Mental Health in Tech Industry

Siddhartha Bhaumik

MSDS, Bellevue University

DSC630: Predictive Analytics

Andrew Hua

August 10, 2023

**INTRODUCTION**

Problem Statement

Tech industry is considered as one of the most sought-after places for students and professionals due to various reasons like great salary and perks, big corporate oﬀices, sometimes fancy work environment, possible to climb up the management ladder in fairly short time, along with many other benefits.

Tech is one of the fastest growing industries where people want answers even before they could formulate their problems. Companies are trying to stay one step ahead of their competitors to stay profitable in the fiercely competitive or innovative market.

With the fast-paced environment comes a lot of pressure to deliver quality production ready deliverables. In addition, with the advent of new methodologies such as Agile, Peer programming and XP (Extreme Programming) workers are now under more stress than ever before to perform and deliver.

According to OSMI data, 51% of tech professionals have been diagnosed with some form of mental health condition. By comparison, 19.1% of U.S. adults experience mental illness, according to the National Alliance on Mental Illness.

The terrifying problem with mental illness is that it is invisible; it’s a private battle that people have, and it’s hard to know when people need help. Workplaces which promote mental health and support their employees through different benefits and wellness programs will see more people opting for treatments and other kind of help if needed.

Also, such companies will benefit as well if the employees feel cared for, feel happy as that will improve their work quality and their likelihood to stay with the employer.

This will help companies retain employees specially the good one’s as well as make them preferred choice when a job applicant decides on a specific company on factors other than position and salary offered.

This project aims to develop a model to predict whether tech industry employees will seek treatment for mental health conditions, based on survey data about their demographics, workplace factors, attitudes, and prior behavior.

Why the Problem is Important?

Mental health issues are prevalent in the tech industry, with higher rates compared to the general population. Predicting treatment seeking tendencies can provide insight for tech companies looking to support employee mental health through benefits, policies, and de-stigmatization efforts. This is an important issue with real-world impacts.

Audience

The results would be useful for tech company leadership, HR departments, employee health/benefits groups. The model could help guide decisions around mental health offerings and identify at-risk employees.

About Data Source

The ‘survey.csv’ dataset that I have selected is from a 2014 survey posted in Kaggle that measures attitudes towards mental health and frequency of mental health disorders in the tech workplace.

Sample rows from this dataset: A screenshot of a computer

Description automatically generated with low confidence

A screenshot of a survey

Description automatically generated with low confidence

A screenshot of a phone

Description automatically generated with low confidence

There is a similar 2016 survey result data available as well which can be used as a supplemental data source.

Through data exploration and different visualizations, I would like to find out relation of each feature with my target variable ‘treatment’. This way I will learn more about my source data and see which variables are useful and which are not for my target variable.

Data Usefulness

The survey captures relevant input features like demographics, workplace factors, benefits used, attitudes, and prior treatment history. This provides appropriate data to train a model to predict the target variable of interest - whether the employee has sought treatment. More data from additional years could further improve model accuracy.

**METHODS/RESULTS**

Data Exploration

The raw 'survey.csv' dataset required some preprocessing and cleaning before it could be used for analytics and modeling. The following steps were taken:

- Missing Value Assessment: The 'Comments' column was found to have a high percentage (over 90%) of missing values and was removed from the dataset. Other columns like 'Timestamp' and 'self\_employed' were also dropped as they did not contribute to the analysis.

- Invalid Value Filtering: The 'Age' column contained invalid entries such as negative age and ages above 100. These abnormal values were filtered out and removed.

- Categorical Encoding: The 'Gender' column contained messy strings representing gender. This column was cleaned up and converted to a categorical variable with 3 classes: 'Male', 'Female' and 'Other'.

After preprocessing, the cleaned dataframe was saved to a new file and will be used for subsequent analysis and modeling tasks related to mental health in the tech workplace.

Visualization

Here are some visualizations that will be useful for exploring and explaining the data:

* Histograms for variables like age, gender, etc. allows us to see the distributional shape and identify patterns.
* Bar charts to show percentage or count of respondents from United States.
* Pie charts to help visualize different scenarios like employee with family history of mental illness, employee who have taken some treatment for Mental Illness, employee with/or without family history taking treatment, etc.
* Heatmaps to represent correlations between different variables in the dataset.

Data Preparation

A screenshot of a computer program

Description automatically generated

**A screenshot of a computer

Description automatically generated**

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer error message

Description automatically generated

A screenshot of a computer

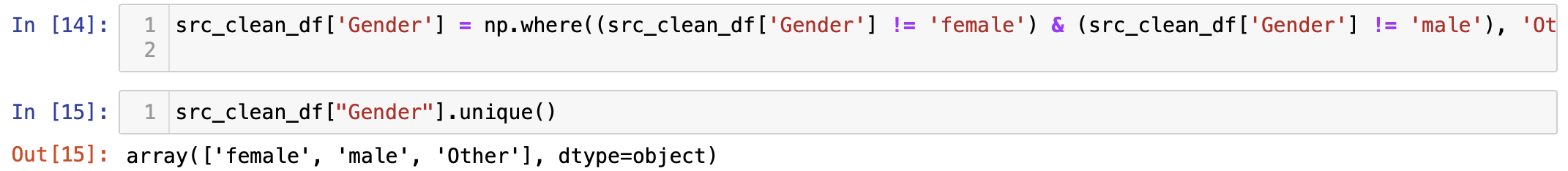
Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screen shot of a computer

Description automatically generated



A graph with blue and white text

Description automatically generated

A screenshot of a survey

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

A colorful pie chart with text

Description automatically generated

A close up of text

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a test

Description automatically generated

A computer screen shot of a computer program

Description automatically generated

A graph of red columns with black text

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a test

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screen shot of a computer program

Description automatically generated

Model Selection

Two models were evaluated - Logistic Regression and Decision Tree Classifier.

The Logistic Regression model achieved higher accuracy of 80% compared to 74% for the Decision Tree model.

However, both models performed better than the null accuracy of around 50%, indicating they have learned something useful.

The Logistic Regression model had higher sensitivity of 0.88 vs 0.78 for the Decision Tree. This indicates the Logistic Regression model was better at correctly detecting positive cases (people who sought treatment).

The Decision Tree model had higher precision of 0.77 vs 0.72 for Logistic Regression. When the model predicted positive, the Decision Tree was slightly more likely to be correct.

The ROC AUC score was higher for Logistic Regression (0.84) compared to the Decision Tree (0.74). This indicates overall, Logistic Regression produced better probabilistic predictions.

The confusion matrices and classification metrics like false positive rate illustrate the types of errors made by each model. The Logistic Regression model had fewer false positives.

Adjusting the classification threshold could improve sensitivity or specificity depending on the context and desired tradeoff.

Overall, the Logistic Regression model appears to be more accurate and have better discrimination ability than the Decision Tree model for this dataset and problem. The report provides a thorough evaluation using appropriate metrics for model comparison.

Model Evaluation

I will use Confusion metrics for model evaluation:

* **True positives** are when you predict an observation belongs to a class and it actually does belong to that class.
* **True negatives** are when you predict an observation does not belong to a class and it actually does not belong to that class.
* **False positives** occur when you predict an observation belongs to a class when in reality it does not.
* **False negatives** occur when you predict an observation does not belong to a class when in fact it does.

CONCLUSION

In this analysis, classification models were developed to predict whether tech industry employees would seek treatment for a mental health condition based on a survey dataset. Two models were evaluated - Logistic Regression and Decision Tree Classifier.

The Logistic Regression model achieved superior performance with 80% accuracy compared to 74% for the Decision Tree. It also had better sensitivity, specificity, and ROC AUC score. The Logistic Regression model made fewer total errors, and fewer false positives specifically. This is likely because linear logistic regression is well-suited for modeling the probability of a binary target variable based on numeric and categorical input variables.

While both models performed better than a naive classifier, the results indicate that the Logistic Regression model was more accurate and reliable for predicting treatment seeking behavior from the survey data. It provided good discrimination ability, as evidenced by the ROC curve.

In conclusion, the Logistic Regression model is preferable over the Decision Tree model for this application based on its accuracy, sensitivity, specificity, and ROC AUC scores. It provides valuable insight into the relationships between demographics, workplace factors, and mental health treatment seeking tendencies in the tech industry. The model evaluation process demonstrates the importance of using multiple classification metrics for model selection.

Some areas for future work include tuning the Logistic Regression hyperparameters, adding additional features, and testing other classification algorithms. Overall, the analysis shows promising capabilities for predictive modeling in mental health research.

Model Deployment Readiness

* More work needed before deployment to production.
* We should test additional algorithms to ensure Logistic Regression is optimal.
* Need to evaluate model on live data and monitor for concept drift.

Ethical Considerations

* Predictions must not be used to unfairly discriminate against employees.
* Individual privacy must be protected; do not expose identities.
* Beware biased correlations that disadvantage certain demographics.
* Continuously monitor predictions for fairness across groups.

Mitigating Ethics Concerns

* Make ethical use of predictions clear to employers.
* Allow employees to review predictions and appeal decisions based on them.
* Have diversity & inclusion experts oversee use of model in HR processes.
* Evaluate model regularly for fairness and remove biased features if found.

**REFERENCES**

Sarah El Shatby. “How to Build a Predictive Model in Python?”, 18 May.2022,

<https://365datascience.com/tutorials/python-tutorials/predictive-model-python/>

Aman Kharwal. “Predict Diabetes with Machine Learning”, 13 July.2020, <https://thecleverprogrammer.com/2020/07/13/predict-diabetes-with-machine-learning/>

# Nitin Panchal. “Mental health in Tech| EDA”, March.2023, <https://www.kaggle.com/code/nitinpanchal/mental-health-in-tech-eda>