Slideshow

*Slides 1 + 2*: Introduce the group

*Slide 3*: Introduction to what data we are looking for

*Slide 4*: During the presentation we will be looking at the following: the questions we were asking of the dataset and why? Where we found the dataset and how we used it to answer our questions. The clean up and data analysis process and finally the conclusions that we’ve drawn from our data and the implications they may have.

*Slide 5:*

*Slide 6 - 11*: Firstly, we wanted to look at if we can compare age and gender to the attrition status of customers, to identify any possible trends. To do this we first selected which columns we would need from the data set and created a separate data frame to allow for easier analysis.

Initially we made pie charts to allow for easier visualization of the percentage of genders that make up the dataset, and of the percentages of attrition flag, which shows us whether a customer is still with the bank or has left. [Show pie charts]

Following this, the .loc method was used to create two data frames that are locked onto Female gender and Male gender, which were then used to get value counts of the attrition flag. This allowed us to create a stacked column chart, showing the numbers of male and female customers that are still with and have left the bank. [Show “customer status by gender” plot]. This allows for easier comparison between the genders.

To look at the customer ages, we first had to create bins to put them in. This allowed for grouping of the ages into a new column, which made the analysis easier. A pie chart was then created to visualize how the age groups make up the data. [Show pie chart]

Using the new data frame with the age group column, the .loc method was used again, this time to create two data frames that are locked onto “Attrited Customer” and “Existing Customer”. These are then used to get the value counts of the age groups, which then had to be sorted using sort\_index, as value counts automatically puts it into descending order.

This was then used to create a stacked column chart showing the number of customers within each age group that are existing customers and have left the bank. [Show “Customer status by Age” plot]

We then created a data frame just to show the mode, mean, median, standard deviation and variance of the age to allow us to compare.

So what did we actually learn from this question? In regards to identifying whether a customer will churn or not, not a lot. The information from this data doesn't really identify an age group or gender that is more likely to churn. However, it does identify the most populous age group (40-49), which may allow the bank to target customers within that age group to ensure they stay with the bank.

*Slides 14 – 17:*

For the third question, we were looking to see if there was any way to predict which existing customers will leave using how long they have been inactive for and how long they have been a customer of the bank. Firstly, we established which columns were needed and created a smaller data frame with the relevant columns to work with.

We then put the months inactive into a pie chart, again for easy visualization. The first figure shows us the percentage of customers and the months they have been inactive for. We can see from this pie chart that 3 months makes up the largest percentage of customers. This is then further confirmed in figures 2 and 3.

When using .describe() on the data frame, we can see that the mean for months inactive is 2.34, the standard deviation is 1.01 and the variance is 1.02. These values indicate that the data for months inactive are closer to the mean and are not spread out over a wide range.

We then create a new data frame, using .loc to lock the Attrition Flag onto Attrited Customer, which then allows us to use a value count on the months inactive of attrited customers. The same was then repeated with existing customers.

These two plots are then combined, as can be seen in figure 4. This plot shows a sharp drop off after 3 months for both existing and attrited customers. This may indicate that 3 months without activity is perhaps a trigger point at which the bank may contact the customer, which will result in either the customer starting to use their card again, or cancelling the account completely.

We then created bins to put the months on book data into, as this data was quite varied and grouping would allow for easier analysis. This data frame with the binned groups was then used to create two new data frames with the attrition flag locked onto attrited and existing customer respectively. These two data frames were then used to get value counts of the grouped months on book column, before using sort index to put them back into order.

This allowed us to produce a stacked bar plot showing the number of months a customer has been with the bank for both existing and attrited customers. This plot shows us that for both existing and attrited customers, the highest number of customers were in the 30-39 months category. This data could support one of two ideas, the first being that the bank does not have the high churn rate that it anticipates. The second idea is that the credit card may have a certain length of time it has to be had for. 3 years (36 months) is within the 30-39 months category, and this can be supported by the fact that the mean for months on the books is 35.9. It may be that for customers to benefit from the rewards scheme, that they must sign up to have the credit card for 3 years.

To reiterate, this data indicates that 3 months inactive seems to be a trigger point. The number of customers that are inactive following 3 months drops rapidly. If this is because the bank sends notifications out at 3 months of inactivity, perhaps the notification could be changed to include a further bonus or reward, to incentivize the customer to use their credit card again.

It may also be worth the bank looking at customers who are approaching the 3 year mark with the bank and trying to ensure that they remain customers.

*Conclusion/Summary*

Although the aim was to identify whether customers will churn or not, we believe with the data we are unable to clearly identify whether they will. We have been able to identify relationships between certain data, and predict perhaps which groups should be targeted for advertising or further incentives to stay with the bank.

In order to be able to predict whether a customer will churn, we would need further information in regard to whether they have had multiple accounts. Possibly look at the same customer in regard to other banks to see if they have had accounts with them as well.

*If we had more time?*

If we had more time, we could have looked at other relationships between the data given. We also would have liked to add an API of average incomes in the United States (from the census data) to create a map of which states the bank would be most likely to succeed in based on the average income.