Predicting Customer Churn
Group 13 - Lance Elson
STAT 331 001 Winter 20/21

I. Introduction

Despite that many customers are loyal to their telecommunications companies, they remain loyal only as long as they are satisfied. Even low customer churn rates can lead to several million dollars in monthly variable costs. Predictive modeling for customer churn can help identify aggregated characteristics of customers expected to leave, and subsequently, help minimize churn. Analyzing demographic and contract-specific data for existing and past customers produces models for predictive implementation.

II. Data Overview

customerID: Unique customer identifier
    Type: Nominal, 2,114 levels
    Note: *Omitted from analysis*
gender: Customer Gender
    Type: Nominal, 2 levels (Female, Male)
SeniorCitizen: Indicates whether a customer is a senior citizen
    Type: Nominal, 2 levels (0, 1)
    Note: 0 = No, 1 = Yes
Partner: Indicates whether the customer has a partner
    Type: Nominal, 2 levels (No, Yes)
Dependents: Indicates whether the customer has dependents
    Type: Nominal, 2 levels (No, Yes)
tenure: The length of time in months that the customer has been a customer
    Type: Numeric, range 0 – 72 inclusive
PhoneService: Indicates whether the customer has phone service with the company
    Type: Nominal, 2 levels (No, Yes)
InternetService: Indicates the type of internet service the customer has with the company
    Type: Nominal, 3 levels (DSL, Fiber optic, No)
    Note: No = No internet service with the company
Contract: The type of contract that the customer has with the company
    Type: Ordinal, 3 levels (1-Month-to-month, 2-One year, 3-Two year)
PaperlessBilling: Indicates whether the customer is enrolled in paperless billing
    Type: Nominal, 2 levels (No, Yes)
PaymentMethod: The most recent payment method used by the customer to pay the company
    Type: Nominal, 4 levels (Bank transfer (automatic), Credit card (automatic), Electronic check, Mailed check
MonthlyCharges: The most recent dollar amount that the customer is charged per month
    Type: Numeric, range 18.25 – 117.8 inclusive
TotalCharges: The total dollar amount that the customer has been charged
    Type: Numeric, range 18.8 – 8684.8 inclusive
Churn: Whether the customer has left the company

Type: Nominal, 2 levels (No, Yes)

```
  customerID            gender     SeniorCitizen Partner    Dependents     tenure       PhoneService
 Length:2114        Female:1032   0:1741        No :1079   No :1495   Min.   : 0.00   No : 203
 Class :character   Male  :1082   1: 373        Yes:1035   Yes: 619   1st Qu.: 9.00   Yes:1911
 Mode  :character                                                     Median :30.00
                                                                      Mean   :33.05
                                                                      3rd Qu.:56.00
                                                                      Max.   :72.00
   InternetService           Contract     PaperlessBilling                   PaymentMethod MonthlyCharges
 DSL        :708    Month-to-month:1152   No : 828         Bank transfer (automatic):437   Min.   : 18.25
 Fiber optic:959    One year      : 467   Yes:1286         Credit card (automatic)  :480   1st Qu.: 38.33
 No         :447    Two year      : 495                    Electronic check         :733   Median : 71.17
                                                           Mailed check             :464   Mean   : 65.42
                                                                                           3rd Qu.: 89.95
                                                                                           Max.   :117.80
  TotalCharges      Churn
 Min.   :  18.8   No :1553
 1st Qu.: 429.6   Yes: 561
 Median :1484.1
 Mean   :2335.2
 3rd Qu.:3921.2
 Max.   :8684.8
```

Figure I: Summary statistic information

Nr	ColName	Class	NAs	Levels
1	customerID	character	.	
2	gender	factor	.	(2): 1-Female, 2-Male
3	SeniorCitizen	factor	.	(2): 1-0, 2-1
4	Partner	factor	.	(2): 1-No, 2-Yes
5	Dependents	factor	.	(2): 1-No, 2-Yes
6	tenure	integer	.	
7	PhoneService	factor	.	(2): 1-No, 2-Yes
8	InternetService	factor	.	(3): 1-DSL, 2-Fiber optic, 3-No
9	Contract	ordered, factor	.	(3): 1-Month-to-month, 2-One year, 3-Two year
10	PaperlessBilling	factor	.	(2): 1-No, 2-Yes
11	PaymentMethod	factor	.	(4): 1-Bank transfer (automatic), 2-Credit card (automatic), 3-Electronic check, 4-Mailed check
12	MonthlyCharges	numeric	.	
13	TotalCharges	numeric	.	
14	Churn	factor	.	(2): 1-No, 2-Yes

Figure II: Abstract information

Data Cleaning

405	1371-DWPAZ	Female	0	Yes	Yes	0	No	DSL	Two year	No	Credit card (automatic)	56.05	*NA*	No
2040	2775-SEFEE	Male	0	No	Yes	0	Yes	DSL	Two year	Yes	Bank transfer (automatic)	61.9	*NA*	No

2 new customers (tenure = 0) had empty TotalCharges fields, so their MonthlyCharges were imposed into the missing field. This matched the total charges for other new customers, whose TotalCharges and MonthlyCharges were equal.

III. Naïve Bayes Modeling

This probabilistic classification model estimates the likelihood of a customer leaving the company across multiple trials. It is simple and quick, and is robust to potentially irrelevant data. However, it relies on estimated probabilities, independency across all variables, and normally distributed numerical data. Since this model can handle both numeric and non-numeric data, and since the numeric data can be normalized using a Box-Cox transformation, it is a candidate for application.

a. Model assumptions

The Naïve Bayes assumes that all variables in the dataset are equally important and independent. This means that it assumes that the values of one class do not depend on the values of another class across all events. It also assumes that numerical variables are normally distributed. To conform to this assumption, the numeric variables (MonthlyCharges, TotalCharges, tenure) were converted to standardized and approximately normal using Box-Cox transformation. Box-Cox transformation assumes that its numeric inputs are non-negative and continuous.

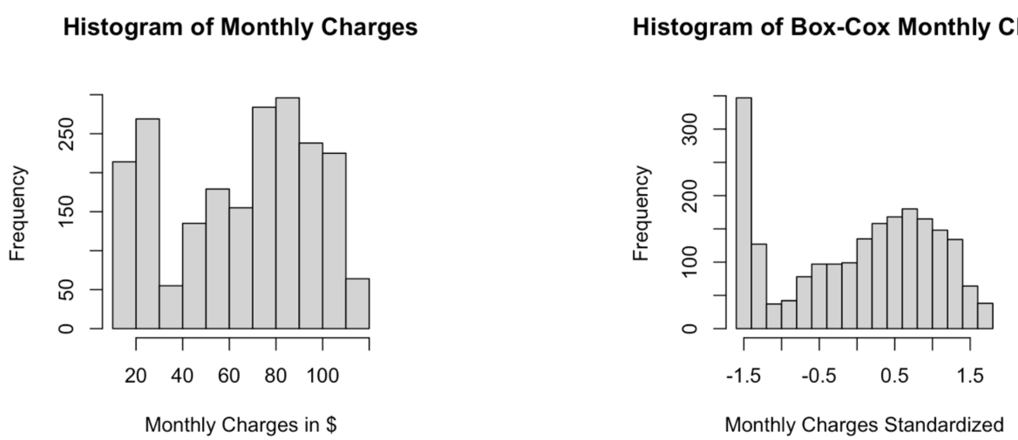

Figure III. Pre-transformed (left) and transformed (right) monthly charges

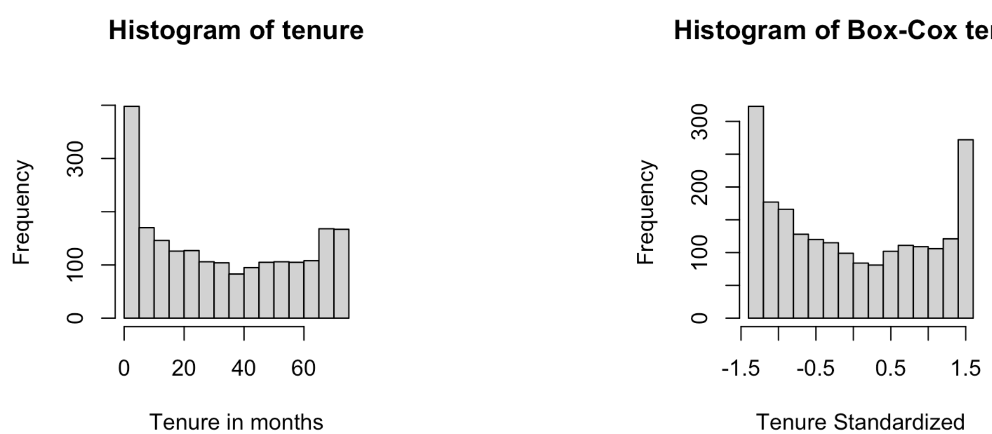

Figure IV. Pre-transformed (left) and transformed (right) tenure

Monthly charges have a right skewed transformed distribution, and tenure has both left and right skews, meaning that this could lead to the Naïve Bayes model being insufficient for application.

To assist with the independency assumption, highly correlated variables are identified and excluded from analysis. This model assumes an adjustable correlation cutoff of 75%.

```
               MonthlyCharges TotalCharges    tenure
MonthlyCharges      1.0000000    0.6525447 0.2356645
TotalCharges        0.6525447    1.0000000 0.8201822
tenure              0.2356645    0.8201822 1.0000000
```

Figure V. Correlation matrix for numeric variables

TotalCharges meets the correlation cutoff with respect to tenure. Subsequently, it needs to be removed from analysis.

b. Naïve Bayes Model

```
A-priori probabilities:
NB_dataTrain$Churn
       No       Yes
0.7346336 0.2653664

Conditional probabilities:
                  gender
NB_dataTrain$Churn    Female      Male
               No  0.4907482 0.5092518
               Yes 0.5144766 0.4855234

                  SeniorCitizen
NB_dataTrain$Churn         0         1
               No  0.8648431 0.1351569
               Yes 0.7104677 0.2895323

                  Partner
NB_dataTrain$Churn        No       Yes
               No  0.4714401 0.5285599
               Yes 0.6102450 0.3897550

                  Dependents
NB_dataTrain$Churn        No       Yes
               No  0.6645213 0.3354787
               Yes 0.8285078 0.1714922

                  PhoneService
NB_dataTrain$Churn         No        Yes
               No  0.09814964 0.90185036
               Yes 0.09799555 0.90200445

                  InternetService
NB_dataTrain$Churn        DSL Fiber optic         No
               No  0.37087691  0.36283186 0.26629123
               Yes 0.27394209  0.67706013 0.04899777
```

```
                  PaperlessBilling
NB_dataTrain$Churn        No       Yes
               No  0.4416734 0.5583266
               Yes 0.2383073 0.7616927

                  PaymentMethod
NB_dataTrain$Churn Bank transfer (automatic) Credit card (automatic) Electronic check Mailed check
               No                  0.2397426               0.2606597        0.2622687    0.2373290
               Yes                 0.1180401               0.1425390        0.5723831    0.1670379

                  Contract
NB_dataTrain$Churn Month-to-month   One year   Two year
               No      0.42477876 0.27272727 0.30249397
               Yes     0.88641425 0.09576837 0.01781737

                  MonthlyCharges
NB_dataTrain$Churn       [,1]      [,2]
               No  -0.1104606 1.0364161
               Yes  0.2804474 0.8003579

                  tenure
NB_dataTrain$Churn       [,1]      [,2]
               No   0.2015214 0.9764952
               Yes -0.5815960 0.8018825
```

Figure VI. Naïve Bayes model

To interpret the conditional probabilities, for example, let’s take the Churn | gender output.
* 0.49 is the probability that Churn = No given that the customer gender = Female.
* 0.51 is the probability that Churn = Yes given that the customer gender = Female.

* 0.51 is the probability that Churn = No given that the customer gender = Male.
* 0.49 is the probability that Churn = Yes given that the customer gender = Male.

Notably, the Naïve Bayes model is a poor estimator. The resulting conditional probabilities are not useful for interpretation.

c. Naïve Bayes Performance

```
          Reference
Prediction  No Yes
       No  979 129
       Yes 264 320

               Accuracy : 0.7677
                 95% CI : (0.7469, 0.7877)
    No Information Rate : 0.7346
    P-Value [Acc > NIR] : 0.0009858

                  Kappa : 0.4565

 Mcnemar's Test P-Value : 1.386e-11

            Sensitivity : 0.7127
            Specificity : 0.7876
         Pos Pred Value : 0.5479
         Neg Pred Value : 0.8836
              Precision : 0.5479
                 Recall : 0.7127
                     F1 : 0.6196
             Prevalence : 0.2654
         Detection Rate : 0.1891
   Detection Prevalence : 0.3452
      Balanced Accuracy : 0.7502

       'Positive' Class : Yes
```

```
          Reference
Prediction  No Yes
       No  246  30
       Yes  64  82

               Accuracy : 0.7773
                 95% CI : (0.7345, 0.8161)
    No Information Rate : 0.7346
    P-Value [Acc > NIR] : 0.0252787

                  Kappa : 0.4792

 Mcnemar's Test P-Value : 0.0006648

            Sensitivity : 0.7321
            Specificity : 0.7935
         Pos Pred Value : 0.5616
         Neg Pred Value : 0.8913
              Precision : 0.5616
                 Recall : 0.7321
                     F1 : 0.6357
             Prevalence : 0.2654
         Detection Rate : 0.1943
   Detection Prevalence : 0.3460
      Balanced Accuracy : 0.7628

       'Positive' Class : Yes
```

Figure VII. Performance measures for the model applied to training (left) and testing (right) data

Performance measures are nearly consistent between the Naïve Bayes model applied to both the training (data used to train the model) and testing (data excluded from training the model) datasets. Overall, this model exhibits good performance.

Sensitivity across both models = ~0.72, meaning roughly 72% of examples are correctly. Since specificity is higher, at ~79%, model is slightly better at predicting customers who remain loyal. Accuracy = ~0.77 meaning that the proportion of correct: incorrect predictions is ~77%. Kappa, or accuracy under random example assumption, is ~46%, resulting in fair agreement.

d. Naïve Bayes Goodness of Fit

	Training	Testing
Accuracy	7.677305e-01	0.7772511848
Kappa	4.564706e-01	0.4792312136
AccuracyLower	7.468594e-01	0.7344933705
AccuracyUpper	7.876687e-01	0.8160715457
AccuracyNull	7.346336e-01	0.7345971564
AccuracyPValue	9.858144e-04	0.0252786503
McnemarPValue	1.385588e-11	0.0006648213

	Training	Testing
Sensitivity	0.7126949	0.7321429
Specificity	0.7876106	0.7935484
Pos Pred Value	0.5479452	0.5616438
Neg Pred Value	0.8835740	0.8913043
Precision	0.5479452	0.5616438
Recall	0.7126949	0.7321429
F1	0.6195547	0.6356589
Prevalence	0.2653664	0.2654028
Detection Rate	0.1891253	0.1943128
Detection Prevalence	0.3451537	0.3459716
Balanced Accuracy	0.7501527	0.7628456

Figure VII. Overall (left) and class-level (right) goodness of fit information

The Naïve Bayes model is balanced. Training and testing performance measures are approximately equal at both overall and class-levels.

IV. Logistic Regression Modeling

This model fits a curve against a dichotomous dependent variable. In this case, churn, a 2-level factor, is the dependent variable. Logistic regression, unlike Naïve Bayes, assumes neither interdependency nor normally distributed numeric variables. It can handle irrelevant and redundant variables, works well with high-dimensional data, and produces interpretable results. The output enables inferencing customer characteristics related to churn.

a. Model Assumptions

Logistic regression assumes that the data has no missing values. Missing values were addressed during data cleanup, so the model is readily applicable without further transformation.

b. Logistic Regression Model

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7650  -0.7087  -0.2792   0.7804   3.2000

Coefficients:
                                       Estimate Std. Error z value Pr(>|z|)
(Intercept)                          -0.6977770  0.4951563  -1.409 0.158774
genderMale                           -0.0042651  0.1308812  -0.033 0.974003
SeniorCitizen1                        0.4148197  0.1646188   2.520 0.011739 *
PartnerYes                            0.2222572  0.1534837   1.448 0.147594
DependentsYes                        -0.2541215  0.1795994  -1.415 0.157087
tenure                               -0.0742777  0.0133224  -5.575 2.47e-08 ***
PhoneServiceYes                      -0.5502431  0.2932391  -1.876 0.060596 .
InternetServiceFiber optic            0.6968372  0.2646235   2.633 0.008456 **
InternetServiceNo                    -1.0715928  0.3944911  -2.716 0.006600 **
Contract.L                           -1.3231331  0.2909912  -4.547 5.44e-06 ***
Contract.Q                           -0.1320203  0.2053656  -0.643 0.520318
PaperlessBillingYes                   0.3195706  0.1530935   2.087 0.036850 *
PaymentMethodCredit card (automatic)  0.2137331  0.2324190   0.920 0.357781
PaymentMethodElectronic check         0.5437870  0.1960115   2.774 0.005533 **
PaymentMethodMailed check             0.1667801  0.2397548   0.696 0.486662
MonthlyCharges                       -0.0049766  0.0081677  -0.609 0.542324
TotalCharges                          0.0005027  0.0001520   3.307 0.000944 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1958.0  on 1691  degrees of freedom
Residual deviance: 1426.4  on 1675  degrees of freedom
AIC: 1460.4

Number of Fisher Scoring iterations: 7
```

Figure VIII. Log regression model

- The relationship between a customer having phone service and not churning is slightly statistically significant.
- The relationship between a customer being a senior citizen and churning is statistically significant.
- The relationship between a customer using paperless billing and churning is statistically significant.
- The relationship between a customer having fiber optic internet service and churning is more statistically significant.
- The relationship between a customer not having internet service and not churning is more statistically significant.
- The relationship between a customer paying via electronic check and churning is more statistically significant.
- The relationship between a customer having a long tenure and not churning is very statistically significant.

- The linear relationship between a customer having a longer contract and not churning is very statistically significant.
- The relationship between a customer having more total charges and churning is very statistically significant.

```
                        (Intercept)                         genderMale
                          0.4976904                          0.9957440
                     SeniorCitizen1                         PartnerYes
                          1.5140977                          1.2488926
                      DependentsYes                             tenure
                          0.7755975                          0.9284139
                    PhoneServiceYes        InternetServiceFiber optic
                          0.5768096                          2.0073937
                  InternetServiceNo                         Contract.L
                          0.3424626                          0.2662997
                         Contract.Q                PaperlessBillingYes
                          0.8763232                          1.3765366
PaymentMethodCredit card (automatic)     PaymentMethodElectronic check
                          1.2382921                          1.7225177
          PaymentMethodMailed check                     MonthlyCharges
                          1.1814944                          0.9950358
                       TotalCharges
                          1.0005028
```

Figure IX. Log regression model – odds ratios

For odds above 1, churn has a greater chance of occurring when the factor variable equals the value mapped, (ex. Senior Citizen = 1, Partner = Yes, etc.) or when the numeric variable increases in value.

For odds below 1, the customer is more likely to remain for the same reasons. For example, when the customer has dependents, they are less likely to churn.

c. Logistic regression performance

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1111  212
       Yes  132  237

               Accuracy : 0.7967
                 95% CI : (0.7767, 0.8156)
    No Information Rate : 0.7346
    P-Value [Acc > NIR] : 1.685e-09

                  Kappa : 0.4471

 Mcnemar's Test P-Value : 2.050e-05

            Sensitivity : 0.5278
            Specificity : 0.8938
         Pos Pred Value : 0.6423
         Neg Pred Value : 0.8398
              Precision : 0.6423
                 Recall : 0.5278
                     F1 : 0.5795
             Prevalence : 0.2654
         Detection Rate : 0.1401
   Detection Prevalence : 0.2181
      Balanced Accuracy : 0.7108

       'Positive' Class : Yes
```

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  271  48
       Yes  39  64

               Accuracy : 0.7938
                 95% CI : (0.7521, 0.8314)
    No Information Rate : 0.7346
    P-Value [Acc > NIR] : 0.002863

                  Kappa : 0.4574

 Mcnemar's Test P-Value : 0.391064

            Sensitivity : 0.5714
            Specificity : 0.8742
         Pos Pred Value : 0.6214
         Neg Pred Value : 0.8495
              Precision : 0.6214
                 Recall : 0.5714
                     F1 : 0.5953
             Prevalence : 0.2654
         Detection Rate : 0.1517
   Detection Prevalence : 0.2441
      Balanced Accuracy : 0.7228

       'Positive' Class : Yes
```

Figure X. Performance measures for the model applied to training (left) and testing (right) data

Considering that sensitivity is extremely low compared to specificity for the model applied to both the testing and training data sets, logistic regression performs better when used to predict customers that remain. Accuracy and kappa are comparable to the naïve bayes model.

d. Logistic regression goodness of fit

	Training	Testing
Accuracy	7.966903e-01	0.793838863
Kappa	4.470872e-01	0.457359071
AccuracyLower	7.767111e-01	0.752055112
AccuracyUpper	8.156348e-01	0.831445230
AccuracyNull	7.346336e-01	0.734597156
AccuracyPValue	1.685233e-09	0.002863156
McnemarPValue	2.049814e-05	0.391063648

	Training	Testing
Sensitivity	0.5278396	0.5714286
Specificity	0.8938053	0.8741935
Pos Pred Value	0.6422764	0.6213592
Neg Pred Value	0.8397581	0.8495298
Precision	0.6422764	0.6213592
Recall	0.5278396	0.5714286
F1	0.5794621	0.5953488
Prevalence	0.2653664	0.2654028
Detection Rate	0.1400709	0.1516588
Detection Prevalence	0.2180851	0.2440758
Balanced Accuracy	0.7108225	0.7228111

Figure XI. Overall (left) and class-level (right) goodness of fit information

The Log regression model is balanced. Training and testing performance measures are approximately equal at both overall and class-levels.

V. Random Forest Modeling

This is an efficient ensemble method that builds decision trees considering a random sample of m predictors, where m is usually the square root of the number of predictors, or independent variables considered in the model. This model is less interpretable than the linear regression and naïve bayes models.

a. Model Assumptions

No formal assumptions for random forest modeling.

b. Random Forest Model

```
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 20.69%
Confusion matrix:
      No Yes class.error
No  1108 135   0.1086082
Yes  215 234   0.4788419
```

Figure XII. Random forest model

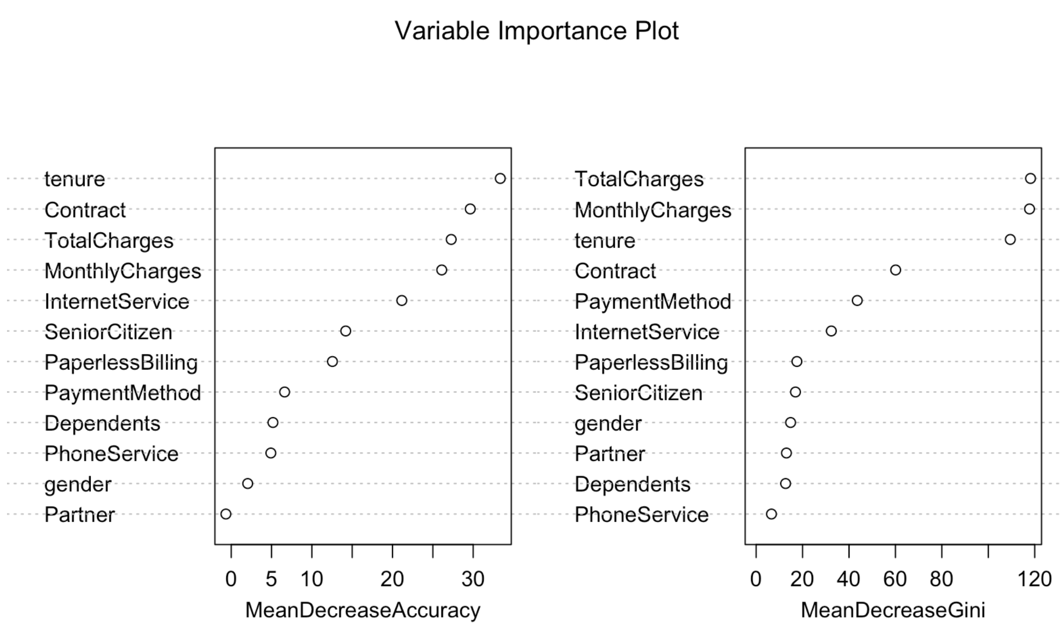

Figure XIII. Variable importance plot using mean decrease in accuracy & Gini impurity

Contract, tenure, TotalCharges, and MonthlyCharges are the most important variables in influencing customer churn based on the above criteria.

c. Random forest performance

```
          Reference
Prediction   No  Yes
       No  1108  215
       Yes  135  234

               Accuracy : 0.7931
                 95% CI : (0.773, 0.8122)
    No Information Rate : 0.7346
    P-Value [Acc > NIR] : 1.328e-08

                  Kappa : 0.4374

 Mcnemar's Test P-Value : 2.414e-05

            Sensitivity : 0.5212
            Specificity : 0.8914
         Pos Pred Value : 0.6341
         Neg Pred Value : 0.8375
              Precision : 0.6341
                 Recall : 0.5212
                     F1 : 0.5721
             Prevalence : 0.2654
         Detection Rate : 0.1383
   Detection Prevalence : 0.2181
      Balanced Accuracy : 0.7063

       'Positive' Class : Yes
```

```
          Reference
Prediction  No Yes
       No  268  52
       Yes  42  60

               Accuracy : 0.7773
                 95% CI : (0.7345, 0.8161)
    No Information Rate : 0.7346
    P-Value [Acc > NIR] : 0.02528

                  Kappa : 0.412

 Mcnemar's Test P-Value : 0.35326

            Sensitivity : 0.5357
            Specificity : 0.8645
         Pos Pred Value : 0.5882
         Neg Pred Value : 0.8375
              Precision : 0.5882
                 Recall : 0.5357
                     F1 : 0.5607
             Prevalence : 0.2654
         Detection Rate : 0.1422
   Detection Prevalence : 0.2417
      Balanced Accuracy : 0.7001

       'Positive' Class : Yes
```

Figure XIV. Performance measures for the model applied to training (left) and testing (right) data

Like the linear regression model, the random forest model performs better when predicting customers who do not churn. Specificity is higher than sensitivity. Accuracy and kappa are comparable to othe previous models.

d. Random forest goodness of fit

	Training	Testing
Accuracy	7.931442e-01	0.77725118
Kappa	4.374434e-01	0.41197747
AccuracyLower	7.730480e-01	0.73449337
AccuracyUpper	8.122182e-01	0.81607155
AccuracyNull	7.346336e-01	0.73459716
AccuracyPValue	1.328212e-08	0.02527865
McnemarPValue	2.413634e-05	0.35326280

	Training	Testing
Sensitivity	0.5211581	0.5357143
Specificity	0.8913918	0.8645161
Pos Pred Value	0.6341463	0.5882353
Neg Pred Value	0.8374906	0.8375000
Precision	0.6341463	0.5882353
Recall	0.5211581	0.5357143
F1	0.5721271	0.5607477
Prevalence	0.2653664	0.2654028
Detection Rate	0.1382979	0.1421801
Detection Prevalence	0.2180851	0.2417062
Balanced Accuracy	0.7062750	0.7001152

Figure XV. Overall (left) and class-level (right) goodness of fit information

The random forest model is balanced. Training and testing performance measures are approximately equal at both overall and class-levels.

VI. Discussion and Conclusion

All 3 models displayed good fit, and approximately equal Accuracy and Kappa. The Naïve Bayes model performed best for predicting customers who churned, having the highest sensitivity. Notably, Naïve Bayes preforms best using categorical data, and since most of the independent variables are categorical, it works well with the given data. However, the predictors are not independent. For example, if a customer does not have a partner, they are highly unlikely to have dependents. A customer with higher tenure will have higher total payments. Additionally, a Naïve Bayes model is a poor estimator, meaning that the conditional probabilities it provides are not useful for interpretation. Therefore, the Naïve Bayes will produce a good-performing, well-fitting, robust, and quick model for predicting whether a customer will leave, but without producing good class-level estimations.

In order to produce an interpretable model that includes total customer charges, while sacrificing sensitivity, a logistic regression model can be used. It can supplement Naïve Bayes predictions by highlighting variables that increase or decrease churn odds. For example, most of the customers that leave have a month-to-month contract, and/or had fiber internet service with the company. Since the Naïve Bayes model assumes no interdependence, it does not account for customers who, for example, have both a month-to-month contract and fiber internet services as well as a logistic regression model does. For this reason, a logistic regression model can serve as a cross-validator or class-level predictor supplementing a Naïve Bayes-based prediction.

Moving forward, prioritizing increasing satisfaction for slightly newer customers, with fiber optic internet service, month-to-month contracts, and payments via e-check will address churn. Prioritize investigating issues with the fiber optic options, and customer payment issues with primarily paperless, monthly e-checks.

VII. References

https://github.com/lance-elson/Drexel-Projects/blob/master/STAT331%20-%202021/CustomerChurn/CustomerChurn.csv