Churn Prediction and Customer Segmentation

Süleyman Tuğrul Dinçer

Middle East Technical University

Abstract

Attracting new consumers is no longer a good strategy in the telco industry because the cost of keeping existing customers is significantly lower. In the telecommunications sector, churn management is critical. Because there has been little research integrating churn prediction and customer segmentation, the goal of this paper is to provide a churn prediction and segmentation of who are predicted as churn. There are 4 parts in the framework of the project, which are data-preprocessing, churn prediction, and customer segmentation. This methodology combines churn prediction and customer segmentation to give telco operators with a comprehensive churn analysis for better customer management. IBM Telco Company dataset is used in the project with different machine learning classifiers. These classifiers are implemented on the test data coming out from dataset. After churn status of the specific customer is determined with 10-fold cross validation, accuracy is used for comparing machine learning classifiers as determining metric. Analysis indicates that Logistic classifier is performed best to identify customer who may be churn with the accuracy score 80%. Churn customer segmentation is then carried out using K-means clustering. Customers are segmented into different groups, which allows marketers and decision makers to adopt retention strategies more precisely.

Introduction

A telecommunication is facing a large customer churn rate. Churn might bankrupt a real-world business. We are using the IBM Telco Customer Churn simulated data set, which provides labeled churn for 7,043 customers, due to a lack of publicly available customer data. Churn prediction and customer segmentation are two major components of customer analytics in the telco sector. Because the telco market is saturated, putting too much emphasis on gaining new users is no longer appropriate. It has been proven that the cost of recruiting new consumers by investing large resources is significantly higher than the cost of retaining existing ones.

It is suggested that operators identify consumers who are about to churn before they actually churn. Customer churn prediction gives operators a window of opportunity to remediate and adopt a number of tactical retention initiatives before their existing customers migrate to another provider [1]. Customer segmentation, on the other hand, is a significant tool for performing customer analytics since it divides customers into many categories based on various criteria. Churn prediction should be paired with customer segmentation to enable telco marketers make better judgments. Furthermore, telco operators frequently require more than just projections about client churn.

Literature Review

Most of the earlier research on churn prediction in the telco industry concentrated on a few machine learning classifiers. To make comparisons and find the best predictive model, researchers ran three trials with seven prediction methods: Logistic Regression, Random Forests, Decision Tree, XGBoost, and Support Vector Machine (SVM).

One of the studies is prepared in a way that data mining techniques are used to construct a model for "churn prediction," which determines which customers are likely to stay as subscribers and which are likely to turn to a competitor. In the research CRISP methodology is presented firstly. 6 phases are formed for CRISP methodology. These are as follows: Business

Understanding, Data Understanding, Data preparation, Modeling, Evaluation, Deployment [2]. According to this framework, there are two approach to detect churn customers. In the first approach, K-Nearest Neighbours, Naive Bayes, Decision Trees and Artificial Neural Network are implemented directly to the dataset. However, in the second approach, K-Means clustering is implemented firstly to identify different clusters with custoers who have the similar characteristics. After obtaning these clusters, classification techniques are implemented seperately to the each of the cluster.

Another paper suggests that K-Means algorithm needs to be used to detect different customer clusters who will be churn in near future. These segmentation enables companies to prepare promotions for their customer before they are switching to competitor. This step is implemented after prediction of the customer is determined. This framework is better to act qucikly to respond the customer demands.

Customer Analytics Framework

An integrated consumer analytics framework is provided in this study, as indicated in Figure 1. The pre-processing process begins with data cleansing, data transformation, and data normalization. If there are missing values in the data, they are checked and it is decided whether these data will be filled or deleted. In this study, we decided to delete this data because customers have just entered into a new agreement and have not made their first payment yet. Some features in the data need to be changed and transferred to the model in this way. As can be seen in Figure x, these features with categorical data must be converted to binary data. Therefore, this data undergoes a binary transformation. As can be seen in Figure y, the data in some of the features have a change that does not prevail over each other. In this case, this data was made ready for modelling by applying one hot encoding method.

These prepared data are tested with some machine learning algorithms. 6 different machine learning algorithms such as Logistic Regression, Random Forest, Decision Tree, AdaBoost Classifier, Gradient Boosting Classifier, Support Vector Machines were applied on the data. With these algorithms, churn prediction was targeted. At this stage, the algorithms scikit.learn algorithms are applied one by one, but there is another easy library where we can use different applications such as time elapsed and k-fold cross validation together. After the algorithms we apply one by one, we run the algorithm that gives the best value with this ready library.

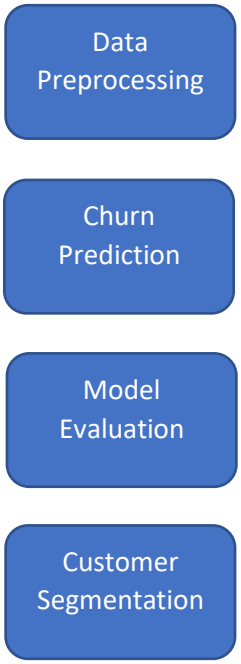

Figure 1 - Analysis Framework

Implementation

Data Setup

For the project scope, Jupyter Notebook is used with Python 3 environment. Sample dataset is used from IBM imaginary customer churn data. Features and format of the data can be seen in Table 1. Before machine learning algorithms are implemented, data cleaning and transformation steps are provided. Dataset included, firstly, 7043 rows and 33 features. These features are grouped into demographic parts, location, services provided by company, payment methods etc. Most of the demographic information is not needed in this case. Morevoer, location information such as country, city, latitude and longitude features is not important to decide whether the customer

will be churn or not. Hence, we can eliminate these columns before machine learning algorithms are implemented and tested.

Table 1 - Data Features and Format

#	Features	Format
1	Gender	Male/Female
2	SeniorCitizen	Yes/No
3	Partner	Yes/No
4	Dependents	Yes/No
5	Tenure	Numerical
6	PhoneService	Yes/No
7	MultipleLines	Yes/No/No phone service
8	InternetService	DSL/Fiber optic /No
9	OnlineSecurity	Yes/No/No internet service
10	OnlineBackup	Yes/No/No internet service
11	DeviceProtection	Yes/No/No internet service
12	TechSupport	Yes/No/No internet service
13	StreamingTV	Yes/No/No internet service
14	StreamingMovies	Yes/No/No internet service
15	Contract	Month-to-Month /One Year /Two Year
16	PaperlessBilling	Yes/No
17	PaymentMethod	Bank Transfer /Credit Card /Electronic Check / Mailed Check
18	MonthlyCharges	Numerical
19	TotalCharges	Numerical
20	Churn	Yes/No

Performance Metrics

To evaluate the performance of models, accuracy, precision, recall and F1-score are used. The ratio of the number of samples properly predicted by the model to the total number of samples is known as accuracy. The ratio of really positive samples to ones predicted to be positive is referred to as precision. How to calculate these metrics are shown below:

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

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - Score = \frac{2 * Recall * Precision}{Recall + Precision}$$

Moreover, ROC curves can be drawn by FPR and TPR. These values are axes of the graphs. Area under the curve (AUC) represents the ROC curve area. Larger AUC values are better for evaluation.

$$FPR = \frac{FP}{FP + TN}$$

$$TPR = \frac{TP}{TP + FN}$$

Results of Dataset

It can be observed from Table 2 that the highest accuracy is obtained as 80% from Logistic Regression. When the other metrics are examined such as F1-Score, Precision, Logistic Regression is dominated other algorithms. Precision is 75%, Recall is 72%, and F1-Score is 73%. AUC from ROC curve for Logistic Regression is also the highest with 84.6% as it can be seen in Figure 2. AdaBoost and Gradient Boosting Classifiers are also giving close results to the Logistics Regression. However, to continue Logistic Regression is selected. By applying K-Fold Cross Validation with the help of LuciferML library, accuracy is going up to 81.05%.

Table 2 - Model Evaluations

Classifiers	Accuracy	Precision	Recall	F1-Score
Logistic Regression	**0.8**	**0.75**	**0.72**	**0.73**
Random Forest	0.79	0.73	0.69	0.70
Decision Tree	0.74	0.68	0.68	0.68
AdaBoost Classifier	0.79	0.74	0.71	0.72
Gradient Boosting Classifier	0.79	0.74	0.71	0.72
Support Vector Machines	0.73	0.37	0.5	0.42

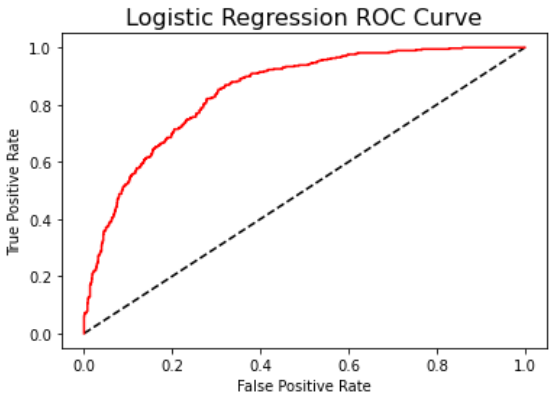

Figure 2 - ROC Curve of Logistic Regression

Customer Segmentation

After prediction of churn customer, these customers are separated from dataset to analyze. Customers who are churn need to be clustered to see the pattern within them. To be able to do that, K-Means algorithm is used to cluster these customers. Firstly, Elbow Method is implemented to see how many clusters are seemed within the churn customers. From this method, elbow criteria shows that optimal number of clusters is set to K=3. Then, K-Means algorithm is used to data to be grouped into 3. The Elbow Method result can be seen in Figure 3 as below.

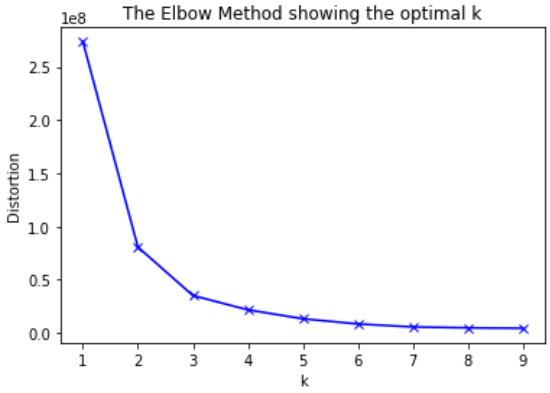

Figure 3 - Elbow Method

PCA is implemented to data to see the clusters with two features. 3 clusters are obtained, and they can be seen as blue, green and orange dots below in Figure 4.

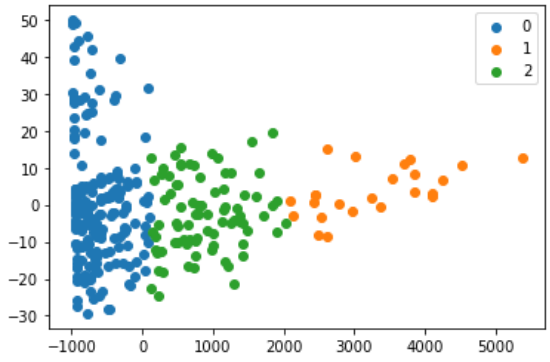

Figure 4 - Cluster of Churn Customers

These clusters are obtained and features within these clusters are evaluated. Which features are more important within the data, which features are more counted within the clusters are reported and following table is prepared.

Table 3 - Clusters with Features

Features	C1	C2	C3
senior_citizen	1 0.24	1 0.45	1 0.38
partner	1 0.19	1 0.45	1 0.45
Tenure_months	4,5	42,6	21,2
multiple_lines	1 0.41	1 1.0	1 0.67
online_backup	1 0.16	1 0.54	1 0.30
device_protection	1 0.19	1 0.62	1 0.29
streaming_tv	1 0.32	1 1.0	1 0.60
streaming_movies	1 0.38	1 0.87	1 0.59
monthly_charges	78.75	95.70	74.95
total_charges	426	4816	2869
payment_method_Bank transfer	1 0.07	1 0.16	1 0.25
payment_method_ Electronic check	1 0.74	1 0.79	1 0.60

Over others, these features which are written in Table 3 has significant difference from each other. Hence, other features are eliminated from this part to determine how the company approaches its customers.

Tenure months, monthly_charges and total charges are numerical features and the number represents the average value for each customer within the clusters. Tenure months in cluster 1 is less than the other clusters. These customers are stayed average 4.5 months with the company which is very less, and churn reason should be examined further. Cluster 2 has an average 42.6 months which is nearly 4 years, and cluster 3 has an average 21.2 months which is nearly 2 year. These customers are loyal for a long time, and they may wait for different campaigns due to loyalty. They may also wait for better care policy when compared to opponents. Other features are represent ratios of customers within the data with the label 1. First row shows the number 1, and this represent Yes for features. Second row represents the ratios. Cluster 1 customers do not generally have partners. They do not have partner, however they are churn. All clusters have higher value for electronic check for payment. However, they are churn and this payment method should be investigated or company could give surveys to its customer how much they are pleased with this method. Cluster 1 has lower ratio for streaming TV and movies. Company could present packages with these features.

Discussion and Summary

Because the telecom market is always saturated with customers, it is more advantageous for operators to propose retention measures for clients who are about to depart. An integrated customer analytics framework is proposed and different machine learning algorithms are implemented in this study.

Logistic Regression gave the best records in terms of evaluation metrics. After these result, customers who are predicted as churn are seperated from actual dataset to implement K-Means algorithm. The aim of this framework to see the patterns within clusters with customers who are predicted as churn. Some features are concentrated in some clusters. These results provide companies with information on how to approach customer groups in churn status. Some examples were examined in this study and some suggestions were made. However, different information and working methods will emerge as a result of data analysis.

References

Utterback, C., 2021. *Predict Customer Churn With Precision*. [online] Medium. Available at: <https://towardsdatascience.com/predict-customer-churn-with-precision-56932ae0e5e3> [Accessed 1 July 2021].

Girgin, S., 2021. *K-Means Clustering Model in 6 Steps with Python*. [online] Medium. Available at: <https://medium.com/pursuitnotes/k-means-clustering-model-in-6-steps-with-python-35b532cfa8ad> [Accessed 1 July 2021].

Wu, S., Yau, W.-C., Ong, T.-S., & Chong, S.-C. (2021). Integrated Churn Prediction and Customer Segmentation Framework for Telco Business. *IEEE Access*, *9*, 62118–62136. https://doi.org/10.1109/access.2021.3073776

Mahnaz Sharafkhani, Hamidreza Koosha (2016). Churn Prediction in Telecommunication Industry: A Data Mining Approach. *ICIE*,