Predicting Mental Health Treatment in the Tech Industry: A Project Report

Introduction

Mental health is a critical yet often stigmatized issue within the fast-paced tech industry. The goal of this project was to leverage machine learning to address this challenge by building an accurate and interpretable model to predict whether an individual in the tech sector has sought treatment for a mental health condition.

The solution is an end-to-end data product, beginning with raw survey data and culminating in a deployed web application. This tool is designed not only to make predictions but also to provide actionable insights that can help tech companies foster more supportive and informed workplace environments.

Methodology

Our approach followed a structured, industry-standard machine learning workflow, from data preprocessing and feature engineering to model selection, tuning, and deployment.

1. Data Preprocessing

The initial dataset required significant cleaning to be suitable for modeling. Our key preprocessing steps included:

- **Handling Invalid Data:** The Age column contained impossible values (e.g., negative numbers), which were corrected by replacing them with the median of a valid age range (18-75). The Gender column contained numerous non-standard free-text entries, which were standardized into three categories: 'Male', 'Female', and 'Other'.
- **Imputing Missing Values:** Missing entries in categorical columns like work_interfere and self_employed were filled using the mode (the most frequent value) to preserve the data distribution.
- **Dropping Unnecessary Columns:** The Timestamp, comments, and state columns were dropped as they were either irrelevant for prediction, mostly empty, or had excessive missing values that would introduce bias.

2. Feature Engineering

To enhance the model's predictive power, we engineered several new features from the existing data:

- **Composite Scores:** We created a support_score by aggregating an employee's responses to questions about benefits, wellness programs, and leave policies.

Similarly, a stigma_score was created to quantify the perceived negative consequences of discussing mental health.

- **Binning:** The continuous Age feature was binned into categorical age_groups (e.g., '18-30', '31-40') to help the model capture non-linear relationships.
- **Encoding:** All categorical features were converted to a numerical format. Ordinal mapping was used for features with an inherent order (e.g., company size), while one-hot encoding was applied to nominal features like Country and Gender.

3. Model Selection and Justification

We evaluated a suite of classification algorithms to find the best performer for this task:

1. **Baseline Models:** We began with Logistic Regression as a simple, interpretable baseline, which achieved a strong 76% accuracy. We then tested more complex ensemble models like Random Forest, XGBoost, and LightGBM.
2. **Champion Model:** Initially, the advanced models did not outperform the baseline. However, after hyperparameter tuning, the Random Forest Classifier matched the baseline's performance while being a more powerful and nuanced model. It was selected as our final champion model.

4. Hyperparameter Tuning

To optimize our selected Random Forest model, we used Optuna, an advanced Bayesian optimization framework. This was chosen over a traditional Grid Search for its efficiency in intelligently searching the hyperparameter space. The process involved running 100 trials with 5-fold cross-validation to find the optimal combination of parameters like n_estimators and max_depth, successfully boosting the model's performance to 76% on the validation set.

5. Deployment Procedure

The final model was deployed as a live web application on Hugging Face Spaces.

- **Modular Codebase:** The project was structured into a modular pipeline with separate scripts for preprocessing (preprocess.py), training (train.py), and prediction (predict.py).
- **Web Framework:** A Flask application (app.py) was built to serve as the backend, handling requests and processing user input.
- **Frontend:** A user-friendly HTML form (templates/) was created and styled with external CSS (static/) to provide a professional user interface.
- **Deployment:** The entire application was deployed using Git. Key configuration files like requirements.txt (to define the server environment) and README.md (to

configure the Hugging Face Space) were created to ensure a reproducible and successful build.

Results

Model Performance

The final, tuned Random Forest model was trained on the full training dataset and evaluated on the completely unseen test set. The official performance is as follows:

- **Overall Accuracy: 73.02%**

Class	Precision	Recall	F1-Score
0 (Did Not Seek Treatment)	**0.71**	**0.77**	**0.74**
1 (Sought Treatment)	**0.75**	**0.70**	**0.72**

General Insights & Model Interpretation

To move beyond simple feature importance and understand *why* the model makes its decisions, we employed SHAP (SHapley Additive exPlanations). This advanced technique reveals how each feature influences individual predictions. The SHAP analysis confirmed that the model learned a logical and clinically relevant set of rules:

1. **family_history:** This is the most dominant predictor. A high value (i.e., having a family history) strongly and consistently pushes the model's prediction towards "Sought Treatment."
2. **work_interfere:** The severity of work interference is the second most critical factor. When a respondent indicates their condition "Often" or "Sometimes" interferes with work, it significantly increases the probability of predicting they sought treatment.
3. **care_options:** Awareness of care options is highly influential. A high value ('Yes') pushes the prediction towards seeking treatment, while a low value ('No' or 'Not sure') strongly pushes the prediction in the opposite direction, highlighting that a lack of awareness is a major barrier.

In summary, the model's logic is transparent: it prioritizes an individual's background and the severity of their condition, then heavily weighs the perceived support system within their workplace.

Discussion

Challenges Faced and Solutions

- **Data Quality:** The primary challenge was the messy and inconsistent nature of the raw survey data. This was overcome through a systematic preprocessing pipeline involving robust cleaning and imputation techniques.
- **Deployment:** The deployment phase presented several technical hurdles, including Python pathing issues, library dependency conflicts, and binary version mismatches between local and server environments. These were systematically debugged and resolved by creating a modular codebase, pinning exact library versions in requirements.txt, and synchronizing the model artifacts with the server environment.

Limitations and Potential Improvements

- **Geographical Bias:** The dataset is heavily skewed towards respondents from the United States (~60%). Consequently, the model may not generalize well to tech employees in other parts of the world.
- **Performance Bias:** Our fairness analysis revealed a performance disparity, with the model being approximately 5% more accurate for female respondents than for male respondents.
- **Future Work:** To improve the model, future efforts should focus on gathering a more geographically diverse dataset and implementing advanced bias mitigation techniques to ensure more equitable performance across demographic groups.

Real-World Implications

This project demonstrates the tangible value of applying data science to workplace wellness. The key insights derived from the model have direct real-world implications:

- **Tech companies can use these findings to focus their mental health initiatives. For example, the high importance of care_options and benefits proves that proactive and clear communication of available resources is a highly effective intervention strategy.**
- **The tool can serve as a prototype for more sophisticated systems that help organizations understand the factors driving mental health treatment-seeking in their specific workforce, enabling them to create a more supportive and destigmatized environment.**

Conclusion

This project successfully delivered an end-to-end machine learning solution, from raw data to a deployed, interactive web application. We developed a predictive model with 73% accuracy that can identify whether a tech employee is likely to have sought mental health treatment. The key takeaway is that an individual's decision is most strongly influenced by their personal history and the level of support and awareness within their

immediate work environment. This work serves as a powerful proof-of-concept for how data-driven insights can be used to foster healthier and more supportive workplaces in the tech industry.