Project 2 White Paper

To Churn or Not to Churn… That is the Question

OPIM 5604 - B12: Predictive Modeling

Team 2

This report was developed by Team 2:

Meghna Ashok, Luke D’Agostino, Mohit Khanna, Logan Miller, Stuti Viyulie

Table of Contents

Executive Summary	**3**
Problem Statement	**3**
Methodology	**4**
Sample	5
Explore	5
Modify	6
Model	8
Assess	9
Results	**10**
Conclusions & Recommendations	**11**
References	**13**
Appendix	**14**
Appendix A: Dataset Description	14
Appendix B: Data Partitions	16
Appendix C: Data Exploration	17
Appendix D: Outlier Exploration	22
Appendix E: Model Results - Group 1 (All Variables)	23
Appendix F: Model Results - Group 2 (Subset of Variables)	31
Appendix G: Model Ranking - Assessment	37

Executive Summary

The purpose of this report is to build a predictive model to predict the Target Variable - **Churn- "Yes/No".** The chosen dataset was explored for innate trends using various unsupervised learning techniques like Data Visualization, Affinity Analysis etc.18 out of the 20 columns in our data set were of the categorical type. We converted these variable types to continuous variables by using dummy variables in order to make the data ready to feed into particular models. Our dataset had no outliers due to it being heavy in categorical variables. There were 11 missing values in the dataset which we excluded from our analysis. Several pureplay and ensemble models were utilized on the dataset and the best one was selected.We concluded that some of the strongest models we recommend our business partners to use would be the Group 2 significant variables models, specifically the nominal logistic model, with the ensemble model and the neural network model as runner ups.

Based on our analysis we concluded the following:

- Customers with multiple service lines are more likely to churn out
- Senior Citizens are more likely to churn out compared to middle aged customers
- Tenure is the most significant variable based on our logistic regression model; customers are more likely to churn if they have been recently on boarded by the Telco
- Customers are more likely to churn out if they subscribe to a monthly payment plan

Problem Statement

A telecom company has expanded its product offerings to offer more than just traditional ISP and phone services. The company has been offering digital services like smart TV, OTT bundled services along with their phone lines. These services have helped them capture a more young and tech savvy customer demographic. This company believes that these customers have a higher upselling and cross selling potential. However, the company has recently noticed an increase in their customer churn rate.The company wants to answer the following questions:

1. Is there a specific customer type that is churning out?
2. For customers that have churned out, are they customers of a specific product?
3. Is there a correlation between churned out customers and the number of lines that they are subscribed to?
4. If a customer is subscribed to additional security and general services (such as device protection, tech support, etc.) are they less likely to churn out?
5. If a customer has dependents on their plan are they less likely to churn?

We are hoping to answer these questions by building a predictive model based on this classification problem. Through this model we will try to predict the target variable **Churn- "Yes/No".**

By understanding the different features of the historical customers who have churned out like demographic, product types and additional subscriptions, the company will be able to work towards improving those products and reducing churn rate.

If the company is losing a certain demographic of customer, say the telco is experiencing heavy churn rate among the older generation of customers, the company can better focus on those customers to get them to subscribe back to their plans. Maybe they can come up with new products that better suit their needs. The company could also tweak their existing product portfolio to better suit the needs of this demographic of customers

Methodology

We have undertaken the five-step SEMMA process (Sample, Explore, Modify, Model, and Assess) as our methodology for this project.

Sample

The dataset we have chosen is 'Telcom Customer Churn data' from Kaggle. The dataset was a part of IBM Cognos Analytics 11.1.3. It contains information about a telecommunication company that provided home phone and Internet services to 7043 customers in California. The original data contains 7043 rows (customers) and 21 columns (features). The "Churn" column is our target, which is a categorical variable that indicates whether the customer has churned or not (Yes or No). The columns and their data types can be found in Appendix A [A]. As a part of the Sampling process in SEMMA, we have also partitioned our data into training, validation and test sets, in the ratio of 60:20:20 [B].

Explore

Some of the relevant patterns and correlations we found through visualization are listed below. Please see screenshots for the findings below under Appendix C [C].

1. **Senior Citizen vs Churn:** Out of all customers, 6.8% of customers are senior citizens and have churned, and 476 out of 1142 senior citizens have churned i.e almost 42% of Senior Citizens have churned. Hence, Senior Citizens are more likely to churn than middle aged customers and this is an interesting predictor.
2. **Partner vs Churn**: 17% of customers churned, having no partner. Among customers without partners, 33% churned. i.e. Customers without partners are more likely to churn as compared to customers with partners.
3. **Dependent vs Churn:** Overall, Customers without dependents churned more than customers with dependents. Among customers without dependents, 31% of them churned. This implies that customers without dependents are more likely to churn compared to customers with dependents.
4. **Phone service vs Churn:** Customers with phone service are more likely to churn as

compared to customers without phone service.

5. **Multiple Lines vs Churn:** Customers with multiple lines have a slightly higher churn rate. Only a small percentage of customers don't have a phone service connection.
6. **Internet Service vs Churn:** 44% of total customers have Fiber optic InternetService and 34.4% of the customers have DSL InternetService. Customers who have not taken up the Internet service have a very low possibility of churning. Customers who have taken FiberOptic Internet service are more likely to churn than customers who have taken DSL. This is interesting because DSL connections are slower and more expensive than FiberOptic, but are still less likely to churn.
7. **Contract vs Churn**: Month to Month contract seems to be the most popular among customers, and has the highest chance of churning.
8. **Paperlessbilling vs Churn:** Customers who have taken up paperless billing are more likely to churn than customers who have not.
9. **Payment method vs Churn**: Electronic check seems to be the most popular payment method among customers, and also has the highest churn rate.
10. **Services taken up by the customer vs Churn rate:** Customers who have taken up all the 4 services- Online security, Online backup, Device protection and Tech support are less likely to churn, compared to customers who have not taken up these services.
11. **Tenure vs Churn:** There is slight differentiation between the variables for certain ranges. Customers having Tenure<15 months are highly likely to Churn, i.e. Newer customers are more likely to churn.

Modify

Dummy Variables

The following columns were converted from categorical variables to dummy variables, with values 1 (Indicating ‘Yes’/presence in column) and 0 (Indicating a ‘No’). This was done using Indicator columns, to be able to perform math on our categorical variables more easily. Note that we have made new indicator columns one fewer than the number of categories in each column.

- **Gender**: New column created Gender (Female).
- **Partner:** New column created Partner (Yes).
- **Dependents**: New column created Dependants (Yes).
- **Phone Service**: New column created Phone Service (Yes).
- **Multiple Lines**: New columns created: Multiple Lines (No phone service) and Multiple Lines (Yes). When both have the value 0, it indicates Multiple Lines (No).
- **Internet Service**: New columns created: Internet Service (DSL) and Internet Service (Fiber optic). When both have the value 0, it indicates Internet Service (No).
- **Online Security:** New columns created: Online Security (Yes) and Online Security (No). When both have the value 0, it indicates Online Security (No Internet Service).
- **Online Backup:** New columns created: Online Backup (Yes) and Online Backup (No). If both have the value 0, it indicates Online Backup (No Internet Service).
- **Device Protection:** New columns created: Device Protection (Yes) and Device Protection (No). When Device Protection (Yes) and Device Protection (No) both have the value 0, it indicates Device Protection (No Internet Service).
- **Tech Support:** New columns created: Tech Support (Yes) and Tech Support (No). When both have the value 0, it indicates Tech Support (No Internet Service).
- **Streaming TV**: New columns created: Tech Support (Yes) and Tech Support (No). When both have the value 0, it indicates Tech Support (No Internet Service).
- **Streaming Movies:** New columns created: Streaming Movies (Yes) and Streaming Movies (No). When Streaming Movies(Yes) and Streaming Movies(No) both have the value 0, it indicates Streaming Movies (No Internet Service).

- **Contract:** New columns created: ‘Month-to-month’ and ‘One year’. When Month-to-month and One year both have the value 0, it indicates a ‘Two year’ contract.
- **Paperless Billing:** New column: Paperlessbilling. 1 indicates Yes, 0 indicates No
- **Payment method:** New columns created: Bank transfer (automatic), Credit card (automatic) and Electronic check. When all three have the value 0, it indicates ‘Mailed check’ is the payment method.

Missing variables

TotalCharges column had 11 missing values. Since the percentage of missing values was very small, we excluded these 11 values as it wouldn’t have a significant impact on retained data.

Outliers

The dataset has no outliers for continuous variables [D]. Moreover, the continuous variables are not skewed and hence do not need standardization. We have also chosen not to perform a Principal Component Analysis on our dataset since there is not much complexity in our data. Also, since we have many categorical predictor variables, explainability is of the utmost importance with respect to making business decisions with our results.

Model

The team took a robust approach to exploring the different modeling methods which could be used to forecast future data in this field. The modeling approach was, in essence, to throw everything at the wall and see what sticks. The team ran every applicable model that was taught in the course, meaning that the models used for this project were: logistic regression, decision tree, boosted tree, bootstrap forest, neural network, discriminant analysis, k nearest neighbors, naive bayes, and an averaged ensemble model.

The team used every one of these modeling techniques with every predictor variable in the dataset in order to attempt to predict churn. Some of these predictive models were great, while others didn’t quite meet the mark. Many of the models gave very similar results. In the case of the ensemble model, it was an average of every other model built. The results for the individual models can be found in the appendix [E].

The team also noticed, however, that the initial run of the logistic regression model indicated that about half of the variables were insignificant in predicting the target variable. The team eliminated all of the predictors which were not significant, based on the threshold of a p-value <= 0.05. The results of the new logistic regression were a much better performing model. Due to this, Team 1 decided to rebuild every one of the models, this time only including the predictor variables deemed significant by the logistic regression method. These variables, in order of importance, were: tenure, Month-to-Month, Internet Service (Fiber Optic), Electronic check, Paperlessbilling, Multiple Lines (Yes), Online Security (No), TotalCharges, Phone Service (Yes), Tech Support (No), Dependents (Yes), StreamingMovies (Yes), and SeniorCitizen. Once again, for the ensemble model it was an average of every other model built with just these variables. The results for the individual models can be found in the appendix [F].

Assess

To rank the model’s performance we separated the models into two groups. Group 1 included the models with all the variables while Group 2 included the models with only the significant variables. Then the accuracy of the performance was sorted by partition in ascending order. We also checked for overfitting from the training partition to the validation partition. A model was considered overfit if its misclassification rate’s growth rate exceeded a 15% threshold from training to validation. Models that were overfit received the lowest ranking overall. This allowed us to better assess the performance of our models and interpret the results to draw conclusions. This spreadsheet can be found in the appendix [G].

Results

As mentioned in the Modeling section of the paper, the team ran the models using two sets of data, one with all of the predictor variables and the other with only those deemed significant by the logistic regression. This effectively created two groups of models to analyze.

Group 1 consisted of all the predictor variables, regardless of their significance value. The sizes of our Training, Validation and Test groups were 4,226, 1,398 and 1,408 respectively. Overfitting was accounted for by comparing the changes in misclassification rates between the Training and Validation groups. With a growth rate of 15% ((Validation Misclassification Rate - Training Misclassification Rate)/Training Misclassification Rate) or more being the identifier of an overfit model, it was found that the Ensemble Model Average, Partition model, Boosted Tree model and Bootstrap Forest model were overfit with growth rates of 16.6%, 17.7%, 31.7% and 97.2% respectively. In regards to the validation group, the best performing model for Group 1 was the Fit Nominal Logistic model with a misclassification rate of 19.6%, overfitting growth rate of only 1.3% and a RASE score of .37. Following close behind, the second best performing model for Group 1 was the Neural Model NTanH(3)NTanH2(3) with a misclassification rate of 20.1%, overfitting growth rate of 3.2% and RASE score of .37. The rest of the rankings went as follows; Neural Model NTanH(3), KNN, Discriminant, Naive Bayes and then the overfit models, Ensemble Model Average, Boosted Tree, Bootstrap Forest and Partition.

Group 2 consisted of only the predictor variables that were deemed significant. The sizes of our Training, Validation and Test groups were 4,226, 1,398 and 1,408 respectively. Overfitting was accounted for in the same manner as for Group 1. The models that were found to be overfit were the Boosted Tree model and the Bootstrap Forest model with growth rates of 22.6% and 65.6% respectively. When looking at the validation group, the best performing model for Group 2 was the Fit Nominal Logistic model with a misclassification rate of 19.6%, overfitting growth rate of 1.3% and a RASE score of .37. It is interesting to note that there was no change in the measurement results for this model between Group 1 and Group 2. The second best performing model for Group 2 was the Ensemble Model Average with a misclassification rate of 19.7%, overfitting growth rate of 14.1% and RASE score of .37. The rest of the rankings went as follows; Neural Model NTanH(3), Neural Model NTanH(3), Partition, KNN, Naive Bayes, Discriminant and then the overfit models, Boosted Tree and Bootstrap Forest.

With these results, we decided to go with Group 2 where only the significant predictor variables were used. Even though the Nominal Logistic Model performed the same in both groups, Group 2's second best performing model was the Ensemble Model Average whereas in Group 1 this model was overfit. This shows that using Group 2 was a better performer as it also helped the issue of overfitting.

Conclusions & Recommendations

Based on the results from our modeling efforts to predict whether a customer will churn we have concluded that some of the strongest models we recommend our business partners to use the Group 2 significant variables models, specifically the nominal logistic model with the ensemble model and the neural network model as runner ups. These models had the most accuracy with the least amount of overfitting. After interpreting the results from these models we recommend the following for our business partners:

We found that customers with multiple lines are more likely to churn than customers with only one service. More services leads to more opportunities to experience dissatisfaction. In addition, senior citizens were slightly more likely to churn than middle aged customers. Telco should cater enhanced customer service and deals and discounts to combat unpleasurable experiences to prevent customer churn. These customers should be flagged in their customer demographics system because they are more susceptible to churn. Additionally, the company can promote and offer more Family plans and Group-centric offers since customers with partners or dependents are less likely to churn.

The nominal logistic regression revealed that customer tenure is one of the most important predictors of customer churn, specifically newer customers are more likely to churn. We recommend that Telco recognizes newer customers have more risk of churning and combating urges to depart. For example, this can be achieved by offering better customer service experiences and deals or discounts to keep newer customers engaged longer. Telco's responsiveness to customer dissatisfaction will be crucial in maintaining customer retention. They can offer a discount or a deal to show that they care about their customers and are committed to maintaining their business. Discounts and Incentives can also be offered for DSL internet connections, as customers taking up DSL are less likely to churn as compared to customers taking up FiberOptics. One possible reason for this is that DSL is more widely available than FiberOptics, and is hence more practical for the average customer.

Telco can promote packages/bundle deals consisting of additional services like Online security, Online backup, Device protection and Tech support, at a discount for new customers and existing customers having a phone service or Internet connection, since customers who have taken up all 4 additional facilities are less likely to churn, and will continue with their services. Another recommendation is to push yearlong payment agreements rather than month to month arrangements. Customers that pay month to month have an easier way to cancel services rather than a customer who is contractually obligated and will incur a fee for breaching a contract. Longer term plans combined with better customer service will discourage customers from seeking services from other competitors. In conclusion, we recommend focusing on the "at risk" customer demographics and the characteristics that usually lead to churning to decrease their churn rate in the future.

References

BlastChar. “Telco Customer Churn.” *Kaggle*, 23 Feb. 2018, https://www.kaggle.com/blastchar/telco-customer-churn?select=WA_Fn-UseC_-Telco-Customer-Churn.csv.

Shmueli, Galit, et al. *Data Mining for Business Analytics: Concepts, Techniques, and Applications with JMP Pro*. John Wiley & Sons, Inc., 2017.

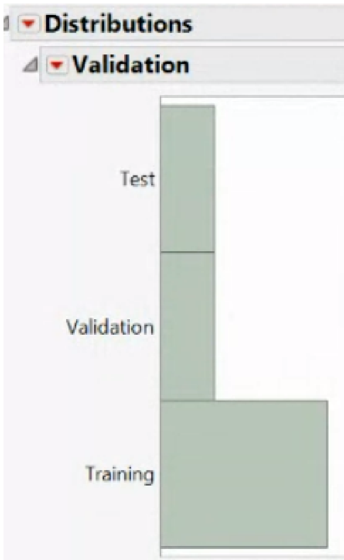
Appendix

Appendix A: Dataset Description

Column Name	Data Type	Description
CustomerID	Categorical	Unique ID for each customer
Gender	Categorical	Customer’s Gender (Male/Female)
Age	Categorical	Customer’s age at the time of sampling (in years)
SeniorCitizen	Categorical	Indicates whether or not the customer is a senior citizen, 65+ (1=Yes, 0=No)
Partner	Categorical	Indicates if the customer is married (Yes/No)
Dependents	Categorical	Indicates if the customer lives with dependents (children, parents, grandparents, etc.) with values (Yes/No)
Tenure	Continuous	Total # of months the customer has been with the Telco company
Phone Service	Categorical	Indicates if the customer subscribes to home phone service (Yes/No)
Multiple Lines	Categorical	Indicates if the customer has subscribed to multiple phone lines (Yes/No)... ‘No Phone Service’ means that this customer doesn’t subscribe to phone service (i.e. NA)
Internet Service	Categorical	Indicates if the customer subscribes to internet service w/ the Telco company & the type of service (‘DSL’/’Fiber Optic’/No)
Online Security	Categorical	Indicates if the customer subscribes to additional online security service (Yes/No)... ‘No Internet Service’ means that this customer doesn’t subscribe to internet service (i.e. NA)
Online Backup	Categorical	Indicates if the customer subscribes to additional online backup service (Yes/No/’No Internet Service’)
Device Protection Plan	Categorical	Indicates if the customer subscribes to additional device protection for their internet equipment (Yes/No/’No Internet Service’)
Tech Support	Categorical	Indicates if the customer subscribes to additional tech support service with their internet plan (Yes/No/’No Internet Service’)
Streaming TV	Categorical	Indicates if the customer uses their internet service to stream television programming (Yes/No/’No Internet Service’)

Streaming Movies	Categorical	Indicates if the customer uses their internet service to stream movies (Yes/No/’No Internet Service’)
Contract	Categorical	Indicates the customer’s current contract type (‘Month-to-Month’/‘One Year’/‘Two Year’)
Paperless Billing	Categorical	Indicates if the customer has chosen paperless billing (Yes/No)
Payment Method	Categorical	Indicates the mode which the customer uses to pay their bills (‘Electronic Check’/’Mailed Check’/’Bank Transfer (automatic)’/’Credit Card (automatic)’)
Monthly Charge	Continuous	The customer’s current monthly charge for all their services
Total Charges	Continuous	The customer’s total amassed charges
Churn	Categorical	Indicates whether or not the customer has churned (Yes/No) **This is the target variable**

Appendix B: Data Partitions



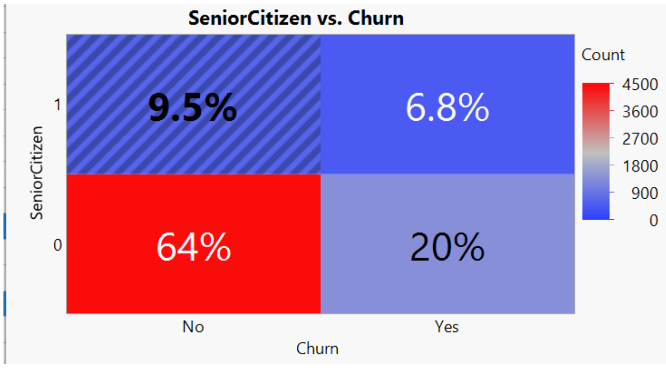
Frequencies

Level	Count	Prob
Training	4219	0.59997
Validation	1409	0.20037
Test	1404	0.19966
Total	7032	1.00000

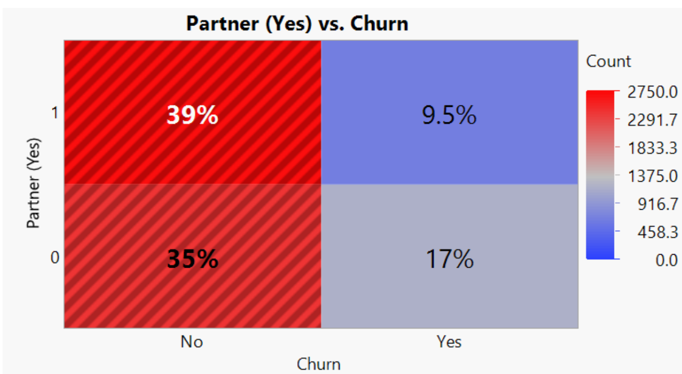
N Missing 0
3 Levels

Appendix C: Data Exploration

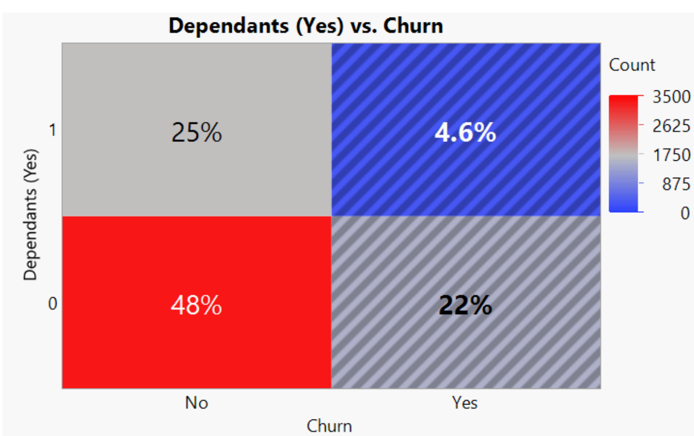
[1] **Senior Citizens** are more likely to churn than middle aged customers.



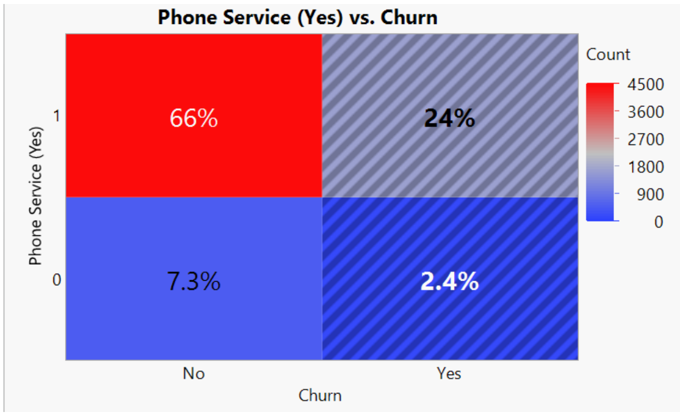
[2] Customers **without partners** are more likely to churn as compared to customers with partners.



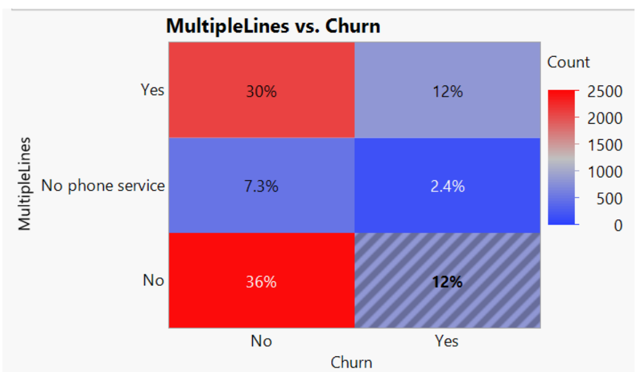
[3] Customers **without dependents** are more likely to churn compared to customers with dependents.



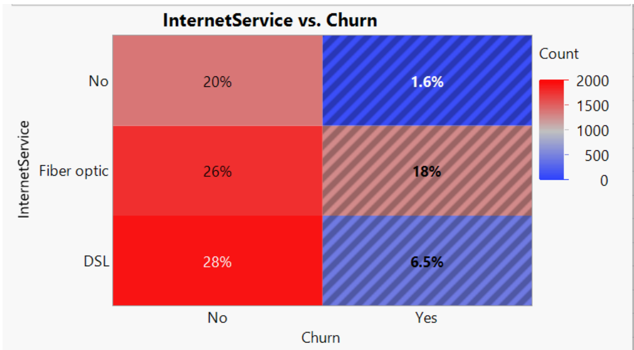
[4] Customers **with phone service** are more likely to churn as compared to customers without phone service. Only a small percentage of customers don't have phone service.



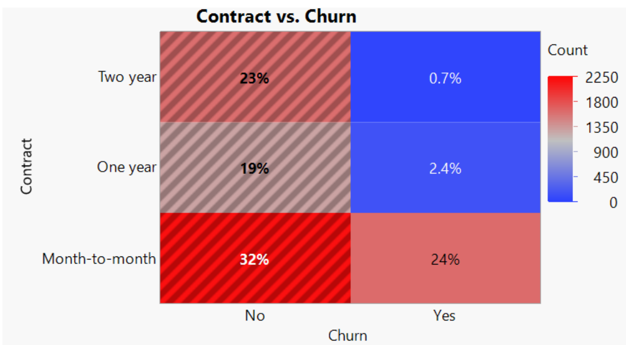
[5] Customers with **Multiple lines** have a slightly higher churn rate, compared to customers with 1 service.



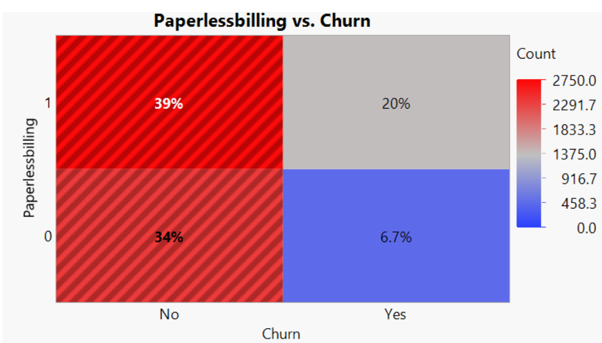
[6] Customers who have taken **FiberOptic Internet service** are more likely to churn than customers who have taken DSL. Customers who have not taken up the Internet service have a very low possibility of churning.



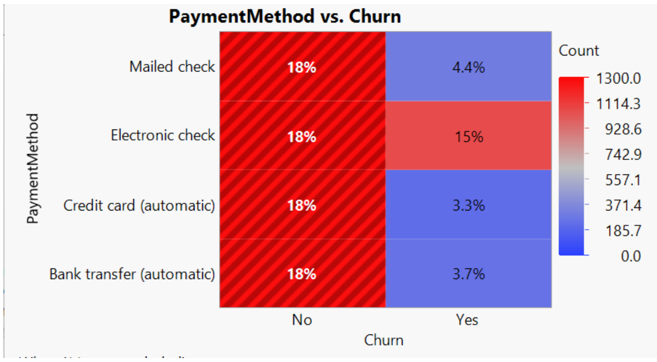
[7] **Month to Month** contract seems to be the most popular among customers, and has the highest chance of churning.



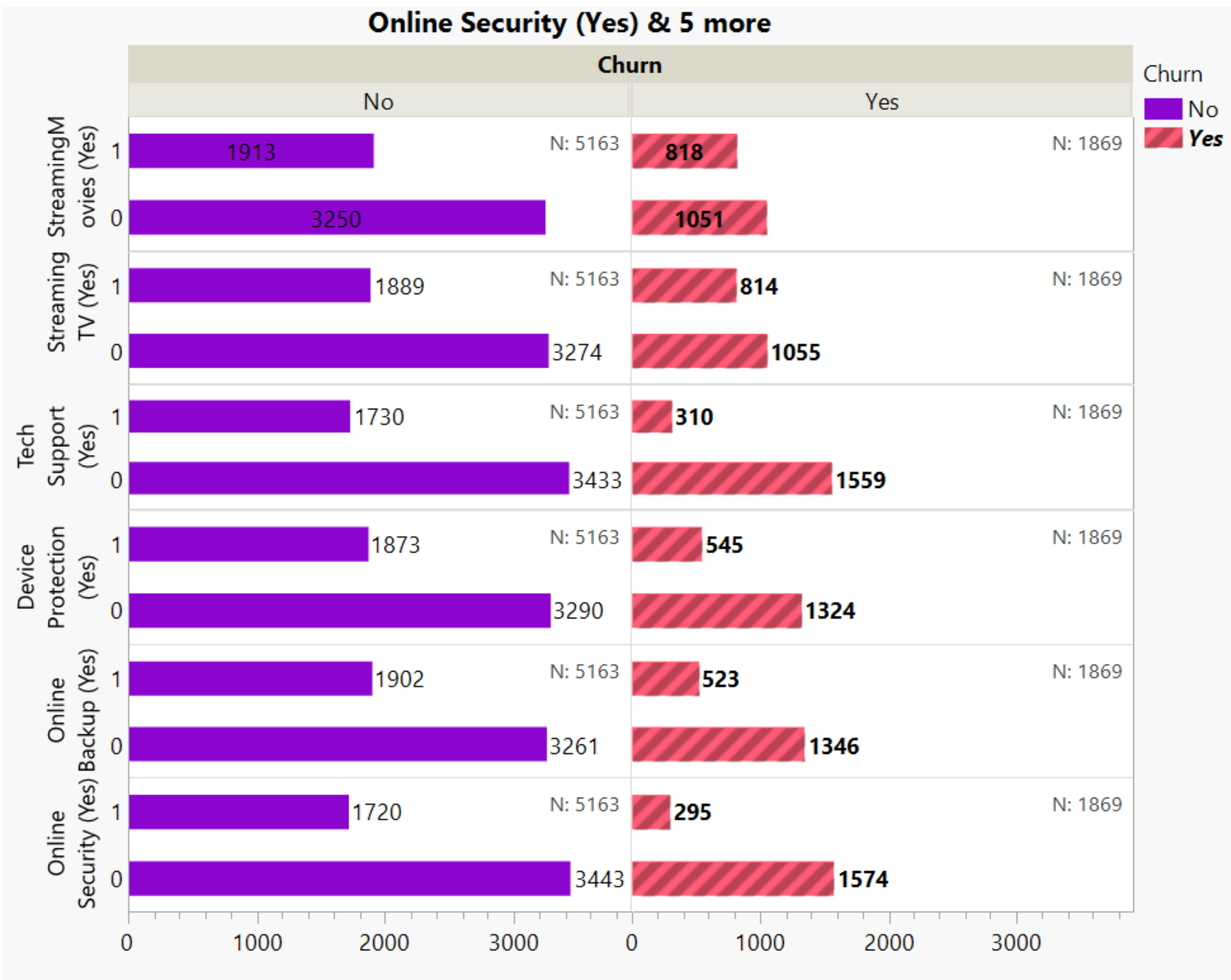
[8] Customers who have taken up **Paperless billing** are more likely to churn than customers who have not.



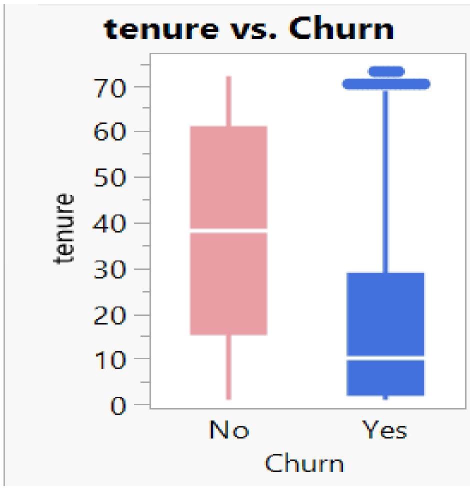
[9] **Electronic check** seems to be the most popular payment method among customers, and also has the highest churn rate.



[10] Customers who have taken up all the **4 services- Online security, Online backup, Device protection and Tech support** are less likely to churn, compared to customers who have not taken up these services.



[11] Newer customers (**Tenure**<15 months) are more likely to churn.


tenure vs. Churn
70
60
50
40
30
20
10
0
tenure
No
Yes
Churn

Appendix D: Outlier Exploration

Column	10% Quantile	90% Quantile	Low Threshold	High Threshold	Number of Outliers	Outliers (Count)
tenure	2	69	-199	270	0	
MonthlyCharges	20.05	102.65	-227.75	350.45	0	
TotalCharges	84.53	5978.86	-17598	23661.9	0	

Appendix E: Model Results - Group 1 (All Variables)

Logistic Regression

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.2928	0.2782	0.2382	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.4189	0.4024	0.3515	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.4084	0.4207	0.4415	$\sum -\mathrm{Log}(\rho[j])/n$
RASE	0.3637	0.3707	0.3774	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.2653	0.2715	0.2759	$\sum \lvert y[j]-\rho[j]\rvert/n$
Misclassification Rate	0.1936	0.1960	0.2038	$\sum (\rho[j] \neq \rho\mathrm{Max})/n$
N	4226	1398	1408	n

Confusion Matrix

Training

Actual Churn	Predicted Count Yes	Predicted Count No
Yes	620	497
No	321	2788

Actual Churn	Predicted Rate Yes	Predicted Rate No
Yes	0.555	0.445
No	0.103	0.897

Validation

Actual Churn	Predicted Count Yes	Predicted Count No
Yes	203	174
No	100	921

Actual Churn	Predicted Rate Yes	Predicted Rate No
Yes	0.538	0.462
No	0.098	0.902

Test

Actual Churn	Predicted Count Yes	Predicted Count No
Yes	200	175
No	112	921

Actual Churn	Predicted Rate Yes	Predicted Rate No
Yes	0.533	0.467
No	0.108	0.892

Decision Tree

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.2969	0.2351	0.2278	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.4238	0.3483	0.3382	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.4061	0.4459	0.4476	∑ -Log(ρ[j])/n
RASE	0.3638	0.3830	0.3812	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.2652	0.2806	0.2766	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1933	0.2275	0.2159	∑ (ρ[j]≠ρMax)/n
N	4226	1398	1408	n

Confusion Matrix

Training

Actual Churn	Predicted Count No	Predicted Count Yes
No	2780	329
Yes	488	629

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.894	0.106
Yes	0.437	0.563

Validation

Actual Churn	Predicted Count No	Predicted Count Yes
No	881	140
Yes	178	199

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.863	0.137
Yes	0.472	0.528

Test

Actual Churn	Predicted Count No	Predicted Count Yes
No	899	134
Yes	170	205

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.870	0.130
Yes	0.453	0.547

Boosted Tree

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.4041	0.2602	0.2170	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.5445	0.3801	0.3240	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3441	0.4313	0.4538	∑ -Log(ρ[j])/n
RASE	0.3307	0.3730	0.3807	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.2393	0.2708	0.2755	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1564	0.2060	0.2088	∑ (ρ[j]≠ρMax)/n
N	4226	1398	1408	n

Confusion Matrix

Training

Actual Churn	Predicted Count No	Predicted Count Yes
No	2913	196
Yes	465	652

Validation

Actual Churn	Predicted Count No	Predicted Count Yes
No	920	101
Yes	187	190

Test

Actual Churn	Predicted Count No	Predicted Count Yes
No	930	103
Yes	191	184

Training

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.937	0.063
Yes	0.416	0.584

Validation

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.901	0.099
Yes	0.496	0.504

Test

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.900	0.100
Yes	0.509	0.491

Bootstrap Forest

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.5182	0.2547	0.2155	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.6575	0.3732	0.3221	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.2783	0.4345	0.4547	∑ -Log(ρ[j])/n
RASE	0.2896	0.3758	0.3815	√ ∑(y[j]-ρ[j])^2/n
Mean Abs Dev	0.2097	0.2739	0.2777	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1060	0.2117	0.2145	∑ (ρ[j]≠ρMax)/n
N	4226	1398	1408	n

Confusion Matrix

Training

Actual Churn	Predicted Count No	Predicted Count Yes
No	3002	107
Yes	341	776

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.966	0.034
Yes	0.305	0.695

Validation

Actual Churn	Predicted Count No	Predicted Count Yes
No	913	108
Yes	188	189

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.894	0.106
Yes	0.499	0.501

Test

Actual Churn	Predicted Count No	Predicted Count Yes
No	919	114
Yes	188	187

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.890	0.110
Yes	0.501	0.499

Neural Network

Training

Churn

Measures	Value
Generalized RSquare	0.4436072
Entropy RSquare	0.3135743
RASE	0.3578331
Mean Abs Dev	0.2565787
Misclassification Rate	0.1836252
-LogLikelihood	1675.2996
Sum Freq	4226

Confusion Matrix

Actual	Predicted Count	
Churn	No	Yes
No	2786	323
Yes	453	664

Confusion Rates

Actual	Predicted Rate	
Churn	No	Yes
No	0.896	0.104
Yes	0.406	0.594

Validation

Churn

Measures	Value
Generalized RSquare	0.3818809
Entropy RSquare	0.2615969
RASE	0.3760086
Mean Abs Dev	0.2727438
Misclassification Rate	0.2052933
-LogLikelihood	601.75275
Sum Freq	1398

Confusion Matrix

Actual	Predicted Count	
Churn	No	Yes
No	907	114
Yes	173	204

Confusion Rates

Actual	Predicted Rate	
Churn	No	Yes
No	0.888	0.112
Yes	0.459	0.541

Test

Churn

Measures	Value
Generalized RSquare	0.3449256
Entropy RSquare	0.2330232
RASE	0.3796913
Mean Abs Dev	0.2736474
Misclassification Rate	0.2073864
-LogLikelihood	625.88991
Sum Freq	1408

Confusion Matrix

Actual	Predicted Count	
Churn	No	Yes
No	912	121
Yes	171	204

Confusion Rates

Actual	Predicted Rate	
Churn	No	Yes
No	0.883	0.117
Yes	0.456	0.544

Discriminant Analysis

Score Summaries

Source	Count	Number Misclassified	Percent Misclassified	Entropy RSquare	-2LogLikelihood
Training	4226	979	23.1661	0.14287	4183.84
Validation	1398	343	24.5351	0.12671	
Test	1408	365	25.9233	0.07707	

Training

Actual Churn	Predicted Count No	Predicted Count Yes
No	2380	729
Yes	250	867

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.766	0.234
Yes	0.224	0.776

Validation

Actual Churn	Predicted Count No	Predicted Count Yes
No	773	248
Yes	95	282

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.757	0.243
Yes	0.252	0.748

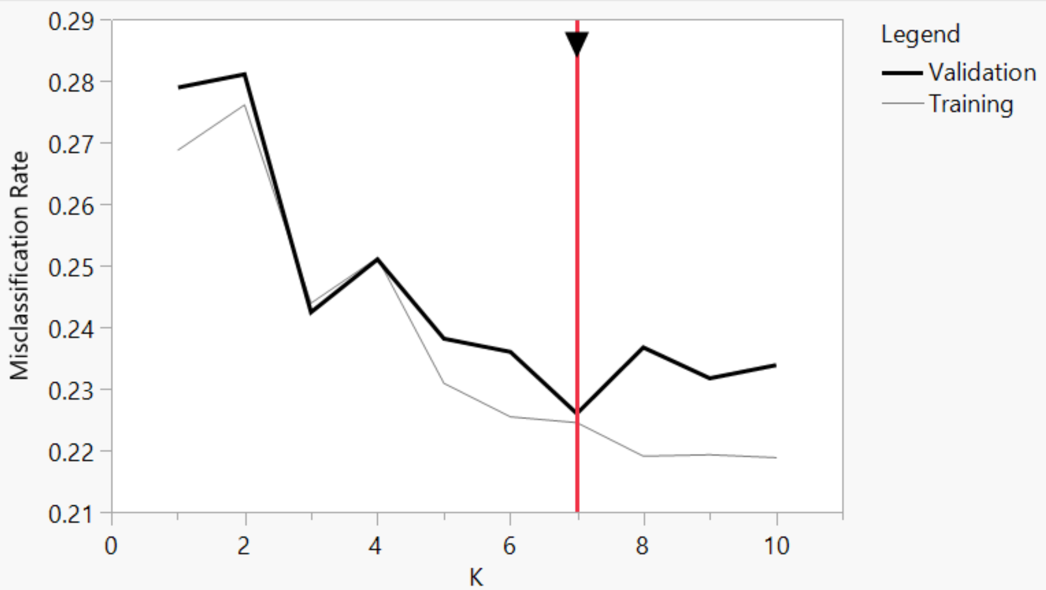
Test

Actual Churn	Predicted Count No	Predicted Count Yes
No	773	260
Yes	105	270

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.748	0.252
Yes	0.280	0.720

K Nearest Neighbors

Model Selection



Training

K	Count	RSquare	Misclassification Rate	Misclassifications	
1	4226	-0.0648	0.26881	1136	
2	4226	0.0117	0.27615	1167	
3	4226	0.0621	0.24397	1031	
4	4226	0.11993	0.25130	1062	
5	4226	0.14114	0.23095	976	
6	4226	0.16861	0.22551	953	
7	4226	0.18124	0.22456	949	
8	4226	0.19585	0.21912	926	
9	4226	0.20391	0.21936	927	
10	4226	0.21232	0.21888	925	*

Validation

K	Count	RSquare	Misclassification Rate	Misclassifications	
1	1398	-0.083	0.27897	390	
2	1398	0.01703	0.28112	393	
3	1398	0.06507	0.24249	339	
4	1398	0.09101	0.25107	351	
5	1398	0.13248	0.23820	333	
6	1398	0.15855	0.23605	330	
7	1398	0.17575	0.22604	316	*
8	1398	0.19607	0.23677	331	
9	1398	0.20717	0.23176	324	
10	1398	0.20967	0.23391	327	

Test

K	Count	RSquare	Misclassification Rate	Misclassifications	
1	1408	-0.1666	0.30682	432	
2	1408	-0.0427	0.28622	403	
3	1408	-0.0057	0.27273	384	
4	1408	0.04919	0.26918	379	
5	1408	0.08018	0.26349	371	
6	1408	0.10451	0.25568	360	
7	1408	0.11064	0.24645	347	
8	1408	0.12799	0.24574	346	
9	1408	0.1358	0.24077	339	
10	1408	0.14355	0.23793	335	*

Confusion Matrix for Best K=7

Training

Actual Churn	Predicted Count No	Predicted Count Yes
No	2672	437
Yes	512	605

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.859	0.141
Yes	0.458	0.542

Validation

Actual Churn	Predicted Count No	Predicted Count Yes
No	880	141
Yes	175	202

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.862	0.138
Yes	0.464	0.536

Test

Actual Churn	Predicted Count No	Predicted Count Yes
No	865	168
Yes	179	196

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.837	0.163
Yes	0.477	0.523

Naive Bayes

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	-0.647	-0.629	-0.903	1-Loglike(model)/Loglike(0)
Generalized RSquare	-1.622	-1.573	-2.695	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.9510	0.9498	1.1031	∑ -Log(ρ[j])/n
RASE	0.4452	0.4536	0.4657	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.2385	0.2458	0.2543	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.2343	0.2468	0.2472	∑ (ρ[j]≠ρMax)/n
N	4226	1398	1408	n

Confusion Matrix

Training

Actual Churn	Predicted Count No	Predicted Count Yes
No	2401	708
Yes	282	835

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.772	0.228
Yes	0.252	0.748

Validation

Actual Churn	Predicted Count No	Predicted Count Yes
No	782	239
Yes	106	271

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.766	0.234
Yes	0.281	0.719

Test

Actual Churn	Predicted Count No	Predicted Count Yes
No	793	240
Yes	108	267

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.768	0.232
Yes	0.288	0.712

Averaged Ensemble Model

Model Comparison Validation=Validation

Target Churn missing a predictor for category No

Predictors

Measures of Fit for Churn

Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RASE	Mean Abs Dev	Misclassification Rate	N
Boosted Tree		0.2602	0.3801	0.4313	0.3730	0.2708	0.2060	1398
Boosted Tree		0.2602	0.3801	0.4313	0.3730	0.2708	0.2060	1398
Neural Model NTanH(3)		0.2802	0.4048	0.4196	0.3702	0.2709	0.1974	1398
Boosted Tree		0.2602	0.3801	0.4313	0.3730	0.2708	0.2060	1398
Neural Model NTanH(3)		0.2690	0.3910	0.4261	0.3731	0.2724	0.2031	1398
Partition		0.2351	0.3483	0.4459	0.3830	0.2806	0.2275	1398
Fit Nominal Logistic		0.2782	0.4024	0.4207	0.3707	0.2715	0.1960	1398
Bootstrap Forest		0.2524	0.3704	0.4358	0.3761	0.2737	0.2060	1398
K Nearest Neighbors		.	.	.	.	.	0.2246	1398
Discriminant		0.1267	0.1995	0.5091	0.4099	0.3156	0.2454	1398
Naive Bayes		.	.	.	.	.	0.2468	1398
Model Averaged		0.2754	0.3990	0.4224	0.3713	0.2789	0.2003	1398

Appendix F: Model Results - Group 2 (Subset of Variables)

Logistic Regression

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.2928	0.2782	0.2382	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.4189	0.4024	0.3515	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.4084	0.4207	0.4415	∑ -Log(ρ[j])/n
RASE	0.3637	0.3707	0.3774	$\sqrt{}$ ∑(y[j]-ρ[j])^2/n
Mean Abs Dev	0.2653	0.2715	0.2759	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1936	0.1960	0.2038	∑ (ρ[j]≠ρMax)/n
N	4226	1398	1408	n

Confusion Matrix

Training

Actual Churn	Predicted Count Yes	Predicted Count No
Yes	620	497
No	321	2788

Actual Churn	Predicted Rate Yes	Predicted Rate No
Yes	0.555	0.445
No	0.103	0.897

Validation

Actual Churn	Predicted Count Yes	Predicted Count No
Yes	203	174
No	100	921

Actual Churn	Predicted Rate Yes	Predicted Rate No
Yes	0.538	0.462
No	0.098	0.902

Test

Actual Churn	Predicted Count Yes	Predicted Count No
Yes	200	175
No	112	921

Actual Churn	Predicted Rate Yes	Predicted Rate No
Yes	0.533	0.467
No	0.108	0.892

Decision Tree

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.3011	0.2576	0.2238	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.4288	0.3769	0.3330	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.4036	0.4328	0.4499	$\sum -\mathrm{Log}(\rho[j])/n$
RASE	0.3607	0.3753	0.3840	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.2607	0.2731	0.2773	$\sum \lvert y[j]-\rho[j]\rvert/n$
Misclassification Rate	0.1872	0.2024	0.2102	$\sum (\rho[j]\neq\rho\mathrm{Max})/n$
N	4226	1398	1408	n

Confusion Matrix

Training

Actual Churn	Predicted Count No	Predicted Count Yes
No	2785	324
Yes	467	650

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.896	0.104
Yes	0.418	0.582

Validation

Actual Churn	Predicted Count No	Predicted Count Yes
No	905	116
Yes	167	210

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.886	0.114
Yes	0.443	0.557

Test

Actual Churn	Predicted Count No	Predicted Count Yes
No	903	130
Yes	166	209

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.874	0.126
Yes	0.443	0.557

Boosted Tree

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.3769	0.2738	0.2319	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.5153	0.3971	0.3434	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3598	0.4233	0.4452	$\sum -\mathrm{Log}(\rho[j])/n$
RASE	0.3391	0.3711	0.3800	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.2452	0.2707	0.2755	$\sum \lvert y[j]-\rho[j]\rvert/n$
Misclassification Rate	0.1656	0.2031	0.2053	$\sum (\rho[j]\neq\rho\mathrm{Max})/n$
N	4226	1398	1408	n

Confusion Matrix

Training

Actual Churn	Predicted Count No	Predicted Count Yes
No	2899	210
Yes	490	627

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.932	0.068
Yes	0.439	0.561

Validation

Actual Churn	Predicted Count No	Predicted Count Yes
No	928	93
Yes	191	186

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.909	0.091
Yes	0.507	0.493

Test

Actual Churn	Predicted Count No	Predicted Count Yes
No	926	107
Yes	182	193

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.896	0.104
Yes	0.485	0.515

Bootstrap Forest

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.4829	0.2625	0.2235	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.6241	0.3830	0.3326	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.2986	0.4299	0.4500	∑ -Log(ρ[j])/n
RASE	0.3054	0.3752	0.3802	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.2189	0.2706	0.2735	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1306	0.2103	0.2081	∑ (ρ[j]≠ρMax)/n
N	4226	1398	1408	n

Confusion Matrix

Training

Actual Churn	Predicted Count No	Predicted Count Yes
No	2958	151
Yes	401	716

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.951	0.049
Yes	0.359	0.641

Validation

Actual Churn	Predicted Count No	Predicted Count Yes
No	919	102
Yes	192	185

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.900	0.100
Yes	0.509	0.491

Test

Actual Churn	Predicted Count No	Predicted Count Yes
No	923	110
Yes	183	192

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.894	0.106
Yes	0.488	0.512

Neural Network

Model NTanH(3)

Training

Churn

Measures	Value
Generalized RSquare	0.4164688
Entropy RSquare	0.2907597
RASE	0.3644257
Mean Abs Dev	0.2659424
Misclassification Rate	0.1942735
-LogLikelihood	1730.9811
Sum Freq	4226

Confusion Matrix

Actual Churn	Predicted Count No	Predicted Count Yes
No	2789	320
Yes	501	616

Confusion Rates

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.897	0.103
Yes	0.449	0.551

Validation

Churn

Measures	Value
Generalized RSquare	0.3895373
Entropy RSquare	0.2677514
RASE	0.3733085
Mean Abs Dev	0.2740722
Misclassification Rate	0.1974249
-LogLikelihood	596.73719
Sum Freq	1398

Confusion Matrix

Actual Churn	Predicted Count No	Predicted Count Yes
No	920	101
Yes	175	202

Confusion Rates

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.901	0.099
Yes	0.464	0.536

Test

Churn

Measures	Value
Generalized RSquare	0.3355164
Entropy RSquare	0.2257559
RASE	0.3807302
Mean Abs Dev	0.2786729
Misclassification Rate	0.2052557
-LogLikelihood	631.82038
Sum Freq	1408

Confusion Matrix

Actual Churn	Predicted Count No	Predicted Count Yes
No	920	113
Yes	176	199

Confusion Rates

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.891	0.109
Yes	0.469	0.531

Neural Network (2 layers)

Model NTanH(3)NTanH2(3)

Training

Churn

Measures	Value
Generalized RSquare	0.4200724
Entropy RSquare	0.2937548
RASE	0.3635013
Mean Abs Dev	0.2637391
Misclassification Rate	0.1916706
-LogLikelihood	1723.6713
Sum Freq	4226

Confusion Matrix

Actual	Predicted Count	
Churn	No	Yes
No	2780	329
Yes	481	636

Confusion Rates

Actual	Predicted Rate	
Churn	No	Yes
No	0.894	0.106
Yes	0.431	0.569

Validation

Churn

Measures	Value
Generalized RSquare	0.3945645
Entropy RSquare	0.2718166
RASE	0.3725551
Mean Abs Dev	0.2718134
Misclassification Rate	0.2010014
-LogLikelihood	593.42425
Sum Freq	1398

Confusion Matrix

Actual	Predicted Count	
Churn	No	Yes
No	903	118
Yes	163	214

Confusion Rates

Actual	Predicted Rate	
Churn	No	Yes
No	0.884	0.116
Yes	0.432	0.568

Test

Churn

Measures	Value
Generalized RSquare	0.3409144
Entropy RSquare	0.2299177
RASE	0.3794557
Mean Abs Dev	0.2755812
Misclassification Rate	0.2002841
-LogLikelihood	628.42422
Sum Freq	1408

Confusion Matrix

Actual	Predicted Count	
Churn	No	Yes
No	916	117
Yes	165	210

Confusion Rates

Actual	Predicted Rate	
Churn	No	Yes
No	0.887	0.113
Yes	0.440	0.560

Discriminant Analysis

Discriminant Scores

Score Summaries

Source	Count	Number Misclassified	Percent Misclassified	Entropy RSquare	-2LogLikelihood
Training	4226	1007	23.8287	0.13994	4198.16
Validation	1398	341	24.3920	0.13079	
Test	1408	361	25.6392	0.08615	

Training

Actual	Predicted Count	
Churn	No	Yes
No	2354	755
Yes	252	865

Actual	Predicted Rate	
Churn	No	Yes
No	0.757	0.243
Yes	0.226	0.774

Validation

Actual	Predicted Count	
Churn	No	Yes
No	774	247
Yes	94	283

Actual	Predicted Rate	
Churn	No	Yes
No	0.758	0.242
Yes	0.249	0.751

Test

Actual	Predicted Count	
Churn	No	Yes
No	774	259
Yes	102	273

Actual	Predicted Rate	
Churn	No	Yes
No	0.749	0.251
Yes	0.272	0.728

K Nearest Neighbors

Model Selection

Training

K	Count	RSquare	Misclassification Rate	Misclassifications
1	4226	-0.0688	0.27023	1142
2	4226	0.0393	0.26455	1118
3	4226	0.09186	0.22977	971
4	4226	0.1273	0.23497	993
5	4226	0.15765	0.22101	934
6	4226	0.17201	0.22196	938
7	4226	0.18715	0.21486	908
8	4226	0.19949	0.21533	910
9	4226	0.21448	0.20776	878 *
10	4226	0.21845	0.21226	897

Validation

K	Count	RSquare	Misclassification Rate	Misclassifications
1	1398	-0.0395	0.26323	368
2	1398	0.05845	0.27897	390
3	1398	0.10768	0.23677	331
4	1398	0.13315	0.23319	326
5	1398	0.16041	0.22961	321
6	1398	0.17485	0.23176	324
7	1398	0.17491	0.21817	305 *
8	1398	0.1873	0.23605	330
9	1398	0.18973	0.22604	316
10	1398	0.19058	0.22961	321

Test

K	Count	RSquare	Misclassification Rate	Misclassifications
1	1408	-0.0897	0.27912	393
2	1408	0.02344	0.26705	376
3	1408	0.08714	0.23509	331
4	1408	0.11688	0.23580	332
5	1408	0.14033	0.22585	318
6	1408	0.15995	0.23438	330
7	1408	0.15662	0.22443	316
8	1408	0.16946	0.23082	325
9	1408	0.18782	0.21946	309 *
10	1408	0.1899	0.22940	323

Confusion Matrix for Best K=7

Training

Actual Churn	Predicted Count No	Predicted Count Yes
No	2750	359
Yes	549	568

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.885	0.115
Yes	0.491	0.509

Validation

Actual Churn	Predicted Count No	Predicted Count Yes
No	889	132
Yes	173	204

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.871	0.129
Yes	0.459	0.541

Test

Actual Churn	Predicted Count No	Predicted Count Yes
No	893	140
Yes	176	199

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.864	0.136
Yes	0.469	0.531

Naive Bayes

Churn

Training

Count	Misclassification Rate	Misclassifications
4226	0.22646	957

Validation

Count	Misclassification Rate	Misclassifications
1398	0.23391	327

Test

Count	Misclassification Rate	Misclassifications
1408	0.24219	341

Confusion Matrix

Training

Actual Churn	Predicted Count No	Predicted Count Yes
No	2453	656
Yes	301	816

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.789	0.211
Yes	0.269	0.731

Validation

Actual Churn	Predicted Count No	Predicted Count Yes
No	804	217
Yes	110	267

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.787	0.213
Yes	0.292	0.708

Test

Actual Churn	Predicted Count No	Predicted Count Yes
No	810	223
Yes	118	257

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.784	0.216
Yes	0.315	0.685

Averaged Ensemble Model

Model Comparison Validation=Validation

Target Churn missing a predictor for category No

Predictors

Measures of Fit for Churn

Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RASE	Mean Abs Dev	Misclassification Rate	N
Boosted Tree		0.2602	0.3801	0.4313	0.3730	0.2708	0.2060	1398
Boosted Tree		0.2602	0.3801	0.4313	0.3730	0.2708	0.2060	1398
Neural Model NTanH(3)		0.2802	0.4048	0.4196	0.3702	0.2709	0.1974	1398
Boosted Tree		0.2738	0.3971	0.4233	0.3711	0.2707	0.2031	1398
Neural Model NTanH(3)		0.2720	0.3948	0.4244	0.3718	0.2709	0.2003	1398
Neural Model NTanH(3)NTanH2(3)		0.2724	0.3952	0.4242	0.3721	0.2735	0.1974	1398
Partition		-0.000	-0.000	0.583	0.4438	0.3914	0.2697	1398
Bootstrap Forest		0.2591	0.3787	0.4319	0.3762	0.2709	0.2089	1398
Discriminant		0.1308	0.2055	0.5067	0.4097	0.3197	0.2439	1398
Naive Bayes		.	.	.	.	.	0.2339	1398
K Nearest Neighbors		.	.	.	.	.	0.2167	1398
Model Averaged		0.2707	0.3931	0.4252	0.3702	0.2921	0.1974	1398

Appendix G: Model Ranking - Assessment

Group 1 - All Variables

Training (N = 4226)										
Rank	Model Name	Misclassification Rate	Total Accuracy	Accuracy of the "1"	True Positives	True Negatives	False Positives	False Negatives	RASE	Overfitting
1	Bootstrap Forest	10.55%	89.45%	88.97%	766	3014	95	351	0.29	
2	Boosted Tree	15.64%	84.36%	76.89%	652	2913	196	465	0.331	
3	Ensemble Model Average	17.18%	82.82%	69.93%	686	2814	295	431	0.343	
4	Neural Model NTanH(3)	18.86%	81.14%	66.10%	657	2772	337	460	0.363	
5	Partition	19.33%	80.67%	65.66%	629	2780	329	488	0.364	
6	Fit Nominal Logistic	19.36%	80.64%	65.89%	620	2788	321	497	0.364	
7	NTanH(3)NTanH2(3)	19.47%	80.53%	65.34%	626	2777	332	491	0.362	
8	K Nearest Neighbors	22.46%	77.54%	58.06%	605	2672	437	512		
9	Discriminant	23.17%	76.83%	54.32%	867	2380	729	250	0.403	
10	Naïve Bayes	23.43%	76.57%	54.12%	835	2401	708	282		

Validation (N = 1398)										
Rank	Model Name	Misclassification Rate	Total Accuracy	Accuracy of the "1"	True Positives	True Negatives	False Positives	False Negatives	RASE	Overfitting
1	Fit Nominal Logistic	19.60%	80.40%	67.00%	203	921	100	174	0.371	
2	NTanH(3)NTanH2(3)	20.10%	79.90%	64.91%	209	908	113	168	0.37	
3	Neural Model NTanH(3)	20.46%	79.54%	64.26%	205	907	114	172	0.37	
4	K Nearest Neighbors	22.60%	77.40%	58.89%	202	880	141	175		
5	Discriminant	24.54%	75.46%	53.21%	282	773	248	95	0.41	
6	Naïve Bayes	24.68%	75.32%	53.14%	271	782	239	106		
7	Ensemble Model Average	20.03%	79.97%	64.92%	211	907	114	166	0.371	Overfit
8	Boosted Tree	20.60%	79.40%	65.29%	190	920	101	187	0.373	Overfit
9	Bootstrap Forest	20.82%	79.18%	64.43%	192	915	106	185	0.375	Overfit
10	Partition	22.75%	77.25%	58.70%	199	881	140	178	0.383	Overfit

Test (N = 1408)										
Rank	Model Name	Misclassification Rate	Total Accuracy	Accuracy of the "1"	True Positives	True Negatives	False Positives	False Negatives	RASE	Overfitting
1	Neural Model NTanH(3)	20.31%	79.69%	63.05%	215	907	126	160	0.377	
2	Fit Nominal Logistic	20.38%	79.62%	64.10%	200	921	112	175	0.377	
3	NTanH(3)NTanH2(3)	20.88%	79.12%	62.54%	202	912	121	173	0.376	
4	K Nearest Neighbors	24.64%	75.36%	53.85%	196	865	168	179		
5	Naïve Bayes	24.72%	75.28%	52.66%	267	793	240	108		
6	Discriminant	25.92%	74.08%	50.94%	270	773	260	105	0.42	
7	Ensemble Model Average	19.96%	80.04%	64.33%	211	916	117	164	0.378	Overfit
8	Bootstrap Forest	20.67%	79.33%	64.09%	191	926	107	184	0.381	Overfit
9	Boosted Tree	20.88%	79.12%	64.11%	184	930	103	191	0.381	Overfit
10	Partition	21.59%	78.41%	60.47%	205	899	134	170	0.381	Overfit

Group 2 - Significant Variables

Training (N = 4226)										
Rank	Model Name	Misclassification Rate	Total Accuracy	Accuracy of the "1"	True Positives	True Negatives	False Positives	False Negatives	RASE	Overfitting
1	Bootstrap Forest	12.92%	87.08%	82.38%	720	2995	154	397	0.305	[redacted]
2	Boosted Tree	16.56%	83.44%	74.91%	627	2899	210	490	0.331	[redacted]
3	Ensemble Model Average	17.30%	82.70%	71.69%	638	2857	252	479	0.347	[redacted]
4	Partition	18.72%	81.28%	66.74%	650	2785	324	467	0.361	[redacted]
5	Neural Model NTanH(3)	19.24%	80.76%	66.49%	613	2800	309	504	0.363	[redacted]
6	Fit Nominal Logistic	19.36%	80.64%	65.89%	620	2788	321	497	0.364	[redacted]
7	NTanH(3)NTanH2(3)	19.85%	80.15%	64.82%	608	2779	330	509	0.364	[redacted]
8	K Nearest Neighbors	21.49%	78.51%	61.27%	568	2750	359	549	[redacted]	[redacted]
9	Naïve Bayes	22.65%	77.35%	55.43%	816	2453	656	301		[redacted]
10	Discriminant	23.83%	76.17%	53.40%	865	2354	755	252	[redacted]	[redacted]

Validation (N = 1398)										
Rank	Model Name	Misclassification Rate	Total Accuracy	Accuracy of the "1"	True Positives	True Negatives	False Positives	False Negatives	RASE	Overfitting
1	Fit Nominal Logistic	19.60%	80.40%	67.00%	203	921	100	174	0.371	
2	Ensemble Model Average	19.74%	80.26%	67.60%	194	928	93	183	0.37	
3	Neural Model NTanH(3)	19.81%	80.19%	66.45%	202	919	102	175	0.37	
4	NTanH(3)NTanH2(3)	19.81%	80.19%	66.45%	202	919	102	175	[redacted]	
5	Partition	20.24%	79.76%	64.42%	210	905	116	167	0.375	
6	K Nearest Neighbors	21.82%	78.18%	60.71%	204	889	132	173	[redacted]	
7	Naïve Bayes	23.39%	76.61%	55.17%	267	804	217	110	**	
8	Discriminant	24.39%	75.61%	53.40%	283	774	247	94	0.41	
9	Boosted Tree	20.31%	79.69%	66.67%	186	928	93	191	0.373	Overfit
10	Bootstrap Forest	21.39%	78.61%	63.36%	185	914	107	192	0.376	Overfit

Test (N = 1408)										
Rank	Model Name	Misclassification Rate	Total Accuracy	Accuracy of the "1"	True Positives	True Negatives	False Positives	False Negatives	RASE	Overfitting
1	Ensemble Model Average	19.89%	80.11%	65.89%	197	931	102	178	0.377	
2	Fit Nominal Logistic	20.38%	79.62%	64.10%	200	921	112	175	0.377	
3	Neural Model NTanH(3)	20.74%	79.26%	63.34%	197	919	114	178	0.378	
4	NTanH(3)NTanH2(3)	20.95%	79.05%	62.74%	197	916	117	178	[redacted]	
5	Partition	21.02%	78.98%	61.65%	209	903	130	166	[redacted]	
6	K Nearest Neighbors	22.44%	77.56%	58.70%	199	893	140	176	**	
7	Naïve Bayes	24.22%	75.78%	53.54%	257	810	223	118	**	
8	Discriminant	25.64%	74.36%	51.32%	273	774	259	102	0.419	
9	Boosted Tree	20.53%	79.47%	64.33%	193	926	107	182	0.381	Overfit
10	Bootstrap Forest	20.74%	79.26%	63.43%	196	920	113	179	0.273	Overfit