Executive Summary

Overview

In this project, a machine learning model was developed to predict whether individuals would opt for insurance policies based on a variety of input features. The workflow encompassed multiple stages, including data preprocessing, feature selection, model training with an XGBoost classifier, and thorough evaluation of the model's performance. Key preprocessing techniques, such as target encoding for high-cardinality categorical variables and log transformation for skewed numerical features, were employed to optimize model accuracy. The model was evaluated using comprehensive performance metrics, including accuracy, precision, recall, F1-score, and ROC-AUC score, along with insights derived from confusion matrix.

Results

The XGBoost classifier achieved an accuracy of 63.1% (0.6306), indicating that approximately 63.1% of its predictions were correct. However, it exhibited challenges in effectively predicting the minority class (Class 1) due to the imbalanced data distribution.

- **Precision:** Class 0 (Not Taken) achieved a precision of 0.83, whereas Class 1 (Taken) reached 0.30, indicating limited reliability in predicting the minority class.
- **Recall:** Class 0 had a recall of 0.66, while Class 1 achieved 0.54, showing that the model identified over half of the positive cases but struggled with precision.
- **F1-Score:** The F1-score for Class 0 was 0.74, demonstrating a balanced performance between precision and recall. For Class 1, the F1-score was 0.39, reflecting the model's difficulty in balancing true positives and false positives for this class.

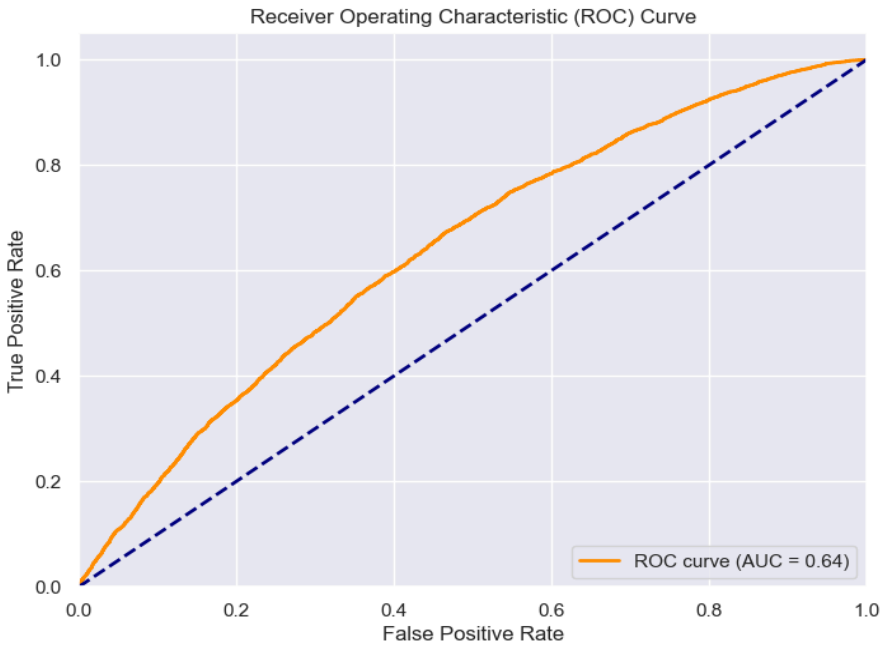
On an unseen test set uploaded to a leaderboard, the model produced the following results:

- Test Set Accuracy: 59%
- False Negative Rate: 40%
- False Positive Rate: 40%
- Advertising Revenue: 16 cents per person

The prediction distribution revealed that Class 0 accounted for 56.3% of predictions, while Class 1 represented 43.7%, suggesting a considerable number of individuals were predicted to opt for insurance policies.

ROC Curve Analysis

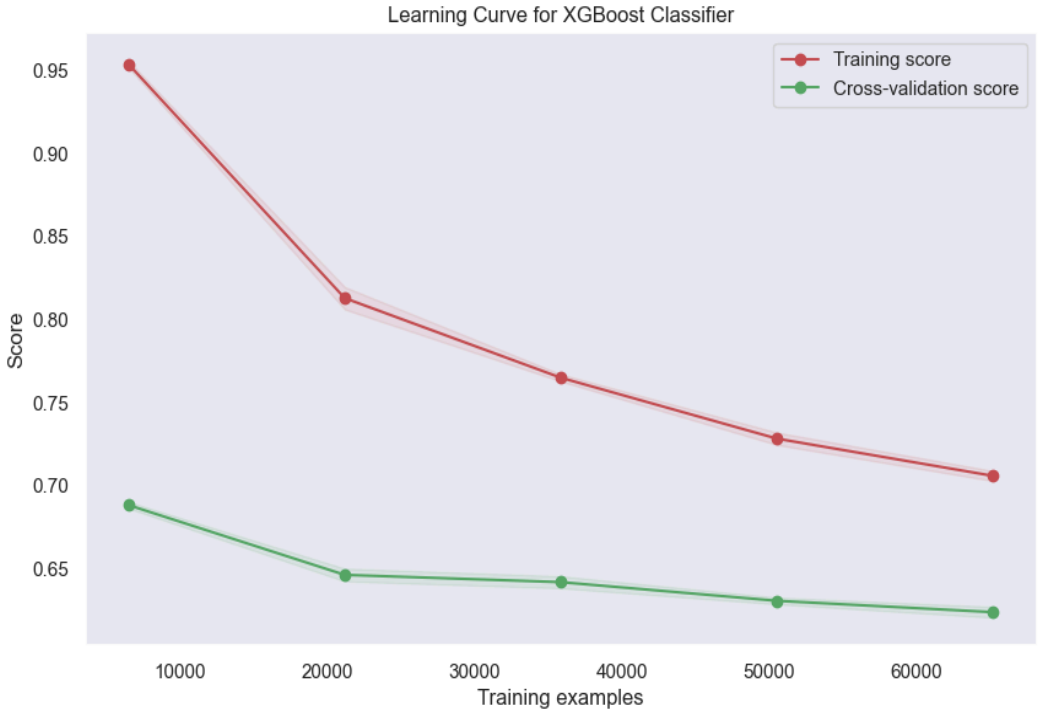
The ROC curve presented an Area Under the Curve (AUC) of 0.64, indicating that the model has moderate discriminative ability between the two classes. While the AUC is above random guessing (AUC = 0.5), it highlights room for



improvement in the model's capacity to distinguish between individuals likely to take out insurance policies and those who are not. The ROC curve reflects the model's ability to capture meaningful patterns in the data despite the challenges posed by class imbalance.

Underfitting and Overfitting Observations

The learning curve revealed that the model initially experienced overfitting, as evidenced by high training accuracy and low cross-validation accuracy when fewer training examples were used. Over time, as the training dataset size increased, the training accuracy declined, and overfitting was mitigated. However, the cross-validation accuracy plateaued at a relatively low value, suggesting potential underfitting. This plateau indicates that the model's capacity to generalize to unseen data is constrained, likely due to limitations in feature representation, hyperparameter tuning, or model complexity.



Final Model

The final model leveraged the XGBoost classifier, known for its robustness in capturing non-linear relationships and handling imbalanced datasets. The `scale_pos_weight` parameter was used to address class imbalance. The evaluation metrics—accuracy, ROC-AUC score, precision, recall, and F1-score—were systematically analyzed to validate the model's ability to generalize effectively.

In conclusion, while the XGBoost classifier demonstrated reasonable performance, the analysis underscored areas for improvement, particularly in enhancing precision and recall for the minority class. The model provides a solid foundation for understanding customer behavior and offers actionable insights for further optimization and refinement.