Customer Churn: California Telecom Company Case Study

Vinita Kumari Jhakra
Department *of Computer Science*
City, University of London

***Abstract*—This paper focuses on analyzing the customer historical information data from a Telecommunication company using data exploration to identify the customer churn behavior and build a churn prediction model that predicts the customer tends to churn. This predicted classification would be utilized by the company to reduce customer outflow and revenue growth by customer retention. The Ensemble technique Random Forest is used for Churn Prediction by effectively selecting the predictive features and appropriate balancing techniques for reducing the cost. Considering the wide clientele, it is not an effective business growth strategy to practice the retention of every individual customer. Henceforth, the outcome of the predictive model would benefit the company focusing on the particular customers classified as future churners.**

I. Introduction

With the advancement in technology and the highly competitive market in particular, customers are more likely to switch to other telecom operators easily if they get better offers and rewards, which could cause a huge loss to the business. There are many researches done in the past which have outlined that retaining the customer is more profitable than onboarding new customers and according to them - the cost required to retain a customer is 16 times less than to get a new customer [3]. Also, F. Reichheld and W. E. Sasser [2] projected that controlling 5% of the customer churn rate can increase the company's profit by between 25% and 85%. Henceforth, service provider companies focus more on retention than on acquiring new customers. Data Science plays a crucial role in the area of customer churn where historical data is analysed to illuminate patterns and insights of churning and identify the customers who are prone to switch to another service provider.

II. Anlytical questions

For any product or service-based company, customer churn aka customer attrition is one of the major challenges where multiple alternatives are available. Along with offering the services to new customers, it is equally important to retain the existing ones for continuous business growth. There can be several reasons why a customer might choose to switch from their present service provider and if the company can effectively predict customer churn then they can segment such customers and build strong retention strategies for different segments of churner.

With this study, we aim to investigate the patterns in the customers churning from the telecom service provider and predict the customer who tends to churn which would allow the company to target customer retention using more effective ways in market strategies to resolve customer churn.

This problem can be divided into two sets of analytical categories. To start with, focus on analysing the characteristics of the possible churns using visual analytics and data mining approaches to address the below questions.

- Major reason why Customers churning?
- Is there any pattern of people churning?
- Is there any certain group of people who are churning more?
- who are the loyal customers?

This then drives the rest of the research by investigating which customers tend to churns:

- How machine learning model can help businesses to reduce the loss?
- Who are the people at high risk of churn and whom to give more attention to retaining?

III. Data(materials)

Data has been taken from the public repository Kaggle which was originally scraped from the Maven Analytics website platform. This dataset has overall 38 features and 7043 observations from a Telecommunication company in California in Q2 2022.

Each observation represents a customer and contains features like demographics, tenure, subscription services, monthly and total charges, status for the quarter (joined, stayed, or churned), and more. Not all the features are important and relevant to the analysis of the problem. There is some missing and negative data. Data is observed biased among its classes (in the ratio of 26.54% ‘Churned’, 67.02% ‘Stayed’, and 6.45% 'Joined’). We have a mixture of numerical (discrete, continuous) and categorical (binary) types. Each observation contains various charges which is one of the key features that would be used in the ML model to anticipate if the customer tends to churn or not. Also, “churn reason” can help to observe the behaviour of customers leaving. Subscription services-related features can give a glimpse of internet, streaming, and phone services subscribed by the customer segment as ‘churned’ and ‘stayed’ in particular.

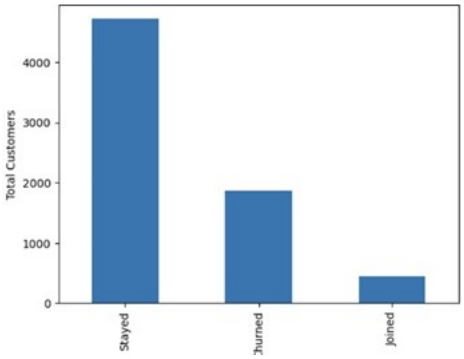


Fig. 1. Distribution of customer status.

IV. Analysis plan

The aim is to anticipate the customer churn from the company and maximize retention and customer satisfaction thus loyalty to the company. Visual analytics are done on the data to illustrate the insights and a classification model is developed for prediction. Therefore, the following analytical components would outline the overall analysis plan execution.

A. Exploratory data Analysis

EDA is the initial step of data science where raw data is investigated by finding the patterns among the features that can help to understand the data. Hidden trends of data were explored using descriptive statistics and different data visualisations, for example, histogram, Pareto graph, and count plot which illustrated the insights of important features and patterns. These analysis outcomes helped to take out the decision on the ML model we need to build in the later stage. The following aspects were examined with the help of EDA. A histogram was used to examine the distribution of numerical features and count plot for distribution by categorical features, Fig. 1 shows the distribution of the target variable "Customer Status".

B. Data Preparation

Data required to be cleaned before processing for the model otherwise could lead to misleading results, cleaning data is one of the major challenges in the field of data science. Data had missing and negative values and no duplicates. Missing values were based on the type of subscription services. For those customers who did not subscribe to the phone or internet services, their respective subscription services-related features were null, we imputed them with "No" for binary features(Yes or No). "Not applicable" for other categorical values and 0 for numerical values. Monthly charges had negative values with outlier values (for example - 4) that were imputed with the formula (*Monthly charges = Total charges/Tenure in Months*). Different encoding is performed on categorical data, label encoding for binary classification and binary encoding for rest [4]. Numerical data is normalized to get better results.

C. Data derivation

Manipulating a variable to create a new variable is called data derivation. Our dataset has already some features that were derived from existing features(Total Revenue, Total long-distance charges). We have binned a variable called Age (continuous variable) into different age groups (categorical) to visualise the pattern better and to see if this could improve the results. One more variable "Tenure in Months" is manipulated and converted into years which may provide more insights to the data. The calculated churn rate of each categorical data with their respective percentage population.

D. Feature selection

Feature selection plays a vital role in churn prediction as there are n number of features available related to customers and finding the best for good results is a challenging task. Important features were selected by considering hypothesis test results (ANOVA test for numerical features and chi-square test for categorical data). Features with a p-value of more than 0.05 were discarded. Important features with p-value less than 0.05 were selected by analysing them with the help of visualisation. Along with that, we have considered the feature importance obtained by the model. So, the best features for training were selected by considering these three.

Also, features with high correlation have linear relation among them so any one of the such features is considered for model training.

E. Construction of Model

As per many literature reviews, the ensemble technique provides better accuracy for customer churn prediction [5]. We have used the Random Forest algorithm for the classification of customer status. 80% of the original data was used to train the model and 20% for testing. The model was trained with selected important features and optimized the number of trees to be 200 as a hyperparameter and five-fold cross-validation was performed. As our data is unbalanced we experimented with different balancing sampling techniques i.e., SMOTE, SMOTEENN, and WRF. The SMOTEENN technique performed better in handling unbalanced data. SMOTEENN is a balancing technique that is the combination of over and under-sampling.

F. Validation results

The model obtained a maximum validation score of 97.11. In our model, we gave more attention to improving Type II error along with Accuracy. Type II error (False Negatives) needs to be minimized as customer acquisition is more expensive than customer retention with rewards or discounts. Results show that balancing data in customer prediction is a challenging task.

Incorrect predictions could result in a company losing profits because of the discounts offered to continuous subscribers. Therefore, the right predictions of the churn customers have become highly important for companies [6]. Here we have considered the results with the SMOTEENN technique. Performance metrics are shown in Fig. 2 and Fig. 3.

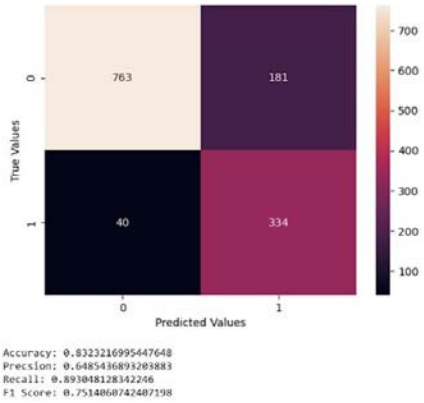


Fig. 2. Confusion Matrix and performance parameters.

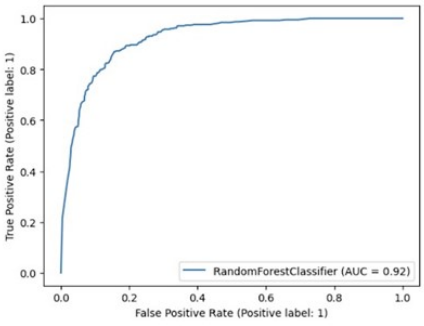


Fig. 3. ROC curve.

V. FINDINGS, REFLECTIONS AND FURTHER WORK

A. *Customer churn key reasons*

Based on the feedback provided by the customers churned, the top reason for customer churn is the market competition. The second most reason is dissatisfaction of services offered (i.e. related to product, network etc.) and the third is the attitude of customer services or service provider shown in Fig. 4. Better customer services and support are the key to keeping customer satisfaction high.

B. *Churning Pattern*

It is shown in Fig. 5a and Fig. 5b that customers with contract "Month-to-Month" are the majority of population(~49%) with high churning rate (~52%) compared to "one-year" and "two-year" with (~11%) and (~2%) respectively. We can say that the latter ones are loyal customers and month-to-month are high churners. Based on the pattern illustrated in Fig. 6a and Fig. 6b, the majority of customers did not receive any offer. Also, those who received offer E have high churn rate (~ 68%) whereas the customer who received offer A have lowest churn rate ~ 8%.Company needs to revise its reward strategy inline to offer A. It is necessary for the company to announce the deals and special offers to the customers for better engagement resulting in high customer satisfaction.

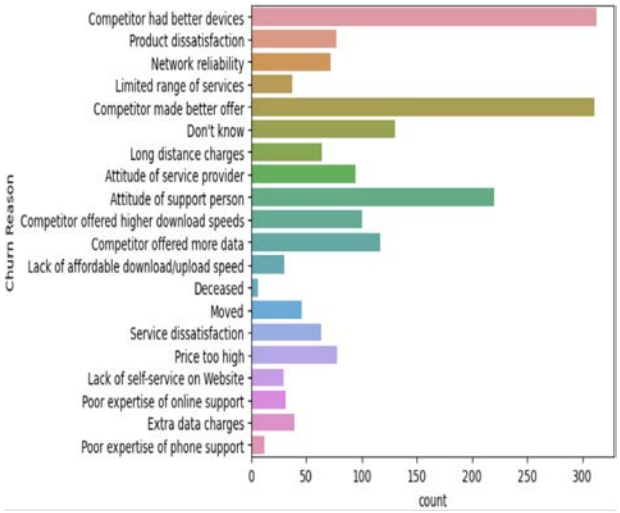
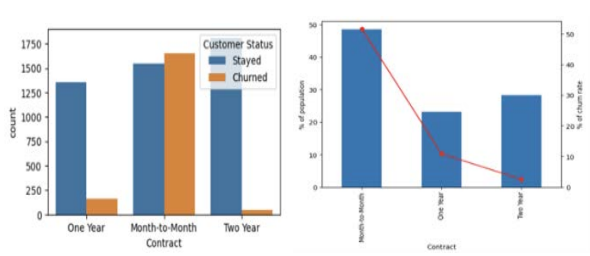
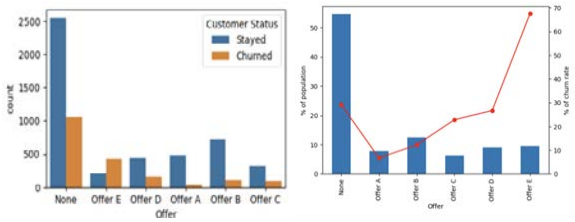


Fig. 4. Total count by Churn reason.



(a) (b)

Fig. 5. (a) Total count by contract. (b) Contract Churn rate against population.



(a) (b)

Fig. 6. (a) Total count by offer. (b) Offer Churn rate against population.

C. *Churn impact on revenue*

It was observed that Telecom company had a churn rate of 26.53% in Q2 2022. Total company revenue was $21,316,851.94 out of which 17.28% revenue generated by churned customers. Avg monthly charge for customer churned was $73.35, higher compared to the customer stayed $61.73. In Fig. 7, customer with contract 'Month-to-Month' had highest population 88.55% among all the churn contracts and contributed ~68% in the overall revenue generated by churned customers.

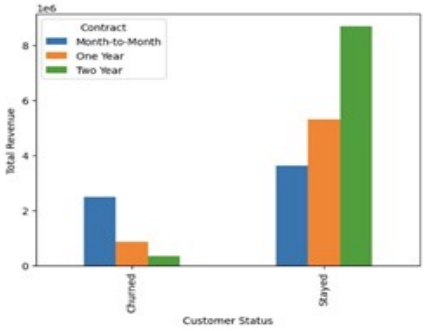
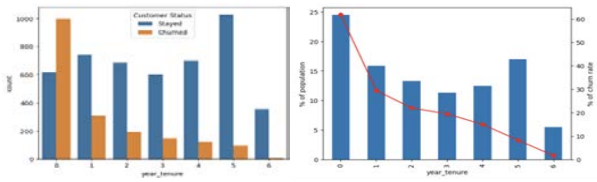
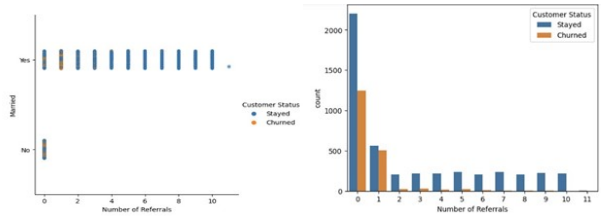


Fig. 7. Total revenue generated by Churn status and contract.



(a) (b)

Fig. 8. (a) Total count by offer. (b) Offer Churn rate against population.



(a) (b)

Fig. 9. (a) Referral given by married and unmarried. (b) Total count by referral.

D. *Loyal customers*

Customers who have been associated with the company for 5 years or more were very less tend to leave and hence loyal to the company (Fig. 8a, 8b). Fig. 9a, 9b illustrate that only married customers shared the referral with others and the customers among those with high referrals were very less tend to leave. Such customers were low maintenance to company

and their loyalty helped in marketing to the business by referring their knowns, dears and other connections [3].

E. Customer requires most attention

From Fig. 11, churn was highest (in Red) and most when monthly charges were high and tenure was less than a year. It implies that recent customers (within 1 year tenure) with high monthly charges are more tend to leave.

F. Revenue oriented ML model

A churn prediction ML model would help the company to predict and take necessary actions to retain the customers that tend to leave. Fig. 10 illuminates the features with the most information gained during model training and similar top results we obtained with EDA and hypothesis testing. We therefore propose to focus on these features most but not limited to.

G. Reflections and further work

A churn considering the scope and timelines of the project, we believe that this research was sufficiently thorough to address our analytical questions and apply the relevant techniques to support data-driven decision-making regarding customer churn. The limitations found in the dataset make it challenging to provide concrete and specific recommendations. Data is unbalanced. Additional features like time series, customer behavior, network behavior, survey details - customer satisfaction (CSAT) etc. would have helped to make more sophisticated predictive model to provide precise predictions.

More advanced sampling methods are required for balancing the data to get more accurate prediction compared to the currently applied methods i.e. with improved TYPE II error.

Finally, successful churn prevention system must consider the retention action that outlines the right strategies to retain the valuable customers to company [7]. Further, inline to that we should have recommendation model following the predictive model outcomes.

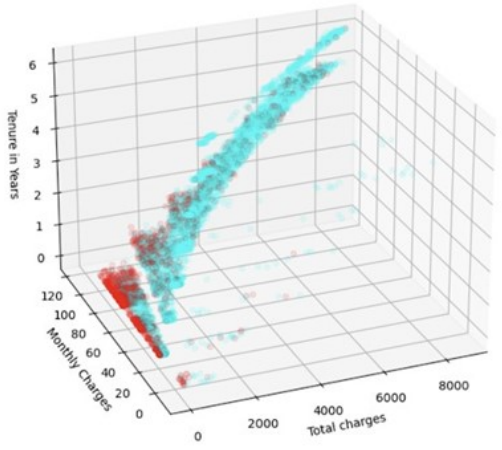


Fig. 11. Monthly charges and Total charges by tenure.

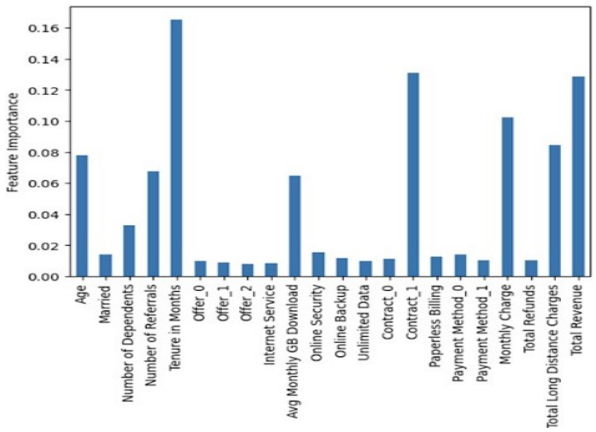


Fig. 10. Total revenue generated by Churn status and contract.

REFERENCES

[1] Ahmed, Ammara, and D. Maheswari Linen. “A Review and Analysis of Churn Prediction Methods for Customer Retention in Telecom Industries.” In 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS), 1–7, 2017.

[2] Reichheld, F., and W. E. Sasser Jr. "Zero Defections: Quality Comes to Services." Harvard Business Review 68, no. 5 (September–October 1990): 105–111.

[3] Almana, Amal M, Mehmet Sabih Aksoy, and Rasheed Alzahrani. “A Survey On Data Mining Techniques In Customer Churn Analysis For Telecom Industry” 4, no. 5 (2014).

[4] Reilly, Denis, Mark Taylor, Paul Fergus, Carl Chalmers, and Steven Thompson. “The Categorical Data Conundrum: Heuristics for Classification Problems—A Case Study on Domestic Fire Injuries.” IEEE Access 10 (January 1, 2022): 1–1.

[5] Mishra, Abinash, and U. Srinivasulu Reddy. “A Comparative Study of Customer Churn Prediction in Telecom Industry Using Ensemble Based Classifiers.” In 2017 International Conference on Inventive Computing and Informatics (ICICI), 721–25, 2017.

[6] Yildiz, Mumin, and Songul Varlı. “Customer Churn Prediction in Telecommunication.” 2015 23rd Signal Processing and Communications Applications Conference, SIU 2015 - Proceedings, June 19, 2015, 256–59.

[7] Hung, Shin-Yuan, David C. Yen, and Hsiu-Yu Wang. “Applying Data Mining to Telecom Churn Management.” Expert Systems with Applications 31, no. 3 (October 1, 2006): 515–24

WORD COUNTS

TABLE I. WORD COUNTS OF EACH SECTION

Serial no.	Word counts		
	Section	***Maximum***	***Actual***
1	Abstract	150	119
2	Introduction	300	161
3	Analytical questions	300	243
4	Data(Materials)	300	179
5	Analysis	1000	748
6	Findings, reflections and further work	600	593
	Total	2650	2049