Mental Health in Tech Industry

Christine Yao
Brown University
Github

12.12.24

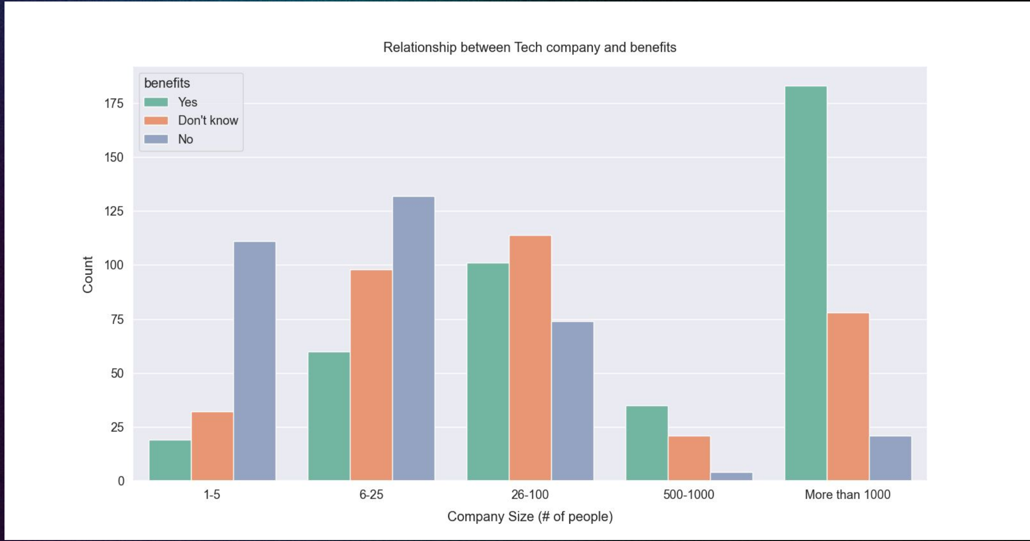
Introduction + Recap

- Predict whether people sought mental health help/therapy based on their survey answers

- This could be used to help **improve mental health resources** in tech companies, and help create a **better atmosphere** for workers

- **Source of data:** OSMI 2014 Survey https://osmhhelp.org/research.html

- Survey answers for target variable sought treatment was yes (translates to 1 after encoding) and no (translates to 0), so this is **Classification Problem**



Insights

Data shows that more mental health support and clearer benefits would help employees feel more confident on taking action and reaching out for help

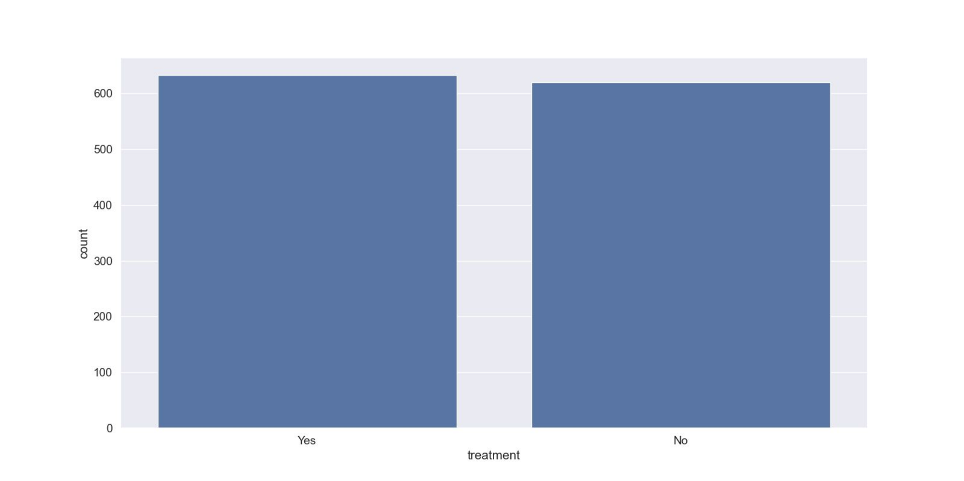
From EDA:

- Identified patterns related to mental health factors such as job satisfaction, and work-life balance
- Noticed an unbalanced gender ratio
- Larger tech companies → resources for mental health benefits

- Cleaning of the dataset
  - Age
  - Gender
  - Dropping columns like comments (although I made a word cloud with comments-more on this later)
- Preprocessing
  - OHE on categorical features
  - Ord Encoder on Ordinal features
  - Minmax only on Age
  - Pipeline automatically applies standard scaler later on

Baseline Accuracy

- Target Variable is balanced
- Baseline accuracy ~ 0.51
- Since dataset is balanced and there is no high cost of predicting positives or negatives, I chose to use accuracy as the metric



Cross Validation

My pipeline consisted of the following in a loop for each model. Ran 10 times on 10 different random states.

- Created a function
  - ```
    def MLpipe_KFold_Accuracy (X, y, preprocessor, ML_algo, param_grid, nr_states  =10):
    ```
    - Returns test scores of all 10 random states, best models and hyperparameters for one ML algo when called

Inside the function:

- Splitting
  - **Initial split** with train-val (80%) test set (20%)
- CV Pipeline
  - **KFold** (n_splits=5), ensures robust evaluation
  - **make_pipline**(preprocessor, ML_algo) which applies standard scaler to model
  - **Grid Search** to tune **parameters** (shown on the next slide), assess all parameter combinations for the model and evaluate with accuracy as a metric, then use the best model for that random state to predict on test set.
  - **Test scores** added to a list for future use

Before this I had already collected features

Output: 10 different models, with 10 best parameter combinations.
    ```

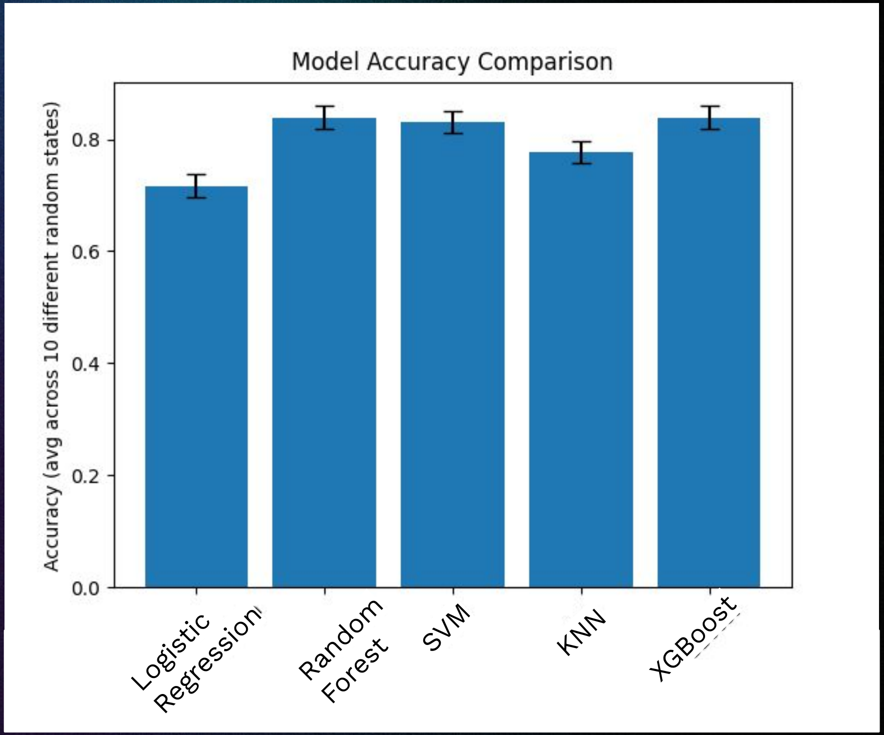
Model	Parameters Tuned	Optimal Parameters
Logistic Regression	C- regularization inverse (log scale) penalty (elastic net, l1, l2) Solver (elastic net requires saga,	C: 1.0 penalty: l1 solver: liblinear
Random Forest	N_estimators (linear) Max_depth (linear) min_samples _split min_samples_leaf	max_depth: 3 max_features: 0.75
K-Nearest Neighbors	metric n_neighbors weights	metric: 'euclidean', n_neighbors': 11 weights: distance
Support Vector Machine	C-regularization inverse (log scale) Gamma (kernel coefficient)	C: 1.0 Gamma: 0.1
XGBoost	n_estimators (linear) Max_depth (linear scale) Learning_rate (log scale)	learning_rate: 0.01 max_depth: 3 n_estimators: 50

Table of Models with average accuracies across 10 random states

Model	Average Accuracy	Standard Deviation
Logistic Regression	0.7160	0.0209
Random Forest	0.8389	0.0204
KNN	0.7767	0.0198
SVM	0.8297	0.0197
XGBoost	0.8389	0.0197

Best Model: XGBoost
Accuracy: 0.8389
std 0.02

Model comparison visual representation



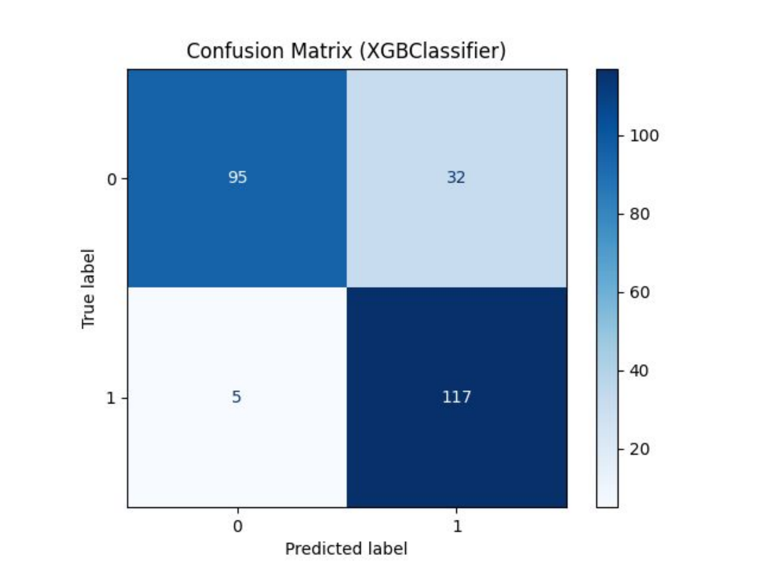
Confusion Matrix for the most accurate model:

Accuracy: 0.8514

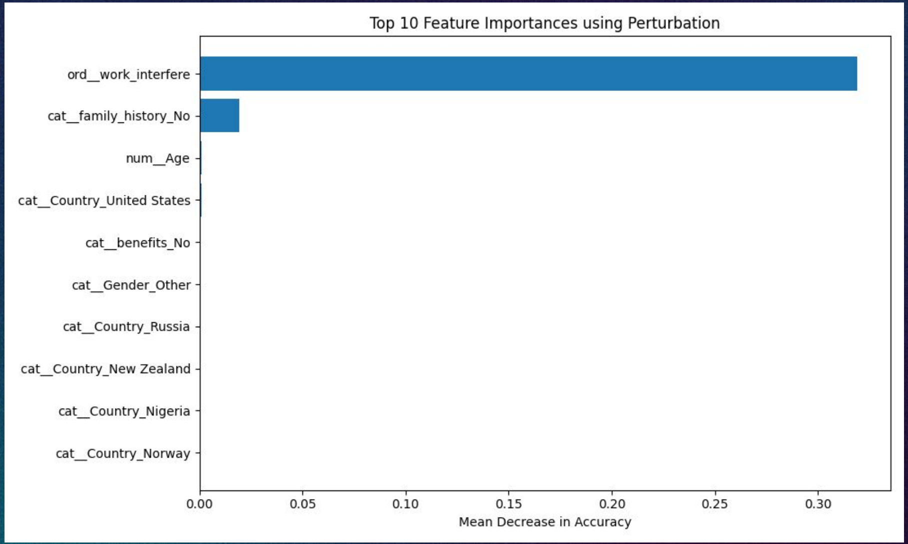
Precision: 0.7852

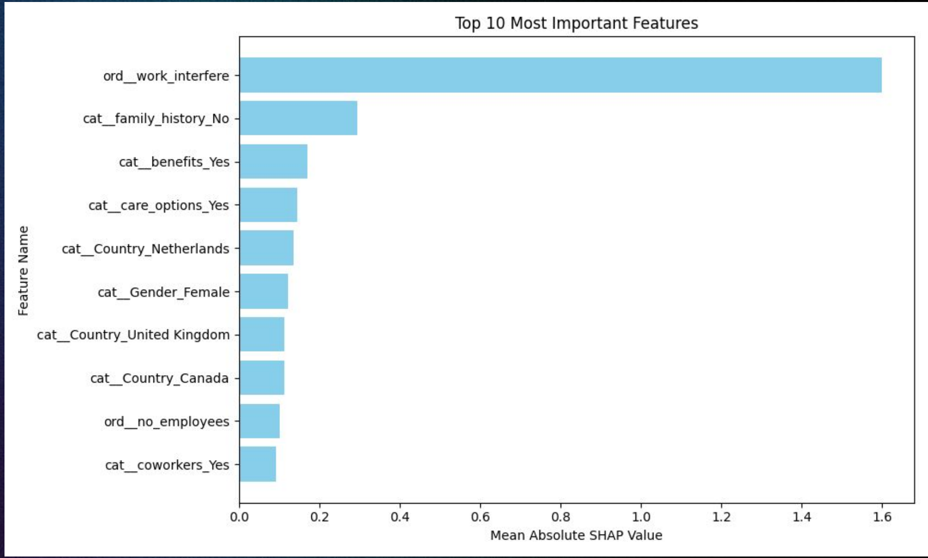
Recall: 0.9590

F1 score: 0.8634



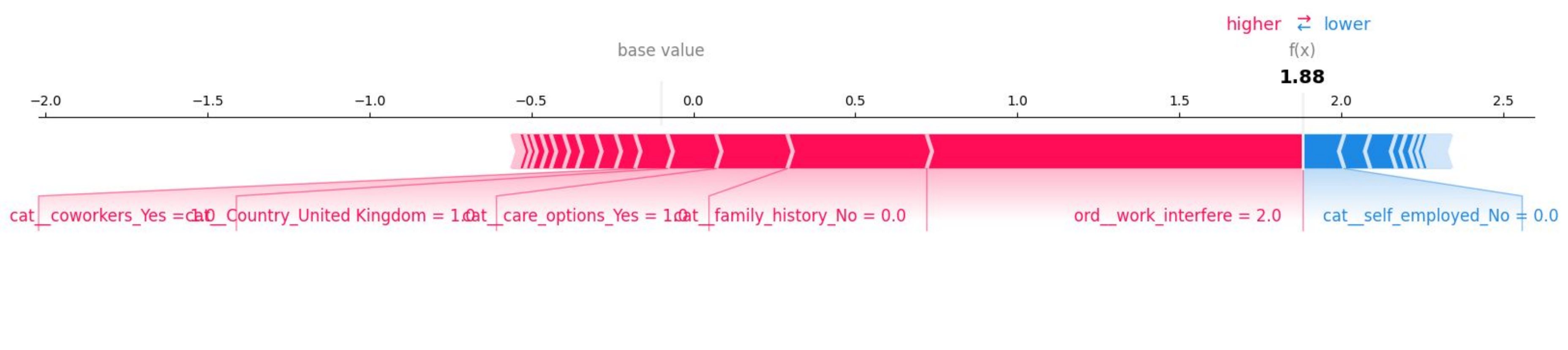
Global Feature Importances



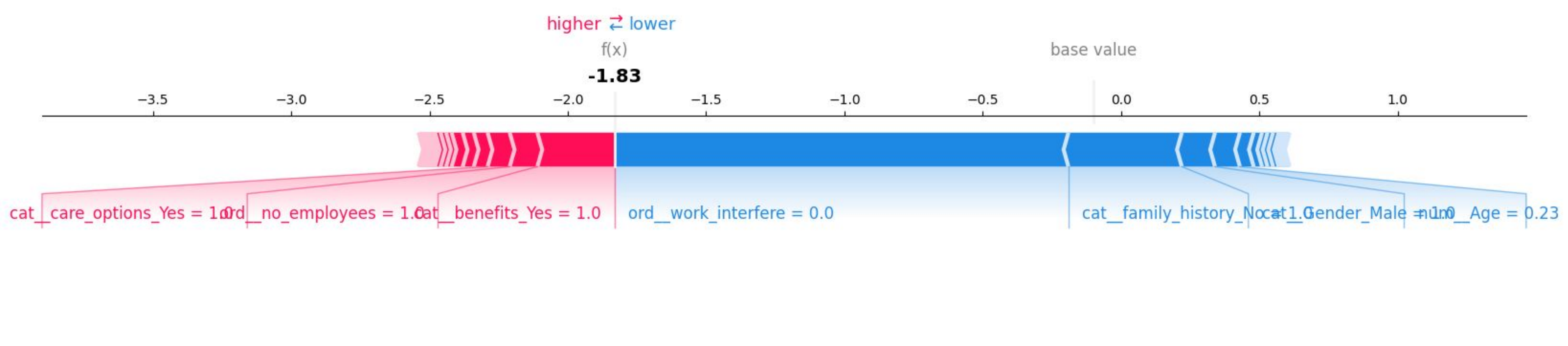


SHAP Local Feature Importances for two people

SHAP Force Plot for index 0:



SHAP Force Plot for index 100:



Word Cloud created from Comments

Future Considerations

- Tuning more hyperparameters to improve test score
- Understand and analyze the comments feature section
    - Sentiment analysis
    - Trigger words in the comments
- Collect more data points and possibly more survey questions on what mental health issues people had such as schizophrenia, bipolar disorder, or anxiety.