

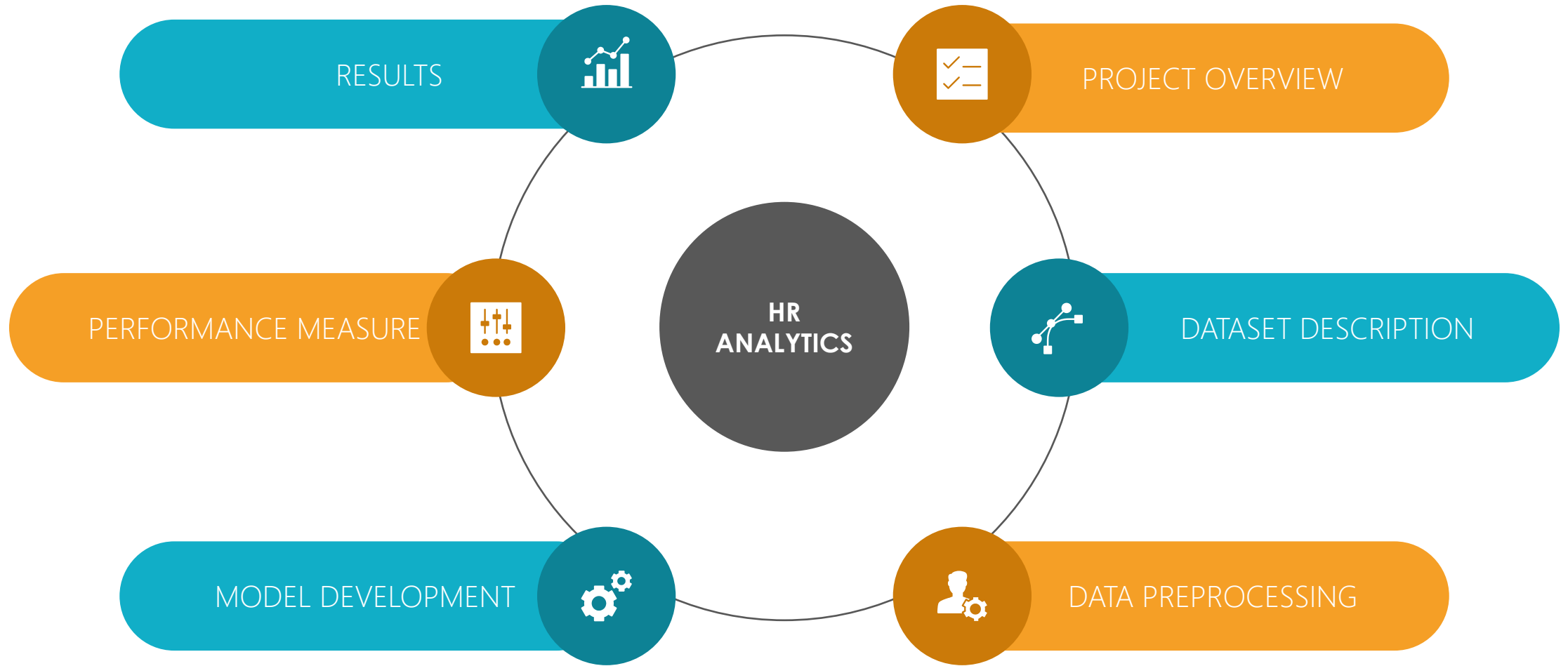


HR Analytics : Job change







Presented by
Adina Dingankar
Pranay Bhakthula
Rehapiadarsini Manikandasamy

Submitted for
DATS 6103


• Presentation Outline •



Project Overview

-  The focus of this project is to predict the probability of a candidate to look for a new job or who will continue to work for the company
-  It will be demonstrated using three machine learning algorithms:
 -  Decision Tree Classifier
 -  Random Forest Classifier
 -  Support Vector Classifier
-  Developed a GUI based application to display the end-to-end modelling

Dataset Description

 The dataset used has educational and professional records of various candidates who have completed training in a company

 The dataset has 19158 observations and 14 features, most of them are categorical

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

 8 amongst 14 features have missing values


```
These are the list of features in the dataset :  
city  
city_development_index  
gender  
relevent_experience  
enrolled_university  
education_level  
major_discipline  
experience  
company_size  
company_type  
last_new_job  
training_hours  
target
```

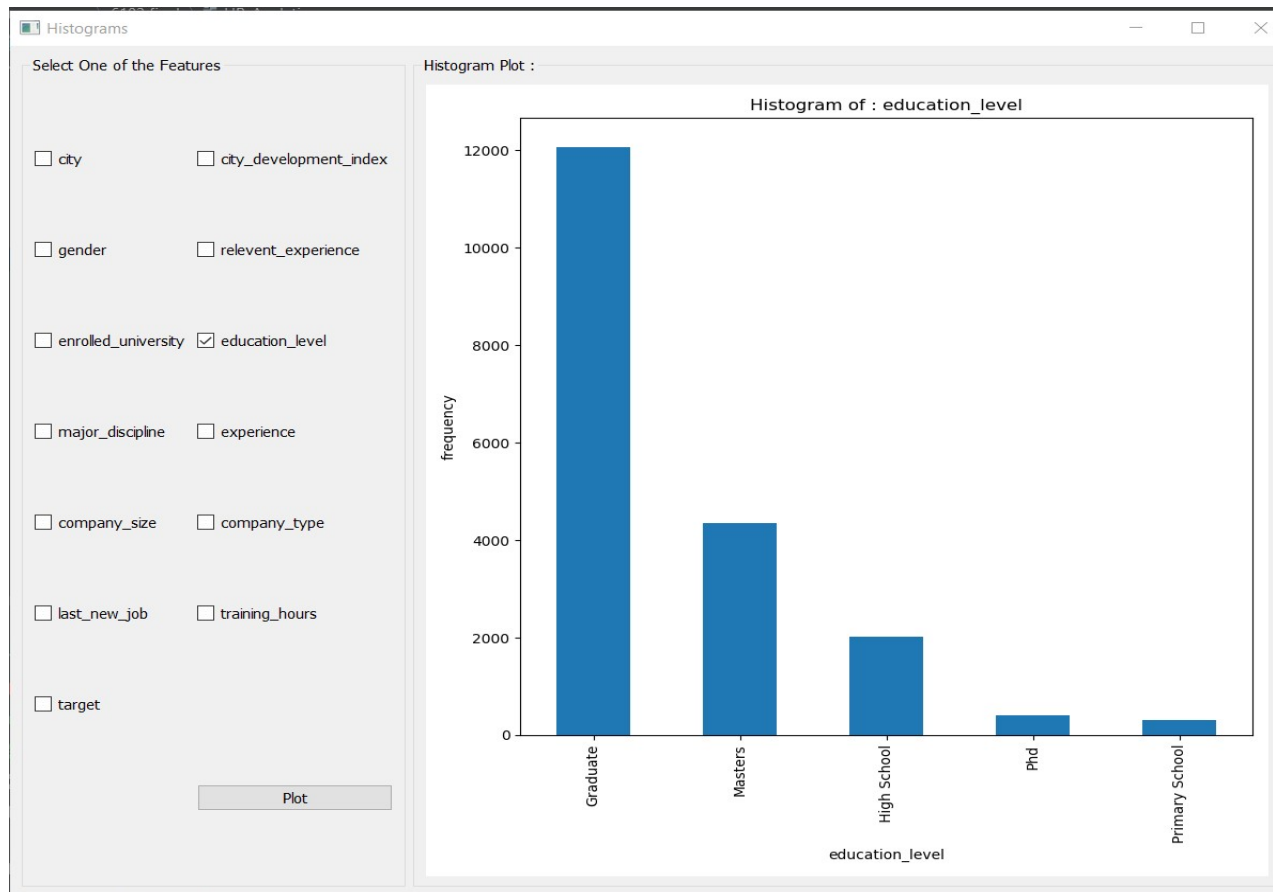
Data Preprocessing


#	Column	Non-Null Count	Dtype
0	enrollee_id	19158 non-null	int64
1	city	19158 non-null	object
2	city_development_index	19158 non-null	float64
3	gender	14650 non-null	object
4	relevent_experience	19158 non-null	object
5	enrolled_university	18772 non-null	object
6	education_level	18698 non-null	object
7	major_discipline	16345 non-null	object
8	experience	19093 non-null	object
9	company_size	13220 non-null	object
10	company_type	13018 non-null	object
11	last_new_job	18735 non-null	object
12	training_hours	19158 non-null	int64
13	target	19158 non-null	float64


- Features with null values are updated with maximum value count of their respective columns
- The column enrollee_id is dropped , since it doesn't have much influence on target
- Label Encoder is applied to the features as our use case being the classification problem
- Encoding is done to decide in a better way on how these labels must be operated and labels are converted into numeric form

EDA Analysis

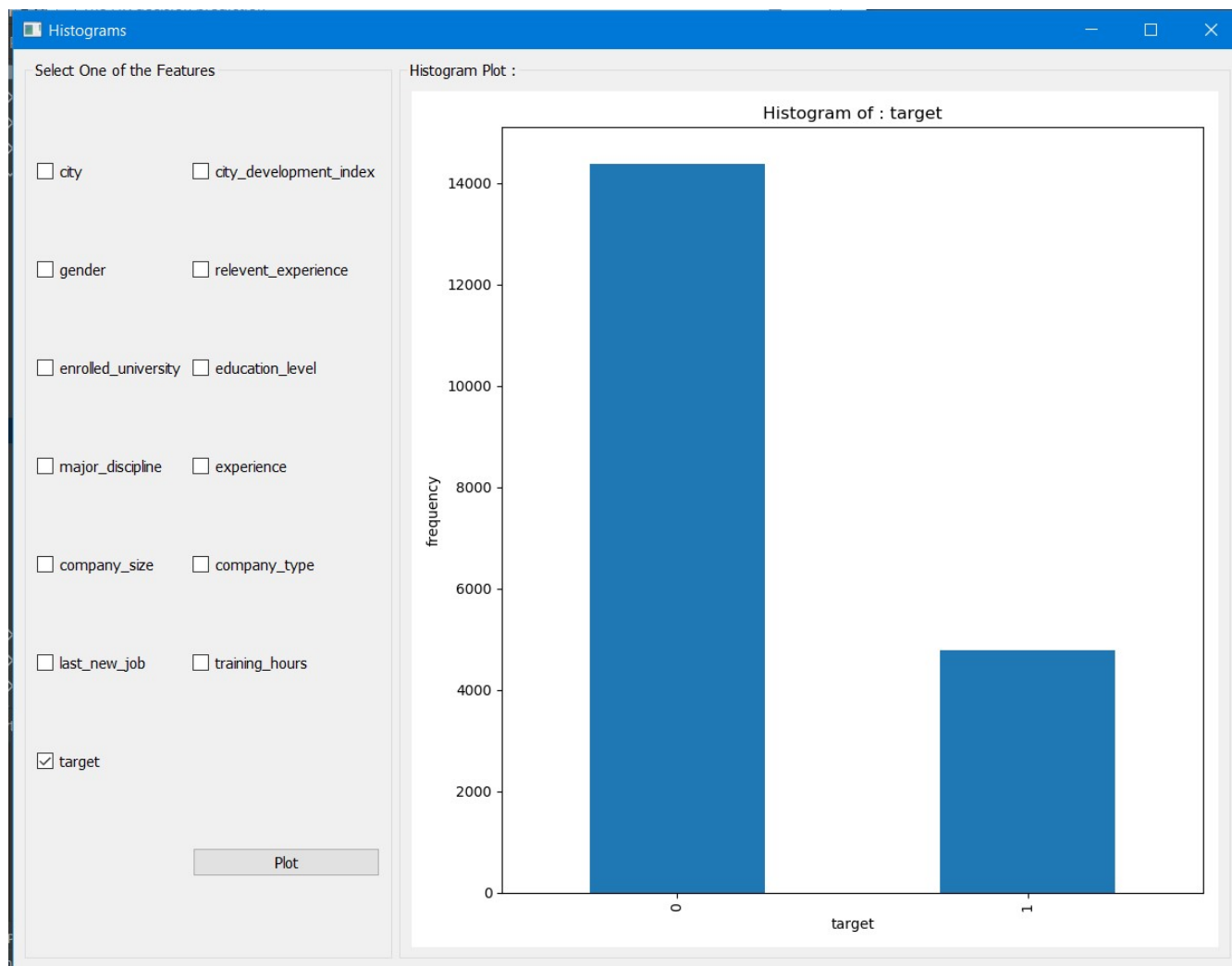
 EDA analysis option allows user to visualize histograms and scatter plots of variables



 This histogram sample provides the distribution of employee's based on their education level

 Likewise, the user can view the distribution of other variables

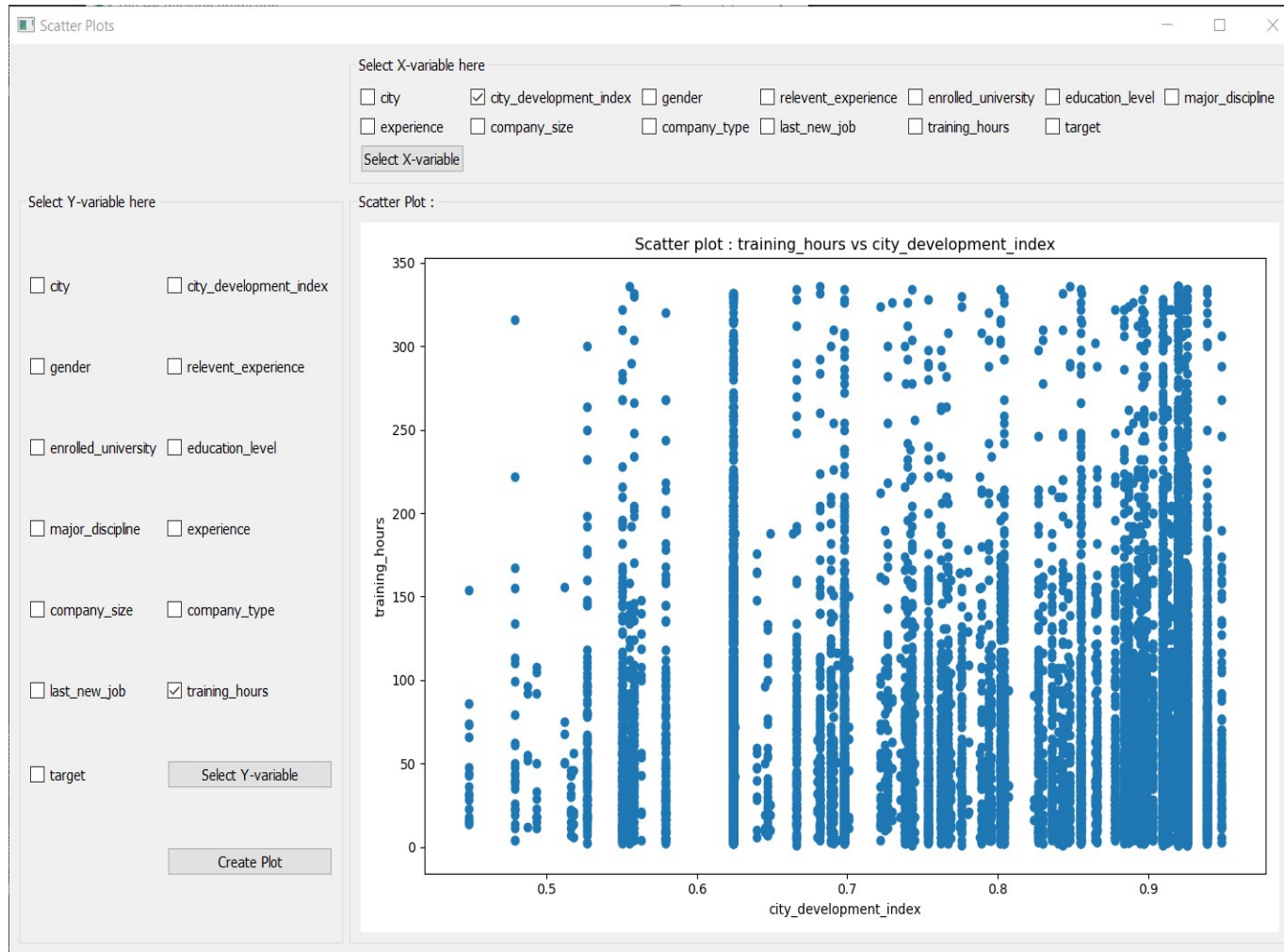
EDA Analysis



- There are observations with more number of 0's than 1's in the dataset
- Since there is a huge difference between the number of 0's and 1's there is chance of high false negatives
- And less precision on predicting on 1's

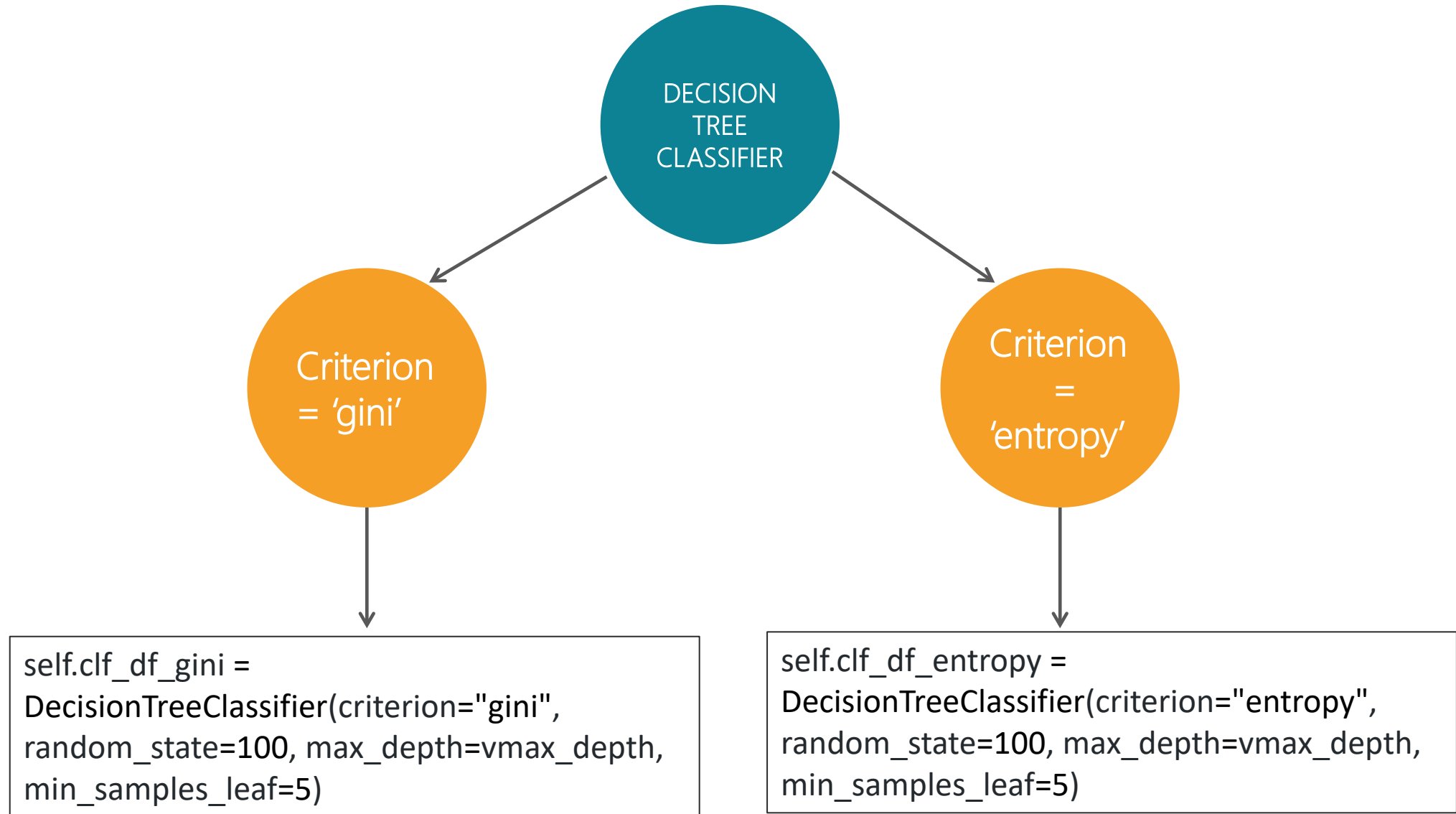
EDA Analysis

Scatter Plot







The scatterplot sample visualizes the training hours of employee against the city development index

• Model Development •




• Model Development •

RANDOM FOREST CLASSIFIER

-  The Random forest classifier creates a set of decision trees from a randomly selected subset of the training set. It is basically a set of decision trees (DT) from a randomly selected subset of the training set and then It collects the votes from different decision trees to decide the final prediction.
-  The dashboard incorporates the feature selection , no-of estimators selection and plotting the roc-auc graphs.
-  Models for gini and entropy are been built respectively by selecting the associated criterion during the model development.
-  The random forest classifier algorithm is being built using the Scikit Learn Package.

• Model Development •





SUPPORT VECTOR CLASSIFIER

 Support Vector Machines (SVM) is a widely used supervised learning method and it can be used for regression, classification, anomaly detection problems. The SVM based classifier is called the SVC (Support Vector Classifier) and which can be used it in classification problems.

 The kernel has been set to rbf for our data by default.

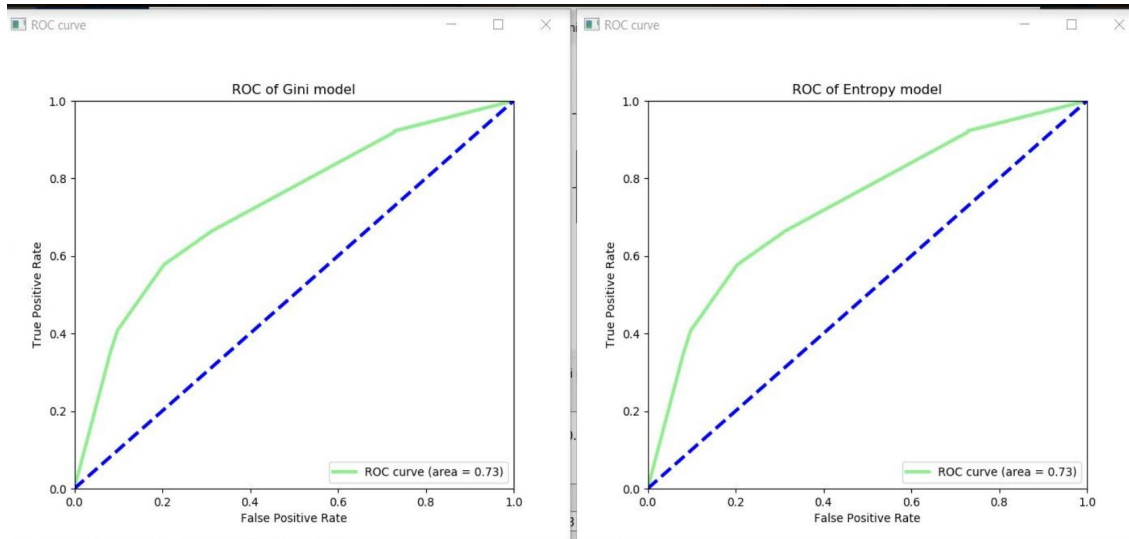
• Performance Measure •

The performance of the models are measured by :

-  Confusion matrix
-  Classification report
-  Accuracy score
-  Roc_auc curve

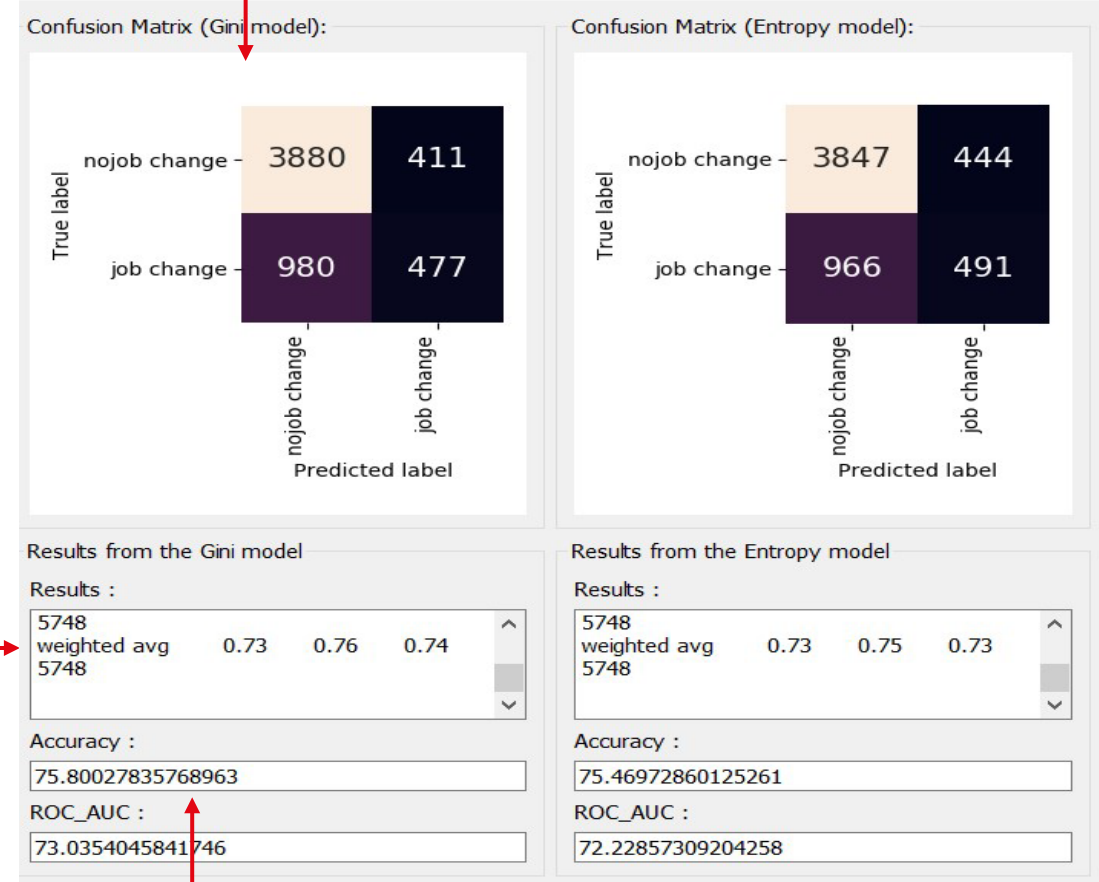
Performance Measure

The number of correct and incorrect predictions are summarized with count values and broken down by each class



ROC is a probability curve and AUC represents the degree or measure of separability

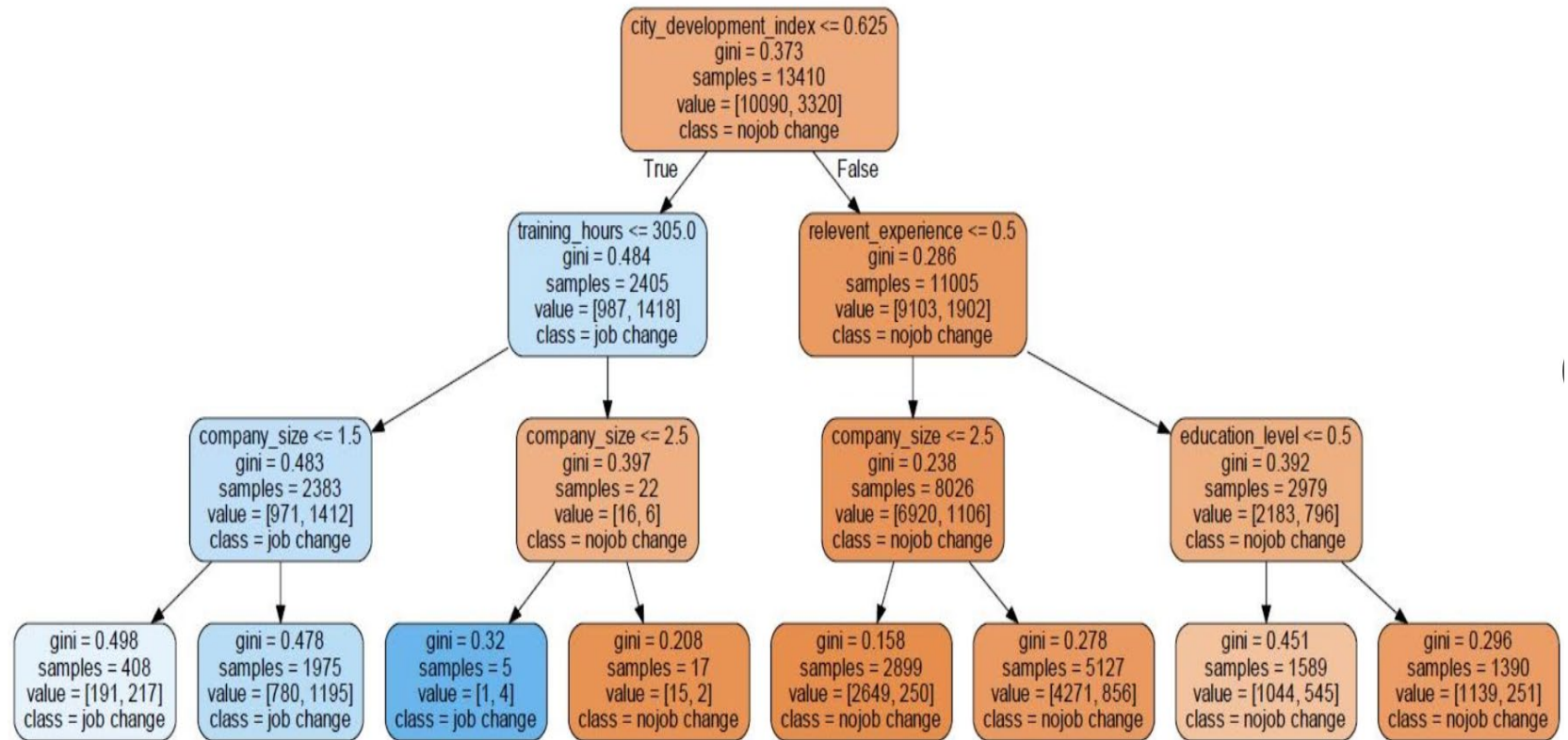
The report shows the main classification metrics precision, recall and f1-score on a per-class basis



Accuracy is the fraction of predictions our model got right

Performance Measure

Decision tree is visualized by using Graphviz



• Structure of the Application •

File

- **Exit** – It quits the entire application

Load Dataset

- **Upload Data** – It takes up dataset from user and displays the features of the dataset

EDA Analysis

- **Histogram** - This option presents a distribution of each feature in the processed dataset
- **Scatter plot** - This option displays a dot plot that shows the relation of features

ML Models

- **Decision Tree Classifier** -This option creates a dashboard with the results from the Decision Tree algorithm developed using the Sklearn Decision Tree Classifier module
- **Random Forest Classifier** – This option creates a dashboard of results generated for Random forest algorithm
- **Support Vector Machine** – This option allows user to generate a SVC model with selected features

Results

Decision Tree Classifier:

- Accuracy of model = 77.6% (Test size=30%, Max_depth=3)
- ROC_AUC value = 73.01 Precision of 0's=0.82 Precision of 1's=0.59
- The gini and entropy models have similar accuracy





Random Forest Classifier:

- Accuracy of model = 75.80% (Test size=30, No. of estimators = 10, Criterion = Gini)
- ROC_AUC value = 73.03 Precision of 0's=0.80 Precision of 1's=0.53
- Gini model has better accuracy than the entropy model

Support Vector Classifier:

- Accuracy of model = 74.65%
- ROC_AUC value = 71.64 Precision of 0's=0.77 Precision of 1's=0.55

Conclusion

-  Comparing the results of models, almost all the three models has accuracy value more than 70%
-  Decision Tree Classifier tops the list by having the highest accuracy of 77%
-  The decision tree and random forest models suffer when their parameter values like depth and estimators are changed
-  The models in future enhancement needs to be tuned to predict the job change class correctly



Video link:

https://drive.google.com/file/d/1Y_u4un0_inFmVXfVGb4GQ2kmbNNgqJ7d/view?usp=sharing



Any Questions?

GitHub Link: https://github.com/adingankar/FINAL_PROJECT_GROUP7