Predictive Analysis of the Churning of the Customers in the Telecom Service Industry

Hardik Prabhu

Chennai Mathematical Institute
October 12, 2020

Abstract

Customer churn occurs when a customer or subscriber stops doing business with a company or service. It is believed that the prior service data of the customer is often reflective of the churning of the customer. With this project, we aim at creating a predictive model to identify if a particular customer is at risk of churning. We begin by considering the churning of a customer as a binary classification problem. For a company that sells monthly plans, we look at who's at risk of canceling now, $Y = 1$(Churn), based on last month's usage. For the customer at risk, the company can offer a loyalty bonus upfront like discounts, upgrades, etc. This remains a business decision and not the focal point of our project. In our project, based on the customer profile and the service usage data, we figure out the driving factors responsible for churning and also come up with a model for predicting whether a customer will churn or not. To achieve our goals, first, we do an exploratory data analysis, followed by logistic regression for binary classification.

1 Introduction

Customer retention is one of the most important metrics for a growing business. Not only, the existing customers have better conversion rate, but once a customer becomes a churn, the loss incurred by the company is not just the lost revenue due to the lost customer but also the costs involved in additional marketing in order to attract new customer. Reducing customer churn is a key business goal of every business.

Over the last few decades, the telecom industry has witnessed enormous developmental changes in terms of an increase in the level of competition and the competitors, opening to new services, and the booming technology industry. A churn of a customer severely hits the company's revenue and its marketing expenses. Predicting churning of the customer in advance provides an opportunity for cutting marketing expenses and proactive customer retention.

In this project, we look into the customer profile and service usage database of a telecom company. Based on last month's data of a customer, we answer the following questions.

1. Which variables influences whether the client will leave?
2. Which clients are likely to leave? (classified as churn)

2 Data set

In this project, we use the **Telco Customer Churn dataset** to study customer behavior prior to churning. This data set has 7043 samples and 21 features, including both categorical and numeric attributes. The features also include demographic information about the client like gender, age range, and if they have partners and dependents, the services that they have signed up for, account information, etc. The "Churn" column is our target. It takes binary values "YES" and "NO".

This data set is publicly available on kaggle.

2.1 Data preprocesssing

After inspection, we find that the data has no entry with missing values. Is our data error-free? No, the datatype for ”TotalCharges” is string instead of float. The conversion from string to float is not possible because some entries contain an empty string. All such entries have the value for tenure = 0. It implies that the customers had just adopted the service and were not charged yet. So, we appropriately replace the empty strings with 0.

3 Exploratory Data Analysis

Before doing some exploratory analysis, the data set is first split into training and testing, we split it as 70:30, that is 70% training and 30% is kept for model evaluation. The exploratory data analysis is being done on the training data. The training data has

3.1 Demographics

3.1.1 Gender

We look at the distribution of the churn based on the gender of the customer.

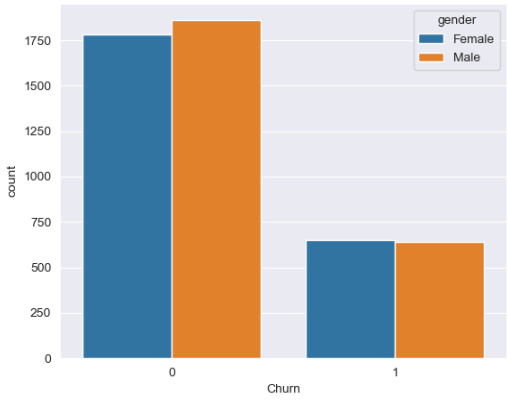

Figure 1: Churn distribution based on gender. On the Y axis we have the counts, on the X axis 0, 1 for churn and not churn respectively.

The bar graph plot indicates that gender doesn't have much impact on churning. The ratio of male to female in both, churn and not churn is similar.

3.1.2 Senior citizen

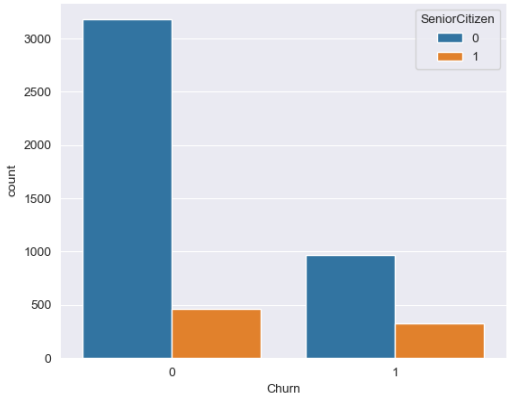

Figure 2: Churn distribution Based on age group. On the Y axis we have the counts, on the X axis 0, 1 for churn and not churn respectively.

The plot indicates that a person who is not old is less likely to terminate the service.

3.2 Service charges, tenure and contract length

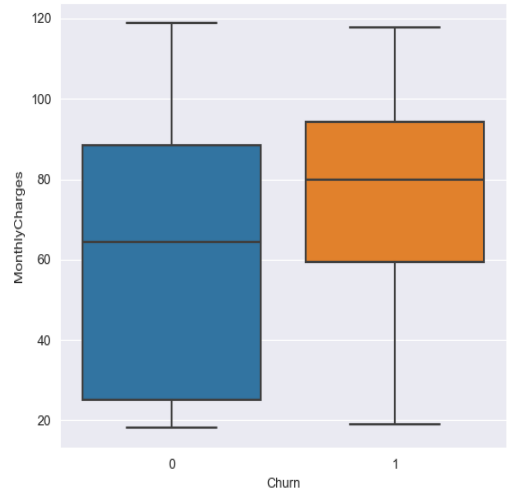

Figure 3: Box plot for Monthly Charges. It captures the spread of distribution. The line in the middle indicates the mean value. The box encloses all the values lying within one standard deviation from the mean.

The box plot indicates, the higher the monthly charges, the more likely the customer will churn. On the other hand, the total charges don't indicate the same. It could be due to the fact that the total charges also depend on the tenure. The total charges could be accumulated over the years, and as we will see that, the tenure is a good measure of loyalty of the customer and hence less likely to churn.

Figure 5 suggests that the period of tenure could be one of the major factor for determining the churn.

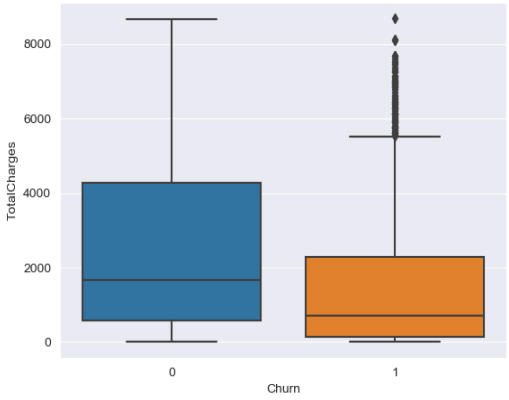

Figure 4: Box plot for Total Charges

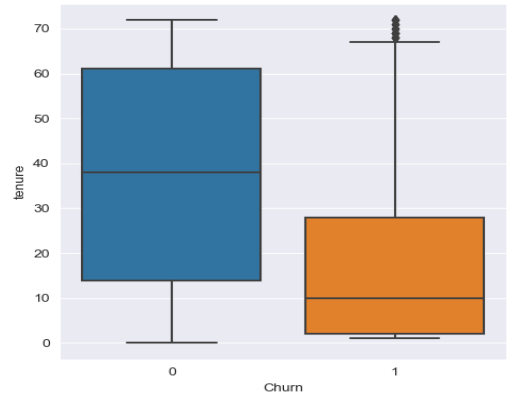

Figure 5: Box plot for Tenure

Another indication of customer loyality could be the commitment in terms of contract length. **Figure 6** indicates that customers who commit to month to month plans are more likely to churn.

3.3 Type of service

3.3.1 Phone Service

Over 90% of customers have opted for phone service. About 47% of them have opted for service on multiple-lines. The proportion of churned customer is around 25% for both, with and without multiple-lines.

3.3.2 Internet Service

Only about 22% of the total customers don't opt for internet service. **Figure 7** shows the distribution of the churn among the service users.
We can clearly see that the quality of internet service has a major impact on churning of customers.

- **Fiber optic:** Customers opting for fibre optic cable are more likely to churn.
- **Online Security:** Customers not opting for online security are more likely to churn.

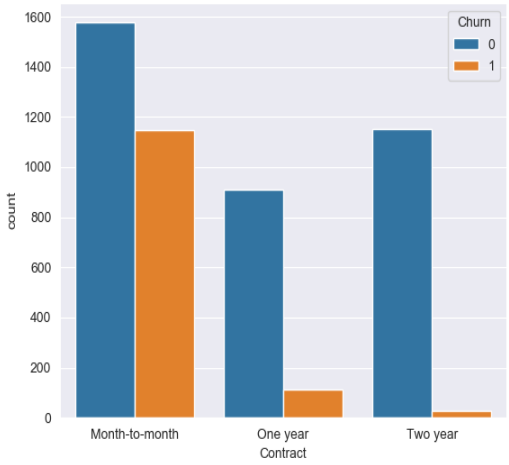

Figure 6: Contract length and churn. On the Y axis we have the counts, on the X axis we have the different contract lengths. Orange and blue hues for churn, not churn respectively.

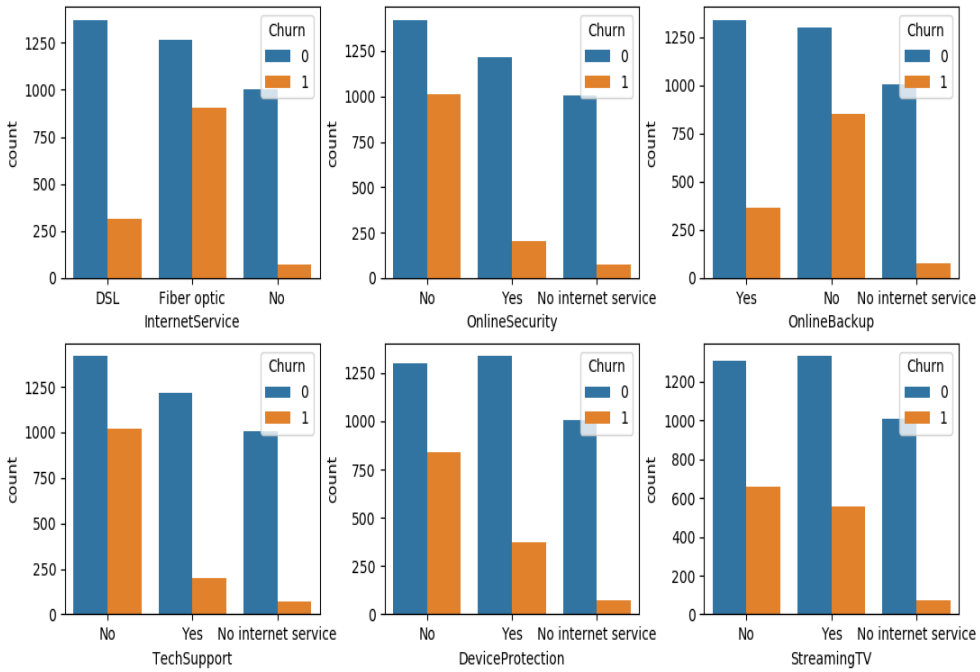

Figure 7: Plots Of the churn counts for the types of the internet service provided.

- **Tech Support:** Customers not opting for tech support are more likely to churn.
- **Device Protection:** Customers not opting for device support are more likely to churn.
- **Online Backup:** Customers not opting for online backup are more likely to churn.

3.4 Partners and dependents

Partners and dependents are two categorical variables with values ”Yes” and ”No”.

We find that a customer with no partners and no dependents is more likely to churn. To illustrate this, we engineer a new feature ”close contact” which is a categorical variable with 3 degrees. Degree 0: No dependents and No partners. Degree 1: Either dependents or partners. Degree 2: Both dependents and partners.

4 Models

4.1 First Model

We create our logistic regression model by selecting the relevant features found during the exploratory data analysis.

4.1.1 Model Features

The features selected for our model are listed below.
Numeric: SeniorCitizen(0-1 binary), Tenure (float), MonthlyCharges(float), TotalCharges (float), closecontact(0,1,2)
Categorical: InternetService, OnlineSecurity, OnlineBackup , DeviceProtection, TechSupport

We perform one-hot encoding for categorical variables.

4.1.2 Addresing multicolinearity

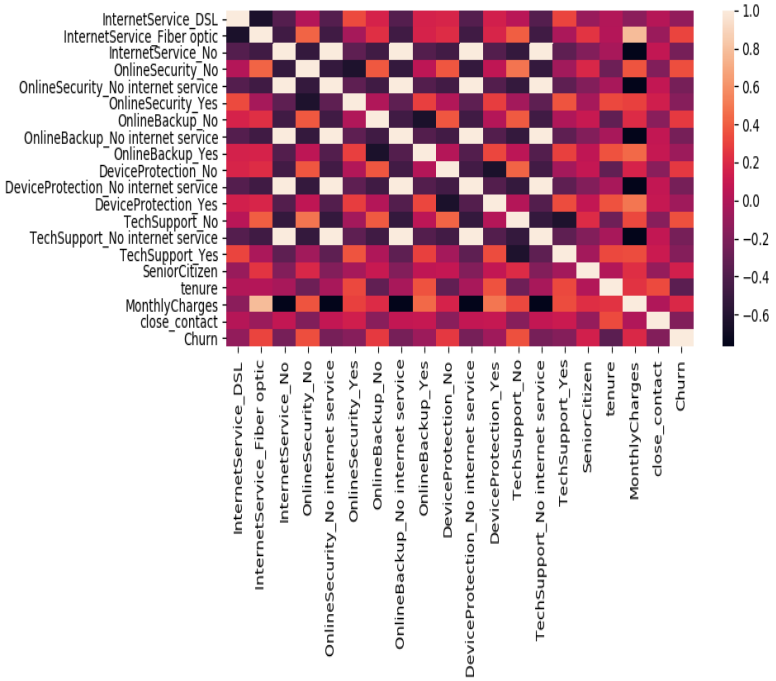

Figure 8: Correlation coeffiecient heatmap between different features. All the off-diagonal 1s indicate multicolinear relationship between our features.

We first look at the correlation between the variables. We want to avoid high multicollinearity. Correlation coefficient value above 0.7 and below -0.7 may be concerning. We will also look at the respective VIF values. From taking

a glance at figure 8 we can clearly see the issue of multicollinearity. Some of the variables are redundant and we can remove them easily. For example, If a person has not opted for internet service, then the "OnlineSecurity-No internet service" variable will be 0 (No) as well. We appropriately make some changes and then see the VIF values.

VIF Factor	**Features**
6.621942	InternetService-DSL
22.186879	InternetService-Fiber optic
1.945856	OnlineSecurity-Yes
2.088946	OnlineBackup-Yes
2.200815	DeviceProtection-Yes
2.020580	TechSupport-Yes
1.304932	SeniorCitizen
4.270455	tenure
41.351913	MonthlyCharges
2.103782	close-contact

The vif values for monthly charges and InternetService-Fibre optic is high. We can clearly see the high correlation between the two by comparing mean of monthly charges with respect to internet service. We drop the monthly charges because of the high VIF factor.

InternetService	**Monthly Charges** (avg.)
DSL	58.592191
Fiber optic	91.529086
No	21.118611

4.1.3 Model Summary

Logit Regression Results

Dep. Variable:	Churn	No. Observations:	4930
Model:	Logit	Df Residuals:	4920
Method:	MLE	Df Model:	9
Date:	Tue, 22 Dec 2020	Pseudo R-squ.:	0.2588
Time:	20:43:46	Log-Likelihood:	-2099.6
converged:	True	LL-Null:	-2832.7
Covariance Type:	non robust	LLR p-value:	3.957e-310

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.6999	0.130	-13.064	0.000	-1.955	-1.445
InternetService_DSL	1.6220	0.150	10.804	0.000	1.328	1.916
InternetService_Fiber optic	2.7906	0.143	19.549	0.000	2.511	3.070
OnlineSecurity_Yes	-0.5079	0.099	-5.113	0.000	-0.703	-0.313
OnlineBackup_Yes	-0.0794	0.090	-0.882	0.378	-0.256	0.097
DeviceProtection_Yes	-0.0332	0.090	-0.368	0.713	-0.210	0.144
TechSupport_Yes	-0.5962	0.100	-5.985	0.000	-0.791	-0.401
SeniorCitizen	0.3544	0.098	3.598	0.000	0.161	0.547
tenure	-0.0385	0.002	-16.983	0.000	-0.043	-0.034
close_contact	-0.0881	0.053	-1.678	0.093	-0.191	0.015

4.1.4 Statistical Inference

For each input, the model gets a score(Z) between 0 and 1. Z is interpreted as the probabilty of churn for the given input(customer).

$$log\frac{Z}{1-Z} = -1.66 + 1.6DSL + 2.7Fiberpotic - 0.5OnlineSecurity - 0.07OnlineBackup$$

$$-0.03DeviceProtection - 0.58TechSupport + 0.3SeniorCitizen - 0.03Tenure - 0.08CloseContact$$

By looking at P values for features, we can say that any feature is statistically significant if P value < 0.05. Features such as InternetService, SeniorCitizen has a postive impact on churn, while features such as OnlineSecurity, Tenure, Tech Support has negative impact on the churn. We also learned that the major factor behind the rise in the monthly cost is that Internet service, especially Fibreoptics is expensive.

4.2 Full Model

We create our 2nd logistic regression model by selecting from all the available features. The categorical are converted to numeric by introducing dummy variables. we iteratively calculate vif values for features, and remove the features with very high vif values. We repeat this process till all the vif values are less than 5.

Logit Regression Results

Dep. Variable:	Churn	No. Observations:	4930
Model:	Logit	Df Residuals:	4908
Method:	MLE	Df Model:	21
Date:	Thu, 23 Dec 2020	Pseudo R-squ.:	0.2891
Time:	13:43:53	Log-Likelihood:	-2013.8
converged:	True	LL-Null:	-2832.7
Covariance Type:	non robust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.2019	0.183	-1.101	0.271	-0.561	0.158
SeniorCitizen	0.1634	0.102	1.606	0.108	-0.036	0.363
tenure	-0.0340	0.003	-11.749	0.000	-0.040	-0.028
gender_Male	-0.0302	0.078	-0.385	0.700	-0.184	0.123
Partner_Yes	0.0330	0.095	0.348	0.728	-0.153	0.219
Dependents_Yes	-0.1569	0.109	-1.442	0.149	-0.370	0.056
PhoneService_Yes	-0.5008	0.156	-3.220	0.001	-0.806	-0.196
MultipleLines_Yes	0.3219	0.094	3.432	0.001	0.138	0.506
InternetService_Fiber optic	0.9164	0.109	8.383	0.000	0.702	1.131
InternetService_No	-0.8851	0.167	-5.313	0.000	-1.212	-0.559
OnlineSecurity_Yes	-0.3580	0.102	-3.504	0.000	-0.558	-0.158
OnlineBackup_Yes	-0.1052	0.092	-1.142	0.253	-0.286	0.075
DeviceProtection_Yes	-0.0358	0.094	-0.380	0.704	-0.220	0.149
TechSupport_Yes	-0.4511	0.104	-4.329	0.000	-0.655	-0.247
StreamingTV_Yes	0.2808	0.096	2.919	0.004	0.092	0.469
StreamingMovies_Yes	0.2868	0.096	3.000	0.003	0.099	0.474
Contract_One year	-0.6440	0.130	-4.957	0.000	-0.899	-0.389
Contract_Two year	-1.4690	0.223	-6.592	0.000	-1.906	-1.032
PaperlessBilling_Yes	0.2951	0.090	3.295	0.001	0.120	0.471
Payment_Credit card	-0.1176	0.138	-0.851	0.395	-0.389	0.153
Payment_Electronic check	0.3467	0.113	3.060	0.002	0.125	0.569
Payment_Mailed check	0.0089	0.137	0.065	0.948	-0.260	0.278

4.2.1 Statistical Inference

For each input, the model get a score(Z) between 0 and 1. Z is interpreted as the probability of churn for the given input (customer). By looking at P values for features, we can say that a feature is statistically significant if $p < 0.05$.

The statistically significant features that positively impacts churning are given below:
1. Multiple lines 2. Fibre Optic 3.Streaming TV 4. Streaming Movies 5. Paperless Billing 6. Electronic Check.

The statistically significant features that negatively impacts churning are given below:
1. Tenure 2. Phone Service 3. No Internet Service 4. Online Security 5. Tech Support 6. Yearly Contracts

4.3 Evaluation of the predictions made over the test data

The model was trained on 4930 data points. We use the remaining 2113 data points for testing. First, we get the probability of churn over each test point by using the model. We then convert those to binary 0,1 based on setting a probability threshold. The customers with probabilities less than the threshold are labelled 0 (Not Churn) and above the threshold are labelled 1 (Churn).

The metrics used for evaluating our models are Precision, Recall, F-score, and Accuracy. For varying values of probability thresholds, we plot the curves of the evaluation metrics.

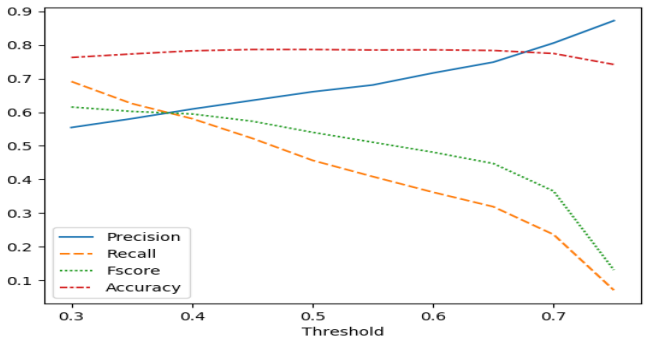

Figure 9: Model evaluation based on varying probability thresholds (between 0.3 and 0.8) for our first logistic regression model. As the threshold value increases, the precision increase and recall decrease. The F-score is the harmonic mean between the the precision and the recall.

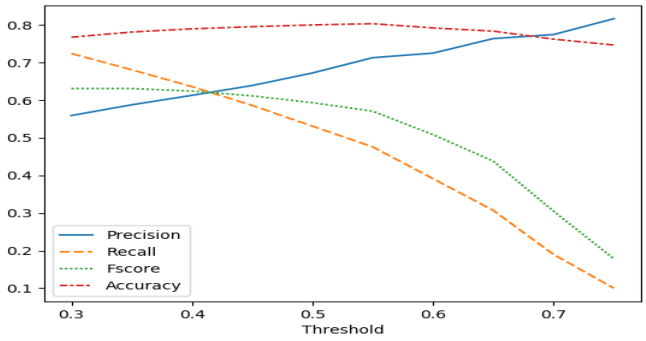

Figure 10: Model evaluation based on varying probability thresholds(between 0.3 and 0.8) for the final logistic regression model.

5 Conclusion

Both the models perform well on the training data with around 0.8 accuracies. The full model is slightly better in terms of performance. Based on business needs, the probability threshold value can be adjusted to either have high precision or recall (between 0.3 and 0.8).
In conclusion, the quality of internet service is a major contributor to churning. Fiber optics is very expensive and increases the monthly charges. The streaming quality further dissatisfies the customer. Long term customers are less likely to churn while customers with monthly contracts are more prone to churn. Older citizens are less likely to churn, it could be because they are reluctant to opt for internet services.

References

Hastie, T., Tibshirani, R., Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer.

James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert (2017). An Introduction to Statistical Learning (8th ed.). Springer Science+Business Media New York. ISBN 978-1-4614-7138-7.