SPRINGBOARD CAPSTONE SLIDEDECK

UNDERSTANDING AND PREDICTING BANK CUSTOMER CHURN

Momin Asadullah Khan

WHY?

My motivation:
- Interest in Cognitive Science, Human Behavior and Psychology

Promotional offers:
- Always get offers to join banks in the mail

Curious about why customers leave and if it could be predicted

OBJECTIVE

THE IMPLEMENTATION OF THIS MODEL WILL ALLOW THE BANK TO MAKE INFORMED DECISIONS REGARDING CUSTOMERS.

1. To understand what factors contributed most to customers leaving the bank.
2. To create a model that predicts if a customer will leave the bank or not.
3. To create or improve existing customer retention strategies e.g. incentive offers

ANALYTICAL PIPELINE

1. First approach in solving the problem is obtaining the relevant data.
2. Cleaning the data is the next step. This includes checking for null values, imputing missing values and checking column names.
3. Next, we explore the data to gain insights into our dataset and what it contains. This includes looking for outliers or unusual data as well as using a correlation matrix to understand the relationship between the response and predictor variables.
4. Modeling using Machine Learning gives us our predictive power on whether the bank will lose a customer or not.
5. Lastly, we interpret the results and see what can we conclude about bank customers leaving the bank. For example, what feature variable was most indicative of customers leaving the bank?

THE PROBLEM

The bank is at risk of losing billions from customers so it wants to understand why they are leaving and if they can reach them before they exit the bank.

The bank could lose up to $50 billion if the customers decided to take their business to another bank. Some common excuses for leaving are:

- Excessive/hidden fees
- Bad customer service
- Checks/funds bouncing
- Most expensive debits charged first
- Loyalty means nothing

SOLUTION

Retention Plan

The goal is to create a **retention plan**!

We can help identify who is at risk of leaving the bank and provide outreach to prevent churn.

The model will predict and calculate the likelihood of each customer exiting the bank.

THE DATASET

- RowNumber: Row Number
- CustomerId: Customer ID
- Surname: Customer's Surname
- CreditScore: Customer's Credit Score
- Geography: Country of Customer
- Gender: Customer's Gender
- Age: Customer's Age
- Tenure: Customer's Tenure
- Balance: Customer's Balance
- NumOfProducts: How many accounts, bank account affiliated products the person has?
- HasCrCardDoes: Does the customer have a credit card through the bank?
- IsActiveMember: Subjective, but for the concept
- EstimatedSalary: Customer's Estimated Salary?
- Exited: Did they leave the bank after all?

THE TABLE

	Exited	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary
0	1	15634602	Hargrave	619	0	1	42	2	0.00	1	1	1	101348.88
1	0	15647311	Hill	608	2	1	41	1	83807.86	1	0	1	112542.58
2	1	15619304	Onio	502	0	1	42	8	159660.80	3	1	0	113931.57
3	0	15701354	Boni	699	0	1	39	1	0.00	2	0	0	93826.63
4	0	15737888	Mitchell	850	2	1	43	2	125510.82	1	1	1	79084.10

SUMMARY – EXITED VS NOT EXITED

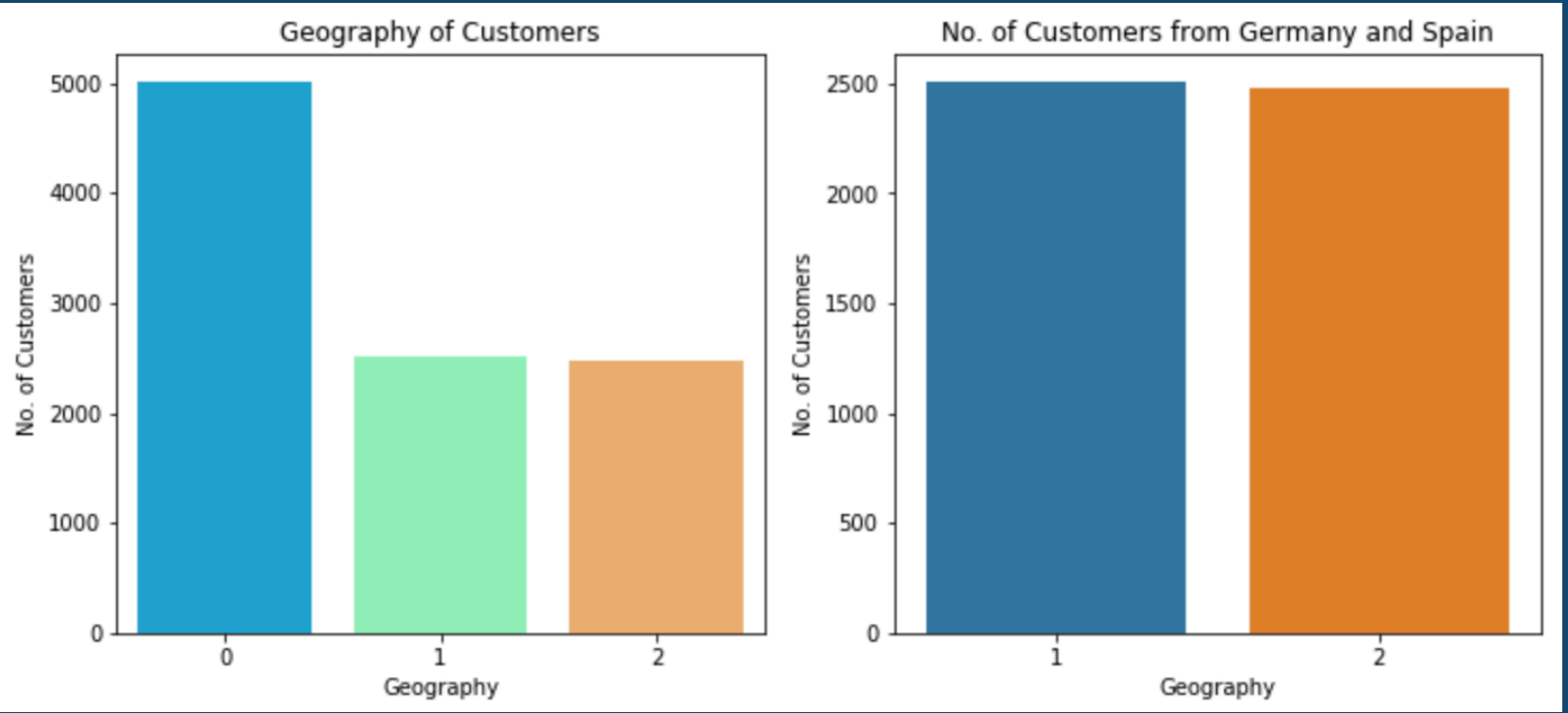
- Dataset has 10,000 entries
- Class Imbalance Problem (Exited vs Non Exited)
- The bank had an exit rate of 20.4%
- The mean credit score of customers was 650.52
- 0 = France, 1 = Germany and 2 = Spain
- 0 = Male and 1 = Female

	CustomerId	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary
Exited											
0	1.569117e+07	651.853196	0.731257	0.427477	37.408389	5.033279	72745.296779	1.544267	0.707146	0.554565	99738.391772
1	1.569005e+07	645.351497	0.805106	0.559156	44.837997	4.932744	91108.539337	1.475209	0.699067	0.360825	101465.677531

CUSTOMER BREAKDOWN - GEOGRAPHY

0 = France, 1 = Germany and 2 = Spain

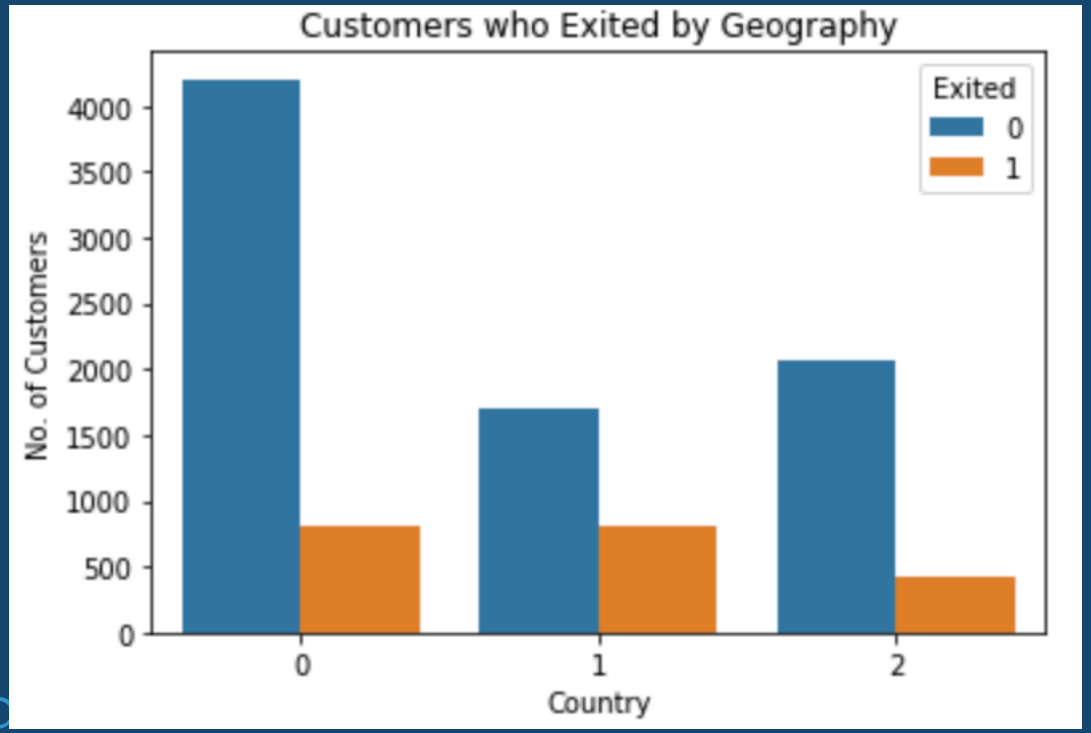
There were twice as many customers from France than from Spain or Germany.



CUSTOMERS WHO EXITED - GEOGRAPHY

0 = France, 1 = Germany and 2 = Spain

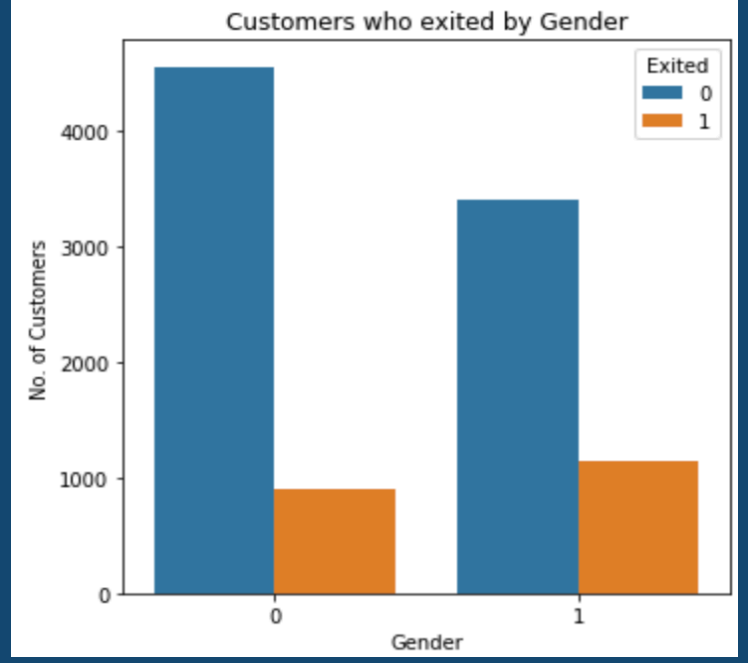
We can see Germany had the most customers who exited followed by France and Spain.

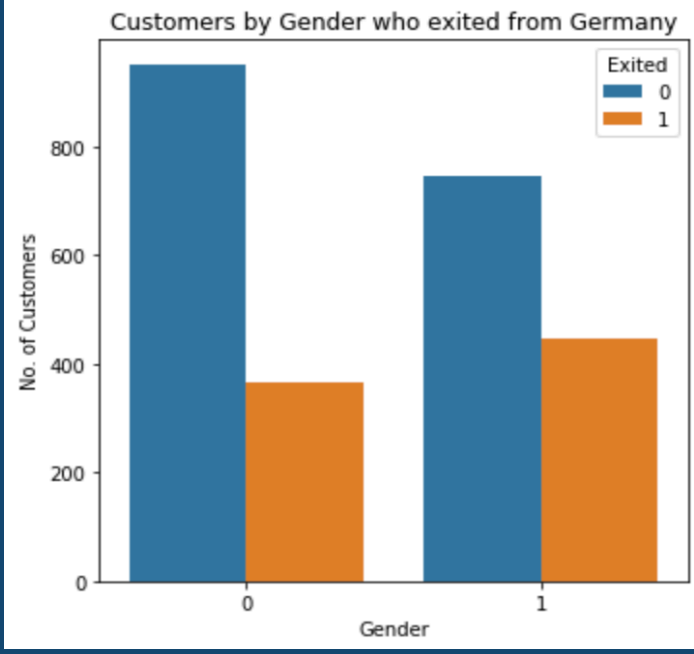


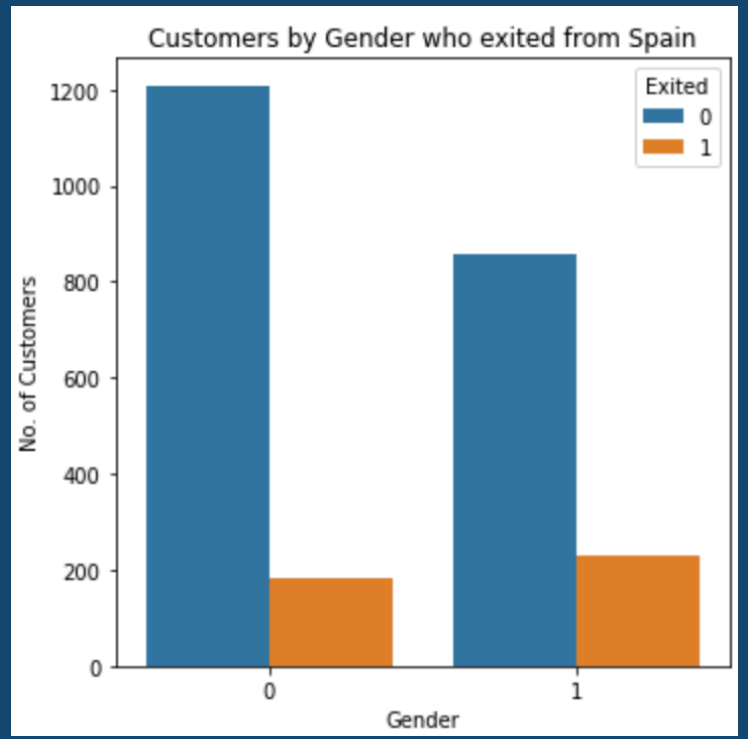
CUSTOMERS WHO EXITED – GENDER & GEOGRAPHY

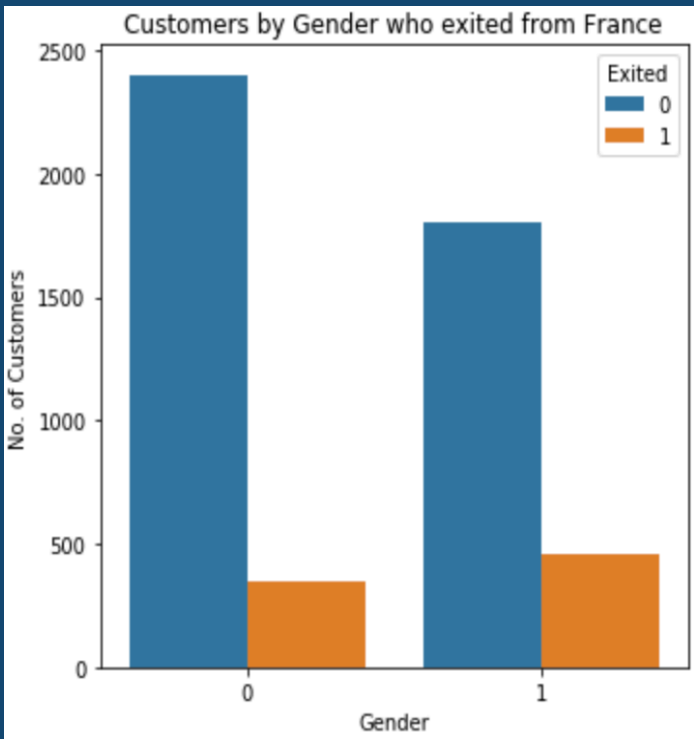
The first plot shows the breakdown of customers who exited by their gender and the rest of the plots look at customers by gender who exited from France, Germany and Spain.

From these plots, we can see that there were a higher number of female customers who exited, France had the highest number of female customers who exited the bank and Germany had the highest number of Male customers who exited the bank.



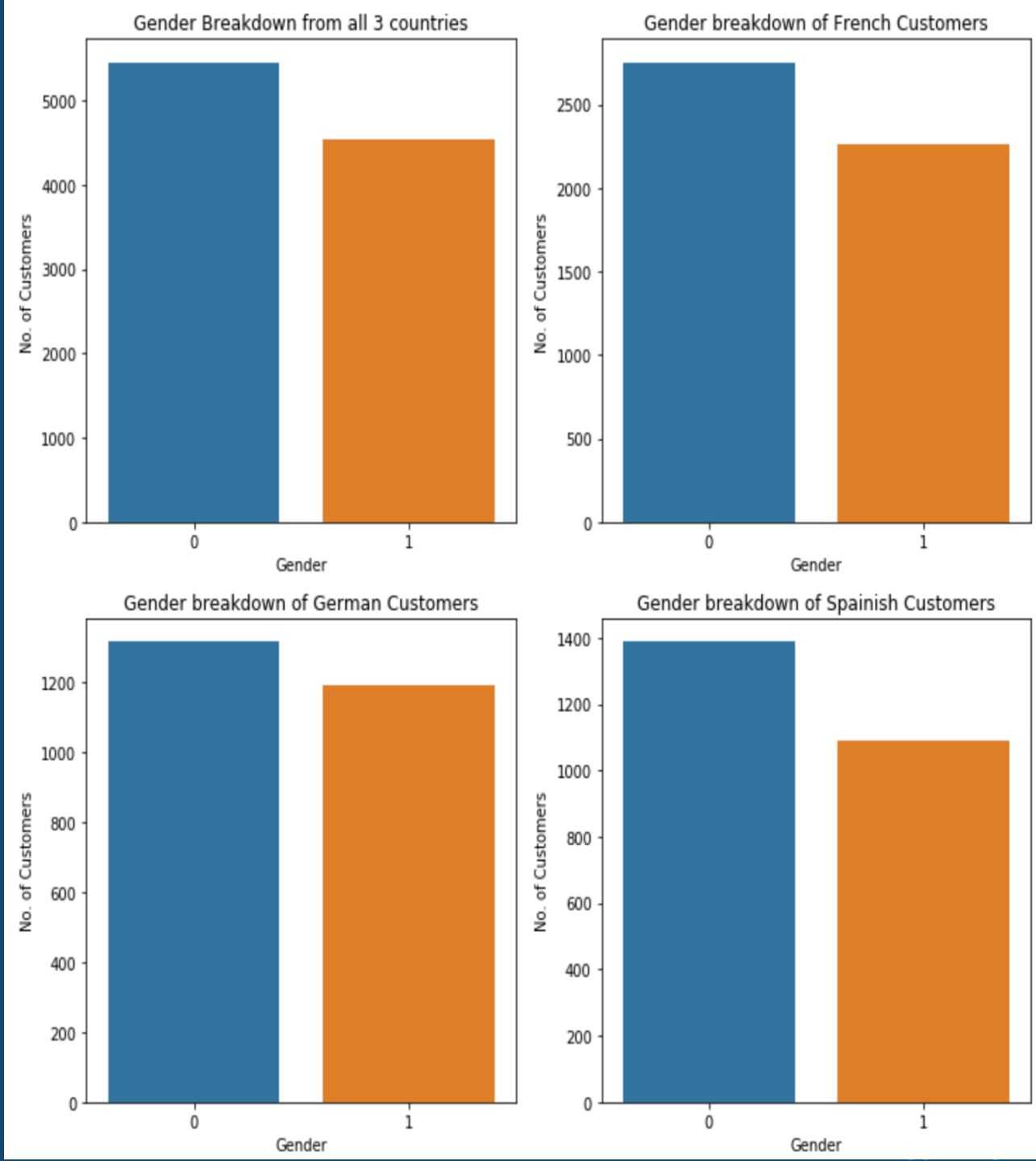






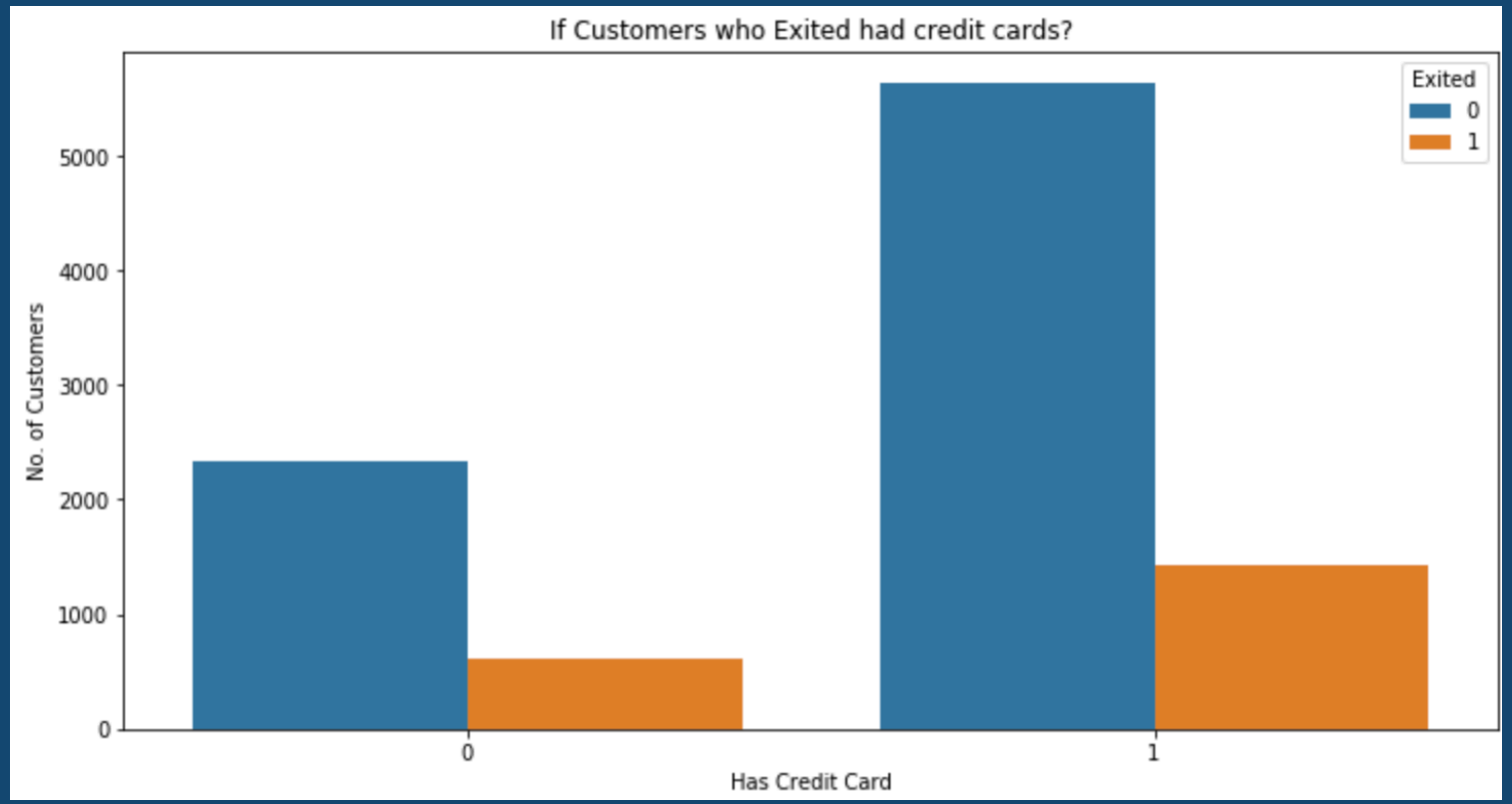
CUSTOMERS BY GENDER AND COUNTRY

The plots look at the gender breakdown of all customers and customers from each country. We can see that France has the highest female and male customers



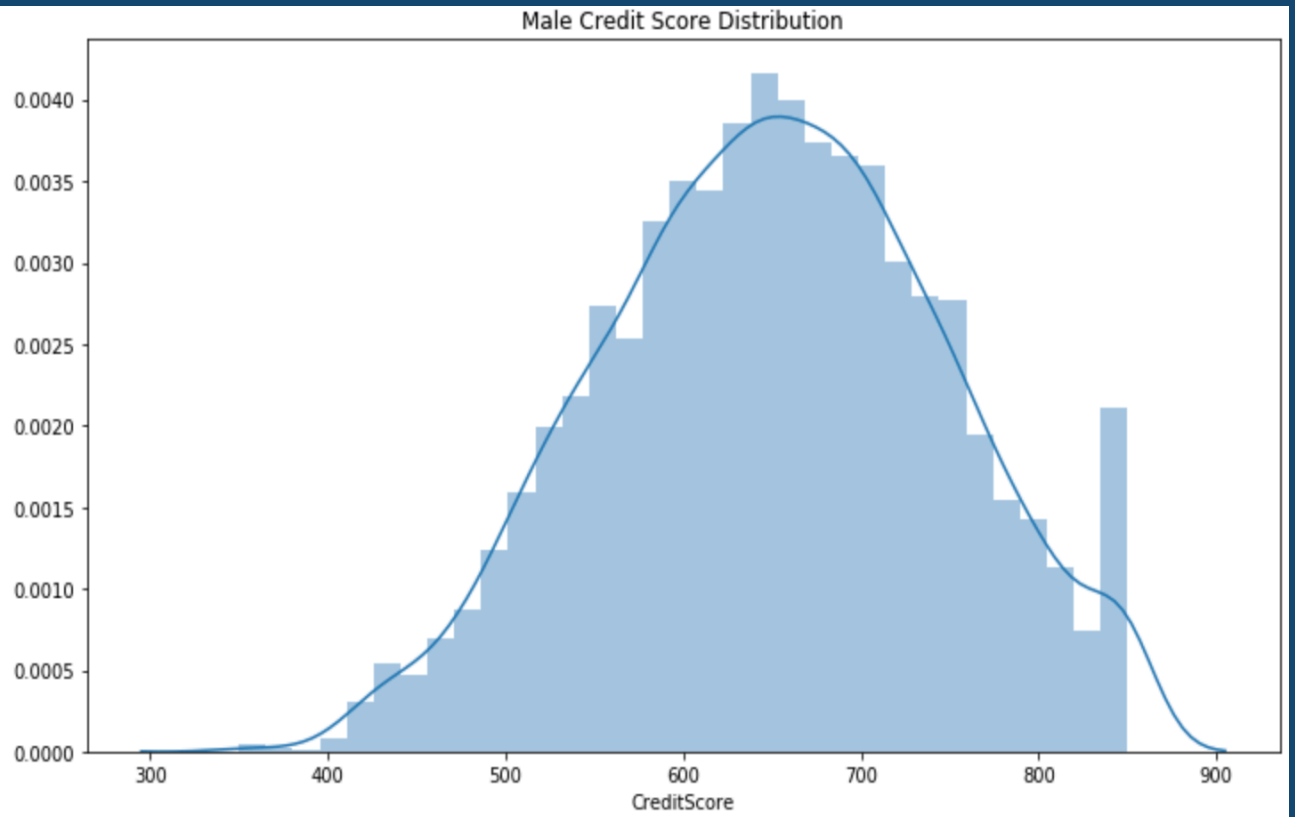
CUSTOMERS EXITED – CREDIT CARD

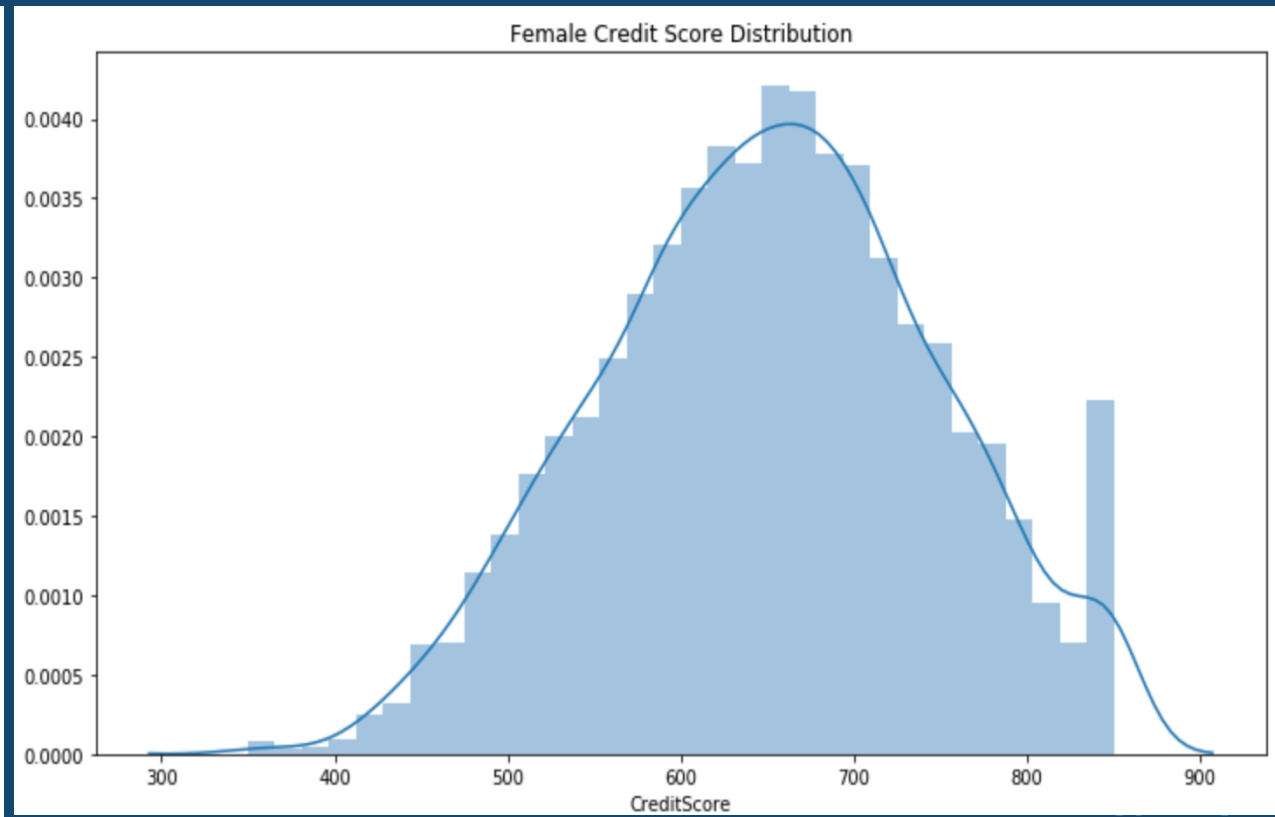
The plot shows whether customers had a credit card and if they exited. Customers who had a credit card exited more than customers who did not possess a credit card



CREDIT SCORE DISTRIBUTION – BY GENDER

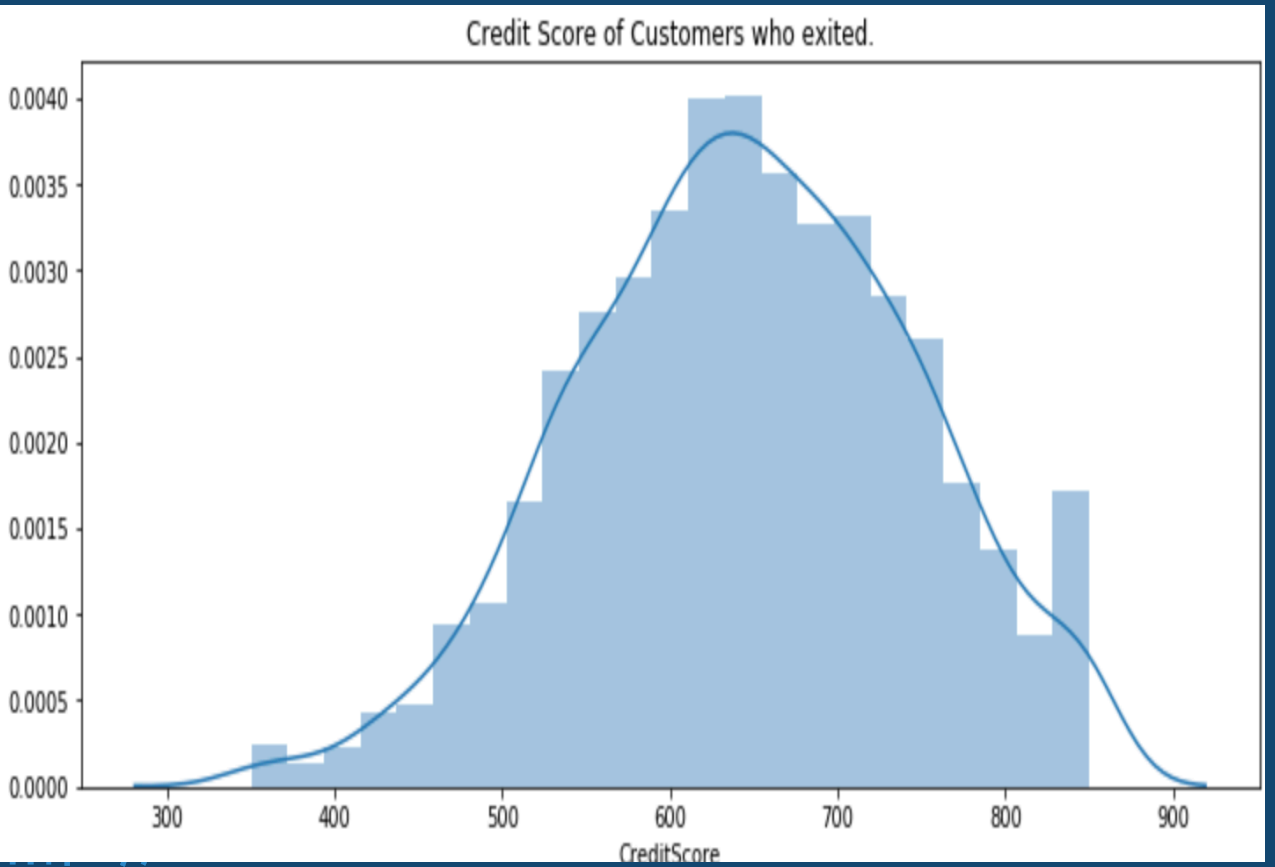
The plots above look at the credit score distribution of female and male customers. There is no significant difference in the distribution of both gender's credit scores

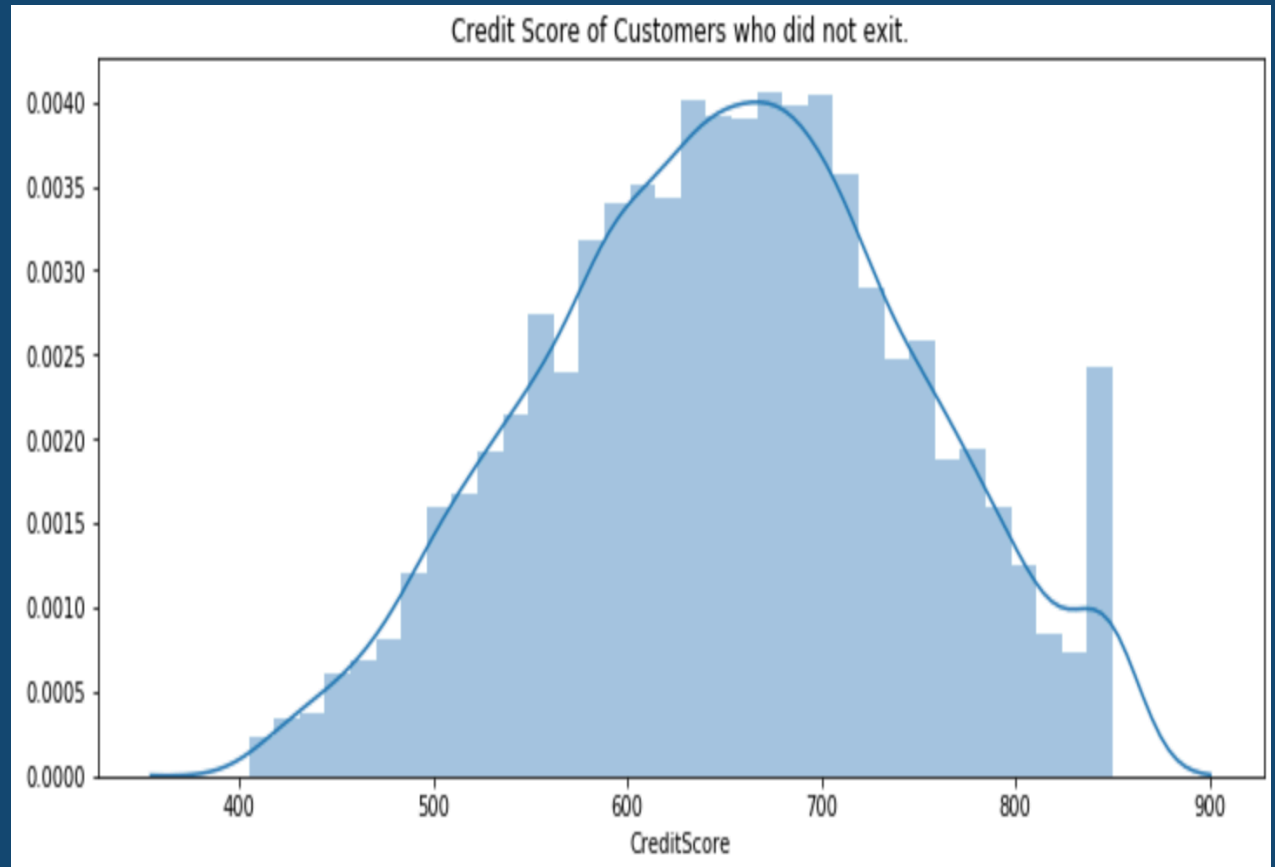




CREDIT SCORE DISTRIBUTION – EXITED VS NOT EXITED

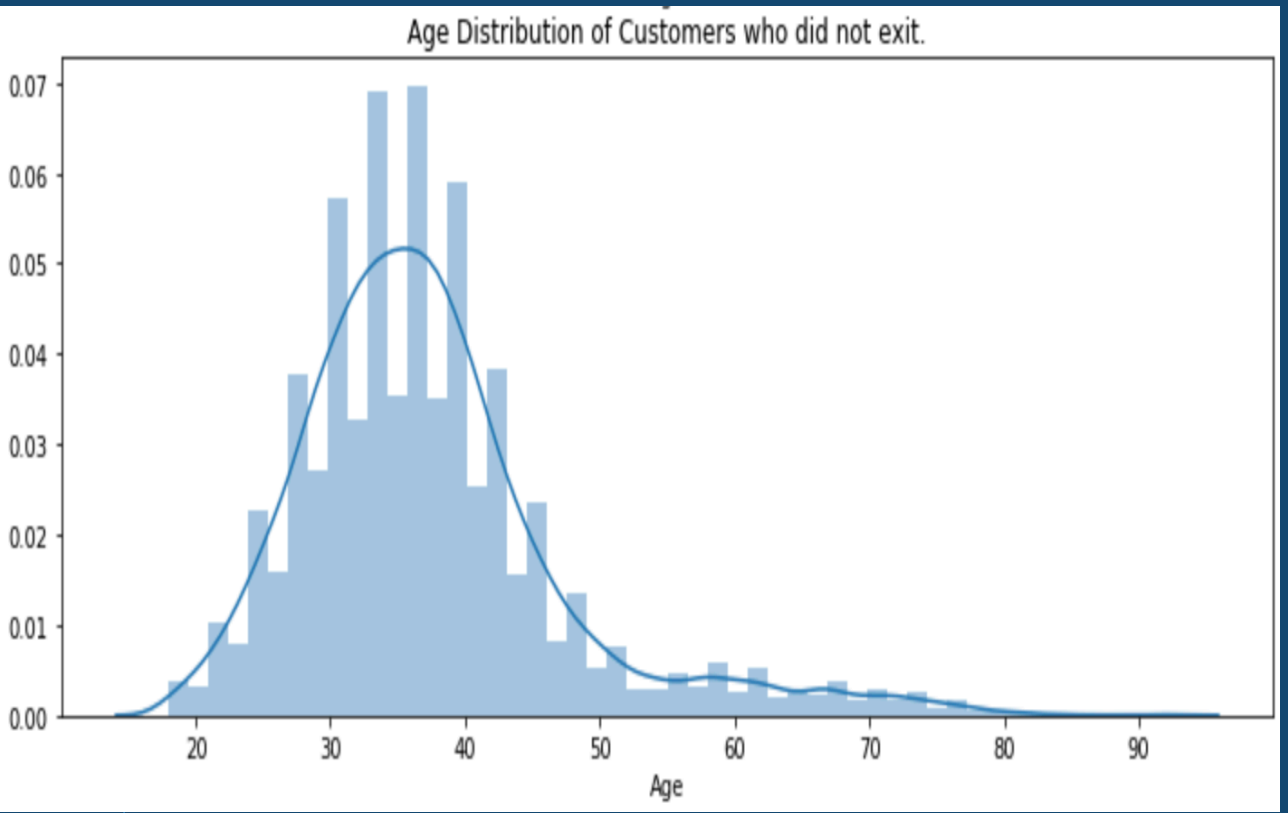
The plots above show the credit score distribution of customers who exited and did not exit. Customers who exited had a roughly lower average credit score.

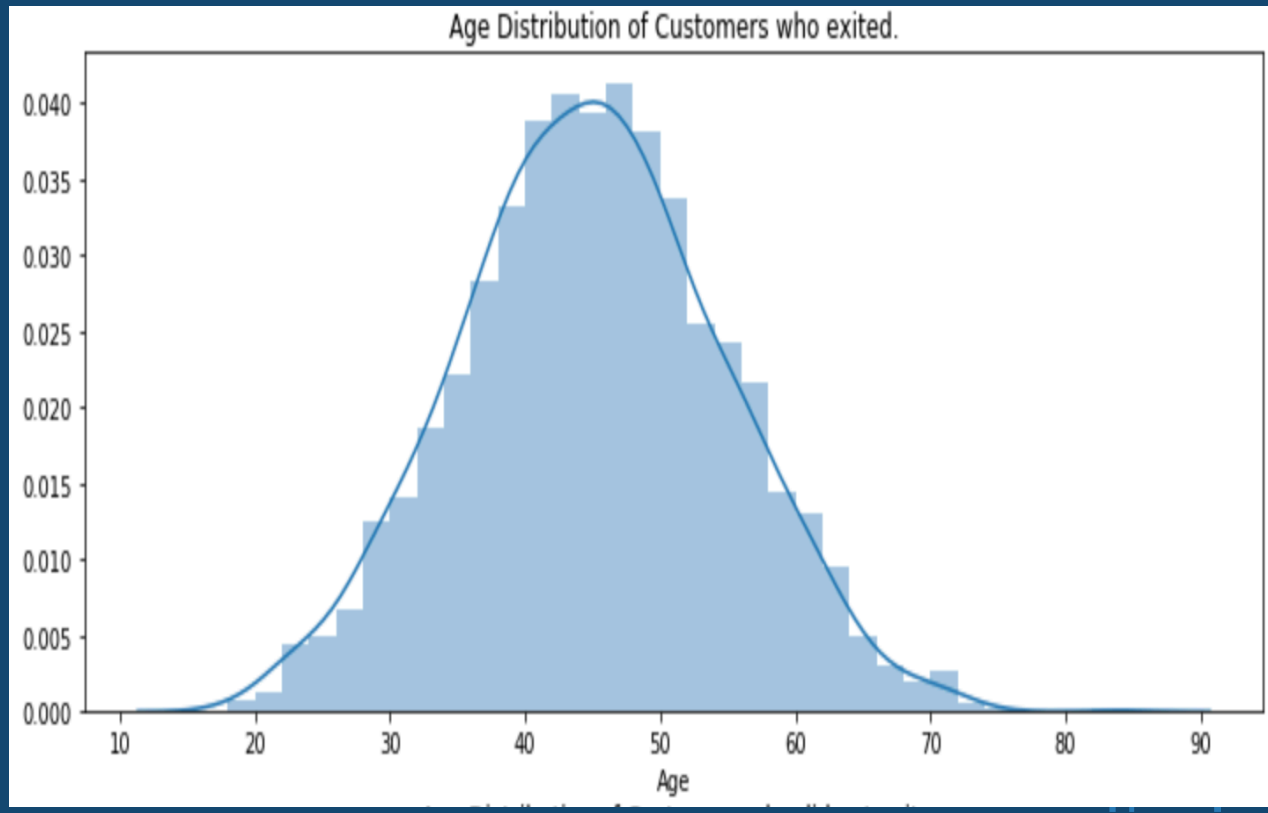




AGE DISTRIBUTION – EXITED VS NOT EXITED

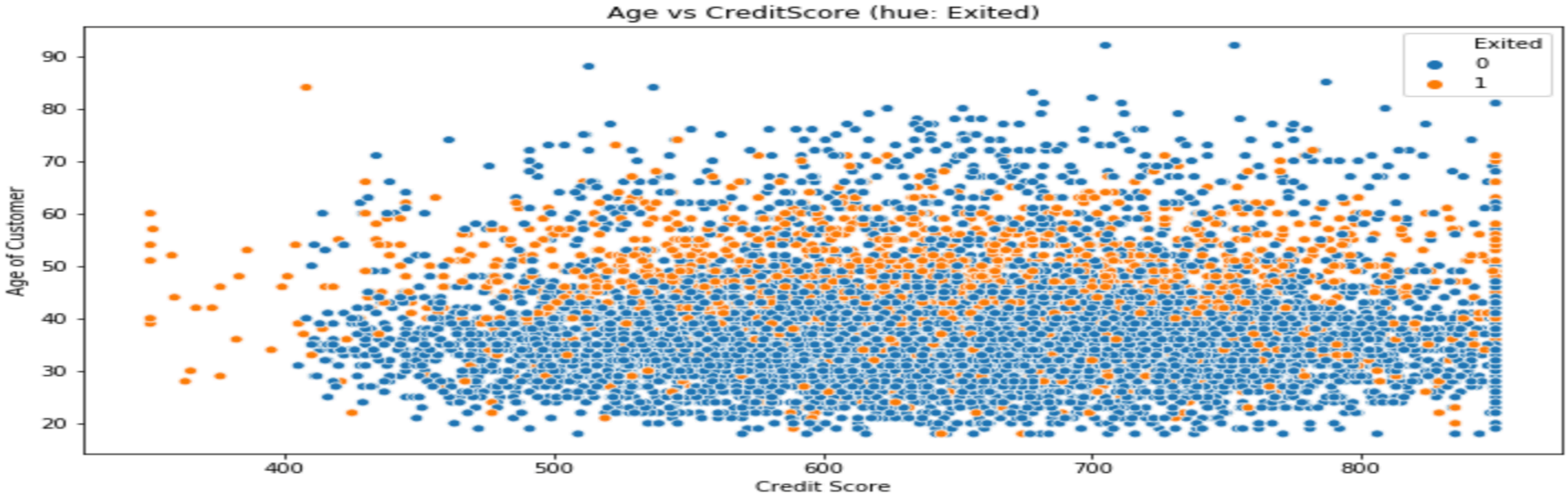
The plots below show the age distribution of customers who exited and who did not exit. The age distribution of customers who exited looks close to normally distributed while the age distribution of customers who did not exit is positively skewed.



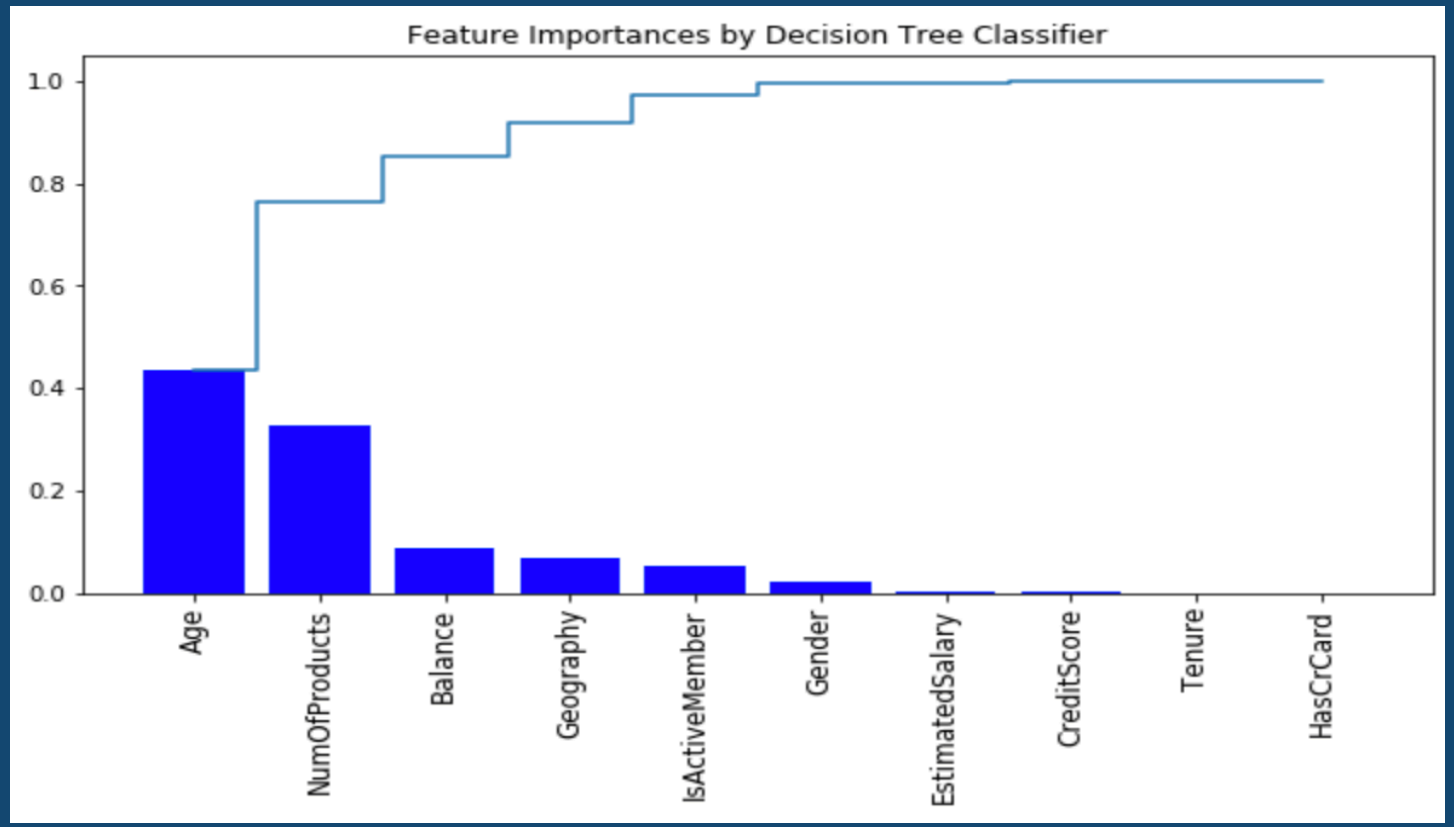


SCATTER PLOT – AGE VS CREDIT SCORE (HUE: EXITED)

The plot below is a scatterplot of Credit Score and Age with the points color coded by whether customers exited or not. It can be seen that the majority of the people who exited were between 40 and 60 years old



DECISION TREE – FEATURE IMPORTANCE

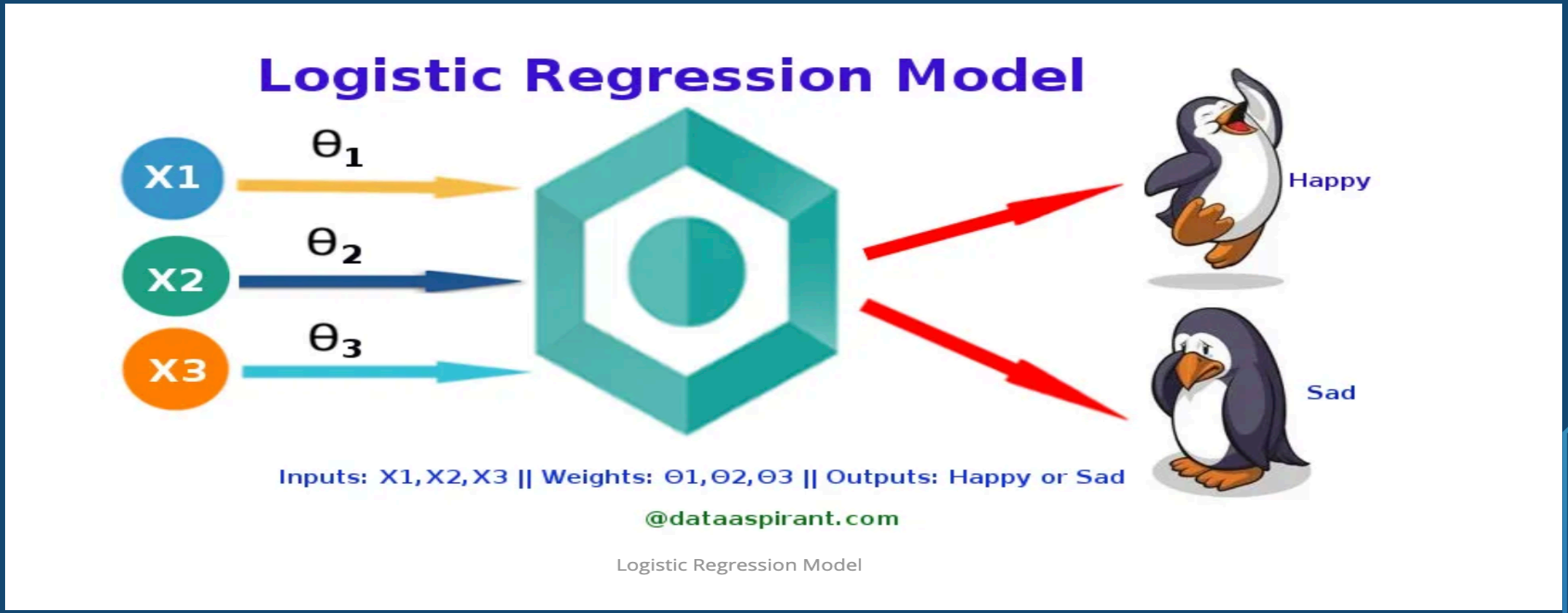


Top 3 Features:

1. Age
2. Number of Products
3. Balance

INTRODUCTION TO LOGISTIC REGRESSION – BASE RATE MODEL

The Logistic Regression Classifier was chosen as the base rate model. A base rate model is used to compare how other models perform against it.

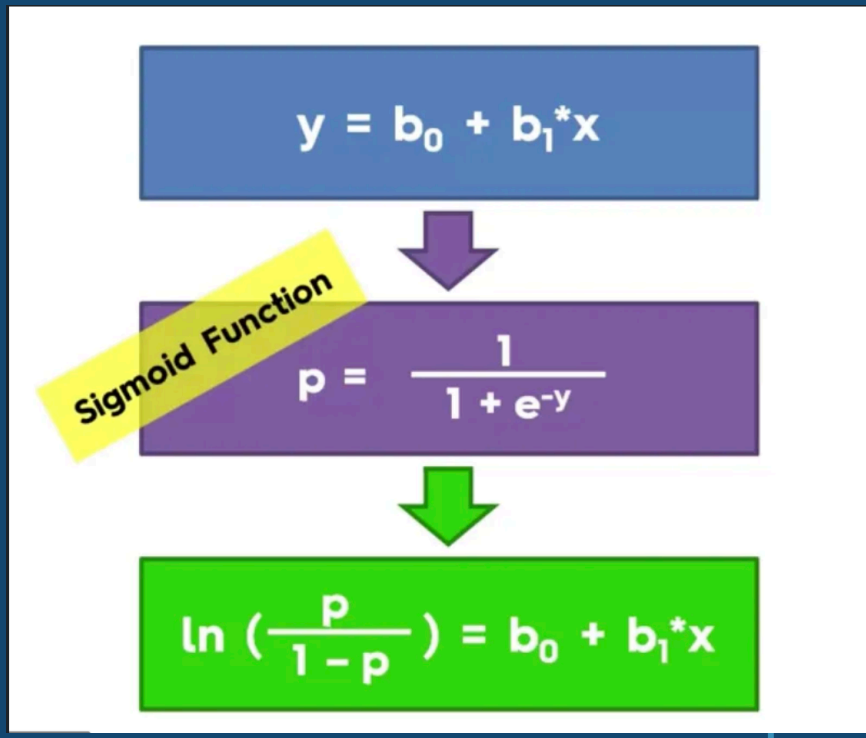


Logistic Regression Model

LOGISTIC REGRESSION – HOW IT WORKS

$$\text{logit}\left[\theta(x)\right] = \log\left[\frac{\theta(x)}{1-\theta(x)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

- Θ(x) – Dependent Variable (in our analysis, 'Exited')
- (xi) – Independent Variable (the variable that predicts the event)
- α - Constant
- B – coefficient of predictor variable
- Logit function (aka sigmoid function)



CLASS IMBALANCE – EVALUATION METRIC

- This dataset is an example of a class imbalance problem because of the uneven distribution of customers who did and did not exit the bank. To handle this, use the SMOTEENN method which combines over- and under-sampling using SMOTE and Edited Nearest Neighbors
- In this case, evaluating our model's algorithm based on accuracy is the wrong thing to measure. We will have to take into consideration the False Positive and False Negative Errors and use that as a metric to evaluate our model's performance.
- False Positives (Type I Error): You predict that the customers will leave, but do not.
- False Negatives (Type II Error): You predict that the customer will not leave, but does leave.
- Which error should we be more concerned about?

HYPERPARAMETER TUNING

AdaBoost Classifier :

```
{'learning_rate': 0.1, 'n_estimators': 1000}
```

SVM :

```
{'C': 0.1, 'class_weight': 'balanced'}
```

Decision Tree :

```
{'class_weight': 'balanced', 'criterion': 'gini', 'max_depth': 8, 'min_weight_fraction_leaf': 0.001}
```

Random Forest:

```
{'class_weight': 'balanced', 'max_depth': None, 'min_samples_split': 10, 'min_weight_fraction_leaf': 0.001, 'n_estimators': 1000}
```

ALL MODEL EVALUATIONS – CLASSIFICATION REPORT

-- Logistic Regression Model --
-- Logistic Regression Model AUC = 0.68 --

	precision	recall	f1-score	support
0	0.89	0.71	0.79	1593
1	0.36	0.65	0.47	407
accuracy			0.70	2000
macro avg	0.63	0.68	0.63	2000
weighted avg	0.78	0.70	0.72	2000

---Decision Tree Model---
Decision Tree AUC = 0.77

	precision	recall	f1-score	support
0	0.93	0.77	0.84	1593
1	0.46	0.78	0.58	407
accuracy			0.77	2000
macro avg	0.70	0.77	0.71	2000
weighted avg	0.84	0.77	0.79	2000

---SVM Model---
SVM AUC = 0.76

	precision	recall	f1-score	support
0	0.92	0.77	0.84	1593
1	0.45	0.74	0.56	407
accuracy			0.76	2000
macro avg	0.69	0.76	0.70	2000
weighted avg	0.83	0.76	0.78	2000

---AdaBoost Tree Model---
AdaBoostClassifier AUC = 0.77

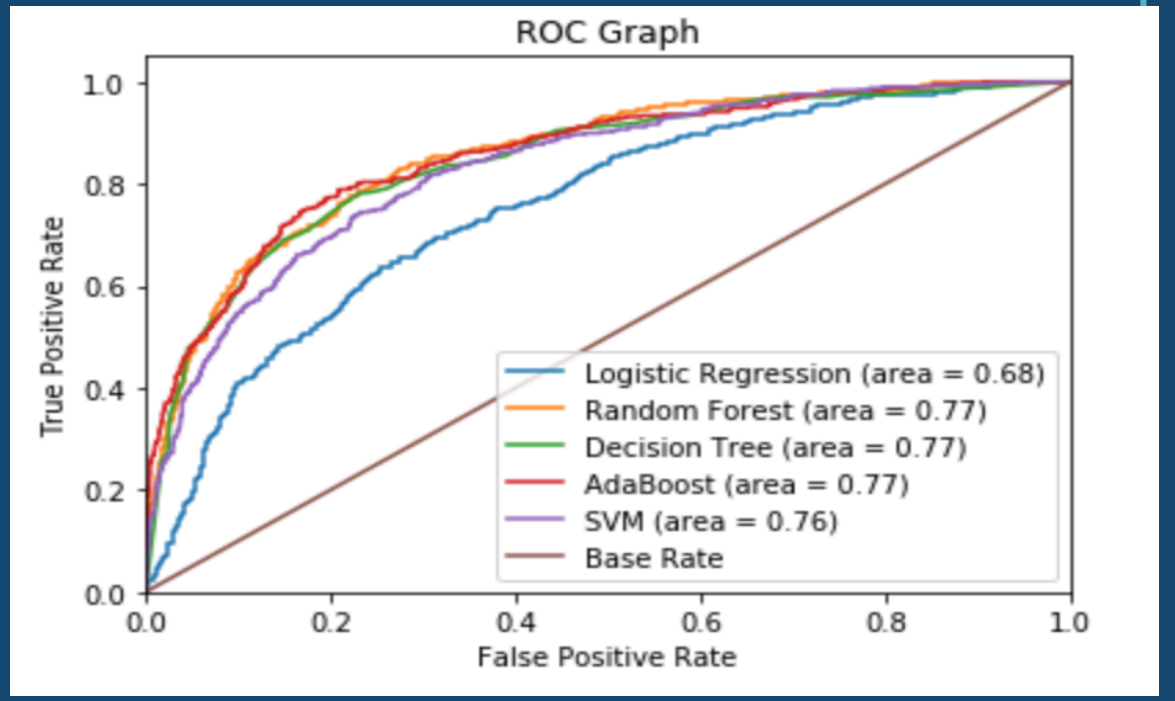
	precision	recall	f1-score	support
0	0.94	0.72	0.82	1593
1	0.43	0.81	0.56	407
accuracy			0.74	2000
macro avg	0.68	0.77	0.69	2000
weighted avg	0.83	0.74	0.76	2000

---Random Forest Model---
Random Forest AUC = 0.77

	precision	recall	f1-score	support
0	0.92	0.80	0.86	1593
1	0.48	0.74	0.58	407
accuracy			0.79	2000
macro avg	0.70	0.77	0.72	2000
weighted avg	0.83	0.79	0.80	2000

MODEL COMPARISON

TYPE	NAME	DESCRIPTION	ADVANTAGES	DISADVANTAGES
Linear	Linear regression	**The "best fit" line** through all data points. Predictions are numerical.	Easy to understand -- you clearly see what the biggest drivers of the model are.	✗ Sometimes too simple to capture complex relationships between variables. ✗ Tendency for the model to "overfit".
Linear	Logistic regression	The adaptation of **linear regression** to problems of classification (e.g., yes/no questions, groups, etc.)	Also easy to understand.	✗ Sometimes too simple to capture complex relationships between variables. ✗ Tendency for the model to "overfit".
Tree-based	Decision tree	A graph that uses a **branching method** to match all possible outcomes of a decision.	Easy to understand and implement.	✗ Not often used on its own for prediction because it's also often too simple and not powerful enough for complex data.
Tree-based	Random Forest	Takes the average of many decision trees, each of which is made with a sample of the data. Each tree is weaker than a full decision tree, but **by combining them we get better overall performance.**	A sort of "wisdom of the crowd". Tends to result in very high quality models. Fast to train.	✗ Can be slow to output predictions relative to other algorithms. ✗ Not easy to understand predictions.
Tree-based	Gradient Boosting	Uses even weaker decision trees, that are increasingly **focused on "hard" examples.**	High-performing.	✗ A small change in the feature set or training set can create radical changes in the model. ✗ Not easy to understand predictions.



The ROC graph allows you to classify your errors for your true labels and false labels.

SUMMARY

- Germany had the most customers who exited followed by France and Spain
- There were a higher number of female customers who exited. France had the highest number of female customers who exited and Germany had the highest number of male customers who exited the bank.
- France had the highest female and male customers
- There is no significant difference in the distribution of both gender's credit scores
- Customers who exited had a roughly lower average credit score.
- Customers who had a credit card exited more than customers who did not possess a credit card.
- The majority of the people who exited were between 40 and 60 years old

THE ANALYSIS

Descriptive Analytics - What's Happening? :

Generally, bank customers who are 40-60 years old, have a lower credit score and have a credit card are more prone to leaving the bank.

Diagnostic Analytics – Why is it happening? :

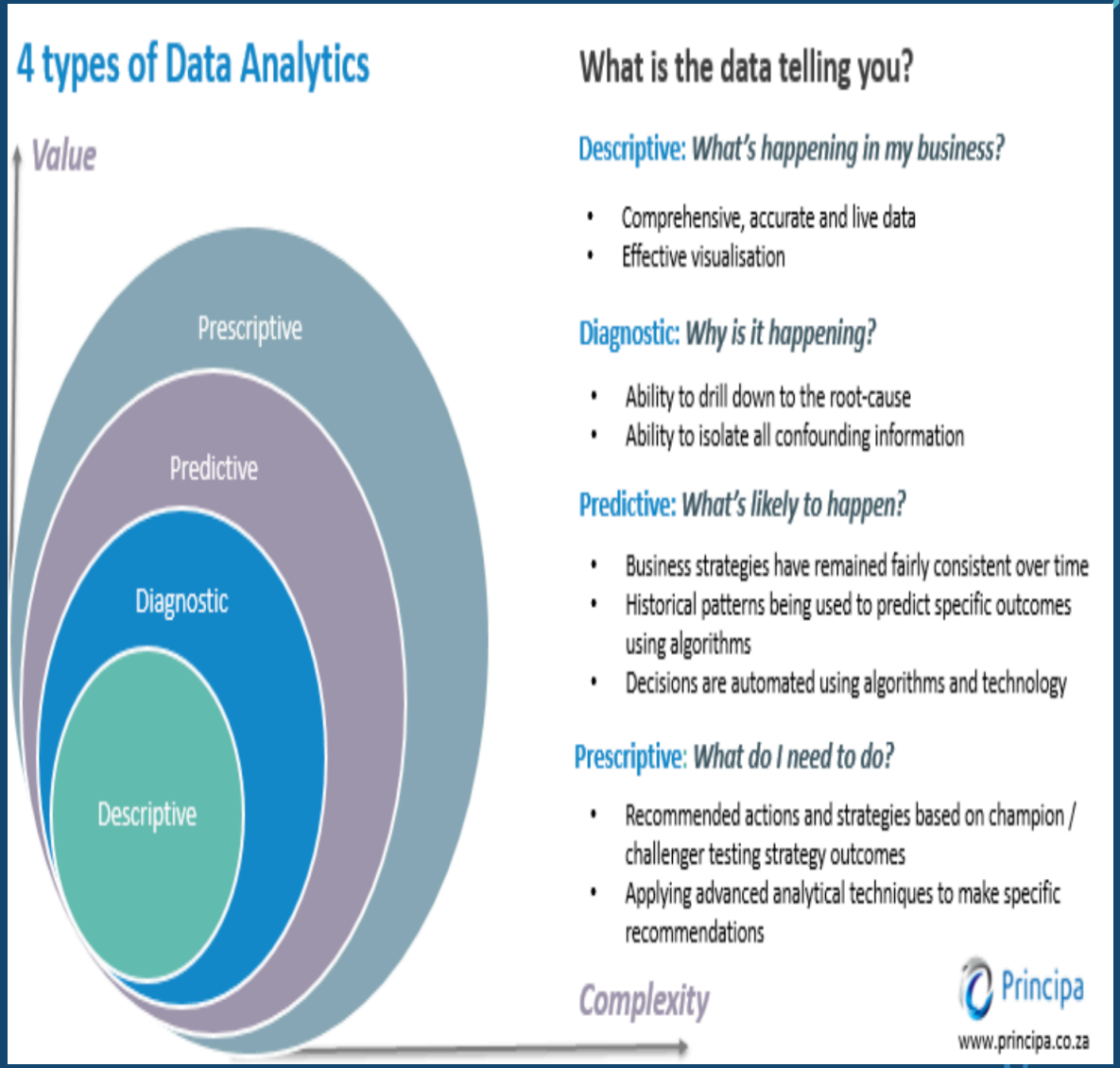
We will need to collect more information and ask further questions. From the data and our decision tree model, we know that customers age is a major contributor in them leaving the bank.

Predictive Analytics – What's likely to happen? :

Using our Logistic Regression or Decision Tree model, not only can we predict if a customer will exit the bank or not, we can also get the probability of them leaving.

Prescriptive Analytics – What do I do? :

We can use our probability scores of customers exiting to devise strategies of customer retention plans i.e. providing incentives to customers at risk of exiting the bank.



PROBLEM STATEMENT & RECAP

- **Problem:** The bank is losing customers and is interested in gaining insights about why they are leaving and if they can predict who is at risk of leaving so they can offer them incentives to stay. Ultimately, this reduces to a **binary classification problem**.
- **Need for Application:** We need to prevent customers from leaving the bank.

It is important that rather than just predicting if the bank customer will exit the bank or not, we would rather have an estimate of the probability that the customer will exit the bank or not. We could rank the customers by their probability of leaving, and allocate an incentive budget to the highest probabilities in hopes of retaining them.

Two problems with this approach:

1. Predicting a customer will leave but does not. This is called a false positive and can be expensive, inconvenient and time consuming for all parties.
2. No incentive offers provided but the customer leaves. This is called a false negative and we would like to prevent these as it results in lost revenue to the bank. Exactly, what we are trying to prevent!

POTENTIAL SOLUTIONS:

Solution 1:

- We can rank the customers by their probability of leaving the bank and set aside a budget to incentivize those customers not to leave the bank.
- Flipping this approach around, we can rank the customers by highest expected loss to the bank if they lose their business and reach out to those customers.

Solution 2:

- Provide training to the bank staff to improve their customer services. Track and measure their performances using analytics.

QUESTIONS TO ASK GOING FORWARD

- Why are more female customers leaving the bank?
- Are customers with credit cards moving their business to other banks with better rates?
- Why are customers in the 40 years - 60 years age range exiting the bank?
- Does the bank have a bigger market in France than Spain and Germany based on their customers?
- Is it possible to gather data on metrics like customer satisfaction score that can provide us with more insights?

POTENTIAL ADDITIONS

The best way to improve customer retention is to train the bank staff in customer relation best practices and investing in their skills. Some suggestions that can be explored are:

- Empowering employees
- Allow consumers to self-service
- Stay consistent across all touch points
- Educate customers on financial literacy
- Embrace financial technology
- Become an advisor, not just a lender, for small businesses
- Segment your client base and create personalized customer experiences
- Keep iterating on processes

LOOKING TO THE FUTURE

This problem was about equipping the bank with actionable knowledge regarding their customers. When modeling the data, we should not use the predictive metric as our final solution. Instead, we should use the information we get from modeling and arm the bank staff so they can carry out informed decision making.

Another thing the bank could do is start collecting more data on more features for e.g. how long the customer has been with the bank, satisfaction score with the bank etc. These things might help us improve our model, especially collecting more data as this bank customer dataset was relatively small. Once we have more data, we can go back and improve our predictions as well as gain further insights to see if anything has changed now that we have more customer information.

After our attempts to understand why customers are leaving the bank, we can flip the problem around and ask ourselves:

- What features contribute most to customers retaining their services with the bank?
- What features cause employees not to quit?
- What is the most valued thing about the bank by the customers?