Final Project

Customer churn analysis in the telecommunication industry

Niharika Reddy(nb7ug)

Problem

Acquiring new customers can be more expensive for companies in comparison to retention of existing customers. The telecommunication industry (TCI) is facing a lot of stress due to fierce competition, saturated markets and reduced product lifecycles. All of these powerful trends are forcing telecom companies to respond through more competitive offers, bundles, and price cuts. Given the dynamics of this industry, it is of utmost importance that the companies manage their customer base effectively as acquiring new customers can be more expensive than retaining existing ones. Churn rate for wireless carriers in the USA, ranges from about 1% to 4% every quarter[i]. Targeted customer retention programs are needed to tackle this problem. Increasing customer retention rates by just 5% can increase profits by 25% to 95%[ii].

Objectives and Metrics

The objective of this project is to accurately predict which customers will switch providers and understand the behavior of likely-to-churn customers so that appropriate retention programs can be undertaken. The metric used to assess the outcome of the project is the misclassification rate which needs to be minimized.

State-of-the-art Prior Work

A risk prediction technique that utilizes Generalized Additive Models was presented by Coussement et al[iii]. This technique is exhibited to improve marketing decisions by identifying the risky customers and also providing visualizations of non-linear relationships. A neural network-based customer profiling technique was presented by Tiwari et al[iv]. This technique differs from the other proposed techniques by the fact that most of the techniques are only able to identify the customers who will instantaneously churn. However, the neural network-based churn prediction model proposes to predict customer's future churn behavior, providing the buffer for the organizations to perform prevention activities. Many challenges can arise in this analysis such as target leakage, unavailable or missing information, or the need for optimal feature transformations. Even constructing the target variable for the churn event may not always be straightforward. For example, in a setting where customers cancel and renew frequently, how can we define churn? What about in a setting where customers can subscribe and purchase multiple product lines?

Hypothesis and Approach

In order to improve the predictive performance of the classification model, an ensemble technique is proposed which is hypothesized to give better results than using just one technique. A combination of random forest, SVM and logistic regression are herewith proposed to produce better accuracy compared to a logistic regression. The methods will entail engineering of novel features (such as subscription to multiple product lines) not used in prior work and will aim for better predictive accuracy.

The dataset being investigated for this analysis will contain customer demographic data, usage data and billing data for an anonymized telecom service provider in the U.S. This data will be

sourced from Kaggle[v]. There is a risk of bias in the data due to the fact that only one service provider is included in the analysis. Also, the class imbalance could result in added bias.
The evaluation setup to assess the hypothesis, which is the misclassification error, would objectively measure the fit of the model to the data.

Execution and Results

Exploratory Data Analysis

Before proceeding with any analysis, the data was explored to find any patterns or oddities in the data. The column of TotalCharges was observed to have missing values(a count of 11) and the respective rows were dropped(Fig. 1). Next the proportion of customers that churned was explored. About 25% of customers are shown to have churned (Fig. 2). Also, monthly charges of customers who churned seems to be a lot higher than those that haven't. Compared to this, total charges of customers that churned seems to be lesser(Fig. 3). This shows that customers that have not churned are subscribed to more services from provider compared to just the phone service alone. The other categorical variables were analysed similarly (Fig. 4) and the following insights were obtained

1. Customers with month-to-month contract churned the most in their category.
2. Customers with FiberOptic internet service churned quite a lot as well.
3. Customers that didn't opt for internet security churned way more than others.
4. Customers that didn't have tv streaming services churned a lot. About 60% of customers who did not have streaming TV have churned

Model-Building and Cross Validation

1. A single decision tree
   A single decision tree was plotted and the following variables were chosen to create the splits in the tree: "Contract", "MonthlyCharges", "OnlineSecurity" and "tenure". The tree resulted in 5 terminal nodes and on cross-validation using the misclassification error was found to have an accuracy of 77.2%. Pruning the tree did not seem to improve the accuracy. Refer to Fig.5 and 6 in Appendix.

   Confusion matrix:

	0	1
0	**1182**	**293**
1	**108**	**174**

2. Random forest
   The next approach was to build a random forest in order to reduce the bias in the model. On cross-validation with a validation set, the best method appeared to be to grow 250 trees using 3 variables at every split. This gave an accuracy of 80.3%. Fig. 6 details the chart of variable importance.

   Confusion matrix:

	0	1
0	**1178**	**234**
1	**112**	**233**

3. Support vector classifier
   The next approach was to explore the results of a support vector classifier using both a linear kernel and a radial kernel. The linear kernel with a cost tuning parameter of 0.1 performed better than a radial kernel at different cost and gamma parameters. The accuracy of the support vector classifier with linear kernel was 77.8%.

   Confusion matrix:

	0	1
0	**1170**	**270**
1	**120**	**197**

4. Logistic Regression
   The final methodology employed in this paper is a logistic regression framework. Features were chosen by using AIC to extract variables of significance. The variables that were chosen to be significant were:
   SeniorCitizen + Dependents + tenure + PhoneService + PaperlessBilling + TotalCharges + Contract + OnlineSecurity + InternetService

   The accuracy of the logistic regression model using the above features was found to be 79.7%

   Confusion matrix:

	0	1
0	**1167**	**233**
1	**123**	**234**

5. Ensemble Model
   An ensemble of all the above models was created, by taking the mode of predictions. The accuracy on performing cross validation was found to be 80.2% which wasn't better than the accuracy of the random forest.

   Confusion matrix:

	0	1
1	**1183**	**241**
2	**107**	**226**

Conclusion

With respect to the initial hypothesis, the ensemble method did not produce the best results. Random forest seems to have performed the best upon cross-validation. This proves that linear models have not performed well to classify the response variable. However, achieving an accuracy of 80% is no mean feat could lead companies on their way to retaining more of their customers.

[i] https://www.statista.com/statistics/283511/average-monthly-churn-rate-top-wireless-carriers-us/
[ii] https://www.getcloudcherry.com/blog/improve-customer-retention-in-telecom/

[iii] K. Coussement, D.F. Benoit, D. Van den Poel, Preventing customers from running away! Exploring generalized additive models for customer churn prediction. The Sustainable Global Marketplace, Springer International Publishing (2015)

[iv] A. Tiwari, J. Hadden, C. Turner, A new neural network based customer profiling methodology for churn prediction.Computational Science and Its Applications–ICCSA 2010, Springer, Berlin Heidelberg (2010)

[v] https://www.kaggle.com/danwheble/churn-prediction-telco-customer-churn/data

Appendix

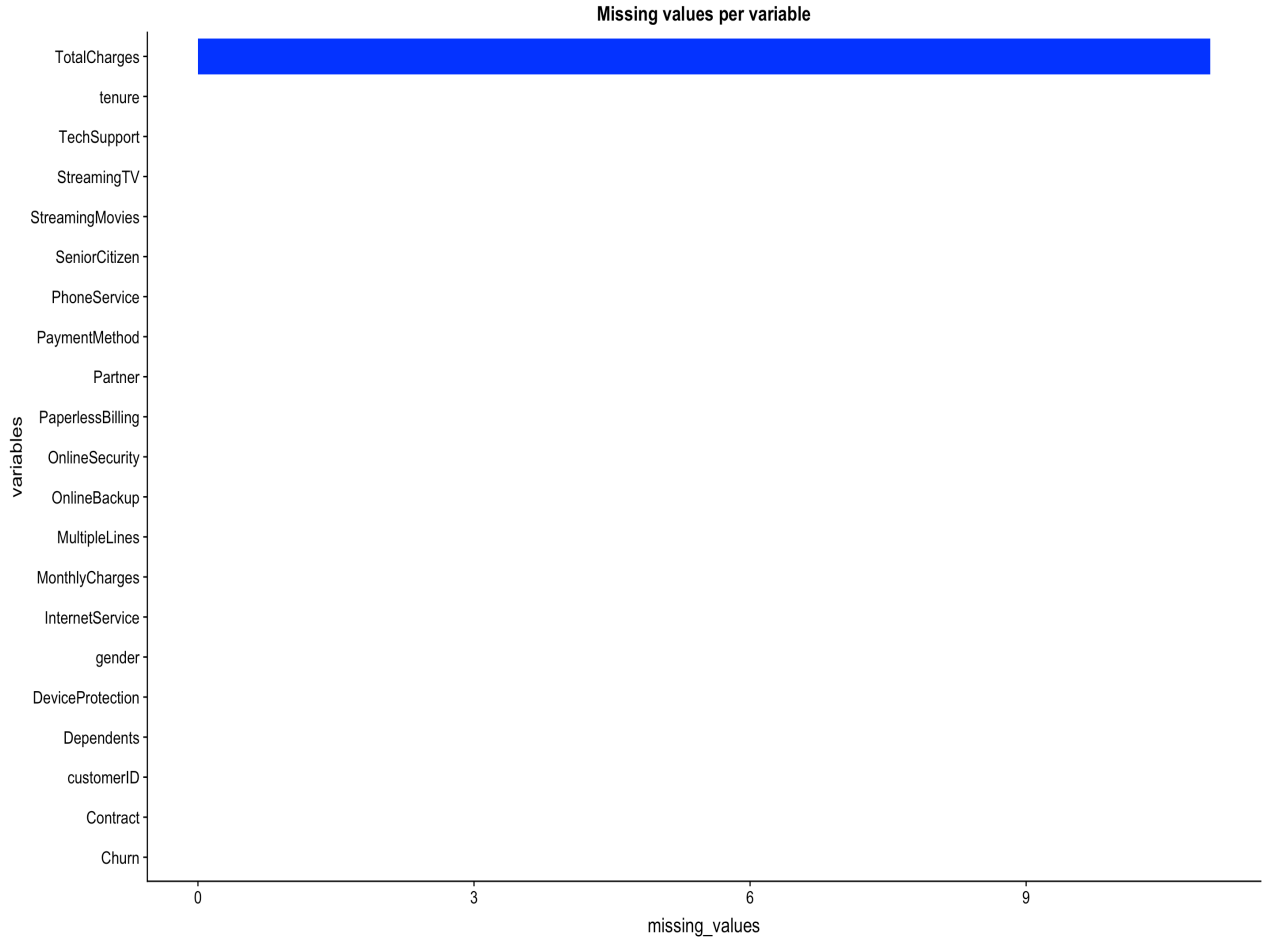


Fig. 1

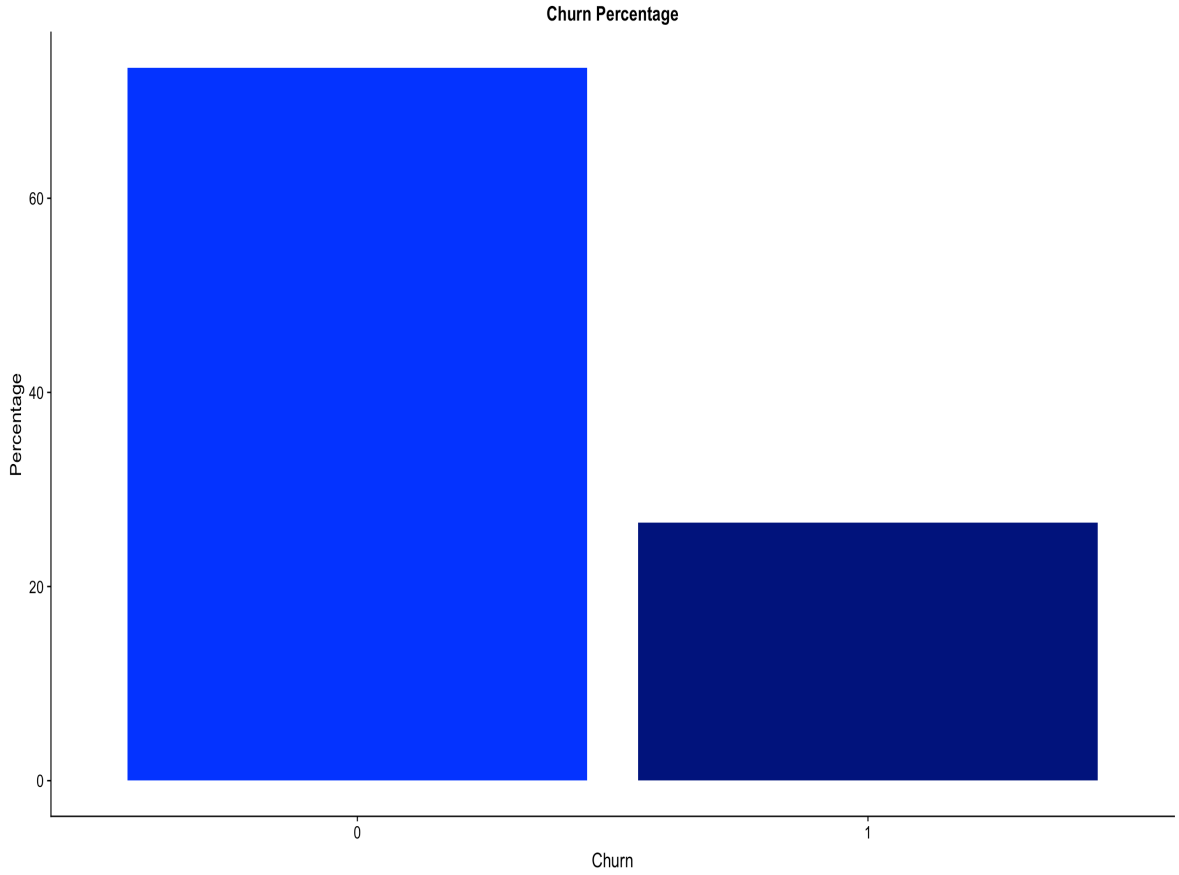
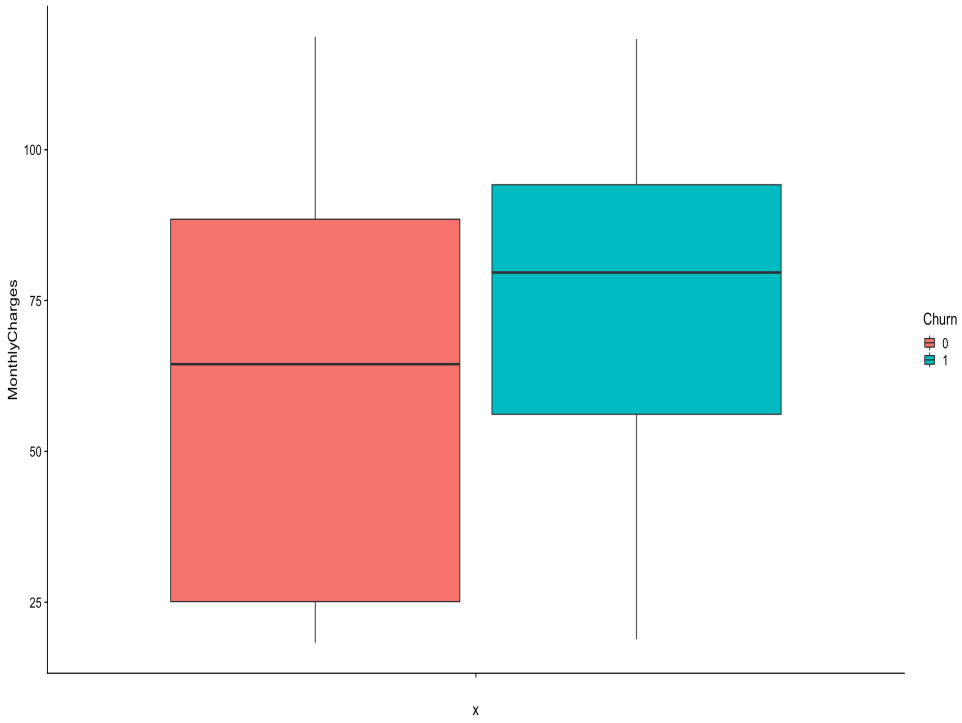


Fig. 2



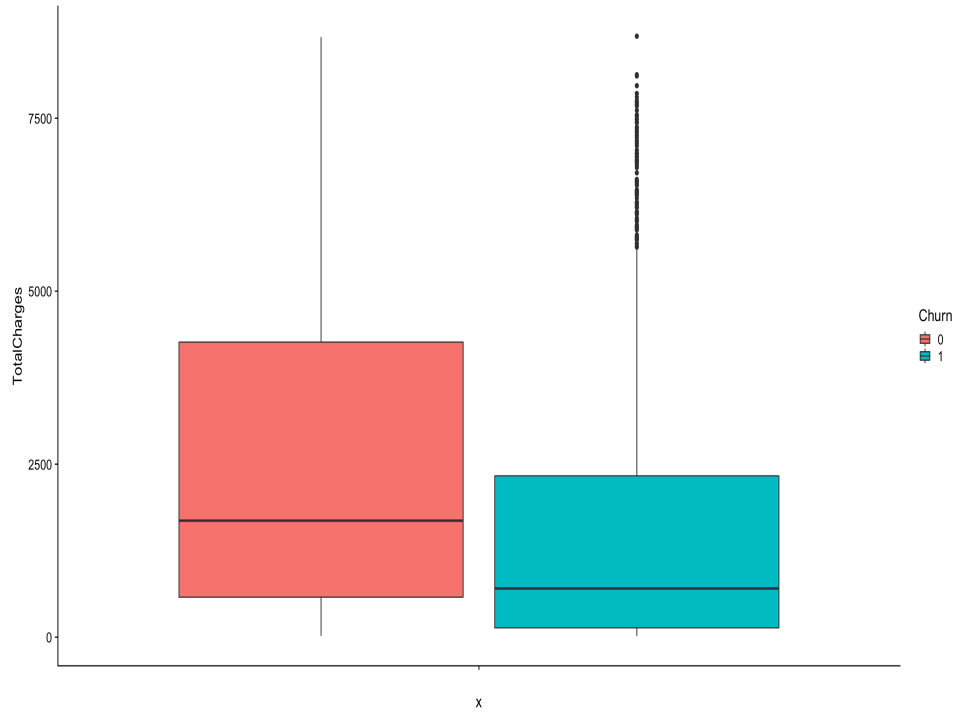


Fig.3(Analysis of Monthly charges vs total charges)

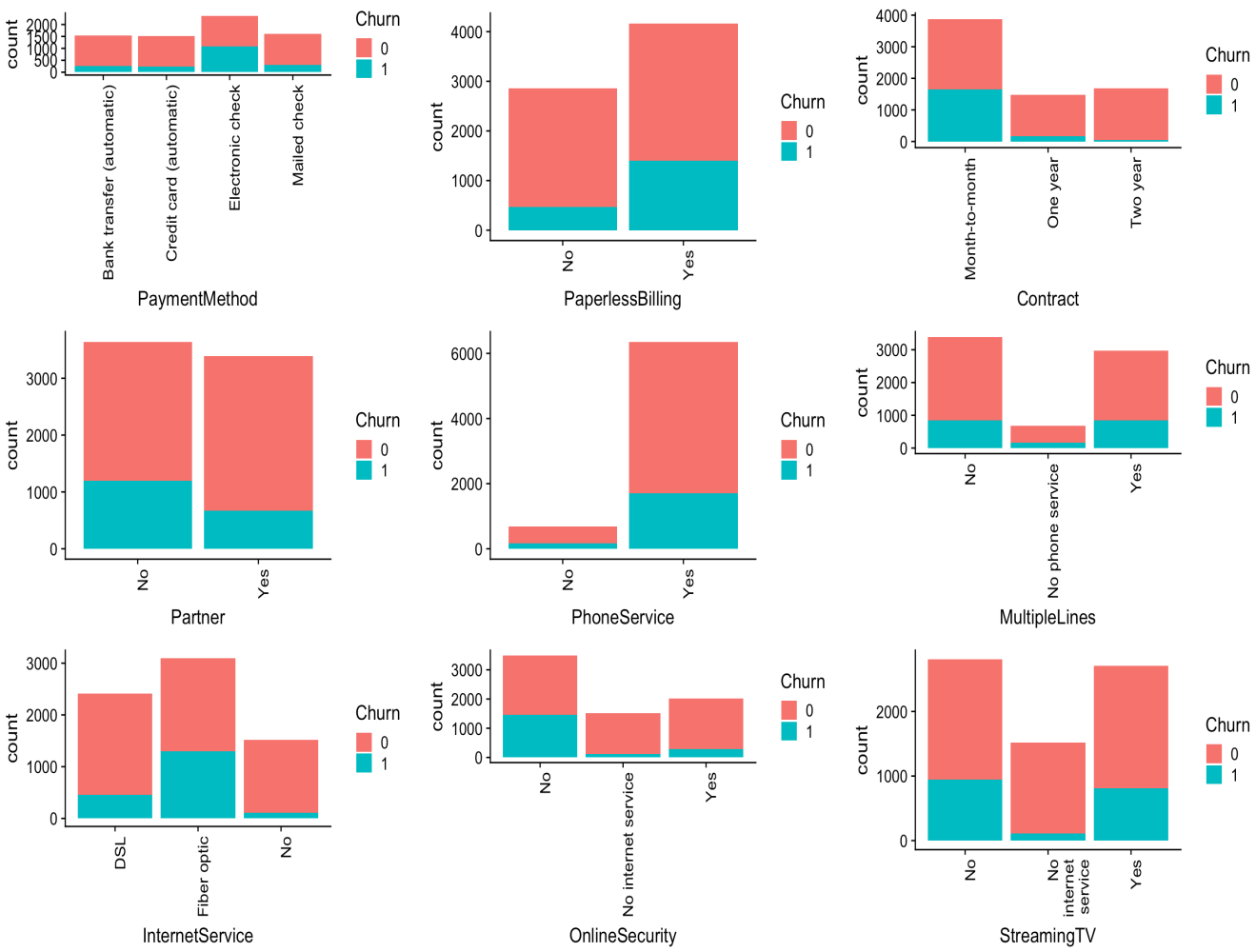

count
Churn
0
1
Bank transfer (automatic)
Credit card (automatic)
Electronic check
Mailed check
PaymentMethod
No
Yes
PaperlessBilling
Month-to-month
One year
Two year
Contract
Partner
PhoneService
No phone service
MultipleLines
DSL
Fiber optic
InternetService
No internet service
OnlineSecurity
No internet service
StreamingTV

Fig. 4

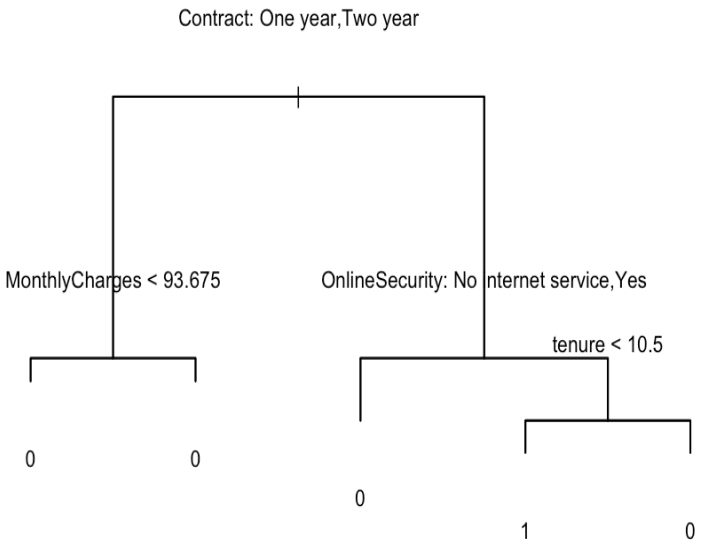

Contract: One year,Two year
MonthlyCharges < 93.675
OnlineSecurity: No internet service,Yes
tenure < 10.5
0
0
0
1
0

Fig.5

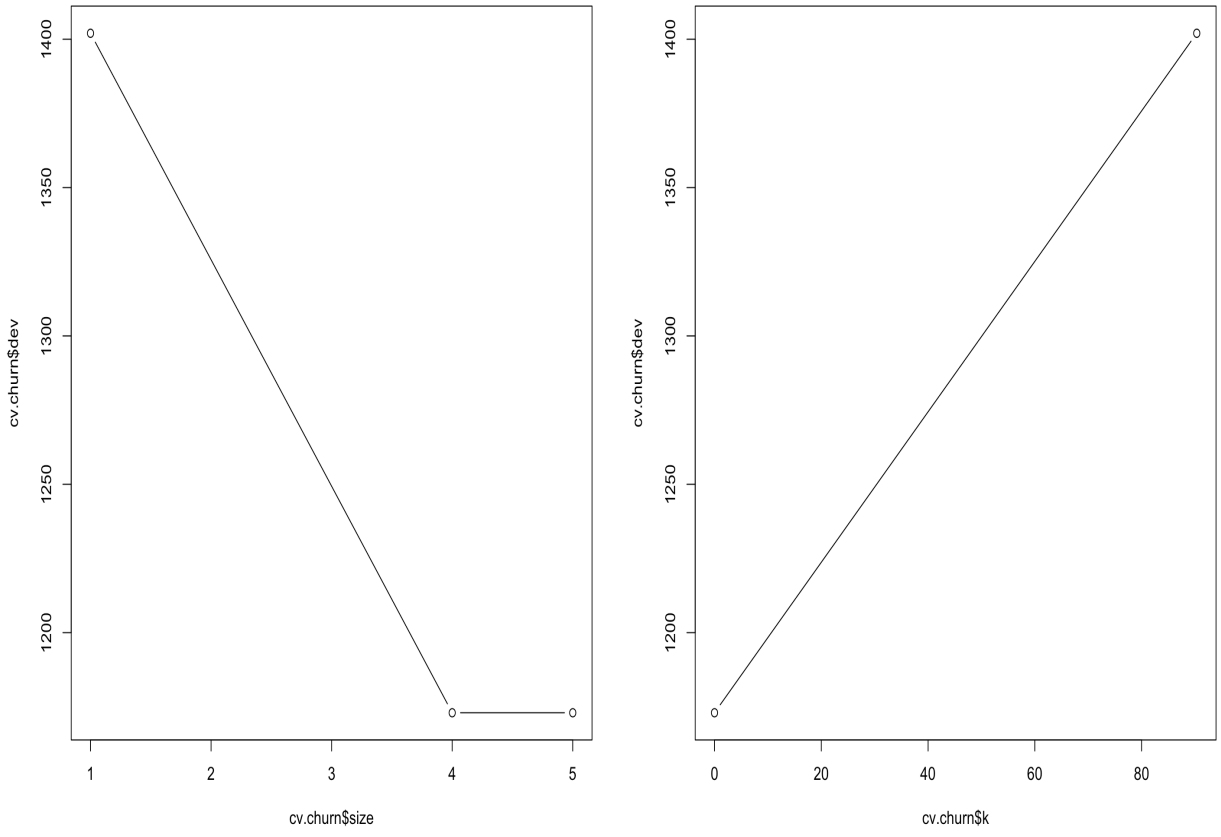

cv.churn$dev
1400
1350
1300
1250
1200
1
2
3
4
5
cv.churn$size
cv.churn$dev
1400
1350
1300
1250
1200
0
20
40
60
80
cv.churn$k

Fig. 6