DATS 6501-Data Science Capstone
Aluya Omofuma

Evaluating Churn Models for Telecom

Introduction

Customer Churn (or customer attrition) refers to the loss of clients or consumers of a product/service. “The churn rate, also known as the rate of attrition or customer churn, is the rate at which customers stop doing business with an entity. It is most commonly expressed as the percentage of service subscribers who discontinue their subscriptions within a given time period.” (Investopedia) Churn rate is a key performance metric in service companies because executives in service companies are beginning to see the value of retaining their customers by providing quality service.. “Customer defections have a surprisingly powerful impact on the bottom line. They can have more to do with a service company’s profits than scale, market share, unit costs, and many other factors usually associated with competitive advantage...Companies can boost profits by almost 100% by retaining just 5% more of their customers.” (HBR) The nature and magnitude of service companies has changed radically in the digital age which creates a unique challenge for companies. They need to create models that inform them when a customer is likely to churn. Customer retention leads to more financial success and customers now have more options and information available to them i.e. customers are exposed to similar products offering similar perks at a high frequency, companies pay a great deal of attention and invest a lot in customer retention programs in order to gain competitive advantage in the market. Given that companies now collect information about their individual clients they can monitor their behavior i.e. how they use the service and check against important variables to determine if the

customer is displaying behavior that is in line with historical information about customers who have churned in the past.

The purpose of this study is to analyze and evaluate the effectiveness of churn prediction models and determine how they can be used to inform customer retention programs. This study will focus on Machine learning algorithms and a statistical model, Survival Analysis to predict churn in customers of a telecom company.

Literature Review

This segment focuses on the related studies in the field of customer attrition in various service industries like telecom companies, banking and mobile services. It is important because we will investigate modelling techniques used and their effectiveness in predicting churn.

Keramati used the decision tree algorithm to develop a prediction model that identified features of churners using data for electronic banking services. Due to a dearth of data in the bank database they used a small dimensional dataset consisting of variables like customer dissatisfaction, service usage and other customer related variables. The feature determinations are done by using the technique of forwarding selection and backward elimination. An exactness of 85% is acquired with this model when the model is presented to the set number of data. They used the CRISP-DM to build their model and validated using the receiver operating characteristic curve. The bank's database imposed some limitations on the study where they could examine only the factors that were recorded in the bank's database.

Churn analysis has been used in telecom and they made use of varying machine learning algorithms Naive bayes, Logistic regression, Decision Tree and an artificial neural network(multilayer perceptron). Their data was obtained from 5 service providers and they used the rapid miner simulator to build the model. In preprocessing, the variables are changed to numeric types for predictive analysis. Their goal was to determine the churn rate of customers in telecom companies, cluster them into different categories and monitor the relevant patterns in the data that had an effect on revenue and growth.

Lai and Zeng did a study focused on digital libraries to demonstrate the successful application of Survival Analysis for understanding customer churn status and relationship duration distribution between customers and libraries. They used the non-parametric method of survival analysis to discuss customer churn behavior in Digital Libraries, which include the Life Table Analysis and Kaplan-Meier method. The former is more suited to analyze the overall situation of given samples, since it summarizes terminal events during a specific period of time. The kaplan Meier method is a more popular method for estimating survival time, and is the fulcrum of this study. The study was able to show that the median survival period for the total population was less than 2 years while the probability of churn was highest within the first 3 months of subscription. The empirical analysis showed that customer churn rate of the study is very high at 65.7 percent of the 8,054 investigated customers at the point of analysis.

Methodology

In order to evaluate the models, the machine learning process was followed. This section delves into the approach used in understanding and preparing the data, building and evaluating the models.

Data Preprocessing

The data used in this study from IBM and was entered in a Kaggle competition. It is called Telco customer churn which consists of historical records of customer churn. It shows what customers churned and how long they subscribed to the business before churn. The nature of the dataset allows for censorship which will be explored in the survival analysis model. This dataset consists of 7,043 observations(subscribers) and 21 variables. For each subscriber we have information that is split between 3 numeric variables (duration of subscription, monthly and total charges), string variables and some that were converted to ordered factors or categories.

For the data preprocessing, I dropped observations that were missing total charges because from the Exploratory Data Analysis it seemed that that variable was important. Some discoveries made during EDA were the average tenure for subscribers in the dataset was 32 months and the average monthly bill was about $65. Looking a little deeper at customers who were retained we see that they had an average tenure of 37 months (over 3 years) and paid about $61 compared to the averages for churned customers who were subscribed for only 17 months and paid a higher monthly charge of about $71. We can infer that the average monthly charge played a significant role in causing customers to churn.

This telecom company provides Internet and phone services so we looked at the types of services and the internet services provided contained 3 different values; fiber optics, DSL and some other service and we can tell whether or not the subscriber in question subscribed to the phone service. Majority of the observations have fiber optic internet service and also have phone service provided by the same company which is similar for the churned customers so it's not apparent if this plays a role in the larger monthly fee of the churned customers.

Machine Learning Algorithms

We move forward without dropping any variables and we set the predictors i.e. the information about the subscribers to an array and set the target variable to a different array. In addition, one-hot encoding was used to convert the categorical variables to a form accessible for the machine learning algorithms and we also used a label encoder to set the labels to numerical values. "Since we intend to build a classification model based on learning algorithms we must train them first and then test them, and therefore we must partition the dataset in two: a training and a testing set" (Brandusoiu) We randomly partition the training set to be approximately

80% of the original data set, consisting of 5617 subscribers, and the testing set to be approximately 20%, consisting of 1405 subscribers. We also used a stratify method that ensured the training and test sets had even proportions of churned customers.

The models used for this analysis were logistic regression, decision tree, random forest, naive bayes and k-nearest neighbor. For each model we created a pipeline where the features would go

through a standard scalar which would normalize the values and then they would be fit on the model/classifier. The fit will then be tested on the test set and we can evaluate the performance.

In order to improve the performance of the model we did hyperparameter tuning using GridSearchCV. "In one line: cross-validation is the process of splitting the same dataset in K-partitions, and for each split, we search the whole grid of hyperparameters to an algorithm, in a brute force manner of trying every combination." This process takes different segments of the training and test data and switches them up to obtain an all encompassing view of the data to determine which hyperparameters are best. "In an iterative manner, we switch up the testing and training dataset in different subsets from the full dataset. We usually split the full dataset so that each testing fold has 10% (K=10) or 20% (K=5) of the full dataset...what we do for the grid search is the following; for each iteration, test all the possible combinations of hyperparameters, by fitting and scoring each combination separately."(Hanson) Hyperparamener tuning helped to improve the performance of the classifiers and we subsequently obtained the random forest as the best performing classifier. We further tested it on the confusion matrix to ensure that the results were indeed apt.

Survival Analysis

Churn prediction is useful in having an idea about if a customer is seeming likely to stop subscribing to a business, determining the point in time at which they will make this decision is valuable to a business in terms of avoiding this event from happening, so we use survival analysis to predict the survival time of a customer. "The survival time was then added to the labeled data for the training set. Survival time is defined as the period between the date of last

activity from the provided data and the date of most recent activity, which was not provided but instead calculated when all predictions were submitted." (Lee)

Survival Analysis is a statistical measure used to determine the time before an event occurs. Some important concepts to consider are censorship, survival function and hazard functions. We usually take a continuous random variable to represent the time from subscription to the time of churn for an individual customer. We obtain the survival function for each individual customer by determining the probability of a customer surviving past a determined time t. We are also interested in the rate at which churn is taking place, out of the surviving population(i.e. Customers who continued to subscribe) at any given time t. This value is the ***hazard function*** which we can simply define as the rate of churn of the customers remaining at time t.

Let $T \geq 0$ be a random variable representing the survival (or event) time. The *survival* (or *survivor*) *function* is the probability that an individual survives beyond time t,

$$S(t) = \mathbb{P}(T > t), \quad 0 < t < \infty.$$

The *probability density function* $f(t)$ is the frequency of events per unit time. The probability density function is related to the survival function,

$$f(t) = -\frac{dS(t)}{dt}.$$

The *hazard function* is the instantaneous rate at which events occur for individuals which are surviving at time t,

$$h(t) = \lim_{\delta t \to 0^+} \frac{\mathbb{P}(t \leq T < t + \delta t | T \geq t)}{\delta t}$$

and the *cumulative hazard function* is

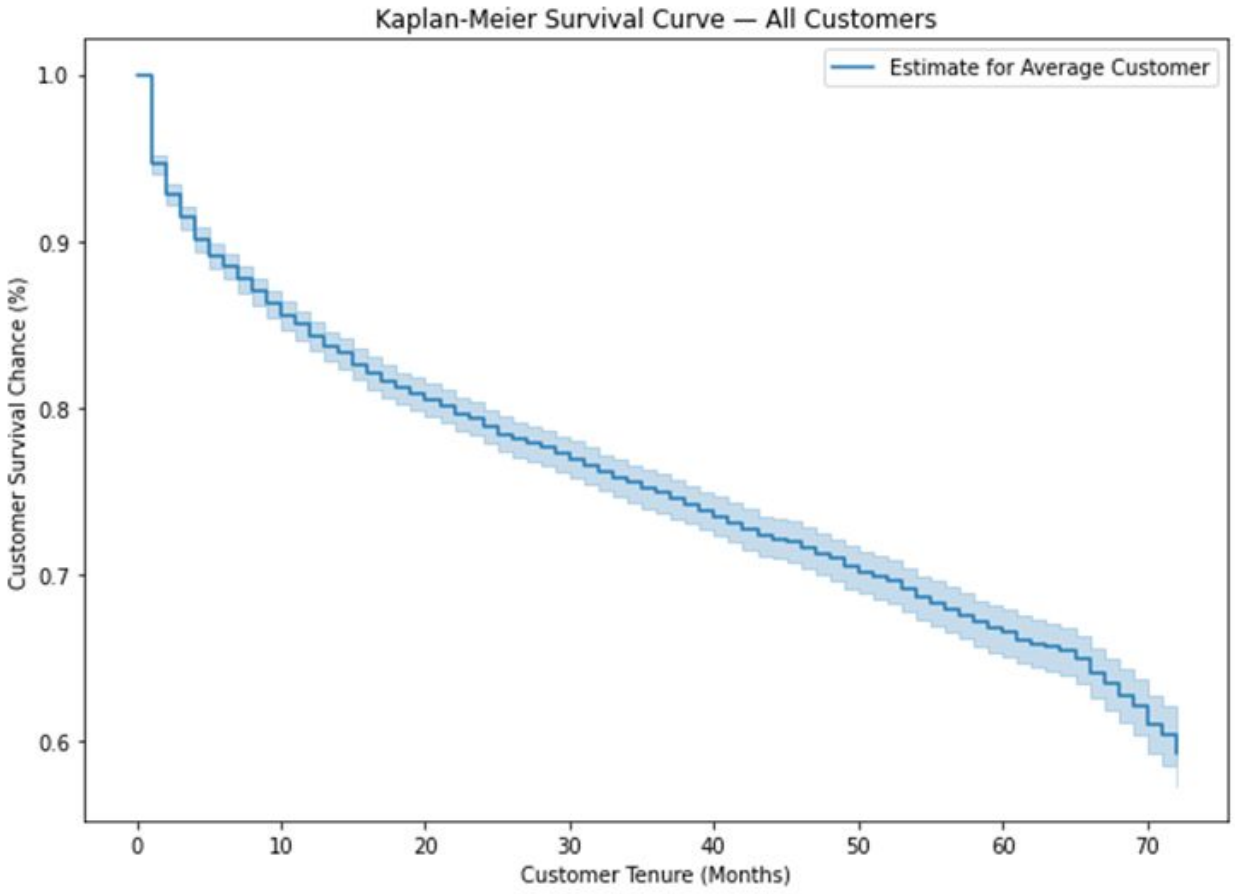
$$H(t) = \int_0^t h(u)\, du.$$

The cumulative hazard function is related to the survival function as follows:

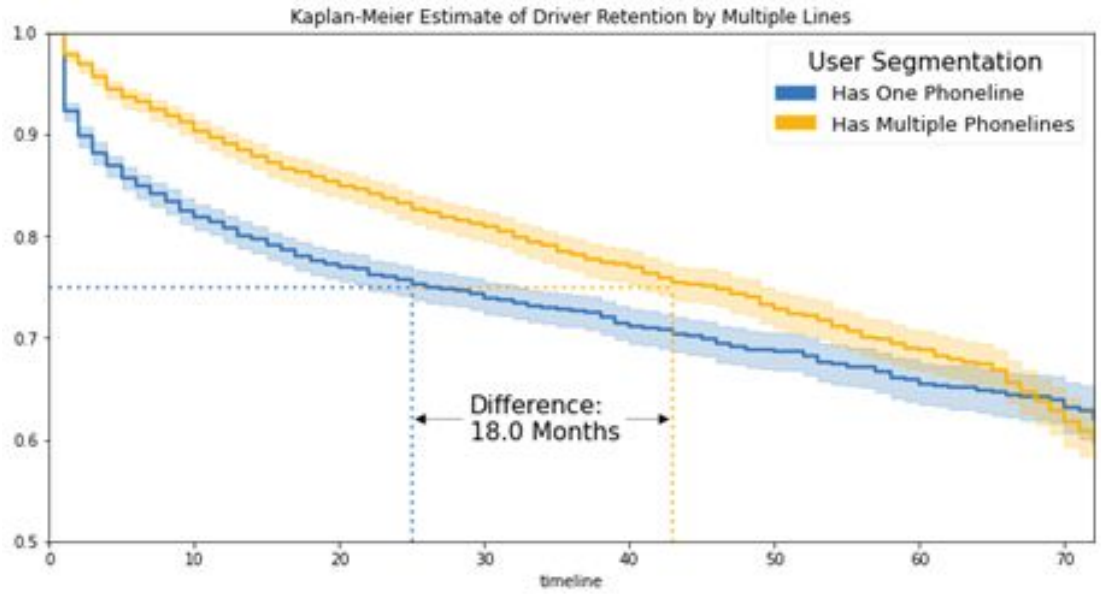
$$S(t) = e^{-H(t)}.$$

(Lee)

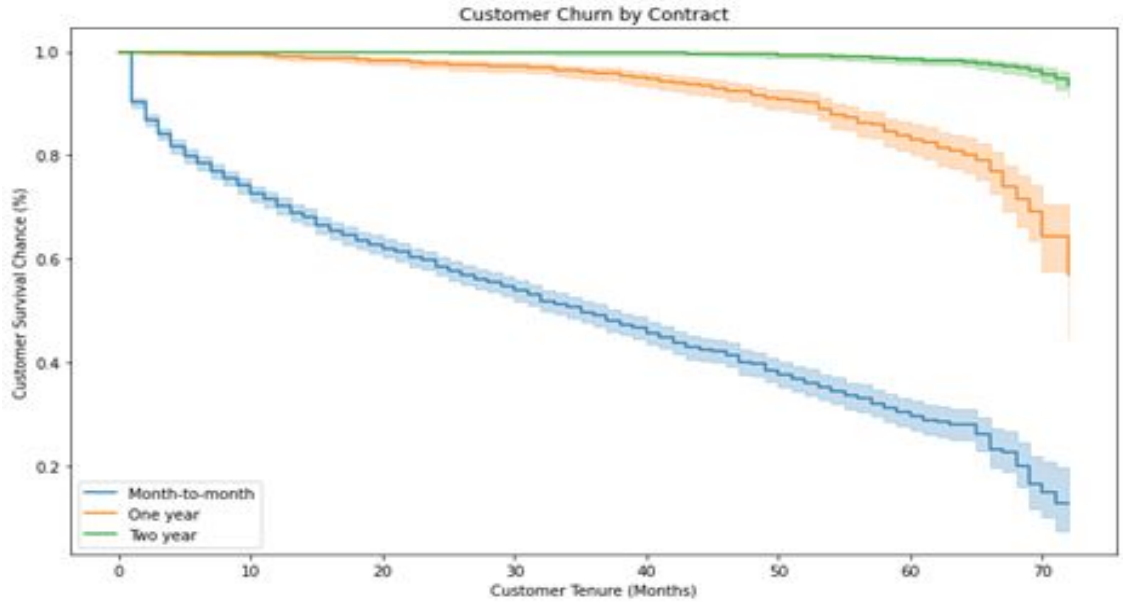
We use the KaplanMeier fitter to fit and model the survival function of all the customers in the dataset and we can see that after 32 months (the mean tenure of customers in the dataset) the survival chance of the clientele is slightly below 80%.



We can use this fitter as an exploratory tool to investigate survival rates in features that we suspect to have great effects of chun rates. For instance, In our exploratory data analysis we saw that customers who have multiple phone lines are more likely to have a long standing subscription, so we investigate further to determine the survival chances of customers with multiple phone lines.



We can see that the survival chances of customers with one line drops to 75% after about 25 months and it takes 18 more months for customers with multiple phone lines to reach the same chances of survival. We also investigate the contract types to see how the contract lengths affect a customer's chance of survival.



We can see that after about 30 months customers with month - to - month contracts have a survival chance of about 60% while customers on other contracts don;t churn at the same rate for over double that time (~65 months).

We use a parametric model, the Cox Proportional Hazard model is used to determine the impact of the features on the survival function.´It initially models the hazard function assuming the covariates have a linear multiplicative effect on the hazard function.´The idea behind the model is that the log-hazard of an individual is a linear function of their static covariates, *and* a population-level baseline hazard that changes over time.

Results and Analysis

To assess the prediction accuracy of the proposed machine learning models, the dataset is split in training and test sets with the proportion of 80:20 i.e. The model is trained using 80% of the data. The test sets are passed to the proposed model in order to test or to validate the models. The results obtained through prediction are compared with actual figures. The validation of the model is done using the parameters F1 score, recall, precision and confusion matrix to avoid overfitting or underfitting the data and biased result. “The F1 score is viewed as critical to identify the biased prediction result with the given Model. It works based on the false positive and false positive statistics of the prediction model.The equation to ascertain the F1 score is given underneath.”

F1_score = 2 * (Precision * Recall) / (Precision+Recall)

(Bilal Zorić)

The confusion matrix is another measure to evaluate the accuracy of the machine learning model. The outcome with the confusion matrix is a false negative, false positive, true positive and true negative. (Mundada)

Random Forest Confusion matrix

	Precision	Recall	f1-score	support
0	0.82	0.90	0.86	1038
1	0.61	0.46	0.52	369
micro avg	0.78	0.78	0.78	1407
macro avg	0.72	0.68	0.69	1407
weighted avg	0.77	0.78	0.77	1407

The random forest model correctly predicted non-churn customers 82% of the time and correctly predicted churned customers 61% of the time. The weighted precision score i.e. how many times the model made the right prediction is 77% which is a very useful score. The recall value shows that the model did a good job of correctly predicting loyal customers. However, in predicting churned customers, only 46% of the predictions were correct. Thus, there is room for improvement.

Decision Tree confusion Matrix

	Precision	Recall	f1-score	support
0	0.82	0.79	0.81	1038
1	0.46	0.50	0.48	369
micro avg	0.72	0.72	0.72	1407

macro avg	0.64	0.65	0.64	1407
weighted avg	0.72	0.72	0.72	1407

The random forest model correctly predicted non-churn customers 82% of the time and correctly predicted churned customers less than 50% of the time. The weighted precision score i.e. how many times the model made the right prediction is 72% which is a very useful score but is boosted by the success of predicting loyal customers. The recall value shows that the model did a good job of correctly predicting loyal customers. However, in predicting churned customers, only 50% of the predictions were correct. Thus, there is room for improvement and maybe the decision tree is not the ideal classifier.

Logistic Regression confusion Matrix

	Precision	Recall	f1-score	support
0	0.85	0.90	0.87	1038
1	0.66	0.55	0.60	369
micro avg	0.81	0.81	0.81	1407
macro avg	0.75	0.72	0.74	1407
weighted avg	0.80	0.81	0.80	1407

The results of the logistic regression confusion matrix are best. It's precision recall and f1-scores outperform those of other models in the weighted average and for the individual labels. It's recall in correctly predicted churn customers is 55% and maintains consistency among all the other metrics.

K-Nearest neighbors confusion matrix

	Precision	Recall	f1-score	support

0	0.83	0.84	0.84	1038
1	0.54	0.51	0.52	369
Micro avg	0.76	0.76	0.76	1407
Macro avg	0.68	0.68	0.68	1407
Weighted avg	0.75	0.76	0.75	1407

The knn model correctly predicted non-churn customers 83% of the time and correctly predicted churned customers 54% of the time. The weighted precision score i.e. how many times the model made the right prediction is 75% which is a very useful score. The recall value shows that the model did a good job of correctly predicting loyal customers. However, in predicting churned customers, only 51% of the predictions were correct.

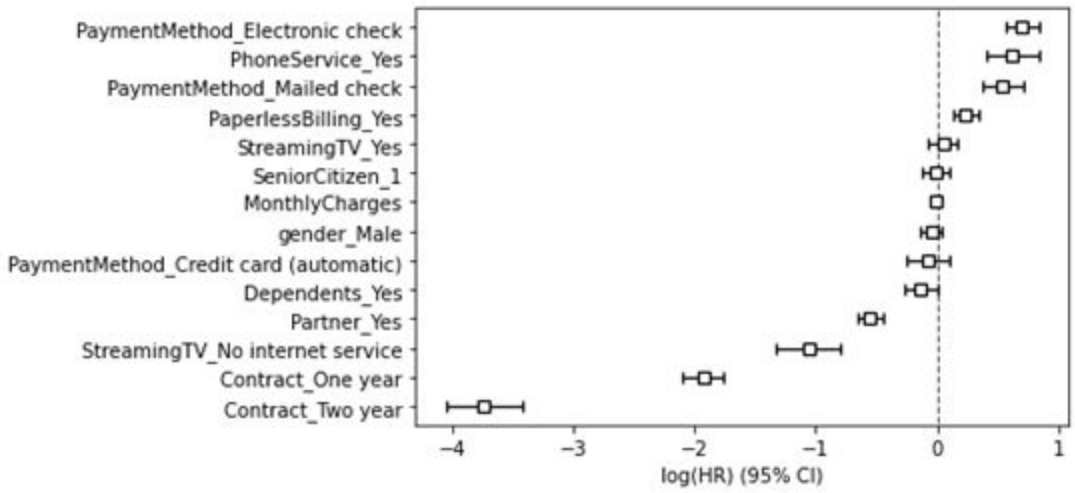
Naive bayes confusion matrix

	precision	recall	f1-score	support
0	0.92	0.65	0.76	1038
1	0.46	0.85	0.60	369
Micro avg	0.70	0.70	0.70	1407
Macro avg	0.69	0.75	0.68	1407
Weighted avg	0.80	0.70	0.72	1407

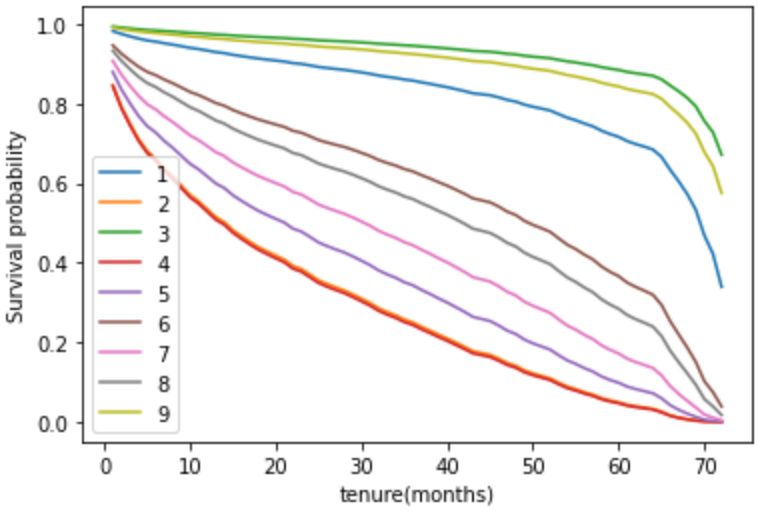
The Naive Bayes model correctly predicted non-churn customers 92% of the time and correctly predicted churned customers only 46% of the time. The weighted precision score i.e. how many times the model made the right prediction is 80% which is a very useful score. The recall value shows that the model did a good job of correctly predicting loyal customers.It shows that of the customers who were predicted to churn, the model got the prediction right 85% of the time.

Survival Analysis

In an attempt to determine the survival curve for each customer we use the Cox proportional hazard (cph) model to derive the survival function from the hazard ratios. We use the cph to obtain hazard ratios for each feature to determine how they impact the customers survival function. ´Monthly charges play a significant role in predicting churn(i.e. it has a p-value smaller than 0.005) although it's coefficient is 0. This is because it's a continuous variable and varies in order.



We are able to estimate survival functions for the first10 customers in the dataset where we notice that after 3 years, 3 of the observations have a survival probability of less than 50%.



Conclusion

In this study, we built multiple classification models to evaluate which classifier returned the best results for predicting which customer will churn and created a Survival Analysis model to calculate the chance of "survival" for each customer at certain periods.

The logistic regression proved to return more reliable results as the model features are balanced and therefore the results suffer when other classification models are used. The survival analysis model proved to be more versatile as it can produce a time frame in which a customer is likely to churn.

Some limitations with the dataset are the lack of diverse groupings i.e. when we determine the survival chance of each individual client we can't fit them into demographics. Being able to fit customers to demographics will give the company an idea of the target audience and develop requisite marketing strategies to retain their business.

References

Harrison, T., Ansell, J. Customer retention in the insurance industry: Using survival analysis to predict cross-selling opportunities. *J Financ Serv Mark* 6, 229–239 (2002). https://doi.org/10.1057/palgrave.fsm.4770054

´J.K. Rogers, S.J. Pocock, J.J. McMurray, *et al.*Analysing recurrent hospitalizations in heart failure: a review of statistical methodology, with application to CHARM-Preserved Eur J Heart Fail, 16 (2014), pp. 33-40

´Wang, Q., Xu, M. & Hussain, A. Large-scale Ensemble Model for Customer Churn Prediction in Search Ads. *Cogn Comput* 11, 262–270 (2019). https://doi.org/10.1007/s12559-018-9608-3

´Kaya, E., Dong, X., Suhara, Y. *et al.* Behavioral attributes and financial churn prediction. *EPJ Data Sci.* 7, 41 (2018). https://doi.org/10.1140/epjds/s13688-018-0165-5

Reichheld, Fredrick F and Sasser W. Earl. Zero Defections: Quality Comes to Services. Boston Harvard Business Publishing, 1990. Web. 11 february 2020.

Christiana Kartsonaki, Survival analysis, Diagnostic Histopathology, Volume 22, Issue 7, 2016, Pages 263-270, ISSN 1756-2317, https://doi.org/10.1016/j.mpdhp.2016.06.005.
(http://www.sciencedirect.com/science/article/pii/S1756231716300639)