Assignment 2

Group Assignment 2

Rishi Ashok Kumar (560527) Nicolas Gonzalez (780037)
Aleksandra Tatko (648925) André van der Meij (589994)

September 30, 2025

Question 1

Run a single logistic regression model on the full sample of customers. The logistic regression model should predict the probability that a customer churns by using the following variables: monthly charges, total charges, gender, senior citizen status, partner, dependents, tenure, phone service, streaming movies, contract type, paperless billing, and payment method. Make sure to convert all categorical variables to factors for appropriate analysis. Note that some variables are already excluded to avoid multicollinearity. Report only the variance inflation factors (VIFs) for each variable. What do you conclude? (No need to report the final model output.)

To test for multicollinearity, the variance inflation factors (VIFs) of all predictors in the logistic regression model were examined.

Table 1: VIF Values for each variable

Variable	VIF
MonthlyCharges	2.68
TotalCharges	4.51
gender	1.00
SeniorCitizen	1.06
Partner	1.17
Dependents	1.13
tenure	3.98
PhoneService	1.43
StreamingMovies	1.43
Contract	1.11
PaperlessBilling	1.06
PaymentMethod	1.05

Most predictors show a GVIF value below 2, indicating that multicollinearity is not a concern for the majority of the variables as it is way below the accaptable treshold of 3. However, three predictors—**MonthlyCharges**, **TotalCharges**, and **tenure**—exhibit GVIF values above 2. This outcome is expected, as these variables are inherently related:

$$\text{TotalCharges} \approx \text{MonthlyCharges} \times \text{tenure}$$

Among them, **TotalCharges** shows the highest degree of multicollinearity. Keeping it alongside **tenure** may reduce the clarity of the model due to overlapping effects. To enhance interpretability, we will exclude **TotalCharges** from the final specification.

Question 2

Exclude total charges from the model in Question 1 and re-run the logistic regression. Present the final model output. Select three significant variables and interpret their coefficients (i.e., sign, significance, exact relationship between independent and dependent variables).

TotalCharges is excluded and the logistic regression is re-ran. Table 2 shows the results.

Table 2: Results from the Logistic Model

Term	Estimate	Std. Error	z value	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

Interpretation:

1. **PhoneServiceYes**: Customers with phone service have 57% (exp(-0.8447) 0.43) lower odds of churning compared to those without phone service, holding all other factors constant.
2. **Contract two year**: Customers on a two-year contract have 82% (exp(-1.6995) 0.18) lower odds of churning compared to those on a month-to-month contract, ceteris paribus. This suggest that customers with long- term contracts have higher probability of staying in the company.
3. **PaymentMethodElectronic check**: Customers who pay by electronic check have 49% (exp(0.3986) 1.49) higher odds of churning compared to those paying by automatic bank transfer, ceteris paribus. This suggest that customers with electronic check as a payment method are in a high-risk group for churn.

Question 3

Provide two actionable recommendations in light of the model results of Question 2 (i.e., the one excluding total charges). Provide brief justification for each recommendation.

1. Promote long-term contracts with added benefits. Since customers on two-year contracts have 82% lower odds of churning compared to month-to-month customers, the company should encourage longer contracts. This can be done by offering special discounts, loyalty perks, or bundled services (e.g. internet + streaming) to incentivize signing longer contract. Strengthening these offers will help lock in customers and reduce churn risk.
2. Improve retention strategies for high-risk payment methods. Customers paying by electronic check have 49% higher odds of churning compared to those using automatic bank transfers. The company should

target this group with retention campaigns, such as promoting automatic payment options, offering small discounts for switching payment methods, or providing reminders and support for electronic check users. Addressing this high-risk segment directly can reduce churn.

Question 4

Use the model of Question 2 (i.e., the one excluding total charges). Generate and provide two lists: one with the 10 customers having the highest predicted churn probabilities and another with the 10 customers having the lowest predicted churn probabilities. Each list should include the customer ID, the predicted churn probability, and the actual churn outcome. Present these results in a well-organized table.

Comparison between Top and Bottom.

Top 10

ID	Prediction	Churn
1400−MMYXY	0.868	Yes
6496−SLWHQ	0.864	Yes
7216−EWTRS	0.860	Yes
3292−PBZEJ	0.849	No
3389−YGYAI	0.843	Yes
2265−CYWIV	0.841	Yes
5178−LMXOP	0.838	Yes
2081−VEYEH	0.835	No
8884−ADFVN	0.834	Yes
9300−AGZNL	0.833	Yes

Bottom 10

ID	Prediction	Churn
2848−YXSMW	0.002	No
0784−ZQJZX	0.002	No
6928−ONTRW	0.002	No
4086−WITJG	0.002	No
3173−WSSUE	0.002	No
1052−QJIBV	0.002	No
3279−DYZQM	0.002	No
0831−JNISG	0.002	No
1403−GYAFU	0.002	No
4957−SREEC	0.002	No