ST 563
Project

Modeling Customer Churn Using Supervised Machine Learning Models

By

Lilian Ngonadi

1 Introduction

In this project, we used the supervised learning approach to analyze whether a customer is going to churn or not based on various attributes. By churn, we mean whether a customer closes their account or leaves the bank. If we are able to accurately identify customers that are at risk of churning, it can help the financial sector to implement proactive retention measures. The churn dataset used for this project is sourced from the Kaggle website and contains 10,000 observations and it includes variables such as CreditScore, Geography, Gender, Age, Tenure, Balance, Number of products, HasCrCard, IsActiveMember and Estimated salary. The response variable used in the project is Exited(which tells whether a customer churns or not).

This task is a classification problem and this is because our response variable is categorical(churn or not churn). The predictors are mixed of both categorical and numeric predictors. Age, Tenure, Balance, Number of products and Estimated salary are all Numerical while gender, age, HasCrCard and IsActiveMember are all categorical. For the analysis we explored different methods that were taught in class which includes the k-Nearest Neighbors (kNN), regularized logistic regression, generalized additive models (GAMs), decision trees, ensemble tree methods, and support vector machines (SVMs).

Goal: The objective of the work is to build a predictive model that help classify whether a customer will churn or not based on the various features.

2 Methods

2.1 Data Preparation

The dataset used for this project was imported into R using the read_csv() function, then i excluded three identifier columns which include (RowNumber, CusomerId and Surname). I went ahead to encode the categorical variables as factors. Also, the response variable Exited was cast as a factor for classification. A sample of the dataset is shown in table 2

CreditScore <dbl>	Geography <fctr>	Gender <fctr>	Age <dbl>	Tenure <dbl>	Balance <dbl>	NumOfProducts <dbl>	HasCrCard <fctr>	IsActiveMember <fctr>	EstimatedSalary <dbl>	Exited <fctr>
619	France	Female	42	2	0.00	1	1	1	101348.88	1
608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
502	France	Female	42	8	159660.80	3	1	0	113931.57	1
699	France	Female	39	1	0.00	2	0	0	93826.63	0
850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

Table 1: Sample of the Churn dataset

2.2 Handling Missing Data

For the project we made sure to check if there is any missing data but it showed that the dataset did not have any missing entries. You will notice for the Balance we have some zeros that isnt considered as missing values but it just meant the account has no money in it.

CreditScore	Geography	Gender	Age	Tenure	Balance
0	0	0	0	0	0
NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	
0	0	0	0	0	

Table 2: Check for Missing values

2.3 Outliers detection

We used boxplot as shown in figure 1 to check for the presence of outliers and noticed that Age and Credit score has some extreme values but we decided to retain that since they are valid customers and not due to some entry errors. Also, since i am working with churn data , the values could help in the analysis

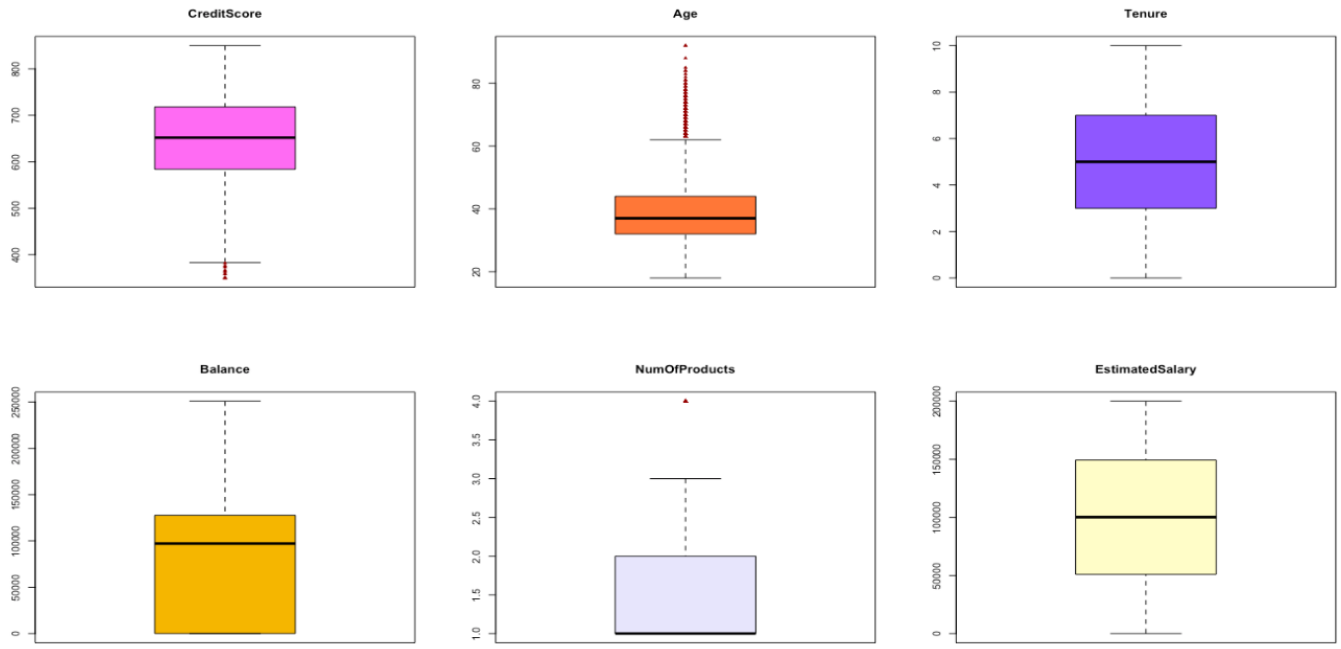

Figure 1: Check for presence of outliers

2.4 Data Exploration

The Exited variable have a case of imbalanced data with approximately 80% of the customers not leaving the bank and 20% leaving the bank. For future work we are expected to use Smote to address the class imbalance but since it wasnt covered in class we will work with the original dataset.

From table 3 we can see that customers have credit scores ranging from 350 to 850 and the mean age of customers is 39years. The median account balance of the customers is high with a value of 97199 although many of the customers dont have any balance in their bank account. We can also see that many of the customers have 1 or 2 products with their salay ranging from a very low value of 11.58 to nearly 200,000.

Variable	Min	Q1	Median	Mean	Q3	Max
CreditScore	350	584	652	650.5	718	850
Age	18	32	37	38.92	44	92
Tenure	0	3	5	5.01	7	10
Balance	0	0	97199	76486	127644	250898
NumOfProducts	1	1	1	1.53	2	4
EstimatedSalary	11.58	51002	100194	100090	149388	199992

Table 3: Summary statistics for key numeric variables in the dataset

Also, the result from table 4 shows that the bank has more male customers(5457) than the female customers(4543). Also most of the customers are residing in france. High number of customers have credit card(7055) when compared to those without credit card(2945) and they are also more active members. We have more customers that did not churn than those that churned.

Category	Level	Count
Gender	Female	4543
	Male	5457
Geography	France	5014
	Germany	2509
	Spain	2477
HasCrCard	No	2945
	Yes	7055
IsActiveMember	No	4849
	Yes	5151
Exited	No	7963
	Yes	2037

Table 4: Frequency counts for categorical variables in the dataset

From the scatter plot we can see that customer churn has strong association with age than

variables like the credit score or estimated salary. Older customers are more likely to churn than customers that are young, this can be seen in figure 2 and figure 3

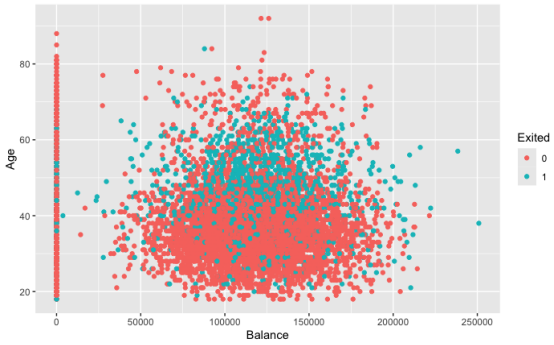

Figure 2: Scatter plot showing the distribution of Balance vs Age colored by Churn status

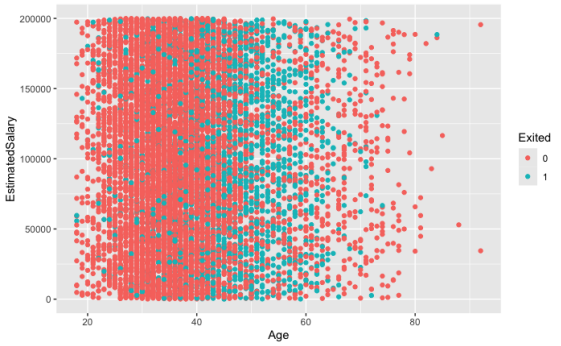

Figure 3: Scatter plot showing the distribution of Age vs Estimated Salary colored by Churn status

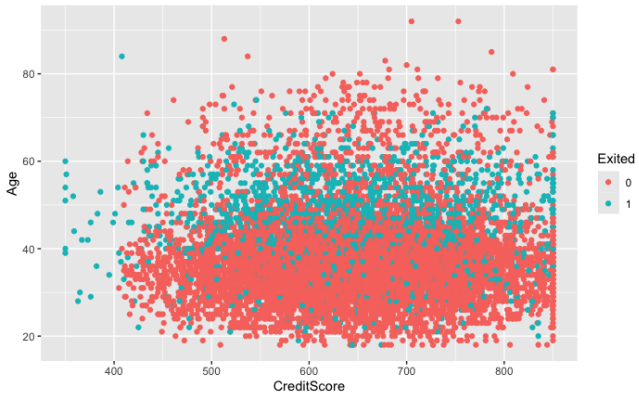

Figure 4: Scatter plot showing the distribution of CreditScore vs Age colored by Churn status

Figure 5: Scatter plot showing the distribution of Age vs Estimated Salary colored by Churn status

From figure figure 6 we can see that age has a strong influence to the churn staus than other features that are less influential. But overall age and balance seems to be have an influence on whether a customer will churn or not

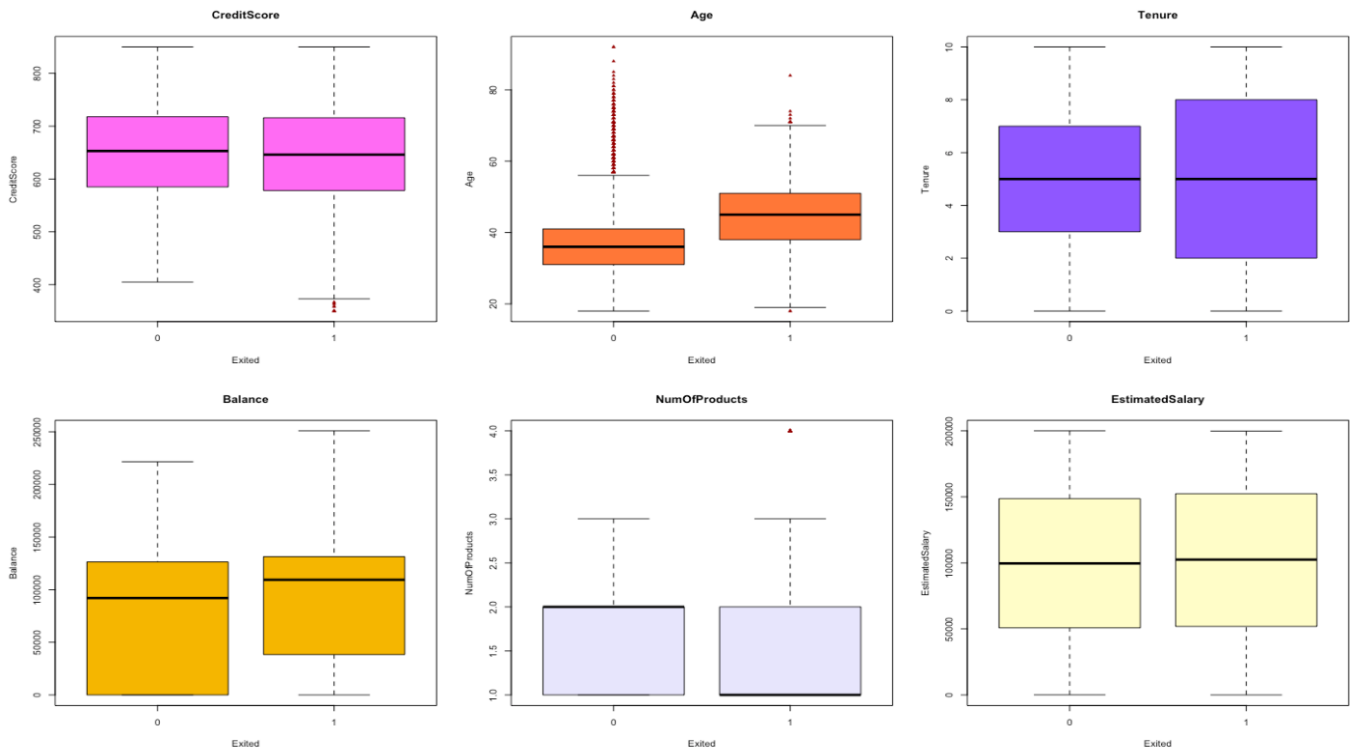

Figure 6: Distribution of the various features by Churn status

2.5 Data Splitting

The data was splitted into a training set using 70% of the data and test set using 30% of the data. We made sure to use stratified sample inorder to preserve the distribution of the Exited variable in both subsets.

Dataset	Class 0 (No Churn)	Class 1 (Churn)
Training Set	5575	1426
Test Set	2388	611

Table 5: Class distribution in training and test datasets

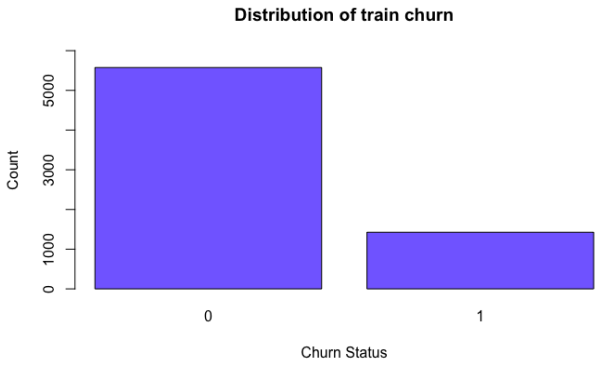

Figure 7: Distribution of train churn

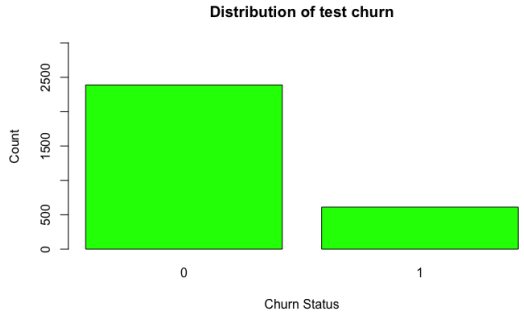

Figure 8: Distribution of test churn

2.6 Model Training

For the analysis we explored 6 different methods which includes the k-Nearest Neighbors (kNN), regularized logistic regression, generalized additive models (GAMs), decision trees, ensemble tree methods, and support vector machines (SVMs). Each of the model was trained using the training set(70%) that was obtained after the data was splitted.

2.6.1 K-Nearest Neighbors (kNN)

The KNN model is said to be a non parametric classification method. We standardized the predictors before training since the model is sensitive to the feature scale and inorder to keep all predictors on the same scale i had to standardize. We used a 5 fold cross validation over a range of k values from 1 to 51 inorder for us to determine the value of the optimal k. After obtaining the optimal K we had to retrain the model using the optimal k obtained. We fitted four different models on their training set and then evaluate using the test set(30%). The four models that were fitted used different subset of the predictors. Model 1 used the entire predictors, model 2 used only age and balance variables, model 3 used age, balance and the interation between age and balance while the final model used geography, gender, IsActiveMember, age, balance and Number of products. We cannot use the KNN model for variable selection and inference but rather for prediction

From the result in table 6 we can see that model 4 outperformed other models with an accuracy of 85%, sensitivity of 87.1% and balanced acuracy of 79.0% which shows it provides the best trade off between customrs that churn and customers that did not churn.

Model	Accuracy	Sensitivity	Specificity	Balanced Accuracy	Selected k
Model 1	82.4%	83.7%	67.8%	75.7%	17
Model 2	79.9%	82.8%	51.6%	67.2%	33
Model 3	80.3%	82.8%	53.6%	68.2%	49
Model 4	**85.0%**	**87.1%**	**70.9%**	**79.0%**	11

Table 6: Performance comparison of kNN models using different subsets of predictors

From the result in figure 9, 10, 11 and 12 the optimal tuning parameter k obtained for model 1, 2, 3 and 4, respectively, are 17, 33, 49 and 11 with the best KNN model being model 4 at optimal k=11.The confusion matrix of our final model 4 is given in table 7

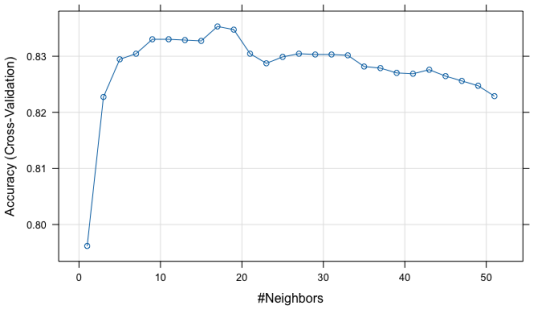

Figure 9: Optimal at K=17

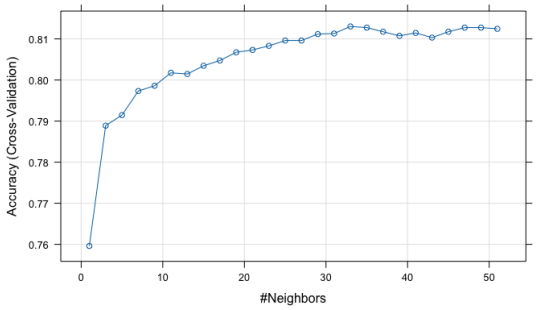

Figure 10: Optimal at K=33

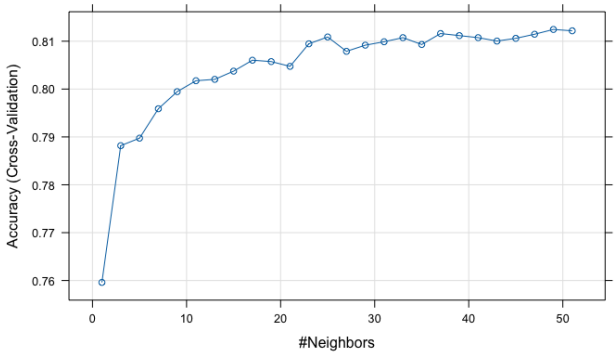

Figure 11: Optimal at K=49

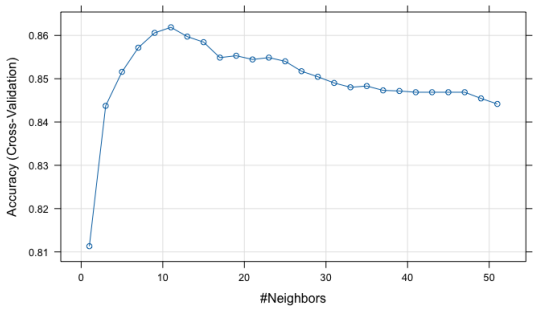

Figure 12: Optimal at K=11

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2275  113
         1  336  275

               Accuracy : 0.8503
                 95% CI : (0.837, 0.8629)
    No Information Rate : 0.8706
    P-Value [Acc > NIR] : 0.9995

                  Kappa : 0.466

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8713
            Specificity : 0.7088
         Pos Pred Value : 0.9527
         Neg Pred Value : 0.4501
             Prevalence : 0.8706
         Detection Rate : 0.7586
   Detection Prevalence : 0.7963
      Balanced Accuracy : 0.7900

       'Positive' Class : 0
```

Table 7: Confusion Matrix to evaluate a Classifier

	Predicted Class		
True Class	**0**	**1**	**Row Sum**
0	0.953	0.047	0.047
1	0.550	0.450	0.550
Col Sum	0.748	0.252	0.150

Table 8: proportional error matrix with predicted classes shown across columns

Note that for the first point, in table 9 probabilities associated with classes 0 and 1 are quite similar (47% vs 53%). So while we are quite confident about the predicted class of the second data point, there is some uncertainty about the first prediction! Note that for the second point, has about a 81% probability associated with class 0, and hence we are quite confident about our final class prediction of 0 that says customer will not churn.

However, for the second point, has about an 81% probability associated with class 0, and hence we are quite confident about our final class prediction of 0 that says customer will not churn.

Observation	Class 0	Class 1
1	0.4667	0.5333
2	0.8182	0.1818

Table 9: Predicted probabilities for two-class classification

2.6.2 Regularized Logistic Regression and Logistic Regression

So we want to explore the performance of the logistic regression for predicting whether a customer will churn or not so the first thing we did was to fit two different models using the LASSO regularization technique on the training set and then retrained the model using the optimal penalty parameter λ . One of the model has all the predictor variables included (CreditScore, Geography, Gender, Age, Tenure, Balance, Number of products, HasCredit-Card, IsActiveMember and Estimated salary) while the second included some variables(Age, Balance, EstimatedSalary, CreditScore, NumOfProducts, IsActiveMember) that we feel are strong predictors of whether a customer will churn or not.

This is a parametric model and it does have a tuning parameter λ, also we can use the model for variable selection since we regularized using lasso and also the glmnet handled the standardization of the predictors.

We implemented the regularization process using the 10-fold cross validation to select the optimal penalty parameter λ as seen in figure 13 and figure 14 which gave us 0.0152 and 0.0201 respectively using the 1-SE rule. The glmnet() automatically scales the predictors before estimating the regression coefficients, and then outputs the coefficients in the original scale. Both model 1 and model 2 did not actually perform well in detecting customers that actually churn (class 1), even though they both achieved a high accuracy.But model 1 is said to be preferred to model 2 since it has a higher balanced accuracy, higher specificity, higher kappa and lower p value than that of model 2. We went further to build two standard logistic regression model without the regularization using only the variables that were retained by the LASSO for model 1 as seen in table 10 to know whether there is going to be an improvement in the model.

Variable	Coefficient
(Intercept)	-3.575660
CreditScore	–
GeographyFrance	–
GeographyGermany	0.623145
GeographySpain	–
GenderMale	-0.289919
Age	0.060605
Tenure	–
Balance	1.14e-06
NumOfProducts	–
HasCrCard1	–
IsActiveMember1	-0.820441
EstimatedSalary	–

Table 10: Model Coefficients from LASSO Logistic Regression (Model 1)

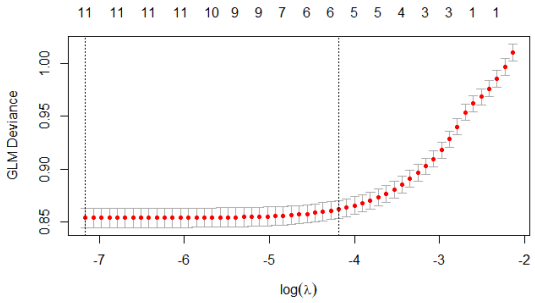

Figure 13: Cross-Validation Plot for LASSO Penalty Selection for model 1

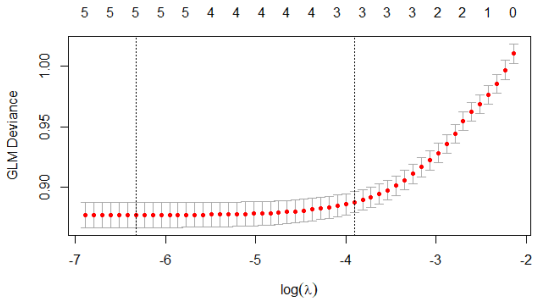

Figure 14: Cross-Validation Plot for LASSO Penalty Selection for model 2

Model 3 was built using Geography, Gender, Age, Balance, and IsActiveMember variables while Model 4 was built using Geography, Gender, Age, Balance, IsActiveMember and interaction between age and gender variables. The table 11 shows four differnt models that were fitted where model 1 and model 2 is the regularized logistic regression and model 2 is the standard logistic regression

Table 12 summarizes our final logistic model which is that of model 4(Geography, Gender, Age, Balance, IsActiveMember and interaction between age and gender variables). The result shows that being from Germany increases the log-odds of churn by 0.793 compared to France, while being male decreases the log-odds by 0.447. Each additional year of age increases the log-odds of churn by 0.075, but this effect is slightly less for males due to a negative interaction between Gender and Age. Being an active member significantly reduces the log-odds of churn by 1.119. Balance has a very small positive effect, and Spain shows little

Metric	Model 1	Model 2	Model 3	Model 4
Accuracy (%)	80.39	79.89	80.83	**80.89**
Kappa	0.13	0.07	0.21	**0.21**
Sensitivity(%)	98.16	**98.83**	96.78	96.82
Specificity (%)	10.97	5.89	18.49	**18.66**
Balanced Accuracy (%)	54.56	52.36	57.63	**57.74**
P-value (Accuracy > NIR)	0.15	0.37	0.05	**0.04**

Table 11: Comparison of Logistic Regression Models for Churn Prediction

difference from France. Overall, the model shows that age, geography, active membership, and account balance are all significant and are the major driving force of whether a customer will churn or not churn.

Variable	Estimate	Std. Error	z value	$\Pr(>\|z\|)$	Signif.
(Intercept)	-4.136	0.1907	-21.686	< 2e-16	***
GeographyGermany	0.7933	0.0807	9.827	< 2e-16	***
GeographySpain	0.0637	0.0845	0.754	0.4507	
GenderMale	-0.4468	0.2573	-1.736	0.0825	.
Age	0.0745	0.0042	17.619	< 2e-16	***
IsActiveMember1	-1.119	0.0695	-16.117	< 2e-16	***
Balance	2.67e-06	5.93e-07	4.496	6.93e-06	***
GenderMale:Age	-0.0016	0.0059	-0.273	0.7846	

Table 12: Logistic Regression Coefficient Summary (Model 4)

The confusion matrix in table 13 of our final model shows that the model 4 correctly identified 2312 customers that will not churn and also correctly identified 114 customers that will churn while 497 were incorrectly classified as churners when they are actually non churners and missed to classify 76 churner by predicting them as non churners.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2312  497
         1   76  114

               Accuracy : 0.8089
                 95% CI : (0.7944, 0.8229)
    No Information Rate : 0.7963
    P-Value [Acc > NIR] : 0.04373

                  Kappa : 0.2081

 Mcnemar's Test P-Value : < 2e-16

            Sensitivity : 0.9682
            Specificity : 0.1866
         Pos Pred Value : 0.8231
         Neg Pred Value : 0.6000
             Prevalence : 0.7963
         Detection Rate : 0.7709
   Detection Prevalence : 0.9366
      Balanced Accuracy : 0.5774

       'Positive' Class : 0
```

Table 13: Confusion Matrix to evaluate a Classifier

2.6.3 Generalized Additive models (GAMs) using Smoothing Splines

So for me to be able to explore the non linear relationships in some of the predictors, i fitted a gam model using smoothing splines. The response variable i used here is Exited which is binary that tells whether a customer will churn or not churn. This gam model is a semi parametric models because it combines both the parametric and non parametic (smooth functions). So we were able to handle both linear effect using the categorical variables (Geography, Gender, and IsActiveMember) while the complex non linear effect was handled using the predictors like Age, Balance, and NumOfProducts. I worked with three different models, one of the modlel has a smooth term of order 3, the second a smooth term of order 4 and the third a smooth term of order 5. So i fitted each of this model on the training set and assesed the performance on the test set. so after each model was fitted i assesed the performance of the model using the confusion matrix and area under the curve (AUC) of the ROC curve.

Six plots are shown in figure 15, 16 and 17 respectively , one for each predictor in the model. The categorical predictor, geography, shows box plots with Germany showing higher effect which means it has more likelihood of churn when compared to that of france and spain. The other three plots below show the polynomial fits for each predictor.These are smoothed lines in each case but that of Number of product shows its discrete nature. The figure 16 is chosen as the best since it gives the best balance for both smoothness and

overfitting.

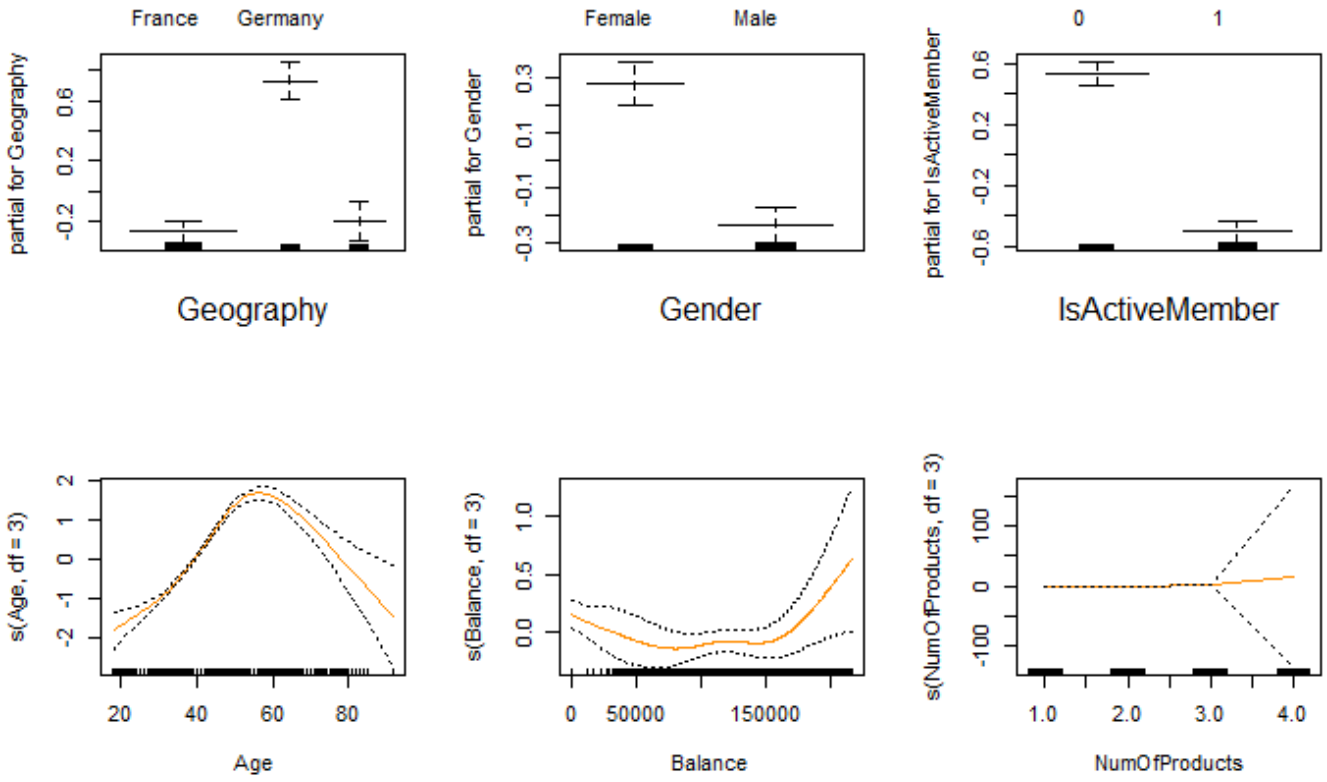

Figure 15: GAM model with smooth terms of degree 3

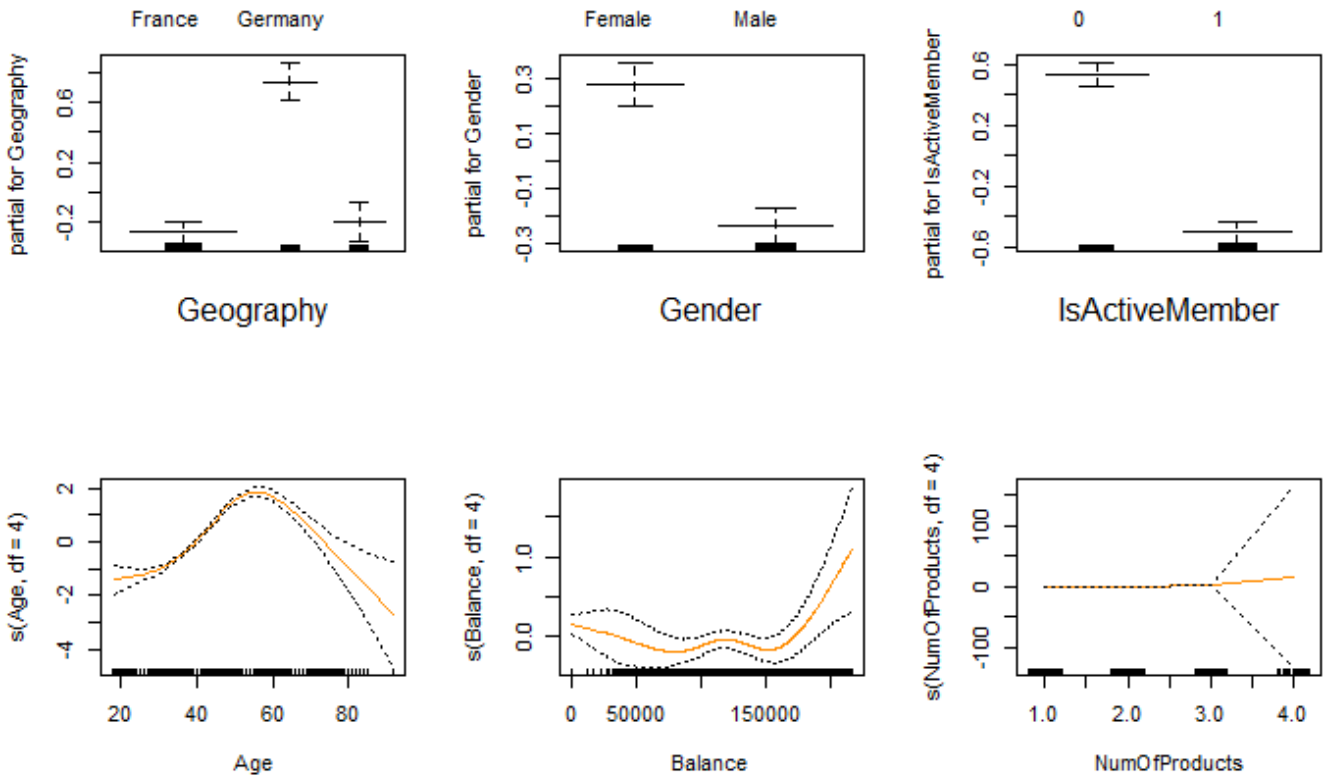

Figure 16: GAM model with smooth terms of degree 4

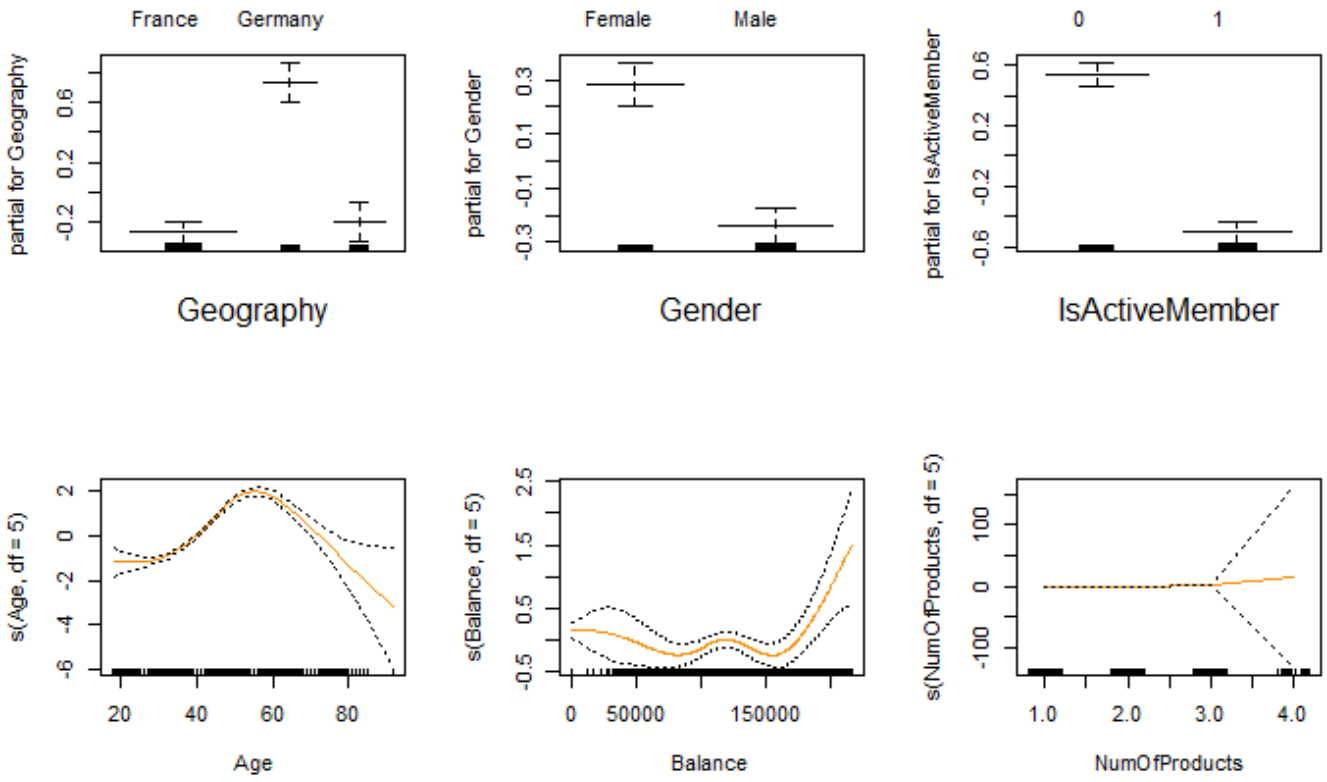

Figure 17: GAM model with smooth terms of degree 5

Metric	**df = 3**	**df = 4**	**df = 5**
Residual Deviance	4846.29	4809.31	4795.83
AIC	4874.29	4841.31	4831.83
Accuracy	84.93%	84.83%	84.73%
AUC	0.8418	0.8441	0.8453
Balanced Accuracy	69.22%	69.71%	69.58%
Sensitivity	95.73%	95.23%	95.14%
Specificity	42.72%	44.19%	44.03%
Kappa	0.4528	0.4574	0.4542

Table 14: Comparison of GAM models with varying smooth term degrees of freedom (df)

As seen in table 15, all the predictors also have parametric effect except for balance, also for the Nonparametric effects, we noticed that all three predictors have non-linear relationships with the response(Exited)

Parametric Effects (ANOVA)

Term	Df	Sum Sq	F value	Pr(>F)
Geography	2	109.2	50.61	< 2.2e-16
Gender	1	40.3	37.40	1.02e-09
IsActiveMember	1	76.0	70.42	< 2.2e-16
s(Age)	1	494.5	458.30	< 2.2e-16
s(Balance)	1	0.1	0.09	0.7653
s(NumOfProducts)	1	57.1	52.93	3.83e-13

Nonparametric Effects (Smooth Terms)

Term	Npar Df	Npar Chisq	P(Chi)
s(Age)	3	263.77	< 2.2e-16
s(Balance)	3	24.20	2.27e-05
s(NumOfProducts)	2	470.88	< 2.2e-16

Model Fit Statistics

Null Deviance	7077.58 on 7000 df
Residual Deviance	4809.31 on 6985 df
AIC	4841.31

Table 15: GAM Model Summary with Smoothing Splines (df = 4)

2.6.4 Decision trees(Single tree Model)

We used a training data set to fit a classification tree and this model is a non parametric model that makes use of recursive binary splitting inorder to partiition the predictors into different regions. So for the tree model it does an automatic variable selection and there is no need standardizing the predictors and this is not typically used for inference. So we set the complexity parameter to 0 and then did cross validation of 10 folds inorder to help us evaluate model performance and select the optimal tree size. The plot of the Cross-validated errors vs com- plexity parameter values.is shown in figure 18. The optimal cp value (using a 1 SE rule) obtained is 0.00421 as shown in table 16 which corresponded to a pruned tree that balanced complexity and predictive accuracy. So our final pruned tree for the churn data is shown in figure 19

CP	Number of Splits	Relative Error	X-Error	X-Std
0.07153	0	1.00000	1.00000	0.02363
0.03997	2	0.85694	0.85694	0.02227
0.03296	3	0.81697	0.82539	0.02194
0.02945	5	0.75105	0.76438	0.02127
0.02454	6	0.72160	0.72581	0.02083
0.01227	7	0.69705	0.70126	0.02053
0.00596	9	0.67251	0.68093	0.02028
0.00561	12	0.65288	0.67812	0.02025
0.00421	13	0.64727	0.66900	0.02013
0.00327	15	0.63885	0.67461	0.02020
0.00316	18	0.62903	0.67391	0.02019
0.00281	23	0.61150	0.67321	0.02018
0.00245	25	0.60589	0.67111	0.02016
0.00210	29	0.59607	0.66900	0.02013
0.00200	30	0.59397	0.67041	0.02015
0.00164	41	0.57083	0.67672	0.02023
0.00140	44	0.56592	0.68583	0.02034
0.00117	46	0.56311	0.69986	0.02051
0.00105	53	0.55470	0.70547	0.02058
0.00084	59	0.54839	0.71318	0.02068
0.00070	68	0.53366	0.74474	0.02105
0.00058	82	0.52104	0.74474	0.02105
0.00055	93	0.51122	0.75245	0.02114
0.00035	105	0.50351	0.76367	0.02127
0.00023	109	0.50210	0.76999	0.02134
0.00012	112	0.50140	0.77489	0.02139
0.00010	124	0.50000	0.77560	0.02140
0.00000	131	0.49930	0.77770	0.02142

Table 16: Cross-Validation Results for Classification Tree

The Root node error is: $1426/7001 = 0.20369$.

This shows that 20.37% of the training data would be misclassified.

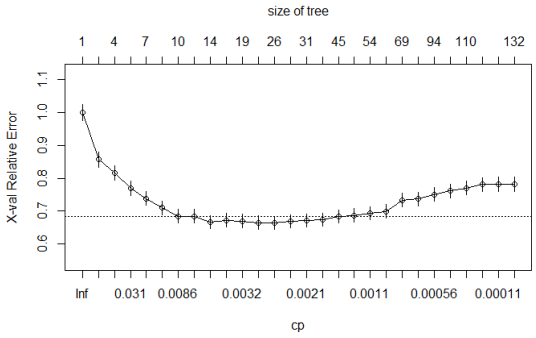

Figure 18: Cross-validated errors vs complexity parameter values.

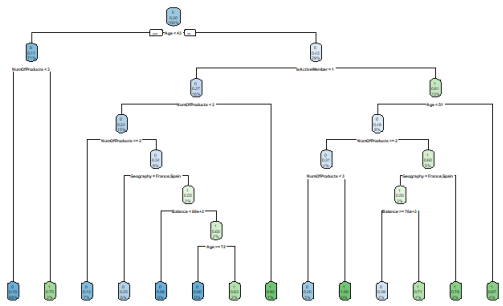

Figure 19: Final pruned tree for the Churn data

From table 17 we were able to correctly classify 95.7% of the customers as non churners and also correctly classified 43.86% as churners. 4.3% of the non churners were incorrectly classified as churners and 56.1% were incorrectly classified as churners when they are churners

	Predicted Class		
True Class	**0**	**1**	**Row Sum**
0	0.9569	0.0431	0.0431
1	0.5614	0.4386	0.5614
Col Sum	0.7691	0.2309	0.1487

Table 17: Relative Confusion Matrix for Pruned Classification Tree

From the result of the confusion matrix in table 18 we can see that we have an accuracy of 85.23% with kappa value of 0.4679.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2285  340
         1  103  271

               Accuracy : 0.8523
                 95% CI : (0.8391, 0.8648)
    No Information Rate : 0.7963
    P-Value [Acc > NIR] : 1.542e-15

                  Kappa : 0.4679

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9569
            Specificity : 0.4435
         Pos Pred Value : 0.8705
         Neg Pred Value : 0.7246
             Prevalence : 0.7963
         Detection Rate : 0.7619
   Detection Prevalence : 0.8753
      Balanced Accuracy : 0.7002

       'Positive' Class : 0
```

Table 18: Confusion Matrix for the decision tree

From the plot of the variable in fig 20 importance we can see that age, number of products, is active member and balance are important variables for predicting whether a customer will churn or not churn

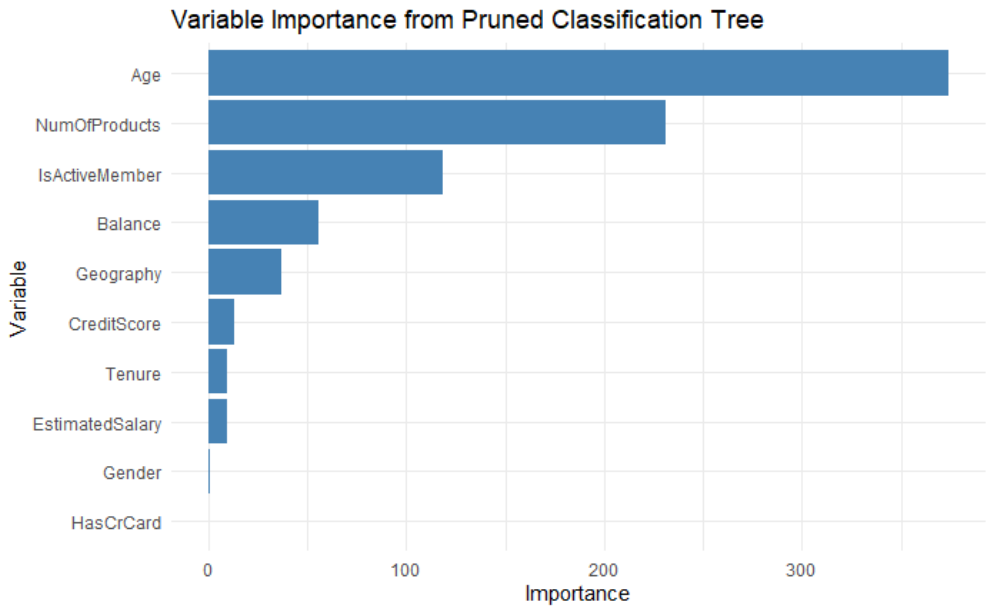

Figure 20: Variable importance for classification tree

2.6.5 Random Forest(Ensemble Tree)

We fitted a random forest to a training dataset and this is an ensemble learning technique. This method is a non parametric ensemble learning as it builds many decision trees and at the end aggregates the predictions obtained. This cannot be used for inference but for predictions and it performs variable selection. We do not need to standardize the predictor variables.

So for the random forest model we have a tuning parameter known as the mtry which is the number of parameter that are randomly selected during each split. We used differnt values of mtry which ranges from 1 to 10. For each value, the classification accuracy and Cohen's Kappa statistic were calculated on the test data. The selected mtry was found at 3 based on both accuracy and Kappa as seen in table 19 with an accuracy of 86.89% and kappa value of 0.5289

mtry	Accuracy	Kappa
1	0.8032	0.0526
2	0.8622	0.4783
3	**0.8689**	**0.5289**
4	0.8652	0.5214
5	0.8642	0.5200
6	0.8623	0.5139
7	0.8609	0.5121
8	0.8616	0.5173
9	0.8604	0.5116
10	0.8606	0.5108

Table 19: Random Forest Accuracy and Kappa for Different `mtry` Values

From the table in 20 and the figure in 21, we can see that age, number of products, creditscore, balance, estimated salary are very important variables for predicting whether a customer will churn or not churn

Variable	Importance Score
Age	100.000
NumOfProducts	53.643
CreditScore	51.520
Balance	50.888
EstimatedSalary	50.103
Tenure	26.014
IsActiveMember1	13.883
GeographyGermany	6.554
GenderMale	2.627
HasCrCard1	1.819
GeographySpain	0.000

Table 20: Variable Importance from Random Forest Model (`mtry = 3`)

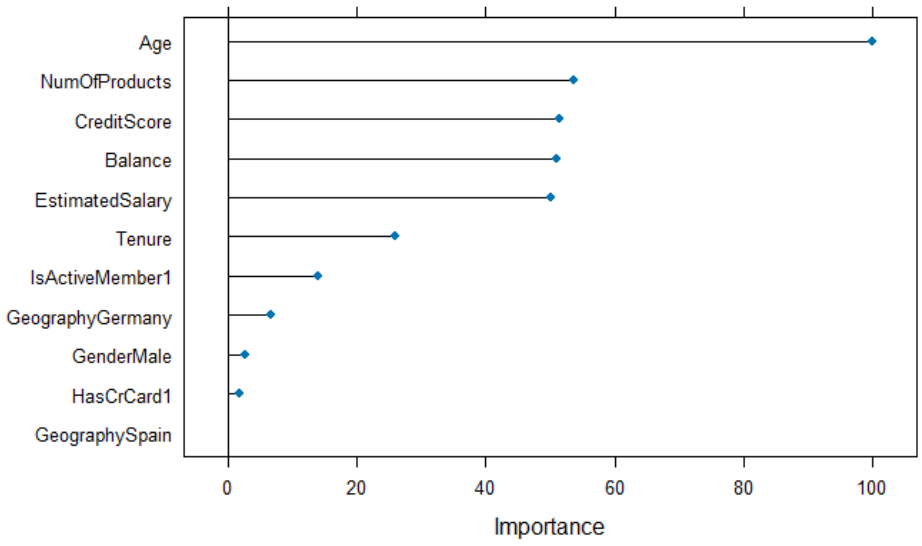

Figure 21: Variable importance for Random forest model

2.6.6 Support Vector Machine(SVM)

Here, we fitted a SVM with an svmRadial function kernel using the training dataset. SVM is a non parametric method and this is use strictly for prediction and not for inference. Although the SVM is robust to scaling it still performs that by default, it can also be used for variable selection since it does automatic variable selection. We tuned the model using a grid of cost and sigma values with the help of a 5 fold cross validation helping us select the one that best maximize accuracy. After getting that we retrained the model using the best Cost value and Sigma value. We then went ahead to evaluate the model performance using the test set and the confusion matrix is given

From table 21 we were able to correctly classify 97.15% of the customers as non churners and also correctly classified 38.95% as churners. 2.85% of the non churners were incorrectly classified as churners and 61.05% were incorrectly classified as churners when they are churners

	Predicted Class		
True Class	**0**	**1**	**Row Sum**
0	0.9715	0.0285	0.0285
1	0.6105	0.3895	0.6105
Col Sum	0.8458	0.1542	0.1470

Table 21: Relative Confusion Matrix for a Support Vector Machine with Radial Kernel

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2320  373
         1   68  238

               Accuracy : 0.853
                 95% CI : (0.8398, 0.8654)
    No Information Rate : 0.7963
    P-Value [Acc > NIR] : 6.975e-16

                  Kappa : 0.4434

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9715
            Specificity : 0.3895
         Pos Pred Value : 0.8615
         Neg Pred Value : 0.7778
             Prevalence : 0.7963
         Detection Rate : 0.7736
   Detection Prevalence : 0.8980
      Balanced Accuracy : 0.6805

       'Positive' Class : 0
```

Figure 22: Confusion matrix for SVM

The SVM achieved an accuracy of 85.30%

2.6.7 Best Model Selection

From table 22 we can see that the overall best model is that of random forest with an accuracy of 86.52% and kappa value of 0.5214.

Model	Accuracy	Kappa
kNN	85.03%	0.466
Logistic Regression	80.89%	0.2081
GAM	84.83%	0.4574
Decision tree(Single tree)	85.2%	0.4746
Random Forest	86.52%	0.5214
SVM	85.3%	0.4434

Table 22: Model Comparison based on model accuracy and Kappa

2.6.8 Final Model Fit on the entire dataset

So for the final model, we fitted a random forest to the entire dataset and this is an ensemble learning technique. This method is a non parametric ensemble learning as it builds many

decision trees. This cannot be used for inference but for predictions and it performs variable selection. We do not need to standardize the predictor variables.

So for the random forest model we have a tuning parameter known as the mtry which is the number of parameter that are randomly selected during each split. We used differnt values of mtry which ranges from 1 to 10. For each value, the classification accuracy and Cohen's Kappa statistic were calculated on the test data. The selected mtry was found at 3 based on both accuracy and Kappa as seen in table 23 with an accuracy of 86.51% and kappa value of 0.5104

mtry	Accuracy	Kappa
1	0.8024	0.0469
2	0.8627	0.4743
3	**0.8651**	**0.5104**
4	0.8624	0.5085
5	0.8609	0.5074
6	0.8597	0.5040
7	0.8590	0.5046
8	0.8578	0.5016
9	0.8566	0.4974
10	0.8567	0.4987

Table 23: Final Model(Random Forest) Accuracy and Kappa for Different `mtry` Values

Variable	Importance Score
CreditScore	51.794
GeographyGermany	6.987
GeographySpain	0.000
GenderMale	2.348
Age	100.000
Tenure	26.226
Balance	54.537
NumOfProducts	56.924
HasCrCard1	1.872
IsActiveMember1	13.377
EstimatedSalary	52.968

Table 24: Variable Importance from the final model (Random Forest) (`mtry = 3`)

From the table in 24 and the figure in 23, we can see that age, number of products, balance, estimated salary and credit score are very important variables for predicting whether a customer will churn or not churn

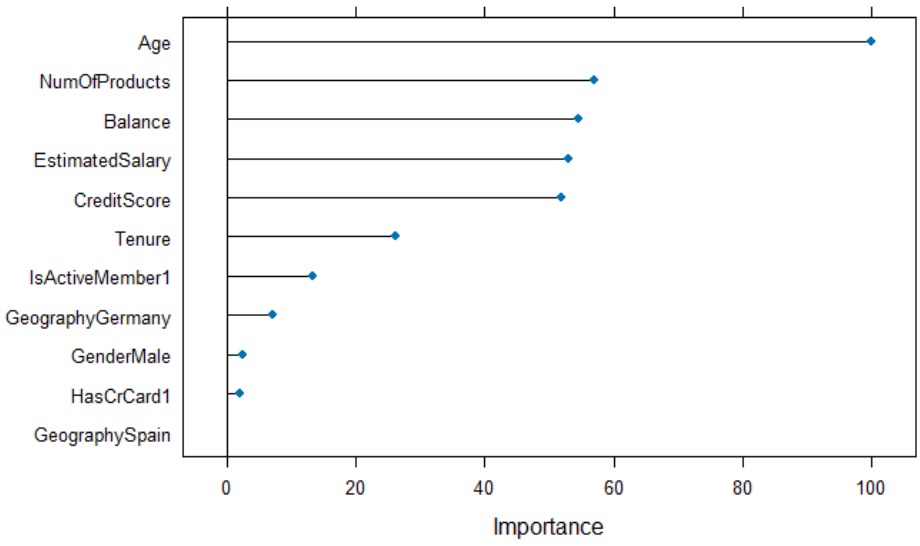

Figure 23: Variable importance for the Final model

2.6.9 Conclusion

In conclusion, we can see that the Random forest gave the most accurate prediction of customer churn, making it the best candidate. Though less interpretable than logistic models, it significantly outperformed others in terms of accuracy and Kappa. The predictors influencing whether a customer will churn or not includes different variables with age, number of products, balance, estimated salary and credit score having the highest influence.

2.6.10 Further Work

One limitation of my project is that class imbalance may affect the model's sensitivity to the minority class. So for future work i would explore resampling methods using SMOTE