UCC

Coláiste na hOllscoile Corcaigh, Éire
University College Cork, Ireland

MSC DATA SCIENCE AND ANALYTICS

CS6405 DATA MINING

"CHURN PREDICTION"

PRESENTED TO
SEBASTIAN SCHEURER

Submitted by

Vikrant Siwal	**118220030**
Mohammad Azeem Mohammad Rafique Edrisi	**118220338**
Ankit Talwar	**118220956**
Sherif Baruwa	**118220341**
Anurag Kumar Sinha	**118220658**

Submitted on
April 13, 2019

Table of Contents

1. Introduction

Churn prediction is a technique used in business for detecting customers who are likely to unsubscribe from their services. Customer churn happens when customers unsubscribe from plans or packages, which can be stated as the loss of business. The telecommunications industry is one of the most volatile industries in which churn rates are particularly useful because most service consumers have different options for choosing services from different organizations.

Our Mission:

The objective of this project is to analyze the churn data provided and develop a supervised binary classification model for predicting customers churn rate on unseen data.

Dataset: Churn Dataset

We are provided with the churn-train dataset and churn holdout datasets separately, in which we developed an algorithm/model using the churn-train data and predicted the class of churn holdout data with the developed model.

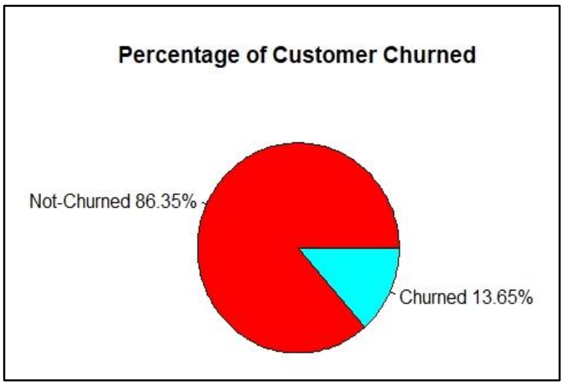

Figure 1: Percentage of customer churned & non-churned

Dimensions:

The dimension of train dataset is 4000×21, where 13.65% (546) of the total observations are classified as “1” (churned) and 86.35% (3454) of the total records are classified as “0” (non-churned). As seen in Figure 1 the data is imbalanced.

Table 1 is the description of the variables provided in the churn-train dataset.

Table 1: Description of the churn-train dataset

Data Field	Data Type	Definition
state	Number	Ranges from 0 to 50
account_length	Integer	Ranges from 1 to 243
area_code	Number	Has: 408 , 415, 510
phone_number	Integer	Unique phone numbers
international_plan	Number	Has: 0 or 1
voice_mail_plan	Number	Has: 0 or 1
number_vmail_messages	Integer	Ranges from 0 to 52
total_day_minutes	Number	Ranges from 0 to 346.8
total_day_calls	Integer	Ranges from 0 to 165
total_day_charge	Number	Ranges from 0 to 58.96
total_eve_minutes	Number	Ranges from 0 to 361.8
total_eve_calls	Integer	Ranges from 0 to 170
total_eve_charge	Number	Ranges from 0 to 30.75
total_night_minutes	Number	Ranges from 0 to 395

total_night_calls	Integer	Ranges from 0 to 175
total_night_charge	Number	Ranges from 0 to 17.77
total_intl_minutes	Number	Ranges from 0 to 20
total_intl_calls	Integer	Ranges from 0 to 20
total_intl_charge	Number	Ranges from 0 to 5.4
number_customer_service_calls	Integer	Ranges from 0 to 9
class	Number	Has: 0 or 1

2. Methods

2.1 Preprocessing:

Any predictive modelling requires that we look at the data before we start modelling. It demands to explore the data, finding the correlation between the predictor variables and choosing the set of important variables.

Data Exploration:

Some of the variables (“area_code”, “international_plan”, “voice_mail_plan”, “class”) were changed to categorical variables which were incorrectly classified as integers. Predictor variable “phone_number” has unique values and acts as a customer identifier. We will make “phone_number” as a row name in our train dataset.

Correlation:

Best-correlated features are identified before proceeding to modelling stage to avoid overfitting hence a correlation plot was used to identify features that are highly correlated as seen in Figure 2.

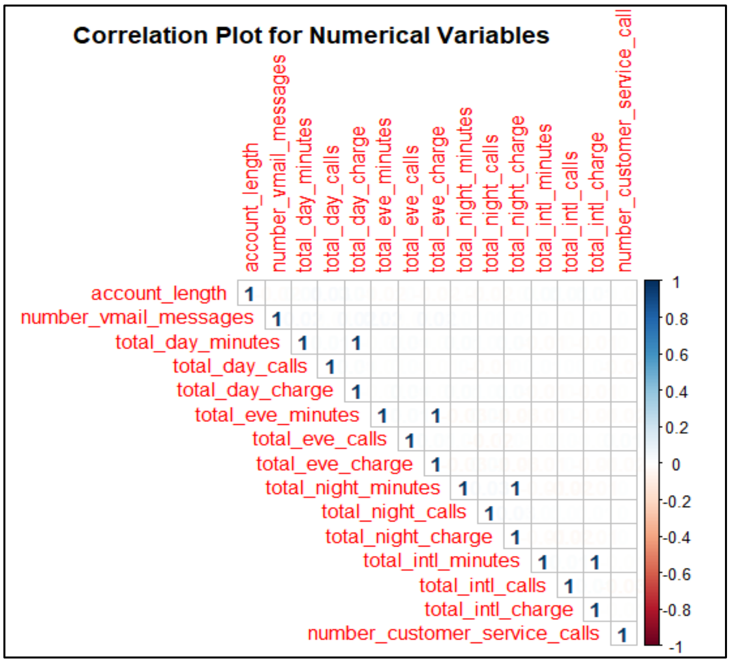

Figure 2: Correlation Plot for Numerical Variable

Based on the results from the correlation plot Figure 2, the following variables "total_day_charge", "total_eve_charge", "total_night_charge", "total_intl_charge" are dropped because they are identified to be highly correlated with “total_day_minutes”, “total_eve_minutes”, “total_night_minutes”, “total_intl_minutes” respectively.

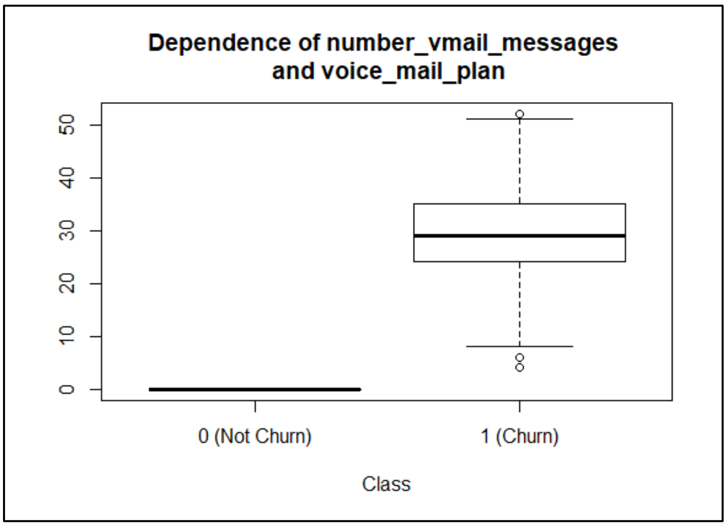

Figure 3: Boxplot of Number of voice mail versus voice mail plan

From Figure 3, it is evident that "number_vmail_messages" are highly dependent on "voice_mail_plan" hence we drop the variable "number_vmail_messages" and make "voice_mail_plan" as a factor predictor in our dataset.

Feature Selection:

Forward and backward stepwise selection approach was used to identify the best combination of features which has a minimum value of Bayesian Information Criterion (BIC). The best combination indicates that a further increase in the number of variables in the model will not make a positive impact in making the model better.

Both approaches provide the same set of variables which gives minimum value of BIC in regards to the stepwise selection procedure.

List of important variables selected by stepwise selection :
"international_plan", "voice_mail_plan", "total_day_minutes", "total_eve_minutes", "total_night_minutes ", "total_intl_minutes", "total_intl_calls", "number_customer_service_calls".

2.2 Modeling Methodology:

Data Splitting:

The Churn-train dataset was split into a 70:30 ratio for training and validation respectively, ensuring each split accounts for the same percentage of churn as in original train dataset. Fitting of the model was done on the training data. The performance of the model was evaluated using the validation data.

Modelling:

The two learning algorithms that were used are **Support Vector Machine and Random Forest** and they will now be discussed more in depth.

Model 1: Support Vector Machine (SVM)

SVM is effective in classification problems and adapts well to imbalanced data.

Note: Kappa score is used for model selection criteria because the data is imbalanced and when considering accuracy for choosing the best values of parameters it can lead to selecting a bias model.

Tuning SVM:

A 10 fold repeated cross-validation with three repeats was set up to limit and reduce overfitting on the training model. Before fitting the model, three different kernels were evaluated to find out which kernel gives the highest value of Kappa. The kernel with the best Kappa value was then used to further tune the model.

Best Kernel Search:

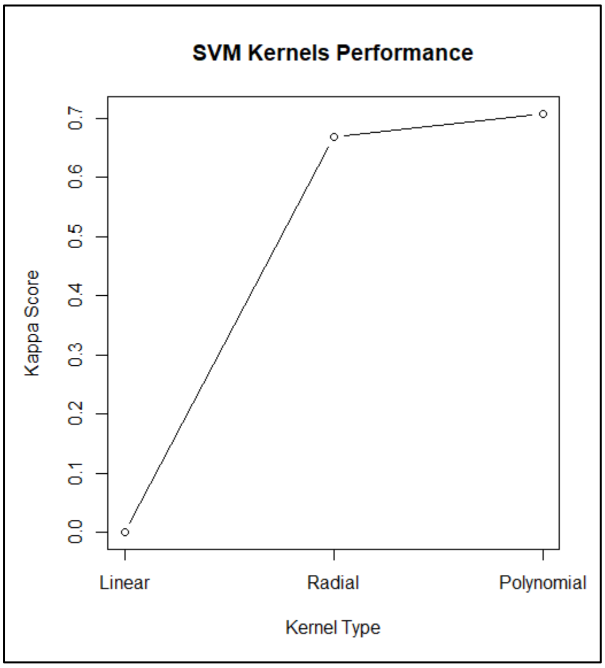

Figure 4: Kernel Type Vs Kappa

From Figure 4, the polynomial kernel was identified with the best Kappa value. Hence proceeding to fit the model using the polynomial kernel.

Note: In the default polynomial model, the final parameters used are degree = 3, scale = 0.1 and C (cost)= 1. Therefore, these parameters were tuned using within a feasible range to identify the best tuning parameter.

Tuning SVM-Polynomial Model:

The parameters were set within the ranges listed below and the best value for each parameter was identified.

Degree: range(2 – 5) Scale: range(0.08 – 0.15) C(**cost**): range(0.4 – 1.4)

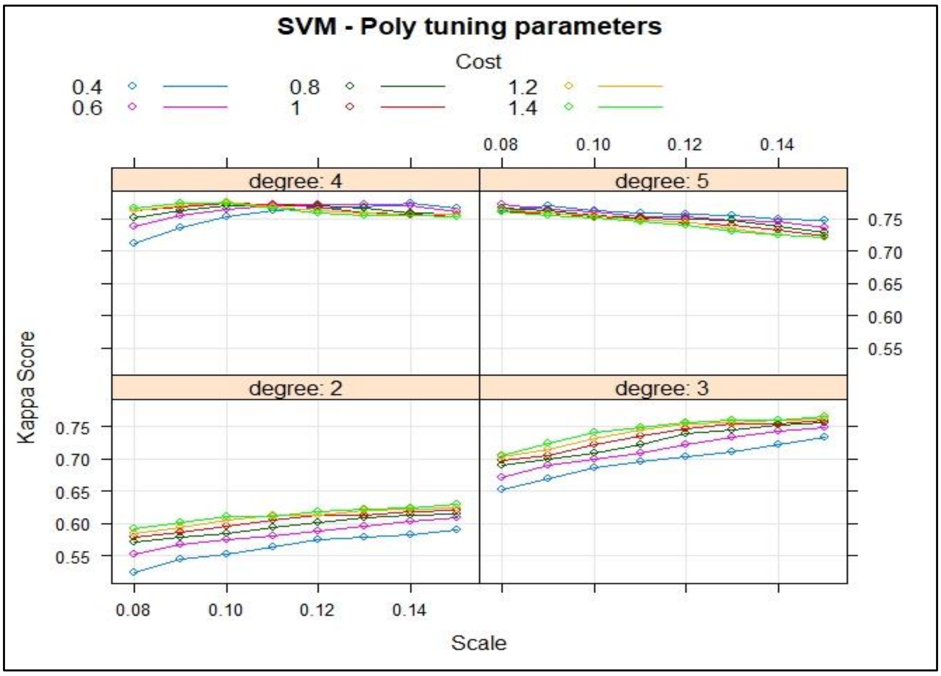

Figure 5: Kappa score for different tuning parameters for SVM-Polynomial model of training dataset

Figure 5 shows that after tuning the polynomial parameters, the values for which Kappa is maximum are **Degree = 4, Scale = 0.1 and Cost = 1.2**

Next, we checked the Kappa and accuracy of the fitted model on validation dataset:

Table 2: Kappa and accuracy for SVM-Polynomial model of validation dataset

	Kappa	Accuracy
Score (validation dataset)	0.7491	0.9458

From Table 2, the value of Kappa and accuracy of validation dataset agree with the 95% Confidence Interval (CI) value of the fitted (training) model as seen in Table 3, **therefore the model is a good fit.**

Table 3: 95% CI for SVM-Polynomial model of training dataset

	Lower Limit	Score (Validation dataset)	Upper Limit
Kappa	0.7085018	0.7491	0.7896021
Accuracy	0.9359431	0.9458	0.9535298

Model 2: Random Forest (RF)

Random Forest is a robust and easy to interpret algorithm for classification problem. For imbalanced data, RF is one of the best algorithms to work on. It has some useful features and techniques which take into account the weight of each class. The first technique bootstrapping is the method of combining weak learner trees inside RF with high learner trees, which average out the variance and bias. It also takes into consideration the weight of each class.

Similar to SVM, a 10 fold repeated cross-validation with 3 repeats was set up to limit and reduce overfitting on the training model.

Tuning Random Forest:

Model is tuned with different combinations of mtry and ntree parameters to find out the best tuning parameters which give the highest value of Kappa score.

Considering a "too" small value or "too" large value of mtry and ntree can lead to bias or an overfit model. Therefore identification of the best combination of two parameters which can provide the tradeoff between bias and variance was implemented.

Identifying the best value of mtry and ntree:

At a time, keeping one of the parameters at a constant value so as to identify the influence of the other parameter on the kappa score.

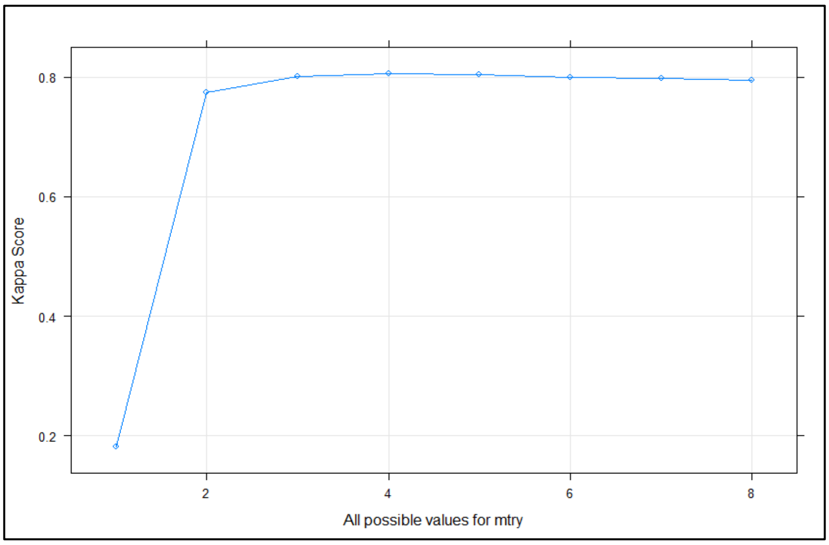

Figure 6: Kappa Score vs Different values of mtry

From Figure 6, it is observed that Kappa is highest when the value of mtry is 4, and ntree is a default value.

Next, Keeping mtry = 4 and trying range of ntree = c(500,1000,1500,2000,2500)

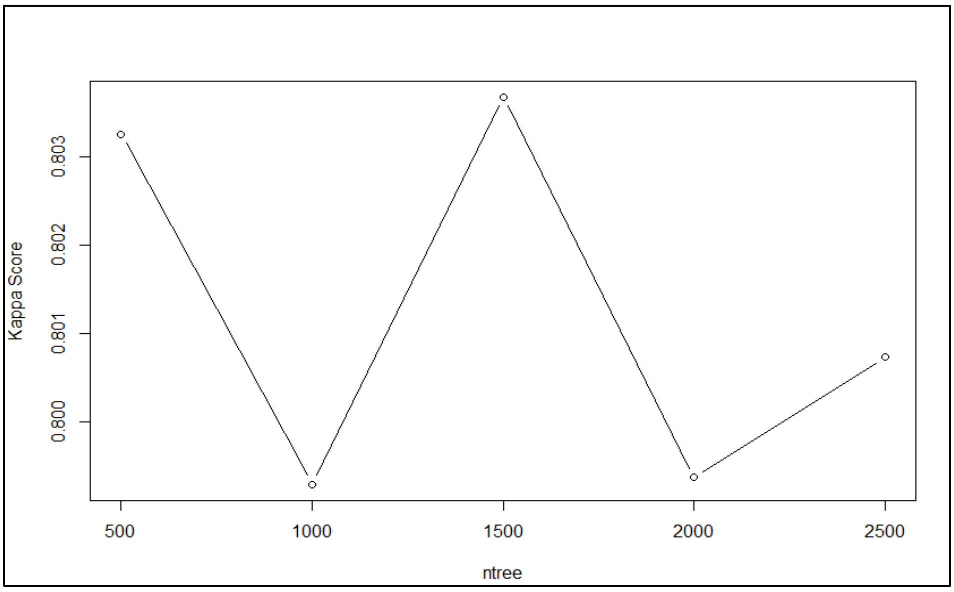

Figure 7: Kappa Score vs Different values of ntree

Note: Considering the dimension of training dataset, a feasible range of ntree was chosen which can possibly give a trade-off.

From Figure 7, it is observed that Kappa score is highest at ntree =1500, keeping mtry as 4.

The model was then tuned on all possible combinations of mtry and ntree.

The possible range for mtry and ntree:
mtry= c(1,2,3,4,5,6,7,8)
ntree= c(500,1000,1500,2000,2500)

Kappa Score table:

Table 4: Kappa score for different values of mtry and ntree, maximum value of Kappa is highlighted

ntree/mtry	500	1000	1500	2000	2500
1	0.1652	0.1776	0.1773	0.1847	0.1798
2	0.7765	0.7703	0.7726	0.7769	0.7736
3	0.8049	0.7966	0.8003	0.7964	0.799
4	0.8025	0.8040	0.8034	0.7948	0.795
5	0.7966	0.8044	0.7998	0.7976	0.8037
6	0.7885	0.793	0.7995	0.796	0.795
7	0.7929	0.7918	0.7996	0.7952	0.798
8	0.7905	0.7926	0.7921	0.7965	0.7978

Table 4 shows that for all feasible possible combinations of mtry and ntree, Kappa is highest at mtry=3 and ntree=500.

Accuracy table:

As per Table 5, mtry=3 and ntree=500 which were identified as the best tuning parameters were used to fit a random forest model on the training dataset.

Table 5: Accuracy for different values of mtry and ntree, score for mtry=3 and ntree=500 is highlighted, as kappa is maximum for this combination

ntree/mtry	500	1000	1500	2000	2500
1	0.8776	0.8786	0.8787	0.8793	0.879
2	0.9532	0.952	0.9522	0.9529	0.9524
3	0.9576	0.9561	0.9569	0.956	0.9567
4	0.9573	0.9575	0.9574	0.9556	0.9557
5	0.956	0.9578	0.9568	0.956	0.9575
6	0.9543	0.9553	0.9566	0.9557	0.9557
7	0.9553	0.9553	0.9564	0.9557	0.9565
8	0.9545	0.955	0.955	0.9557	0.9562

Table 6 shows the Kappa and accuracy of the fitted model on validation dataset:

Table 6: Kappa and accuracy for the RF model of validation dataset

	Kappa	Accuracy
Score (validation dataset)	0.8098	0.9591

From Table 6, the value of Kappa and accuracy agree with the 95% CI value of the fitted(training) model as seen in Table 7, therefore the model is a good fit.

Table 7: 95% CI for RF model of training dataset

	Lower Limit	Score (Validation dataset)	Upper Limit
Kappa	0.7564732	0.8098	0.8377107
Accuracy	0.9481727	0.9591	0.9642857

3. Results and Discussion

The estimated predictive test performance of the fitted models:

RF Model:

Table 8: Estimated value of Kappa and Accuracy for RF Model

	Lower Limit	Upper Limit
Kappa	0.7564732	0.8377107
Accuracy	0.9481727	0.9642857

From Table 8, there is 95% certainty that the value of Kappa and accuracy of the RF model on test-dataset would be between above intervals (Lower limit, Upper limit).

SVM Model:

Table 9: Estimated value of Kappa and Accuracy for SVM Model

	Lower Limit	Upper Limit
Kappa	0.7085018	0.7896021
Accuracy	0.9359431	0.9535298

From Table 9, there is a 95% certainty that the value of Kappa and accuracy of SVM model on the test dataset would be between above intervals (Lower limit, Upper limit).

Model performance criterion used for model comparison are:

- Area Under the Curve (AUC).
- Sensitivity.
- Kappa score.
- F1 Score

RF vs SVM on validation dataset:

Table 10: Comparison table for RF and SVM model's metrics

	Kappa	Sensitivity	F1	ROCAUC
RF	0.8098	0.7485	0.8328	0.928
SVM	0.7491	0.70552	0.77966	0.914

From Table 10, the RF model outperforms the SVM model on the validation dataset for all performance criteria stated in the table. Hence, the RF model was used to make predictions on the holdout dataset.

There is a 95% certainty that the performance of the model on holdout dataset will fall within intervals stated in Table 8.

Table 11: Confusion matrix of RF and SVM model of validation dataset

Note: 'Positive' Class is Churn

SVM Model		
	Reference	
Prediction	NoChurn	Churn
NoChurn	1019	48
Churn	17	115

RF Model		
	Reference	
Prediction	NoChurn	Churn
NoChurn	1028	41
Churn	8	122

Table 11, shows the confusion matrix of SVM and RF model of validation dataset, RF model's prediction of churn customers is more accurate than SVM.

As per Figure 8, AUC of Random forest is better than AUC of SVM.

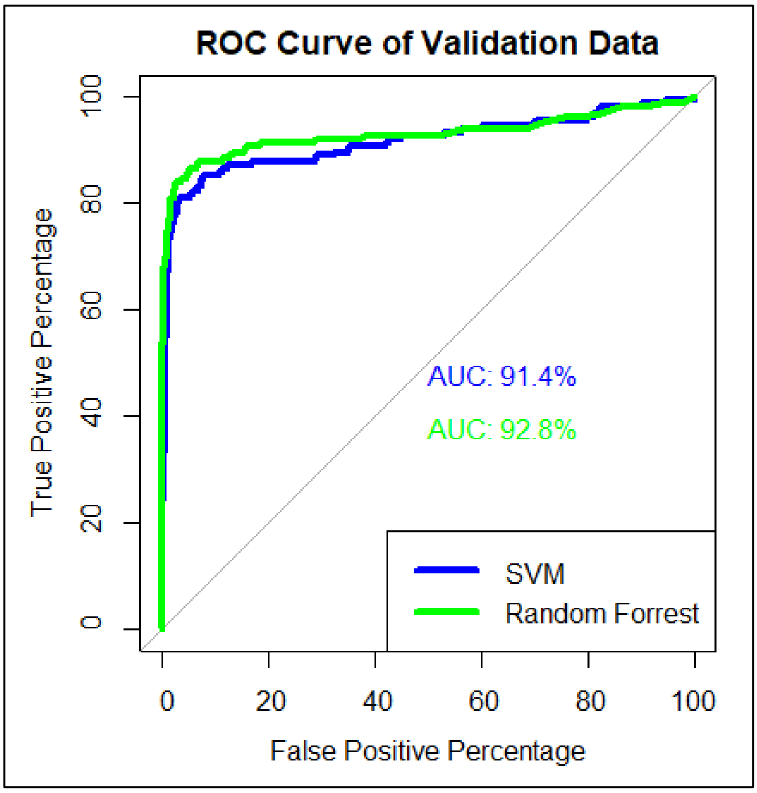

Figure 8: SVM vs RF- ROC curve of validation dataset

From this modelling technique:

- The telecommunication company could predict which customers are more likely to churn and the company can then focus on these customers.
- Customers with high total day call charges are more prone to churn the company. To stay ahead, the company must focus more on daytime call rates and should beware of other competitors call rates.
- If the company wants to plan a promotion campaign, there are some costs associated with it. Therefore, promotion can either be to retain in house customers or generate new prospect customers for profitability. Promotional campaigns are highly effective but at the same time is very expensive for the company. It is highly important that the promotion is presented only to customers that are more likely to churn, as it will have a two-way impact on the company's performance. First, saving the campaign cost and second, retaining the in-house business.
- There is a statistically significant difference between the two groups for the number of customer service calls. It can be interpreted that customers receiving more number of service calls are more likely to churn which can be due to either of the following two factors or both: customers are unhappy with the promotion being done by customer service operators, or customer queries not getting resolved in the desired turnaround time.

Limitations:

- The class variable is imbalanced as the probability of customers churning is very low.
- The data available to train the model is small.

4. Conclusion

In this project, the analyzing of churn customers using two statistical models (SVM and RF) was done. The efficiency of this model was evaluated and compared on the basis of ROC-AUC(Area Under The Curve - Receiver Operating Characteristics), sensitivity, Kappa and F1 score. Overall accuracy is not a good performance metric when the data is imbalanced.

The performance of different kernel functions in SVM was investigated, and the result showed that the polynomial kernel function got the highest Kappa score of 74.9%. The RF model was tuned with different combinations of mtry and ntree parameters to find the best tuning parameters which gave the highest value of the Kappa score of 80.9%.

The RF model outperforms the SVM model on all performance criteria. Whereby, the churn class contains the fewer number of samples as compared to the non-churn class and this makes it difficult for the models to accurately identify the minority class. Although both models might have achieved high overall accuracy, that does not give a clear picture of how well the model did for the minority class. Finally, both classes are considered equally important while deciding on the best model.

The effective churn prediction model supports companies to know which customers are about to churn. Successful churn model must also include effective retention actions. Mobile service providers need to develop attractive retention programs to satisfy these customers.

Prediction summary of holdout dataset:

RF was used to predict the churn holdout dataset. The summary of the prediction can be seen in Table 12.

Table 12: Prediction summary of holdout dataset

0-Class (No Churn)	1-Class (Churn)
873	127

5. Future Scope

- More data on churn customers should be collected to make prediction model more robust, which can be useful in making a more precise prediction about churn customers.
- More explanatory variable like customer's profile(age, income, gender), should be added which can help in making a precise prediction about churn customers.
- Customers predicted to be churned should be given more importance so as to retain the company's business and evaluate the effectiveness of a marketing scheme. Also, these customers' data should be checked periodically to enhance business strategies.

6. References

1. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning, second edition: Data mining, inference, and prediction. New York: Springer.
2. L. Wang, Support Vector Machines: Theory and Applications, Studies in Fuzziness and Soft Computing, Volume 177, Springer, 2005.
3. C. Kang and S. Pei-ji, "Customer churn prediction based on svm-rfe," in Businessand Information Management, 2008. ISBIM '08. International Seminar on, vol. 1,pp. 306–309, Dec 2008.
4. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). An introduction to statistical learning with applications in R. New York: Springer.