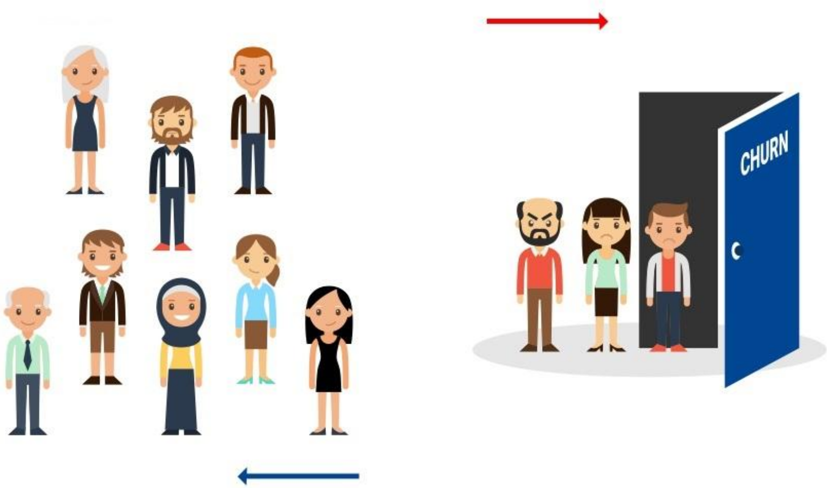
STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Predicting Customer Churn using Logistic Regression and Random Forest



Final project presentation

BIA 652-B

(MULTIVARIATE DATA ANALYSIS)

Under the guidance of Prof. Khasha Dehnad

Presented By:


Ameya Swar


Suguna Bontha


Rashmi Khurana


Rushabh Vakharia

Objectives/Business needs

- Customer churn has become one of the top issues for most banks.

- It costs significantly more to acquire new customers than it costs to retain existing ones, and it costs far more to re-acquire defected customers.

- In fact, several empirical studies and models have proven than churn remains one of the biggest destructors of enterprise value for banks and other customer intensive companies.

- Churn is important because it directly affects the organization's profitability and it is common to assume that the profitability of a service is directly related to the growth of its customer base.

- Using the available information in the dataset, we have used logistic regression and random forest to predict which customers are most likely to exit the bank in the near future.

Dataset Source and Variables

- The dataset was obtained from Kaggle and it had a total of 11458 records.
- The dataset was divided into a training dataset (10000 records) and a test dataset (1000 records).
- The remaining 458 records had to be removed since they either contained missing values or were duplicates.
- The dataset had a total of 14 variables/features.
- The dependent variable is “Exited”.

Dataset Source and Variables

- The list of variables is as follows:
  1. RowNumber (Numerical)
  2. CustomerId (Numerical)
  3. Surname (Categorical)
  4. CreditScore (Discrete Numerical)
  5. Geography (Categorical)
  6. Gender (Binary Categorical)
  7. Age (Discrete Numerical)
  8. Tenure (Discrete Numerical)
  9. Balance (Continuous Numerical)
  10. NumOfProducts (Discrete Numerical)
  11. HasCrCard (Binary Categorical)
  12. IsActiveMember (Binary Categorical)
  13. EstimatedSalary (Continuous Numerical)
  14. Exited (Binary Categorical)

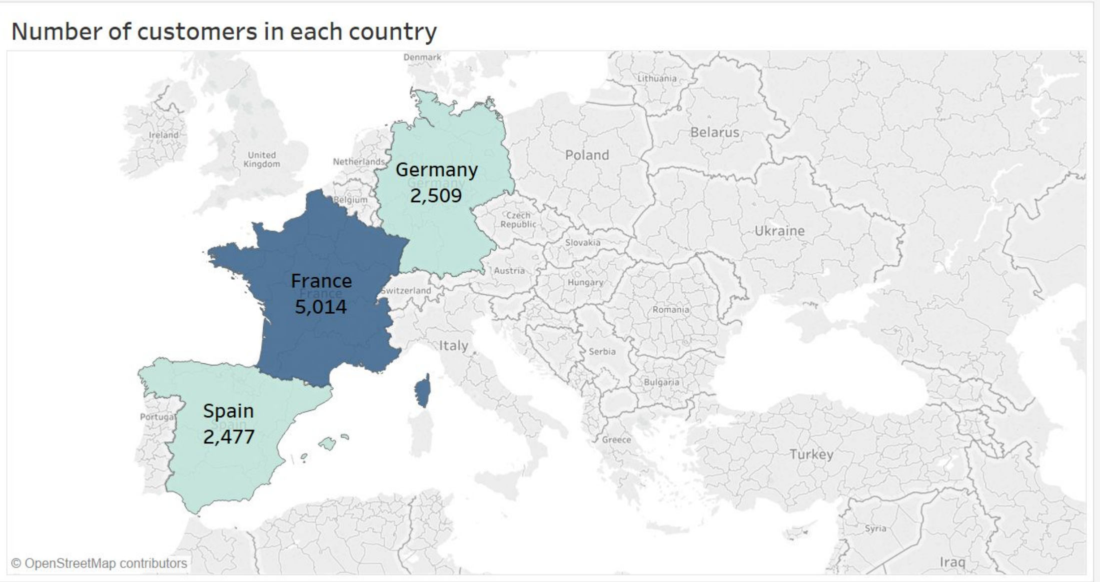
Tools Used for Analysis

- Multiple tools were used for the analysis.
- The dataset was initially in a .csv format and we used MS Excel to process the same.
- SSIS (SQL Server Integration Services) was used for the ETL (data pre-processing) stage.
- We made use of Tableau for creating visualizations, dashboards, and storyline.
- SAS and R were used for data modeling.
- We even used GRETL to do the prediction of data.

Exploratory Data Analysis

- In order to get an initial idea of the data, we created a few visualizations. Some of them are as follows:

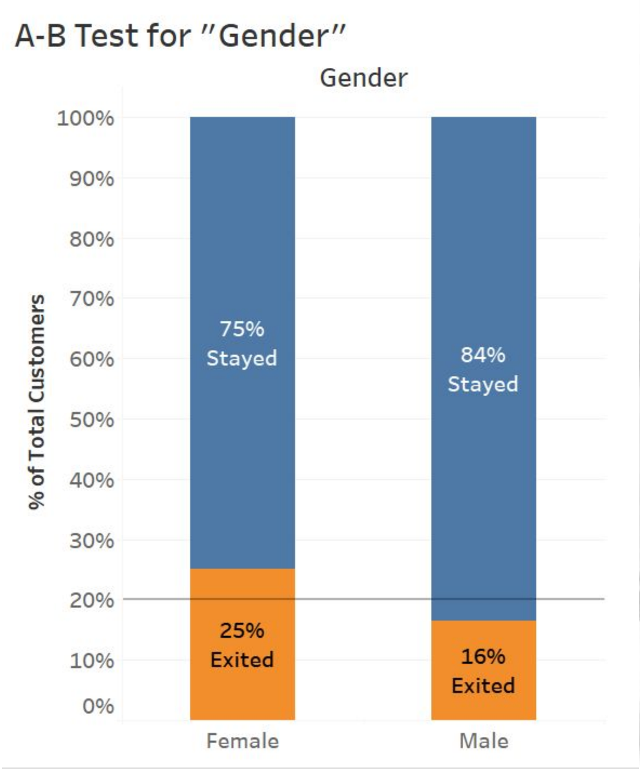
1. **Number of customers in each country.**



Takeaway: Distribution of customers across the three countries shows that France has the maximum number of customers followed by Germany and Spain.

Exploratory Data Analysis

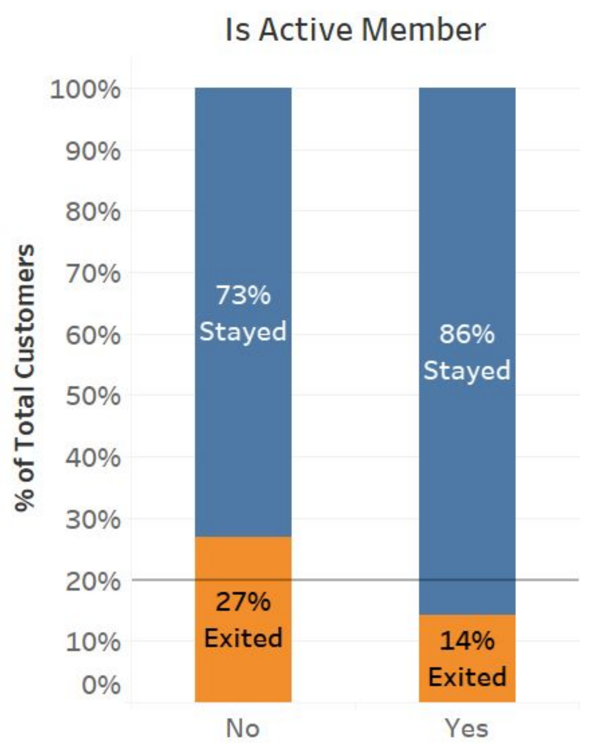
2. A-B Test for "Gender"



Takeaway: The graph shows that out of the total number of females, 25% exited the bank whereas only 16% of the total number of males exited in comparison.

Exploratory Data Analysis

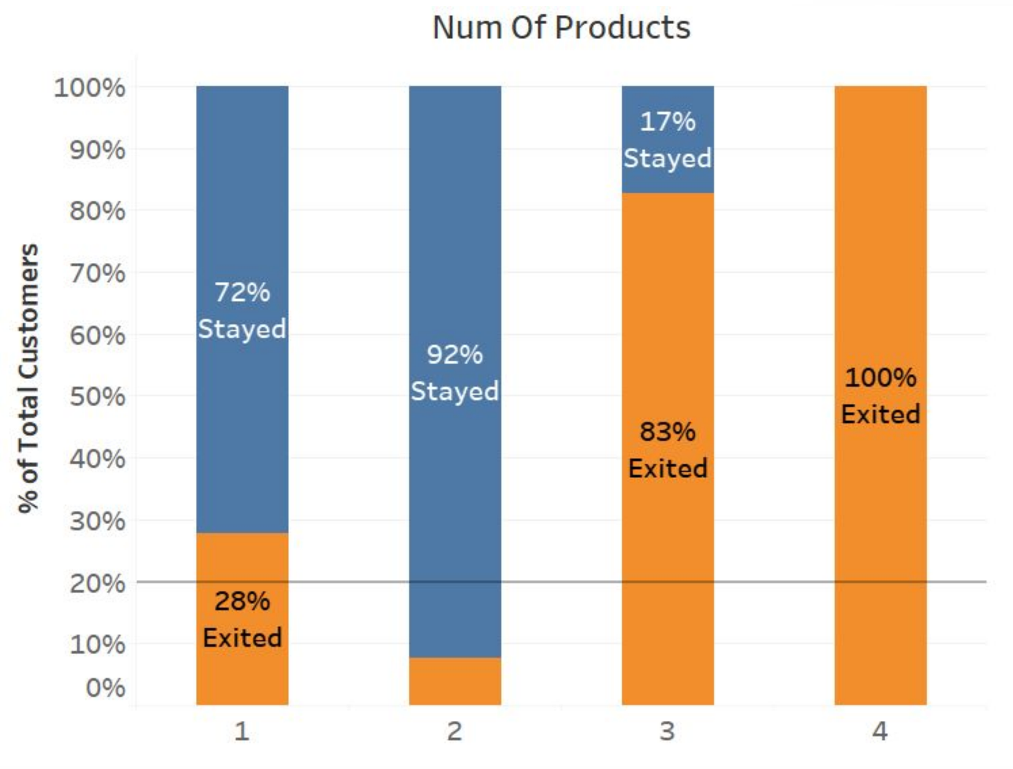
3. A-B Test for “IsActiveMember”



Takeaway: A-B Test for "IsActiveMember" shows that members who are active (make more number of transactions) are less likely to leave the bank.

Exploratory Data Analysis

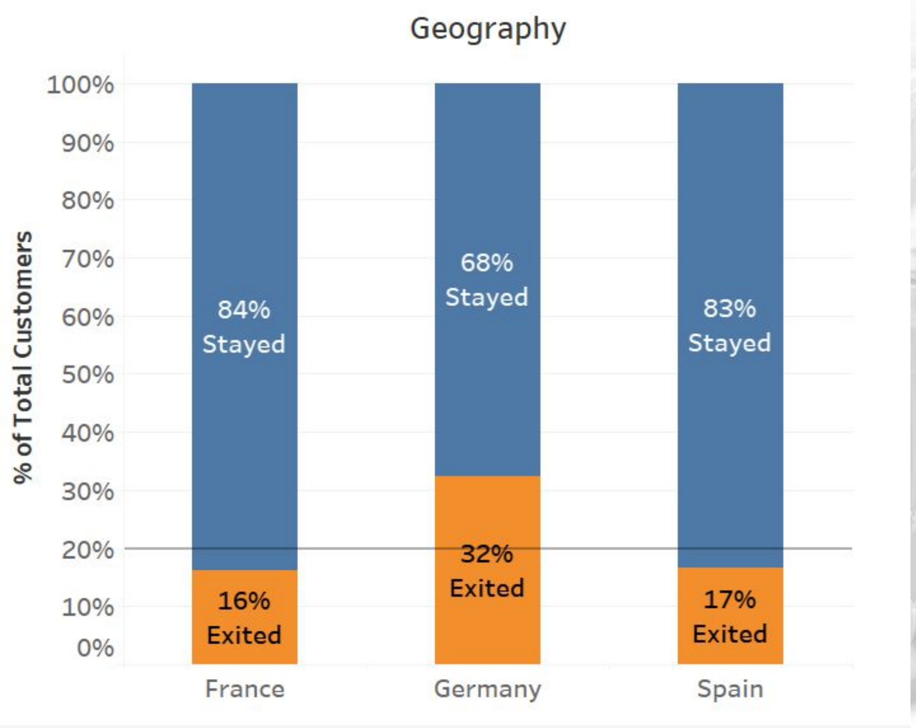
4. Classification Test for “NumOfProducts”



Takeaway: Classification Test for "NumOfProducts" shows that customers having 3 or 4 products with the bank are more likely to leave. But this result is insignificant since the number of observations are very less for the last 2 bars (data is skewed).

Exploratory Data Analysis

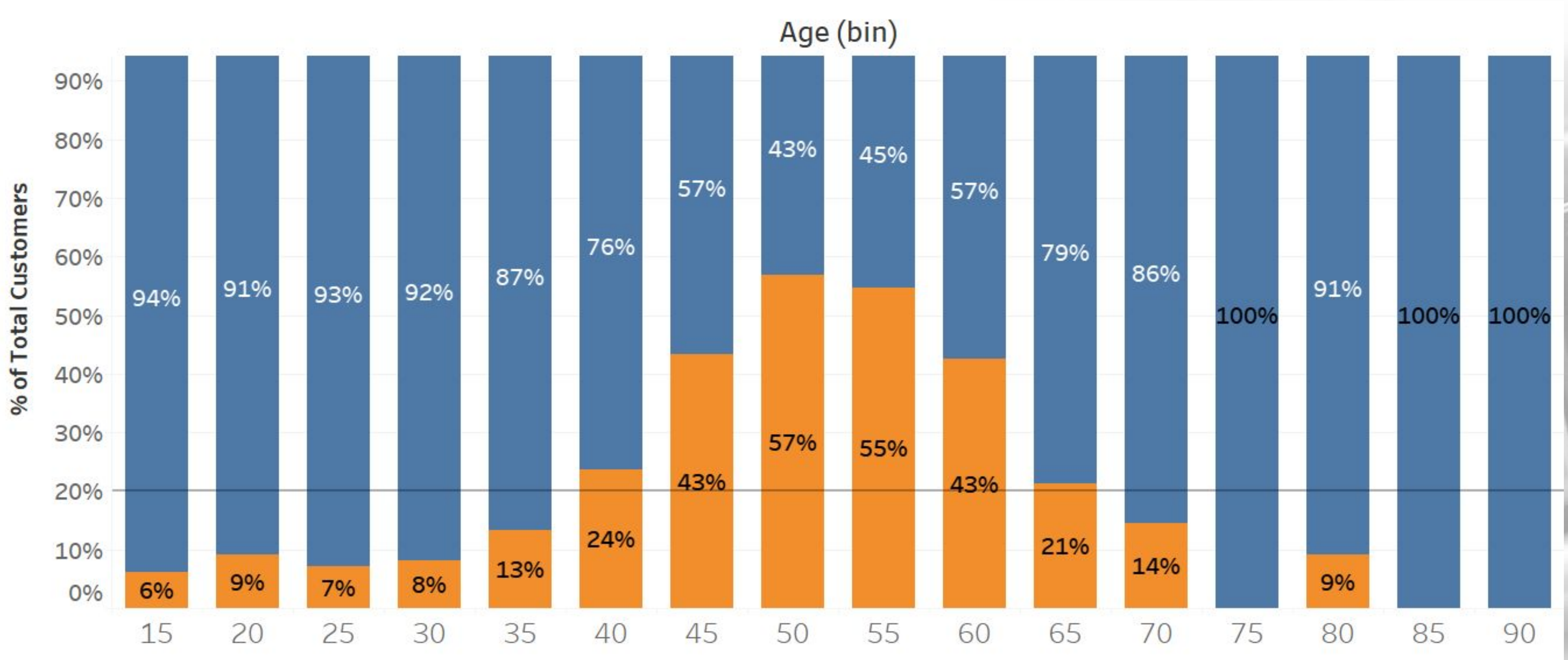
5. Classification Test for “Geography”



Takeaway: Classification Test for "Geography" shows that customers in Germany are more likely to leave.

Exploratory Data Analysis

6. Classification Test for “Age”



Takeaway: Classification Test for “Age” shows that people in the age group of 45 to 60 are more likely to leave the bank.

Principal Component Analysis

- Principal Component Analysis (PCA) is used to explain the variance-covariance structure of a set of variables through linear combinations. It is often used as a dimensionality-reduction technique.

- As can be seen on the next slide, the variables have very low correlation.

- The correlation between variables does not bring about a redundancy in the information that can be gathered by the dataset and thus, there is no need to use PCA to transform the original variables to the linear combination of variables which are independent.

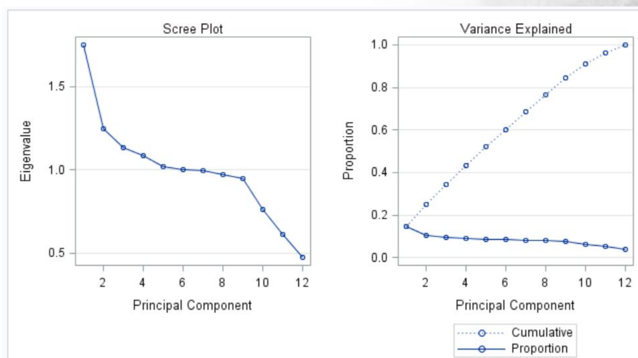
Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	1.75038598	0.50183591	0.1459	0.1459
2	1.24855007	0.11223849	0.1040	0.2499
3	1.13631158	0.05329479	0.0947	0.3446
4	1.08301679	0.06315015	0.0903	0.4349
5	1.01986663	0.01767278	0.0850	0.5198
6	1.00219385	0.00878234	0.0835	0.6034
7	0.99341151	0.02039710	0.0828	0.6861
8	0.97301440	0.02769193	0.0811	0.7672
9	0.94532248	0.18418878	0.0788	0.8460
10	0.76113369	0.14770536	0.0634	0.9094
11	0.61342833	0.14006365	0.0511	0.9606
12	0.47336469		0.0394	1.0000

Eigenvectors												
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12
Exited	0.363134	0.587698	0.006410	-.118944	0.033956	0.024325	0.030644	-.034973	-.168474	0.076410	0.679804	0.091070
CreditScore	-.015008	-.082584	-.011808	0.291790	-.082784	0.430912	0.794505	0.058870	-.276325	-.058692	0.024971	-.002603
Age	0.197647	0.575606	-.262522	0.321732	0.242458	0.135555	-.136482	-.041072	-.108009	-.174514	-.563500	-.007477
Tenure	-.023210	-.035316	0.164755	-.250545	0.469442	0.575445	-.000071	-.433189	0.409695	-.011157	0.032793	0.009043
Balance	0.538632	-.268435	-.259297	-.095725	-.042630	0.051653	0.040191	0.035677	0.077525	0.368768	-.174644	0.617694
NumOfProducts	-.261192	0.176569	0.631728	0.324877	0.012316	-.022743	-.091379	-.129596	-.161343	0.454944	-.103350	0.355635
HasCrCard	-.003651	-.046822	0.164137	-.128953	0.661679	-.069536	0.080502	0.706715	-.061719	0.036233	-.001931	0.022746
IsActiveMember	-.089420	-.157726	-.298350	0.719171	0.149141	0.081097	-.209682	0.107242	0.340752	0.094804	0.388875	0.017182
EstimatedSalary	0.021129	0.014731	0.131863	-.076856	-.410526	0.648985	-.421201	0.449973	-.063282	-.045876	-.007597	-.016583
Female	0.072736	0.329656	0.210548	0.010032	-.285011	-.131792	0.320623	0.257969	0.747662	-.012995	-.120260	0.005283
Germany	0.559972	-.169662	0.267650	0.165897	0.019989	-.012987	-.031736	-.056024	-.022140	0.353426	-.086225	-.648798
Spain	-.377683	0.217353	-.433745	-.228808	-.025269	0.104374	0.086982	0.069234	0.016979	0.692596	-.065353	-.247963

Principal Component Analysis

Pearson Correlation Coefficients, N = 10000
Prob > |r| under H0: Rho=0

	Exited	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Female	Germany	Spain
Exited	1.00000	-0.02709 0.0067	0.28532 <.0001	-0.01400 0.1615	0.11853 <.0001	-0.04782 <.0001	-0.00714 0.4754	-0.15613 <.0001	0.01210 0.2264	0.10651 <.0001	0.17349 <.0001	-0.05267 <.0001
CreditScore	-0.02709 0.0067	1.00000	-0.00396 0.6918	0.00084 0.9329	0.00627 0.5308	0.01224 0.2211	-0.00546 0.5852	0.02565 0.0103	-0.00138 0.8899	0.00286 0.7752	0.00554 0.5798	0.00478 0.6327
Age	0.28532 <.0001	-0.00396 0.6918	1.00000	-0.01000 0.3175	0.02831 0.0046	-0.03068 0.0022	-0.01172 0.2412	0.08547 <.0001	-0.00720 0.4715	0.02754 0.0059	0.04690 <.0001	-0.00169 0.8662
Tenure	-0.01400 0.1615	0.00084 0.9329	-0.01000 0.3175	1.00000	-0.01225 0.2205	0.01344 0.1789	0.02258 0.0239	-0.02836 0.0046	0.00778 0.4364	-0.01473 0.1407	-0.00057 0.9548	0.00387 0.6989
Balance	0.11853 <.0001	0.00627 0.5308	0.02831 0.0046	-0.01225 0.2205	1.00000	-0.30418 <.0001	-0.01486 0.1374	-0.01008 0.3133	0.01280 0.2007	-0.01209 0.2268	0.40111 <.0001	-0.13489 <.0001
NumOfProducts	-0.04782 <.0001	0.01224 0.2211	-0.03068 0.0022	0.01344 0.1789	-0.30418 <.0001	1.00000	0.00318 0.7503	0.00961 0.3365	0.01420 0.1555	0.02186 0.0288	-0.01042 0.2975	0.00904 0.3661
HasCrCard	-0.00714 0.4754	-0.00546 0.5852	-0.01172 0.2412	0.02258 0.0239	-0.01486 0.1374	0.00318 0.7503	1.00000	-0.01187 0.2354	-0.00993 0.3206	-0.00577 0.5642	0.01058 0.2903	-0.01348 0.1777
IsActiveMember	-0.15613 <.0001	0.02565 0.0103	0.08547 <.0001	-0.02836 0.0046	-0.01008 0.3133	0.00961 0.3365	-0.01187 0.2354	1.00000	-0.01142 0.2534	-0.02254 0.0242	-0.02049 0.0405	0.01673 0.0943
EstimatedSalary	0.01210 0.2264	-0.00138 0.8899	-0.00720 0.4715	0.00778 0.4364	0.01280 0.2007	0.01420 0.1555	-0.00993 0.3206	-0.01142 0.2534	1.00000	0.00811 0.4173	0.01030 0.3032	-0.00648 0.5169
Female	0.10651 <.0001	0.00286 0.7752	0.02754 0.0059	-0.01473 0.1407	-0.01209 0.2268	0.02186 0.0288	-0.00577 0.5642	-0.02254 0.0242	0.00811 0.4173	1.00000	0.02463 0.0138	-0.01689 0.0912
Germany	0.17349 <.0001	0.00554 0.5798	0.04690 <.0001	-0.00057 0.9548	0.40111 <.0001	-0.01042 0.2975	0.01058 0.2903	-0.02049 0.0405	0.01030 0.3032	0.02463 0.0138	1.00000	-0.33208 <.0001
Spain	-0.05267 <.0001	0.00478 0.6327	-0.00169 0.8662	0.00387 0.6989	-0.13489 <.0001	0.00904 0.3661	-0.01348 0.1777	0.01673 0.0943	-0.00648 0.5169	-0.01689 0.0912	-0.33208 <.0001	1.00000



Data Modeling

- We were not really sure if our data had a linear or non-linear decision boundary.

- Hence, we first decided to start with logistic regression, and then we tested out a random forest model.

- We later compared the performance of the two models using the ROC curves and other parameters.

Logistic Regression

- Logistic regression is a linear classifier, which makes it easier to interpret than non-linear models.
- At the same time, because it's a linear model, it has a high bias towards this type of fit, so it may not perform well on non-linear data.
- We developed a logistic regression model by splitting our dataset into a training set (10000 records), and test set (1000 records).
- We removed the CustomerId, RowNumber and Surname features because they were unique for each observation, and probably didn't add valuable information to our model.
- For our categorical variables "Geography" and "Gender", we created dummy variables. "Geography" was split into France, Germany and Spain. "Gender" was split into Male and Female.

Logistic Regression

- While modeling, we considered France and Male as the baseline and hence, they were omitted from analysis.

- We used the backward elimination method (MAXR option isn't available for logistic regression) to model our training data and the results were as follows:

Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	Spain	1	10	0.2486	0.6181
2	HasCrCard	1	9	0.5733	0.4489
3	EstimatedSalary	1	8	1.0346	0.3091
4	Tenure	1	7	2.9227	0.0873

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr> ChiSq
Intercept		1	3.9760	0.2312	295.8376	<.0001
CreditScore		1	0.000666	0.000280	5.6501	0.0175
Age		1	-0.0727	0.00257	797.3454	<.0001
Balance		1	-2.65E-6	5.139E-7	26.6299	<.0001
NumOfProducts		1	0.1010	0.0471	4.5985	0.0320
IsActiveMember	1	1	1.0718	0.0576	346.0674	<.0001
Female	1	1	-0.5306	0.0545	94.8958	<.0001
Germany	1	1	-0.7608	0.0633	144.3322	<.0001

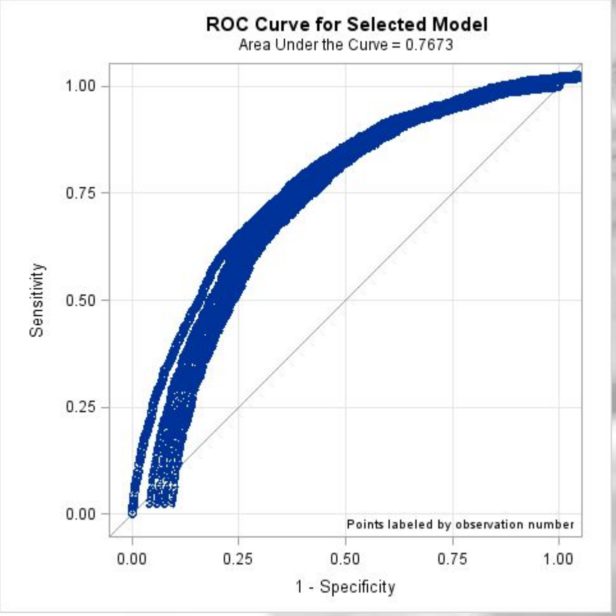
- The four variables removed were insignificant at the 5% level. Finally, we were left with only 7 factors which were influential.

Logistic Regression (Assessment)

- Using GRETL, we predicted the probability of customers exiting the bank. Probability of more than 50% was classified as 1 and less than 50% was classified as 0.

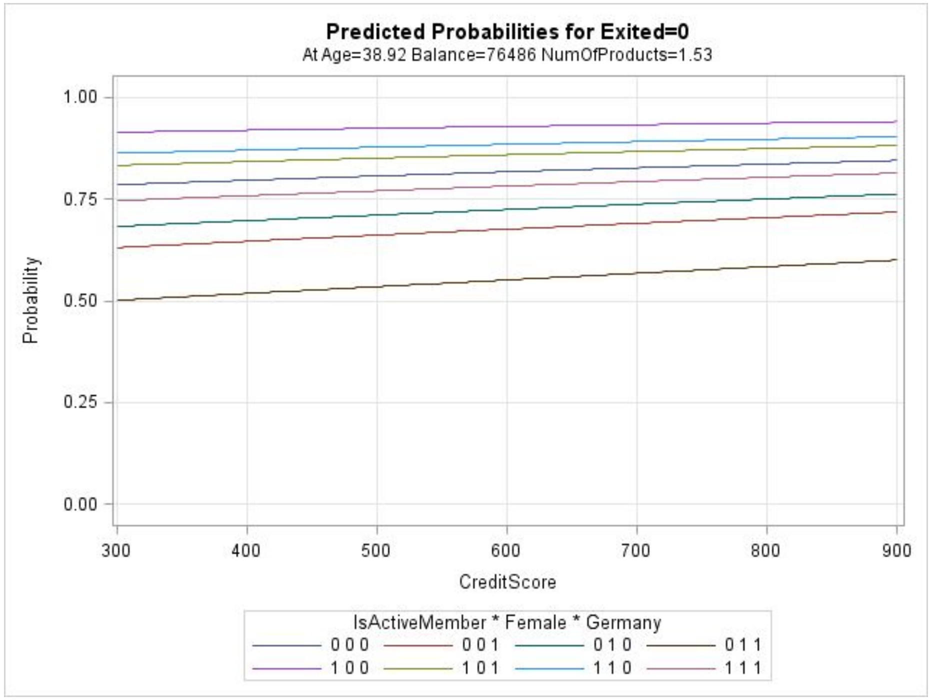
- The ROC Curve and the confusion matrix for our test data are as below.

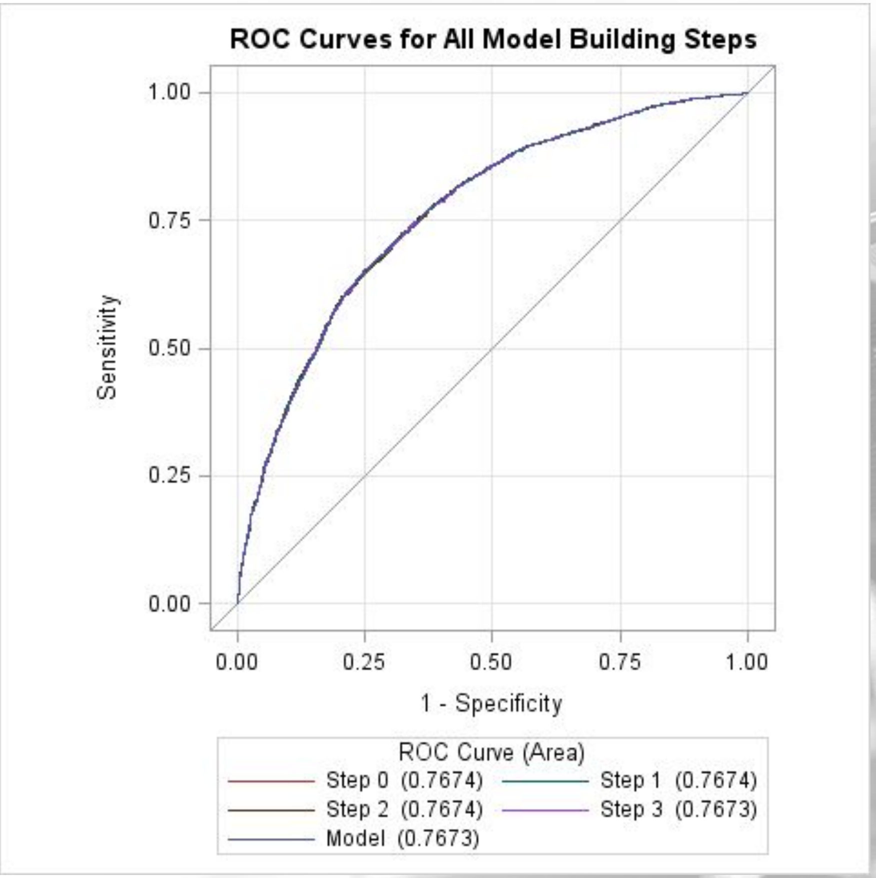
		Predicted	
		0	1
Actual	0	713	27
	1	216	44



- The model gave us an accuracy rate of 76% (as can be seen from the area under the curve)

Logistic Regression (Assessment)





Random Forest

- Random forest is another popular classification method.
- Unlike logistic regression, random forest is better at fitting non-linear data. It can also work well even if there are correlated features, which can be a problem for interpreting logistic regression.
- If random forest performs better than logistic regression, we can probably assume that there was non-linearity in data.
- Here, we see that the mean of squared residuals after fitting the model as 0.103551. It's quite small which means the fit of the model is quite tight.

```
Call:
 randomForest(formula = Exited ~ ., data = churn_train, importance = T)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 0.103551
                    % Var explained: 36.16
```

Random Forest (Assessment)

- Here, we see the confusion matrix and the random forest error rate.

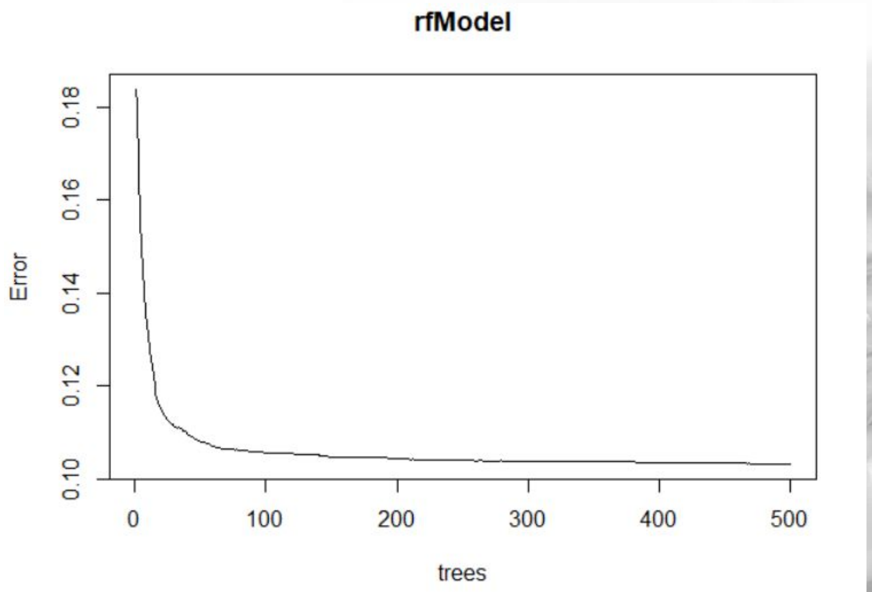
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 718 158
         1  22 102

               Accuracy : 0.82
                 95% CI : (0.7948, 0.8433)
    No Information Rate : 0.74
    P-Value [Acc > NIR] : 1.325e-09

                  Kappa : 0.4367
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9703
            Specificity : 0.3923
         Pos Pred Value : 0.8196
         Neg Pred Value : 0.8226
             Prevalence : 0.7400
         Detection Rate : 0.7180
```

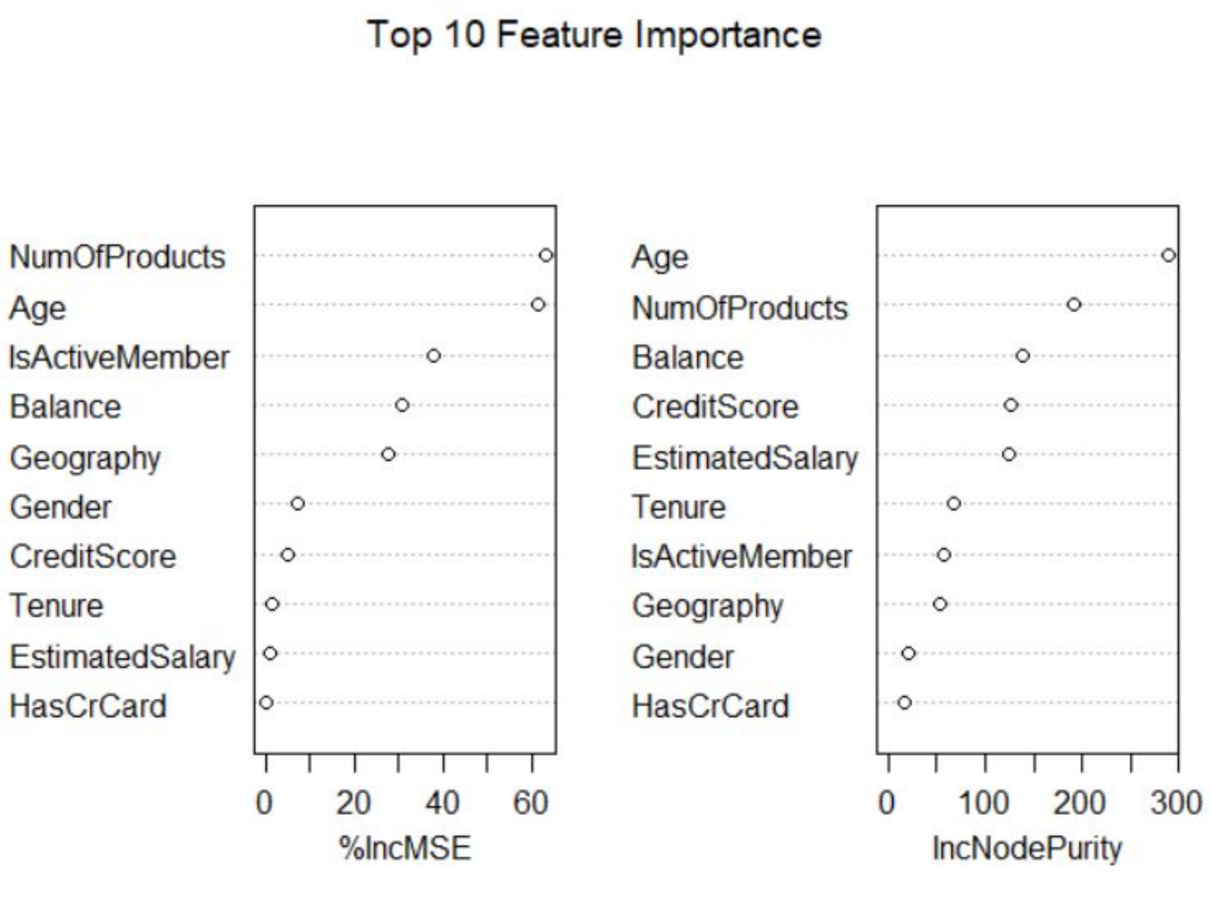


- We see that the accuracy is 82% which is better than logistic regression.

- We use this plot to help us determine the number of trees. As the number of trees increases, the OOB (out-of-bag) error rate decreases, and then becomes almost constant. We are not able to decrease the OOB error rate after about 100 to 150 trees.

Random Forest (Assessment)

- Here, we see the top 10 important features.



Conclusion

- We can see that Logistic Regression and Random Forest can be used for customer churn analysis for this particular dataset and Random Forest works a little better.

- Throughout the analysis, we have learnt several important things:

  - Features such as “HasCrCard”, “EstimatedSalary”, and “Spain” (“France”) aren’t really important and have to be removed.

  - Although, “Tenure” had to be removed, it actually makes sense to keep it in the model since the loyalty of a customer towards a bank usually increases with time.

  - Females who live in Germany, who are not active, who have only 1 or 2 products with the bank, is in the age group of 40-60, and has a low credit score is most likely to leave the bank.

THANK YOU