Mental Health in the Technology Industry: Who goes to therapy in the industry?

Mythri Partha

Springboard Data Science Career Track

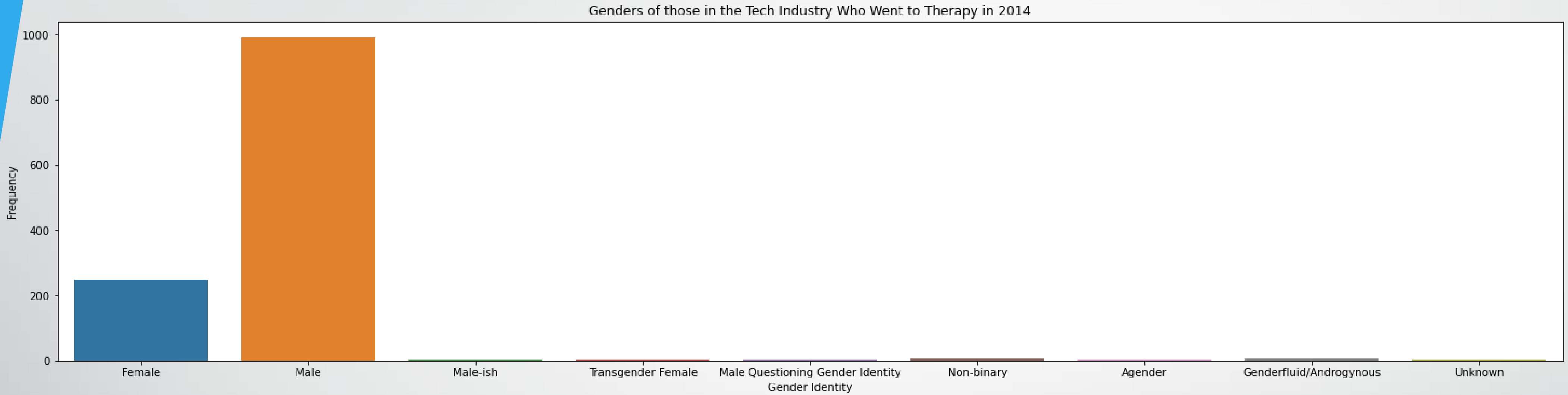
The Problem and the goal

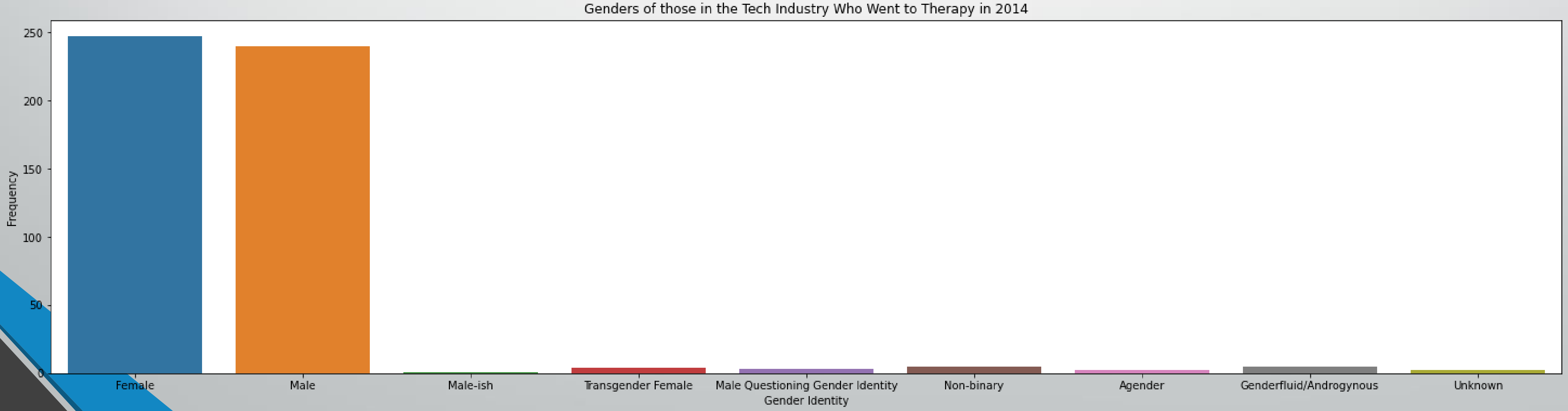
- Almost half (about 46.6%) of American adults struggle with mental illness at some point in their lives
- Only about 41% of people in the US who struggle with their mental health actually seek help
- What are the factors that lead to a person choosing to go to therapy or seek mental health help?
- GOAL: Create a model that can predict who in the technology industry seeks counseling or therapy

Data Wrangling

- Source: Open Sourcing Mental Illness, https://www.kaggle.com/osmi/mental-health-in-tech-survey
- Dataset had ages of survey participants that were in the thousands or negative, so those were replaced with NaN values then filled with the average of the ages
- The dataset contained misspelled genders and inconsistencies in labels for the genders, so those were changed to make the data more consistent
- Data had way more males than females in the survey so I resampled randomly to even out the males and females in the study.

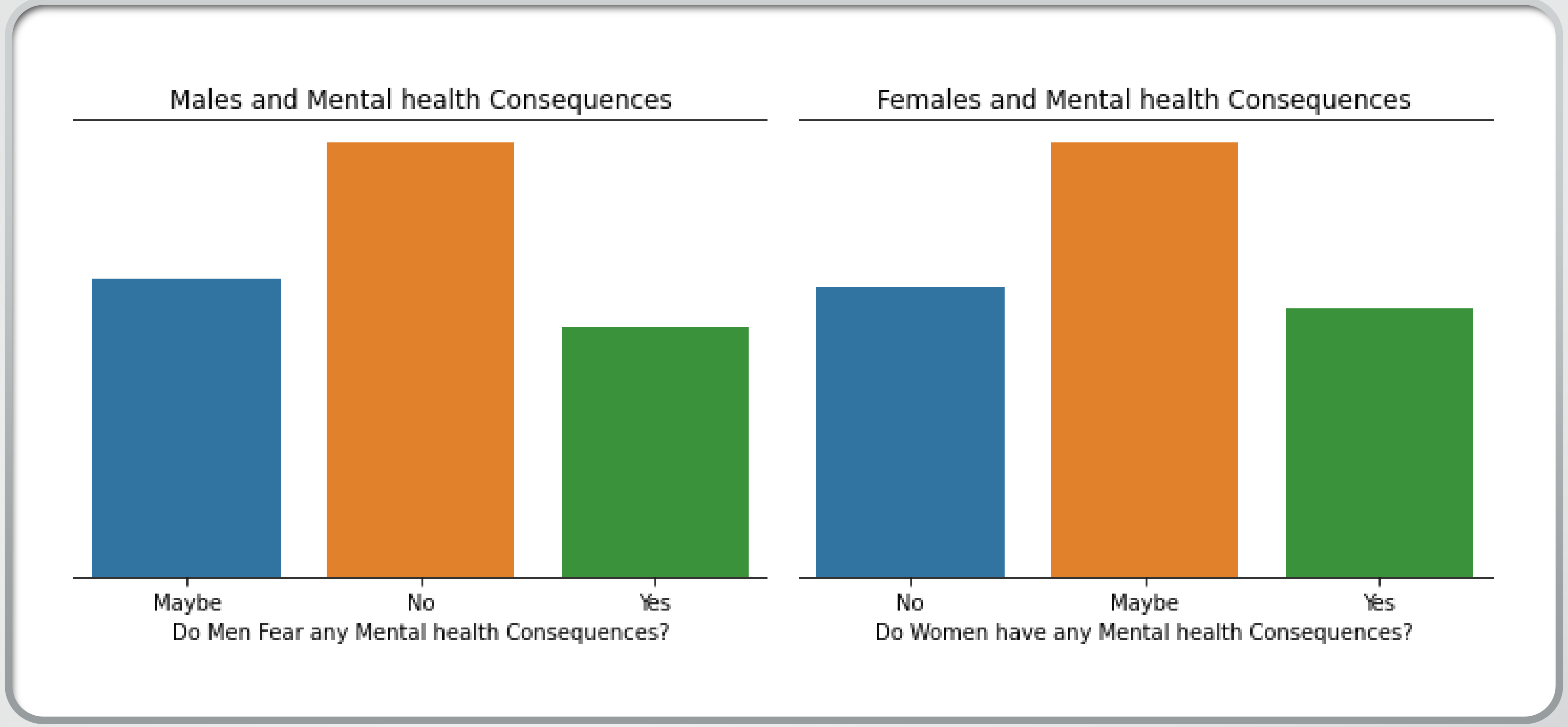
Gender Distribution (before and after resample)





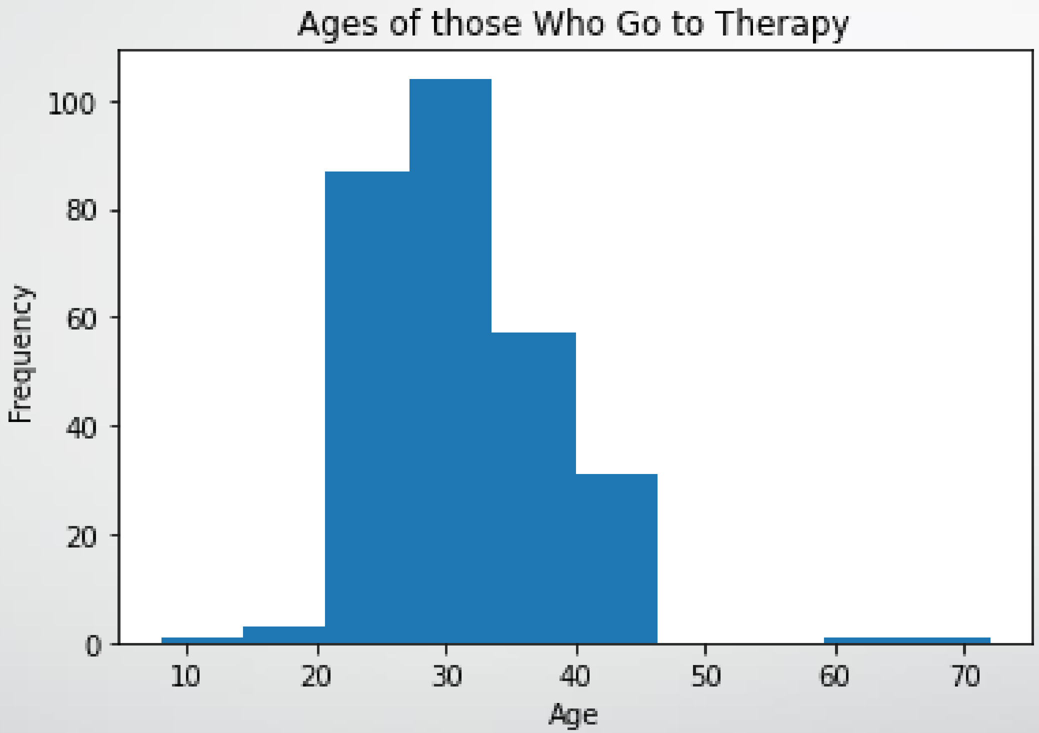
Exploratory Data Analysis

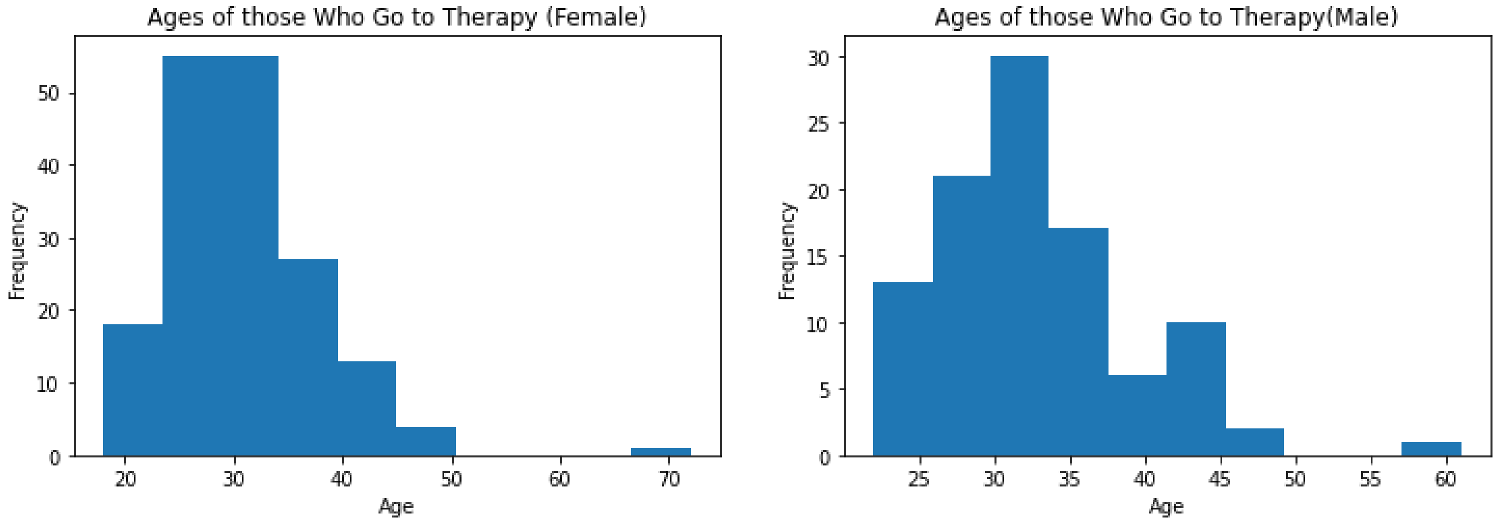
- Does gender and age affect the decision to go to therapy?
- Who fears that there are consequences to talking about mental health with superiors/coworkers more?
- Is there a normal distribution in the ages of those who go to therapy?



Who fears that there will be consequences if they speak about mental health?

Distribution of those who go to therapy





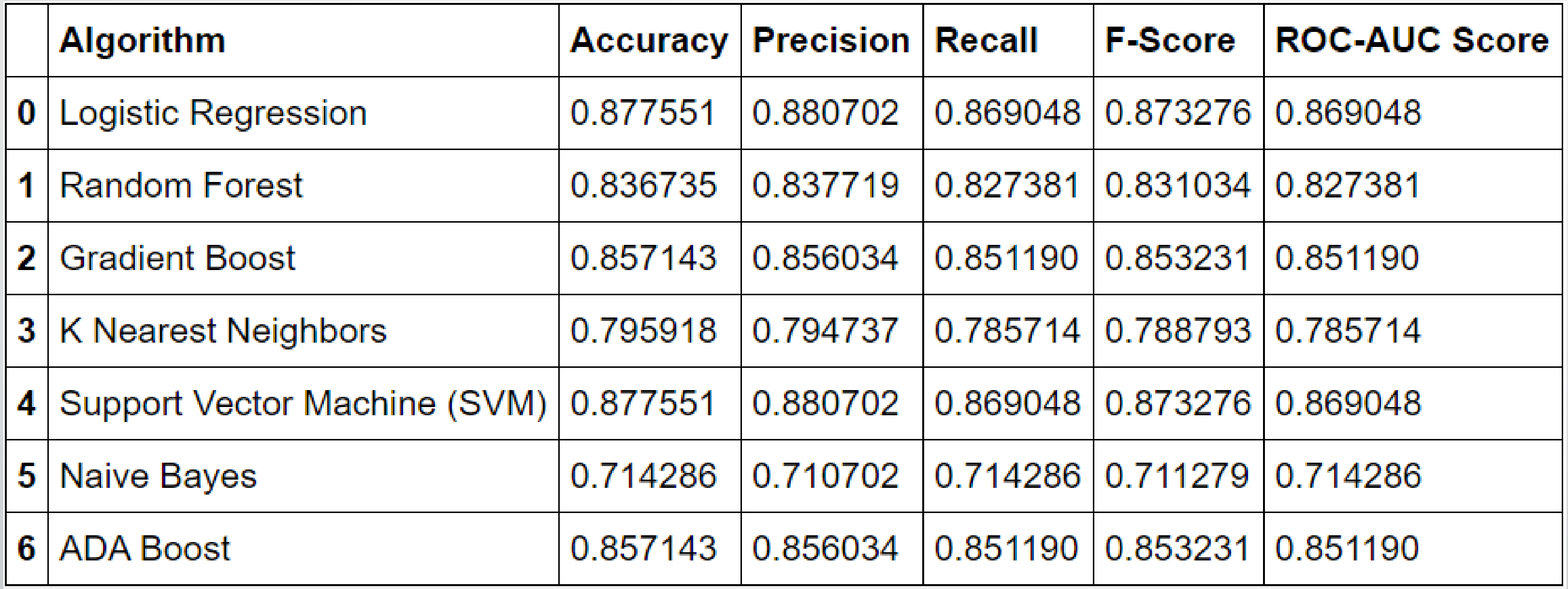
Males and Females who go to therapy by age

Modeling

Goals and Questions

- Create Machine learning models and see which one is best equipped to answer the questions asked
- Which is the most accurate and precise machine learning method?
- Which features are most important in finding out the answers to the main questions we have?

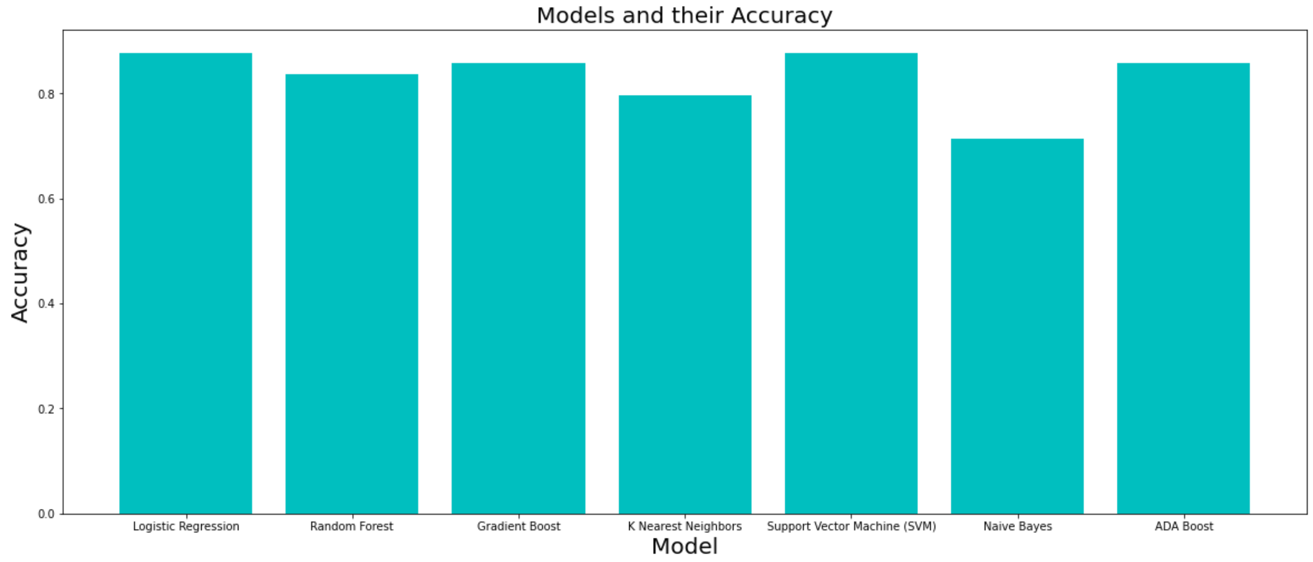
Models completed and their data



	Algorithm	Accuracy	Precision	Recall	F-Score	ROC-AUC Score
0	Logistic Regression	0.877551	0.880702	0.869048	0.873276	0.869048
1	Random Forest	0.836735	0.837719	0.827381	0.831034	0.827381
2	Gradient Boost	0.857143	0.856034	0.851190	0.853231	0.851190
3	K Nearest Neighbors	0.795918	0.794737	0.785714	0.788793	0.785714
4	Support Vector Machine (SVM)	0.877551	0.880702	0.869048	0.873276	0.869048
5	Naive Bayes	0.714286	0.710702	0.714286	0.711279	0.714286
6	ADA Boost	0.857143	0.856034	0.851190	0.853231	0.851190

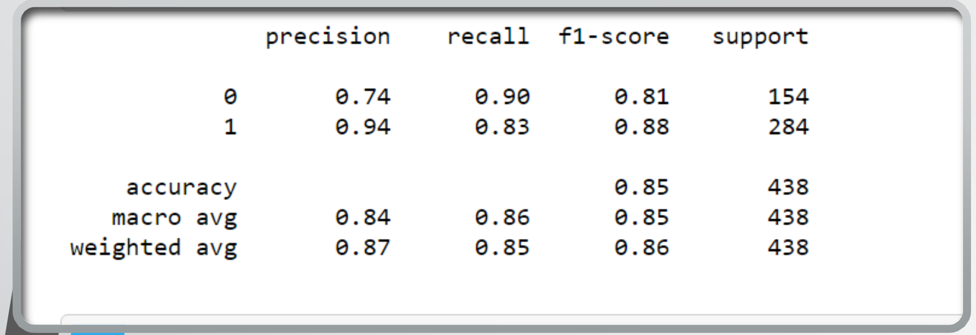
Model Accuracy Comparison

- The best models are logistic regression and support vector machine (SVM), from here, I will optimize the parameters.

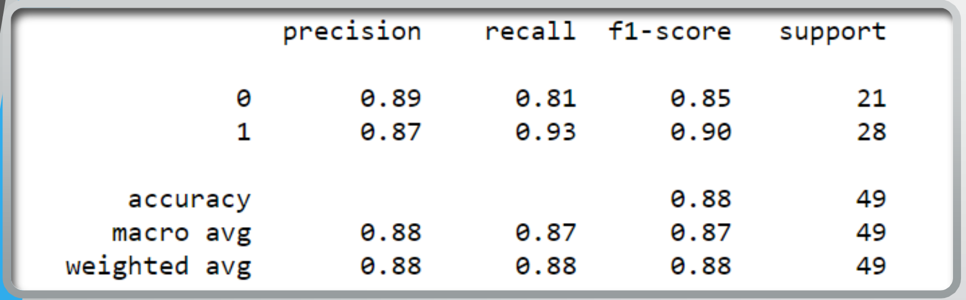


Best Models, Hyperparameter tuning

- The models with highest accuracy and precision are SVM and Logistic regression, I then completed hyperparameter tuning on both. Here are their classification reports.
- SVM accuracy stayed the same, while Logistic regression accuracy decreased after tuning

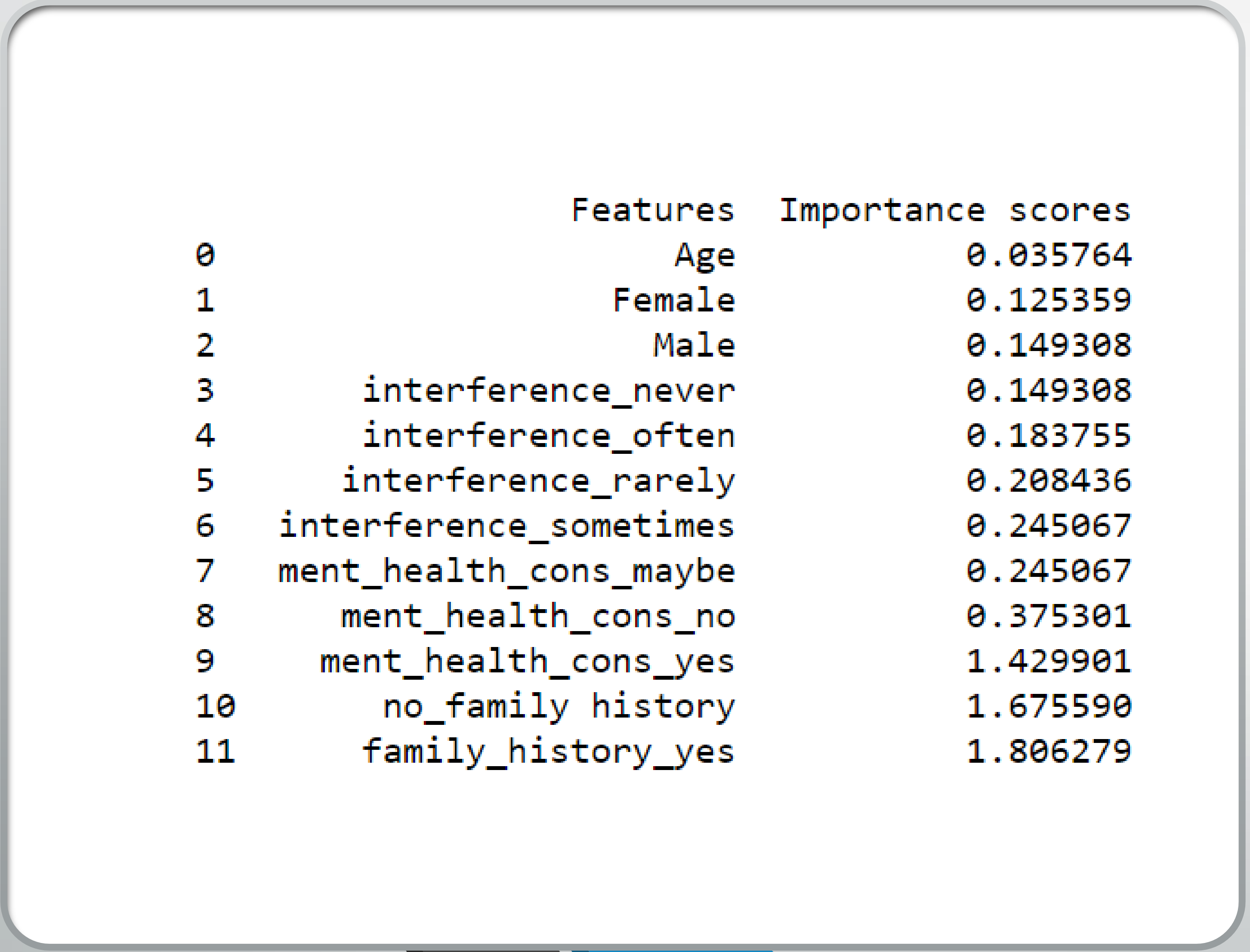


```
              precision    recall  f1-score   support

           0       0.74      0.90      0.81       154
           1       0.94      0.83      0.88       284

    accuracy                           0.85       438
   macro avg       0.84      0.86      0.85       438
weighted avg       0.87      0.85      0.86       438
```



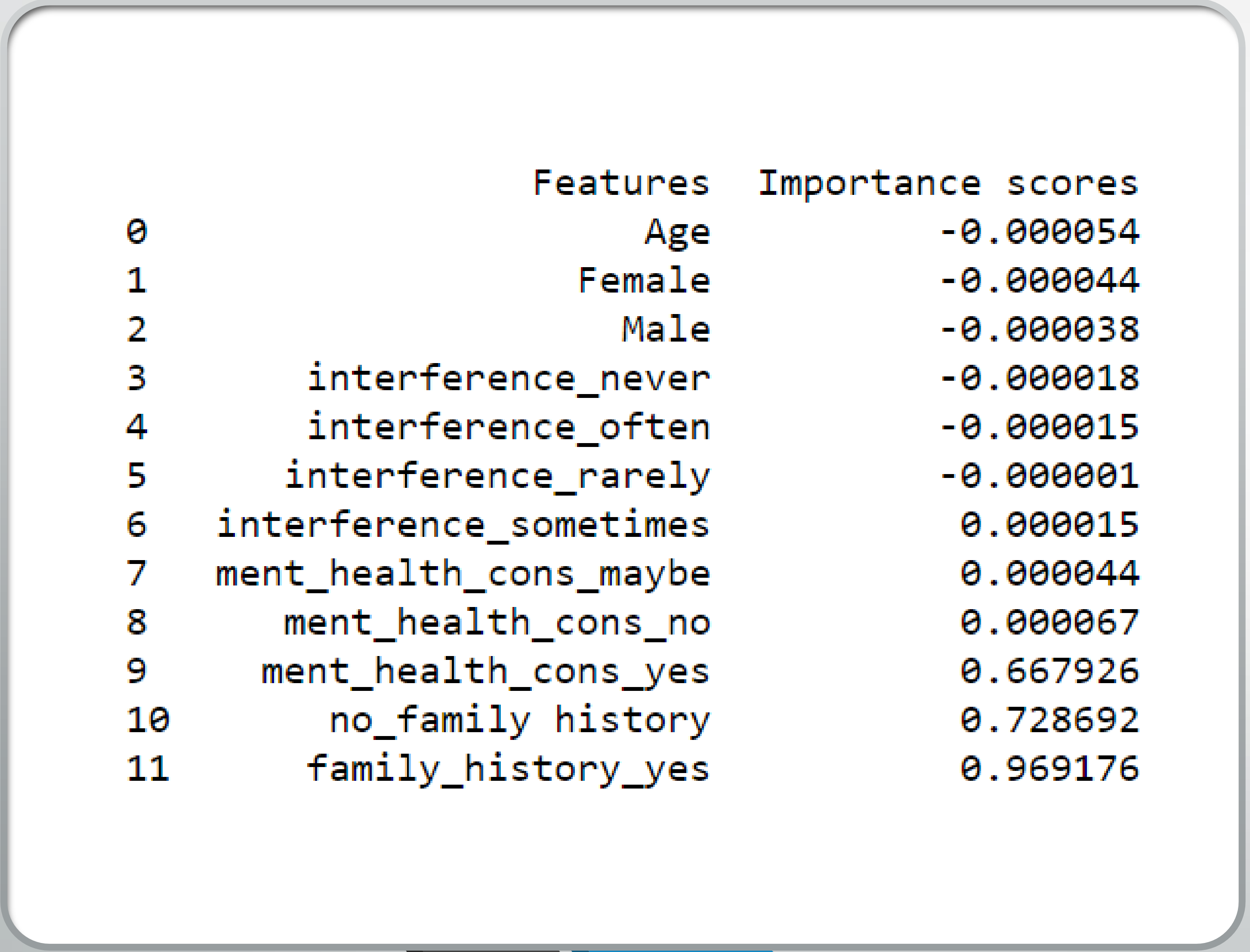
```
              precision    recall  f1-score   support

           0       0.89      0.81      0.85        21
           1       0.87      0.93      0.90        28

    accuracy                           0.88        49
   macro avg       0.88      0.87      0.87        49
weighted avg       0.88      0.88      0.88        49
```

Feature Importance: Logistic Regression



```
                 Features  Importance scores
0                     Age           0.035764
1                  Female           0.125359
2                    Male           0.149308
3      interference_never           0.149308
4      interference_often           0.183755
5     interference_rarely           0.208436
6  interference_sometimes           0.245067
7  ment_health_cons_maybe           0.245067
8     ment_health_cons_no           0.375301
9    ment_health_cons_yes           1.429901
10      no_family history           1.675590
11     family_history_yes           1.806279
```

Feature Importance: SVM



```
                  Features  Importance scores
0                      Age          -0.000054
1                   Female          -0.000044
2                     Male          -0.000038
3       interference_never          -0.000018
4       interference_often          -0.000015
5      interference_rarely          -0.000001
6   interference_sometimes           0.000015
7   ment_health_cons_maybe           0.000044
8      ment_health_cons_no           0.000067
9     ment_health_cons_yes           0.667926
10       no_family history           0.728692
11      family_history_yes           0.969176
```

Limitations and Assumptions

- This survey only covers 2014
- We focused mainly on just men and women, not over those with other gender identities
- We only really utilized 5 major features

Improvements

- In order to improve the model further, I could have used a different train-test split to explore other issues with mental health in technology
- Extract more specific information about mental health of each patient
- Use other sources to improve diversity of gender identities

Conclusion

- The best models were logistic regression and SVM
- After hyperparameter tuning, the best features to use when predicting who in the tech industry goes to therapy are work interference, family history, and whether or not the participant fears talking to co-workers/superiors about their mental health
- The model can still be improved

References

- https://www.mentalhealthfirstaid.org/2019/02/5-surprising-mental-health-statistics/#:~:text=In%20the%20United%20States%2C%20almost,equivalent%20to%2043.8%20million%20people.
- Survey I used: https://www.kaggle.com/osmi/mental-health-in-tech-survey