Challenge Model Report

Challenge Overview

Customer churn, also known as customer attrition, occurs when customers stop doing business with a company or stop using a company's services. For mobile phone companies like Verizon, ATT, and Mint Mobile, customer churn will cost a considerable sum of money. Therefore, identifying the customers who are more likely to churn becomes a significant task to improve the profitability of phone companies.

Executive Summary

We built two models to predict customers who are more likely to churn and decided the KNN model is better based on model evaluations. This model has an accuracy of 0.976 and an AUC of 0.914 (further explanation is shown below).

Using this KNN model, we can successfully identify 60% of customers who churned, and by sending out vouchers to those customers who are likely to churn according to this model, we can avoid 86% of the loss.

Except that, we should dig deeper into what's happened to customers using bank transfers and 12GB stream plans. We should also encourage them to use paperless billing, spend more money on our plans, and choose 3GB or 6GB streaming plans.

Problem Statement

The challenge is creating machine learning models to predict which customers are more likely to churn.

Metrics - Accuracy

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

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

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

Metrics - Area Under the ROC Curve

We will also evaluate the models using Area Under the Curve (AUC), which represents the area under the ROC curve. For an effective model, the AUC should be higher than 0.5, which means the model is better than random guessing. The higher AUC, the better the model. So our goal is to get the AUC as closer as to 1.

Key Findings

1. 40% of bank transfer users quit.

Compared to other payment methods, bank transfer users are much more likely to churn. According to our dataset, nearly 40% of customers who uses bank transfer quit. There might be an inconvenience in using bank transfers to pay for their phone bills. Further reasons should be navigated since bank transfer is a very popular payment method.

2. churners don’t spend as much.

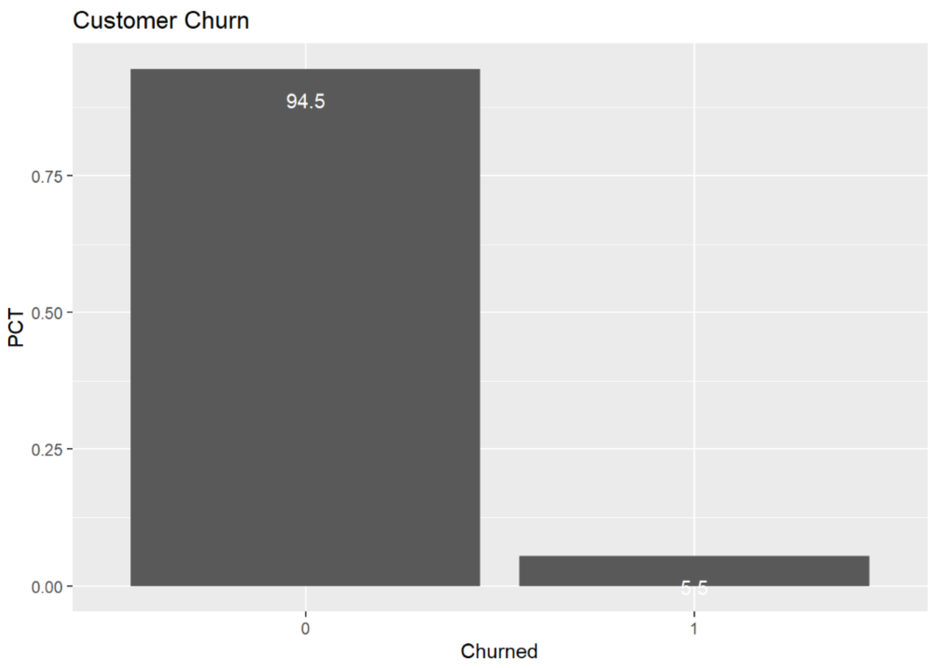
People with lower total bills and previous balances are more likely to churn, which indicates that they are not spending as much.

3. churners stream more.

Customers with a 12GB streaming plan and who spend more time streaming are more likely to churn.

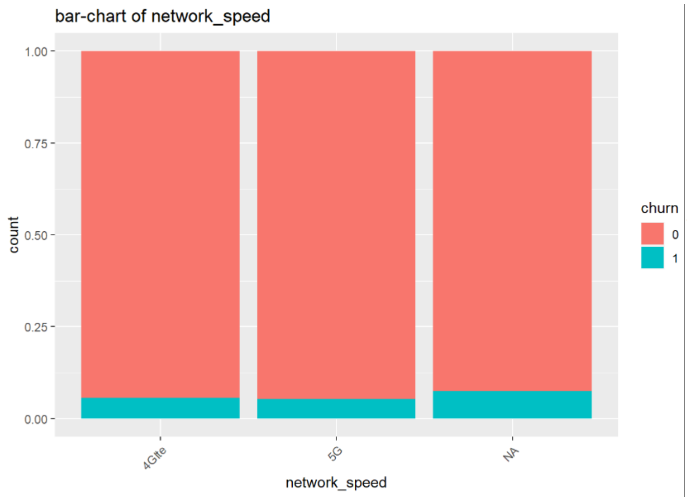
Exploratory Analysis

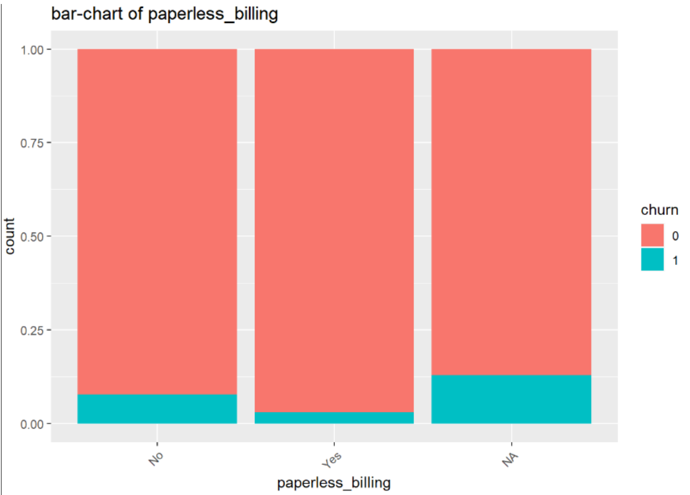
Target Exploration

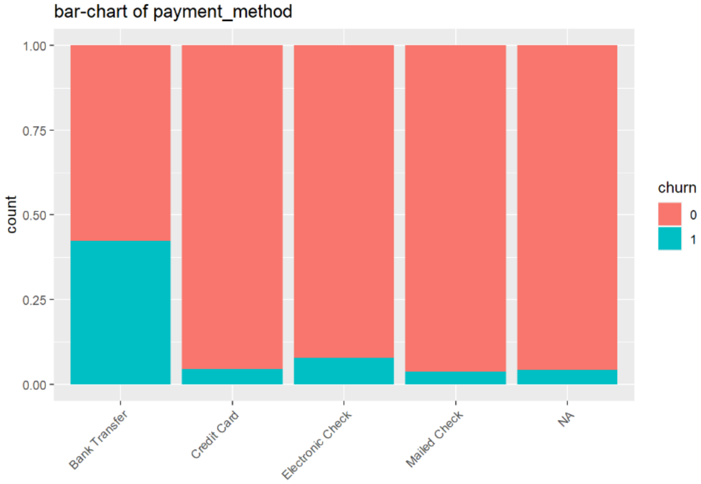


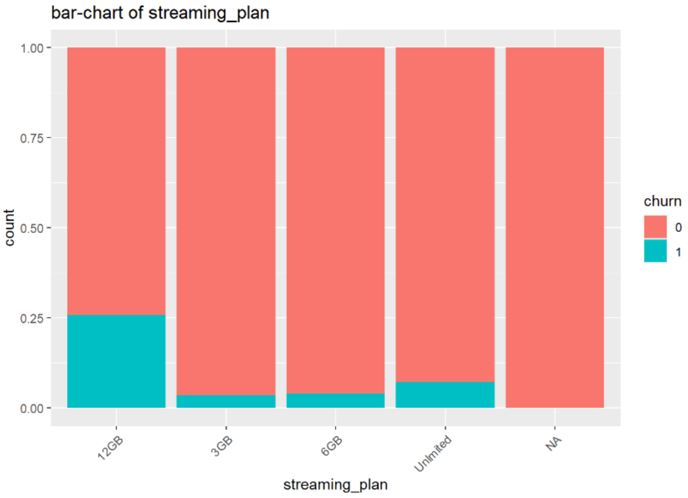
The chart above shows that 5.5% of customers churned, while 85901 people chose to stay within their current phone company. The default accuracy would be the majority case that everyone chooses to stay with their current company which is 94.5%.

Characteristic Predictors Exploration







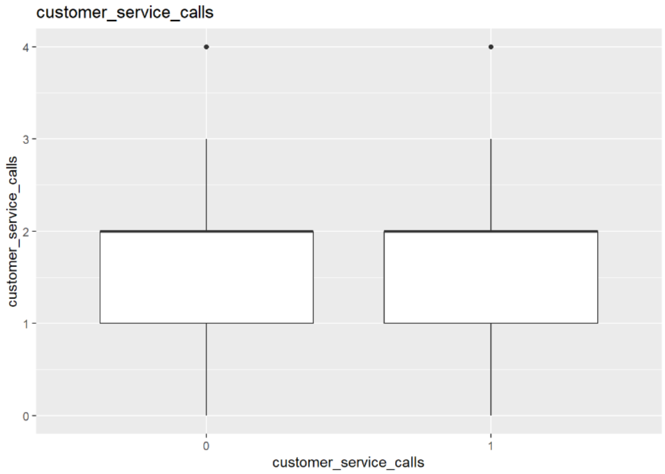


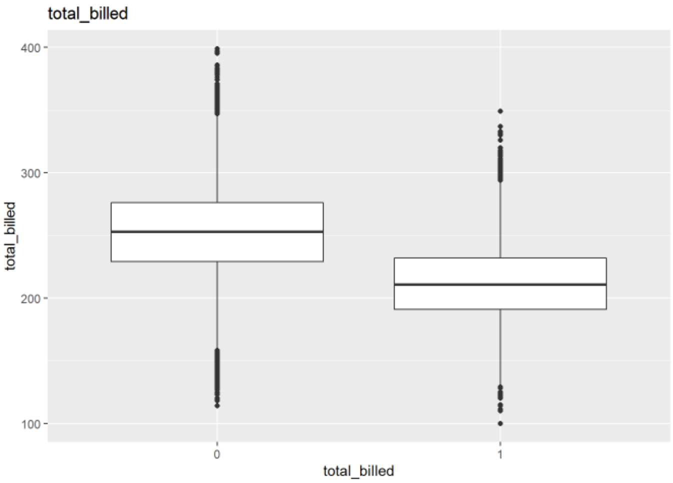
From the bar charts above, we can identify that people with a 4G or 5G network speed are equally likely to churn, as the percentage of 1 in the “yes” and the “no” bars is equal, which means network speed may not be a useful predictor.

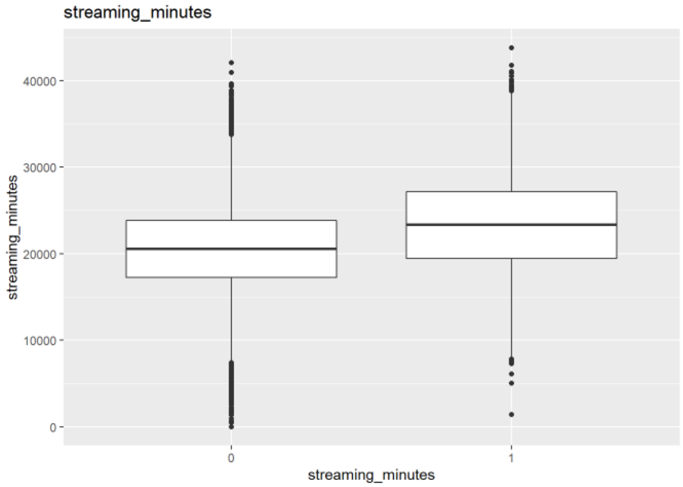
Customers who opt for paperless billing are less likely to churn, as the percentage of 1 in the “yes” bar is lower than the “no” bar in the paperless_billing chart, which indicates that paperless billing could be an effective predictor.

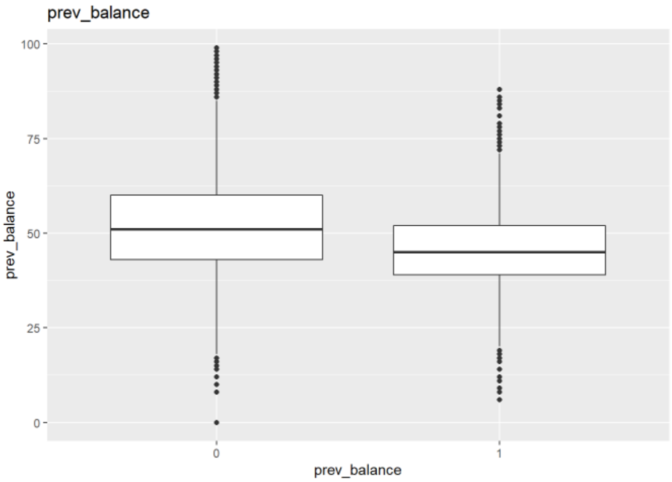
The same logic may apply to all characteristic variables as well, and this step will help us identify useful variables, including payment methods and streaming plans. Apparently, people using bank transfers and 12GB streaming plans are more likely to churn.

Numeric Predictor Exploration









From the boxplot above, we can see that the distributions of customer service calls from people churned and not churned are identical. This means that the number of customer service calls is irrelevant to churn, for those who churned and didn't churn tend to make the same amount of service calls.

However, we can see that people who didn't churn have a higher median total bill, which means that the total bill will likely be a useful predictor.

The same logic may apply to all numeric variables as well, and this step will help us identify useful variables, including streaming minutes and previous balance. Apparently, people with lower streaming minutes and higher previous balances will be more likely to churn.

Data Dictionary

Variable	Definition	Keep	Key
customer_id	unique customer id	ignore	
churn	Whether the customer switched phone company		0 = Not Churn, 1 = Churn

Billing_address	Customer’s billing address	ignore	
Ip_address_asn	Customer’s ip address	ignore	
Gender	Sex		Male, female
Monthly_minutes	Monthly minutes using the phone		
Customer_service_calls	# of service calls customer made		
Streaming_minutes	Streaming minutes		
Total_billed	Total bill		
Prev_balance	Previous balance		
Late_payments	Late payments made		
Phone_area_code	The code of phone area		
Number_phones	# of phones customer owns		
Senior_citizen	Whether the customer is senior		Yes, No
Email_domain	Customer’s email domain		Gmail, Hotmail, Yahoo
Phone_model	Customer’s phone model		
Billing_city	The city on the customer’s billing address	Ignore	
Billing_postal	The postal code of the customer’s billing address		
Billing_state	The state of the customer’s billing address	ignore	
Partner	Whether the customer is a partner		Yes, No
Phone_service	Whether the customer has phone service		Yes, No
Multiple_lines	Whether the customer uses multiple lines		Yes, No
Streaming_plan	The streaming plan customer uses		3GB, 6GB, 9GB, Unlimited
Mobile_hotspot	Whether the customer use hotspot		Yes, No
Wifi_calling_text	Whether the customer uses wifi calling texts		Yes, No
Online_backup	Whether the customer uses online backup		Yes, No
Device_protection	The kind of device protection the customer use		From A to Z
Contract_code	The contact the customer use		From A to Z
Currency_code	The currency the customer use		USD, EUR, CAD
Maling_code	The code of maling		From A to Z
Paperless_billing	Whether the customer adopted paperless billing		Yes, No
Payment_method	The payment method the customer use		Bank Transfer, Credit Card, Mailed Check, Electronic Check.
Network_speed	The network speed the customer chooses		5G, 4GLTE
Customer_reg_date	Customer’s registration date		As dates

In this case, we are ignoring variables that are likely passenger identifiers and those variables unique for each customer, including customer_id, billing_address, and ip_address_asn, and churn is our response variable.

Other than that, we ignore billing_city and billing_state, as the information is contained in billing_postal.

Methodology

1. Data partitioning
   - Split the data into 70/30 train/test split using random sampling
2. Data preprocessing
   - Formula
     - i. Churn ~ monthly_minutes + number_phones + prev_balance + streaming_minutes + phone_area_code + total_billed + email_domain + streaming_plan + mobile_hotspot + paperless_billing + payment_method + wifi_calling_text + multiple_lines + partner + phone_service
   - Numeric Predictor Pre-Processing
     - i. Replaced missing numeric variables with the median
   - Categorical Predictor Pre-Processing
     - i. Replaced missing categorical variables with the mode
     - ii. Dummy encoded categories with 1s and 0s
3. Model specification
   - Train a K-Nearest Neighbors(KNN) with K = 10
   - Train a Logistic Regression model

Model Metrics & Evaluation

Model Summary

model	part	accuracy	roc_auc	precision	recall
KNN model **With K =10**	training	0.983	0.999	0.980	0.698
	testing	0.976	0.914	0.924	0.608

model	part	accuracy	roc_auc	Precision	recall
Logistic **Regression**	training	0.965	0.921	0.809	0.463
	testing	0.965	0.925	0.800	0.483

Judging from the model summary above, we believe that the KNN model is better. The KNN model has higher accuracy, precision, and recall. Even though the AUC is a little bit below the AUC of the Logistic Regression model, it is still very close to the best one and very close to 1.

KNN model has an accuracy of 97.6%, which means that 97.6% of the total customers are identified correctly, and much higher that the default accuracy, of 94.5%.

Precision is TP/(TP+FP), Where TP = True Positives, FP = False Positive. In other words, precision measures how many customers actually churned in those who are predicted to churn. In this case, 92.4% of customers churned in those we predicted to churn.

Recall is TP/(TP+FN), Where TP = True Positives, FN = False Negatives. In other words, recall measures how many actually churned customers are successfully identified. In this case, 60.8% of customers who actually churned are identified using our KNN model.

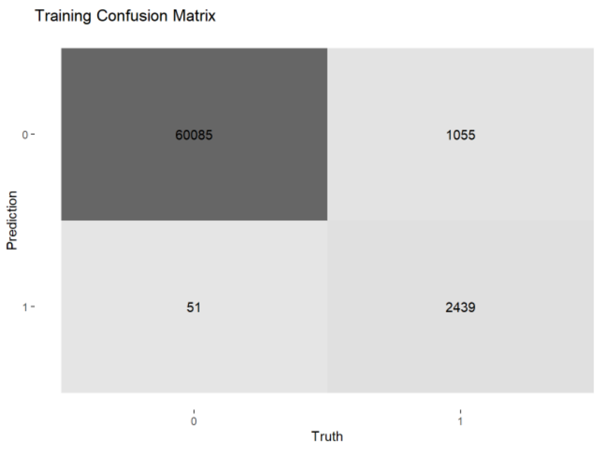
Using our KNN model, the loss we can expect is $249200 (901*500-74*50-580*1200). If we do nothing, the loss we can expect is $1777200 (901*1200+580*1200). Therefore, we can avoid 86% of the loss using the KNN model.

TP = 901, with profit of $500. FP = 74, with a loss of $50. FN = 580, with a loss of $1200.

The training data set is used to build the model while the testing dataset is used to evaluate the validity of the model. The metrics difference between those two is from other influencers which did not include in the model and some random errors.

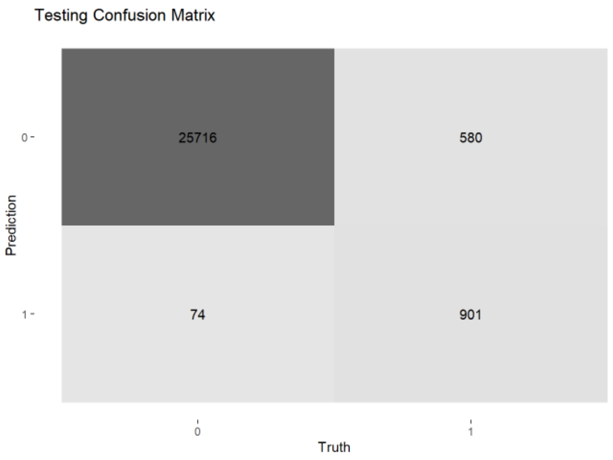
Confusion Matrices

KNN model Matrices



Training Confusion Matrix

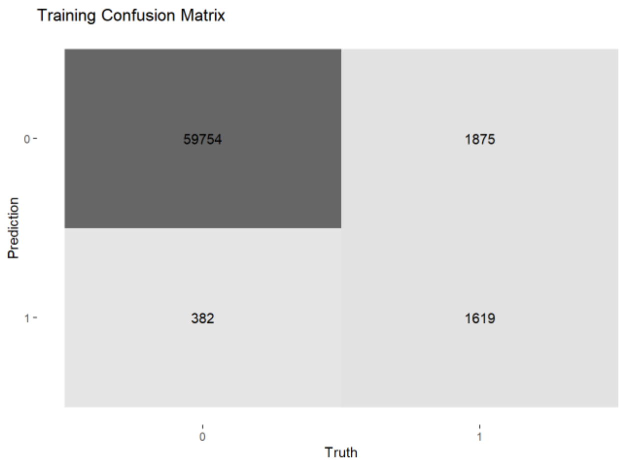
	0	1
0	60085	1055
1	51	2439



Testing Confusion Matrix

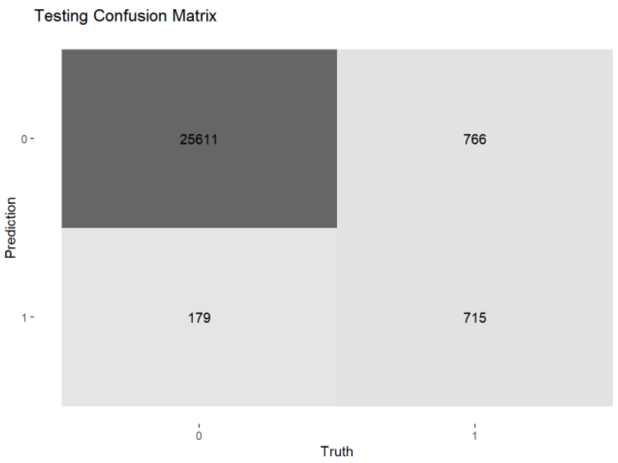
	0	1
0	25716	580
1	74	901

Logistic Regression model Matrices



Training Confusion Matrix

	0	1
0	59754	1875
1	382	1619



Testing Confusion Matrix

	0	1
0	25611	766
1	179	715

Recommendation

Key Business Questions

1. Does network speed influence churn?

According to the bar chart above, and our models, network speed doesn't have an influence on churn.

2. Does phone model influence churn?

Even though the bar chart showed that among customers with different phone models, the likelihood of churning is different, which indicates that the phone model could be useful. However, during our modeling process, we detected that the estimate of phone_model in logistic regression is not significant, and after deleting phone_model from the KNN model, the metrics of the KNN model didn't decrease. Therefore, the phone model does not influence churn.

3. Does paperless billing influence churn?

Yes, whether the customer opts for paperless billing influence churn. Those who use paperless billing are less likely to churn.

Actionable Recommendations

1. send $50 vouchers to those identified to churn based on the KNN model
As calculated above, this will help us avoid 86% of the loss.

2. encourage paperless billing

As stated before, those customers with paperless billing are less likely to churn. Therefore, encouraging the use of paperless billings will be helpful to keep our customers stay. We can achieve this by sending extra discounts or credits to those customers without a paperless billing service in exchange for them choosing this service.

3. encourage 3GB & 6GB streaming plan

Customers with a 12GB streaming plan are more likely to churn. Except for figuring out the reason, we can advertise and prompt our 3GB, 6GB, and unlimited plans.

4. encourage spending more for those who haven't

Customers with lower previous balances and total bills are more likely to churn. We can send out promotions to those who have a lower bill to incentivize them to spend more money.

Kaggle Submission

Kaggle Name: Emma Wang (Jiawen)

Kaggle reported a score: of 0.97524

Kaggle reported the position at the time of submission: #13

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