Project

Fang Yu Lim(Fiona)

2024-03-06

```
churn_data = read.csv("../data/teleco_with_county.csv")
churn_original = read.csv("../data/Telco_customer_churn_cleaned.csv")

library(dplyr)
```

Case 3 and 4

```
# Create the best case (not churned) and worst case (churned group)
unknown_churn = filter(churn_data, churn_data$tenure_months < 12 & churn_data
$churn_value== 0) #1070
unknown_churn_best = filter(churn_data, churn_data$tenure_months  < 12 & chur
n_data$churn_value== 0) #1070
unknown_churn_worst = filter(churn_data, churn_data$tenure_months  < 12 & chu
rn_data$churn_value== 0) #1070
known_churn = churn_data %>%
  filter(! customerid %in% unknown_churn$CustomerID) #5973
unknown_churn_best[ , 'churn_12month'] = 0
unknown_churn_worst[ , 'churn_12month'] = 1
known_churn[ , 'churn_12month'] = known_churn$churn_value

best_case = rbind(known_churn, unknown_churn_best)
worst_case = rbind(known_churn, unknown_churn_worst)



# CASE 4 for each data set
best_case$churn_12month[best_case$Tenure > 12 & best_case$Churn_val == 1] = 0
worst_case$churn_12month[worst_case$Tenure > 12 & worst_case$Churn_val == 1]
= 0

# eliminate churn_val since we substitute them with 12 month churn_val using
two cases
best_case = best_case[-best_case$churn_val]
worst_case = worst_case[-worst_case$churn_val]
```

1) Is there a specific county with notably high or low churn rates.

```
library(ggplot2)
library(reshape2)

# Best case
proportion_churn = best_case%>%
  group_by(county) %>%
  summarize(proportion_churned = sum(churn_12month == 1) / n())

proportion_churn_table = as.data.frame(proportion_churn)

proportion_table_1 = proportion_churn_table %>%
  mutate(proportion_not_churned = 1 - proportion_churned)
```

```
#worst case
proportion_churn = worst_case%>%
  group_by(county) %>%
  summarize(proportion_churned = sum(churn_12month == 1) / n())

proportion_churn_table = as.data.frame(proportion_churn)

# Proportion table.
proportion_table_2 = proportion_churn_table %>%
  mutate(proportion_not_churned = 1 - proportion_churned)

# Best case
proportion_melted_1 = melt(proportion_table_1, id.vars = "county", variable.n
ame = "status", value.name = "proportion")

#Proportion plot
ggplot(proportion_melted_1, aes(x = county, y = proportion, fill = status)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(title = "Proportion of Churned vs. Not Churned Individuals by County",
       x = "County", y = "Proportion") +
  scale_fill_manual(values = c("lightcoral","lightblue"))+  # Using Set2 pale
tte from
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
# Worst case
#Transforming to suitable format for visualization
proportion_melted_2 = melt(proportion_table_2, id.vars = "county", variable.n
ame = "status", value.name = "proportion")

ggplot(proportion_melted_2, aes(x = county, y = proportion, fill = status)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(title = "Proportion of Churned vs. Not Churned Individuals by County",
       x = "County", y = "Proportion") +
  scale_fill_manual(values = c("lightcoral","lightblue")) +  # Automatic lege
nd labels
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

chi-square test of independence. H0: There is not association between the proportion of churned and not churned individuals by county. Ha: There is an association between the proportion of churned and not churned individuals by county.

```
# Best case
contingency_table_1 = table(best_case$county, best_case$churn_12month)
expected_counts_1 = outer(rowSums(contingency_table_1), colSums(contingency_t
able_1)) / sum(contingency_table_1)
sum(expected_counts_1 < 5)

## [1] 3

#Worst case
contingency_table_2 = table(worst_case$county, worst_case$churn_12month)
expected_counts_2 = outer(rowSums(contingency_table_2), colSums(contingency_t
able_2)) / sum(contingency_table_2)
sum(expected_counts_2 < 5)

# Best case
fisher_test_result_1 = fisher.test(contingency_table_1, simulate.p.value=TRUE
)
fisher_test_result_1



# Worst case
fisher_test_result_2 = fisher.test(contingency_table_2, simulate.p.value=TRUE
)
fisher_test_result_2
```

2) Is there a difference in churn (0,1) based on the usage of different services? (Including Phone Service, Multiple Lines, Internet Service, Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, and Streaming Movies.

```
# Best case
results_1 = list(
  phone_service = chisq.test(table(best_case$phone_service, best_case$churn_1
2month)),
  multiple_lines = chisq.test(table(best_case$multiple_lines, best_case$churn
_12month)),
  internet_service = chisq.test(table(best_case$internet_service, best_case$c
hurn_12month)),
  online_security = chisq.test(table(best_case$online_security, best_case$chu
rn_12month)),
  online_backup = chisq.test(table(best_case$online_backup, best_case$churn_1
```

```r
2month)),
  device_protection = chisq.test(table(best_case$device_protection, best_case
$churn_12month)),
  tech_support = chisq.test(table(best_case$tech_support, best_case$churn_12m
onth)),
  streaming_tv = chisq.test(table(best_case$streaming_tv, best_case$churn_12m
onth)),
  streaming_movies = chisq.test(table(best_case$streaming_movies, best_case$c
hurn_12month))
)

# Print the results
for (name in names(results_1)) {
  res = results_1[[name]]
  if (inherits(res, "htest")) {
    cat("Variable:", name, "\n")
    cat("Chi-Square Statistic:", res$statistic, "\t", "p-value:", res$p.value
, "\n")
    cat("\n")
  } else {
    cat("Error: Unable to perform chi-square test for", name, "\n")
  }
}

# worst case
results_2 = list(
  phone_service = chisq.test(table(worst_case$phone_service, worst_case$churn
_12month)),
  multiple_lines = chisq.test(table(worst_case$multiple_lines, worst_case$chu
rn_12month)),
  internet_service = chisq.test(table(worst_case$internet_service, worst_case
$churn_12month)),
  online_security = chisq.test(table(worst_case$online_security, worst_case$c
hurn_12month)),
  online_backup = chisq.test(table(worst_case$online_backup, worst_case$churn
_12month)),
  device_protection = chisq.test(table(worst_case$device_protection, worst_ca
se$churn_12month)),
  tech_support = chisq.test(table(worst_case$tech_support, worst_case$churn_1
2month)),
  streaming_tv = chisq.test(table(worst_case$streaming_tv, worst_case$churn_1
2month)),
  streaming_movies = chisq.test(table(worst_case$streaming_movies, worst_case
$churn_12month))
)

# Print the results
for (name in names(results_2)) {
  res = results_2[[name]]
  if (inherits(res, "htest")) {
    cat("Variable:", name, "\n")
    cat("Chi-Square Statistic:", res$statistic, "\t", "p-value:", res$p.value
, "\n")
    cat("\n")
  } else {
    cat("Error: Unable to perform chi-square test for", name, "\n")
  }
}
```

3) Are users using more services less likely to churn?

```
# best case
num_services = numeric(nrow(best_case))
for (i in 1:nrow(best_case)){
  num =0

  #Phone Service
  if (best_case[i, "phone_service"] == "Yes"){
    num = num +1
  }

  #Multiple Lines
  if (best_case[i, "multiple_lines"] == "Yes"){
    num = num +1
  }

  #Internet Service
  if (best_case[i, "internet_service"] == "DSL" | best_case[i, "internet_serv
ice"] == "Fiber optic") {
    num = num + 1
  }

  #Online Security
  if (best_case[i, "online_security"] == "Yes"){
    num = num +1
  }

  #Online Backup
  if (best_case[i, "online_backup"] == "Yes"){
    num = num +1
  }

  #Device Protection
  if (best_case[i, "device_protection"] == "Yes"){
    num = num +1
  }

  #Tech Support
  if (best_case[i, "tech_support"] == "Yes"){
    num = num +1
  }

  #Streaming movies
  if (best_case[i, "streaming_movies"] == "Yes"){
    num = num +1
  }

  #Streaming TV
  if (best_case[i, "streaming_tv"] == "Yes"){
    num = num +1
  }

   num_services[i] = num
}

best_case$num_service = num_services

chisq.test(table(best_case$num_service, best_case$churn_12month))

# Worst case
num_services = numeric(nrow(worst_case))
for (i in 1:nrow(worst_case)){
  num =0

  #Phone Service
  if (worst_case[i, "phone_service"] == "Yes"){
    num = num +1
  }

  #Multiple Lines
  if (worst_case[i, "multiple_lines"] == "Yes"){
    num = num +1
  }

  #Internet Service
  if (worst_case[i, "internet_service"] == "DSL" | worst_case[i, "internet_se
rvice"] == "Fiber optic") {
    num = num + 1
  }

  #Online Security
  if (worst_case[i, "online_security"] == "Yes"){
    num = num +1
  }

  #Online Backup
  if (worst_case[i, "online_backup"] == "Yes"){
    num = num +1
  }
  #Device Protection
  if (worst_case[i, "device_protection"] == "Yes"){
    num = num +1
  }

  #Tech Support
  if (worst_case[i, "tech_support"] == "Yes"){
    num = num +1
  }

  #Streaming movies
  if (worst_case[i, "streaming_movies"] == "Yes"){
    num = num +1
  }

  #Streaming TV
  if (worst_case[i, "streaming_tv"] == "Yes"){
    num = num +1
  }

   num_services[i] = num
}

worst_case$num_service = num_services

chisq.test(table(worst_case$num_service, worst_case$churn_12month))

library(ggplot2)

ggplot(worst_case, aes(x = factor(num_service), fill = factor(churn_12month))
) +
  geom_bar(position = "dodge", color = "black") +
  labs(x = "Number of Services Used", y = "Count", fill = "Churn Status") +
  scale_fill_manual(values = c("lightblue", "lightcoral"),
                    labels = c("Not Churned", "Churned")) +  # Specify colors
 directly
  theme_minimal() +
  scale_x_discrete(breaks = as.character(1:9))  # Specify breaks for x-axis
```

```
# Best case
table(best_case$churn_12month)

##
##    0    1
## 6244 1869
```

```r
churned_data_1 = best_case[best_case$churn_12month == 1, "num_service"]
not_churned_data_1 = best_case[best_case$churn_12month == 0, "num_service"]

# Perform independent samples t-test
t_test_result_1 = t.test(churned_data_1, not_churned_data_1)

# Print the result
print(t_test_result_1)

# Worst case
table(worst_case$churn_12month)

churned_data_2 = worst_case[worst_case$churn_12month == 1, "num_service"]
not_churned_data_2 = worst_case[worst_case$churn_12month == 0, "num_service"]

# Perform independent samples t-test
t_test_result_2 = t.test(churned_data_2, not_churned_data_2)

# Print the result
print(t_test_result_2)

# best case : high < low < medium
# Separate to low medium high

best_case$service_level = ifelse(best_case$num_service >= 1 & best_case$num_service <= 3, "low",
                                   ifelse(best_case$num_service >= 4 & best_case$num_service <= 6, "medium",
                                          ifelse(best_case$num_service >= 7 & best_case$num_service <= 9, "high", NA)))

# Calculate total number of observations in each group
total_low_1 = sum(best_case$service_level == "low")
total_medium_1 = sum(best_case$service_level == "medium")
total_high_1 = sum(best_case$service_level == "high")

# Calculate number of churners in each group
churners_low_1 = sum(best_case$churn_12month[best_case$service_level == "low"] == 1)
churners_medium_1 = sum(best_case$churn_12month[best_case$service_level == "medium"] == 1)
churners_high_1 = sum(best_case$churn_12month[best_case$service_level == "high"] == 1)

# Calculate proportion of churners in each group
proportion_churners_low_1 = churners_low_1 / total_low_1
proportion_churners_medium_1 = churners_medium_1 / total_medium_1
proportion_churners_high_1 = churners_high_1 / total_high_1

# Print the proportions
```

```
prop_table_1 = matrix(c(proportion_churners_low_1, proportion_churners_medium
_1, proportion_churners_high_1),
                     nrow = 1, byrow = TRUE)

# Convert the matrix to a data frame for better visualization
prop_table_df_1 = as.data.frame(prop_table_1)

# Assign column names
colnames(prop_table_df_1) = c("Low", "Medium", "High")

# Print the proportion table
print(prop_table_df_1)

# Perform chi-square test of independence
chi_square_result_1 = chisq.test(best_case$service_level, best_case$churn_12m
onth)

# Print the test result
print(chi_square_result_1)

pairwise.prop.test(table(best_case$service_level, best_case$churn_12month))

# Create a bar plot
barplot(c(proportion_churners_low_1, proportion_churners_medium_1, proportion
_churners_high_1),
        names.arg = c("Low", "Medium", "High"),
        xlab = "Service Level",
        ylab = "Proportion of Churners",
        main = "Proportion of Churners by Service Level")

# Worst case: high < medium < low
# Separate to low medium high

worst_case$service_level = ifelse(worst_case$num_service >= 1 & worst_case$nu
m_service <= 3, "low",
                                   ifelse(worst_case$num_service >= 4 & worst
_case$num_service <= 6, "medium",
                                          ifelse(worst_case$num_service >= 7
& worst_case$num_service <= 9, "high", NA)))

# Calculate total number of observations in each group
total_low_2 = sum(worst_case$service_level == "low")
total_medium_2 = sum(worst_case$service_level == "medium")
total_high_2 = sum(worst_case$service_level == "high")

# Calculate number of churners in each group
churners_low_2 = sum(worst_case$churn_12month[worst_case$service_level == "lo
w"] == 1)
churners_medium_2 = sum(worst_case$churn_12month[worst_case$service_level ==
"medium"] == 1)
```

```r
churners_high_2 = sum(worst_case$churn_12month[worst_case$service_level == "h
igh"] == 1)

# Calculate proportion of churners in each group
proportion_churners_low_2 = churners_low_2 / total_low_2
proportion_churners_medium_2 = churners_medium_2 / total_medium_2
proportion_churners_high_2 = churners_high_2 / total_high_2

# Print the proportions
prop_table_2 = matrix(c(proportion_churners_low_2, proportion_churners_medium
_2, proportion_churners_high_2),
                     nrow = 1, byrow = TRUE)

# Convert the matrix to a data frame for better visualization
prop_table_df_2 = as.data.frame(prop_table_2)

# Assign column names
colnames(prop_table_df_2) = c("Low", "Medium", "High")

# Print the proportion table
print(prop_table_df_2)

# Perform chi-square test of independence
chi_square_result_2 = chisq.test(worst_case$service_level, worst_case$churn_1
2month)

# Print the test result
print(chi_square_result_2)

pairwise.prop.test(table(worst_case$service_level, worst_case$churn_12month))

# Create a bar plot
barplot(c(proportion_churners_low_2, proportion_churners_medium_2, proportion
_churners_high_2),
        names.arg = c("Low", "Medium", "High"),
        xlab = "Service Level",
        ylab = "Proportion of Churners",
        main = "Proportion of Churners by Service Level")
```

An explanation of the results of the best case in real-world scenario is that users who have higher service usage usually have a good experience then starting using more services. At the same time, this group of users might receive exceptional service, and personalized attention. On the other hand, customers with low service may have chosen a company to minimize expenses and are willing to tolerate the trade-offs, have limited expectations and are satisfied with basic services. As for customers with medium service level, there might be multiple options available for them at a similar price range resulting in these customers to churn easier than customers with low service usage.