Customer Churn Prediction

Ilyas Aaqaoui, S3502317

University of Groningen
Faculty of Economics and Business

Abstract

Customer retention has gotten great attention by companies in terms of resources allocation. It is a justified investment due to the induced high cost of lost customers. One way orgnizations can act proactively to prevent such incidents and increase the retention is predicting customer churn ahead. In this paper, we first examine existing litterature on methods used in predicting and factors influencing customer churn. Then We compare The following three methods: K-Nearest Neighbours, Support Vector Machines and Random Forest. Results indicate a clear outperformance for Random Forest due to its robustness and speed of computation.

1 Introduction

Companies, at various development stages and serving all kinds of markets, recognize the necessity to stay close to customers and keep a positive relationship on a long period of time. In their effort to be sustainable in a constantly changing environment, Data is often used all Along the customer journey - Acquisition, development and retention - to keep track of the consumers' behaviour [25]. This drives companies to make budget allocation decisions for each part of the journey. Evidence suggests that customer acquisition costs for companies are much higher than retention costs [10]. Besides, long term customers prove to be more profitable [13]. Companies therefore put more effort and focus on valuing existing customers than on acquiring new ones. However, This marketing effort put by firms into retention strategies is an investment that should engender high Return On Investment and keep control on costs. Still, Stages in customer journey are not disconnected. In fact channels through which customers were acquired (e.g., website) also affect customer loyalty [27].

Retention rates, components of long-term customer value and ability to calculate the long-term customer value are the main determinants of a company's profitability [18]. In fact, Customer Lifetime Value (CLV) is a pertinent indicator for overall firm value. And increasing the CLV relies heavily on the customer churn [21], Which is how likely a customer is to switch, unsubscribe or stop purchasing a service or product.

All in all, Firms should go beyond a simple binary classification when it comes to churn, and eventually include unsupervised machine learning (clustering) as a preprocessing step to get insights on different categories of customers before running predictions. Firms should also account for the distribution of churn classes, which are often not balanced and thus challenge the predictive power of most algorithms [28].

The focus of this paper will be on customer churn prediction models. Particularly, comparing the three methods: Random Forest, Support Vector Machines and K-Nearest Neighbours. Logistic regression will be the only statistical method used firstly as a mean to understand independent variables' influence on churn for the dataset used in this paper and secondly in comparison to the other methods.

2 Review of Churn Prediction Models and implementation Methods

As part of customer retention strategies, customer churn can be caused or influenced by various factors. Rust & Zahovik (1993)[20] were among the first to analytically illustrate how satisfaction impacts financial returns in bank retailing. Specifically, Satisfaction can be viewed in different dimensions: Service quality, service features, service problems and service recovery [17]. A positive effect of affective commitment and loyalty programs was also demonstrated on customer retention [26]. One of the comprehensive models was proposed by Schweider, Fader & Bradlow (2008) [22], in addition to promotional effects, subscriber heterogeneity, it also included time components such as duration dependence and calendar time effects.
From a behavioural Aspect, some behavioural manifestation of customer engagement, such as Word Of Mouth (WOM), co-creation and complaining behaviour have also an impact on customer retention [25].

In terms of implementation, many experiments of different algorithms were applied in the case of customer churn prediction. The most used algorithms are Logistic regression and Decision trees [13]. Other algorithms such as Neural Networks and Random Forest were also used but did not gain much exposure due to models complexity and low accuracy sometimes [23]. However, most of experiments done using these models date back 10 years ago. In the recent years, computational power has significantly increased and is now more and more accessible and affordable, particularly for companies.

According to H.Risselada et Al (2010), Neural Networks were tried the first time for churn prediction in 1997. In the context of assessing targeting issues of mailing, it was found that the neural networks predictive power - after testing dozens of configurations - was slightly different from Logistic regression, and that it was not encouraging to use it due to the complexity of setting up and configuring a neural

network [12]. However, this is no longer a problem since most configurations are now automated. Similarly, in an experiment in the Consumer Goods industry, no significant difference was found between Neural Networks, Logistic Regression and Random Forest [4]. Neural Networks are often accused of instability and overfitting. However, with the right parameters tuning, Neural Networks' performance significantly increased in churn prediction [11]. They can also be useful in the case of identifying the influencing factors on customer churn[1].

Eventhough Logistic Regression and Decision trees remain popular for predicting customer churn, the focus of this paper will be on Random Forest - an ensemble of decision trees-, Support Vector Machine and k-Nearest Neighbours.

- Random Forest is a combination of Decision trees (Which explains the term "forest"), where it is taking random decision trees and combining them together. Random Forest was found to outperform Support Vector Machine in customer churn prediction [5][7]. Similarly, it was showed that Random Forest lands better results than linear or logistic regression [16]. Improved Balanced Random Forest - a variation of RF taking data unbalancing into consideration - also outperforms LR [29].
  These results favouring Random forest are expected. It is leveraging the already favourable results of decision trees by combining multiple trees, in the order of hundreds, in one algorithm.

- Support Vector machine (SVM) is a discriminant based algorithm, it is not based on maximum likelihood estimation, but on defining a hyperplane – in multidimensional case, aka, multiple predictors - that separates the two output classes, this hyperplane is defined by vectors called support vectors, which are input points that are close to or belong to the hyperplane. SVM generally outperforms Logistic Regression when the problem at hand is not linearly separable. And is also suitable when the number of variables is large compared to the number of observations. There is enough evidence on the success of SVM over other methods in the customer churn prediction context, and in the particular case of scrupulous selection of parameters, SVM performed better than Logistic Regression. But could not perform better that Random Forest [29]. Even if Logistic Regression and Decision Trees gained popularity for their relative simplicity, SVM can in some cases outperform Logistic Regression [7].

- K-Nearest Neighbours is a method based on distance calculation. Observations that are close together, are considered part of the same class. KNNN is a non-parametric method, it does not take into account the distribution of data. The K in Knn stands for the number of neighbours to be counted to define a class. Knn was first used in churn prediction as a hybrid combination with Logistic Regression (LR), it has improved the performance of LR in the case of non-linearity between factors and output [30]. Another Hybrid model was also used in combination with Artificial Neural Networks, decision trees and support vector machines, however, when compared individually, KNN underperformed compared to other models [15].

When it comes to prediction accuracy, two issues have been documented. First, staying power of customer churn prediction models is low [19]. This finding is understandable. Prediction Models should be continuously updated and adapted to changing patterns in data. This is the essence of machine learning. Papers coming from marketing literature stream have a data mining modeling approach, and rarely mention machine learning. Second issue deals with measurement of robustness and power of a model. This issue has two levels. First, most datasets accessible have an unbalanced distribution of churn/non churn [8], sometimes amounts to 98%. This means that an automatic prediction of non-churn will land a 98% accuracy, and therefore the use of a relevant performance measure such as Area under Curve is more suitable. The second level is the cost of misclassification: for missclassified cases, the cost of false positives is much higher than the cost of false negatives [6]. In other terms, wrongly predicting a non churner while in fact he/she is does more harm than wrongly predicting a churner. Customer prediction models need to take this into consideration. By doing so, Bahnsen, Aouada and Ottersten [3] developed a framework to account for cost and were able to save up to 26.4% in costs.

To improve customer churn prediction, the above mentioned methods can be enhanced by methods such as Boosting, Bragging and Ensembling. Recently, The Extreme Gradient Boosting (XGBoost) has been widely used in competitions submissions, and it has also proven to yield better results compared to KNN and Neural Networks [9].

3 Methodology

Most churn experiments done in the past tend to dismiss or overlook relatively simple methods as K-nearest neighbours and Support Vector Machines. Theses algorithms have the potential for very positive results in the context of churn prediction. One reason is the existence of categorical (ordinal or nominal) variables. Provided that the data is scaled, KNN is more likely to produce highly accurate reseults in this case. Another reason is the high number of variables, often generated by the exshaustive data collection by companies, that makes SVM better suited due to its robust hyperplane separation method.

In order to assess KNN and SVM, Random Forest will be used as a baseline in the experiment for comparison.

3.1 Data Preparation

3.1.1 Data Description

The dataset used was retrieved from IBM Watson Analytics datasets, it contains 20 independent variable for 7043 unique customer of a telecommunications company in the United States. the independent variables fall roughly into the following three categories:

1. Demographics
    - *customerID*
    - *gender*: female, male.
    - *SeniorCitizen*: Whether the customer is a senior citizen or not (1, 0).
    - *Partner*: Whether the customer has a partner or not (Yes, No).
    - *Dependents*: Whether the customer has dependents or not (Yes, No).
2. Service Subscription Information
    - *tenure*: Number of months the customer has stayed with the company.
    - *PhoneService*: Whether the customer has a phone service or not (Yes, No).
    - *MultipleLines*: Whether the customer has multiple lines or not (Yes, No, No phone service).
    - *InternetService*: Customer's internet service provider (DSL, Fiber optic, No).
    - *OnlineSecurity*: Whether the customer has online security or not (Yes, No, No internet service).
    - *OnlineBackup*: Whether the customer has online backup or not (Yes, No, No internet service).
    - *DeviceProtection*: Whether the customer has device protection or not (Yes, No, No internet service).
    - *TechSupport*: Whether the customer has tech support or not (Yes, No, No internet service).
    - *streamingTV*: Whether the customer has streaming TV or not (Yes, No, No internet service).
    - *streamingMovies*: Whether the customer has streaming movies or not (Yes, No, No internet service).
3. Paiment Informations

- *Contract*: The contract term of the customer (Month-to-month, One year, Two year).
- *PaperlessBilling*: Whether the customer has paperless billing or not (Yes, No))
- *PaymentMethod*: The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)).
- *MonthlyCharges*: A numeric variable describing the amount charged to the customer monthly.
- *TotalCharges*: A numeric variable describing the total amount charged to the customer.

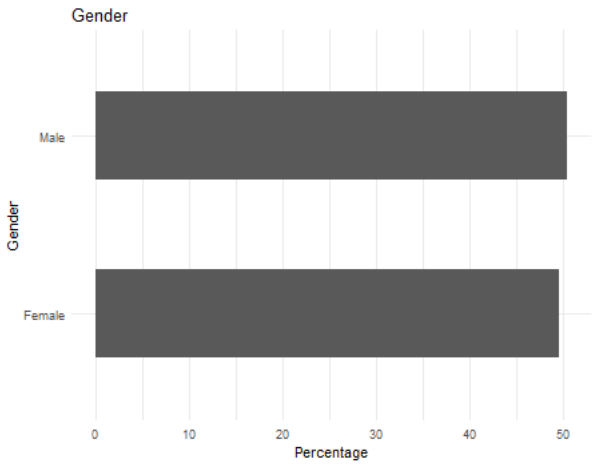
The dependent variable is "Churn", with two categories: 'Yes' and 'NO' describing weather the customer has churned or not.

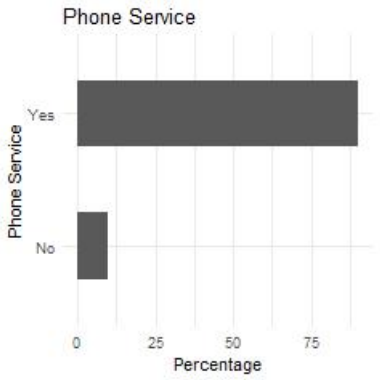
$$\Pr(Churn) = f(Demographics, ServiceSubscriptionDetails, PaymentEase)$$

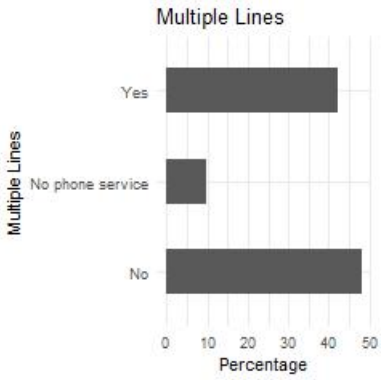
3.1.2 Data Preprocessing

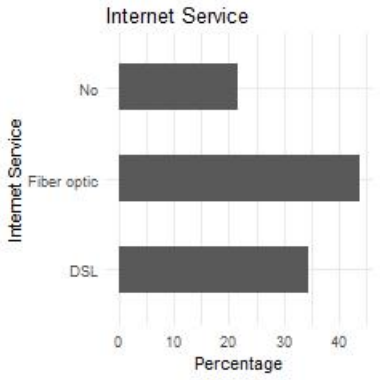
The dataset contains only eleven missing values in one variable: Total Charges. The rows containing these missing values will subsequently be removed.

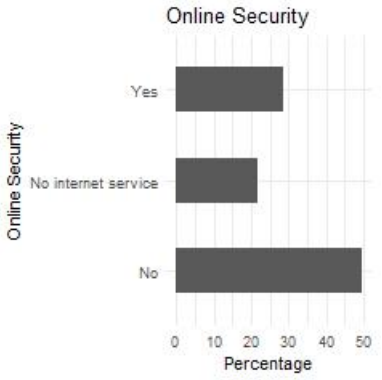
The following barplots illustrate the distribution of each category within independent variables:

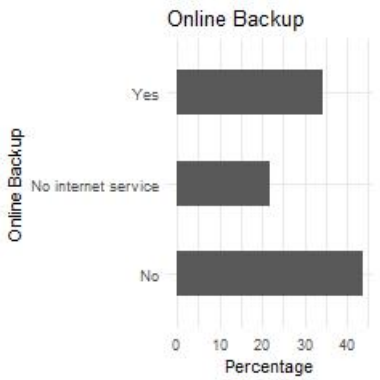


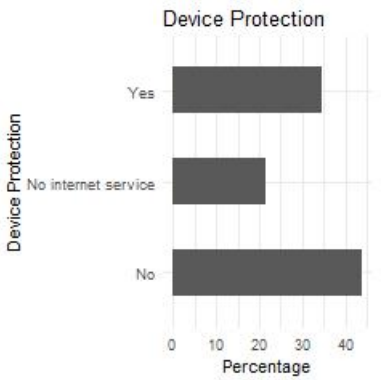

Phone Service
Phone Service
Yes
No
0
25
50
75
Percentage

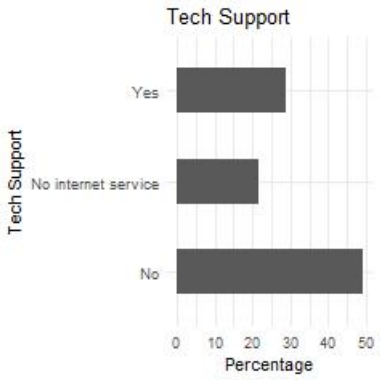

Multiple Lines
Multiple Lines
Yes
No phone service
No
0
10
20
30
40
50
Percentage

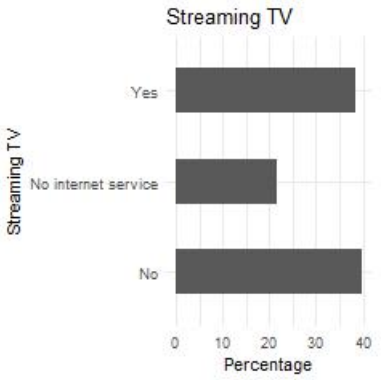

Internet Service
Internet Service
No
Fiber optic
DSL
0
10
20
30
40
Percentage

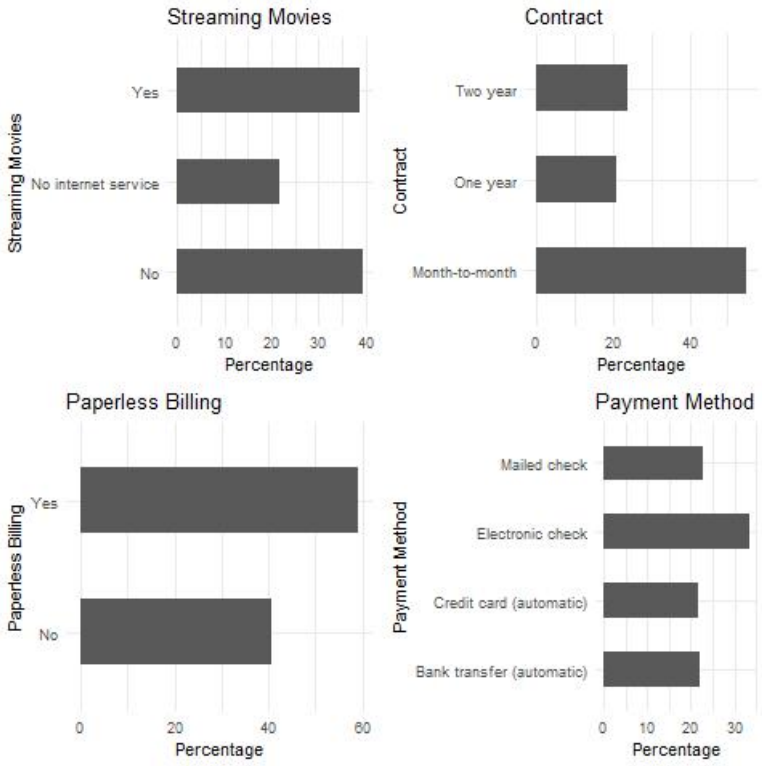

Online Security
Online Security
Yes
No internet service
No
0
10
20
30
40
50
Percentage


Online Backup
Online Backup
Yes
No internet service
No
0
10
20
30
40
Percentage


Device Protection
Device Protection
Yes
No internet service
No
0
10
20
30
40
Percentage


Tech Support
Tech Support
Yes
No internet service
No
0
10
20
30
40
50
Percentage


Streaming TV
Streaming TV
Yes
No internet service
No
0
10
20
30
40
Percentage



- Fitting a Logistic regression model:

To better understand the influence of predictors on the churn outcome, We will fit a logistic regression model on the dataset at hand. we find that *Tenure* (How long have customers been subscribed to the service) is the most relevant independent variable having a positive influence on Churn with a high significance level. Followed by The *Contract* type and *Paperless Billing*. Logistic regression's predictive power will also be compared to the three other methods.

```
Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            -1.245727   1.012736  -1.230  0.21867
genderMale                              0.005215   0.078054   0.067  0.94673
SeniorCitizenYes                        0.268218   0.100590   2.666  0.00767 **
PartnerYes                             -0.031401   0.094273  -0.333  0.73907
DependentsYes                          -0.126373   0.108965  -1.160  0.24615
PhoneServiceYes                         0.039976   0.796526   0.050  0.95997
MultipleLinesYes                        0.395405   0.217701   1.816  0.06933 .
InternetServiceFiber optic              1.568899   0.977972   1.604  0.10866
InternetServiceNo                      -1.445689   0.992016  -1.457  0.14503
OnlineSecurityYes                      -0.215432   0.218837  -0.984  0.32490
OnlineBackupYes                        -0.061062   0.213932  -0.285  0.77532
DeviceProtectionYes                     0.018193   0.215515   0.084  0.93272
TechSupportYes                         -0.210097   0.220270  -0.954  0.34018
StreamingTVYes                          0.575121   0.398781   1.442  0.14925
StreamingMoviesYes                      0.580613   0.400970   1.448  0.14761
ContractOne year                       -0.884153   0.132181  -6.689 2.25e-11 ***
ContractTwo year                       -1.625461   0.212907  -7.635 2.27e-14 ***
PaperlessBillingYes                     0.274043   0.089526   3.061  0.00221 **
PaymentMethodCredit card (automatic)   -0.222822   0.136348  -1.634  0.10221
PaymentMethodElectronic check           0.175685   0.112318   1.564  0.11778
PaymentMethodMailed check              -0.108172   0.136507  -0.792  0.42811
MonthlyCharges                         -0.025536   0.038938  -0.656  0.51195
tenure_group0-12 Month                  1.788663   0.204671   8.739  < 2e-16 ***
tenure_group12-24 Month                 0.888372   0.200765   4.425 9.65e-06 ***
tenure_group24-48 Month                 0.434733   0.184778   2.353  0.01864 *
tenure_group48-60 Month                 0.204885   0.198250   1.033  0.30138
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 1: Variables coefficients

These finding are also confirmed through a plot of decision trees illustrating the importance of these independent variables:

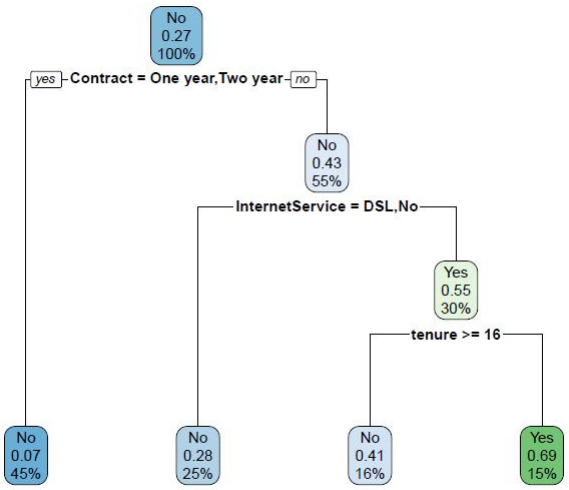


Figure 2: Variable importance

Seventeen categorical variable are encoded into dummy variables, this results in forty three dummy variable. The three remaining numerical variables were scaled to mean 0 and a standard deviation of value one. Scaling is necessary for SVM and Knn as these methods are distance based, an unscaled dataset will influence the weight of variables with large numbers.

20% of the dataset will be randomly taken as test data, and will only be used at the end to compare the algorithms. the remaining 80% will be divided into 90% for training and 10% for cross-validation to fine tune each algorithm individually (e.g., number of trees for random forest, number of neighbours for K-nn). Furthermore, each base-learner is trained over a slightly different dataset, sampled randomly from the training set with replacement. In Ensemble Learning, Where results of different models are combined - average or majority vote - to improve prediction, bagging models can be used in two ways. First, By randomly selecting a subset of the data for each base learner and training the model on that subset, then averaging results. Second, Bagging can also be used individually for each base learner, where the random selection of a subset is repeated for one algorithm separately. results are then averaged to keep the balance between variance and bias. Because the purpose of this paper is to compare different methods, Ensemble learning will not be used.

3.2 Accuracy Measures and Evaluation

One way to evaluate the performance of the three methods is to determine the predictive accuracy. The predictive accuracy could be used as a statistical measure of how many observations in the testing set are correctly classified. This would amount to counting the true positives and true negatives compared to the total, including the false positives and false negatives. Nevertheless, the predictive accuracy must be approached with caution, because the it does not reflect the underlying class distribution of the dependent variable.

Conequently, we consider another measure for overall model performance which is based on the ROC-Curve (Receiver Operating Characteristic) [14]. In the ROC graph the True Positive Rate (Number of True Positives divided by the number of Positive cases) is plotted against the False Positive Rate (number of False Positives divided by the number of Negative cases) for a threshold which varies from 0 to 1. A good model will yield a high true positive rate whereas the false positive rate will remain small. Thus, for a good model the ROC-curve will rise steeply close to the origin, and flatten at a value near the maximum of 1. On the other hand, the ROC-curve for a poor model will lie adjacent to the

diagonal where the true positive rate equals the false positive rate, this implies that the model makes random predictions. Subsequently, the Area Under the ROC-Curve (AUC) is a more accurate measure for the methods' performance [24]. Good models achieve an AUC approximating 1, while poor models will have an AUC near 0.5.

4 Results

As mentionned, 20% of the dataset will be withheld until the end to make the comparison betwen the three methods, and to ensure that our validation accuracy is consistent with data we have never seen during the iterative process. The remaining 80% is used individually in each method to find the best parameters. In the following part we will discuss performances of base learners individually, then we will use the best models found in the final comparison on the same test set.

4.1 Base learners

1. Statistical method: Logistic Regression

   Logistic Regression was fast to converge and returned unexpectedly good results, with an AUC of 0.846. creation of dummy variables and scaling of numerical predictors increased the speed of convergence, the fitting and prediction on the test data.

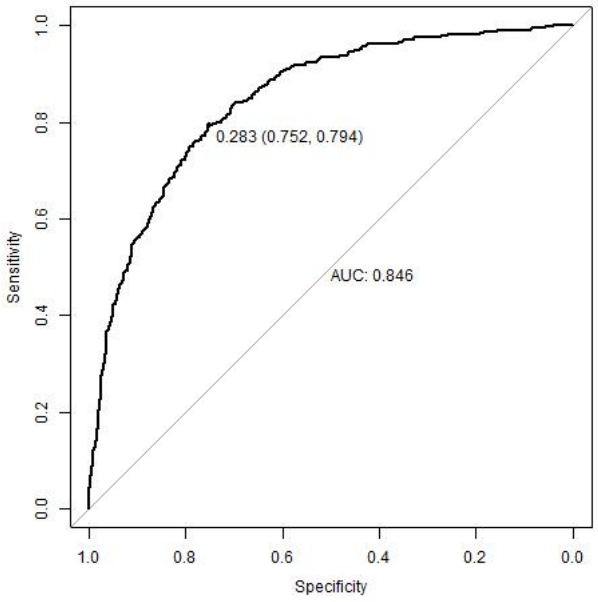


Figure 3: Area Under the Curve for Logistic Regression

2. KNN

   The distance measure and the number of neighbours (k) are the two parameters that we can experiment with. For the present dataset, Euclidian distance is the best fit, and after experimenting with different values of *k*, k=30 yields the best result with an Area Under Curve (AUC) of 0.739.

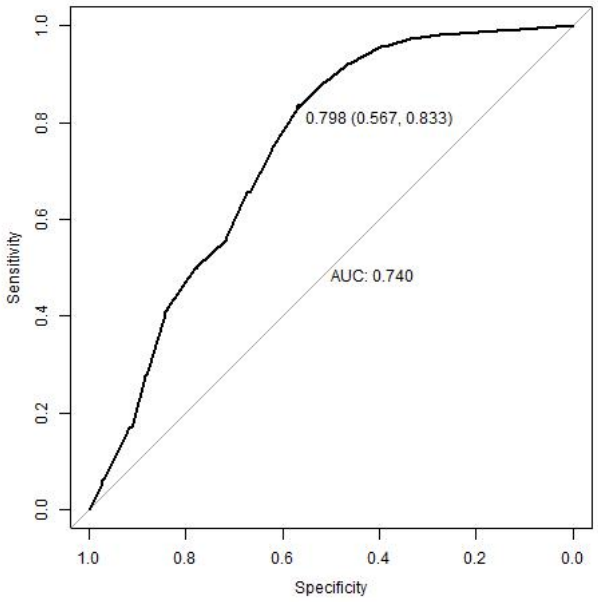


Figure 4: Area Under the Curve for Knn

3. SVM

   There are two methods to apply Support Vector Machines, either with a *linear* or *radial* kernel. For each, *cost* and *gamma* are two parameters to be tuned. Using a sequence of Gamma between 0.005 and 0.05 and sequence of cost between 2^-1 and 2^5, to find the best combination results in the best AUC of 0.702 for the Radial kernel.

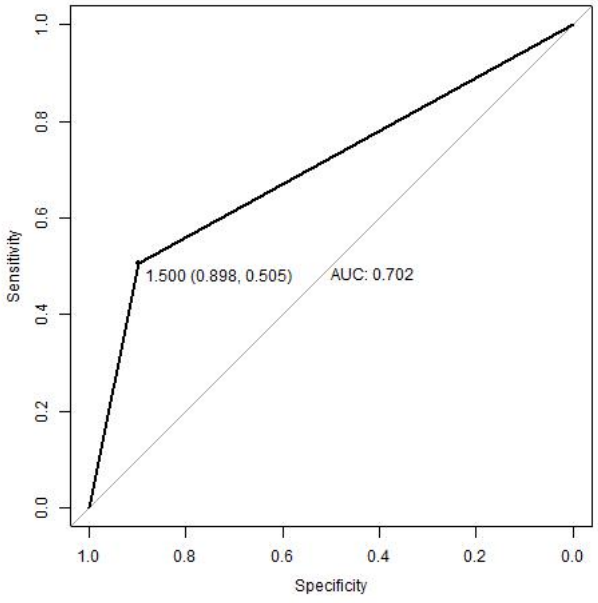


Figure 5: Area Under the Curve for SVM

4. Random Forest

Two main parameters are fine tuned to obtain the best model on the validation set: the number of trees and the number of variables to select for each decision tree. a number of trees of 150 combined with 5 variables for each decision tree resulted in the best model with an AUC of 0.841.

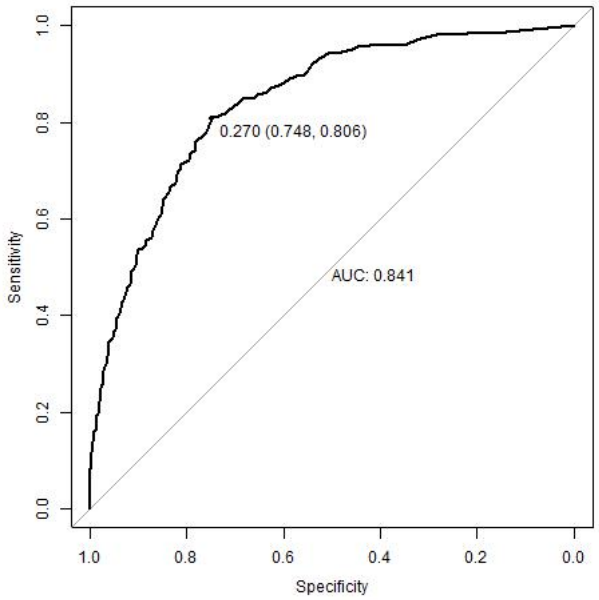


Figure 6: Area Under the Curve for Random Forest

So far, Random Forest is fitting the data the best, however, these methods need to be compared on the same data set.

4.2 Comparison on the Testing Set

The resulting models from each method will be used to make predictions on a new data set.

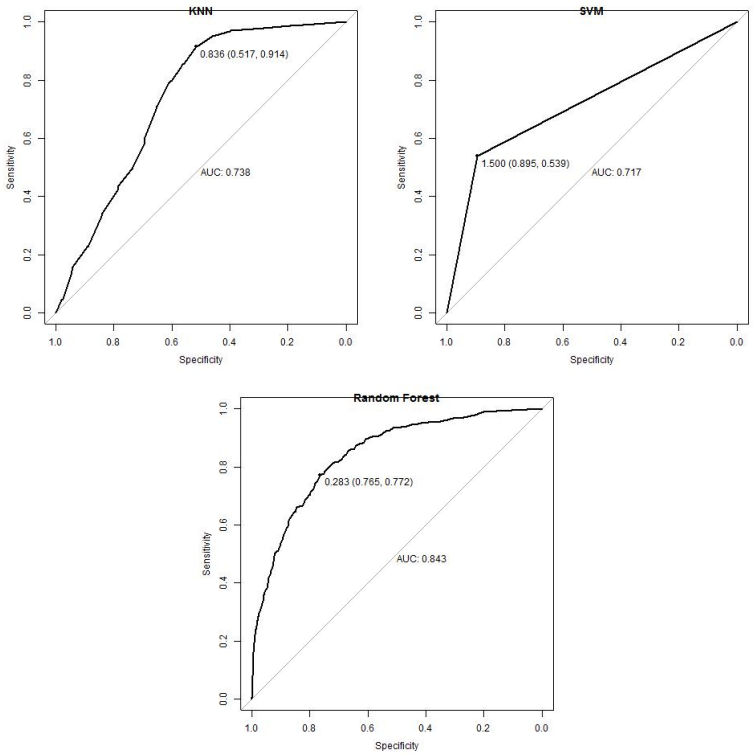


Figure 7: Area Under the Curve for the three methods

As illustrated in the figures, neither Knn nor SVM could outperform random forest for this dataset. However, The difference in performance is not significant. Table 1 summarizes results and computation time for the three methods:

Table 1: Results Summary

	KNN	SVM	Random Forest
Individually	0.739	0.702	0.841
On the same set	0.738	0.717	0.843
Duration	few seconds	one hour	few seconds

5 Managerial Implications

Correctly detecting churn behaviour allows firms to act proactively with regards to both maintaining a close customer relationship and increasing retention, making it easier for a firm to transform into a customer centric organization. In the particular case of subscription based companies - illustrated by the dataset used in our experiment -, missing on churn behaviour can have tremendous financial impact, not only from loosing one customer revenue stream, but also for the high acquisition efforts and costs that the company would have to engage in. Unlike one-time-payment models, in a subscription based setting, a customer is much less likely to return after churning. Managers need to allocate the necessary ressources to look closely at customer behaviour in terms of *complaint behaviour*, *ease of payment* and *loyalty*.

Building a strong analytical and predictive model for customer churn is a first step into efficient retention, and subsequently effective marketing funnel. Models used must be continuously improved and tweaked. Any action taken with a retention goal in mind (e.g., emails, special offer or discount) need to be data based to increase the fit with the customer expectation. Customer churn prediction is not merely a classification task, but a process with the goal of understanding consumer behaviour in its

heart. Only then a firm can become data driven, when most decisions involving customers are based on evidence.

6 Discussion and Conclusion

The Analysis of Churn factors revealed the positive influence of the duration of the contract, the longer a customer is subscribed to the company's service, the less likely he/she is to Churn. When implementing a churn prediction model, Companies should not limit their observations to the predictive power, but should also keep an eye on the practicality of implementation at scale. For instance, eventhough SVM results are usable, training a SVM model takes significantly more time than Random forest or Knn (approximately one hour). In a setting with larger customer base and real time data input, SVM would absolutely not be recommended. Besides, Complexity can be penalizing as illustrated by logistic regression results (best results with a simple model). Any team wowrking on customer churn need to find a balance between scalability and interpretability.
Eventhough bagging and boosting were not used for this experiment, including them wouldn't have changed the results significantly. However, addressing the output class unbalancement could have an impact on predictive performance of the three methods.

Predictive models in this paper have nevertheless not taken into account the financial costs of wrongly prediction of non-churners as churners or failing to predict real churners [2]. This limitation hinders the alignment with commercial goals of the firm.

References

[1] Jaradat K Harfoushi O Ghatasheh N. Adwan O, Faris H. Predicting customer churn in telecom industry using multilayer preceptron neural networks: modeling and analysis. *Life science journal*, 11(3):75–81, 2014.

[2] Djamila A. Alejandro C. B. and Björn O. A novel cost-sensitive framework for customer churn predictive modeling. *Decision Analytics*, 2(1):1–15.

[3] Ottersten B.r Bahnsen A.C., Aouada D. A novel cost-sensitive framework for customer churn predictive modeling. *Decision Analytics*, 2(1):1–15, 2015.

[4] W. Buckinx and Van den Poel D. Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual fmcg retail setting,. *European Journal of Operational Research*, 164(1):252–68, 2005.

[5] Wouter Buckinx and Dirk Van den Poel. Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual fmcg retail setting,. *European Journal of Operational Research*, 164(1):252–68, 2005.

[6] Kristof Coussement. Improving customer retention management through cost-sensitive learning. *European Journal of Marketing*, 48(3/4):477–495, 2014.

[7] Kristof Coussement and Dirk Van den Poel. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques,. *Expert Systems with Applications*, 34(1):313–27, 2008.

[8] Klyueva I. Demidova L. Predicting customer retention and profitability by using random forests and regression forests techniques,. *6th Mediterranean Conference on Embedded Computing*, 2017.

[9] Phuong V. Duyen D., Phuc H. Customer churn prediction in an internet service provider. *IEEE International Conference on Big Data*, 2017.

[10] Bendle N. T. Pfeifer P.E. Reibstein D.J. Farris, P.W. *Marketing Metrics: The Definitive Guide to Measuring Marketing Performance*. Pearson FT Press, New Jersey, USA, 2010.

[11] Sungzoon Cho Ha, Kyoungnam and Douglas MacLachlan. Response models based on bagging neural networks,. *Journal of Interactive Marketing*, 19(1):17–30, 2005.

[12] Zahavi J. and levin N. Applying neural computing to target marketing,. *Journal of Interactive Marketing*, 11(1):5–22, 1997.

[13] K.E. Reynolds. J. Ganesh, M.J. Arnold. Understanding the customer base of service providers: An examination of the differences between switchers and stayers. *Journal of Marketing*, 64(3):65–87, 2000.

[14] Hjalmar R. Bouma Fred Geus Anne H. Epema Jose Castela Forte, Marco A. Wiering. Predicting long-term mortality with first week post-operative data after coronary artery bypass grafting using machine learning models. *Journal of Machine Learning Reaserch*, 68, 2017.

[15] Aliannejadi M Ahmadian I Mozaffari M Abbasi U. Keramati A, Jafari-Marandi R. Improved churn prediction in telecommunication industry using data mining techniques. *Applied soft computing journal.*, 24:994–1012, 2014.

[16] Bart Larivière and Dirk Van den Poel. Predicting customer retention and profitability by using random forests and regression forests techniques,. *Expert Systems with Applications*, 29(2):472–84, 2005.

[17] Gordon H.G. McDougall Levesque T. Determinants of customer satisfaction in retail banking. *International journal of bank marketing*, 14(7):12–20, 1996.

[18] Gordon McDougal. Customer retention strategies,. *Services Marketing Quarterly*, 22(1):39–55, 2001.

[19] Bijmolt THA. Risselada H, Verhoef PC. Staying power of churn prediction models. *Journal of Interactive Marketing*, 24(3):198–208, 2010.

[20] Zahorik A.J. Rust R.T. Customer satisfaction, customer retention, and market share. *Journal of Retailing*, 69(2):193–215, 1993.

[21] B. Hardie W. Kahn V. Kumar N. Lin N. Ravishanker S. Sriram S. Gupta, D. Hanssens. Modeling customer lifetime value,. *Journal of Service Research*, 9(2):139–155, 2006.

[22] Bradlow E.T. Schweidel D.A., Fader P.S. Understanding service retention within and across cohorts using limited information. *Journal of Retailing*, 72(1):82–94, 2008.

[23] W. Kamakura L. Junxiang Scott A.N., S. Gupta and Mason C.H. Defection detection: Measuring and understanding the predictive accuracy of customer churn models,. *Journal of Marketing*, 43(2):204–11, 2006.

[24] Ewout W Steyerberg, Andrew J Vickers, Nancy R Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J Pencina, and Michael W Kattan. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128, 2010.

[25] Frank Block et al Tammo H. A. Bijmolt, Peter S. H. Leeflang. Analytics for customer engagement. *Journal of service research*, 13(3):341–356, 2010.

[26] Peter C. Verhoef. Understanding the effect of customer relationship management efforts on customer retention and customer share development. *Journal of Marketing*, 67(4):30–45, 2003.

[27] Donkers B. Verhoef P.C. The effect of acquisition channels on customer loyalty and cross-buying. *Journal of interactive marketing*, 19(2):31–43, 2005.

[28] Xiao J. Wang Y. Transfer ensemble model for customer churn prediction with imbalanced class distribution. *International Conference on Information Technology Computer Engineering and Management Sciences (ICM)*, pages 177–181, 2011.

[29] Xiu Li E.W.T. Ngai Xie, Yaya and Weiyun Ying. Customer churn prediction using improved balanced random forests,. *Expert Systems with Applications*, 36(3):5445–9, 2009.

[30] Jiayin Q. Yangming Z. and Huaying S. A hybrid knn-lr classifier and its application in customer churn prediction. *IEEE International Conference on Systems Man and Cybernetics*, pages 3265–3269, 2007.

7 Appendix

Table 2: Independent Variables Descriptive Statistics

Independent Variable	Class Proportion	Mean	St.Dev.
gender	Female: 0,505 Male: 0,495		0,50
Senior Citizen	Yes: 0,16		0,37
Partner	Yes: 0,48		0,49
Dependent	Yes: 0,29		0,46
Tenure	–	32,37	24,56
Phone Service	Yes: 0,90		0,29
Multiple Lines	No: 0,48 No Phone service: 0,09 Yes: 0,42		0,94
Internet Service	DSL: 2421 Fiber Optic: 3096 No: 1526		0,74
Online Security	NO: 3498 No internet service: 1526 Yes: 2019		0,86
OnlineBackup	NO: 3088 No internet service: 1526 Yes: 2429		0,88
Device Protection	NO: 3095 No internet service: 1526 Yes: 2422		0,87
TechSupport	NO: 3473 No internet service: 1526 Yes: 2044		0,86
StreamingTV	NO: 2819 No internet service: 1526 Yes: 2707		0,88
StreamingMovies	NO:2785 No internet service: 1526 Yes: 2732		0,88
Contract	NO: 3875 No internet service: 1473 Yes: 1695		0,83
PaperlessBilling	No: 2872 Yes: 4171		0,49
Paymnet Method	bank transfer: 1544 credit card: 1522 electronic check: 2365 mail check: 1612		1,06
MonthlyCharges	–	64,76	30,1
TotalCharges	–	2283,3	2266,7
Churn	No: 5174 Yes: 1869		0,44

Total number of observations, $N = 7,032$