Bellarmine University

Churn EDA analysis

Ashley Ridley

aridley@bellarmine.edu

2/2/2024

Introduction:

Previously, I was going to use predictive analysis to make a prediction model regarding company bankruptcy that would accurately predict whether a company can go bankrupt or not. Unfortunately, the data from the model was very imbalanced with only 3% of the instances being bankruptcies. This led to conducting further research to find a better dataset and stumbling upon this churn dataset, which can be found on the Kaggle website. This dataset is a collection of customer data that focuses on the tendency of customers to stop using a company's products or services. The dataset will be used to analyze and understand factors that contribute to customer churn and to build a predictive model to identify customers at risk of churning. I chose this dataset because it met all the requirements for this project, while also being a very interesting topic.

Data set description:

The dataset contains various features that describe each customer, such as their credit score, country, gender, age, tenure, balance, number of products, credit card status, active membership, estimated salary, and churn statues. There are 12 columns and 10,000 data values, which will help to make a more accurate prediction model. Below, is a table with name, data type, range of values and percentage of data missing to help us understand more about the data. Looking at the table below, this dataset contains both categorical and numerical variables. Categorical variables are a type of variable that represent categories or groups, in this case country and gender. While numerical variables are quantitative or continuous variables that represent measurable quantities, such as age, tenure, and credit score. These variables can help to determine which data type each data value belongs to depending on the characteristics and properties.

Name:	Data type:	Range of values:	% of missing data
Customer_id	nominal	1.55e+07-1.581e+07	0
Credit_score	ratio	350-850	0
country	nominal	n/a	0
gender	nominal	n/a	0

age	Ratio	18-92	0
tenure	ratio	0-10	0
balance	ratio	0-250,898.09	0
Products_number	ratio	1-4	0
Credit_card	nominal	0-1	0
Active_member	nominal	0-1	0
Estimated_salary	ratio	11.580-199,992.48	0
churn	Nominal	0-1	0

Looking at the data type column, there are 2 different data types, which consists of nominal, and ratio. Nominal data such as churn, country, and gender, is data that consist of categories without any order or ranking. This type of data does not have any meaningful numeric values. On the other hand, ratio data such as age, tenure, and balance, have a specific order with meaningful intervals and a true zero point. Lastly, we will look at the rage of values since there are no missing values. It is important to look at the range of values because it provides information about the spread and data distribution. Unusually large or small values can significantly impact the range from the rest of the dataset. These outliers can contribute to an increased range. In the table above, the range of values was taken from the maximum and minimum values from each column of the dataset. Country and Gender do not have any values because they are categorical, while churn, credit card, and active member have ranges from 0-1 to because they are represented by binary numbers, 0 for no and 1 for yes.

Data set summary and statistics:

Listed below, are some tables that helped to further investigate the data using exploration techniques.

Numerical summarization:

Listed below Is the numerical summarization of the data. This table provides a concise summary of the numerical data aspects of the dataset, which involves various statistical measures to describe the central tendency, dispersion, and shape of the data. This can help to identify trends and patterns within the dataset and make data-driven decisions.

	customer_id	credit_score	age	tenure	balance	products_number	credit_card	active_member	estimated_salary	churn
count	1.000000e+04	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000	10000.000000
mean	1.569094e+07	650.528800	38.921800	5.012800	76485.889288	1.530200	0.70550	0.515100	100090.239881	0.203700
std	7.193619e+04	96.653299	10.487806	2.892174	62397.405202	0.581654	0.45584	0.499797	57510.492818	0.402769
min	1.556570e+07	350.000000	18.000000	0.000000	0.000000	1.000000	0.00000	0.000000	11.580000	0.000000
25%	1.562853e+07	584.000000	32.000000	3.000000	0.000000	1.000000	0.00000	0.000000	51002.110000	0.000000
50%	1.569074e+07	652.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	1.000000	100193.915000	0.000000
75%	1.575323e+07	718.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	1.000000	149388.247500	0.000000
max	1.581569e+07	850.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	1.000000	199992.480000	1.000000

Correlation table:

This is a correlation matrix that displays the correlation coefficients between pairs of variables in the dataset. Each cell in the table represents the correlation between two variables, which indicates the strength and direction of their linear relationship ranging from -1 to 1. 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates there is no linear relationship. This table helps to gain insights into the relationships between variables.

```
Correlation Matrix:
                 customer_id  credit_score       age    tenure   balance
customer_id         1.000000      0.005308  0.009497 -0.014883 -0.012419
credit_score        0.005308      1.000000 -0.003965  0.000842  0.006268
age                 0.009497     -0.003965  1.000000 -0.009997  0.028308
tenure             -0.014883      0.000842 -0.009997  1.000000 -0.012254
balance            -0.012419      0.006268  0.028308 -0.012254  1.000000
products_number     0.016972      0.012238 -0.030680  0.013444 -0.304180
credit_card        -0.014025     -0.005458 -0.011721  0.022583 -0.014858
active_member       0.001665      0.025651  0.085472 -0.028362 -0.010084
estimated_salary    0.015271     -0.001384 -0.007201  0.007784  0.012797
churn              -0.006248     -0.027094  0.285323 -0.014001  0.118533

                 products_number  credit_card  active_member  \
customer_id             0.016972    -0.014025       0.001665
credit_score            0.012238    -0.005458       0.025651
age                    -0.030680    -0.011721       0.085472
tenure                  0.013444     0.022583      -0.028362
balance                -0.304180    -0.014858      -0.010084
products_number         1.000000     0.003183       0.009612
credit_card             0.003183     1.000000      -0.011866
active_member           0.009612    -0.011866       1.000000
estimated_salary        0.014204    -0.009933      -0.011421
churn                  -0.047820    -0.007138      -0.156128

                 estimated_salary     churn
customer_id              0.015271 -0.006248
credit_score            -0.001384 -0.027094
age                     -0.007201  0.285323
tenure                   0.007784 -0.014001
balance                  0.012797  0.118533
products_number          0.014204 -0.047820
credit_card             -0.009933 -0.007138
active_member           -0.011421 -0.156128
estimated_salary         1.000000  0.012097
churn                    0.012097  1.000000
```

Missing values:

This table shows the missing values from the dataset for each column. This provides a convenient way to quickly assess the extent of the missing data in the dataset and decide how to handle them, such as dropping rows with missing values or imputing missing values. As we can see there is no missing data from the dataset.

```
df.isnull().sum()
```

```
credit_score        0
country             0
gender              0
age                 0
tenure              0
balance             0
products_number     0
credit_card         0
active_member       0
estimated_salary    0
churn               0
dtype: int64
```

Unique values

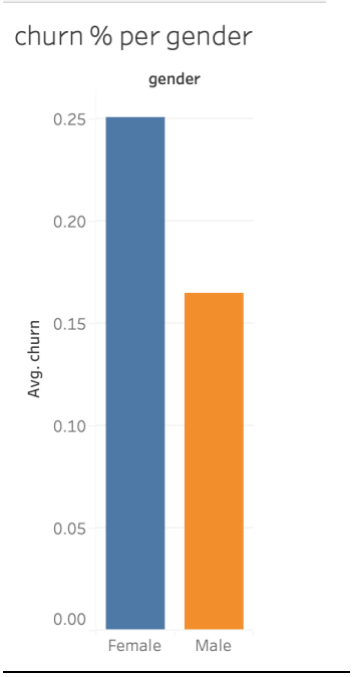
Lastly, this table shows the count of unique values of each column in the dataset. This is useful for understanding the diversity and distribution of values for assessing the uniqueness of categorical variables.

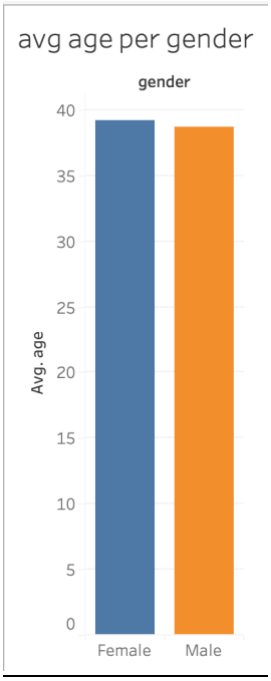
```
print(df.nunique())
```

```
credit_score         460
country                3
gender                 2
age                   70
tenure                11
balance             6382
products_number        4
credit_card            2
active_member          2
estimated_salary    9999
churn                  2
dtype: int64
```

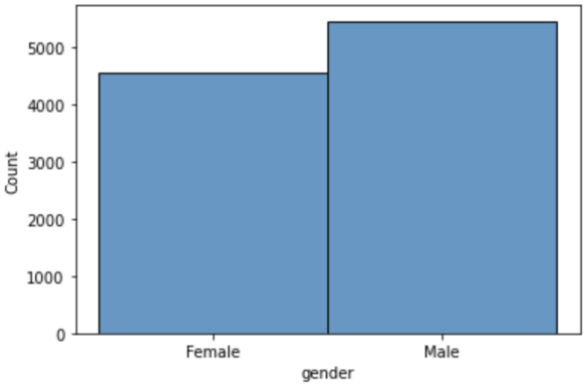
Data Set Graphical Exploration:

Next, we will look at some helpful graphs to further interpret the data.

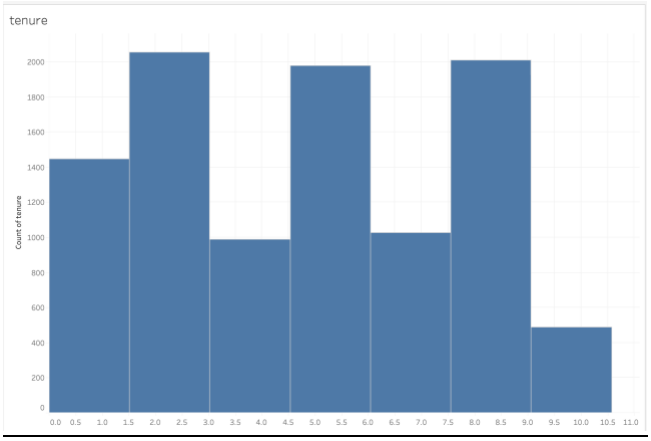




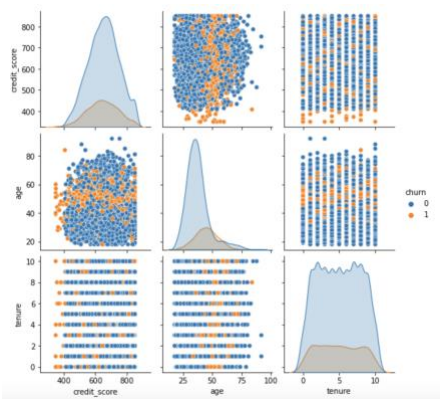
These two charts display the churn percentage per gender (left), and the average age per gender (right). The graph on the left shows that the female churn percentage is 25%, while the male is 16%. The graph on the right shows that the average age for females is 39.24, while for males its 38.66.

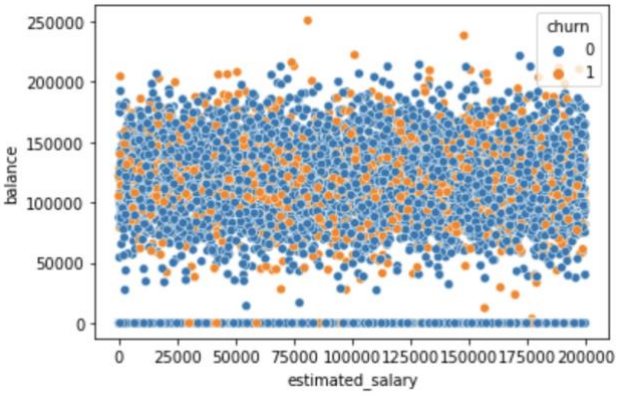


The chart above is a histogram for the female to male ratio. Here we can see there are more males than females. This can help to later determine if gender influences customer churn.

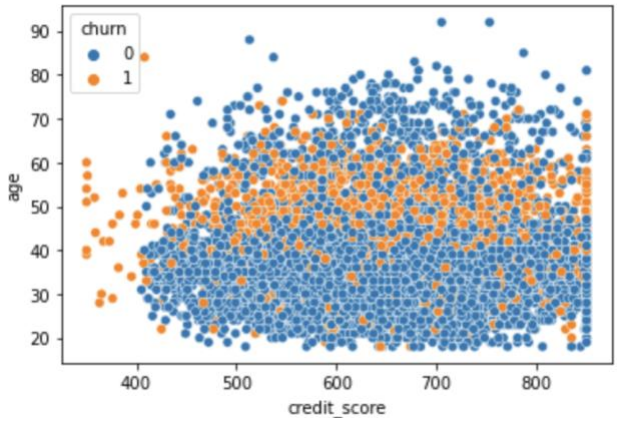


The graph above shows the tenure by bin by the count of tenure. For example, bin 2 has a tenure count of 2,057.





The graph on the left displays a pair plot that analyzes the relationship between the numerical features and churn status, while the graph on the right is a scatterplot of the estimated salary vs balance. Both graphs are valuable to understand the structure of the data, and the relationships between them.



The graph above is another scatterplot between credit score and age. This shows the correlation and outliers from the data. Since this graph is grouped by age, this can help to identify whether certain age groups tend to have higher or lower credit scores on average.

Lastly, these two graphs are heatmaps. The graph on the left displays a correlation between all the columns, while the one on the right displays the correlation between churn, age, gender, country, tenure, and estimated salary.

Summary of findings:

The dataset consists of 12 columns and a total of 10,000 observations. Since there are no missing values this indicates the dataset is complete. All the columns have appropriate data types, with numerical variables represented as integers or floats and categorical variables represented as strings or categories. Since the dataset is complete, looking at the categorical variables, gender and country will need to be imputed to binary numbers. To do this, I will replace female with 0 and male with 1. The customer ID column is not helpful for interpretation so I will drop this column. For the EDA analysis, the distribution of age is approximately normal, with a mean of 39 years. Credit scores range from 350 to 850, with a median of 650, and tenure has a median of 5 years. The churn status indicates that approximately 20% of customers have churned. As for the correlation, there is a moderate positive correlation between age and credit score, which suggest that older customers tend to have higher credit scores. There is no significant correlation between tenure and credit score, but there is a slight negative correlation between churn status and credit score which implies that customers with a lower credit score may be more likely to churn.