Prepared by **Yuqing Chen, Heejae Roh**
Nikshita Ranganathan, Archit Barua
Professor Paromita Guha

Feb 14th, 2023

BANK CUSTOMER CHURN

Northeastern University

slido

What factors do you think drives the bank churn?

ⓘ Start presenting to display the poll results on this slide.

slido

Which customer segments are most likely to churn?

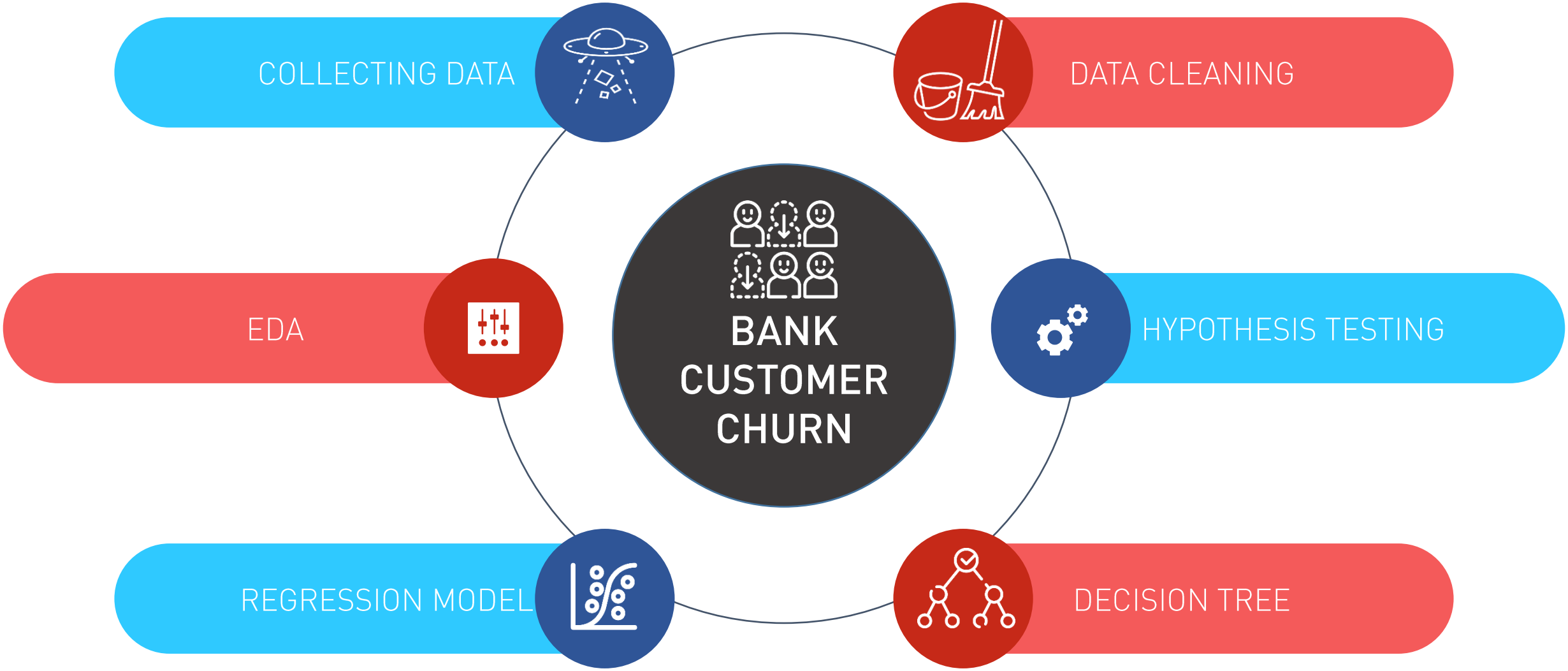
ⓘ Start presenting to display the poll results on this slide.

slido

What strategies will you recommend to increase customer loyalty?

ⓘ Start presenting to display the poll results on this slide.

“Churn butter, not customers!”


COLLECTING DATA
DATA CLEANING
EDA
BANK CUSTOMER CHURN
HYPOTHESIS TESTING
REGRESSION MODEL
DECISION TREE

PRIMARY QUESTIONS

1. Main Factors affecting churn?

2. Customer Segments most affected?

3. Effective Strategies to reduce churn?

4. Best Methods make loyalty & engagement?

UNDERSTANDING THE DATASET

Exited = Target Variable

Contains details of 10,000 customers & 14 features

Customer's demographics, financial status, & banking behavior

7 categorical, 1 character and 6 numerical variables

Descriptive statistics

DATA CLEANING AND MANIPULATION

01

Checking for NA/null values and Duplicate rows

02

Dropping & Adding column

03

Recoding columns

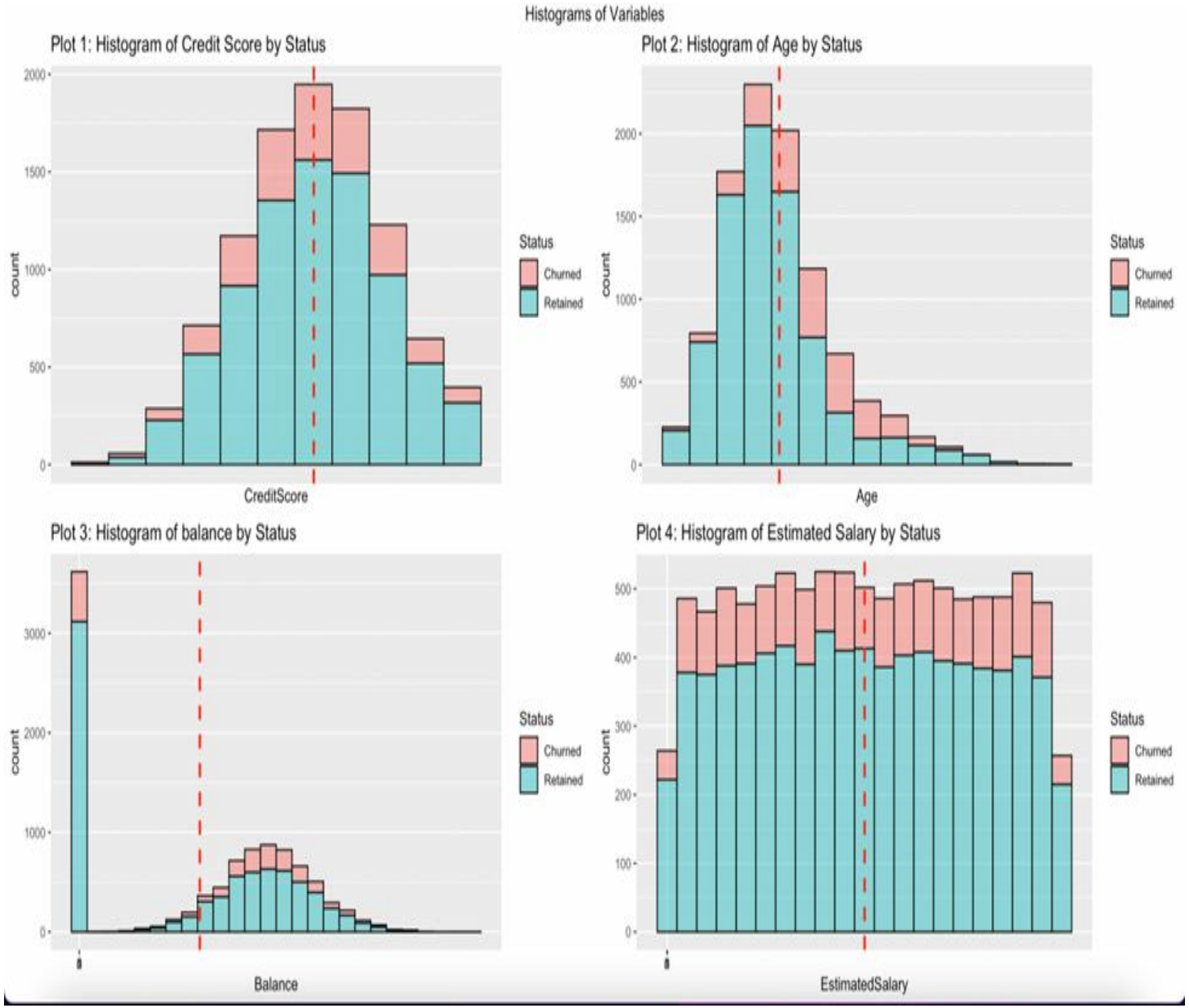
04

Changing column names

05

Modifying the datatypes

HISTOGRAMS



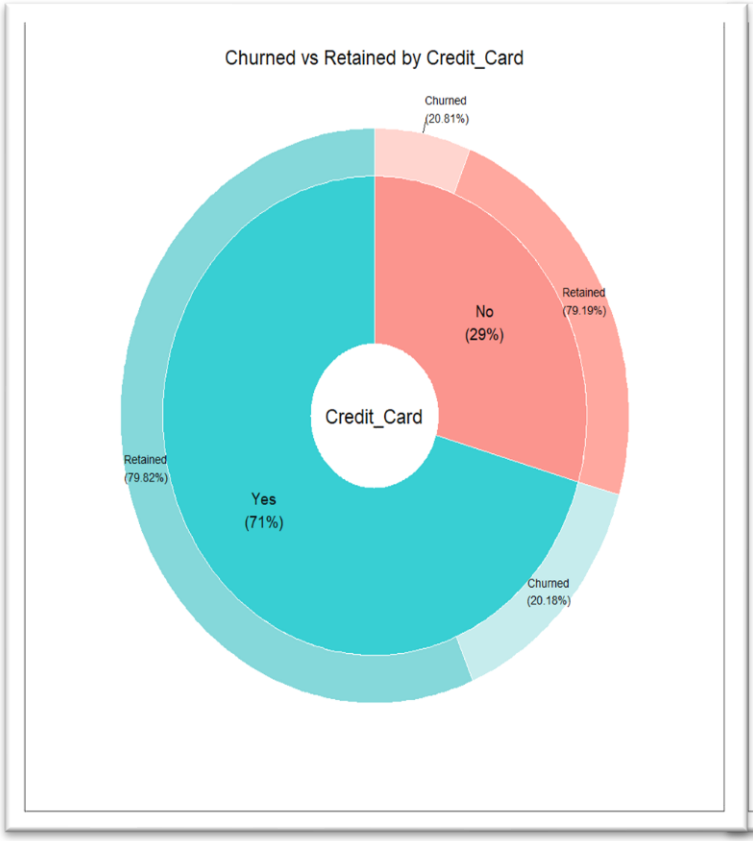
Credit score - Normally distributed
Average score of 650

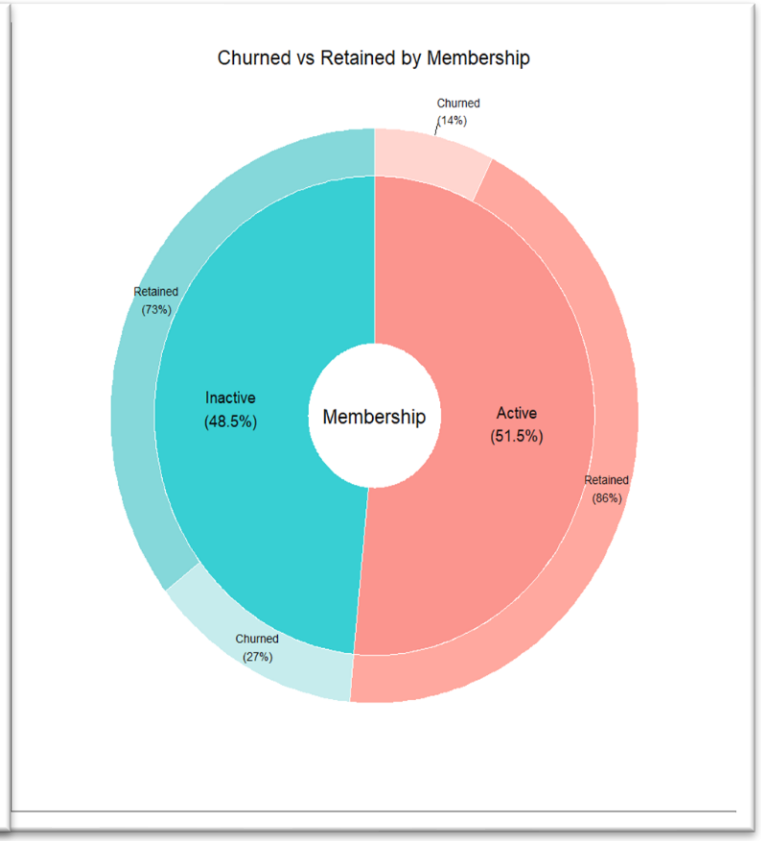
Age - Right skewed

Balance - Normal distribution except zero

Estimated salary - Uniform distribution
Mean of 100k

DONUT PIE CHARTS



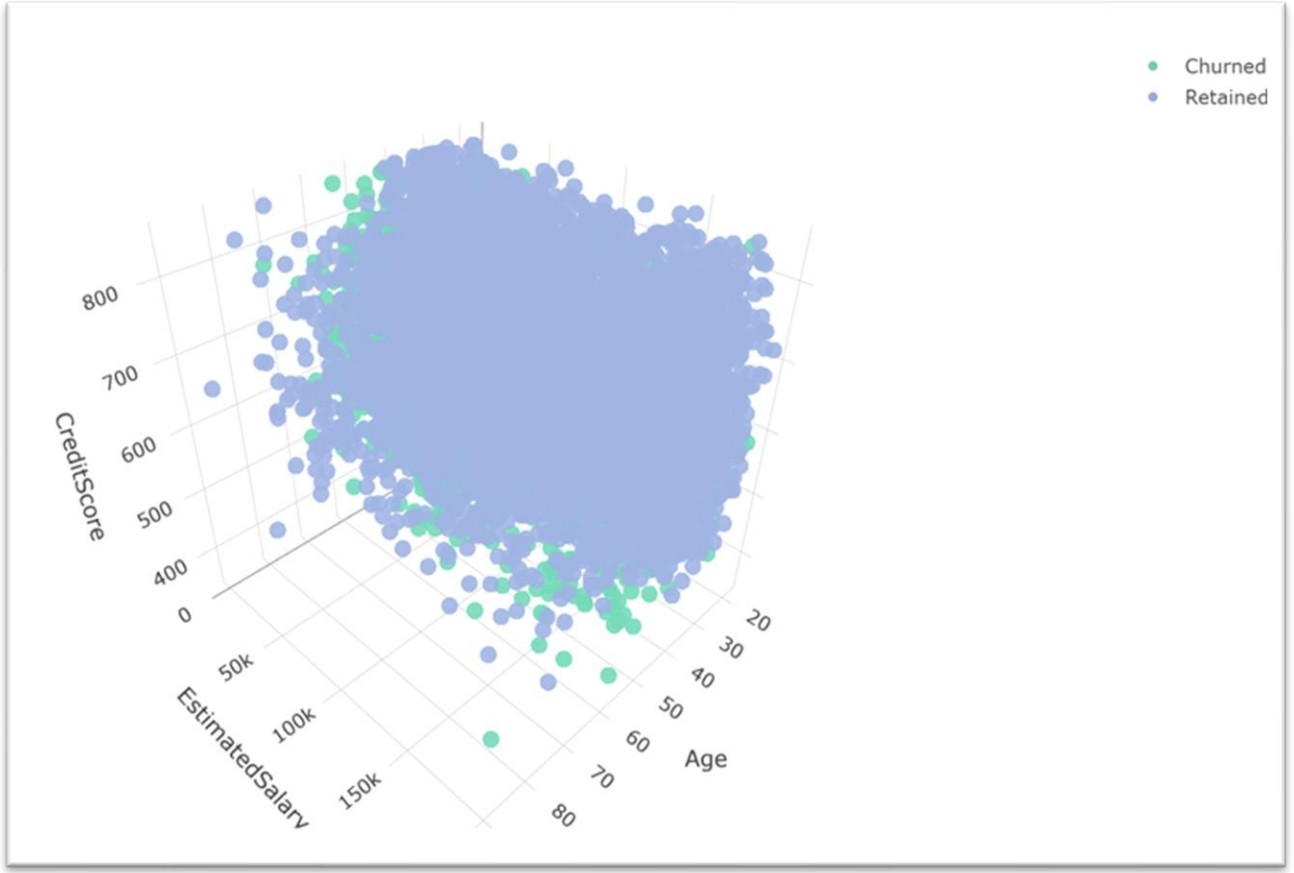


Credit cards do not have a significant effect on churn rate

Membership status is an important factor in customer retention.

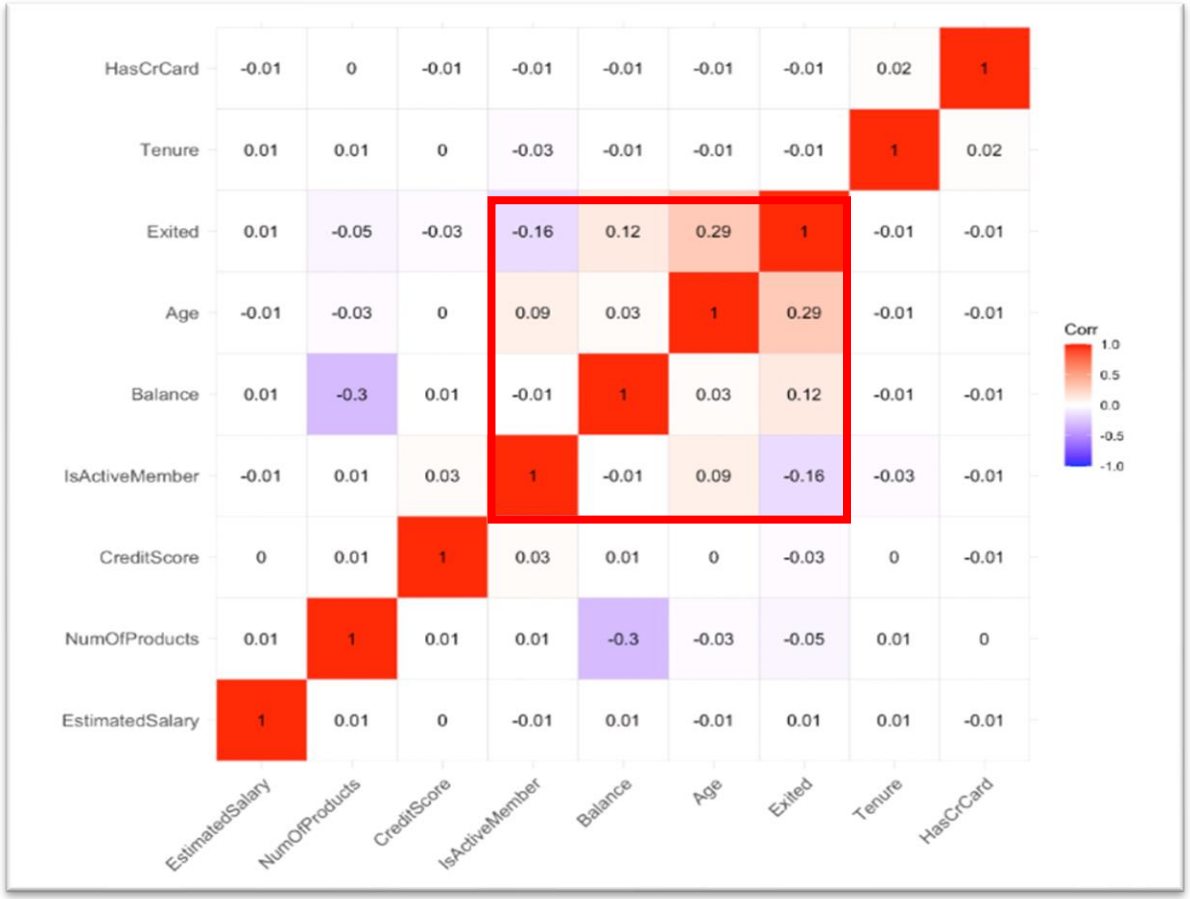
3D PLOT

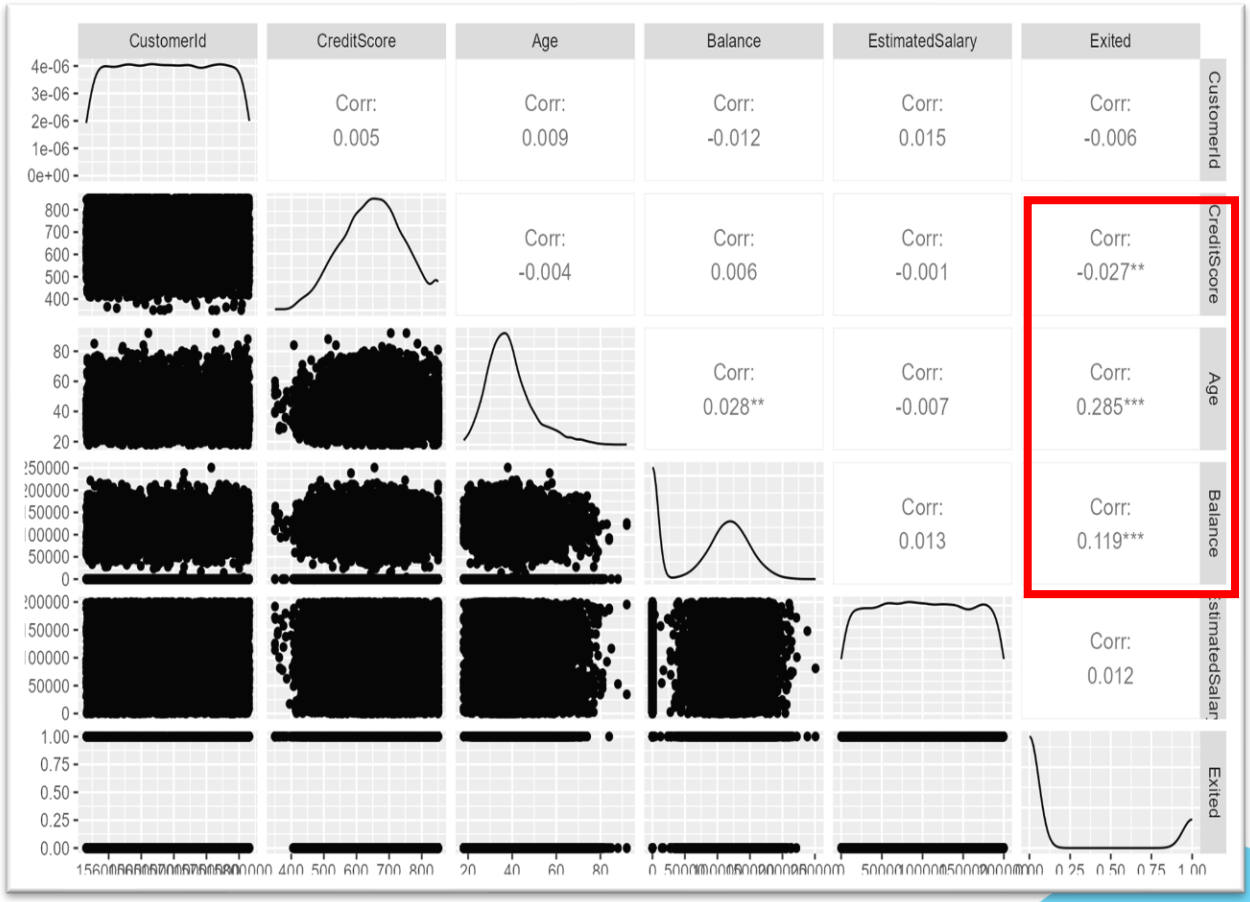
Relationship between
Credit score, Salary & Age.



CORRELATION

Exited column is most correlated with CreditScore, Age, and Balance





One Sample t-test

STEP 0 Finding a Key Metrics

< 600

Credit Score

H_1, claim

Step 1 Hypothesis.

Null: The mean of CreditScore is greater or equal to 600
Alternative: The population mean of CreditScore is less than 600 (claim)

- H0: μ1 ≥ 600, H1: μ1 < 600 (claim)

Step 2 Find the critical value.

- The p-value is α = 0.05

Step 3 Compute the test value.

churn$CreditScore, mu=600, alternative = "less"

t = 52.278 df = 9999 p-value = 1

95 percent confidence interval: -Inf 652.1188

sample estimates: mean of x 650.5288

Step 4 Make the decision

There is not enough evidence to reject null hypothesis since 1 (p-value) > 0.05

Step 5 Summarize Results

Credit Score

H_0

≥ 600

Two Sample t-test

STEP 0 Finding a Key Metrics

H_0, claim

Credit Score by Gender

Step 1 Hypothesis.

Null: The mean of CreditScore is equal between Males and Females(claim)
Alternative: The mean of CreditScore differs between Male and Female

H0: μ1=μ2 (claim) H1: μ1≠μ2

Step 2 Find the critical value.

- The p-value is α = 0.05

Step 3 Compute the test value

Data: Male$CreditScore and Female$CreditScore		
t = -0.28563	df = 9998	p-value = 0.7752
95 percent confidence interval:	-4.359797	3.250804
sample estimates:	mean of x 650.2769	mean of y 650.8314

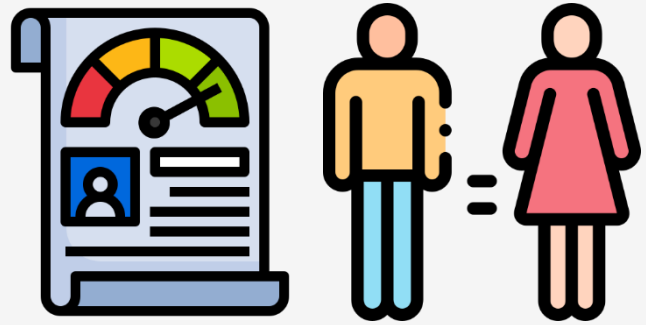
Step 4 Make the decision

There is not enough evidence

to reject null hypothesis

since 0.7752 (p-value) > 0.05

Step 5 Summarize Results



H_0

Credit Score by Gender

F-test

STEP 0 Finding a Key Metrics

H0, claim

Salary Var. by Gender

Step 1 Hypothesis.

Null: No difference in the variance of Salary between Male and Female (claim)
Alternative: Difference in the variance of Salary between M and F

H0: σ2m = σ2f (claim) H1:σ2m ≠ σ2f

Step 2 Find the critical value.

- The p-value is α = 0.05

Step 3 Compute the test value

F test to compare two variances: Male$EstimatedSalary and Female$EstimatedSalary

F = 1.009	num df = 5456	denom df = 4542	p-value = 0.7536

95 percent confidence interval: 0.954268 1.066681
sample estimates: ratio of variances 1.008983

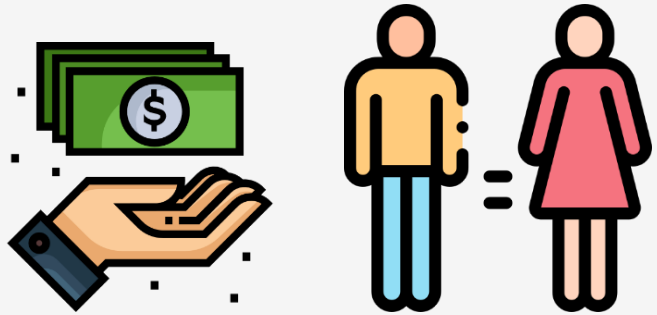
Step 4 Make the decision

There is not enough evidence

to reject null hypothesis

since 0.7536 (p-value) > 0.05

Step 5 Summarize Results

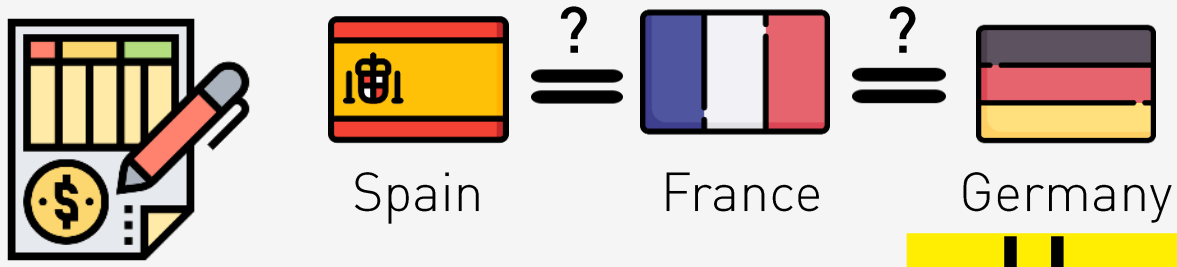


H0

Salary Var. by Gender

One-way ANOVA

STEP 0 Finding a Key Metrics



H_1, claim

Balance by Geography

Step 1 Hypothesis.

Null: There is no difference in mean of Balance according to Geography
Alternative: At least one mean is different from the others (claim).
H0: μ1 = μ2 = μ3

Step 2 Find the critical value.

- The p-value is α = 0.05

Step 3 Compute the test value

Balance ~ Geography, data=churn

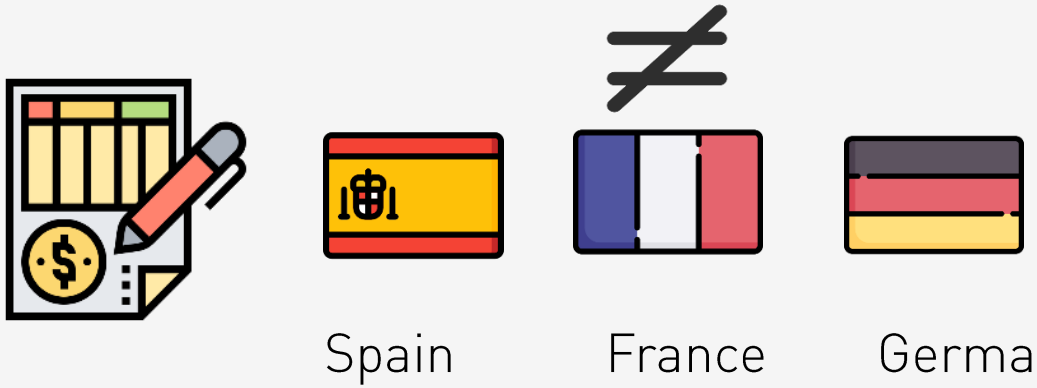
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Geography	2	6.264e+12	3.132e+12	958.4	<2e-16 ***
Residuals	9997	3.267e+13	3.268e+09		

Step 4 Make the decision

There is enough evidence to reject null hypothesis since 2e-16 (p-value) < 0.05

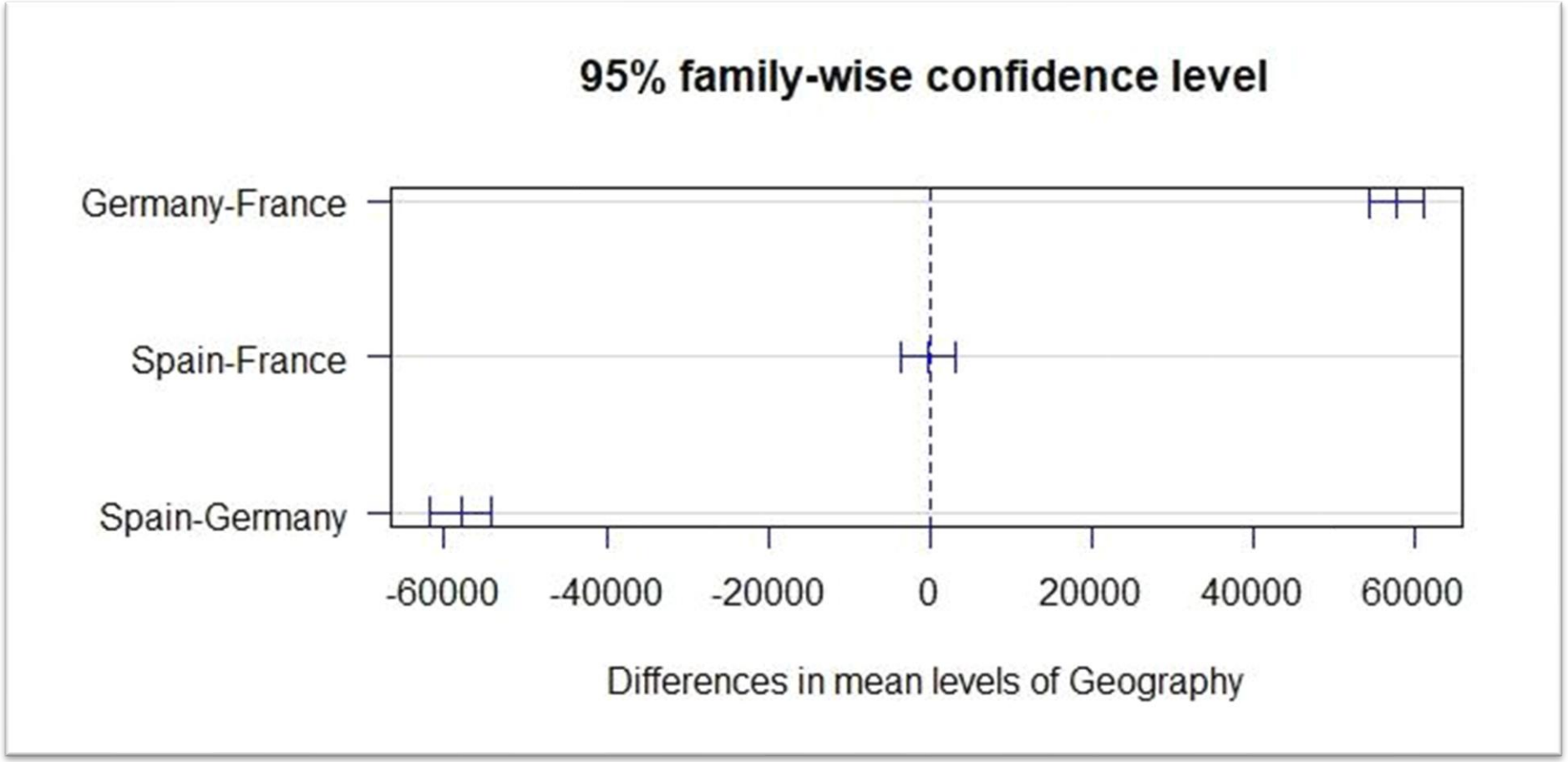
H_1

Step 5 Summarize Results

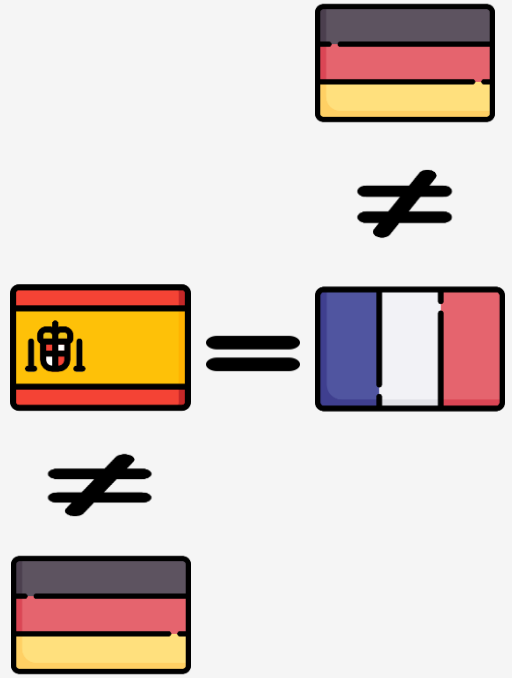


Balance by Geography

Tukey test



Results

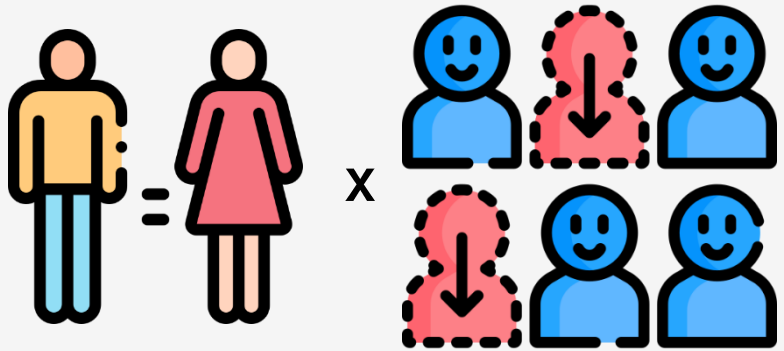


Two-way ANOVA

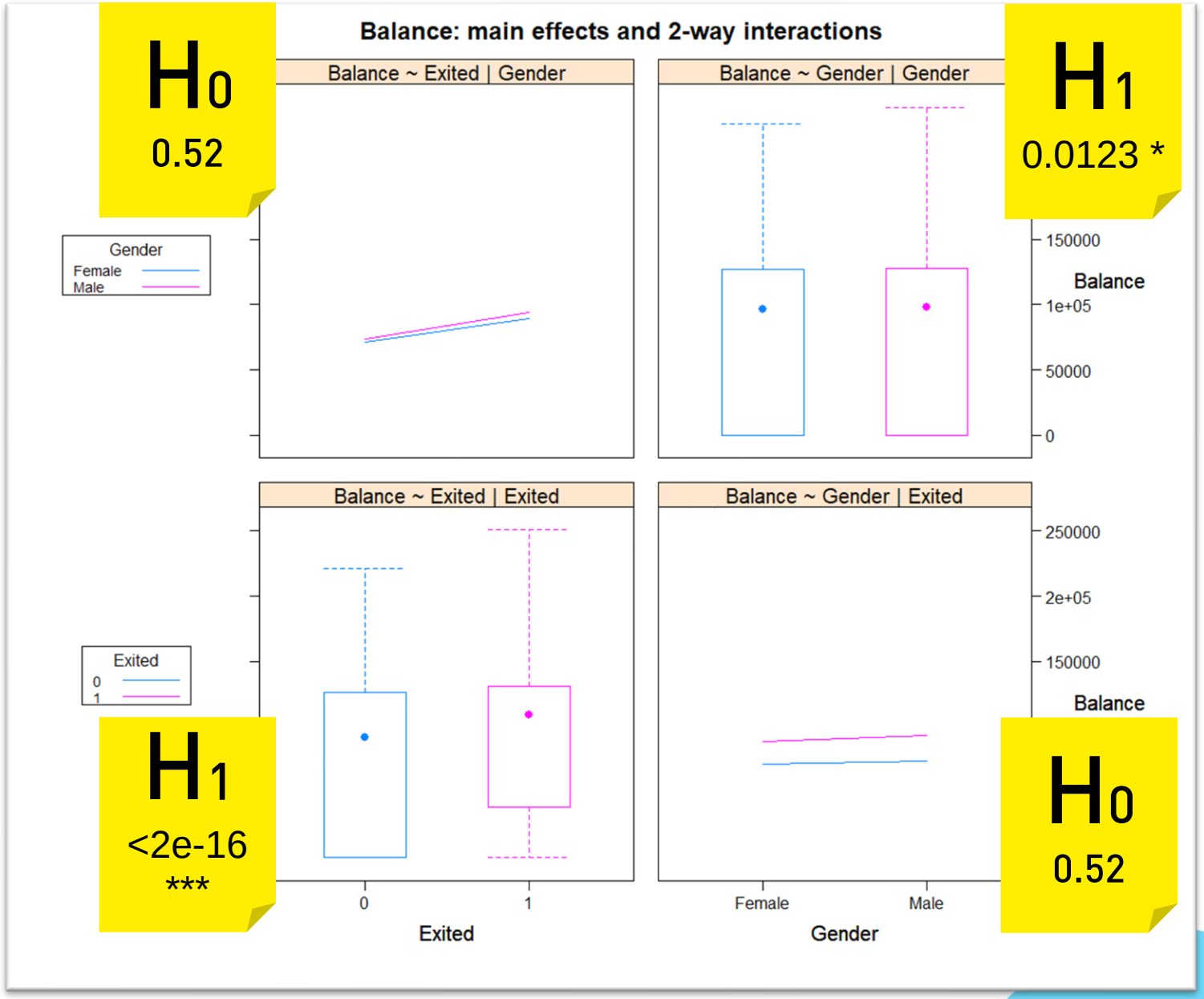
STEP 0 Finding a Key Metrics

Balance

H0, claim



Gender x Churn



H0
0.52

H1
0.0123 *

H1
<2e-16
***

H0
0.52

Linear Regression

Call: lm(formula = Exited ~ Balance + Age + NumOfProducts, data = churn)

Residuals

Min	1Q	Median	3Q	Max
-0.81248	-0.22060	-0.13807	-0.02975	1.07857

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.651e-01	1.975e-02	-13.422	<2e-16 ***
Balance	7.016e-07	6.453e-08	10.872	<2e-16 ***
Age	1.083e-02	3.659e-04	29.602	<2e-16 ***
NumOfProducts	-4.227e-03	6.923e-03	-0.611	0.542

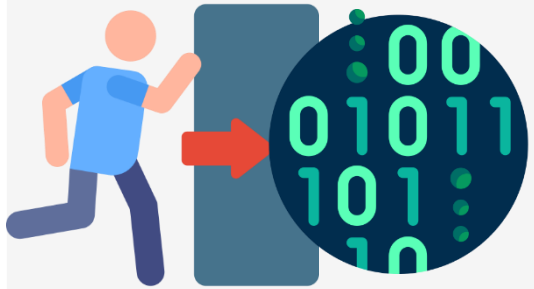
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3835 on 9996 degrees of freedom

Multiple R-squared: 0.09365, Adjusted R-squared: 0.09338

F-statistic: 344.3 on 3 and 9996 DF p-value: < 2.2e-16

CUSTOMER CHURN



BINARY

Logistic Regression- Model 1

Call: glm(formula = Exited ~ CreditScore + Geography + Gender + Age + Tenure
+Balance + NumOfProducts + Credit_Card + Membership
+EstimatedSalary, family = "binomial", data = data_train)

Deviance Residuals

Min	1Q	Median	3Q	Max
-2.2809	-0.6601	-0.4621	-0.2763	2.8899

Coefficients:

	Estimate	Std. Error	z value	Pr(>|t|)
(Intercept)	-3.386e+00	2.929e-01	-11.559	< 2e-16 ***
CreditScore	-5.343e-04	3.332e-04	-1.604	0.10876
GeographyGermany	7.708e-01	8.105e-02	9.511	< 2e-16 ***
GeographySpain	6.620e-02	8.352e-02	0.793	0.42803
GenderMale	-5.188e-01	6.490e-02	-7.993	1.31e-15 ***
Age	7.044e-02	3.071e-03	22.941	< 2e-16 ***
Tenure	-1.129e-02	1.115e-02	-1.012	0.31149
Balance	2.370e-06	6.162e-07	3.846	0.00012 ***
NumOfProducts	-1.273e-01	5.685e-02	-2.238	0.02519 *
Credit_Card1	-1.557e-02	7.093e-02	-0.219	0.82627
Membership1	-1.070e+00	6.876e-02	-15.569	< 2e-16 ***
EstimatedSalary	5.209e-07	5.623e-07	0.926	0.35423

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null deviance: 7063.5 on 6999 degrees of freedom
Residual deviance: 6032.4 on 6988 degrees of freedom
AIC: 6056.4

Number of Fisher Scoring iterations: 5

- Several variables in the logistic regression model are statistically significant at the 5% level, indicated by the "***" in the Pr(>|z|) column.
- The statistically significant variables include CreditScore, Geography (specifically Germany), Gender, Age, Balance, and Membership (specifically Inactive).
- Variables with p-values greater than 0.05 are not statistically significant, which means that their effect on the outcome is not significant enough to be distinguished from chance alone.
- The variables with p-values greater than 0.05 in this model are NumOfProducts, Credit_Card, and estimated salary.

Logistic Regression- Model 2

Call: glm(formula = Exited ~ Balance + Age + Membership + Gender, family = "binomial", data = data_train)

Deviance Residuals

Min	1Q	Median	3Q	Max
-2.1415	-0.6766	-0.4737	-0.2862	2.9134

Coefficients:

	Estimate	Std. Error	z value	Pr(>|t|)
(Intercept)	3.886e+00	1.396e-01	-27.842	<2e-16 ***
Balance	4.870e-06	5.324e-07	9.148	<2e-16 ***
Age	7.031e-02	3.037e-03	23.153	<2e-16 ***
Membership1	-1.079e+00	6.814e-02	-15.833	<2e-16 ***
GenderMale	-5.306e-01	6.423e-02	-8.261	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null deviance: 7063.5 on 6999 degrees of freedom
Residual deviance: 6135.0 on 6995 degrees of freedom
AIC: 6145

Number of Fisher Scoring iterations: 5

- There are trade-offs to consider when comparing the two models: the first model with all the predictors has a lower residual deviance and lower AIC, indicating a better fit to the data.
- However, the second model with fewer predictors may be preferred if the goal is to reduce the number of predictors in the model or to simplify interpretation.
- The second model may be easier to interpret because it has fewer predictor variables to consider and may be more parsimonious or efficient, especially if some predictors have limited impact on the outcome or are redundant with other predictors.

Confusion Matrix- Train

		Actual Values	
		No	Yes
Predicted Values	No	5398	1343
	Yes	188	71

Accuracy	0.7813	95% CI	(0.7714, 0.7909)
No Information Rate	0.798	P-Value [Acc > NIR]	0.9997
Kappa	0.0238	Mcnemar's Test P-Value	<2e-16
Sensitivity	0.05021	Specificity	0.96634
Pos Pred Value	0.27413	Neg Pred Value	0.80077
Prevalence	0.20200	Detection Rate	0.01014
Detection Prevalence	0.03700	Balanced Accuracy	0.50828
'Positive' Class	Yes		

- The model predicted "No" 5,398 times correctly and "Yes" 71 times correctly, but misclassified "No" 1,343 times and "Yes" 188 times.
- The overall accuracy of the model is 0.7813, which means it correctly predicted the outcome 78.13% of the time.
- The sensitivity of the model is very low, at 0.05021, indicating that the model is not very good at predicting positive cases.
- The specificity of the model is high, at 0.96634, meaning that it is good at identifying negative cases.
- False negatives would be more damaging in this case, as they represent customers who have actually left the bank, but the model predicted they would not.

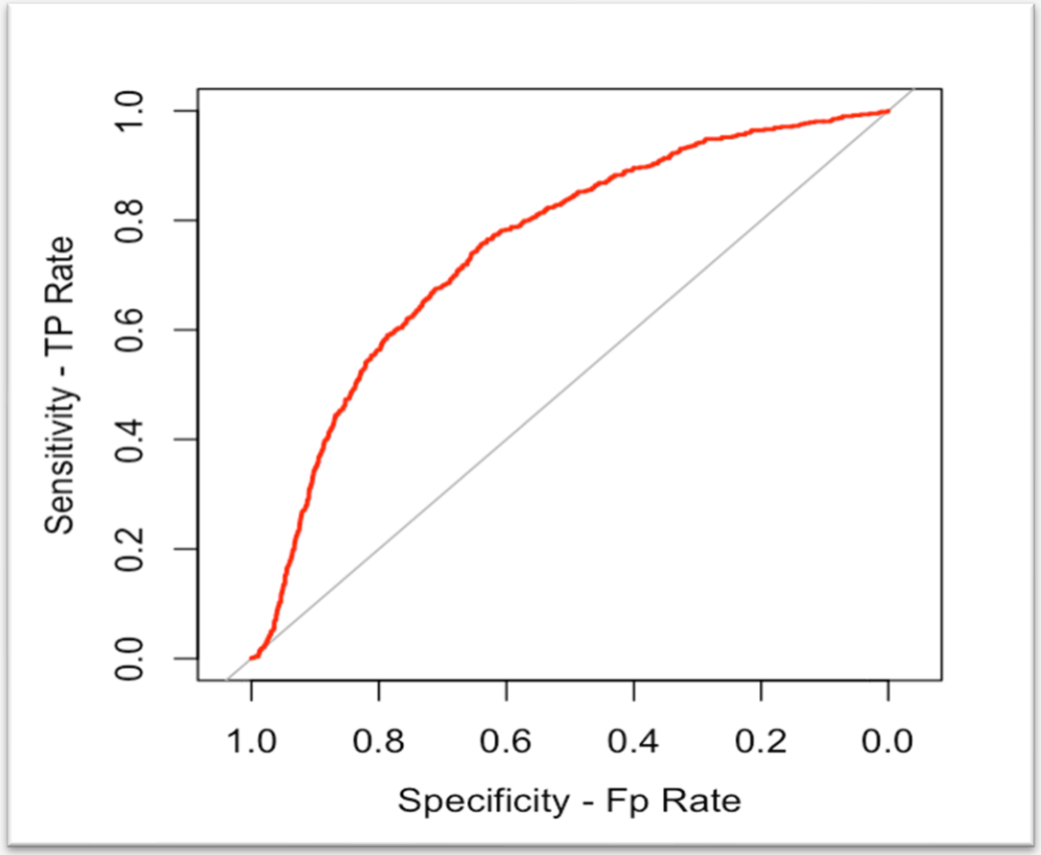
Confusion Matrix- Test

		Actual Values	
		No	Yes
Predicted Values	No	2298	591
	Yes	79	32

Accuracy	0.7767	95% CI	(0.7613, 0.7915)
No Information Rate	0.7923	P-Value [Acc > NIR]	0.9831
Kappa	0.026	Mcnemar's Test P-Value	<2e-16
Sensitivity	0.05136	Specificity	0.96676
Pos Pred Value	0.28829	Neg Pred Value	0.79543
Prevalence	0.20767	Detection Rate	0.01067
Detection Prevalence	0.3700	Balanced Accuracy	0.50906
'Positive' Class	Yes		

- Out of 3,000 customers in the test data, 2,298 were correctly classified as "No".
- 591 customers were incorrectly classified as "No", when they actually left the bank.
- 79 customers were incorrectly classified as "Yes", when they actually did not leave the bank.
- 32 customers were correctly classified as "Yes".
- The accuracy of the classifier is 0.7767, which means that 77.67% of the time, the classifier correctly predicts if a customer will leave the bank or not.

ROC & AUC

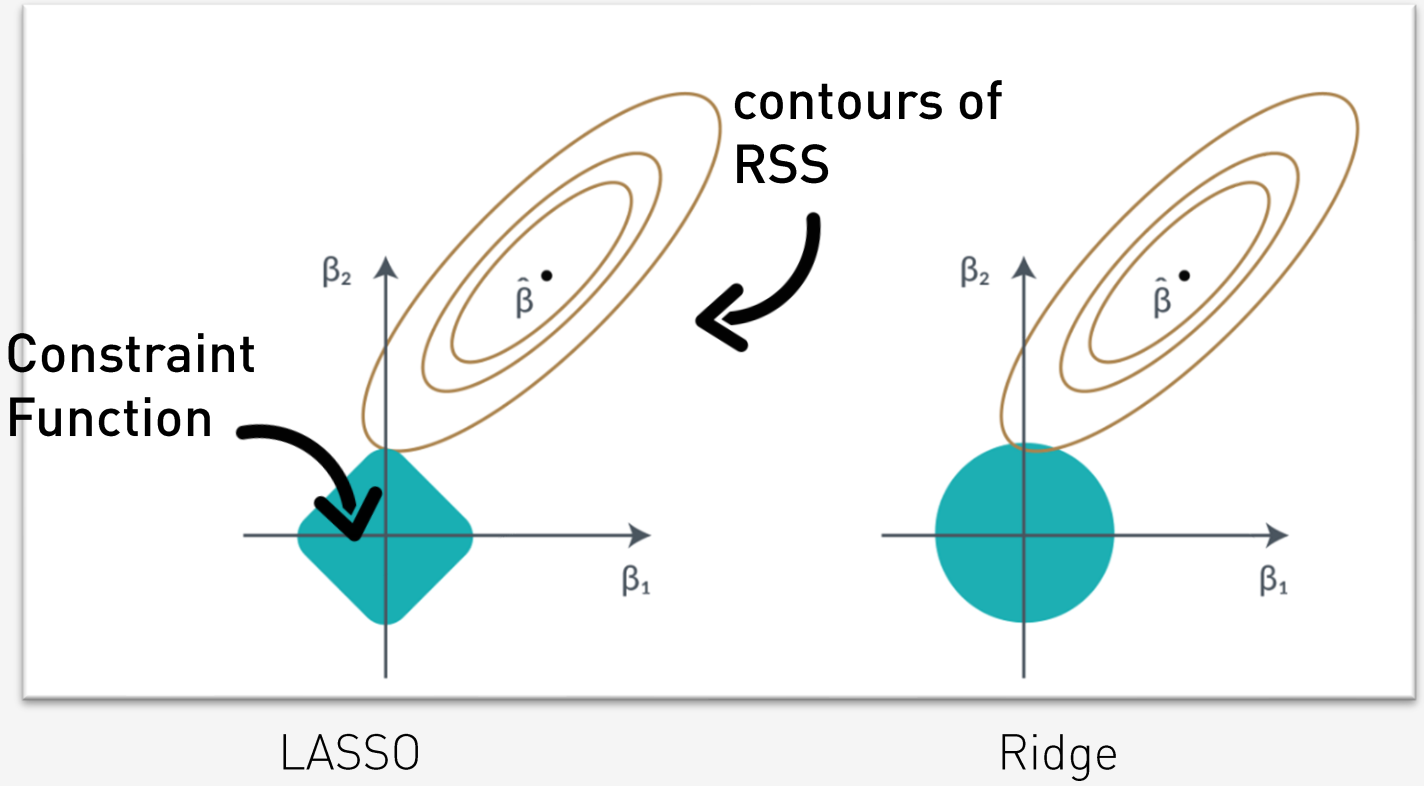


- The area under the curve (AUC) is a common metric used to evaluate the performance of binary classifiers.

- The AUC value ranges from 0 to 1, where a value of 0.5 indicates that the classifier is performing no better than random chance, while a value of 1 indicates perfect performance.

- In this case, the AUC value of 0.7485 suggests that the binary classifier has an average performance in distinguishing between the positive and negative classes.

Regularization



LASSO
Logistic Regression(L1)

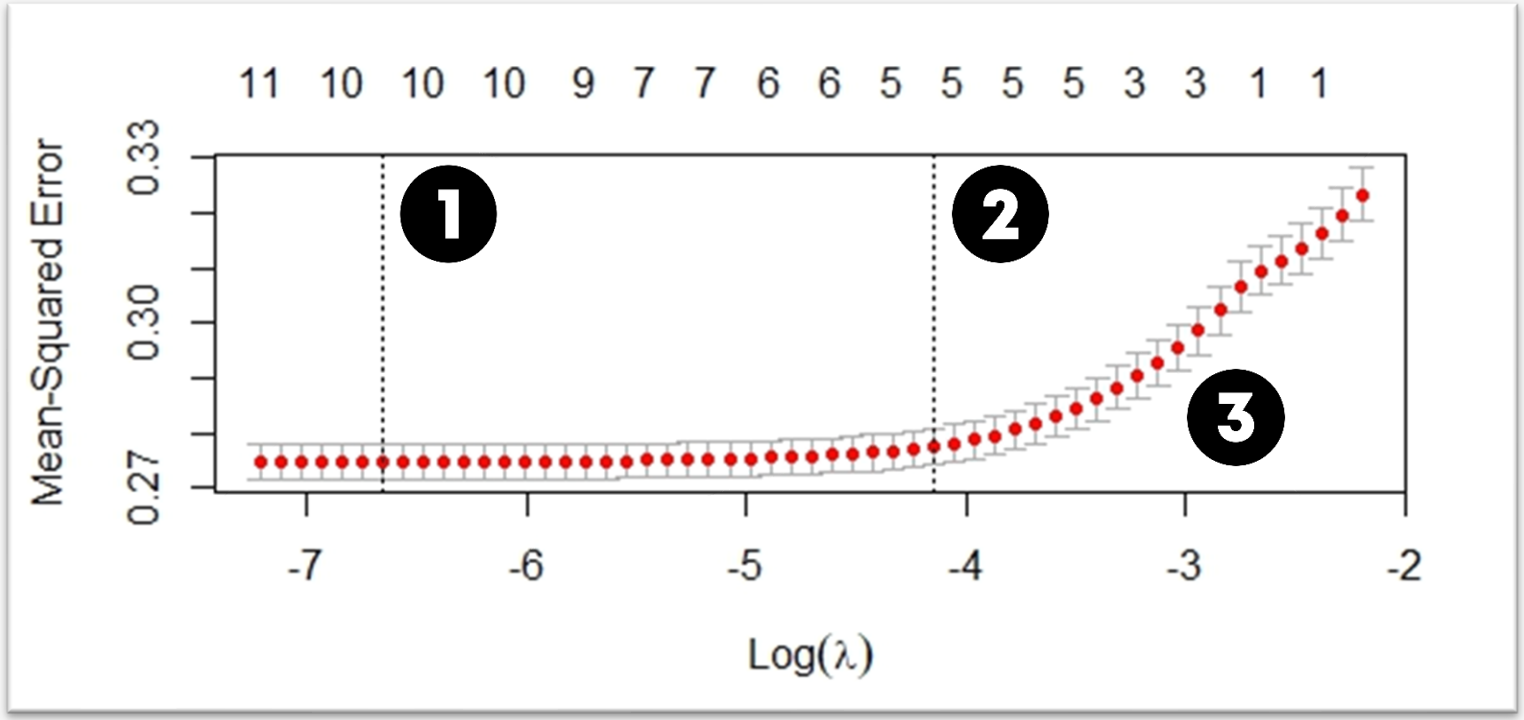
- penalty = absolute magnitude
- May shrink coefficients to zero

Ridge
Logistic Regression(L2)

- Penalty = square of the magnitude
- Shrunk by the same factor.

Gupta, P. (2017, November 16). Regularization in Machine Learning
https://towardsdatascience.com/regularization-in-machine-learning-76441ddcf99a

Lasso Logistic Regression



The Best Lambda

log(λ$min) ①	-6.66026412	λ$min	0.00128080
log(λ$1se) ②	-4.14835311	λ$1se	0.01579039

1. Left dot line
   = minimum lambda
2. Right dot line
   = 1SE of lambda
3. Red dots
   = loss metrics

Lasso Logistic Regression

glmnet(x,y= train, alpha = 1, lambda = cv.lasso$lambda.min)

	Df	%Dev	Lambda
1	10	14.38	0.001281

coef(model with min Lambda)

13 x 1 sparse Matrix of class "dgCMatrix"	s0	λ$min	0.00128080 s0
(Intercept)	-8.837045e-02	CreditScore	-6.587130e-05
GeographyGermany	1.213995e-01	GeographySpain	3.299456e-03
GenderMale	-7.108533e-02	Age	1.071278e-02
Tenure	-7.654091e-04	Balance	2.711338e-07
NumOfProducts	-1.765865e-02	HasCrCard	.
IsActiveMember	-1.387397e-01	EstimateSalary	6.167052e-08

Minimum Lambda

= 0.001281

Lasso with 1se of the Min. lambda

glmnet(x,y = train, alpha = 1, lambda = cv.lasso$lambda.1se)

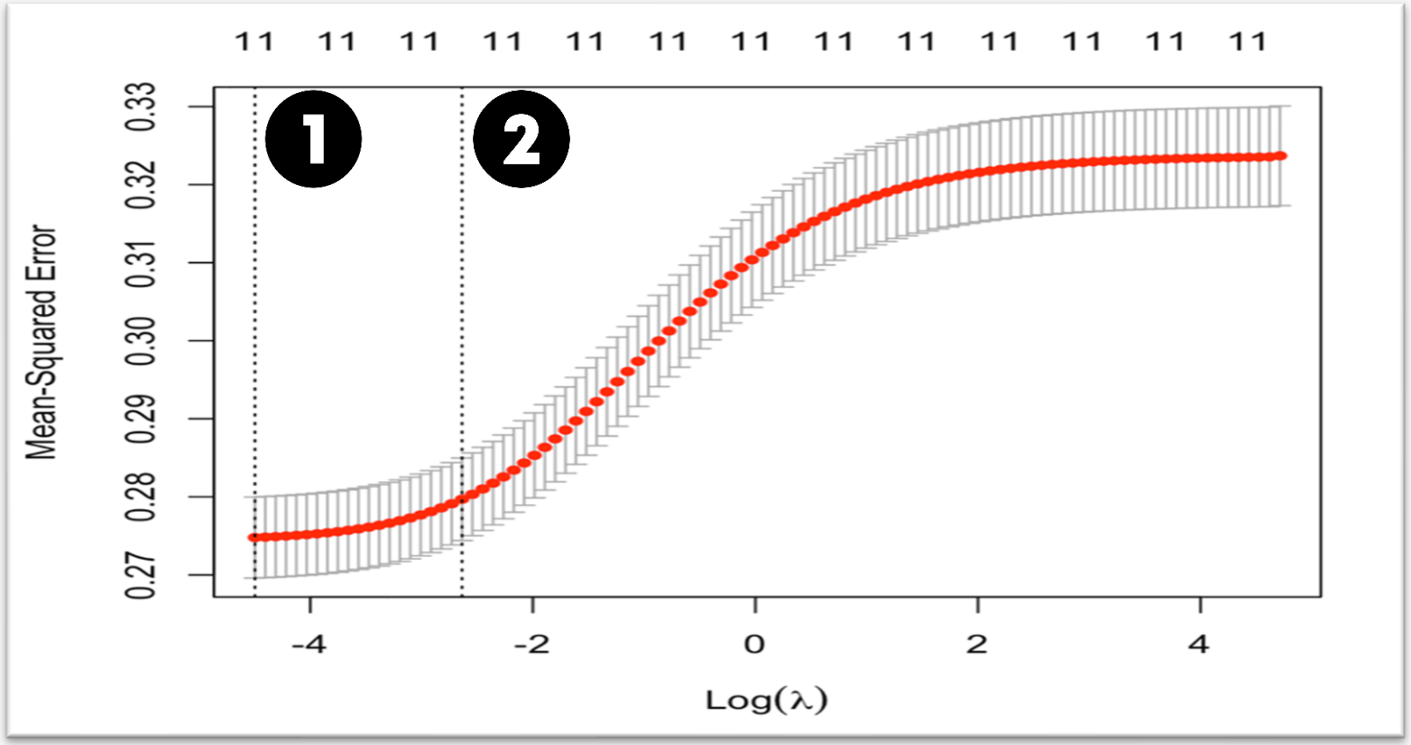
	Df	%Dev	Lambda
1	5	13.59	0.01579

coef(model with 1se Lambda)

13 x 1 sparse Matrix		λ$min	0.01579
	s0		s0
(Intercept)	-1.1772e-01	CreditScore	.
GeographyGermany	9.5137e-02	GeographySpain	.
GenderMale	-4.3096e-02	Age	9.344845e-03
Tenure	.	Balance	1.661237e-07
NumOfProducts	.	HasCrCard	.
IsActiveMember	-1.0889e-01	EstimateSalary	.

One standard Error Lambda = 0.01579

Ridge Logistic Regression



The Best Lambda

log(λ$min) ①	-4.49722296	λ$min	0.01113981
log(λ$1se) ②	-2.63655482	λ$1se	0.07160754

1. Left dot line
   = minimum lambda

2. Right dot line
   = 1SE of lambda

Ridge Logistic Regression

glmnet(x,y=train, alpha = 0, lambda = cv.ridge$lambda.min)

	Df	%Dev	Lambda
1	11	14.38	0.01113981

coef(model with min Lambda)

13 x 1 sparse Matrix of class "dgCMatrix"		λ$min	0.01113981
	s0		s0
(Intercept)	-7.3332e-02	CreditScore	-7.7012e-05
GeographyGermany	1.2156e-01	GeographySpain	6.3492e-03
GenderMale	-7.1908e-02	Age	1.0529e-02
Tenure	-1.1684e-03	Balance	2.8352e-07
NumOfProducts	-1.9075e-02	HasCrCard	-7.5599e-04
IsActiveMember	-1.3721e-01	EstimateSalary	8.1356e-08

Minimum Lambda

= 0.01113981

Variables

Ridge with 1se of the Min. lambda

glmnet(x,y = train, alpha = 0, lambda = cv.ridge$lambda.1se)

	Df	%Dev	Lambda
1	11	14.07	0.07160754

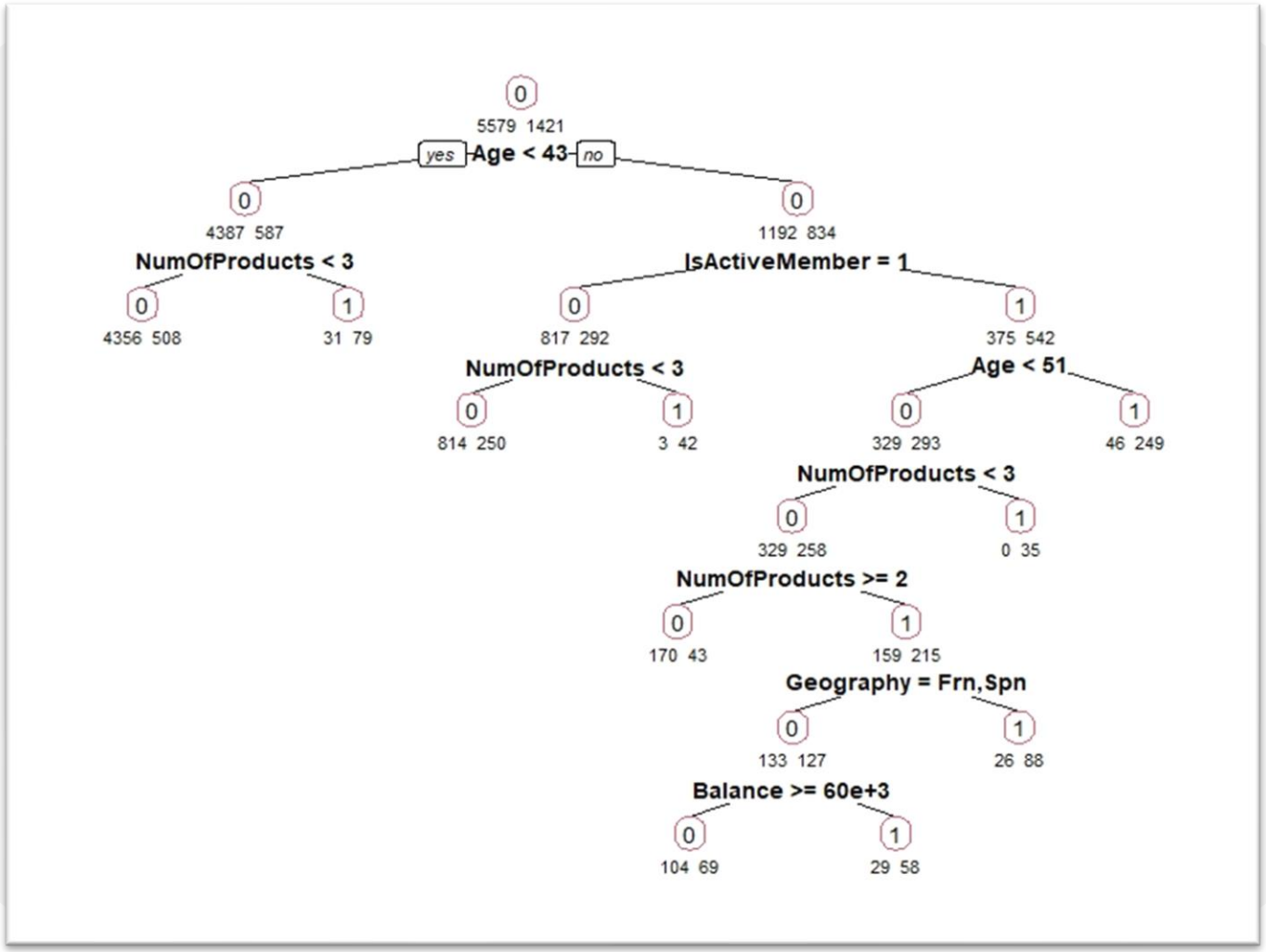
coef(model with 1se Lambda)

13 x 1 sparse Matrix	s0	λ$min	0.07160754 s0
(Intercept)	-3.766242e-02	CreditScore	-6.852981e-05
GeographyGermany	1.056159e-01	GeographySpain	6.888855e-04
GenderMale	-6.390416e-02	Age	9.154446e-03
Tenure	-9.792571e-04	Balance	2.934445e-07
NumOfProducts	-1.713984e-02	HasCrCard	-7.200222e-04
IsActiveMember	-1.179581e-01	EstimateSalary	6.933143e-08

One standard Error Lambda = 0.07160754

Variables

Decision Tree



1. A decision tree is a visual representation of a decision-making process that uses a tree-like structure to show different possible outcomes or decisions based on a set of criteria or conditions.

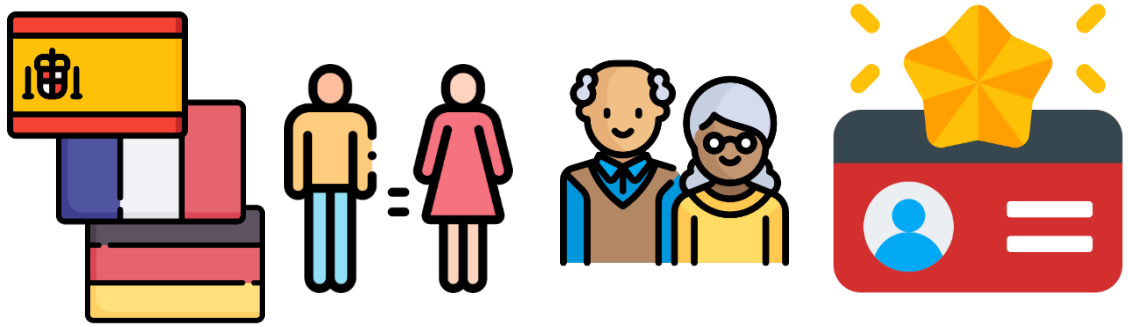
2. Each node in the tree represents a question or condition, and the answer leads you to the corresponding branch until you reach a final outcome or decision at the end of a branch.

3. 0= RETAINED 1= CHURNED

Answer for Primary Questions

1

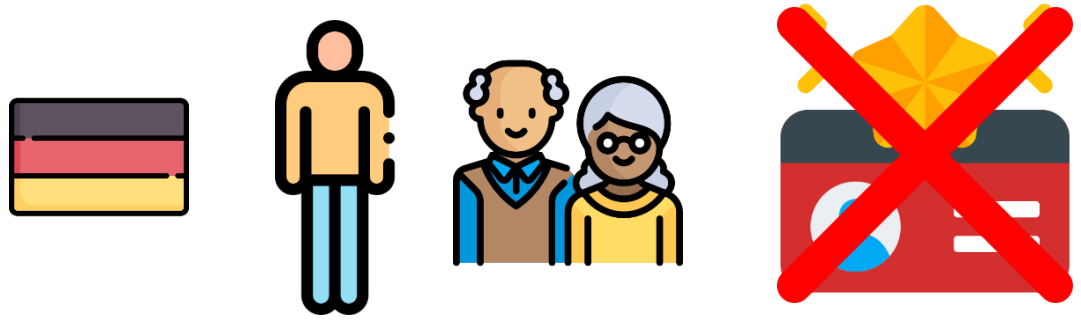
Main Factors affecting churn?



Geography Gender Age Membership

2

Customer Segments most affected?



Germany Male Age High No Active Membership

Answer for Primary Questions

3

Effective Strategies to reduce churn?

More
Active
Membership

4

Best Methods make loyalty & engagement?

Using
Personalization

Rewards
& Loyalty program

Improving
Service

Offering
New Products

REFERENCE

- Northeastern. (2023, Jan). ALY 6015 - lesson 4-1,4-2, 4-3, 4-4, 4.5 — regularization. Retrieved fro m https://northeastern.instructure.com/courses/131212/pages/lesson-4-1-regularizati on?module_it em_id=8227336

- Rosane Rech. (2021, June 28). Pie-donut chart in R. Retrieved from https://statdoe.com/pie-donut-chart-in-r/

- STHDA. (n.d.). ggplot2 legend : easy steps to change the position and the appearance of a graph l egend in R software. Retrieved from http://www.sthda.com/english/wiki/ggplot2-legend-easy-steps-to-change-the-position-and-the-appearance-of-a-graph-legend-in-r-software

- Vedda. (2018, May 31). Remove all of x axis labels in ggplot [duplicate]. Retrieved from https://stackoverflow.com/questions/35090883/remove-all-of-x-axis-labels-in-ggplot

REFERENCE

- STHDA. (n.d.). ggplot2 histogram plot : Quick start guide - R software and data visualization. Retrieved from http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-visualization

- jmp. (n.d.). The one-sample t-Test. Retrieved from https://www.jmp.com/en_sg/statistics-knowledge-portal/t-test/one-sample-t-test.html

- K. (2017, July 11). A gentle introduction to logistic regression and lasso regularisation using R. wordpress.com. Retrieved from https://eight2late.wordpress.com/2017/07/11/a-gentle-introduction-to-logistic-regression-and-lasso-regularisation-using-r/

- finnstats. (2021, April 19). Decision trees in R. R-bloggers. Retrieved from https://www.r-bloggers.com/2021/04/decision-trees-in-r/

Q&A


WHY?
WHAT?
WHERE?
WHEN?
WHO?
HOW?
WHICH?
WHOSE?
WHAT? HOW? WHY?
WHO? WHERE? WHAT? HOW?
WHERE? WHICH? WHOSE? WHEN? WHY?
WHY? HOW? WHERE?
WHO? WHOSE?
WHERE? WHAT? HOW?