PITAO, ANGELICA JAYNE L.

At first, I used Logistic Regression to predict whether or not a customer would churn. I transformed the train data into binary values and created dummy variables to categorize the answers. By running it in SPSS, I was able to form an equation that can predict the probability of churning.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	GenderDum	-.013	.067	.035	1	.851	.987
	SeniorCitizen	.191	.088	4.727	1	.030	1.210
	PartnerDum	.008	.081	.009	1	.923	1.008
	DependentsDum	-.131	.093	1.987	1	.159	.877
	Tenure	-.062	.007	90.995	1	.000	.940
	PhoneServiceDum	.470	.674	.486	1	.486	1.600
	MultipleLinesDum2	.513	.184	7.755	1	.005	1.670
	InternetServiceDum1	2.145	.838	6.546	1	.011	8.538
	InternetServiceDum2	4.329	1.656	6.834	1	.009	75.902
	OnlineSecurityDum2	-.125	.185	.457	1	.499	.882
	OnlineBackupDum2	.098	.182	.289	1	.591	1.103
	DeviceProtectionDum2	.258	.183	1.979	1	.160	1.294
	TechSupportDum2	-.127	.188	.456	1	.499	.881
	StreamingTVDum2	.791	.340	5.432	1	.020	2.206
	StreamingMoviesDum2	.766	.340	5.081	1	.024	2.150
	ContractDum1	-.660	.112	34.841	1	.000	.517
	ContractDum2	-1.404	.186	56.867	1	.000	.246
	PaperlessBillingDum	.379	.078	23.902	1	.000	1.461
	PaymentMethodDum1	-.352	.100	12.408	1	.000	.703
	PaymentMethodDum2	-.315	.098	10.238	1	.001	.730
	PaymentMethodDum3	-.429	.101	17.876	1	.000	.651
	MonthlyCharges	-.057	.033	3.003	1	.083	.944
	TotalCharges	.000	.000	22.508	1	.000	1.000
	Constant	-.267	.208	1.653	1	.199	.765

a. Variable(s) entered on step 1: GenderDum, SeniorCitizen, PartnerDum, DependentsDum, Tenure, PhoneServiceDum, MultipleLinesDum2, InternetServiceDum1, InternetServiceDum2, OnlineSecurityDum2, OnlineBackupDum2, DeviceProtectionDum2, TechSupportDum2, StreamingTVDum2, StreamingMoviesDum2, ContractDum1, ContractDum2, PaperlessBillingDum, PaymentMethodDum1, PaymentMethodDum2, PaymentMethodDum3, MonthlyCharges, TotalCharges.

Logistic Regression Equation:

ln(odds) = -.267 - .013(GenderDum) + .191(SeniorCitizen) + .008(PartnerDum) - .131(DependentsDum) -.062(Tenure) + .470(PhoneServiceDum) + .513(MultipleLinesDum2) + 2.145(InternetServiceDum1) + 4.329(InternetServiceDum2) - .125(OnlineSecurityDum2) + .098(OnlineBackupDum2) + .258(DeviceProtectionDum2)- .127(TechSupportDum2) + .791(StreamingTVDum2) + .766(StreamingMoviesDum2) - .660(ContractDum1) – 1.4004(ContractDum2) + .379(PaperlessBillingDum) - .352(PaymentMethodDum1)- .315(PaymentMethodDum2) - .429(PaymentMethodDum3) - .0057(MonthlyCharges)

Based on the significance or p-value of each coefficient, variables such as Senior Citizen, Tenure, MultipleLinesDum2, InternetServiceDum1, InternetServiceDum2, StreamingTVDum2, ContractDum1, ContractDum2, PaperlessBillingDum, PaymentMethodDum1, PaymentMethodDum2, PaymentMethodDum3, and TotalCharges are significant because they are less than alpha=.05. All other variables are insignificant.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	21.348	8	.006

Therefore, transforming the test dataset to include dummy variables and making use of the equation formulated by SPSS, I attempted to predict whether or not a customer would churn. However, the Hosmer-Lemeshow Test in SPSS which measures how good the logistic regression is in predicting the odds presented a significance level of only 0.006 which tells us that the regression equation is very inaccurate.

Finally, I looked into using KNIME to formulate a decision tree. The program created a predictor for churning based on the train dataset which I applied to the test dataset, allowing me to come up with my predictions in the excel file.