Employment Data, Exploratory Data Analysis and Modeling on Fellow Placement

Mohamad A. Quteifan, Data Scientist, M.S. in Data Science

Bellevue University

Author Note

Mohamad Quteifan, Department of Science and Technology, Bellevue University

Abstract

This paper provides a comprehensive introduction to Exploratory Data Analysis, Classification, and Regression Modeling, focusing on insights around if a fellow will ultimately be placed at a company and the length it will take a fellow to be placed. The research for the study has been done within interactive Jupyter Notebook, an integrated development environment (IDE) for Python programming language. The goal is to create two highly effective models that will predict when and if a fellow will be placed through the examination of the relationships between a collection of variables. Throughout the research, I discovered a few significant insights. One with the greatest significance is the lack of impact that education had on placement. Before conducting the research, I assumed there would be a positive correlation between education and placement, but the assumption turned out to be false. Another significant discovery was the lack of correlation between the number of applications and placement. This means that the high number of applications does not increase your chances of employment. The other insights retrieved from the research are provided later in the paper in the conclusion. To take full advantage of the data imputing was incorporated into the study on *days_program* feature. Removing the empty values would reduce the data we have and lead to a misrepresented model. Linear regression was utilized to impute the missing values for *the days_program* feature. The median imputing technique was employed to impute for the feature *number_of_*interviews. There are two types of models implemented in this research, the first is a classification model (using logistic regression), and the other is a regression model. The purpose of the first model is to classify IF a fellow will be placed and the purpose of the second is to determine WHEN a fellow will be placed. The implementation of a logistic regression model for the classification of IF a fellow would be placed yielded an AUC score of .74. In the

research, I utilized an XGBoost model to attempt to enhance the logistic regression model. Unfortunately, the XGBoost model produced an AUC score of .70, and the original model provided more effective classifications. Although an AUC score of .74 is not great, it is still deemed acceptable. The F1 score of the logistic regression model is .70, which is also acceptable. The implementation of linear regression was utilized to determine WHEN a fellow would be placed. The linear regression model yielded poor results, unlike the logistic regression model. An MSE of 2.6 means squared error is the best result of the model, meaning the predictions and actual results were 2.6 (log) apart. An XGBoost model was implemented as well but yielded worse results (2.7). Although the regression model was deemed ineffective in making perditions, it still provided essential insights into the data.

Keywords: *Exploratory Data Analysis, Data Cleaning, Predictive Modeling, Regression, Correlation, Model*

Employment Data, Exploratory Data Analysis and Modeling on Fellow Placement

Mohamad A. Quteifan, Data Scientist, M.S. in Data Science

Exploratory Data Analysis

Exploratory data analysis is the first step in the study, and the approach is utilized to analyze the data. The focus for the Exploratory Data Analysis(EDA) is going to be on the insights of placements and the relationship between variables. The CSV file was imported into the IDE, converted into a data frame, and stored as, *df.* There is a total of 16 features within the data frame, *id, pathrise_status, primary_track, cohort_tag, program_duration_days, placed, employment_status , highest_level_of_education, length_of_job_search, biggest_challenge_in_search, professional_experience, work_authorization_status, number_of_interviews, number_of_applications, gender, and race.* Out of the 16 features, 5 are quantitative, and the rest are categorical.

The following is brief description of each variable:

1. *id*: The identification of the individual, each id represents an individual.
    - There are 2543 different values.

2. *pathrise_status*: The feature represents the pathrise status of the individual.
    - Categorical Feature
    - Independent variable
    - There are 9 different values.

3. *primary_track*: The feature represents the individuals career path/track.
    - It is a Categorical feature (values consists of textual data)
    - Independent variable
    - There are 6 different values.

4. *cohort_tag*: The feature represents the class(cohort) the individual joined.
    - It is a datetime variable, consists of text and numeric values.
    - Independent variable
    - The variable contains 47 different values.

5. *program_duration_days*: The feature is a dependent variable for the regression model. The length that the individual was part of the program.
    - Quantitative data.
    - Dependent variable.
    - Column name changed from proram_duration_days to days_program.
    - There is a total of 411 different values.

6. *placed*: The feature represents if the individual was successfully placed in a position.
    - binary variable
    - Dependent variable
    - There are 2 different values, 0 = not placed 1 = placed.

7. *employment_status*: The feature represents the employment status of the individual prior to joining pathrise.
    - Categorical Feature
    - Independent variable
    - There are 5 different values.
      **The column name contains a space and that needs to be corrected in the following steps.

8. *highest_level_of_education*: The feature represents the education of the individual prior to joining pathrise.
    - Categorical feature.
    - Independent variable.
    - Column name converted from highest_level_of_education to education.
    - There are 7 different values.

9. *length_of_job_search*: The feature represents the length of the job search of the individual.
    - Categorical feature.
    - Will be converted to quantitative data.
    - Independent variable.
    - There are 5 different values.

10. *biggest_challenge_in_search*: The feature represents the biggest challenge the individual had prior to joining pathrise.
    - Categorical feature.
    - Independent variable.
    - There are 10 different values.

11. *professional_experience*: The feature represents the professional experience of the individual prior to joining pathrise.
    - Categorical feature.
    - Will be converted to a quantitative feature.
    - Independent variable.

- There are 4 different values.

12. *work_authorization_status*: The feature represents the work authorization status(citizenship/authorization).
    - Categorical feature.
    - Independent variable.
    - There are 9 different values.

13. *number_of_interviews*: The feature represents the number of interviews an individual had prior to joining pathrise.
    - Quantitative feature.
    - Independent variable.
    - There are 21 values.

14. *number_of_applications*: The feature represents the number of applications that the individual filled prior to joining pathrise.
    - Quantitative feature.
    - Independent variable.
    - There are 41 different values.

15. *gender*: The feature represents the gender(sex) of the individual.
    - Binary feature.
    - Independent variable.
    - There are 4 different values.

16. *race*: The feature represents the race of the individual.
    - Binary feature.
    - Independent variable.
    - There are 9 different values.

Analysis – Dependent Variables

The first portion of the study, as I mentioned earlier, is focused solely on the analysis of the data and concludes Pathrise fellow placement and the length it took for prior Pathrise fellows to be placed. I breakdown the study into two different portions, placement, and duration until placement. Reviewing the problem statement, we determine two models that can provide more insights into the data. One model is a regression model, and another being a classification

problem. Based on the analysis conducted, the two dependent variables are *placed* (classification) and *days_program* (regression).

Placement – Dependent Variable

Looking through the data, we can conclude that the analysis will not represent the data entirely effectively; mainly due to the data containing individuals who dropped out of the program or failed to complete the program. In figure 1, it is evident that there were more individuals that were not placed. This information may be misleading since many individuals withdraw from the program or did not participate in the program. Furthermore, 37.6% of the individuals were placed and that is based on the complete dataset, including individuals who did not enroll in the program (withdrawn). In figure 2, we can see that many individuals included in the data were/are enrolled in the program. It is important to note that not all were enrolled. Figure 1 and figure 2 present similar distribution, but they are not to be mistaken as anything more than visuals of distribution.

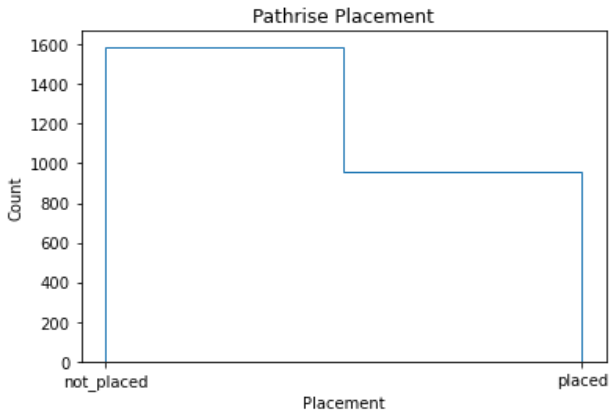

FIG. 1. Histogram of placement distribution.

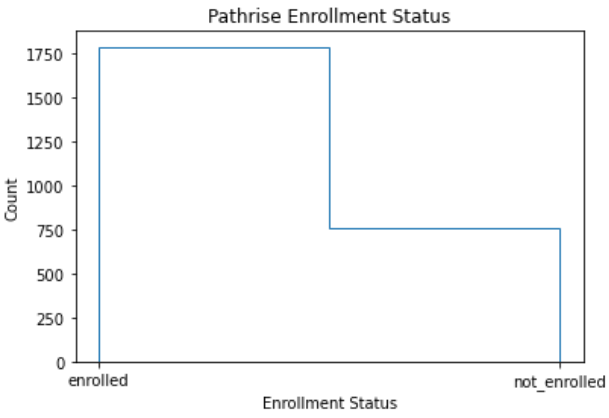

FIG. 2. Histogram of Enrollment Status distribution.

Education and Placement

Throughout the research, we discovered a weak correlation between education and placement. During the first portion of the analysis, we dove into the research without removing

empty values or filtering out the individuals who did not actively participate in the program. This was mainly done to compare the results. Out of the 956 fellows that were placed, ~55% were fellows who had Bachelor's Degree, ~30% received their Masters, ~6% received their Doctorate or Professional Degree, ~5% had some college experience but did not receive a degree, ~.4% were fellows without a GED or a high school diploma, ~1.5% were fellows who received either their high school diploma or a GED certificate. We can conclude from these statistics is that the higher the education of the fellow, the more likely you will be placed. Once cleaning the education variable and removing all empty values we determine that ~54% of the fellows that received a Bachelor's Degree got placed, ~52% of the fellows who received a Master's got placed, ~58% of fellows who received either a Doctorate Degree or Professional Degree were placed, ~55 of the fellows with some college, but no degree got placed, ~77% of the fellows with high school diploma got placed, ~50% of the fellows without a high school diploma got placed, and ~71% of the fellows who received their GED (or equivalent) got placed. The statistics presented above provide evidence that education level does not impact placement, but it does have an impact on entering the program. The majority of the fellows either have a college degree or higher, which means the individuals with lower education criteria are likely to struggle to enter the program during the application process. Another possibility is that many individuals with lower education criteria simply do not apply for the program. What can be concluded is the flawed relationship between education and placement. Education is not significant when it comes to placement, but it may have an impact on admissions into the program.

Duration of Placement – Dependent Variable

As I mentioned above, the data contains an ample amount of missing values in the data. The feature, *days_program* (changed from *program_duration_days*), contained the most missing

values. This is not surprising considering the data contains individuals who either did not continue with the program(failed/withdrew) or currently active in the program. Although I did remove the empty values to calculate the average days it takes for a fellow to be placed, I will be implementing imputing to compensate for the values. Given that the data set is small, removing the missing values would limit the research and may lead to misrepresented results. When analyzing the feature, *days_program*, I made a few discoveries. For one, it took ~161 days on average for a fellow to be placed in a company. Another discovery made was the duration that fellows were taking in the program, with one fellow taking part in the program for 548 days. Also, the majority of fellows spent 0 days in the program and that causes an issue with the distribution of values within the feature. In figure 4, we can see the data is skewed to the right, and the distribution is unimodal, meaning the feature contains a single highest value.

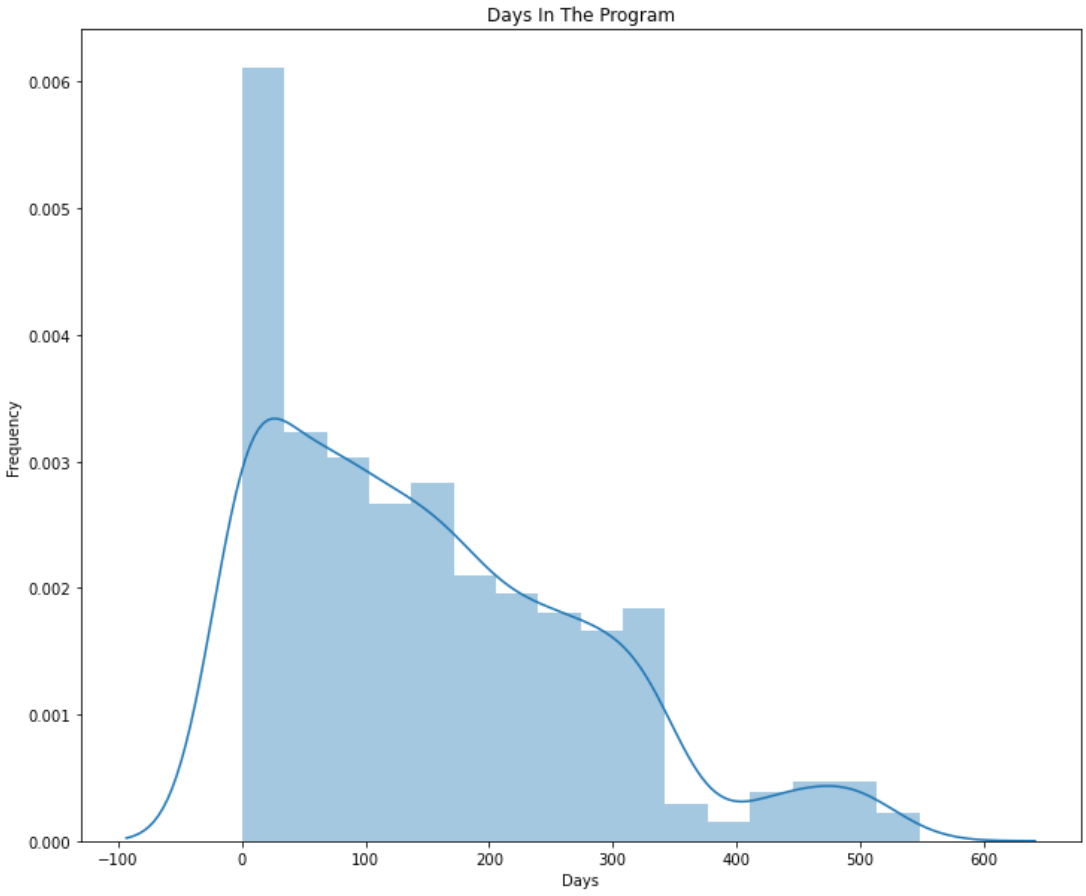

FIG. 4. Histogram of distribution of Days in the program.

Analysis – Independent Variables

We will be using the majority of the other features as independent variables in the study. The feature, *pathrise_status,* is one that is not going to be employed in the research. The feature contains only one value, enrolled, and the not_enrolled individuals have been completely removed from the data. The research is focused on fellows who are enrolled in the program, and including fellows who ended up not enrolling in the program is going to misrepresentation the data during the analysis process. The independent variables in the research are, *primary_track,* cohort_tag, *employment_status , education, length_of_job_search, professional_experience, work_authorization_status, number_of_interviews, number_of_applications, gender, and race.* All but two of the variables are categorical.

.

Independent Variables- Categorical Features

Primary Track – Independent Variable

The primary track feature describes the career path of the fellow. There are six different career paths, data science, design, marketing, PSO, SWE, and Web. In figure 5, we can clearly see that majority of the enrolled fellows are Software Engineers (SWE). Another important detail to note is that there are no missing values for the primary track feature.

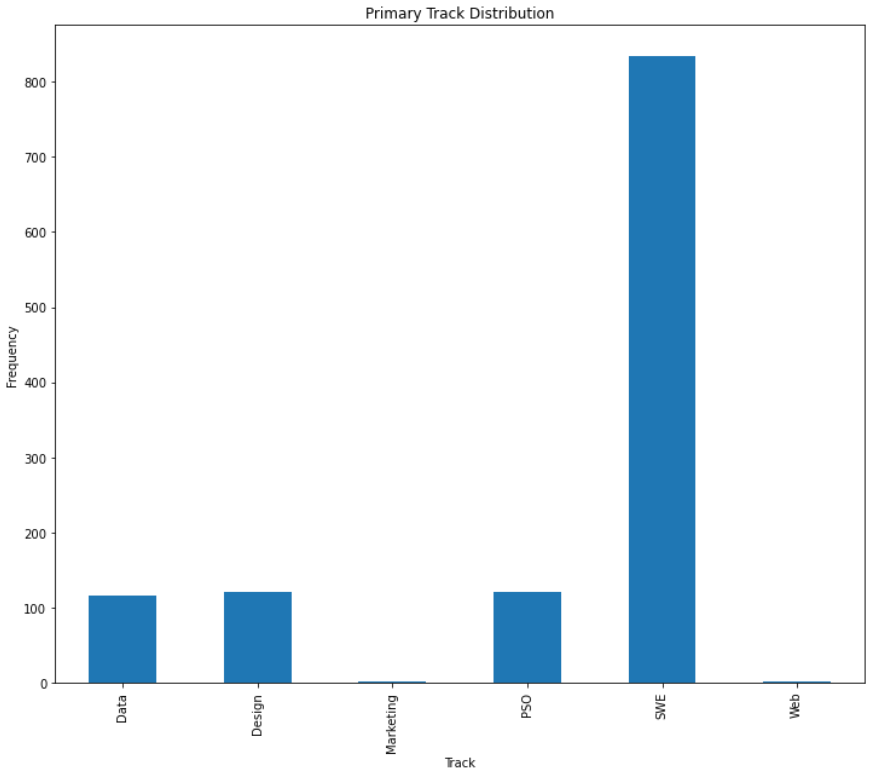

FIG. 5. Bar Chart visualizing the distribution of the fellows primary career track.

Cohort_tag – Independent Variable

The feature, *cohort_tag*, describes the class of the fellows. The feature is nearly distributed evenly, as you can see in figure 6. The visual of the distribution highlighted an issue in feature, there seems to be a cohort value with a lowercase a. Further investigation made it evident that it was a simple typo, and replacing the FEB20a with FEB20A fixed the issue. Another issue I encountered in the feature is the value representing October 2021(OCT21A). The typo is alarming since October 2021 is eight months in the future. Including OCT21A with cohort OCT19A or creating a new value (OCT20A) may lead to misrepresented data since it is not clear where it belongs. I did find one odd trend; it seems to be the later cohorts in 2020 seem to contain fewer fellows.

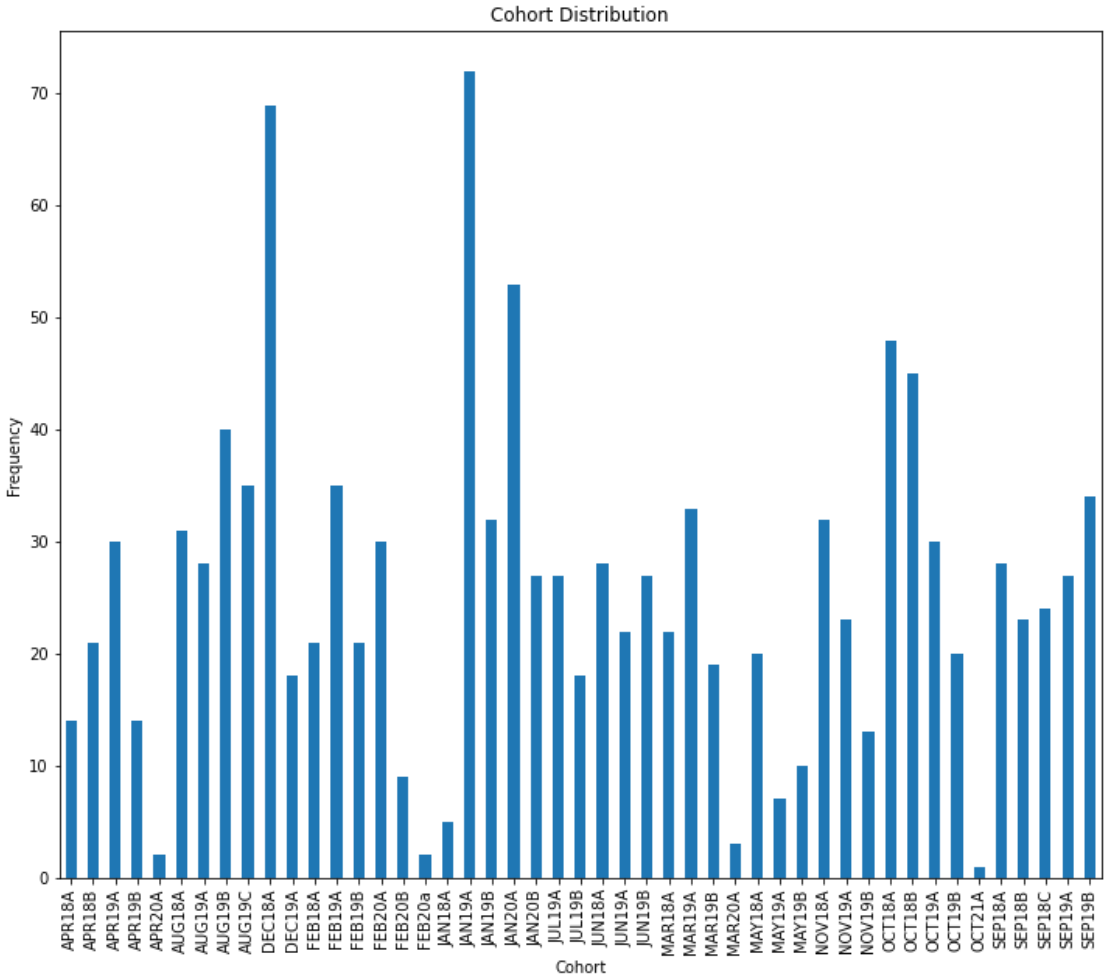

FIG. 6. Bar Chart visualizing the distribution of the fellows Cohort Tag.

Employment_status – Independent Variable

The feature, *employment_status*, describes the fellow's employment status before joining Pathrise. We can see the majority of the fellows are either students or unemployed. This is not entirely unexpected because students and unemployed individuals are the ones actively looking for employment. I was surprised to see that the third most employment status among fellows is full-time employees. I did not expect that and I assumed either part-time or contractors would be seeking to join Pathrise at a higher rate due to their employment status. The independent does not provide many insights.

.

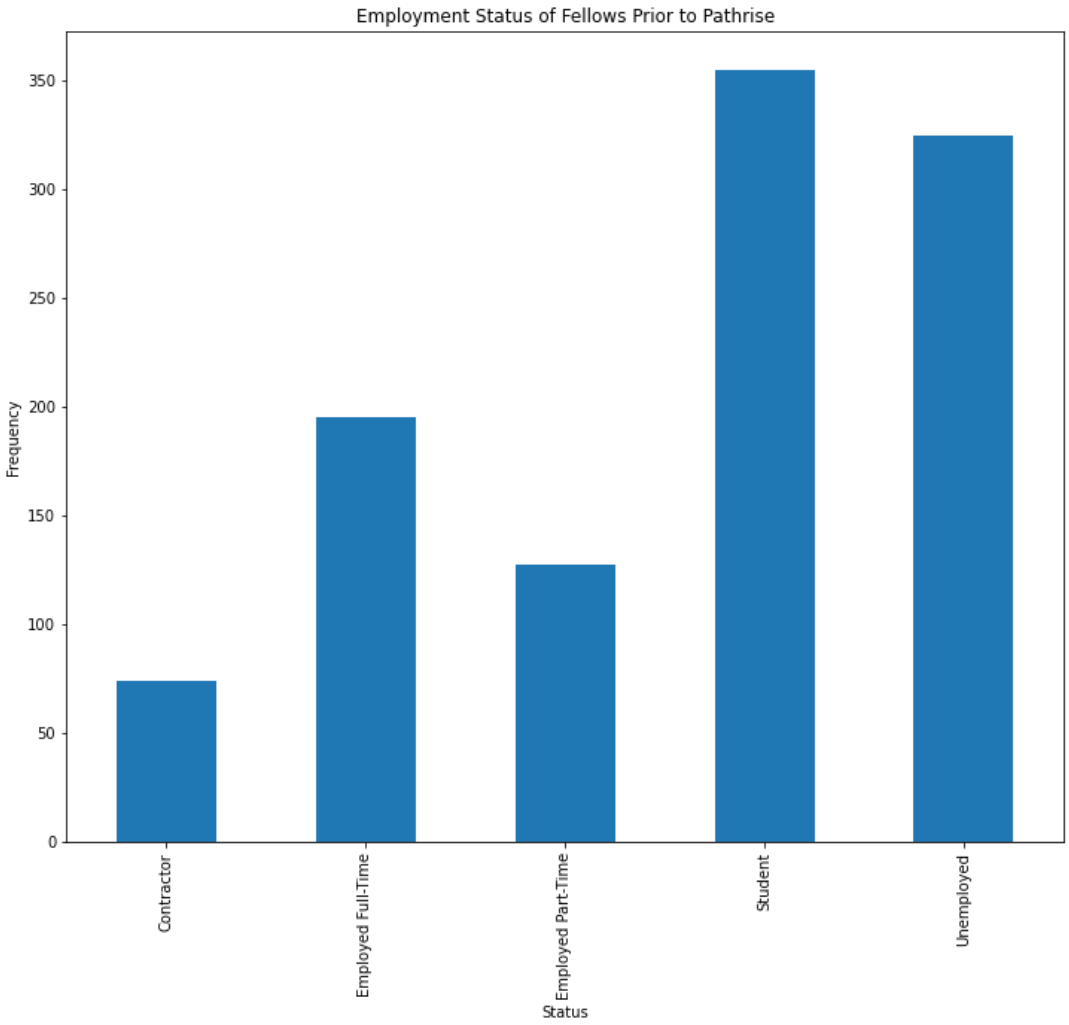

FIG. 7. Bar Chart visualizing the distribution of the fellows Employment status.

Education – Independent Variable

The feature, *education*, describes the highest level of education a fellow has received. We can see in figure 8 that the majority of the fellows in the program received either a Bachelor's Degree or a Master's Degree. In our earlier analysis of placement and education, we determine that education level did not affect placement. The data, however, suggests that fellows with higher education are more likely to gain admissions to the program. Before the analysis, I assumed the higher the education level of a fellow the more likely a fellow will be placed, but the data makes it evident that there is no strong relationship between placement and education.

.

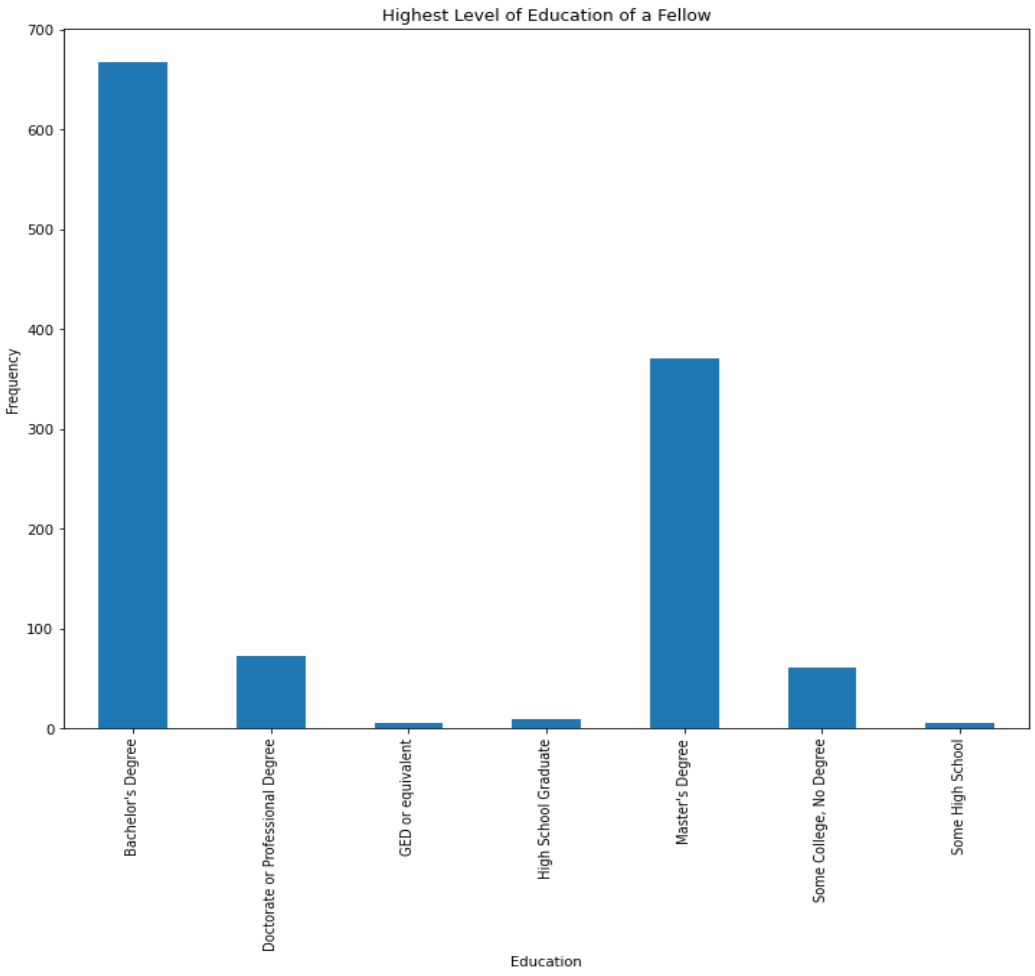

FIG. 8. Bar Chart visualizing the distribution of the fellows Highest Level of Education.

Length_of_job_search – Independent Variable

The feature, *length_of_job_search*, represents the length of a fellow until placement in a company. In figure 9, we can see the majority of the fellows are placed within two months, some even placing less than a month of joining Pathrise. The rest either day 3-5 months to a year with a few fellows taking more than a year to be placed.

.

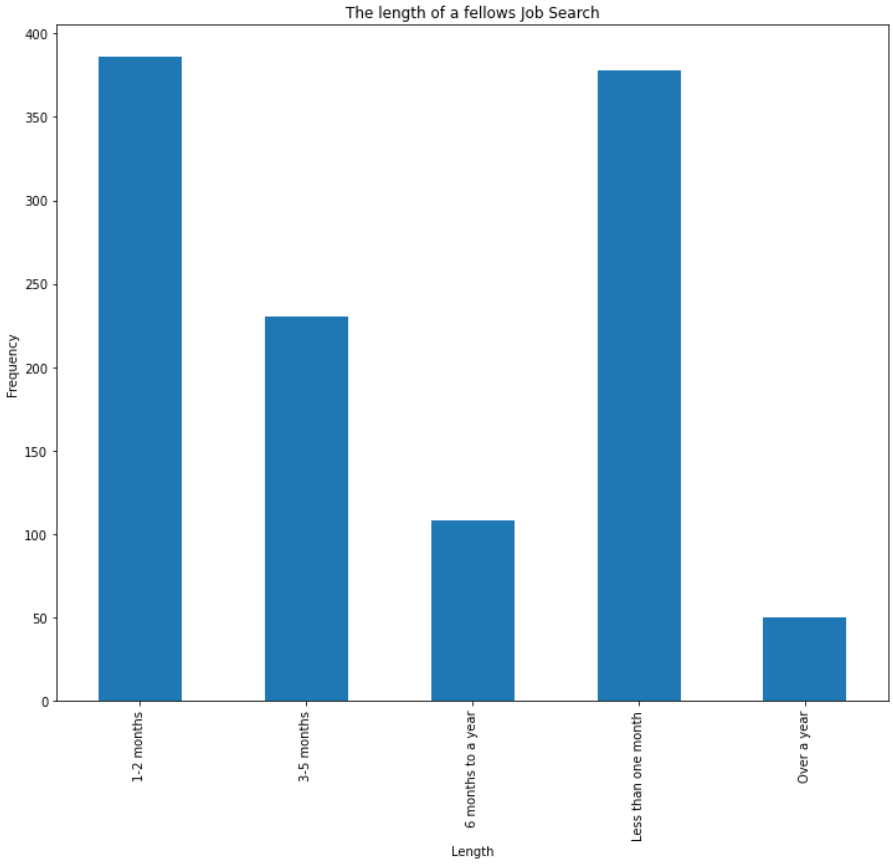

FIG. 9. Bar Chart visualizing the distribution of the fellows job search length.

Work_authorization_status – Independent Variable

The feature, *work_authorization_status*, represents the work authorization of the fellows. In figure 10, we can see the clear majority of the fellows are authorized to work in the United States. Work authorization can be a factor when applying to positions and can be an obstacle for fellows looking to be placed with a company. Many companies hesitate to sponsor an applicant, and they prefer an applicant who has the authorization to work in the United States.

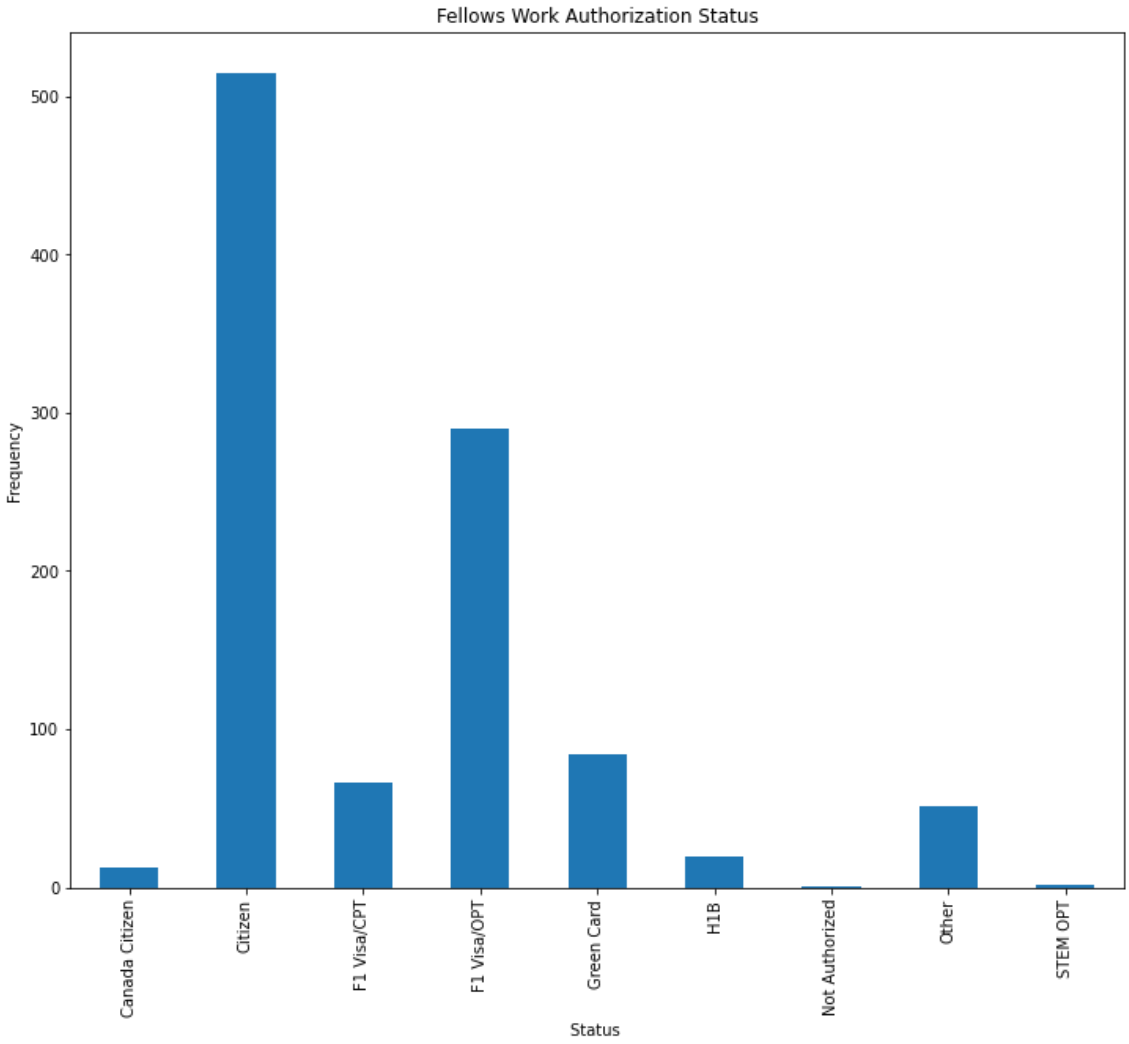

FIG. 10. Bar Chart visualizing the distribution of the work authorization status of Pathrise fellows.

Professional_experience – Independent Variable—will be converted to quatitiative variable

The feature, *professional_experience,* represents the work experience of the fellow before joining Pathrise. In figure 11, we can see the majority of the fellows have less than two years of professional work experience. This makes sense because many of the fellows are students or unemployed. Further analysis is required for insights into the fellow's work experience.

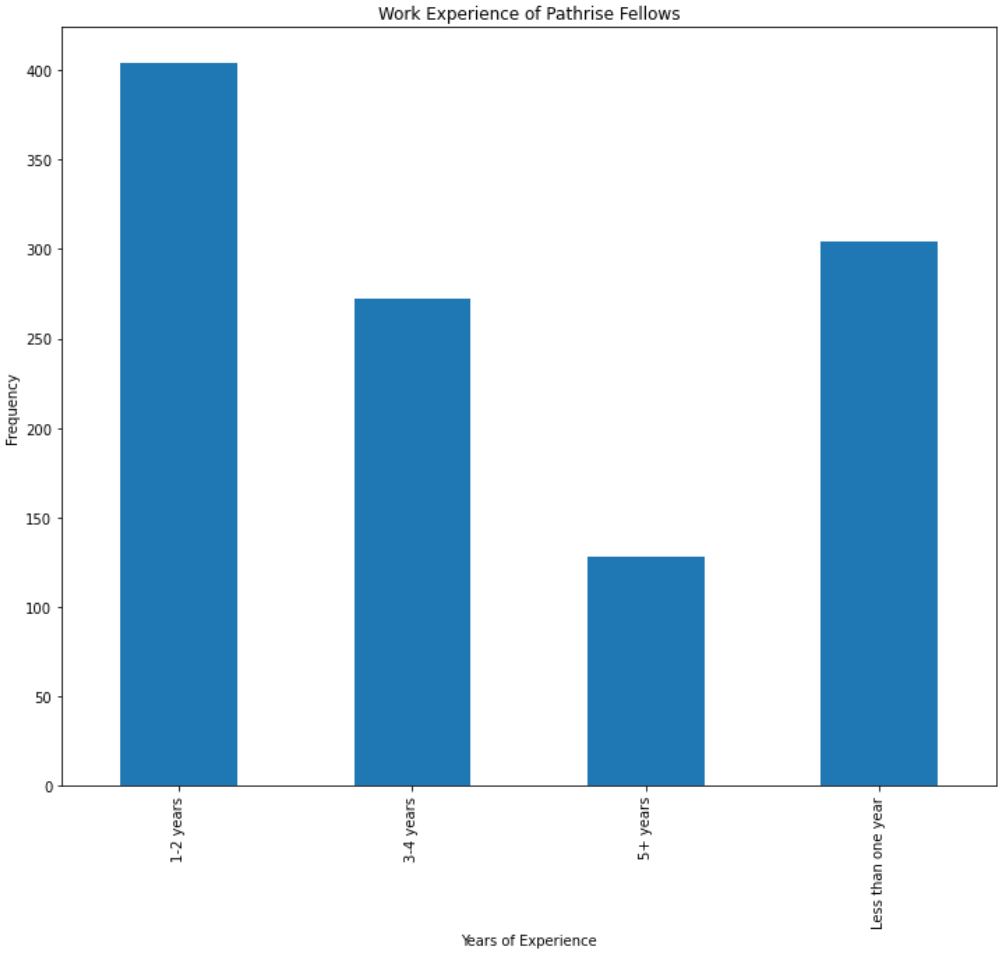

FIG. 11. Bar Chart visualizing the distribution of fellows professional work experience.

Gender – Independent Variable

The feature, *gender*, represent the gender of the fellow. The feature contains an ample amount of missing values. It makes sense, many people are uncomfortable disclosing their gender. In figure 12, we can see that majority of the fellows were males. The technology industry is dominated by males, and seeing that most of the fellows are software engineers it comes as no surprise that the majority of the fellows were males.

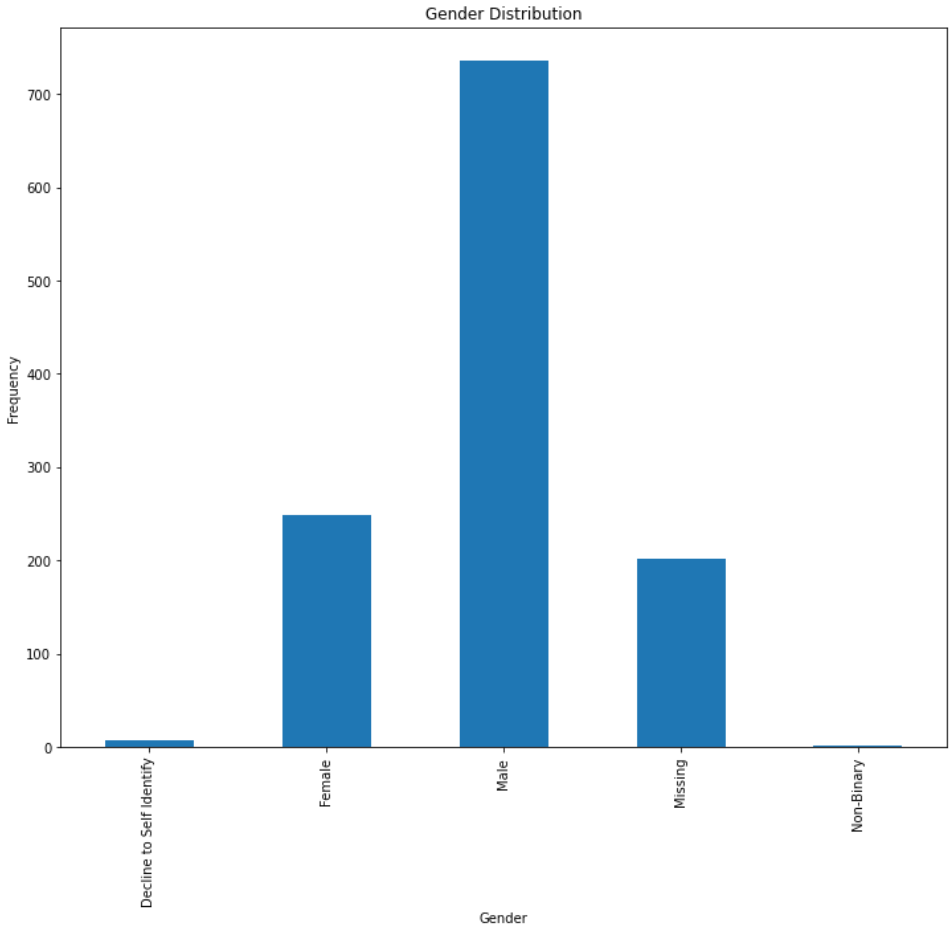

FIG. 12. Bar Chart visualizing the distribution of fellows gender.

Race – Independent Variable

The feature, *race*, represents the race of the fellow. To my surprise, the feature did not contain as many missing values as the *gender* feature. I expected fellows to be equally uncomfortable disclose that information. In figure 13, we can see that majority of the fellows are East Asian/Asian American, non-Hispanic white/euro-American, or south Asian/Indian-American.

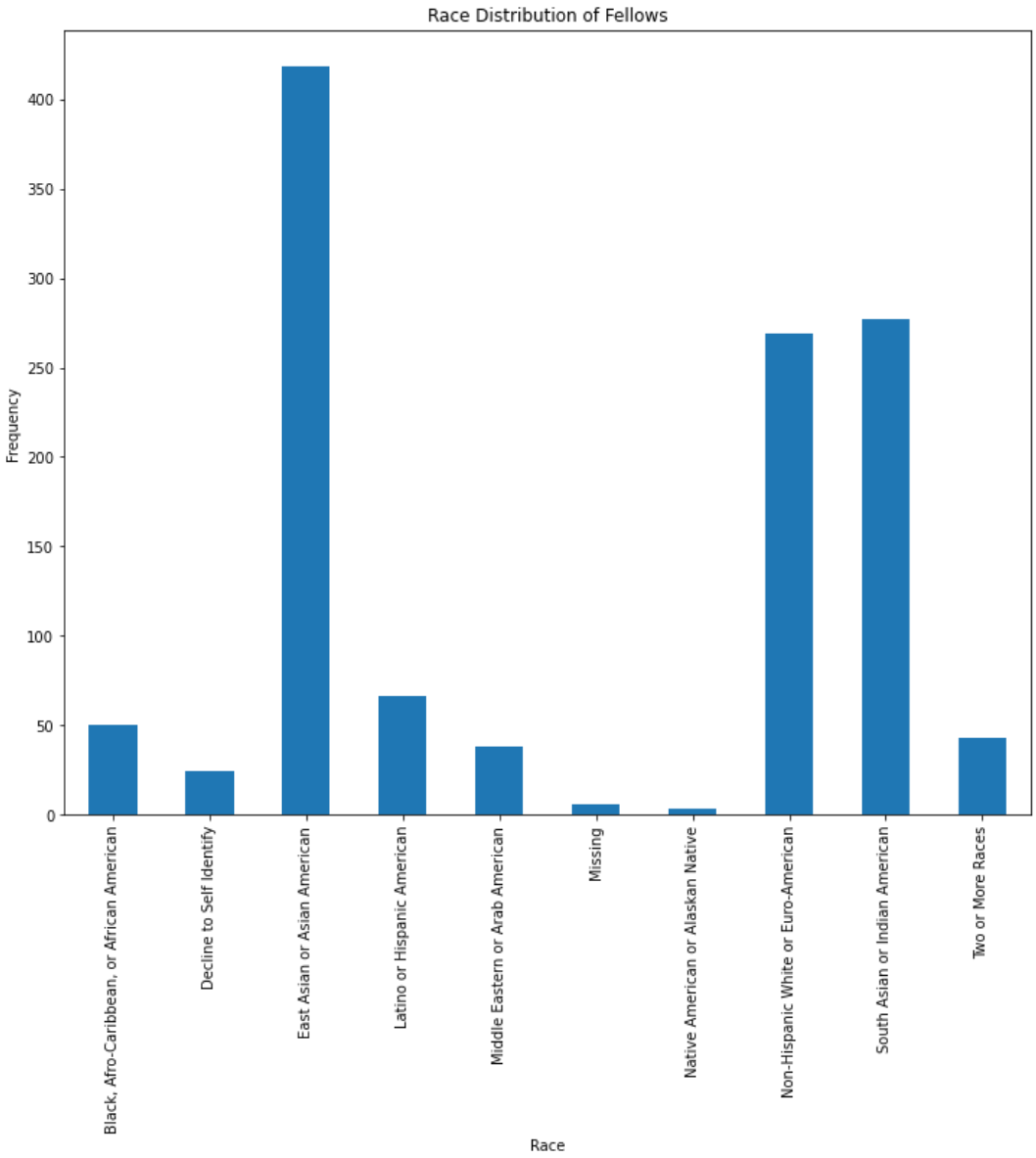

FIG. 13. Bar Chart visualizing the distribution of fellows race.

Independent Variables- Quantitative Features

Number_of_Interviews– Independent Variable

The feature, *number_of_interviews*, represents the number of interviews a fellow has been in before joining Pathrise. In figure 14, we can that majority of the fellows have not had any interviews. This should not surprise anyone because most of the fellows are students and unemployed individuals.

.

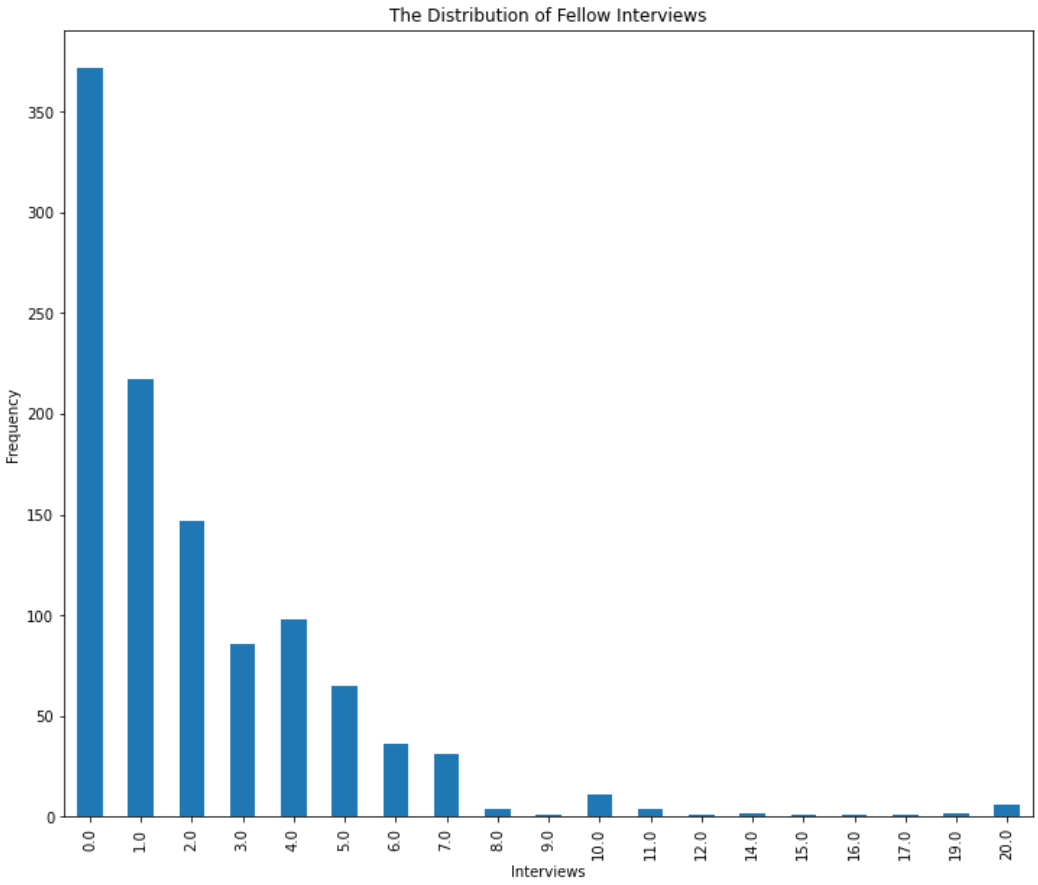

FIG. 14. Bar Chart visualizing the distribution of fellow interviews.

Number_of_applications– Independent Variable

The feature, *number_of_applications,* represents the number of applications a fellow has submitted before joining Pathrise. In figure 15, we can that many fellows submitted 15 applications.

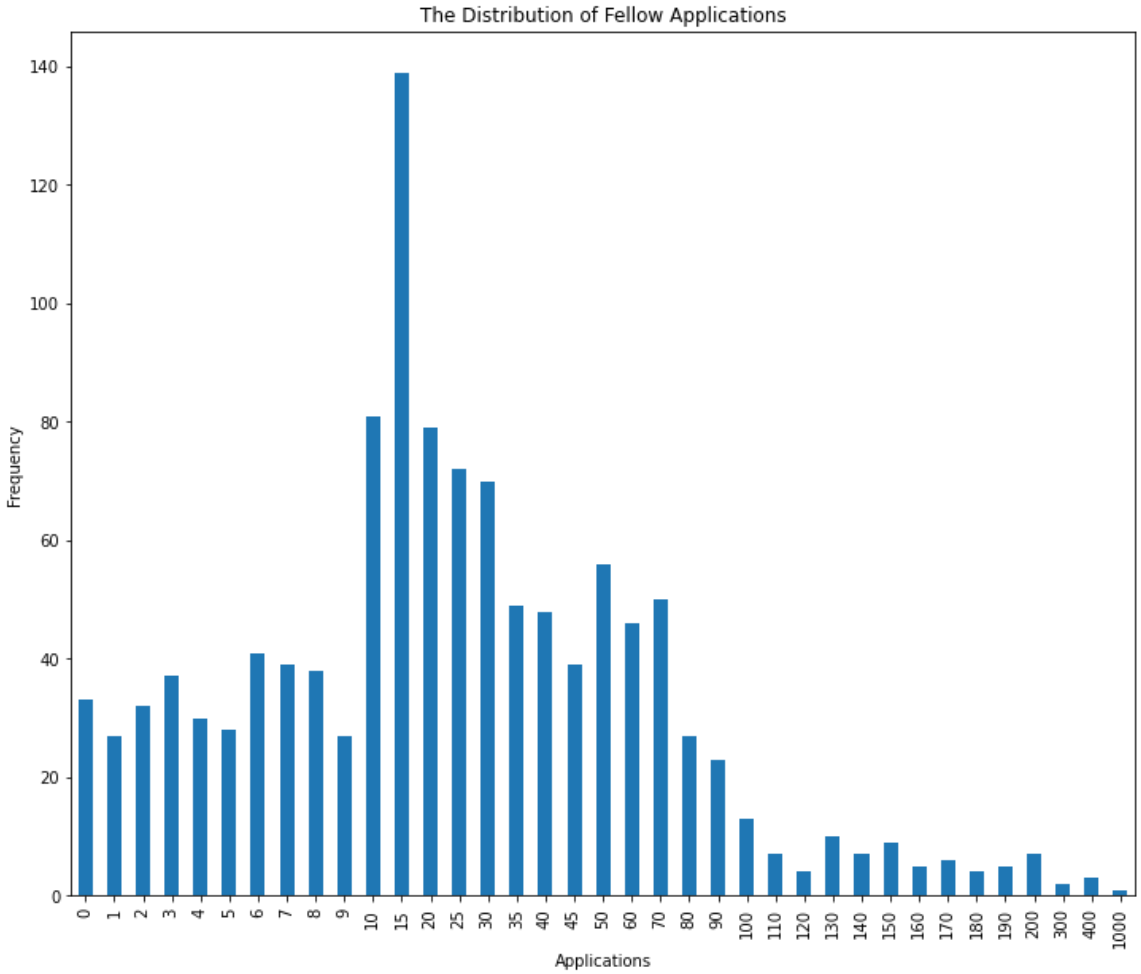

FIG. 15. Bar Chart visualizing the distribution of fellow applications.

Modeling

Before the Exploratory Data Analysis that was conducted earlier in the research, we had an idea of the types of models that we are implementing to provide insightful predictions on when and if a fellow will be placed. The two models are a regression model and a classification model. The regression model is going to be utilized to determine approximately when a fellow will be placed. The classification model, however, will be implemented to determine if a fellow will even be placed based on their credentials (independent variables).

Chi-Square Test

A Chi-Square test is a method that was implemented in the research to find relationships between the categorical data and the dependent variables. The Chi-Square test assumes the frequency of the categorical variable matches the expected frequencies for that particle categorical variable. That is known as the null hypothesis. If we run a Chi-Square test and the p-value is less than the significance level, which I set to 0.05 then we reject the null hypothesis. This means the categorical variable has a statistical significance in determining the dependent variable, which in our case is *placed* and *days_program.* It is simple if the p-value is less than .05, the categorical variable there is a relationship between the variables. If the p-value is greater than .05 then we retain the null, and the categorical variable does not interact with the dependent variable (no significant relationship between the two variables).

Classification Model – significant interaction with placement

The Chi-square test provided the classification with four significant interactions, *primary_track, cohort_tag, gender, race, and days_program.* The expectation was the features, *number_of_applications,* and *number_of_interviews* would play a bigger impact on the model. The lack of significance led to the exclusion of the two variables in the classification model.

Regression Model – significant interaction with placement time

The regression model only had three features that present significance to the *days_program.* The three features: cohort_tag, work_authorization_status, and gender. The important thing to note is that there are no quantitative features for the regression model. All the independent variables are categorical.

Conclusion

Classification model

The Classification model yielded the best results. The logistic regression with an AUC of .74 **(XGBoost had an AUC .70)**. Although it is not perfect, it is still deemed as an effective placement classifier. There are still a few things that can be done to further enhance the model, such as implementing a few feature engineering techniques. the insights of the model are provided below.

* The features with significance for classification model:

1. primary_track
2. cohort_tag
3. gender
4. race

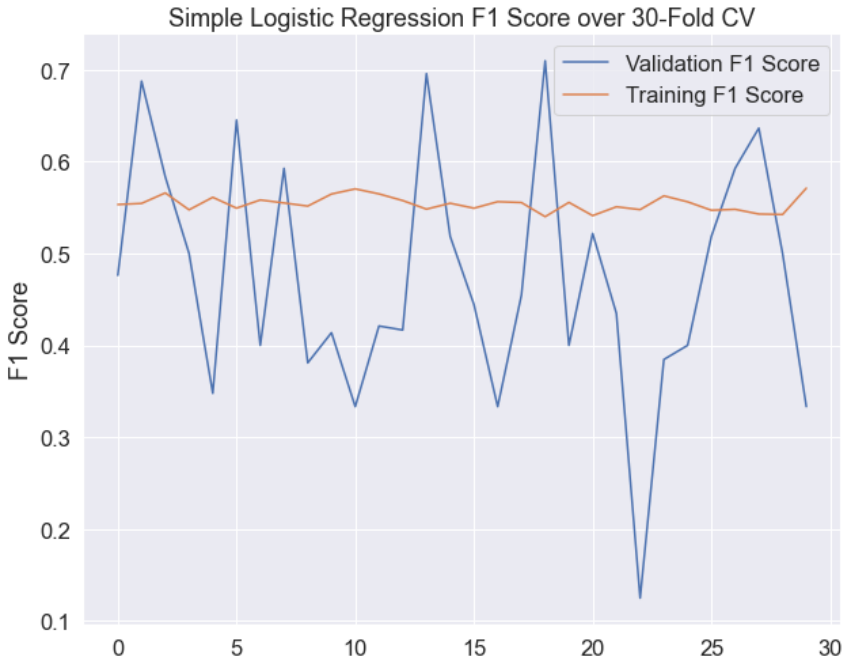

FIG. 16 Classification Model, highest F1 score is .70

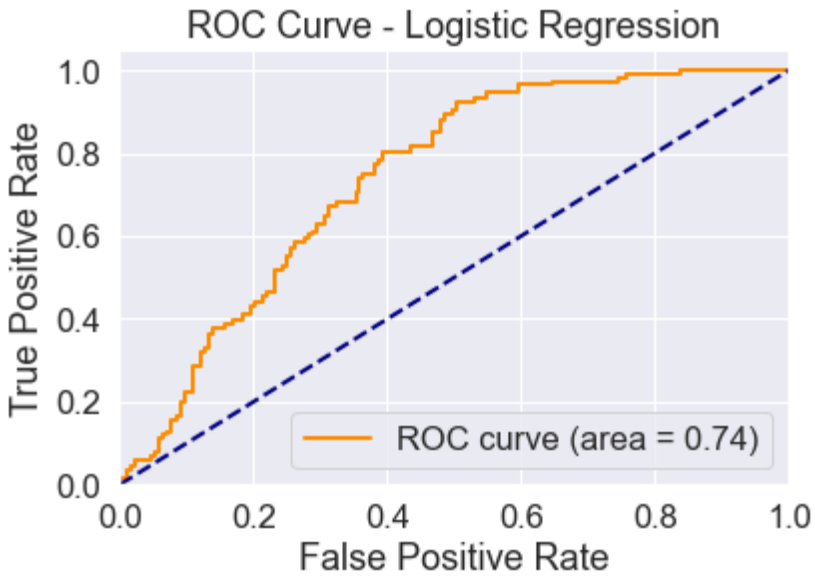

FIG. 17 Classification Model, ROC Curve indicating a AUC of .74

Regression Model

The Linear Regression model was not a great one, yielding an MSE of 2.6 and 2.7(XGBoost model). Considering the value is in the log, the predictions are hugely inaccurate. Although the creation of the model did not provide effective predictions on placement time, it still gave us beneficial insights that could be beneficial for future advancements.

The features with significance for Linear Regression Model:

1. cohort_tag
2. work_authorization_status
3. gender

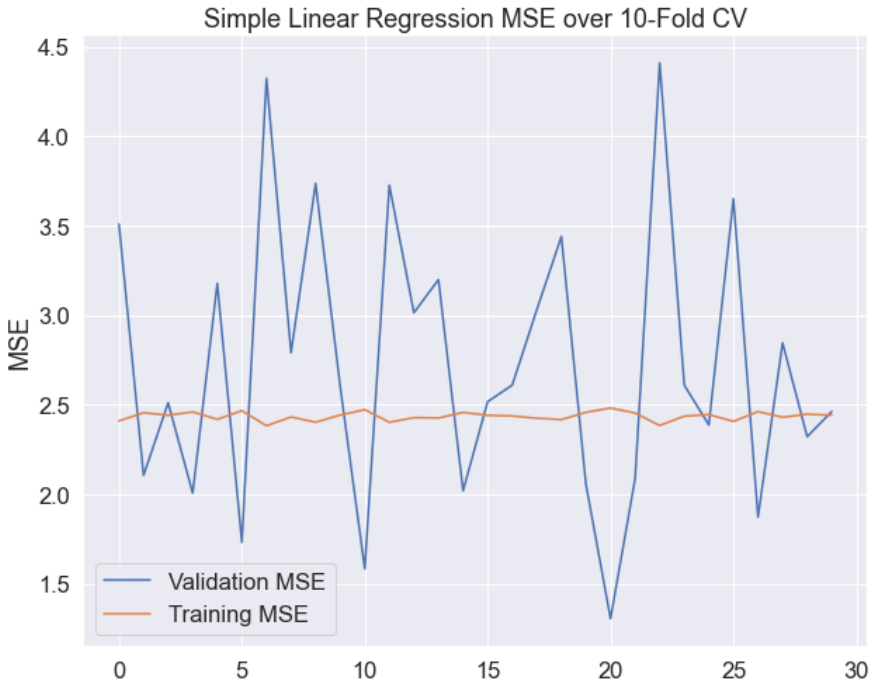

FIG. 18 Regression Model, best MSE value 2.6 log_days.

Key insights of the research

1. Education did not have a big impact on placement. It did not matter what type of education you had to get placed, getting a less than an HS diploma might've had an impact on being able to enter to program but did not have an impact on placement. The model provided even more evidence that the education did not impact placement, education was not in the top 10 significant features in the model. This allows us to conclude that:
2. Higher education did not mean higher placement rate!

2. Many of the fellows had a less than 2 years of experience (many were college students)

3. Majority of the fellows were legally allowed to work in the United States.

4. ON average each fellow sent out an average of 20 applications.

5. Many of the fellows who applied to Pathrise for were struggling to hear back from recruiters.

6. Unemployed and they were Male

7. The median time a fellow stays in the program is around 111 days, that is around the amount that Pathrise states.

8. Sending out many Applications is not correlated with getting placed faster and had a minimal impact on placement.

9. Later cohort contain less and less individuals. It would be interesting

References

Bangdiwala, S. I. (2018). Regression: multiple linear. *International Journal of Injury Control and Safety Promotion*, *25*(2), 232–236. https://doi.org/10.1080/17457300.2018.1452336

Pandis, N. (2016). Multiple linear regression analysis. *American Journal of Orthodontics and Dentofacial Orthopedics*, *149*(4), 581. https://doi.org/10.1016/j.ajodo.2016.01.012