Customer Attrition Analysis

***Abstract*—This project forecasts the risk of customer attrition and determines the point of intervention for each customer to retain them. We deploy the XGBoost algorithm which takes selected features as input and calls Cox proportional model to obtain a hazard score. We elaborate on how the technique is implemented and evaluate the model performance.**

***Keywords*—XGBoost, Survival Analysis, Cox's Hazards Function, Intervention Point, Retention**

I. Introduction

Gone are the days when companies used to rely on the word of mouth for spreading awareness of their products and services. In the digital age, firms invest heavily in advertisements and marketing to sign up more customers who could subscribe to as many services as possible. However, such companies focus extensively on onboarding potential customers rather than concentrating on retaining present clients. As per figures put together by Invesp, 44% of companies look in the direction of acquisition, while only 18% stress retention.

This trend results in an exorbitant capital going towards enrolling new customers which could instead be diverted in retaining existing customers. It has been discovered that bringing in a new customer can cost five times more than retaining an existing customer. More surprisingly, increasing customer retention by 5% can lead to a rise in profits by 25 to 95%. Undoubtedly, there are a variety of reasons that reinforce these statistics. Firstly, loyal customers are more likely to rebuy multiple products and endorse services by referring to new clients. Secondly, a loyal customer base is a testimony of the quality of service offered by an organization. As per the American Express study, 33% of customers will lay off a commodity right after one poor experience. Therefore, customer retention will not only bolster the quality of service but also boost the brand reputation in the market. For an organization to stay competitive, it has to analyze the expectation of the masses and articulate their needs for enhancing service experience. Another interesting fact which comes to light is the success rate of selling to an existing customer is 60-70%, while the same figure goes down to 5-20% for a new customer. Needless to say, we have substantial evidence to reiterate that customer retention will drive growth and optimize the expenditure cost of an organization.

II. Motivation

While customer retention is fruitful in any domain, one industry that caught our eye was the telecom sector. As per the American Customer Satisfaction Index (ACSI), telecom scored just 72 points finishing 9th in the overall customer satisfaction index.

TABLE I.
ACSI Index Score

Rank	Sector	Score
1	Manufacturing (Non-durable goods)	80.4
2	Manufacturing (Durable goods)	79.1
3	Hospitality & Food services	78.9
4	Finance & Insurance	77.8
5	Retail	77.3
6	Transportation	75.1
7	Healthcare & social assistance	74.7
8	Energy Utilities	73.2
9	Telecommunications	72
10	Public assistance/Government	66.7

Source:https://www.europeanbusinessreview.com/how-costly-is-customer-churn-in-the-telecom-industry/

Being a telecom user for over the last 10 years, we were surprised with these results and couldn't resist understanding the rationale behind such a low ranking. We couldn't help but wonder, how can we help the telecom business in retaining their customers and make a difference. Therefore, our first step was to pick up a dataset of a telecom provider and perform creative analysis for bringing out the critical features, understanding trends, loopholes in policies and other essentials for customer satisfaction. As we go down the line, we will elaborate on the remediation approach and look into several parameters that can assist telecom providers better in retaining their customers.

III. Data Description

A. Dataset

The data is taken from the University of California repository, and it consists of 5000 customer call details of an Iranian Telecom Company. It has 21 attributes with five attributes are categorical and the rest 16 are numerical variables. All of the attributes except for churn is the aggregated data of the first 9 months. The churn labels are the state of the customers at the end of 12 months. The three months is the designated planning gap. The total number of churners is 2498 (50% of total customers). So, it is balanced on target.

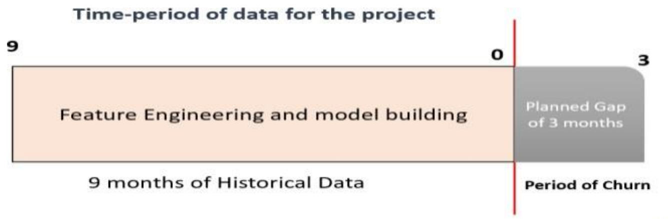

Time-period of data for the project
9
0
3
Feature Engineering and model building
Planned Gap of 3 months
9 months of Historical Data
Period of Churn

Fig. 1. Time-period of Data for Project

Target Features: As the model predicts, both time and risk of attrition/churn, two target variables are required to achieve it.

TABLE II.
Target Features Description

Target Feature	Description
Account length	Duration of customers stay
Churn?	Whether the customer left the service or not

Dataset Attributes Type: The summary of attributes data-type is displayed below

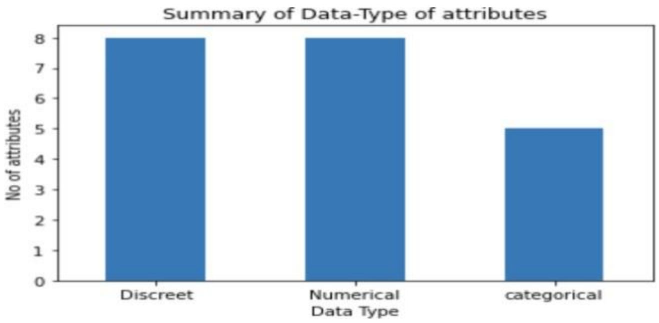


Fig. 2. Summary of Data Type of Attributes

Data Dictionary: The Name and description of important features are listed below

TABLE III.
Data Attributes Description

Attribute Name	Description
State	A two-letter abbreviation of US state
Int'l Plan	Flag showing that whether the customer has taken international plan
Email Plan	Flag showing that customers took Vmail message plan
Day Mins	Average Minutes of day calls
Day Calls	The average number of daily calls
Day Charge	The average charge of day calls
Eve Mins	Avg. Minutes of evening calls
Eve Calls	Avg. number of evening calls
Eve Charge	Avg. charge of evening calls
Night Mins	Avg. Minutes of night calls
Night Calls	Avg. number of night calls
Night Charge	Avg. charge of night calls
Intl Mins	Avg. min of international calls
Intl Calls	Avg num of international calls
Intl Charge	Avg. charge of international calls
CustServ Calls	Avg. number of calls made to Customer service

IV. Exploratory Data Analysis (EDA)

The data is analyzed and investigated to get the main characteristics, understand the underlying data pattern and test our hypothesis. Some important insights and attribute characteristics are given below

a) The day calling time and call charges are higher for customers that left the company.

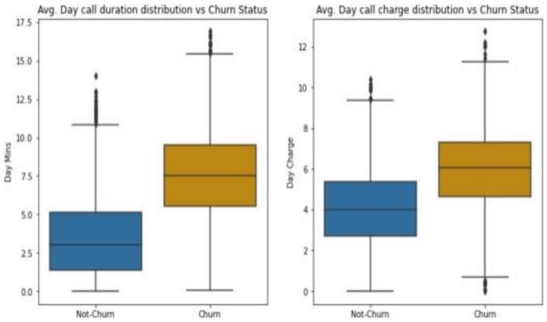


Fig. 3. Boxplot of Churn vs Avg. Day Call Durations and Charges

b) The avg number of night calls are lower, but night call charges are higher for churn customers.

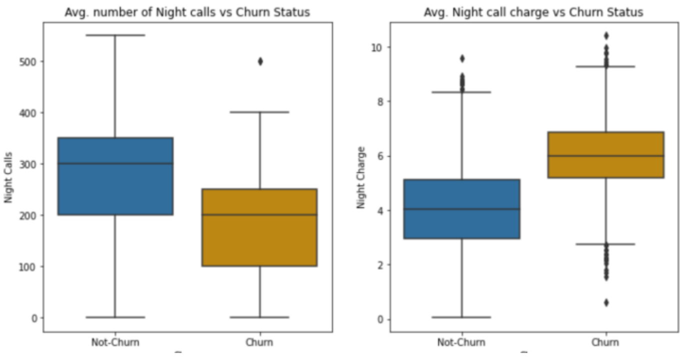


Fig. 4. Boxplot of Churn vs Avg. number of night calls and Charges

c) The evening calling time and call charges are higher for customers that left the company

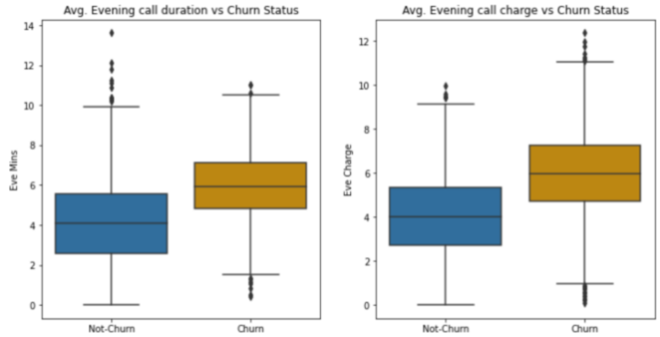


Fig. 5. Boxplot of Churn vs Avg. Evening Call Durations and Charges

d) The night call charges are higher for churned customers even though the time duration is almost the same.

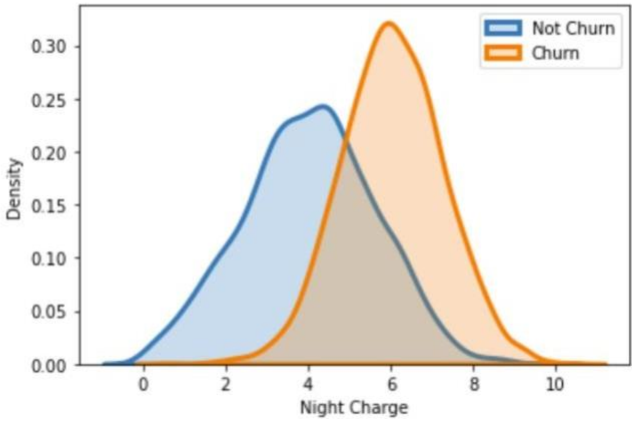
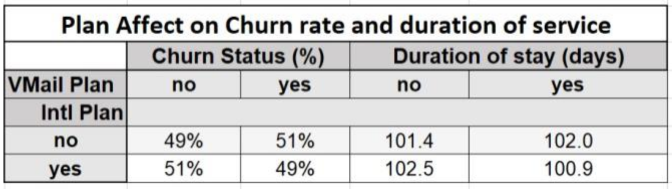


Fig. 6. Area chart between Churn vs Night Charge

e) It appears that company plans are not working as the avg. churn rate and duration of service is the same without a plan



Plan Affect on Churn rate and duration of service				
	Churn Status (%)		Duration of stay (days)	
VMail Plan	no	yes	no	yes
Intl Plan				
no	49%	51%	101.4	102.0
yes	51%	49%	102.5	100.9

Fig. 7. Effect on Churn Rate and Duration Services

f) The customers sending lesser messages are more likely to leave the services.

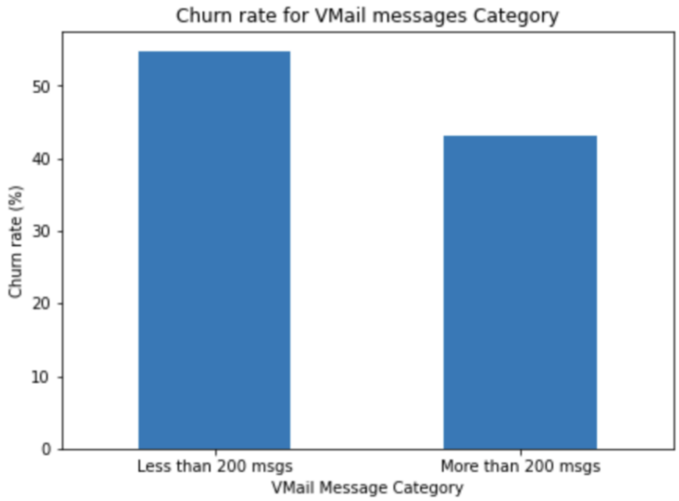


Fig. 8. Bar plot of Churn rate vs Messages group

g) NY has the highest churn rate (~62%) among all states and, apparently in NY, those who have not taken Vmail Plan are leaving more likely to leave (65% vs 58%).

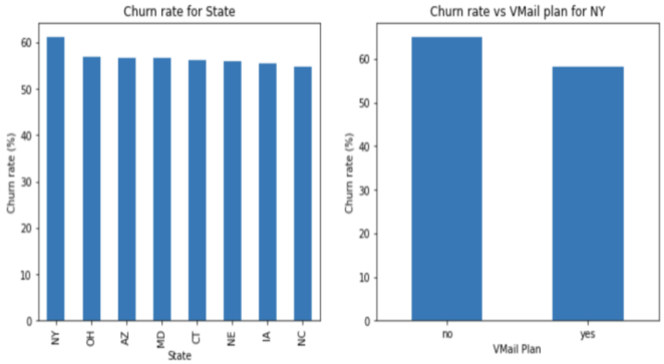


Fig. 9. State vs Churn Rate

V. Feature Engineering

In Machine learning “garbage in, garbage out” is the concept that if the erroneous data goes in the model, the output follows the same fate. Thus, the data needs to be analyzed, visualized and processed to get the best features for the model. It’s the most time taking and challenging part. It is divided into the following stages

A. *Feature elimination*

The following features are removed in the beginning as they don’t explain the target.

- Phone: each customer has a unique phone number, and it has no bearing on churn rate and duration of stay.
- Area Code: there are 33 different area codes, and each area code has almost the same churn rate (~50%) so it constant variance.

B. *Feature Creation*

The following features are created to explain the target and uncover the hidden pattern.

- min_per_day_call: Avg. num of mins per call in the daytime.
- min_per_eve_call: Avg. num of mins per call in eve-time
- min_per_night_call: Avg. num of mins per call in the nighttime.
- min_per_Intl_call: Avg. num of mins per international call.
- charge_per_day_call: Avg. charge per call in the daytime.
- charge_per_eve_call: Avg. charge per call in the evening time.
- charge_per_night_call: Avg. charge per call in nighttime
- charge_per_Intl_call: Avg. charge per international call.

C. *Feature Transformation*

Feature transformation may not have a significant on the overall performance as XGBoost is used which is a tree-based boosting technique. However, the data attributes are normalized to get the best results.

D. *Feature Selection*

Before running the model, it is necessary to find the relevant features. It is required to avoid overfitting and reduce computation time. For this purpose, two methods are used to get the most appropriate variables The techniques used for this feature importance are:

- Feature importance using Random Forest
- Univariate Selection
- Recursive Feature Elimination
- Feature related to duration using Linear Reg.

Feature importance using Random Forest: Identified the following attributes as significant.

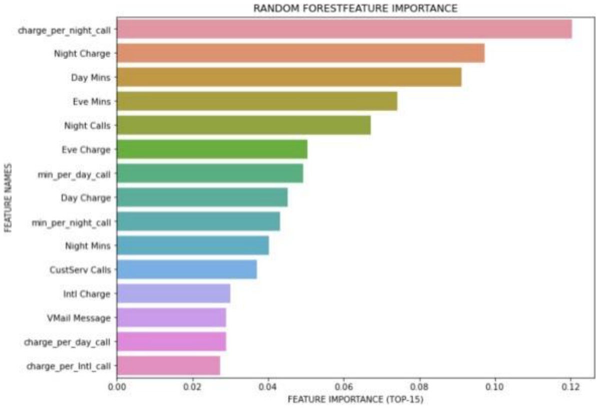
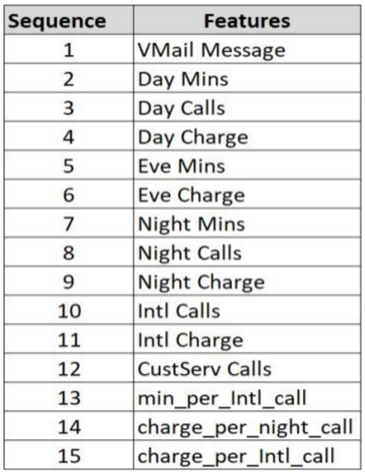


Fig. 10. Feature Importance using Random Forest

Recursive Feature Selection: identified the following attributes as significant.



Sequence	Features
1	VMail Message
2	Day Mins
3	Day Calls
4	Day Charge
5	Eve Mins
6	Eve Charge
7	Night Mins
8	Night Calls
9	Night Charge
10	Intl Calls
11	Intl Charge
12	CustServ Calls
13	min_per_Intl_call
14	charge_per_night_call
15	charge_per_Intl_call

Fig. 11. Feature Importance using Recursive Feature Elimination

Univariate Selection: the features that are highly correlated to the target (churn) are selected. The following criteria are used based on the variable data type.

- Categorical variable: **Chi-Squared** statistics are used, and it has identified '<u>State</u>' as an important feature
- Numerical variable: **Spearman correlation** are used for numerical variables, and it has identified the following as important

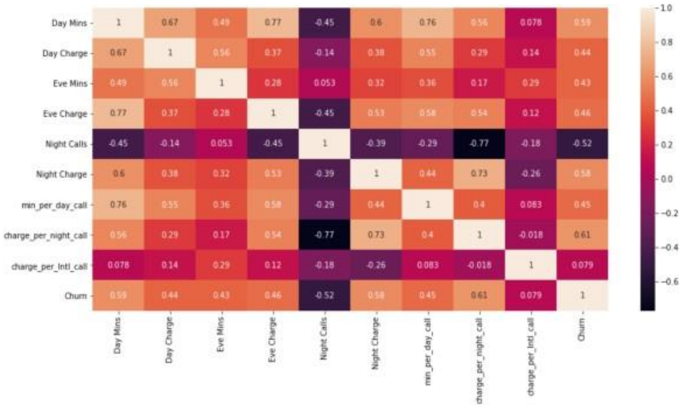


Fig. 12. Correlation Heat Map

Duration related Features: the features that are highly correlated to the duration (time of churn) are selected. The input features are regressed on the duration of churn and the best fifteen features are selected as given below

Features selected using Linear Regression
VMail Plan, Day Mins, Eve Calls, Night Calls, Night Charge, Intl Mins, min_per_eve_call, min_per_Intl_call, charge_per_day_call, charge_per_Intl_call, charge_per_Intl_min, State, VMail Message Category

VI. ATTRITION MODEL

As both time and risk of churn need to be predicted for each customer, a more informative modelling approach is implemented. The main parts and steps are explained below

A. *Survival Function*

- Survival function *S(t)* gives the probability of attrition beyond a point of time *S(t) = P(T > t)*
- It is created to predict the risk of attrition over time. There are two ways to obtain it:
    - ○ Parametric approach: Choose an appropriate distribution such as exponential, geometric and test it with data.
    - ○ Non-Parametric approach: An estimator like - Kaplan Meier is used to create survival function

  The estimator is defined as:

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}$$

  n_i – The number of customers is at risk of churn before time t

  d_i- The rate of customer churning at time t

- It gives an estimate of when the intervention is required at an overall level, not at the customer level. As shown in fig below, by about 152 days half of the customers will churn. But it doesn't give information about every customer.

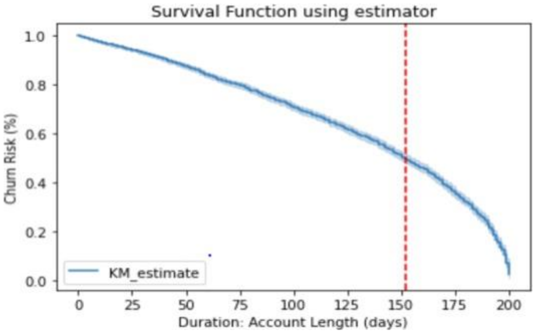


Fig. 11. Survival Function using Estimator

B. *Cox proportional hazards (PH) regression*

- To resolve the issue with the Survival function, Cox (PH) regression method is used in the model. It will give the risk of churn for each customer at a different point in time.
- It uses a log-risk function called Hazard function (*h(t))* and input features to perform regression.
- It maximizes the partial likelihood function to obtain the coefficients in regression.

C. *Hazard Function*

- Hazard function *h(t)* gives the instantaneous probability of attrition at *t* given the customer stays up to time *t.* Mathematically, it is given as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t};$$

- Survival can also be calculated using the Hazard function by simply integrating it from 0 to time *t* and taking the exponential of it. S(*t*) = exp(-$\int_0$h(*t*))

D. *XGBoost*

- XGBoost is an ensemble boosting algorithm that combines weak base learners sequentially to minimize the loss function.
- The same boosting method is used to maximize the partial likelihood function of Cox (PH).
- The data is prepared in DMatrix format for XGBoost to run through the regular, non-scikit API.
- The model predicts hazard score (HR) for each customer at a specified time interval. The hazard score estimates the Customer's probability to leave the Company. For ex.

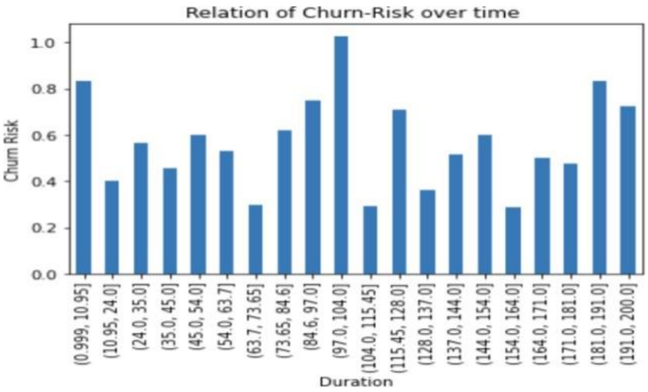


Fig. 12. Churn Risk over time by base model

E. *Model Performance Evaluation*

Two performance metrics were used to evaluate the model's performance:

- **Harrell's Concordance Index (C):** is a goodness of fit for models which gives a risk score, in our case Hazard score. C=0.5 is equivalent to a random guess. The formula is shown below

$$c = \frac{\#\ \text{concordant pairs}}{\#\ \text{concordant pairs} + \#\ \text{discordant pairs}}.$$

- **Brier Score:** is used to measure the accuracy of a predicted survival function at a given time *t* it represents the average squared distances between the observed survival status and the predicted survival probability.

$$\mathrm{BS}^c(t) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \le t \land \delta_i = 1)\frac{(0 - \hat{\pi}(t|\mathbf{x}_i))^2}{\hat{G}(y_i)} + I(y_i > t)\frac{(1 - \hat{\pi}(t|\mathbf{x}_i))^2}{\hat{G}(t)},$$

where $\pi(t|\mathbf{x})$ is the predicted probability of remaining event-free up to time point t for a feature vector $\mathbf{x}$, and $1/\hat{G}(t)$ is a inverse probability of censoring weight, estimated by the Kaplan-Meier estimator.

F. *Methodology*

The model is tested on five different features sets obtained during feature engineering

- Base model: All the variables are used to get the performance of the base model.
- Filter based: Features created using Univariate method are used for this model.
- RFE based: The model is built using features selected using Recursive Feature Elimination.
- Random Forest based: The top features picked by Random Forest are used for this model.
- Linear Regression based: The best features that predicts the duration of churn is applied for the model.

Target creation: The data is right censored (some customers are not churned during the observation window), so the target is set accordingly.

- Right censored data: The duration is set to negative for the customers who have not churned yet.
- Uncensored data: The duration is set to positive for the customers who have churned during the observation window.

ML model: DMatrix format of XGBoost is used for the model building.

- Survival function: The cox proportional hazard function is optimized to get the best results
- Hyperparameter Tuning: The hyperparameter related to boosting are tuned to optimize the loss function. Some important parameters are

  eta: It is the step size shrinkage used in update to prevent overfitting

  max_depth: The maximum depth of the tree. This is also used to control overfitting.

  Subsample: It denotes the fraction of the observations to be randomly sampled for each tree.
- Pipeline implementation: The python pipeline is used for preprocessing and model creation so that it can be deployed to cloud or any other platform easily.

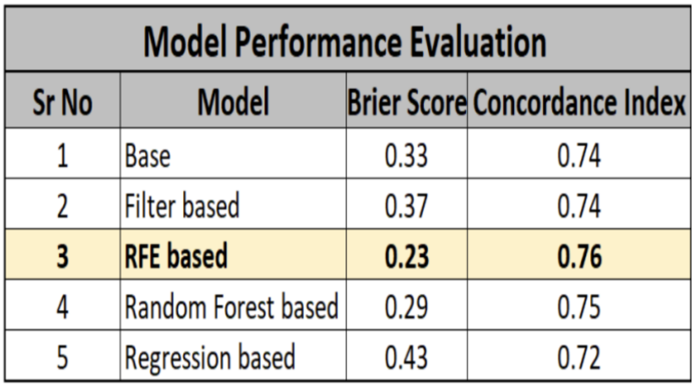
G. *Result:*

As mentioned before, Brier score and concordance index (C-Index) are used to measure the performance of the models.

Brier score: ranges between 0 to 1, 0 being the best and 1 being the worst. The score around 0.25 is considered to be satisfactory and acceptable.

Concordance Index: is equivalent to AUC score and ranges between 0 and 1. 0 is the worst and 1 is the best. However, score of 0.7-0.8 is considered to be adequate and suitable for the model acceptance.

The results for the five models run on different set of features are shown below.



Model Performance Evaluation

Sr No	Model	Brier Score	Concordance Index
1	Base	0.33	0.74
2	Filter based	0.37	0.74
3	**RFE based**	**0.23**	**0.76**
4	Random Forest based	0.29	0.75
5	Regression based	0.43	0.72

Fig. 13 Table showing the model performance on different features

Observation: RFE based feature is giving the best Brier score and Concordance index. So, it is the final model.

Random Forest based model is also very close to RFE and it can be fine-tuned for better performance

Linear Regression based model displayed the worst performance implying that features related to churn duration are not substantial and insightful.

Global Prediction: The point in time churn- probability predicted by the model can be used to see how overall customer base churn appears over the observation period. As per the plot, the chances of churn is maximum during 80-102 days.

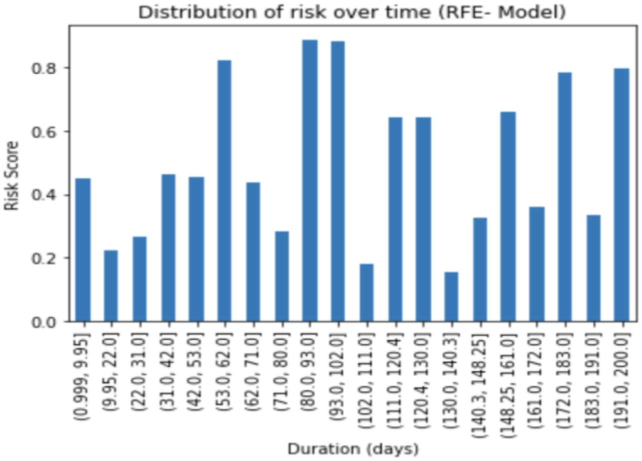


Fig. 14 The bar plot show the probability of churn over time of stay.

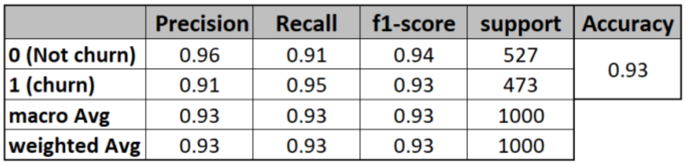
H. *Additional use of Model:*

The model can be utilized for many purposes. It has numerous benefits such as using it as classification model to predict churn or non-churn customers, customer segmentation based on risk score.

Used as Classification model:

The probability of churn can also be used to predict the customer who is going to churn or not. It is similar to the classification model. And it shows that we can utilize the model for classification purpose as well.

Results: The probability of RFE based model is used to predict the churn. The Precision, Recall, accuracy and AUC score are shown below. It proves that the model classifies with high precision and accuracy.



	Precision	Recall	f1-score	support	Accuracy
0 (Not churn)	0.96	0.91	0.94	527	0.93
1 (churn)	0.91	0.95	0.93	473	
macro Avg	0.93	0.93	0.93	1000	
weighted Avg	0.93	0.93	0.93	1000	

Fig. 15 Classification performance matrices using the final model

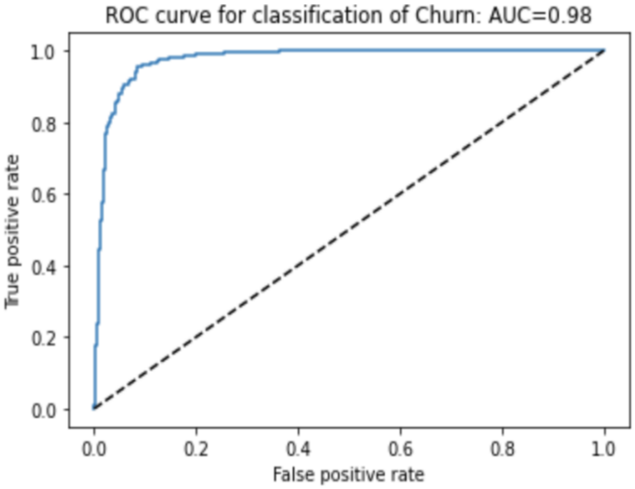


Fig. 16 Area Under Curve (AUC) on test data using the final model

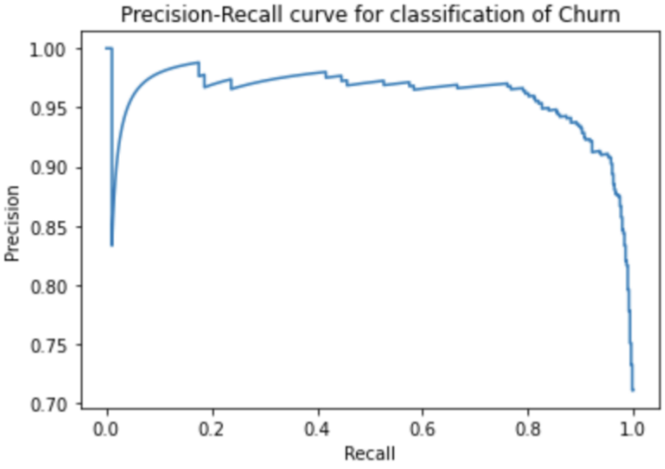


Fig. 17 Precision-Recall curve on test data using the final model

Customer Segmentation:

The probability of churn or the risk score can be used to group the customer base. After understanding the groups attributes and behavior, customized strategy could be planned to stop them from quitting. To cite an example, the test data is segregated into two groups based on the risk-score. The K-means algorithm is applied to perform the clustering exercise. The results are depicted below

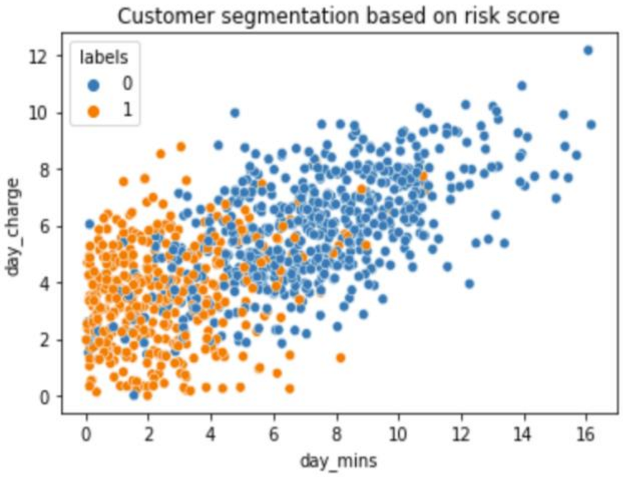


Fig. 18 Customer segmentation on test data using K-Means

Observation: The customer base is clearly partitioned into lower usage in daytime and higher user in daytime. Based on different usage a customized plan can be operated for higher retention.

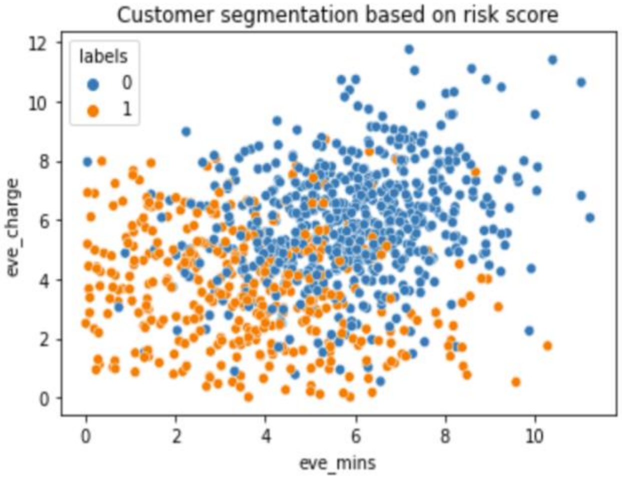


Fig. 19 Customer segmentation on test data using K-Means

Observation: The customer base is distinctly separated into lower usage in evening time and higher user in evening time. Based on customers' behavior an efficient strategy could be proposed to suit customer profile.

VII. BROADER IMPACT OF RESEARCH RESULTS

- **Churn Prevention:** Based on the model risk-score of customers, we can identify the demographic of customers and their point of time in future when they are likely to leave. This will significantly ameliorate the attrition rate of customers and help the organization in devising right strategies and plans to move forward in right direction.

- **Exponential Growth with increased revenue:** Retained customers can drive up the profits of a company by promoting the products and spreading awareness in the market. This accelerates the rate of growth for a company and save efforts invested in the form of endorsements and sales. Word of mouth referrals will bring in more customers in less span of time and will optimize expenditure cost for the company as well.

- **Customer Segmentation:** Every company is in constant pursuit of improvement in terms of policies or strategies that could raise the bar of quality delivered to their clients. And what could be a better approach than reaching out to the dissatisfied customers. Through this model, we can categorize and reach out to customers that are likely to leave for their constructive feedbacks. Such surveys will assist the firm in recognizing the loopholes and will pave way for bringing about a change in the system.

- **Consistent Sales:** When it comes to rolling out a new product in the market, companies rely on their loyal fan base for the success of that product. Since new customers are oblivious or apprehensive to try, it is this fan base that instills confidence and trust in the quality. Apart from it, loyal customers also boost the sales figures as the probability of purchasing breadth of products is higher.

- **Acquisition is expensive:** Targeting new customers and onboarding is not a cheap process and entails a lot of different stages. Companies often focus on simply adding subscribers without realizing how much they are spending in the process. Therefore, this model will ensure that only essentials funds are allocated for acquisition of customers and rest are invested in right departments.

- **Upsell and Cross-sell:** Customer retention presents the organization with the opportunity of upselling and cross-selling services to their customers. For instance, by collecting data on the customers activities, we can analyze, predict and offers services based on customer needs. This will help in articulating the customer needs and offering services that meet those requirements. Eventually, it will raise the revenue of the company by driving up sales figures by upselling & cross-selling products and services.

- **Brand Value Creation:**
  Of late, there is a lot of buzz around creation of brand value among customers. Clients tend to feel pride in using services of a brand that is well reputed. Customer retention leads to a higher brand value in the market of an organization because their consumers feel as a part of a family. This emotional engagement creates a strong sense of community which is not only forgiving in the long run but also strongly advocate the services of the organization in the society.

- **Outshine Competition:**
  In order to outperform the competition, corporates are always in the race to win over the market share. They resort to planning, strategizing, and testing their produce before rolling out so as to live up to their customer expectations. Hence, organizations, that are able to hold on to their clients strongly over a period of time, prove their excellence in the quality of service. Customers only stick to products that excel their expectations and meet even the unarticulated needs.

VIII. Discussion

Summary

This paper addresses the customer attrition problem of a telecom organization by predicting the churn and probability of churn of each customer for next 200 days, which can help the business to chalk out customized approaches to mitigate the churn by intervening at appropriate time. The organization can launch Customer Satisfaction (CSAT) surveys to the group of customers based on their churn probability beforehand to understand their exact problems. Business also can coordinate with organization's customer call centers to see if high risk customers have reached out to call center and check whether their issues have been addressed or not.

We have used survival function with XGBoost to model the data, the performance metrics of survival function like brier score and concordance index have optimal values for RFE based feature selection.

Once the probabilities of churn have been predicted, we have segmented the customers basis the risk score and observed that the churn rate is higher in customers with longer call duration one probable reason could be that they are not satisfied their current talk time charges as these customers are frequent users, the company has to devise new plans according to the average talk time of customers.

Future work

As we know that customer attrition is very dynamic and needs real time attention to prevent any losses, we need to automate the entire process of data collection, cleaning, model implementation, insights generation and recommendations.

To achieve the above objective, the organization should have to build a robust data pipeline for real time data collection and good tech infrastructure (production servers) to process the data. We can integrate data collection process with data processing, modelling and visualization components for automated insights generation. The models performance can be evaluated in real time and can be improvised if necessary.

As a future work, we can demonstrate the above capability through proof of concept (POC).

From domain point of view, we can integrate analysis like Customer attrition, CSAT (Customer Satisfaction) scores, NPS (Net Promoter Scores) and organization's call centers operations as these very much inter linked. Functioning of call center plays crucial role in retaining the customers as they are first line of defense because the customer reaches out call centers most of the time for any issues that they are facing. We need to ensure that call centers operate effectively by meeting targeted AHT (Average Handling Time)

References

https://www.huify.com/blog/acquisition-vs-retention- customer-lifetime-value

https://smallbiztrends.com/2014/09/increase-in-customer- retention-increases-profits.html

https://www.elasticpath.com/blog/customer-acquisition-vs- retention-infographic

https://www.theacsi.org/national-economic- indicator/national-sector-and-industry-results

Raw Data Link

https://github.com/awslabs/aws-customer-churn-pipeline/tree/main/data