Introduction

The telecoms churn dataset consists of 512 observations of unique customers. This data contains recorded information such as a categorical variable under the payment method used column, a continuous variable under the monthly charge's column, do they live with a partner indicated with a yes or no, what kind of internet service do they have recorded as a categorical variable, gender denoted by a 1 or 0, customer id and if they stayed with the provider or not indicated by a binary value under the churn column.

Table 1: Telecoms churn data sample

churn	customerid	gender	internetservice	partner	monthlycharges	paymentmethod
0	7619-PLRLP	Female	DSL	Yes	74.10	Bank transfer (automatic)
0	7847-GYYKX	Female	No	Yes	20.35	Bank transfer (automatic)
0	3906-MKIMJ	Male	DSL	Yes	41.95	Electronic check
0	8561-NMTBD	Female	Fiber optic	Yes	112.35	Credit card (automatic)
0	6198-PNNSZ	Female	Fiber optic	Yes	109.80	Bank transfer (automatic)
0	6838-HVLXG	Female	No	No	20.30	Mailed check
0	4378-MYPGO	Male	Fiber optic	Yes	105.25	Bank transfer (automatic)
0	5882-CMAZQ	Female	DSL	Yes	34.25	Electronic check
0	1866-ZSLJM	Male	No	No	20.50	Credit card (automatic)
0	1090-PYKCI	Female	Fiber optic	Yes	105.10	Credit card (automatic)
0	7608-RGIRO	Male	No	No	24.40	Mailed check
0	3551-GAEGL	Male	DSL	Yes	30.40	Bank transfer (automatic)
0	5223-UZAVK	Male	Fiber optic	No	100.30	Credit card (automatic)
0	5787-KXGIY	Male	No	Yes	19.30	Credit card (automatic)
0	8979-CAMGB	Male	Fiber optic	No	102.10	Electronic check
0	2603-HVKCG	Male	Fiber optic	No	101.40	Electronic check
0	2581-VKIRT	Female	DSL	Yes	65.50	Mailed check
0	2954-PIBKO	Female	DSL	Yes	64.15	Credit card (automatic)

The goal of this report is to synthesise a model that could provide a probability for a customer who is male, living with a partner, has monthly charges equal to 70, has fibre optic internet and uses a credit card as a payment method. Along with this provide a report of the relationship between each predictor and the response churn. There are 2 more versions of the

dataset described above; one only contains users that have internet access whereas the other has those that don't. Users without internet make up 100 of the 512 observations and users with internet make up the other 412 observations. The goal here is to investigate if there is a separate trend between these two groups. A density plot was used to split users with and without internet based on monthly charges, as seen in Figures 1, 2 and 3. Internet users seem to also have two separate groups of customers divided by monthlycharges indicatd by the bimodal distribution.

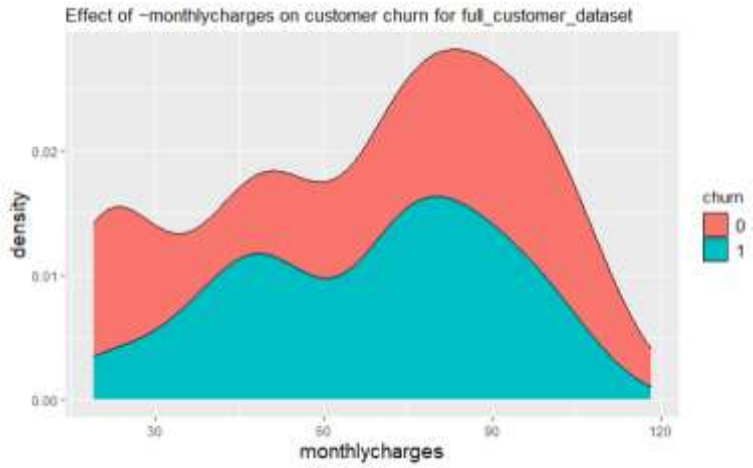

Figure 1: Density graph of monthly charges for all users

Figure 2: Density graph of monthly charges, users with internet

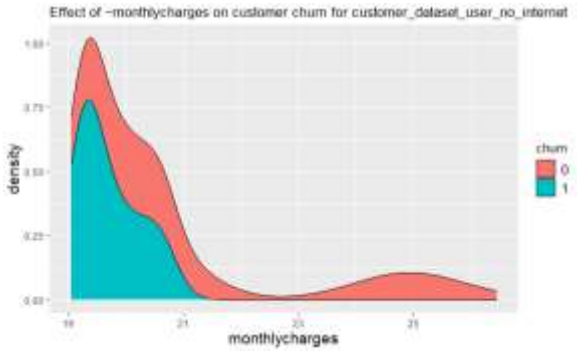

Figure 3: Density graph of monthly charges, users without internet

Table 2 presents the values and distribution of values per column without flooding the report with a plethora of plots and graphs. 16.6 % of customers in this data set have chosen to leave. There is a ~5% difference between customer genders, 55.66% being men, 44.34 being women. The most popular internet offering is fibre optic at 46.88 % then 19.53% of users don't have either DSL or fibre optic. There is an even split of users that live with and without a partner at 50.59 and 49.41 respectively. The preferred payment method is electronic check at 35.35%.

Table 2: Unique values and % of representation in the original telecoms churn data set

Column Name	Values	% Representation of Column
churn	0	83.40
churn	1	16.60
gender	Female	44.34
gender	Male	55.66
internetservice	DSL	33.59
internetservice	Fiber optic	46.88
internetservice	No	19.53
partner	No	49.41
partner	Yes	50.59
paymentmethod	Bank transfer (automatic)	20.90
paymentmethod	Credit card (automatic)	22.07
paymentmethod	Electronic check	35.35
paymentmethod	Mailed check	21.68

Table 3: Unique values, % of representation, users with internet

Column Name	Values	% Representation of Column
churn	0	80.58
churn	1	19.42
gender	Female	44.42
gender	Male	55.58
partner	No	48.30
partner	Yes	51.70
paymentmethod	Bank transfer (automatic)	19.90
paymentmethod	Credit card (automatic)	22.82
paymentmethod	Electronic check	41.99
paymentmethod	Mailed check	15.29

Table 4: Unique values, % of representation, users without internet

Column Name	Values	% Representation of Column
churn	0	95
churn	1	5
gender	Female	44
gender	Male	56
partner	No	54
partner	Yes	46
paymentmethod	Bank transfer (automatic)	25
paymentmethod	Credit card (automatic)	19
paymentmethod	Electronic check	8
paymentmethod	Mailed check	48

Model 1

A full interaction general linear model was fitted using *gender, internetservice, partner, monthlycharges, paymentmethod* as factors with churn as the response. The full equation would have been:

$$y_i = \beta_0 + \beta_1(\delta_{\text{i,male}}) + \beta_2(\delta_{\text{i,No}}) + \beta_3(\delta_{\text{i,FIber optic}}) + \beta_4(\delta_{\text{i,No}}) + \beta_5(\delta_{\text{i,Electronic Check}}) + \beta_6(\delta_{\text{i,Mailed } Check}) + \beta_7(\delta_{\text{i,Bank } transfer}) + \cdots + \beta_n(\text{i}_{\text{ Gender}} * \text{i}_{\text{ Internet Service}} * \text{i}_{\text{ Partner}} * \text{i}_{\text{ Payment Method}}) + \epsilon_i$$

Where y_i is the churn for sample i , each δ dummy variable is replace with 1 if target variable is present and 0 otherwise. $\beta_1(\delta_{\text{i,male}})$ = 1 if customer is male, 0 otherwise, repeat for all other variables such as *Internet Service* in β_2 *1* if No 0 otherwise or in $\beta_3$1 if *fibre optic 0 otherwise etc*, *Partner* (β_4) and *Payment Method (*$\beta_5, \beta_6, \beta_7$*)* where each variables respected β_n value is attributed to the categorical value in each column that isn't covered by the intercept. The additional interaction equation above denoted by β_n would have to be filled out for each combination of each variable listed by name, excluding what's covered by the intercept term, which is

Table 5: Values, covered by intercept β_0

Gender	Internet service	Partner	Payment method
Female	DSL	Yes	Bank Transfer

Where n in β_n is the next beta to be appended to the equation for each combination of variables. Gender would be male, as female is covered in the intercept. Internet service would be No or Fiber Optic as DSL is covered by the intercept. Partner would be No as Yes is covered by the intercept. Finally, payment method would be electronic check, mailed check or Credit card check as Bank transfer is covered by the intercept. A combination of each categorical variable listed would need to be fitted to the β_n equation above.

This hypothesis was tested using the standard formulation for general linear hypothesis, results in Table 5. Using the anova( ) function in R with the F-statistic as the test, *internetservice* had a p-value of 0.0007 and a f-value of 7.35, *partner* had a p-value of 9.600e-05 with a f-value of 15.47, *monthlycharges* had a p-value of 2.631e-08 and a f-value of 6.6, *paymentmethod* had a p-value of 0.0002 and a f-value of 6.6.

There were two interactions found using the anova( ) function, *partner* and *monthlycharges* yielded a p-value of 0.043 with a f-value of 4.1.The second interaction is between *monthlycharges* and *paymentmethod* with a p-value of 0.00946 and a f-value of 3.8.

When using the drop1( ) function in R using the F-statistic as a test there are 2 similar interactions, which were *monthlycharges* and *paymentmethod* electronic check with a p-value of 0.0024 the other being between *partner* and *monthlycharges* with a p-value of 0.0429.

Upon examining the summary of the model, *paymentmethod* using electric check had a p-value of 0.0009, having a partner yielded a p-value of 0.04. An interaction between having a partner while using electronic check as a payment method had a p-value of 0.008, the p-value for the intercept of this model is 0.8713. As one could expect *monthlycharges* had a p-value of 0.0059 with a f-value of 8.531, the interaction between *monthlycharges* and *paymentmethod* yielded a p-value of 0.0078. All other values were not statistically significant to include in the report. When using the two sanitised datasets, the above statistics were found to be statistically significant.

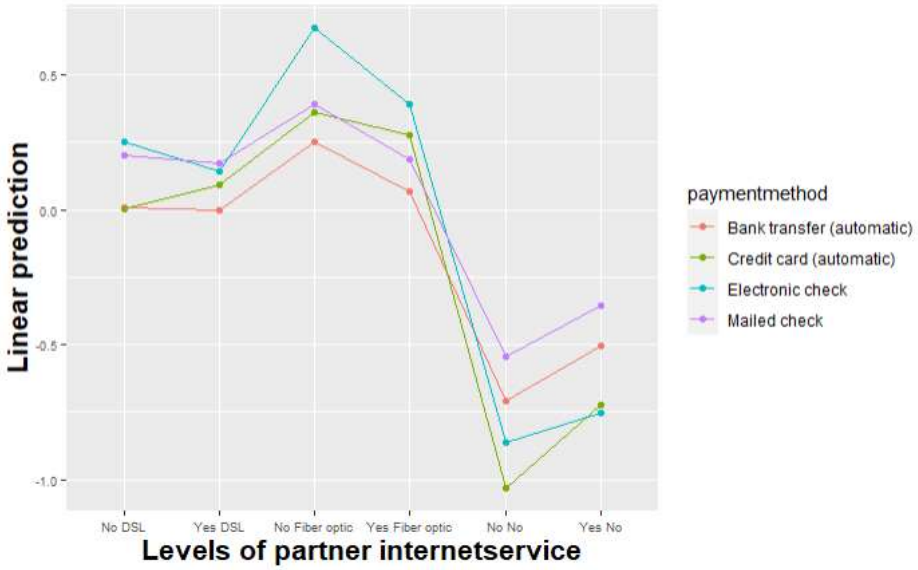

Figure 4: Interaction plot for model 1 of partner status and internet service used

Conclusion

The goal of this analysis was to see the probability of a specific client leaving meanwhile reviewing any predictors and possible interactions between them.

Based on Table 4 'users without internet' only 5% (5 users) of 100 customers contributed to churn in said group, meanwhile in Table 3 'users with internet' almost 20% (80 users) of 412 customers contributed to churn in said group. Based on data gathered in Tables 2, 3 and 4 customers without internet contribute 0.0097% (5/512) to **total customer churn** which is **16.6%**, whereas the latter contributes 15.6% (80/512) of total customer churn.

When trying to predict the probability whether a male living with a partner, who's monthly charges are 70 for fiber optic internet paid for with a credit card will leave the telecommunication provider it was found that the probability is 35% using the general linear model documented under model 1 with an upper bound of 0.53 and lower bound 0.17.

Table 6: Lower and upper confidence intervals for model 1 predictions on requested customer profile

upper_ci	lower_ci	prediciton_probability
0.53	0.17	0.35

It has been found that both monthly charges were the highest predictor of churn with a p-value of 0.0009, however when adjusting monthly charges for the specified user the higher the cost the lower the probability of them leaving, with an inverse effect as costs went down, albeit that could be a correlation. Based on the interaction plot which didn't include monthly charges, living without a partner while paying for fiber optic internet was the highest predictor of churn for all payment methods, meanwhile there was very little difference on this combination when fiber optic was swapped out for DSL in conjunction with payment method. This could infer those customers who pay for their internet services with electronic checks would have less use for it and/or income to pay for fiber optic internet as they may be retired, have it's only a speculation that can't be teste because age has not been included in this dataset. Living with a partner and monthly charges had an interaction effect with a p-value of 0.043 which would

Code sample

```
---
title: "Assigment 3 Linear and Logistic Regression Models - C15311966 Maks Drzezdzon"
output: html_notebook
notebook_link: https://github.com/Maks-Drzezdzon/Masters-Classes-L-
O/blob/master/Applied%20Statistics/assignment/assignment%203/assign_3_C15311966_Maks_Drzezdzon.R
md
---
```

```
library(pastecs) #For creating descriptive statistic summaries
library(ggplot2) #For creating histograms with more detail than plot
library(psych) # Some useful descriptive functions
library(semTools) #For skewness and kurtosis
library(FSA) # For percentage
library(car) # For Levene's test for homogeneity of variance
library(effectsize) # To calculate effect size for t-test
library(kableExtra) # Used to generate report ready tables
library(tidyverse) # data wrangling
library(gtsummary) # generate table for model results
library(multcomp) # needed for glht
library(emmeans)
require(ggiraph) # to use ggPredict
require(ggiraphExtra) # to use ggPredict
require(plyr)
library(plotly) # density plot
library(gridExtra)
library(PCAmixdata)


# Different subsets of data for experiments if time allows for it
untouched_dataset = read.csv("telecoms_churn.csv")
full_customer_dataset = subset(untouched_dataset, select = -c(X))
customer_dataset_user_has_internet =
full_customer_dataset[which(full_customer_dataset$internetservice != 'No'), ]
customer_dataset_user_has_internet = subset(customer_dataset_user_has_internet, select = -
c(internetservice))
customer_dataset_user_no_internet =
full_customer_dataset[which(full_customer_dataset$internetservice == 'No'), ]
customer_dataset_user_no_internet = subset(customer_dataset_user_no_internet, select = -
c(internetservice))


anovatab =
  function(mod){
    tab=as.matrix(anova(mod))
    rows=dim(tab)[1]
    moddf=sum(tab[,1])-tab[rows,1]
    ssmodel=sum(tab[,2])-tab[rows,2]
    msmodel=ssmodel/moddf
    f=msmodel/tab[rows,3]
    p=1-pf(f,moddf,tab[rows,1])
    tab2=tab[(rows-1):rows,]
    tab2[1,1:5]=c(moddf,ssmodel,msmodel,f,p)
    tab2=rbind(tab2,c(moddf+tab2[2,1],ssmodel+tab2[2,2],rep(NA,3)))
    rownames(tab2)=c('Model','Error','Total')
    colnames(tab2)[1]='df'
    return(print(tab2,na.print = "" , quote = FALSE,digits=3))
  }

get_var_name = function(var) {
  deparse(substitute(var))
}

get_plots = function(df, df_name, columns_to_assess, target){

  for (col in columns_to_assess){
      bar_plot = ggplot(df, aes_string(x=col,
                                       y=target,
```

```
                                    show.legend = T)) +
                  geom_bar(stat="identity") +
                  ggtitle(paste0("Effect of ", aes_string(x=col), " on customer churn for ",
df_name)) +
                  theme(plot.title = element_text(hjust = 0.5))
      print(bar_plot)

  }

    conditional_density_plot = ggplot(df, aes(x=monthlycharges, fill=factor(churn))) +
          geom_density(position = "stack") +
          xlab("monthlycharges") + labs(fill='churn') +
          ggtitle(paste0("Effect of ~monthlycharges on customer churn for ", df_name)) +
          theme(legend.text=element_text(size=12),
                axis.title=element_text(size=14),
                plot.title = element_text(size=11))


  return(conditional_density_plot)


}

combine_dfs_freq_table = function(columns, df){
  # hard coded data set full_customer_dataset
  # @columns => what columns will be present in the frequency table
  df_list = list()
  i=1
  for (col in cols){
    tmp_df = data.frame(round(prop.table(table(df[[col]])) * 100, 2))
    colnames(tmp_df)[1] = "Values"
    tmp_df['Column Name'] = col
    df_list[[i]] = tmp_df
    i=i+1
  }
  final_df = do.call(rbind, df_list)
  colnames(final_df)[2] = "% Representation of Column"
  final_df = final_df %>% relocate('Column Name', .before = "Values")
  return(final_df)
}

has_0 = full_customer_dataset[which(full_customer_dataset$churn == '0'), ]
has_1 = full_customer_dataset[which(full_customer_dataset$churn == '1'), ]

cols_list = colnames(full_customer_dataset)
cols = cols_list[cols_list != "churn" & cols_list != "customerid" & cols_list !=
"monthlycharges"]

get_plots(full_customer_dataset, get_var_name(full_customer_dataset), cols, "churn")
get_plots(customer_dataset_user_has_internet, get_var_name(customer_dataset_user_has_internet),
cols, "churn")
get_plots(customer_dataset_user_no_internet, get_var_name(customer_dataset_user_no_internet),
cols, "churn")

df = full_customer_dataset
kbl(df) %>%
 kable_classic(full_width = F)

cols_list = colnames(df)
cols = cols_list[cols_list != "customerid" & cols_list != "monthlycharges"]
df = subset(df, select=c(cols))
data=df
table_df = combine_dfs_freq_table(cols, df)
kbl(table_df) %>%
 kable_classic(full_width = F)

df = customer_dataset_user_has_internet
cols_list = colnames(df)
cols = cols_list[cols_list != "customerid" & cols_list != "monthlycharges"]
df = subset(df, select=c(cols))
```

```
table_df = combine_dfs_freq_table(cols, df)
kbl(table_df) %>%
 kable_classic(full_width = F)

df = customer_dataset_user_no_internet
cols_list = colnames(df)
cols = cols_list[cols_list != "customerid" & cols_list != "monthlycharges"]
df = subset(df, select=c(cols))
table_df = combine_dfs_freq_table(cols, df)
kbl(table_df) %>%
 kable_classic(full_width = F)

# (.)^2 maps an interaction between all variables such as x1:x2 + x1:x3 + x2:x3... etc etc
df = customer_dataset_user_no_internet
df = customer_dataset_user_has_internet

df = full_customer_dataset
cols_list = colnames(df)
cols = cols_list[cols_list != "customerid"]
data = subset(df, select=c(cols))

glm1 = glm(churn ~ (.)^2, data = data)

anova(glm1, test='F')
drop1(glm1, test='F')
summary(glm1)
vcov(glm1)
confint(glm1)

emmeans(glm1, ~monthlycharges+partner+paymentmethod)
lsmeans(glm1, ~paymentmethod:partner, adjust="fdr")

# interaction plot
emmip(glm1, paymentmethod~partner+internetservice) + theme(axis.text=element_text(size=7),
        axis.title=element_text(size=16, face="bold"))

glm1 %>% tbl_regression()

new_data = data.frame(gender="Male", partner="Yes", monthlycharges=70.0, internetservice="Fiber
optic", paymentmethod="Electronic check")
preds = predict(glm1, new_data, interval = "confidence")
preds
# or
preds = predict(glm1, new_data, se.fit = TRUE, type="response")
critval = qnorm(0.975) ##  95% CI
upr = preds$fit + (critval * preds$se.fit)
lwr = preds$fit - (critval * preds$se.fit)
fit = preds$fit
conf_inter = data.frame(upper_ci=round(upr,2), lower_ci=round(lwr,2),
prediciton_probability=round(fit,2))

kbl(conf_inter) %>%
 kable_classic(full_width = F)


scatter_plot = ggplot(df, aes(x=monthlycharges,
                              y=churn,
                              show.legend = T)) +
               geom_point(size=2) +
               ggtitle("Effect of temperature/additive on carbon fibre strength") +
               theme(plot.title = element_text(hjust = 0.5))
scatter_plot
```