Cian Kieffer Lim

The first thing I did for predicting the results is I created dummy variables for all non categorical variables. Initially, I planned to create tables and charts so that I can see which variables are the most important. However, I did not understand the code that much.

Afterwards, I decided to run the model with all the independent variables excluding customer ID and the dependent variable. The following dependent variables and independent variables are listed below.

Independent Variables		Dependent Variables
Gender	Senior Citizen	Churn
Parnter	Dependents	
Tenure	Phone Service	
Multiple Lines	Internet Service	
Online Security	Online Backup	
Device Protection	TechSupport	
StreamingTV	StreamingMovies	
Contract	Paperless Billing	
Payment Method		

After the code ran, I noticed that there were some NaN in my logistic regression result as seen below.

nk transfer (automatic)	-0.2392	7.55e+07	-3.17e-09	1.000	-1.48e+08	1.48e+08
Credit card (automatic)	-0.3160	nan	nan	nan	nan	nan
Electronic check	-0.1982	nan	nan	nan	nan	nan
ParnterYes	0.0462	0.089	0.518	0.605	-0.129	0.221

I tried to fix these errors but I could not do it. I decided to remove these variables or factors in the model because these factors did not seem important in deciding on churning, logically. Moreover, I also removed constants who has a confidence interval from a negative number to a positive number. I was left with these independent variables: senior citizen, tenure, monthly

charges, phone service, multiple lines, contract, and paperless billing. Moreover, I also did a quick training and testing set in the training set with 80%-to-20% ratio to see how reliable my model is. I ran it a couple of times and the testing accuracy and training accuracy lies between 78% - 80%. Below are photos of the final results of the model.

```
Training Accuracy: 0.7937631528601492
Test Accuracy: 0.8056618209640398
```

	coef	std err	z	P>|z|	[0.025	0.975]
SeniorCitizen	0.3347	0.094	3.578	0.000	0.151	0.518
tenure	-0.0389	0.002	-15.661	0.000	-0.044	-0.034
MonthlyCharges	0.0266	0.002	13.834	0.000	0.023	0.030
PhoneServiceYes	-1.9384	0.124	-15.609	0.000	-2.182	-1.695
MultipleLinesNPS	-1.0204	0.137	-7.429	0.000	-1.290	-0.751
MultipleLinesYes	0.2283	0.092	2.487	0.013	0.048	0.408
One year	-0.9976	0.118	-8.463	0.000	-1.229	-0.767
Two year	-1.9755	0.207	-9.558	0.000	-2.381	-1.570
PaperlessBillingYes	0.4452	0.084	5.284	0.000	0.280	0.610

The final logistic regression equation I utilized was:

Log(x)=0.33(SC)-0.04(tenure)+0.03(MC)-1.94(PS)-1.02(MLNoPhoneService)+0.2283(MLYes)-0.99(ContractOneYear)-1.98(ContractTwoYear)+0.4452(PaperBillingYes)

As we can see, there are less people who churn with longer contracts, no phone service. On the other hand, there are more people who churn with higher monthly charges, a senior citizen, multiple lines, and has paperless billing. I would recommend the company to have more monthly initiatives and promotions for the senior citizens. At the same time, they should be more wary and meticulous on their paperless billing process since I believe that a lot of people churns because paperless billing is not quite as effective as paper billing in reminding the customers to pay, thus incurring other miscellaneous fees.