ACKNOWLEDGMENTS

First, I would like to thank my advisor, Professor …………, for being a part of this project since its beginning, and supporting me in all the challenges that I have faced. As well, my co-advisor, Professor ……………, for helping me go further with his remarkable vision and insights. Not only by all the patience and support they have provided, but also, by all the knowledge they have passed during the lectures in the Master’s degree.

Moreover, I want to express my gratitude to my supervisor …………….., for trusting me with this project and for all the support, knowledge, and opportunities he has given me for these months. Lastly, I would like to thank my friends and family for believing in me and being a part of my academic journey and life.

## INTRODUCTION

Customer Churn (Customer attrition) is the most challenging problem for businesses such as credit cards or telecommunication companies. It would be efficient to build models that can predict if a customer would churn or continue their service. This is the case that we would like to look and understand the reason for the churn of customers using credit card services. The phenomenon of customer churn is not limited to just banking and financial industries. It is very much prevalent in other service industries too, such as mobile telecommunications, and television viewership. The goal of this research is to build up a predictive model for companies that can recognize clients who are probably going to churn from the services provided by the companies.

Credit card churn prediction can assist a bank to know which of its customers is going to retain and can also assist the bank to know on which customer it needs to emphasize more so that it can earn the highest gains. Future churn of customers of the bank can also be estimated by evaluating recent industry patterns and customer behavior, as well as potential innovations. The total transaction amount, number of transactions, and total years of relationship of the customer with the bank matters a lot in retaining the customer. The functionality is based on a banking website that recognizes customer requirements and then incorporates the use of data mining's methods and machine learning. This study aims to use classification analysis using machine learning to predict Customer Churn based.

**Research Questions to be answered:**

* What are the main reasons for a customer to churn from the usage of a credit card?
* Can the important features be obtained from the bank and customer information stored which drives the churn?
* What are the things that can be done to retain customers to a bank?
* Can promotional or other offers be provided to a set of customers who are on verge of being churn?

**Why are these questions important?**

* Banks generate a lot of revenue from customers using credit cards. It is evident that 30% of total revenue comes from the usage of cards so it is important for the banks to maintain a healthy relationship with the customers who use credit cards.
* Also customers tend to churn from the credit card usage when there is a decrease in the offers that they receive from their cards so it is important to decide over the required promotional offers that can be given to customers.
* Also the offers over credit cards seem to decrease the rate of churn and increase its usage over various online shopping or purchases so this can also be a trigger for banks to maintain good relationships with ecommerce companies.

## Literature Review

* 1. LITERATURE REVIEW IN CHURN MODELING IN BANKING

Günther et al. (2014) defined customer churn in business finance as *“When a customer cancels all his/her bank account or credit card, either to switch competitive account provider or because the need of bank account is no longer present*”. However, there is a differentiation in the types of customers’ churns, and not all are suitable for analysis. A churner can be classified as voluntary when the customer deliberately decides to switch to another service provider or involuntary (usually not included for modeling) if circumstances like death or bankruptcy occur (X. Zhao et al., 2009).

Nowadays, customers can easily compare other credit card providers in terms of offers and interest rates just by using the internet. As a result, this transparency leads many customers to voluntarily switch to other providers and cancel the existing product they have in the company. This transparency has a significant impact on the financial market's competitiveness. Moreover, customer defections can have more to do with a service company’s profits than many factors associated with a competitive advantage, given that profit can be boosted by almost 100% by retaining just 5% more of their customers (Sasser & Reichheld, 1990). Furthermore, attracting new customers can be twelve times more costly than retaining the company’s current ones (Price, 2002).

Given this, especially now, to enhance competitiveness, companies should focus on reducing customer churn. Although Sasser & Reichheld (1990) are the grounders of the “zero defections” theory, they defend that companies should not try to eliminate all defections but be prepared to spot customers who leave and act accordingly to their findings. Hence, different industries such as telecom, game, finance, and insurance use churn analysis to detect early churn signals and identify customers who are highly likely to leave voluntarily (Vafeiadis et al., 2015).

Although the goal among the industries is the same, the preferred models are different due to the differences in types and cycles of the log data. For the insurance sector case, the log data is relatively small. There are no significant changes in the customer’s information, so many statistical approaches using traditional machine learning models or survival analysis are seen (Ahn et al., 2020).

Given this, a quick overview of the most commonly used techniques in the insurance sector was done. The most recent literature on the topic (last five years (2015-2020)) was selected for this analysis. Based on the results, the most used algorithms present in literature for insurance churn modeling are Logistic Regression, Decision Tree, Support Vector Machine, and Neural Networks. However, Gradient Boosting Model has shown the ability to attain good results, being one of the optimal models in Y. He et al. (2020).

| S.No | Model | Used in Literature |
| --- | --- | --- |
| 1 | Logistic Regression | Y. He et al. (2020) Vafeiadis et al. (2015)  Sundarkumar & Ravi (2015) Bolancé et al. (2016)  Spiteri & Azzopardi (2018) |
| 2 | Decision Tree | Vafeiadis et al. (2015) Sundarkumar & Ravi, 2015) Bolancé et al.(2016) Dolatabadi et al. (2017) Spiteri & Azzopardi (2018)  Scriney et al. (2020) |
| 3 | Support Vector Machine | Y. He et al., (2020) Vafeiadis et al. (2015)  Sundarkumar & Ravi (2015) Bolancé et al. (2016) Dolatabadi et al. (2017) Spiteri & Azzopardi (2018) Scriney et al. (2020) |
| 4 | Neural Network | Y. He et al. (2020) Vafeiadis et al. (2015)  Sundarkumar & Ravi (2015) Bolancé et al. (2016) Dolatabadi et al. (2017) Scriney et al. (2020) |
| 5 | Gradient Boosting | Y. He et al. (2020) |
| 6 | Naïve Bayes Classifier | Vafeiadis et al. (2015) Dolatabadi et al. (2017) Spiteri & Azzopardi (2018) Scriney et al. (2020) |
| 7 | Random Forest | Y. He et al. (2020) Spiteri & Azzopardi (2018) |
| 8 | Extra Trees Classifier | Y. He et al. (2020) |

Table 2.1 Models used in literature for insurance churn modeling

* 1. MACHINE LEARNING

#### Unsupervised Learning

Unsupervised learning is a form of machine learning that aims to group data into segments based on similar attributes, or naturally occurring trends, patterns, and relationships hidden in the data (McCue, 2015). Standard algorithms used for this kind of task are clustering, anomaly detection, neural networks, and approaches for learning latent variable models (El Bouchefry & De Souza, 2020).

The main goal of clustering analysis is to segment the finite unlabeled dataset into a finite and discrete set of hidden data structures (Xu & Wunsch, 2005). The result of this analysis is several groups of the data named clusters. These are subsets of data grouped when, according to the studied criteria, have similar characteristics and, when not, separated into different groups (Rokach & Maimon, 2005).

Moreover, two types of clustering techniques are partitional clustering and hierarchical clustering. In the first, groups are created by partitioning the space into a pre-defined number of subspaces. On the other hand, in hierarchical clustering, the data objects are grouped in sequence by a hierarchical structure.

#### Supervised Learning

Another form of Machine Learning is supervised learning. Contrary to unsupervised learning, it uses a dataset that has already been classified (labeled) as a basis for predicting the classification of other unlabeled data (Talabis et al., 2015). The labeled dataset is a training set composed of input variables (features) and an output variable (label). The used features will influence the model’s ability to correctly classify the predicted variable, the output variable (L. Wang et al., 2021).

Supervised Learning can be divided into classification when the label is categorical and regression when the label is continuous. Therefore, the studied problem is a classification task since churn modeling can be translated to a binary classification problem: assuming 1 when the customer churns and 0 when the customer does not churn.

### Data Imbalance

For many real supervised learning problems involving a binary response variable, datasets present a skewed distribution, having one class with a much lower representation than another. When the dataset shows this type of behavior, having an underrepresented class, the data is said to be unbalanced (S. Wang & X. Yao, 2012). Researchers have concluded that this imbalance causes a suboptimal classification performance (Chawla et al., 2004), since classifiers tend to give much higher importance to the larger classes.

H. He & Garcia (2009) bring that. Usually, classifiers, when dealing with an imbalanced dataset, tend to “*provide a severely imbalanced degree of accuracy, with the majority class having close to 100 percent accuracy and the minority class having accuracies of 0-10 percent*”, such consequence can represent high costs to some industries so, therefore, is vital to construct a model that “*will provide high accuracy for the minority class without severely jeopardizing the accuracy of the majority class*”.

In order to deal with imbalanced datasets, three possible approaches can be taken: data level, algorithmic level, and combining or ensemble methods, for which the first involves resampling to reduce the class skewness (Yap et al., 2014). Resampling is done either by removing instances from the

majority class (undersampling) or adding instances to the minority class (oversampling) by using algorithms such as SMOTE.

#### Synthetic Minority Oversampling Technique

Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) is one of the most popular and influential data pre-processing algorithms to deal with the data imbalance problem (García et al., 2016).

This technique is an oversampling approach, said that new instances from the smaller class are introduced into the dataset. Unlike basic approaches such as random oversampling (ROS), which only duplicates samples from the minority class, SMOTE generates new synthetic samples, overcoming the overfitting caused by approaches like ROS (Fernández et al., 2018).

The first step of this technique is to define the amount of oversampling. Here, it is possible to either set up this value to approximate a balanced class distribution or discover it via a wrapper process (Chawla et al., 2008). Then, based on k nearest neighbors and linear interpolation ideas, the synthetic samples are created. SMOTE operates in the feature space rather than in the data space, each minority class sample is considered along with its k nearest neighbors, and the new samples are introduced along the line segments joining them (considering any/all of the k neighbors) (Chawla et al., 2002).

### Model Calibration

When using such techniques of artificially rebalancing the dataset or even by consequence of the dataset’s characteristics, the training and test sets have different distributions. This difference in training and test sets distribution violates the basic assumption in machine learning that both are drawn from the same underlying distribution (Pozzolo et al., 2015). By violating this assumption, the predictions obtained in the test set will be biased and, therefore, enhance the need for probability calibration to obtain unbiased predictions.

Furthermore, some methods tend to bias predicted probabilities by pushing away or closer to 0 and 1, enhancing the need for calibration (Niculescu-Mizil & Caruana, 2005). Dormann (2020) even states that “*it should be applied to any model type as part of the prediction process, before predicting, cross- validating and making effect plots and maps or using predictions in any other probabilistic interpretation*”.

Two model calibration methods that can be used to correct these biased probabilities are Platt Scaling and Isotonic Regression. The first is more effective when the distortion is sigmoid-shaped, and the latter is a more robust method that can correct any monotonic distortion but, more prone to overfitting (Niculescu-Mizil & Caruana, 2005).

### Threshold-Moving Method

A technique that should be considered when dealing with class imbalance is changing the decision threshold (model’s continuous output cut-off) and adapting it to a performance metric. The main difference between rebalancing (using techniques like SMOTE) and threshold-based methods is that the latter relies on manipulating the continuous output of a learned model instead of relying on data pre-processing before the learning happens (Collell et al., 2018).

Provost (2008) even states that “*The bottom line is that when studying problems with imbalanced data, using the classifiers produced by standard machine learning algorithms without adjusting the output threshold may well be a critical mistake*”.

The threshold moving method uses the original training set to train and tunes or shifts the decision threshold by adapting it to a performance metric. One possible approach is to use the ROC evaluation procedure and move from where misclassifications attain their maximum on the positive class to the point where the maximum in the negative class is attained, selecting the point where the curve attains its maximum (H. He & Garcia, 2009).

### Feature Selection

A well-known problem is the “curse of finite sample size”, for which the relationship among the number of samples available and the features considered for modeling needs to be considered (Jain & Chandrasekaran, 1982). Each new feature introduced to the model will represent a new dimension. The higher the dimensionality, the sparser the dataset becomes and, thus, lower the feature space coverage (Verleysen & François, 2005). Consequently, the problem’s complexity rapidly grows with the introduction of more dimensions. Bellman (1966) introduced the term Curse of Dimensionality to explain such phenomena.

Therefore, and given the nowadays existing high-dimensional data, feature selection is one of the essential techniques in data preprocessing by eliminating irrelevant, redundant, or noisy features (Kalousis et al., 2007). Performing such techniques allows faster algorithms and, besides improving predictive power, also improves comprehensibility (Kumar & Minz, 2014).

It is possible to broadly classify feature selection methods into filter and wrapper methods, where the first ranks features based on statistical measures, independently of the learning algorithm (Kohavi & John, 1997). One example of this type of method is to use as importance measure (score) the variable’s correlation with the target. On the other hand, wrappers evaluate each candidate subset of features' impact on a particular learning algorithm. The latter approach usually allows achieving better results given their close interaction with the classifier (El Aboudi & Benhlima, 2016).

Some examples of wrapper methods are forward selection, backward elimination, and recursive feature elimination. The first keeps adding new features which improve the model performance until no other feature respects the criteria. The second works similarly but oppositely, starts with all features, and removes the least significant effect that does not meet the model’s staying criteria until all features are significant. In both (forward and backward elimination), the feature stays in the model once added/removed. Lastly, recursive feature elimination is similar to the forward selection method. In this case, effects are added and removed into the model such that one or more backward elimination steps can happen after a forward selection step (Bursac et al., 2008).

Another family derived from the two previous method families (filter and wrappers) are embedded methods. These methods combine the classifier development with the search of the optimal subset of features, capturing dependencies at a lower computational cost than wrappers but (like wrappers) also have a risk of overfitting (Seijo-Pardo et al., 2017). Examples of embedded methods are Lasso and Ridge regression, which have built-in feature selection methods that employ L1 and L2 regularization (respectively).

### Ensemble Learning for Feature Selection

Ensemble Learning is a type of learning where multiple models are trained and combined to solve the same problem (Polikar, 2006). Ensemble Learning is based on the assumption that combining the solution of multiple experts is better than using the solution of a single one. In this way, a set of hypotheses is constructed rather than using only one single hypothesis to explain the data, making it possible to reduce bias and variance from the learning algorithms (Dietterich, 2002).

Although usually employed to improve classification results, it is also possible to use ensemble learning as a feature selection technique. Combining multiple feature selection methods (instead of relying on

just one) makes it possible to attain more robust feature subsets, showing a great promise to high- dimensional datasets with small sample sizes (Saeys et al., 2008).

The approach benefits from diversity and control of variance and is possible to employ in two ways: one is to use the same algorithm to retrieve the features’ importance using different subsets of the data (data perturbation / homogeneous) and, other, is to use different feature selection techniques in the same dataset (function perturbation / heterogenous) (Chiew et al., 2019).

Said this, different levels can be varied and shall be chosen when employing ensemble learning in feature selection, following Bolón-Canedo & Alonso-Betanzos (2019) can be defined as follows:

* + - * + Dataset Level: use different subsets of data (or not)
        + Feature Level: use different subsets of features (or not)
        + Learner Method Level: use or design different learning algorithms (or not)
        + Combination Level: Use or design different combination/aggregation methods
        + Threshold Level: use of different thresholding methods (in case of using ranker methods)

Bolón-Canedo & Alonso-Betanzos (2019) also verified that heterogeneous feature selection ensembles are more commonly used than homogeneous ones. However, it is possible to obtain either a feature subset or a feature ranking in both cases, depending on the type of feature selectors. For the latter, a threshold method needs to be defined.

When using feature rankers, several feature ranking algorithms of the ensemble are combined, creating a final ranked list of the features, given the features’ relevance to prediction. The approach for combining the ensemble members' results (ranks) has different proposals in the literature, from simpler to more complex solutions (Seijo-Pardo et al., 2015). Some of the most popular straightforward methods to combine such ranks attributed to each feature are: minimum (best) rank, median rank, arithmetic mean rank, and geometric mean (Bolón-Canedo & Alonso-Betanzos, 2019).

Given the threshold method decision, the most common approach is defining a fixed percentage of the top features, but this percentage depends on the used dataset (Bolón-Canedo & Alonso-Betanzos, 2019). Therefore, this technique is not optimal since it is prone to overstating or understating this cut- off value (Chiew et al., 2019).

#### Permutation feature importance

Permutation feature importance (PFI) is a model-agnostic feature selection method. Therefore, it is possible to perform a heterogeneous approach by using this algorithm as a base learner for the ensemble and introduce variability by using different base models in the algorithm.

This permutation feature importance measurement was firstly introduced for random forests in 2001 (Breiman, 2001) but can be used in any model. PFI assesses the variable’s importance for the given model when its relationship with the target is broken by observing the decrease in the model’s score when random noise replaces a variable (introduced by randomly shuffling the variable’s values) (McGovern et al., 2019). Therefore, it is possible to understand how important a given feature is to the model’s ability to predict the target correctly.

PFI’s operation method makes it sensible to correlated features (Strobl et al., 2008) and is, therefore, essential to address this issue primarily. However, PFI has advantages such as robustness not to bias the measures favoring high cardinally features over binary features.

## METHODOLOGY

* 1. Research Method   
     The CRISP-DM (cross-industry process for data mining) methodology was applied in the execution of this project. This methodology suggests that the life-cycle of a data mining project can be broken into six phases that can be further segmented into several tasks (Wirth & Hipp, 2000). Said that, at the top-level, it is possible to organize a data-mining project into the following six primary phases:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

Diagram

Description automatically generated

Figure 1.1 Phases of CRISP-DM (Wirth & Hipp, 2000)

It is not precisely necessary to perform the suggested tasks entirely in the proposed sequence. As the methodology suggests ([Figure](#_heading=h.1t3h5sf) [1.1](#_heading=h.1t3h5sf)), moving back and forward is usually required. It is possible to find more details about the methodology in Wirth & Hipp (2000).

Therefore, these main phases are covered in this report and further segmented into the methodology's different suggested tasks. This report is organized as follows. Section [2](#_heading=h.gjdgxs) provides the theoretical framework - first, a customer churn and a modeling literature review; then, a theoretical review of machine learning theory and some of the used techniques. Section [3](#_heading=h.2grqrue) provides an overview of the collected data for modeling the portfolio’s customer churn. Next, in Section [4](#_heading=h.tyjcwt), the applied methodology and all its phases and tasks are explained. Section [5](#_heading=h.vx1227) presents the churn prediction model results and its discussion. Finally, Section [6](#_heading=h.3fwokq0) contains the conclusions from this project. As for used software, Kaggle was used to collect the data, and Python was used in the remaining tasks.

* 1. Tools/Technologies used  
     There are several tools or technologies used in this project as shown below:

| **Technology** | **Version** | **Use** |
| --- | --- | --- |
| Python | python3.7 | * Whole project is done using python language |
| Anaconda Environment | Anaconda 2022.10 | * Whole project is done in jupyter notebook which is from Anaconda |
| ML libraries like Pandas, Numpy, Scikit-learn, & Matplotlib | pandas 1.5.2,  numpy 1.23.2,  scikit-learn 1.1.3 and matplotlib 3.6.2 | * Pandas is used for reading data and data manipulation * Numpy is used for mathematical calculations * scikit-learn is used for training models * Matplotlib is used for data visualization |

* 1. BUSINESS UNDERSTANDING  
     Business organizations and finance companies face an important problem of customer churn where the customers churn from their services and this has also been an important problem in the banking industry where customers churn from the usage of credit card of the bank and choose cards of other banks. So Data Science and machine learning will play an important role in building churn models using the customer data.  
     These churn models would help the organization to foretell the behavior of a customer who is willing to churn. Further analysis with this model will also talk about the important attributes that makes a customer unhappy about the services and that leads to churn. Also, it is very important to retain an existing customer than getting new customers, additionally the expense of procuring another customer is more time consuming than holding a current customer. For this project, a prototype model was developed to improve prediction of customers who are on the verge of churn. Such a model would help businesses to keep up their consumer base with zero or low maintenance expenses.
  2. Model architecture  
     In this section, we would look into the architecture to better understand the flow of the model. The whole architecture will have several stages like data collection, Exploratory Data Analysis, Data Cleaning and Feature Engineering, Model building, Model evaluation and model deployment.



* 1. DATA Collection and Data Understanding

The data used is obtained from Kaggle. This is a CC0(Creative Commons Zero) licensed dataset, and it is obtained from a banking firm named Thera bank. They have analyzed this data after observing a steep decline in the number of users of their credit card.

After having the data available, a data exploration was carried out having the first insights on the data and discovering possible data quality issues and activities that need to be dealt with in the Data Preparation phase. These tasks were also crucial to secure a robust project plan in the Business Understanding phase. While performing the data exploration, it was observed that the columns like customer and age, and credit limit segments presented different characteristics and behaviors. Therefore, it was decided that, in the problem context, it would make sense to separate them and construct three different models (one for each customer segment).

The different characteristics and behaviors of the segments also impact the relevance of the observed variables. For example, by segmenting the data, some of the gathered variables’ information was no longer valuable and were, therefore, directly excluded.

For this project, the different types of attributes used in the data are listed and described as:

* CLIENTNUM: Client number. Unique identifier for the customer holding the account
* Attrition\_Flag: Internal event (customer activity) variable - if the account is closed then "Attrited Customer" else "Existing Customer"
* Customer\_Age: Age in Years
* Gender: Gender of the account holder
* Dependent\_count: Number of dependents
* Education\_Level: Educational Qualification of the account holder - Graduate, High School, Unknown, Uneducated, College (refers to a college student), Post-Graduate, Doctorate
* Marital\_Status: Marital Status of the account holder
* Income\_Category: Annual Income Category of the account holder
* Card\_Category: Type of Card
* Months\_on\_book: Period of relationship with the bank (in months)
* Total\_Relationship\_Count: Total no. of products held by the customer
* Months\_Inactive\_12\_mon: No. of months inactive in the last 12 months
* Contacts\_Count\_12\_mon: No. of Contacts in the last 12 months
* Credit\_Limit: Credit Limit on the Credit Card
* Total\_Revolving\_Bal: Total Revolving Balance on the Credit Card
* Avg\_Open\_To\_Buy: Open to Buy Credit Line (Average of last 12 months)
* Total\_Amt\_Chng\_Q4\_Q1: Change in Transaction Amount (Q4 over Q1)
* Total\_Trans\_Amt: Total Transaction Amount (Last 12 months)
* Total\_Trans\_Ct: Total Transaction Count (Last 12 months)
* Total\_Ct\_Chng\_Q4\_Q1: Change in Transaction Count (Q4 over Q1)
* Avg\_Utilization\_Ratio: Average Card Utilization Ratio

Data for any project plays an essential role in training a model. Here we have used a data with several such important attributes as shown above that contains the details about customer transactions, balances in terms of money , credit limit that they have and several such important attributes as shown above. Now we have read about the different attributes that are used in the data, additionally there are some attributes which are specific to the domain of banking and finance and these need to be understood in depth like the revolving balance, average open to buy and average utilization ratio, so these columns are explained below:

What Is a Revolving Balance?

If we don't pay the balance of the revolving credit account in full every month, the unpaid portion carries over to the next month. That's called a revolving balance

What is the Average Open to buy?

Open to Buy' means the amount left on your credit card to use. Now, this column represents the average of this value for the last 12 months.

What is the Average utilization Ratio?

The Avg\_Utilization\_Ratio represents how much of the available credit the customer spent. This is useful for calculating credit scores.

There is also an important relation between the discussed three attributes, which can be observed below:

Relation Between Avg\_Open\_To\_Buy, Credit\_Limit and Avg\_Utilization\_Ratio:

(Avg\_Open\_To\_Buy / Credit\_Limit) + Avg\_Utilization\_Ratio = 1

Now we would further look into the different attributes of customer data that we would likely use in predicting the reason for churn in the usage of credit cards.

First, we will try to look into the statistical and numerical analysis of the different columns present in the data and will try to gain some analysis from this statistical information of the data.

A picture containing table

Description automatically generated

Figure 4.1 Statistical Information about data

We can have several observations from the above chart:

* CLIENTNUM: It is a unique identifier for customers and can be dropped as it wouldn't add any information to our analysis.
* Customer\_Age: Average age of customers is 46 years; the age of customers has a wide range from 26 to 73 years.
* Dependent\_count: On average the customers in the data have 2 dependents and a maximum of 5 dependents.
* Months\_on\_book: All the customers of the bank have at least been with them for a year and 50% of the customers for at least 3 years.
* Total\_Relationship\_Count: All customers use at least one product of the bank, whereas 75% of customers use 5 or fewer products of the bank.
* Months\_Inactive\_12\_mon: On average customers were inactive for two months in the past 12 months - this shows that the bank customers are active in transactions or usage of cards it would be interesting to see if high inactivity leads to churning of a customer.
* Contacts\_Count\_12\_mon: On average banks and customers interacted twice in the past 12 months.
* Credit\_Limit: There's a huge difference between the third quartile and maximum value. The range of credit limits is very wide from 1438 to 34516, customers with high credit limit might be outliers.
* Total\_Revolving\_Bal: Average revolving balance of customers is 1162, there's not much difference in the third quartile and maximum value.
* Avg\_Open\_To\_Buy: Average amount that goes unused by the customers is 7469, the range is very wide for this variable and the extreme values (min and max) might be outliers.
* Total\_Amt\_Chng\_Q4\_Q1: For 75% of the customers the transaction amount in Q4 was less than the transaction amount in Q1 (as value is equal to ~0.9).
* Total\_Trans\_Amt: Average transaction amount of last 12 months is 4404, some customers spent as little as 510 while some customers made the transaction of more than 18k.
* Total\_Trans\_Ct: On average customers made 64 or fewer transactions while 75% of the customers made 81 transactions.
* Total\_Ct\_Chng\_Q4\_Q1: For 75% of the customers the number of transactions in Q4 was less than the transactions made in Q1.
* Avg\_Utilization\_Ratio: On average customers used ~27% of the available credit amount of their card, with 75% of the customers utilizing 50% or less of their available credit amount.

Now when we try to have a look into the unique values present in the object category

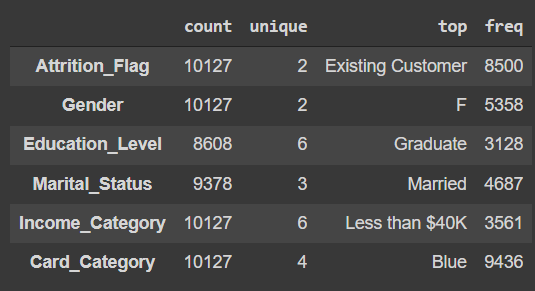


Figure 4.2 Unique values in each category

We can observe that:

* Most of the records are for existing customers.
* Most of the bank's customers are female
* Most customers are graduates.
* Most customers are married.
* Most customers lie in the income group of less than $40k
* Most customers have a blue card.
* 'abc' value of Income\_Category can be considered and treated as missing values.
  1. DATA PREPARATION

The data preparation phase covered all activities from the initial raw data to the final dataset used for modeling. The tasks managed in this phase can be separated into Data Selection, Data Cleansing, Feature Engineering, and Data integration and Format. These tasks are described below.

### Data Selection

By performing some statistical analysis and visualizations, some information was considered to be irrelevant. Subsequently, features that presented only or mainly one single value (with low variance) were considered irrelevant, having no (or not enough) predictability power for the given customer segment and, therefore, excluded.

Redundant information was present as well. Correlated features unnecessarily increase the feature space and can impact the model interpretation, masking meaningful interactions, and therefore, essential features appear irrelevant (Toloşi & Lengauer, 2011). Therefore, such features were identified and eliminated, avoiding such negative impacts in processes like feature selection.

Although the Pearson correlation coefficient is the most common measure of correlation, the correlation measure used was Spearman’s rank correlation coefficient, which measures the monotonic relationship between the two variables. Although this latter is calculated similarly to Pearson’s, Spearman’s rank correlation coefficient ranks both 𝑥 and 𝑦 to transform them to values between 1 and 𝑁, allowing this coefficient to be relatively robust to heavily tailed distributions and outliers (De Winter et al., 2016).

Said this, all groups of features presenting a Spearman’s rank correlation higher than 0.6 were analyzed. With the acquired business knowledge, correlations were eliminated by removing the less relevant feature.

### Data Cleansing and Analysis(Exploratory Data Analysis)

Data Cleansing is the process of improving the data quality in an existing system and may be directly tied to data acquisition (Maletic & Marcus, 2000). In this project, this was as well the approach. Firstly, as mentioned above, inconsistencies found on the data by performing some designed validations tests were already dealt with within the data acquisition process. Then, two other essential tasks to guarantee data quality were performed: outlier detection and handling missing data.

Finally, the missing data were handled. Given that the missing data present in the dataset is not considered NMAR (not missing at random) (Rubin, 1976), missing values imputation methods can be considered. Replacing the missing values with a new value is usually preferred over eliminating the variable with missing values or even the row, keeping the maximum information possible. However, supposing that the number of missing values in a variable is too high, not dropping such variable might lead to the risk of providing misleading information to the model given the high number of synthetic values. Therefore, the approach used to deal with the missing values varied accordingly with each variable’s characteristics. Moreover, a threshold was defined, to consider a variable’s removal depending on the variable’s significance.

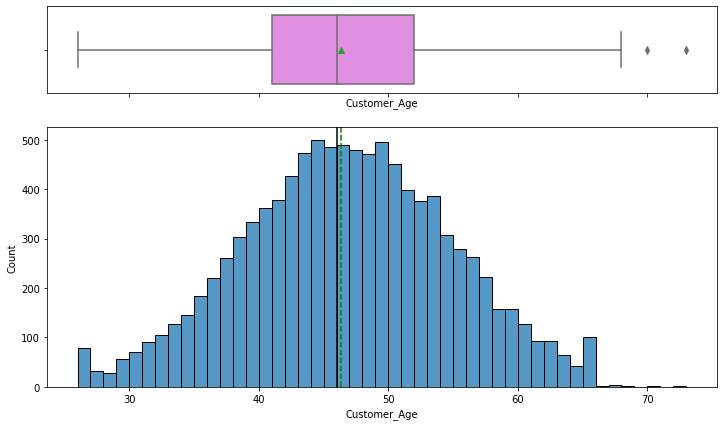
#### Missing Values in Numerical Variables

For numerical variables with missing values was decided to use the other variables present in the dataset and take advantage of their relations with the variables. These relations can be captured by computing the correlations among the variables with missing values and other variables present in the dataset and predicting the missing value’s value. Constructing a prediction model can maintain those relations, avoiding more straightforward techniques like mean imputation, which reduces variance in the dataset. Here, one widely used machine learning technique for missing data imputation, the K- nearest neighbors (Batista & Monard, 2003), was used.

#### Missing Values in Categorical Variables

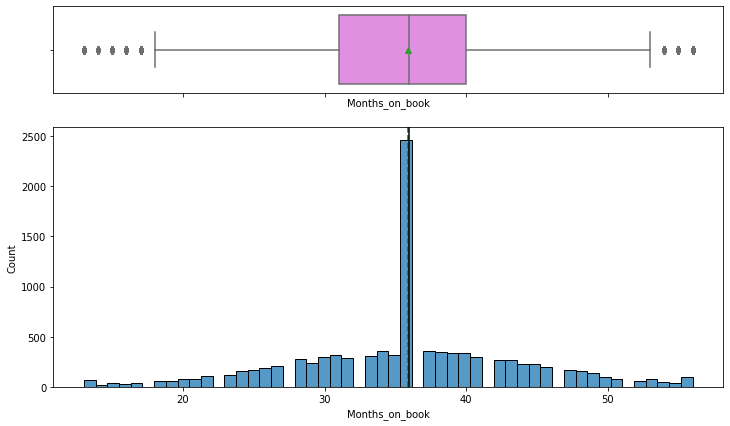
Regarding the missing values found on categorical variables, a special value characterizing the information’s absence already existed for some cases, so this value was assigned to the missing ones. For other variables, for which the percentage of missing values was relatively low (<1%), the mode was inputted. A more straightforward method would not introduce a significant bias given its low exposure. At last, according to the acquired business knowledge, other categorical variables with missing values were highly related to the customer’s location. Therefore, the customer’s municipality mode was introduced, always tracking this imputation's impact on the final variable’s distribution.

Observations on Customer\_age:



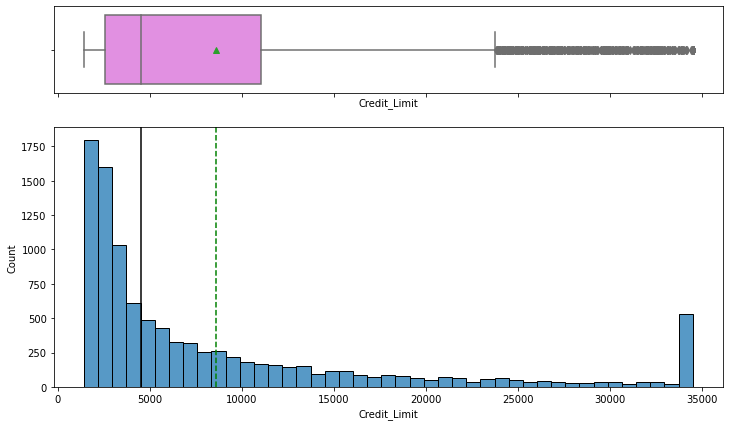
* The distribution of Customer\_Age is normally distributed with mean and median at 46 years.
* From the boxplot, we can see that there are a few outliers.

Observations on Months\_on\_book:



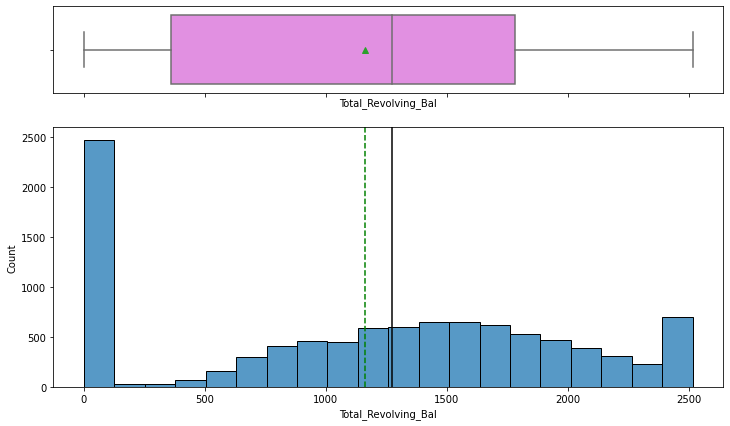
* Most customers are with the bank for 3 years.
* From the boxplot, we can see that there are outliers on both sides of the whiskers.

Observations on Credit\_Limit:



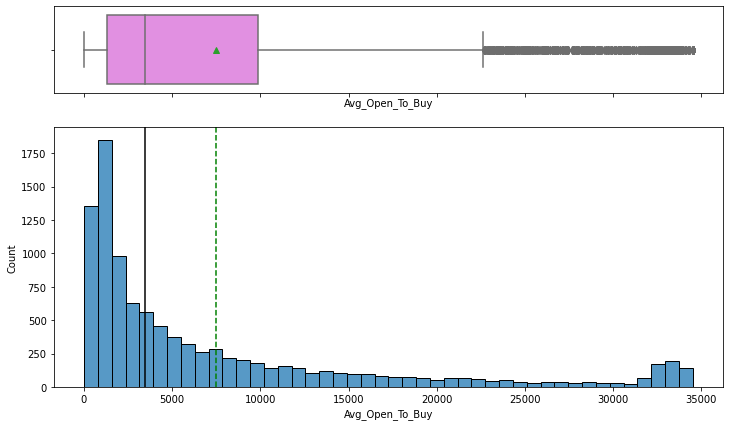
* The distribution of the Credit\_Limit is skewed to the right.
* There are quite a few customers with a maximum Credit Limit of 35000.
* 50% of the customers of the bank have a credit limit of less than <5000.

Observations on Total\_Revolving\_Bal:



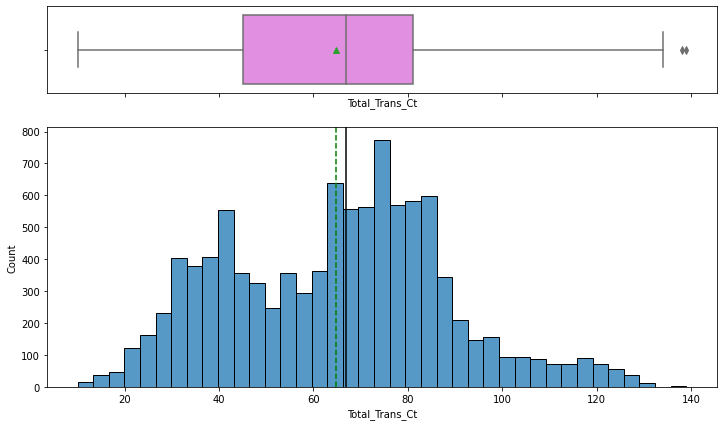
* Most customers pay the complete dues of credit card and have 0 revolving balance.
* There are quite a few customers with a revolving balance of 2500.

Observations on Avg\_Open\_To\_Buy:



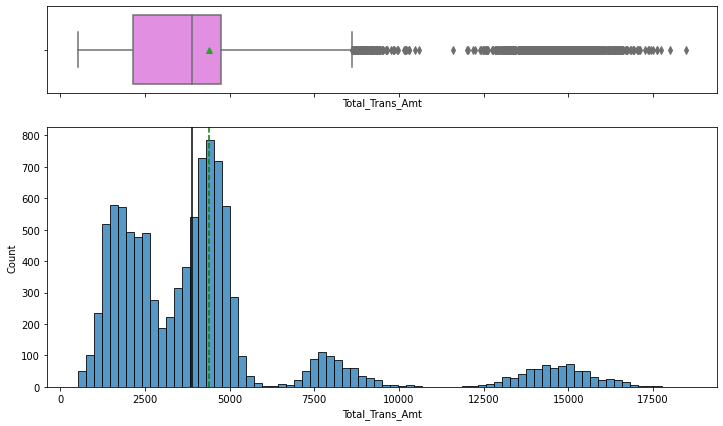
* The distribution of the Avg\_Open\_To\_Buy column is right-skewed.
* A right-skewed distribution indicates that most customers used a big part of their limit while only a few customers (on the right tail) were left with a majority of their credit amount.

Observations on Total\_Trans\_Ct:



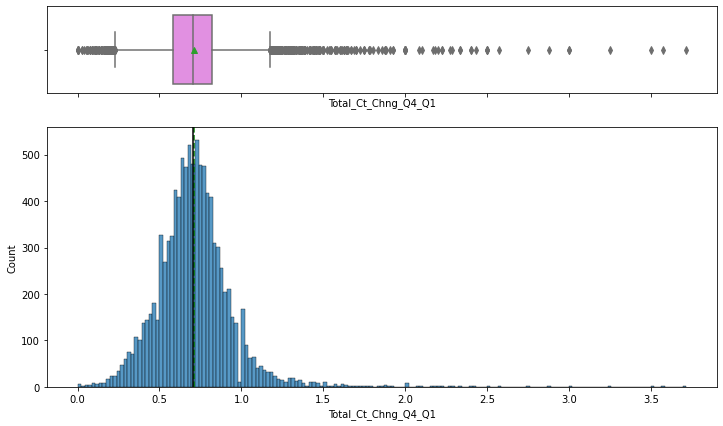
* The distribution of Total\_Trans\_Ct shows two peaks on 40 and 80 transactions in a year which indicates that customers used credit cards 3 to 6 times a month to make transactions.

Observations on Total\_Trans\_Amt:



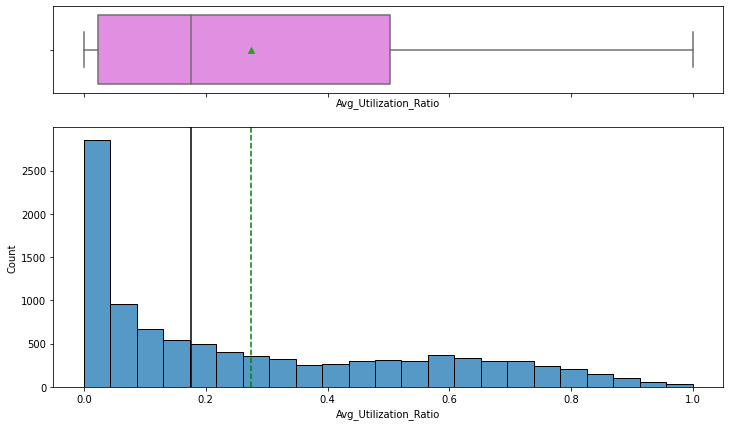
* The distribution of Total\_Trans\_Amt is skewed to the right.
* There are two peaks in data at total transaction amounts of one around 2500 and the second around the mean value of ~4500.
* From the boxplot, we can see that there are outliers - customers with more than ~8000 total transaction amounts are being considered as outliers.
* It would be interesting to check if the customers spending less with the card are the ones churning or the ones spending more are churning, if the latter is the case, then there is a problem for the bank as it is losing valuable customers.

Observations on Total\_Ct\_Chng\_Q4\_Q1:



* The distribution of Total\_Ct\_Chng\_Q4\_Q1 looks normally distributed but there's a slight skew towards the right.
* From the boxplot, we can see that there are outliers on both sides of the whiskers.

Observations on Avg\_Utilization\_Ratio:



* The distribution of Avg\_Utilization\_Ratio is skewed to the right.
* This distribution is not a positive sign for the bank as most of the customers are not utilizing their credit amount.

Some of the important observations from the univariate analysis

Credit limit, Average open to buy, and Average utilization ratio is right-skewed

1. Open to buy means how much credit a customer is left with
   * Low values of Open to buy could represent either

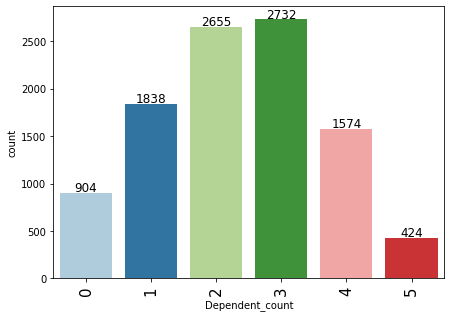
* Customers have low credit limits
* Customers are spending a lot so they are left less open to buy

1. Average utilization ratio = (1 - (open to buy/credit limit))
   * Low values of the Average utilization ratio represent
   * (Open to buy/credit limit) is nearly equal to 1
     + Open to buy is nearly equal to the credit limit
     + Customers are spending less using their credit cards
2. Credit limit is also right-skewed which represents - most of the customers have low credit limits.

Looking at the 3 variables, we can conclude that most of the customers have low credit limits and are not utilizing their credit cards much

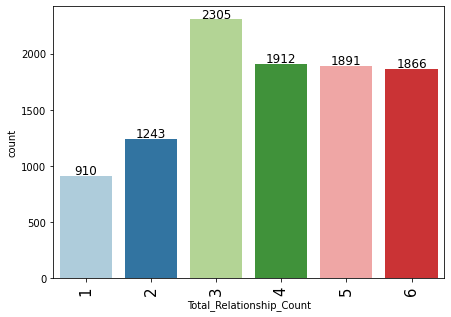
Now this statement justifies the right skewness for all 3 variables

Observations on Dependent\_count:



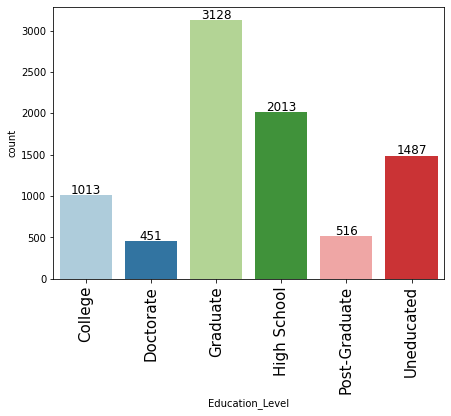
* The distribution of dependents is very realistic with most customers having 2 or 3 dependents.

Observations on Total\_Relationship\_Count:



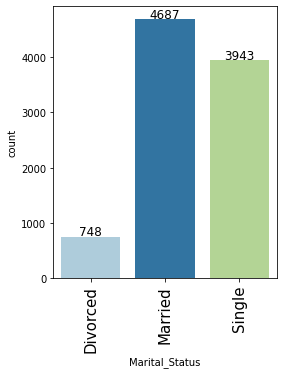
* 22.8% of the customers are using 3 products offered by the bank while an equal percentage ~19% of customers use 4 or more than 4 products.
* Equal percentage of customers using 4,5 and 6 products might suggest that customers who opt or buy the 4th product will also be ready to buy more products.

Observations on Education\_Level:



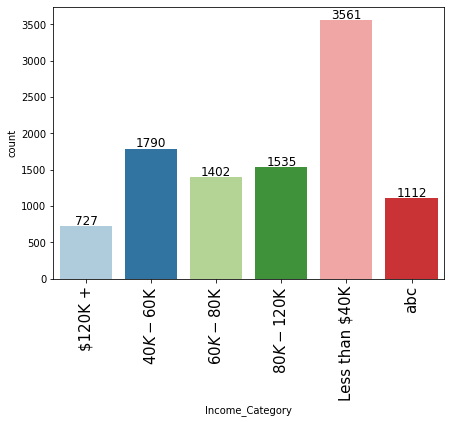
* 30.9% of the customers are graduates, followed by 19.9% of the customers who completed high school.
* Percentage of missing value in Education\_Level column - 15%.

Observations on Marital\_Status:



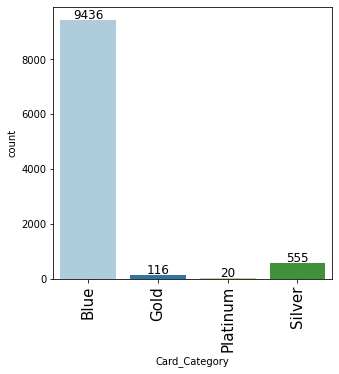
* 46.3% of the customers are Married, followed by 38.9% of Single customers.
* Percentage of missing value in Marital\_Status column - 7.4%.

Observations on Income\_Category:



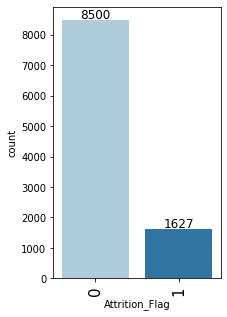
* 35.2% of the customers lie in the Less than 40k income category group, followed by 17.7% of the customers in the 40k-60k income group.
* Percentage of missing value in Income\_Category column - 11%.

Observations on Card\_Category:



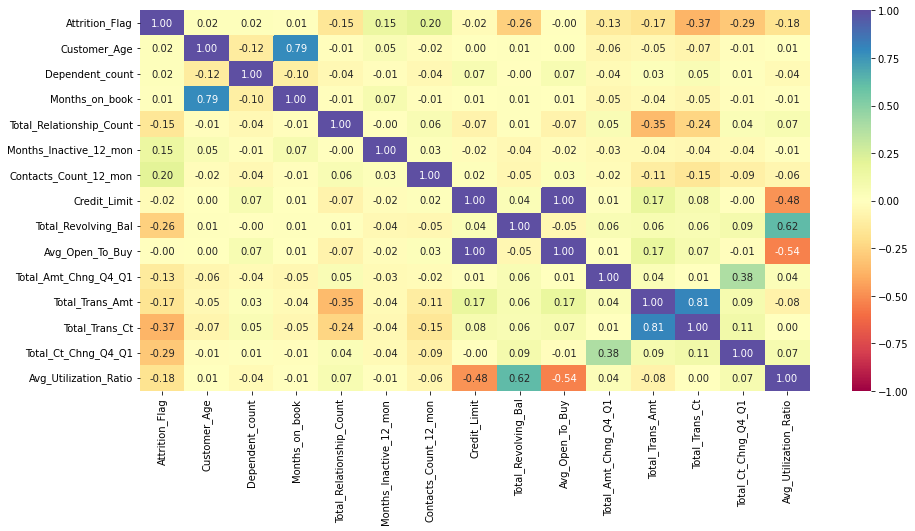
* 93.2% of the customers have the blue card.
* Blue card would be a standard card given by the bank to all its customers.

Observations on Attrition\_Flag:



* 16.1% of the customers attrited.
* This indicates an imbalance in the data.

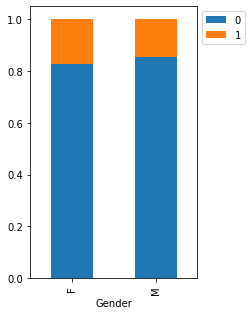
Bivariate Analysis



Observations from the heatmap

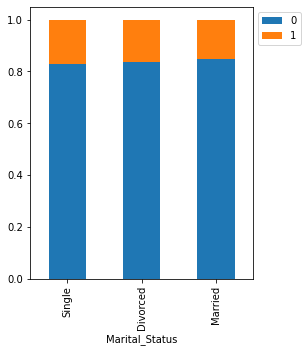
* Attrition\_Flag shows a bit of a negative correlation with Total\_Trans\_Ct (total transactions) and Total\_Trans\_Amt (total transaction amount).
* There's a strong positive correlation between Months\_on\_book and Customer\_Age, Total\_Revolving\_Bal and Avg\_Utilization\_Ratio, Total\_Trans\_Amt and Total\_Trans\_Ct.
* There's a negative correlation of Total\_Relationship\_count with Total\_Trans\_Amt and Total\_Trans\_Ct, Avg\_Utilization\_Ratio with Credit\_Limit and Avg\_Open\_To\_Buy.

Attrition\_Flag vs Gender



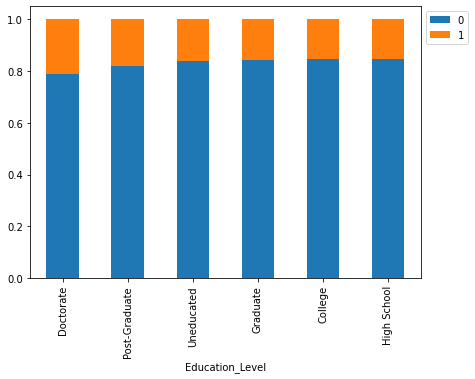
* There's not much difference in attrition percentages for Males and Females.
* ~20% of both Males and Females attrite.

### Attrition\_Flag vs Marital\_Status



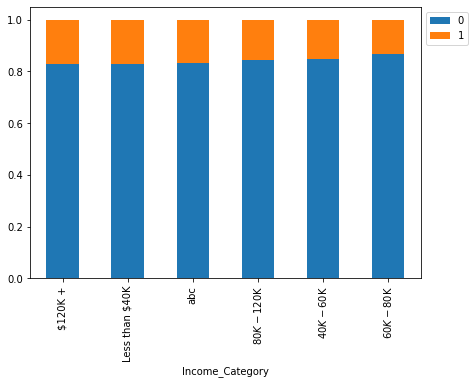
* There's not much difference in attrition percentages for Marital\_Status.
* ~20% of Singles, Divorced attrite.
* Married customers attrite the least.

Attrition\_Flag vs Education\_Level



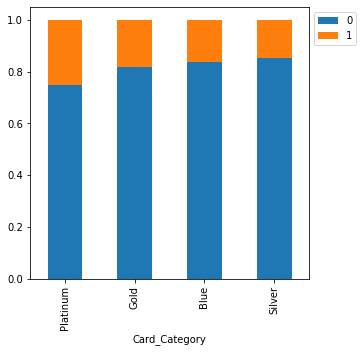
* Customers with higher education - Doctorates and Post Graduates are the ones most (~20% for both education levels) attiring.

### Attrition\_Flag vs Income\_Category



* The customers from two extreme income groups - Earning less than 40K and earning more than 120k+ are the ones attriting the most

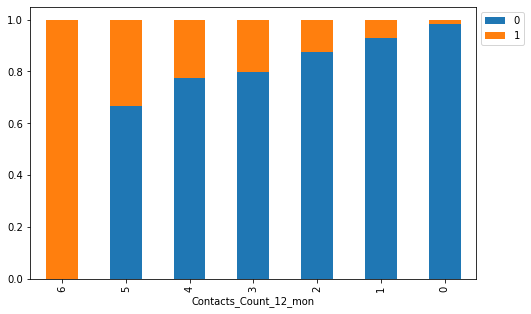
### Attrition\_Flag vs Card\_Category



* ~35% of attrition is amongst the customers with platinum cards followed by ~30% attrition in gold cards.
* Customers with Platinum and Gold cards are our premium customers and the highest attrition for these customers is alarming as they are using the premium card provided by the bank.

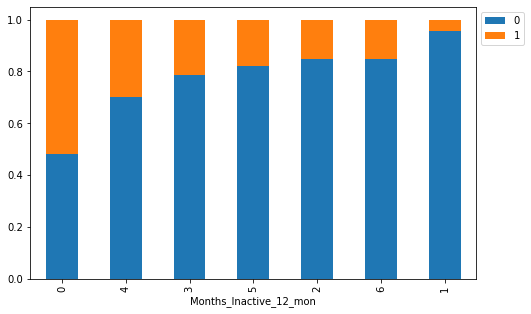
### 

### Attrition\_Flag vs Contacts\_Count\_12\_mon



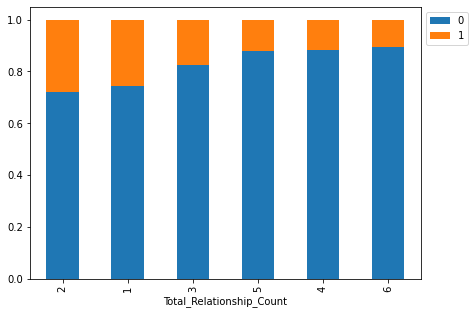
* Highest attrition is among the customers who interacted the most with the bank.
* This signifies that the bank is not able to resolve the problems faced by customers leading to attrition
* A preliminary step to identify attiring customers would be to look out for customers who have reached out to them repeatedly.

### Attrition\_Flag vs Months\_Inactive\_12\_mon



* As inactivity increases attrition also increases (2-4 months)
* The interpretation from here for 0 months and 6 months is difficult as customers who recently used the card attired the most while those who were inactive for 6 months attired less.

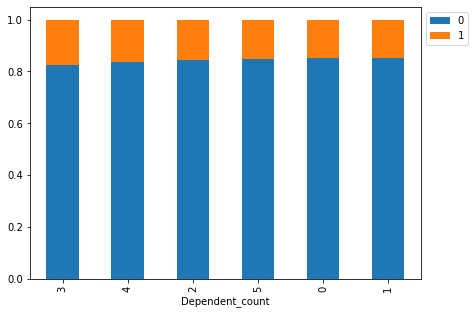
### Attrition\_Flag vs Total\_Relationship\_Count



* Attrition is highest among the customers who are using 1 or 2 products offered by the bank - together they constitute ~55% of the attrition.
* Customers who use more than 3 products are the ones least attiring, such customers might be more financially stable and actively invest in different services provided by the bank.

### 

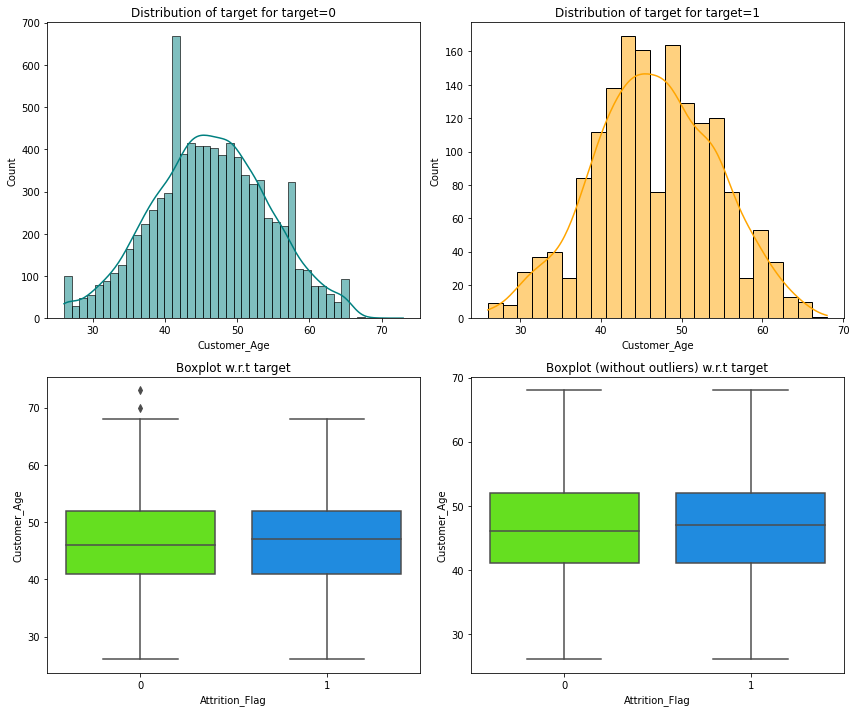
### Attrition\_Flag vs Dependent\_count



* More the number of dependents more is the attrition, more responsibilities might lead to financial instability in such customers.
* Attrition is fairly low for customers with 0 or 1 dependents.

### 

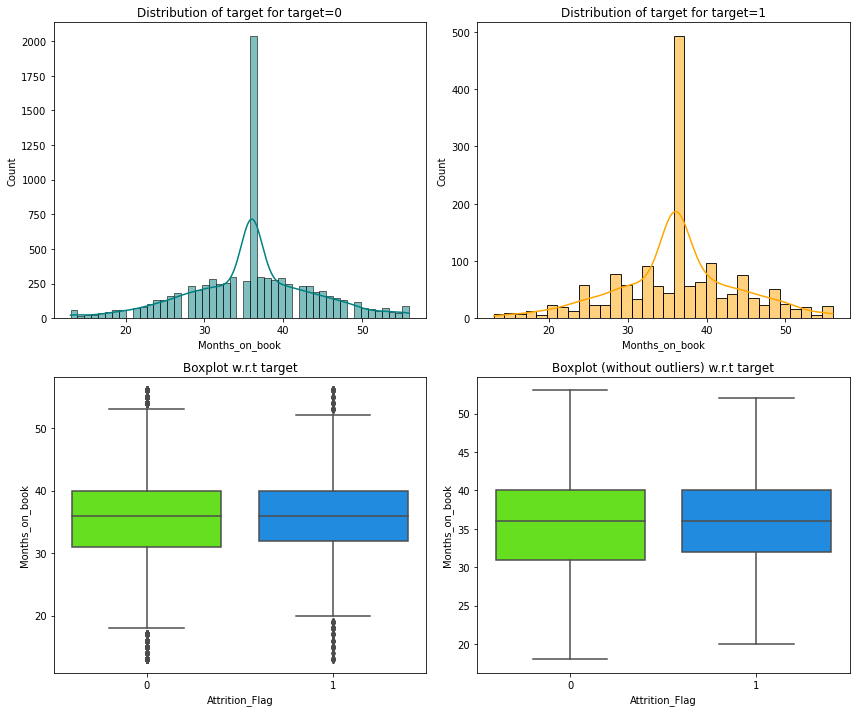
### Attrition\_Flag vs Customer\_Age



* There's no difference in the age of customers who attired and who didn't.

### 

### Attrition\_Flag vs Months\_on\_book



* Tenure of relationship with the bank doesn't seem to have an impact on attrition.

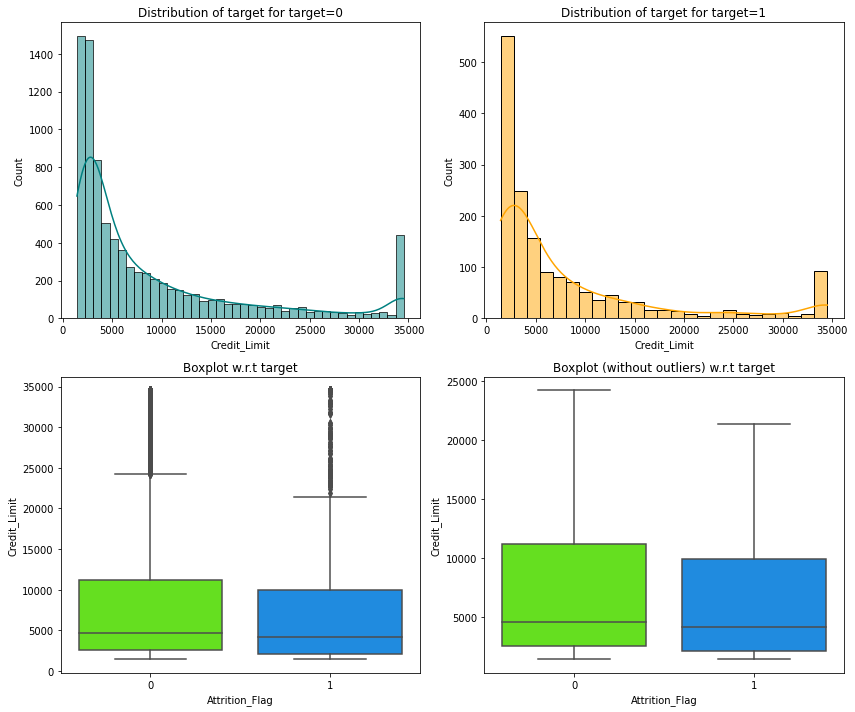
### 

### 

### 

### 

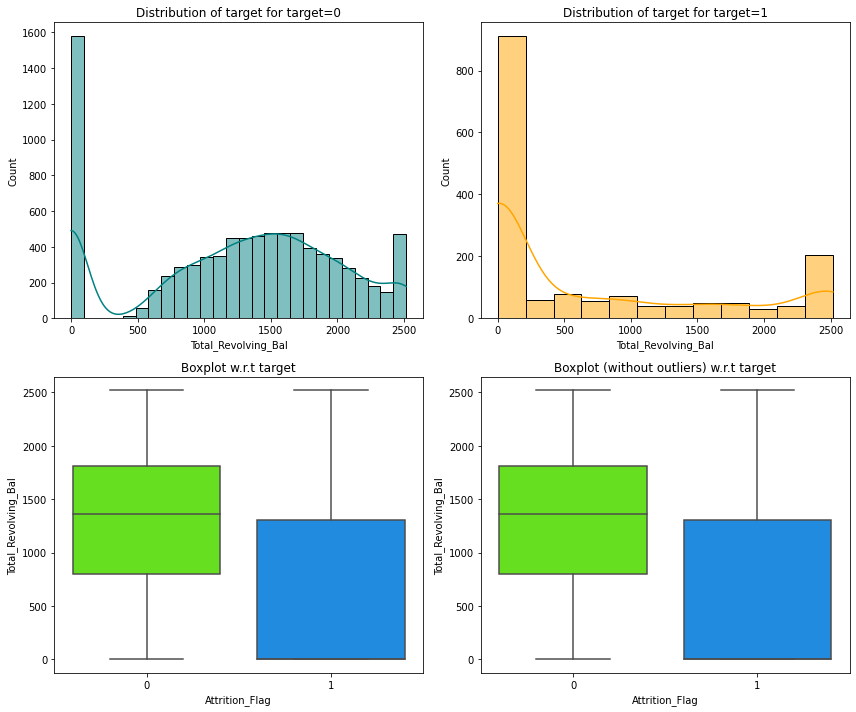
### Attrition\_Flag vs Credit\_Limit



* Customers with lower credit limits are the ones who attired.

### 

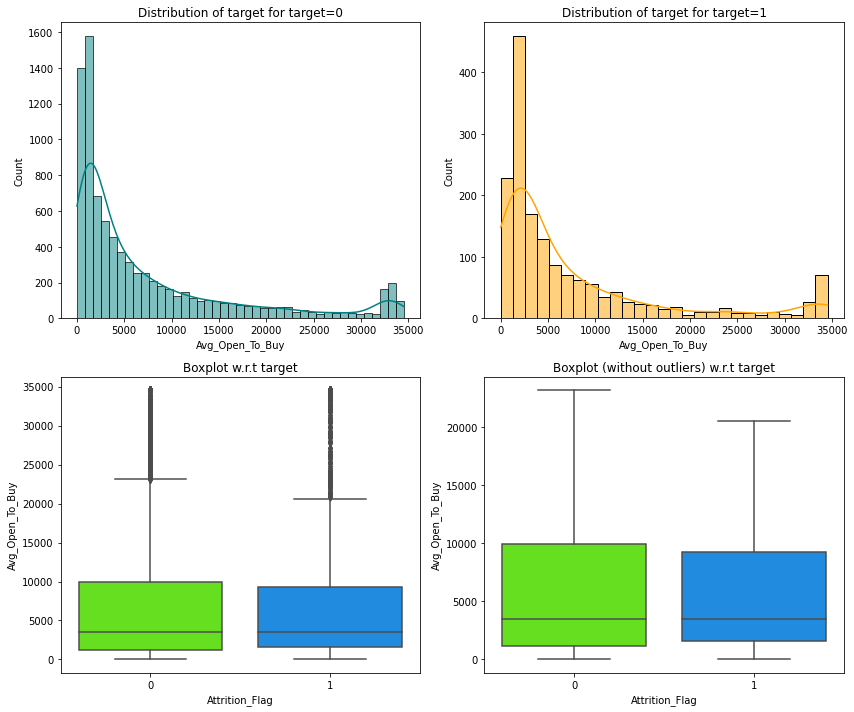
### Attrition\_Flag vs Total\_Revolving\_Bal



* Customers with less total revolving balance are the ones who attired, such customers must have cleared their dues and opted out of the credit card service.

### 

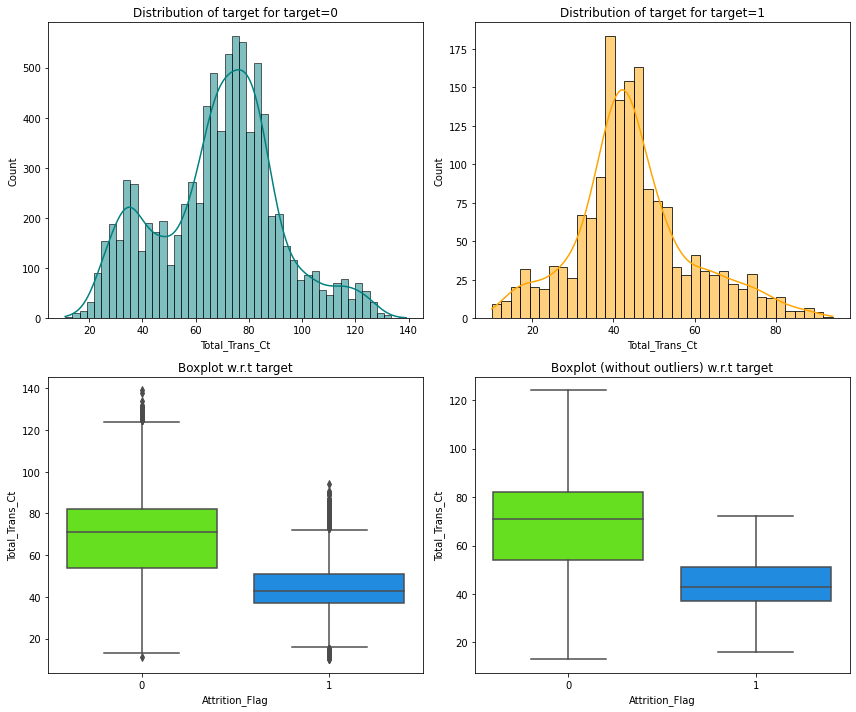
### Attrition\_Flag vs Avg\_Open\_To\_Buy



* There's not much difference in the distribution for an attired and existing customer.

### 

### Attrition\_Flag vs Total\_Trans\_Ct



* Less number of transactions lead to higher attrition.
* Customers with less than 80 to 100 transactions (or median transactions equal to 40) in a year should be more focused upon.

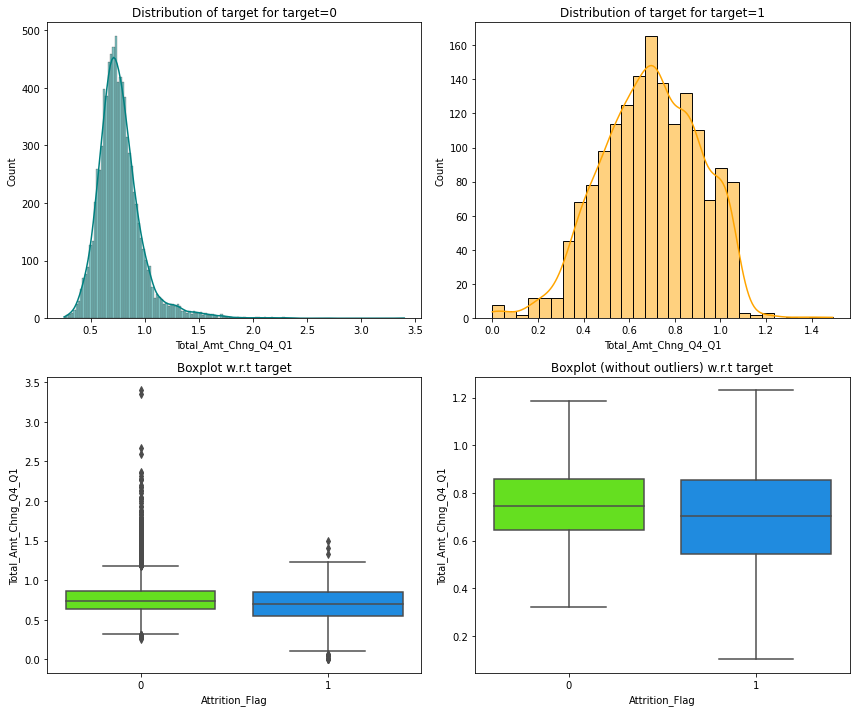
### Attrition\_Flag vs Total\_Trans\_Amt

### 

* Less number of transactions might lead to a less transaction amount and eventually leading to customer attrition

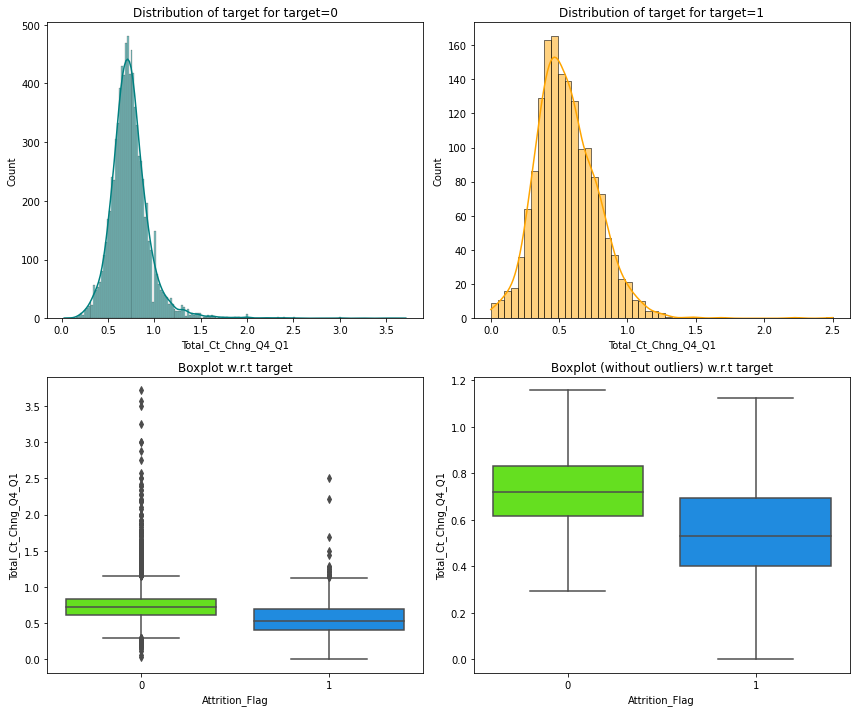
### 

### Attrition\_Flag vs Total\_Amt\_Chng\_Q4\_Q1



* Customers who didn't attrite showed less variability across Q4 to Q1 as compared to the ones who attrited.

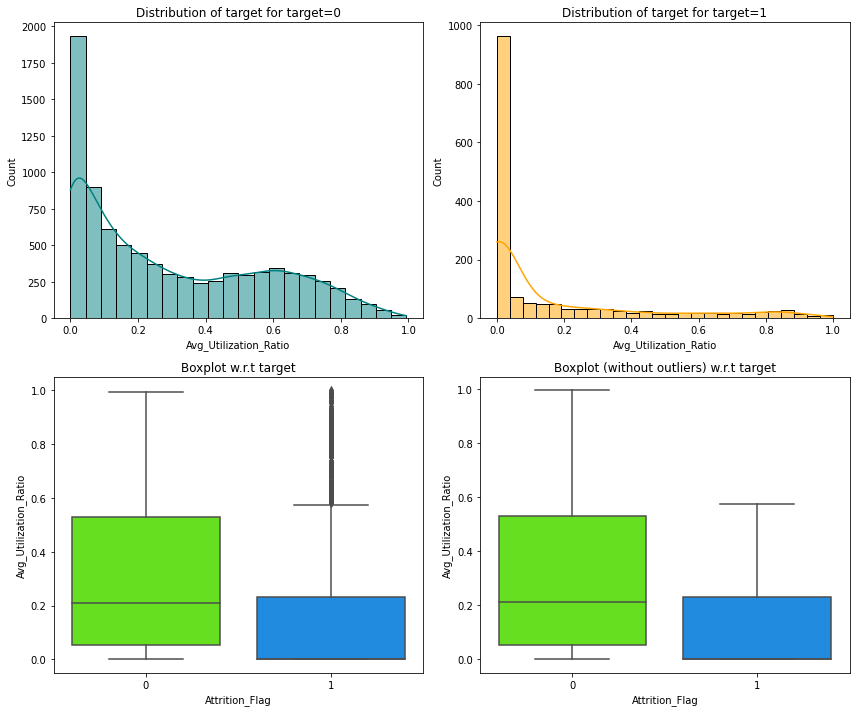
### Attrition\_Flag vs Total\_Ct\_Chng\_Q4\_Q1



* Customers who didn't attrite showed less variability across Q4 to Q1 as compared to the ones who attired.

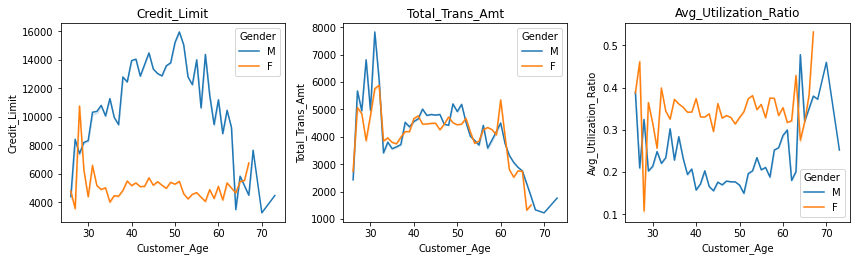
### 

### Attrition\_Flag vs Avg\_Utilization\_Ratio



* Customers utilizing their full credit limit are the ones who didn't attrite.
* Less utilization of the available credits indicates the inactivity of customers.

Analyzing "Credit\_Limit", "Total\_Trans\_Amt", "Avg\_Utilization\_Ratio" using a plot



* With age the credit limit of male customers increased (till 50 years) but for the female customers the credit limit constant throughout.
* Although the credit limit for female customers is less the total transactions made by them for all age groups is the same.
* Female customers utilized their credits more as compared to the male customers where utilization dropped from 30-50 years but increased after 60 years.

### Let's find the percentage of outliers, in each column of the data, using IQR.

Attrition\_Flag 16.066

Customer\_Age 0.020

Dependent\_count 0.000

Months\_on\_book 3.812

Total\_Relationship\_Count 0.000

Months\_Inactive\_12\_mon 3.268

Contacts\_Count\_12\_mon 6.211

Credit\_Limit 9.717

Total\_Revolving\_Bal 0.000

Avg\_Open\_To\_Buy 9.509

Total\_Amt\_Chng\_Q4\_Q1 3.910

Total\_Trans\_Amt 8.848

Total\_Trans\_Ct 0.020

Total\_Ct\_Chng\_Q4\_Q1 3.891

Avg\_Utilization\_Ratio 0.000

* After identifying outliers, we can decide whether to remove/treat them or not. It depends on one's approach, here we are not going to treat them as there will be outliers in real case scenario (in age, the total amount of transactions, number of transactions, etc.) and we would want our model to learn the underlying pattern for such customers.

### Missing value imputation

* We will first replace 'abc' values with 'np.nan' in Income\_Category.
* We will impute missing values in all 3 columns using mode/ most repeating category

We will encode the target variable using one-hot encoding using get\_dummies() function

For categorical variables where no such ordinal relationship exists, the integer encoding is not enough.

In fact, using this encoding and allowing the model to assume a natural ordering between categories may result in poor performance or unexpected results (predictions halfway between categories).

In this case, a one-hot encoding can be applied to the integer representation. This is where the integer encoded variable is removed and a new binary variable is added for each unique integer value.

### Feature Engineering

As a result of the previous tasks, new attributes are created derived from the dataset’s existing features. Also, in this task, generating new records to assist the modeling phase and improve the model’s performance on the dataset is considered.

* 1. MODELING

In this phase, the used modeling techniques are selected. Also, a test design is created to build the different models and correctly assess them, understanding their value for resolving the problem at hand.

Summary

To summarize, for the finding of this exploration, we can recommend organizations use data-mining strategies to change the current client data in their databases to exploitable information that can help them in their showcasing plans. Besides, it would be valuable for them to construct a prescient churn model by the utilization of data mining which assumes the part of a cautioning framework for the organizations, and it can assist them with spending their maintenance financial plan proficiently. Every year, many customers buy credit cards, and many customers leave so we need to forecast future churn in the credit card company.