TELECOM CHURN

By BUAN6356.006.11

ABSTRACT

"The cost of acquiring a new customer can be higher than that of retaining a customer by as much as 700%. Increasing customer retention rates by a mere 5% could increase profits by 25% to 95%."

Tony John
Haw-Jan Hwang
Neeladri Mohapatra
Samarth Khare
Under Guidance of Prof. Ling Ge

EXECUTIVE SUMMARY

President (US) Bill Clinton once phrased “The price of doing the same old thing is far higher than the price of change”. Over the ages, the masses all over the world have taken for granted “the phone company” aka the telecom companies. And it turns out, in turn, the phone companies have taken its customers for granted, as well. That was effective and went as planned for the telecom companies till telecommunications was elementary: lines carried just the voice, rates seldom spiked, and customers remained loyal for a lifetime, they didn’t “churn”.

Then the corporate monolith splintered. Technology took gargantuan steps in every field. Rivalry among the service givers exacerbated, and customers quickly mastered the idea of churning — to gleefully shift from one provider to another and often back again or over to a third one. The sundry of scions of “The Phone Company” aka the modern day telecom giants did not goad anymore with the idea of taking the customers for granted.

Over time, telecom providers have reciprocated to churn in diverse manners. While, some strived to woo and preserve already existing subscribers and patrons with discounts and extra minutes, megabytes and messages, while others have undertaken further structured and thoughtful measures. The recipe for success in order to retain customers and patrons, it seems, is intimate, up-to-date knowledge of the customer segments. But that’s a tough nut to crack when the proclivity fluctuates by age group, region, income, and other factors, and when little of it remains unfazed over passage of time. Marketing teams scramble and find it arduously challenging to keep up with the everchanging dynamic customer segments.

PROJECT MOTIVATION/BACKGROUND:

“The cost of acquiring a new customer can be higher than that of retaining a customer by as much as 700%. Increasing customer retention rates by a mere 5% could increase profits by 25% to 95%.” [From Research conducted by Bain and Company]

These figures above, are pretty extraordinary and crucial. Ergo, these significant figures indicates, towards a foregone conclusion, that patron or customer retention carries most prevalence and is the pivotal objective for most businesses. Retained clients are more probable to be further involved and open-minded towards cross-selling and up-selling.

All these perceptions and insight emanate from data — “data that is unbound by silos, that is integrated by those who need the insight, and that is analysed quickly and in a dimension-free manner”.[TIBC Research pages] The telecom service-provider, that, reliably and consistently determines the customer-segments and client necessities, is the one that commands the majority of customer’s faithfulness and scoops a greater share of new customers.

In all these, the entity that holds vital importance, is analysing the right data using the right platform and also using the right analytics techniques. Prompt, swift and precise data analysis takes utmost precedence, followed by analysis of granular customer data without imposing time constraints. Predictive and event-driven analytics reveals trends, hidden relationships, and unseen patterns.

BUSINESS PROBLEM FORMULATION

Customers keep churning out from any given network .The most impending concern is churning out of High valued and esteemed customers, owing to the fact that their share in revenue generation is much more in comparison to others under a particular service provider.

Generally, service providers are conservative or reactionary in addressing the discretionary churn. “But supposing we can use the unstructured data for creating a prediction-model, that will discern or recognize the most profitable as well as most bankable telecom clients and aid in generating a dynamic and proactive program for securing the customer’s loyalty?” We, in this project have tried, using analytics and predictions of R language in order to create a model that predicts the churning from an unstructured dataset. Also, we have tried answering some of the following impending research questions which can be broadly categorised as follows:

- What attributes or variables can we employ so as to create a predictive model for customer churn in the telecom arena?
- Which predictive models should we use to best predict the customer churn?
- What are the different cost savings while employing various predictive models?

DATA DESCRIPTION

```
Classes 'tbl_df', 'tbl' and 'data.frame':       7043 obs. of  21 variables:
 $ customerID      : chr  "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
 $ gender          : chr  "Female" "Male" "Male" "Male" ...
 $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Partner         : chr  "Yes" "No" "No" "No" ...
 $ Dependents      : chr  "No" "No" "No" "No" ...
 $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
 $ PhoneService    : chr  "No" "Yes" "Yes" "No" ...
 $ MultipleLines   : chr  "No phone service" "No" "No" "No phone service" ...
 $ InternetService : chr  "DSL" "DSL" "DSL" "DSL" ...
 $ OnlineSecurity  : chr  "No" "Yes" "Yes" "Yes" ...
 $ OnlineBackup    : chr  "Yes" "No" "Yes" "No" ...
 $ DeviceProtection: chr  "No" "Yes" "No" "Yes" ...
 $ TechSupport     : chr  "No" "No" "No" "Yes" ...
 $ StreamingTV     : chr  "No" "No" "No" "No" ...
 $ StreamingMovies : chr  "No" "No" "No" "No" ...
 $ Contract        : chr  "Month-to-month" "One year" "Month-to-month" "One year" ...
 $ PaperlessBilling: chr  "Yes" "No" "Yes" "No" ...
 $ PaymentMethod   : chr  "Electronic check" "Mailed check" "Mailed check" "Bank transfer
(automatic)" ...
 $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
 $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn           : chr  "No" "No" "Yes" "No" ...
```

I. customerID
II. gender (female, male)
III. SeniorCitizen (Whether the customer is a senior citizen or not (1, 0))
IV. Partner (Whether the customer has a partner or not (Yes, No))
V. Dependents (Whether the customer has dependents or not (Yes, No))
VI. tenure (Number of months the customer has stayed with the company)
VII. PhoneService (Whether the customer has a phone service or not (Yes, No))
VIII. MultipleLines (Whether the customer has multiple lines r not (Yes, No, No phone service)
IX. InternetService (Customer's internet service provider (DSL, Fibre optic, No)
X. OnlineSecurity (Whether the customer has online security or not (Yes, No, No internet service)
XI. OnlineBackup (Whether the customer has online backup or not (Yes, No, No internet service)
XII. DeviceProtection (Whether the customer has device protection or not (Yes, No, No internet service)
XIII. TechSupport (Whether the customer has tech support or not (Yes, No, No internet service)
XIV. streamingTV (Whether the customer has streaming TV or not (Yes, No, No internet service)
XV. streamingMovies (Whether the customer has streaming movies or not (Yes, No, No internet service)
XVI. Contract (The contract term of the customer (Month-to-month, One year, Two year)
XVII. PaperlessBilling (Whether the customer has paperless billing or not (Yes, No))
XVIII. PaymentMethod (The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)))
XIX. MonthlyCharges (The amount charged to the customer monthly — numeric)
XX. TotalCharges (The total amount charged to the customer — numeric)
XXI. Churn (Whether the customer churned or not (Yes or No))

The raw data contains 7043 rows (customers) and 21 columns (features). The "Churn" column is our target.

DATA WRANGLING/PREPARATION

This stage for us was slightly mechanical and cumbersome. At the same time, it was very pivotal for us to undergo this step as converting the categorical values to numerical wouldn't have given us any sensible outcome. We did data wrangling in the following ways:

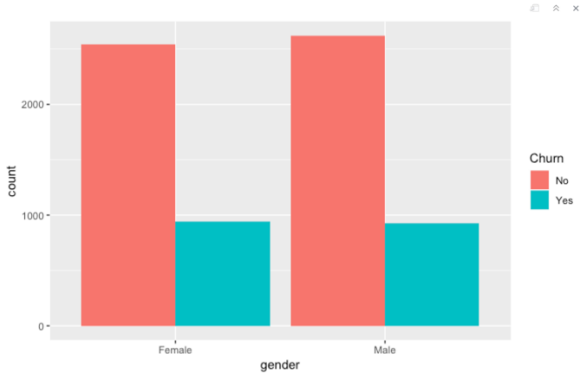
1. Cleaning up data: Some spaces, typos and unwanted symbols in the attributes which were biasing the analysis, were cleaned and the data was formatted wherever it wasn't uniform.
2. Handling missing data: Some attributes had value which were missing or NA. While we could have deleted all the 11 rows which were having 'NA' or missing values, but instead we replaced them with value 0 through imputation. All missing value in Total Charges are found when the tenure is 0 indicating the customer might not have been billed till then. So, we can impute value "0" to replace the missing data.
3. Reducing variability in the categorical field: We changed "No internet service" to "No" for six columns, they are: "OnlineSecurity", "OnlineBackup", "DeviceProtection", "TechSupport", "streamingTV", "streamingMovies".
   We also Change the values in column "SeniorCitizen" from 0 or 1 to "No" or "Yes" respectively. We also changed the datatype of column "SeniorCitizen" from 'integer' to 'factor'.
4. Data Partitioning: We split the Training data into 80% Training Set and 20% Validation set.
5. Pre-processing using the recipe library: The Recipe library basically saves the series of steps used in our pre-processing and allows us to reuse the sets for any new data. Down the line afterwards if we want to add other steps, we just add it to the existing recipe so that the data need not be retrained from the start (helpful for big datasets). Steps involved in our process are:
   - We convert the tenure into different groups to compare the probability to churn among the groups.
   - We convert SeniorCitizen to a factor.
   - We apply BoxCox transformation to our TotalCharges variable to reduce the skewness and normalise the variable.
   - We add dummy variables to the categorical variables.
   - We standardize data(Subtract mean and divide by SD), to improve the prediction power, eg. improve the KNN algorithm so that TotalCharges distance doesn't shadow the categorical variable distance.

EXPLORATORY DATA ANALYSIS

To assist our analysis, we did some deeper exploration of the customer segments, trying to answer the following questions:

- Are men more likely to churn than women?
- Are senior citizens more like to churn?
- Do individuals with a partner churn more than those without a partner?
- Do people with dependents churn more than people that do not have dependents?

We explored "gender" as we didn't expect one gender to be more likely than another to churn.

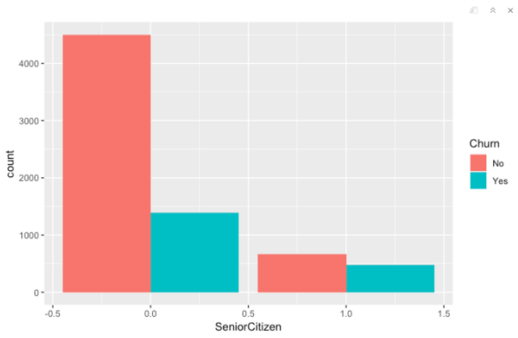


gender <chr>	Churn <chr>	n <int>
Female	No	2544
Female	Yes	939
Male	No	2619
Male	Yes	930

4 rows

Taking a look, the results are similar. Roughly one quarter of the male customers churn, and roughly one quarter of the female customers churn. We can also take a look at exactly how many people from each gender churned as shown above.

Next we take a look at senior citizens.

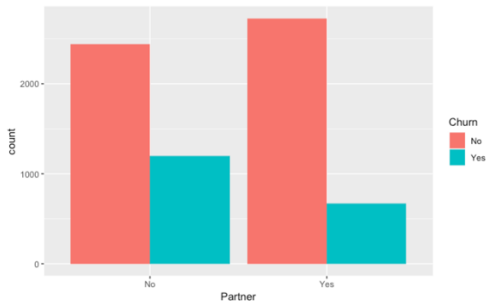


SeniorCitizen <int>	n <int>	freq <dbl>
0	5890	0.8375995
1	1142	0.1624005

SeniorCitizen <int>	Churn <chr>	n <int>	freq <dbl>
0	No	4497	0.7634975
0	Yes	1393	0.2365025
1	No	666	0.5831874
1	Yes	476	0.4168126

This variable shows a much more meaningful relationship. Roughly 16% of the customers are senior citizens, and roughly 42% of those senior citizens churn. On the other hand, of the 84% of customers that are not senior citizens, only 24% churn. These results show that senior citizens are much more likely to churn.

Now we consider people with partners.

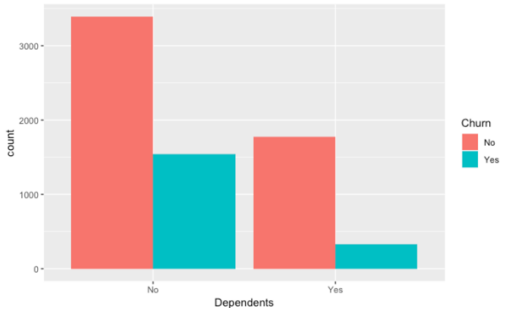


Partner <chr>	n <int>	freq <dbl>
No	3639	0.5174915
Yes	3393	0.4825085

Partner <chr>	Churn <chr>	n <int>	freq <dbl>
No	No	2439	0.6702391
No	Yes	1200	0.3297609
Yes	No	2724	0.8028294
Yes	Yes	669	0.1971706

We observe that roughly half of the people have partners. Of the people with partners, 20% churn. For people without partners, approximately 33% churn.

Next, we take a look at the Dependents category.

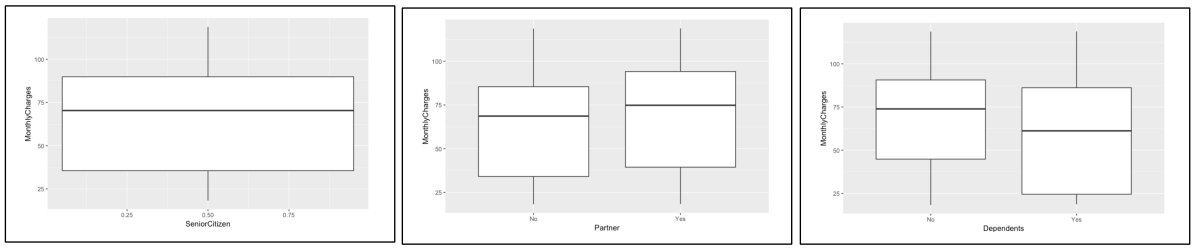


Dependents <chr>	n <int>	freq <dbl>
No	4933	0.7015074
Yes	2099	0.2984926

Dependents <chr>	Churn <chr>	n <int>	freq <dbl>
No	No	3390	0.6872086
No	Yes	1543	0.3127914
Yes	No	1773	0.8446879
Yes	Yes	326	0.1553121

Approximately 30% of the people have dependents, of which 15% churn. For the other 70% that don't have dependents, 31% churn.

Another useful visualization is the box and whisker plot. This gives us a little bit more compact visual of our data, and helps us identify outliers. We also take a look at some box and whisker plots for total charges of the different customer segments.



After looking at these initial results, we must compare the total charges of senior citizens, people without partners, and people without dependents.

n <int>	total <dbl>	avg_tenure <dbl>
476	38419.6	21.03361

1 row

n <int>	total <dbl>	avg_tenure <dbl>
1200	85741.15	13.17667

1 row

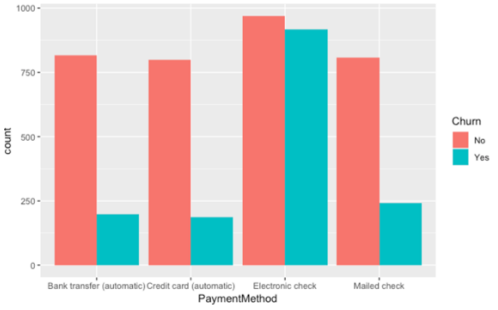
n <int>	total <dbl>	avg_tenu <db
1543	115376.5	17.123

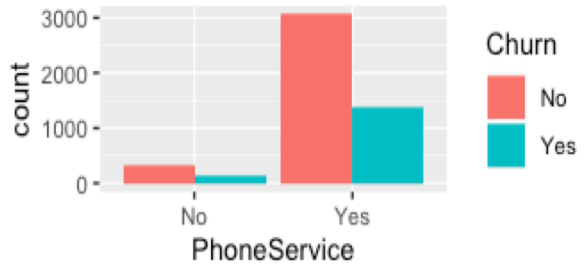
1 row

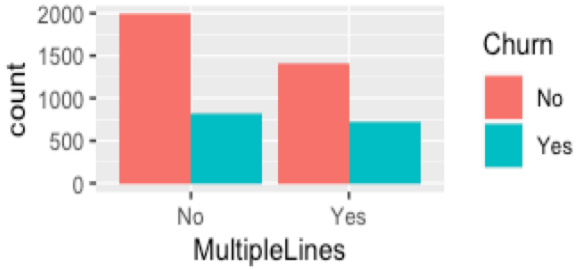
These seem to be the subsets of people most likely to churn within their respective customer segments. Let's compare them so that we can identify where we would potentially focus our efforts.

Based on these above results, we should focus our efforts on people without dependents. This customer segment that churned had nearly 2.3MM in total charges compared to 1.3MM for people without partners, and only 900K for senior citizens.

We go even further and find what services the customer segment with no dependent uses using the following GRIDPLOTS:



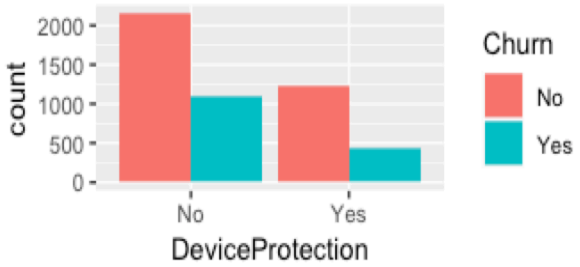

count
3000
2000
1000
0
No
Yes
PhoneService
Churn
No
Yes

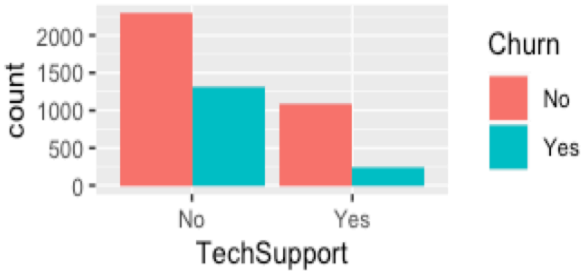

count
2000
1500
1000
500
0
No
Yes
MultipleLines
Churn
No
Yes

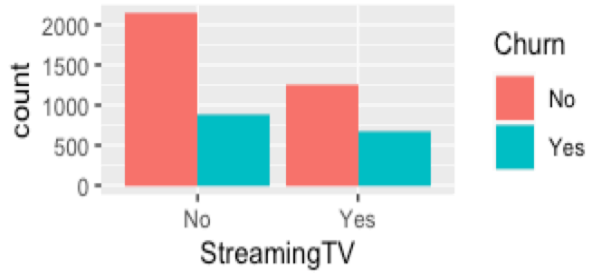

count
1000
500
0
DSL
Fiber optic
No
InternetService
Churn
No
Yes

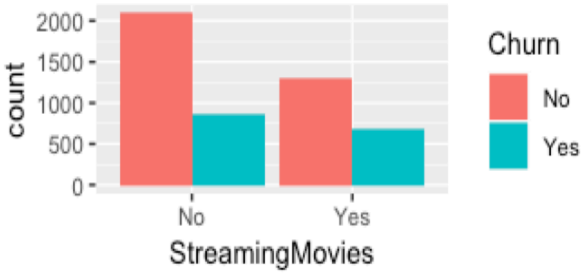

count
2000
1500
1000
500
0
No
Yes
OnlineSecurity
Churn
No
Yes


count
2000
1500
1000
500
0
No
Yes
OnlineBackup
Churn
No
Yes

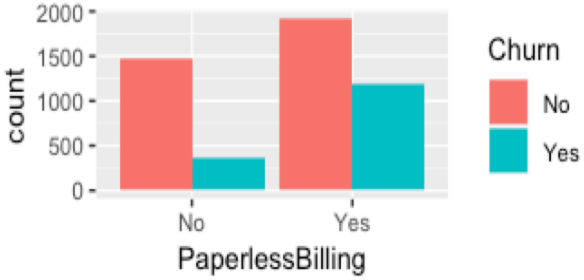

count
2000
1500
1000
500
0
No
Yes
DeviceProtection
Churn
No
Yes


count
2000
1500
1000
500
0
No
Yes
TechSupport
Churn
No
Yes


count
2000
1500
1000
500
0
No
Yes
StreamingTV
Churn
No
Yes


count
2000
1500
1000
500
0
No
Yes
StreamingMovies
Churn
No
Yes


count
1500
1000
500
0
Month-to-month
One year
Two year
Contract
Churn
No
Yes


count
2000
1500
1000
500
0
No
Yes
PaperlessBilling
Churn
No
Yes

Taking a look at this initial data exploration results, we gain some potential insights. Based on these insights, here are some recommendations for improving customer retention:

- A lot of people with phone service churned. Maybe these people don't really use the phone service. Moving them to a plan without phone service to save them some money on their bill might help retain them.
- People with fibre optic internet churned much more than people with DSL or no internet at all. Maybe moving some of those people to DSL or eliminating their internet service would be an option. Another option could be some sort of price reduction to their fibre optic plan as some sort of a promotion for being a loyal customer.
- People without online backup, device protection, and online security churn fairly frequently. Maybe their devices have crashed, causing them to lose valuable files. They may have also experienced fraud or identity theft that has left them very unhappy. Moving these people to some of these services may help safeguard their systems, thus preventing a lot of unwanted headaches.
- Similarly to online backup and security, those without device protection tended to churn more than those that subscribed to the service. Adding device protection to their plans may be a good way to prevent churn.
- Those without tech support tend to churn more frequently than those with tech support. Moving customers to tech support accounts might be another potential way to prevent churn.

MODELS AND ANALYSIS

Some steps involved in our process are:

1. We convert the tenure into different groups to compare the probability to churn among the groups.
2. We also convert Senior-Citizen to a factor
3. We apply log- transformation to our Total-Charges variable to reduce the skewness and normalise the variable.
4. We add dummy variables to the categorical variables
5. We standardize data(Subtract mean and divide by SD), to improve the prediction power, eg. improve the KNN algorithm so that total-Charges distance doesn't shadow the categorical variable distance.

We use the following predictive models:

1. Logistic regression is a linear classifier, which makes it easier to interpret than non-linear models. At the same time, because it's a linear model, it has a high bias towards this type of fit, so it may not perform well on non-linear data.
2. Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically "decision trees". It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. (although shrinkage methods like the Lasso and Ridge Regression can help with correlated features in a logistic regression model).
3. Decision Tree visualization is used by us for illustration purpose of the variables used in our analysis.

MODELS, IMPROVEMENTS AND EVALUATIONS:

1.Logistic Regression

```
Call:
glm(formula = Churn ~ . - TotalCharges, family = "binomial",
    data = train_tbl)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0034  -0.6745  -0.2900   0.6802   3.1132

Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -1.9243504  0.4120766  -4.670 3.01e-06 ***
genderMale                       -0.0007526  0.0683452  -0.011 0.991214
SeniorCitizenYes                  0.2222187  0.0888093   2.502 0.012342 *
PartnerYes                       -0.0279834  0.0818638  -0.342 0.732480
DependentsYes                    -0.0504977  0.0945567  -0.534 0.593309
tenure2-3 years                  -0.3167477  0.1300390  -2.436 0.014859 *
tenure3-4 years                  -0.3111175  0.1426000  -2.182 0.029128 *
tenure4-5 years                  -0.5717138  0.1510955  -3.784 0.000154 ***
tenureLess than year              0.9025773  0.1019633   8.852  < 2e-16 ***
tenureMore than 5 years          -0.8700813  0.1781758  -4.883 1.04e-06 ***
PhoneServiceYes                   0.0890950  0.3078467   0.289 0.772265
MultipleLinesYes                  0.3560057  0.0927701   3.838 0.000124 ***
InternetServiceFiber optic        1.4089438  0.2397507   5.877 4.19e-09 ***
InternetServiceNo                -1.8912888  0.5447996  -3.472 0.000517 ***
OnlineSecurityYes                -0.2794919  0.0998476  -2.799 0.005123 **
OnlineBackupYes                  -0.1041458  0.0908639  -1.146 0.251723
DeviceProtectionYes               0.0755758  0.0935008   0.808 0.418923
TechSupportYes                   -0.2882864  0.0997571  -2.890 0.003854 **
StreamingTVYes                    0.4287998  0.1196512   3.584 0.000339 ***
StreamingMoviesYes                0.4721448  0.1207045   3.912 9.17e-05 ***
ContractOne year                 -0.7948494  0.1135713  -6.999 2.58e-12 ***
ContractTwo year                 -1.6399564  0.1913974  -8.568  < 2e-16 ***
PaperlessBillingYes               0.3150633  0.0787831   3.999 6.36e-05 ***
PaymentMethodCredit card (automatic) -0.0821744  0.1194682  -0.688 0.491557
PaymentMethodElectronic check     0.3000517  0.0996149   3.012 0.002594 **
PaymentMethodMailed check        -0.0340306  0.1203183  -0.283 0.777301
MonthlyCharges                   -0.7936019  0.3684497  -2.154 0.031248 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7322.8  on 6328  degrees of freedom
Residual deviance: 5275.3  on 6302  degrees of freedom
AIC: 5329.3

Number of Fisher Scoring iterations: 6
```

Analysing the summary table, we can understand the variable significance and beta coefficients of the logistic model... Positive coefficients of variables like internet-services (fibre-optics service), payment method (electronic check), SeniorCitizen indicate higher probability of churn for these customer groups. Negative coefficients like in higher year tenure groups and online services groups and yearly contract groups indicates a decreasing relationship with probability of churn.

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  916 198
       Yes  98 194

               Accuracy : 0.7895
                 95% CI : (0.7672, 0.8105)
    No Information Rate : 0.7212
    P-Value [Acc > NIR] : 2.627e-09

                  Kappa : 0.4321
 Mcnemar's Test P-Value : 8.702e-09

            Sensitivity : 0.4949
            Specificity : 0.9034
```

We observe from above confusion matrix that Logistic regression model has accuracy of 78.95%. The model has better interpretability and evaluating predictor significance.
Assessing the predictive ability of the Logistic Regression model, we find the following result:

```
log.predDir  No Yes
        No  397  37
        Yes 116 153
```

2.LDA/QDA

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  913 188
       Yes 101 204

               Accuracy : 0.7945
                 95% CI : (0.7724, 0.8153)
    No Information Rate : 0.7212
    P-Value [Acc > NIR] : 1.667e-10

                  Kappa : 0.4515
 Mcnemar's Test P-Value : 4.219e-07

            Sensitivity : 0.5204
            Specificity : 0.9004
```

LDA

- We observe from the confusion matrix that LDA model's accuracy is 79.45%.
- We may interpret a LDA model to have Improved accuracy but less interpretability.

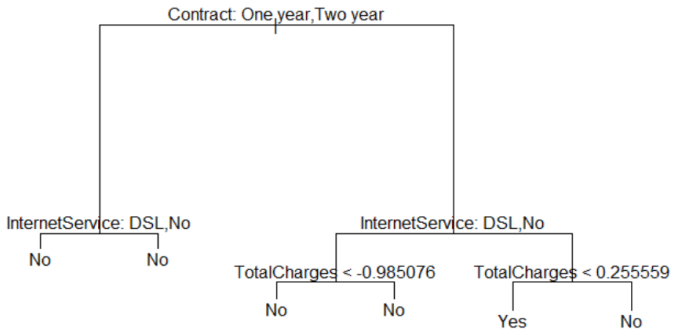
```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  773  91
       Yes 241 301

               Accuracy : 0.7639
                 95% CI : (0.7408, 0.7859)
    No Information Rate : 0.7212
    P-Value [Acc > NIR] : 0.0001627

                  Kappa : 0.4745
 Mcnemar's Test P-Value : 2.899e-16

            Sensitivity : 0.7679
            Specificity : 0.7623
```

QDA

- We observe from the confusion matrix that QDA model's accuracy is 76.39%.
- We may interpret a QDA model to have degraded accuracy but better sensitivity (76.79%).

3.Decision Tree & KNN

Decision tree model




```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  903 206
       Yes 111 186

               Accuracy : 0.7745
                 95% CI : (0.7518, 0.7961)
    No Information Rate : 0.7212
    P-Value [Acc > NIR] : 3.055e-06

                  Kappa : 0.3943
 Mcnemar's Test P-Value : 1.295e-07

            Sensitivity : 0.4745
            Specificity : 0.8905
```

From the above confusion matrix for the decision tree model, the accuracy is 77.45%.

Pruned Decision tree model

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  914 216
       Yes 100 176

               Accuracy : 0.7752
                 95% CI : (0.7525, 0.7968)
    No Information Rate : 0.7212
    P-Value [Acc > NIR] : 2.272e-06

                  Kappa : 0.3853
 Mcnemar's Test P-Value : 9.849e-11

            Sensitivity : 0.4490
            Specificity : 0.9014
```

- Decision tree pruned by limiting the number of terminal nodes to 5
- We were able to Improve the model accuracy to 77.52%, by pruning the decision tree

KNN (K=100)

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  903 176
       Yes 111 216

               Accuracy : 0.7959
                 95% CI : (0.7738, 0.8167)
    No Information Rate : 0.7212
    P-Value [Acc > NIR] : 7.299e-11

                  Kappa : 0.4652
 Mcnemar's Test P-Value : 0.0001582

            Sensitivity : 0.5510
            Specificity : 0.8905
```

- From the confusion matrix, the KNN model accuracy is 79.52%.
- We further tune K possible to improve the model accuracy.

Faster modelling, no data normalization are some of the advantages of using decision tree method but it is prone to overfitting. We can further improve the decision tree possibly though ensemble methods like boosting.

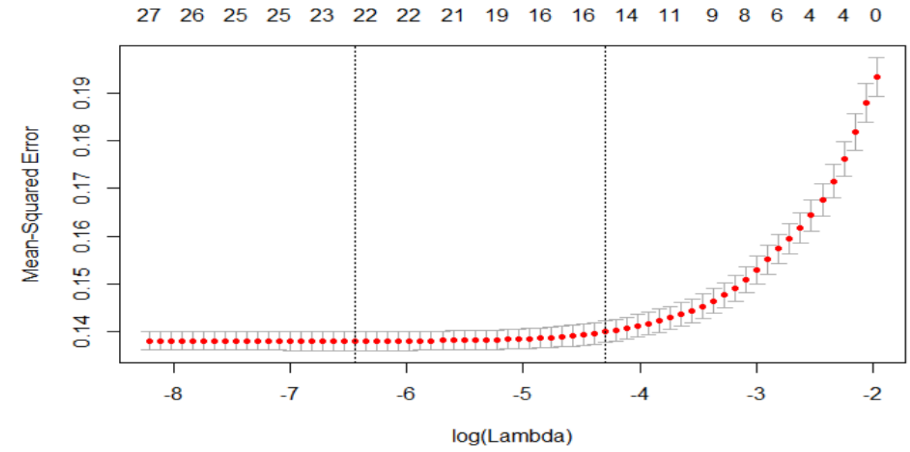
4.Improvement Models

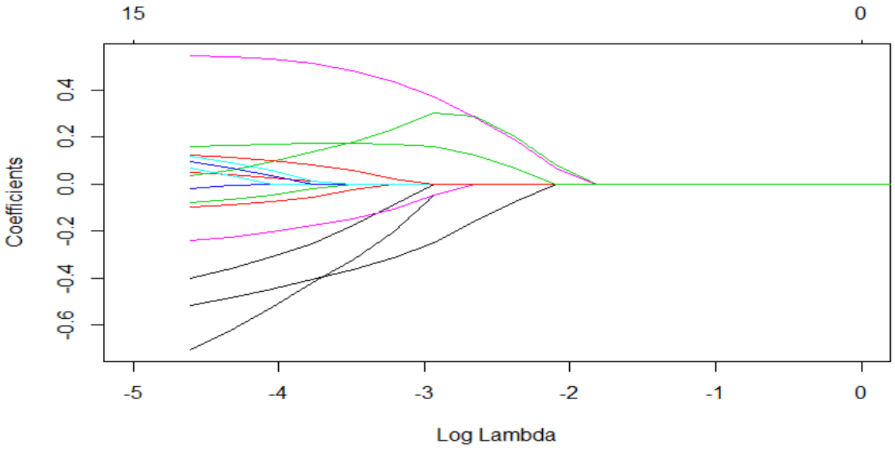
Shrinkage method (Binomial Lasso model)

```
(Intercept)                          -1.45694812
MonthlyCharges                        .
TotalCharges                         -0.70451216
gender_Male                           .
SeniorCitizen_Yes                     0.04790992
Partner_Yes                           .
Dependents_Yes                        .
tenure_X2.3.years                     .
tenure_X3.4.years                     .
tenure_X4.5.years                     .
tenure_Less.than.year                 0.03505287
tenure_More.than.5.years             -0.02093087
PhoneService_Yes                      .
MultipleLines_Yes                     0.06825893
InternetService_Fiber.optic           0.54813079
InternetService_No                   -0.40335716
OnlineSecurity_Yes                   -0.09737925
OnlineBackup_Yes                      .
DeviceProtection_Yes                  .
TechSupport_Yes                      -0.08131494
StreamingTV_Yes                       0.09361545
StreamingMovies_Yes                   0.11883289
Contract_One.year                    -0.24226256
Contract_Two.year                    -0.51672897
PaperlessBilling_Yes                  0.12455352
PaymentMethod_Credit.card..automatic.  .
PaymentMethod_Electronic.check        0.15766886
```

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  938 207
       Yes  76 185

               Accuracy : 0.7987
                 95% CI : (0.7768, 0.8194)
    No Information Rate : 0.7212
    P-Value [Acc > NIR] : 1.329e-11

                  Kappa : 0.4423
 Mcnemar's Test P-Value : 1.095e-14

            Sensitivity : 0.4719
            Specificity : 0.9250
```

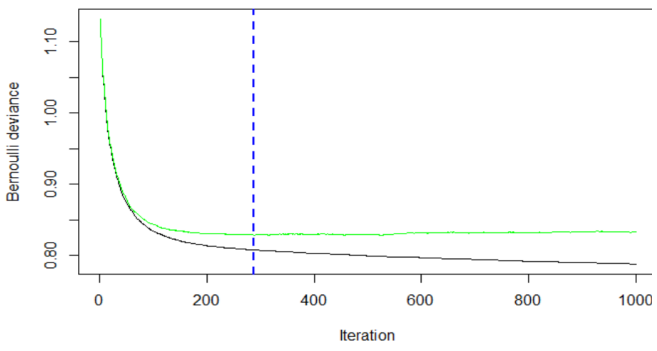
- Here, the coefficients of non-significant predictors are reduced to zero.
- Decrease in variables reduces overfitting, thereby decreasing variance while testing with unknown data.
- Model accuracy has improved to 79.87% but with reduced sensitivity on testing.





Gradient boosted model

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  932 187
       Yes  82 205

               Accuracy : 0.8087
                 95% CI : (0.7871, 0.8289)
    No Information Rate : 0.7212
    P-Value [Acc > NIR] : 1.968e-14

                  Kappa : 0.4817
 Mcnemar's Test P-Value : 2.283e-10

            Sensitivity : 0.5230
            Specificity : 0.9191
```

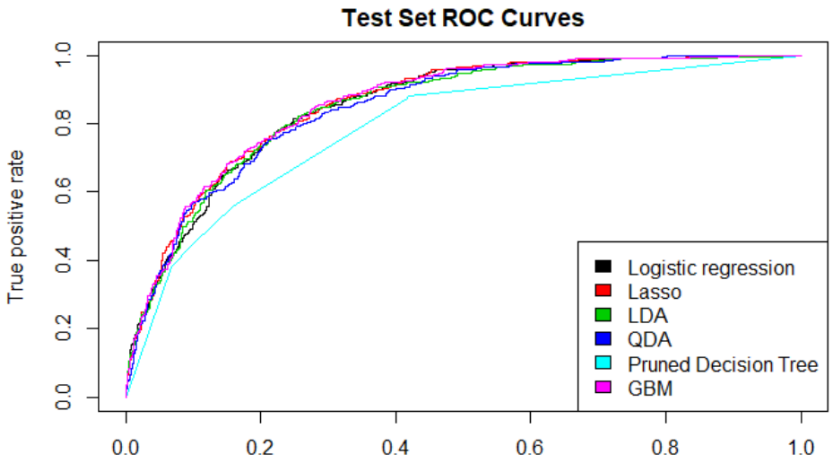


The Hyperparameters Used for Gradient Boost are:

- interaction.depth=1 (Maximum depth of each tree)
- bag.fraction = 0.4 (fraction of training test used to propose for next tree)
- cv.folds=10 (10- fold cross validation)
- n.trees =1000 (Total #trees to fit)
- shrinkage = 0.1 (learning rate)

From Gradient boosting, we get best in class accuracy of 80.87%. The Hyperparameter tuning was required to improve the model accuracy. Early stopping using 10 fold cross-validation was performed to eliminate overfitting problem with GBM models.

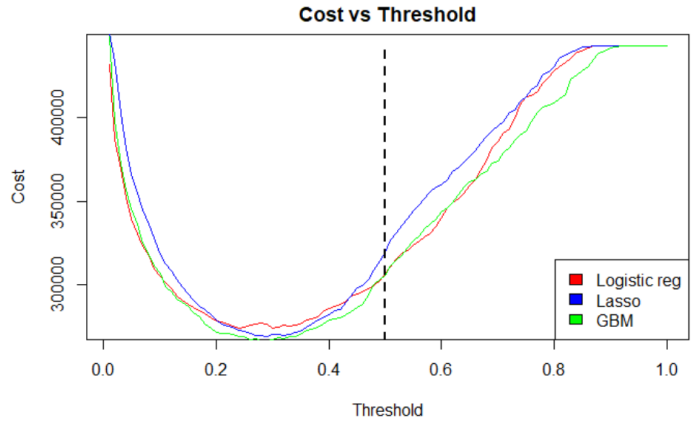
4. Model Evaluation – ROC and AUC



	auc
PruneTree	0.7902
QDA	0.8477
LDA	0.8505
logistic	0.8529
lasso	0.8585
GBM	0.8595

We are taking AUC as measure of model evaluation – higher the better. GBM is having the highest area under the ROC curve – 0.8595.

5. Cost Considerations




```
"Cost without model: $ 539400"
"Cost with model: $ 68440"
"Total cost savings: $ 470960"
```

The following assumptions are taken for the cost analysis of our models:

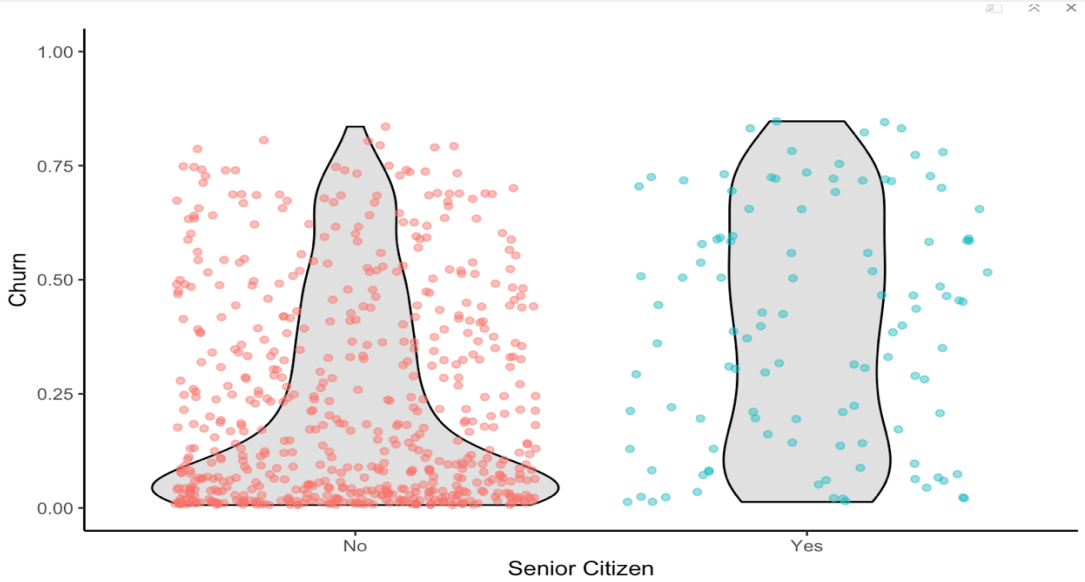
- The cost of a customer churn = $300
- The cost of offers offered to churnable customer = $80
- All customers offered a offer was retained.

We get better cost savings by fixing a threshold near 0.30 rather 0.5, while taking into account cost of TP and FP/FN predictions.
We get Cost savings of 87.31% ($470960) using a prediction model on test dataset (~1400 user).

INFERENCE GRAPHS

#Senior Citizen

```
ggplot(aes(test_tbl$SeniorCitizen,log.prob),data=test_tbl)+
  geom_violin(col="black", fill = "lightgrey", alpha = 0.7)+
  geom_jitter(aes(col=test_tbl$SeniorCitizen), alpha =0.5,show.legend = FALSE)+
  scale_x_discrete(name="Senior Citizen")+
  scale_y_continuous(name="Churn", limits = c(0,1))+
  theme_classic()
```

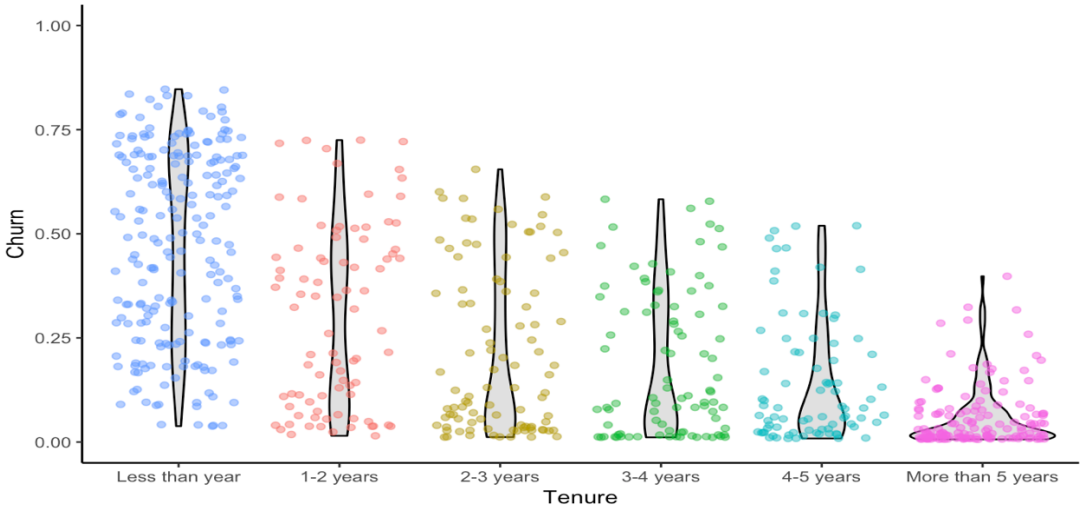


Inference: Senior citizen appeared in several of the cases indicating it was important to the model for all samples. However, it was not highly correlated to Churn, which may indicate that the model is using in an more sophisticated manner (e.g. as an interaction). It's difficult to say that senior citizens are more likely to leave, but non-senior citizens appear less at risk of churning.

Opportunity: Target users in the lower age demographic.

#Tenure

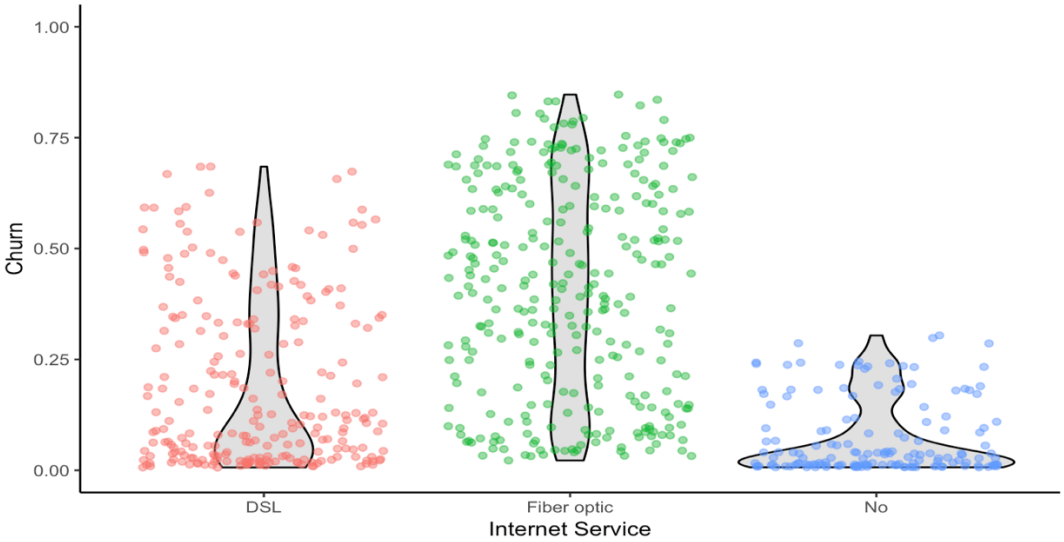
```
ggplot(aes(test_tbl$tenure,log.prob),data=test_tbl)+
  geom_violin(col="black", fill = "lightgrey", alpha = 0.7)+
  geom_jitter(aes(col=test_tbl$tenure), alpha =0.5,show.legend = FALSE)+
  scale_x_discrete(name="Tenure", limits = c("Less than year","1-2 years","2-3 years","3-4
years","4-5 years","More than 5 years"))+
  scale_y_continuous(name="Churn", limits = c(0,1))+
  theme_classic()
```



Inference:
There is an indication that the predictive model is using this feature frequently and high correlation agrees that this is important. Investigating the feature distribution, it appears that customers with lower tenure (bin 1) are more likely to leave.
Opportunity: Target customers with less than 12 month tenure.

#InternetService

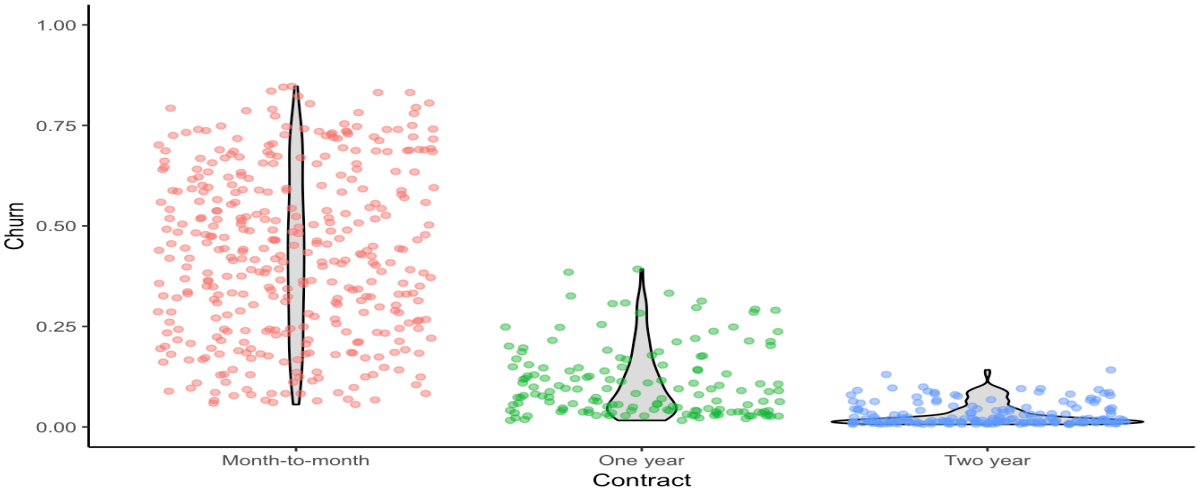
```
ggplot(aes(test_tbl$InternetService,log.prob),data=test_tbl)+
  geom_violin(col="black", fill = "lightgrey", alpha = 0.7)+
  geom_jitter(aes(col=test_tbl$InternetService), alpha =0.5,show.legend = FALSE)+
  scale_x_discrete(name="Internet Service")+
  scale_y_continuous(name="Churn", limits = c(0,1))+
  theme_classic()
```



Inference:
While the predictive models did not indicate this as a primary feature in the sample cases, the feature is clearly correlated with those electing to stay. Customers with fibre optic service are more likely to churn while those with no internet service are less likely to churn. **Improvement Area:** Customers may be dissatisfied with fibre optic service.

#Contract

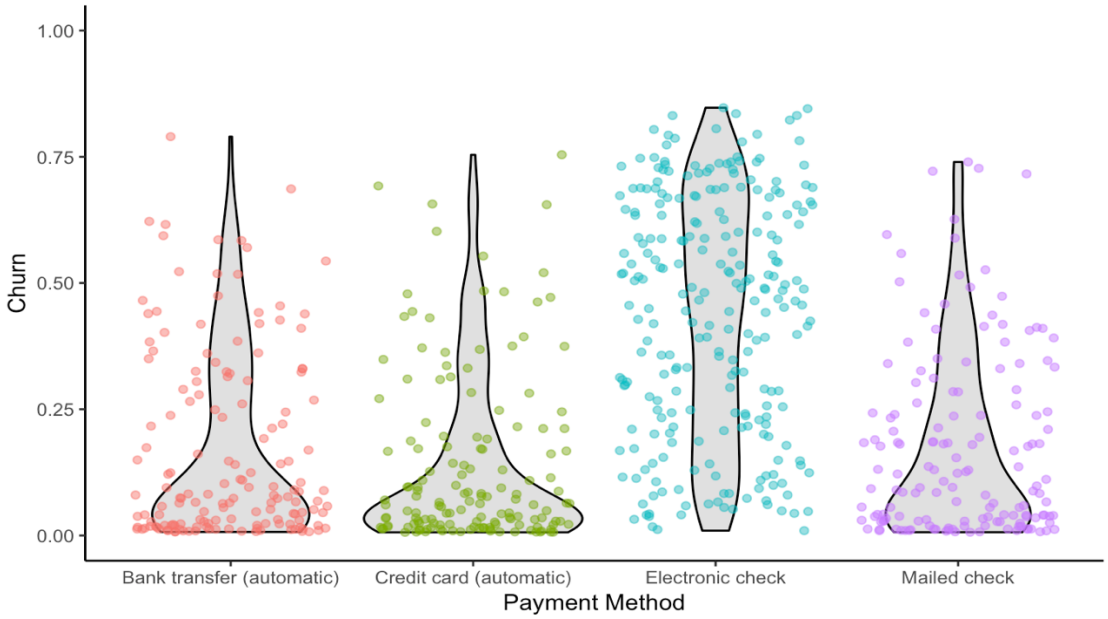
```
ggplot(aes(test_tbl$Contract,log.prob),data=test_tbl)+
  geom_violin(col="black", fill = "lightgrey", alpha = 0.7)+
  geom_jitter(aes(col=test_tbl$Contract), alpha =0.5,show.legend = FALSE)+
  scale_x_discrete(name="Contract")+
  scale_y_continuous(name="Churn", limits = c(0,1))+
  theme_classic()
```



Inference:
While the predictive models did not indicate this as a primary feature in the sample cases, the feature is clearly correlated with those electing to stay. Customers with one and two year contracts are much less likely to churn.
Opportunity: Offer promotion to switch to long term contracts.

#PaymentMethod

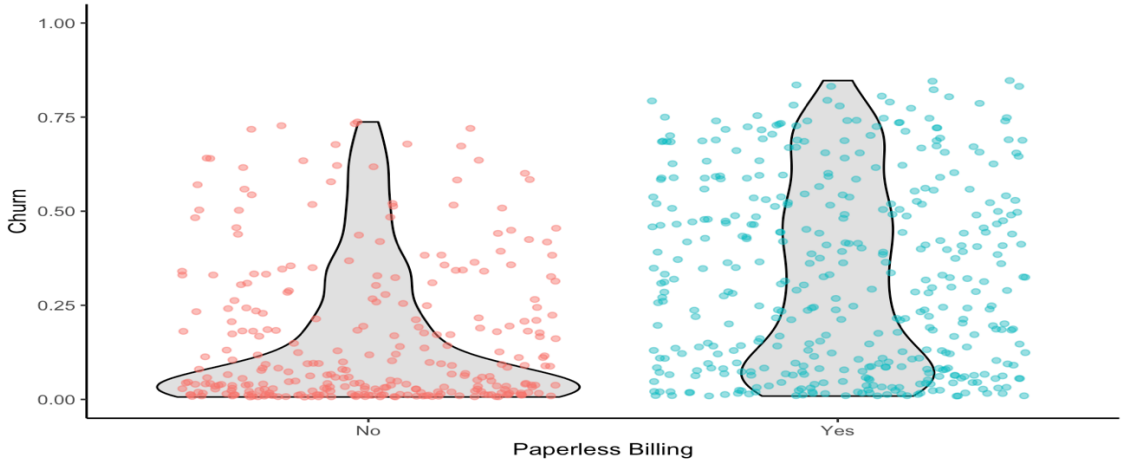
```
ggplot(aes(test_tbl$PaymentMethod,log.prob),data=test_tbl)+
  geom_violin(col="black", fill = "lightgrey", alpha = 0.7)+
  geom_jitter(aes(col=test_tbl$PaymentMethod), alpha =0.5,show.legend = FALSE)+
  scale_x_discrete(name="Payment Method")+
  scale_y_continuous(name="Churn", limits = c(0,1))+
  theme_classic()
```



Inference:
While the predictive model did not indicate this as a primary feature in the sample cases, the feature is clearly correlated with those electing to stay. Customers with electronic check are more likely to leave.
Opportunity: Offer customers a promotion to switch to automatic payments.

#PaperlessBilling

```
ggplot(aes(test_tbl$PaperlessBilling,log.prob),data=test_tbl)+
  geom_violin(col="black", fill = "lightgrey", alpha = 0.7)+
  geom_jitter(aes(col=test_tbl$PaperlessBilling), alpha =0.5,show.legend = FALSE)+
  scale_x_discrete(name="Paperless Billing")+
  scale_y_continuous(name="Churn", limits = c(0,1))+
  theme_classic()
```

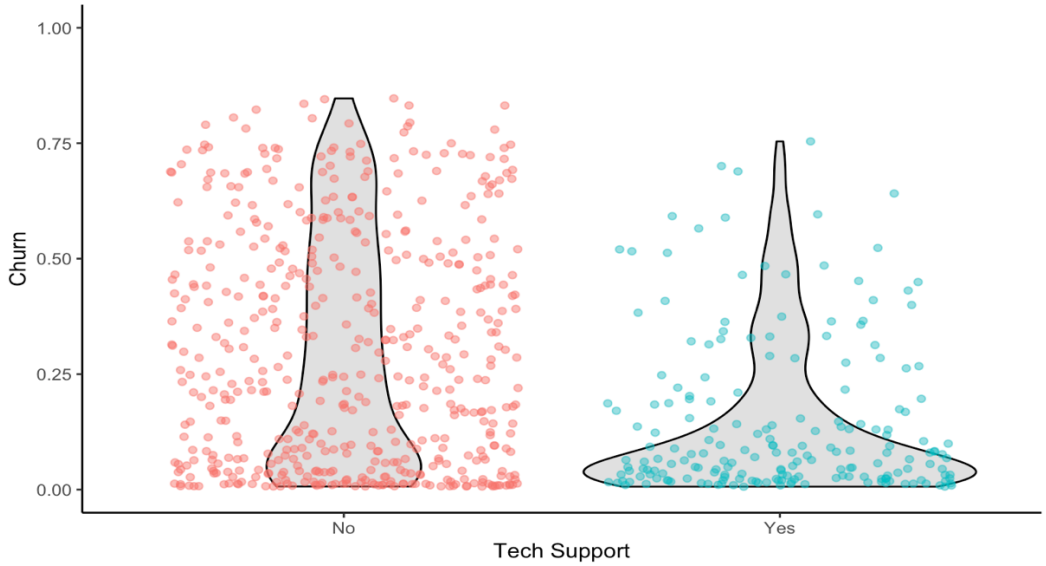


Inference: Paperless Billing appeared in several of the cases indicating it was important to the model for all samples. However, it was not highly correlated to Churn. So it's tough to tell whether customers preferring Paperless Billing are more likely to leave, but customers not preferring Paperless Billing appear less at risk of churning.

Opportunity: This criteria may not provide much opportunity for the company.

#TechSupport

```
ggplot(aes(test_tbl$TechSupport,log.prob),data=test_tbl)+
  geom_violin(col="black", fill = "lightgrey", alpha = 0.7)+
  geom_jitter(aes(col=test_tbl$TechSupport), alpha =0.5,show.legend = FALSE)+
  scale_x_discrete(name="Tech Support")+
  scale_y_continuous(name="Churn", limits = c(0,1))+
  theme_classic()
```



Inference:

Customers that did not get proper tech support were more likely to leave while customers who were given the right tech support for their problems were less likely to leave. **Opportunity:** Promote better tech support for customers that increase retention rates.

SUMMARY

Throughout the analysis, we learnt several important things:

- Features such as tenure_group, Contract, PaperlessBilling, MonthlyCharges and InternetService appear to play a role in customer churn.
- There does not seem to be a relationship between gender and churn.
- Customers in a month-to-month contract, with PaperlessBilling and are within 12 months tenure, are more likely to churn; On the other hand, customers with one or two year contract, with longer than 12 months tenure, that are not using PaperlessBilling, are less likely to churn.

Telecommunication industry always suffers from a very high churn rates when one industry offers a better plan than the previous there is a high possibility of the customer churning from the present due to a better plan in such a scenario it is very difficult to avoid losses but through prediction we can keep it to a minimal level. In this paper the method used is Logistic Regression (backward logistic regression) and this helps to identity the probable churn customers and then make the necessary business decisions. Using a decision tree would give a more appropriate result, by using logistic regression the result achieved is 80.02% accurate.

REFERENCES

1. https://towardsdatascience.com/ai-101-understanding-customer-churn-management-514416c17643
2. https://blogs.rstudio.com/tensorflow/posts/2018-01-11-keras-customer-churn/
3. https://www.tutorialspoint.com/r/
4. http://www.adobe.com/in/solutions/digital-analytics/customer-churn-analysis.html
5. Business Intelligence and Insurance, White Paper, Wipro Technologies,Bangalore,2001
6. http://www.galitshmueli.com/sites/galitshmueli.com/files/Predicting%20Consumer%20Churn%20Report.pdfJiawei Han and Micheline Kamber, Data mining, concept and techniques" http://www.cs.sfu.ca.
7. https://jtsulliv.github.io/churn-eda/
8. http://www.alteryx.com/solutions/customer-churn-analytics
9. L. Yangi , C. Chiu , Subscriber Churn Prediction in Telecommunications
10. http://www.rdatamining.com/
11. http://www.ats.ucla.edu/stat/r/dae/logit.htm
12. Telecommunication Subscribers' Churn Prediction Model Using Machine Learning Saad Ahmed Qureshi, Ammar Saleem
13. https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/