Don’t Bank On It: Logistic Regression, Random Forest, SVM and Their Performance in Predicting Bank Customer Churn

Introduction and Background:

The purpose of this project is to compare the performance of Logistic Regression, Random Forest, and SVM in predicting whether a customer will leave a bank or not. The methods were selected because they are the most popular classification methods out there. Because banks generally make money by using customer savings, having customers leave is detrimental to their business. By creating a model that can predict who is to leave, banks can design business initiatives to cater to those customers and better retain them.

Data Structure:

Per Kaggle [1], the dataset utilized is originally from a US bank, though the geography suggests this project only considered its European clients. The dataset contains 10,000 observations. A description of the variables can be found in ***Table 1***.

Table 1: Variables and their Description

Variable	**Description**
RowNumber	The number of the row
CustomerId	Customer ID
Surname	Client's last name
CreditScore	Client's credit score
Geography	Location (Germany, France or Spain)
Gender	Male/Female

Age	Age of customer
Tenure	How many years they have been bank customers
Balance	Average balance of customer
NumOfProducts	Number of bank product facilities customer is using
HasCrCard	Whether they have a credit card or not
IsActiveMember	Whether they are an active member or not
EstimatedSalary	Estimated Salary
Exited	Whether they left the bank or not

The first 3 variables represent an ID in some form so there were automatically removed leaving 10 variables to try to predict the dependent variable Exited, that is whether a customer left the bank or not. To get a feel of the data, some visualizations were created. ***Fig1*** shows a bar plot of exited where to the left you will find the proportion of people that didn't exit the bank and to the right the ones that did. Overall, most of the customers didn't exit the bank as 79.63% of all values belong to that category. While this makes the distribution not degenerate, it is rather unbalanced so a ROC curve will need to be implemented to address this.

Fig1: Churn Proportions with left representing the percentage that left the bank and right the ones that did not.

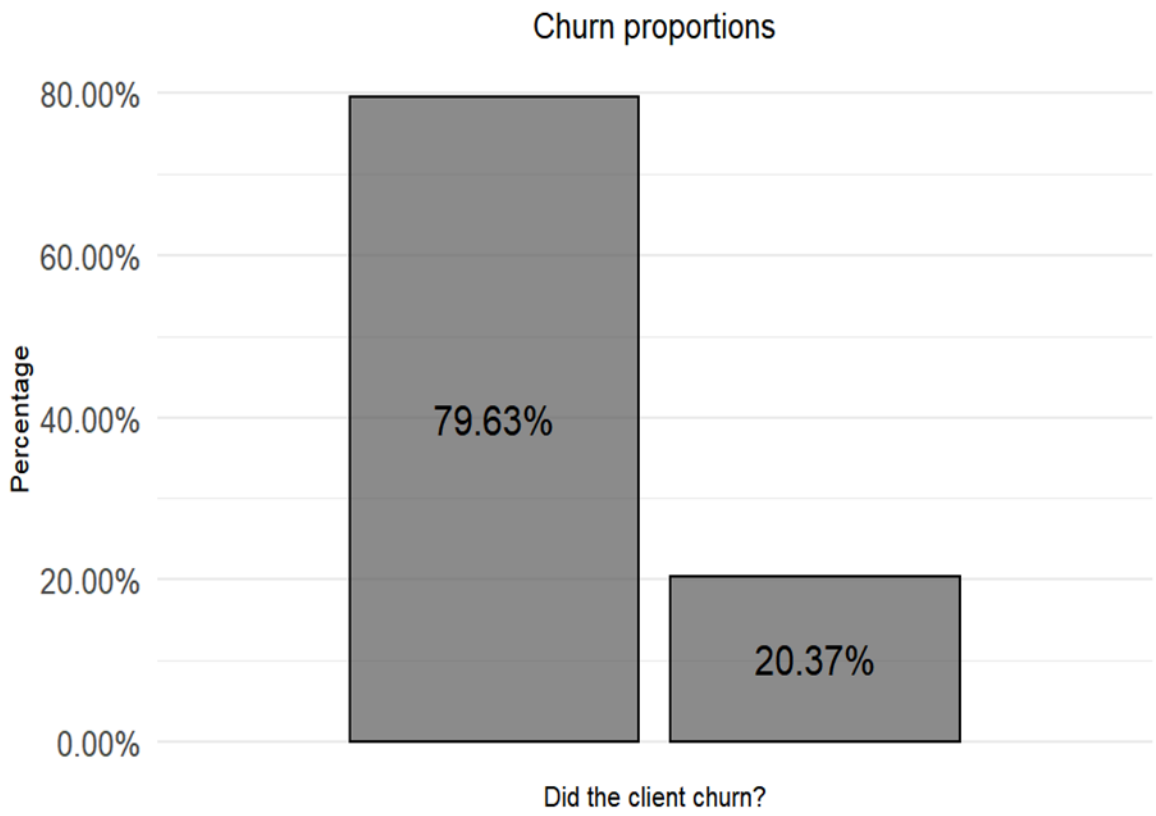


Fig2 just shows a breakdown by both gender and geography. What's interesting to see here is that women tend to leave their banks more often than men do for any of the three countries considered. Even without proper modeling, this already provides some guidance to banks on some demographics they might want to cater to. As far as geography, Germany has more churning than either France or Spain. ***Fig3*** provides some insights on some of the initial trends previously mentioned. We see that Females earn either more than men or just as much in all the countries. It is possible that higher earnings lead customers to seek banks that offer better ways to grow their money which is why churning is far more common in women than in men. Another noteworthy thing is that the average salary is high, so this is not your regular earner.

Fig2: Churn Proportions by Gender and Geographical Location

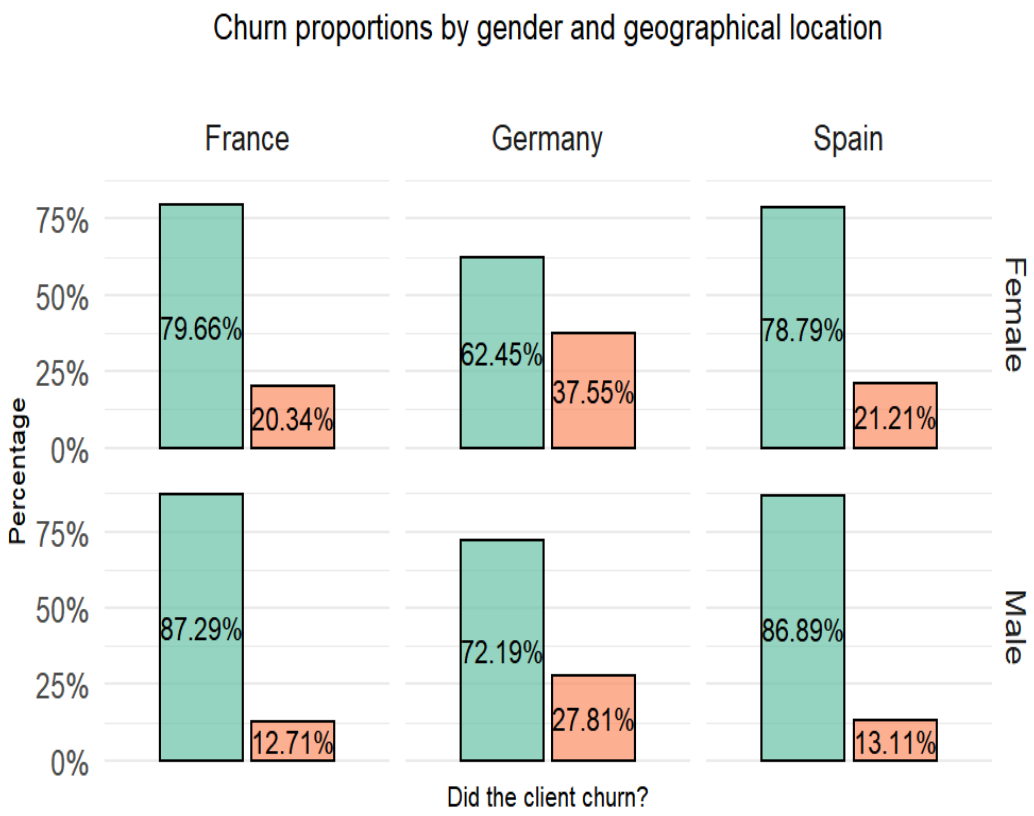
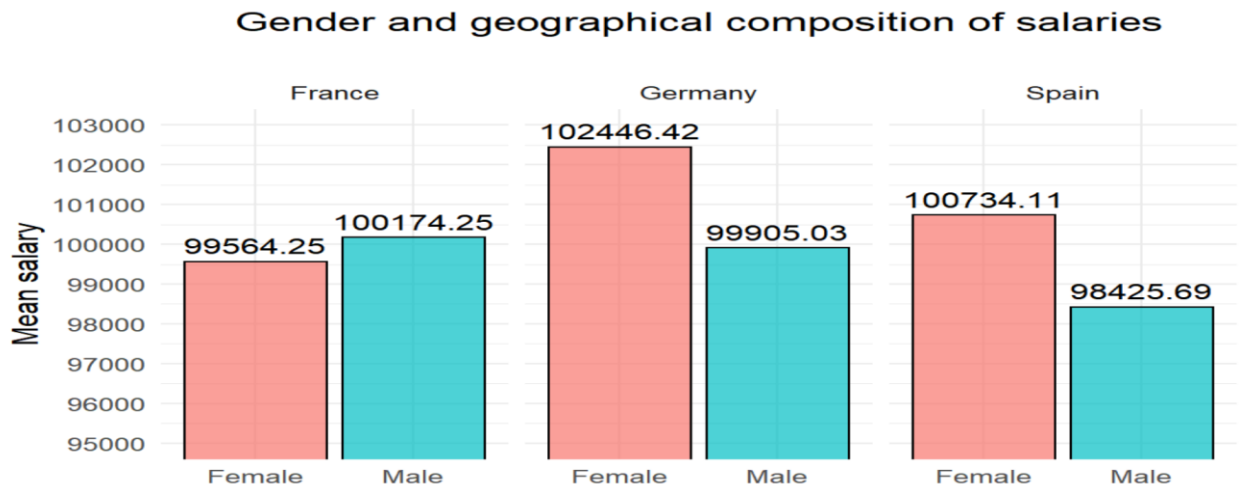


Fig3: Average Salary by Geography and Gender



All character variables were turned to categorical variables and summary was applied to see if check any oddities in the data. ***Fig4*** shows the output of the summary. Nothing really sticks out except for the minimum of EstimatedSalary being $11.58. This seems rather low but upon inspection of the lower ranges of the data, this value is not extraordinary as there are others close to it. Maybe these people work odd jobs or side business or perhaps they put the wrong

information in the bank application. Given this possibility, the small minimum was not a concern.

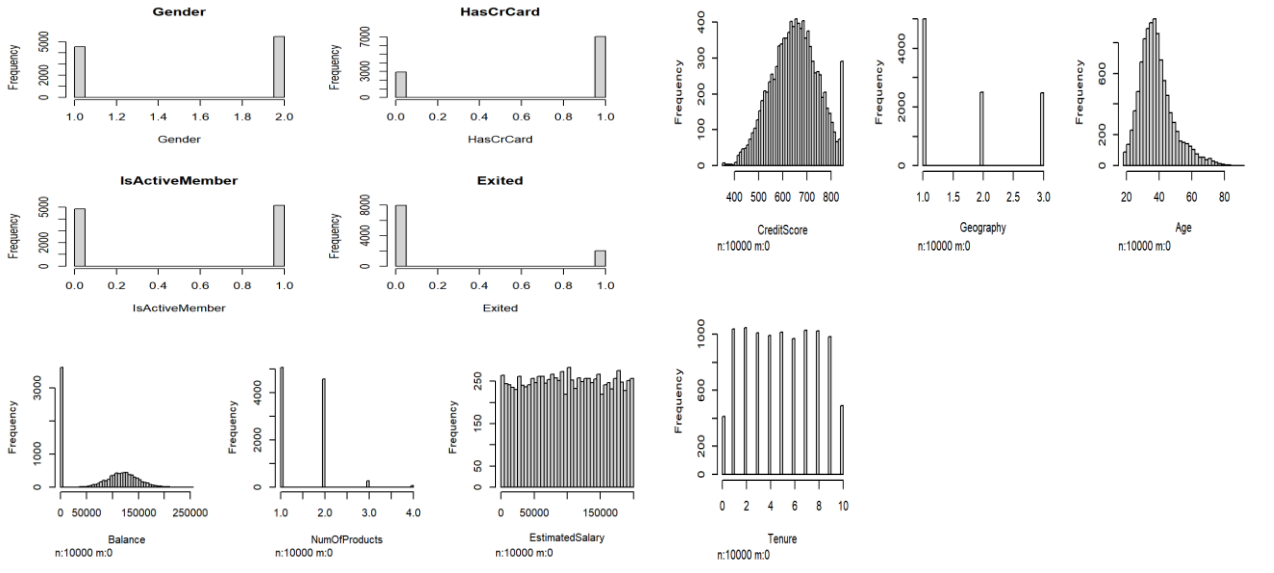
Fig4: Summary for all variables

```
summary(Churn)
```

```
##   CreditScore      Geography         Gender          Age
##  Min.   :350.0   Min.   :1.000   Min.   :1.000   Min.   :18.00
##  1st Qu.:584.0   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:32.00
##  Median :652.0   Median :1.000   Median :2.000   Median :37.00
##  Mean   :650.5   Mean   :1.746   Mean   :1.546   Mean   :38.92
##  3rd Qu.:718.0   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:44.00
##  Max.   :850.0   Max.   :3.000   Max.   :2.000   Max.   :92.00
##      Tenure          Balance       NumOfProducts    HasCrCard
##  Min.   : 0.000   Min.   :     0   Min.   :1.00   Min.   :0.0000
##  1st Qu.: 3.000   1st Qu.:     0   1st Qu.:1.00   1st Qu.:0.0000
##  Median : 5.000   Median : 97199   Median :1.00   Median :1.0000
##  Mean   : 5.013   Mean   : 76486   Mean   :1.53   Mean   :0.7055
##  3rd Qu.: 7.000   3rd Qu.:127644   3rd Qu.:2.00   3rd Qu.:1.0000
##  Max.   :10.000   Max.   :250898   Max.   :4.00   Max.   :1.0000
##  IsActiveMember   EstimatedSalary         Exited
##  Min.   :0.0000   Min.   :    11.58   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.: 51002.11   1st Qu.:0.0000
##  Median :1.0000   Median :100193.91   Median :0.0000
##  Mean   :0.5151   Mean   :100090.24   Mean   :0.2037
##  3rd Qu.:1.0000   3rd Qu.:149388.25   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :199992.48   Max.   :1.0000
```

Degenerate distributions could cause issues so the distribution for all variables was checked in ***Fig5*.** Something that could be of concern is that many people have a credit score of 850. While this might seem odd, these people are generally high earners, so it is very likely that they tend to make payments on time. Balance has a lot of zeroes but zeroes, in general, are quite common in people's bank accounts. With no more concerns to consider, the data was checked for missing values and as none were found, the data was ready for modeling.

Fig5: Distributions for all variables



Statistical Learning Methods:

- **Logistic Regression Model**

This statistical analysis (also referred to as a logit model) is frequently used for predictive analytics and modeling, and it has applications in machine learning. The dependent variable in this analytics technique is finite or categorical: either A or B (binary regression) or a range of finite possibilities A, B, C or D (multinomial regression) [2]. It is used in statistical software to estimate probabilities and explain the relationship between the dependent variable and one or more independent variables using a logistic regression equation. These models assist us in understanding the relationships, predicting and forecasting the outcomes, which allows us to take action to enhance decision-making.

- **Random Forest**

A well-known machine learning (ML) model for data classification is Random Forest (RF). This algorithm is widely used in industries including investment, customer relationship management, and branding. The RF is supported by a forest composed of many decision trees. It is supplemented by an average of the prediction's mean value, which is generated at the end of each tree, minimizing the single tree's lack of robustness [3]. Each tree is constructed using a subset of input variables chosen at random.

The estimated model may be expressed as follows:

$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n} g_k(x)$$

where g(x) is a collection of the kth learner random trees, and x is a vector of input features. The final RF estimate is the average of all the trees' results. As a result of such weights, each individual tree influences RF estimate.

The Random Forest model outperforms conventional machine learning algorithms. This is owing to the former's consistency in gathering training data from subsets automatically and shaping trees utilizing random procedures. Additionally, because the Random Forest model is trained via bootstrapping on a randomly selected independent subset of datasets, the overfitting quantity is retained.

- **Support Vector Machines (SVMs)**

SVM is a popular supervised machine learning technique for classification. SVM training algorithm generates a model for a given collection of training data, each designated as belonging to one of two categories, by finding a hyperplane that classifies the provided data as properly as practicable by optimizing the distance between two data clusters [4].

A classification problem's class prediction is based on determining the best boundary between classes. The optimal boundary of an SVM model may be determined based on the values of accuracy, sensitivity, and specificity.

Fig6: Support Vector Machine and their decision boundary [5]

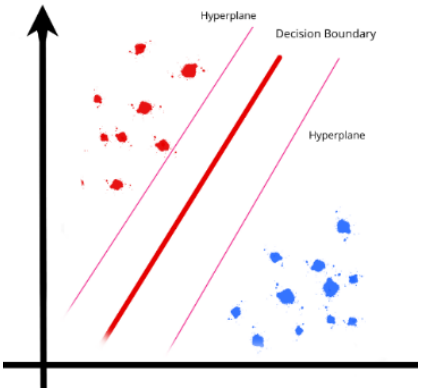
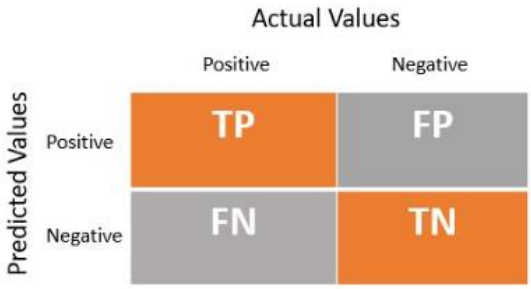


Fig7: Confusion Matrix



From the confusion matrix Accuracy, Sensitivity and Specificity are evaluated using the following equations:

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

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

Where TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative

To make the mathematics possible, Support Vector Machines use ‘Kernel Function’ to systematically find support vector classifiers in higher dimensions and the support vector classifiers can be used to classify new observations.

The Kernels are mapping functions to map the data from one space to a new higher-dimensional space. These functions can be different types, i.e., linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid and their equations are as follows,

Polynomial Kernel: $k(x,y)= tanh(\gamma xTy + r)d\ ,\ \gamma>0$

Gaussian Radial Basis Function (RBF) : $k(x,y)=exp(-\gamma||\ x\ -y\ ||2)$

Sigmoid Function: $k(x,y) = tanh(\gamma xTy + r)$

where γ, r are kernel parameters.

Analysis Results:

- **Logistic Regression Model**

The most statistically significant variables were Geography, Gender, Age, Balance and whether a client is an active member or not.

Coefficient Interpretation:

Geography|Germany: Given that, the country of residence for a client is Germany, the hazard to churn increases by a factor of 2.17 or by 159 % for German residents compared to French residents. Germans are more likely to decide to churn in contrast to French and Spanish clients.

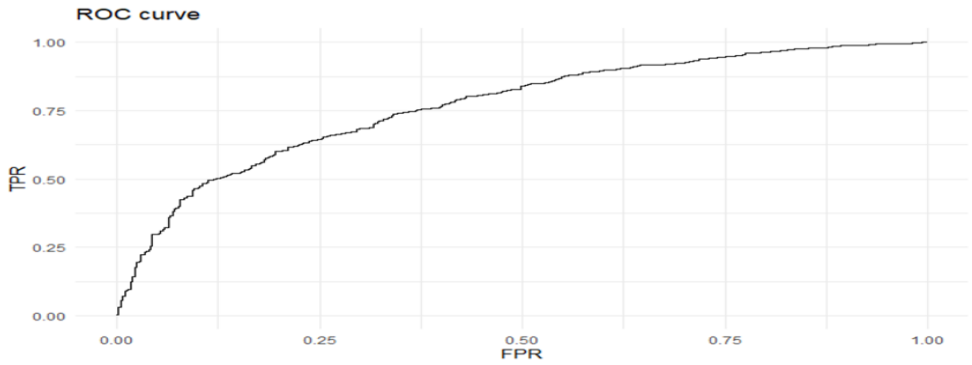
Gender|Male: Given that the client is male, the risk of churning decreases by a factor of 0.59 or by 41% compared to female clients. Men are 41% more likely to stick to one bank than women are.

Age: A one-year increase in age of a client increases the hazard to churn by a factor of 1.08 or by 8%.

IsActiveMember|yes: If a person is an active member of a bank system, the hazard to churn decreases by a factor of 0.33 or by 77%. Active bank clients are 77% more likely to stay.

Confusion Matrix, Statistics, and ROC curve

Fig8: ROC Curve | Area under the curve: 0.7682. The ROC and AUC metrics confirm the result from confusion matrix.

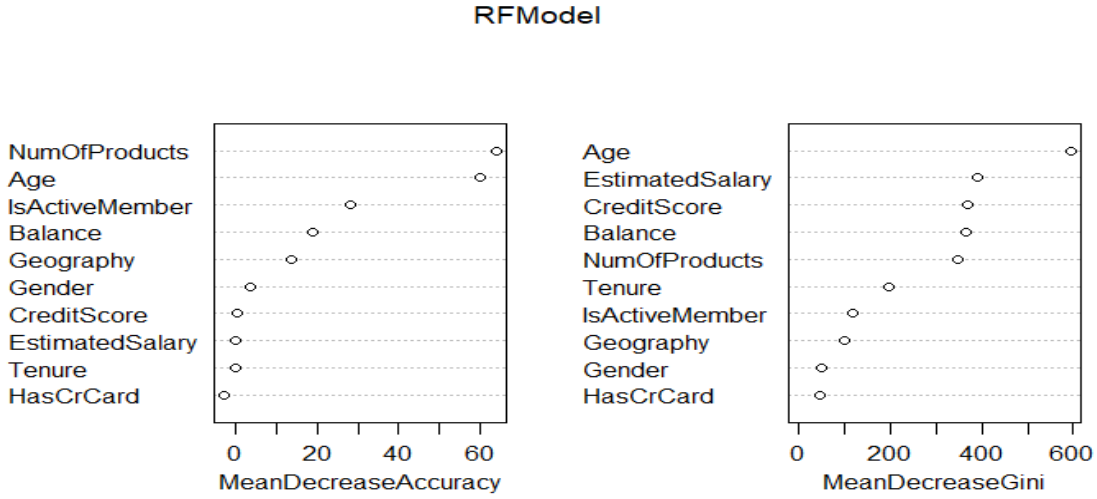


After we have partitioned the data into train and test datasets, we have found that model accuracy on the test subset (Accuracy = 80.76%) is not that different compared to the model on the train subset (Accuracy = 80.92%), which means that the model was trained properly (no overfitting) and performs almost as equally well as on the training set, which is great. However, sensitivity was rather low at about 20%.

- **Random Forest**

The Mean Decrease Accuracy plot shown in ***Fig9*** expresses how much accuracy the model losses by excluding each variable. The more the accuracy suffers, the more important the variable is for the successful classification. The variables are presented in descending importance. The mean decrease in the Gini coefficient is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. The higher the value of mean decrease accuracy or mean decrease Gini score, the higher the importance of the variable in the model.

Fig9: Variable Importance for Random Forest



Confusion Matrix and Statistics:

```
## Confusion Matrix and Statistics
##
##          Reference
## Prediction    0    1
##          0 1524  226
##          1   53  197
##
##               Accuracy : 0.8605
##                 95% CI : (0.8445, 0.8754)
##    No Information Rate : 0.7885
##    P-Value [Acc > NIR] : < 2.2e-16
##
##                  Kappa : 0.5082
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9664
##             Specificity : 0.4657
##          Pos Pred Value : 0.8709
##          Neg Pred Value : 0.7880
##              Prevalence : 0.7885
##          Detection Rate : 0.7620
##    Detection Prevalence : 0.8750
##       Balanced Accuracy : 0.7161
##        'Positive' Class : 0
```

The Random Forest model with 10-Fold repeated cross-validation produces a prediction accuracy of 86% with a Sensitivity and Specificity of 97% and 47%, respectively.

- **SVM**

Table 2 provides a comparison of the results of the different parameters that the models are being evaluated. It also shows that for polynomial SVM, sensitivity is 36.52%, which indicates that the model has a recall of 36%, and it correctly identifies more than 36% of all churns. Since the specificity for polynomial SVM is 98.43%, it indicates that the model has a precision of more than 98%. When this model predicts whether a customer will churn or not, it will be correct more than 98% of the time.

Table 2: Comparison of Model Statistics

	Test accuracy	**Sensitivity**	**Specificity**
SVM-Linear	0.7961	0.0000	1.0000
SVM-Polynomial	0.8581	0.36520	0.98431
SVM-RBF kernel	0.8571	0.38971	0.97677
SVM-Sigmoid Kernel	0.6952	0.17402	0.82863

While predicting whether a customer will leave a bank or not, our application of SVMs models achieved good classification performance with a polynomial function accuracy of around 86% and RBF function accuracy of around 86%.

Conclusion:

Selecting the best model, in this case, is not just as simple as selecting the most accurate model. The main issue here is that the concern is with the people that left the bank as they are the type of customers banks should focus on. Given this, the focus should really be on the specificity, that is how good is the model at predicting whether a customer will leave. While this distinction needed clarification, the clear winner in both accuracy and sensitivity (specificity in the output as the matrix was specified backwards) is the random forest at 86% and 47% respectively.

References

1. Kaggle dataset. https://www.kaggle.com/datasets/shantanudhakadd/bank-customer-churn-prediction Accessed 25 April 2022
2. Logistic Regression. IBM. https://www.ibm.com/topics/logistic-regression Accessed 6 May 2022
3. Random Forest. Wikipedia. https://en.wikipedia.org/wiki/Random_forest Accessed 6 May 2022
4. Nello Cristianini and Elisa Ricci, Support Vector Machines. Encyclopedia of Algorithms. SpringerLink. DOI: https://doi.org/10.1007/978-0-387-30162-4_415
5. Image Source: https://towardsdatascience.com/breaking-down-the-support-vector-machine-svm-algorithm-d2c030d58d42 Accessed 6 May 2022