PREDICTING CUSTOMER CHURN –

AND TAKING STRATEGIC MEASURES TO **STOP THE LOSS**

Carolyn Massa

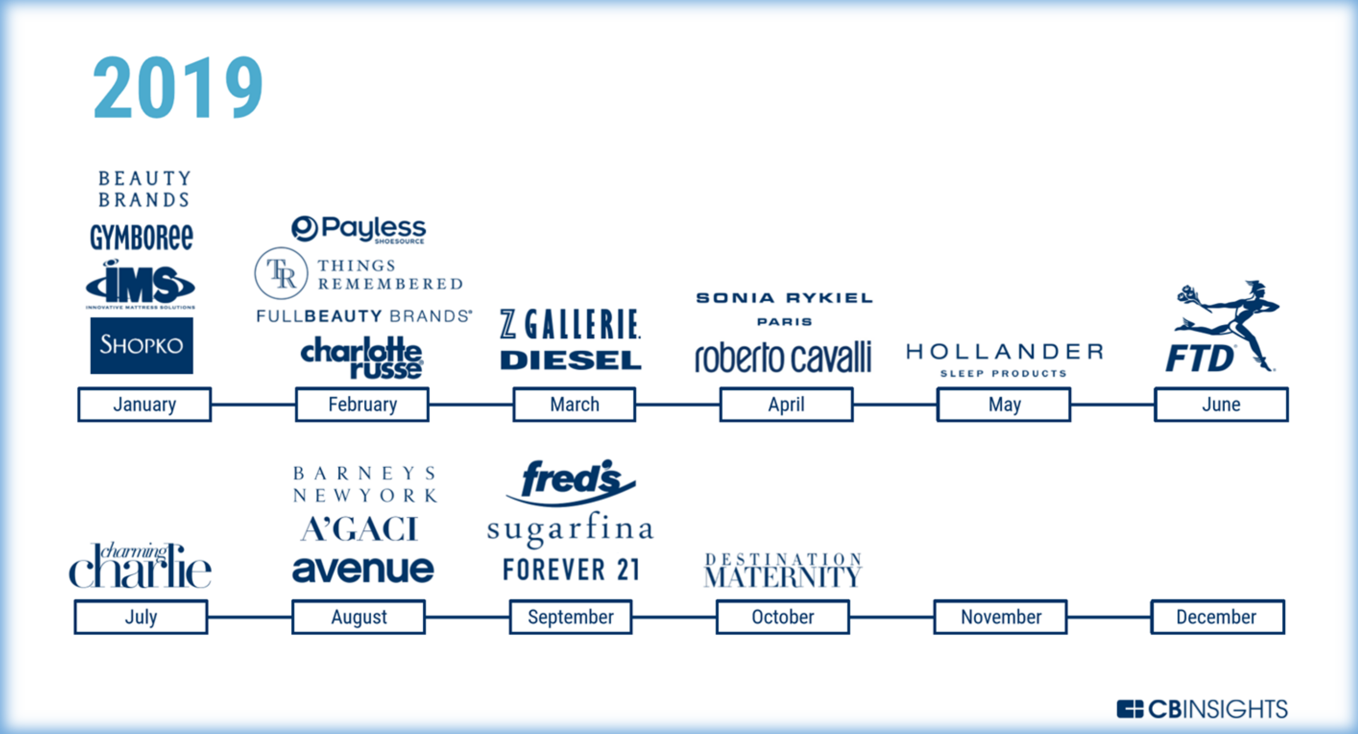
Carolyn@revinformatics.com

3.2020

WE CARE BECAUSE CUSTOMER CHURN IMPACTS EVERY BUSINESS:

- Lost jobs
- Lost Profit
- Increased Sales and Marketing Expenses
- Additional Work Created/Lost Productivity
- Increased stress

JUST A SMALL SAMPLE OF "BETTER KNOWN" BUSINESSES THAT FILED BANKRUPCY IN 2019! (WE ALL KNOW MANY MORE!)



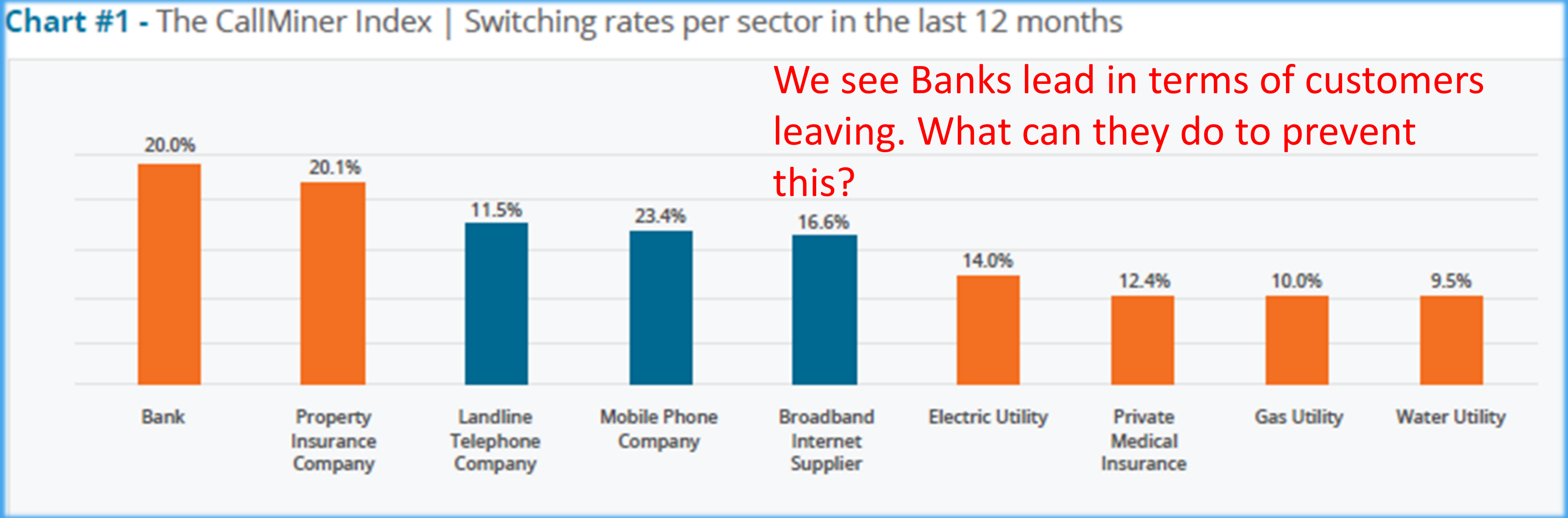
WE CARE BECAUSE CUSTOMER CHURN IMPACTS EVERY BUSINESS

According to research firm, CallMiner, avoidable customer churn is costing US businesses $136 billion a year. CallMiner surveyed 1,000 US adults to find out why consumers contact suppliers, how they were feeling when they contacted a call centers and which communications channels they preferred. What we uncovered is a switching epidemic – and that call centers play a pivotal role in whether consumers stay loyal or decide to switch.

OTHER FRIGHTENING FACTS THAT HAVE US RUNNING TO EXAMINE CUSTOMER CHURN:

- U.S. companies lose **$136.8 billion** per year due to avoidable consumer switching. (CallMiner)
- More than half of Americans have scrapped a planned purchase or transaction because of bad service. (American Express)
- 33% of Americans say they'll consider switching companies after **just a single instance** of poor service. (American Express)
- Companies with great experiences have a **16% price premium** on products and services. (PWC)
- 63% of U.S. consumers say they'd share more personal information with a company that offers a great experience. (PWC)
- After having a positive experience with a company, **77% of customers** would recommend it to a friend. (Temkin Group)

2017 TO 2018 CUSTOMER CHURN BY SECTOR



We see Banks lead in terms of customers leaving. What can they do to prevent this?

LET’S TAKE THE RIGHT ACTION!

We attempt to predict churn rate and
Beat the accuracy of other potentially ineffective approaches such as:

- Faulty business changes such as advertising that produces little measurable results
- Basic marketing and competitive analyst which overlooks deep analysis insights

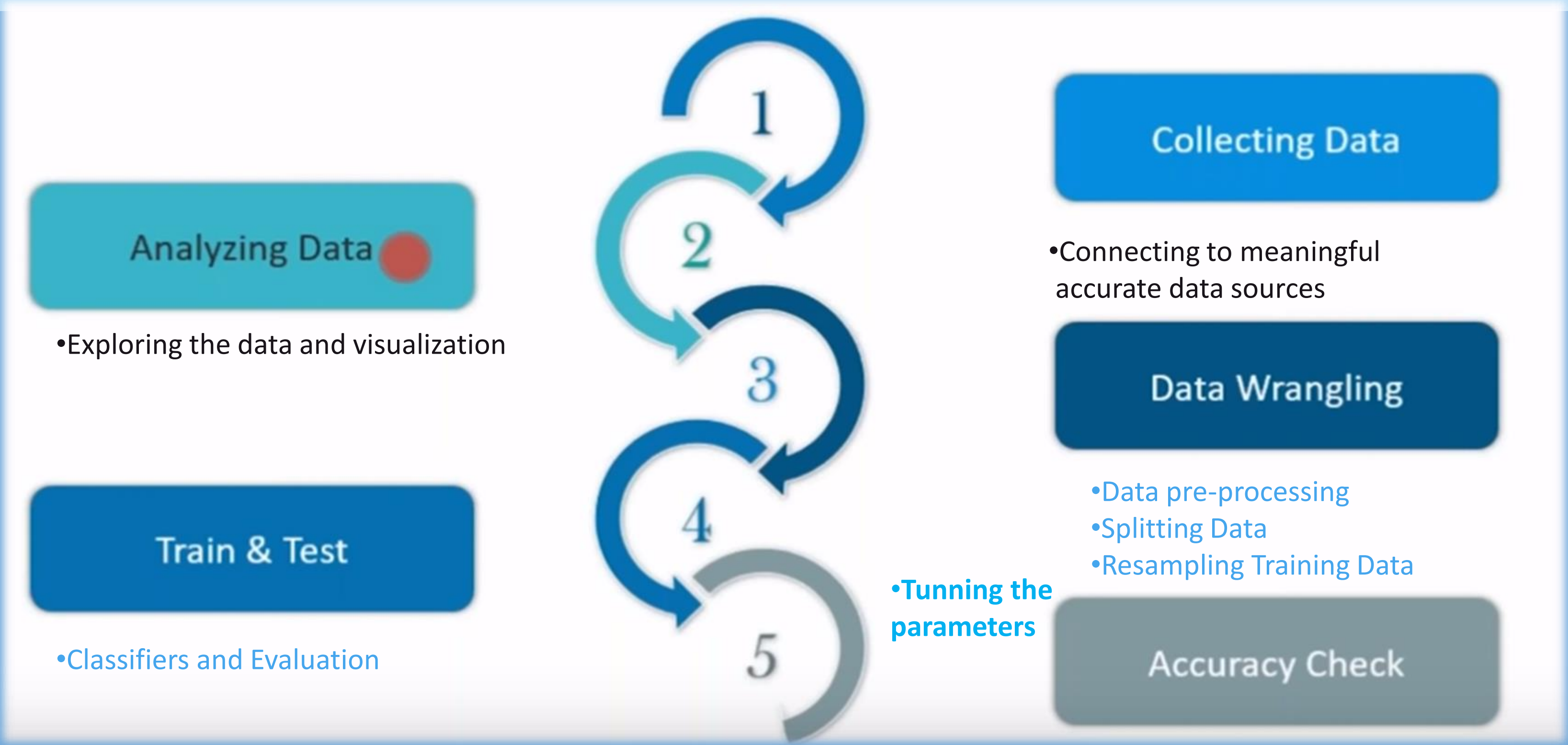
PROJECT STEPS

Analyze a Customer Churn Data Set to:

- **Discuss/determine goals and metrics for the project.** Is reducing churn the right problem to solve? How specifically shall we measure performance improvement? What is our success goal? (Assume here: it is the right problem, we measure performance overall by reducing customer churn, success is reducing customer churn by 10% in next 6 months).

- **Understand what deliverables are useful for internal stakeholders** (Assume it is churn prediction factors, later a spreadsheet of customer churn predictions, production pipeline and perhaps an internal dashboard).

THE 5 STEP PROCESS TO ANALYZE WORLD BANK DATA SET



- Connecting to meaningful accurate data sources
- Exploring the data and visualization
- Data pre-processing
- Splitting Data
- Resampling Training Data
- Classifiers and Evaluation
- **Tunning the parameters**

PROCESS 1: COLLECTING DATA

The World Bank Data have provided me with an 11 year .csv file for me to work with.

The Data set below has 13 variables to from which to develop a predictive churn model with 10,000 customer observations

CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0
...	...	...	...	...	...	...	...	...	...	...	...	...
15606229	Obijiaku	771	France	Male	39	5	0.00	2	1	0	96270.64	0
15569892	Johnstone	516	France	Male	35	10	57369.61	1	1	1	101699.77	0
15584532	Liu	709	France	Female	36	7	0.00	1	0	1	42085.58	1
15682355	Sabbatini	772	Germany	Male	42	3	75075.31	2	1	0	92888.52	1

PROCESS 2: - ANALYZING & EXPLORING THE DATA

THE GREATEST VALUE OF A PICTURE IS WHEN IT FORCES US TO NOTICE WHAT WE NEVER EXPECTED TO SEE.

- JOHN TUKEY -

PROCESS 2: - ANALYZING & EXPLORING THE DATA

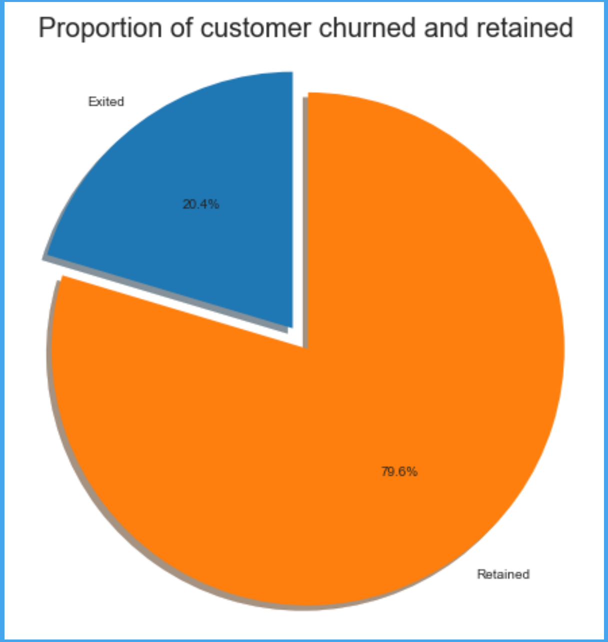
I notice the data is incomplete and leaves a lot of *unanswered questions.*

- Would it be possible to obtain balances over a period of time as opposed to a single date?
- What date did the customer exit?
- What types of products are the customers in?
- Could they have exited from a product and not the bank? What is an “Active Member”?

For this exercise, we proceed to model without context even though typically having context and better understanding of the data extraction process would give better insight and possibly lead to better and contextual results of the modelling process

PROCESS 2: - ANALYZING & EXPLORING THE DATA

To begin, I need to verify the type of data, what % is useable and look for patterns; I notice that of the 10,000 customers of Word Bank, 2,037 have churned in the past 11 years which is a 20% Churn Rate; which is considered an average to high rate among industry experts. I also review the averages in the variables where customers have churned.

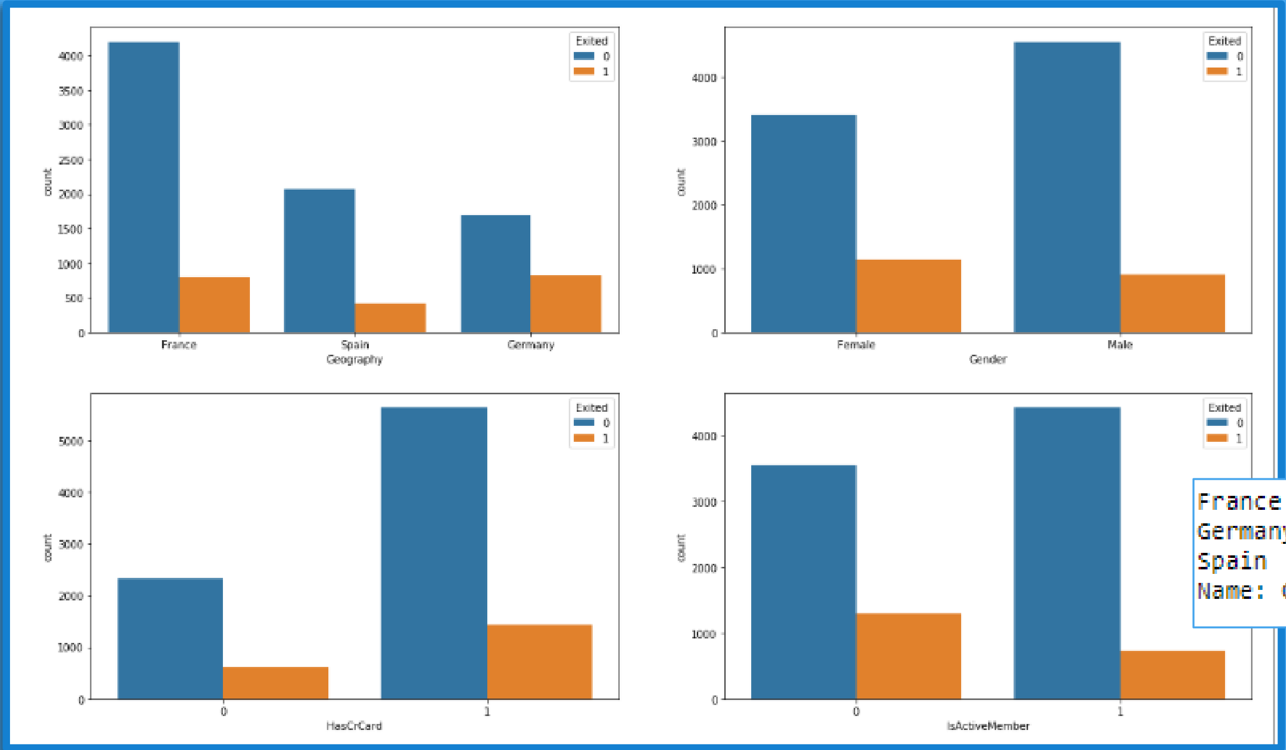



```
In [86]:    1  code['Exited'].value_counts()

Out[86]: 0    7963
         1    2037
         Name: Exited, dtype: int64
```

```
In [11]:    1  code[code['Exited'] == 1].mean()

Out[11]: CreditScore           645.351497
         Age                    44.837997
         Tenure                  4.932744
         Balance             91108.539337
         NumOfProducts           1.475209
         HasCrCard               0.699067
         IsActiveMember          0.360825
         EstimatedSalary    101465.677531
         Exited                  1.000000
```

PROCESS 2: - EXPLORING THE "CATEGORICAL DATA" VARIABLES

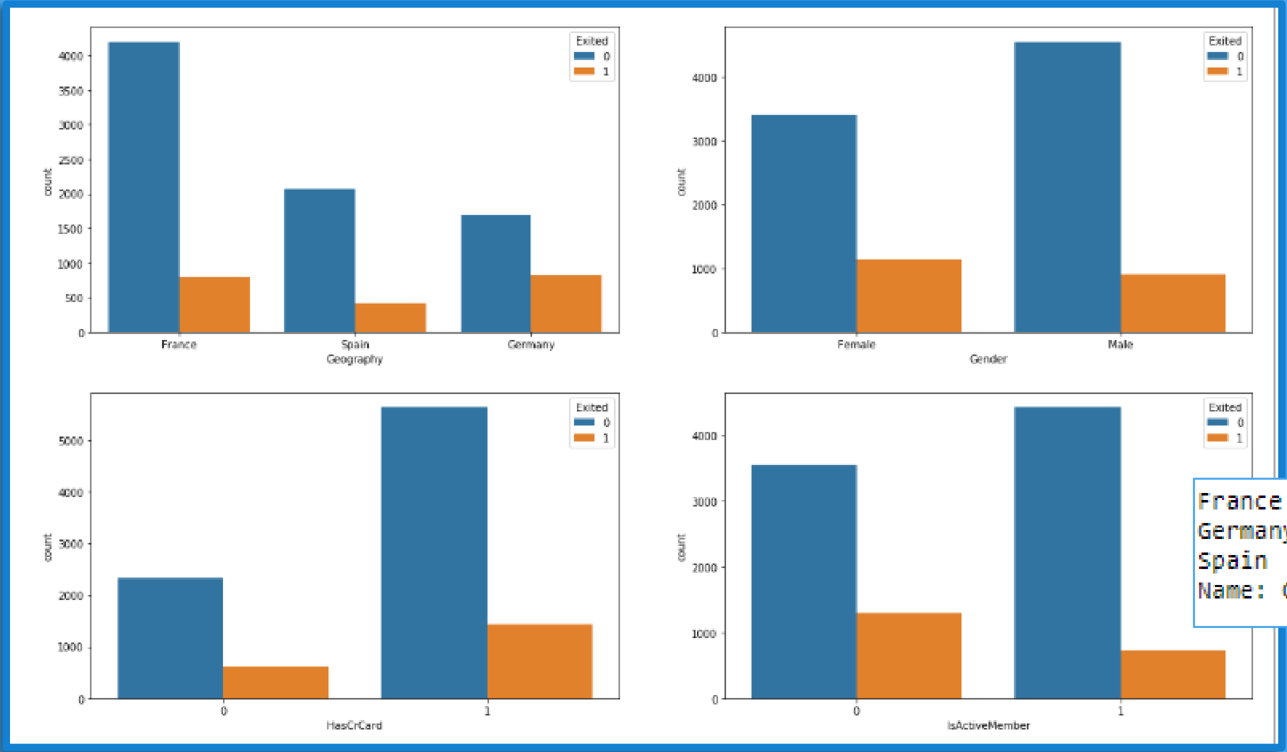
I Compare the Churn rates by Gender, Geography, Credit Card Holder, and Product Participation:




```
France      5014
Germany     2509
Spain       2477
Name: Geography, dtype: int64
```

PROCESS 2: - EXPLORING THE "CATEGORICAL DATA" VARIABLES

I Compare the Churn rates by Gender, Geography, Credit Card Holder, and Product Participation:




```
France     5014
Germany    2509
Spain      2477
Name: Geography, dtype: int64
```

FINDINGS FROM CATEGORICAL VARIABLES

We discovered the following:

•The country of France has the least churn rate and Germany the highest. However, the proportion of churned customers is with inversely related to the population of customers alluding to the bank possibly having a problem (maybe not enough customer service resources allocated) in the areas where it has fewer clients.

•The proportion of female customers churning is slightly higher than that of male customers

•Oddly, the majority of the customers that churned are those with credit cards. The majority of customers have credit cards so this could just be a coincidence.

•The inactive members have a greater churn. The overall proportion of inactive members is quite high suggesting that the bank may need a program implemented to turn this group to active customers as this can have a positive impact on the customer churn.

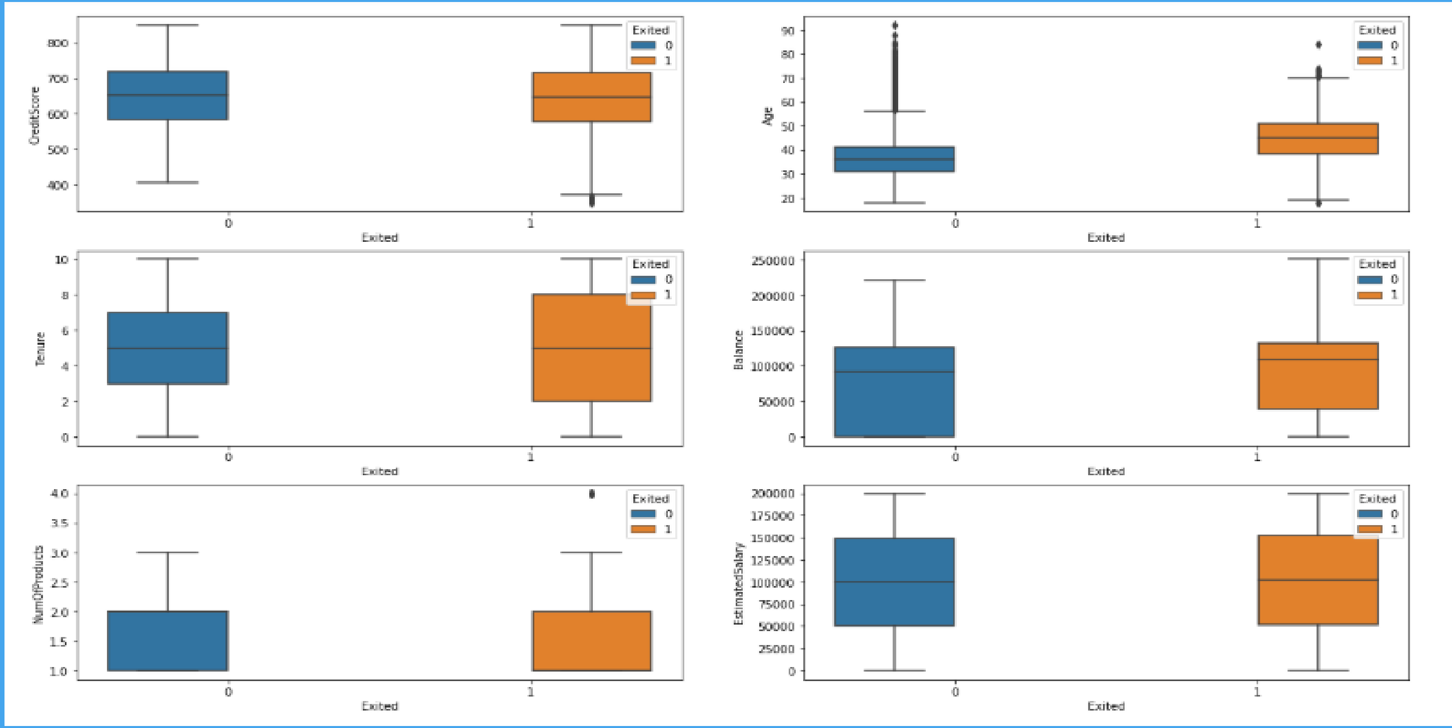
LET'S MAKE OUR "NULL HYPOTHESIS"

My "Null Hypothesis" (H0) is that "Since an Average Churn of Banks, according to "Call Miner" is 20% annually then there is few factors that can be determined that lead to whether a customer stays with the bank or not."

My "Alternative Hypothesis" (HA) is that "There are distinct factors and characteristics" that lead to a customers to leave a bank."

Through my analysis I will reach a conclusion on my hypothesis.

NOW THE "CONTINUOUS VARIABLES" ARE COMPARED AGAINST CHURN



FINDINGS FOR "CONTINUOUS DATA"

Interestingly, neither the product nor the salary has an impactful effect on the likelihood to churn.

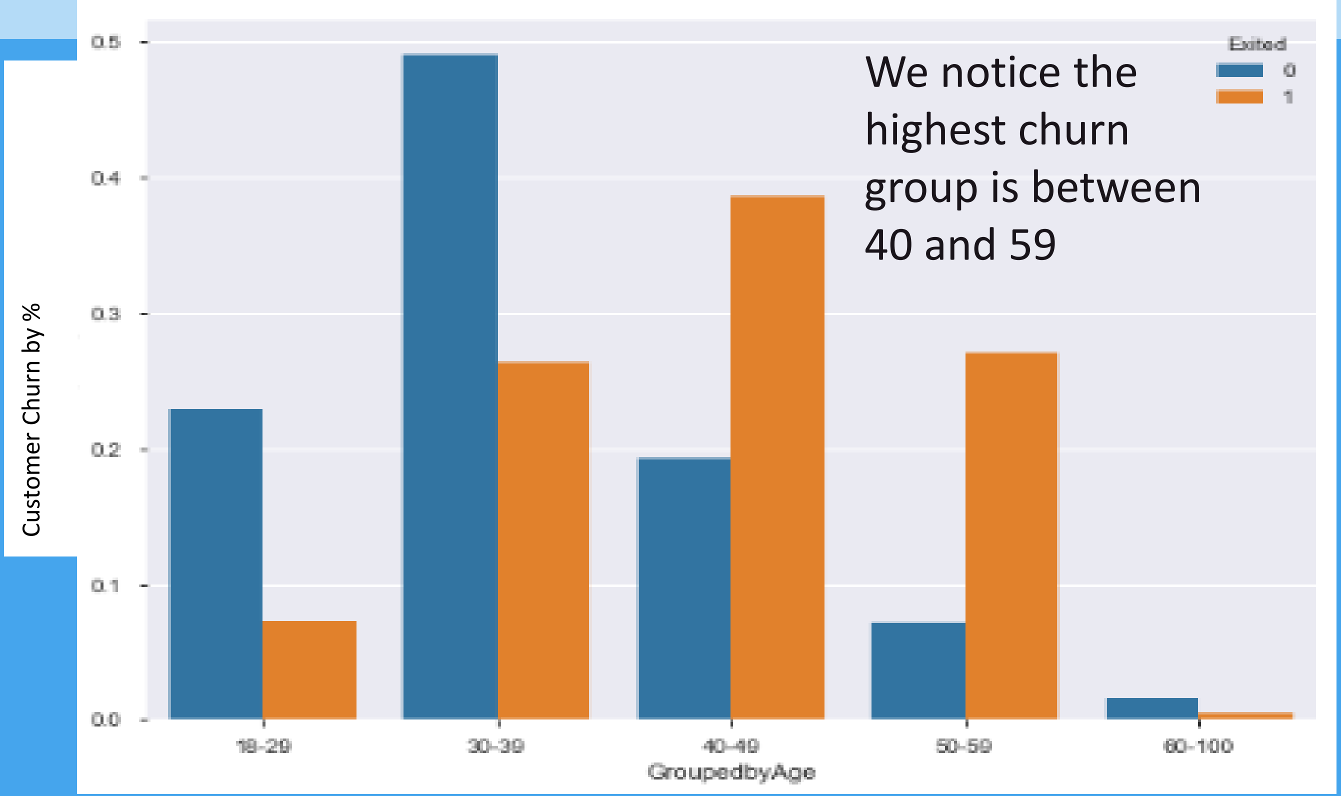
- There is no significant difference in the credit score distribution between retained and churned customers.

- With regard to the tenure, the average tenured client which is 5 years had a lesser likelihood to churn where those customers on opposing spectrums (spent little time with the bank to a lot of time with the bank) were **more likely to churn.** The highest Churn years in the 11 years examined were years 2, 4 and 10.

- Worryingly, the bank is losing customers with **significant bank balances** which is likely to hit their available capital for lending. The average bank balance for a churned customer is $91,000 with an average bank balance of $101,465.

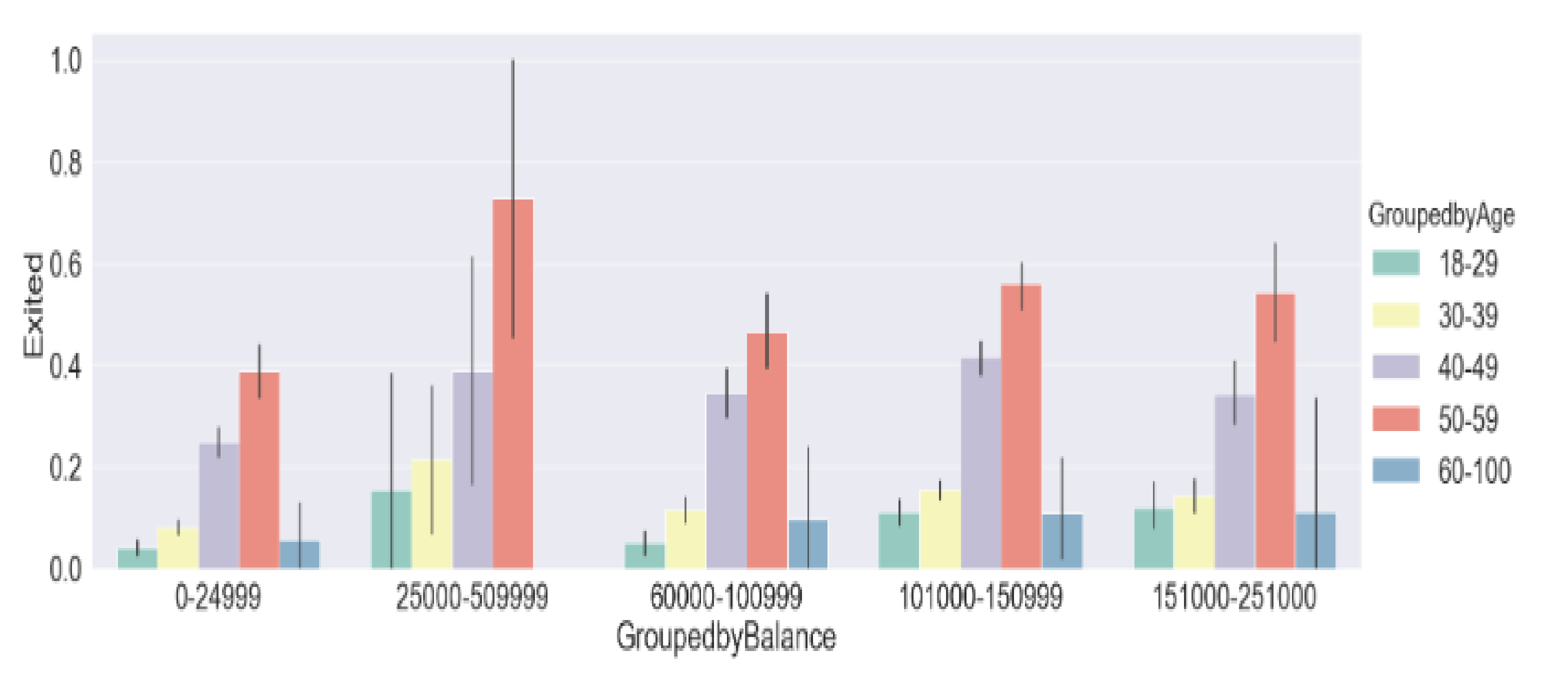
One interesting find is the **older customers are churning at more than the younger ones** which alludes to the fact that the bank may not adequate service standards that meet customer service expectations of older clients. Another conclusion is that they may be retiring and consolidating their assets elsewhere. The bank may gain from creating additional services plans for this client base.

AGE GROUPING AND TREND INVESTIGATION



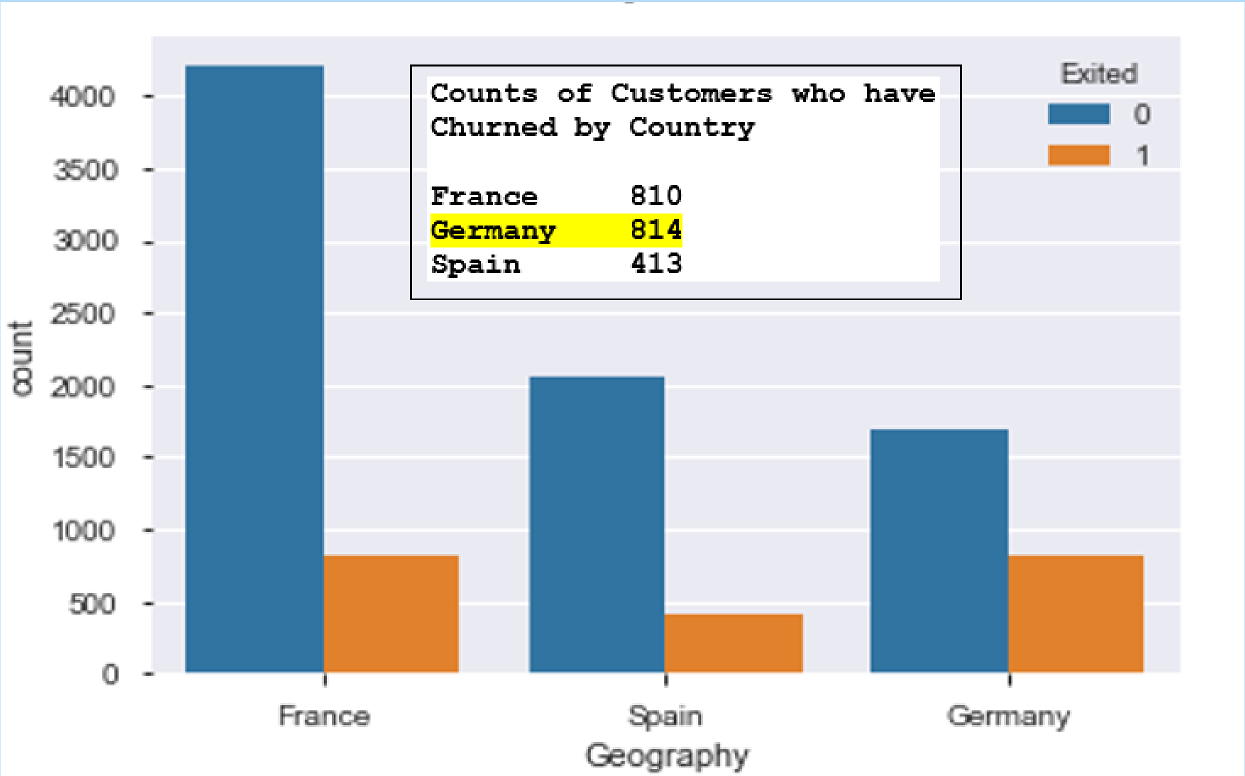
WE SEE A CLEAR TREND IN OUR BANK MEMBER AGES AND CHURN

BAR CHART OF EXITED BANK MEMBERS GROUPS BY THEIR BANK BALANCE



WE SEE A CLEAR TREND IN OUR BANK MEMBER AGES AND CHURN

BAR CHART OF EXITED BANK MEMBERS GROUPS BY THEIR ESTIMATED SALARY



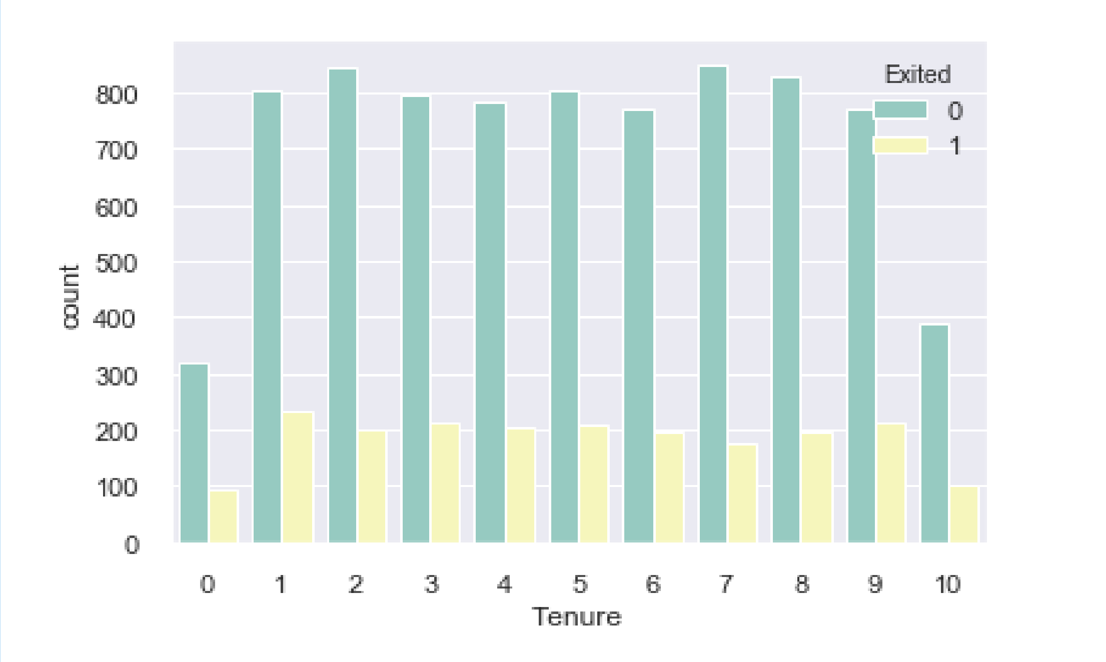
GERMANY HAD THE HIGHEST CHURN RATE
AND FRANCE THE LOWEST AS FRANCE HAS THE HIGHEST % OF MEMBERS

Avg Bank Balance by Country

Churn versus No Churn

0	France	60339.275678
	Germany	119427.106696
	Spain	59678.070470
1	France	71192.795728
	Germany	120361.075590
	Spain	72513.352446

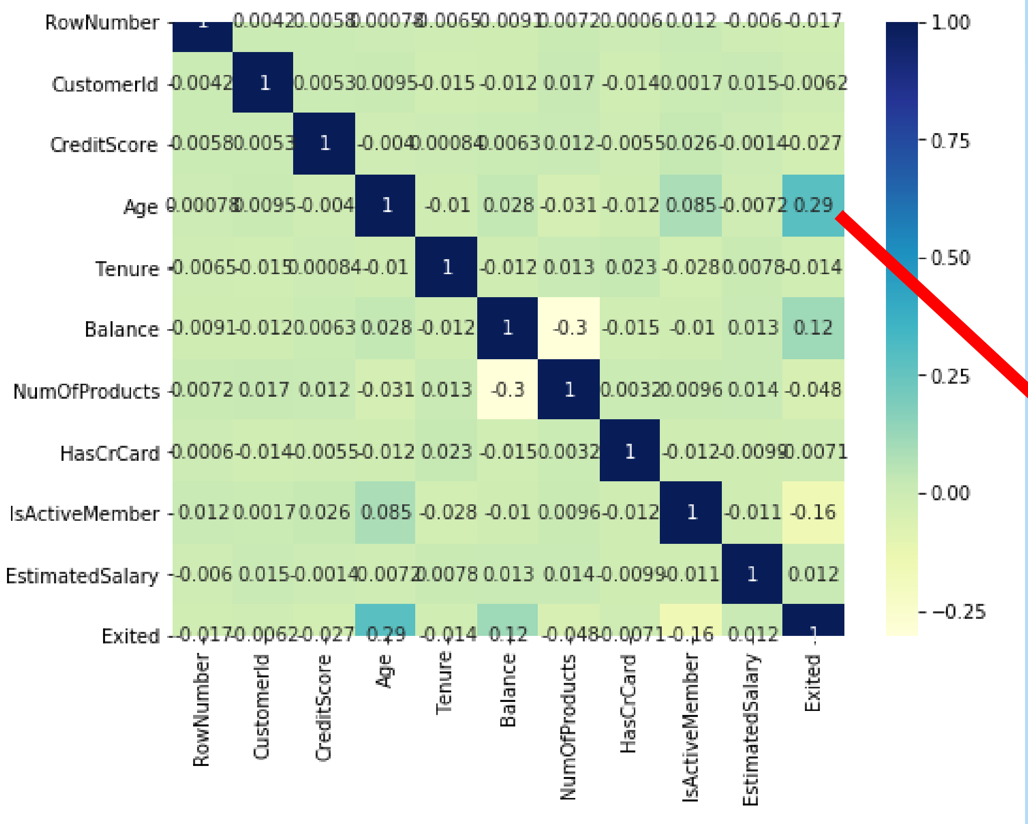
Churn by Country and Bank Balance



We notice that our graph shows highest Churn rates in years labeled 1, 3 & 9

CUSTOMER CHURN BY YEAR 1 THROUGH 11

DOES THIS CONNECT? LET'S LOOK OUR RELATIONSHIPS USING CORRELATION



The strongest correlation I noticed was the Age variant (.29) which corresponds to my Age box Plot.

STEP 3 – DATA WRANGLING

```
In [13]: code.isnull().sum()

Out[13]: RowNumber          0
         CustomerId         0
         Surname            0
         CreditScore        0
         Geography          0
         Gender             0
         Age                0
         Tenure             0
         Balance            0
         NumOfProducts      0
         HasCrCard          0
         IsActiveMember     0
         EstimatedSalary    0
         Exited             0
         dtype: int64
```

My Dataset had no “Null” values
so no need to fill or add mean data
From my Data Set I dropped
Columns that had
zero impact on my results:

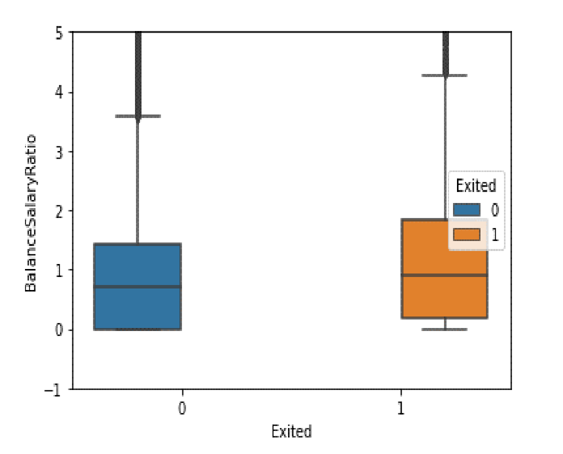
```
code = pd.read_csv('bank_churn.csv')
code.drop(['Surname', 'CustomerId', 'RowNumber'], axis=1,inplace=True)
```

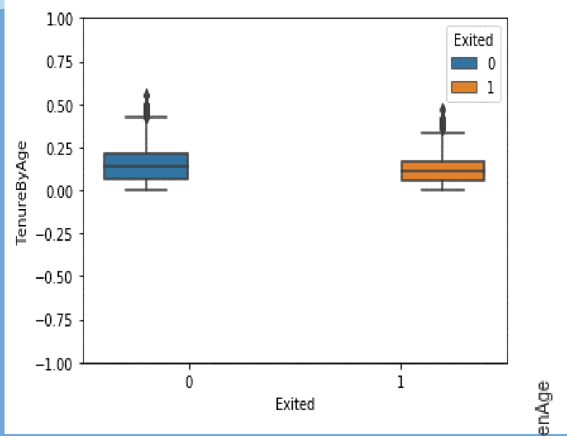
STEP 3 – DATA WRANGLING

I introduce **3 new variables** by combining **6 existing variables** as they had correlations to each other and would improve my analysis.

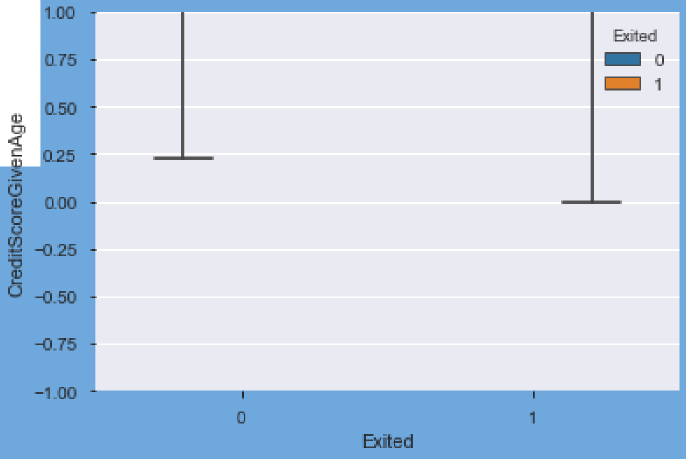
I created 3 new variables that correlated to each other by transforming them into ratios: **Balance/Salary, TenurebyAge, and CreditScore Given Age**

RESULTS OF 3 NEW COMBINED VARIABLES

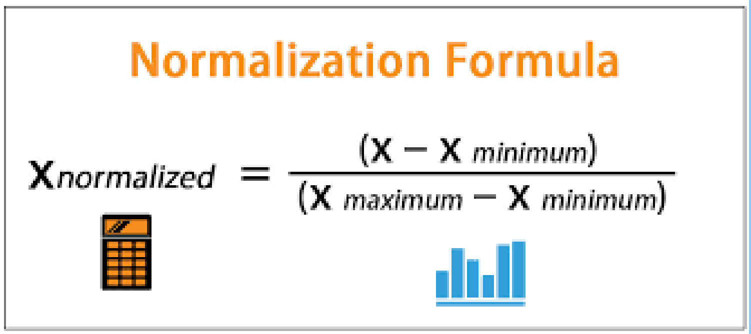




We see visit evidence of a slightly higher Churn rate among those with a higher Balance/Salary Ratio



Step 3 – Data Wrangling - Normalization



I use the min/max operations to scale my "continuous variables" to eliminate unnecessary variances. Min/Max is also known as **"Normalization".** This formula behind this is below: These "normalization" techniques help in comparing corresponding normalized values from two or more different data sets in a way that it eliminates the effects of the variation in the scale of the data sets i.e. a data set with large values can be easily compared with a data set of smaller values.

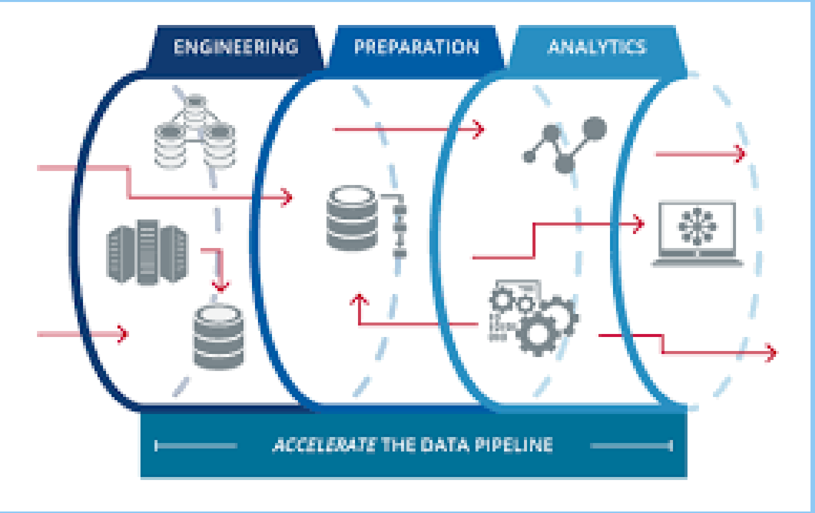
Step 3 – Data Wrangling

Hot Key encoding: In the dataset, there are some variables with numerical values, some variables with categories and some variables with binary values (0 and 1). For numerical and binary variables, we do not worry about labeling. However, we perform label encoding for the categorical variables. This step is carried out on the whole dataset. I "Hot Key" encoded the following variables: Gender and Geography to transform them to binary using a "for"/"if" statement. I performed **"Hot Label Encoding**" where I changed the value "0" (no churn) in the two categorical variables "Has Credit Card" and "Is Active Member" to a -1 to show a negative relationship more clearly.

Step 3 – Data Wrangling

Data splitting: This involves splitting the label encoded dataset into train and test datasets. In this project I separated the data to a 70/30 ratio. The fractions of both classes remain the same in train (70) and test (30) datasets. This is to avoid what is known as “overfitting” of data which related to applying my newly created Churn Model to general unused data (test data) after I train my models on my “train” data set.

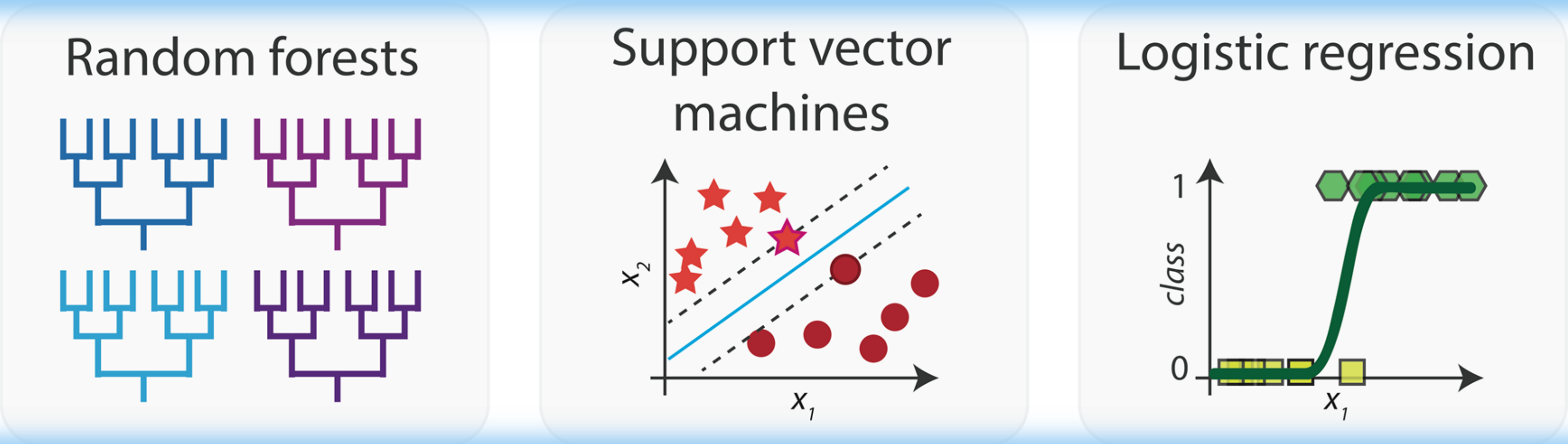
Step 4 – Train and Test the Data



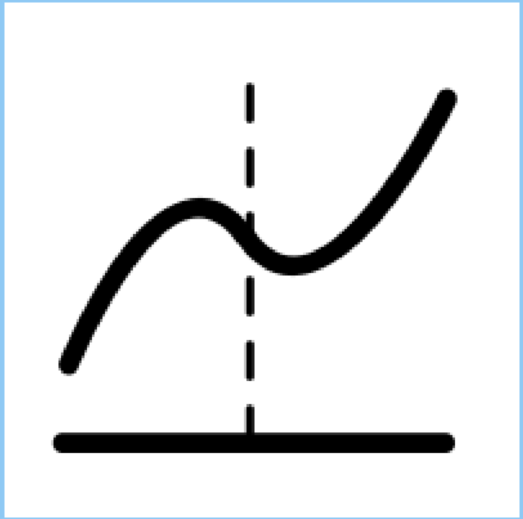
Modeling Pipeline

Next, I build a **“Data Pipeline” In Python** which allows the me to transform data from one representation to another through a series of steps. In other words, to ensure my hot encoding, min/max normalization and both my categorical and continuous variables continue in the test and train modeling

Step 4 – Train and Test the Data – Feeding the Data into an Algorithm

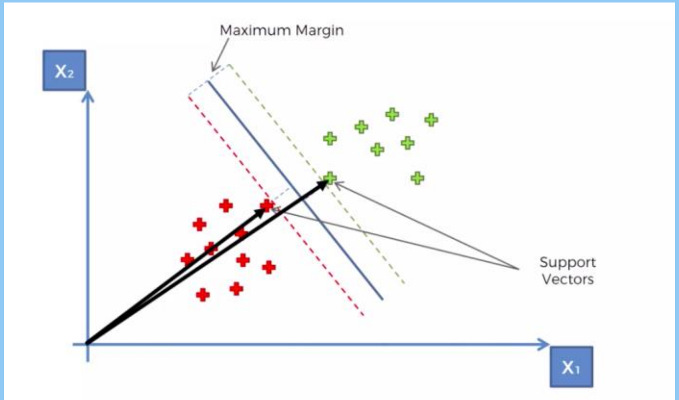


Step 4 – Train and Test the Data – Logistic Regression – Model 1



Logistic Regression: Logistic Regression is one of the basic and popular algorithm to solve a classification problem. It is named as **'Logistic Regression'**, because it's underlying technique is quite the same as Linear Regression. The term "Logistic" is taken from the **Logit function** that is used in this method of classification which uses the Sigmoid function. Since our target variable is **Binary** (either the customer churned or they did not churn) I chose Logistic Regression.

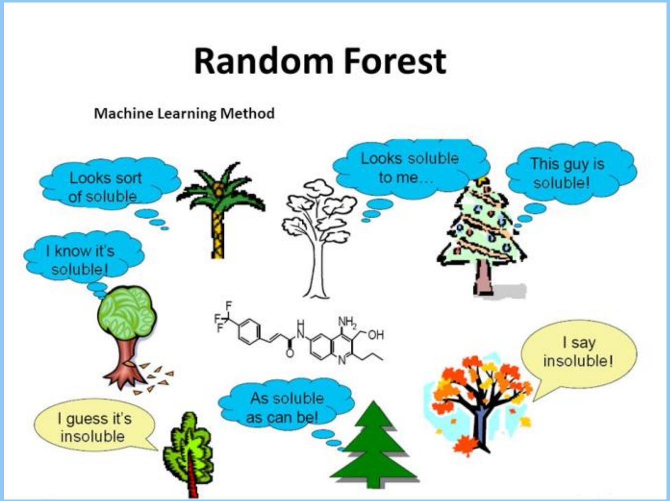
Step 4 – Train and Test the Data – SVM



Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. SVM is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a coordinate. SVM is better suited as I need a way to separate my data into **CHURN or NO CHURN** and Logistic Regression uses a straight line. SVM also maximizes margin, so the model is slightly more robust, but more importantly: SVM supports kernels, so you can model even non-linear relations. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier which my data requires to build the **Churn/No Churn** model as mentioned earlier.

Source: Wikipedia

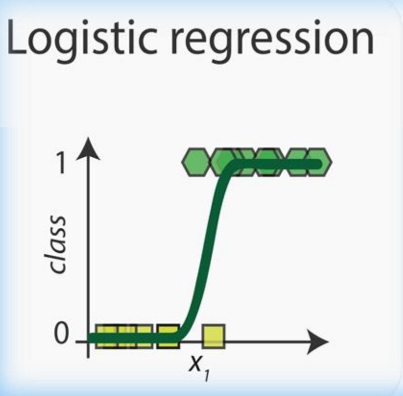
Step 4 – Train and Test the Data – Random Forest



Random Forest works well with a mixture of numerical and categorical features which the Bank Churn data has. When features are on the various scales, it is also fine. Roughly speaking, with Random Forest you can use the data as it is. Random Forest uses a large # of trees, works with missing values and is often considered to be a highly accurate model for both regression and classification problems.

Source: Wikipedia

Step 4 – Lets see both Train and Test Results Side by Side



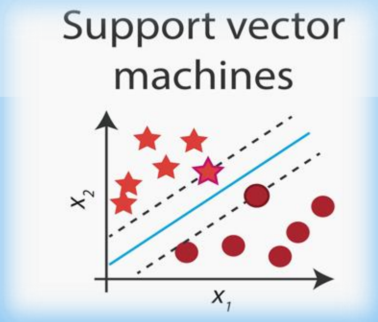
TEST DATA RESULTS

	precision	recall	f1-score	support
0	0.83	0.96	0.89	2411
1	0.57	0.20	0.29	584
accuracy			0.81	2995
macro avg	0.70	0.58	0.59	2995
weighted avg	0.78	0.81	0.78	2995

TRAIN DATA RESULTS

	precision	recall	f1-score	support
0	0.83	0.96	0.89	5547
1	0.61	0.24	0.34	1453
accuracy			0.81	7000
macro avg	0.72	0.60	0.61	7000
weighted avg	0.78	0.81	0.78	7000

Step 4 – Lets see both Train and Test Results Side by Side



TEST DATA RESULTS

	precision	recall	f1-score	support
0	0.88	0.98	0.93	2411
1	0.86	0.45	0.59	584
accuracy			0.88	2995
macro avg	0.87	0.71	0.76	2995
weighted avg	0.88	0.88	0.86	2995

TRAIN DATA RESULTS

	precision	recall	f1-score	support
0	0.87	0.90	0.89	5547
1	0.57	0.50	0.53	1453
accuracy			0.82	7000
macro avg	0.72	0.70	0.71	7000
weighted avg	0.81	0.82	0.81	7000

Step 4 – Lets see both Train and Test Results Side by Side

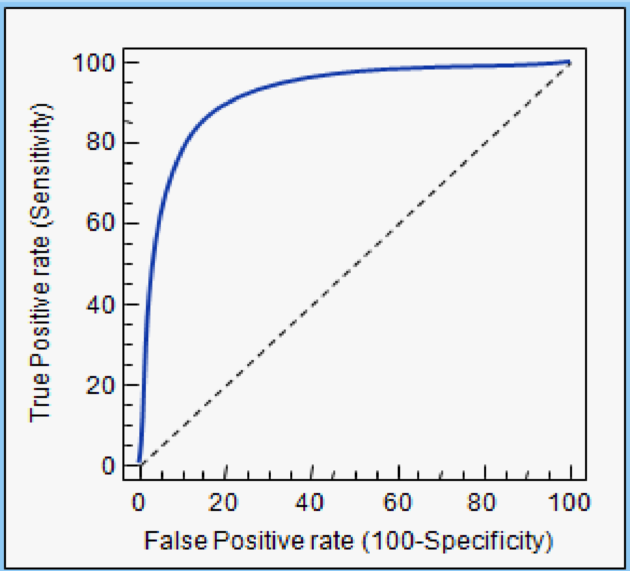
TEST DATA RESULTS

	precision	recall	f1-score	support
0	0.86	0.98	0.92	2411
1	0.85	0.36	0.50	584
accuracy			0.86	2995
macro avg	0.85	0.67	0.71	2995
weighted avg	0.86	0.86	0.84	2995

TRAIN DATA RESULTS

	precision	recall	f1-score	support
0	0.87	0.95	0.91	5547
1	0.72	0.47	0.57	1453
accuracy			0.85	7000
macro avg	0.80	0.71	0.74	7000
weighted avg	0.84	0.85	0.84	7000

Step 5 – Accuracy Check – Did the 3 models make it?



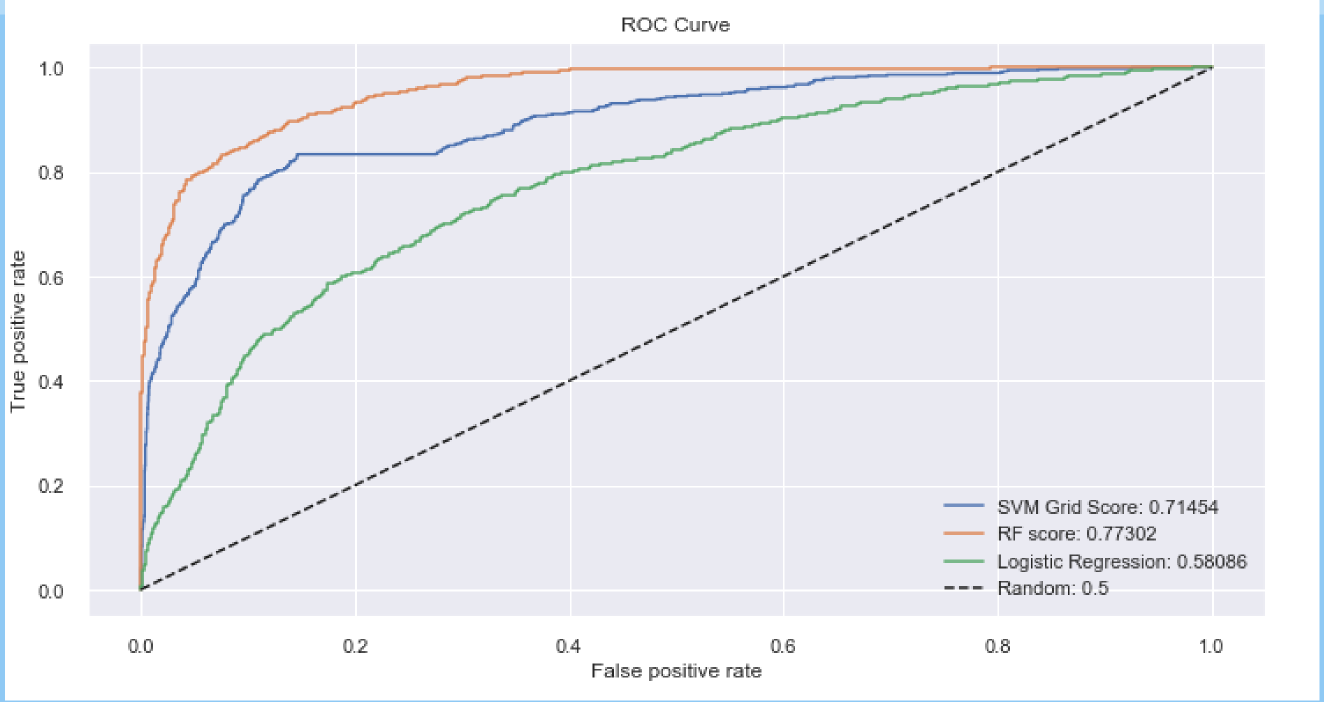
A ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives.

AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and the AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. The Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better the model is at distinguishing between patients with disease and no disease.

The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis.

Source: Toward Data Science

Step 5 – Accuracy Check – Bank Churn Results

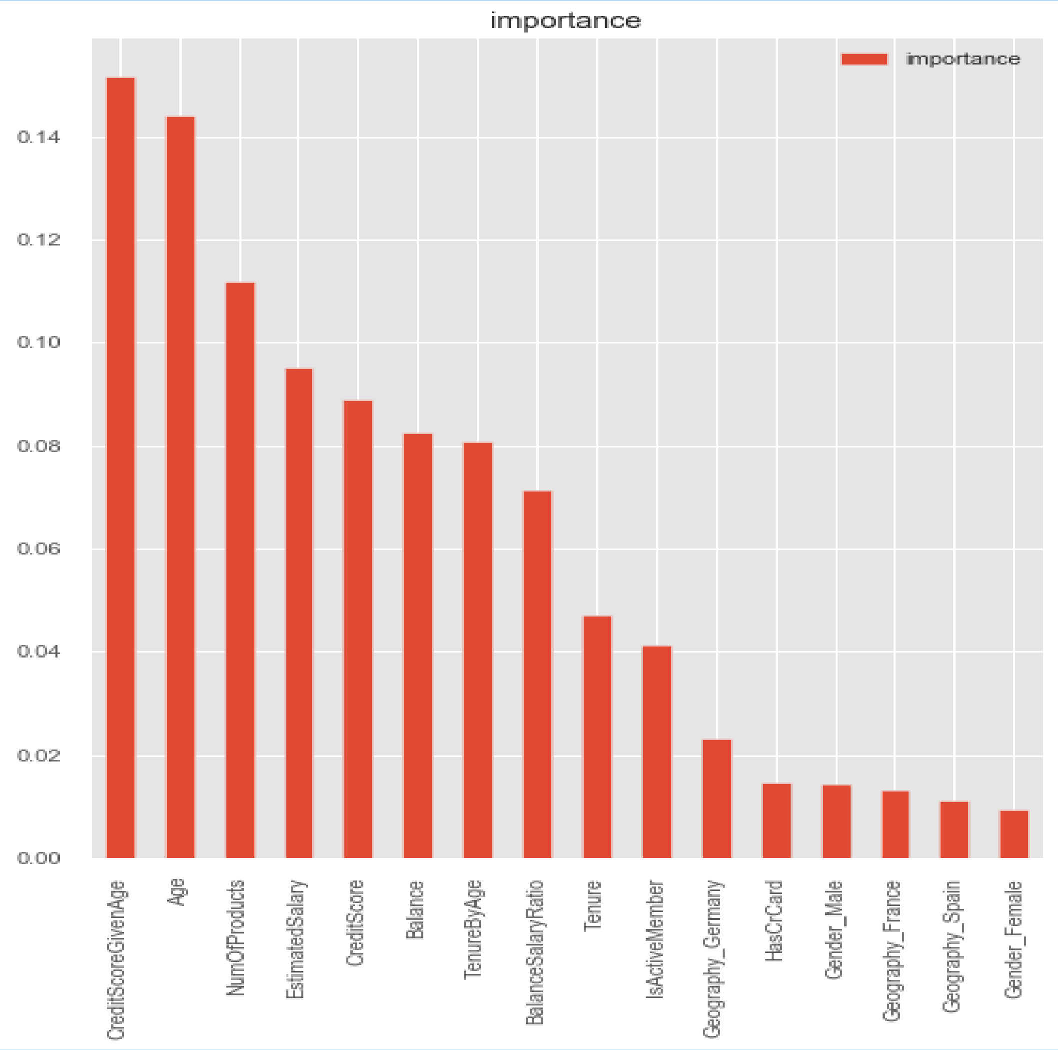


We see that the RF Score had the best results with a .77 which means it is able to correctly classify the client data into a "churn" or "no churn" category .77 of the time.

	Confusion Matrix for Random Forest	
	Predicted by Model	
Actual	Positive	Negative
TRUE	TP = 2373	FN = 38
FALSE	FP = 375	TN = 209

We then add the TP and the TN then divide by our sample of 2995 to get .86 which matches our weighted average in our RF Classification Report.

STEP 5 – ACCURACY CHECK – RANDOM FOREST CONFUSION MATRIX



STEP 5 – FEATURE IMPORTANCE– RANK WHAT MATTERS

We see that Age played a key component in Customer Churn along with Balance.

"I believe that the Gender of each customer has little impact on whether the bank customer will leave or not."

Observed Values : These are the values that were validated from our test (TRUE VALUES)
Stayed | Left
Female 3404 |1139 *3404 females stayed with the bank but 1139 left the bank
Male 4559|898 *4559 Males Stayed with the Bank but 898 left
Expected Values : Below is the expected given outcomes given our data so more females left in reality than what was predicted according to our data. More males were expected to leave though in reality fewer left the bank.
Stayed Left
Female 3617.5909 | 925.4091
Male 4345.4091 |1111.5909
Degree of Freedom:- 1
chi-square statistic:- 113.44910030392086
critical_value: 3.841458820694124
p-value: 0.0
Significance level: 0.05
Degree of Freedom: 1
chi-square statistic: 113.44910030392086
critical_value: 3.841458820694124
p-value: 0.0 - we see our P-value is less than our .05 Significance level so we reject our H0
Reject H0,There is a relationship between 2 categorical variables

STEP 5 – NULL HYPOTHESIS STATEMENT – ACCEPT OR REJECT?

The **Chi-Square test** is intended to **test** how the **Chi-square test** is intended to **test** how likely it is that an observed distribution is due to chance. It is also called a **"goodness of fit"** statistic, because it measures how well the observed distribution of data fits with the distribution that is expected if the variables are Independent.

CHI-SQUARE TEST RESULTS

REJECTED!

We see from the results below the Chi-Square Statistic is high as well as the Critical Value. The p-value is less than our significance of .05 so we reject the HO and go with the HA which is **"Gender has a determining factor if the customer leaves the bank"**

```
Gender
Female   3404   1139
Male     4559    898  We notice more females have left the bank than males
Observed Values :- Our observed values (actual values are different from
our "Expected" Values if there were no relationship)
 [[3404 1139]
 [4559  898]]
Expected Values :-
 [[3617.5909  925.4091]
 [4345.4091 1111.5909]]
Degree of Freedom:- 1
chi-square statistic:- 113.44910030392086
critical value: 3.841458820694124
p-value: 0.0
Significance level:   0.05
Degree of Freedom:
```

STEP 5 – A 2ND NULL HYPOTHESIS STATEMENT – ACCEPT OR REJECT?

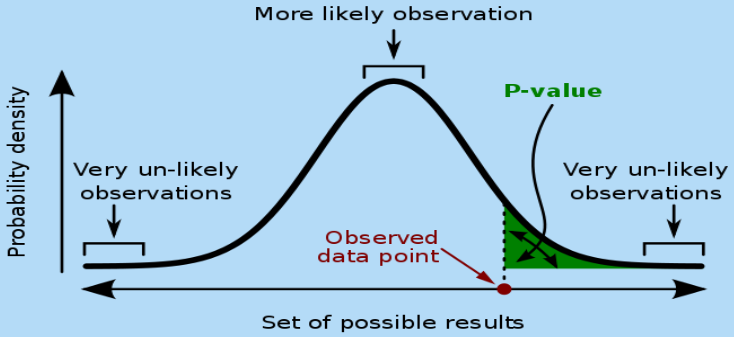
"I believe that Bank Balance has no impact to whether the bank customer will leave or not."

Important:

Pr (observation | hypothesis) ≠ Pr (hypothesis | observation)

The probability of observing a result given that some hypothesis is true is *not equivalent* to the probability that a hypothesis is true given that some result has been observed.

Using the p-value as a "score" is committing an egregious logical error: **the transposed conditional fallacy.**



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Using the "p-value" test I will set .05 as my threshold and run a test to verify that I am correct about my Null Hypothesis (H0)

Below is the best explanation of "p-value"

CLICK HERE

NOW LET'S LOOK DEEPER AT BANK BALANCE

T-Test

I performed an unpaired t-test to compare the mean Bank Balance for Retained vs. Churned customers. Retained customers have an average bank balance of *$91,00, vs. $101,465* among Churned customers. The test yielded a t-statistic of 4.10 with a p-value of 0.00067. The null hypothesis is that the two groups have the same mean, and the t-statistic and p-value tell us it is very unlikely that we would observe such large differences in the means assuming the two groups have the same mean. Thus, with a p-value < 0.05, we reject the null hypothesis in favor of the alternative hypothesis that there is a relationship between bank balance and whether the customer churned or not.

FINAL THOUGHTS AND CONCLUSIONS

Through my examination of the small data set I did discover a few significant findings:

- There is no significant difference in the credit score distribution between retained and churned customers.
- The older customers (over 35) are churning at a higher rate than the younger ones alluding to a difference in service preference in the age categories. The bank may need to review their **target market** or review the strategy for retention between the different age groups
- Bank members with an average tenure are slightly less likely to churn than those with either low or high number of tenure years.

FINAL THOUGHTS AND CONCLUSIONS

Continued….

- The data shows that customers with higher balances are churning at a higher rate which is cause for concern for their lending capability. The bank could benefit from offering special programs when, say, a balance of $75,000 and offer a higher rate of interest on a savings account or special investment privileges.
- Neither the product nor the salary has a significant effect on the likelihood to churn.
- More females have churned than males
- More credit card holders churn though most of the bank customers possess credit cards. The bank can benefit from increasing incentives in keeping credit card holders.

FINAL THOUGHTS AND CONCLUSIONS

Model Review

My Random Forest Model performed the best as it was able to capture both the True Positives and True Negative outcomes .86 out of 100. This is a strong result and in this model an analyst is not required to scale/normalize the data.

According to my Feature Selection results from my Random Forest model, the most important determinants of Churn are Age, Gender and Bank Balance followed by Tenure and Geography.

I think developing an accurate predictive model should be incorporated in every bank's business plan as the closer a bank can get to modeling the characteristics of those who may leave, the more detailed a retention plan the bank can create and implement. Insights carefully extracted from data are invaluable to the health of the competitive nature of banks, especially credit cards, bank-balance specific programs and other key ancillary features a bank provides to keep their clients.

OTHER MODELS

I would like to train banking data using the Naïve Bayes Classifier as this assumes independence between every pair of features and accounts for new changes relating to the features and is quick to run. I would also like to try K nearest neighbors as it is a classification is a type of lazy learning as it does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the k nearest neighbors of each point and is robust against noisy data. It will also work well for a Bank Data set with over 100,000 observations.

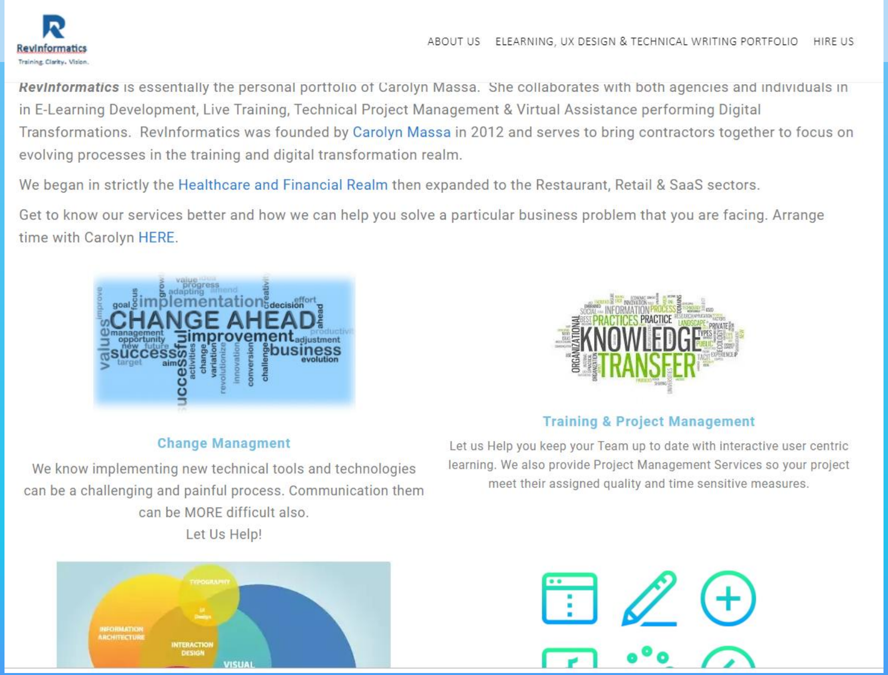
FURTHER RESEARCH….

The study can be greatly improved with the following data since there are many unanswered questions:

- Would it be possible to obtain balances over time as opposed to a single date?
- What date did the customer exit?
- What types of products are the customers in?
- Could they have exited from a product and not the bank?
- Does the bank have an investment division?
- Did the customer retire and consolidate assets elsewhere?
- Are there NPS scores to factor into the data?

Of course, every business needs to perform analysis and take measures to prevent Customer Churn; considering the cost of acquiring each customer, a study should be an annual requirement and perhaps creating ABM (Accounts Based Marketing) plans to target key client groups.

Carolyn Massa



- *Technical Trainer/Project Manager entrepreneur*
- *Masters in Information Technology*

Personal site: revinformatics.com

GitHub: https://github.com/CarolynMassa

Email: carolyn@revinformatics.com

LinkedIn:

https://www.linkedin.com/in/carolynmassa

Portfolio