Abrar Imon STAT 724 Final Project

Introduction:

The aim of this paper is to create models that will help predict whether future customers will exit the bank's services or not. Banks can use this information to better understand which of their customers are more likely to leave. The algorithms used in this paper are Logistic Regression, Quadratic Discriminant analysis, Decision Tree classifier, Random Forest classifier, Gradient Boosting classifier, and XGBoost classifier. These algorithms will offer ways to interpret/understand the effects of the different features on the response variable as well as make predictions about future customers' actions.

This paper includes three rounds of analysis. The first round is about setting up the important features and investigating how these features affect the response variable using Logistic Regression, and Decision Tree classifier. The second round will be about comparing how the most powerful classification algorithms perform on predicting the behavior of new customers. In the third round, the algorithms that performed the best in the second round will be further optimized with the help of Grid Search CV and compared. In this paper, the primary method for comparing different algorithms will be with the use of accuracy score, but other metrics like AUC & precision will be reported along with visuals such as confusion matrices and ROC.

Description of Data:

This is a dataset about a bank's churn rate. This dataset has information about many of their customers and whether they ended their business with the company or not. Each row represents a customer. The columns include information about the individuals' credit score,

geography, gender, age, tenure, balance, number of products, and estimated salaries. These are also the features that were selected to be used training the models. This dataset is from Kaggle.

Data Setup:

The dataset contains information about the geography and gender of the customers. Since these were categorical variables, they needed to be modified before the machine learning algorithms could be trained on them. The column for geography had three unique values and they were unordered. Therefore, the column was one-hot encoded into two new columns: "geo_Germany", and "geo_Spain". These columns have the value 1, when a customer is from that country and the value 0, when a customer is not from that country. The column on gender was changed to a column called "Male", which has values 1 and 0 to represent whether an individual is male or not. The data was checked for null values and for placeholders for missing values. The dataset does not contain any such values.

The dataset contains 10,000 rows for 10,000 customers. These were randomly split so that 7,000 of the rows will be used for training (and validation when required) and the other 3,000 rows will be used for testing the models. In the comparing phase, all models will be trained on the same training data and tested on the same test data.

Analysis Round 1 (Interpretable Models):

Decision Tree Classifier:

This model was trained so that the tree does not have any more than 3 levels. This was done to get an interpretable visual of the tree in which only the most important features are shown. This limitation also prevents the tree from overfitting.

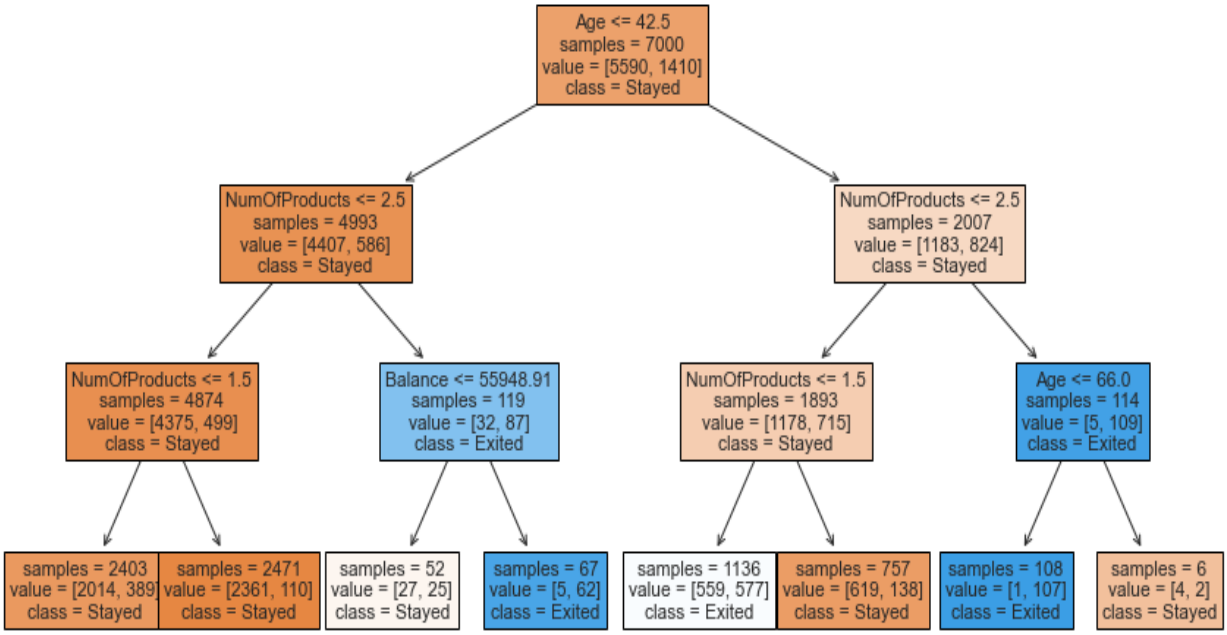

Figure 1: Decision Tree with max_depth=3

The number of products that a customer utilizes, is an important predictor for whether a customer will exit or not. Customers who used 2 or less products are more likely to stay than customers who used more than 2. Customers who utilize exactly 2 products are most likely to stay. The feature that initially divides the two classes of customers the best, is the age feature. Customers who are younger than 43 years, are more likely to stay than the customers who are older. The customers that are most likely to leave are the ones in the second to last leaf node. This is because that node represents 108 customers of which 107 have exited, making it the node with the highest ratio of customers who exited to customers who stayed. The conditions that lead to the highest probability of a customer exiting, are that the customer is between and including the ages 43 and 66 and they utilize less than or equal to 2 products.

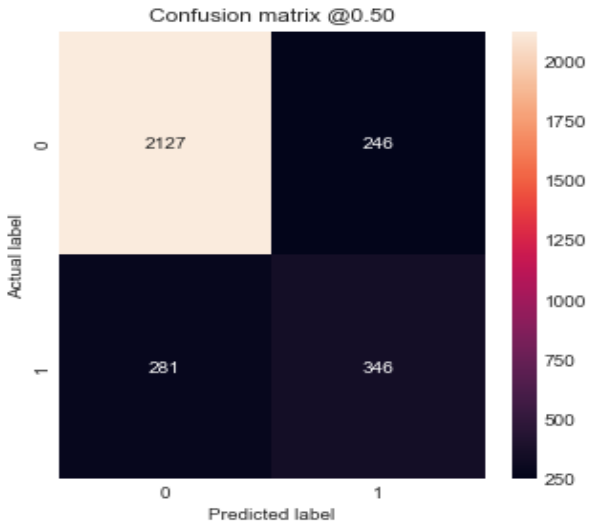

Figure 2: Decision Tree; Accuracy on test data: 0.824333

Logistic Regression:

Logistic Regression is applied to the training data. Below, the features are mentioned in order from most significant to least significant. The cutoff between whether a feature is significant or not, is at a p-value of 0.1.

The features that are the most significant predictors of whether a customer will exit or not, are: gender, geography, balance, age, and credit score. Males are less likely to exit than females are. Of the different nationalities, someone who is from Germany is more likely to exit than someone from Spain or France. Customers with higher balance and higher age, are more likely to exit. Customers with higher credit scores are less likely to exit.

Below, the features that are not significant are mentioned in order from least significant to most significant. The p-values for estimated salary and tenure are the highest. This implies that information about the estimated salary of a customer and the amount of time that a customer has spent being a customer, are not useful when predicting whether the customer will exit. Also,

information about whether a customer has a credit card or not, does not help the logistic regression mode in predicting whether the customer will exit.

The p-value for the number of products that a customer uses, is 0.117. This suggests that the "NumOfProducts" is not a useful feature for predicting whether a customer will exit or not. According to the interpretation of this feature from the Decision Tree section, a value of "NumOfProducts" = 2, minimizes the probability of a customer exiting. So, the relationship between this feature and the response variable is likely non-monotonic. The Logistic Regression model did not capture this relationship, because it is a linear model.

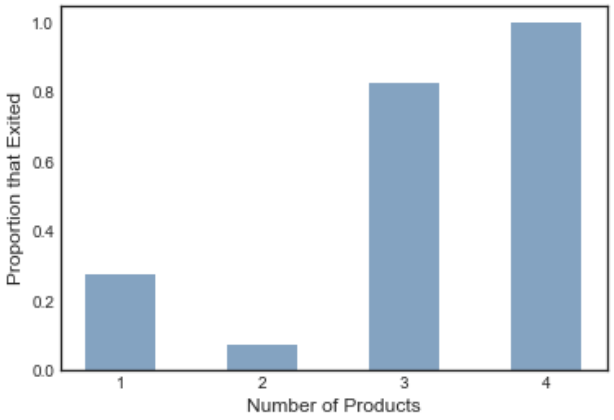

Figure 3: Effect of the number of products on churn rate

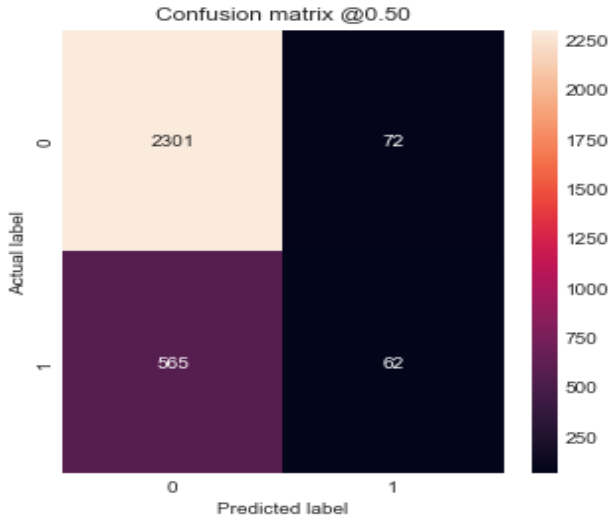

Figure 4: Logistic Regression; Accuracy on test data: 0.787667

Analysis Round 2 (Finding the Best Models):

More classification algorithms are trained on the training data. These algorithms are Quadratic Discriminant Analysis, Random Forest classifier, Gradient Boosting classifier, and XGBoost classifier. These models are applied either with the default parameters or with minor adjustments to the default parameters. Below are their performance metrics on the test data at 0.5 threshold:

Algorithm	**TP**	**TN**	**FP**	**FN**	**Accuracy**
QDA	195	2265	108	432	0.820000
Random Forest	305	2247	126	322	0.850667
Gradient Boost	262	2287	86	365	0.849667
XGBoost	298	2245	128	329	0.847667
Decision Tree	346	2127	246	281	0.824333
Logistic Regression	62	2301	72	565	0.787667

Judging by the accuracy metric, the algorithms that performed the best in this round are the ensemble algorithms. These algorithms (in order from most accurate to least accurate) are Random Forest classifier, Gradient Boosting classifier, and XGBoost classifier. These 3 algorithms will move on to the next round.

Analysis Round 3 (Hyperparameter Tuning):

For each of the three most accurate models from last round, the Grid Search CV algorithm is applied on the training data many times. This is to zero in on the best combination of hyperparameters for each of the algorithms, while minimizing the number of combinations of hyperparameter values that are cross validated. Below are the performance metrics of each of the ensemble algorithms on the test data at 0.5 threshold, before and after optimization:

Algorithm	**TP**	**TN**	**FP**	**FN**	**Accuracy**
Random Forest	305	2247	126	322	0.850667
Random Forest (Optimized)	258	2303	70	369	0.853667
Gradient Boost	262	2287	86	365	0.849667
Gradient Boost (Optimized)	267	2287	86	360	0.851333
XGBoost	298	2245	128	329	0.847667
XGBoost (Optimized)	268	2287	86	359	0.851667

All three of the ensemble algorithms improved slightly, after hyperparameter tuning. The Random Forest classifier improved by about 0.003 accuracy points. The Gradient Boosting classifier improved by about 0.002 accuracy points. The XGBoost classifier improved by about 0.04 accuracy points.

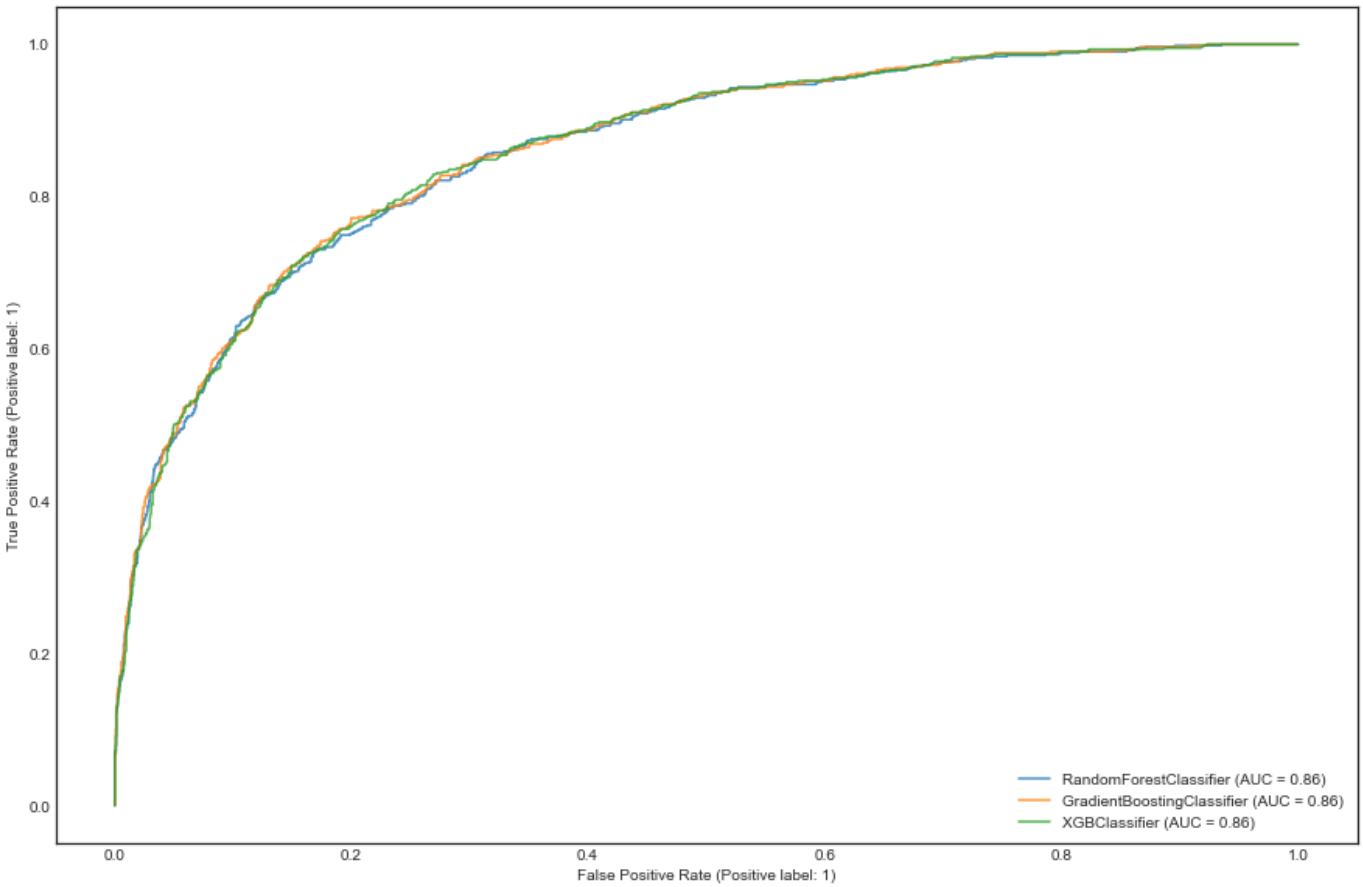

Figure 5: ROC plot of the three ensemble models

These three ensemble models have similar accuracy scores at 0.5 threshold. To compare them in a threshold independent way, the ROC is plotted, and the AUC is calculated.

The ROC looks identical for the different models, with each model having some interval of false positive rate, at which it maximizes the true positive rate. The AUC for all three models, are 0.86. They are calculated with more precision and the new AUC are in the table below:

Algorithm	**AUC**
Random Forest (Optimized)	0.859906
Gradient Boost (Optimized)	0.862324
XGBoost (Optimized)	0.861566

Conclusion:

In this paper, the way the different features affect the response variable is analyzed with interpretable machine learning algorithms such as Logistic Regression classifier and a Decision Tree classifier. The Decision Tree classifier suggests that customers between and including the ages 43 and 55 are most likely to exit and the customers who use exactly 2 different products are most likely to stay. This suggests that the bank should invest in attracting customers outside of that age range as they are more likely to stay. The plot of the Decision Tree classifier shows other effects such as how some conditions affect how other features affect the chances of a customer exiting. The interpretability that this model provides, can help the bank retain more customers. For example, the bank should try to change the way they interact with customers of ages between 42 and 67 and test which combination of interactions allow them to retain most of these customers. The bank should further investigate and try to understand why customers who use exactly 2 different products are most likely to stay.

The Logistic Regression algorithm suggests that males are more likely to stay than females. The bank should experiment and see what they can do to attract more female customers. The significance of geography, balance, age, and credit score in predicting whether customers will exit or not also provides the bank with information they can use to decide what kind of customers they invest more into attracting.

The bank might also want to predict whether a future customer will exit or not. The algorithms that provide the best way to do that are the hyperparameter tuned ensemble algorithms. Out of those algorithms, the Random Forest classifier has the highest accuracy, and the Gradient Boosting algorithm has the highest AUC.

Below is a summary of the best algorithms that were trained and their respective AUC and accuracy (ordered from highest AUC to lowest AUC):

Algorithm	**AUC**	**Accuracy**
Gradient Boost (Optimized)	0.862324	0.851333
XGBoost (Optimized)	0.861566	0.851667
Random Forest (Optimized)	0.859906	0.853667
Quadratic Discriminant Analysis	0.803756	0.820000
Decision Tree	0.799236	0.824333