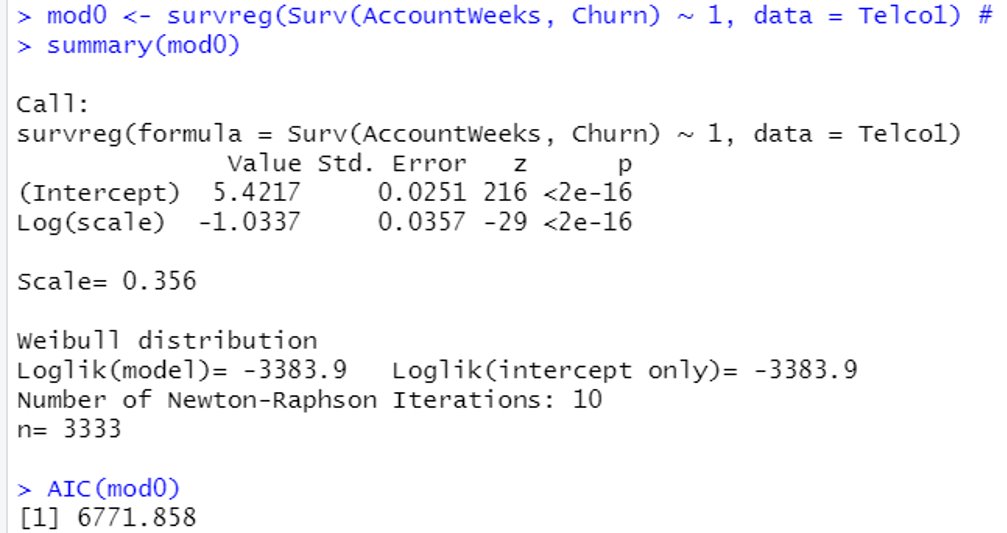
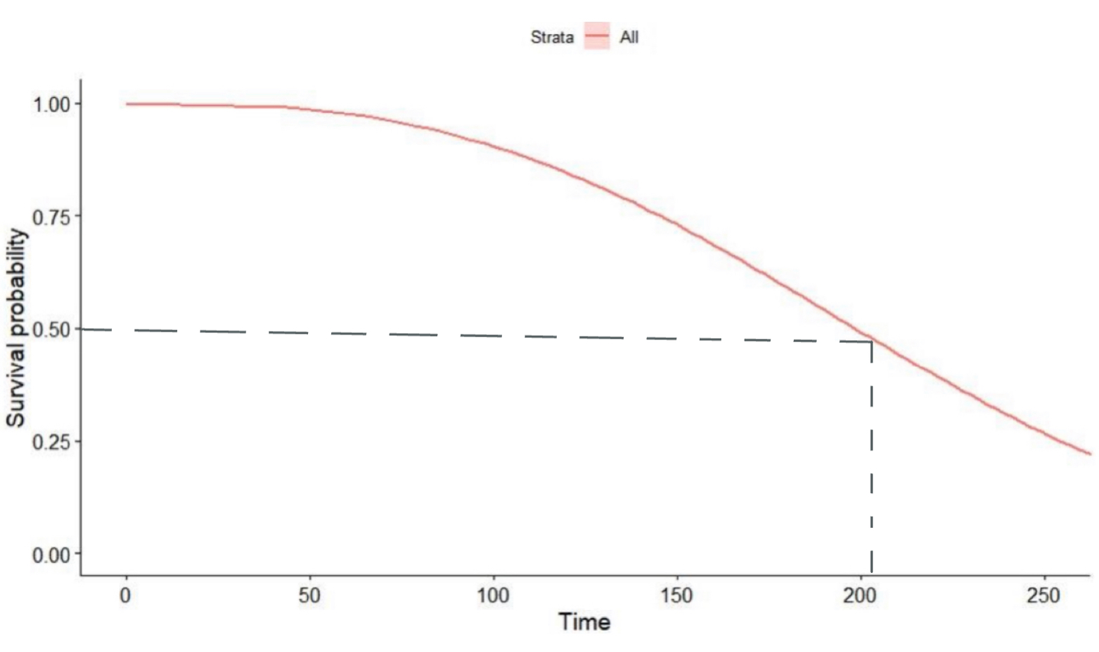
**Task 1) Estimate a base survival model for an average customer. Call this mod0. Please provide the model summary and visualize the survival curve (screenshots suffice; when visualizing the survival curve use breaks of 50 so that the x-axis captions can be clearly read).**

Model 0’s summary is shown in (figure 1) we compared model fitness between models looking at log-likelihood and AIC. For model 0 log-likelihood is –3383.9 the higher this number the better and AIC is 6771.858, the lower this number the better in model performance. It is a baseline form that we can improve by adding explanatory variables.

The survival curve of the basic model 0 without any explanatory variable (figure 2) illustrates the probability for customers to stay with the company for a certain period. For example, in model 0 there is a probability of 50% of customers can survive (not churn in our case) until 200 weeks.



*(Figure 1. Summary of model 0)*

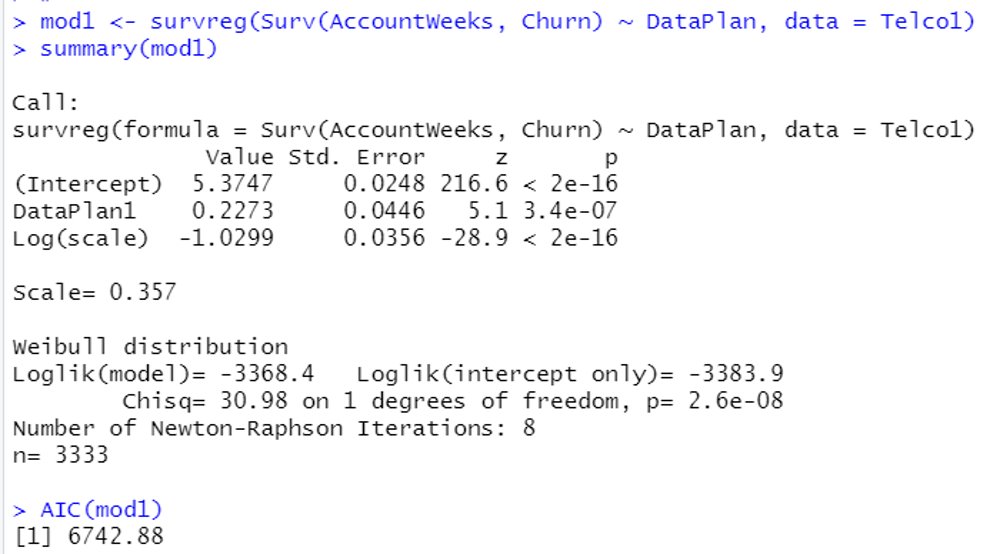


*(Figure 2. Survival curve for model 0)*

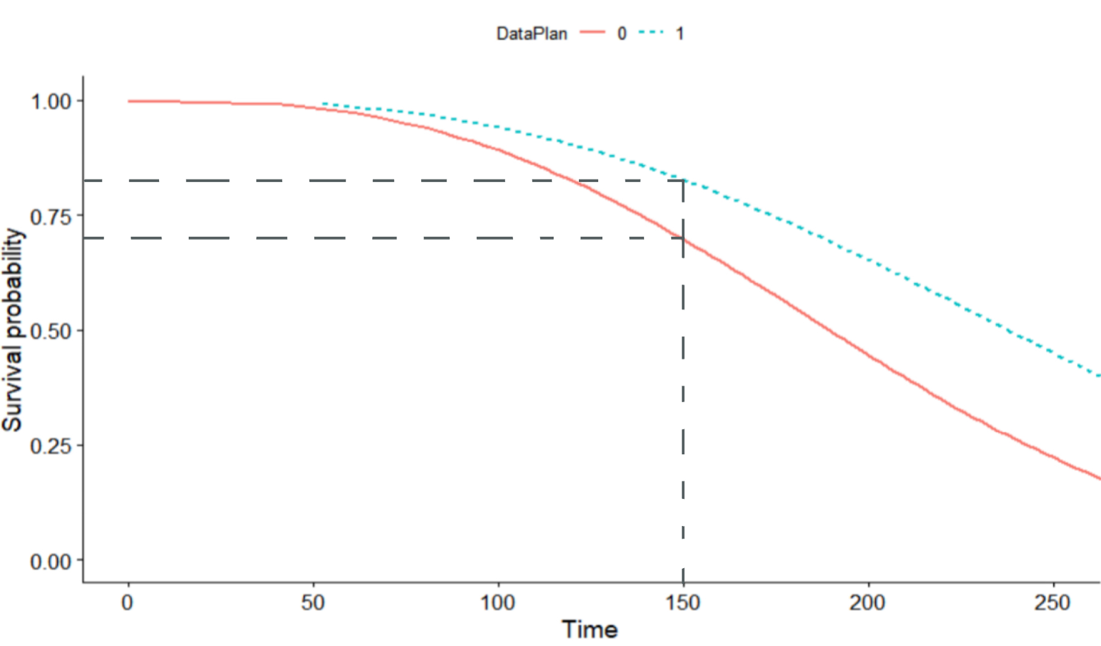
**Task 2) Please estimate a model that includes DataPlan as an explanatory variable. Call it mod1. Please provide the model summary and visualize the survival curve (screenshot suffices).**

Model 1 contains one explanatory variable “Data Plan”. The local performance for model 1 is the significance of the p-value for Data Plan is smaller than 0.05 (figure 3), which indicates the independent variable “Data Plan” has a significant impact on the dependent variable “churn”. Furthermore, the Chi-square test has a P-value < 0.05. For further estimation with other models, Log–likelihood for model 1 is –3368.4 and for AIC it is 6742.88. In summary, the global performance of model 1 can be used for further analysis.

The survival curves of model 1 (figure 4) show the survival probability for customers with and without Data Plan. The red line is “DataPlan=0” which means for customers who do not have Data Plan and the blue line stands for “DataPlan= 1” which means for customers who do have Data Plan. The shape of the curve illustrates that there is an obvious difference in the probability of “0” and “1” because the differences become larger follow with time period. For customers who can survive up to 150 weeks, the probability for those with Data Plan is approximately 78% and for those without Data Plan is 73%.



*(Figure 3. Summary of model 1)*

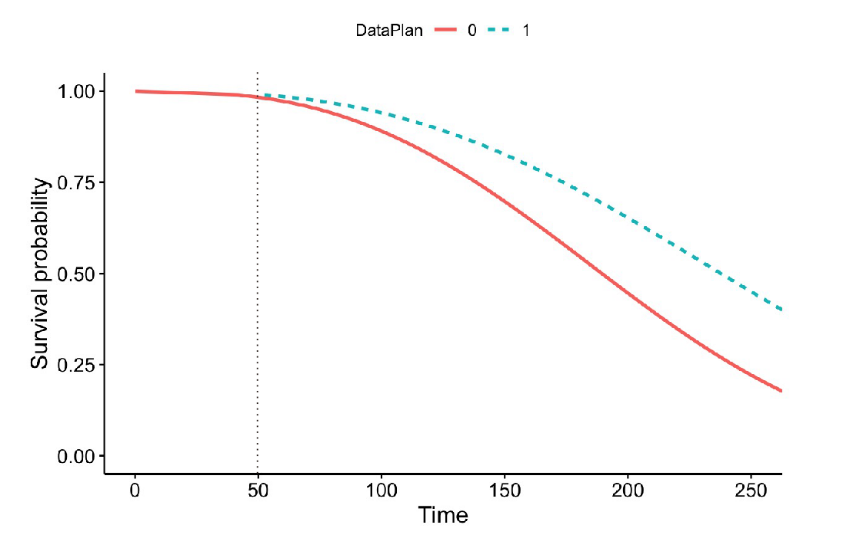


*(Figure 4. Survival curve for model 1)*

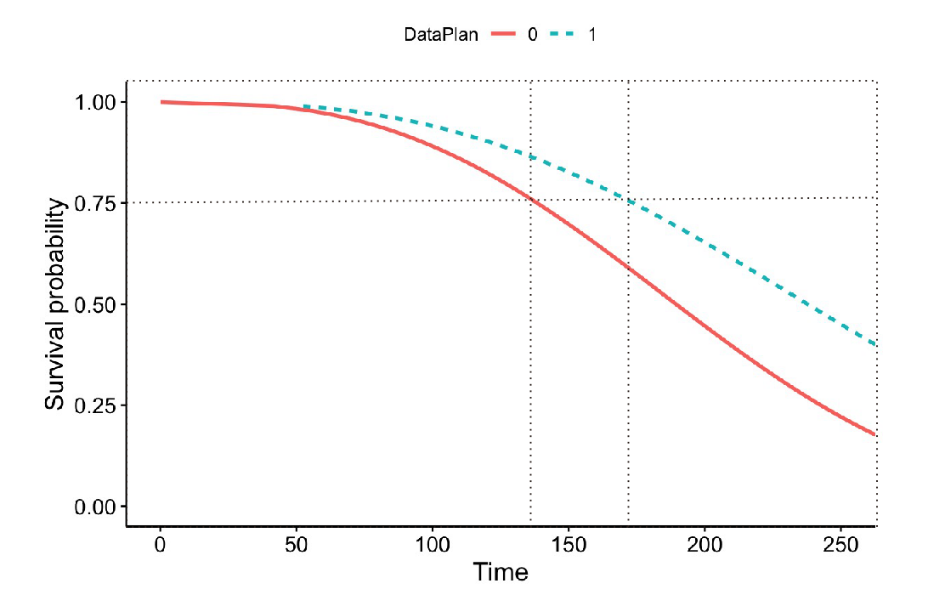
**Task 3) You decide to move on with mod1. Critically evaluate the predicted curve. Do you see any reason for concern?**

The impact of Data Plan on survivability is negligible until 50 weeks, and the consumer retention rate (not churn) is close to 100% (figure 5). The retention rate of consumers with Data Plan decreases more slowly than the retention rate of consumers without Data Plan until 75%. At this point, the retention time for consumers without Data Plan is 140 weeks, and the retention time for consumers with Data Plan is 170 weeks (figure 6). When the survival probability is 50%, the retention time is 190 weeks for consumers without Data Plan and 230 weeks for consumers with Data Plan. The downward trend is similar in both cases (figure 7).

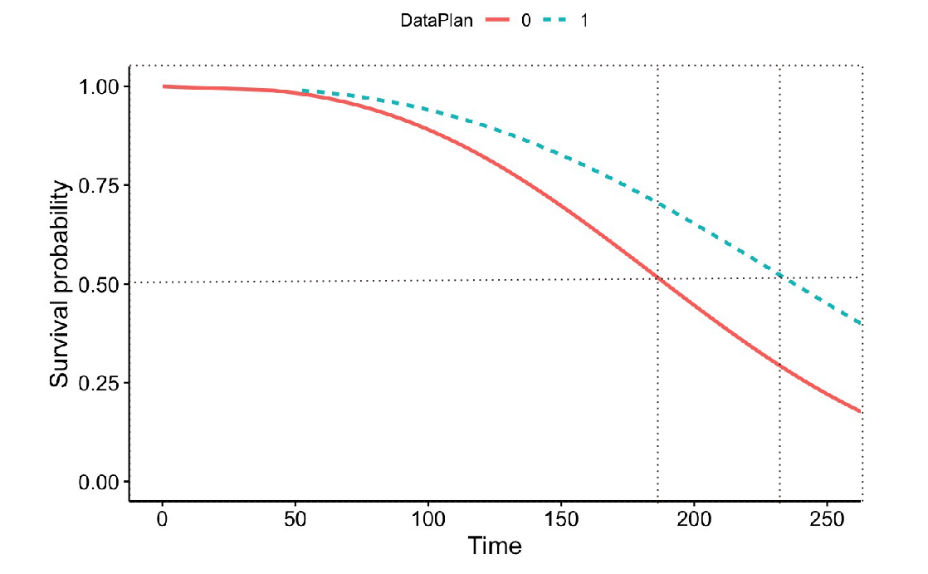
However, Model 1 only has a data plan as the explanatory variable, which couldn’t reflect the influence of other elements on the model, and the precision is not enough, thus may lead to inaccurate results.



*(Figure 5. Survival curve for model 1)*



*(Figure 6. Survival curve for model 1)*



*(Figure 7. Survival curve for model 1)*

**Task 4) You want to use the model to make predictions of survival probabilities to improve the targeting of existing customers (e.g., which customers should be offered promotions etc.). Do you see a chance to improve model performance given the data at hand? Please explain your answer.**

1. Since in model1 we had not enough model variables, and we found that there were no meaningful variables at all, so we decided to select all the variables for calculating in model2.

2. Model2 has a better loglikhood than Model1, which is –3233.4. And there are many significant variables, so they can be useful for further analysis. But there are several relatively large VIFs, so we decided to remove the **monthly charge** with the largest VIF (more than 2800) for the next step of the model calculation.

3. The loglikhood of Model3 is better than model2 after removing the monthly charge, and the variables are also significant. However, there are still several large VIFs, so we decided to remove the largest datausage for the next step of the model calculation, which is close to 15.

4. After removing datausage, both loglikhood and AIC of model4 are better than model3 (loglikehood is larger and AIC is smaller), and the variables are also significant, and there are no more particularly large VIFs. But we still try to remove the largest custservicecall for the next step of model calculation.

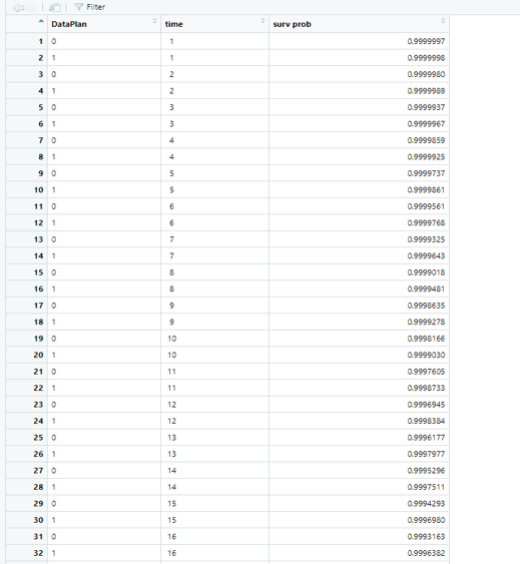
5. The VIF of model5 after removing custservcall does not have particularly large variables and is simpler than model4. But its AIC value is larger than model4, 6586.742 vif for Model 5 and 6483.009 vif for model 4. Because smaller VIC would be better, model5 is discarded and model4 is chosen.

**Task5) You decide to use mod1 to calculate the expected CLV for customers who have a data plan and those who do not have a data plan. The annual interest rate is 5% (note that you have to translate this into weekly discount rates by dividing it by 52). For an assumed customer lifetime of 500 weeks, please calculate the CLV and the probability corrected CLV for both customer types.**

CLV aims to determine the long-term customers’ value in the whole relationship with the firm whether then focus on their short-term performance. It was calculated using model 1 which had been analyzed in previous task, we distinguish the CLV into two customers prototypes, first is customers with data plan and second is customers without a data plan. The use of period is in the range of 500 weeks and has an annual interest rate of 5%, because the use of time period is count in week, so it needs to be translated into weekly discount rates by dividing the annual interest rate 5% by 52, which is 0.096%.

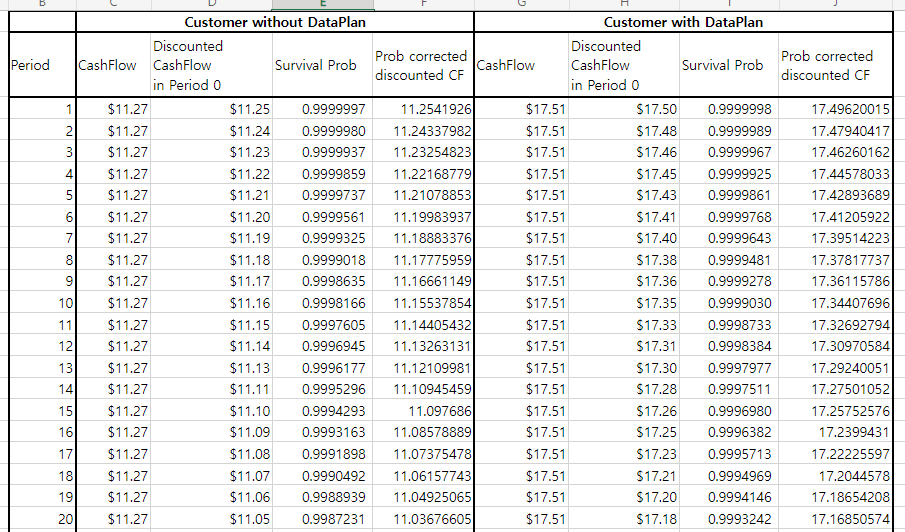
1. We derived weekly average cash flow first. The number of customers without data plan is 2411, and the number of customers with data plan was 922. Therefore mean of monthly charge who have data plan was 48.81501 and mean of monthly charge who don’t have data plan was 75.89165. Average weekly cash flow is 11.27, 17.51 for each group of customers that was derived from dividing mean of monthly charge with 12/52.

2. Then we derived the predicted survival probabilities for each customer using R. (figure 8)



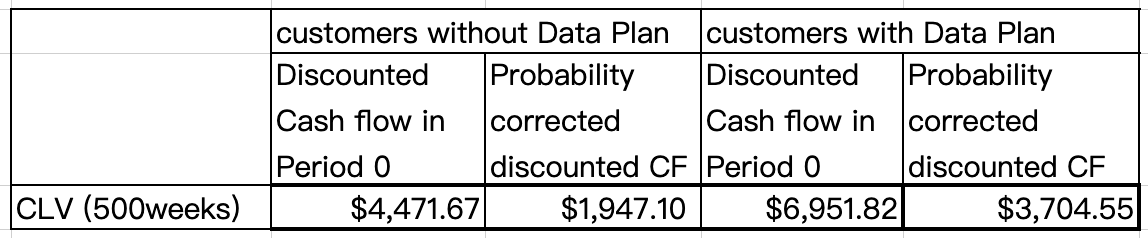
*(Figure 8. the first 20 rows of survival probabilites )*

3. We calculated CLV and probability corrected CLV using excel. (figure 9)



*(Figure 9. the first 20 rows of the results of* CLV and probability corrected CLV calculation*)*

4. The total discounted cash flow for customers without a data plan is 4471.67 dollars (figure 8), and the probability corrected discounted cash flow that takes all period’ s survival probabilities into account is 1947.10 dollars. Second, looking at customers who have data plans the discounted cash flow is 6951.82 dollars and the probability corrected discounted cash flow is 3704.55 dollars. In the comparison of these two prototypes, the overall performance of discounted cash flow of the first prototype (customers with a data plan) is about 1757 dollars more than Second prototype (customers without a data plan). So according to CLV calculation, our firm should focus more on customers that have a data plan because they can provide the firm with more returns in each future period.

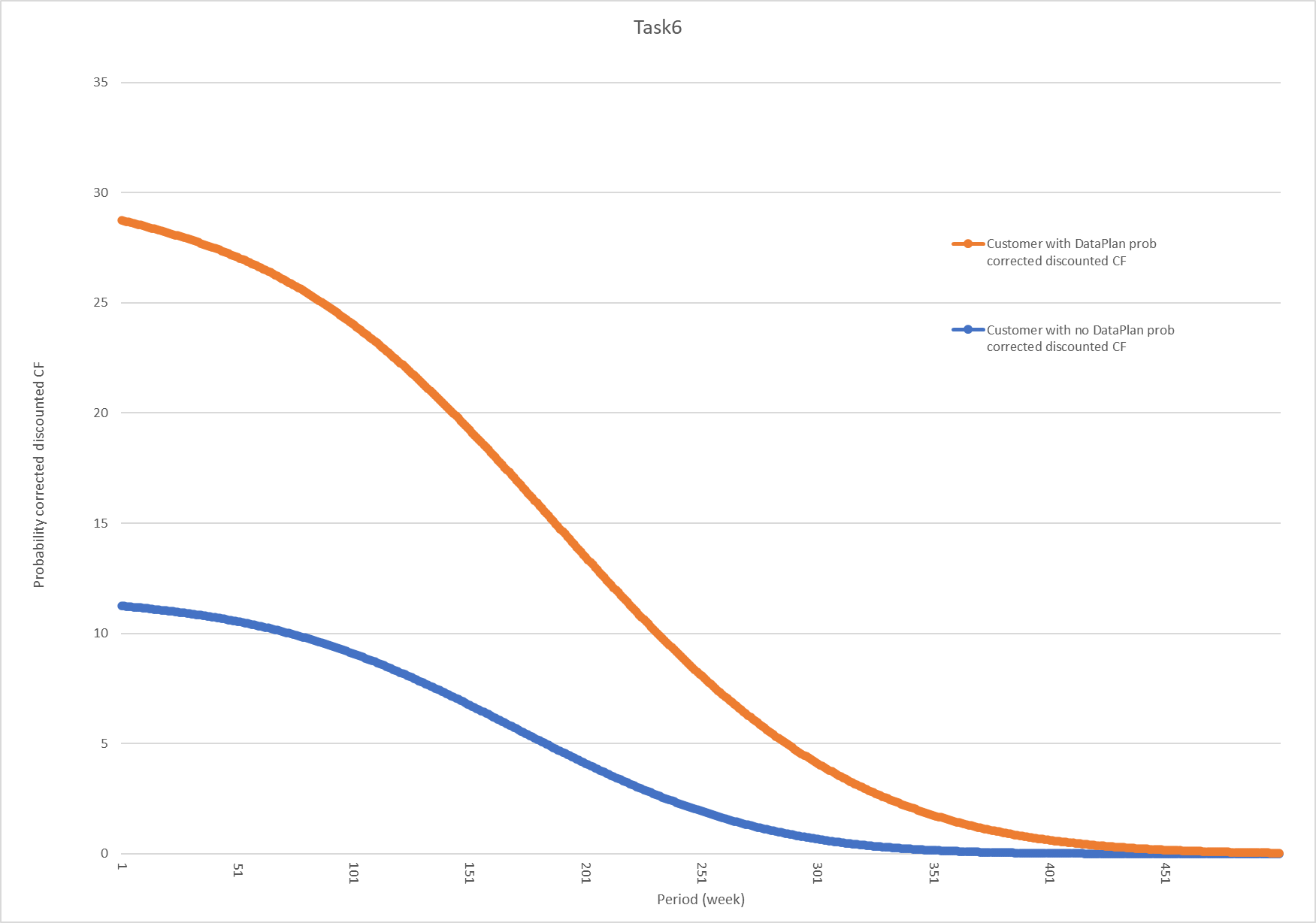


*(Figure 8. CLV calculation for model 1)*

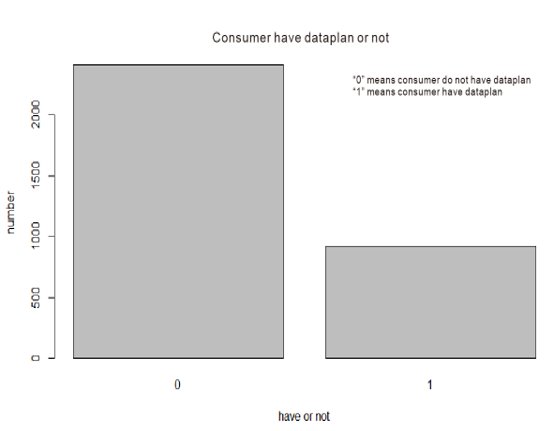
**Task6) Please present a simple visualization that demonstrates your key insight from the probability corrected CLV to managers. (It is easiest to use PowerPoint to provide an appropriate chart.)**

We first plotted our results as a visual line graph, as shown in figure 10, blue shows the expected probablity discount cash flow for consumers without dataplan and red shows the expected probablity discount cash flow for consumers with dataplan. We can see the difference between two line is relatively large until 400 weeks and close to convergence after 400 weeks. We can see that the red line has a consistently higher cashflow than the blue one. Also, consumers with dataplan tend to have more cash flow, by categorising consumers according to whether they have a dataplan or not, we can see that more consumers do not have a dataplan, which is about 2.6 times more than those who do (figure 11). So companies need to help consumers to set up a dataplan in order to help them to be more loyal and increase the consumer CLV. For example, company can encourage consumer to build a dataplan at the begin of the contracts. Or the company can give some award to consumer whom make dataplan, the cost will lower than the benefit so it should be a success method.

As the model only takes into account the relevant scenarios for dataplan, companies will still need to improve the results in the future for other factors that may affect consumer CLV such as Gigabytes of average monthly data usage ，Average daytime minutes (calling time) per month, Average number of daytime calls ， Average monthly bill， Largest overage fee in the last year, Average number of roaming minutes And Whether or not a customer has called the service line more than once  
  
In conclusion, the problem of company is to improve long-term revenue and determine how to use different marketing approaches for different users, in this way, we use survival model to show that dataplan is a Explanatory variables of the relation between time and churn, we found that consumer have dataplan will more likely to stay and has higher CLV, so the company can encourage consumer to build the dataplan, and we build another model to improve it based on model one, we found the explanatory variable can be 6 in model 4, and company can use it to do the furthur calculate.



*(Figure 10. line graph of probability corrected discounted CF)*



*(Figure 11. The number of customers with data plan and without data plan)*