WEEK 6

For this week, we built a logistic regression model to predict the binary outcome of customer churn. Logistic regression is a popular model for binary classification as it estimates the probability of the outcome being one class versus the other class. The complexity of the logistic regression model depends on the number of features used and the size of the dataset. The model is relatively simple, and the coefficients can be interpreted to understand the impact of each feature on the outcome. From the table, we noticed that age, being active member, geography, gender, balance were the most important features in predicting whether a customer would churn or not.

```
Logistic Regression (L1) Coefficients
Age : 0.7307
IsActiveMember : -0.5046
Geography_Germany : 0.3121
Gender : -0.2409
Balance : 0.1509
CreditScore : -0.0457
NumOfProducts : -0.0438
Tenure : -0.0271
EstimatedSalary : 0.0092
Geography_France : -0.0043
HasCrCard : -0.0022
Geography_Spain : 0.0
```

We selected performance metrics of accuracy, precision, recall and F1 score to evaluate how well the logistic regression model preformed in predicting the binary outcome of whether a customer churned or not. We paid more attention on precision and recall because our dataset was imbalanced, with class 0 being much more frequent than class 1. Therefore, the accuracy metric could be misleading. We tried both resampling and SMOTE to handle the imbalanced data, but the results did not meet our expectation and the datasets were still imbalanced. We would continue working on this part and try to solve this problem next week.

After selecting the metrics, we split the dataset into a training and validation set and calculated the performance metrics for both sets. We also evaluated three variations of the model using 12, 9 and 6 features. We selected all features for model 1 except the useless ones such as surname. We picked the most important 9 features for model 2 and 6 crucial attributes for model 3 according to L1 coefficients. The performance metrics for the training and validation datasets for all three variations were shown in the table below. By looking at the table, we noticed that model 2 and 3 had a better performance than model 1 because they had higher accuracy, precision and recall for both training and validation sets than model 1.

	Variation	Accuracy (Training)	Accuracy (Validation)	Precision (Training)	Precision (Validation)	Recall (Training)	Recall (Validation)
0	Model 1	0.79	0.79	0.37	0.38	0.05	0.05
1	Model 2	0.81	0.81	0.60	0.62	0.21	0.21
2	Model 3	0.81	0.81	0.61	0.62	0.21	0.21

When comparing how the three models performed on the test dataset, we noticed that they had similar performance. After we tuned the hypermeters, the results were slightly different. Based on the performance metrics and the small difference between the variations, we would select the model using 6 features (model 3) as the best model for the week. Based on the information provided in the classification report, the model achieved a high precision of 0.83 for class 0, which indicated that when the model predicted that a customer was likely to stay at the same bank, it is correct 83% percent of the time. The high recall of 0.96 for class 0 also indicated that the model was able to identify the majority of customers who were likely to stay at the same bank, as 96% of actual customers who stayed at the same bank were correctly identified by the model. The f1-score of 0.89 for class 0 indicated a good balance between precision and recall.

However, the model performed relatively poorly on class 1 with a precision of 0.59, which indicated that when the model predicted that a customer was likely to churn, it was correct 59% of the time. The recall of 0.20 for class 1 indicated that the model was only able to identify 20% of the customer who actually churned. The f-1 score of 0.30 for class 1 was relatively low, indicating that the model was not able to balance precision and recall as well for class 1.

	precision	recall	f1-score	support
0	0.83	0.96	0.89	1593
1	0.59	0.20	0.30	407
accuracy			0.81	2000
macro avg	0.71	0.58	0.60	2000
weighted avg	0.78	0.81	0.77	2000

: 0.8095

As mentioned above, we evaluated two hyperparameters for the logistic regression model; the penalty(L1 or L2 regularization) and the regularization strength (C). To evaluate these hyperparameters, we used grid search, which exhaustively searched over a range of values to find the best combination that maximized the model performance. Based on the grid search results, the best hyperparameters for model 3 were C=0.01 and penalty = L2. This suggested that the model used L2 regularization with a regularization strength of 0.01 and prioritized simplicity over fitting the training data very closely.

Overall, the model performed well in predicting customers who were likely to stay at the same bank but struggled in predicting customers who were likely to churn. The results suggested that the model might benefit from further tuning or more data to improve its performance on class 1