2767 - Machine Learning - Group Project

Predict Anxiety Attacks: An Insurer's Perspective

Group W
Marco Piccolo - 63996
Marvin Schumann - 63529
Philipp Goetting - 64737
Julian Fuchs - 63539
Giorgia Glorio - 65948
Maria Teresa Daffan - 66143

Table of Contents

1. Business Problem Definition

Health insurance companies are experiencing increasing costs due to rising claims related to mental health disorders, particularly anxiety disorders. Anxiety attacks often result in emergency room visits, therapy, medications, and even long-term disability claims. These costs put financial strain on both the insurance provider and the insured individuals. By leveraging machine learning models to analyze client health data, insurers can predict the severity of anxiety attacks in patients that have shown indicators of anxiety. This predictive capability shifts the focus from costly interventions to preventive care, reducing long-term healthcare expenses and improving client well-being. By assessing the severity of an individual's panic attack, the insurance company can take proactive measures to prevent it, thereby mitigating the associated costs and challenges. This creates a win-win model, clients receive support to improve their mental health, while insurers mitigate financial risks associated with anxiety-related claims.

Drawing from a dataset of patients with a history of anxiety, this report outlines the development of a machine learning model that predicts the severity of an anxiety attack based on daily habits and physiological indicators. The initial section will conduct an exploratory analysis of the data. Following this, three models will be introduced a Logistic Regression, a Random Forest, and an XGBoost and they will be evaluated and compared. Lastly, the report will provide the business implications and recommendations.

2. Exploratory Data Analysis

In line with the business goal of identifying customers at risk of severe anxiety attacks, we categorized the severity variable into low, medium, and high. This simplification enhances clarity and helps prioritize high-severity cases for targeted interventions.

To develop a foundational understanding of the dataset, we began by performing a univariate analysis across all features, followed by bivariate and multivariate analyses to investigate correlations among the variables. The dataset contained no missing or duplicate values. Most boolean features were imbalanced, with the *No (0)* class being dominant. Behavioral patterns were also identified by analyzing features such as *Smoking*, *History of Anxiety* or *Life Events*, *Dizziness*, and *Medication*, which may potentially influence the target variable. Ordinal features showed notable variation, generally indicating high stress levels, while Diet suggested a more moderate balance. Continuous variables revealed important trends as well, with outliers identified beyond the third quartile range. Table 1 in the Appendix shows summary statistics for continuous features.

Moreover, key factors influencing anxiety severity were identified through correlation and statistical tests. *High respiratory* and *heart rates, family history*, *therapy sessions*, and *stress* were positively linked to anxiety, while *sleep*, *medication*, and *physical activity* showed

protective effects. ANOVA and chi-square tests confirmed significant variable differences across severity levels, supporting targeted interventions based on physiological and behavioral patterns. Refer to the attached notebook to have a better overview of the EDA and how specific variables impact the severity of an anxiety attack.

In summary, the Exploratory Data Analysis highlighted important variables that may influence anxiety severity. Features such as *Stress Level*, *Sleep Hours*, *Caffeine Intake*, *Dizziness*, and *Family History* stood out due to their distributions and plausible predictors of anxiety.

3. Models Development

As aforementioned, the original anxiety severity scale (1-10) was converted into three categories low, medium, and high-to simplify the prediction task and facilitate clearer interpretation and actionable results. The "high" severity category was intentionally designed to be broader since accurately identifying severe anxiety attacks carries greater practical and clinical importance. Following the same rationale, hyperparameter tuning was conducted optimizing for weighted recall rather than simple accuracy. This approach was chosen specifically to prioritize accurately identifying medium and high severity anxiety cases, as failing to detect these more serious cases poses greater risks than misclassifying lower-risk instances. By optimizing weighted recall, the models emphasize minimizing false negatives for critical severity categories, ensuring more individuals who need intervention receive timely support. Moreover, considering feature selection, *ID* and *Heart Rate during panic attack* were not used as predictors. The first is intuitively irrelevant, and the latter reflects a physiological response measured during an anxiety attack. Since this information is not available beforehand, it would not be useful for predicting future attacks and could lead to misleading results if included.

Hence, considering that our task involves multiclass classification, we developed three models - Logistic Regression, Random Forest, and XGBoost - to capture different aspects of the problem. Together, these models allow us to balance interpretability and predictive power, ultimately guiding us in selecting the most effective approach for identifying customers at risk of low, medium and intense anxiety attacks.

For a more detailed understanding of the models, feature importance and model explainability refer to the attached notebook.

3.1 Logistic Regression

Logistic Regression, used as a simple and interpretable baseline model, was developed by first applying feature engineering to encode binary indicators, generate occupation dummies, and remove irrelevant columns. Then, a pipeline applies these transformations and scales numerical values before fitting a multinomial logistic regression model, which leverages

the Softmax function to handle multiple classes. A grid search systematically explores different hyperparameter values such as regularization strength (C) and maximum iterations to identify the best-performing model through cross-validation. Finally, a Logit regression was performed to understand the linear impact of each feature through coefficients and p-values. However, Logistic Regression only captures linear relationships, hence, some features like smoking and dizziness might not appear significant even though they could be valuable when modelling more complex interactions. Therefore, we include all features in the following models, which can capture these non-linear relationships and potentially improve overall performance.

3.2 Random Forest

Random forest excels in multiclass problems because it naturally captures complex non-linear interactions and aggregates predictions from multiple trees, thereby reducing variance and overfitting. Similarly to the previous model, the process begins with feature engineering. As aforementioned, no additional features were dropped compared to the Logistic Regression. Moreover, to address class imbalance, SMOTE is applied, ensuring that minority classes receive adequate representation during training. Finally, a grid search tunes the Random Forest hyperparameters such as the number of estimators, maximum depth, and minimum samples per split to optimize weighted recall. Feature importance shows that Breathing Rate has the greatest influence on the model's predictions, followed by Physical Activity, Sleep Hours, and Caffeine Intake, indicating that physiological and lifestyle factors play a crucial role in determining the attack severity. Occupation-related features appear less influential; hence, the model finds them less predictive.

3.3 XGBoost

The model training process begins with fitting an initial XGBoost classifier, using multi-class log loss as the evaluation metric. XGBoost was chosen for its high performance with structured data, ability to handle imbalanced classes, and built-in feature selection. Its gradient boosting framework makes it particularly effective for capturing complex relationships in healthcare data while maintaining computational efficiency. Feature importance is then analyzed, and low-importance features are removed based on a predefined threshold of 0.03 to improve model interpretability and efficiency. Additionally, grid search with cross-validation is performed to optimize hyperparameters, testing multiple combinations of tree depth, learning rate, and sampling strategies. Finally, the model's feature importance shows that sleep hours, physical activity, and caffeine intake strongly predict anxiety severity, highlighting the importance of behavioural and lifestyle factors compared to immediate physiological indicators like breathing rate and sweating.

4. Models Evaluation

Table 2 – Models Performance Comparison

	Logistic Regression			Random Forest			XGBoost		
	L	M	H	L	M	H	L	M	H
Accuracy		69%			67%			70%	
Precision	0.73	0.56	0.45	0.82	0.52	0.37	0.78	0.58	0.42
Recall	0.91	0.34	0.24	0.76	0.44	0.65	0.86	0.40	0.47
F-1 Score	0.81	0.42	0.32	0.79	0.48	0.47	0.82	0.47	0.44
AUC	0.78	0.75	0.78	0.82	0.76	0.87	0.85	0.80	0.88

The logistic regression model achieves an overall accuracy of approximately 69%, with strong performance on the “Low” class but weaker precision and recall for “High” and “Medium.” While the model is effective at identifying “Low” cases, it struggles to detect “High” and “Medium” severity levels accurately. The Random Forest model achieves a similar overall accuracy of 67%, performing better on the “Low” class with an 0.82 precision. Notably, it improves recall for the “Medium” category to 0.44 (vs. 0.34), and for “High” category to 0.65 (vs. 0.24), meaning it captures more individuals in critical groups than the earlier model. Although precision for “High” remains relatively low (0.37), the better recall indicates a greater ability to correctly identify severe cases, which is critical in the context of this business problem as the company aims to identify and help those with more severe anxiety attacks. On the other hand, the XGBoost model yields a 70% accuracy and effectively identifies low-risk cases. However, its lower recall and F1 scores for Medium and High severity classes indicate it produces more false negatives, meaning it incorrectly classifies some severe cases as less severe. In this context, false negatives are particularly problematic, as failing to identify a severe anxiety case could lead to a lack of necessary intervention. Additionally, although the XGBoost model has a marginally higher AUC for all three classes, indicating good discriminatory ability, the model misses more severe cases than the Random Forest. Consequently, from a practical standpoint, it is preferable to accept more false positives, because incorrectly classifying milder cases as severe ensures that truly severe cases are not missed. Considering this trade-off, the preferred model to predict the level of anxiety attack of a person is the Random Forest. The pickle file to run the model on future clients will be attached to this document with the necessary additions.

5. Findings

SHAP analyses of both the XGBoost and Random Forest models consistently highlight breathing rate, stress levels, family history, physical activity, medication use, and sleep quality as influential predictors of anxiety severity. The strong presence of these predictors across

both models reinforces their reliability as key factors insurers should focus on to proactively manage anxiety risks. Although the exact ranking of features varies slightly between models, insurance companies can leverage these consistent insights to design effective interventions.

5.1 Business Implications

The predictive model benefits insurance companies by enabling early identification of individuals at risk for severe anxiety attacks. By accurately assessing the severity of potential anxiety episodes among clients who already exhibit anxiety indicators, insurance providers can adopt a proactive stance. This preventative approach reduces the frequency and severity of anxiety-related emergencies, ultimately decreasing healthcare expenditures associated with emergency room visits, prolonged therapy, medication use, and long-term disability claims. In addition to financial benefits, this model supports clients by promoting improved mental health outcomes through tailored preventive measures. Additionally, according to the SHAP analysis, several key factors significantly influence anxiety severity, highlighting clear opportunities for actionable interventions.

5.2 Actionable Recommendations

Breathing rate, stress levels, and family history emerged as influential predictors of anxiety severity, suggesting insurers can proactively address risk of medium and high severe anxiety attacks by offering specialized programs. Interventions could include mindfulness and breathing techniques supported by wearable devices, personalized counseling for individuals with a family history of anxiety, and resilience training to build emotional strength. Additionally, insurers should encourage physical activity through incentives like subsidized gym memberships and emphasize improved sleep through educational initiatives and digital health tools to effectively lower anxiety severity. Other predictive indicators include sweating levels, therapy session frequency, caffeine and alcohol consumption, and medication use. Educating clients on managing caffeine and alcohol consumption, coupled with medication reviews and targeted counseling, will further mitigate anxiety-related risks, enhancing client well-being while reducing insurance costs.

Although the identified predictors impact low, medium, and high severity anxiety categories somewhat uniformly, insurance companies can still adopt tailored strategies based on severity. Clients displaying low-severity anxiety indicators might primarily benefit from general wellness initiatives, while those with medium-severity symptoms could receive targeted coaching and structured mental health programs. High-severity cases require intensive intervention such as frequent therapy, specialized medical support, and dedicated resources to manage critical stressors effectively.

6. Appendix

Table 1 – Summary Statistics Continuous Features

	Min	Max	Mean	Median	Std.	Skew	Kurt.
Sleep_Hours	3.0	12.0	6.5	6.5	1.9	0.2	-0.5
Physical_Activity_(hrs/week)	0.0	30.0	5.1	4.2	3.9	1.9	5.9
Caffeine_Intake_(mg/day)	0.0	800.0	257.3	188.0	213.7	1.2	0.6
Heart_Rate_(bpm_during_attack)	60.0	179.0	93.3	91.0	20.0	1.1	2.3
Breathing_Rate_(breaths/min)	11.0	39.0	17.1	17.0	4.3	2.1	7.0