Telco Customer **Churn**

Focused customer retention programs

Forough Mofidi | Kate Pferdner |
Samantha Vega | Zainab Sunny
March 2024 | University of Chicago

Table of **contents**

01

Problem **Statement**

Business Use Case

Reducing churn is a key focus for many businesses, as retaining existing customers is often more cost-effective than acquiring new ones. Strategies for reducing churn include improving customer service, enhancing product or service offerings, implementing loyalty programs, and addressing issues that may be causing dissatisfaction among customers. Our goal through this exploratory analysis is to gain deeper insights into strategies that are more effective in keeping customers. For instance: what incentives should the company implement or what customer profiles should be targeted in marketing campaigns.

02

Exploratory Data Analysis

Data Investigation

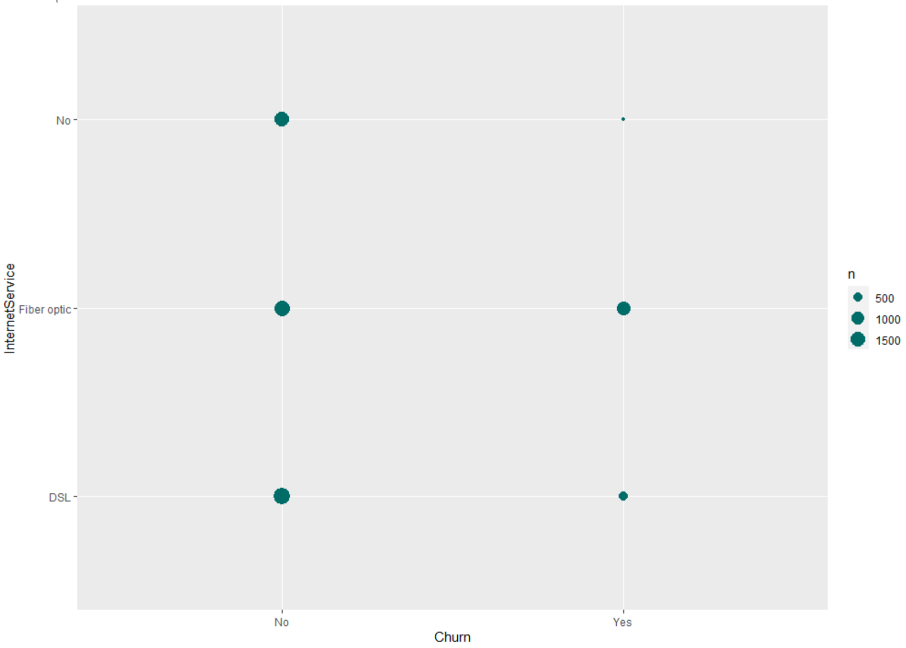
Variables	
Data Type	**Count**
Numeric	3
Categorical	12
Boolean	6
Total	**21**

Distributions of Binary Variables		
Column Name	**Yes Count**	**No Count**
Partner	3,402	3,641
Phone Service	6,361	682
Paperless Billing	4,171	2,872
Churn	**5,174**	**1,869**

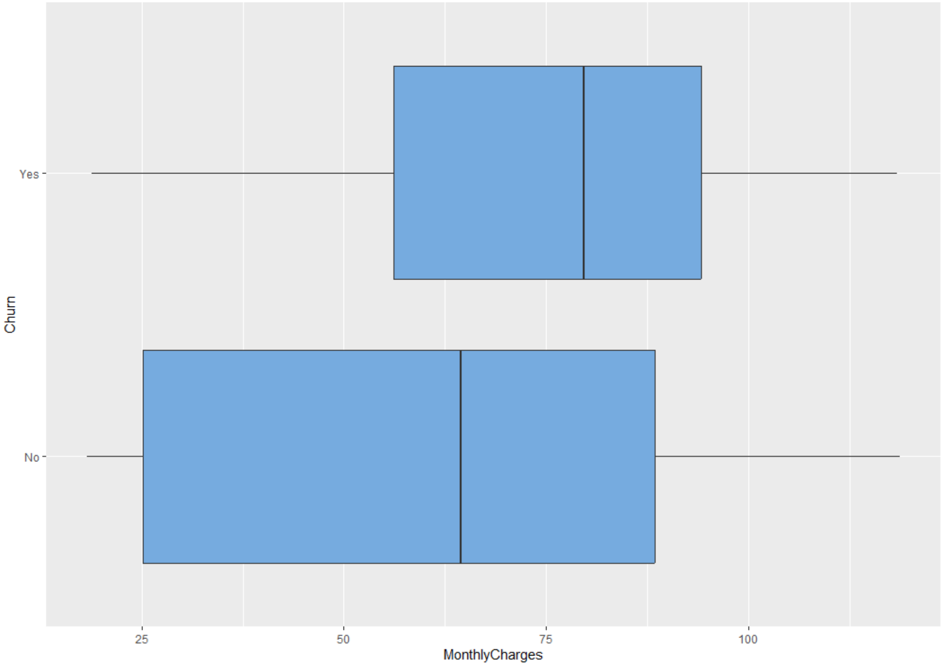
Numerical Summaries					
Column Name	**Minimum**	**Lower-Hinge**	**Median**	**Upper- Hinge**	**Maximum**
Tenure	0	9	29	55	72
MonthlyCharges	$18.25	$35.50	$70.35	$89.85	$118.75
TotalCharges	$18.80	$401.40	$1,397.48	$3,794.98	$8,684.80

Data Investigation

Churn by Internet Service Type

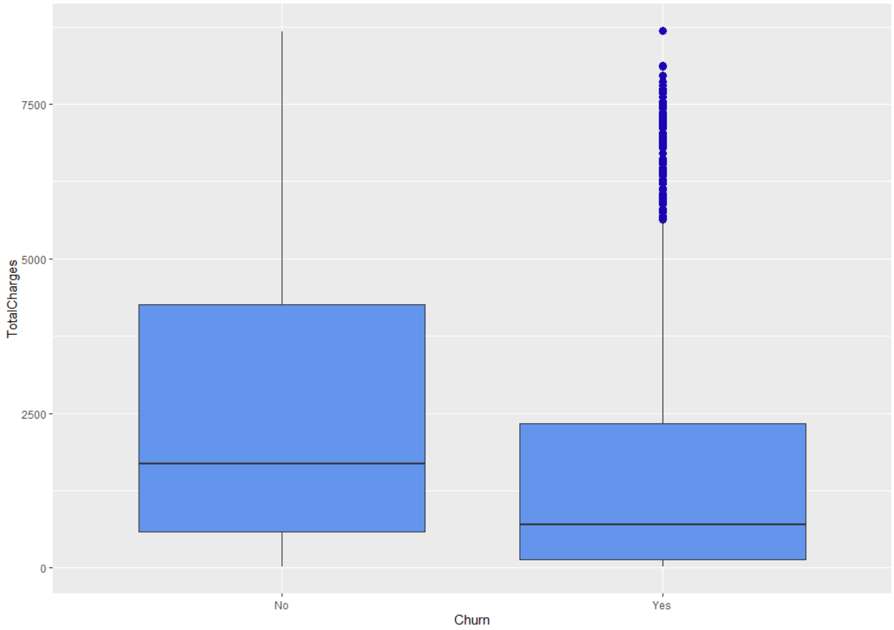


Churn by Monthly Charges

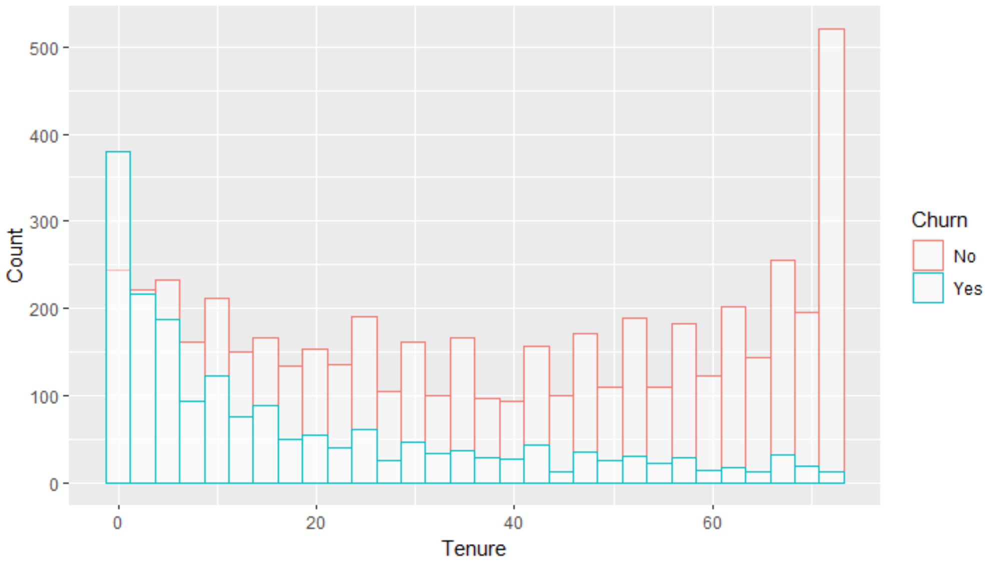


Data Investigation

Churn by Total Charges



Churn by Tenure



03

Modeling

Logistic Regression
Classification Tree
Random Forest
Support Vector Machine
K-Nearest Neighbors

Logistic Regression

Confusion Matrix		
Train Data	**No**	**Yes**
No	3,470	403
Yes	612	262

Accuracy of Train: 0.8075829

Confusion Matrix		
Test Data	**No**	**Yes**
No	1,157	133
Yes	205	262

Accuracy of Test: 0.8076266

```
Call:
glm(formula = Churn ~ SeniorCitizen + tenure + MultipleLines +
    InternetService + OnlineBackup + DeviceProtection + StreamingTV +
    StreamingMovies + Contract + PaperlessBilling + PaymentMethod +
    MonthlyCharges + TotalCharges, family = binomial, data = train)

Coefficients: (4 not defined because of singularities)
                                       Estimate Std. Error z value Pr(>|z|)
(Intercept)                           3.333e+00  6.440e-01   5.175 2.28e-07 ***
SeniorCitizenYes                      2.204e-01  9.616e-02   2.292 0.021908 *
tenure                               -6.610e-02  7.351e-03  -8.992  < 2e-16 ***
MultipleLinesNo phone service        -1.002e+00  3.063e-01  -3.271 0.001071 **
MultipleLinesYes                      5.940e-01  1.101e-01   5.395 6.85e-08 ***
InternetServiceFiber optic            2.863e+00  3.237e-01   8.847  < 2e-16 ***
InternetServiceNo                    -2.853e+00  3.942e-01  -7.237 4.59e-13 ***
OnlineBackupNo internet service              NA         NA      NA       NA
OnlineBackupYes                       2.713e-01  1.109e-01   2.447 0.014422 *
DeviceProtectionNo internet service          NA         NA      NA       NA
DeviceProtectionYes                   3.724e-01  1.126e-01   3.306 0.000947 ***
StreamingTVNo internet service               NA         NA      NA       NA
StreamingTVYes                        1.017e+00  1.611e-01   6.311 2.77e-10 ***
StreamingMoviesNo internet service           NA         NA      NA       NA
StreamingMoviesYes                    1.121e+00  1.603e-01   6.993 2.69e-12 ***
ContractOne year                     -7.589e-01  1.257e-01  -6.035 1.59e-09 ***
ContractTwo year                     -1.402e+00  2.018e-01  -6.947 3.74e-12 ***
PaperlessBillingYes                   3.712e-01  8.584e-02   4.325 1.53e-05 ***
PaymentMethodCredit card (automatic) -7.997e-02  1.321e-01  -0.605 0.544991
PaymentMethodElectronic check         2.605e-01  1.088e-01   2.395 0.016637 *
PaymentMethodMailed check            -1.117e-01  1.323e-01  -0.845 0.398374
MonthlyCharges                       -8.492e-02  1.301e-02  -6.527 6.71e-11 ***
TotalCharges                          3.873e-04  8.298e-05   4.667 3.06e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Kaplan-Meier

```
Call: survfit(formula = Surv(tenure, Churn) ~ StreamingTV, data = df)

                                       n nevent     rmean*
StreamingTV=No, (s0)                2809      0 47.440355
StreamingTV=No internet service, (s0) 1520    0 66.459816
StreamingTV=Yes, (s0)               2703      0 56.068951
StreamingTV=No, Yes                 2809    942 24.559645
StreamingTV=No internet service, Yes 1520   113  5.540184
StreamingTV=Yes, Yes                2703    814 15.931049
   *restricted mean time in state (max time = 72 )
```

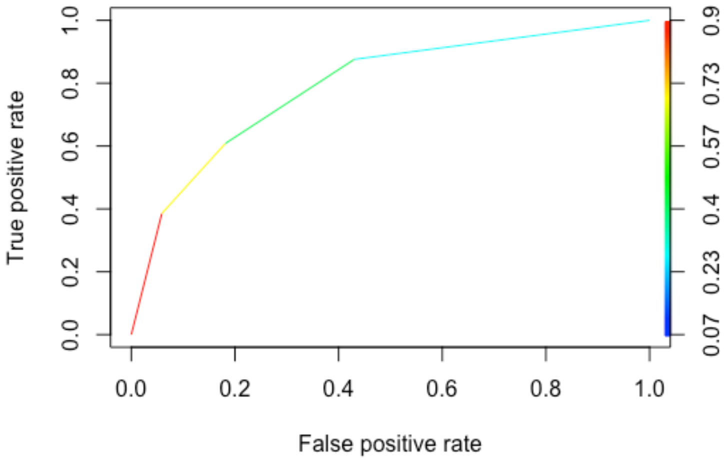
Customers who are not subscribed to streaming TV or have no internet service have longer average times until churn compared to those who are subscribed to streaming TV or have internet service.

Customers who are subscribed to both streaming TV and internet service tend to have shorter average times until churn compared to other groups.

Tree Classification

Model Results		
Accuracy	**Sensitivity**	**Specificity**
79.34%	94.11%	38.54%

ROC Curve - Classification Tree



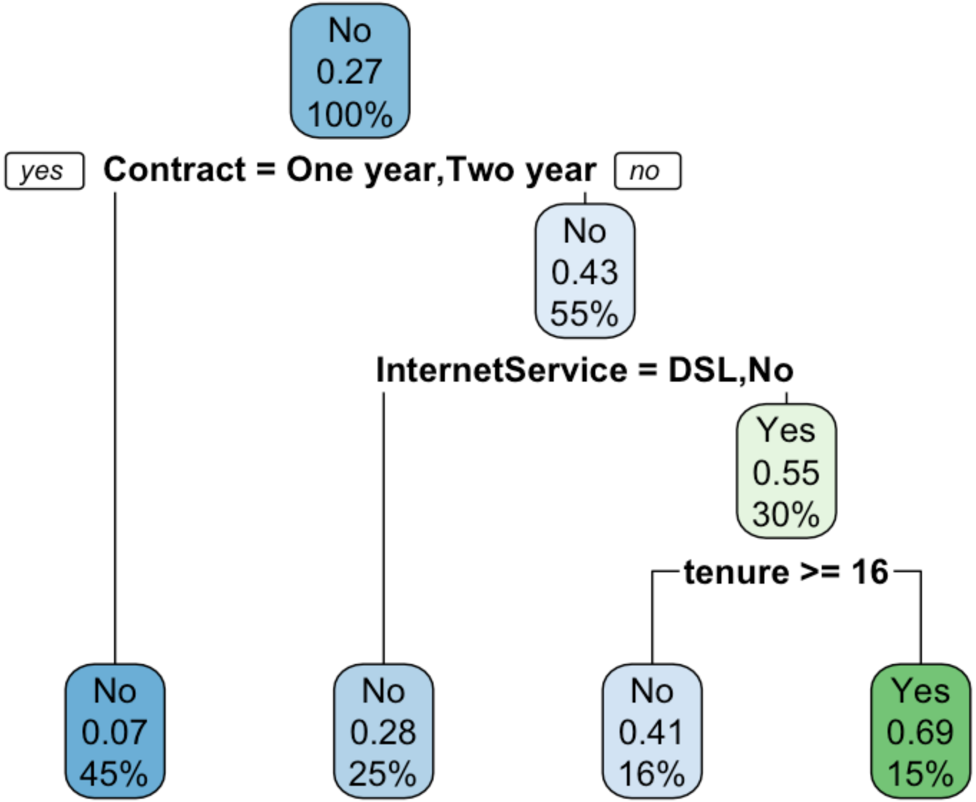
AUC Value is: 0.7912

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1214  287
       Yes   76  180

               Accuracy : 0.7934
                 95% CI : (0.7737, 0.8121)
    No Information Rate : 0.7342
    P-Value [Acc > NIR] : 4.882e-09

                  Kappa : 0.3815

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9411
            Specificity : 0.3854
         Pos Pred Value : 0.8088
         Neg Pred Value : 0.7031
             Prevalence : 0.7342
         Detection Rate : 0.6910
   Detection Prevalence : 0.8543
      Balanced Accuracy : 0.6633

       'Positive' Class : No
```

Tree Classification

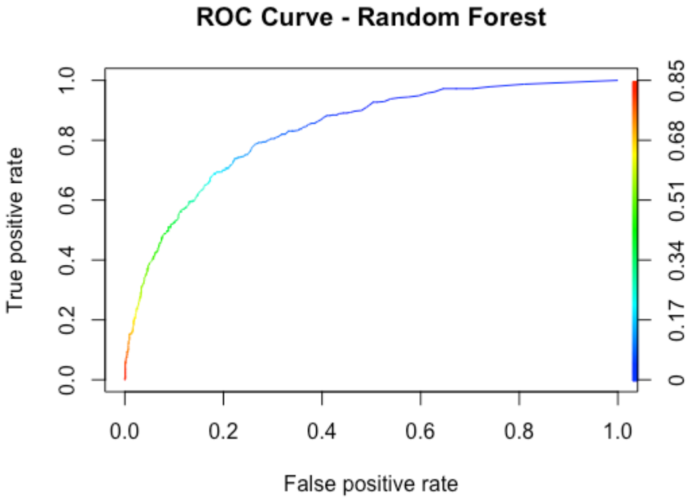
Customers with month-to-month contracts and fiber optic internet service, particularly those with shorter tenures, are more likely to churn.
In contrast, customers with longer-term contracts exhibit lower churn rates.

Results: Offering incentives to switch to longer contracts or addressing service quality issues for fiber optic customers with shorter tenure's could help reduce churn.



Random Forest

Model Results		
Accuracy	**Sensitivity**	**Specificity**
80.08%	95.04%	38.7%



AUC Value is: 0.8331

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1226  286
       Yes   64  181

               Accuracy : 0.8008
                 95% CI : (0.7813, 0.8192)
    No Information Rate : 0.7342
    P-Value [Acc > NIR] : 4.581e-11

                  Kappa : 0.3984

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9504
            Specificity : 0.3876
         Pos Pred Value : 0.8108
         Neg Pred Value : 0.7388
             Prevalence : 0.7342
         Detection Rate : 0.6978
   Detection Prevalence : 0.8606
      Balanced Accuracy : 0.6690

       'Positive' Class : No
```

Support Vector Machine

Model Results		
Accuracy	**Sensitivity**	**Specificity**
80.6%	66.2%	84.7%

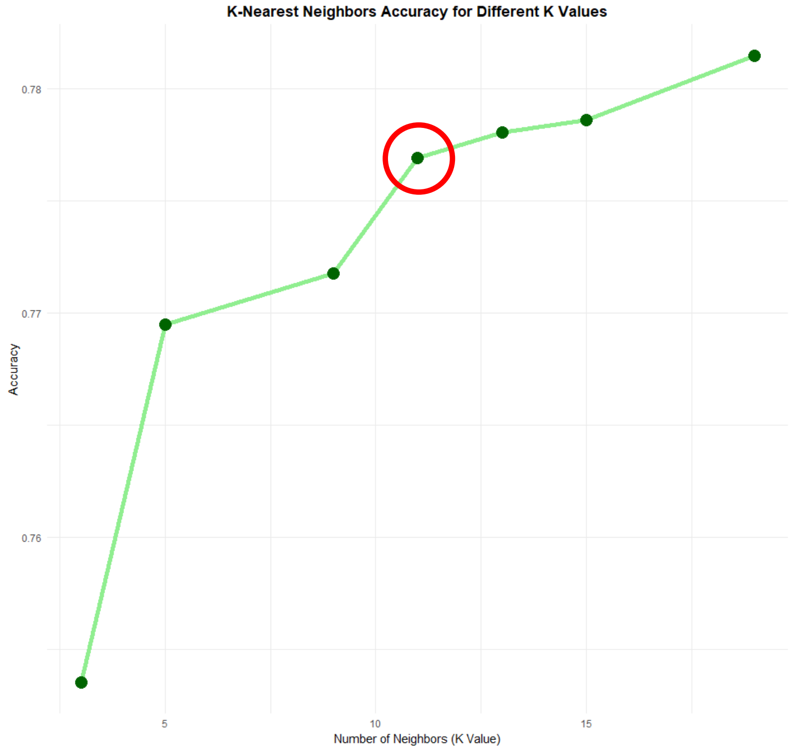
Based on the SVM model results:

- Churn is accurately predicted 80.6 % of the provided instances
- True Negatives are detected 84.7% of the time.
- True Positives are captured 66.2% of the time.

```
Call:
svm(formula = Churn ~ ., data = train, kernel = "linear", cost = 0.1)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  0.1

Number of Support Vectors:  2454
```

K–Nearest Neighbor

- K-Nearest Neighbor is a non-parametric supervised classification algorithm.
- Because our dataset, is primarily composed of categorical values, we were limited to a subset of the variables originally available, numerical and categorical with a maximum of 2 levels.
- K = 11 produced the best results with an accuracy of **77.7%**



K-Nearest Neighbor

Model Results		
Accuracy	**Sensitivity**	**Specificity**
77.7%	88.4%	48.0%

Based on the model results:

- Churn is accurately predicted 77.7% of the provided instances
- True Negatives are detected 48.0% of the time.
- True Positives are captured 88.4% of the time.

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1141  243
       Yes  149  224

               Accuracy : 0.7769
                 95% CI : (0.7567, 0.7962)
    No Information Rate : 0.7342
    P-Value [Acc > NIR] : 2.123e-05

                  Kappa : 0.3891

 Mcnemar's Test P-Value : 2.637e-06

            Sensitivity : 0.8845
            Specificity : 0.4797
         Pos Pred Value : 0.8244
         Neg Pred Value : 0.6005
             Prevalence : 0.7342
         Detection Rate : 0.6494
   Detection Prevalence : 0.7877
      Balanced Accuracy : 0.6821

       'Positive' Class : No
```

04

Evaluation

Model Comparison
Recommendations

80.76%

Logistic Regression

79.34%

Decision Tree

80.08%

Random Forest

80.64%

Support Vector Machine

77.7%

K-Nearest Neighbor

Modeling Comparison

Model Results			
	Accuracy	Sensitivity	Specificity
Logistic Regression	80.76%	56.10%	89.69%
Classification Tree	79.34%	94.11%	38.54%
Random Forest	80.08%	95.04%	38.7%
SVM	80.60%	66.2%	84.7%
K-Nearest Neighbor	77.7%	88.4%	48.0%

Recommendations

High Accuracy & Interpretability

Considering the need for both high accuracy and interpretability in the telecommunication business, Logistic Regression seems like a strong candidate. It provides a transparent way to understand why customers might leave and allows for easy communication of the results to non-technical stakeholders, which is valuable for implementing strategic business decisions.

Predictive Power

However, if we value predictive power more and have the capacity to handle a more complex model, Random Forest or Tree Classification might be better, as they offer slightly higher accuracy and can capture more complex relationships in the data.

Thanks

Any questions?