2023 3rd International Symposium on Computer Technology and Information Science (ISCTIS) | 979-8-3503-2538-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISCTIS58954.2023.10213032

Application of machine learning to predict mental health disorders and interpret feature importance

Yifan LI
Department of Statistics and Data Science, Faculty of Science and Technology
Beijing Normal University-Hong Kong Baptist University United International College (UIC)
Zhuhai, China
evelinee6611@outlook.com

Abstract—**The mental health and mental illness crisis has become increasingly acute in recent years, and many digital solutions with artificial intelligence at their core offer hope for reversing the deterioration of our mental health. Machine deep learning techniques can be used to analyse big data to build predictive models for psycho-education, assessment and screening to assess the mental health status of subjects, and can help the clinical community discover information that is not available to many traditional psychological research tools. This paper presents an in-depth analysis of a mental health survey and examines how it can be applied to the AI/ML domain of mental health research and how machine learning models can be used in this domain for fitting and prediction. Based on this, the importance of the presence or absence of current mental health disorders on other characteristics of respondents is assessed and visualised. It was found that the Cross Gradient Booster (Random Forest) model gave the best prediction fit among the various types of machine learning models, and the Grid Search algorithm was used to confirm that the final model had the highest accuracy of 0.79784 at a learning rate of 0.1. The Permutation Importance analysis revealed that the most important characteristic is whether or not the person has suffered from a mental health disorder in the past.**

Keywords: machine learning, data mining, artificial intelligence, feature importance, model selection, interpretative machine learning, prediction, mental health, visualisation

I. Introduction

The World Health Organisation (WHO) conceptualises mental health as "a state of well-being in which individuals are aware of their capabilities, able to cope with the normal stresses of life, able to work productively, and able to contribute to their communities"[1]. Mental disorders and psychoactive substance-related disorders are highly prevalent worldwide and are a major contributor to morbidity, disability and premature death. People with severe mental illness die prematurely - up to 20 years - due to preventable physical conditions. Mental health conditions can have a significant impact on all areas of life, such as school or work performance, relationships with family and friends, and the ability to participate in the community. The two most common mental health conditions, depression and anxiety, cost the global economy US$1 trillion each year. Many mental health conditions can be treated effectively at relatively low cost, but the gap between those who need treatment and those who can access it remains wide. Coverage of effective treatment remains extremely low[2].

Artificial intelligence has greatly improved the effectiveness of prevention, measurement, diagnosis and treatment of psychiatric disorders through big data mining, natural language processing and deep learning algorithms for mental illness. Machine learning and artificial intelligence tools can enable healthcare stakeholders to access life-saving innovations. Machine learning is an artificial intelligence technique that allows a machine to perform a task well and autonomously when it is given a large amount of data and examples of good behaviour. It can also help identify meaningful patterns that humans might not be able to find as quickly without the help of a machine[3]. Combining machine learning with the mental health field, this related technology can quantify patient needs, patient quality of life, patient affordability, patient needs, healthcare costs, healthcare savings, healthcare investments, patient and income returns to healthcare, healthcare savings and healthcare measures, determine an individual's risk of developing a mental health condition, and can greatly assist in the early detection and diagnosis of mental health problems, and related measures for early preventive interventions, thus achieving the potential to reduce the risk of mortality[4].

In this paper, the results of the OSMH/OSMI Mental Health in Tech Survey conducted by Open Source Mental Health, a non-profit company in 2016, are used as the dataset, which was pre-processed and attempted to be fitted and predicted using different machine learning models to select the model that tested best with the dataset. A visual analysis of the importance of the data features was carried out to help employers to recognise the areas of mental health that are important to their employees and to provide advice and assistance to prevent them from being affected.

The following is a flow chart of the specific steps in this article for the processing and analysis of this data set:

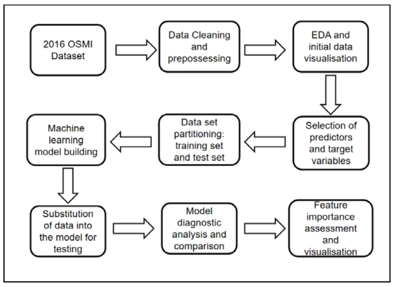

2016 OSMI Dataset
Data Cleaning and prepossessing
EDA and initial data visualisation
Selection of predictors and target variables
Data set partitioning: training set and test set
Machine learning model building
Substitution of data into the model for testing
Model diagnostic analysis and comparison
Feature importance assessment and visualisation

Figure 1. Flow chart of paper structure

II. Methods

A. Data Source

The dataset used in this study is from the Open Sourcing Mental Illness (OSMI) organization's questionnaire: OSMH/OSMI Mental Health in Tech Survey. Resources created by OSMI volunteers provide direction in These works are licensed under a Creative Commons Attribution-ShareAlike 4.0 International. The annual questionnaire has been ongoing in recent years and has collected over 1200 responses in 2014, 1400 in 2016, 756 in 2017, 400 in 2018, 350 in 2019, 180 in 2020 and 131 in 2021[5]. To ensure the accuracy and generalisability of the predictions and analysis results, we preferred a dataset with a large sample size that better characterised the overall population: the OSMI Mental Health in Tech Survey 2016. With 1570 responses, the 2016 survey aimed to measure attitudes towards mental health in the tech workplace, and examine the frequency of mental health disorders among tech workers.

B. Data Cleaning

In the original unprocessed dataset of the 2016 OSMI Mental Health in Tech Survey, it had 1433 rows and 63 columns, excluding the horizontal and vertical headers, with the horizontal coordinates being the respondent number and the vertical coordinates being the questionnaire questions. There are a large number of missing values in the dataset, and as the questions are not multiple choice, but open-ended, there is no standardised format for the responses, e.g. m, M, male, Male, etc. for gender. In addition, there are other problems such as invalid data and abnormal data. In order to subsequently process the data for analysis and to implement relevant visualisations, we need to perform data cleansing on this original dataset.

The following is a summary of specific aspects of data cleansing[6]:

- **Rename columns**: rewrites the original question stem into a short string.
- **Recode sex columns & company size**: normalises the answers uniformly.
- **Remove outliers from age:** replaces outliers with the average of non-abnormal ages.
- **Delete missing value listwise**: for variables where missing observations were more than half Delete the variable.
- **Pad missing value**: for other missing values take a padding using the plurality of that column.
- **Encode column**: converts columns of non-contiguous variables to numeric encoding.
- **Filter country**: only replies with more than 30 are retained.
- **Create technical column with flag 1/0**: create new variable virtual technical column by job position variable.

C. Exploratory Data Analysis

Exploratory data analysis (EDA) is used by data scientists to analyze and examine data sets and summarize their key characteristics, often using data visualization methods. It helps find the best way to manipulate data sources to get the answers they need and makes it easier for data scientists to spot patterns, find anomalies, test hypotheses, or test hypotheses[7].

The main purpose of EDA is to help you examine the data before forming hypotheses. It can help you identify obvious errors, better understand patterns in your data, spot outliers or unusual events, and find interesting relationships between variables. This is the Baseline Characteristic Table for this dataset, which disaggregates the quartiles of the variable of age and the results are shown below:

TABLE I. FEATURE CLINICAL CHARACTERISTICS OF THE STUDY POPULATION DISAGGREGATED BY QUARTILES OF AGE FROM OSMI 2016(N=1433)

CHARACTERISTIC	**AGE QUARTILES**				
	Q1	**Q2**	**Q3**	**Q4**	**P-VALUE**
n	313	334	315	273	
self_empl_flag (mean (SD))	0.10 (0.30)	0.18 (0.38)	0.19 (0.39)	0.29 (0.45)	<0.001
comp_no_empl (mean (SD))	2.58 (1.50)	2.58 (1.42)	2.65 (1.50)	2.86 (1.55)	0.08
tech_comp_flag (mean (SD))	0.83 (0.38)	0.85 (0.36)	0.77 (0.42)	0.77 (0.42)	0.02
mh_coverage_flag (mean (SD))	1.75 (1.28)	2.05 (1.28)	2.08 (1.27)	2.27 (1.19)	<0.001
mh_resources_provided (mean (SD))	0.93 (0.67)	1.02 (0.67)	0.95 (0.71)	1.07 (0.66)	0.043
mh_medical_leave (mean (SD))	1.07 (0.84)	1.26 (0.85)	1.23 (0.88)	1.30 (0.82)	0.004
mh_discussion_cowork (mean (SD))	0.78 (0.42)	0.92 (0.27)	0.95 (0.21)	0.90 (0.29)	<0.001
prev_mh_benefits (mean (SD))	1.37 (0.90)	1.57 (0.99)	1.38 (1.05)	1.57 (1.01)	0.01
prev_mh_discussion (mean (SD))	1.09 (0.46)	1.22 (0.61)	1.21 (0.60)	1.19 (0.56)	0.014
prev_mh_resources (mean (SD))	0.20 (0.44)	0.41 (0.57)	0.40 (0.58)	0.44 (0.59)	<0.001
prev_mh_anonimity (mean (SD))	0.42 (0.92)	0.64 (1.07)	0.62 (1.08)	0.72 (1.14)	0.005
prev_mh_conseq_coworkers (mean (SD))	0.33 (0.55)	0.40 (0.59)	0.44 (0.59)	0.50 (0.61)	0.004
sex (mean (SD))	0.39 (0.56)	0.28 (0.51)	0.23 (0.44)	0.31 (0.48)	0.001
age (median [IQR])	26.00 [24.00, 27.00]	31.00 [30.00, 32.00]	36.00 [35.00, 38.00]	45.00 [42.00, 49.00]	<0.001

a. This table only shows variables with p-values less than or equal to 0.05.
b. Q1 (age ≤28), Q2 (28< age ≤33), Q3 (33< age ≤39), Q4 (age >39).

III. Machine Learning Model Prediction

In order to predict whether a patient is currently suffering from a mental health disorder, the "has_current_mental_health_disorder" variable was used as the predictor target variable and all non-numeric variables in the dataset were removed and the remaining variables were used as predictor features variables.

Before building the prediction and classification models, we also need to partition the dataset into a training set and a test set. We randomly partitioned the dataset with a total of

1433 observations in a ratio of 80:20, resulting in 1146 training sets and 287 test sets.

Supervised machine learning models can be classified as single models and ensemble models. A single model is a machine learning model that consists of only one model and is trained and validated independently based on a particular model, including logistic regression, k-nearest neighbour, decision tree, neural network, support vector machine and naive Bayesian in this paper. The opposite of a single model is an integrated model. An integrated model is a combination of several single models into a single strong model that takes the strengths of all the single models and achieves relatively optimal performance. Commonly used integrated models in this paper include GBRF, XGBoost, Stochastic Gradient Descent and Random Forest[8].

Using several machine learning models described above, we put the test set into the model for fitting as well as prediction and obtained the corresponding results according to the information in the Confusion Matrix of the model as shown in the following table:

TABLE II. SUMMARY OF MACHINE LEARNING MODELS PREDICTION RESULTS

Machine Learning	Accuracy	Precision	Sensitivity	Specificity	F1 Score
Naive Bayes	0.75202	0.71274	0.71108	0.71108	0.71172
KNN	0.71429	0.70689	0.62386	0.62386	0.62134
Decision trees	0.69542	0.63875	0.66171	0.64299	0.62475
Random Forest	0.78167	0.75138	0.72633	0.72633	0.73325
Support Vector Machine	0.77628	0.75493	0.72476	0.72476	0.73367
Logistic Regression	0.77628	0.71978	0.71175	0.71175	0.71512
Neural Nets	0.71968	0.67928	0.68002	0.68002	0.67963
Stochastic Gradient Descent	0.64151	0.66933	0.60512	0.60512	0.59511
Cross Gradient Booster	0.76011	0.71829	0.71012	0.71012	0.71302
Cross Gradient Booster (Random Forest)	0.78706	0.71829	0.74116	0.74116	0.74915

In the table above, the meaning of these indicators is[9]:

- **Accuracy:** (True Positive + True Negative) / Total Predictions, which measures how often the model is correct.
- **Precision:** True Positive / (True Positive + False Positive),which measures what the percentage is truly positive of the positives predicted.
- **Sensitivity (Recall):** True Positive / (True Positive + False Negative), which measures what percentage are predicted positive of all the positive cases.
- **Specificity:** True Negative / (True Negative + False Positive), which measures how well the model is at predicting negative results.
- **F-score:** 2 * ((Precision * Sensitivity) / (Precision + Sensitivity)), which is the "harmonic mean" of precision and sensitivity, considering both false positive and false negative cases and is good for imbalanced datasets.

The premise for calculating these indicators: Target is multiclass so we choose another average setting when calculating these indicators: macro. Macro calculates the indicators for each label and finds their unweighted average[10]. This does not take into account label imbalance. That is, we first find the precision of each class separately and then its arithmetic average.

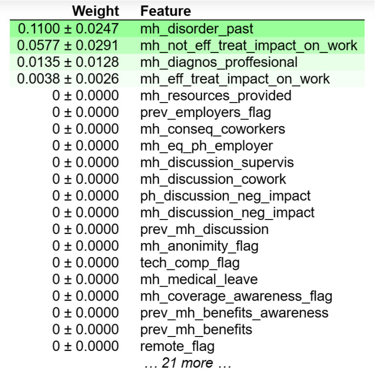
From the table above, we can see that the Cross Gradient Booster (Random Forest) machine learning model for this dataset performs the best among the models mentioned in the table above in terms of the indicators Accuracy, Precision, Sensitivity, Specificity and F1 Score, with values of 0.78706, 0.71829, 0.74116, 0.74116, and 0.74915 respectively.

In machine learning models, the parameters that need to be selected manually are called hyperparameters. And when selecting hyperparameters, Grid Search can be used: a circular traversal through all candidate parameter choices, trying each possibility, with the best-performing parameter being the final result. The Grid Search algorithm uses each set of hyperparameters to train the model and selects the combination of hyperparameters with the lowest validation set error[11]. With our previously chosen Cross Gradient Booster (Random Forest) model, the best parameters were selected using Grid Search tuning to obtain the best learning rate of 0.1 and the final best model with an accuracy of 0.79784.

IV. APPLICATION OF INTERPRETABLE MACHINE LEARNING: FEATURE IMPORTANCE

A. *Permutation Importance*

There are several ways to measure the importance of a feature, one of which is Permutation Importance (PI), where the principle is that after the model has been estimated, if the data in one column of the Validation data is randomly swapped and the other columns are held constant, see how the final prediction changes[12]. If, for a particular column, the predicted outcome changes dramatically, then the data for that particular column is very important. Conversely, if the prediction does not change much, then this column is less important. We can use the ELI5 library in Python to calculate permutation importance.



Weight	Feature
0.1100 ± 0.0247	mh_disorder_past
0.0577 ± 0.0291	mh_not_eff_treat_impact_on_work
0.0135 ± 0.0128	mh_diagnos_proffesional
0.0038 ± 0.0026	mh_eff_treat_impact_on_work
0 ± 0.0000	mh_resources_provided
0 ± 0.0000	prev_employers_flag
0 ± 0.0000	mh_conseq_coworkers
0 ± 0.0000	mh_eq_ph_employer
0 ± 0.0000	mh_discussion_supervis
0 ± 0.0000	mh_discussion_cowork
0 ± 0.0000	ph_discussion_neg_impact
0 ± 0.0000	mh_discussion_neg_impact
0 ± 0.0000	prev_mh_discussion
0 ± 0.0000	mh_anonimity_flag
0 ± 0.0000	tech_comp_flag
0 ± 0.0000	mh_medical_leave
0 ± 0.0000	mh_coverage_awareness_flag
0 ± 0.0000	prev_mh_benefits_awareness
0 ± 0.0000	prev_mh_benefits
0 ± 0.0000	remote_flag
	… 21 more …

Figure 2. PERMUTATION IMPORTANCE

At the top of the list, the most important characteristic is whether or not the person has suffered from a mental health disorder in the past. The next second, third and fourth most important characteristics are: whether it would interfere with work if treatment is not effective, whether it has been professionally diagnosed, and whether it would interfere with work if treatment is effective. When a replacement is done for a particular column, it is usually done several times. The plus and minus signs in the weight indicate the magnitude of the accuracy of each replacement change.

B. SHAP Value

In addition to understanding which features have the greatest impact and how much each feature affects the final result. Sometimes, we also want to know how the different features combine to produce the final prediction. For example, in the OSMI 2016 dataset, the final predicted outcome for a person is "currently suffering from a mental health disorder", so there may be many features that influence this outcome, so how do these features combine to influence the final predicted outcome. SHAP(SHapley Additive exPlanations) addresses this problem. It is based on the effect of a given feature, at a given value, compared to a baseline value[13]. The same applies to the model results of the PI example.

C. Interpretability of model predictions

1) Explanatory visualisation of individual forecasts

The SHAP force plot provides interpretability of single model predictions and can be used for error analysis to find explanations for instance-specific predictions. Each observation has its own plot of forces[14]. We can analyse a single observation.

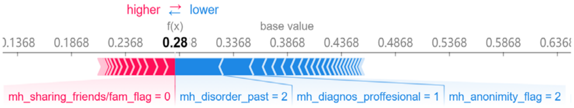


Figure 3. SHAP FORCE PLOT

As shown in the figure above, we look at the prediction results for the first observation.The model output value is 0.28. The longer the arrow, the greater the effect of the feature on the output. The amount of reduction or increase in impact can be seen by the value of the scale on the x-axis.The predictions for this observation are a little higher compared to the baseline value. “mh_sharing_friends/fam_flag” pushes these predictions up.“mh_disorder_past”,“mh_diagnos_professional”,“mh_anonimity_flag” reduce the predictions.

2) Explanatory visualisation of multiple predictions

By combining the plots of all the forces, rotating them by 90 degrees and stacking them horizontally, we will obtain a plot of the forces for the entire data set[15]. This is the plot for the whole data set:

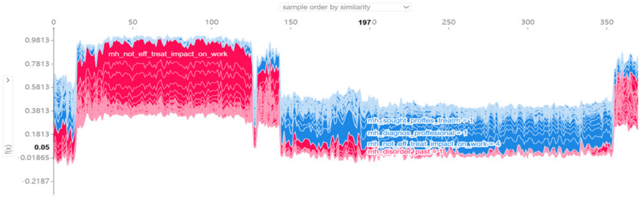


Figure 4. SHAP FORCE PLOT FOR THE WHOLE DATA SET

The top and left tabs in the figure above allow you to arbitrarily select the effect of multiple samples of a single variable on the model output results.

D. SHAP Summary Plots

While Permutation Importance describes the importance of features, the shap.summary_plot provided in SHAP values also provides information on the importance of features and is more detailed[16].

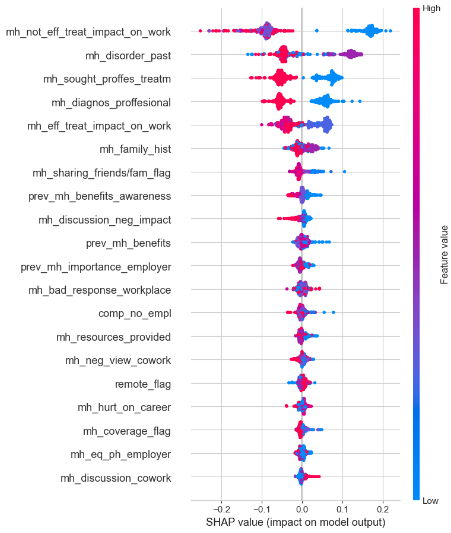


Figure 5. SHAP SUMMARY PLOTS

In this diagram, each point represents an observation. The diagram conveys the following information[12]:

- **Feature importance**: the variables are ranked in order of importance from highest to lowest;
- **Impact**: the horizontal axis indicates whether the variable has a positive or negative impact on the predicted outcome;

- **Raw values**: the shade of colour of each point, indicating the size of the value taken by that observation for this feature;
- **Correlation**: provides information on the correlation between the variable and the predicted outcome. For example, the higher the value of "mh_diagnos_professional", the more negative the predicted outcome is; "mh_discussion_cowork", the higher the value, the more positive the predicted outcome is.

V. Conclusion

In this paper, we found that based on the research survey dataset in this paper, using Naive Bayes, KNN, Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, Neural Network, Stochastic Gradient Descent, Cross Gradient Booster, Cross Gradient Booster (Random Forest) machine learning models, we found that the Cross Gradient Booster (Random Forest) model was the most effective and best performing model, with metric accuracy, precision, sensitivity, specificity and F1 scores of 0.78706, 0.71829, 0.74116, 0.74116, and 0.74915 respectively. Based on this, we used a Grid Search algorithm to determine the optimal model parameters. With our previously selected Cross Gradient Booster (Random Forest) model, the best parameters were selected using grid search adjustments to obtain the best learning rate of 0.1 and the final best model with an accuracy of 0.79784.

More generally, we analysed feature importance, using an analytical method that is used in the field of interpretable machine learning: Permutation importance. We found that the most important feature was whether the person had suffered from a mental health disorder in the past. The next second, third and fourth most important features were whether the treatment would affect work if it was ineffective, whether it had been professionally diagnosed, and whether it would affect work if it was effective. Based on the SHAP values, we used the Shap Force Plots and separate explanatory visualisations for single and multiple predictions, and also used the shap sammary plot to show more content and information about the importance of the features.

Therefore, the model results and the visualisation of the importance of the characteristics presented in this paper will be useful to the clinical and commercial community in screening and identifying people at risk of mental health disorders and those at risk of developing them, in determining the importance of any factors that affect people's mental health, and in providing measures and interventions for mental health problems in the population based on the information held, which will have far-reaching implications and reference value for the mental health field.

References

[1] Pan American Health Organisation, "Mental Health - PAHO/WHO | Pan American Health Organization," www.paho.org, 2022. https://www.paho.org/en/topics/mental-health

[2] World Health Organization, "Mental Health," WHO, 2022. https://www.who.int/health-topics/mental-health#tab=tab_1

[3] A. Gold and D. Gross, "Deploying machine learning to improve mental health," MIT News | Massachusetts Institute of Technology, Jan. 26, 2022. https://news.mit.edu/2022/deploying-machine-learning-improve-mental-health-rosalind-picard-0126

[4] K. M. Mitravinda, D. S. Nair, and G. Srinivasa, "Mental Health in Tech: Analysis of Workplace Risk Factors and Impact of COVID-19," SN Computer Science, vol. 4, no. 2, Feb. 2023, doi: https://doi.org/10.1007/s42979-022-01613-z.

[5] "Research :: Open Sourcing Mental Health - Changing how we talk about mental health in the tech community - Stronger Than Fear," osmhhelp.org. https://osmhhelp.org/research.html.

[6] "Model and Visualize Mental Health in Tech," kaggle.com. https://www.kaggle.com/code/andradaolteanu/model-and-visualize-mental-health-in-tech.

[7] A. Biswal, "What is Exploratory Data Analysis? Steps and Market Analysis | Simplilearn," Simplilearn.com, Feb. 17, 2023. https://www.simplilearn.com/tutorials/data-analytics-tutorial/exploratory-data-analysis

[8] J. Zhang, "Top 5 common machine learning model types in general,"blog.csdn.net, Mar. 29, 2022. https://blog.csdn.net/junhongzhang/article/details/123813842?utm_medium=distribute.pc_relevant.none-task-blog-2~default~baidujs_baidulandingword~default-0-123813842-blog-124143500.235

[9] "Confusion Matrix in Machine Learning - Javatpoint," www.javatpoint.com. https://www.javatpoint.com/confusion-matrix-in-machine-learning

[10] S. Saxena, "Multi-class Model Evaluation with Confusion Matrix and Classification Report," Medium, Sep. 30, 2022. https://pub.towardsai.net/multi-class-model-evaluation-with-confusion-matrix-and-classification-report-c92a74d5e908.

[11] H. Mujtaba, "An Introduction to Grid Search CV | What is Grid Search," GreatLearning, Sep. 29, 2020. https://www.mygreatlearning.com/blog/gridsearchcv/

[12] P. Pandey, "Interpretable Machine Learning," Medium, Apr. 10, 2019. https://towardsdatascience.com/interpretable-machine-learning-1dec0f2f3e6b

[13] V. Trevisan, "Using SHAP Values to Explain How Your Machine Learning Model Works," Medium, Jul. 05, 2022. https://towardsdatascience.com/using-shap-values-to-explain-how-your-machine-learning-model-works-732b3f40e137

[14] C. K. Dataman, "Explain Your Model with the SHAP Values," Dataman in AI, Aug. 11, 2022. https://medium.com/dataman-in-ai/explain-your-model-with-the-shap-values-bc36aac4de3d

[15] C. K. Dataman, "Explain Any Models with the SHAP Values — Use the KernelExplainer," Medium, Nov. 24, 2021. https://towardsdatascience.com/explain-any-models-with-the-shap-values-use-the-kernelexplainer-79de9464897a

[16] L. Monigatti, "How to Easily Customize SHAP Plots in Python," Medium, Oct. 04, 2022. https://towardsdatascience.com/how-to-easily-customize-shap-plots-in-python-fdff9c0483f2.