Career Path (CP) Prediction Using (AI/ML) Machine Learning

Project/Problem Statement

The development of information technology over the past several years has paved the way for a wide range of employment opportunities. This made it challenging for those looking for work to locate positions that would utilize their skill sets. Additionally, the traditional method of hiring takes a long time to sort through the many applications from candidates. Therefore, a novel strategy is required to fix the problems with the current system. The goal of the job prediction application is to categorize occupations with a high level of recall and precision by combining recent machine learning algorithms with sophisticated data cleaning procedures. Support Vector Machines and sophisticated curve fitting methods are a couple of the key algorithms utilized in this application as test/pilot versions (v 1.0)

Sub Problem/Module Statement: Refer project table **(for your project description)**

Version 1.0 (Dataset & Results) Description:

The dataset used for the application is a part of HR Analytics on the Kaggle platform. The dataset has 13 key features like enrollee_id, city, city_development_index, gender, relevent_experience, enrolled_university, education_level, major_discipline, experience, company_size, company_type, last_new_job and training_hours.

After applying data cleaning techniques on the dataset, a report was generated with the number of samples of each feature which is used for training the model.

```
=======gender=======
Male      13221
Female     1238
Other       191
Name: gender, dtype: int64

=======enrolled_university=======
no_enrollment       13817
Full time course     3757
Part time course     1198
Name: enrolled_university, dtype: int64

=======education_level=======
Graduate          11598
Masters            4361
High School        2017
Phd                 414
Primary School      308
Name: education_level, dtype: int64

=======major_discipline=======
STEM               14492
Humanities           669
Other                381
Business Degree      327
Arts                 253
No Major             223
Name: major_discipline, dtype: int64

=======experience=======
>20    3286
5      1430
```

4 1403
3 1354
6 1216
2 1127
7 1028
10 985
9 980
8 802
15 686
11 664
14 586
1 549
<1 522
16 508
12 494
13 399
17 342
19 304
18 280
20 148
Name: experience, dtype: int64

=======company_size=======
50-99 3083
100-500 2571
10000+ 2019
10/49 1471
1000-4999 1328
<10 1308
500-999 877
5000-9999 563
Name: company_size, dtype: int64

=======company_type=======
Pvt Ltd 9817
Funded Startup 1001
Public Sector 955
Early Stage Startup 603
NGO 521
Other 121
Name: company_type, dtype: int64

=======last_new_job=======
1 8040
>4 3290
2 2900
never 2452
4 1029
3 1024
Name: last_new_job, dtype: int64

Methodology

In this application, By analyzing the data gathered from the applicants either through online forms or extracting data from the submitted resumes, a supervised learning approach is used in this application to tackle the use case of forecasting the job change trend.

The data is mostly reviewed for missing values when it is obtained via the online forms. Data cleaning techniques are used on the data after such rows have been identified and removed. The data is then divided into training, validation, and test sets after this stage.

Data is synthesized to balance out the various classes in order to improve the training data so that the model won't be biased toward one class over another and overfit. Based on the sample values supplied, fictitious data is generated using the SMOTE function in the imblearn library.

The model for forecasting the status of a job transition is then developed using a support vector classifier. This strategy is used because SVMs can handle noisy data by utilizing the maximum margin hyperplane idea and have a high tolerance for outlier data.

Additionally, this method was combined with GridSearchCV and the cross-validation score to compare many model iterations and choose the model that best fits the data. Cross validation score is a crucial consideration since it provides information on the variations of many algorithm models and aids in the decision of which algorithm is best suited for the supplied data.

Version 1.0 Data & Collaboration: Kaggle Platform & SEET Lab (Client: Divya D K)

Results

Linear SVC Results

Training accuracy score: 0.8360281718083523

Testing accuracy score: 0.7653096729297146

SVC Classifier after

Training accuracy score: 0.7891871737509322

Testing accuracy score: 0.7712247738343772

Version 2.0 Data & Collaboration: Bright Initiative & SEET Lab (Client: Divya D K)

Career Path (CP) 2.0 Prototype: