2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)
Amity University, Noida, India. June 4-5, 2020

Application of Machine Learning to Predict Hospital Churning

Shweta Chauhan
Amity Institute of Information Technology, Sector-125, Noida, U.P
shweta3101995@gmail.com

Sonia Saini
Amity Institute of Information Technology, Sector-125, Noida, U.P
ssani2@amity.edu

Ruchika Bathla
Amity Institute of Information Technology, Sector-125, Noida, U.P
rbathla@amity.edu, bathla.ruchika@gmail.com

Ajay Rana
AIIT, Amity University Uttar Pradesh
Noida, India
ajay_rana@amity.edu

***Abstract*- Nowadays, with options in health amenities, healthcare is going through a digital transformation to find ways to be able to retain patients and reduce patient churn, which consequently is crucial for their success and growth. In the healthcare system, patients are seen as treasured customers, and hospitals are trying to re-orient around providing enhancements in treatments for achieving this, but patient churn is an ongoing challenge for this. Therefore, hospitals are increasingly relying on AI and Machine Learning based insights to find those factors that are most likely to lead to patient churn and then take appropriate actions on it to reduce them. To predict the churn ratio is the main aim of our work which will assist hospitals to identify the patients who are most likely subject to churn. We identify the factors to predict the churn ratio and see how effectively we can use machine learning techniques which will help us to strengthen prediction performance by identifying, attract, and retain preferred patients.**

Keywords- Churn, Machine Learning, Patient churn,Churn Prediction, Readmission.

I. INTRODUCTION

Patient churn is a basic problem for hospitals which means loss of patients because they move out to competitors. Examining the history of a patient's behavior including various other affecting variables will help us determine what factors increase patient churn.

A. Churn Prediction

It is the method of predicting, which users will stop using the platform. These predictions are utilized to pro-actively take retention actions on churning customers by marketers.

Firstly, we will get to know why patients are leaving and then we will take appropriate actions to prioritize patient engagement and satisfaction to stop them before its too late [4].

The work done in the TME industry can benefit health care sector to reduce churn and strengthen business performance by identification, attraction and retention of patients.

B. Identification of patients

To better recognize patient populations and then provide better care to them health care researchers had been segmenting the marketplace for decades .To pick out satisfactory clients, Organizations want to start with growing significant market segmentations to pick out the best quality and sub-corporations of clients that force the enterprise to key exercise areas [6]. Meaningful market segments for enhancing acquisition and churn may be established to respond otherwise to marketing variables inclusive of price, promotion, channels, and many others. This allows the organization to check a number of incentives and strategies to optimize the overall performance through the years.

C. Attract New Patients

Compiling preferred patient profiles allows the organization to transport on to the assignment of tailoring messaging and advertising greenbacks to target audiences to draw their enterprise. Here the task is figuring out which participants of these desired consumer segments fall in the maximum "winnable" [5]. Criteria will vary from segment to segment, but most organizations would be conscious on clients who're more likely to:

- Benefit from online resources associated with their health care, options for brand new remedies, or value saving ideas
- Be referred to your exercise via every other doctor or organization
- Respond to surveys regarding care and the treatment by the service provider
- Bring or refer circle of relative's participants in your facility [7].
- Conduct studies and comparison on optional surgeries and services earlier than selecting their carrier issuer e.g., knee or hip joint substitute

D. Retaining Patients or Reducing Churn

In the TME enterprise, lowering churn is all about analytics. They put money into quantity-crunching algorithms to arm their retention teams with the facts they want to recognize which customers are most valuable and what will work excellent to preserve them [8]. As in health care, there was a wave of investment in constructing large, luxurious internal guide systems, including: data lakes, new warehouses, and new software and hardware, to perform those superior analytics. However, building information analytics internally take giant time, finances and human resources, and that first wave of funding gave way to a 2nd wave of TME players going with an Insights-as-a Service model, wherein rather than spending to construct their personal potential to reply questions

about patron conduct, they paid 0.33-events for the solutions [1].

Analyzing and understanding the healthcare customer in a mile deeper and extra unique manner is critical to a goal favored potential and existing clients primarily based on their desires and traits. The TME industry has confirmed a movement of this type of insight to broaden greater, and effective outreach campaigns, increase models to decrease the churn charges of present clients, and interact with patients on their cases. The quicker the groups can apprehend and adapt those strategies for the health care market, the bigger an advantage they'll have on this rapidly evolving and evermore purchaser-centric industry [3].

II. STEPS FOR PREDICTING CHURN PREDICTION

A. Choosing a dataset

Before performing any machine learning technique, one of the major tasks is to find a good dataset. In today's healthcare world we can easily find datasets that contain useful information but have messy schemas or unstructured content or datasets that are very clean but contain sterile information [2]. Using a dataset from UCI repository containing Personally Identifiable Information (PII) removed from it.This is diabetes patient data from various U.S. hospitals (1999–2015) which containing 91000 observations over a 16-year period. The dataset contains 54 features which include patient conditions, characteristics, 23 medications and tests. Only diabetic data are included. We use this dataset to infer the likelihood of diabetic patient churn during the treatment.

B. Methodology

Before obtaining actual modeling, some data wrangling was required. We implemented 3 types of techniques for achieving this:

- Cleaning tasks such as dealing with missing values and dropping unwanted data.
- Modification of existing features.
- From existing features creation or derivation of new features.

1) Dealing with missing values: Firstly, we find out missing values that are presented into dataset

The variables with the missing values do not contribute to the results and hence we removed these values from the dataset.

Missing estimation of variable weight is over 98%. In view of exceptionally helpless interpretability of missing qualities and minimal prescient generalizability to different patients, the best thing is to simply drop it.

Clinical Specialty and Payer code of rewarding doctor likewise have 41–51% missing qualities. Thus, we chose to drop these, yet additionally there are some different approaches to manage these missing qualities.

Essential (diag_1), Secondary (diag_2) and Additional (diag_3) analyze were having not very many missing qualities.

In fact, if key highlights are missing qualities, that is develops as terrible information. Along these lines, we just drop those records where every one of the 3 judgments are absent.

Likewise, another cleaning step that thoroughly relies upon understanding the brought information and is that: since we are attempting to anticipate readmissions, and refute the zero likelihood of readmission, we sift through those patients who kicked the bucket during emergency clinic confirmation. In this way, we should evacuate those records (release air = 11).

We additionally saw that for two medications, named citoglipton and examide, all records had a similar worth. Basically, these can't give any biased or interpretive data for anticipating readmission, thus we drop these sections too. Actually, this is a missing data issue not a missing worth issue [9].

2) Collapsing of Multiple Encounters for same patient: A few patients in the dataset had more than one experience. We were unable to consider them autonomous experiences. Along these lines, we attempted multi-strategies to merge and crumple multi-experiences for same patient, for example,

- Considering multiple readmissions across multi-experiences as readmission for fallen record.
- Considering normal remain at emergency clinic across multi-experiences.
- Considering the level of the drug changes across multi-experiences
- Considering the all out number of the experiences to supplant the experience one of a kind ID
- Considering the blend of findings across multi-experiences as a rundown

In any case, taking the highlights, for example, 'analysis', for example, we didn't discover this as significant to join multi-straight out qualities into an exhibit for building an information model. We at that point considered first experience and last experience independently as potential portrayals of multi-experiences. Be that as it may, last experiences gave amazingly imbalanced information for readmissions (96/4 Readmissions versus No Readmissions) and consequently, we chose to utilize first experiences of patients with multi-experiences. This will be decreased to around 60,000 experiences into the dataset.

3) Standardization: We chose to normalize our numerical highlights utilizing the equation:

New value = (Value – Mean (Values))/ (Standard Deviation (Values))

4) Removal of Outliers: We guaranteed that all the taken factors were almost ordinary after log change, the dispersions should be less or progressively typical. For ordinary conveyance by utilizing the inclusion rule, which can be summed up by the outline given below:[12]

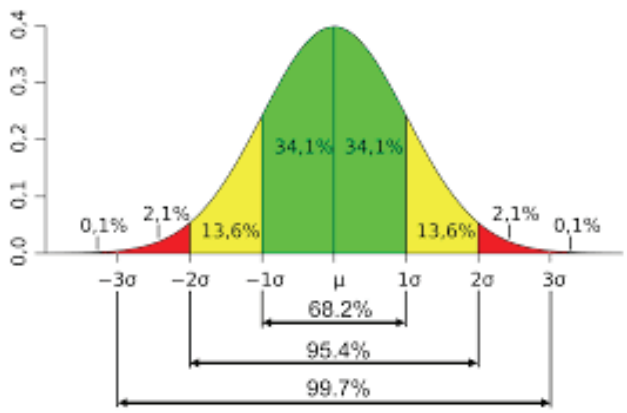

Fig. 1. Standard deviation diagram

Anything inside 3 Standard Deviations on either side of the mean would incorporate 99.7% of the information and the rest 0.3% we can treat as exceptions. Utilizing this rationale, we limited the information to inside 3 Standard Deviations on either side from the mean for each numeric section.

5) Data Balancing: Information is exceptionally imbalanced w.r.t readmissions (just 10% records for 30-day readmissions), prompting high exactness. In addition, the high exactness could be ascribed not to the generalizability of the model to assorted patient records yet to the pattern precision of 90%: anticipating that no patient would be readmitted. This was clear from the helpless accuracy and review of the model in foreseeing persistent readmissions. We utilized manufactured minority over-examining method (SMOTE) to oversample the underrepresented class of readmissions and acquire equivalent portrayal of the overrepresented and underrepresented classes [10].

The underneath screen capture of the forecasts from the model when information adjusting shows the impact as altogether lower type 2 errors:[12]

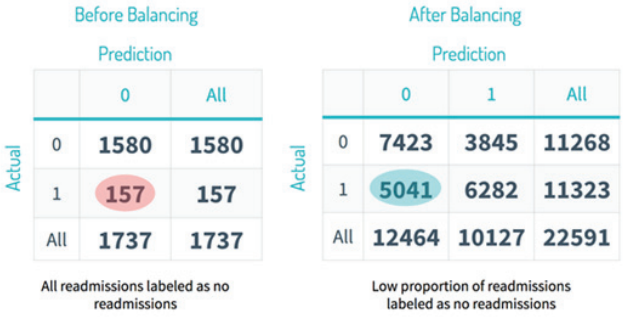

Fig. 2. Balancing prediction [12]

As should be obvious, the sort 2 diabetes related re-affirmation blunders are essentially decreased post adjusting, showing a superior review for the model. This means a lower extent of readmitted individuals anticipated as not readmitted: a basic measurement for clinics and protection offices both from a money related and preemptive human services viewpoint.

6) Feature sets for modeling: While there are such a large number of potential combinations of highlights one could test for, we needed to test two forms — one moderately an easier model with less and some composite highlights we had made and a perplexing one containing point by point highlights for example supplanted with composite element in the basic list of capabilities and number of outpatient, crisis visits and inpatient were remembered for complex element set[12].

Presently, every one of these highlights are available in the information outline. Along these lines, in python, characterizing a specific arrangement of highlights is basically proclaiming a rundown variable containing those segments.

At that point, at whatever point for demonstrating we have to utilize a specific set, we simply need to choose that arrangement of segments from the total information outline.

7) Data Split and Ensuring Robust Predictions: We randomized and partitioned the spotless information that we acquired from the procedure information cleaning into two sections: Test and Training Data, in a proportion of 20:80, which permitted us to utilize the 20% of the information to evaluate the exhibition of models and rest 80% to prepare models. To guarantee that models were both powerful and versatile and weren't over-fitting on the preparation information, so we utilize 10-overlap cross-approval on preparing information. Likewise, we guaranteed that models were assessed and prepared on similar information to guarantee execution correlation across models for precisely the same information being assessed and prepared upon.

C. How to obtain and interpret the results of models

1) Logistic Regression: For testing any model in a hearty manner, we simply need to cross-approve and analyze test versus train exactness and different measurements. We are utilizing a basic strategy from statsmodel.api to show the particular results of a calculated relapse [7].

Since there are such a large number of coefficients and factors to see, we are getting just those coefficients that have p-esteem < 0.01 and have in any event 0.2 extents to make it somewhat simpler.

As a general note for calculated relapse coefficients announced, the translation is to be done while considering both the changes we have applied to the information before displaying and the idea of strategic forecast. For instance, in deciphering the impact old enough on readmissions, we may follow these means:

EXP (Coefficient old enough) = EXP (Log of unit chances change) = EXP (0.25) = 1.28

Be that as it may, recall Age was normalized and 1 SD of Age = 15 years

In this manner: For at regular intervals increment in age, there is 28% expansion in Odds of being readmitted as opposed to not being readmitted!

Looking at results from mind boggling and straightforward capabilities, we can make sense of that the coefficient esteems are close enough with the exception of that intricate set gives us some extra coefficients for a few medications. It is to be deciphered that after a great deal of pre-preparing of the information we have arrived at these numbers and their translation ought to consistently be while taking other factors into consideration.

Rather than simply taking a gander at coefficients, we likewise need to perceive how well the model is acting as far as mistakes and exactness. For this reason, we test it by utilizing 20% test, and furthermore utilize a 10-crease cross-approval. We can likewise utilize scikit-learn measurements and library and take a gander at the disarray lattice, to compute different proportions of precision.

TABLE I. RESULT OF LOGISTIC REGRESSION

Cross Validation Score	62.10%
Dev Set score	61.62%
Precision	61%
Accuracy	61%
AUC	60%
Recall	56%

As should be obvious, with a Recall of 55% the model precision is 61% and Classification blunder is of 39% for complex list of capabilities. Utilizing the decreased list of capabilities comparative examination gives nearly similar outcomes. Now, it appears as though model isn't performing admirably as far as blunder and precision measurements.

2) Decision Trees: We ought to preferably expel the collaboration factors from the list of capabilities since this strategy endeavors a wide range of connections between factors characteristically [4].

At that point utilizing scikit-learn we applied a choice tree classifier.

TABLE II. RESULT OF DECISION TREES

Cross Validation Score	91.03%
Dev Set score	90.17%
Precision	92%
Accuracy	90%
AUC	90%
Recall	89%

Contrasted with strategic relapse here in all measurements we can see a lot of progress. For tweaking there are a few model boundaries, for example, kind of data gain work (gini versus entropy), least example size, profundity of tree, etc. To locate the ideal boundaries for best execution one ought to in a perfect world actualize a matrix search.

In any case, that is not everything we can get from choice trees. Every hub of the tree into choice trees demonstrates an information parting choice that gets us closer to anticipating readmission. The test is imagining profound trees with loads of hubs.

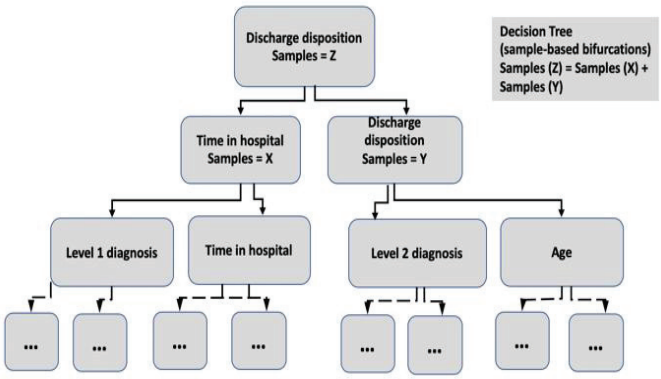

Fig. 3. Decision Tree

By looking the graph [12], we can say that the first component utilized in choosing is whether the patient is released to another clinic or office or whether a patient will get readmitted or not. Trees next level shows this equivalent element rehashed on one hand and days spent in emergency clinic on the other. However, this can undoubtedly get befuddling, so we summed up this data in such manner. "Highlight Importance" is one normal approach to do this.

Highlights having significance higher than 0.01 are thought of, and in this, the time went through in clinic with most noteworthy estimation of 0.41 shows up as the most noticeable component. These are relative numbers, determined by contrasting the presentation of the model and without a given element. Along these lines, this would imply that in the event that we evacuate the time spent in medical clinic from the model, we lose the most prescient worth.

3) Random Forest: Into this technique as opposed to relying upon a solitary choice tree, we attempt a wide range of trees with arbitrarily alloted subsets of highlights. By utilizing casting a ballot across expectations made by all the trees in the backwoods the last forecast is determined [11].

In the wake of applying this technique, we get an exactness of 94% paying little heed to utilizing Gini or Entropy work. Highlight significances are just marginally extraordinary utilizing these two capacities, and we show the one with Gini work beneath:

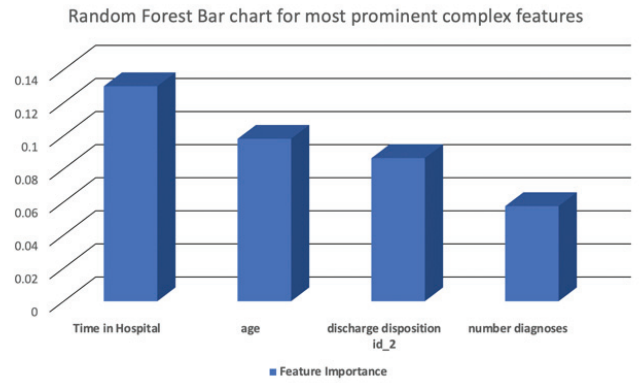

Fig. 4. Random Forest Bar Chart for complex features

At the point when we rehash this procedure for the less difficult model, we get comparative precision i.e., around 93%, yet contrasted with the unpredictable model the top component significance's varied somewhat:

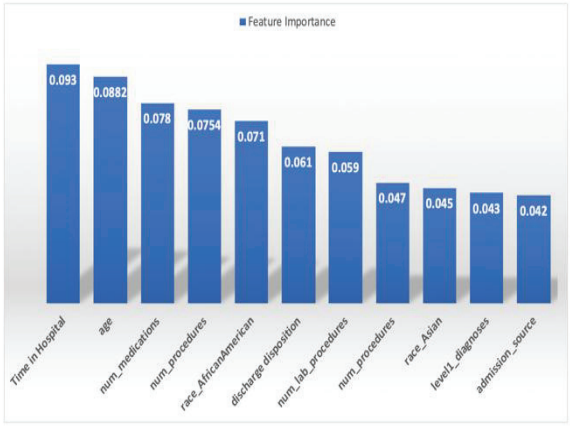

Fig. 5. Random Forest Bar Chart for simple features

III. CONCLUSION

The choice tree model shows profoundly significance of time spent in the age, medical clinic and release to another emergency clinic for both straightforward form just as unpredictable rendition. In the event that we plot these two against the coefficients from strategic relapse, they don't connect completely. Thus, we utilized cross-approval and accomplished same test and train exactness to keep away from over fitting; this may recommend a lower clarified difference in calculated model or distinctive relationship structure of the factors in the model. Similar highlights had high significance for arbitrary woods gathering, despite the fact that the dispersion was more in contrast with choice tree. This is because of adjustment of significance across numerous trees.

REFERENCES

[1] T.VafeiadisaK.I.DiamantarasbG.SarigiannidisaK.Ch.Chatzisavvasa, ”A comparison of machine learning techniques for customer churn prediction”, Simulation Modeling Practice and Theory Volume 55, June 2015, Pages 1-9.

[2] Abdelrahim Kasem Ahmad, Assef Jafar & Kadan Aljoumaa, “Customer churn prediction in telecom using machine learning in big data platform”, Journal of Big Data volume 6, Article number: 28 (2019)

[3] Ramsey, Gregory & Bapna, Sanjay. (2019). “Predicting Patient Turnover” 10.4018/978-1-5225-8244-1.ch006

[4] Akshara Santharam, Siva Bala Krishnan,” Survey on customer churn prediction technique”, International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395-0056 Volume: 05 Issue: 11 | NOV 2018 p-ISSN: 2395-0072

[5] Anne Merel Sternheim,” Predicting Patient Churn: Features that Predict when Breast Cancer Patients Leave their Online Community”, Utrecht University, July 9, 2018

[6] Syed Muzamil Basha, Avanti Khare, Jyothi Gadipalli,” Training and Deploying Churn Prediction Model using Machine Learning Algorithms” International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 5, Issue 4, April 2018.

[7] Tan Yi Fei; Lam Hai Shuan; Lai Jie Yan; Guo Xiaoning; Soo Wooi King,” Prediction on Customer Churn in the Telecommunications Sector Using Discretization and Naïve Bayes Classifier.”, International Journal of Advances in Soft Computing & Its Applications. Dec2017, Vol. 9 Issue 3, p23-35. 13p.

[8] Sabbeh, Sahar F. “Machine-Learning Techniques for Customer Retention: A Comparative Study.” (2018), Procedia Computer Science, Volume 141, pages 484-489.

[9] Malathi, D. and R. Mythili. “Preserving the Data Privacy and Prediction of Hospital Readmission using Machine Learning in Data Mining.”, ISSN 2321 3361 2019 IJESC, Volume 9 Issue No.9.

[10] Bhardwaj, Rohan et al. “A Study of Machine Learning in Healthcare.” 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC) 02 (2017): 236-241.

[11] Khan, Atiya and Tabrez Nafis. “Neoteric Breast Cancer through Machine Learning Algorithms.” (2018), International Journal of Computer Applications (0975 – 8887) Volume 179 – No.49, June 2018.

[12] Usman Raza, Harman Shah Singh, Ching-Yi Lin and Rohan Kar. “Predicting hospital readmission using machine learning.”2018, www.medium.com