Will a Pathrise fellow get placed at a company?

Introduction

Pathrise is an online program that provides 1-on-1 mentorship, training, and advice to help job seekers get the best possible jobs in tech. Every two weeks, Pathrise welcomes a new cohort of fellows. If a candidate is interested in joining their program and successfully passes all stages of their admission process, they receive an offer to join Pathrise and become a fellow. The first 2 weeks in the program are called a free trial period and a fellow can withdraw within this free trial period without any penalty. After 2 weeks, a fellow needs to sign an ISA (Income Share Agreement) with them if they want to stay in the program. The entire program lasts up to a year, including 8 weeks of the core curriculum. If a fellow is unable to find a job within a year after joining Pathrise, his/her contract is terminated. However, there might be some exceptions. For instance, if someone was on a break, pathrise may extend their contract for the period of the break.

On average, pathrise fellows who stay with them after their free trial period, it takes about 4 months to receive a final job offer. However, there is a lot of variation in fellows' outcomes. Being able to predict how fast every single fellow is going to find a job is crucial for their business. The main goal of our analysis is to derive insights around if a fellow will ultimately be placed at a company

Data

To solve the above problem, I am going to use a sample of information Pathrise collected from their fellows from the moment they joined their program.

Methodology

To solve the above problem, we need to build a classification model which predicts whether a fellow gets placed at a company or not.

Exploratory Data Analysis

The dataset has 2544 rows and 16 features. Most of the features are categorical in nature with a few being numeric. Most of the features have missing values which will need to be treated appropriately.

Data Preprocessing

The data comprises information about 4 categories of applicants
1 - Applicants who got placed through the program
2 - Applicants who failed to get placed through the program
3 - Applicants who did not accept admission offer or did not continue after the free 14-day trial period
4 - Applicants who are currently enrolled in the program
To train our model we need to analyze the data comprising the first 2 categories as stated above. I filtered the dataset based on the pathrise_status column having values('MIA', 'Placed', 'Withdrawn' and 'Withdrawn (Failed)'). There are 4 columns (id, pathrise_status, cohort_tag and program_duration_days) which do not provide any valuable information so we can get rid of them. So now we are left with 12 features including the target variable 'placed'. Out of these 12 features 9 of them have missing values. Columns highest_level_of_education, length_of_job_search, biggest_challenge_in_search and race have a low proportion of missing values so we can simply drop the corresponding rows. In the gender column ~75% of the rows contain the value 'Male' so it's good enough to impute the missing values in this column with 'Male'. There were 2 rows having a value of 'Non-Binary' and 6 rows having a value of 'Decline to Self-Identify'. We can eliminate these rows to limit the cardinality of the column to 2. The employment_status column has 5 labels ('Contractor', 'Employed Full-Time', 'Employed Part-Time', 'Student', 'Unemployed'). We can combine the labels ('Contractor', 'Employed Full-Time', 'Employed Part-

Time') into a new label called 'Employed'. 'Student' label can be clubbed together with 'Unemployed'. By doing this we end up with just 2 labels for the employment_status column. The work_authorization_status column has 9 labels ('Canada Citizen', 'Citizen', 'F1 Visa/CPT', 'F1 Visa/OPT', 'Green Card', 'H1B', 'Not Authorized', 'Other', 'STEM OPT'). We can club similar visas together. 'F1 Visa/CPT', 'F1 Visa/OPT' and 'STEM OPT' were clubbed together into a new label called 'F1'. 'Other' and 'H1B' were clubbed together into a new label called 'H1B'. So now the cardinality of this column was reduced to 6. The professional_experience and number_of_interviews columns were imputed with mode and median respectively.
Since most of the ML models accept only numeric features, we need to convert the categorical variables to numbers such that the model is able to understand and extract valuable information.

- Use ordinal encoder for the ordinal columns highest_level_of_education, professional_experience and length_of_job_search.
- Use binary encoding for employment_status and gender.
- Use dummy encoding for the remaining categorical features.

Feature Selection

We can use the chi-squared test to find the most important features. The test returned very low values for most of the features, so it was hard to say which features were most important. So, let's include all the features in our model.

Model Training, Selection and Tuning

I used 4 algorithms to build my classifier with default parameters. Used training set (70%) and test set (30%).

- Logistic Regression – Achieved an accuracy of 65.62 %
- Support Vector Machines - Achieved an accuracy of 65.62 %
- Random Forest - Achieved an accuracy of 62.95 %
- Gradient Boosting - Achieved an accuracy of 62.71 %

After comparing the above 4 models we can see that Logistic Regression and Support Vector Machines performed the best. Let's try to tune the SVM model. Changed the kernel to 'linear' and achieved a slight improvement in accuracy of 67.07 %.

Results

The Support Vector Machine classifier turned out to be the best model with an accuracy of 67.07%.

Conclusion

There are many factors which decide whether a fellow would get placed at a company or not. A few include how much effort a student puts into the program, how consistent is the student, how much interest does a student show after enrolling into the program, personal life etc. These factors are very uncertain and can never be predicted accurately. Our model is not trained on any of these factors and that could be a possible reason it's not able to achieve a high classification accuracy.