TELCO CHURN PREDICTION MODEL

Ng Chen Yong (Bryan)

AGENDA

EXPLORATORY DATA ANALYSIS

DATA PREPROCESSING

MODEL DEVELOPMENT AND RESULTS

INSIGHTS AND RECOMMENDATIONS

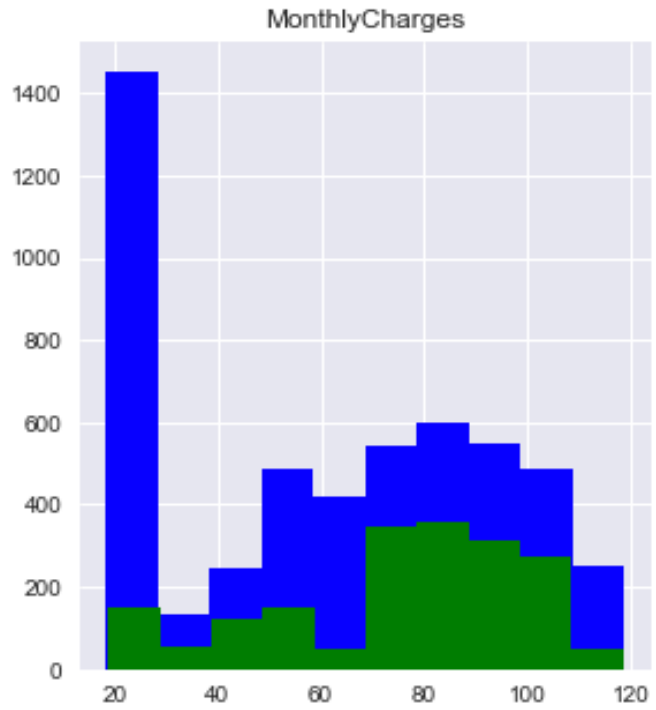
EXPLORATORY DATA ANALYSIS

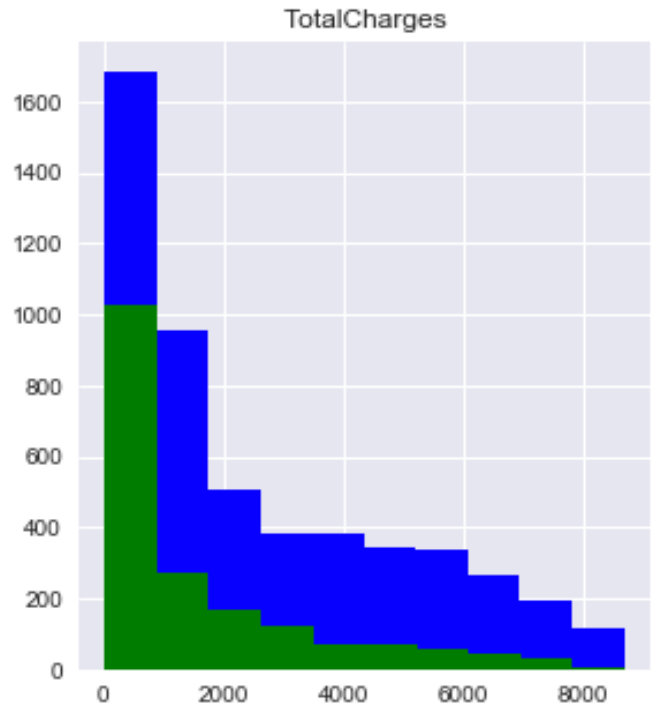
COLUMNS TYPE

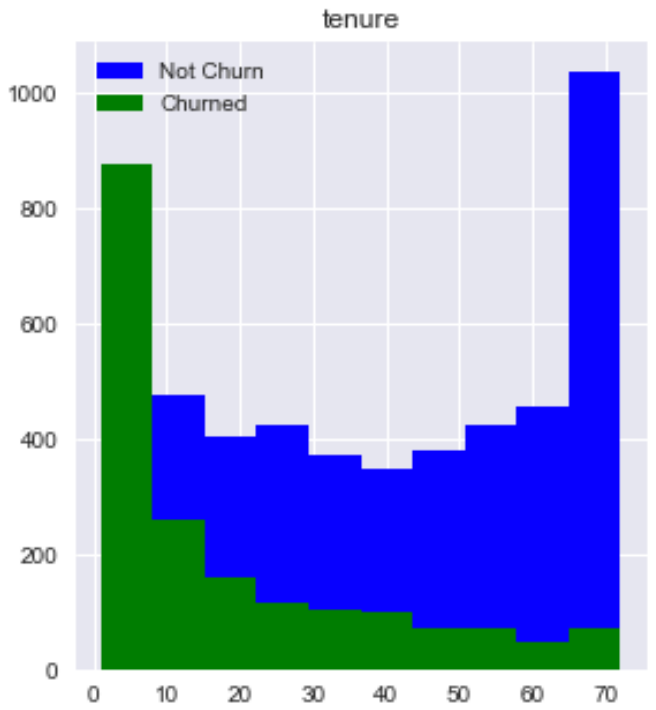
Numerical Variable	Categorical Variable	
• Tenure • MonthlyCharges • TotalCharges	• Gender • SeniorCitizen • Partner • Dependents • PhoneService • MultipleLines • InternetService • OnlineSecurity • OnlineBackup	• DeviceProtection • TechSupport • StreamingTV • StreamingMovies • Contract • PaperlessBilling • PaymentMethod • Churn (response variable)

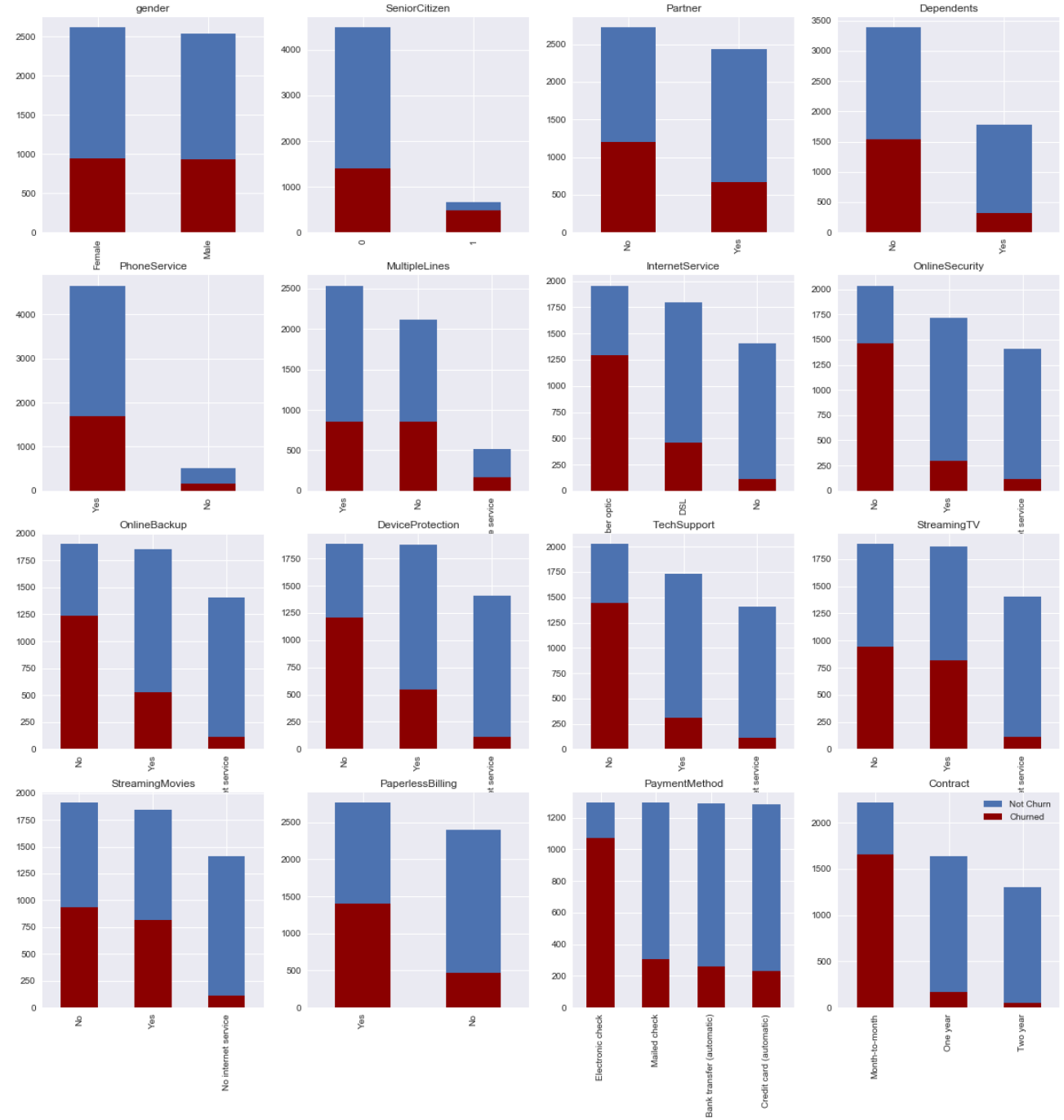
NUMERICAL VARIABLES

- Most of the churned customers are from
  - Slightly higher Monthly Charges (May feel overprice, switch to competitor with competitive price)
  - Smaller Total Charges (No staying long enough with the business)
  - Shorter tenure (No staying long enough with the business)







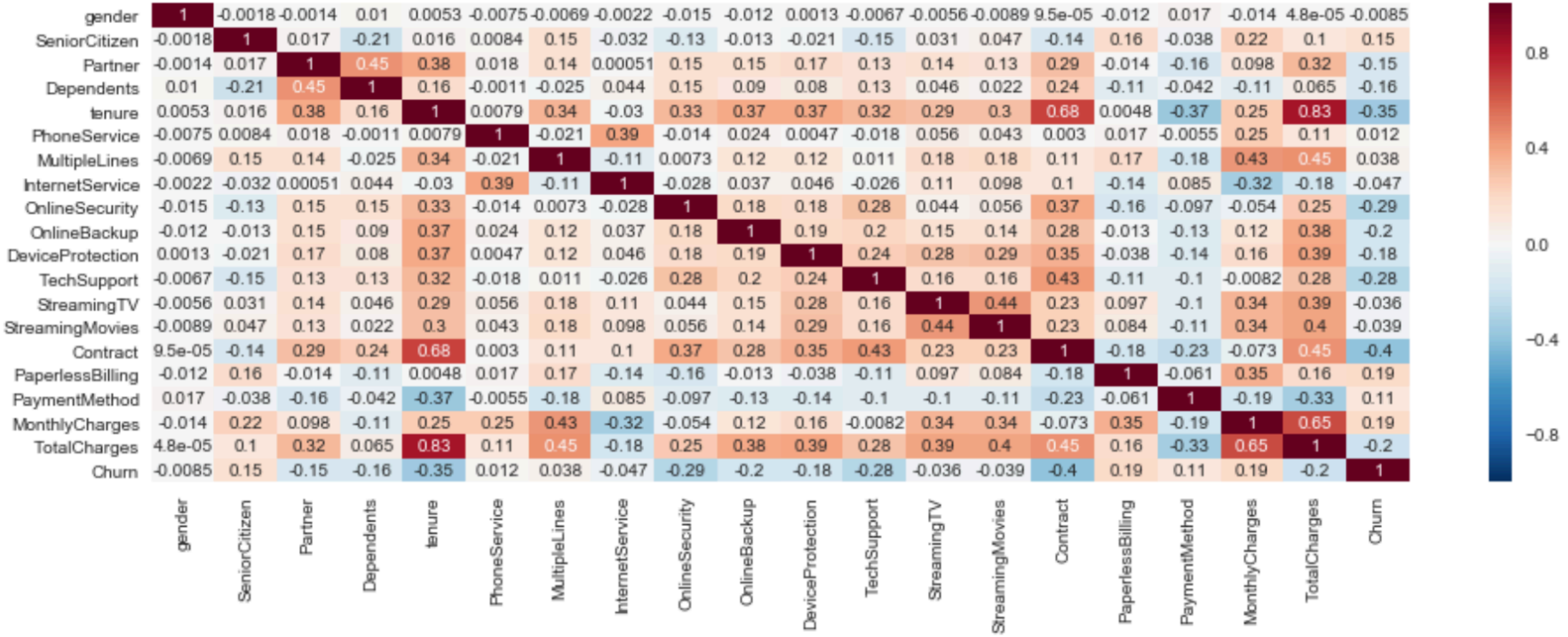


CATEGORICAL VARIABLES

- Key Characteristics of churned customer
  - No dependents
  - Using fiber optic internet service
  - Not using online security, online back up, tech support, device protection
  - Paperless billing
  - Payment via Electronic Check
  - Month-to-month Contract
  - Phone Service
- Hypothesis of Customer Churn Reason
  - Churned customer are mostly young generation who is savvy to find information online
    - No dependents
    - Paperless billing
    - Payment via Electronic Check
  - Not engaged with other services provided
    - Not using online security, online back up, tech support, device protection
  - Shorter contract which is easy to terminate

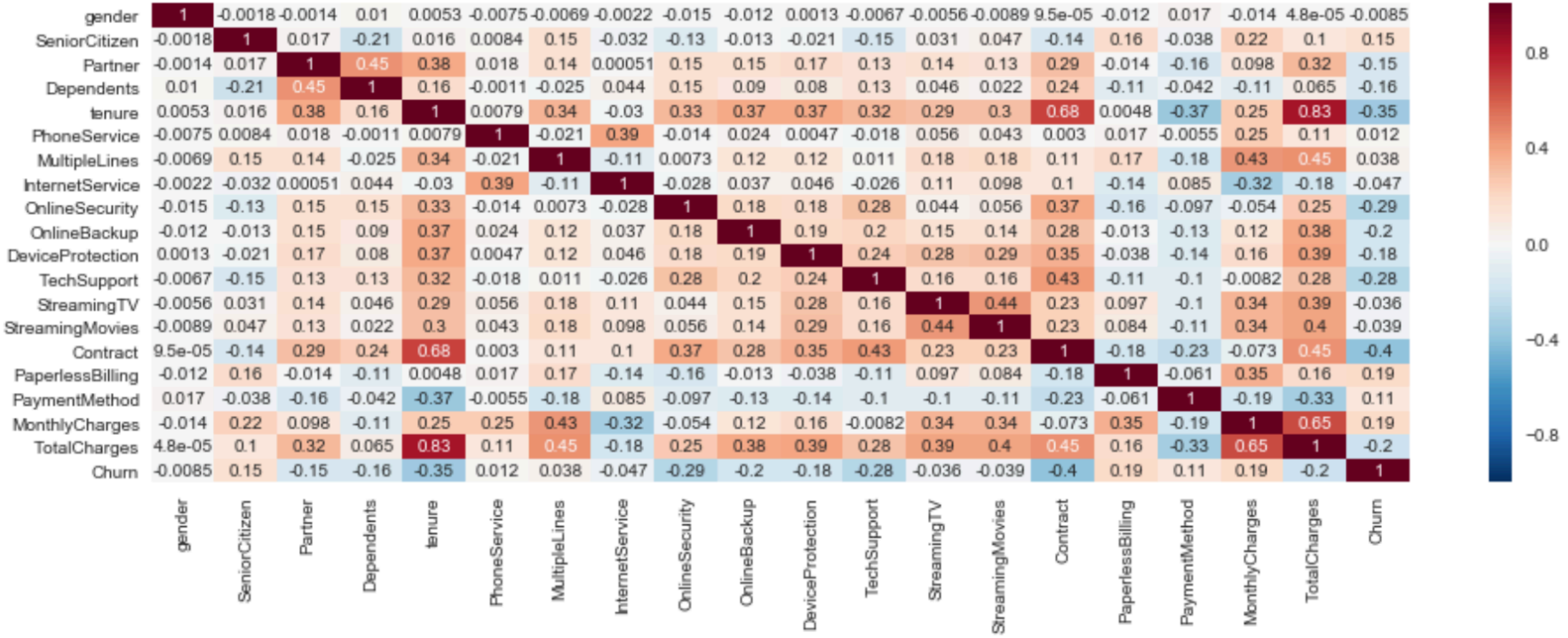
CORRELATIONS BETWEEN VARIABLES

- High positive correlation between
  - TotalCharges and Tenure (charge accumulated if stay longer with business)
  - Contract and Tenure (shorter contract period, shorter tenure)
  - TotalCharges and MonthlyCharges (Higher MonthlyCharges, higher TotalCharges)



CORRELATIONS BETWEEN VARIABLES

- Churn variable has negative correlation with
  - Tenure (shorter tenure, easier to churn)
  - Contract (shorter contract, easier to churn)
  - Other services (online security, tech support) (not using other services, easier to churn)



DATA PREPROCESSING

DATA CLEANING

- Data Duplication Check
  - No duplicate customerID in the dataset

- Null Values
  - 11 empty values in TotalCharges column
  - Remove the rows because it is small (0.1% of dataset) and the response is not churned

FEATURE ENGINEERING

- Label Encoding
  - To convert the string to numerical variable so that the models can read the data

MODEL DEVELOPMENT AND RESULTS

1. LOGISTIC REGRESSION (ALL VARIABLES)

- Accuracy – 80.3%
- Confusion Matrix

	Predict: Churn	Predict: Not Churn
Actual: Churn	337	273
Actual: Not Churn	184	1,527

- Precision Recall Matrix

Precision (% predicted churned users predicted correctly)	0.65
Recall (% actual churned users predicted correctly)	0.55
F-score	0.60

2. LOGISTIC REGRESSION (RESCALE)

- Rescale the variable to in between 0 and 1 (to prevent large value difference in variable affect the results)
- Accuracy – 80.3%
- Confusion Matrix

	Predict: Churn	Predict: Not Churn
Actual: Churn	325	285
Actual: Not Churn	173	1,538

- Precision Recall Matrix

Precision (% predicted churned users predicted correctly)	0.65
Recall (% actual churned users predicted correctly)	0.53
F-score	0.59

3. LOGISTIC REGRESSION (REMOVE TENURE)

- Remove tenure variable as it is highly correlated with total charges (to prevent collinearity issue)
- Accuracy – 79.6%
- Confusion Matrix

	Predict: Churn	Predict: Not Churn
Actual: Churn	320	290
Actual: Not Churn	182	1,529

- Precision Recall Matrix

Precision (% predicted churned users predicted correctly)	0.64
Recall (% actual churned users predicted correctly)	0.52
F-score	0.57

4. LOGISTIC REGRESSION (REMOVE INSIGNIFICANT VARIABLE)

- Remove insignificant variable (gender, MultipleLines, StreamingTV, StreamingMovies, PaymentMethod) based on p-value (refer to Appendix 1,2)
- Accuracy – 80.3%
- Confusion Matrix

	Predict: Churn	Predict: Not Churn
Actual: Churn	336	274
Actual: Not Churn	185	1,526

- Precision Recall Matrix

Precision (% predicted churned users predicted correctly)	0.64
Recall (% actual churned users predicted correctly)	0.55
F-score	0.59

5. XGBOOST

- Accuracy – 79.8%
- Confusion Matrix

	Predict: Churn	Predict: Not Churn
Actual: Churn	268	342
Actual: Not Churn	127	1,584

- Precision Recall Matrix

Precision (% predicted churned users predicted correctly)	0.68
Recall (% actual churned users predicted correctly)	0.44
F-score	0.53

MODEL SELECTION

- Model 4 (logistic regression with insignificant variables removed) because
  - Comparably high accuracy and F-score
  - Use lesser variables and computation power
- To increase accuracy of the model,
  - Collect data on customer online behavior, usage pattern as the features of the model
    - Online time usage
    - Online frequency
    - Calling time
    - Internet speed
    - No of times using tech support
  - Acquisition source and channel, etc

INSIGHTS AND RECOMMENDATIONS

CHARACTERISTICS OF CHURN CUSTOMERS

- Young generation who is savvy with technology
- Not engaged with internet services
- Short month-on-month contract

RECOMMENDATIONS (BUSINESS)

- Encourage customers with internet service to use other services, like online security, online back up, device protection, tech support and so on, to increase stickiness of the customers with business
- Create competitive advantages that keep younger generation
  - For example, faster speed and wider coverage
  - Add-on to the data
- Provide incentives to convert monthly contract to yearly contract

APPENDIX

APPENDIX 1

Optimization terminated successfully.
Current function value: 0.415349
Iterations 8

Logit Regression Results

Dep. Variable:	Churn	No. Observations:	4711
Model:	Logit	Df Residuals:	4692
Method:	MLE	Df Model:	18
Date:	Sat, 08 Feb 2020	Pseudo R-squ.:	0.2845
Time:	14:34:01	Log-Likelihood:	-1956.7
converged:	True	LL-Null:	-2734.7
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
gender	-0.0583	0.078	-0.751	0.452	-0.210	0.094
SeniorCitizen	0.1985	0.102	1.937	0.053	-0.002	0.399
Partner	0.1281	0.095	1.351	0.177	-0.058	0.314
Dependents	-0.2778	0.111	-2.509	0.012	-0.495	-0.061
tenure	-0.0702	0.007	-10.489	0.000	-0.083	-0.057
PhoneService	-1.1045	0.169	-6.549	0.000	-1.435	-0.774
MultipleLines	0.0456	0.049	0.925	0.355	-0.051	0.142
InternetService	0.2236	0.080	2.782	0.005	0.066	0.381
OnlineSecurity	-0.2973	0.050	-5.955	0.000	-0.395	-0.199
OnlineBackup	-0.1253	0.046	-2.723	0.006	-0.216	-0.035
DeviceProtection	-0.0896	0.048	-1.880	0.060	-0.183	0.004
TechSupport	-0.2610	0.051	-5.116	0.000	-0.361	-0.161
StreamingTV	0.0002	0.050	0.004	0.997	-0.097	0.098
StreamingMovies	-0.0067	0.050	-0.134	0.894	-0.105	0.091
Contract	-0.7980	0.096	-8.272	0.000	-0.987	-0.609
PaperlessBilling	0.2151	0.088	2.441	0.015	0.042	0.388
PaymentMethod	-0.0046	0.038	-0.120	0.904	-0.080	0.070
MonthlyCharges	0.0207	0.002	8.560	0.000	0.016	0.025
TotalCharges	0.0004	7.24e-05	6.110	0.000	0.000	0.001

APPENDIX 2

Optimization terminated successfully.
Current function value: 0.415502
Iterations 8

Logit Regression Results

Dep. Variable:	Churn	No. Observations:	4711
Model:	Logit	Df Residuals:	4697
Method:	MLE	Df Model:	13
Date:	Sat, 08 Feb 2020	Pseudo R-squ.:	0.2842
Time:	14:33:57	Log-Likelihood:	-1957.4
converged:	True	LL-Null:	-2734.7
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
SeniorCitizen	0.2002	0.102	1.956	0.050	-0.000	0.401
Partner	0.1288	0.095	1.360	0.174	-0.057	0.315
Dependents	-0.2810	0.110	-2.543	0.011	-0.497	-0.064
tenure	-0.0704	0.007	-10.797	0.000	-0.083	-0.058
PhoneService	-1.1449	0.156	-7.331	0.000	-1.451	-0.839
InternetService	0.2256	0.077	2.937	0.003	0.075	0.376
OnlineSecurity	-0.2976	0.050	-5.976	0.000	-0.395	-0.200
OnlineBackup	-0.1264	0.046	-2.758	0.006	-0.216	-0.037
DeviceProtection	-0.0914	0.047	-1.936	0.053	-0.184	0.001
TechSupport	-0.2644	0.051	-5.230	0.000	-0.363	-0.165
Contract	-0.8056	0.096	-8.400	0.000	-0.994	-0.618
PaperlessBilling	0.2168	0.088	2.469	0.014	0.045	0.389
MonthlyCharges	0.0211	0.002	10.008	0.000	0.017	0.025
TotalCharges	0.0005	6.89e-05	6.567	0.000	0.000	0.001

APPENDIX 3 - ASSUMPTIONS OF THE DATA

- SeniorCitizen = 1 means senior citizen, else not senior citizen