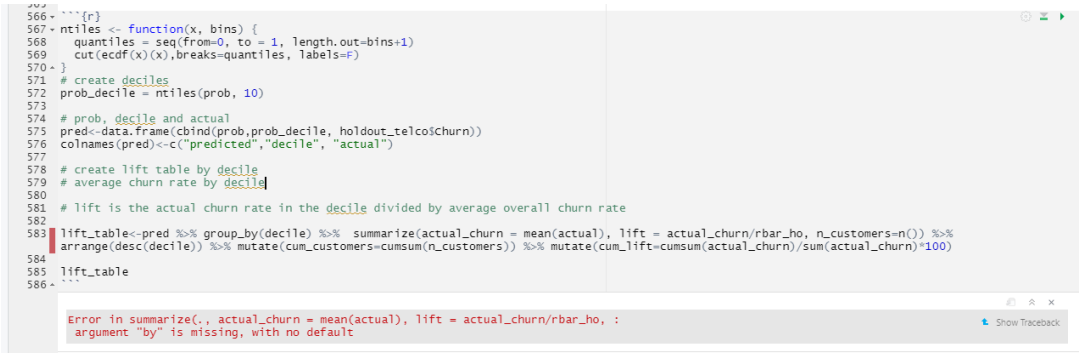
Questions Q&A 3 – Tuesday Nov 16th

Coding Questions – Gijs Breakout Room

- [Mihai];

  I'm trying to get the lift using the tesco_holdout data from the assignment, but I'm getting an error message that I do not know how to resolve.



```r
```{r}
ntiles <- function(x, bins) {
  quantiles = seq(from=0, to = 1, length.out=bins+1)
  cut(ecdf(x)(x),breaks=quantiles, labels=F)
}
# create deciles
prob_decile = ntiles(prob, 10)

# prob, decile and actual
pred<-data.frame(cbind(prob,prob_decile, holdout_telco$Churn))
colnames(pred)<-c("predicted","decile", "actual")

# create lift table by decile
# average churn rate by decile

# lift is the actual churn rate in the decile divided by average overall churn rate

lift_table<-pred %>% group_by(decile) %>%  summarize(actual_churn = mean(actual), lift = actual_churn/rbar_ho, n_customers=n()) %>%
arrange(desc(decile)) %>% mutate(cum_customers=cumsum(n_customers)) %>% mutate(cum_lift=cumsum(actual_churn)/sum(actual_churn)*100)

lift_table
```
```
Error in summarize(., actual_churn = mean(actual), lift = actual_churn/rbar_ho, :
 argument "by" is missing, with no default
```

  Naturally, I ran

  holdout_telco<-read.csv('data/telco_holdout.csv')
  holdout_telco$gender<-as.factor(holdout_telco$gender)
  holdout_telco$PaymentMethod<-as.factor(holdout_telco$PaymentMethod)

  # Change Churn from "no" "yes" to 0 1
  holdout_telco <- holdout_telco %>%
  mutate(Churn = ifelse(Churn == "No",0,1))

  and followed along the steps from the L3 sign. so I'm not sure what I'm doing wrong.

Questions Week 3 – Logistic Regression

- [Efe];

  In the 40th slide and 07.05 minute of the webclip 5, I didn't understand why we are changing the threshold. And also, I couldn't understand its relation between HIT rate and predicted positive or negative. Can you make brief explanation about that part?

- [Mieke];

  Usually, if we want to determine how much the odds of churning (dependent variable) changes by an extra unit of an explanatory variable, then we could read the coefficient from the table and use (exp(beta_k)-1)*100 as on slide 13. However, when this explanatory variable is also included in the model as an interaction term and we now want to determine the change in odds for customers in a specific state of the interacted variable, how to proceed?
```

- [Sterre];

  In the file of the quiz the following is done:

  Churn.num <- as.numeric(asfactor(telco$Churn))-1

  model_1 <- glm(Churn.num ~ gender + SeniorCitizen + tenure, data=telco, family="binomial")

  model_2 <- glm(Churn.num ~ gender*SeniorCitizen + tenure, data=telco, family="binomial")

  model_3 <- glm(Churn.num ~ . - Churn, data=telco, family="binomial")

  Why is it necessary to change 'Churn' into a numeric and factor variable?
  And why do we need to delete 'Churn' (-Churn) in model_3, but not in model_1 and model_2?

- [Noa];

  Practice Quiz Q1: in the script provided, we create Churn.num. But why? Can't we just recode Churn by using as.factor? Additionally, the formula used to create Churn.num is Churn.num<-as.numeric(as.factor(telco$Churn))-1, I was wondering where the -1 comes from?