Predicting Philippine Undergraduate Employability in Mock Interviews using XGBoost, CatBoost, and LightGBM

Gary Louis P. Garcia
South Philippine Adventist College
Matanao, Davao del Sur, Philippines
garylouis.p.garcia@gmail.com

Jane T. Antoque
South Philippine Adventist College
Matanao, Davao del Sur, Philippines
janeantoque7@gmail.com

ABSTRACT

Employability encompasses well-rounded talents beyond technical skills. Mock interviews offer insights into employability, and we propose a predictive model using mock interview ratings and gradient boosting algorithms such as LightGBM, XGBoost, and CatBoost. We employed a rigorous validation process, using 5-times repeated 10-fold cross-validation on 70% of the dataset and a separate 30% for testing. After hyperparameter optimization, LightGBM and XGBoost attained an accuracy of 91.5%. All models demonstrated a precision of 93.7%. Both LightGBM and XGBoost achieved a recall rate of 91.2%. Notably, XGBoost exhibited the highest AUC of 98.1%. Feature importance analysis reveals a combination of key factors enhancing employability including cognitive abilities, physical presentation, communication skills, practical experience, and self-confidence.

KEYWORDS

Undergraduate, employability, interview, LightGBM, XGBoost, CatBoost

1 INTRODUCTION

Employability is the ability of an individual to secure the right job that matches their education [8]. Despite the importance of technical skills and experience, true employability goes beyond by cultivating a well-rounded set of talents and achievements that make graduates stand out. Not only does this increase your chances of landing a good job, but it also paves the way for long-term fulfillment and success, benefiting not only yourself, but also your future employers, the local community, and even the national economy [12].

Mock interviews are simulated job interviews that provide individuals the opportunity to practice answering common interview questions and gain experience interacting with potential employers in a formal setting [7].

Recent studies have been conducted to predict the employability of undergraduate students using mock interview ratings. A study by Casuat and Festijo in 2019 [2] predicted the overall employability of undergraduate students where classifiers employed are decision tree (DT), random forest (RF), and support vector machines (SVM). Among the three classifiers, SVM had the highest score in all classification metrics, namely accuracy, recall, precision, and F1-score, of 91 2%. In 2020, they have extended their study [3]. They identified the most predictive attributes among the employability signals of undergraduate students using the scores generated by three feature reduction techniques with SVM with SMOTE such as recursive feature elimination (RFE), univariate selection (US), and principal component analysis (PCA).

In the present effort, we utilized a dataset of mock interview ratings collected from the Kaggle website for undergraduate students across various disciplines. We employed gradient boosting algorithms, namely XGBoost, LightGBM, and CatBoost, to build predictive models of student employability based on mock interview performance. Our analysis also includes key findings about employability indicators, which confirm some, but also provide novel insights compared to previous studies [2, 4].

The general objective of this study is to predict the employability of undergraduate students based on mock interview ratings using LightGBM, XGBoost, and CatBoost. Specific objectives of this study are as follows.

- Evaluate the accuracy, precision, recall, and f1-score of LightGBM, XGBoost, and CatBoost in predicting the employability of undergraduate students.
- Compare the performance of LightGBM, XGBoost and CatBoost in predicting student employability, measured by the key metric categories of accuracy, precision, recall, f1-score, and area under the curve (AUC).
- Generate feature importance plot for LightGBM, XGBoost, and CatBoost to identify the most predictive features for undergraduate students' employability.

The remainder of this paper is organized as follows. Section 2, the Methodology, presents the experimental setup. Section 3, the Results and Discussions presents our findings and their interpretations. Finally, Section 4, Recommendations and Future Works summarize our key insights and pave the way for future research directions.

2 METHODOLOGY

2.1 Dataset Collection

This study uses the publicly available data set titled 'Students' Employability Dataset-Philippines' hosted on Kaggle https://www.kaggle.com/datasets/anashamoutni/students-employability-dataset. The dataset comprises mock job interview results for 2,982 students from various university agencies across the Philippines. It adheres to the Data Privacy Act, ensuring participant anonymity and confidentiality. However, biases inherent in mock interview evaluations may limit its representativeness of the entire student population. Table 1 shows included features and descriptions, based on Casuat and Festijo's work [3], except for 'Self-confidence,' which is described by the authors of this study.

Table 1: Students' Employability Dataset

Feature Name	Description
General Appearance	The way a person looks in general
Manner of Speaking	Appropriate style of expressing oneself
Physical Condition	The condition or state of the body
Mental Alertness	The state of active attention of the mind
Self-confidence	Manifests as poise, conviction, and clear, persuasive communication.
Ability to Present Ideas	The ability to present the ideas clearly
Communication Skills	The ability to convey ideas to others effectively and efficiently
Internship Student Performance Rating	The performance assessment conducted by the immediate superior of OJT

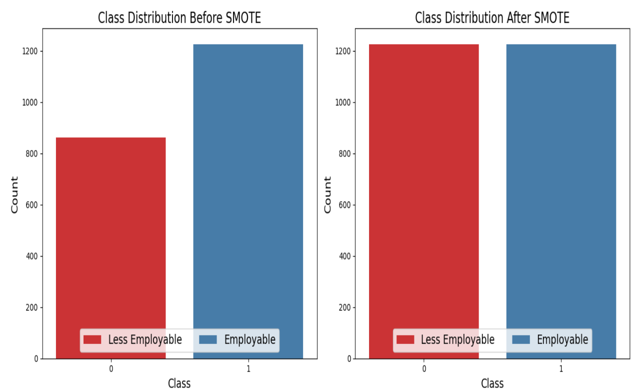

Figure 1: Class distribution before and after SMOTE.

2.2 Dataset Pre-Processing

The dataset contained no missing values. The 'Student number' feature was deemed irrelevant and removed. The 'Class' feature labels were transformed to numerical values: 'employable' as 1 and 'less employable' as 0 for machine learning algorithms. A 70/30 split was used to divide the dataset into training and testing sets, allowing for training three different algorithms while ensuring a fair evaluation on unseen data.

2.3 Handling Class Imbalance

The dataset had a class imbalance, with 42% labeled 'less employable' and 58% 'employable'. To address this without reducing overall observations, Synthetic Minority Oversampling Technique (SMOTE) was employed. SMOTE generated new data for 'less employable', balancing classes to a 50/50 ratio as shown in Figure 1. This mitigated bias towards the majority class, enhancing model performance.

2.4 Model Selection

Previous study [2] predicting student employability using similar features have primarily focused on none-boosting machine learning algorithms. This study explores the potential of boosting algorithms for this task. Three prominent boosting algorithms were chosen for analysis: LightGBM [6], XGBoost [5], and CatBoost [9]. This selection of models allows for a comprehensive evaluation of boosting algorithms compared to none-boosting methods for predicting student employability.

2.5 Hyperparameter Tuning

We have implemented grid search hyperparameter optimization to improve the performance of LightGBM, XGBoost and CatBoost models. Our goal was to find the best parameter values for each model and assess its impact on classification metrics including AUC, accuracy, precision, recall and F1 scores. The optimized hyperparameters for each model are shown in Table 2.

Table 2: LightGBM, XGBoost, and CatBoost Hyperparameters

Models	*Hyperparameter*	*Value*
LightGBM	subsample	0.8
	reg_lambda	0
	reg_alpha	0.5
	num_leaves	36
	n_estimators	200
	min_child_samples	20
	max_depth	7
	learning_rate	0.05
	colsample_bytree	1.0
XGBoost	max_depth	6
	alpha	1
	learning_rate	0.1
	n_estimators	500
	subsample	0.6
	colsample_bytree	0.8
	min_child_weight	1
	gamma	0.1
	reg_alpha	1
	reg_lambda	1
CatBoost	depth	6
	iterations	150
	l2_leaf_reg	3
	learning_rate	0.5

2.6 Model Evaluation

We employed repeated k-fold cross-validation by splitting the 70% training data into 10 folds. Each boosting model was trained on nine folds and evaluated on the held-out fold which is repeated five times for robust performance estimation. The remaining 30% served as the test set to gauge model generalization. Performance was assessed using the following performance metrics.

- **Accuracy:** It computes the ratio of correctly classified instances to the total number of instances [10].

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

- **Precision:** It is the ratio of true positive instances divided by the total number of instances predicted as positive [11].

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

- **Recall:** Given as the ratio of relevant instances that are recovered [11].

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

- **F1-Score:** It is the combination of both precision and recall used to get the average value of them [1].

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Additionally, ROC curves and associated AUC were used to evaluate the model's class discrimination ability across various thresholds, offering a comprehensive performance view.

2.7 Predictive Features Identification

To understand which factors or features drive decision-making among the models, we analyzed feature importance scores. Visualizing these scores as column charts helped us identify key predictors for each model. Comparing the results across models revealed features consistently highlighted as top predictors by multiple models.

3 RESULTS AND DISCUSSIONS

3.1 Evaluation and Comparison of Models' Performance

LightGBM, XGBoost, and CatBoost all have very similar performance metrics with respect to accuracy, precision, recall, and F1-score. Their accuracy rates are around 91.5%, 91.5%, and 91.4%, respectively, which indicates that the three models are highly effective in predicting the employability status of individuals, as shown in Table 3.

The precision of 93.7% for all models suggests that when they predict an individual as employable, there is a high chance of this prediction being accurate. This is crucial for avoiding false positives in employability predictions, where predicting someone as employable when they are not could lead to inefficient use of resources.

Recall values are slightly different, with LightGBM and XGBoost at 91.2% and CatBoost at 91.0%, showing that LightGBM and XGBoost are marginally better at identifying all actual employable cases. However, this difference is minimal. The F1-scores are also very close, with LightGBM and XGBoost at 92.4% and CatBoost at 92.3%, indicating a balanced performance between precision and recall across the models.

The AUC (Area Under the ROC Curve) values are impressive across all models, ranging from 0.979 to 0.981. This suggests that the three models have excellent capability in distinguishing between the employable and less employable individuals. The slight differences in AUC values (LightGBM at 0.979, CatBoost at 0.980, and XGBoost at 0.981) may not be practically significant, considering the overall high performance.

Furthermore, in the previous study of Casuat and Festijo [2] , an SVM model was used, achieving a 91.22% accuracy rate. Although the experimental setups differ, the current models' performance remains comparable and on par with the SVM's performance. The models' performance are collectively shown in Table 3.

Table 3: Models' Performance Comparison

Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*AUC*
SVM[2]	0.9122	0.9115	0.910	0.910	–
LightGBM	0.915	0.937	0.912	0.924	0.979
XGBoost	0.915	0.937	0.912	0.924	0.981
CatBoost	0.914	0.937	0.910	0.923	0.980

3.2 Predictive Features Analysis

In the quest to better understand factors influencing student employability, the feature importance rankings provide insights into

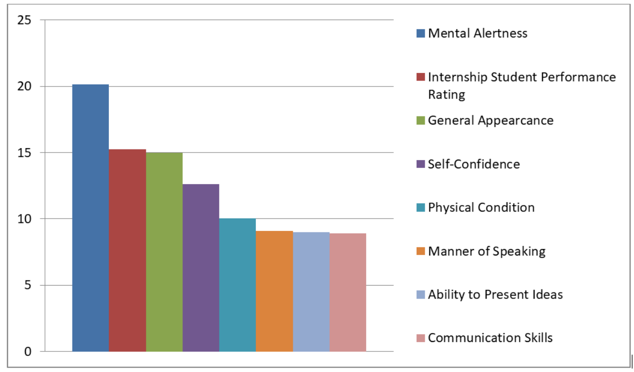

Figure 2: CatBoost Feature Importance Chart

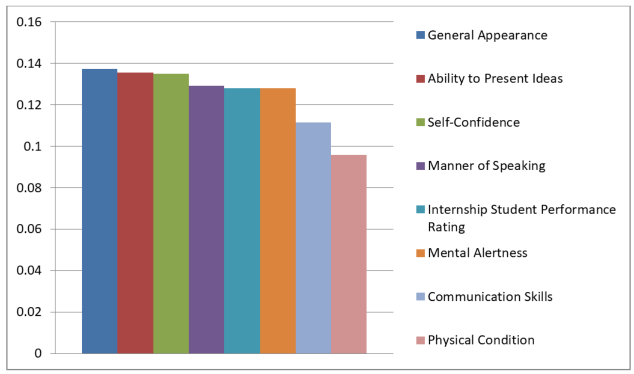

Figure 3: XGBoost Feature Importance Chart

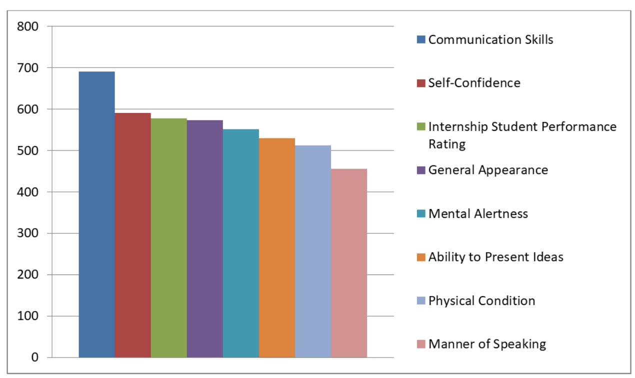

Figure 4: LightGBM Feature Importance Chart

the aspects that each machine learning model (i.e., LightGBM, XGBoost, and CatBoost) considers most influential for predicting a student's likelihood of securing employment, in the context of job interviews.

In the CatBoost model, as shown in Figure 2, Mental Alertness is identified as the most crucial feature according to CatBoost. This suggests that the model places high importance on cognitive abilities, attentiveness, and quick thinking. The internship student performance rating is also important, indicating that the model considers past OJT performance as a significant factor in predicting employability. General Appearance and Self-Confidence are also ranked high, suggesting that the model gives importance to how students present themselves, both physically and in terms of confidence. Soft skills such as communication skills and the ability to present ideas are also considered important, but are ranked lower than the aforementioned factors.

On the other hand, XGBoost, as shown in Figure 3, places the highest importance on General Appearance, suggesting that the overall appearance of a student plays a significant role in predicting employability. The ability to Present Ideas and Self-Confidence follow closely in importance, indicating that the model values students' abilities to articulate and express themselves. Mental Alertness and Student Performance Ratings are also considered important but are ranked slightly lower. Communication Skills and Physical Condition are relatively lower in importance according to XGBoost.

Furthermore, LightGBM, as shown in Figure 4, places the highest importance on Communication Skills, suggesting that effective communication is a key factor in predicting student employability. Self-confidence and Internship Student Performance Ratings follow closely, indicating the significance of confidence and OJT performance. General Appearance and Mental Alertness are also considered important. The ability to Present Ideas and Physical Condition are ranked lower in importance according to LightGBM.

Comparing across models, we observe that Mental Alertness, General Appearance, and Communication skills appeared as top-ranked qualities. Moreover, employability qualities such as Self-Confidence, Internship Performance Rating, and General Appearance are common in at least two models. This suggests their potential universal importance in employability prediction. Additionally, each model prioritizes unique features. CatBoost focuses on cognitive abilities, XGBoost emphasizes physical presentation and communication skills, while LightGBM highlights communication skills and confidence.

In the previous study by Casuat and Festijo [3] , the identified key predictors of student employability are Manner of Speaking, Mental Alertness, and Ability to Present Ideas. We can observe an overlap between the current study and previous work regarding Mental Alertness being a crucial factor in both. This reinforces its potential universal importance in employability prediction. Interestingly, both studies also identify communication skills as significant, although emphasized differently by the chosen models in this study and directly mentioned as a key predictor in the previous work. Exploring further, the current study identifies a broader range of influential features, including General Appearance, Communication Skills, and Self-Confidence, which are not explicitly mentioned in the previous work.

In light of this, we can emphasize the qualities undergraduates should need to be employable such as Mental Alertness, Communication Skills, General Appearance, Self-Confidence, Internship Student Performance Rating, and Ability to Present Ideas as shown in Figure 5. This means that a combination of cognitive abilities, physical presentation, communication skills, practical experience, and self-confidence are key determinants of student employability, reflecting the multifaceted nature of readiness for the workplace. Candidates who possess and demonstrate these qualities are more likely to stand out to employers and succeed in securing employment opportunities.

Figure 5: Employability Qualities

4 RECOMMENDATIONS AND FURTHER WORK

All models such as LightGBM, XGBoost, and CatBoost perform similarly well in predicting employability status, with high accuracy, precision, recall, and AUC values. Given their comparable performance, the choice between them can be based on factors like ease of implementation, computational efficiency, or specific project requirements. Feature importance may influence the choice of a model for the task at hand. Feature importance rankings can validate or challenge existing domain knowledge about crucial employability factors. For example, if educators prioritize fostering cognitive abilities, CatBoost's emphasis on Mental Alertness might make it a preferred choice for training programs. Further efforts can focus on continuously collecting data and refining models to better account for evolving job market dynamics and educational trends. These enhancements have the potential to improve the accuracy and relevance of employability predictions.

ACKNOWLEDGMENTS

We are grateful to Ms. Neila M. Paglinawan-Muñez for her insightful comments and helpful suggestions, which significantly improved this work.

REFERENCES

[1] A. Alhassan, B. Zafar, and A. Mueen. 2020. Predict students' academic performance based on their assessment grades and online activity data. *International Journal of Advanced Computer Science and Applications* 11, 4 (2020).
[2] C. D. Casuat and E. D. Festijo. 2019. Predicting students' employability using machine learning approach. In *2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*. 1–5.
[3] Cherry D Casuat and Enrique D Festijo. 2020. Identifying the most predictive attributes among employability signals of undergraduate students. In *2020 16th IEEE international colloquium on signal processing & its applications (CSPA)*. IEEE, 203–206.
[4] C. D. Casuat, E. D. Festijo, and A. S. Alon. 2020. Predicting students' employability using support vector machine: a smote-optimized machine learning system. *International Journal* 8, 5 (2020).
[5] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.
[6] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30 (2017).
[7] Rhiannon Lord, Ross Lorimer, John Babraj, and Ashley Richardson. 2019. The role of mock job interviews in enhancing sport students' employability skills: An

example from the UK. *Journal of Hospitality, Leisure, Sport & Tourism Education* 25 (2019), 100195.
[8] National Committee of Inquiry into Higher Education (Great Britain), Ron Dearing, and Sir Ron Garrick. 1997. *Higher education in the learning society.* The Committee.
[9] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems* 31 (2018).
[10] O. Saidani, L. J. Menzli, A. Ksibi, N. Alturki, and A. S. Alluhaidan. 2022. Predicting student employability through the internship context using gradient boosting models. *IEEE Access* 10 (2022), 46472–46489.
[11] P. Thakar and A. Mehta. 2017. Role of Secondary Attributes to Boost the Prediction Accuracy of Students Employability Via Data Mining. *arXiv preprint* (2017). arXiv:1708.02940
[12] Mantz Yorke. 2006. *Employability in higher education: what it is-what it is not.* Vol. 1. Higher Education Academy York.