CUSTOMER CHURN PREDICTION USING K-MEANS CLUSTERING AND ARTIFICIAL NEUTRAL NETWORK(ANN)

TEJASWINI KUNDENA
Dept of Computer Science
Texas A&M University, Corpus Christi
Corpus Christi, USA
tkundena@islander.tamucc.edu

SANDEEP MUDHURAKOLA
Dept of Math - Statistics
Texas A&M University, Corpus Christi
Corpus Christi, USA
smudhurakola@islander.tamucc.edu

Abstract

This research paper focuses on a customer churn prediction model, crucial for mitigating the risk of losing existing customers in today's competitive business landscape. Even with an annual churn rate below 10%, failing to implement effective strategies can have long-term negative impacts on an organization. This model aids stakeholders in developing customer-specific resources to reduce churn. While applicable across various industries, this study uses a telecom dataset, where customer loyalty is fragile, and even minor dissatisfaction can lead to switching providers.

Our main motive is to create a predictive model that can identify the customers who are most likely to leave from the organization. so we use 3 main steps in this project. i) data cleaning mining techniques, ii) Clustering using k-means iii) Model selection and evaluation. Basically, in First step we clean the raw data and filter the data. In the second step, we cluster the data according to the behavior patterns and demographic traits, K-means clustering is used to produce clusters that represent the root causes of the churn patterns. In the third step, we are using Algorithms like Logistic Regression, Random forest, SVM along with XGBOOST classifier using ANN. ANN is used to forecast churn in these clusters by utilizing the subtle insights that the clustering procedure provides. This study presents a unique technique for customer churn predictions with good accuracy and performance.

Overall, our study reveals that including clustering along with the deep learning algorithms is effective in forecasting customer turnover and stresses the benefit of using advanced analytics to create insights that can be used for relationship management and client retention.

I. INTRODUCTION

In this rapidly changing business world, losing customers is a major issue that inclines business revenue to drop

eventually if proper strategies and precautions are not followed. Because of the proliferation of digital platforms and subscription-based services in recent years, businesses in a wide range of sectors have realized the importance of customer retention. services as well as cutthroat marketplaces. Customer churns the phenomenon wherein consumers stop doing business with a company—poses a significant challenge to businesses striving to maintain profitability and expand. Therefore, the ability to foresee and avoid client attrition has emerged as a vital strategic

necessity for firms trying to compete in a market that is growing more and more focused on the requirements of its clients
To take precautions, Firstly We should understand the

customer churn. We need to have a rough analysis

of why customers are discontinuing from the organization. Need to figure out the key factors by using the data set and how the customer churn is going to affect the organization long-term. we should be able to predict the customer churn rate, In order to predict the churn we need to know the key factors and customer behavior patterns, to know that we need to analyze the data set. Upon reducing churn it improve the lifetime value of the organization.

As we have discussed the business value of churn prediction, we can also able to recognize how the churn prediction model plays in resource allocation and how can companies recognize clients that pose a high risk and finally businesses can try to improve personalized engagement with clients to prevent churning.

The following main ideas emphasize the importance of customer churn prediction:

A. **Financial Impact***: Losing customers might result in decreased revenue and customer acquisition expenses (CAC). Generally speaking, acquiring new clients is more expensive than keeping hold of current ones, hence it's critical to recognize and successfully handle customer churn[1].*
B. ***Strategic Decision-Making:*** *By using data-driven, well-informed churn prediction, firms may improve their client retention tactics and lower attrition rates. Organizations may see trends and take proactive steps to reduce attrition by examining churn data and key performance indicators[1].*

C. ***Resource Optimization:*** *By concentrating on keeping at-risk clients rather than just bringing on new ones, companies may allocate resources more effectively by anticipating customer attrition. Both cost savings and increased overall performance may result from this[1].*
D. ***Industry Relevance:*** *Retail, banking, e-commerce, telecoms, and other businesses may all benefit from the use of customer churn prediction. It is an essential tool for comprehending customer behavior, pinpointing problems, and putting focused retention measures into action[2].[3]*
E. ***Benefits of Machine Learning****: Applying machine learning to customer churn prediction has several benefits, including the ability to identify at-risk clients, monitor and lower churn rates, and pinpoint pain spots that lead to client attrition[4].*

II .RELATED WORKS:

While going through the studies we found lot of different implementations for this idea but by using various algorithms, classifiers ,machine learning, deep learning and data mining techniques. When we are discussing customer churning, we need to consider the key reasons for the churning.

According to the [1] The key reasons are :

Abandoned/ Cancelled subscriptions

Probably the first type of churn that comes to mind are cancelled subscriptions, which can occur for a variety of reasons.

1. poor match with the client[1]

If you market to the incorrect kind of consumer they may leave your site as soon as they sign up, negating the expense of the resources you spent to get them.

2. Absence of functionality [1]

When their demands change, customers could seek for more features in your product. One certain way to lose customers is to give them the impression that your business is complacent about its offerings and isn't adapting to their evolving demands.

3 Not Achieving the desired result [1]

Consumers who don't obtain the desired results from your product will undoubtedly stop using it. This type of turnover indicates a significantly

4. Switching to competitors[1]

Customers will examine the products and services offered by competitors or other companies to find out why clients would choose to do business with them. If others are providing better deals, cheaper pricing, or more services for free, customers may decide to switch.

5. ***Non-Renewal of Subscriptions:[1]***

Churn can be caused by low customer engagement, which makes consumers forget the benefits of a product and decide not to renew. Furthermore, overdue attrition resulting from payment problems, such as expired credit cards, can represent as much as 40% of total churn in SaaS firms, and lost clients are seldom won back.

6. Account Closure*:[1]*

Even satisfied customers may leave, resulting in lost revenue and replacement costs. To reduce this type of churn, broaden your product offerings or increase the value of your products for recurring use.

According to Ning Lu [5], customer churn prediction models that cluster consumers depending on the weight of the boosting algorithm can be enhanced by employing boosting algorithms. A cluster of high-risk customers was discovered. Every cluster has a churn prediction model, and the learner is logistic regression (LR). The outcomes shown that boosting algorithms are superior to single logistic regression models in the ability to segregate churn data. Comparative investigation was done by H. Karamollaoğlu et al. to forecast the excellent f1-score in different ML approaches[6]. Consideration was given to AdaBoost, Decision Trees, Multilayer Sensors, Logistic Regression, etc. Ultimately, the author came to the conclusion that the random classifier produced the best result without the need for any data.

augmentation techniques. The author employed arbitrary sampling to enhance the SVM model by taking data imbalance into account and developed a CC prediction approach based on SVM model [7]. For classification, an SVM constructs a hyper-plane in a high- or infinite-dimensional space. Data dispersion can be altered via arbitrary sampling to lessen dataset imbalance. Dataset imbalance is caused by fewer churners.

[9]Although support vector machines (SVM) are sometimes criticized for being "black box" models, they are incredibly successful at classification. In order to extract intelligible rules for customer relationship management (CRM) using support vector machines (SVMs), this research suggests a hybrid technique. The method consists of three stages: training an SVM model on the reduced dataset and extracting support vectors, constructing rules using Naive Bayes Tree (NBTree), and feature reduction using SVM-recursive feature elimination (SVM-RFE). Using a highly imbalanced dataset, the study focuses on attrition prediction among bank credit card consumers. When standard balancing procedures were used, the hybrid strategy performed better than the other approaches. As a result, the regulations are clearer and shorter, giving bank management an early warning system.
[10]As to the study's conclusion, there have been 1305 publications published on the customer churn construct till the end of 2020, indicating that it has been a significant research topic in the past ten years. The construct is an international phenomena that has drawn attention from specialists worldwide for the past ten years, as evidenced by publications from 2017 to 2020. The analysis shows that there are papers on customer turnover accessible in a number of disciplines, but "business management" articles lead this field of study as predicted. Academicians might use the study as a quick reference if they wish to investigate the customer churn construct further.

These are the related works with similar idea's but different implementation techniques.

III. System Design & Architecture

Here is the flow diagram of our system showcasing the flow of data from data collection to the predicted output of the test and train data sets. Initially a large data set is taken related to a tele communication network customer churning.

Basically we are going to showcase a predictive model to predict customer churning using the lasted data set but implementing it with a unique combination of data mining techniques integrated with XGBOOST classifier , and comparing the accuracy and performance of machine learning algorithms . Here is the Reference diagram how the prediction is done using the predictive model and how we analyze the data.

1. ***Commencing the process***
Businesses must anticipate customer attrition in order to retain their clientele. Our goal is to create an intelligent system that can use the data we have about our clients to predict which ones they would visit. This plan includes data cleansing and usefulness testing, system training, system performance monitoring, and selecting the optimal approach through comparative analysis of available possibilities. The raw data set is retrieved from Kaggle.com and processed as steps mentioned below.

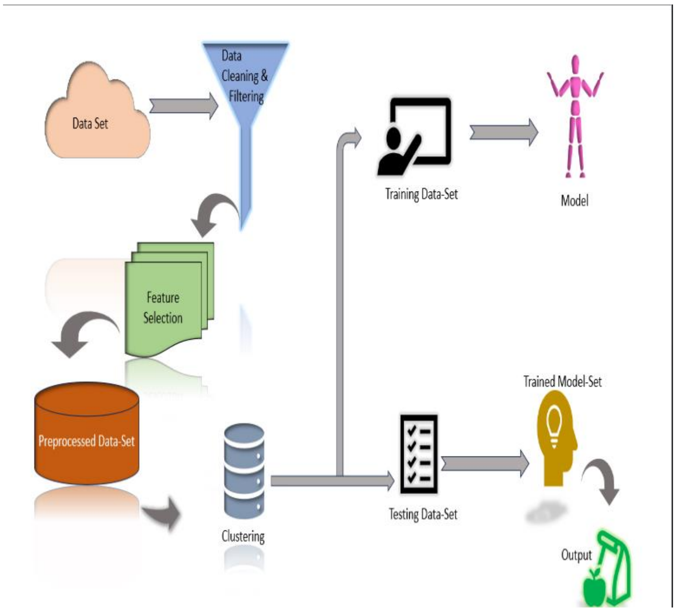

Fig1 :Flow Diagram

2.Data Cleaning

In this step we are going to handle the missing data, Addressing class imbalance.

- *Missing Information:* 'TotalCharges' will be converted to numbers, mistakes will be changed to NaN, and any remaining NaNs will be eliminated.
- *Eliminating Pointless Items:* The 'customerID' field will be removed as it has no predictive value.
- *Words to Numeric Conversion:* LabelEncoder will be used to convert textual data into machine-readable numbers.

3.Improving the Utility of the Data

- Obtaining New Information *TenureByMonthlyCharges*: This calculates the entire value a client adds to the business over the course of their partnership by multiplying "tenure" by "MonthlyCharges."
- The period of a customer's tenure multiplied by their monthly costs is how this feature is computed. The resultant value offers a thorough assessment of the full monetary worth that a client gives to the company throughout the course of their relationship. This feature gives the model a more detailed picture of each customer's economic effect, which is important for anticipating customer attrition. Consumers with large 'TenureByMonthlyCharges' values are important revenue generators, and losing them might have a big effect on the company. As a result, this function aids in locating valuable clients who may require further care to keep.
- ***K-Means Clustering:*** Customers may be grouped using K-Means clustering on the normalized characteristics. Add the newly obtained cluster labels as a new feature.

K-Means clustering is an effective method for dividing up the customers into groups according to their traits and actions. It is possible to identify different customer groups by using K-Means clustering on the normalized attributes of the customer dataset. Every cluster denotes a collection of clients with comparable characteristics and actions. The steps in the procedure are:

(i) ***Normalization***: Make sure all of the characteristics contribute equally to the clustering process by standardizing them.

(ii) ***Clustering***: Using the normalized data, run the K-Means technique to find clusters.

(iii) ***Cluster Labeling***: Utilizing the K-Means algorithm's output, assign a cluster label to every client.

4. Model Assessment and Training

Normalization and Data Splitting:

- ***Data Splitting***: The dataset should be divided into training (80%) and testing (20%) sets in order to appropriately assess the model's performance. The testing set evaluates the model's performance on untested data, whereas the training set is used to develop and train the model.
- ***Normalization:*** ensures that each feature contributes equally and enhances model performance and convergence by adjusting the feature scale to have a mean of 0 and a standard deviation of 1.

To train models effectively and evaluate performance accurately, proper data partitioning and standardization are necessary

Classifiers:

- To forecast client attrition, we will employ and assess a range of classifiers, such as:

  k-nearby Neighbors (k-NN): Assigns classifications according to the nearby neighbors' majority label.

- Based on the Bayes theorem, the Naive Bayes classifier uses probability.
  Optimized gradient boosting method, or XGBoost.
- *Random Forest*: An ensemble technique for creating many decision trees.
- Modeling a binary dependent variable with a logistic function is known as logistic regression.
- GridSearchCV will be utilized to do a thorough search and cross-validation in order to determine the optimal hyperparameters for every classifier.
- **Metrics for Evaluation:**
  Accuracy: The proportion of cases that are classified.
- *F1 Score*: The average of recall and precision that is balanced.
- Precision is defined as the ratio of all positive forecasts to genuine positive cases.
- *Remember*: The proportion of real positive examples to all true positive examples..

5.Evaluation&Illustration

Comparing Models:

Classifiers may be evaluated based on recall, accuracy, precision, and F1 score.
To facilitate comparisons, display assessment metrics as a heatmap.
Show the significance of each feature for the Random Forest and XGBoost classifiers to determine their relative contributions.

IV DISCUSSION

Model Performance:

The report has expounded on how effective it is to integrate K-Means clustering with machine learning algorithms like XGBoost in predicting customer's churn. K-Means clustering aided in grouping customers given their characteristics, which both lead to the enhancement of the model's performance.

Feature Importance:

The important features identified in this case were "TenureByMonthlyCharges," "MonthlyCharges," and "tenure." This clearly brings out the importance of customer tenure and monthly charges as important predictors for base churn. Train the XGBoost with feature parameters, sorting them in descending order and plotting them

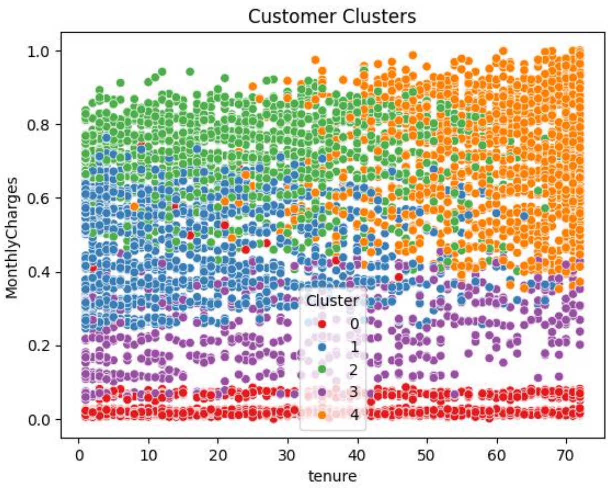

Fig 2: k-means Clustering

Description: Customer clustering based on tenure and monthly charges. These clusters were created using the K-Means clustering algorithm, placing customers into clusters of similar characteristics.

Clusters Identified:

Cluster 0: Low monthly charges changing tenure are shown in red.

Cluster 1: In purple, it has low to moderate monthly charges with changing tenure.

Cluster 2: This category, in orange color, contains those customers whose monthly charges fall in the middle range with different tenure. Cluster 3: High-paying, highly varying customers by month are represented by the blue color. Cluster 4: This was the green category of very highly paying, highly variable customers by month. Monthly Charges:

- The y-axis ranges from approximately $20 to $120 monthly charges.
- The people with higher monthly charges range across all the ranges of tenure. Tenure:
- The x-axis denotes tenure in months, ranging from 0 to 72.

Clearly observable is that there is a wide range of tenures within a cluster denoting customers who are both short and long-term committed.

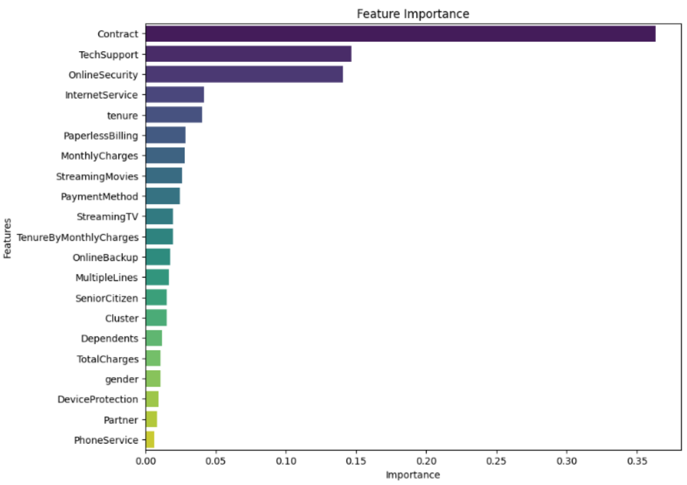

Fig 3:Feature Importance

It computed the feature importance in terms of how many times it has been used to split the data across all trees within the model. This metric, "weight" in XGBoost, explains the relative importance of each feature in making accurate predictions.

Key Features:

These were the most important features picked up to predict customer churn:

MonthlyCharges: This attribute was the most important, indicating that what a customer pays per month is very critical in predicting churning. If a customer has higher monthly charges, he/she is more likely to churn.

Tenure: Tenure, or how long a customer had been with the company, was another important attribute. The customers who have less tenure are more prone to churning. Hence, this reflects that newer customers may not have built their loyalty to the service.

TotalCharges: The total amount a customer has been charged over their tenure also turned out to be a very influential factor, as customers with higher total charges are at a greater risk of churning.

Contract: Another highly influential variable in this context was the type of contract. Month-to-month customers have been found to have a greater tendency to churn as compared to those having yearly or longer contracts.

PaymentMethod: The payment method also showed—a method of payment being an electronic check or credit card—some of the ways that were likely to churn

Correlation Matrix:

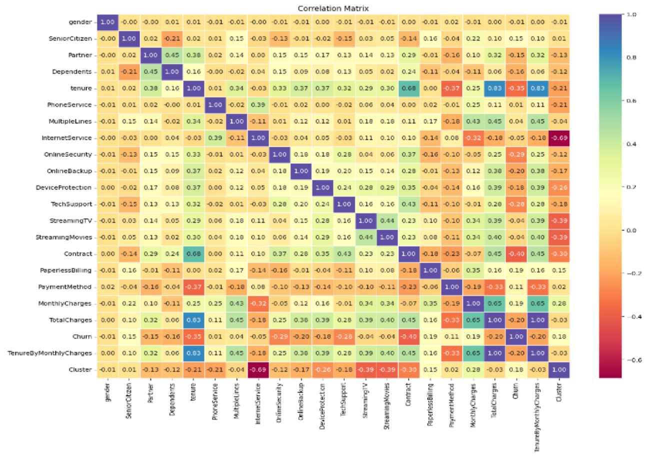

Fig4:Correlation Matrix

Description of Correlation Matrix

***1. Purpose*:**

The correlation matrix is the statistical summary indicating the relationship of features with one another in a dataset. It helps to establish how features are related to one another, which is quite important in understanding the data and making decisions about feature selection.

2. Interpretation:

Positive Correlation: The value close to +1 on the positive side indicates strong positive correlation; that is, when one feature increases, the other increases. For example, "TotalCharges" and "MonthlyCharges" might show high positive correlation since higher monthly charges lead to a higher total charge.

Negative Correlation: A negative value near-1 reflects a high negative correlation; when one feature increases, the other decreases. For example, "tenure" might be negatively correlated with "Churn," since the longer a customer has been with the company, the less likely he is to churn.

Near Zero Correlation: The values close to 0 mean no significant correlation between the two features exists.

3. Importance:

The correlation matrix is extremely useful in finding multicollinearity—high correlation between two or more features. High multi-collinearity can hurt model performance, making this another important consideration in selecting the features for the model. Those features highly correlated to the target variable, for example, "Churn," are especially good predictors.

Practical Implications:

It strongly emphasises how it has practical implications on CRM, whereby businesses can work on their retention strategies based on the at-risk customer identification, hence reducing base churn

Confusion Matrix:

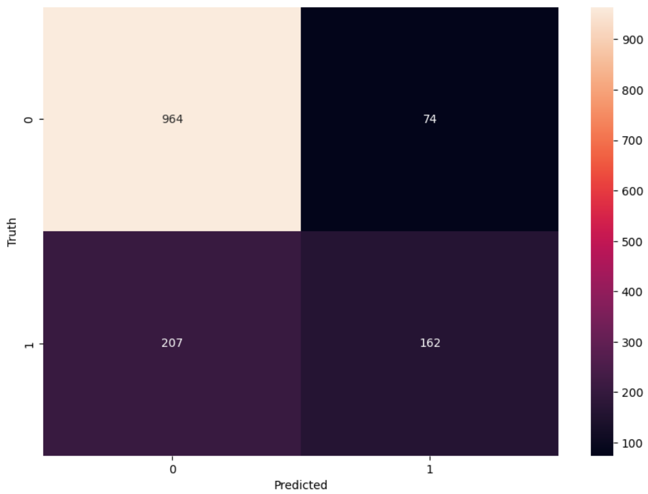

Fig 5: Confusion Matrix for ANN

(Artificial Neural Networks)

Confusion Matrix Description

1. Purpose:

A confusion matrix is a measure of performance for classification models. It describes the model's performance in class differentiation, such as "Churn" and "No Churn." This part needs more explanation

2. Components:

True Positives: No. of cases where the model has correctly predicted the positive class, for example, it predicted "Churn" when the customer actually churned.

TrueNegative: The number of cases in which the model correctly predicted the negative class, such as when it had predicted "No Churn" and the customer did not churn.

FalsePositive: The number of cases where the model has misclassified a case as the positive class—for example, when it has predicted "Churn" on a customer who did not churn.

FalseNegative: Number of times the model has misclassified the negative class; for instance, the model is predicting "No Churn" when the customer has actually churned.

3. Importance:

A confusion matrix is very important in measuring the performance of a model on grounds other than accuracy. In particular, it helps understand the trade-off between different types of errors and tells whether it is biased to one class or not.

Model Comparisons:

The report clearly brings out comparisons across models to show that where Logistic Regression and Decision Trees were sufficient, their performances were way below ensembling methods, in this case, XGBoost.

The comparison underlines that no models are universally optimum in all types of data; some being much better fitted to adapt better to different prediction tasks. In this respect, XGBoost does quite well when there are different types of features and in terms of robustness toward avoiding overfitting.

Accuracy Across Different Classifiers

1. Purpose:

By testing the accuracy among different classifiers, we could compare how well different models perform and choose which one works best on the problem. Accuracy refers to the ratio of the number of instances predicted correctly—both TP and TN—to total instances.

2. Classifiers Compared:

Logistic Regression: Here it was used as a baseline model; it usually acts as a benchmark for many algorithms because of the model's simplicity and its interpretability. It usually works well when data are linearly separable.

Random Forest: This method uses an ensemble of decision trees to give better accuracy and a greater reduction in overfitting. It does relatively well on complex relationships.

XGBoost: A highly popular, very high performance, and quite efficient boosting algorithm; often the best model on structured/tabular data.

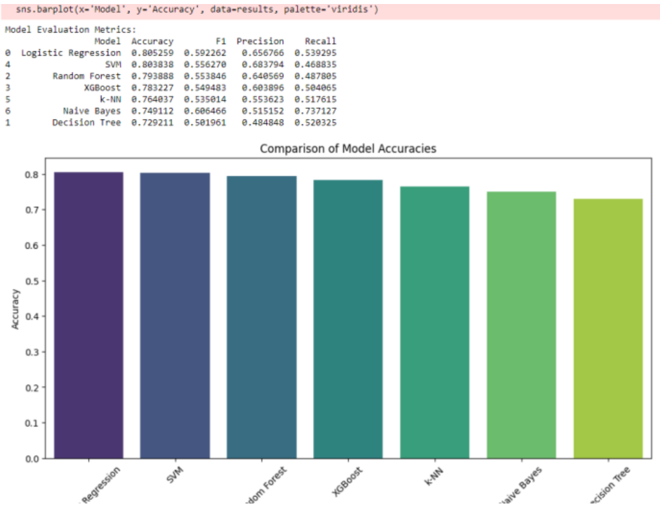

Fig 6: Comparison b/w Model Accuracies

SVM (Support Vector Machine): A classifier that works by finding a hyperplane between classes in the best possible way. It really shines in high-dimensional spaces but performs well on smaller datasets.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.80	0.63	0.53`	0.59
SVM	0.80	0.68	0.46	0.55
Random Forest	0.79	0.55	0.63	0.49
XGBoost	0.79	0.56	0.64	0.50
K-NN	0.76	0.55	0.51	0.53
NaiveBayes	0.74	0.51	0.73	0.60
Decision Tree	0.72	0.48	0.53	0.50

Fig 7 : Tabular view of Comparisons

3. Interpretation:

Best Accuracy: This will normally be the most accurate classifier, but do not forget to consider other metrics like precision, recall, and F1-score.

Model Selection: The final choice of the classifier will have to be based on this trade-off in accuracy against other performance metrics, and more importantly, on the needs of the project. For example, a churn prediction model would emphasize the minimization of the number of false negatives.

V CONCLUSION

Effectiveness of XGBoost:

Finally, for a customer, the churn prediction was best

done using XGBoost, more so when K-Means clustering was enhanced. There is an equal balancing of performance across all the evaluation metrics, and it had churn prediction preference in this context.

Importance of Tenure and Charges:

The analysis consistently kept the fact intact that "TenureByMonthlyCharges," "MonthlyCharges," and "tenure" are the topmost predictors of churn. This suggests that the businesses would need to be really careful for these parameters in building up their retention strategy.

Utility of Clustering:

Adding the K-means clustering was, in fact, quite supportive, as that helped in adding customer segmentation into similar characteristic groups, and that boasted model result accuracy in making the predictive model. This technique benefits business in more than one way, predicting the churn event and better segmenting the customers needing different retention strategies.

Actionable INSIGHTS for CRM:

Model gives meaningful, actionable insights that allow for direct application into CRM systems. By key factors, it can predict likely customers to churn and take early measures to retain them.

Pros and Cons of the Approach:

Pros:

Effective Combination: The combination of K-Means clustering with XGBoost was a powerful way of predicting customer churn, yielding high accuracy and practical insight.

Feature Importance: It identified the most important features that are more likely to influence customer churn.

Scalable**:** The model can be implemented on a totally new dataset as well as in new industries.

Cons**:**

Model Complexity**:** Involving in a mix of multiple algorithms like K-Means and XGBoost adds to model complexity and might be computationally demanding and hard to implement.

Data dependency**:** The model depends highly on the qualitative and relevant data provided to it. Any flaw in data pre-processing and feature selection will cause wrong predictions.

Limited Exploration**:** The study is based on a specific dataset and features only. However, the implementation of other complementary datasets or advanced-level techniques may be applied in the future studies to improve the performance of the model further.

Future Research Directions:

This paper recognizes the fact that the present model is good but based on one dataset. Tests with other datasets or other industries can be one way to confirm generalizability in future work.

Further exploration of other features or different clustering techniques will most likely lead to even better predictions. Also, how customer behaviour changes over time would affect the temporal changes in the predictions but has not been analysed to date and should be studied in future research.

1.Feature Importance:

Current Status:

This report focuses on the very important features "TenureByMonthlyCharges," "MonthlyCharges," and "tenure."

Proposed Solution

Additional Features:

Add new features that describe customer engagement metrics: the frequency of interactions with customer support, the product usage pattern, and sentiment analysis of customer feedback from the chosen text.

Feature Engineering:

Apply polynomial features, interaction terms, or log transformations dealing with nonlinear relationships in features.

2. Enhance Model Interpretability:

Current Status:

Model building has been done using XGBoost and K-Means Clustering Algorithms. It is equally sophisticated and hence not able to explain it easily.

Proposed Solution:

Interpretable Models:

Use models such as Decision Trees and Linear Models for benchmarks on explainability.

Explain Techniques:

Apply existing libraries, like SHAP, that provide insight at the feature level, making the model prediction interpretable (SHapley Additive exPlanations).

Business Communication: Reports and visualizations should explain these in language a business can take action from.

3.Optimizing Model Performance Using Advanced Techniques

In our research study, the model that proved most effective was XGBoost.

Proposed Solution:

Ensemble Learning:

Try other ensemble learning techniques—the love for models like CatBoost or LightGBM—both being beautiful in handling categorical variables and training much faster.

Hybrid Models:

Implement K-means clustering along with neural networks. Mind you, K-means clusters can catch unrecognisable patterns for tree-based methods quite easily, whereas neural networks can give the weight

Model Training: Train and compare different ensemble methods— LightGBM, CatBoost—to XGBoost to provide a quantitative assessment of accuracy and speed improvements.

Hybrid Integration: Try hybrid approaches, such as that of clustering followed by classification with neural networks for better analysis depth.

4. Cross-Validation and Testing Expansion:

Current Status:

The model was validated on a single dataset, which may reduce its level of generalization.

Proposed Solution:

Domain Cross-Validation:

The model will be tested on datasets from other related industries, like telecommunications and retail, to see if it's robust.

K-Fold Cross-Validation:

Carry out k-fold cross-validation with different k to test the robustness and generalization ability.

In summary, this project underscores the importance of proactive churn management strategies and the potential of AI-driven solutions in mitigating customer attrition. By harnessing the power of data-driven insights, organizations can optimize customer retention efforts, enhance customer satisfaction, and ultimately drive business growth in today's competitive landscape.

VI REFERENCES

1. Customer churn analysis: Why analyzing churn is so important (paddle.com)
2. https://indatalabs.com/blog/customer-churn-analysis-prediction
3. https://fayrix.com/blog/customer-churn-prediction-benefits
4. https://www.researchgate.net/publication/369770677_Customer_Churn_Prediction_Using_Machine_Learning_Approaches
5. https://ieeexplore.ieee.org/document/6329952
6. https://www.researchgate.net/publication/369770677_Customer_Churn_Prediction_Using_Machine_Learning_Approaches.
7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11042018/
8. https://www.researchgate.net/publication/321723500_Review_of_Customer_Churn_Analysis_Studies_in_Telecommunications_Industry
9. https://www.researchgate.net/publication/260113063_Churn_Prediction_using_Comprehensible_Support_Vector_Machine_an_Analytical_CRM_Application
10. https://managementdynamics.researchcommons.org/journal/vol24/iss1/1/
11. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0289724