Yinan Guo

guoy221@wfu.edu

2021/10/25

# Executive Summary

## Problem

This task is to build machine learning models to help identify likely churners for mobile phone companies like Verizon, ATT, etc. In this case, the company can incentivize the churners to stay with the company.

## Key Findings

1. Network speed is not a significant factor to affect churn.
2. Customers using Google Pixel 4a have a churn rate which is over 5-times more than churn rate for whom using iPhone XR. There is difference in customers using different phone models so that it is an important factor.
3. Customers who use paper billings are 2-times more likely to churn than those who use paperless billing. Thus, it is a significant factor.
4. Customers who use EURO are 6-times more likely to churn than who use USD and 3-times more likely than who use CAD.

## Model Performance Summary & Interpretation

1. My final model is a logistic regression with an AUC of 0.935 for train and 0.932 for test, which performs great. I select predicters other than some of insignificant factors like customer id, billing address, billing city, etc. The most important predictors are total billed, paperless billing and number of phones. With higher total billed amount, accept paperless billing and higher number of phones, customers are less likely to churn.
2. By changing the threshold to 0.269, where recall is 0.70 and precision is 0.695, the final TPR is 70.6% and the FPR is 1.9%. According to the rule of Earning=(-50(FP+TP)+500TP-150FN), out of 27271 samples in test data of training dataset where 1557 of them have churned, this model can help the company earn $400,350.

## Recommendations

1. Since Customers who use EURO are highly likely to churn, the company can introduce more special offer aiming to European customers to save them from churning.
2. Since customers who have higher monthly calling minutes and streaming minutes tend to have higher churning likelihood, the company could give out more preferential policies or other activities for these customers to prevent them from leaving this mobile phone company.

# Detailed Analysis & Steps

### File(s) Summary

| **File Name** | **Record count** | **Column count** | **Numeric columns** | **Character columns** | **Date columns** |
| --- | --- | --- | --- | --- | --- |
| Churn\_training.csv | 90901 | 34 | 11 | 22 | 1 |
| Churn\_holdout.csv | 10099 | 33 | 10 | 22 | 1 |

### Field Summary

| **Name** | **Data Type** | **Feature Type** | **Count** | **# Distinct** | **% Distinct** | **# Missing** | **% Missing** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| churn | num | target | 90901 | 2 | 0.00% | 0 | 0.00% |
| customer\_reg\_date | date | Date | 90874 | 308 | 0.33% | 27 | 0.00% |
| email\_domain | char | categorical | 90873 | 3 | 0.00 | 28 | 0.03% |
| phone\_model | char | categorical | 90876 | 15 | 0.00 | 25 | 0.03% |
| billing\_city | char | categorical | 90872 | 8140 | 0.09 | 29 | 0.03% |
| billing\_postal | char | categorical | 90873 | 9956 | 0.11 | 28 | 0.03% |
| billing\_state | char | categorical | 90875 | 48 | 0.00 | 26 | 0.03% |
| partner | char | categorical | 90876 | 2 | 0.00 | 25 | 0.03% |
| phone\_service | char | categorical | 90876 | 2 | 0.00 | 25 | 0.03% |
| multiple\_lines | char | categorical | 90877 | 2 | 0.00 | 24 | 0.03% |
| streaming\_plan | char | categorical | 90873 | 4 | 0.00 | 28 | 0.03% |
| mobile\_hotspot | char | categorical | 90865 | 2 | 0.00 | 36 | 0.04% |
| wifi\_calling\_text | char | categorical | 90869 | 2 | 0.00 | 32 | 0.04% |
| online\_backup | char | categorical | 90872 | 3 | 0.00 | 29 | 0.03% |
| device\_protection | char | categorical | 90872 | 26 | 0.00 | 29 | 0.03% |
| contract\_code | char | categorical | 90875 | 26 | 0.00 | 26 | 0.03% |
| currency\_code | char | categorical | 90872 | 3 | 0.00 | 29 | 0.03% |
| maling\_code | char | categorical | 90870 | 26 | 0.00 | 31 | 0.03% |
| paperless\_billing | char | categorical | 90870 | 2 | 0.00 | 31 | 0.03% |
| payment\_method | char | categorical | 90877 | 4 | 0.00 | 24 | 0.03% |
| customer\_id | char | ID | 90901 | 90901 | 1.00 | 0 | 0.00% |
| billing\_address | char | text | 90881 | 90880 | 1.00 | 20 | 0.02% |
| gender | char | categorical | 90874 | 2 | 0.00 | 27 | 0.03% |
| network\_speed | char | categorical | 90874 | 2 | 0.00 | 27 | 0.03% |
|  |  |  |  |  |  |  |  |

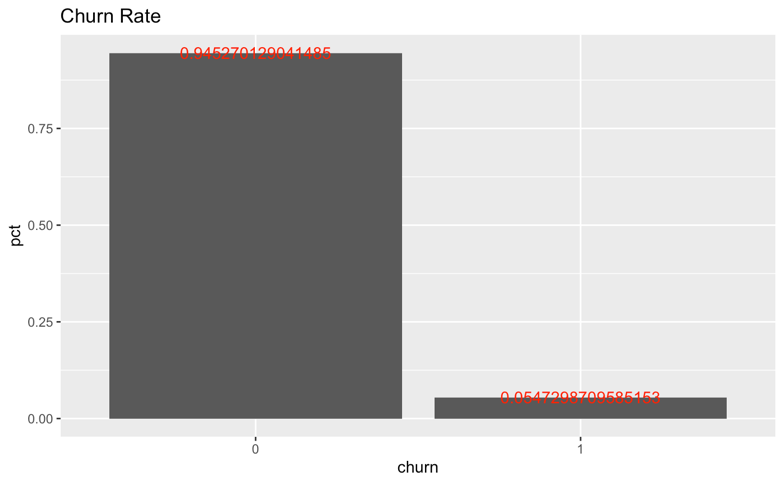
For numeric variables:

| **column** | **n** | **dist** | **nmiss** | **mean** | **min** | **max** |
| --- | --- | --- | --- | --- | --- | --- |
| monthly\_minutes | 90881 | 22100 | 20 | 19852 | 0 | 43799 |
| customer\_service\_calls | 90879 | 5 | 22 | 1.65 | 0 | 4 |
| streaming\_minutes | 90879 | 21935 | 22 | 20697 | 0 | 43799 |
| total\_billed | 90867 | 272 | 34 | 250 | 100 | 399 |
| prev\_balance | 90879 | 93 | 22 | 51.5 | 0 | 99 |
| late\_payments | 90881 | 10 | 20 | 4.8 | 0 | 9 |
| ip\_address\_asn | 90884 | 10570 | 17 | 34847 | 2013 | 65533 |
| phone\_area\_code | 90873 | 85 | 28 | 248 | 200 | 289 |
| number\_phones | 90871 | 11 | 30 | 5.31 | 0 | 10 |
| senior\_citizen | 90866 | 2 | 35 | 0.5 | 0 | 1 |
|  |  |  |  |  |  |  |

## Target Summary

* The churn rate is 5.47% and the default rate is 94.53%.

|  |  |  |
| --- | --- | --- |
| Churn | n | Pct |
| No | 85926 | 0.94527013 |
| Yes | 4975 | 0.05472987 |



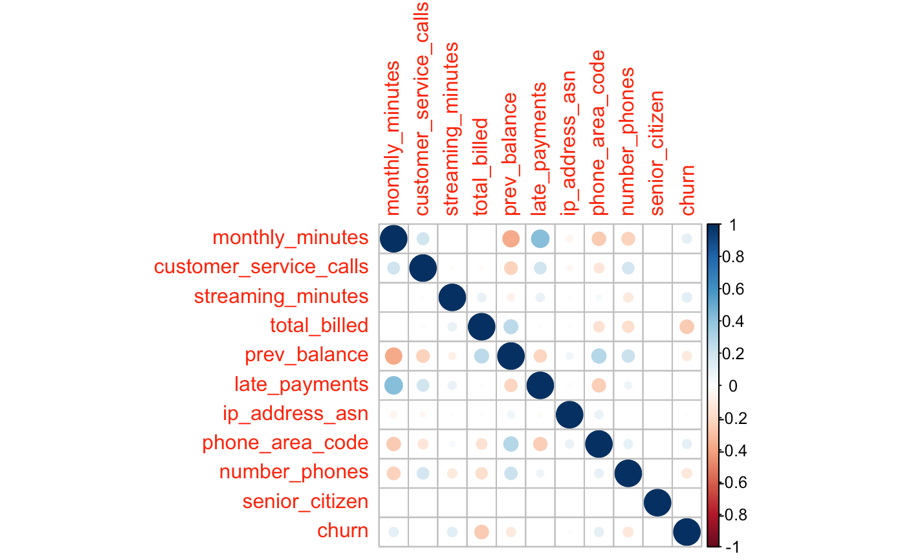
## Exploratory Data Analysis & Screening

### Descriptive Statistics

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

Here are the boxplots for most of the numeric variables. Among all the numeric variables, monthly minutes, prev balance, streaming minutes, number phones phone area code and total billed have different distributions within different groups divided by churn or not, indicating that these may be important factors to affect churning.

### Correlation Analysis



Among all the numeric variables, total\_billed has the largest correlation with churn and they are negatively related. Also, late payments is positively related to monthly minutes and prev balance is negatively related to monthly minutes.

### Frequency Analysis

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

According to the histogram of characters, if different columns in a graph have different proportions of blue part, which indicates the proportion of churn, then the churn rate is strongly correlated with this variable. (i.e. The barplot for mailing code shows a strong difference in proportion of churn in different mailing codes, indicating that the mailing code may affect churning decisions for the customers.)

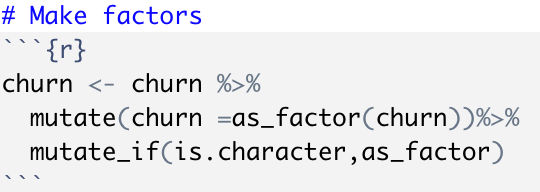
## Initial Screening & Exploration

|  |  |
| --- | --- |
| Customers using hotmail has the largest rate of churn, which has a churn rate 2-times about average whereas churn rate for customers using yahoo is far below the average. Thus, email domain is an important factor. | Nearly 25% of the customers using 12GB streaming plan churn, which is 5-times of average churn rate. Thus, streaming plan is an important factor. |
| Over 15% of the customers using EURO to pay churn, whereas only 3% of the customers using USD churn. Thus, currency code is an important factor. | Churn rate has a great difference in different mailing code. All of the customers to have Z and I churn so it is great signal to indicate churn. |
| Over 40% percent of customers who pay by bank transfer churn. So that payment method is an important factor. | As for gender, since the churn rate in male group and female group are similar, it is not an important factor to help decide churn or not. |
| The distribution for customer service calls in different groups are almost the same so that it is not a useful predictor. | The overall total billed of churn users is smaller. Thus, it may be a useful predictor. |

## Data Preparation & Transformation

1.make factors

The target and all of the characters are turned into factors.



2.using recipe to deal with variables

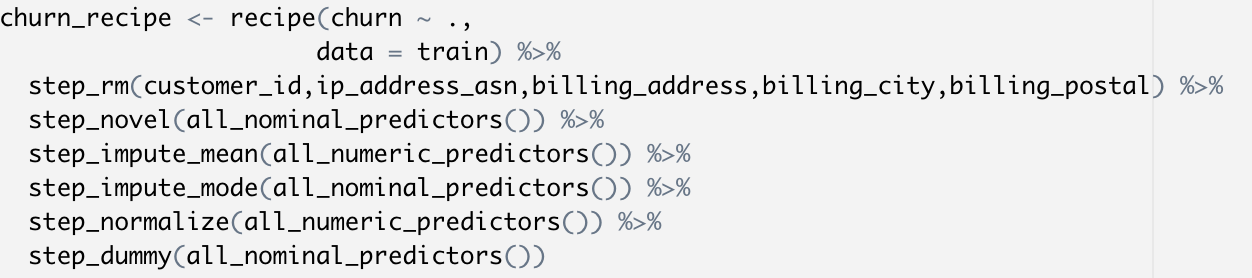
By step\_rm, i remove some unimportant variables. (This is an example for full logistic model.)

By step\_impute\_mean I deal with missing numeric variables to make all of them to be mean value of that variable.

By step\_impute\_mode, I deal with missing character variables to make all of them to be mode value of that variable.

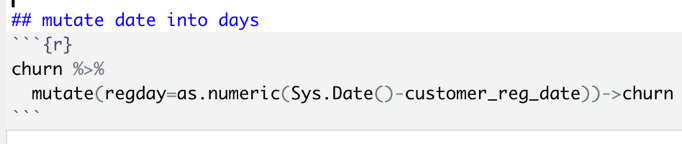
By step\_normalize, I scale and center my data.

By step\_dummy, I deal with character variables to make predictions.

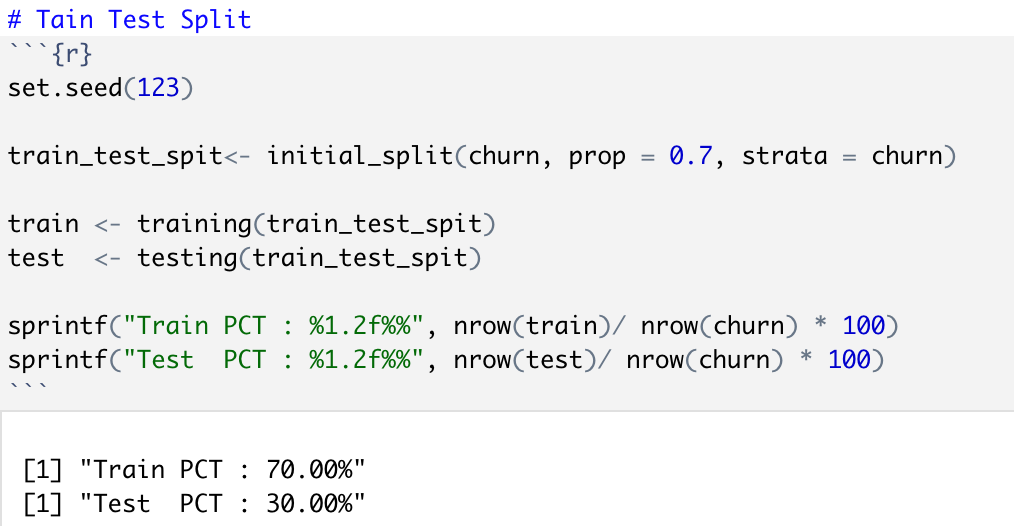


### Derive new variables

To adopt registration date in the models, I transfer registration date into days till now.



## Model Building

70% of data are used to build the model and 30% are used for validation. 

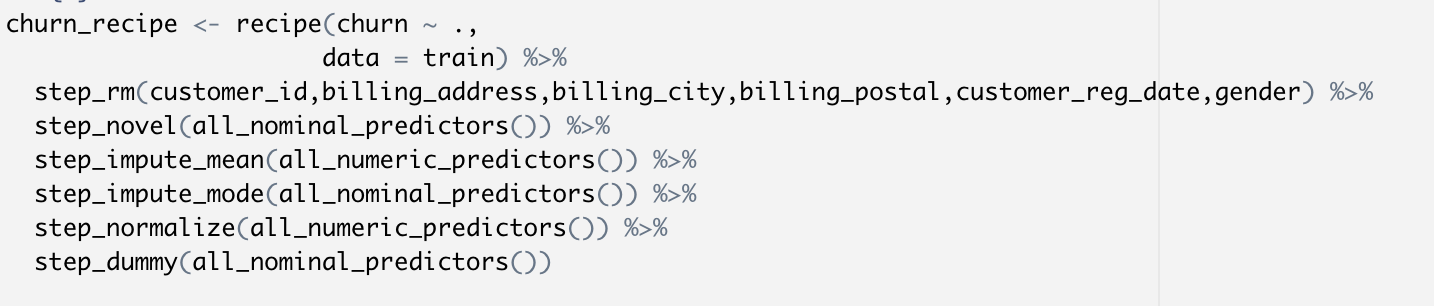
## Model Training

* Document the variables and the roles used

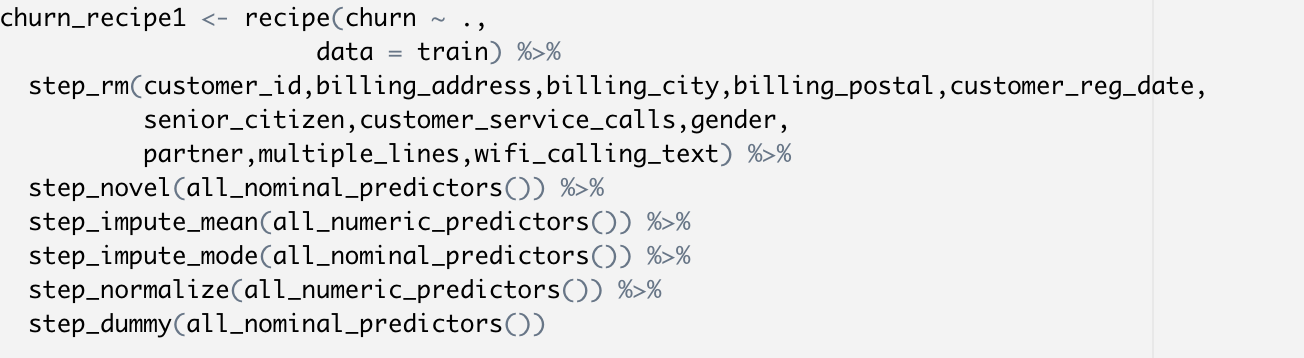
| **Name** | **Data Type** | **Full logistic/default tree/best complexity tree** | **Reduced logistic** |
| --- | --- | --- | --- |
| churn | num(target) | / | / |
| customer\_id | num | 0 | 0 |
| customer\_reg\_date | date | 0 | 0 |
| email\_domain | char | 1 | 1 |
| phone\_model | char | 1 | 1 |
| billing\_city | char | 0 | 0 |
| billing\_postal | char | 0 | 0 |
| billing\_state | char | 1 | 1 |
| partner | char | 1 | 0 |
| phone\_service | char | 1 | 1 |
| multiple\_lines | char | 1 | 0 |
| streaming\_plan | char | 1 | 1 |
| mobile\_hotspot | char | 1 | 1 |
| wifi\_calling\_text | char | 1 | 0 |
| online\_backup | char | 1 | 1 |
| device\_protection | char | 1 | 1 |
| contract\_code | char | 1 | 1 |
| currency\_code | char | 1 | 1 |
| maling\_code | char | 1 | 1 |
| paperless\_billing | char | 1 | 1 |
| payment\_method | char | 1 | 1 |
| billing\_address | char | 0 | 0 |
| gender | char | 1 | 0 |
| network\_speed | char | 1 | 1 |
| monthly\_minutes | num | 1 | 1 |
| customer\_service\_calls | num | 1 | 0 |
| streaming\_minutes | num | 1 | 1 |
| total\_billed | num | 1 | 1 |
| prev\_balance | num | 1 | 1 |
| late\_payments | num | 1 | 1 |
| ip\_address\_asn | num | 1 | 1 |
| phone\_area\_code | num | 1 | 1 |
| number\_phones | num | 1 | 1 |
| senior\_citizen | num | 1 | 0 |
| regtime | num | 1 | 1 |

Since the default tree model and the best complexity tree are using the same recipe with full logistic model, there are only 2 recipes. The first recipe only removed some variables with too many distinct values which can only represent individual situation but not contributive to the prediction. The second recipe for reduced logistic also removed some variables with high pvalues.

* Define Recipes
* Full logistic/default tree/best complexity tree



* Reduced logistic



* Define your Model

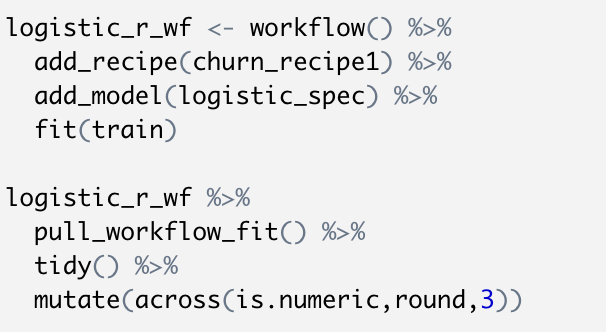
Except what has been removed in the recipes, all of other predictors are used in the model.

The top 5 important factors for each model are:

|  |  |  |  |
| --- | --- | --- | --- |
| Full logistic | Reduced logistic | Default tree | Best complexity tree |
| Total billed | Total billed | Total billed | Total billed |
| Paperless billing | Paperless billing | Streaming minutes | Streaming minutes |
| Number phones | Number phones | Prev balance | Streaming minutes |
| Prev balance | Monthly minutes | Number phones | Prev balance |
| Payment method | Prev balance | Monthly minutes | Payment method |

* Create a workflow and fit the model

Codes are shown as following:



* Evaluate metrics on Train and Test:
  + Classification:
    - table of Metrics:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | | AUC | |
| Train | Test | Train | Test |
| Full logistic | 0.9706 | 0.9700 | 0.9405 | 0.9365 |
| Reduced logistic | 0.9696 | 0.9696 | 0.9353 | 0.9318 |
| Default tree | 0.9761 | 0.9282 | 0.9662 | 0.8997 |
| Best complexity tree | 0.9690 | 0.9658 | 0.8754 | 0.8722 |

* + - ROC comparing trains and tests

|  |  |
| --- | --- |
|  |  |

* + - Confusion matrix comparing train and test

|  |  |
| --- | --- |
| Train | Test |
| Full logistic | |
|  |  |
| Reduced logistic | |
|  |  |
| Default tree | |
|  |  |
| Best complexity tree | |
|  |  |

* + - Test Score Threshold

The rule to select threshold is to find out a threshold where both of recall and precision are relatively high, where I got 0.282 for full logistic model, 0.269 for reduced logistic model, 0.158 for default tree model and 0.119 for best complexity tree model.

|  |  |
| --- | --- |
| Full logistic | Reduced logistic |
| Default tree | Best complexity tree |

* + - chart of variable importance
* Top 10 important variables for models

|  |  |
| --- | --- |
| Full Model | Reduced Model |
| Default tree | **Best complexity tree** |

* + - Score distribution for test dataset

|  |  |
| --- | --- |
| Full Model | Reduced Model |
| Default tree | **Best complexity tree** |

## Model Comparison

* Compare and contrast model 1 and model 2

I chose reduced logistic model and default tree to make final comparation. And here is the chart showing accuracy, auc, precision and recall for both of the models.

* Accuracy is the rate for (TP+TN)/(TP+TN+FP+FN)
* Precision is the rate for TP/(TP+FP), which represents the positive predictive value and important to a model.
* Recall is the rate for TP/(TP+FN), which represents the true positive rate and important to a model.
* AUC is the area under the ROC.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Accuracy | AUC | Precision | Recall |
| Reduced Logistic | train | 0.970 | 0.935 | 0.844 | 0.656 |
| test | 0.970 | 0.932 | 0.849 | 0.568 |
| Default Tree | train | 0.963 | 0.902 | 0.838 | 0.563 |
| test | 0.975 | 0.928 | 0.751 | 0.585 |

* Which model performs better and why?

Reduced logistic model performs better. Though the test accuracy of default tree is a little bit higher and the recall rate is similar to logistic model, the precision rate is much lower than logistic model. Also, the difference between AUC for logistic model is small, so that there is no overfitting problem.

* Compare their auc/accuracy/precision on training and test sets.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Accuracy | AUC | Precision | Recall |
| Reduced Logistic | train | 0.970 | 0.935 | 0.844 | 0.656 |
| test | 0.970 | 0.932 | 0.849 | 0.568 |
| Default Tree | train | 0.963 | 0.902 | 0.838 | 0.563 |
| test | 0.975 | 0.928 | 0.751 | 0.585 |