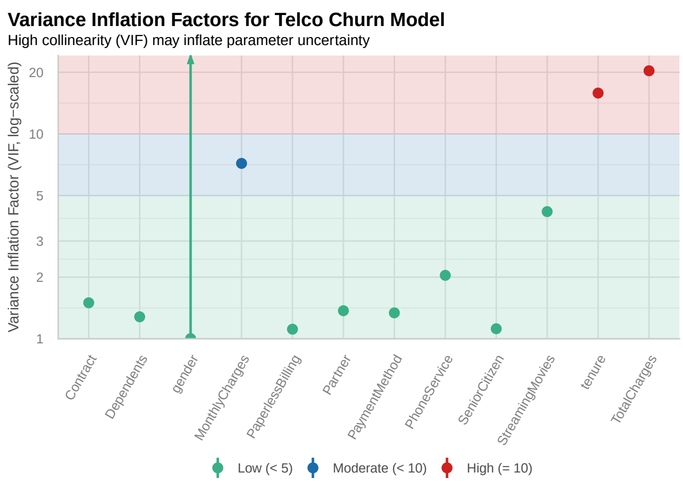
Assignment 2 - Strategic Marketing Decision Making 2025-2026

Gonçalo Passinhas (732518)

2025-09-30

As can be seen in the plot below, the variance inflation factor (VIF) results show that most variables have a Low Correlation with a value of below 5. As can be seen, "gender" has a VIF of 1 , SeniorCitizen 1.12 , Partner 1.37, Dependents 1.28, PhoneService 2.04, StreamingMovies 4.18, Contract 1.5, PaperlessBilling 1.11, and PaymentMethod 1.34. These values all fall below 5, indicating low levels of multicollinearity. 7.18 has a VIF of 7.18, which points to a moderate level of correlation with other predictors.



In contrast, both TotalCharges and tenure stand out with very high VIFs of 20.35 and 15.84, respectively. These elevated values signal severe multicollinearity, suggesting that these variables overlap strongly with other predictors in the model. As a result, there is a risk of inflated standard errors and unstable coefficient estimates. To address this issue, it may be necessary to consider removing or combining these variables, or to apply regularization techniques such as ridge or lasso regression to mitigate the effects of multicollinearity.

The results in Table 1 indicate that tenure, contract type, and phone service all play an important role in explaining customer churn. The coefficient for tenure is negative and statistically significant ($\beta_{Tenure} = -0.038$, $p < 0.001$), which means that customers with a longer history at the company are less likely to leave. Expressed in odds ratios, each additional month of tenure decreases the odds of churn by about 4% ($oddsratio = 0.96$). This finding suggests that customer loyalty grows over time: the longer a customer remains with the company, the more stable the relationship becomes, and the lower the likelihood of churn.

Similarly, the coefficient for being on a one-year contract is also negative and highly significant ($\beta_{ContractOneYear} = -0.862$, $p < 0.001$). The odds ratio of approximately 0.42 implies that customers with a one-year contract are about 58% less likely to churn compared to the baseline group, which is most likely month-to-month contract holders. This demonstrates the strong stabilizing effect of longer contract commitments, as customers who agree to a fixed-term contract tend to be more engaged and less inclined to switch providers.

Finally, the presence of phone service shows a significant negative effect on churn as well ($\beta_{PhoneServiceYes} = -0.845$, $p < 0.001$). The corresponding odds ratio of around 0.43 indicates that customers with phone service are about 57% less likely to churn than those without it. This points to the value of bundled or additional services in customer retention. By offering phone service, the company increases customer stickiness, making it less attractive for customers to leave for competitors.

Table 1: Logistic Regression Results

term	estimate	std.error	statistic	p.value
(Intercept)	-1.293	0.161	-8.045	0.000
MonthlyCharges	0.029	0.003	10.323	0.000
genderMale	-0.008	0.064	-0.123	0.902
SeniorCitizen	0.313	0.083	3.759	0.000
PartnerYes	0.005	0.077	0.065	0.948
DependentsYes	-0.208	0.088	-2.350	0.019
tenure	-0.038	0.002	-16.905	0.000
PhoneServiceYes	-0.845	0.152	-5.572	0.000
StreamingMoviesNo internet service	0.187	0.175	1.069	0.285
StreamingMoviesYes	-0.012	0.087	-0.138	0.890
ContractOne year	-0.862	0.104	-8.312	0.000
ContractTwo year	-1.699	0.170	-9.984	0.000
PaperlessBillingYes	0.396	0.073	5.431	0.000
PaymentMethodCredit card (automatic)	-0.095	0.113	-0.847	0.397
PaymentMethodElectronic check	0.399	0.093	4.285	0.000
PaymentMethodMailed check	-0.061	0.112	-0.545	0.586

Taken together, these results highlight that tenure, contractual arrangements, and service features are strong protective factors against churn. Customers who stay longer, commit to longer contracts, or subscribe to additional services are significantly more likely to remain loyal, underscoring the importance of customer relationship management strategies that foster retention through stability and service bundling.

Recommendation 1. Encourage longer-term contracts through targeted incentives. The regression results show that customers on one-year contracts are significantly less likely to churn compared to those on month-to-month agreements. This suggests that contract length is a strong protective factor against customer attrition. Management should therefore promote longer-term commitments by offering targeted incentives, such as discounted monthly rates, additional service features, or loyalty bonuses for customers who switch to annual or multi-year contracts. By reducing the proportion of month-to-month customers, the company can increase customer stickiness and stabilize its subscriber base.

Recommendation 2. Promote bundled offerings that include phone service. The analysis also indicates that customers with phone service are substantially less likely to churn than those without it. This highlights the value of bundling services to strengthen customer relationships. The firm should develop and market attractive bundle packages that include phone service alongside internet or TV products, potentially at a slight discount relative to purchasing services individually. Such cross-selling not only increases average revenue per user but also enhances customer retention, as customers who adopt additional services become more embedded in the company's ecosystem and less likely to switch providers.

Predicted Churn Probability Tables. Using the logistic regression model from before (excluding TotalCharges), we computed predicted churn probabilities for all customers in the data set. To illustrate the range of predicted outcomes, we report two lists:

- Top 10 customers with the highest predicted churn probabilities which represent customers the model identifies as most at risk of leaving.
- Top 10 customers with the lowest predicted churn probabilities – which represent customers the model identifies as least likely to churn.

Each list includes the customer ID, the predicted churn probability, and the actual churn outcome, which allows us to assess the model's alignment with reality.

Table 2: Top 10 Customers with Highest Predicted Churn Probabilities

customerID	Pred_Prob	Churn
1400-MMYXY	0.86751	Yes
6496-SLWHQ	0.86393	Yes
7216-EWTRS	0.86010	Yes
3292-PBZEJ	0.84946	No
3389-YGYAI	0.84318	Yes
2265-CYWIV	0.84139	Yes
5178-LMXOP	0.83819	Yes
2081-VEYEH	0.83481	No
8884-ADFVN	0.83412	Yes
9300-AGZNL	0.83317	Yes

As can be seen in Table 2, the high-risk group has predicted probabilities above 0.83, and most of these customers actually churned, which suggests the model is capturing meaningful churn risk. There are, however, a few cases where the model predicted a high probability of churn but the customer did not churn. These cases illustrate the probabilistic nature of logistic regression: the model can highlight elevated risk, but it cannot guarantee individual outcomes.

Table 3: Top 10 Customers with Lowest Predicted Churn Probabilities

customerID	Pred_Prob	Churn
2848-YXSMW	0.00221	No
0784-ZQJZX	0.00223	No
6928-ONTRW	0.00225	No
4086-WITJG	0.00231	No
3173-WSSUE	0.00233	No
1052-QJIBV	0.00234	No
3279-DYZQM	0.00238	No
0831-JNISG	0.00240	No
1403-GYAFU	0.00243	No
4957-SREEC	0.00245	No

In contrast, the low-risk group has probabilities around 0.002, and all of these customers indeed remained with the company, which is shown in Table 3. This indicates that the model is especially effective at identifying customers with very low churn likelihood.

Taken together, these results show that while the model is not perfect, it is reasonably effective at separating customers into groups with very different churn risks. The top-10/high-risk list highlights where retention strategies may be most valuable, while the bottom-10/low-risk list confirms that the model reliably identifies stable customers.

Appendix: Full R Code for Reproducibility

```r
# ==================================================
# Telco Churn Analysis
# ==================================================

# ----------------------------------------------
# Load Packages
# ----------------------------------------------
library(gt)
library(ggplot2)
library(tidyverse)
library(knitr)
library(performance)
library(dplyr)
library(broom)
library(kableExtra)

# ----------------------------------------------
# Load Data
# ----------------------------------------------
Telco_Data <- read_csv("~/Downloads/Telco.csv")
str(Telco_Data)

# ----------------------------------------------
# Options
# ----------------------------------------------
options(scipen = 999)

# ----------------------------------------------
# Convert character variables to factors
# ----------------------------------------------
Telco_Data[sapply(Telco_Data, is.character)] <-
  lapply(Telco_Data[sapply(Telco_Data, is.character)], as.factor)

str(Telco_Data)

# ----------------------------------------------
# Logistic Regression Model 1 (with collinearity check)
# ----------------------------------------------
LR.Telco <- glm(
  Churn ~ MonthlyCharges + TotalCharges + gender + SeniorCitizen + Partner +
    Dependents + tenure + PhoneService + StreamingMovies + Contract +
    PaperlessBilling + PaymentMethod,
  data = Telco_Data,
  family = "binomial"
)

summary(LR.Telco)

# ----------------------------------------------
# Check Collinearity
# ----------------------------------------------
```

```r
collinearity_results <- check_collinearity(LR.Telco)
collinearity_results

plot(collinearity_results) +
  ggtitle("Variance Inflation Factors for Telco Churn Model") +
  theme(
    axis.text.x = element_text(angle = 60, hjust = 1, size = 9),
    plot.title = element_text(hjust = 0, size = 13, face = "bold")
  )

# -----------------------------------------------
# Logistic Regression Model 2 (reduced)
# -----------------------------------------------
LR.Telco1 <- glm(
  Churn ~ MonthlyCharges + gender + SeniorCitizen +
    Partner + Dependents + tenure + PhoneService +
    StreamingMovies + Contract + PaperlessBilling + PaymentMethod,
  data = Telco_Data,
  family = binomial
)

summary(LR.Telco1)

tidy(LR.Telco1) |>
  kable(digits = 3, caption = "Logistic Regression Results \\label{tab:logit}")

# -----------------------------------------------
# Predicted Probabilities
# -----------------------------------------------
model <- glm(
  Churn ~ MonthlyCharges + gender + SeniorCitizen + Partner +
    Dependents + tenure + PhoneService + StreamingMovies +
    Contract + PaperlessBilling + PaymentMethod,
  data = Telco_Data, family = "binomial"
)

Telco_Data$Pred_Prob <- predict(model, type = "response")

# Select relevant columns
Results <- Telco_Data[, c("customerID", "Pred_Prob", "Churn")]
Results

# -----------------------------------------------
# Top 10 Highest & Lowest Probabilities
# -----------------------------------------------
top10_high <- Results %>%
  arrange(desc(Pred_Prob)) %>%
  head(10)

top10_low <- Results %>%
  arrange(Pred_Prob) %>%
  head(10)
```

```r
# ------------------------------------------------
# Round and Display Results
# ------------------------------------------------
top10_high_round <- top10_high %>%
  mutate(across(where(is.numeric), ~ round(.x, 5)))

kable(top10_high_round, caption = "Top 10 Customers with Highest Predicted Churn Probabilities") %>%
  kable_styling(
    latex_options = c("hold_position", "scale_down"),
    font_size = 9
  )

top10_low_round <- top10_low %>%
  mutate(across(where(is.numeric), ~ round(.x, 5)))

kable(top10_low_round, caption = "Top 10 Customers with Lowest Predicted Churn Probabilities") %>%
  kable_styling(
    latex_options = c("hold_position", "scale_down"),
    font_size = 9
  )
```