Executive Summary

# Challenge Overview

Customer churn, also known as customer attrition, occurs when customers stop using a company’s services and transfer to somewhere else. In fact, it is always more difficult and expensive to acquire a new customer than it is to retain a current customer. Therefore, we are conducting analysis to collect attributes of churned customers and build models to better predict who is more likely to churn. By doing this, we could perform better to retain our current customers.

# Problem Statement

The challenge is to create two KNN models compared with a classification tree model to predict which customers stop using our company’s services. As the company believes, we will explore if customer churn is related to network speed, phone model, and paperless billing. Furthermore, we will screen all the useful variables and build models based on the given dataset.

# Key Findings

1. Except for paperless billing, the other two variables, network speed and phone model are not closely related to customer churn rate as the company expected.
2. Paperless billing is the most significant predictor used to predict customer churn tendency. Customers who do not choose paperless billing method are more likely to churn.
3. Other significant determiners of customer churn are total\_billed, number\_phones, streaming\_minutes, streaming\_plan, prev\_balance, monthly\_minutes, late\_payments, and payment\_method. We can more accurately predict if a customer will churn using above mentioned variables.
4. By using either of our KNN models, the company could save at least $1,392,100 by reducing 78.33% of loss in churners.

# Recommendations

1. Unlike the company expected, phone models and network speed do not influence customers attrition. The company should not take these two variables as useful predictors.
2. The company should pay attention to customers who do not opt for paperless billing, customers with high monthly minutes and streaming minutes, customers who’s streaming plan is not 3GB, as they are more likely to churn. The company should try to retain those customers with coupons.
3. Even though the model doesn’t perform well in identifying churners (as the low recall showing above), comparing with doing nothing, the company could reduce its loss from $4,192,800 to $358,850, an over 91.44% reduction by using this model.
4. We still need to modify the current model with more churn samples since the dataset we got is highly skewed to non-churn customers, the model couldn’t learn enough attributes of potential churners, resulting in failing to accurately identify churners (high false negative, low recall).

Model Summary

# Metrics

We will evaluate the models using Accuracy, the percentage of customers we correctly predict as churn. Accuracy is one metric for evaluating models. Informally, accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition:

Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

# Analysis

Chart

Description automatically generated

The chart above shows 94.5% of customers didn’t churn, while 4975 people churned. The default accuracy would be the minority case that customers churn which is 5.5%.

Chart, bar chart

Description automatically generated

The chart above shows network speed does not affect the probability of a customer churn. Numbers of customers who use 4Glite and 5G are approximately the same and the churn rate within each category are almost the same as well, which means network speed is not a useful predictor to predict whether a customer will churn.

Chart, bar chart

Description automatically generated

The chart above shows the churn status of customers using different phone models. As we can see, customers who are using Samsung Galaxy S20 Ultra, Samsung Galaxy S10e, OnePlus 8 Pro, iPhone XS, and iPhone SE 2020 seem more likely to churn. But it might because there are lots of people using these phone models (or say, these are just popular phone models). This doesn’t imply phone model is related to customer churn status, in fact, when building a logistic regression or a classification tree, phone model is not a statistically significant variable to predict whether a customer will churn.

Chart, bar chart

Description automatically generated

The chart above shows customer churn status within different billing options. As we can see, customers who do not use paperless billing are more likely to churn.

# Predictor Selection

When building the model, we chose total\_billed, number\_phones, paperless\_billing, streaming\_minutes, streaming\_plan, prev\_balance, monthly\_minutes, late\_payments, and payment\_method as useful predictive variables since these variables perform well in the classification tree model or are statistically significant in the logistic regression model. Furthermore, these variables make sense when considering business logics. For example, if total\_billed is high, meaning that customers frequently use the service, which implies that they are less likely to churn.

Here are some graphs showing customer churn status with different predictors.

Chart, histogram

Description automatically generatedChart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generatedChart, histogram

Description automatically generated

# Data Dictionary

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Definition | Keep | Key |
| email\_domain | Email suffix | Ignore |  |
| phone\_model | Model of customers’ phone | Ignore |  |
| billing\_city | Billing address city | Ignore |  |
| billing\_postal | Billing postal numbers | Ignore |  |
| billing\_state | Billing address state | Ignore |  |
| partner | Phone contract partner company | Ignore |  |
| phone\_service | Type of phone service | Ignore |  |
| multiple\_lines | If a customer has multiple lines | Ignore | Yes or No |
| streaming\_plan | Type of streaming plan |  |  |
| mobile\_hotspot | If a customer has mobile hotspot | Ignore | Yes or No |
| wifi\_calling\_text | If a customer has wifi calling text | Ignore | Yes or No |
| online\_backup | If a customer has online backup | Ignore | Yes or No |
| device\_protection | Type of device protection | Ignore |  |
| contract\_code | Contract type ID code | Ignore |  |
| currency\_code | Currency ID code | Ignore |  |
| maling\_code | Mailing ID code | Ignore |  |
| paperless\_billing | If a customer opts for paperless billing |  | Yes or No |
| payment\_method | Type of payment method |  | Bank transfer, electronic check, credit card, mailed check |
| customer\_id | Customer unique ID | Ignore |  |
| billing\_address | Billing address | Ignore |  |
| gender | gender | Ignore |  |
| network\_speed | Network speed type | Ignore | 4Glte, 5G |
| customer\_reg\_date | Customer registration date | Ignore |  |
| monthly\_minutes | Monthly usage in minutes |  |  |
| customer\_service\_calls | # of customer service calls | Ignore |  |
| streaming\_minutes | Monthly usage in streaming in minutes |  |  |
| total\_billed | Total amount of bill |  |  |
| prev\_balance | Previous balance |  |  |
| late\_payments | # of late payments |  |  |
| ip\_address\_asn | IP address ASN | Ignore |  |
| phone\_area\_code | Phone area code | Ignore |  |
| number\_phones | Number of phones |  |  |
| senior\_citizen | If a customer is senior citizen | Ignore | 0 – No  1 – Yes |
| churn | Customer attrition status |  | 0 – Not churn  1 – churn |

# Methodology

1. Data partitioning
   * Split the data into 70/30 train/test split using random sampling
2. Data preprocessing
   * Formula
     1. churn ~ total\_billed + number\_phones + paperless\_billing + streaming\_minutes + streaming\_plan + prev\_balance + monthly\_minutes + late\_payments + payment\_method
   * Numeric Predictor Pre-Processing
     1. Replaced missing numeric variables with median
     2. Centered and scaled numeric predictors to have a mean of 0 and standard deviation of 1 because we are using a KNN which is sensitive to varying scales of data.
   * Categorical Predictor Pre-Processing
     1. Replaced missing categorical variables with “unknown”
     2. Dummy encoded categories with 1s and 0s
3. Model specification
   * Train 2 K-Nearest Neighbors (KNN) with K=7 & K = 10

# Model Metrics & Evaluation

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| .estimator | part | accuracy | roc\_auc | precision | recall |
| knn = 7 | training | 0.9832783 | 0.9992377 | 0.9764706 | 0.7126503 |
| knn = 7 | testing | 0.9696381 | 0.8758128 | 0.8166828 | 0.5685348 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| .estimator | part | accuracy | roc\_auc | precision | recall |
| knn = 10 | training | 0.9791136 | 0.9981710 | 0.9604424 | 0.6462507 |
| knn = 10 | testing | 0.9707381 | 0.8880196 | 0.8546210 | 0.5557056 |

Overall, model with a k value equals 10 performs better with an accuracy of 0.97 in testing, which means we predict correctly for 97% of customer churn status. Although model with knn = 7 performs better in the training part, model with knn = 10 performs better in the testing part. Also, model with knn = 7 seems more overfitting as the differences between its training and testing results are larger than model with knn = 10.

Although both of knn models perform well in terms of accuracy, they tend to have a problem of overfitting since there are huge gaps between their training and testing outcomes.

Precision is a percentage of true positive among all predicted positive cases, while recall is a percentage of true positive among all actual positive cases. In this case, the gap between precision and recall (between their training and testing metrics) might be caused by the skewed dataset, as most of trials are non-churn. The model couldn’t learn and predict well due to a small sample of churn, which also explains the case that the accuracy is high, while the precision and recall are modest.

An ROC curve entails how capable the model is at distinguishing between results, being true negative, false negative, true positive, or false positive. As roc\_auc approaches 1, the model is more likely to predict a positive as positive, and a negative as negative. The model with knn = 10 has generated a roc\_auc of 0.888, which higher than the model with knn = 7 at a roc\_auc of 0.876 in both testing dataset. Thus, the model with knn = 10 is better at making correct predictions when predicting customer churn.

# Confusion Matrix

Chart, treemap chart

Description automatically generatedChart

Description automatically generated

Chart

Description automatically generatedChart

Description automatically generated

# Kaggle Submission

Kaggle Name: Yi Ding

Kaggle reported score: 0.97213 (Public), 0.96930 (Private)

Kaggle reported position at time of submission: #16

<https://www.kaggle.com/competitions/challenge-1-churn-2021/leaderboard>