

Project:

*Analysis of factors causing a high rate of customers
leaving the bank*

Student: Darko Stankovic

Saint Joseph's University

DSS660 Introducing to Data Mining

Marth 5, 2018

Introduction and Objectives

For most business customers loyalty is the key to success. It is hard to hold customer under your nest if you do not have good product or service. Banking is on the top of the list having a hard time to get consumer loyalty and banks are very competitive worldwide. New global survey of 137,034 consumers in 21 countries by Bain & Company's ¹claims that about 29% of the bank customers globally said they would change primary bank if it were easy to do so.

Banks are collecting different data about their customers and try everything to keep every single of them. This project will analyze different features that banks use to determine their effect on the customer decides to leave the bank.

We are going to see how some statistical analysis can define the facts that bank can use to make better decisions. By using JMP as an analytical software from SAS, we are going to perform Analysis of Variance (ANOVA), and Logistic Regression on the dataset collected in Europe from one of the world banks.

¹ <https://www.forbes.com/sites/baininsights/2016/12/15/many-banks-are-losing-customers-and-dont-even-know-it/#15a187752935> (last visited Marth 5, 2018)

Data

The dataset that we are using for this project is publicly available at Kaggle.com². This dataset is used to make deep learning model and algorithm for the banking system to predict who from customers has potential to leave the bank.

We have fourteen variables in our dataset: Row Number, Customer ID, Surname, Credit Score, Geography, Gender, Age, Tenure, Balance, Number of Products, Has Credit Card, Is Active Member, Estimated Salary, Exited. Our dataset contains 10.000 rows of data.

Screenshot of dataset presented in Appendix 1.

Three binary variables (Has Credit Card, Is Active Member and Exited) and all other factors are a mix of numeric and character variables. The Geography factor contains data from three different countries France, Germany, and Spain. We do not have bank name, but we can see that bank is working in different European countries.

We do not have any missing data in our dataset. The only suspicious factor is "Balance" because we have a lot of "zero" elements, but since nowadays the everyday thing is to have zero balance in the bank we are going to use this data as accurate for our analysis.

² <https://www.kaggle.com/filippoo/deep-learning-az-ann/data> (Last visited Marth 5, 2018)

Analysis and Methodology

After making multivariate correlation matrix to check the correlation between variables we can state that high correlation of 0.3487 is present between Geographical data and Balance in the bank.

See Appendix A for complete correlation matrix.

First, we did One-way ANOVA analysis and confirmed high difference in the mean balances between Germany, Spain, and France.

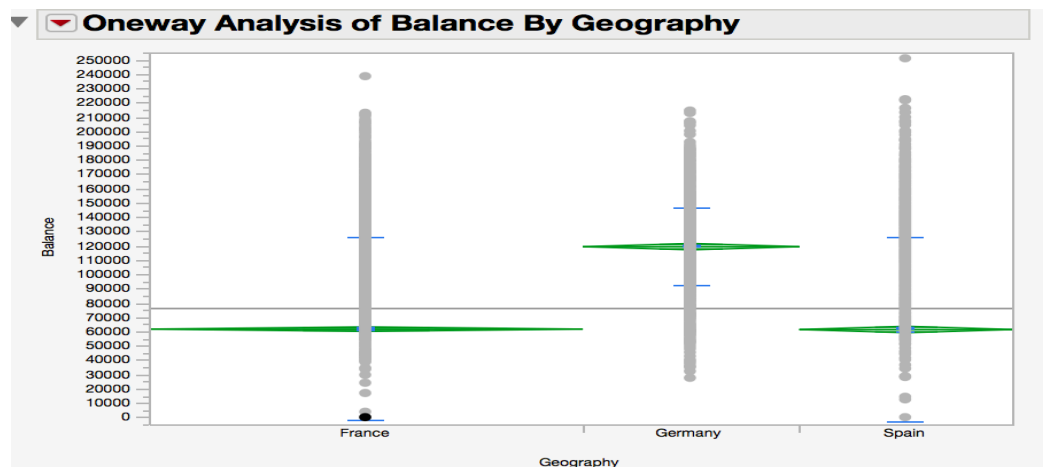
Our hypothesis test is:

Ho: All three countries have same mean money balance in the bank

H1: All three countries do not have same mean money balance in the bank

We can state that bank customers in Germany have on the average much higher amount of money on their bank account graphically comparing them to the customers from France and Spain (customers from France and Spain are on the almost same level).

Graphic is presented below:



Our p-value is <0.001 , and we are rejecting the null hypothesis which means that at least one of the courtiers have a different mean of the money balance in the bank.

See Appendix A for complete analysis and p-value.

Next confirmation of the above statement is coming from Tukey's Multiple Comparison analysis from JMP.

Connecting Letters Report is a good summary. It states that "Levels not connected by same letter are significantly different" which confirms that mean balance of customers from Germany differ from the mean balance of customers from France and Spain (have together B letter).


Report is presented below:

Connecting Letters Report		
Level		Mean
Germany	A	119730.12
France	B	62092.64
Spain	B	61818.15
Levels not connected by same letter are significantly different.		

The last report that we are using from one-way ANOVA analysis is Ordered Differences Report. This report is a very powerful and it shows the difference between levels as well as Lower and Upper Confidence interval level.

The report is presented below:

Ordered Differences Report						
Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
Germany	Spain	57911.97	1619.129	54116.6	61707.29	$<.0001^*$
Germany	France	57637.48	1397.886	54360.8	60914.20	$<.0001^*$
France	Spain	274.49	1403.891	-3016.3	3565.28	0.9791



We can interpret the difference between Germany and Spain by saying that we are 99% confident that the average balance in the bank in Germany is between \$54116.6 and \$61707.29 higher than the mean balance in the bank in Spain.

To confirm accretion of our dataset and given analysis we can make a simple financial comparison of GDP from 2016 between the three countries. The source for the information is from Destatis.de³ which is official European statistical source. We can see that people in Germany, in general, have more money and this confirms that our analysis is correct. The comparison table is given below:

Economy and finance

Country	Gross domestic product at market prices (GDP)	
	bn EUR	per inhabitant (PPS) ¹
	2016	
Germany	3,144	36,000
France	2,229	30,400
Spain	1,119	26,700

³ https://www.destatis.de/Europa/EN/Country/Comparison/GER_EU_Compared.html (Last visited Marth 5, 2018)

The second analysis that we are going to perform is Logistic Regression analysis (Chi-square and Logistic Regression).

The main point of this data collection was to find some way to lower the number of customers leaving the bank. The last column in our data set is "Exited," and it is binary variable represented with:

"0"- customer still with the bank

and

"1"- customer left the bank.

After recoding character variables to numeric (making Dummy variables) we are using the option in JMP to make Correlation Matrix between variables with a Scatter plot, and we can see that the highest correlation with "Exited" factor has next three variables:

1) Age (0.2853)

2) Geography (0.1538)

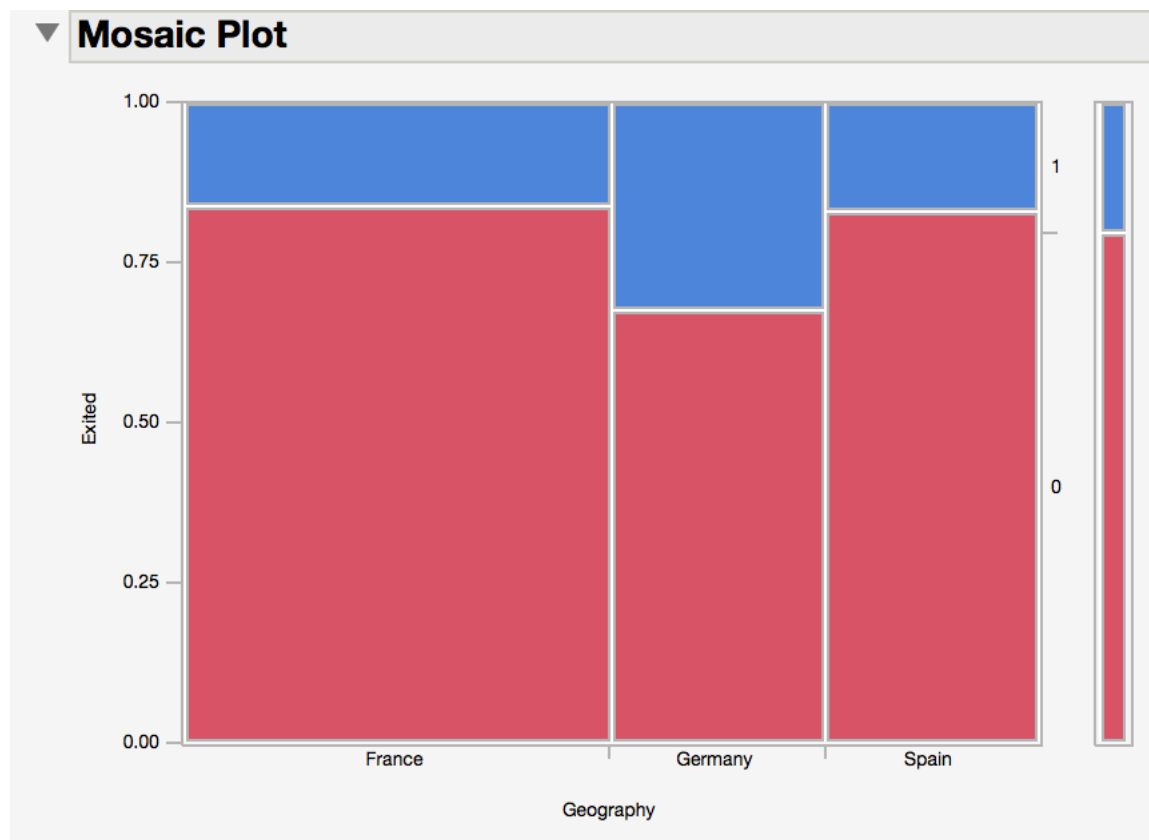
3) Balance (0.1185).

For complete Correlation Matrix and Scatter plot see Appendix B.

By using Fit Y by X, we are selecting as our dependent (y) categorical variable from the last column in our data set "Exited." Independent (x) variable is only categorical variable from three selected "Geographic."

From the Mosaic Plot given by JMP, we can state that Germany has a higher number of people leaving the bank compared to France and Spain.

See the Mosaic Plot bellow:



Besides the plot, JMP gives us Contingency Table as an output where we are finding that Likelihood Ratio is <0.0001 , and it is significant. We can conclude that country where the customer is located is related to the level of staying or leaving the bank.

The odds are showing that bank in Germany is almost losing one customer from every three customers that has.

The Contingency Table and odds explanation are presented in Appendix B.

For the rest of two factors (Age and Balance) we are running Logistic Regression by using Fit Model in JMP. Our y variable is binary variable (0 and 1) column "Exited". We are fitting model with the parameters from Age and Balance and running Logistic Regression.

We can see that both predictors significantly impact whether customer is leaving the bank or not. The parameter Estimates Table show odds of not leaving the bank. P-values for whole model is <0.0001 which means parameters are significant.

Parameter Estimates Table is presented in Appendix C.

The Unit Odds Ratios shows that odds Age ratio of 0.009219.
This can be interpreted as:

$$100 \times (0.009 - 1) = -99.1\%$$

For every increase in 1% Age, the odds of customer Not leaving (Staying = 0) decrease by 99.1%

The same can be calculated for odds Balance Ratio of 0.283469.

$$100 \times (0.283469 - 1) = -71.65\%$$

For every increase in 1% balance on the account, the odd of customer Not leaving (Staying = 0) decrease by 71.65%

If we want to make the prediction based on probability is it the customer going to leave the bank or not we can see that JMP is by default using 50:50 probability cut-off rule.

After saving probability formulas we have new extra 4 columns in our dataset that are presenting probability values.

The rule here is that any prediction above 50% is predicted to be 1 and opposite.

New columns values presented in Appendix C.

If we do cross tabulation on the Exited VS Most Likely Exited, we can see that we have 1916 false positives and 292 false negatives in the model.

This means that our model predicted that 1916 customers are going to leave the bank, but they did not and also 292 customers predicted to stay with a bank and they ended up leaving the bank.

Most Likely Exited	Exited	
	0	1
0	7671	1916
1	292	121

Now we are using better cut off by asking JMP to do ROC Curve which is used to give ideal cut off.

We can find in our ROC Table row selected with asterix. This is our cut off.

0.1995	0.3127	0.6912	0.3785 *	1408	5473	2490	629
0.1994	0.3128	0.6912	0.3784	1408	5473	2491	629

So, we can conclude that our cut of is much lower than 50%. It is actually 19.95%. This means that every prediction over 19.95% means that customer is going to leave the bank.

Discussion

All new modern way of payment, e-banking, and depositing checks to the bank virtually through the phone, tablet or computer force people to be away from the bank as a face to face customer service representative. Many factors can cause customer leaving the bank, but we have to be aware that when he goes from one bank, he is going to another bank. This is why bank system is so complicated and overflowed with factors that can cause losing the customers. Through our dataset example, we can conclude that bank, when making algorithm of potentially "running" customers, should consider customers bank account balance, age, and country where the bank branch is located.

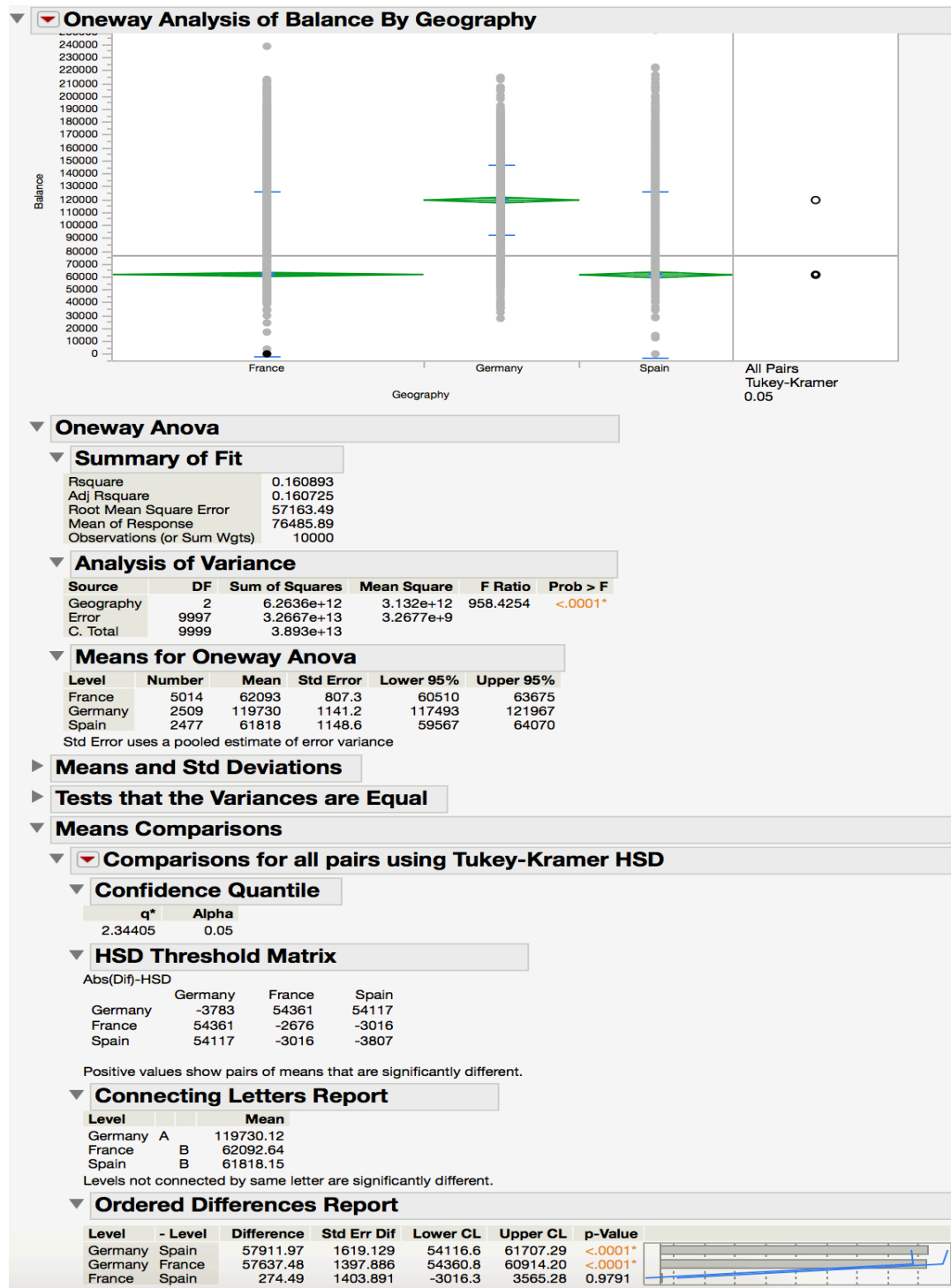
APPENDENCES

Appendix A:

JMP screen shoot of first 45 rows of dataset:

untitled 3													
untitled 3	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
1	15634602	Hargrave	619	France	Female	42	2	0	1 1	1		101348.88	1
2	15647311	Hill	608	Spain	Female	41	1	83807.86	1 0	1		112542.58	0
3	15619304	Onio	502	France	Female	42	8	159660.8	3 1	0		113931.57	1
4	15701354	Boni	699	France	Female	39	1	0	2 0	0		93826.63	0
5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1 1	1		79084.1	0
6	15574012	Chu	645	Spain	Male	44	8	113755.78	2 1	0		149756.71	1
7	15592531	Bartlett	822	France	Male	50	7	0	2 1	1		10062.8	0
8	15656148	Obinna	376	Germany	Female	29	4	115046.74	4 1	0		119346.88	1
9	15792365	He	501	France	Male	44	4	142051.07	2 0	1		74940.5	0
10	15592389	H?	684	France	Male	27	2	134603.88	1 1	1		71725.73	0
11	15767821	Bearce	528	France	Male	31	6	102016.72	2 0	0		80181.12	0
12	15737173	Andrews	497	Spain	Male	24	3	0	2 1	0		76390.01	0
13	15632264	Kay	476	France	Female	34	10	0	2 1	0		28260.98	0
14	15691483	Chin	549	France	Female	25	5	0	2 0	0		190857.79	0
15	15600882	Scott	635	Spain	Female	35	7	0	2 1	1		65951.65	0
16	15643966	Goforth	616	Germany	Male	45	3	143129.41	2 0	1		64327.26	0
17	15737452	Romeo	653	Germany	Male	58	1	132602.88	1 1	0		5097.67	1
18	15788218	Henderson	549	Spain	Female	24	9	0	2 1	1		14406.41	0
19	15661507	Muldrow	587	Spain	Male	45	6	0	1 0	0		158684.81	0
20	15568982	Hao	726	France	Female	24	6	0	2 1	1		54724.03	0
21	15577657	McDonald	732	France	Male	41	8	0	2 1	1		170886.17	0
22	15597945	Dellucci	636	Spain	Female	32	8	0	2 1	0		138555.46	0
23	15699309	Gerasimov	510	Spain	Female	38	4	0	1 1	0		118913.53	1
24	15725737	Mosman	669	France	Male	46	3	0	2 0	1		8487.75	0
25	15625047	Yen	846	France	Female	38	5	0	1 1	1		187616.16	0
26	15738191	Maclean	577	France	Male	25	3	0	2 0	1		124508.29	0
27	15736816	Young	756	Germany	Male	36	2	136815.64	1 1	1		170041.95	0
28	15700772	Nebechi	571	France	Male	44	9	0	2 0	0		38433.35	0
29	15728693	McWilliams	574	Germany	Female	43	3	141349.43	1 1	1		100187.43	0
30	15656300	Lucciano	411	France	Male	29	0	59697.17	2 1	1		53483.21	0
31	15589475	Azikiwe	591	Spain	Female	39	3	0	3 1	0		140469.38	1
32	15706552	Odinakachuk...	533	France	Male	36	7	85311.7	1 0	1		156731.91	0
33	15750181	Sanderson	553	Germany	Male	41	9	110112.54	2 0	0		81898.81	0
34	15659428	Maggard	520	Spain	Female	42	6	0	2 1	1		34410.55	0
35	15732963	Clements	722	Spain	Female	29	9	0	2 1	1		142033.07	0
36	15794171	Lombardo	475	France	Female	45	0	134264.04	1 1	0		27822.99	1
37	15788448	Watson	490	Spain	Male	31	3	145260.23	1 0	1		114066.77	0
38	15729599	Lorenzo	804	Spain	Male	33	7	76548.6	1 0	1		98453.45	0
39	15717426	Armstrong	850	France	Male	36	7	0	1 1	1		40812.9	0
40	15585768	Cameron	582	Germany	Male	41	6	70349.48	2 0	1		178074.04	0
41	15619360	Hsiao	472	Spain	Male	40	4	0	1 1	0		70154.22	0
42	15738148	Clarke	465	France	Female	51	8	122522.32	1 0	0		181297.65	1
43	15687946	Osborne	556	France	Female	61	2	117419.35	1 1	1		94153.83	0
44	15755196	Lavine	834	France	Female	49	2	131394.56	1 0	0		194365.76	1
45	15684171	Bianchi	660	Spain	Female	61	5	155931.11	1 1	1		158338.39	0

One-way ANOVA:

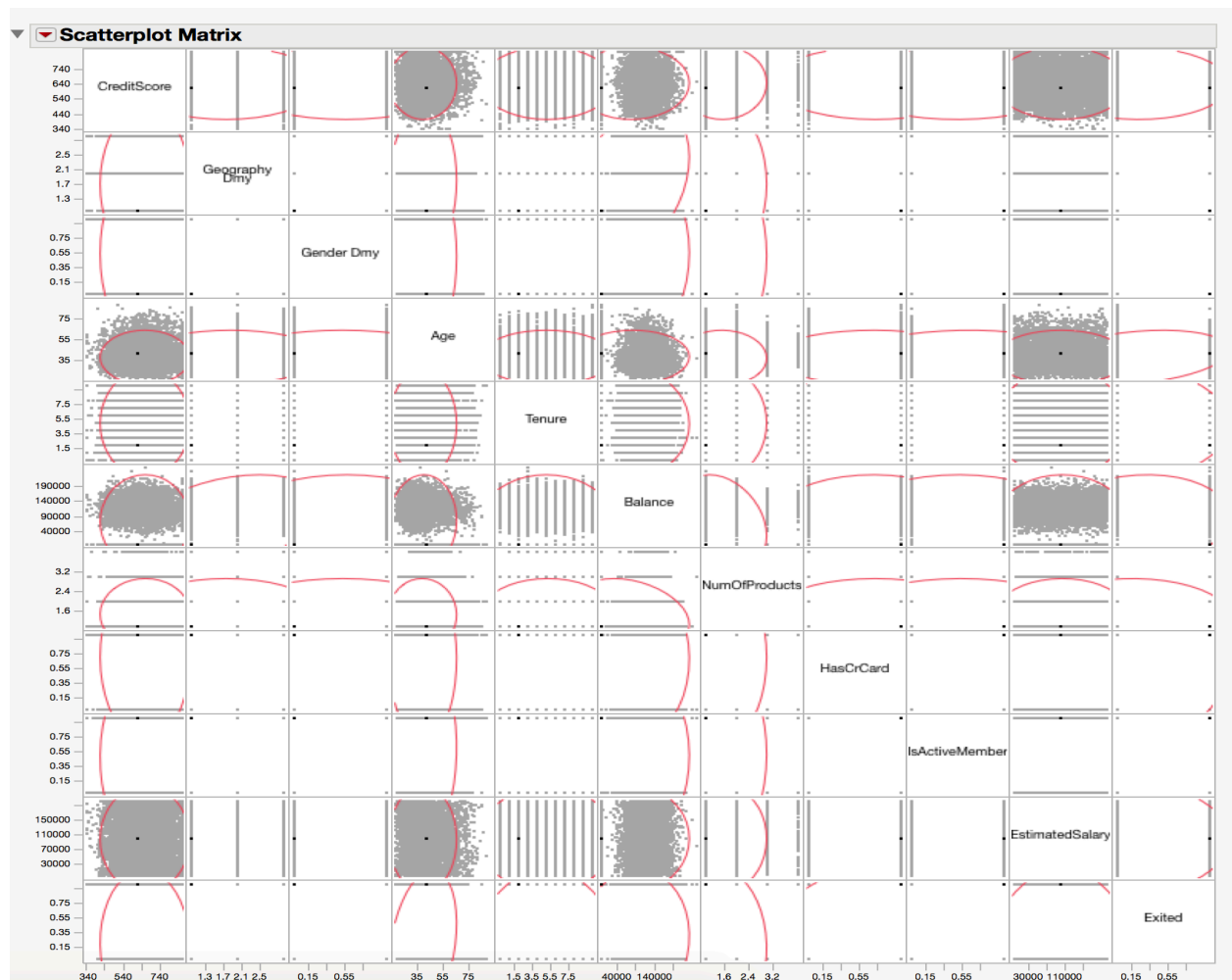


Appendix B:

Correlation Matrix:

Multivariate											
Correlations											
	CreditScore	Geography Dmy	Gender Dmy	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
CreditScore	1.0000	0.0083	-0.0029	-0.0040	0.0008	0.0063	0.0122	-0.0055	0.0257	-0.0014	-0.0271
Geography Dmy	0.0083	1.0000	-0.0169	0.0481	0.0014	0.3487	-0.0062	0.0040	-0.0127	0.0074	0.1538
Gender Dmy	-0.0029	-0.0169	1.0000	-0.0275	0.0147	0.0121	-0.0219	0.0058	0.0225	-0.0081	-0.1065
Age	-0.0040	0.0481	-0.0275	1.0000	-0.0100	0.0283	-0.0307	-0.0117	0.0855	-0.0072	0.2853
Tenure	0.0008	0.0014	0.0147	-0.0100	1.0000	-0.0123	0.0134	0.0226	-0.0284	0.0078	-0.0140
Balance	0.0063	0.3487	0.0121	0.0283	-0.0123	1.0000	-0.3042	-0.0149	-0.0101	0.0128	0.1185
NumOfProducts	0.0122	-0.0062	-0.0219	-0.0307	0.0134	-0.3042	1.0000	0.0032	0.0096	0.0142	-0.0478
HasCrCard	-0.0055	0.0040	0.0058	-0.0117	0.0226	-0.0149	0.0032	1.0000	-0.0119	-0.0099	-0.0071
IsActiveMember	0.0257	-0.0127	0.0225	0.0855	-0.0284	-0.0101	0.0096	-0.0119	1.0000	-0.0114	-0.1561
EstimatedSalary	-0.0014	0.0074	-0.0081	-0.0072	0.0078	0.0128	0.0142	-0.0099	-0.0114	1.0000	0.0121
Exited	-0.0271	0.1538	-0.1065	0.2853	-0.0140	0.1185	-0.0478	-0.0071	-0.1561	0.0121	1.0000

Scatter plot:



▼

Contingency Table

		Exited		
	Count	0	1	Total
Geography	Total %			
	Col %			
	Row %			
	France	4204	810	5014
		42.04	8.10	50.14
		52.79	39.76	
		83.85	16.15	
	Germany	1695	814	2509
		16.95	8.14	25.09
		21.29	39.96	
		67.56	32.44	
	Spain	2064	413	2477
		20.64	4.13	24.77
		25.92	20.27	
		83.33	16.67	
	Total	7963	2037	10000
	79.63	20.37		

▼

Tests

N	DF	-LogLike	RSquare (U)
10000	2	140.17046	0.0277

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	280.341	<.0001*
Pearson	301.255	<.0001*

4204/810=5.19 For France - on every 5.19 customers that stays with the bank one customer is leaving the bank.

$1695/814=2.08$ For Germany - on every 2.08 customers that stays with the bank one customer is leaving the bank.

2064/413= 4.99 For Spain- on every 4.99 customers that stays with the bank one customer is leaving the bank.

15

Appendix C:

Parameter Estimates Table:

Parameter Estimates				
Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	4.35876893	0.1124876	1501.5	<.0001*
Age	-0.0633306	0.0023837	705.89	<.0001*
Balance	-5.0246e-6	4.3317e-7	134.55	<.0001*

For log odds of 0/1

Predicted column in dataset:

tedSalary	Exited	Lin[0]	Prob[0]	Prob[1]	Most Likely Exited
61048.53	0	1.864817436	0.8658574719	0.1341425281	0
111981.19	0	-0.201034155	0.4499100453	0.5500899547	1
62232.6	0	1.8255449952	0.8612301532	0.1387698468	0
156105.03	0	2.0155367905	0.8824187145	0.1175812855	0
150135.38	0	2.4906456678	0.9234834392	0.0765165608	0
93146.11	0	-0.457075208	0.3876798948	0.6123201052	1
33462.94	1	1.1956202136	0.7677447277	0.2322552723	0
144375	0	2.5221815782	0.9256822751	0.0743177249	0
55803.96	1	1.0515964558	0.7410813447	0.2589186553	0
89048.46	1	1.4455614044	0.8093143897	0.1906856103	0
162812.16	0	2.1776494892	0.8982243957	0.1017756043	0
136259.65	0	1.9522061921	0.8756870054	0.1243129946	0
124052.97	0	1.3019880244	0.7861693742	0.2138306258	0
74158.8	0	1.4831260916	0.8150442941	0.1849557059	0
45071.09	1	-0.474322659	0.3835936398	0.6164063602	1
62030.06	0	2.2055285859	0.9007448823	0.0992551177	0
131953.23	1	1.2050241343	0.769417343	0.230582657	0
79414	0	1.9464280395	0.8750566333	0.1249433667	0
199638.56	0	2.7121733736	0.9377411568	0.0622588432	0
157577.29	0	2.3208133539	0.9105861852	0.0894138148	0
134420.75	1	0.3037212303	0.5753519495	0.4246480505	0
99805.99	0	1.6877210572	0.8439242216	0.1560757784	0
97932.68	0	2.2055285859	0.9007448823	0.0992551177	0
170968.99	0	-0.644348345	0.3442642527	0.6557357473	1
57558.95	0	1.614919115	0.8340932212	0.1659067788	0
176713.47	0	2.5855121766	0.9299233276	0.0700766724	0
34283.23	0	1.7622143967	0.8534867809	0.1465132191	0
198637.34	0	2.2055285859	0.9007448823	0.0992551177	0
156917.12	0	2.0182800087	0.8827030412	0.1172969588	0
50457.2	0	1.8157463849	0.8600549437	0.1399450563	0
140075.55	0	1.6763528354	0.8424209822	0.1575790178	0
65323.11	0	1.1524726684	0.7599622701	0.2400377299	0
64323.24	0	0.8169156745	0.6935812315	0.3064187685	0
14956.44	0	1.4042220435	0.8028530084	0.1971469916	0
111879.21	0	2.156185736	0.8962453974	0.1037546026	0
18606.23	0	1.9984807444	0.8806374736	0.1193625264	0
123137.01	0	1.5802276043	0.8292367499	0.1707632501	0
148528.24	0	2.2055285859	0.9007448823	0.0992551177	0
13898.31	0	2.078867389	0.8888321698	0.1111678302	0
81259.25	1	1.8888755936	0.8686272733	0.1313727267	0
161051.75	0	1.8787292137	0.8674650931	0.1325349069	0
191599.67	0	1.7622143967	0.8534867809	0.1465132191	0
48559.19	1	1.7013745991	0.8457141802	0.1542858198	0
186339.74	0	1.4553760658	0.8108244372	0.1891755628	0
33159.37	0	2.4588509797	0.9212063013	0.0787936987	0
93883.53	0	1.8539090718	0.8645854198	0.1354145802	0
110899.3	0	2.0155367905	0.8824187145	0.1175812855	0
11199.04	1	0.4322718291	0.6064160296	0.3935839704	0
189992.97	0	1.0022472152	0.7315001785	0.2684998215	0
164255.69	0	1.7200000828	0.848128847	0.151871153	0
152167.79	1	1.4455614044	0.8093143897	0.1906856103	0
33949.67	0	1.8888755936	0.8686272733	0.1313727267	0
68143.93	0	2.5221815782	0.9256822751	0.0743177249	0

ROC Curve:

