Mental Health Accommodation in Tech Companies

Presented by Xinbei Yu
Institution: DSI at Brown University
Date Presented: 12/06/2022
Link to the project:
https://github.com/XinbeiYu00/project-XinbeiYu.git

Recap

- **Research Question:** Which factors affect the easiness of taking a leave from work due to mental health conditions?

- **Why Important:** With growing attention to mental health issues in every industry, both employer and employee should start exploring their options and responsibilities.

- **Type of Problem:** Classification

- **Target Variable:** Easiness of Taking A Leave ([leave] in the data set)

- **Data Source:** Mental Health in Tech Survey, Kaggle, https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey

- **EDA Recap/Update:**
  - Cleaned gender feature, categorized non-binary and unanswered as other
  - Target variable is imbalanced: 455: 218: 176: 98: 78

- **Preprocessing Recap:**
  - Standardscaler for age as it is skewed
  - Mostly onehotencoder

Splitting

Train_Test_Split

- Implicit Group: Company
- Distributed through website
- Assume i.i.d

StratifiedKFold

- Imbalanced Data: 455: 218: 176: 98: 78

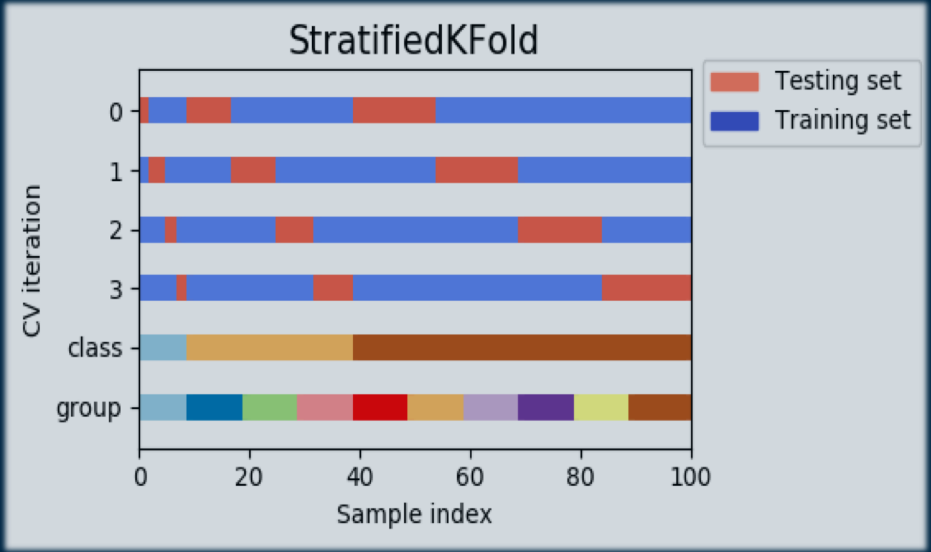


Figure: from class notes week3

CV Pipeline

XGBClassifer:

- Parameter grid
- Hyperparameter tuned: max_depth
- Subsample = 0.5: overfitting
- Evaluation metrics: accuracy_score

Random forest, SVC, KNN, Logistic Regression:

- GridSearchCV
- Hyperparameter tuned:
  - Random Forest: max_depth, max_features
  - SVC: gamma, C
  - KNN: n_neighbors, weights
  - Logistic Regression: penalty
- Evaluation metrics: accuracy_score

Result

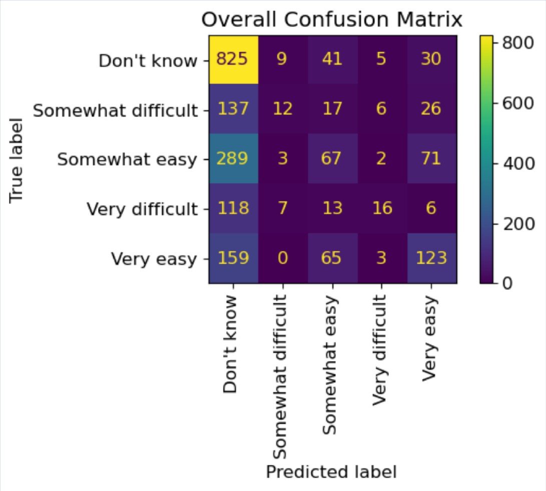
Baseline Accuracy: 0.444

- Assuming all predicted as 'Don't know'

Model	**XGBoost Classifier**	**K-Nearest-Neighbors**	**Random Forest**	**Logistic Regression**	**Support Vector Classifier**
Test Score Mean	0.490	0.493	0.493	0.473	0.509
Test Score Standard dev.	0.020	0.022	0.013	0.024	0.025

Confusion Matrix

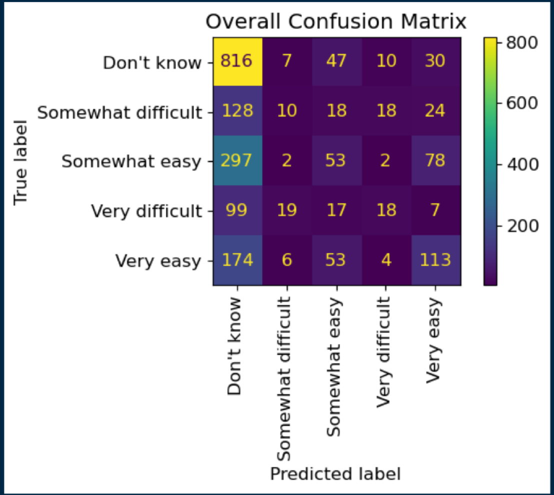
SVC



Overall Confusion Matrix

True label \ Predicted label	Don't know	Somewhat difficult	Somewhat easy	Very difficult	Very easy
Don't know	825	9	41	5	30
Somewhat difficult	137	12	17	6	26
Somewhat easy	289	3	67	2	71
Very difficult	118	7	13	16	6
Very easy	159	0	65	3	123

Random Forest



Overall Confusion Matrix

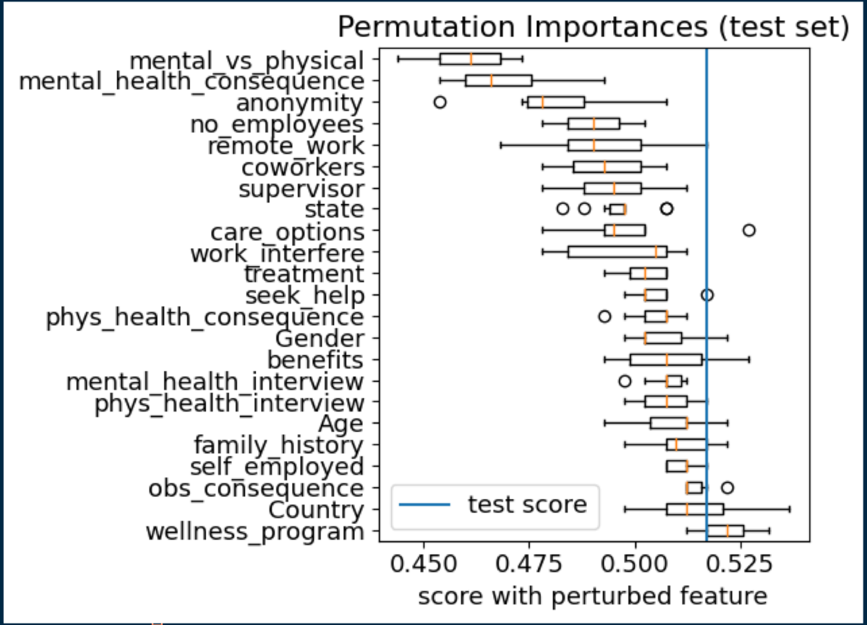
True label \ Predicted label	Don't know	Somewhat difficult	Somewhat easy	Very difficult	Very easy
Don't know	816	7	47	10	30
Somewhat difficult	128	10	18	18	24
Somewhat easy	297	2	53	2	78
Very difficult	99	19	17	18	7
Very easy	174	6	53	4	113

Findings:

- Highest number of correctly predicted don’t know class
- Most correctly predicted points are from don’t know and very easy class

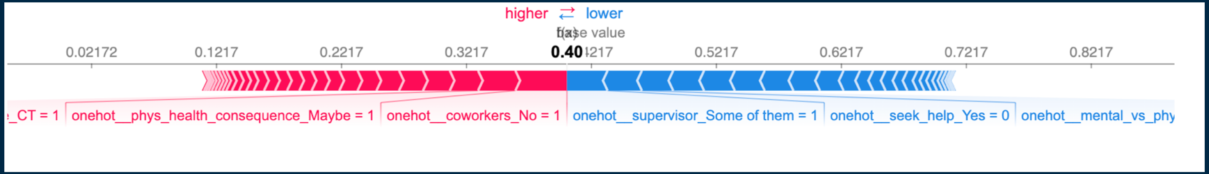
Global Feature Importance

permutation

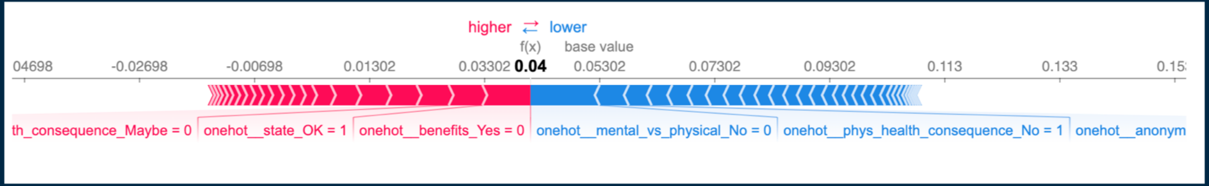


Local Feature Importance

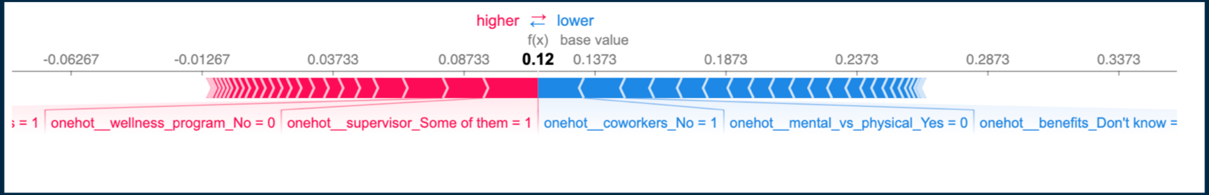
Index = 37, class = ‘Don’t know’



Index = 37, class = ‘Very Difficult’



Index = 37, class = ‘Very Easy’



Outlook

Potential Improvement on Predictive Power

- Try different model: Catboost (mostly categorical features)
- Drop some negative importance feature: Wellness Program

Potential Improvement on Interpretability

- Reduce classes into 2: Easy and Hard
- Use LinearSVM: faster implementation, can be visualized

Q&A