**Project of Advanced SAS Course**

**Project for the Telecom company --XUE LIU**

**Project Title:** Customer Distribution and Deactivation Analyses

**Project Objective:**

The attached data is the CRM data of a wireless company for 2 years. The wireless company would like to investigate customer distribution and business behaviours, and then gain an insightful understanding of the customers, and forecast the deactivation trends for the next 6 months.

**Import Data:**

1. Using the **data step** to read the data, mention dlm=’ ‘ and truncover.

2. Using the column input method, mention the starting and end positions.

3. Format the Actdt and Deactdt with date format and format the Sales variable with dollar format.

**Analysis requests:**

**1.1-1**  Explore and describe the dataset briefly: Is the acctno unique?

I used **2** methods to get the same result.

**Method #1: Use Proc Contents and Proc Freq**

The CONTENTS Procedure shows this data has 102255 Observations.

The FREQ Procedure shows the Acctno has 102255 Levels, so the Acctno is unique.

**Method #2: Use Proc SQL**

Select the distinct Acctno. The result is 102255 is consistent with the total number of observations.

**1.1-2**  Explore and describe the dataset briefly: What is the number of accounts activated and deactivated?

I used **3** methods to get the same result.

**Method #1: Use Proc SQL**

Regarding the active account, the condition is: where Actdt is not null and Deactdt is missing.

Regarding the deactive account, the condition is: where Deactdt is not null.

The number of activated accounts is 82620, and the number of deactivated accounts is 19635.

**Method #2: Use retain**

if not missing(Actdt) and missing(Deactdt) then activated\_accounts + 1, if not missing(Deactdt) then deactivated\_accounts + 1, retain total\_activated\_accounts total and deactivated\_accounts, if last then output.

**Method #3: Use if condition**

if not missing(Deactdt) then Status = 'Deactivated', else if not missing(Actdt) and missing(Deactdt) then Status = 'Activated'.

Segment the accounts by account status “Active” and “Deactivated”, if not missing(Deactdt) then Status = 'Deactivated', if not missing(Actdt) and missing(Deactdt) then Status = 'Activated', else Status = 'Unknown'.

**1.1-3**  Explore and describe the dataset briefly: When are the earliest and latest activation/deactivation dates available?

I used **2** methods to get the same result.

**Method #1: Use Proc Means**

Use Proc Mean to find the min(Actdt), max(Actdt), min(Deactdt)and max(Deactdt), and format them with date format.

**Method #2: Use Proc SQL**

In the select statement, find the min(Actdt), max(Actdt), min(Deactdt)and max(Deactdt), and format them with date format.

The earliest activation date is 01/20/1999. The earliest deactivation date is 01/25/1999.

The latest activation date is 01/20/2001. The latest deactivation date is 01/20/2001.

**1.2**  What are the age and province distributions of active and deactivated customers?

I use **Proc Freq** to find their distributions.

For each province, the proportion of activated accounts to the overall accounts is 81%.

Consistency in proportions could reflect business trends or policies. If there are similar market dynamics, customer demographics, or industry trends across provinces, business decisions might lead to a uniform ratio of activated and deactivated accounts.

I use **Proc Sgplot** to visualize their distributions.

I found active accounts and deactivated accounts are mainly concentrated in the ON province. Among the provinces, the QC province with the smallest proportion of active accounts and deactivated accounts.

And the overall age is mainly concentrated in the middle-aged crowd.

**1.3** Segment the customers based on age, province and sales amount:

First, I use **Proc Format** to format them: Sales segment: < $100, $100---500, $500-$800, $800 and above. Age segments: < 20, 21-40, 41-60, 60 and above.

Then, I used **2** methods to apply the formats.

**Method #1: Use Proc Tabulate**

In the ‘class’ statement: Status Age Province Sales.

Use the ‘format’ statement to format these two formats.

In the ‘tables‘ statement: Status, Age\*(n) Province\*(n) Sales\*(n).

**Method #2: Use Proc Report**

In the ‘column’ statement: Status Age Province Sales.

Use the ‘format’ statement to format these two formats.

and then use the ‘define’ statement: define Status/group; define Age / across 'Age'; define Province / across 'Province'; define Sales / across 'Sales Segment'.

**1.4.**Statistical Analysis:

1) Calculate the tenure in days for each account and give its simple statistics.

I use if not missing(Deactdt) then Tenure = Deactdt – Actdt, else if not missing(Actdt) and missing(Deactdt) then Tenure = '20JAN2001'd – Actdt to calculate the tenure in days for each account.

Because EarliestActivationDate is: 01/20/1999, LatestActivationDate and LatestDeactivationDate both are 01/20/2001 and this CRM data of a wireless company is for 2 years, so I put '20JAN2001'd in there.

There has min=0, check the observation whose Tenure=0.

I use **Proc Sgplot** to visualize the Tenure of active account distribution and Tenure of deactivated account distribution.

These two charts are both descending trends, which might be caused by the following reasons:

1. Account Lifespan:

The descending trend in the chart might suggest that most accounts have relatively short durations of usage.

This could be due to many customers opting to deactivate their accounts within a short period or perhaps the product lifecycle is short, such as temporary subscriptions or events, resulting in accounts becoming inactive quickly.

2. Customer Loyalty:

Short account tenures could imply lower customer loyalty.

Customers deactivating their accounts in a short span might indicate dissatisfaction with the service or product or switching to other providers in a competitive market.

3. Market Dynamics:

In certain cases, short-term account deactivations could reflect the intensity of market competition.

Competition might lead to more frequent switching of providers or services by customers, resulting in shorter account durations.

**1.4.** 2) Calculate the number of accounts deactivated for each month.

I use if not missing(Deactdt) then Tenure\_Month = intck('month',Actdt,Deactdt), else if not missing(Actdt) and missing(Deactdt) then Tenure\_Month = intck('month',Actdt,'21Jan2001'd).

I use **Proc Freq** to find the distribution of ‘Tenure\_Month’ and format the Deactdt with date format.

After that, I forecast the deactivation trends for the next 6 months.

First, I input the time data for the next 6 months and add sequence numbers to the existing months.

Then I use **Proc Corr** and **Proc Reg** to find the Parameter Estimate: B0=-269.78769, B1=87.93231, use these two numbers to predict the deactivation trends for the next 6 months.

**1.4.** 3) Segment the account, first by account status “Active” and “Deactivated”, then by

Tenure: < 30 days, 31---60 days, 61 days--- one year, over one year. Report the

number of accounts of percent of all for each segment.

First, check the missing values of tenure. The data set ALEXIA.MISSINGTENURE has 0 observations and 12 variables, which means there are no missing values.

In the 1.1-2 Method #3, I have already segmented the accounts by account status “Active” and “Deactivated”. The number of activated accounts is 82620. The number of deactivated accounts is 19635.

Next, I segment the accounts by tenure and use **Proc Freq** to report the number of accounts as a percent of all for each segment.

**1.4.** 4) Test the general association between the tenure segments and “Good Credit”

“RatePlan ” and “DealerType.”

I use **Proc Freq** to test the general association with ‘chisq’ option.

The ‘tables’ are:

Tenure\_Segment \* GoodCredit, Tenure\_Segment \* RatePlan, and Tenure\_Segment \* DealerType

H0: There is no significant association between the Tenure segments and GoodCredit, RatePlan, DealerType.

All the Chi-Square Probability are <.0001, indicating that we can reject this null hypothesis, and conclude that there is an ASSOCIATION between the Tenure segments and GoodCredit, RatePlan, DealerType, rather than just random differences.

**1.4.** 5) Is there any association between the account status and the tenure segments?

Could you find a better tenure segmentation strategy that is more associated with the account status?

I use **Proc Freq** to test the association between the account status and the tenure segments with ‘chisq’ option.

H0: There is no significant association between the account status and Tenure segments.

All the Chi-Square Probability are <.0001, which indicates that we can reject this null hypothesis, and conclude that there is an ASSOCIATION between the account status and Tenure segments, rather than just random differences.

The better tenure segmentation is:

if Tenure = 0 then Tenure\_Segment\_B = '0 days ', else if Tenure < 7 then Tenure\_Segment\_B = '1 - 7 days', else if Tenure < 30 then Tenure\_Segment\_B = '7 - 30 days', else if Tenure <= 90 then Tenure\_Segment\_B = '31 - 90 days', else if Tenure <= 180 then Tenure\_Segment\_B = '91 - 180 days', else if Tenure <= 365 then Tenure\_Segment\_B = '180 days - one year', else Tenure\_Segment\_B = 'over one year'.

I use **Proc Freq** to test the association between the account status and the new better tenure segments with ‘chisq’ option and I use **Proc Sgplot** to visualize the distribution.

After dividing the age groups more finely, it can be observed that among all deactivated accounts, the largest proportion of Tenure is '180 days - one year' accounting for 30.82%.

Following this, there are the '91 - 180 days' at 20.84%, and 'over one year' at 19.65%.

This indicates that accounts with a usage duration exceeding 90 days are more prone to churn, demonstrating a lack of strong attachment among long-term users.

Specific marketing strategies should be devised for these long-term users to enhance their loyalty.

**1.4.** 6) Does the Sales amount differ among different account statuses, GoodCredit, and

customer age segments?

I use **Proc GLM** to check the significant difference and use **Proc ANOVA** to check the significant association.

For Sales and Status:

H0: There is no significant difference in the sales amount among different account statuses.

The Levene's Test p-value= 0.0505, higher than 0.05, which indicates that we can not reject the Null Hypotheses, we haven't found strong evidence to conclude that there are significant differences in sales amount among different account statuses.

H0: There is no significant association between Sales and account status.

The p-value= 0.3997, higher than 0.05, which indicates that we can not reject the Null Hypotheses and conclude that there is no association between the Sales and account status.

For Sales and GoodCredit:

H0: There is no significant difference in the sales amount among different GoodCredit.

The Levene's Test p-value= 0.6795, higher than 0.05, which indicates that we can not reject the Null Hypotheses, we haven't found strong evidence to conclude that there are significant differences in sales amount among different GoodCredit.

H0: There is no significant association between Sales and GoodCredit.

The p-value= 0.7788, higher than 0.05, which indicates that we can not reject the Null Hypotheses and conclude that there is no association between the Sales and GoodCredit.

For Sales and Age segments:

H0: There is no significant difference in the sales amount among different age segments.

The Levene's Test p-value= 0.2328, higher than 0.05, which indicates that we can not reject the Null Hypotheses, we haven't found strong evidence to conclude that there are significant differences in sales amount among different age segments.

H0: There is no significant association between Sales and age segments.

The p-value= 0.7583, higher than 0.05, which indicates that we can not reject the Null Hypotheses and conclude that there is no association between the Sales and age segments.

For Sales and Status, GoodCredit, Age segments:

H0: There is no significant difference in the sales amount among different Status, GoodCredit, and Age segments.

All the p-values are higher than 0.05, which indicates that we can not reject the Null Hypotheses, we haven't found strong evidence to conclude that there are significant differences in sales amount among different account statuses, GoodCredit, and age segments.

H0: There is no significant association between Sales and Sales and account status, GoodCredit, or age segments.

All the p-values are higher than 0.05, which indicates that we can not reject the Null Hypotheses and conclude that there is no association between the Sales and account status, GoodCredit, and age segments.

Analysis Data -1: Exploring Data: Categorical Variables

I use **Proc Freq** to find the distribution of categorical variables.

\* About GoodCredit

1 represents good credit, 0 represents bad credit, there have 69.44% of accounts are good credit, and 30.56% of accounts are bad credit.

The proportion of bad credit accounts is too high.

In the future, when opening an account, the customer’s credit rate should be properly checked.

\* About RatePlan

66.69% of accounts are RatePlan#1, 19.74% of accounts are RatePlan#2, 13.57% of accounts are RatePlan#3.

The proportion of RatePlan#1 to the overall data is twice the sum of RatePlan#2 and RatePlan#3.

This means that RatePlan#1 is more popular, and we can consider launching more plans like RatePlan#1 in the future.

\* About DealerType

There 54.87% of accounts are A1, more than all other DealerTypes combined.

It may be because the deals provided by this dealer are cost-effective, or it may be that the dealer provides a variety of deals and covers a wider range of areas.

\* About Province

There 44.11% of accounts are in the ON province, more than all other DealerTypes combined.

This may be because the population of ON Province is the most populous province in Canada.

If we want to analyze the relationship between accounts and provinces, we also need to refer to the population distribution of each province.

Analysis Data -2: Exploring Data: Continuous Variables

I use **Proc Means** and **Proc Univariate** to find the distribution of continuous variables.

\* About Age

The mean of age is 47.6, the overall age is mainly concentrated in the middle-aged crowd.

The maximum age is 99, and the minimum is 0 which is very abnormal and needs further inspection.

Missing data accounted for 7.54% of the total data.

The Skewness value is 0.03053421, very close to zero, suggesting that the data is roughly symmetric and exhibits minimal skewness.

The Kurtosis value is -0.4201847, which is less than 3, indicating a relatively flatter distribution without a pronounced peak. Based on the Skewness and Kurtosis values, the distribution for the "Age" variable appears to be relatively symmetric and flat.

After I used **Proc** **SQL**, I found that the age under 5 has 691 observations.

The above results show that there is a lot of data younger than 5 years old.

This could be due to two reasons:

1. It is mandatory to provide a valid ID when opening an account and many parents open accounts for their children. When the baby is under one year old, the system will input age=0.

Maybe new users have a better deal, Parents open an account in the name of the child, and the parents use it themselves.

2. It is not mandatory to provide a valid ID when opening an account. Many people fill in their age at will or even don’t fill in it.

\* About Sales

The mean of sales is 47.6, the maximum sales are 1200, and the minimum is 0.

Missing data accounted for 8.42% of the total data.

The Skewness value is 2.36652039(positive) which indicates the distribution is skewed to the right, with a longer tail on the right side.

A positive skewness value suggests a right-skewed distribution of sales amounts.

This means that most sales amounts are relatively small, but there are also some larger sales amounts (Potentially representing a few significant transactions or orders).

This could indicate that most transactions have smaller amounts, but a few transactions with higher amounts impact the overall distribution.

The Kurtosis value is 5.28183679, which is significantly greater than 3.

A high kurtosis value could indicate that the distribution of sales amounts has a relatively sharp and pronounced peak.

This suggests the presence of transactions with very high amounts.

It might imply the existence of outliers or large orders within the sales amounts, or that certain transaction amounts are relatively concentrated.

Analysis Data -3: Outliers and Missing values

I use **Proc Sgplot** to visualize the Box Plot to check outliers.

\*Outliers for Sales

These outliers represent transactions with unusually high values, and might have the following practical meanings:

1. Outliers in sales amounts could be indicative of data entry errors, measurement inaccuracies.

2. Outliers on the high end could represent large orders, significant contracts, or high-value customers. These transactions might have strategic importance or unique requirements, which might influence business decisions or sales strategies.

\*Outliers for Age

From this box plot, I can't find any outliers, there may be 2 reasons for this:

1. There are no outliers here.

2. There are Outliers here, but they are affected by other reasons, such as being affected by missing values.

Age has 7708 observations with missing values.

First, I replace the missing values with the mean (mean of age is 47.6) and then find the outliers again.

Outliers in age may reflect data entry errors or exceptional cases in the dataset.

Exceptional cases may have the following two situations:

1. Some parents open accounts for young children if a valid ID is mandatory for account opening.

2. Some people fill in the age at will or even don't fill it in if a valid ID is not mandatory for account opening.

Analysis Data -4: Association between variables

I wrote a Macro to test the association between variables.

H0: There is no significant relationship between them.

Between these variables: (Sales\_Segment, RatePlan), (GoodCredit, RatePlan), (GoodCredit, DealerType), (GoodCredit, Status), (RatePlan, DealerType), (RatePlan, Status), (DealerType, Status)**,** all the Chi-Square Probability are less than 0.0001, indicates that we can reject this null hypothesis, and conclude that there is an ASSOCIATION between them.