**Understanding Churn in Telecom**

Lining Jiang, Wenying Hu, Pooshan Shah

**Contents**

1. Introduction: Background and Problem…………………………..2
2. Data Summary and EDA …………………………………………..3
3. Data Analysis and Key Findings and Conclusions………………….7
4. Marketing Strategy Recommendations, Limitations, and Future Research Directions……………………………………………...10
5. **Introduction: Background and Problem**

In this project, we will be looking at a Telecom company’s churn and retention rate. A telecom company is an organization that provides voice or data transmission services. The major segments within telecommunications are wireless and wired communications, long-distance/international services, satellite communications, etc. Telecommunications may sound like an outdated notion, but in fact these companies are responsible for our day-to-day interactions and communications. In fact, the largest telecom companies today are Verizon, followed by AT&T, China Mobile, Vodafone, etc. Telecommunication is a highly used and important tool for business-- it provides a means of effective communication with customers and allows for greater levels of customer service. With the increase population of video conferencing, video calls, and other forms of communication that bypass telecom companies, it is more important now than ever to better understand how telecom companies can retain their customers and continue to prosper as a business.

The data we are looking at was provided by Kaggle.com. This dataset contains information of over 3300 customers which include the state they are from, account length, area code, whether or not they have an international plan, whether or not they have a voicemail, number of voicemail messages, total day/evening/night minutes used, total day/evening/night calls performed, total day/evening/night charges based on these calls and minutes, total international minutes, calls, and charges, and whether or not that customer churned. By diving deep into this data, we can get a better understanding of how this telecom company’s customers are using and being charged for certain services. Based on this, we can see how these services affect whether or not the customer stayed with the company and we can also look into how each state in the US is related to each of the services and churn differently.

We will start by exploring the surface of this data: we do this by performing logit, probit, and continuous hazard models to the data to get a better understanding of which variables are significant in determining churn or not and with the hazard models, how long before we can expect a customer to churn.

This research is important to understand which of the provided services and informational pieces impact the churn rate the most. As technologies and changing and competition is increasing quickly, it’s important for this telecom company to use its data to understand how to achieve its goals and stay afloat in the market. Depending on the goal of the company- whether it is to retain as many current customers as possible, find new customer segments to market to, understand how to bring back previous customers, or even just to see how certain services are affecting their business, they can use our research accordingly. From here, the telecom company can find their weaknesses and strengths and play to them in the market. They can use our analysis as evidence and direction of which customer segment to market to more and where to spend more resources in order to increase customer retention or just profits in general.

Let’s jump in to our exploratory data analysis and see what kind of patterns, trends, and similarities/differences we find.

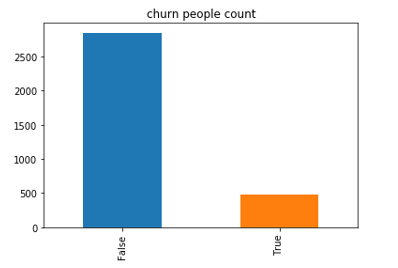
1. **Data Summary and EDA**

This dataset records the information of customers of a Telecom company.

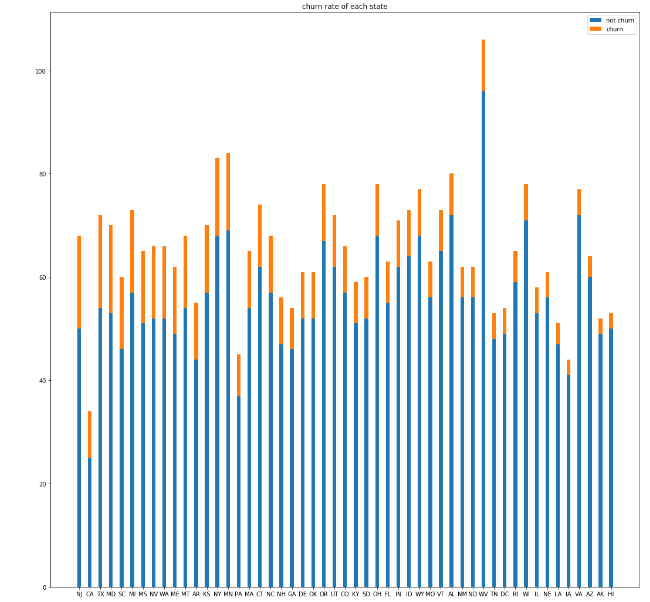
There are total 3333 rows and 21 variables which are: [[1]](#footnote-0)

* state: the state the user lives in
* account length: the number of days the user has this account
* area code: the code of the area the user lives in
* phone number: the phone number of the user
* international plan: true if the user has the international plan, otherwise false
* voicemail plan: true if the user has the voicemail plan, otherwise false
* number vmail messages: the number of voicemail messages the user has sent
* total day minutes: total number of minutes the user has been in calls during the day
* total day calls: total number of calls the user has done during the day
* total day charge: total amount of money the user was charged by the Telecom company for calls during the day
* total eve minutes: total number of minutes the user has been in calls during the evening
* total eve calls: total number of calls the user has done during the evening
* total eve charge: total amount of money the user was charged by the Telecom company for calls during the evening
* total night minutes: total number of minutes the user has been in calls during the night
* total night calls: total number of calls the user has done during the night
* total night charge: total amount of money the user was charged by the Telecom company for calls during the night
* total intl minutes: total number of minutes the user has been in international calls
* total intl calls: total number of international calls the user has done
* total intl charge: total amount of money the user was charged by the Telecom company for international calls
* customer service calls: number of customer service calls the user has done
* churn: true if the user terminated the contract, otherwise false

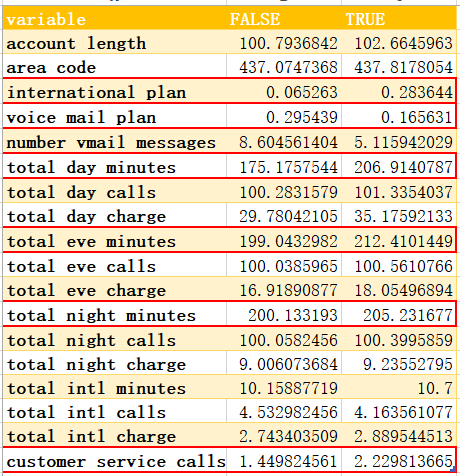
Our mission is to use this dataset to predict if the customer will churn or not and definitely, *churn* is our dependent variable and it is a classification problem.



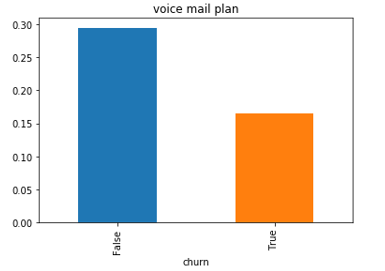
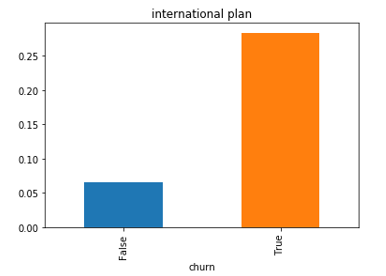
In the dataset, there are total 483 people churn and the churn rate is about 14.49%. If we group people by state that they lived in, we can see that the state with highest churn rate are NJ and CA, 26.47%, the state with lowest churn rate is HI, only 5.67%. However, we also found that the number of people in CA is the least. The following plot shows the churn distribution by state. And it is already ordered by churn rate from left to right.



By group dataset by churn or not, we calculated the mean value of each variable. The result is:

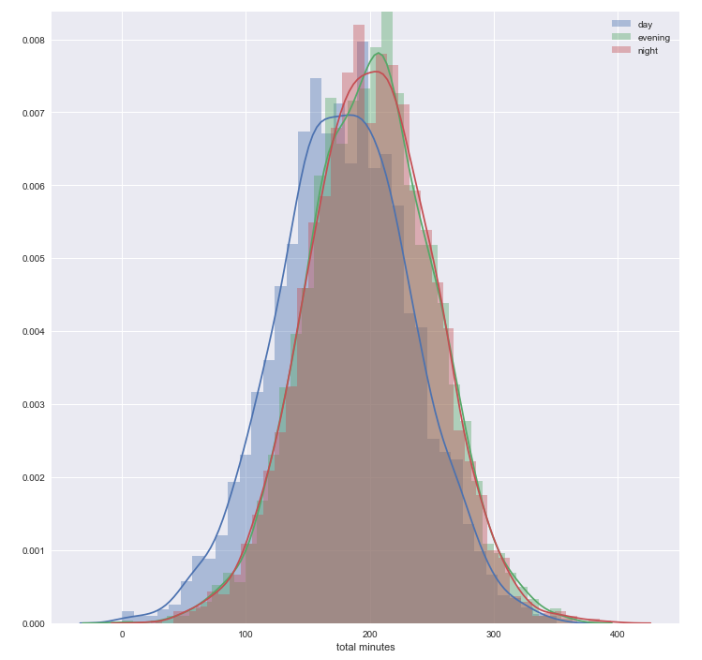


From the table, we can see that the variables with significant difference between churn or not are: international plan, *voicemail plan,number vmail messages,total day minutes,total eve minutes,total night minutes* and *customer service calls.*

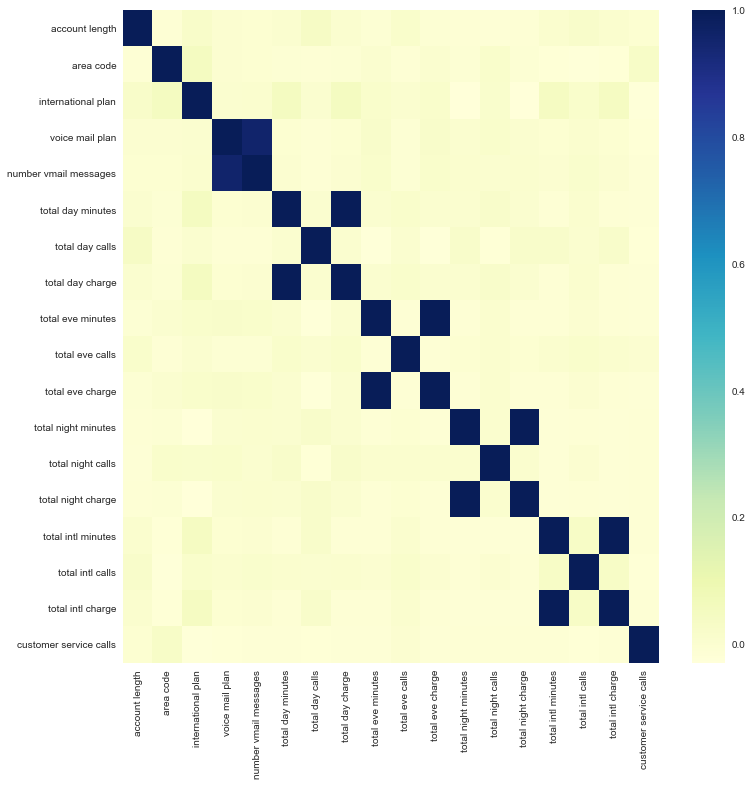
**

It is interesting to see that the churned people prefer to open international plan but not voicemail plan. In addition, their calling minutes are longer but the number of calls are not different. In the other words, it means that average minutes of each calling of these people are longer than people who do not churn.

Additionally, we also noticed that the distributions of number of calls, total minutes and charge in both day,evening and night follow normal distribution. For example, we take total minutes for day, evening and night and the distribution is shown below:



We also plot the correlation between variables. It is not surprised to see that some variables are strongly correlated, such as total minutes and total charge. If a people calls more, definitely he will spend more. So we will drop these variables before modeling.



1. **Data Analysis and Key Findings and Conclusions**

The three main statistical models we used are logistic/probit regression with fixed effect on the region, Hazard models to extract the significant features that impact the attrition of customers. As mentioned above, as the time spent on the phone is directly related to the charges on the phone. In the modeling part, we dropped the total minutes spent feature for day, evening, night and international calls. Besides, we also dropped area code and customer phone number as those are irrelevant variables to our project goals.

One of our objective is to find out the features that impact the probability of customer churning. The data that we were dealing with is the binary data, which allows us to use logistic regression and probit regression to predict the binary class that customer will belongs to and we used AIC to compare models.

* Logistic and Probit Regression

In this section, we tried logistic and probit regression without including state as a variable, and the two methods with fixed effect model on state. It turns out that the simple logistic regression actually perform the better. The AIC adds penalty to the increase in the number of predictors in the model. Therefore, it makes sense that when we add more parameters, (which might not be significant for the model) the AIC increases.

We were able to try general linear model with probit link and logit link (see the summary of the model from appendix figure 1 and figure 2). For simple logistic regression and probit regression. The AIC doesn’t vary that much. For logistic regression, the AIC is 2187.80, whereas for probit regression the AIC is 2187.83. The interpretation of the variables from logistic regression is as follow: (result in figure 1), the significant (p-value smaller than 0.05 significance level) variables are intercept, international plan, voicemail plan, number of voicemail messages, total day charges, total evening charges, total night charges, total international calls, and total international charge, and customer service calls. Within those variables international plan, total day charge, total evening charge, total night charge, and total international charge and customer service call has a positive effect on the churn rate. That means if any of the variables that are mentioned above increase, the probability of a customer to be churn will also increase as well. This makes sense as we could see that most of the variables with positive relation with the churn are prices. As the prices or the service charges increase, the customer are more likely to leave the business. From figure 2, we could also see that significant variables are the same as those in the logistic regression.

When we add the fixed effect of state in the logistic regression model (see figure 3). The significant variables we’ve mentioned are still significant. This time, we add state as fixed effect, so we could see which state has more effect on the probability of churn. Figure 3 provides the level of significance of the state variables. The significant variables are CA, MI, MT, NJ, SC, TX, WA. Furthermore, those variables all have positive coefficients, meaning that customer in those states are likely to leave the business. The AIC of this model is 2200.27. As we stated before, adding variables incurred penalty when we measure the fit of the model using AIC.

Therefore, we can conclude that using the simple logistic regression generate the best fit for the model. The customers’ probability of churn is highly correlation with the charges of the service, and if the customer book the voicemail plan they are less likely to leave the telecom company. However, we can still get some insights from the fixed effect model. For states that are significant in the model, while having a positive coefficient. The customer in those region are more likely to cancel the service. One can possibly guess that the signal in those regions might not as strong as other states or the customer service in those regions less satisfying.

* Hazard Analysis

As this data contained a churning factor, as well as time data, we decided to fit a hazard model to the data to get a better understanding of how people leave/stay with the telecom company. The hazard model would tell us about after how many days or account length time periods would be a customer’s probability of churning. We first tried a regular hazard model, but given the nature of our data set and the limited knowledge of its background, we could fit a discrete model to it. We then decided to fit a continuous hazard model to the data using flexsurv regression. But before we did that, we wanted to understand how the various variables affected churn on their own. So in Figure 3, we see how the probabilities differ for survival when everything is kept constant except for whether or not the customer has a voicemail plan. We see that with a voicemail plan, the lowest survival probability is only around 75% while the lowest survival probability for no plan is less than 25%. That is quite a huge difference. This goes to show that those with voicemail plans are probably planning ahead to be in the plan for the long haul. So one way to maybe increase retention would be to make it much easier or even mandatory to set up the voicemail. In Figure 4, we see the same graph but with whether or not the customer has an international plan or not. Surprisingly, after around 200 days, the probability of surviving WITH an international plan is around 0% while without the international plan, the probability is well above 50%. This leads us to believe that either the international plan is not very competitive or just these kind of customers kinds to leave their plans regularly. But given the data, I would assume that either the charge, quality, or service of the plan is not great. Another aspect we wanted to explore was total number of minutes used by the customers and how that affects survival. We split the minutes into 5 group- 5 being the highest and 1 being the lowest amount. As we can see in Figure 5, only group has a significant difference in survival. Group 5 has total minutes over 675 and these are the people that are churning more often-- so those who use a lot of minutes are probably over charged. For this case of the company, we would need to dive in deeper with more data to figure out how much these many minutes cost the company versus less minutes. Now, moving on to the actual survival model using flexsurv regression model which takes into account continuous data. In Figure 6, we can see the summary output for the ‘weibull’ distribution and Figure 7 has the summary output for the ‘llogis’ distribution. In both of these, we would be able to plot the probability of survival with the time. They are about straight lines with drops in each 10 units of time. In both models, we see that at around 232 minutes is when the churn probability goes to around 50%. So this may be the time period where the company can start remarketing to those customers and give deals/points/programs etc. to increase their probability of staying.

1. **Marketing Strategy Recommendations, Limitations, and Future Research Directions**

* Recommendations

From our analysis on the logistics/probit regression and exploratory analysis phase, we have the following recommendations to the business:

1. We’ve seen that the coefficient of the voicemail plan is negative indicating that people who uses this service are likely to stay with the business. Although we cannot conclude that this service causes the decrease in churn. But increasing the number of people who book this service will definitely help the business for customer retention. For example, the telecom company can offer free voicemail plan for a certain period or offering a binding services to the customer.
2. From our analysis on the fixed effect. The states that have positive coefficients are CA, MI, MT, NJ, SC, TX, WA. The positive coefficient means that customer in those regions are likely to leave the business. This might caused by the insufficient services such as not enough physical store in those regions and so the customer could not get access to support or services that they want to get. There might also be the reason that the signal in those regions are not strong. Actionable items that this telecom company could do is to look at whether the stores could provide enough services to the customers in those area. Furthermore, enhance the telecom services in those regions. If the cost of the second recommendation is too high. The business could try close some of the services in certain region to prevent any future losses.
3. Since we noticed that the total calling minutes for people who churned is higher than people NOT, it is a good idea to provide discounts to customers who have longer calling minutes. Additionally, the company can also change their charging plan. For example, first 100 minutes is regular price, next 100 minutes is 0.9 \* regular price, then next 100 minutes is 0.8\* regular price so on and so forth. We also noticed that people who registered for international plan are likely to churn in the future, which indicates possible drawbacks of this kind of service. Even though we have no further information or detail on this variable, it is worthy for this telecom company to come up some actionable plans for retaining the customer.

From our hazard model analysis, we have the following recommendation to the business:

1. Like previously stated above, we were able to confirm these recommendations through 2 models. We know that voicemail plans increases survival rates so either making it simpler, cheaper, or mandatory to have a voicemail plan may increase probability of survival for the customer. This is turn will keep more loyal customers and a steady business stream
2. Again, high calling customers are more likely to churn- specifically in the above 675 minutes a day region-- in order to understand this a little better- we may want to dive deeper into the cost for the company. If costs are not an issue, then we should maybe give the customers discounts for longer talk times in order to keep them as customers as stated in the previous recommendations.
3. We also saw that international plan customers are more likely to churn so the company should look into changing the international plan to be more affordable, better quality, or more competitive in general.
4. Based on the hazard models, the company can choose at what percentage of survival they believe they should reach back out to the customer and re-engage. We suggested at 50% survival- after 232 days, the customers should be marketed to again and given either loyalty or some sort of promotional marketing strategy.

* Limitations

As we went through the exploratory data analysis phase, we found out that the maximum days that the customer has been with the company is 243 days, which means around 8 months. The max length actually might not be sufficient especially considering the customer subscribing service with a telecom company. Thus, our dataset might have certain deviation from the reality, in which customers might register years of service with a certain telecom service provider. However, we would still be able to capture the churn rate and the trend as the trend goes by.

* Future step

We can collect more data, such as demographic data and can segment customers into different groups based on these data. Then, the company can develop and provide unique products to each group. Since better understanding of customers is the key for us to provide better services and keep them.

**Appendix**

Figure 1 (logistic regression)

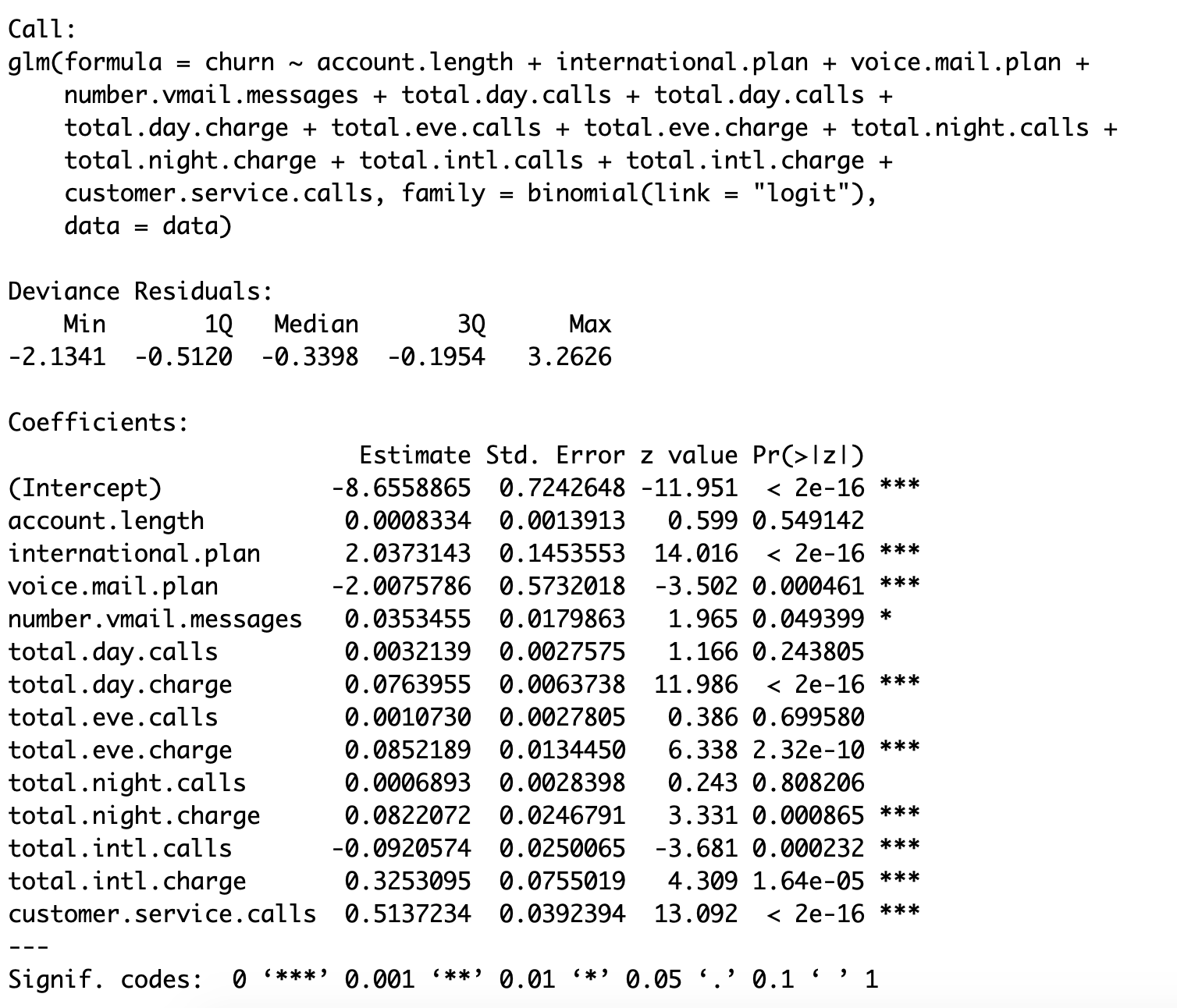
****

Figure 2: (probit regression)

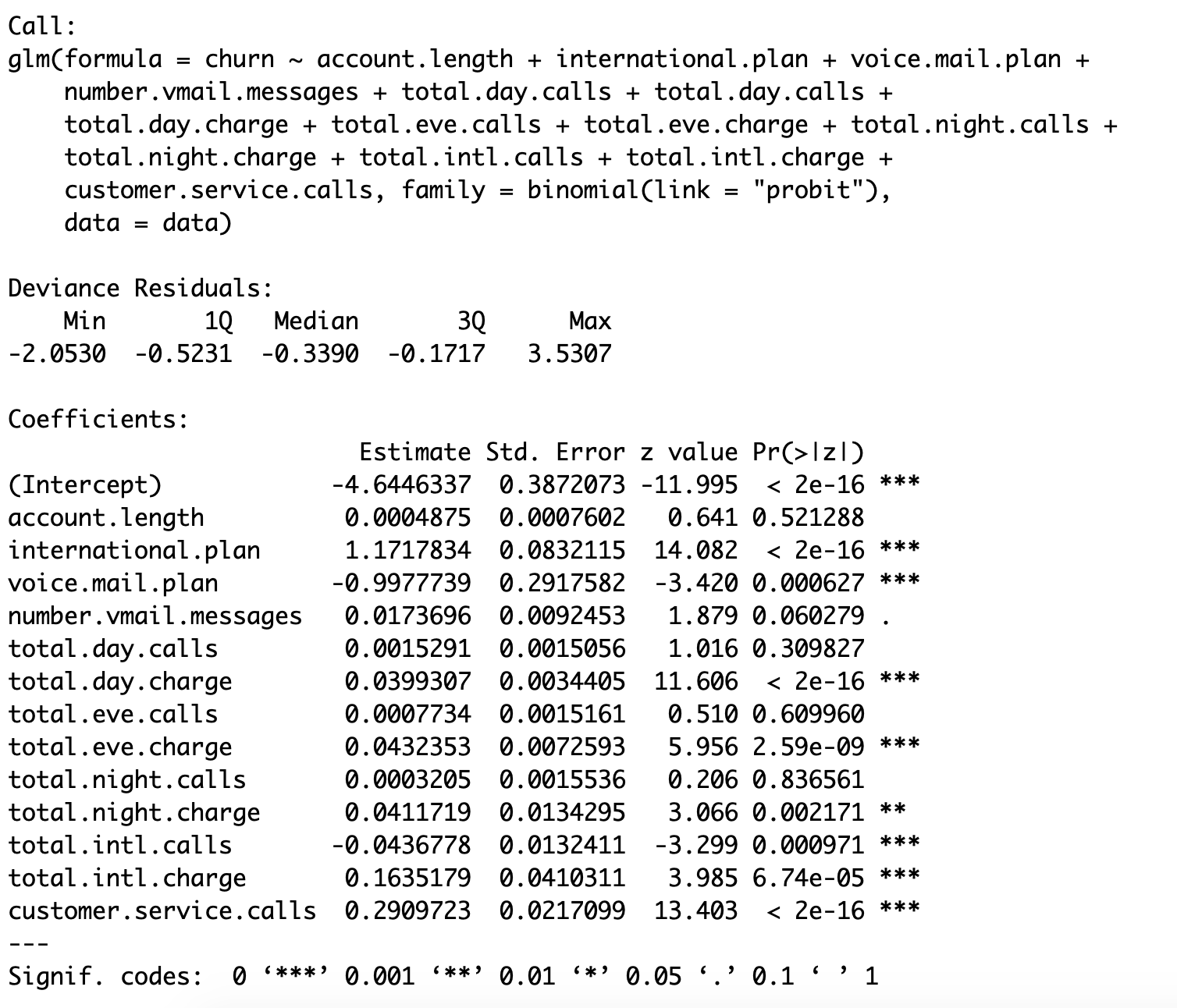


Figure 3 (voicemail plan survival)

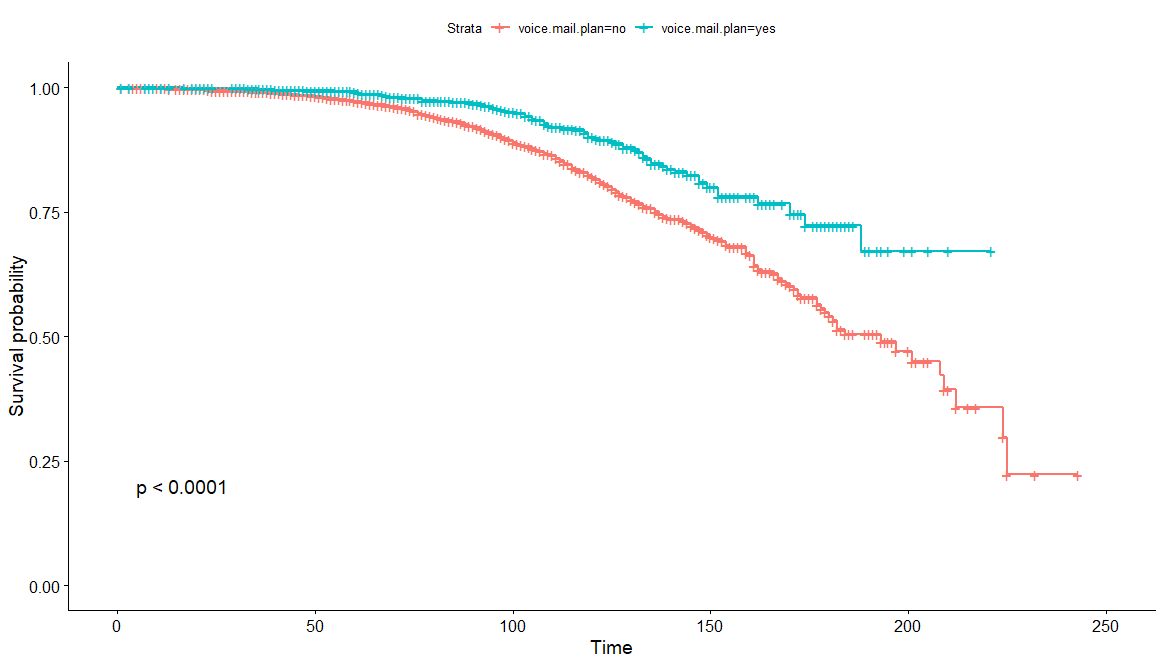


Figure 4 (international plan survival)

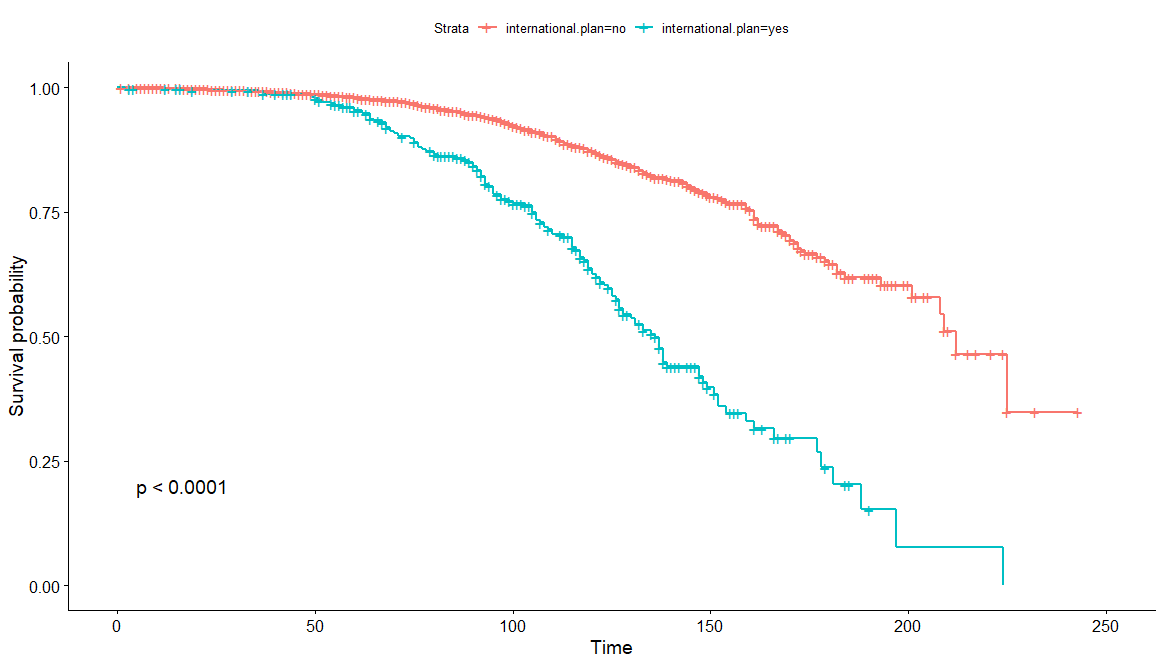


Figure 5 (total minutes used with survival)

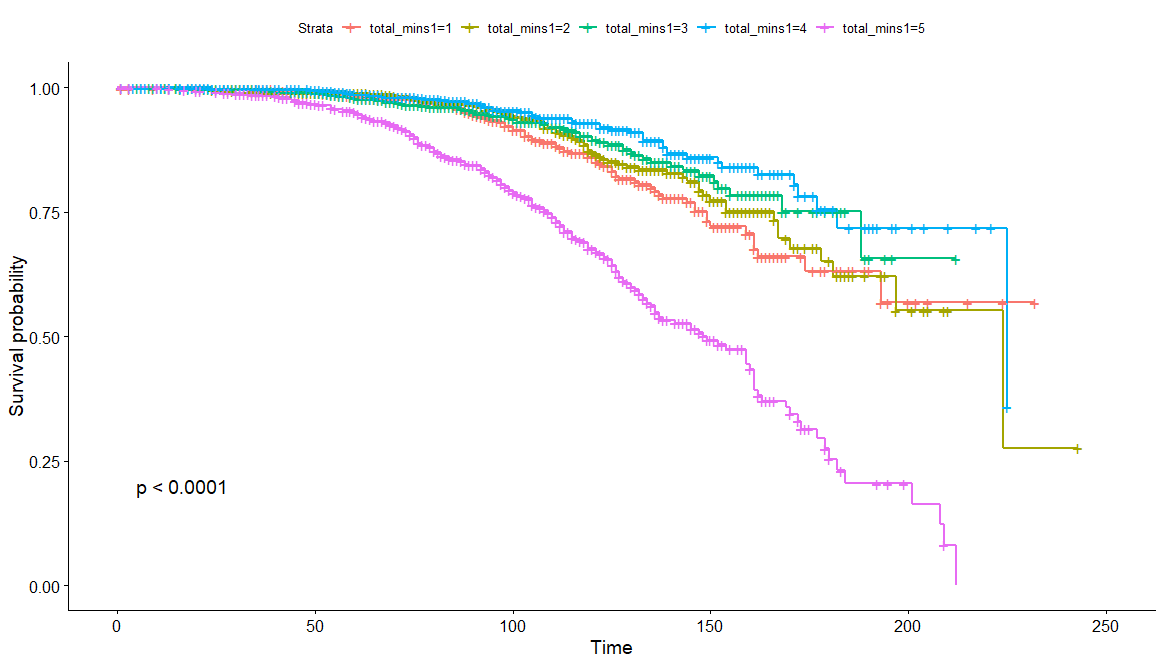
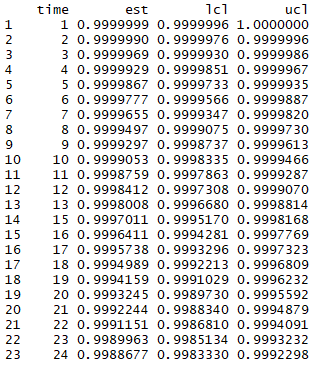
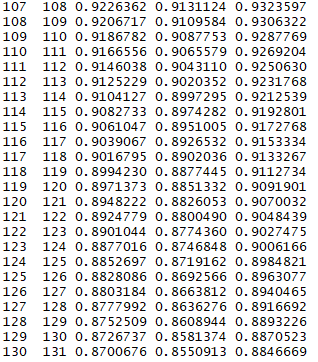


Figure 6 (3 different snippets of flexsurvreg model with weibull)

****

****

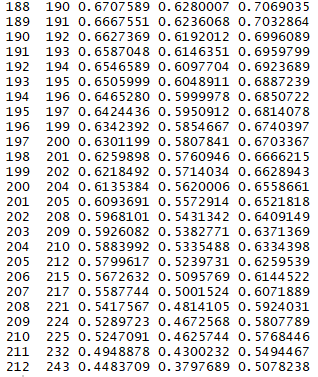
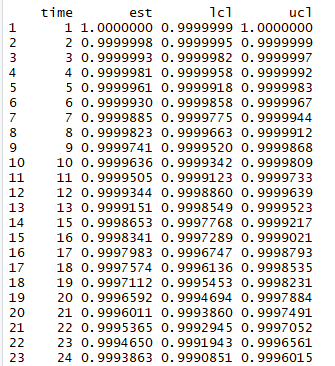
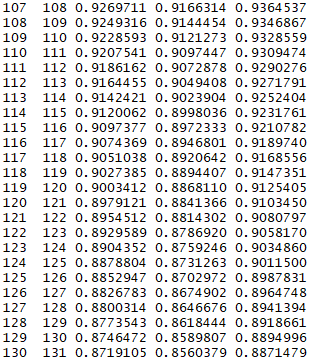
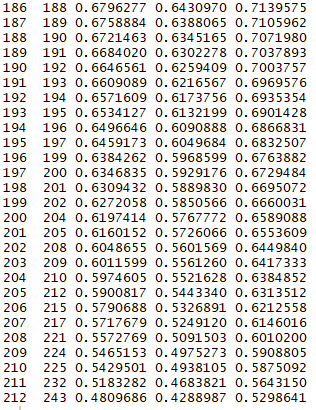
****

Figure 7 (3 different snippets of flexsurvreg with llogis)







1. <https://www.kaggle.com/ambpro/dealing-with-unbalance-eda-pca-smote-lr-svm-dt-rf> The description is from the kenel of this dataset on Kaggle [↑](#footnote-ref-0)