**Visualization**

1. Bar Chart of Attrition Flag

According to the bar chart of attrition flag, we can find that there are 8500 existing customers which occupied 84% of total. It means most of customer attrited.

图表, 条形图

描述已自动生成文本

描述已自动生成

1. Gender Distribution

For gender pie chart, it shows 52.91% female customers and 47.09% male customers. The number of female customers more than male customers slightly.

图表, 饼图

描述已自动生成

1. Customer Age Distribution

This is a histogram of customer age distribution with a kde line. It distributes like a normal distribution, most of customers are between 40 years old and 50 years old. The youngest customer is 26 years old and the eldest customer is 73 years old.

图表, 直方图

描述已自动生成

1. Card Level & Income

We also find most people can only hold the blue card, which up to 93%, only few people can get higher level card. Especially for platinum card, only 20 people in our sample hold the platinum card. For further study, we output the count plot of card category and divide them with customer income. We find that people whose annual income less than $40K

图表, 条形图

描述已自动生成

1. Customer Age & Credit Limit

Based on the scatter plot output from python, we can find that it distribute like a normal distribution, the credit limit increased with people get elder until around 50 years old. After that, the credit limit turns to a downtrend in overall.

图表, 散点图

描述已自动生成

1. Heatmap of Numerical Dataset

According to the heatmap, we can find out there is obvious positive relationship between ‘Customer\_Age’ and ‘Months\_on\_book’, which correlation coefficient is 0.79. Besides, there also exist obvious positive relationship between ‘Total\_Trans\_Ct’ and ‘Total\_Trans\_Amt’, which correlation coefficient is 0.81. On the other hand, the correlation coefficient is -0.54 between ‘Avg\_Utilization\_Ratio’ and ‘Avg\_Open\_to\_Buy’, which means there exist strong negative relationship between them.

图表

描述已自动生成

1. Paired Plot of Financial Variable

There is a strong relationship between ‘Credit\_Limit’ and the ‘Avg\_Open\_to\_Buy’. For existing customers have higher values in ‘Total\_Ct\_Chng’ and ‘Total\_Amt\_Chng’ which means they factor in attrition ‘Total\_Ct\_Chng’ and ‘Total\_Amt\_Chng’ are somewhat related. The bigger the ‘Credit\_Limit’, the lower the ‘Avg\_utilization\_ratio’.

图片包含 背景图案

描述已自动生成

**Modeling**

Linear Regression

We build the linear regression based on normalized data, the normalization can make data more stable, reduce the influence of outliers. We using the k-fold to evaluate the linear regression model which predict the credit limit of customers respect to numerical dataset, we can find the r-squared is 1.0 for 5 models, so that the accuracy of model is 100%, which means this model have pretty good performance.

文本

描述已自动生成

Meanwhile, we test the VIF for the model, we find there are 5 parameters have VIF which is higher than the standard value (10), so we try to remove 'Customer\_Age', 'Months\_on\_book', 'Total\_Amt\_Chng\_Q4\_Q1', 'Total\_Trans\_Ct' based on heatmap and paired plot.

文本

描述已自动生成

After removing the parameter which have high VIF, we also get the accuracy 100% model and the VIF between all parameters have lower than 10. Therefore, we assume our model perform good.

文本

描述已自动生成