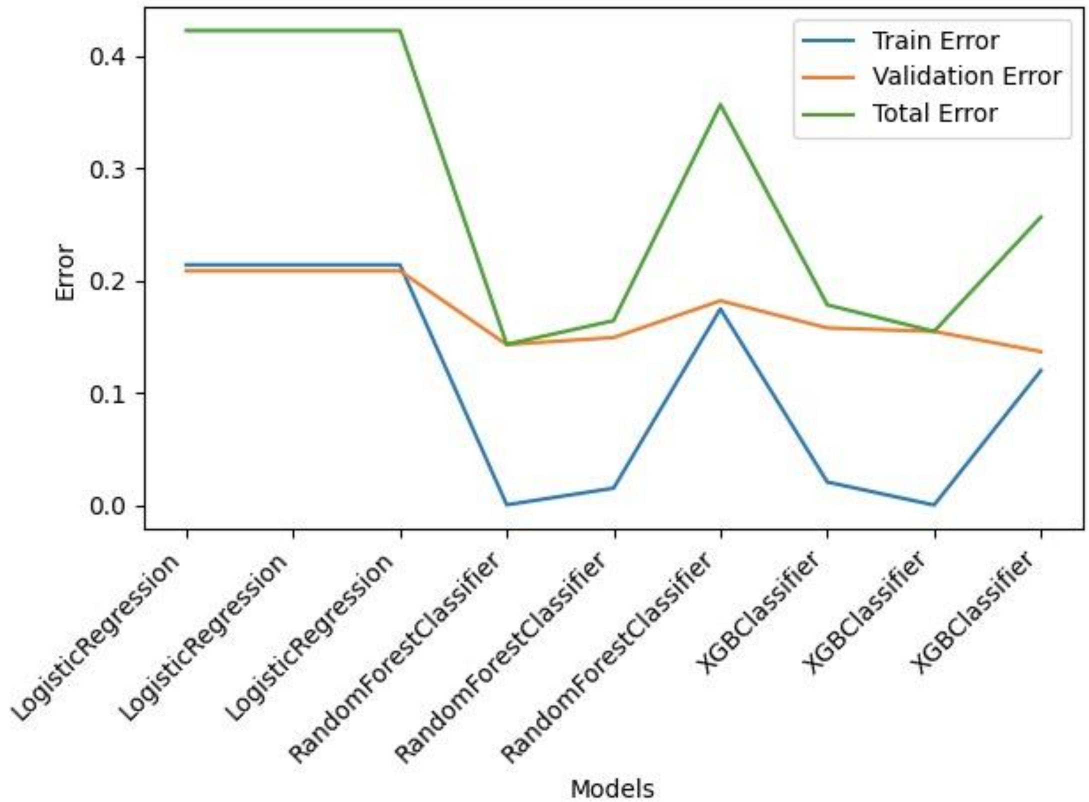
WEEK 9

This week, we evaluated the performance of nine models and selected the winning model based on validation error. Since it was a classification problem, we used 1-accuracy to calculate the validation error. Among the nine models, XGBoost variation 3 with the following parameters (learning rate = 0.1, n_estimator =140, max_depth = 3, min_child_weight=1, gamma=0, subsample=0.8) was identified as the winning model. XGboost is an ensemble method that combines multiple decision tress to enhance predictive performance. Although XGBoost is more complex than logistic regression and random forest, the winning model was not overly complex, with a relatively low depth (max_depth = 3) and some regularization (min_child_weight=1, gamma=0, subsample=0.8) to prevent overfitting.

```
LogisticRegression Validation Error: 0.2086
LogisticRegression Validation Error: 0.2086
LogisticRegression Validation Error: 0.2086
RandomForestClassifier Validation Error: 0.1430
RandomForestClassifier Validation Error: 0.1492
RandomForestClassifier Validation Error: 0.1820
XGBClassifier Validation Error: 0.1578
XGBClassifier Validation Error: 0.1547
XGBClassifier Validation Error: 0.1367
```

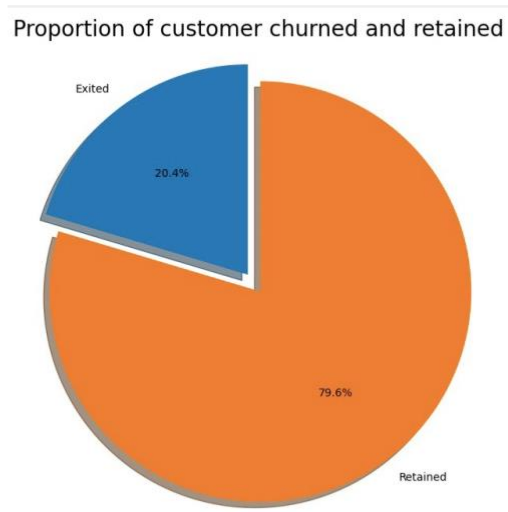
A model's error can be divided into three parts: bias, variance and irreducible error. Bias measures how far off in general the model's predictions are from the correct value. Variance is how much the predictions for a given point vary between realizations of the model. Irreducible error is the error that cannot be reduced by any model. A model with high bias tends to have training error (underfitting) and a model with high variance tends to have high validation error (overfitting). Therefore, it is crucial to find the optimal balance between bias and variance to avoid overfitting or underfitting the data.

The graph presented the training and validation errors of the models, allowing us to infer the bias-variance trade-off. Logistic regression had the highest validation error among the three models, but there was almost no variance for the validation error of the three variations of the logistic model. This high validation error might be due to the model's simplicity, resulting in high bias and underfitting the data. On the other hand, both random forest and XGBoost had lower validation error, indicating that they are capable of modeling the data better. However, the random forest and XGBoost models also showed some variance among their variations, which means that they might have overfitted the data to some extent.




After comparing the validation error of the random forest model variation 1 without tuning any hyperparameters and XGBoost with the following parameters (learning rate = 0.1, n_estimator =140, max_depth = 3, min_child_weight=1, gamma=0, subsample=0.8), we found that XGBoost model performed better due to its high recall. Recall is a critical metric in predicting banking customer churn as it measures the proportion of true positives out of all actual positives. We want a high recall to minimize the number of customers who churn without being identified by our model. We evaluated the performance of our model on test datasets using several metrics, including accuracy, precision, recall and F1 score. The accuracy of our model was over 86%, while the precision was about 75%. However, we obtained a low recall of only 47%. In terms of accuracy, our model is acceptable since 80% of customers stayed with the bank in our dataset, and if we made a random guess and guessed the larger class, we would only get an accuracy of 80%. Nevertheless, a low recall means that our model is not efficiently identifying all customers who churn. Therefore, we need to fine-tune our model further to increase its recall score and accurately identify as many potential churners as possible.

```
Random Forest:
Accuracy: 0.8645
Precision: 0.7563025210084033
Recall: 0.4580152671755725
F1 Score: 0.5705229793977814
XGBoost:
Accuracy: 0.8655
Precision: 0.748
Recall: 0.4758269720101781
F1 Score: 0.5816485225505444
```




We selected XGboost variation 3 as the winning model. The evaluation of our model on the training, validation, and test sets showed that our model's performance remained consistent across in terms of accuracy. However, precision and recall decreased in the test set, indicating that our model may not be capturing all the positive cases, potentially leading to missed potential churners. ROC AUC score, which measures the ability of our model to distinguish between positive and negative cases, also showed a decreasing trend.

In conclusion, XGBoost variation 3 was the winning model for our banking customer churn problem. While the model performed well in terms of accuracy, it had a lower recall score, indicating that it may not be effectively identifying all potential churners. We will need to fine-tune the model further to improve its recall score and ensure that we are capturing as many potential churners as possible. Overall, the XGBoost variation 3 model demonstrated a good balance between bias and variance and had a relatively low complexity, making it a suitable model for this classification problem.

Performance Metrics:

| Metric | Training Set | Validation Set | Test Set |
|---|---|---|---|
| Accuracy | 0.8855 | 0.8545 | 0.8655 |
| Precision | 0.8392 | 0.8248 | 0.7480 |
| Recall | 0.5252 | 0.4748 | 0.4758 |
| F1 Score | 0.6460 | 0.6027 | 0.5816 |
| ROC AUC Score | 0.7501 | 0.7221 | 0.7183 |