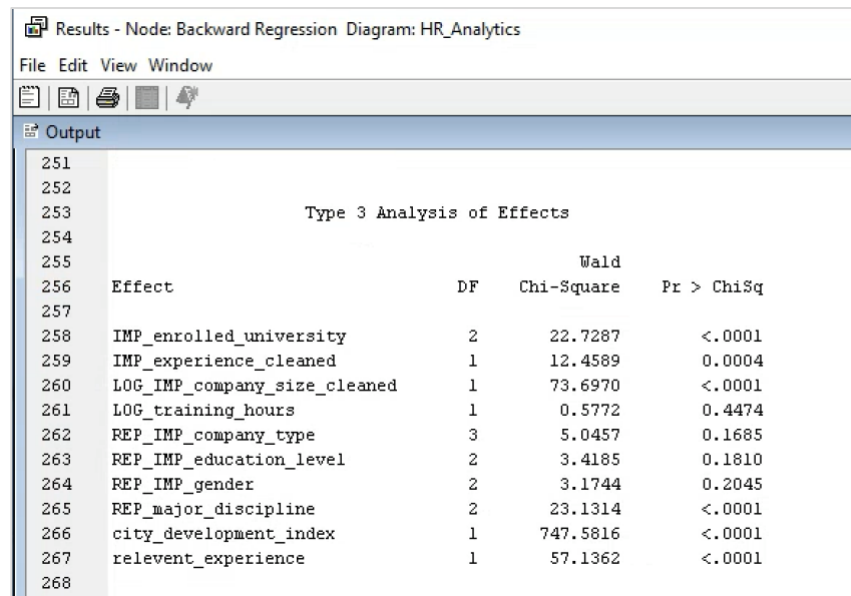
Table 5. Regression – model summary

| Description | Regression models | Average Squared Errors (ASE) |
|--|---------------------------------------|------------------------------|
| Regression models connected to the recode node | Full Regression | 0.159674 |
| | Forward Selection | 0.159724 |
| | Backward Selection | 0.159665 |
| | Stepwise Selection | 0.159724 |
| | Stepwise Polynomial Regression | 0.155537 |

| | | |
|---|--------------------------------|-----------------|
| Regression models connected to the transform node | Full Regression | 0.158557 |
| | Forward Selection | 0.15873 |
| | Backward Selection | 0.158545 |
| | Stepwise Selection | 0.15873 |
| | Stepwise Polynomial Regression | 0.157175 |

As seen from the figure below, the top three most important variables based on the Backward Selection chi-square results are city_development_index, LOG_IMP_company_size_cleaned and relevant_experience.

Figure 26. Regression - Type 3 Analysis of Effects



| Type 3 Analysis of Effects | | | | |
|------------------------------|----|-----------------|------------|--|
| Effect | DF | Wald Chi-Square | Pr > ChiSq | |
| IMP_enrolled_university | 2 | 22.7287 | <.0001 | |
| IMP_experience_cleaned | 1 | 12.4589 | 0.0004 | |
| LOG_IMP_company_size_cleaned | 1 | 73.6970 | <.0001 | |
| LOG_training_hours | 1 | 0.5772 | 0.4474 | |
| REP_IMP_company_type | 3 | 5.0457 | 0.1685 | |
| REP_IMP_education_level | 2 | 3.4185 | 0.1810 | |
| REP_IMP_gender | 2 | 3.1744 | 0.2045 | |
| REP_major_discipline | 2 | 23.1314 | <.0001 | |
| city_development_index | 1 | 747.5816 | <.0001 | |
| relevant_experience | 1 | 57.1362 | <.0001 | |

After creating the logistic regression models and selecting the inputs with the most appropriate selection method and fit statistic, the odds ratios will be used to interpret the model since it specifies each input's effect on the logit score. This will indicate what factors the odds of the candidate looking for a job change for each unit change in the associated input.

Figure 27. Regression - Odds ratio estimates

| | | | |
|---|------------------------------|---|----------------|
| Results - Node: Backward Regression Diagram: HR_Analytics | | | |
| File Edit View Window | | | |
| Output | | | |
| 799 | | | |
| 800 | | | |
| 801 | Odds Ratio Estimates | | |
| 802 | | | |
| 803 | | | |
| 804 | Effect | | Point Estimate |
| 805 | | | |
| 806 | IMP_enrolled_university | Full time course vs no_enrollment | 1.385 |
| 807 | IMP_enrolled_university | Part time course vs no_enrollment | 1.103 |
| 808 | IMP_experience_cleaned | | 0.984 |
| 809 | LOG_IMP_company_size_cleaned | | 1.124 |
| 810 | LOG_training_hours | | 0.979 |
| 811 | REP_IMP_company_type | Other vs Startup | 1.196 |
| 812 | REP_IMP_company_type | Public_NGO vs Startup | 1.010 |
| 813 | REP_IMP_company_type | Pvt Ltd vs Startup | 1.194 |
| 814 | REP_IMP_education_level | Degree_holder vs Primary School | 1.714 |
| 815 | REP_IMP_education_level | High School vs Primary School | 1.608 |
| 816 | REP_IMP_gender | Female vs 0_U | 1.040 |
| 817 | REP_IMP_gender | Male vs 0_U | 0.915 |
| 818 | REP_major_discipline | NON_STEM vs STEM | 1.126 |
| 819 | REP_major_discipline | NO_COLLEGE vs STEM | 0.464 |
| 820 | city_development_index | | 0.003 |
| 821 | relevent_experience | Has relevent experience vs No relevent experience | 0.613 |
| 822 | | | |
| 823 | | | |

Interpretation of the odd ratio estimate:

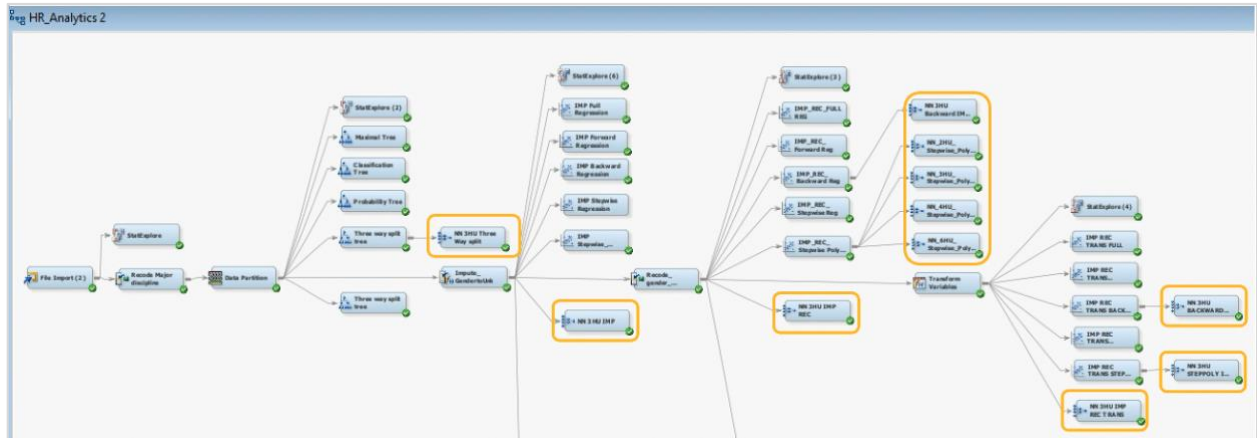
| Effect | Interpretation |
|---|--|
| IMP_enrolled_university (Full time course vs no_enrollment) | Employees enrolled in a full-time course are 38.5% more likely to look for a job change than those who are not enrolled in any course. |
| IMP_enrolled_university (Part time course vs no enrollment) | Employees enrolled in a part-time course are 10.3% more likely to look for a job change than those who are not enrolled in any course. |
| IMP_experience_cleaned | For every unit the employee's total experience in years goes up, the probability that the employee will look for a job change decreases by 1.6%. |
| LOG_IMP_company_size_cleaned | As company size increases by 2.74 times, it is 12.4% more likely that the employee will look for a job change. |
| LOG_training_hours | For every 2.74 times the training hours goes up, it is 2.1% less likely that the employee will look for a job change. |
| REP_IMP_company_type (Other vs Startup) | Employees from Other companies are 19.6% more likely to look for a job change compared to those from start-up companies. |
| REP_IMP_company_type (Public_NGO vs Startup) | Employees from public sector/NGOs are 1% more likely to look for a job change compared to those from start-up companies. |
| REP_IMP_company_type (Pvt Ltd vs Startup) | Employees from Pvt Ltd are 19.4% more likely to look for a job change compared to those from start-up companies. |

| | |
|---|---|
| REP_IMP_education level (Degree holder vs Primary School) | Degree holder employees are 71.4% more likely to look for a job change than those who completed primary school. |
| REP_IMP_education level (High School vs Primary School) | Employees who finished high school are 60.8% more likely to look for a job change than those who completed primary school. |
| REP_IMP_Gender (Female vs O_U) | Females are 4% more likely to look for a job change than Other/Unknown gender. |
| REP_IMP_Gender (Male vs O_U) | Males are 8.5% less likely to look for a job change than Other/Unknown gender. |
| REP_major_discipline (NON_STEM vs STEM) | Employees with non-STEM education major discipline are 12.6% more likely to look for a job change compared to those who took STEM. |
| REP_major_discipline (NO_COLLEGE vs STEM) | Employees with no college are 53.6% less likely to look for a job change compared to those who took STEM. |
| city_development_index | For every unit of city development index that goes up, the probability that the employee will look for a job change decreases by 99.7%. |
| relevant_experience (Has relevant experience vs No relevant experience) | Employees with relevant experience are 38.7% less likely to look for a job change than those with no relevant experience. |

5.3 Neural Network

Neural Network uses a prediction formula based on regression, so probability is also computed and customized for each observation. The results of NN can be more accurate, but it will also be hard to interpret. Since NN can't work on an incomplete dataset, we will be connecting our models after the imputation node. In general, transformation is not needed for NN since extreme and unusual values can be handled with the use of hyperbolic tangent activation function. However, there could be times that results are better when data is less skewed so in this study, we will also include models with transformation.

Figure 28. Diagram: Neural network model



All NNs are using maximum iterations of 100 for the convergence criterion to be met. Each NN model was tested to use preliminary training and will be compared if it has lower ASE when preliminary training is set to No. Whichever is lower will be the final set up for that node.

Full Neural Network

In a full neural network model, all variables will be used in training the data. Three full neural network models were created for this study and will be compared if the results are better with the recode and transform nodes or not. Based on ASE, the best model among the three is the model without any recode and transform, NN 3 HU IMP.

| Full Neural network node | Description | Preliminary training | Average Squared Error (ASE) |
|--------------------------|--|----------------------|-----------------------------|
| NN 3 HU IMP | Directly connected to impute node | Yes | 0.146924 |
| NN 3 HU IMP REC | Connected to impute and recode nodes | No | 0.151555 |
| NN 3 HU IMP REC TRANS | Connected to impute, recode, and transform nodes | Yes | 0.15012 |

Reduced Variables Set Neural Network

Neural network has no internal mechanism to select useful inputs in a model; therefore, we will generate models using the reduced variables from our best decision tree model and regression

models. Among these 5 models, the best one is NN stepwise polynomial connected to impute and recode (NN 3HU Stepwise Polynomial IMP REC).

| Neural network node (Reduced variable) | Description | Preliminary training | Average Squared Error (ASE) |
|---|---|-------------------------|-----------------------------------|
| NN 3HU Three way split | NN from Three-way split decision tree | Yes | 0.174168 |
| NN 3HU Backward IMP REC | NN from backward regression (with recode data) | No | 0.156545 |
| NN 3HU Stepwise Polynomial IMP REC | NN from stepwise polynomial regression (with recode data) | No | 0.146318 |
| NN 3HU Backward IMP REC TRANS | NN from backward regression (with recode and transformed data) | No | 0.156338 |
| NN 3HU Stepwise Poly IMP REC TRANS | NN from stepwise polynomial regression (with recode and transformed data) | No | 0.147195 |

Optimal Neural Network

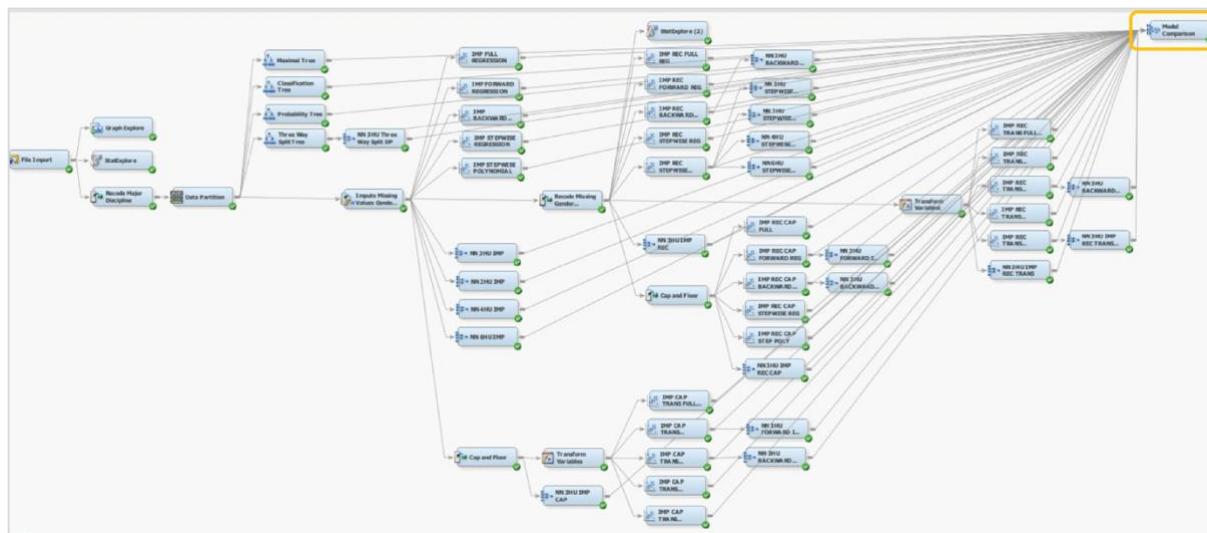
From the neural network models above, the best model is provided by “NN 3HU Stepwise Polynomial IMP REC” three hidden units with ASE of 0.146318. For the final tuning, hidden units were changed to 2, 4 and 6 to check which one will result to a better model. Results showed that the best model created by NN is the 4 hidden units model with ASE of 0.146234.

| Neural Network model | Average Squared Error (ASE) |
|---|--------------------------------|
| NN 2HU Stepwise Polynomial IMP REC | 0.152741 |
| NN 3HU Stepwise Polynomial IMP REC | 0.146318 |
| NN 4HU Stepwise Polynomial IMP REC | 0.146234 |
| NN 6HU Stepwise Polynomial IMP REC | 0.147792 |

6.1 Performance Measure / Model Assessment (ROC/ASE)

To help us summarize the results from all our models, we can connect them to a special node called “Model comparison”. This allows us to cross-model comparisons and predictions from preceding nodes. Our basis for our assessment is based on two criteria: Receiver Operator Characteristic (ROC) index and Average Squared Error (ASE).

Figure 29. Final diagram: Model assessment



Our fit statistics from our model comparison can help us determine which models demonstrates the highest ROC index and lowest Average Squared Error. The results as follows:

Fit Statistics window under Model Comparison node results

| Fit Statistics | | | | | | |
|----------------|------------------|------------|------------------------------------|-----------------|------------------|------------------------------|
| Selected Model | Predecessor Node | Model Node | Model Description | Target Variable | Valid: Roc Index | Valid: Average Squared Error |
| Y | Tree4 | Tree4 | Three Way Split Tree | target | 0.789 | 0.141066 |
| | Tree | Tree | Maximal Tree | target | 0.789 | 0.142565 |
| | Tree2 | Tree2 | Classification Tree | target | 0.779 | 0.143443 |
| | Tree3 | Tree3 | Probability Tree | target | 0.786 | 0.14232 |
| | Neural8 | Neural8 | NN 3HU STEPWISE POLY IMP REC | target | 0.785 | 0.146318 |
| | Neural5 | Neural5 | NN 4HU STEPWISE POLY IMP REC | target | 0.788 | 0.146234 |
| | Neural26 | Neural26 | NN 3HU BACKWARD IMP CAP TRANS | target | 0.784 | 0.14654 |
| | Neural21 | Neural21 | NN 3HU BACKWARD IMP REC CAP | target | 0.785 | 0.146562 |
| | Neural13 | Neural13 | NN 3HU IMP REC TRANS | target | 0.776 | 0.15012 |
| | Neural18 | Neural18 | NN 3HU STEPWISE POLY IMP REC TRANS | target | 0.775 | 0.149904 |
| | Neural | Neural | NN 3HU IMP | target | 0.784 | 0.146924 |
| | Neural11 | Neural11 | NN 6HU IMP | target | 0.774 | 0.15045 |
| | Neural10 | Neural10 | NN 4HU IMP | target | 0.777 | 0.150174 |
| | Neural6 | Neural6 | NN 3HU BACKWARD IMP REC | target | 0.743 | 0.156545 |
| | Neural20 | Neural20 | NN 3HU FORWARD IMP REC CAP | target | 0.742 | 0.156732 |
| | Neural24 | Neural24 | NN 3HU IMP CAP | target | 0.745 | 0.156782 |
| | Neural7 | Neural7 | NN 6HU STEPWISE POLY IMP REC | target | 0.784 | 0.147792 |
| | Neural27 | Neural27 | NN 3HU IMP REC CAP | target | 0.772 | 0.15159 |
| | Neural4 | Neural4 | NN 2HU STEPWISE POLY IMP REC | target | 0.771 | 0.152741 |
| | Neural2 | Neural2 | NN 3HU IMP REC | target | 0.771 | 0.151904 |
| | Neural25 | Neural25 | NN 3HU FORWARD IMP CAP TRANS | target | 0.744 | 0.156941 |
| | Neural9 | Neural9 | NN 2HU IMP | target | 0.772 | 0.152123 |
| | Reg10 | Reg10 | IMP REC TRANS STEPWISE POLYNOMIAL | target | 0.754 | 0.157175 |
| | Neural16 | Neural16 | NN 3HU BACKWARD IMP REC TRANS | target | 0.746 | 0.156338 |
| | Reg7 | Reg7 | IMP REC TRANS FORWARD REG | target | 0.741 | 0.15873 |
| | Reg9 | Reg9 | IMP REC TRANS STEPWISE REG | target | 0.741 | 0.15873 |
| | Reg20 | Reg20 | IMP REC CAP STEPWISE POLY | target | 0.767 | 0.155594 |
| | Reg6 | Reg6 | IMP REC TRANS FULL REG | target | 0.743 | 0.158557 |
| | Reg5 | Reg5 | IMP REC STEPWISE POLYNOMIAL | target | 0.767 | 0.155537 |
| | Reg8 | Reg8 | IMP REC TRANS BACKWARD REG | target | 0.743 | 0.158545 |
| | Reg21 | Reg21 | IMP CAP TRANS FULL REG | target | 0.736 | 0.159482 |
| | Reg11 | Reg11 | IMP FULL REGRESSION | target | 0.736 | 0.159486 |
| | Reg25 | Reg25 | IMP CAP TRANS STEPWISE POLY | target | 0.765 | 0.156022 |
| | Reg13 | Reg13 | IMP BACKWARD REGRESSION | target | 0.736 | 0.159475 |
| | Reg23 | Reg23 | IMP CAP TRANS BACKWARD REG | target | 0.736 | 0.159472 |
| | Reg15 | Reg15 | IMP STEPWISE POLYNOMIAL | target | 0.766 | 0.155964 |
| | Reg22 | Reg22 | IMP CAP TRANS FORWARD REG | target | 0.735 | 0.159633 |
| | Reg24 | Reg24 | IMP CAP TRANS STEPWISE REG | target | 0.735 | 0.159633 |
| | Reg | Reg | IMP REC FULL REG | target | 0.732 | 0.159674 |
| | Reg18 | Reg18 | IMP REC CAP BACKWARD REG | target | 0.732 | 0.159663 |
| | Reg16 | Reg16 | IMP REC CAP FULL REG | target | 0.732 | 0.159671 |
| | Reg3 | Reg3 | IMP REC BACKWARD REG | target | 0.732 | 0.159665 |
| | Reg12 | Reg12 | IMP FORWARD REGRESSION | target | 0.735 | 0.159561 |
| | Reg14 | Reg14 | IMP STEPWISE REGRESSION | target | 0.735 | 0.159561 |
| | Reg17 | Reg17 | IMP REC CAP FORWARD REG | target | 0.731 | 0.159793 |
| | Reg19 | Reg19 | IMP REC CAP STEPWISE REG | target | 0.731 | 0.159793 |
| | Reg2 | Reg2 | IMP REC FORWARD REG | target | 0.731 | 0.159724 |
| | Reg4 | Reg4 | IMP REC STEPWISE REG | target | 0.731 | 0.159724 |
| | Neural12 | Neural12 | NN 3HU Three Way Split | target | 0.712 | 0.174697 |

When arranged accordingly, the top five models based on each respective validation are:

| Top ROC Index | | Top Average Squared Error | |
|------------------------------|------------------|------------------------------|------------|
| Model Description | Valid: Roc Index | Model Description | Valid: ASE |
| Three Way Split Tree | 0.789 | Three Way Split Tree | 0.141066 |
| Maximal Tree | 0.789 | Probability Tree | 0.14232 |
| NN 4HU STEPWISE POLY IMP REC | 0.788 | Maximal Tree | 0.142565 |
| Probability Tree | 0.786 | Classification Tree | 0.143443 |
| NN 3HU STEPWISE POLY IMP RE | 0.785 | NN 4HU STEPWISE POLY IMP REC | 0.146234 |

It is shown that under ROC Index, our best models, Three Way Split Tree and Maximal Tree have both tied with a value of 0.789. Meanwhile, if determined using the Average Squared Error, Three Way Split Tree also leads with the lowest value at 0.141066.

6.2 Feature Importance

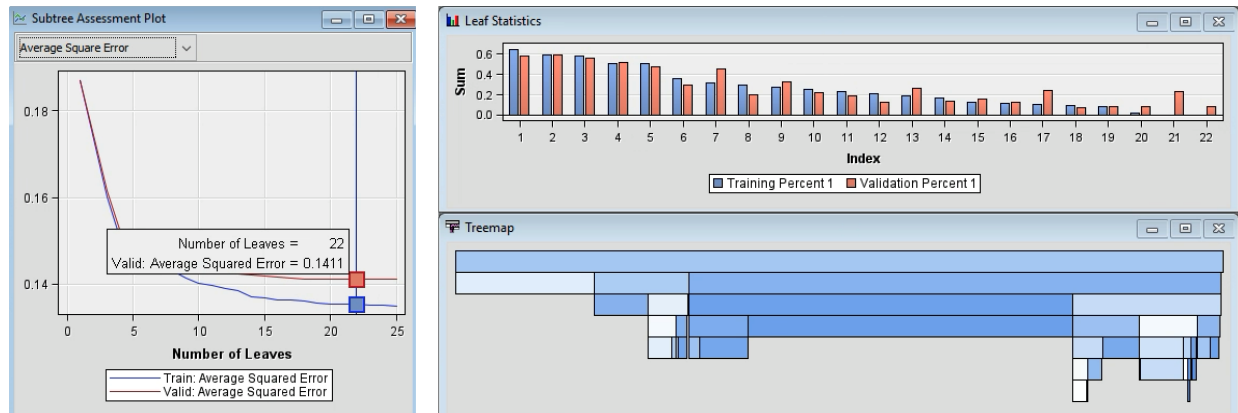
Figure 30. Best model: Three-Way Split Tree - Variable Importance

| Variable Importance | | | | | |
|------------------------|-------------------------------|---------------------------|------------|-----------------------|--|
| Variable Name | Label | Number of Splitting Rules | Importance | Validation Importance | Ratio of Validation to Training Importance |
| city_development_index | | 2 | 1.0000 | 1.0000 | 1.0000 |
| company_size_cleaned | | 2 | 0.5919 | 0.6603 | 1.1157 |
| education_level | | 3 | 0.4069 | 0.3359 | 0.8255 |
| relevent_experience | | 2 | 0.1727 | 0.1464 | 0.8477 |
| last_new_job_cleaned | | 3 | 0.1564 | 0.1380 | 0.8823 |
| company_type | | 2 | 0.1426 | 0.1650 | 1.1570 |
| experience_cleaned | | 1 | 0.1198 | 0.0814 | 0.6797 |
| training_hours | | 1 | 0.1120 | 0.0447 | 0.3995 |
| REP_major_discipline | Replacement: major_discipline | 1 | 0.0729 | 0.0786 | 1.0783 |
| enrolled_university | | 1 | 0.0393 | 0.0256 | 0.6509 |

In determining the important variables, we must first select our overall best model. According to our model comparisons, the Three-Way Split Tree can be considered as this has the highest ROC index and the lowest Average Squared Error. The above image displays the metrics of important input features. The top 3 most important variables concluded are city_development_index, company_size_cleaned, and education_level. The city_development_index is the input shown to have the highest importance. Moreover, company_size_cleaned explains 59.19% of the variability explained by the city_development_index. While for education_level, it's variability relative to the city_development index amounts to 40.69%.

6.3 Best model results

Figure 31. Three-way split tree results



The three-way split decision tree has 22 terminal nodes with ASE of 0.1411. We have selected four nodes (group) that have the highest probability to change jobs and 1 node that has the lowest probability to leave their jobs. The probability and characteristics of these groups are described in Table 6 below. The information will be used to report business results and suggest recommendations which will be discussed in section 7.

Table 6. Leaf nodes considered for business report and recommendations

| Node | Training Observations | Training % | Validation observations | Validation % | Group characteristics |
|------|-----------------------|------------|-------------------------|--------------|--|
| 2 | 1737 | 0.59 | 1690 | 0.59 | Employees from city development index below 0.6245 |
| 7 | 4797 | 0.09 | 4714 | 0.09 | Employees from city development index at least 0.7915 or missing, known company size (11,12,20,21) |
| 42 | 184 | 0.51 | 174 | 0.52 | Employees from city development index between 0.7915 and 0.9175 or missing, Graduate, no relevant experience, Training hours below 126 |
| 26 | 562 | 0.64 | 648 | 0.56 | Employees from company size unknown, city development index between |

| | | | | | |
|---|-----|------|-----|------|---|
| | | | | | 0.9175 and 0.9235, graduate / masters |
| 9 | 362 | 0.53 | 333 | 0.53 | Employees from city development index between 0.6245 and 0.7915, unknown company size, graduate / masters |

7 Conclusion and Recommendations

7.1 Conclusion

This project consists of two major parts: First, the use of intensive modeling techniques such as decision trees, logistic regression, and neural networks to predict the probability of whether data scientists to look for a new job; Second, interpreting affected factors on the employee's decision and finding the optimal solution that can address our business problem. Based on the results from a study of 19,000 data scientists, it can be predicted that those who are looking for a job change are highly influenced by the employee's location (city development index), current company size, and education level. Employees coming from low development index cities (below 0.6245) are more inclined to switch to a new job with a probability of 59%. Moreover, employees who have college or master's degrees from cities with development indexes falling within 0.6245 - 0.7915 or 0.9175 - 0.9235 coming from unidentified company size and have less than 126 training hours are likely to switch jobs as well. On the other hand, employees from cities with a development index of at least 0.7915 or undetermined, and from a company with a specified number of employees are predicted to not look for a new job. This model is tested to be accurate 86% of the time.

7.2 Recommendation

Recruitment and Talent Acquisition

Results showed that from all the factors to be considered, the city development index is the most important determinant of whether a candidate will stay in their current company or switch to a new job. The City Development Index (CDI) can be defined as a measure to examine the socioeconomic development of cities, as well as an individual's access to urban facilities and their average well-being. Through data-driven talent acquisition and HR analytics, the company can reduce costs and improve its hiring process by making decisions based on reliable data by targeting potential candidates who are living in these specific indices in developed cities. Not only can it reduce the drop-off rate during the recruitment process, but it can also eliminate unnecessary costs and time spent in selecting the right employees since they are most likely looking to switch jobs. Practicing evidence-based HR can enable the company to prevent losses before they happen, reduce the risk of turnover and improve retention by designing targeted strategies.

Training and Development

Through machine learning techniques, the HR department now has the ability to pinpoint bottlenecks in their hiring process, streamline the training they currently offer, and improve the structure of candidate pipelines to hire the most motivated and skilled individuals. Offering incentives such as training and certifications can help influence employment and improve retention. As data science constantly evolves, training opportunities can be used as a retention tool that can instill loyalty and commitment from employees to further advance their careers, climb up the management ladder or hone their knowledge of new technologies. By offering access to data science boot camps, certifications (e.g., SAS Advanced Analytics or Azure Data Scientist

Associate from Microsoft), and unlimited access to training platforms (e.g., LinkedIn Learning, Coursera, Skillshare, and Udemy), employees can improve their job performance, as well as their analytical and problem-solving skills. It gives employees the opportunity to upskill, reskill, boost their performance and achieve organizational goals.

7.3 Further Research

Despite obtaining informative results from the modeling techniques used, this study has a huge scope for improvement and extension in the current model. Adding more information and variables can be done to further enhance the results such as performing sentiment analysis or opinion mining to analyze employee feedback and describe how employees feel about their organization. Through this, it can detect changes in employee satisfaction and measure the employee's overall experience. In addition, performing a more thorough analysis of the significant variables used in the study can be done such as learning how to identify the unknown company sizes included in the dataset and understanding why employees from lower CDI are moving to higher CDI locations.

References

- City Development Index*. (n.d.). <https://www.cdindex.net/en/methodology>
- Davenport, T. H., Patil D. (2022). *Is Data Scientist still the sexiest job of the 21st Century*. Harvard Business Review. <https://hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century>.
- Globe News Wire. (2021). *Study Reveals High Turnover Rates Among Data Science Professionals*. <https://www.globenewswire.com/en/news-release/2021/10/15/2314725/0/en/Study-Reveals-High-Turnover-Rates-Among-Data-Science-Professionals.html>
- Kaggle.com dataset. *HR Analytics: Job change of Data Scientists. Predict who will move to a new job*. Retrieved from <https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists/code>.
- Song, Y.-Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, 27(2), 130–135. <https://doi.org/10.11919/j.issn.1002-0829.215044>
- Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica*, 12–18. <https://doi.org/10.11613/bm.2014.003>.
- Omdena. (2022). *3 Reasons why data scientists leave their jobs and how tech companies can change this*. <https://omdena.com/blog/why-data-scientists-leave-their-jobs/>