|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Variables | Description | Domain Type | Entries | Missing Values |
| 1 | RowNumber | The row numbers of each entry | Ratio | 10000 | 0 |
| 2 | CustomerId | An identifier for each customer | Ratio | 10000 | 0 |
| 3 | Surname | The surname of each customer | Nominal | 10000 | 0 |
| 4 | CreditScore | Measure of individual creditworthiness, derived from financial history | Ratio | 10000 | 0 |
| 5 | Geography | Location of customer | Nominal | 10000 | 0 |
| 6 | Gender | Sex of each customer (Male or Female) | Nominal | 10000 | 0 |
| 7 | Age | Each customer’s age | Ratio | 10000 | 0 |
| 8 | Tenure | Duration of customer with bank | Ratio | 10000 | 0 |
| 9 | Balance | Current account balance of customer | Ratio | 10000 | 0 |
| 10 | NumOfProducts | Customer purchases made with the bank | Ratio | 10000 | 0 |
| 11 | HasCrCard | Customer has a credit card in his name | Ratio | 10000 | 0 |
| 12 | IsActiveMember | Customer is an active member of the bank | Ratio | 10000 | 0 |
| 13 | EstimatedSalary | The customers estimated salary | Ratio | 10000 | 0 |
| 14 | Exited | Customer churns or not | Ratio | 10000 | 0 |
| 15 | Complain | Customer makes a complaint or not | Ratio | 10000 | 0 |
| 16 | Satisfaction Score | Satisfaction of customer with the bank’s services | Ratio | 10000 | 0 |
| 17 | Card Type | Type of credit card customer utilizes | Ordinal | 10000 | 0 |
| 18 | Points Earned | Points earned by the customer for using credit card | Ratio | 10000 | 0 |
|  | TOTAL |  |  |  | 0 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Check | Findings | Impact on Solutions |
| 1 | Accuracy and Precision | No issues detected | None |
| 2 | Correctness by Entry | a) We observe some strange values for people’s surnames. Values such as H?, L?, Hs?, Y?, Y?an, K?, and Hs?eh, do not represent real-world names.  b) Estimated salary reports some customers earning less than 10,000 to as low as 11.58. This is strange as many of these reported customers have a bank balance above 100,000. | a) These issues question how reliable the rows with these values are towards our model’s prediction.  b) These errors question how reliable our trained model will be. Setting a threshold below which we consider unreasonable could be one possible solution for this problem. |
| 3 | Completeness | No issues detected | None |
| 4 | Consistency (validity and integrity) | No issues detected | None |
| 5 | Data Source Reliability | While the dataset is gotten from Kaggle, the authenticity of the bank where this dataset is gotten from is not specified. | Difficulty relying on results and findings from dataset as a case for modelling a real-world phenomenon. |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Variables | Mean | Std | Min | Median | Max |
| 1 | CreditScore | 650.53 | 96.65 | 350 | 652 | 850 |
| 2 | Age | 35.92 | 10.49 | 18 | 37 | 92 |
| 3 | Tenure | 5.013 | 2.89 | 0 | 5 | 10 |
| 4 | Balance | 76485.89 | 62397.41 | 0 | 97198.54 | 250898.09 |
| 5 | NumOfProducts | 1.53 | 0.58 | 1 | 1 | 4 |
| 6 | HasCrCard | 0.71 | 0.46 | 0 | 1 | 1 |
| 7 | IsActiveMember | 0.52 | 0.5 | 0 | 1 | 1 |
| 8 | EstimatedSalary | 100090.24 | 57510.49 | 11.58 | 100193.92 | 199992.48 |
| 9 | Exited | 0.2 | 0.4 | 0 | 0 | 1 |
| 10 | Complain | 0.2 | 0.4 | 0 | 0 | 1 |
| 11 | Satisfaction Score | 3.014 | 1.41 | 1 | 3 | 5 |
| 12 | Points Earned | 606.515 | 225.93 | 119 | 605 | 1000 |

Figure 3.1: Card Type

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Diamond | Platinum | Gold | Silver |
| COUNT | 2507 | 2495 | 2502 | 2496 |

Figure 3.2: Geography

|  |  |  |  |
| --- | --- | --- | --- |
|  | France | Spain | Germany |
| COUNT | 5014 | 2477 | 2509 |

Figure 3.2: Gender

|  |  |  |
| --- | --- | --- |
|  | Female | Male |
| COUNT | 4543 | 5457 |

Figure 3.1: Tenure

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| COUNT | 413 | 1035 | 1048 | 1009 | 989 | 1012 | 967 | 1028 | 1025 | 984 | 490 |

Figure 3.2: Satisfaction Score

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 |
| COUNT | 1932 | 2014 | 2042 | 2008 | 2004 |

Figure 3.2: NumOfProducts

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 |
| COUNT | 5084 | 4590 | 266 | 60 |

Figure 3.2: IsActiveMember

|  |  |  |
| --- | --- | --- |
|  | 1 | 0 |
| COUNT | 5151 | 4849 |

Figure 3.1: HasCrCard

|  |  |  |
| --- | --- | --- |
|  | 1 | 0 |
| COUNT | 7055 | 2945 |

Figure 3.2: Exited

|  |  |  |
| --- | --- | --- |
|  | 1 | 0 |
| COUNT | 2038 | 7962 |

Figure 3.2: Complain

|  |  |  |
| --- | --- | --- |
|  | 1 | 0 |
| COUNT | 2044 | 7956 |

# Some potential things to consider

"""

1) What is the impact of our analysis from a specific geographical location?

i.e Does having an imbalanced geographical class impact our analysis

2) Can we create new features that capture the relationship between customers

and churn.

3) Create data binning to help with skewed distributions.

4) When binning the Card Type, it's worth considering if the column is truly Nominal

or should be represented as Ordinal data type. This should be considered, as within

the bank, there is some level of inherent hierarchy among the card type a user holds

and as such, our transformation of the card type column needs to reflect this.

"""

# EDA Findings

"""

After exploratory data analysis was conducted, we make the following findings:

- Without needing to preprocess the data, we highlight 3 features which won't be relevant

for our analysis and drop them. These features include - "RowNumber", "CustomerId", "Surname".

\* Row number here is just an identifier and will not play any role in contributing to the

understanding of why bank customers leave the bank.

\* CustomerID similar to the row number is an identifier and will not contribute to our models

predictive capacity.

\* Surname refers to the lastname of our customers. This doesn't add value to our prediction

as we cannot say in any literature that given a person's surname, we can tell if he or she

will leave the bank. The idea of churn has nothing to do with the customer’s surname. Hence,

we drop this column.

- For our analysis on bank customer churn, our dataset makes use of 5457 males and 4543 females.

- Our data has no duplicate values across each row

- No missing values were found

- The bank customers ages range from 18 to 92

- The data collected spans 3 geographical locations - France, Germany, and Spain.

With the majority of the customers in our analysis coming from France. France has

5014 customers analysed, Germany has 2509 customers, while Spain has 2477 customers.

- Customers within the bank have 4 different possible card types they can hold. These

include - Diamond cards, Gold cards, Silver cards, and Platinum cards.

- When looking at the value counts for our customers age, it is seen that the larger

majority of customers in our dataset are in their 30's.

- Interestingly, we see a 0.996 strong positive correlation between the complain column

and the exited column. This strong linear relationship between the complain feature and

our target exited, suggests that we can use this singular complain column to predict

customer churn in banks. The impact of this feature on our analysis needs to be explored

further.

- No strong linear correlation among other variables outside of the exited and complain

features.

- The Age feature has a left-skewed distribution as seen in our histogram. While

creditscore is slightly skewed to the right. We observe a uniform distribution among the

following features - EstimatedSalary, IsActiveMember, and SatisfactionScore. All other

features have an unbalanced class distribution.

"""

# Data Quality Checks

"""

For our data quality checks, we consider the following checks:

1) Accuracy and Precision Check:

e.g. currency exchange between £ and ¥: £1 ↔ ¥9.10 or £1 ↔ ¥9.056187

2) Correctness by Entry Check:

e.g. entered 9.056287 instead of 9.056187

3) Completeness Check:

e.g. no date of birth is given

4) Consistency (validity and integrity) Check:

e.g. book borrowing date: 02/02/2019, date of return: 03/01/2019

5) Redundancy (unnecessary redundancy) Check:

e.g. simply collecting together data on replication servers

6) Data Source Reliability Checks

"""

"""

HOW IS IT POSSIBLE THAT PEOPLE HAVE ESTIMATED SALARY LESS THAN BALANCE

"""

"""

Do column to column visualization of categorical variables

"""

"""

REMEMBEER YOU DID NORMAL CORRELATION BETWEEN VARIABLES. FOR BINARY TARGET, YOU NEED TO USE

POINT-BISERIAL CORRELATION

"""

# Drop the COMPLAIN COLUMN

"""

During our first iteration in our model building phase, the built model is able to predict accuratly

100% for test data across 5 different algorithms. While at first glance, it may seem

like the best way to go, however, when things are looking too good to be true, it is probably

too good to be true. This means when we have successfully created

a model that predicts the likelihood a bank customer with am accuracy of 100%, the results should be

considered too good to be true and due to chance as predicting accurately a 100% accuracy means

this model never fails.

The possible reason for this behaviour in our model could be the relationship between the complain featrure

and the target exited. After removing outliers using the mahalanobis multivariate outlier technique, which

was able to capture two groups of people between the complain feature and exited. These groups are:

- Those who made complaints but didn't churn

- Those who churned but didn't complain

Upon removing these rows highlighted by our mahalanobis algorithm, correlation between the complain

feature and exited is 1 indicating a perfect positive correlation between the variables. This relationship

therefore crates the illusion that complain and exited are the same. Theoritically, this is misleading while

trying to model the real-world as there exists no perfect correlation between these two variables. This proxy

variable complain is the feature our algorithm captures and uses as the most important feature to learn the

realationship between bank customers and why they churn.

The approach taken to mitigate this was to drop the complain column our dataset and build our predictive model

around the other features.

"""  
  
FOR FEATURE ENGINEERING, CONSIDER CREATING A NEW FEATURE THAT HELPS SOLVE THE ISSUE OF ZEROS. IT CAN BE A BINARY FEATURE THAT WAKENS WHEN THE ACCOUNT BALANCE IS ZERO AND SWITCHES OFF WHEN NOT.