Analysis of customer churn in the telecom industry

Author: F Ferriere

Date: 28/04/2020

Libraries we will need for the analysis

```
library(survival)
library(survivalROC)
library(dplyr)
library(ggplot2)
library(gridExtra)
```

1. Description of the data

General

This dataset contains customer level information from a telecom company. Rows represent individual customers and columns cover customer information, as well as usage information (options included in the contract, average time spent on the phone, etc) and account information (number of days the account has been open, is account still open).

Category	Label	Description	Type
Customer	state	state	STRING
Customer	areacode	area code	STRING
Customer	phonenumber	phone number	STRING
Customer	internationalplan	international option	YES/NO
Customer	voicemailplan	voicemail option	YES/NO
Usage	numbervmailmessages	Number of voicemail messages	INT
Usage	totaldayminutes	Time spent on day calls (minutes)	INT
Usage	totaldaycalls	Number of day calls	INT
Usage	totaldaycharge	Cost of day calls	FLOAT
Usage	totaleveminutes	Time spent on evening calls (minutes)	INT
Usage	totalevecalls	Number of evening calls	INT
Usage	totalevecharge	Cost of evening calls	FLOAT
Usage	totalnightminutes	Time spent on night calls (minutes)	INT
Usage	totalnightcalls	Number of night calls	INT
Usage	totalnightcharge	Number of voicemail messages	FLOAT
Usage	totalintlminutes	Time spent on international calls (minutes)	INT
Usage	totalintlcalls	Number of international calls	INT
Usage	totalintlcharge	Cost of international calls	FLOAT
Usage	customerservicecalls	Number of calls to the customer service desk	INT
Account	accountlength	Number of days the account has been open	INT
Account	churn	Has the customer switched to a competitor ?	0/1

Data Loading & Formatting

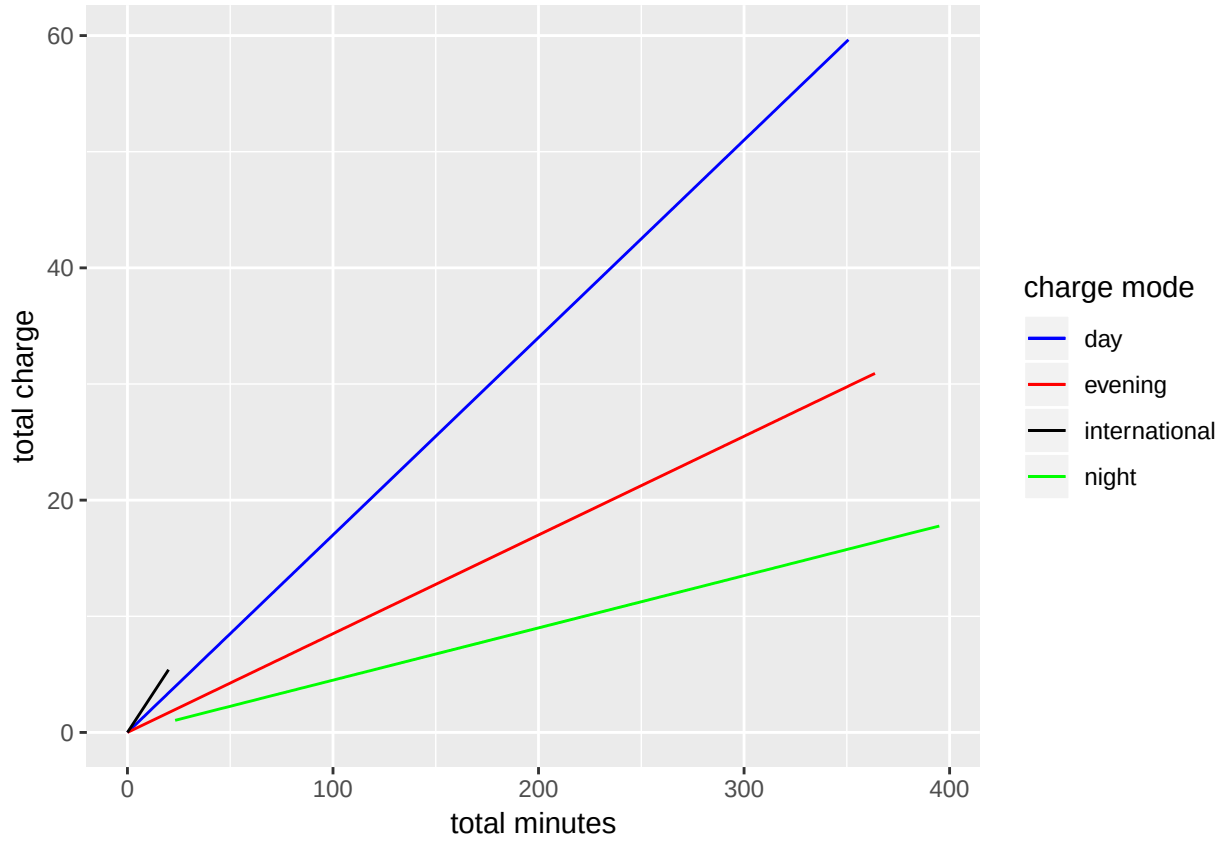
```
dat <- read.csv('telecom_churn.csv', sep=';')
names(dat) <- gsub("\\.", "", names(dat))
dat$areacode <- as.factor(dat$areacode)
dat$churn <- as.numeric(dat$churn)
```

2. Descriptive statistics

The dataset has 3,333 observations and no missing values. We notice some columns exhibit linear dependancies, allowing us to drop some of the features: total day charge, totalevecharge, totalnightcharge, totalintlcharge. The information could nevertheless be useful at an aggregate level, ie totalcharge paid by the customer, so we

will create this new feature as the sum of day, evening, night and international charges.

```
ggplot(dat) +
  geom_line(aes(x=totaldayminutes, y=totaldaycharge, color="day")) +
  geom_line(aes(x=totaleveminutes, y=totalevecharge, color="evening")) +
  geom_line(aes(x=totalnightminutes, y=totalnightcharge, color="night")) +
  geom_line(aes(x=totalintlminutes, y=totalintlcharge, color="international")) +
    xlab('total minutes') + ylab('total charge') +
  scale_color_manual(name='charge mode', values =
                       c("day" = "blue",
                         "evening" = "red",
                         "night" = "green",
                         "international" = "black"))
```




```
dat$totalcharge = dat$totaldaycharge + dat$totalevecharge +
  dat$totalnightcharge + dat$totalintlcharge
```

Categorical variables

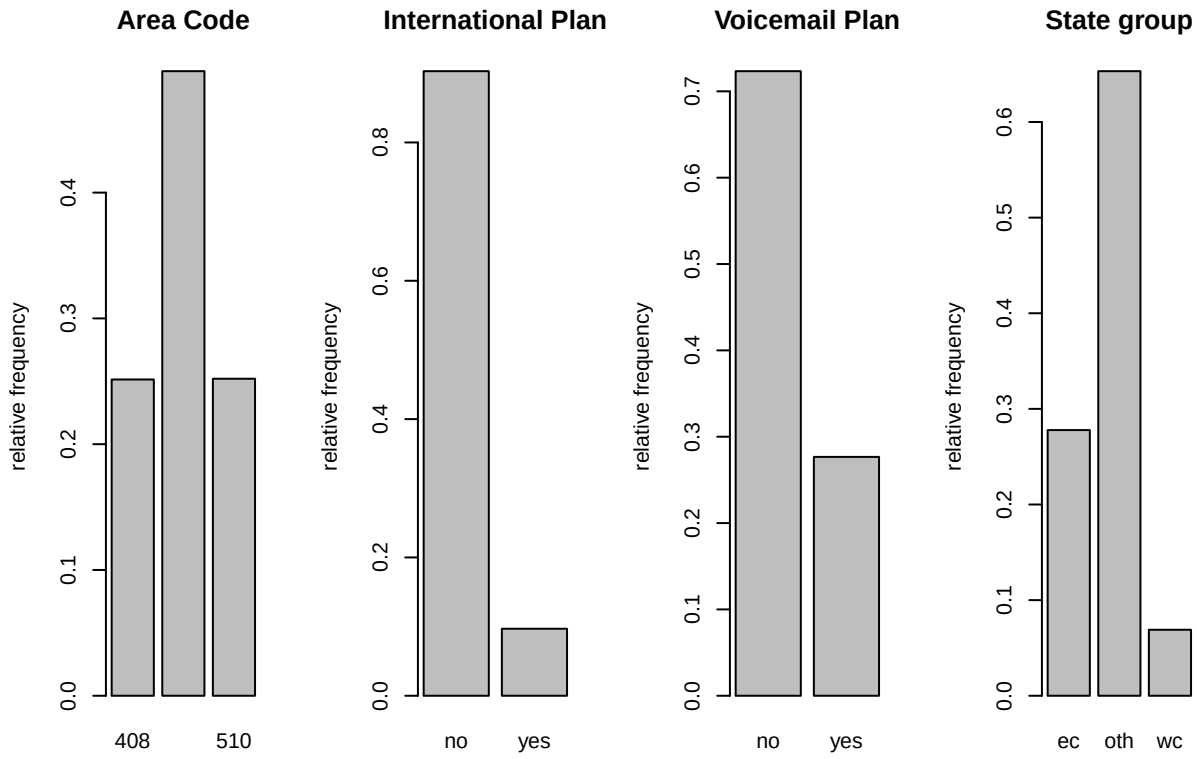
The variable state has 51 different values, it seems reasonable to reduce the dimensionality of this feature. To this end, we will map the state code to one of three groups: east coast, west coast and other. We will apply this function to the state column and populate a new column: stategroup.

```
statemapping <- function(state)
  {
    res <-"oth"
    west <- c("AK","CA","OR","WA")
    east <- c("ME","NH","MA","RI","CT","NY","NJ","DE","MD","VA","NC","SC","GA","FL")
```

```
    if (state %in% west) { res <- "wc" }
    else if (state %in% east) { res <- "ec" }
    return (res)
}
dat$stategroup <- as.factor(sapply(dat$state,statemapping))
```

Distribution of categorical variables

```
countTot <- sum(table(dat$areacode))
par(mfrow=c(1,4))
barplot(table(dat$areacode)/countTot, main='Area Code', ylab='relative frequency')
barplot(table(dat$internationalplan)/countTot, main='International Plan',
        ylab='relative frequency')
barplot(table(dat$voicemailplan)/countTot, main='Voicemail Plan',
        ylab='relative frequency')
barplot(table(dat$stategroup)/countTot, main='State group', ylab='relative frequency')
```

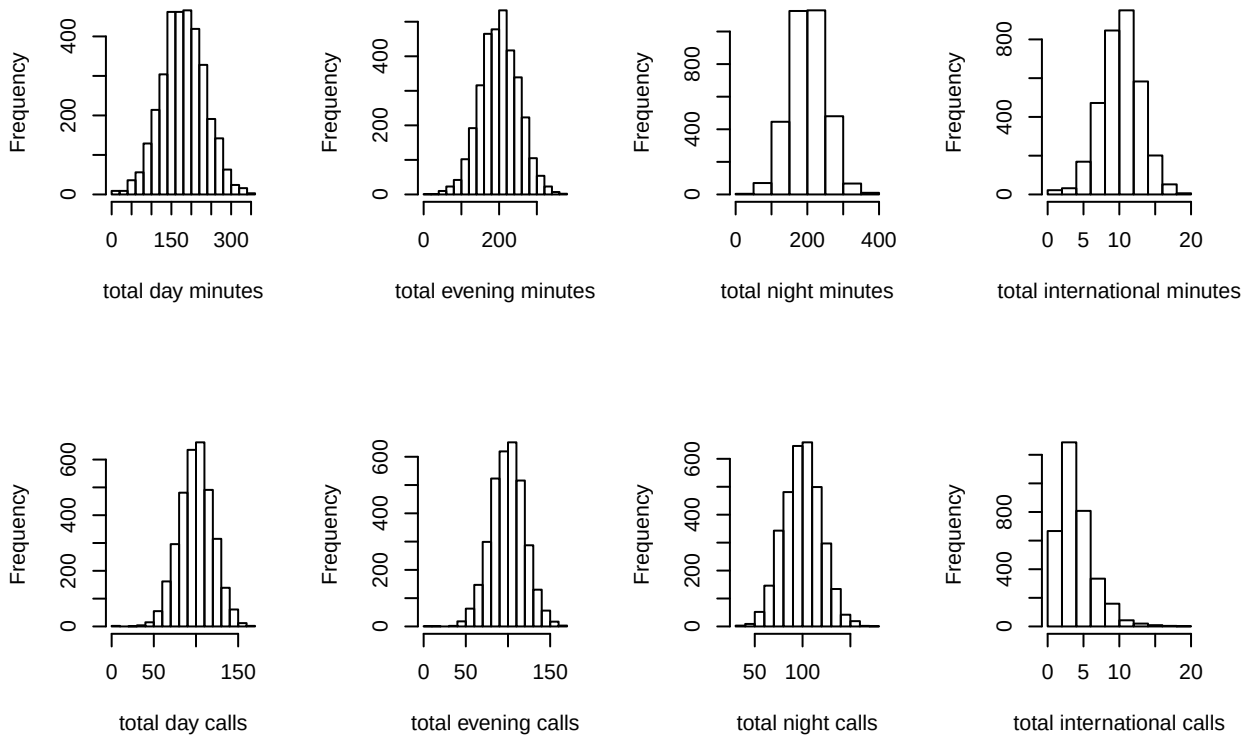


Quantitative variables

Boxplot

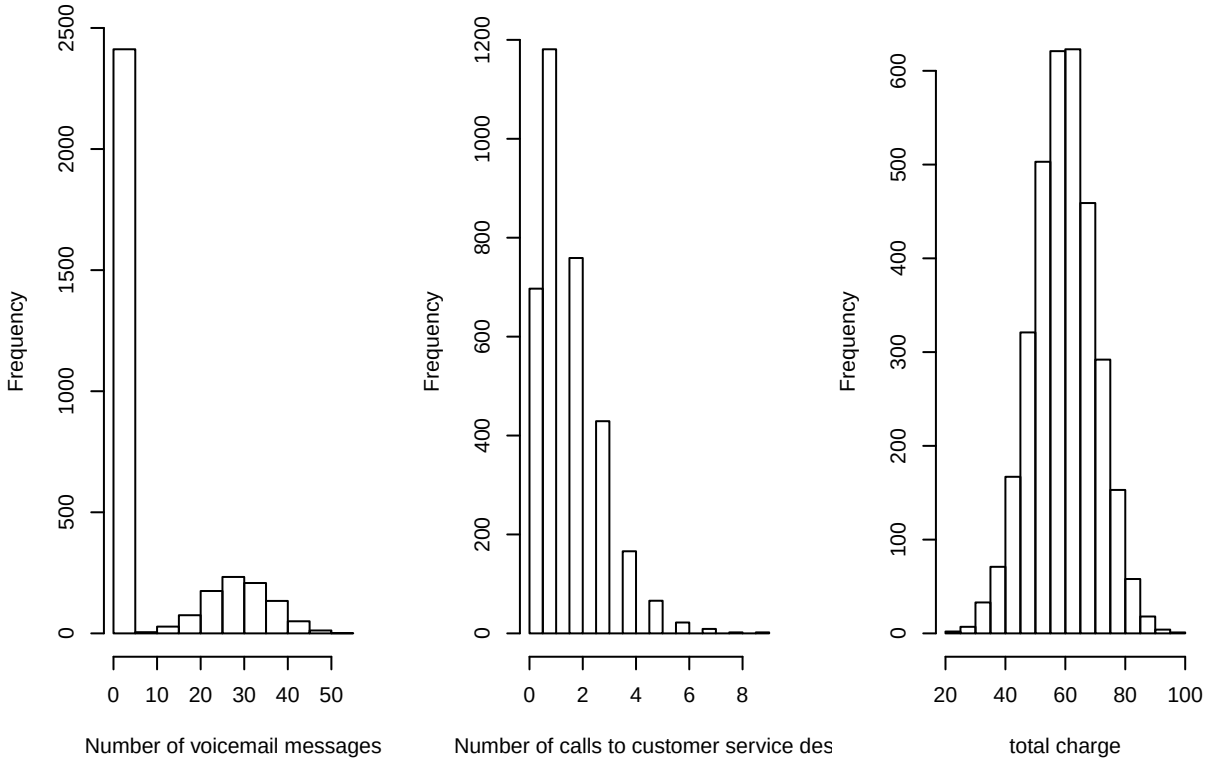
```
par(mfrow=c(2,4))
hist(dat$totaldayminutes, xlab = 'total day minutes', main=NULL)
hist(dat$totaleveminutes, xlab = 'total evening minutes', main=NULL)
hist(dat$totalnightminutes, xlab = 'total night minutes', main=NULL)
hist(dat$totalintlminutes, xlab = 'total international minutes', main=NULL)
hist(dat$totaldaycalls, xlab = 'total day calls', main=NULL)
```

```
hist(dat$totalevecalls, xlab = 'total evening calls', main=NULL)
hist(dat$totalnightcalls, xlab = 'total night calls', main=NULL)
hist(dat$totalintlcalls, xlab = 'total international calls', main=NULL)
```



Observations: Distributions of call duration is of the same order for day, night and evening. Same goes for the number of calls, which is comparable for day, evening and nights. The only distributions that stand out are for international calls, with much lower means and standard deviations.

```
par(mfrow=c(1,3))
hist(dat$numbervmailmessages, xlab = 'Number of voicemail messages', main=NULL)
hist(dat$customerservicecalls, xlab = 'Number of calls to customer service desk', main=NULL)
hist(dat$totalcharge, xlab = 'total charge', main=NULL)
```



3. Customer churn analysis

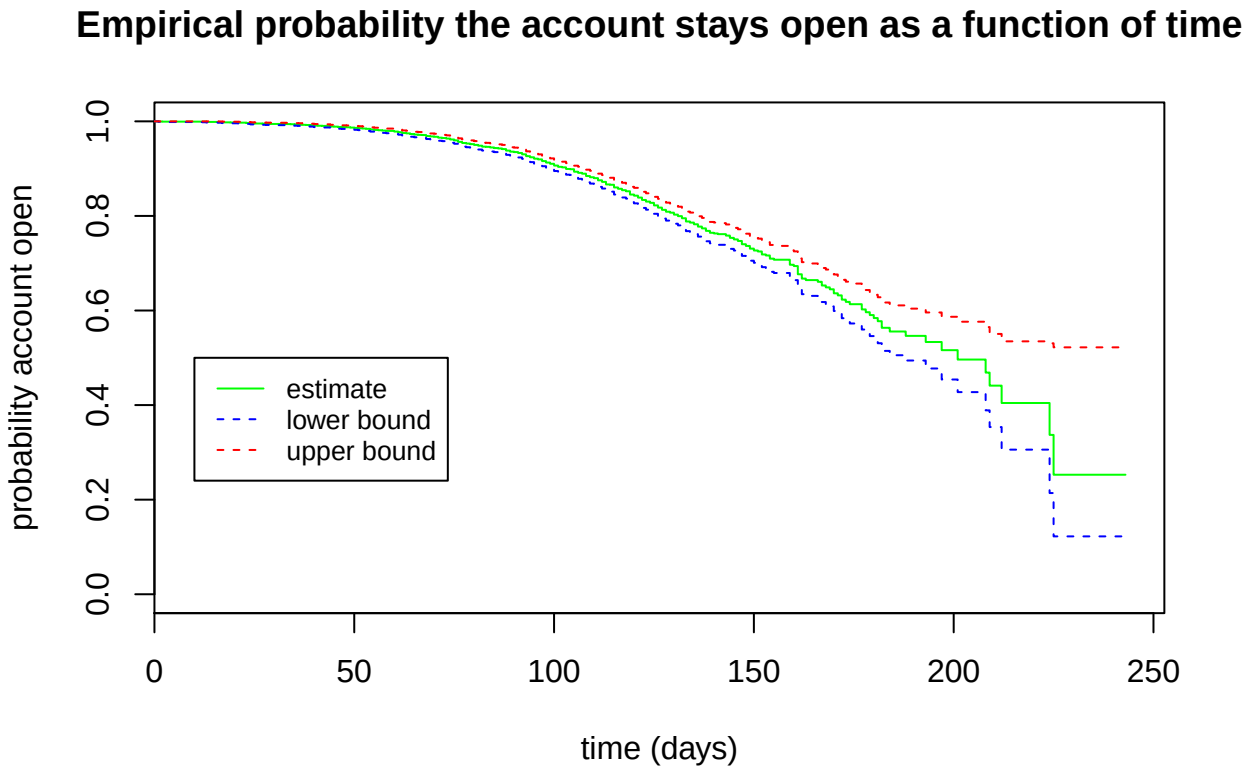
We will try to answer the following questions:
* How long does a customer stay with us on average?
* Which customers are most at risk of terminating their contract?
* Which features seem to influence a customer's decision to leave?

Question 1: How long does a customer stay with us on average?

Method: We will use the Kaplan-Meier estimator

Results:

```
fitKM <- survfit(Surv(accountlength, churn) ~ 1, data = dat)
plot(fitKM, main = 'Empirical probability the account stays open as a function of time',
     xlab = 'time (days)', ylab='probability account open',
     col=c("green", "blue", "red"))
legend(10,0.5,legend=c("estimate", "lower bound","upper bound"),
       col=c("green", "blue", "red"), lty=c(1,2,2), cex=0.8)
```




```
fitKM
```

```
## Call: survfit(formula = Surv(accountlength, churn) ~ 1, data = dat)
##
##       n  events  median 0.95LCL 0.95UCL
##    3333     483     201     188      NA
```

Key take aways:
* Out of 3,333 customers, 483 closed their account
* We estimate that 50% of our customers keep their account open longer that 201 days
* We are 95% certain the median account life is above 188 days.

Does this average estimate across all customers hide strong discrepancies between different customer segments?

Question 2: Which customers are most at risk of terminating their contract?

Impact of International Plan option

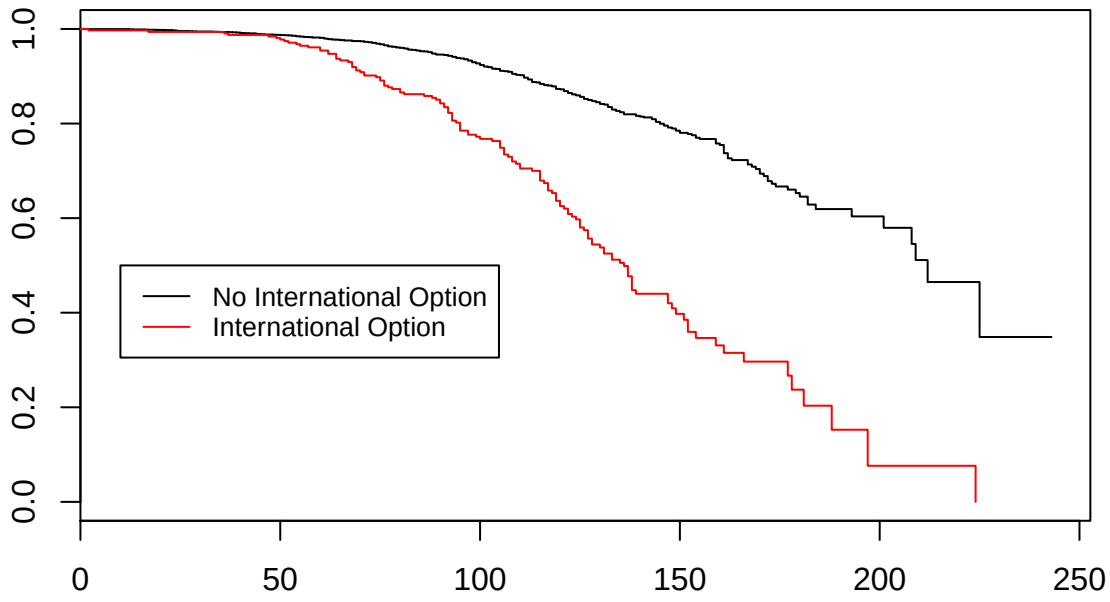
```
lrIntPlan <- survdiff(Surv(accountlength, churn)~internationalplan, data=dat)
lrIntPlan
```

```
## Call:
## survdiff(formula = Surv(accountlength, churn) ~ internationalplan,
##     data = dat)
##
##                           N Observed Expected (O-E)^2/E (O-E)^2/V
## internationalplan=no   3010      346      434      17.9       177
## internationalplan=yes   323      137       49     158.4       177
```

```
##
##  Chisq= 177  on 1 degrees of freedom, p= <2e-16
```

The p-value is very small, we reject the hypothesis that both distributions are the same. There is a significant difference in churn between customers subscribing to the international option and the others.

```
scIntPlan <- survfit(Surv(accountlength, churn) ~ internationalplan, data = dat)
plot(scIntPlan, col = 1:2)
legend(10,0.5,legend=c("No International Option", "International Option"),
       col=c("black", "red"), lty=1, cex=0.8)
```




```
scIntPlan
```

```
## Call: survfit(formula = Surv(accountlength, churn) ~ internationalplan,
##     data = dat)
##
##                           n events median 0.95LCL 0.95UCL
## internationalplan=no   3010    346    212     208      NA
## internationalplan=yes   323    137    136     127     148
```

The median account length is only 136 days for customers with the international option, whereas it is 212 days for other customers. Notice that confidence intervals do not overlap and are far apart.

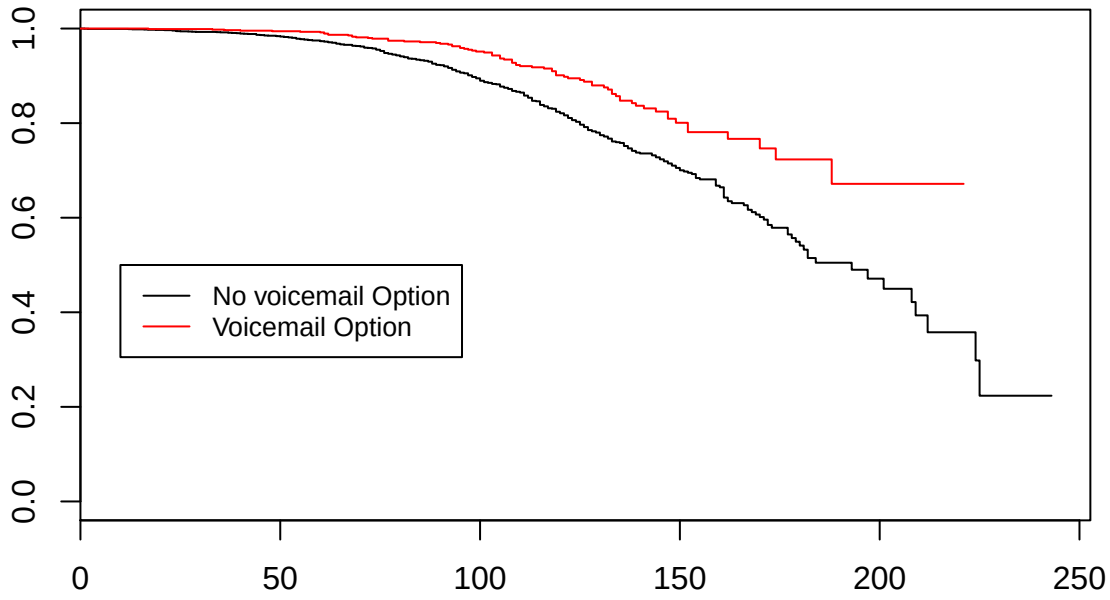
Impact of voicemail Plan option

```
lrVoiceMailPlan <- survdiff(Surv(accountlength, churn)~voicemailplan, data=dat)
lrVoiceMailPlan
```

```
## Call:
```

```
## survdiff(formula = Surv(accountlength, churn) ~ voicemailplan,
##     data = dat)
##
##                       N Observed Expected (O-E)^2/E (O-E)^2/V
## voicemailplan=no   2411      403      351      7.59        28
## voicemailplan=yes   922       80      132     20.25        28
##
##  Chisq= 28  on 1 degrees of freedom, p= 1e-07
```

```
scVoiceMailPlan <- survfit(Surv(accountlength, churn) ~ voicemailplan, data = dat)
plot(scVoiceMailPlan, col = 1:2)
legend(10,0.5,legend=c("No voicemail Option", "Voicemail Option"),
       col=c("black", "red"), lty=1, cex=0.8)
```




```
scVoiceMailPlan
```

```
## Call: survfit(formula = Surv(accountlength, churn) ~ voicemailplan,
##     data = dat)
##
##                       n events median 0.95LCL 0.95UCL
## voicemailplan=no   2411    403    193     179     212
## voicemailplan=yes   922     80     NA      NA      NA
```

The low p-value of the logrank test suggests there is a significant difference between customers subscribing to the voicemail plan and other customers. Indeed, looking at median estimates and confidence intervals, we can see more than 50% of customers with the voicemail option are still customers at the time of survey, and we are unable to calculate a median survival time for this segment.

Number of calls to the customer service center

```
lrCustServ <- survdiff(Surv(accountlength, churn)~customerservicecalls, data=dat)
lrCustServ
```

```
## Call:
## survdiff(formula = Surv(accountlength, churn) ~ customerservicecalls,
##     data = dat)
##
##                            N Observed Expected (O-E)^2/E (O-E)^2/V
## customerservicecalls=0   697       92  101.400     0.871      1.11
## customerservicecalls=1  1181      122  174.152    15.618     24.53
## customerservicecalls=2   759       87  107.927     4.058      5.24
## customerservicecalls=3   429       44   62.312     5.381      6.21
## customerservicecalls=4   166       76   23.313   119.074    125.59
## customerservicecalls=5    66       40    9.223   102.703    105.13
## customerservicecalls=6    22       14    2.684    47.698     48.17
## customerservicecalls=7     9        5    1.510     8.065      8.12
## customerservicecalls=8     2        1    0.133     5.642      5.65
## customerservicecalls=9     2        2    0.346     7.915      7.95
##
##  Chisq= 318  on 9 degrees of freedom, p= <2e-16
```

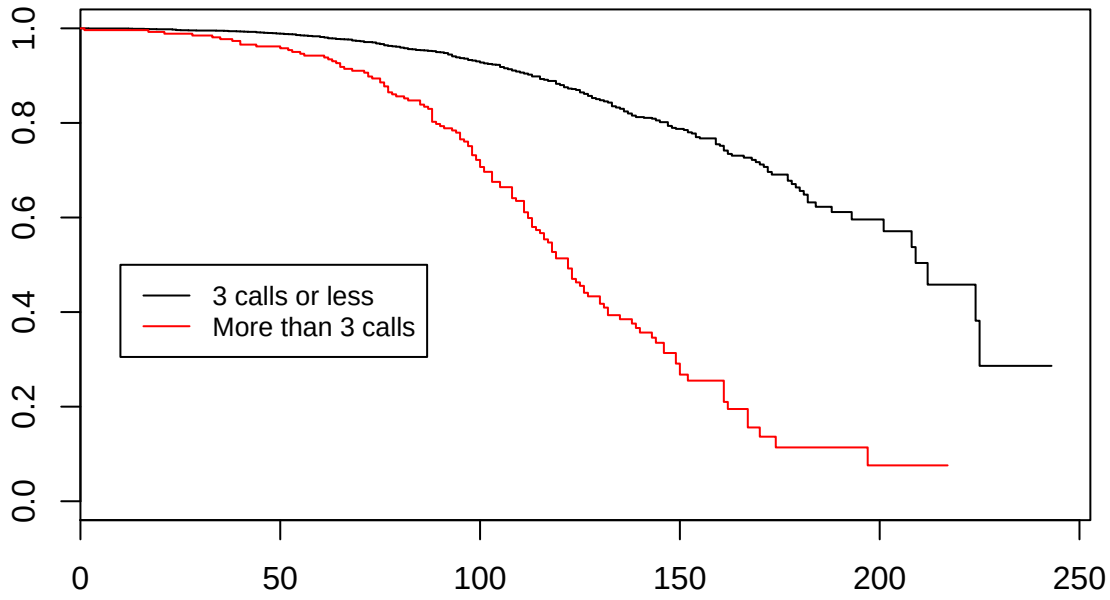
There is a very clear cutoff line: for clients calling 3 times or less, the number of observed events is less than the number of expected events (churn), whereas the number of observed events is higher than anticipated by the model for clients calling 4 times or more. We will use this information to engineer a new feature: custServAbove3=yes if client called customer service 4 times or more, no otherwise.

```
dat$custServAbove3 <- as.factor(ifelse(dat$customerservicecalls > 3, "yes", "no"))
lrcustServ <- survdiff(Surv(accountlength, churn)~custServAbove3, data=dat)
lrcustServ
```

```
## Call:
## survdiff(formula = Surv(accountlength, churn) ~ custServAbove3,
##     data = dat)
##
##                          N Observed Expected (O-E)^2/E (O-E)^2/V
## custServAbove3=no  3066      345    445.8      22.8       297
## custServAbove3=yes  267      138     37.2     273.0       297
##
##  Chisq= 297  on 1 degrees of freedom, p= <2e-16
```

Low p-value indicates the difference is significant.

```
sccustServ <- survfit(Surv(accountlength, churn) ~ custServAbove3, data = dat)
plot(sccustServ, col = 1:2)
legend(10,0.5,legend=c("3 calls or less", "More than 3 calls"),
       col=c("black", "red"), lty=1, cex=0.8)
```




```
sccustServ
```

```
## Call: survfit(formula = Surv(accountlength, churn) ~ custServAbove3,
##     data = dat)
##
##                        n events median 0.95LCL 0.95UCL
## custServAbove3=no   3066    345    212     201      NA
## custServAbove3=yes   267    138    122     116     130
```

The median estimates for both groups are significantly different, with non overlapping 95% confidence intervals. Customers calling customer service 4 times are much more likely to terminate their contract.

General Conclusion for question 2:
* Half the customers stay for at least 201 days
* This estimates hides strong discrepancies between customers, based on their options: customers with the international option seem a lot keener to close their account whereas customers with the voicemail option are more faithful on average.
* The number of calls to customer service also shows a strong impact on the likelihood a client may terminate their contract.

Question 3: Which features seem to influence a customer's decision to leave?

Method

We would like to quantify the impact of the different variables in our dataset on the probability that a customer might terminate their contract.
The framework will be a cox proportional hazard rate model, where the explanatory variables will be some columns from our dataset. We will start from a full model and use stepwise model selection based on AIC to

select the best model.

Results

```
fitfull <- coxph(formula=Surv(accountlength, churn)~areacode + voicemailplan
                 + numbervmailmessages + totaldayminutes + totaldaycalls
                 + totaleveminutes + totalevecalls + totalnightminutes
                 + totalnightcalls + internationalplan + totalintlminutes
                 + totalintlcalls + custServAbove3 + stategroup + totalcharge, data = dat)
aicmodel <- step(fitfull)
```

Here's the final model using backward step selection:

```
summary(aicmodel)
```

```
## Call:
## coxph(formula = Surv(accountlength, churn) ~ voicemailplan +
##     numbervmailmessages + totaldayminutes + totaleveminutes +
##     totalnightminutes + internationalplan + totalintlcalls +
##     custServAbove3 + totalcharge, data = dat)
##
##   n= 3333, number of events= 483
##
##                           coef exp(coef)  se(coef)      z Pr(>|z|)
## voicemailplanyes     -1.705649  0.181655  0.475777 -3.585 0.000337 ***
## numbervmailmessages   0.029971  1.030425  0.014842  2.019 0.043454 *
## totaldayminutes      -0.022080  0.978162  0.010403 -2.123 0.033795 *
## totaleveminutes      -0.012143  0.987931  0.005301 -2.291 0.021985 *
## totalnightminutes    -0.004993  0.995020  0.002896 -1.724 0.084731 .
## internationalplanyes  1.166508  3.210760  0.102446 11.387  < 2e-16 ***
## totalintlcalls       -0.070078  0.932321  0.020991 -3.338 0.000842 ***
## custServAbove3yes     1.551503  4.718556  0.102568 15.127  < 2e-16 ***
## totalcharge           0.185178  1.203432  0.061137  3.029 0.002455 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                      exp(coef) exp(-coef) lower .95 upper .95
## voicemailplanyes        0.1817     5.5050   0.07149    0.4616
## numbervmailmessages     1.0304     0.9705   1.00088    1.0608
## totaldayminutes         0.9782     1.0223   0.95842    0.9983
## totaleveminutes         0.9879     1.0122   0.97772    0.9982
## totalnightminutes       0.9950     1.0050   0.98939    1.0007
## internationalplanyes    3.2108     0.3115   2.62667    3.9247
## totalintlcalls          0.9323     1.0726   0.89474    0.9715
## custServAbove3yes       4.7186     0.2119   3.85925    5.7692
## totalcharge             1.2034     0.8310   1.06753    1.3566
##
## Concordance= 0.807  (se = 0.011 )
## Rsquare= 0.142   (max possible= 0.872 )
## Likelihood ratio test= 509.3  on 9 df,   p=<2e-16
## Wald test            = 568.9  on 9 df,   p=<2e-16
## Score (logrank) test = 652.9  on 9 df,   p=<2e-16
```

```
AIC(aicmodel)
```

```
## [1] 6360.115
```

This naive approach allows us to eliminate a number of features but it has its limits. For example, we know totalcharge is an increasing function of totaldayminutes. Still those two variables have opposite effects in the model, and both appear to be significant. As an alternative, we will estimate a more restrictive model, dropping features relating to time spent on the phone and keeping only totalcharge.

```
fitfinal <- coxph(formula=Surv(accountlength, churn)~ voicemailplan + internationalplan
                  + custServAbove3 + totalcharge, data = dat)
summary(fitfinal)
```

```
## Call:
## coxph(formula = Surv(accountlength, churn) ~ voicemailplan +
##     internationalplan + custServAbove3 + totalcharge, data = dat)
##
##   n= 3333, number of events= 483
##
##                           coef exp(coef) se(coef)      z Pr(>|z|)
## voicemailplanyes     -0.78194   0.45752  0.12351 -6.331 2.44e-10 ***
## internationalplanyes  1.18817   3.28108  0.10166 11.687  < 2e-16 ***
## custServAbove3yes     1.54120   4.67020  0.10123 15.224  < 2e-16 ***
## totalcharge           0.05246   1.05386  0.00425 12.345  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                      exp(coef) exp(-coef) lower .95 upper .95
## voicemailplanyes        0.4575     2.1857    0.3591    0.5828
## internationalplanyes    3.2811     0.3048    2.6883    4.0045
## custServAbove3yes       4.6702     0.2141    3.8297    5.6952
## totalcharge             1.0539     0.9489    1.0451    1.0627
##
## Concordance= 0.806  (se = 0.011 )
## Rsquare= 0.136   (max possible= 0.872 )
## Likelihood ratio test= 486.5  on 4 df,   p=<2e-16
## Wald test            = 559.6  on 4 df,   p=<2e-16
## Score (logrank) test = 637.5  on 4 df,   p=<2e-16
```

```
AIC(fitfinal)
```

```
## [1] 6372.87
```

Model Interpretation

We will use this model as our final model. Even though AIC is slightly worse (6373 vs 6360), we gain a lot in terms of interpretability and simplicity, and all retained variables are very significant.

- The variable with the most impact on customer churn is the indicator variable 'customer called customer service 4 times or more'. All other things being equal, the risk of such a customer closing their contract is 4.7 times bigger than for clients who called customer service 3 times or less.

- Subscription to International option also has a strong influence on the risk a customer may stop their contract. The risk of those customers closing their contract is more than 3 times that of a customer not having this option.

- Next variable of influence is the voicemail plan option. The risk for customers not having the option is twice that of customers subscribing to the voicemail option.

- Finally, totalcharge also has an influence on a customer's risk with a ratio of 1.05. For each dollar

increase in totalcharge, the risk of a customer closing their account increases 5%.

Model validation

We'll look at Schoenfeld's residuals, to check whether the model assumption of proportional hazard rates is satisfied.

```
cox.zph(fitfinal)
```

```
##                          rho chisq     p
## voicemailplanyes      0.0164 0.131 0.718
## internationalplanyes  0.0593 1.703 0.192
## custServAbove3yes    -0.0157 0.124 0.725
## totalcharge           0.0292 0.723 0.395
## GLOBAL                    NA 2.846 0.584
```

Global p-value and individual p-value are all well above 5%, indicating we do not reject the assumption that hazard rates are proportional. The assumptions on which our model relies are satisfied and we can keep it.

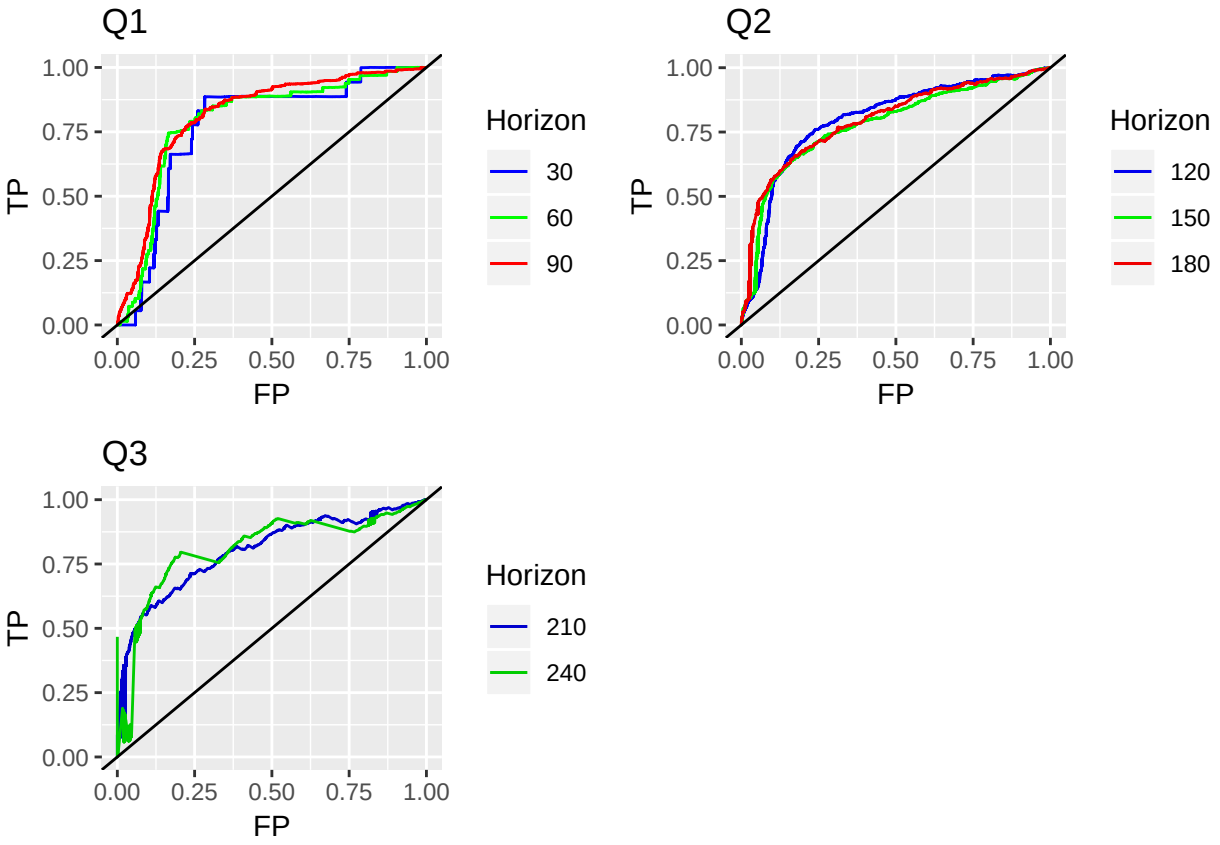
Goodness of fit analysis: ROC

We will plot ROC for our model for different prediction horizons (monthly). This will give us a visualize analysis of the quality of our model.

```
dat_lp <- predict(fitfinal, type="lp")
sroc30 <- survivalROC(Stime = dat$accountlength, status=dat$churn, marker = dat_lp,
                      predict.time = 30, method = "KM")
sroc60 <- survivalROC(Stime = dat$accountlength, status=dat$churn, marker = dat_lp,
                      predict.time = 60, method = "KM")
sroc90 <- survivalROC(Stime = dat$accountlength, status=dat$churn, marker = dat_lp,
                      predict.time = 90, method = "KM")
sroc120 <- survivalROC(Stime = dat$accountlength, status=dat$churn, marker = dat_lp,
                       predict.time = 120, method = "KM")
sroc150 <- survivalROC(Stime = dat$accountlength, status=dat$churn, marker = dat_lp,
                       predict.time = 150, method = "KM")
sroc180 <- survivalROC(Stime = dat$accountlength, status=dat$churn, marker = dat_lp,
                       predict.time = 180, method = "KM")
sroc210 <- survivalROC(Stime = dat$accountlength, status=dat$churn, marker = dat_lp,
                       predict.time = 210, method = "KM")
sroc240 <- survivalROC(Stime = dat$accountlength, status=dat$churn, marker = dat_lp,
                       predict.time = 240, method = "KM")
```

```
t1 <- ggplot(NULL) +
      geom_line(aes(sroc30$FP, sroc30$TP, color='30')) +
      geom_line(aes(sroc60$FP, sroc60$TP, color='60')) +
      geom_line(aes(sroc90$FP, sroc90$TP, color='90')) +
      labs(x= "FP", y = "TP") + labs(title = "Q1") +
      scale_color_manual(name='Horizon',
                         values=c('30'='blue', '60'='green', '90' = 'red')) +
      geom_abline(slope=1, intercept=0)
t2 <- ggplot(NULL) +
      geom_line(aes(sroc120$FP, sroc120$TP, color='120')) +
      geom_line(aes(sroc150$FP, sroc150$TP, color='150')) +
      geom_line(aes(sroc180$FP, sroc180$TP, color='180')) +
      labs(x= "FP", y = "TP") + labs(title = "Q2") +
      scale_color_manual(name='Horizon',
```

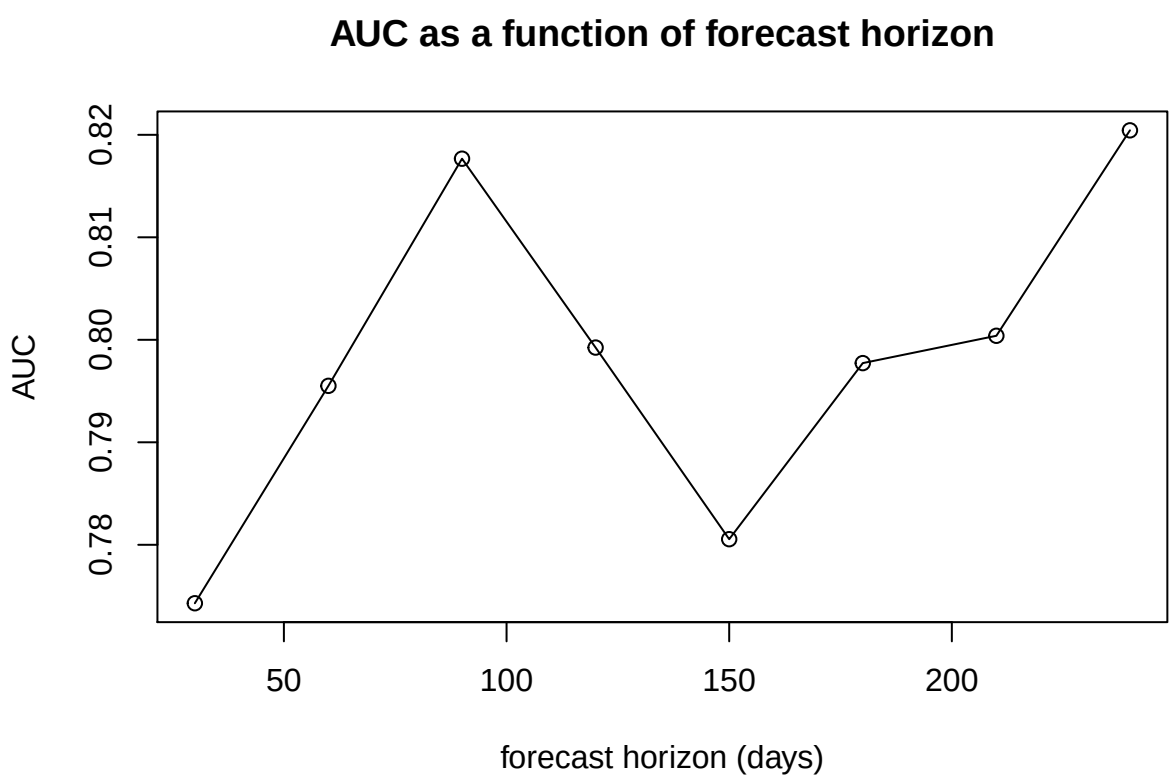
```
                        values=c('120'='blue2', '150'='green2', '180' = 'red2')) +
      geom_abline(slope=1, intercept=0)
t3 <- ggplot(NULL) +
      geom_line(aes(sroc210$FP, sroc210$TP, color='210')) +
      geom_line(aes(sroc240$FP, sroc240$TP, color='240')) +
      labs(x= "FP", y = "TP") + labs(title = "Q3") +
      scale_color_manual(name='Horizon',
                        values=c('210'='blue3', '240'='green3')) +
      geom_abline(slope=1, intercept=0)
grid.arrange(t1, t2, t3, ncol=2, nrow = 2)
```



ROC Analysis: Whatever horizon we select, the model does much better than random guess.

Looking at AUC for the different prediction times will enable us to see for which horizon our model is most reliable.

```
horizons = c(30,60,90,120,150,180,210,240)
AUCs = c(sroc30$AUC,sroc60$AUC,sroc90$AUC,sroc120$AUC,sroc150$AUC,
         sroc180$AUC,sroc210$AUC,sroc240$AUC)
plot(horizons, AUCs, main='AUC as a function of forecast horizon',
     xlab = 'forecast horizon (days)', ylab='AUC')
lines(horizons, AUCs)
```



We notice maximum values at horizons 90 days and 240 days, with AUCs around 82%.

4. Conclusions

The previous analysis highlights a few important business facts, the main takeaways being as follows:
* the number of calls a client places to customer service is a very strong indicator they may end their contract and switch to competition
* Customers subscribed to the international plan option are a lot more likely to terminate their contract than the average customer. This might indicate a poor pricing of this option or excessive cost of international calls compared to competition
* customers subscribing to the voicemail plan are a lot more faithfull than average customers. Again, this may indicate a mispricing vs competition, ie that we price this option too cheap relative to market standards
* totalcharge, though a significant feature, is not the main explanatory variable. It has overall a much more limited impact on the risk a client may terminate their contract.
* AUC oscillates between 78% and 82% for different prediction horizons and indicates the model performs better for predictions at 90 days and 240 days.