Where Are Our Customers Going?

by Corbin Couger

7/14/2023

Introduction:

(Milestone 2)

The problem all started when a bank, Banksy Bank, noticed numerous customers exiting the business for unknown reasons. Maybe they were going to another bank, maybe Banksy Bank wasn't offering/providing what their customers wanted. This problem was something that needed to be solved efficiently and effectively to keep the bank in business. This is where I come in, Banksy Bank needs a Data Scientist to explore their customer database to discover why customers are leaving or staying and build a predictive model to solve this problem for future customers. To me, this is an important and interesting business problem for multiple reasons… the bank needs to be able to predict why customers are exiting and have some sort of idea for churn. There is also a 'scalability' to doing this analysis, as it can apply to different banks and banking corporations bringing answers to many customer retention questions as well as bringing me a product, money, and reputation.

With these things in mind, this problem would be interesting to several parties. The banks and/or banking corporations, myself, and even the customers of the banks. The banks would want to make their customers stick around, so this would lead to making their product more in-favor of the customer. I would want to create this as a product to sell to these numerous banks, after I do analysis on the Banksy Bank's dataset. Which speaking of, I attained this dataset from Kaggle (https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn) and contains 18 different columns with 10,000 customer records to help me get a solid predictive model. The data is described as follows:

- RowNumber – this is a unique number that identifies each row (not useful for the analysis)
- CustomerId – also a unique identifier for each customer in the dataset (wont be useful either)
- Surname – this is the surname of the customer (wont be useful)
- CreditScore – the customer's credit score
- Geography – the customer's location
- Gender – the customer's specified gender
- Age – the customer's age
- Tenure – number of years the customer has been at the bank
- Balance – customer's account balance
- NumOfProducts – number of products the customer has purchased through the bank
- HasCrCard – whether or not the customer has a credit card (0 or 1)
- IsActiveMember – is customer active or not (0 or 1)
- EstimatedSalary – customer's estimated salary
- Exited – whether or not the customer has left the bank (0 or 1, this will be the dependent variable in my model… basically what I'm trying to predict)
- Complain – has the customer complained or not (0 or 1)
- Satisfaction Score – score for how satisfied the customer is with the bank
- Card Type – level of card held by customer
- Points Earned – points earned on customer's credit card

These features, along with the 10k customers in this dataset will make it useful for the problem I'm going to attempt to solve. I'm excited to see what features (independent variables) have the largest impact on predicting my target variable (dependent variable, 'Exited').

For this project, I've decided to move forward, most likely, with a Logistic Regression model. This classification algorithm works well with predictive modeling for bivariate dependent variables. Since the 'Exited' variable is bivariate (0's or 1's) and is what I'm trying to predict, I thought Logistic Regression would be best fit. I will however explore other possible model options to explain some thought process using those. Another model I will be building is a Decision Tree model. This makes a lot of sense for this project because this algorithm works well with bivariate target variables and goes hand in hand with whether a customer is deciding to leave the bank.

I will evaluate each model similarly, for Logistic Regression I will be looking at how well my model fits the actual outcome of the data. I can do this a few ways and will be looking at statistics like; precision (how close the predictions are), recall (a percent of how much data was captured by the model), the accuracy of the model, and an ROC curve to visualize how well the model captured. Similarly for the Decision Tree model, I will be looking at the precision, recall, f1-score, and a visual of the tree itself. These different statistics will help me determine if I can predict a churn rate for the Banksy Bank and then build a fundamental structure for scaling these models to other banks.

By doing EDA, building models, interpreting results, and compiling all of my work into a report, it will help me achieve what I hope to learn. I first want to get more experience with building these models in Python, and secondly, I want to answer this business question and find a way to make this scalable. This is where I will run into ethical concerns… bank information is universally a sensitive topic. People's bank balances, incomes, credit score are all personal information that would need to be anonymous in any type of analysis. If I were to scale this project to other banks, I would need to keep in mind that this is personal information and anonymize any risk that could come about. I would simply just request necessary data from clients (banks) to ensure I never handle this personal data attached to any identifiable datapoints.

I'm excited to get the ball rolling on this project and get to the next step, preliminary analysis. This project seems promising and will hopefully work out, and because I plan on having 2 separate models, there is a good contingency plan there if one of the models does not work out. In the following steps I will be performing; Preliminary Analysis, Finalizing Results, and Creating a report an presentation of my project and findings.

Preliminary Analysis:

(Milestone 3)

To begin my preliminary analysis, I want to take a closer look at each variable in my dataset. I want to discover their datatype, see if I need to make any changes there (i.e. create dummy variables). I will check for missing data, outliers, remove unnecessary variables, etc. and clean up and faults in the dataset. I'm pretty clear and certain that this dataset will be strong with the model choices I have chosen, but will get into more detail about that further on in this section. Let me first start off with some exploration and analysis of my variables.

I first want to remove the unnecessary variables. As I mentioned in my introduction, there are variables that wont be necessary to include in the analysis. These include:

- RowNumber
- CustomerId
- Surname

These are either pointless or are variables that link the data to specific people which is something I want to avoid and anonymize. And, after removing these variables I want to check the data types in my dataset.

```
CreditScore              int64
Geography               object
Gender                  object
Age                      int64
Tenure                   int64
Balance                float64
NumOfProducts            int64
HasCrCard                int64
IsActiveMember           int64
EstimatedSalary        float64
Exited                   int64
Complain                 int64
Satisfaction Score       int64
Card Type               object
Point Earned             int64
```

As I can see, most of my variables are integer or float objects which is what I want for Logistic Regression and Decision Tree modeling. However, there are 3 object variables (Geography, Gender, and Card Type). Let me take a close look at each of these variables.

```
array(['France', 'Spain', 'Germany'], dtype=object)
```

The 'Geography' variable only contains 3 unique values. Now 'Gender' only has 2 types, male or female, but I'll still take a look at it.

```
array(['Female', 'Male'], dtype=object)
```

Finally, I'm going to look at the unique values in the 'Card Type' column.

```
array(['DIAMOND', 'GOLD', 'SILVER', 'PLATINUM'], dtype=object)
```

Since these object variables do not have many unique values, I will just turn them into dummy variables to make them integers.

Now I have data that is almost ready for modeling. I next want to check for any missing values (NAs).

```
customerdummies.shape
```

```
(10000, 21)
```

```
customerdummies.dropna().shape
```

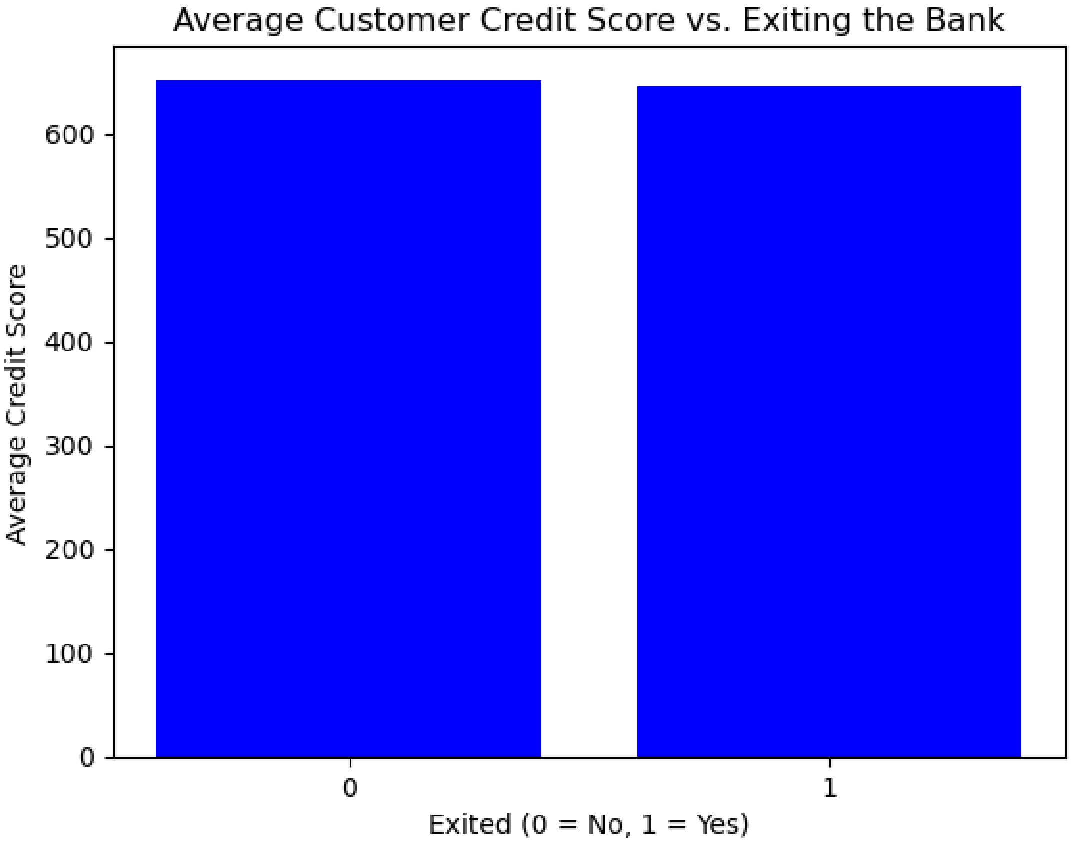
```
(10000, 21)
```

So, looking at the shape of the original dataset and the dataset with the NA's gone, they are the same. This tells me there is no missing data within the dataset.

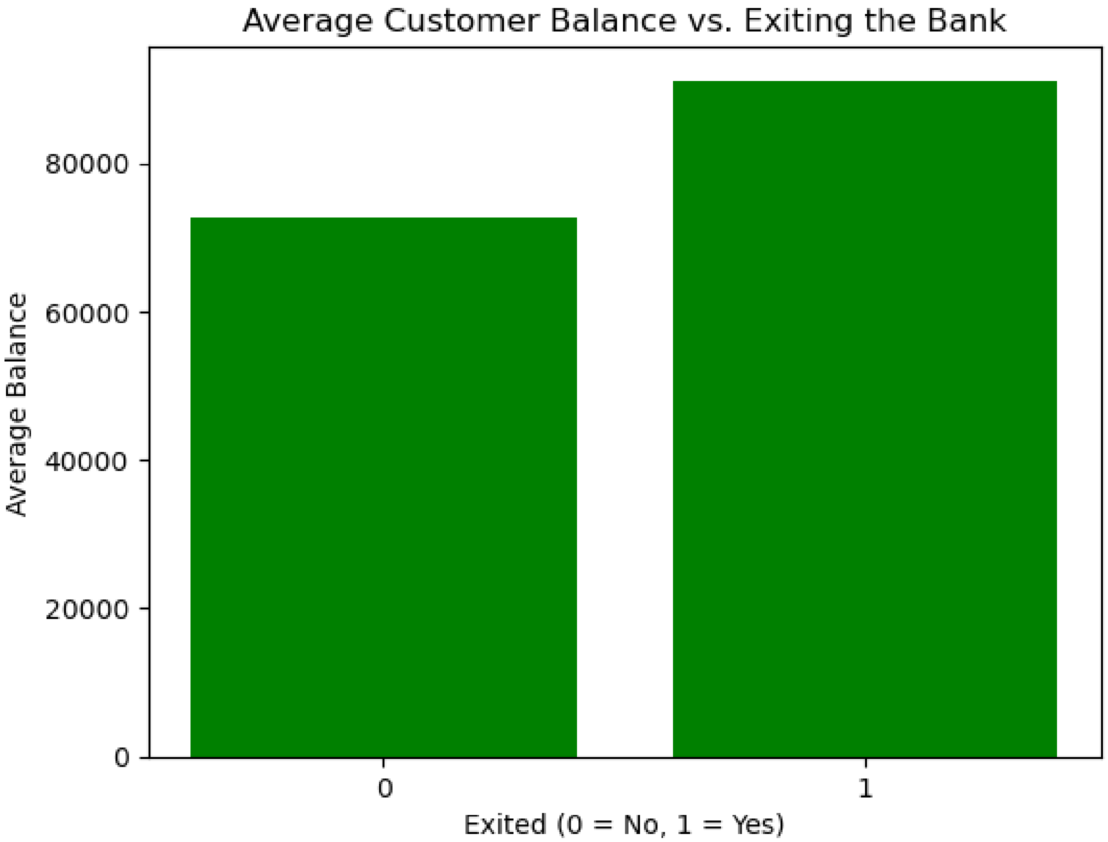
Visualizations:

Before I get to outlier analysis, my next step in preliminary analysis is to take a look a the variables from a visual point of view. To explore each variable, I am deciding to use the non-dummy dataset since those variables are together. This is the only time I will be using this dataset as it is what will be easiest to make visuals, but not the easiest to build a model with.

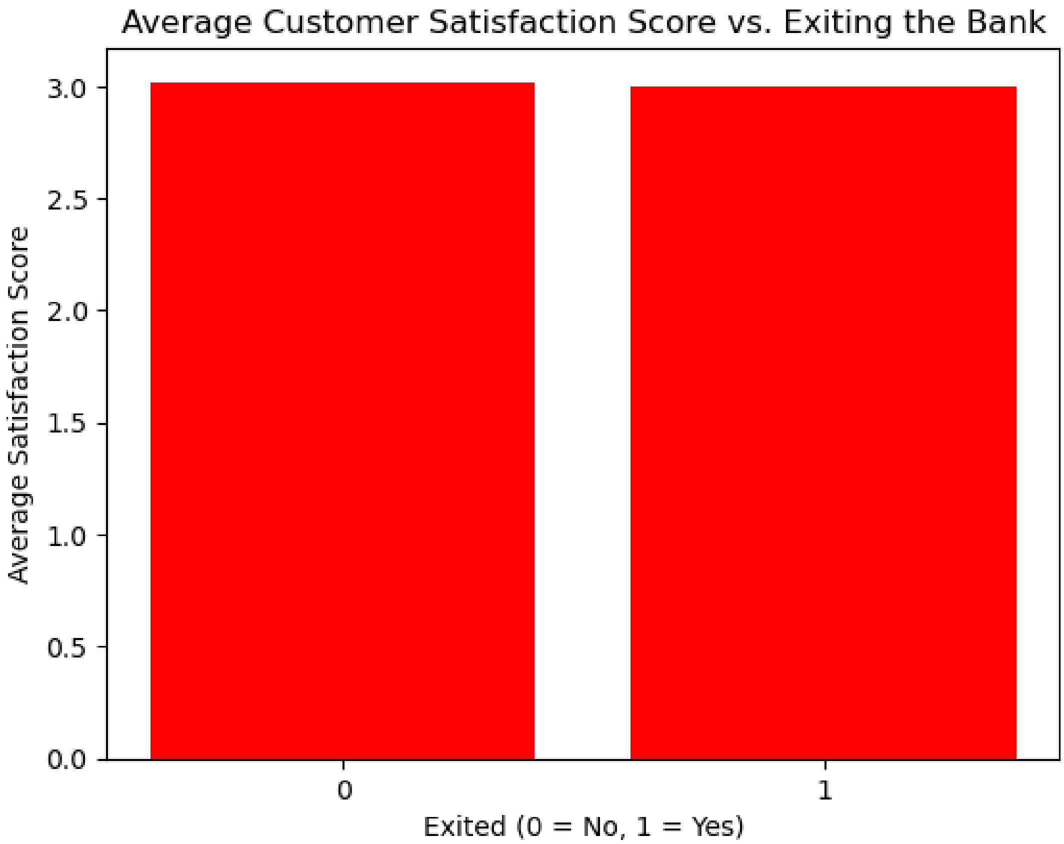
So, without further or do, let's visualize some of this data.



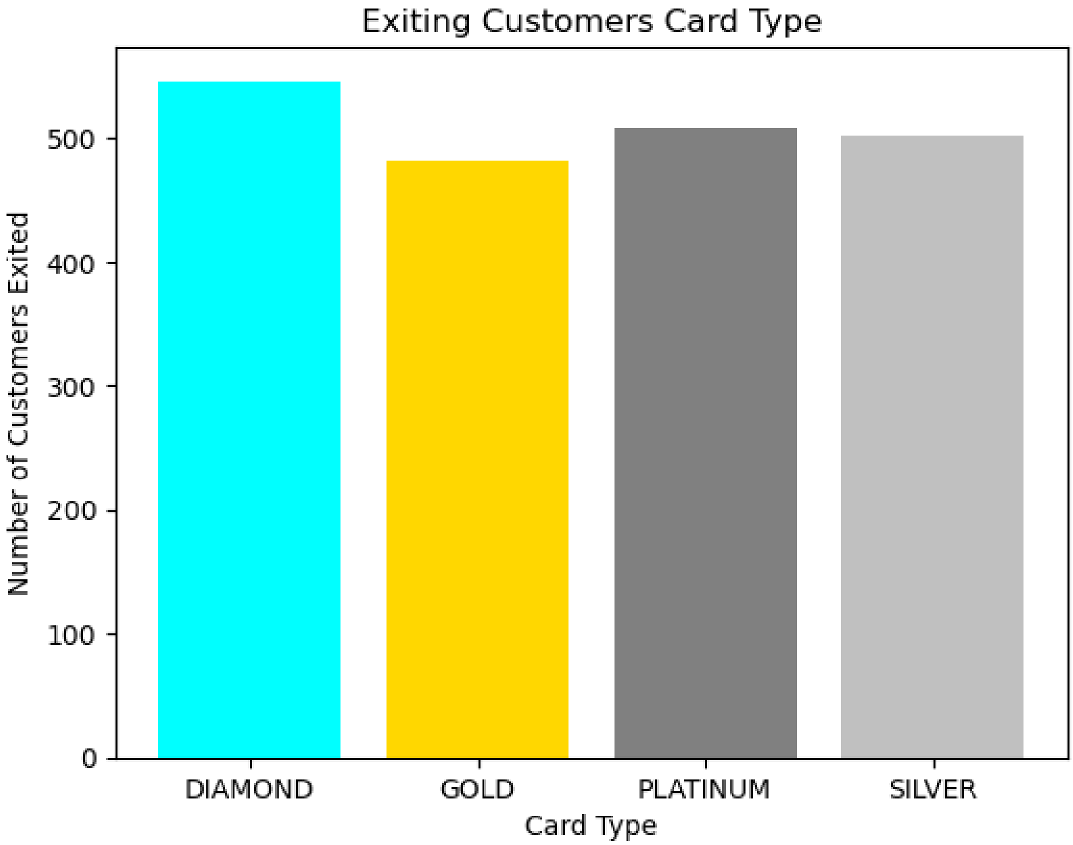
This visual doesn't provide much... the average credit score for those who stayed with the bank and those who exited is almost the same. I don't imagine this variable having much weight in my model.



This bar chart is interesting as it seems those who left the bank, on average, had a higher balance by about $20k. This will be interesting to see how this plays into the models.



This is also something interesting, both sides of this chart are about the same average satisfaction score. I thought this would be a main reason to leave a bank, whether or not your satisfied or not.. but hopefully the model gives us clear answers as to what can predict churn.



Looking at this bar chart, there isn't a clear story between the different card types. It does seem that the Diamond card holders leave more than the other card holders. Hopefully this shows well in the model.

Outlier Analysis:

Continuing on in my preliminary analysis, I want to check for outliers in any of my variables. To do so, I will create a boxplot for all the variables in the dummy dataset.

After looking at the boxplots, I can see outliers in some of the variables, but none that are concerning. I know age would have some outliers as well as credit score. However, none of these concern me so I will not be removing them for my analysis.

Correlation:

Lastly, I just want to check the correlation between my variables. This will help me determine any possible relationships between variables.

```
0    Gender_Female        Gender_Male          -1.000000
1    Geography_Germany    Geography_France     -0.580359
2    Geography_Spain      Geography_France     -0.575418
3    Card Type_DIAMOND    Card Type_GOLD       -0.334134
4    Card Type_DIAMOND    Card Type_SILVER     -0.333599
```

```
5        Card Type_DIAMOND Card Type_PLATINUM-0.333510
6        Card Type_GOLD       Card Type_SILVER     -0.333155
7        Card Type_GOLD       Card Type_PLATINUM-0.333066
8        Card Type_PLATINUMCard Type_SILVER     -0.332534
9        Geography_Germany  Geography_Spain       -0.332084
```

Looking at the top 10 correlated variables, I'm confident that I dont need to get rid of any other column.

Preliminary Analysis Conclusion:

Wrapping up this preliminary analysis, I am confident in this dataset to help me predict churn rate for Banksy Bank. I feel that my original expectations for this project are going to be consistent, but my one not to add (out of curiousity) is how well this data will be able to predict churn. Visually, the relationships didn't seem to obvious, but I do think there will be variables that aid in predicting the exit of a customer. After making my adjustments in this section, I now can move onto the modeling phase where I will first describe the models, create train and test datasets, and then model and analyze the data.

Model Building & Evaluation: Finalizing My Results

(Milestone 4)

Now that I have explored and cleaned my data, I can now start to prepare and build a Logistic Regression model and a Decision Tree model. Like I mentioned in the intro, I will analyze the models a number of ways. I want to check my predictions against the test dataset by calculating precision, recall, and create an ROC curve for each model.

Logistic Regression:

```
x = customerdummies.loc[:, customerdummies.columns != 'Exited']

y = customerdummies.loc[:, customerdummies.columns == 'Exited']
```

My final step in preparing this data for the model is splitting the dataset into test and train datasets.

```
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.25, random_state=16)

logreg = LogisticRegression(random_state=16)

model = logreg.fit(X_train, y_train.values.ravel())

y_pred = logreg.predict(X_test)
```

I have now fitted my train dataset to a Logistic Regression model and have gotten my predictions. I will now interpret these results and calculate my accuracy statistics.

Logistic Regression Evaluation:

lr_matrix = metrics.confusion_matrix(y_test, y_pred)

lr_matrix

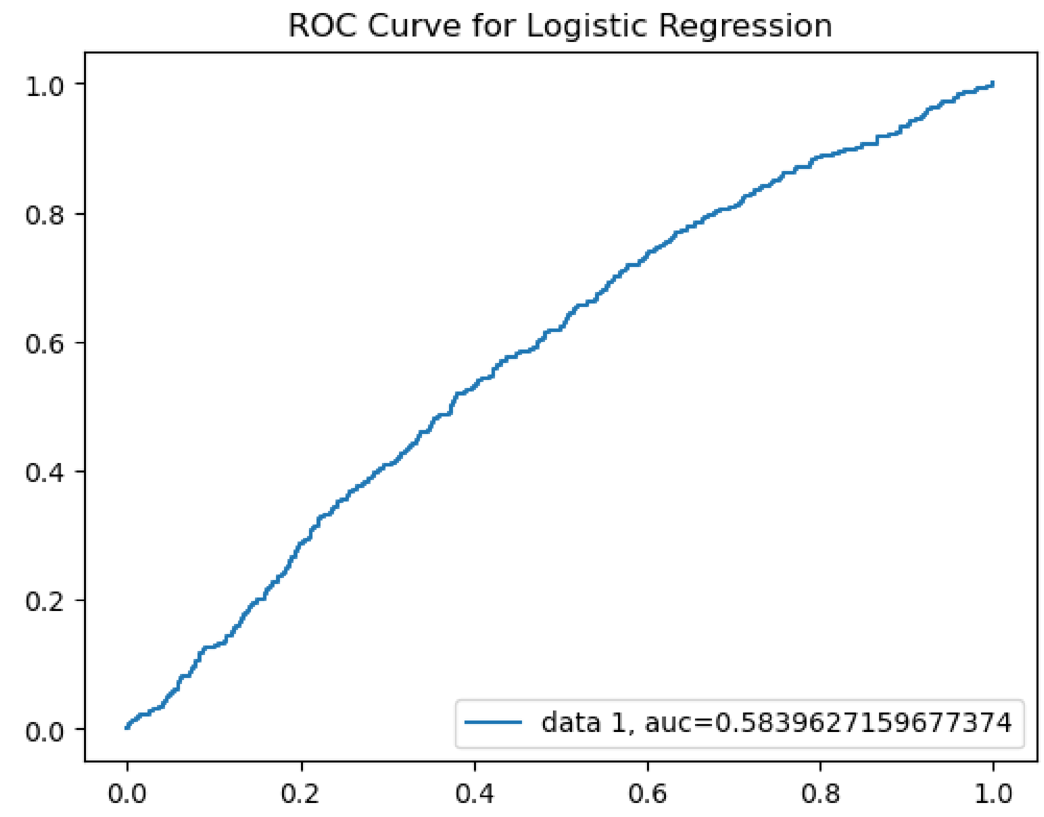
```
array([[1983,    0],
       [ 517,    0]], dtype=int64)
```

To calculate a quick simple accuracy of this model, I can take the confusion matrix and take the actual predictions and divide them by the false predictions.

Logistic Regression Model is: 79.32 % accurate.

Starting off pretty strong with almost 80% accuracy based on the confusion matrix! I hope the other metrics follow the same pattern and validate this model as a good predictor of customer exiting.

```
              precision    recall  f1-score   support

      Exited       0.79      1.00      0.88      1983
  Not Exited       1.00      0.00      0.00       517

    accuracy                           0.79      2500
   macro avg       0.90      0.50      0.44      2500
weighted avg       0.84      0.79      0.70      2500
```



Looking at these metrics, I can see that the Logistic Regression model I made is a good predictor of whether or not a customer will exit the bank or not based on the different variables in the dataset. Looking at the precision for 'Exited', I can see that the model had an 80% success rate at predicting the positive predictions. The recall is even better which is interpreted as being 100% accurate at classifying the positive cases of if a customer exited the bank. The one thing that concerns me is the ROC curve. The other metrics came out to be fairly strong for making this model a good classifier. However, the 'auc' = .584 on the ROC curve is not a good sign and would mean there should be some, but not a lot, of doubt for this classification.

But, the Logistic Regression model turned out great and I'm happy to say that Banksy Bank now has a good, efficient, and accurate way to predict whether or not a customer will exit the business. To even validate and further this experiment, I will now create a different classification model, the Decision Tree.

Decision Tree:

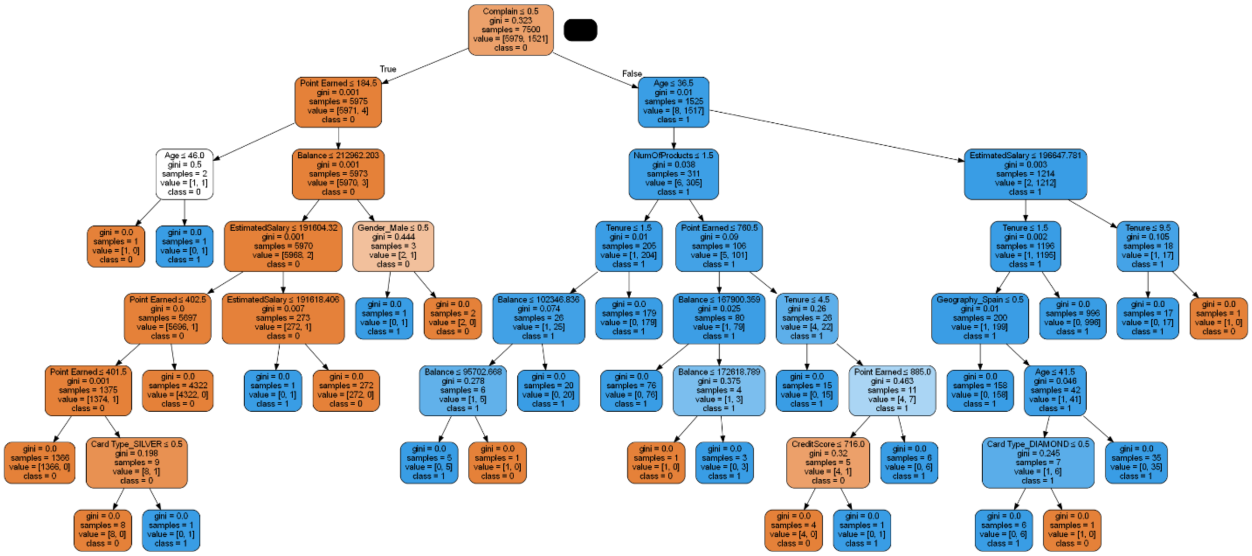
```
clf = DecisionTreeClassifier()

clf = clf.fit(X_train,y_train)

y_pred = clf.predict(X_test)
```

I've created a decision tree model with only a few lines of code, and I will now check its accuracy. Since this model relies on accuracy and its visualization, it does not require a ton of other metrics like my Logistic Regression model did.

Decision Tree Evaluation:

```
Decision Tree Accuracy: 0.9984
```

Wow, this is awesome! 99.84% accuracy for the Decision Tree model. This means that the model can predict a customer's exit nearly 100% of the time. Banksy Bank, based on this quick knowledge so far, I would consider the use of a Decision Tree model for your customer churn. Let me take a closer look at the tree itself.


Complain ≤ 0.5 gini = 0.323 samples = 7500 value = [5979, 1521] class = 0
[redacted]
True
False
Point Earned ≤ 184.5 gini = 0.001 samples = 5975 value = [5971, 4] class = 0
Age ≤ 36.5 gini = 0.01 samples = 1525 value = [8, 1517] class = 1
Age ≤ 46.0 gini = 0.5 samples = 2 value = [1, 1] class = 0
Balance ≤ 212962.203 gini = 0.001 samples = 5973 value = [5970, 3] class = 0
NumOfProducts ≤ 1.5 gini = 0.038 samples = 311 value = [6, 305] class = 1
EstimatedSalary ≤ 196647.781 gini = 0.003 samples = 1214 value = [2, 1212] class = 1
gini = 0.0 samples = 1 value = [1, 0] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1
EstimatedSalary ≤ 191604.32 gini = 0.001 samples = 5970 value = [5968, 2] class = 0
Gender_Male ≤ 0.5 gini = 0.444 samples = 3 value = [2, 1] class = 0
Tenure ≤ 1.5 gini = 0.01 samples = 205 value = [1, 204] class = 1
Point Earned ≤ 760.5 gini = 0.09 samples = 106 value = [5, 101] class = 1
Tenure ≤ 1.5 gini = 0.002 samples = 1196 value = [1, 1195] class = 1
Tenure ≤ 9.5 gini = 0.105 samples = 18 value = [1, 17] class = 1
Point Earned ≤ 402.5 gini = 0.0 samples = 5697 value = [5696, 1] class = 0
EstimatedSalary ≤ 191618.406 gini = 0.007 samples = 273 value = [272, 1] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1
gini = 0.0 samples = 2 value = [2, 0] class = 0
Balance ≤ 102346.836 gini = 0.074 samples = 26 value = [1, 25] class = 1
gini = 0.0 samples = 179 value = [0, 179] class = 1
Balance ≤ 167900.359 gini = 0.025 samples = 80 value = [1, 79] class = 1
Tenure ≤ 4.5 gini = 0.26 samples = 26 value = [4, 22] class = 1
Geography_Spain ≤ 0.5 gini = 0.01 samples = 200 value = [1, 199] class = 1
gini = 0.0 samples = 996 value = [0, 996] class = 1
gini = 0.0 samples = 17 value = [0, 17] class = 1
gini = 0.0 samples = 1 value = [1, 0] class = 0
Point Earned ≤ 401.5 gini = 0.001 samples = 1375 value = [1374, 1] class = 0
gini = 0.0 samples = 4322 value = [4322, 0] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1
gini = 0.0 samples = 272 value = [272, 0] class = 0
Balance ≤ 95702.668 gini = 0.278 samples = 6 value = [1, 5] class = 1
gini = 0.0 samples = 20 value = [0, 20] class = 1
gini = 0.0 samples = 76 value = [0, 76] class = 1
Balance ≤ 172618.789 gini = 0.375 samples = 4 value = [1, 3] class = 1
gini = 0.0 samples = 15 value = [0, 15] class = 1
Point Earned ≤ 885.0 gini = 0.463 samples = 11 value = [4, 7] class = 1
gini = 0.0 samples = 158 value = [0, 158] class = 1
Age ≤ 41.5 gini = 0.046 samples = 42 value = [1, 41] class = 1
gini = 0.0 samples = 1366 value = [1366, 0] class = 0
Card Type_SILVER ≤ 0.5 gini = 0.198 samples = 9 value = [8, 1] class = 0
gini = 0.0 samples = 5 value = [0, 5] class = 1
gini = 0.0 samples = 1 value = [1, 0] class = 0
gini = 0.0 samples = 1 value = [1, 0] class = 0
gini = 0.0 samples = 3 value = [0, 3] class = 1
CreditScore ≤ 716.0 gini = 0.32 samples = 5 value = [4, 1] class = 0
gini = 0.0 samples = 6 value = [0, 6] class = 1
Card Type_DIAMOND ≤ 0.5 gini = 0.245 samples = 7 value = [1, 6] class = 1
gini = 0.0 samples = 35 value = [0, 35] class = 1
gini = 0.0 samples = 8 value = [8, 0] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1
gini = 0.0 samples = 4 value = [4, 0] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1
gini = 0.0 samples = 6 value = [0, 6] class = 1
gini = 0.0 samples = 1 value = [1, 0] class = 0

This Decision Tree visualization helps me and can help Banksy Bank understand what the model is actually doing. The different characteristics of the data are broken down and then you can filter through to see whether or not a customer will exit the bank based on those characteristics.

Based on this analysis, the Decision Tree model is what I would recommend to Banksy Bank. The data drives this model to produce accurate predictions and classifications on customer characteristics, leading to their retention. The implementation of this model would be smooth and quick. I would dive into this visualization and create a series of flags or identifiers on if certain customers consist of exiting customer characteristics. Banksy Bank can then take that information and brainstorm ways to keep those customers around. Maybe by helping the customers save money and build a higher balance, or help them earn points quicker, change their card level, etc. This model proves to be powerful as a strong classifier and I would consider this to be better than the Logistic Regression model I made.

Conclusion

Overall, this data driven analysis will help Banksy Bank implement an accurate way of knowing the exiting customer behavior. It also helped me get more experience and understanding of the inner working of these models in Python. The company and I learned a lot about churn rate without compromising anonymity. My goals for this project were met and I know this is going to be a satisfying answer for Banksy Bank. Using this analysis, I am going to compile a higher level presentation that will describe my findings, results, and recommendations on the data they provided and for the model I built. This process is something I will scale and integrate with other banks and companies reliant on customer retention. The possibilities are endless and I'm excited to take these models elsewhere to discover more answers to more questions.