DS226 : ICAIML Project

Will your employee switch ?

HR Analytics using Machine Learning and Visualization

Aashish (MTech CDS – 19689)

Dyotana Das (MTech CDS – 19223)

Shaurya Tiwari (MTech CDS – 18168)

Soumalya Nandi (MTech CDS – 18016)

Muttaqi Ahmad Alladin (Mtech Res CDS – 18409)

13th December 2021

Problem Statement & Motivation

Problem Statement

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..

Motivation

- □ Improved hiring decisions
- □ Reduced attrition rate
- □ More productive workforce
- □ Better Employee Insights
- □ Improved hiring process

Datasets

Task

- Binary Classification

Data Set

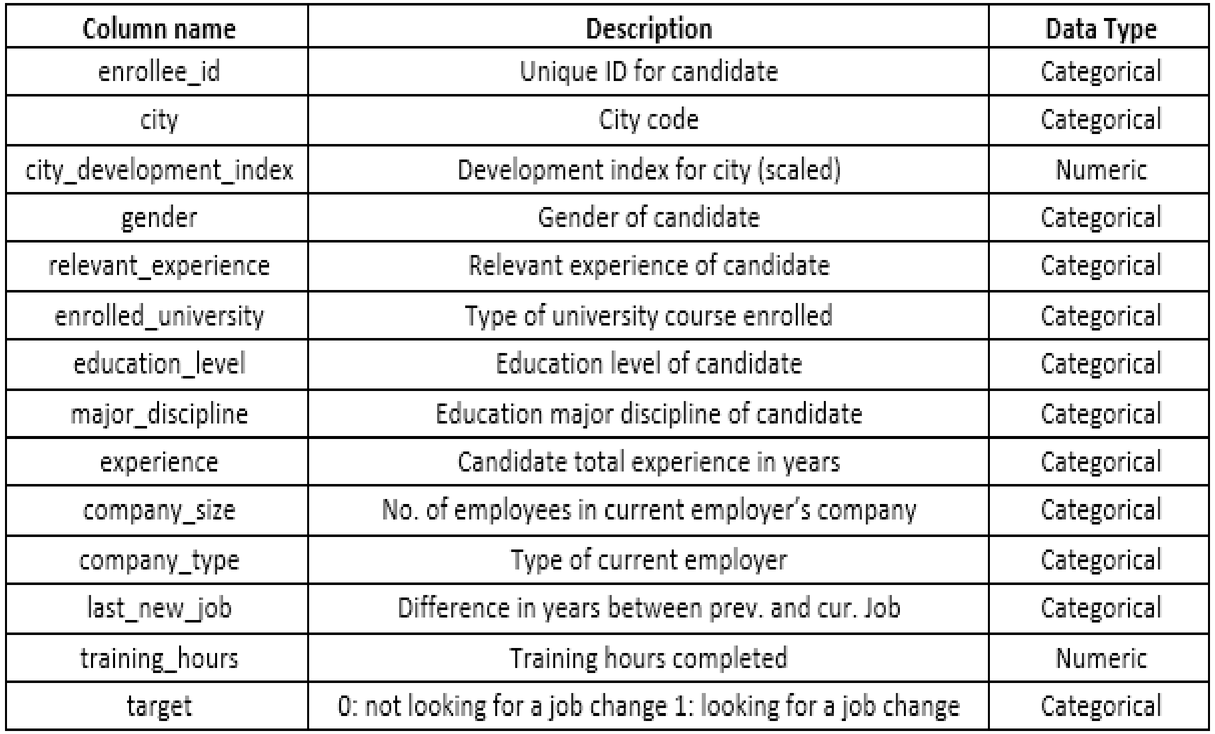
- 19158 rows, 13 predictors, 1 Binary reponse

Response

- Target: 0- Not looking for job change 1- looking for job change

About Dataset

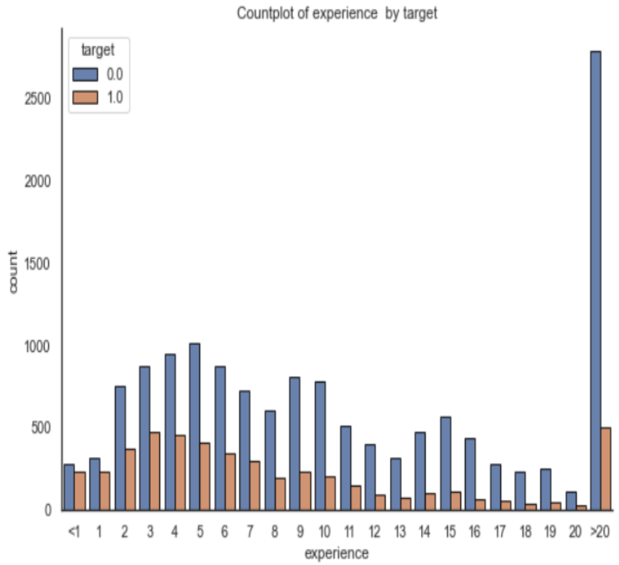
- Imbalanced dataset
- Target:0 – 14381 rows and Target: 1 – 4777 rows
- 12 Categorical and 2 Numeric features
- 8 columns have NA’s or bad data entry points.

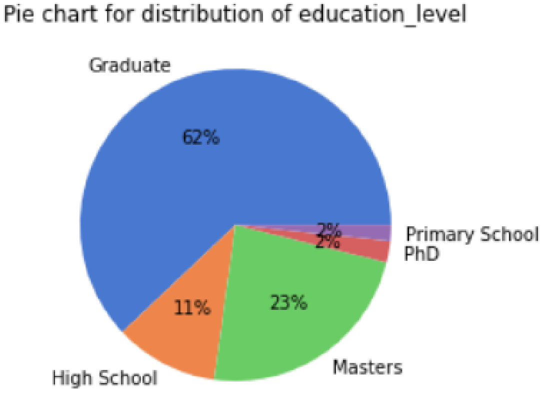


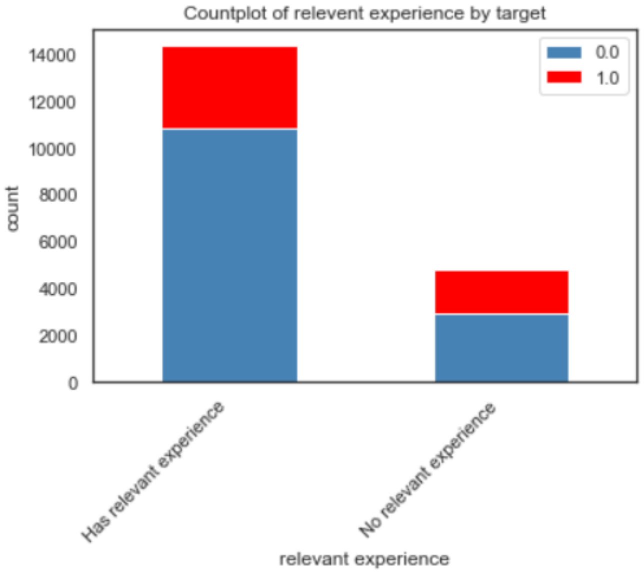
Column name	Description	Data Type
enrollee_id	Unique ID for candidate	Categorical
city	City code	Categorical
city_development_index	Development index for city (scaled)	Numeric
gender	Gender of candidate	Categorical
relevant_experience	Relevant experience of candidate	Categorical
enrolled_university	Type 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 cur. Job	Categorical
training_hours	Training hours completed	Numeric
target	0: not looking for a job change 1: looking for a job change	Categorical

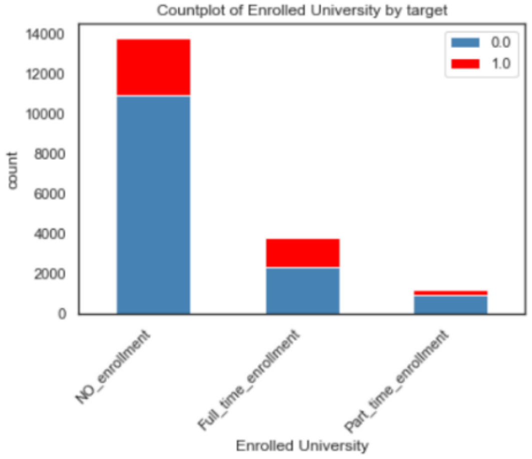
https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists

Features Visualization









- employee having experience >20 years are very less likely to switch the company

- Almost 62% of the employees are from graduate level.

- People having relevant experience are more likely to stay in the same company.

- Employees enrolled fulltime are more likely to switch job. Almost 40% people left their job.

Missing Value Analysis / Imputation (1/2)

Observations

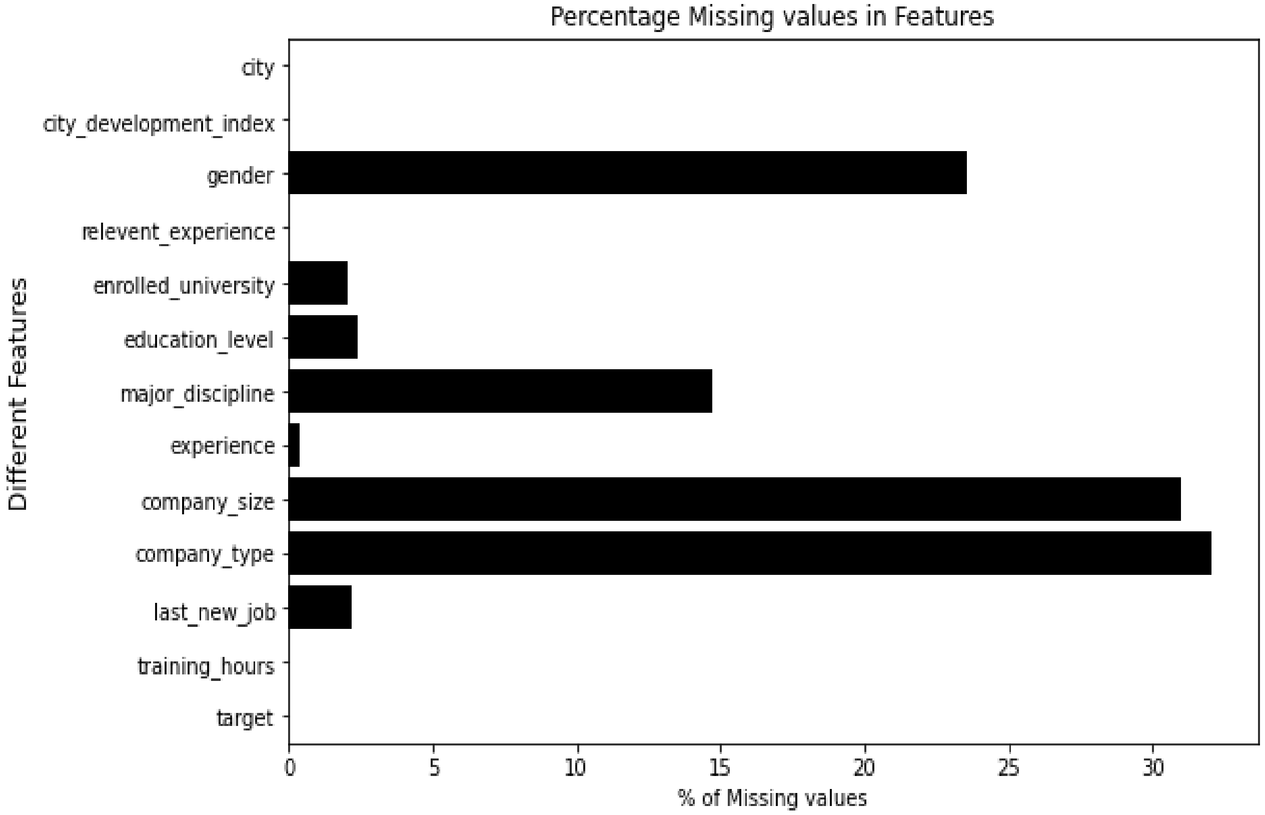
8 Columns have missing values. Out of which 4 have quite high % of missing values

Dropping the rows with missing values will reduce the training data size drastically

So, one option can be to drop the columns full. but for that we need to check whether the columns are significant or not through visualization and some statistical tests.

Chi-square test for independence carried out for these four columns

For the columns with less number of missing values, imputation was done with median and mode accordingly.



Missing Value Analysis / Imputation (2/2)

- H0: The feature and the target variable are independent
- H1: They are associated
- Level of significance is 5%

Feature	Test statistic	dof	p value
Gender	9.04218	2	0.01087
Company_size	45.5318	7	1.078e-07
Company_type	35.0355	5	1.48e-06
Major_discipline	12.207	5	0.03

- No information is lost. We are keeping all the observations
- **Our model is robust to observations where people don’t want to disclose their gender.**

Statistical Analysis : Chi-square test for independence

Conclusion : should not drop all the columns as there is no or very less association between variables (with high missing value %) & target variable

Gender, major_discipline, company_size, company_type

For these 4 columns, make the missing values as separate category as "unknown“ or “not_disclosed”

Education_level, experience, last_new_job

For these features missing values are imputed by median

Enrolled_university

Missing values are imputed by mode

Generalized Linear model

GLM framework: $E[Y|\boldsymbol{X}] = g^{-1}(\boldsymbol{X}\beta)$

- 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.

- In the theory of probability we have a wide range of such functions called the **Cumulative distribution functions (CDF) of probability distributions**

Logistic Distribution

$$f(x) = \frac{e^{-x}}{1+e^{-x}}, x \epsilon \mathbb{R}$$

$$F(x) = \frac{1}{1+e^{-x}}$$

Gaussian Distribution

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, x \epsilon \mathbb{R}$$

$$F(x) = \Phi(x)$$

Cauchy Distribution

$$f(x) = \frac{1}{\pi(1+x^2)}, x \epsilon \mathbb{R}$$

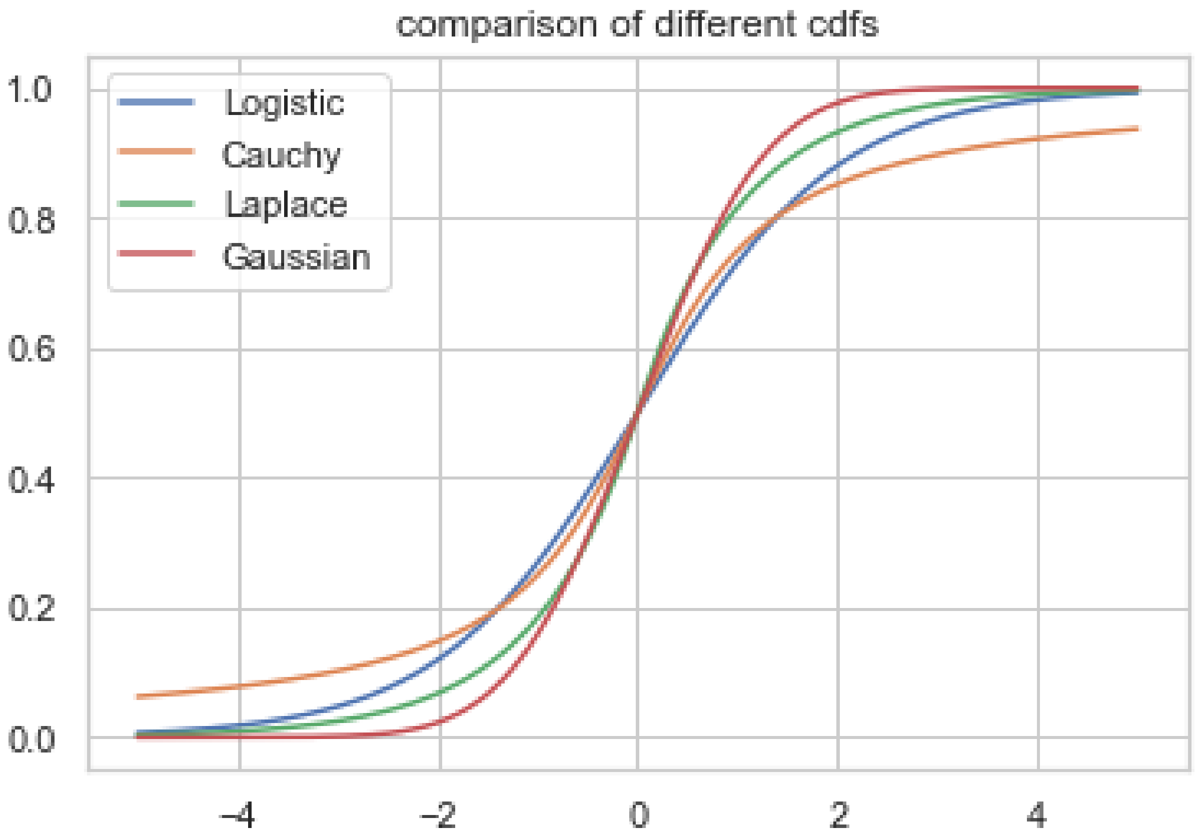
$$F(x) = 0.5 + \frac{\tan^{-1} x}{\pi}, x \epsilon \mathbb{R}$$

Laplace Distribution

$$f(x) = \frac{1}{2} e^{-|x|}, x \epsilon \mathbb{R}$$

$$F(x) = 0.5 + 0.5 * \mathrm{sgn}(\mathrm{x}) * \left(1 - e^{-|x|}\right), x \epsilon \mathbb{R}$$

Classification model building

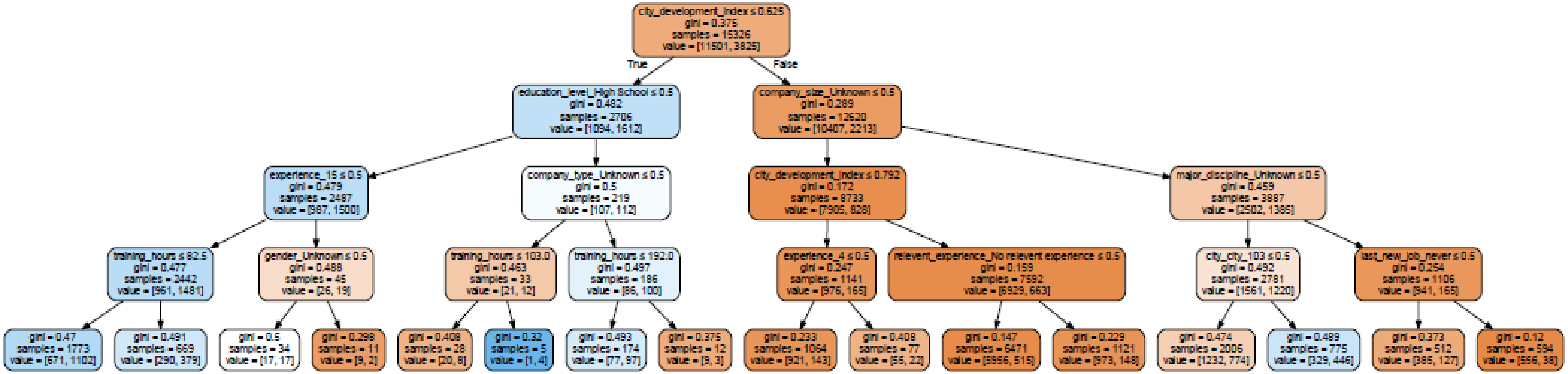


- Logistic regression is the most popular and widely used **because of its interpretability**
- The Gaussian cdf is widely used in econometric models and is termed as **Probit regression.**
- However, the other two are not used out there

Model building strategy

- **Train test splitting with 80:20 ratio.**
- **Dataset is very much imbalanced, hence focus kept on F1 score than accuracy.**
- **Logistic regression, Probit regression, Cauchy Link function based model, Laplace link function based model, random forest, decision tree are applied.**
- **The numeric columns are normalized between 0 and 1**
- **The ordinal columns are encoded with ordinalencoder**
- **The nominal columns are done OneHot encoder.**
- **Final training dataset contains 15,326 rows with 145 columns.**

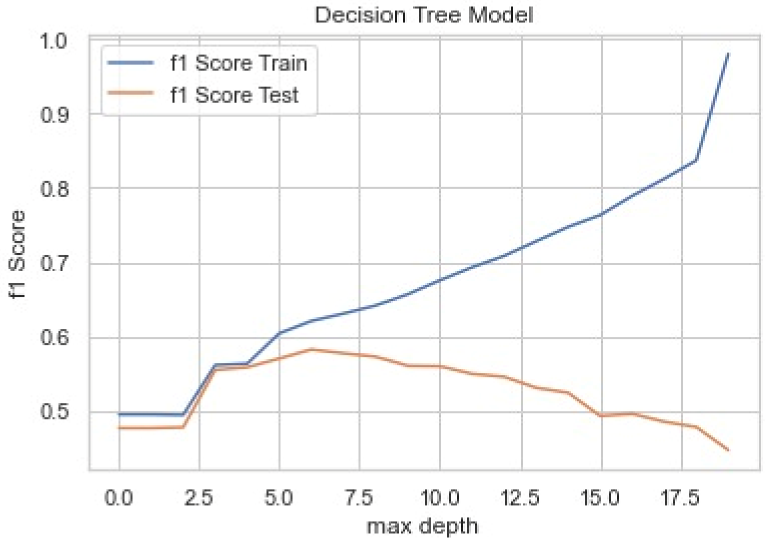
Decision Tree Model



Decision Tree (for max. depth 4)

Tried maximum depth from 1-50 and got the optimal depth at 6.

City development index is the most significant variable to be splited first, which says that people have a tendency to move towards developed cities for better job opportunities



Decision tree f1 score plot

Random Forest Model

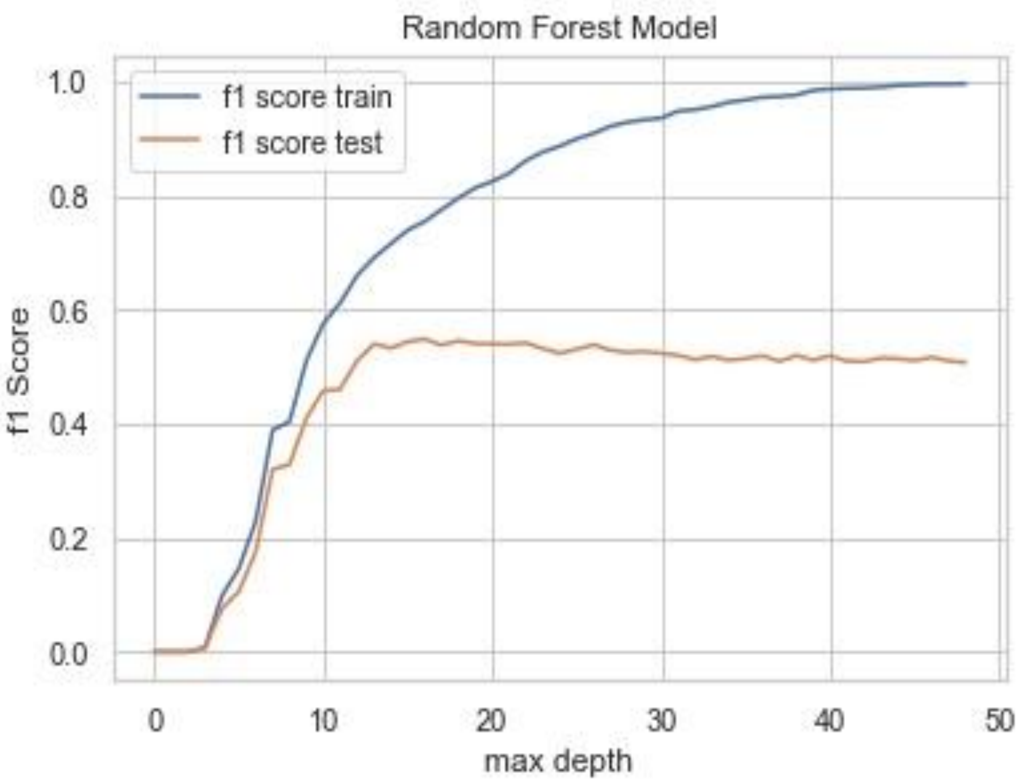
Highly accurate classifier and learning is fast

Optimal depth we got is close to 17

N_trees : 100

Interpretability is lost

F1 score plot for different max depth values



Final results

Logistic Regression

Training f1 Score:	60.59%
Training accuracy:	75.87%
Test f1 Score:	58.53%
Test accuracy:	75.13%

Probit Regression

Training f1 Score:	44.89%
Training accuracy:	78.08%
Test f1 Score:	40.66%
Test accuracy:	77.68%

Decision tree (max depth 6)

Training f1 Score:	60.42%
Training accuracy:	80.36%
Test f1 Score:	57.04%
Test accuracy:	79.33%

Random Forest (max depth 17)

Training f1 Score:	75.52%
Training accuracy:	88.90%
Test f1 Score:	57.88%
Test accuracy:	75.13%

Cauchy Link Function

Training f1 Score:	44.70%
Training accuracy:	76.14%
Test f1 Score:	43.18%
Test accuracy:	76.72%

Laplace Link Function

Training f1 Score:	54.75%
Training accuracy:	76.14%
Test f1 Score:	54.28%
Test accuracy:	76.72%

Random Forest is more accurate among all the models which we experimented**. Hence this is our final model**.

We also tried few new things like Cauchy & Laplace link function

Challenges Faced/ Future Work

Challenges

Checking multivariate outlier detection that too in mixed data

The data being imbalanced, the model performance was not that good.

Future Work

Applying Neural network based classification model

Applying techniques like bootstrapping, SMOTE to tackle the imbalanced problem.

Contribution and References

- Incorporated almost all stages of an ML project
- 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.

References

✓ 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

Thank You