Customer Lifetime Value (CLV) Analysis

CAR INSURANCE

In this project I obtained the retention rate of car insurance customers from one company, these customers are divided into 3 groups according to their corresponding profit value. This annual customer retention rate has been observed for 7 years. Besides, the long-term actuarial data shows that customers are likely to drop insurance coverage at 12 years due to life events, such as marriage, children, college, etc. Thus, the project requires us to predict the customer lifetime value with a span of a potential 12 years expected relationship between this firm and its customers. In consideration of only limited retention data at hand I try to make customer lifetime value prediction using shifted beta geometric distribution method from Fader and Hardie (sBG method) instead of traditional regression method.

Please note that the graph on Chapter 1 uses the actual retention rate values. In later chapters Chapter 2 and Chapter 3, the Fader and Hardie sBG will be used to predict the retention rate, and the values on the graphs/charts are the predicted values.

In this project I use profit for calculation directly, which is provided as follows:

Group 1 with profit of 250 USD/year,

Group 2 with profit of 311 USD/year,

Group 3 with profit of 279 USD/year.

Data source: CarInsurance.csv

Structure of this paper

- Chapter 1 Observation CAR INSURANCE data
- Chapter 2 Design and construction CLV formula
- Chapter 3 Calculating CLV values and verifying different assumptions
- Chapter 4 Conclusion

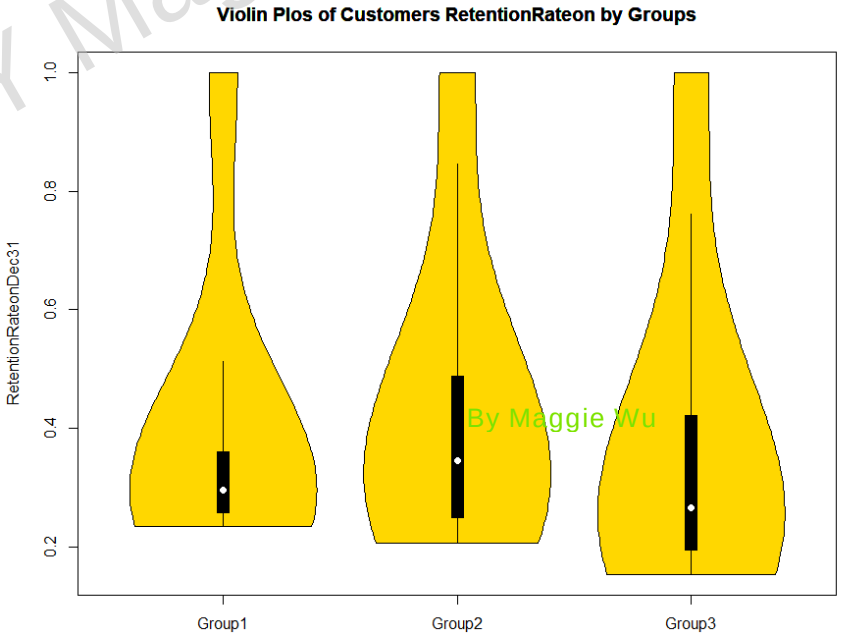
Chapter 1. Observation CAR INSURANCE data

All three Car insurance groups have a 7- year customer retention rate. I know there is a direct relationship between CLV and retention, so what is the difference in retention rates between the three groups so far?

- Overall Data Differences
- Cliff Differences

Violin plot: Declining cliff in retention rates by Groups

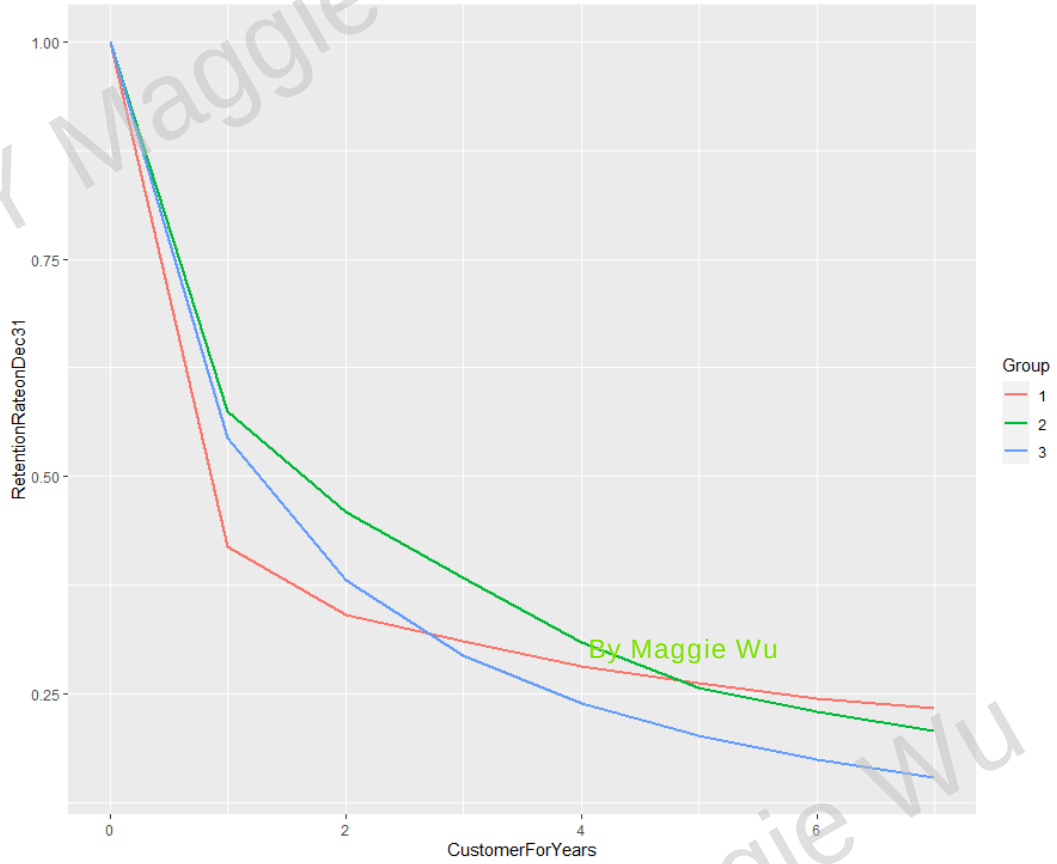
This Violin plot shows the distribution of the data for the 3 groups (the string position contains the mean and the shape reflects the distribution). The distribution of the data for Group 1 is clearly narrower, ranging from approximately 0.25 to 0.55. **Group 2 has the highest mean and more evenly distributed** data.



Line chart: Declining cliff in retention rates by Groups

I know that if I take the customer retention rate at the start time as 1, then the retention rate will change as I move back in time. In the graph below of the change over time, you can see that the Group1 set of data has the highest retention rate (retention) in year 7 of the three sets, but at the same time this set has the lowest retention rate in year 1. Group1 set has the most different performance compared to the Group2 and Group3 sets of data. (Note, first year churn is normal in the US, people will switch their car insurance to another insurer in the second year as the latter will pay their

first-year premiums, but the first-year decline is relatively cliff for Group 1).



Chapter 2. Design and construction CLV formula

2.1 Designing CLV formula for CAR Insurance based on Industry features & Data features

- **CAR Insurance Industry features**
  - ✓ BtoC, Usually, account expansion or account contraction is less relevant to car insurance products. Most car insurance customers are generally on a specific policy (Insurance Choice Program) and remain on that plan for the duration of the insurance contract.
  - ✓ Annual Renewals churn fairly evenly over time, with a larger churn at each contract renewal
- **Data features,From Chapter 1 I can find that**
  - ✓ Churn has been relatively stable in subsequent renewal plans, except for the first year when churn was high. In particular, the Group1.
  - ✓ The profit Annual already given for each group customer.

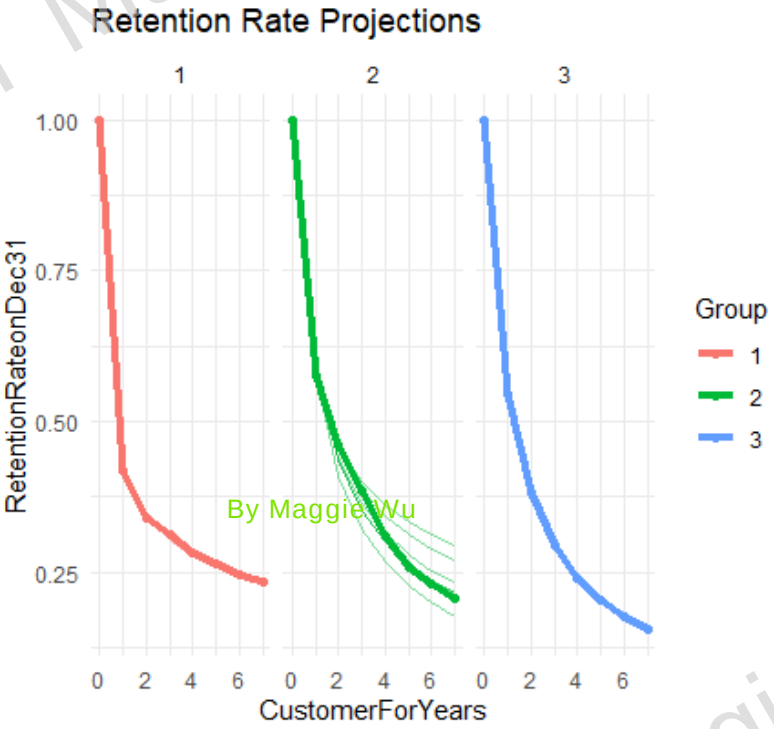
And limited to the sample data and taking into account that Insurance claims are already accounted for in the customer retention costs, the CLV can be calculate by a basic CLV formula.

$$CLV(years) = \sum_{n=0}^{years} (1 - CustomerChurnRate_n) * \text{profitAnnual} \quad \boldsymbol{or} \sum_{n=0}^{years} CustomerRetensionRate_n * \text{profitAnnual}$$

2.2 sBG prediction function

In this part I use retention data and predict outcomes using the Fader-Hardie functions (sBG method). Using ggplot() I plot the retention curves again to see how the predicted curves differ from the observed data curves.

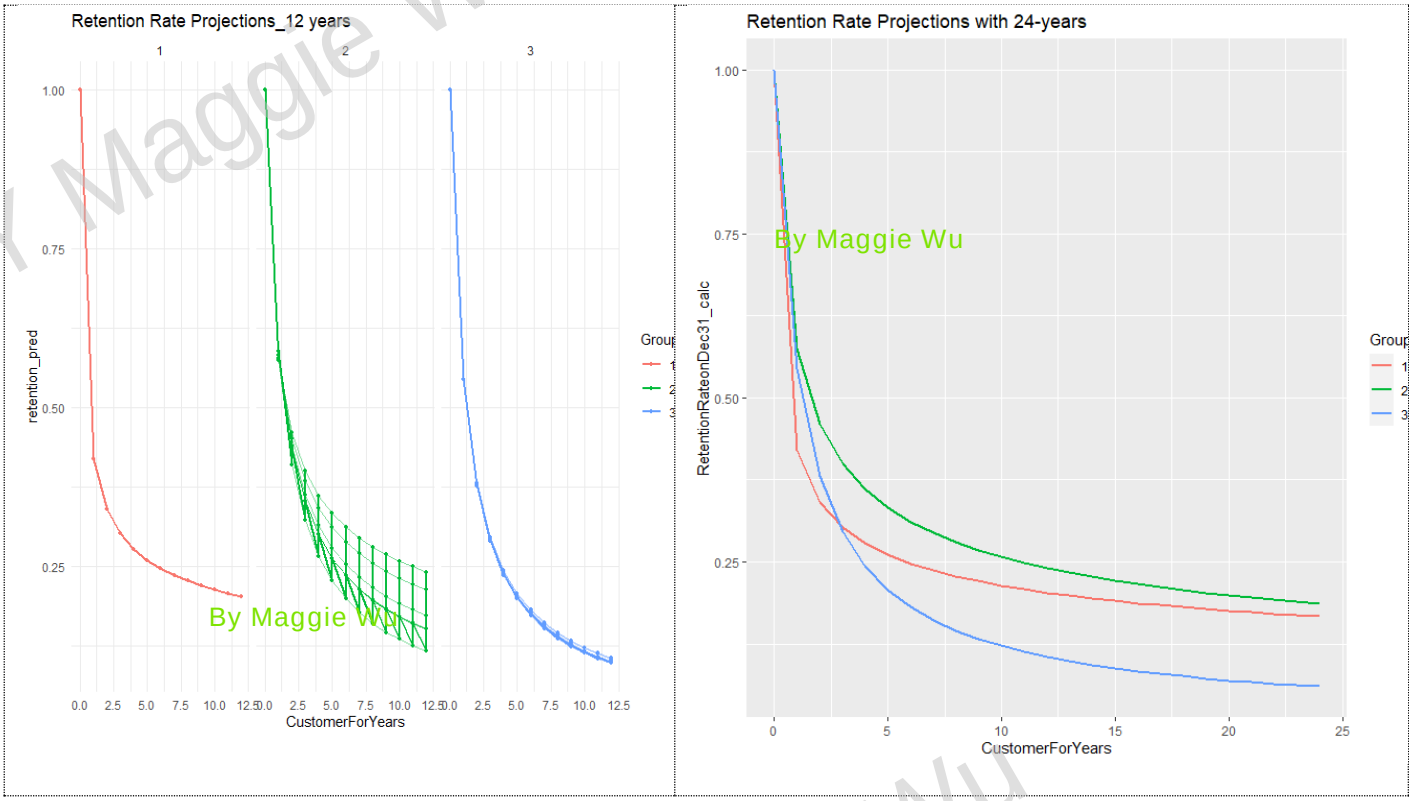
The thick line indicates the observed curve and the fine lines indicate the predicted curves, which mean average percentage error (MAPE). Here it appears that group one has the highest accurate prediction, and group 3 has some small deviation, while group 2 has obvious deviation. However, since I only have very limited data, I think it is still amazing that I got the result using sBG method.



2.3 Retention rate prediction for the 12 years and 24 years

Continue sBG prediction function for 24-years projection, When this period is reached, the current year's customers will have reached middle age or even old age, which also means that their choice patterns will be more stable.

- The downward trend in retention rates for Group 1 starts to slow in year 5, at a slower rate than Group 2 and Group 3.
- I continue to use this method to predict retention rates at 24 years (when customers will be in their middle age) and see that Group 3 will lose all its customers, with Group 1 and 2 having a good 'long tail effect'.



2.4 Cumulative CLV based on Retention rate prediction

Once customer retention projections based on the annual cycle have been obtained, the cumulative CLV for each group of customers can be obtained by combining the cumulative profit calculation

Note: Please refer to Appendix section 2 ~ section 4 for codes covered in this chapter

Chapter 3. Calculating CLV values and verifying different assumptions

3.1 Calculating CLV values and answer the assignment questions

Based on the constructed retention prediction model and the CLV calculation function, I can then directly answer the question of this assignment, which is the cumulative CLV for each group under the 12-year forecast (The result as following table shown).

Group 2's CLV rank the top position. Group 3's CLV is the lowest due to the worst retention rate. Group 1 has the lowest profit value, its retention rate of 20.3% is better than group 3 of 10.5%, as a result, it has higher CLV than group 3. **Therefore, see that "RetentionRateonDec31" is an important factor in the CLV value to Groups' result.**

Group	profit of USD/year ($)	RetentionRateonDec31_calc	CLV_yearly ($)	CLV_cum ($)
1	250	0.203	50.8	1040
2	311	0.241	75	1564
3	279	0.113	31.5	984

3.2 Verifying different assumptions based on "RetentionRate on long time term (24-years)" and "same profit"

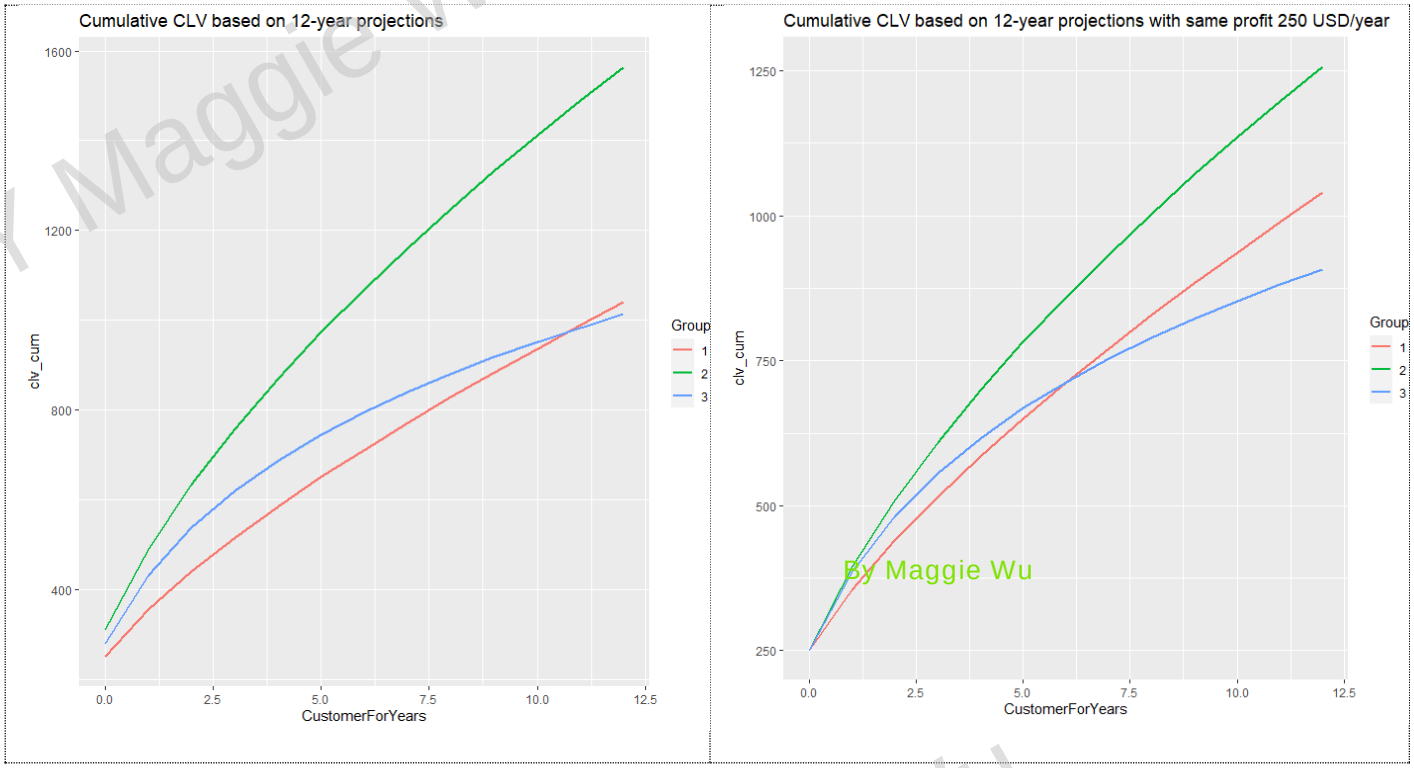
Based on the observation of the long-term (24-years) customer retention forecast data in Chapter 2 and the profit condition given in the title, I will further test the following hypothesis.

- **Assumptions** 1. when the customer lifecycle is extended to 24 years (i.e. in the scenario where customers choose to prefer further consolidation), do the results of the comparison between the three data sets remain consistent?
- **Assumptions** 2. when the product groups are balanced to a consistent customer profit cost scenario, do the values for the three groups of customers change and does this affect the results?

Verification for Assumptions 1.

By modifying the "Year " parameter in the function, I obtained plots for 12 (left-graph) and 24 years (right-graph). Comparing the 3 Groups data trends, I find:

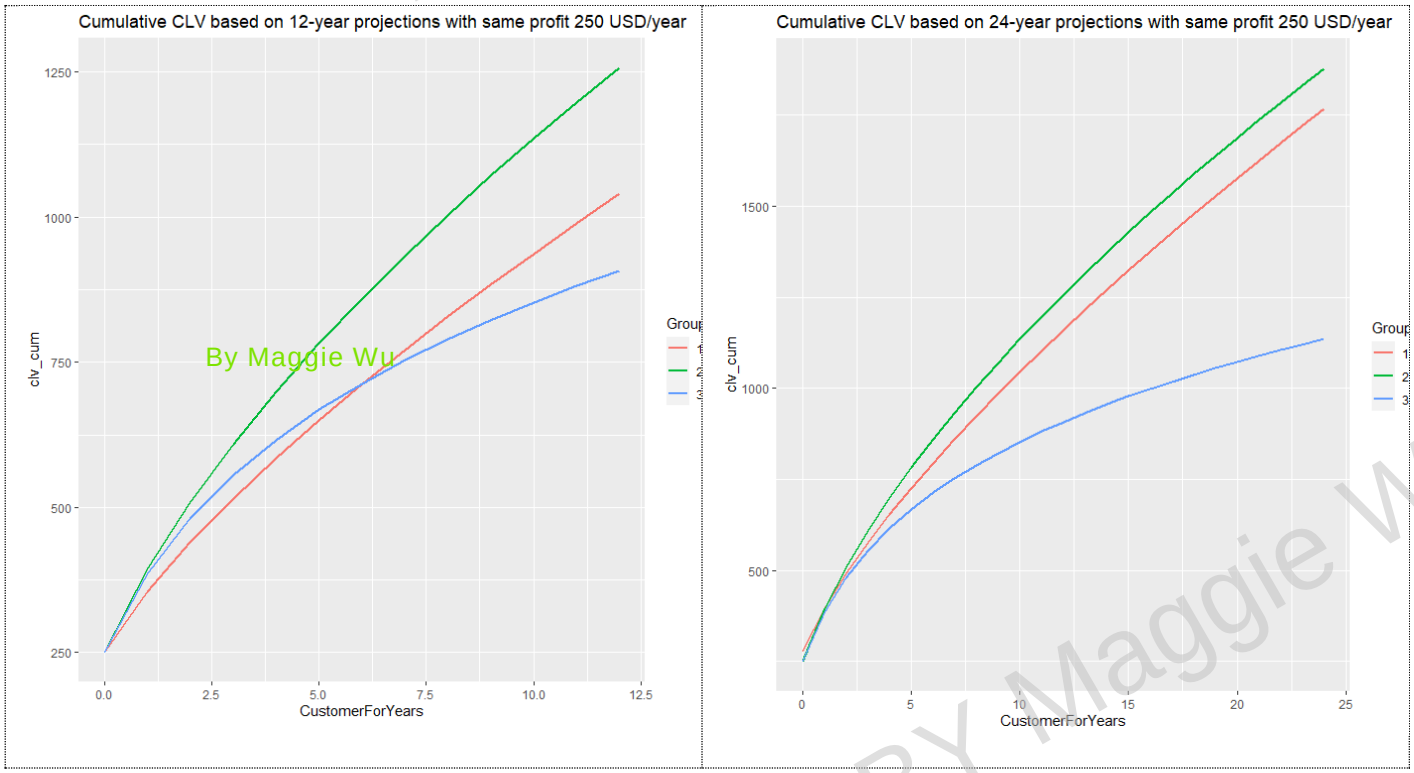
- ✓ Group 1 *(250 USD/year)* complements the product margin disadvantage through retention advantages after year 10 of purchase, with cumulative CLV exceeding that of Group 3 *(279 USD/year)*.
- ✓ Group 2 *(311 USD/year)* CLV has always maintained a profit advantage.



Verification for Assumptions 2.

I now remove the profit variance by setting all `profitAnnual` to 250$ yearly. By modifying both the "Profit " and “Year” parameter in the function, I also obtained plots for 12 (left-graph) and 24 years (right-graph) with the same `profitAnnual` CLV condition. Comparing the 3 Groups trends, I find under the same profit condition:

- ✓ The CLV value of Group 1 moves closer to Group 2 as time elapses (~ 24 years).
- ✓ Customer retention rate plays an absolute role to CLV at the result of this assumption.



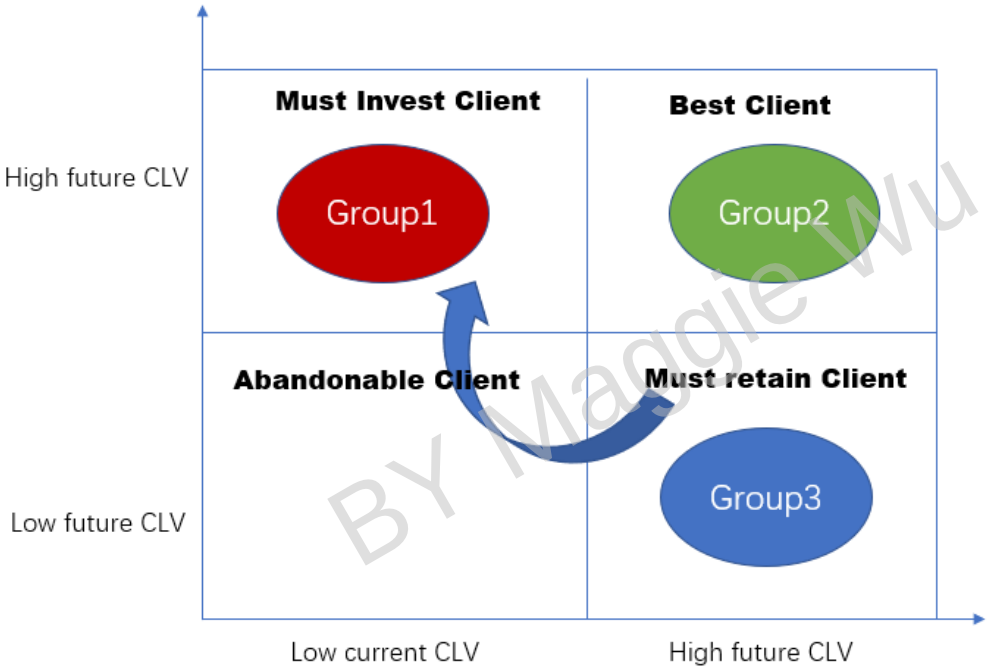
Note: Please refer to Appendix section 5 for codes covered in this Assumptions test.

Chapter 4. Conclusion

Profitability and customer retention are important factors for the company to derive long-term revenue from its customers. In particular, it is clear from the results of the **Assumptions testing in Chapter 3** that in long-term insurance products, customer retention rate would be a more effective indicator of the company's CLV.

Based on the insurer's original profit-driven marketing strategy, I believe that their current focus is on Groups 2 and 3 (because of higher profits). However, based on the CLV analysis, I recommend that insurers change their marketing strategy and focus more on CLV as a figure.On the one hand, company should increase the retention rate of Group 3 by using new marketing strategies, which turned out to be inefficient with their existing marketing strategies. On the other hand, company should focus more on Group 2 and Group3, which the groups with the higher retention and CLV. When product profit differences are removed, Group 1 has a more stable long-term customer retention rate than Group 2, which is higher in terms of customer loyalty value.

I have positioned the three groups of customers based on CLV values (as shown in the diagram), and insurers can apply different strategies of CRM strategies to the three groups of customers based on the current CLV projections.



According to the 80/20 principle, one should focus time on giving special attention to the most valuable customers. For example, by offering better service resources, fostering stronger relationship or giving more discounts in order to retain them for longer, for example, user interaction campaigns through Facebook, WeChat and other social media channel. For younger customers, insurers may try to retain them using more aggressive strategies, such as more positive customer relationship interactions and attention. For older customers, insurers may be able to offer them more discounts on premiums to entice them to stay on.

In conclusion, insurers **should use targeted marketing & CRM strategies for different Customer Clusters based on the CLV measures**. I also recommend that insurers **further analyze how can improve the customer lifecycle based on the attributes of their customer clusters**.

Thank you for reading this report!

*Below is the appendix code, you can find the coding process of the corresponding content in the appendix.

Appendix: Code (R language)

1. **Observation CAR INSURANCE data**

```
> library(reshape2)
> library(ggplot2)
> library(vioplot)
> library(dplyr)
> df_carIns <- read.csv(file.choose())
> str(df_carIns)
'data.frame':     24 obs. of    4 variables:
 $ X                          : int    1 2 3 4 5 6 7 8 9 10 ...
 $ CustomerForYears       : int    0 1 2 3 4 5 6 7 0 1 ...
 $ Group                      : int    1 1 1 1 1 1 1 1 2 2 ...
 $ RetentionRateonDec31: num    1 0.419 0.341 0.311 0.282 ...
# Cloning a data frame “df_ret1” for observational mapping
> df_carIns1 <- df_carIns
> df_carIns1$Group <- as.factor(df_carIns1$Group)
# Dram vioplot chart to present the overall difference between the three “RetentionRateonDec31” sets
> g1 <- df_carIns1$RetentionRateonDec31[df_carIns1$Group==1]
> g2 <- df_carIns1$RetentionRateonDec31[df_carIns1$Group==2]
> g3 <- df_carIns1$RetentionRateonDec31[df_carIns1$Group==3]
> vioplot(g1,g2,g3,names = c("Group1","Group2","Group3"),col="gold")
> title("Violin Plos of Customers RetentionRateon by Groups", ylab="RetentionRateonDec31")
# Dram ggplot chart to present the cliff difference between the three RetentionRateonDec31 sets
  > p2<-ggplot()+geom_line(data=
df_carIns1,aes(x=CustomerForYears,y=RetentionRateonDec31,group=Group,color=Group),size=1)
> p2
```

2. **sBG prediction function**

```
# Refer “7 CLV Predictions Using the sBG Approach in R” to construct the calculation function for sBG distribution
> churnBG <-Vectorize(function(alpha, beta, period) {
+        t1 = alpha / (alpha + beta)
+        result = t1
+        if (period > 1) {
+              result = churnBG(alpha, beta, period -1) * (beta + period -2) / (alpha + beta + period -1)}
+        return(result)
+ }, vectorize.args = c("period"))
> survivalBG <-Vectorize(function(alpha, beta, period) {
+        t1 = 1 -churnBG(alpha, beta, 1)
+        result = t1
+        if(period > 1){
+              result = survivalBG(alpha, beta, period -1) -churnBG(alpha, beta, period)}
+        return(result)
+ }, vectorize.args = c("period"))

> MLL <-function(alphabeta) {
+        if(length(activeCust) != length(lostCust)) {
+              stop("Variables activeCust and lostCust have different lengths: ",
+                    length(activeCust), " and ", length(lostCust), ".")
+        }     t = length(activeCust) # number of periods
+        alpha = alphabeta[1]
+        beta = alphabeta[2]
+        return(-as.numeric(
+              sum(lostCust * log(churnBG(alpha, beta, 1:t))) +
+                    activeCust[t]*log(survivalBG(alpha, beta, t)))))}
```

3. Retention rate prediction for the existing 7 years and predict the date within 12 years

```
# Using the Fader-Hardie functions to transfer original df_carIns to the new df_carIns_pre which contain the active customers num and lost customers num, ready to predict the 12 years RetentionRateonDec31
> df_carIns_pre <-df_carIns %>%group_by(Group) %>%
+       mutate(activeCust = 1000 * RetentionRateonDec31,
+                   lostCust = lag(activeCust) -activeCust,
+                   lostCust = ifelse(is.na(lostCust), 0, lostCust)) %>%
+       ungroup()
# create ret_preds01 for Group1 Retention Rate predication, ret_preds02 for Group2 predication, and ret_preds03 for Group3 predication.
> ret_preds01 <-vector('list', 7)
> for (i in c(1:7)) {
+       df_ret_filt <-df_carIns_pre %>%
+            filter(between(CustomerForYears, 1, i) == TRUE & Group == '1')
+       activeCust <-c(df_ret_filt$activeCust)
+       lostCust <-c(df_ret_filt$lostCust)
+       opt <-optim(c(1, 1), MLL)
+       retention_pred <-round(c(1, survivalBG(alpha = opt$par[1], beta = opt$par[2], c(1:7))), 3)
+       df_pred <-data.frame(CustomerForYears = c(0:7),
+                            Group = '1',
+                            fact_years = i,
+                            retention_pred = retention_pred)}
> ret_preds01[[i]] <-df_pred
> ret_preds01 <-as.data.frame(do.call('rbind', ret_preds01))
> ret_preds02 <-as.data.frame(do.call('rbind', ret_preds02))
> ret_preds03 <-vector('list', 7)
> for (i in c(1:7)) {
+       df_ret_filt <-df_carIns_pre %>%
+            filter(between(CustomerForYears, 1, i) == TRUE & Group == '3')
+       activeCust <-c(df_ret_filt$activeCust)
+       lostCust <-c(df_ret_filt$lostCust)
+       opt <-optim(c(1, 1), MLL)
+       retention_pred <-round(c(1, survivalBG(alpha = opt$par[1], beta = opt$par[2], c(1:7))), 3)
+       df_pred <-data.frame(CustomerForYears = c(0:7),
+                            Group = '3',
+                            fact_years = i,
+                            retention_pred = retention_pred)
+       ret_preds03[[i]] <-df_pred
+ }
> ret_preds03 <-as.data.frame(do.call('rbind', ret_preds03))
# Combine the original df_carIns to ret_preds01, ret_preds02 and ret_preds03
> ret_preds <- bind_rows(ret_preds01, ret_preds02, ret_preds03)
> df_carIns$Group <- as.character(df_carIns$Group)
> df_carIns_pre_all <- df_carIns %>% dplyr::select(CustomerForYears, Group, RetentionRateonDec31) %>% left_join(., ret_preds, by = c('CustomerForYears', 'Group'))
> head(df_carIns_pre_all)
  CustomerForYears Group RetentionRateonDec31 fact_months retention_pred
1                0     1               1.0000           7          1.000
2                1     1               0.4195           7          0.420
3                2     1               0.3405           7          0.340
4                3     1               0.3115           7          0.302
5                4     1               0.2825           7          0.277
6                5     1               0.2635           7          0.260
# Draw ggplot to show the Retention Rate predication between 3 groups by 7-years data..
> ggplot(df_carIns_pre_all , aes(x = CustomerForYears, y = RetentionRateonDec31, group = Group, color= Group)) + theme_minimal() +      facet_wrap(~ Group) +    geom_line(size = 1.5) +geom_point(size = 1.5) + geom_line(aes(y = retention_pred, group = fact_years), alpha = .5) +    ggtitle("Retention Rate Projections")
```

Repeat the Retention rate prediction program for 12-years data and further to 24-years data

*# Repeat the procedure above for creating the prediction data, except that **the predictor** (fact_years **)7 is changed to 12.***
 # For the sake of space, I will not paste the code here -Take ret_preds03 for Group3 for an example as bellow pic.

```
> ret_preds03 <-vector('list', 12)
> for (i in c(1:12)) {
+     df_ret_filt <-df_ret %>%
+         filter(between(CustomerForYears, 1, i) == TRUE & Group == '3')
+     activeCust <-c(df_ret_filt$activeCust)
+     lostCust <-c(df_ret_filt$lostCust)
+     opt <-optim(c(1, 1), MLL)
+     retention_pred <-round(c(1, survivalBG(alpha = opt$par[1], beta = opt$par[2], c(1:12))), 3)
+     df_pred <-data.frame(CustomerForYears = c(0:12),
+                          Group = '3',
+                          fact_years = i,
+                          retention_pred = retention_pred)
+     ret_preds03[[i]] <-df_pred
+ }
> ret_preds03 <-as.data.frame(do.call('rbind', ret_preds03))
```


 # Draw ggplot to show the Retention Rate predication between 3 groups by 12-years data
> ggplot(ret_preds, aes(x = CustomerForYears, y = retention_pred, group = Group, color= Group)) + theme_minimal() + facet_wrap(~ Group) + geom_line(size = 1) +geom_point(size = 1) + geom_line(aes(y = retention_pred, group = fact_years), alpha = .5) + ggtitle("Retention Rate Projections with 12-years")

*# Repeat the procedure above for creating the prediction data, except that **the predictor** (fact_years **) to 24.***
 # Draw ggplot to show the Retention Rate predication between 3 groups by 24-years data
> ret_preds_24 <- bind_rows(ret_preds01_24, ret_preds02_24, ret_preds03_24)
> ggplot()+geom_line(data= ret_preds_24,aes(x=CustomerForYears,y=retention_pred,group=Group,color=Group),size=1)+ ggtitle("Retention Rate Projections with 24-years")

4. Cumulative CLV based on Retention rate prediction

*# Repeat the procedure above for creating the prediction data, except that **the predictor** (fact_years **)7 is changed to 12.***
 # Calculate Car insurance profit with Group1 CLV with profit of 250 USD/year

```
> df_clv_01 <- df_carIns_pre %>%      filter(between(CustomerForYears, 1,2) == TRUE & Group == '1')
> activeCust <- c(df_clv_01$activeCust)
> lostCust <- c(df_clv_01$lostCust)
> opt <- optim(c(1,1), MLL)
> retention_pred <- round(c(survivalBG(alpha = opt$par[1], beta = opt$par[2], c(3:12))), 3)
> df_pred <- data.frame(CustomerForYears = c(3:12), retention_pred = retention_pred)
> df_clv_01 <- df_carIns_pre %>%
+       filter(between(CustomerForYears, 0, 2) == TRUE & Group == '1') %>%
+       dplyr::select(CustomerForYears, RetentionRateonDec31) %>%
+       bind_rows(., df_pred) %>%
+       mutate(RetentionRateonDec31_calc = ifelse(is.na(RetentionRateonDec31), retention_pred,
+                                                 RetentionRateonDec31),
+              clv_yearly = RetentionRateonDec31_calc * 250,    # profit of 250 USD/year
+              clv_cum = round(cumsum(clv_yearly), 2))
> df_clv_01$Group = "1"
> df_clv_01
```

```
# A tibble: 13 x 7
   CustomerForYears RetentionRateonDec31 retention_pred RetentionRateonDec31_calc clv_yearly clv_cum Group
              <int>                <dbl>          <dbl>                     <dbl>      <dbl>   <dbl> <chr>
 1                0                1             NA                         1          250     250  1
 2                1                0.420         NA                         0.420      105.    355. 1
 3                2                0.340         NA                         0.340       85.1   440  1
 4                3               NA              0.302                     0.302      75.5    516. 1
 5                4               NA              0.278                     0.278      69.5    585  1
 6                5               NA              0.261                     0.261      65.2    650. 1
 7                6               NA              0.247                     0.247      61.8    712  1
 8                7               NA              0.237                     0.237      59.2    771. 1
 9                8               NA              0.228                     0.228      57      828. 1
10                9               NA              0.221                     0.221      55.2    884. 1
11               10               NA              0.214                     0.214      53.5    937  1
12               11               NA              0.208                     0.208      52      989  1
13               12               NA              0.203                     0.203      50.8   1040. 1
```