Will Your Employee Switch?
HR Analytics using Machine Learning and Visualization
DS226: Project Report

Aashish (19689) Dyotana Das (19223) Shaurya Tiwary (18168) Soumalya Nandi (18016) Muttaqi Ahmed Alladin (18409)

ABSTRACT

Employees are the key factors for the success of every organization. Organizations all over the world are putting forth enormous efforts to find and, perhaps more importantly, retain valuable people. However, it has been noticed that, in comparison to a few years ago, employees in most IT companies are more likely to switch companies. Therefore, from an employer's perspective, it is important to predict how likely an employee is to leave the organization. Through this project we have built an end-to-end ML model which takes care of data preprocessing, EDA, visualization, efficiently handling missing values along with binary classification. We have also implemented two new variants of logistic regression considering different link function. On the test dataset, we are getting an F_1 score of .

1. Introduction

We are now in the era of Machine Learning (ML) and Artificial Intelligence (AI). The progress in this field in the last few years is unimaginable. It has provided several frameworks/algorithms starting from linear regression, logistic regression, decision tree, support vector machines (SVM) to complex ensemble models like Xgboost, random forest to more sophisticated and complex frameworks like Generative Adversarial Networks (GAN), Recurrent Neural Networks (RNN), Transformers etc. All these theoretically acclaimed algorithms become useless unless they are used efficiently to solve practical real-life problems. One such business problem prevailing out there is the problem of employees leaving a particular company resulting in reduction of the company's value in the global market. The appropriate term to address this problem that is used in the analytics domain is **Churn Analytics**, which is the analysis of employee loss rate using analytical tools in order to reduce it.

This project aims to tackle not exactly the same but something similar. The problem statement is as follows: An organization is looking to hire data scientists who have successfully passed some courses conducted by the company. A large number of people have expressed interest in participating in their training. The organization needs to know which of these candidates are truly interested in working for the company after completing their training or looking for new work. Hence it is highly important and valuable to classify whether a candidate is looking for a new job or not based on several important features of the candidates.

A Data Scientist's job is not to just build classification or regression-based machine learning models but also to extract information as much as possible from data to gain useful insights. In this project also, we have carried out several useful insights from the data through several statistical tests and EDA. We have done proper data preprocessing before building the final classification model. **Along in the line to produce something novel, we have considered two separate link functions other than logit, along with proper theoretical reasoning**.

2. Source of Data

The data source is a public dataset available in Kaggle under the link https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists

3. Related Works

Being an open to public dataset available in Kaggle, many people have done analysis which can be found here: https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists/code. Almost all the works aim just to create a classification model which gives better metrics for classification than others. However, we have kept our focus to do something novel and different from them. The dataset contains missing values, while previous works have mostly replaced the missing ones directly with mean/median/mode, we have performed extensive statistical tests and then have taken decision to select the proper missing value imputation technique. None of the previous works have considered Cauchy and Laplace link functions for this particular dataset. We are the first to use. Therefore our work is very much different from the past works.

4. Dataset Description

The training dataset under consideration has 19,158 observations with 14 columns. Few columns are qualitative or categorical type, while rest are numeric type. The column descriptions are shown in the following table.

Table 1: Table showing the data description

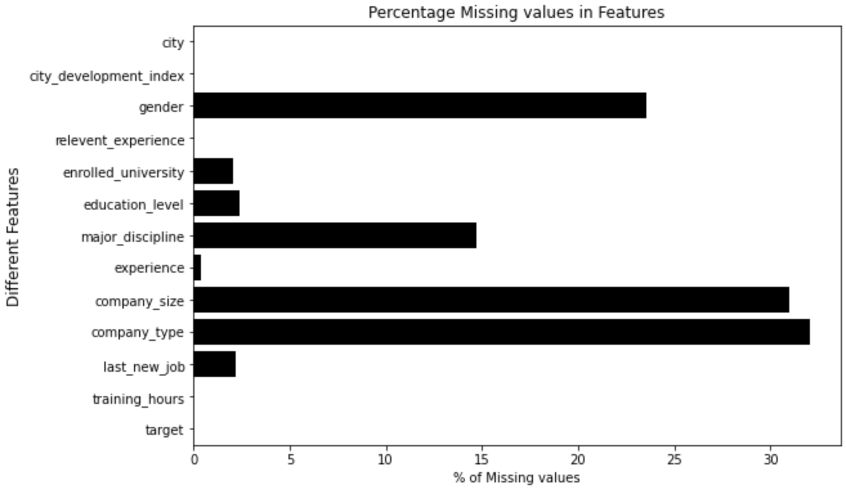
Column name	Description	Data Type
enrollee_id	Unique ID for candidate	Categorical
city	City code	Categorical
city_development_index	Development index for city (scaled)	Numeric
gender	Gender of candidate	Categorical
relevant_experience	Relevent experience of candidate	Categorical
enrolled_university	Type of university course enrolled	Categorical
education_level	Education level of candidate	Categorical
major_discipline	Education major discipline of candidate	Categorical
experience	Candidate total experience in years	Categorical
company_size	No. of employees in current employer's company	Categorical
company_type	Type of current employer	Categorical
last_new_job	Difference in years between prev. and curr. Job	Categorical
training_hours	Training hours completed	Numeric
target	0: not looking for a job change 1: looking for a job change	Categorical

The "target" column is our column of interest. It is the dependent variable while the rest are independent variables. "Enrollee_id" is of no use for our analysis and thus it is discarded from the dataset before doing further analysis.

Out of the 19,158 observations, 14,381 observations have target=0 while the rest 4,777 observations have target=1. This implies that this is an imbalanced dataset. Keeping this in mind, **we will not consider accuracy as a valid metric to evaluate our classification model**.

5. Missing Value Imputation

The very first work to do before diving into model building is to check for missing values in the data and if found then impute them properly. Missing value imputation should be treated with great care. There are many methods out there in the literature. Easy ones include, imputing the missing values by mean (if numeric data) or median (if ordinal data) or mode (if nominal data), whereas the more sophisticated ones include predicting the missing values using K-NN or Random Forest. All these methods have some drawbacks. Therefore, first we need to assess our data.



8 columns have missing values, out of which 4 columns have quite high percentage of missing values. “**gender**” has almost 25% of missing values, “**major_discipline**” has about 15% of missing values, “**company_size**” and “**company_type**”, both have more than 30% of missing values. The rest 4 column have missing value percentage of less than 5%.

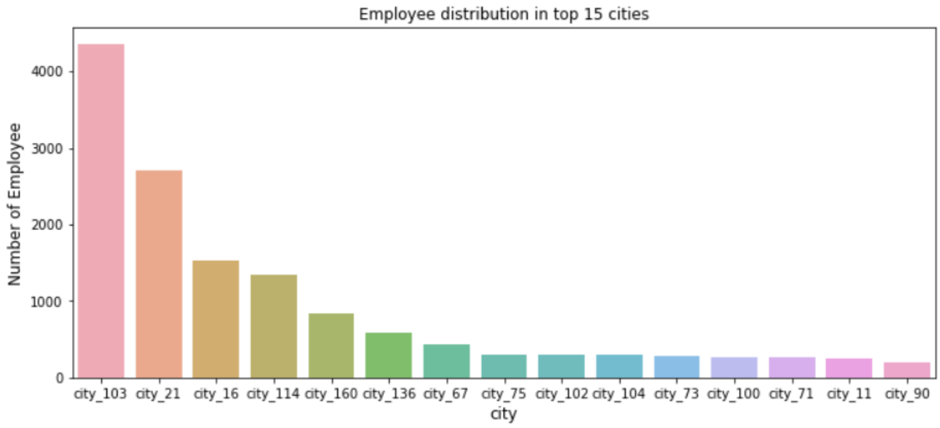
Whatever method we use, that will introduce some amount of bias to the data. Implementing Random Forest like structure to impute the missing values can be tricky here as there are many variables here which also have missing values. Imputing the columns with high number of missing values by mean/median/mode will introduce significant amount of bias which may hamper the model building. We can omit the rows having missing values but that would reduce the training data size drastically. Therefore, to come to a solution,

For the 4 columns with less than 5% of missing values, we will impute by mean/median/mode depending on data type. **For the other 4 columns, we can directly drop them from the dataset to avoid any kind of judgmental biasedness**.

However, first we have to check whether the columns are significant to describe the target variable or not. Through visualization and statistical tests, we will first infer about them, then will take decision accordingly.

6. Visualization and EDA :

- Feature 1– 'city'

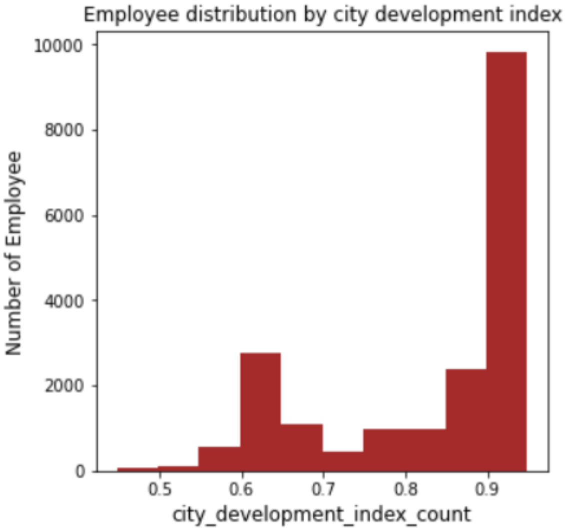


Here we have plotted the no of employees present in a city. As there are many city_index present, in the above graph we have plotted only for top 15 city_index count. City_103 has the maximum no of employees.

From the data we can find that city_103 has city_development_index of 0.92. For city_21 it is 0.624 and for city_16 it is 0.91.

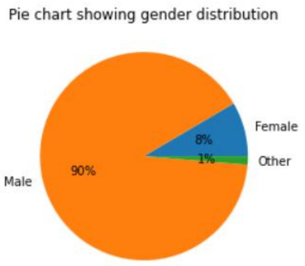
This tells the reason why so many observations are from these three cities, because as there are developed cities, these have more job opportunities. Since our dataset is on job seekers, hence the cities with high development indices dominate.

- Feature 2 - ' city development index'

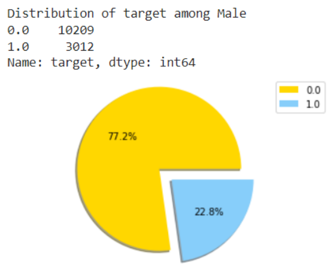


This plot also depicts the similar picture as above. Highly developed cities have more job seekers which is quite natural.

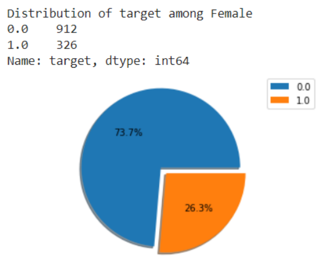
- Feature 3 – ‘Gender’



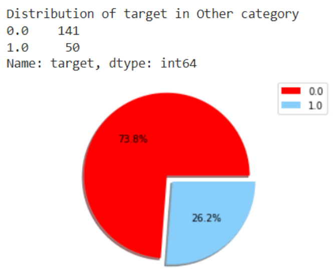
Pie chart to represent the distribution of gender.



```
Distribution of target among Male
0.0    10209
1.0     3012
Name: target, dtype: int64
```



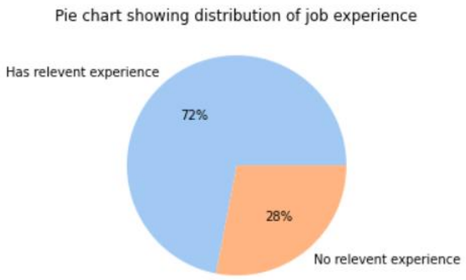
```
Distribution of target among Female
0.0    912
1.0    326
Name: target, dtype: int64
```

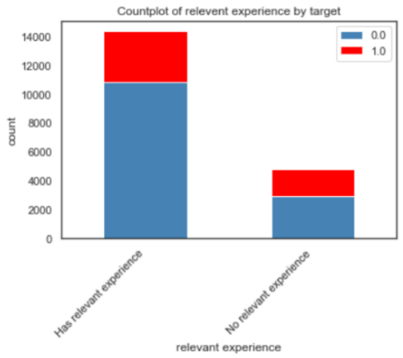


```
Distribution of target in Other category
0.0    141
1.0     50
Name: target, dtype: int64
```

The distribution of target variable within genders are very much similar, but since almost 90% of the filled observations have gender as “male”, we can not conclude anything about the importance of gender on target variable without a proper statistical hypothesis testing.

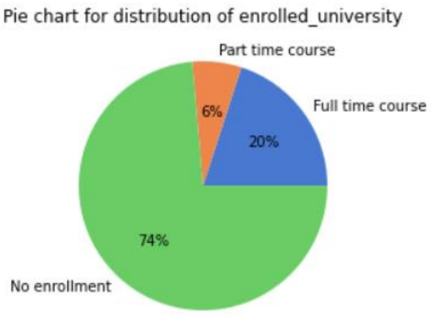
- Feature 4 – ‘Relevant experience’

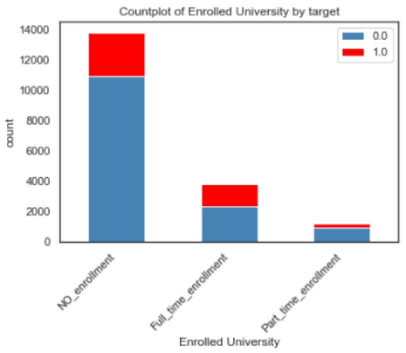




For the above counter plot we can see that people having relevant experience are more likely to stay in the same company and people having no relevant experience will switch more to their domain.

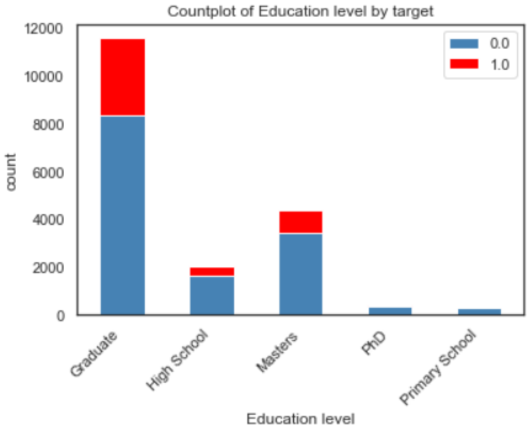
- Feature 5 – ‘Enrolled university’:

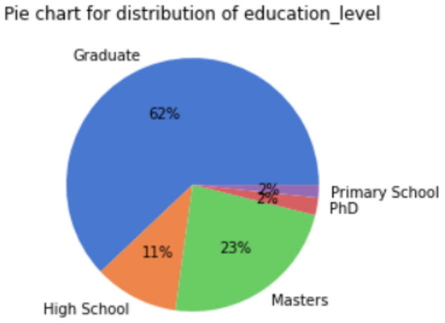




The employees enrolled fulltime are more likely to switch job as almost 40% people left their job from this category. The reason is maybe that as they were enrolled to fulltime university degree, they are more capable to search and land with a better job.

- Feature 6- ‘Education level’

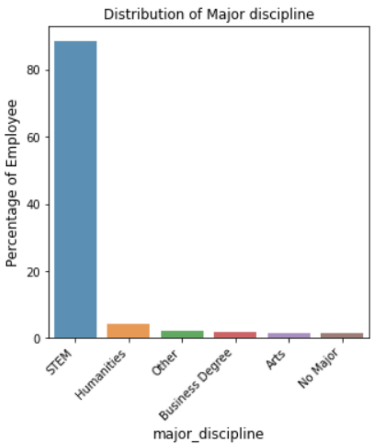


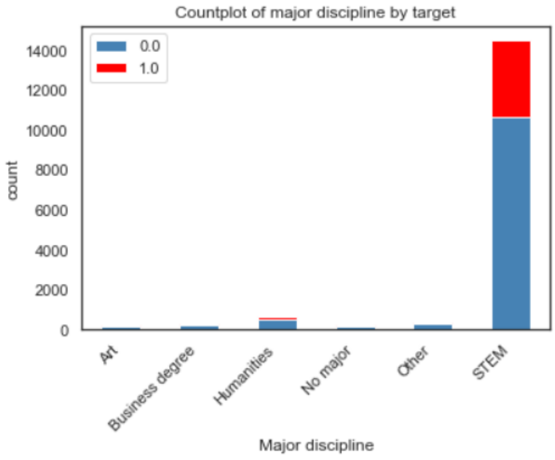


Education level has greater impact on the target value. Employees highly educated such as having degree of PhD or with very low level of education (ex- primary school) will not leave the job. Employee having standard qualification (ex- high school, masters) are less likely to leave the job. But from the about counter plot we have observed that graduate people have higher tendency to switch their job.
The above pie chart shows us that almost 62% of the employees have education_level = graduate. Thus, this feature has high impact on the target value.

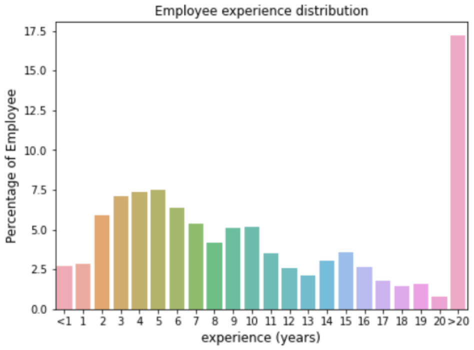
➢ Feature 7 – ‘Major discipline’

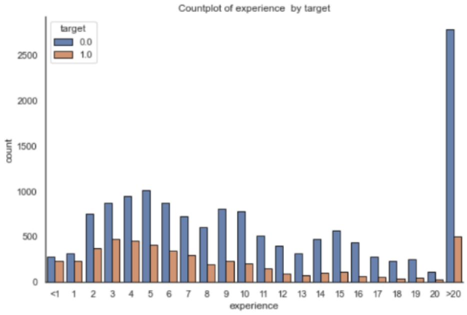
Almost 90% of the employees having major_discipline= ‘STEM’ . This feature does not generalize any quality and will not affect the target value in large amount.





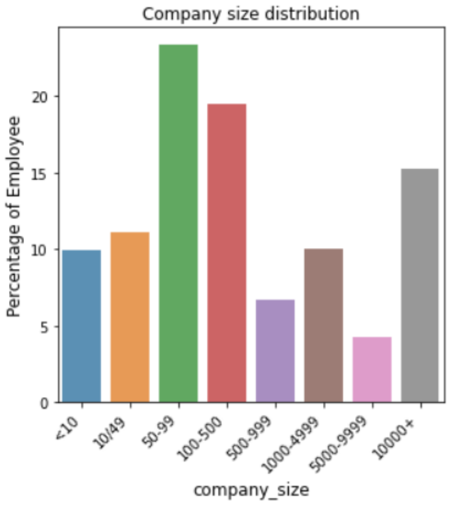
➢ Feature 8– ‘Experience’

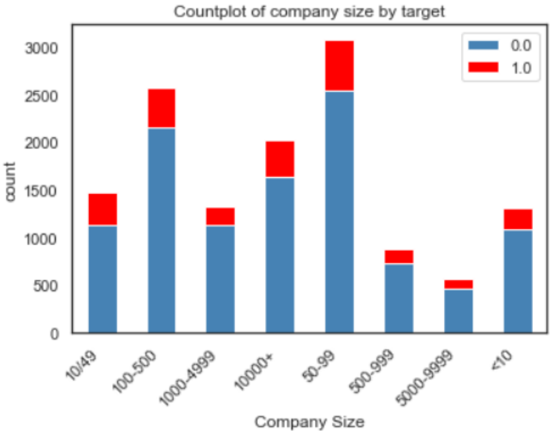




From the counter plot we can interpret that people having experience >20 years are very less likely to switch company. People having less experience are more likely to switch their job. This is indeed an important feature as it affects the target value strongly.

- Feature 9 – ‘Company_size’

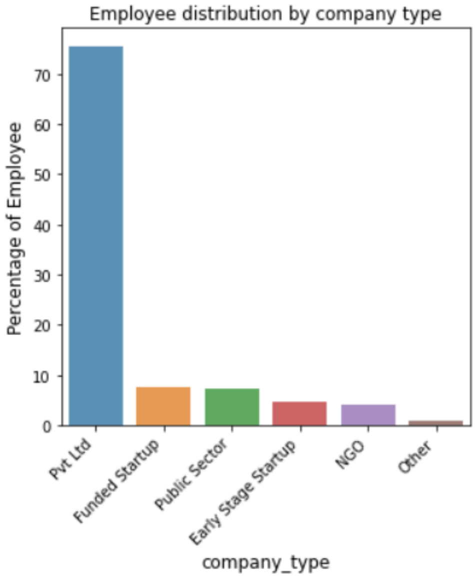


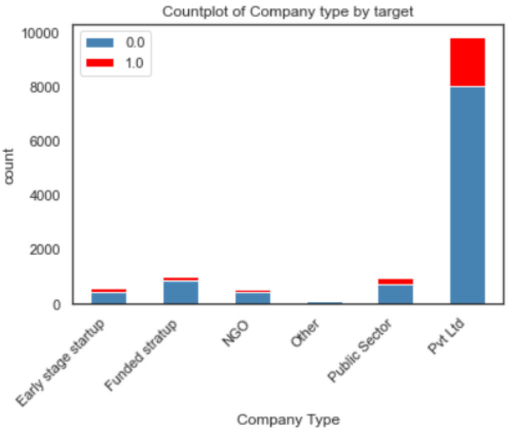


As we can see the ratio of leaving the job vs staying in the same company is almost same for all the companies irrespective of the company_size.

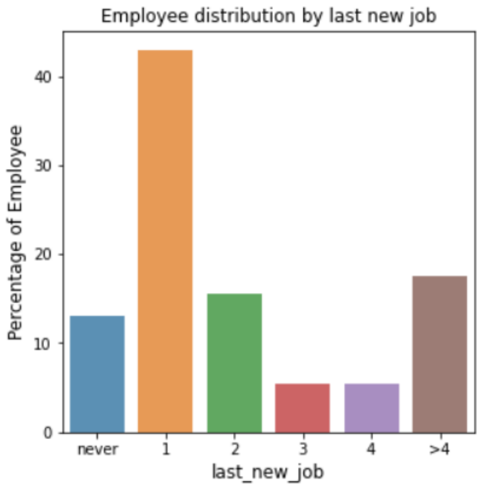
- Feature 10 – ‘Company_type’

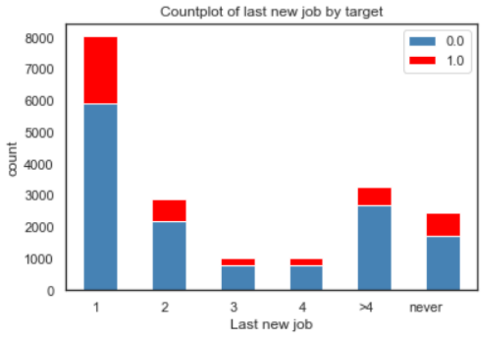
From below plots we can observe that the percentage of company_type=’Pvt_ltd’ is highest .





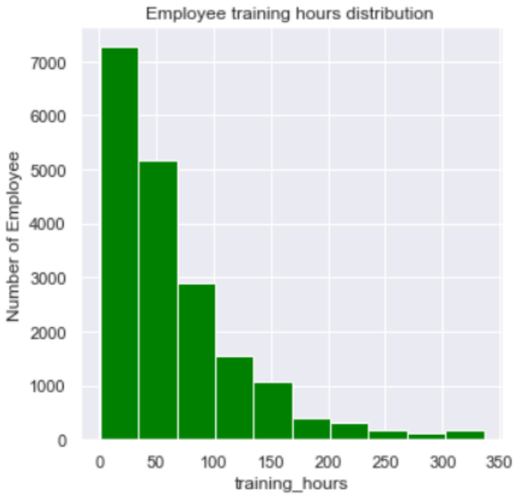
- Feature 11 – ‘Last_new_job ‘





If the employees are having work experience of 1-2 years or they have no prior experience then they more likely to switch.

- Feature 12 – ‘Training hours ‘



A high percentage of employees have undergone very few amount of training. The distribution is highly negatively skewed.

7. Statistical Analysis

Now that we have learnt few aspects of the available features, it is now time to take a deep dive into the variables which have very high number of missing values. As mentioned in section 5, the best solution is to drop these columns from the analysis than imputing 30% observations and introducing bias. However, the question arises, what if these variables are important variables with respect to the target variables? Then maybe omitting them will lead to poor classification.

Therefore, for each of these 4 variables, first we will carry out **chi square test for independence with the "target" variable at 5% level of significance.** If the outcome of the statistical test comes out to be significant, we will keep the variable, otherwise we will remove it.

a) Gender

Table 2: Table showing distribution of Gender with respect to Target

Gender	Target=0	Target=1	Total
Male	10,209	3,012	13,221
Female	912	326	1238
Other	141	50	191
Total	11,262	3,388	14,650

The Null and Alternative hypotheses for the problem are

H_0: Gender variable and target variable are independent vs **H_1: They are associated**

There are 3 classes of Gender and 2 classes of Target variable. Hence the Test statistic would follow a chi-square distribution with degrees of freedom= (3-1) * (2-1) = 2

The observed value of the test statistic is 9.042181 with p-value = 0.01087<0.05

Since p-value<0.05, based on the given data at hand at 5% level of significance we reject the null hypothesis and conclude that "gender" and "target" are associated. **This means we cannot drop the gender variable**.

Just to divert a little, we have seen that;

In our current society, gender equality is a very serious and sensitive issue, especially in the workplace. A person's ability or decision in a workplace should not depend on that person's gender. Everyone should be treated equally irrespective of the sexual orientation. However, we might not be that progressive yet, the problem of gender discrimination still persists in the society which may lead to people being not comfortable to disclose their gender.

Thus, we will tag the missing observations of gender column as a new class of "not_disclosed". This will enable us to take care of the social issue mentioned above as well as making our model robust to observations where gender is missing.

b) Company_size

Table 3: Table showing distribution of Company_size with respect to Target

	<10	10-49	50-99	100-500	500-999	1000-4999	5000-9999	10000+	Total
Target=0	1,084	1,127	2,538	2,156	725	1,128	461	1,634	10,853
Target=1	224	344	545	415	152	200	102	385	2,367
Total	1,308	1,471	3,083	2,571	877	1,328	563	2,019	13,220

The Null and Alternative hypotheses for the problem are

H_0: company_size and target are independent vs **H_1: They are associated**

There are 8 classes of company_size and 2 classes of Target variable. Hence the Test statistic would follow a chi-square distribution with degrees of freedom= (8-1) * (2-1) = 7

The observed value of the test statistic is 45.5318 with p-value = 1.078e-07<0.05

Since p-value<0.05, based on the given data at hand at 5% level of significance we reject the null hypothesis and conclude that “company_size” and “target” are associated. **This means we cannot drop the company_size variable** and will tag the missing observations as “unknown”.

c) company_type

Table 4: Table showing distribution of Company_type with respect to target

	Early stage startup	Funded startup	NGO	Other	Public Sector	Pvt Ltd	Total
Target=0	461	861	424	92	745	8,042	10,625
Target=1	142	140	97	29	210	1,775	2,393
Total	603	1,001	521	121	955	9,817	13,018

The Null and Alternative hypotheses for the problem are

H_0: company_type and target are independent vs **H_1: They are associated**

There are 6 classes of company_type and 2 classes of Target variable. Hence the Test statistic would follow a chi-square distribution with degrees of freedom= (6-1) * (2-1) = 5

The observed value of the test statistic is 35.0355 with p-value = 1.48e-06<0.05

Since p-value<0.05, based on the given data at hand at 5% level of significance we reject the null hypothesis and conclude that “company_type” and “target” are associated. **This means we cannot drop the company_type variable** and will tag the missing observations as “unknown”.

d) major_discipline

Table 5: Table showing distribution of major_discipline with respect to target

	Arts	Business Degree	Humanities	No Major	Other	STEM	Total
Target=0	200	241	528	168	279	10,701	12,117
Target=1	53	86	141	55	102	3,791	4,228
Total	253	327	669	223	381	14,492	16,345

The Null and Alternative hypotheses for the problem are

H_0: major_discipline and target are independent vs **H_1: They are associated**

There are 6 classes of major_discipline and 2 classes of Target variable. Hence the Test statistic would follow a chi-square distribution with degrees of freedom= (6-1) * (2-1) = 5

The observed value of the test statistic is 12.207 with p-value = 0.03<0.05

Since p-value<0.05, based on the given data at hand at 5% level of significance we reject the null hypothesis and conclude that "major_discipline" and "target" are associated. **This means we cannot drop the company_type variable** and will tag the missing observations as "unknown".

Note: After imputation, the chi-square test for independence were again carried for these 4 variables. This time also, all four variables have come out to be significant with respect to target.

8. Feature selection and Data Pre-Processing

As per our analysis in section 5 and section 7, we follow the below mentioned rules for missing value imputation.

Table 6: Table showing the missing value imputation strategy

Column name	Data Type	Imputation Rule
Gender	Categorical (Nominal)	Imputing by a new class "not_disclosed"
Enrolled_university	Categorical (Nominal)	Imputing by mode
Education_level	Categorical (Ordinal)	Imputing by median
Major_discipline	Categorical (Nominal)	Imputing by a new class "unknown"
Experience	Categorical (Ordinal)	Imputing by median
Company_size	Categorical (Ordinal)	Imputing by a new class "unknown"
Company_type	Categorical (Nominal)	Imputing by a new class "unknown"
Last_new_job	Categorical (Ordinal)	Imputing by median

Upon using this strategy, there is no loss of information. The training dataset is intact with 19,158 observations.

Since the features contain both qualitative and quantitative variables, it is difficult to find pairwise association among all of them. **Therefore, we are not performing any feature selection explicitly**.

Only two variables are numeric in type, **city_development_index** and **training_hours**, which have correlation of **0.00192** between them. So, we are keeping both of them.

Thus, our final training dataset will have every mentioned variable with all the observations.

Now, we need to make our data suitable for modelling. There are several categorical features here, which **needs to be encoded properly**.

- For numeric variables we will keep them as it is. We will just normalize the variable “training_hours” as the other numeric variable is already normalized.
- For ordinal variables we will use OrdinalEncoder as we can create a logical ordering of the classes. It will give least value to the lowest ordered class.
- For the rest nominal variables, we will use One-Hot-Encoder, which will create k-1 number of binary variables where k = number of classes of the original variable. Otherwise, the feature matrix would become singular.

Table 7: Table showing the categorical data encoding strategy

Column name	Data_type	Encoder
city	Nominal	OneHot Encoder
city_development_index	Numeric	None
gender	Nominal	OneHot Encoder
relevant_experience	Ordinal (having relevant experience is more preferred)	OrdinalEncoder
enrolled_university	Ordinal (full time course is better than no enrollment)	OrdinalEncoder
education_level	Ordinal	OrdinalEncoder
major_discipline	Nominal	OneHot Encoder
experience	Ordinal (More the experience, better the candidate)	OrdinalEncoder
company_size	Ordinal (“unknown” is given the least preference)	OrdinalEncoder
company_type	Nominal	OneHot Encoder
last_new_job	Ordinal (more the gap, less the preference)	Ordinal Encoder
training_hours	Numeric	None

After carrying out all the above-mentioned pre-processing and encoding, **the final dataset contains 19,158 rows with 145 feature vectors**.

9. Train-Test split and Test Dataset

We split the above-mentioned dataset into train and validation dataset with split ratio as **80:20**. Before splitting we shuffled the dataset randomly. **The training dataset contains 15,326 observations and the rest 3,832 observations are in test dataset**.

10. Classification model building

Finally, we are at the stage of building the classification model. Since the dependent variable “target” has two classes only, it will be a **binary classification model.** Several classification algorithms are out there in the literature, none of which is superior to the other. Every model has some drawbacks. We will use few of them in our analysis and will choose the best one based on the performance on training and validation dataset.

We will mostly apply **Logistic regression, decision tree** and **random forest** for our problem, which are very popular and widely used. However, we will also implement few new things **which have not been used earlier for this problem, as best to our knowledge**.

Statistical aspect of Logistic Regression: We all know what logistic regression is. It uses sigmoid function to bound the outputs within 0 and 1 as its core, the algorithm actually models the probability of dependent variable being 1.

Through the eye of statistical predictive modelling, logistic regression belongs to the wide range of **Generalized linear models (GLM).**

Under the framework of GLM, the dependent variable Y is assumed to be generated from some exponential family of distributions and is related to the feature vector **X** linearly through some **link functions**. Precisely the model can be written as

$$Y = g^{-1}(\boldsymbol{X}\beta) + \varepsilon \text{ , where } \varepsilon \text{ is the unaccountable error with 0 mean}$$

Hence, $E[Y|\boldsymbol{X}] = g^{-1}(\boldsymbol{X}\beta)$. This is the final form of GLM, where we predict the expected value of Y, given the feature vector **X** as a linear function of the parameter β.

This link function g plays a pivotal role here. Depending on the distribution of Y, we choose g accordingly, which eventually gives a wide range of different linear models we use out there.

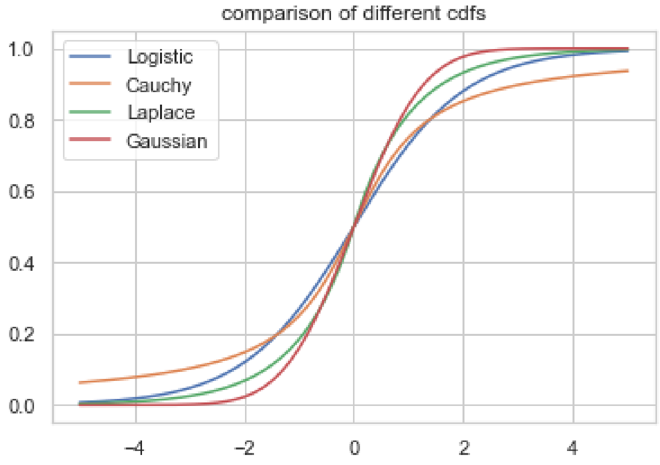
- If **Y is Gaussian i.e., it can take continuous values** over the whole real line, then we can choose $g^{-1}(x) = x$, as the identity function which will give the familiar **linear regression.**
- If Y can take binary values 0 and 1 then Y can be modeled by a **Bernoulli random variable,** and E[Y|**X**] would then imply P[Y=1] which will lie between 0 and 1. But on the right-hand side, $\boldsymbol{X}\beta$ can take any value over the whole real line. We need to select $g^{-1}(.)$, such that its input variable's range is whole real line, and the output is bounded between 0 and 1 only.
  Though it may sound very hard to find, but in the theory of probability we have a wide range of such functions called the **Cumulative distribution functions (CDF) of probability distributions.** One such probability distribution is the **Logistic distribution** with pdf as (for 0 mean and unit standard deviation):
  $$f(x) = \frac{e^{-x}}{1 + e^{-x}}\ , x \in \mathbb{R}$$
  And CDF as,
  $$F(x) = \frac{1}{1 + e^{-x}}$$
  When we take $g^{-1}(x) = F(x) = \frac{1}{1+e^{-x}}$, in the GLM equation, that gives us nothing but the concept of **Logistic regression.**
- Now the question arises can we some other CDFs? There are many continuous probability distributions out there that is valid upon the whole real line. Like **Gaussian distribution, Cauchy distribution, Laplace distribution**. The pdf and cdf of the standard versions of these distributions are as follows,
  1. Standard Gaussian distribution
     $$\text{pdf: } f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, x \in \mathbb{R}$$
     $$\text{cdf: } F(x) = \Phi(x)\ , x \in \mathbb{R}$$
  2. Standard Cauchy distribution
     $$\text{pdf: } f(x) = \frac{1}{\pi(1+x^2)}, x \in \mathbb{R}$$
     $$\text{cdf: } F(x) = 0.5 + \frac{\tan^{-1} x}{\pi}\ , x \in \mathbb{R}$$
  3. Standard Laplace distribution

$$\text{pdf: } f(x) = \frac{1}{2}e^{-|x|}, x \in \mathbb{R}$$

$$\text{cdf: } F(x) = 0.5 + 0.5 * \text{sgn(x)} * \left(1 - e^{-|x|}\right), x \in \mathbb{R}$$

where sgn(x)=1 if x>0 , 0 if x<=0

If we plot these cdfs,



it tells that these 3 cdfs also can be the potential candidates of being link function alternate to logistic. Actually, the Gaussian cdf is widely used in econometric models and is termed as **Probit regression.** However, the other two are not used out there. Logistic regression is the most popular and widely used **because of its interpretability**.

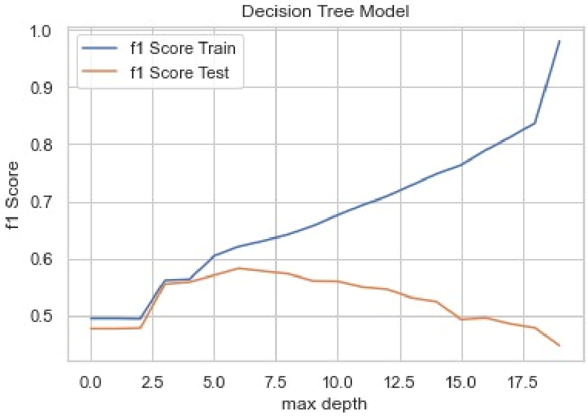
Under Logistic framework, **the log odds ratio** of probability of Y being 1 to probability of Y being 0. Such well-defined interpretability does not exist for the other three cases.

11. Results and Interpretation

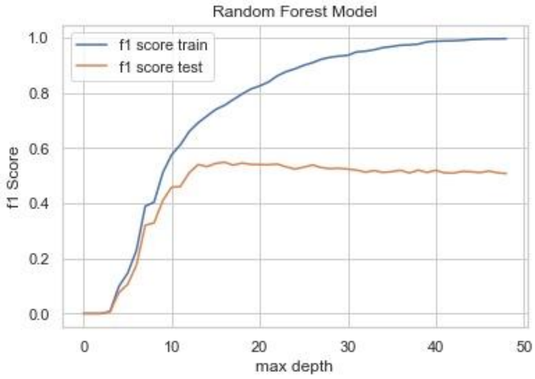
Table 8: Table showing the results of our classification models

Algorithm	Train F_1 score	Test F_1 score	Train Accuracy	Test Accuracy
Logistic Regression	0.60605	0.58992	0.75943	0.75365
Probit Model	0.44897	0.40666	0.78089	0.77688
Cauchy Link function	0.44706	0.43184	0.76145	0.76722
Laplace Link function	0.54753	0.54288	0.76145	0.76722
Decision Tree	0.6042331	0.570455	0.79331941	0.803601
Random Forest	0.755287	0.5788372	0.8890121	0.7974947

- Decision Tree: We have implemented Decision tree, which works well with categorical data.
  Here, we have tried max depth 1-50 and got the optimal depth at 6.



- Random Forest:



12. Conclusion

- Built an end-to-end ML model, consisting of all stages
- Performed EDA explicitly to generate insight from the data
- Demonstrated Logistic regression from a statistical perspective and incorporated two new link functions from our side, which have not been used before for this dataset.

13. Challenges faced/ Future work

- Checking multivariate outlier detection that too in mixed data
- The data being imbalanced, the model performance was not that good.
- Applying Neural network-based classification model
- Applying techniques like bootstrapping, SMOTE to tackle the imbalanced problem.

14. Invidual Contribution

- **Data visualization: Dyotana Das, Aashish, Soumalya Nandi**
- **EDA: Soumalya Nandi, Muttaqi Ahmed Alladin, Shaurya Tiwary**
- **GLM part: Soumalya Nandi**
- **Logistic Regression: Dyotana Das, Aashish**
- **Probit, Cauchy, Laplace: Soumalya Nandi**
- **Random Forest, Decision Tree: Muttaqi Ahmed Alladin, Shaurya Tiwary**
- **PPT and Report creation: Everyone**

14. Reference

- https://towardsdatascience.com/classification-using-different-link-function-than-logit-probit-logistic-trilogy-part-3-df9922b1acf1
- https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d
- https://www.kaggle.com/khotijahs1/predict-who-will-move-to-a-new-job