CHURN -YES/NO | CUSTOMER CHURN PREDICTION IN THE TELECOM INDUSTRY

Dani Alex Parayil
Master of Data Analytics,
Dept. of Science
The University of Western, Ontario
dparayil@uwo.ca

Garima Gambhir
Master of Data Analytics,
Dept. of Science
The University of Western, Ontario
ggambhi@uwo.ca

Ritika Pandey
Master of Data Analytics,
Dept. of Science
The University of Western, Ontario
rpande6@uwo.ca

Semal Shastri
Master of Data Analytics,
Dept. of Science
The University of Western, Ontario
sshastr@uwo.ca

.

Sumedha
Master of Data Analytics,
Dept. of Science
The University of Western, Ontario
sgalgali@uwo.ca

***Abstract*—In the rapidly evolving telecom industry, maintaining a loyal customer base is essential for continuous business growth and contributes to company profits. Customer churn presents a significant threat to revenue streams and long-term sustainability. And for such telecom providers, this project aims to mitigate customer attrition risks by primarily categorizing at-risk customers and identifying underlying factors contributing to customer churn using data analytics and machine learning techniques such as Logistic Regression, Support Vector Machines, advanced decision trees and boosting techniques. Each model was evaluated against appropriate metrics relevant to the needs of the telecom industry.**

I. INTRODUCTION

In the competitive telecommunications industry, retaining customers is essential for sustaining business growth and ensuring profitability. However, customer churn a situation where customers discontinue their services remains a constant challenge. Churn threatens revenue stability and undermines long-term business success by impacting customer loyalty and increasing the cost of acquiring new clients. To overcome this challenge, a proactive approach is essential to detect and retain customers at risk of leaving before churn occurs.

This research focuses on building a predictive model to address the challenge of customer churn competently. By using data analytics and predictive modeling techniques, the study aims to create a consistent framework for identifying customers who are most likely to churn. These insights can strengthen telecommunications companies to develop targeted retention strategies, reduce churn rates, and improve customer relationships.

The importance of this research stems from its ability to provide actionable, data-driven insights that can guide strategic decision-making. By accurately forecasting customer churn, businesses can protect their revenue, enhance customer satisfaction, and improve their competitive edge in an ever-evolving industry. This research aims to reduce the financial risks associated with customer churn and better retention strategies, eventually supporting long-term business growth.

II. BACKGROUND

Customer churn which happens when customers cancel or discontinue their service with the company is an important area of focus in the telecommunications industry due to its correlation to both revenue and the stability of the customer base. Even though definitions of churn may differ based on contractual agreements and customer behavior, it is uniformly acknowledged as a major concern (1). Several factors influence churn, including unsatisfactory service quality, attractive competitive offers, dissatisfaction, and shifting customer needs (2, 3).

Predictive modeling techniques plays an important role in predicting and managing churn. Statistical methods such as logistic regression and non-parametric supervised learning alpgorithms such as decision trees offer interpretability and effectiveness in binary classification tasks (4). Python offers robust tools for logistic regression and decision trees, pivotal for binary classification tasks such as churn prediction (4). Furthermore, the platform supports advanced ensemble methods like Random Forests and Gradient Boosting using Python libraries such as Scikit-learn and XGBoost., promising heightened accuracy in predictive modeling endeavors (5). Studies explore the potential of deep learning techniques in capturing intricate nonlinear relationships and interaction effects (6).

Effective churn prediction is highly reliable on comprehensive data preparation and feature engineering, which reinforces the creation of relevant features such as customer usage patterns and interaction data (7). Incorporating time-series analysis to capture the changing nature of customer behavior can significantly improve the accuracy of predictive models (8). Case studies have shown that applying predictive insights to develop actionable strategies can help reduce churn rates. Customized approaches, based on model results, have proven to be highly productive in retaining customers (9). However, challenges remain in integrating these predictive models within existing CRM systems and ensuring timely, personalized interventions (2).

In the future, the integration of big data technologies and real-time analytics holds promise for dynamic and responsive churn management strategies (10). Moreover, improving personalization in customer interactions and improving overall customer experience management are considered key factors in decreasing churn and building customer loyalty (11).

Customer churn prediction remains a complex yet critical aspect of telecommunications business operations. Using insights from a blend of traditional and modern analytical techniques, informed by the literature's findings, is essential for developing robust churn prediction models that effectively address industry challenges and drive sustainable business growth.

III. Data Pre-Processing

For this project, the dataset sourced from Kaggle comprised of 7043 customer records with 20 features which details information broadly by customer demographics, service usage, customer satisfaction level and contract types.

Features such as Total Charges that were stored in the dataset as string values were converted into numerical datatype and missing values were treated with average of the total charges column. Irrelevant features such as customer IDs were dropped. Highly correlated features such as Phone Service and Monthly Charges were removed as Phone Service failed to explain the variation in the churn rate and Monthly charges were correlated to Total Charges. In context of the customer churn dataset, all tenure values which had the value 0 were replaced with a minimum of 1 as no customer under the telecom provider will logically have 0 tenure if data is available to them. To facilitate model interpretability, binary categorical variables in the dataset such as gender, partner, dependents, phone service, paperless billing, and churn represented by Yes/No were assigned numeric values to label encode them to improve computational efficiency, and prepare the data to seamlessly pass through machine learning algorithms utilized in this project. Multi-class categorical variables namely internet service, contract, payment method, multiple lines, online security, online backup, device protection, tech support, streaming TV, streaming movies were one-hot encoded to avoid implicit ordinal relationships that label encoding would introduce. Encoding techniques were performed to prepare the data for models such as logistic regression, random forest, SVC, and Gradient Boosting.

Feature scaling was applied to numerical features such as tenure, monthly and total charges using the Standard Scaler function to decrease the sensitivity of the selected machine learning algorithms to the magnitude of the feature in use. This was also done in order to enhance model performance and reduce bias induced by features with larger magnitudes. Models such as Logisitc regression, SVC, and Gradient Boosting relies on gradient based optimization techniques. Hence, scaling and standardizing the features improves convergence for the estimated coefficients.

IV. Viszualization

The dataset comprises of a larger proportion of customers on a month-to-month basis subscription. We have also identified that customers having lower contract period, specifically users on a month-to-month contract had a significantly higher risk

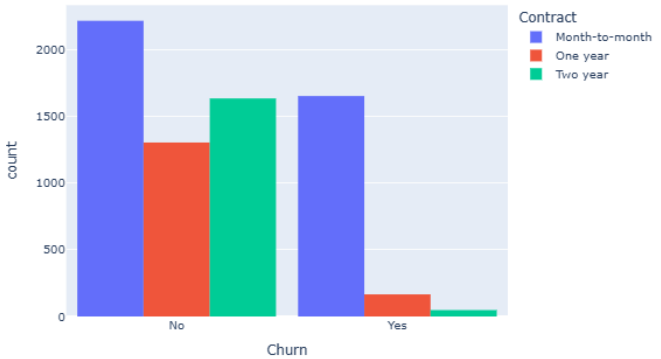

Fig. 1. Churn categories by contract type

of churning. This is also true in cases where customers have subscription to internet service as well.

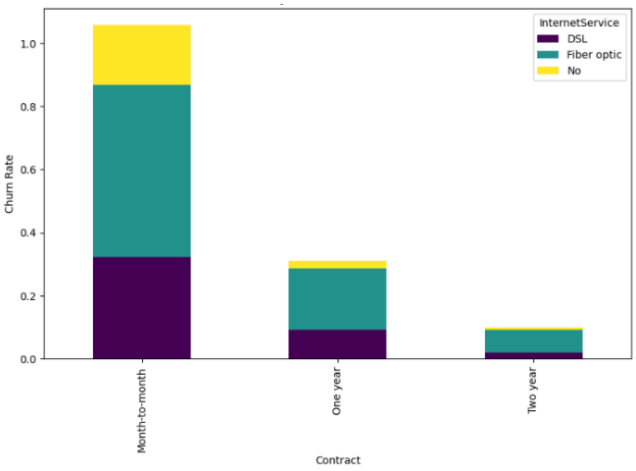

Fig. 2. Churn Rate by Contract and Internet Service Type

However, customers in their first 12 months of their tenure have higher attrition rate compared to users beyond 13 months, with significantly low attrition for customers who have complete 61 months and above.

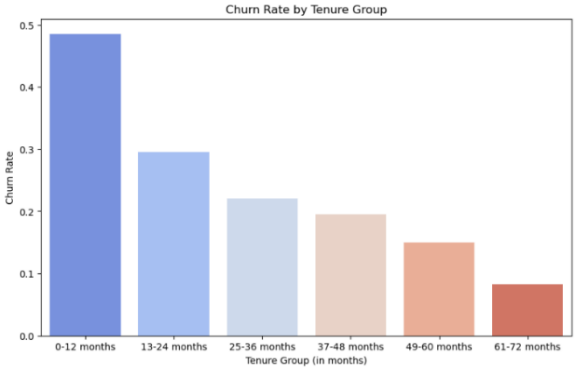

Fig. 3. Churn rate by customer tenure

Based on our demographic analysis, we have uncovered that senior citizens are disproportionately represented among those who churn in this dataset. This demographic potentially feels underserved by current offerings or may require more specialized and targeted strategies that adhere with their needs. Tailoring offerings, such as senior-friendly service plans or enhanced customer service options, could potentially close this gap if our algorithm rightfully classify them.

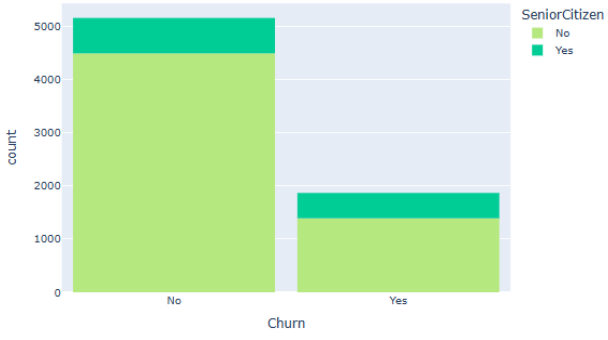

Fig. 4. Disproportionality of senior citizens in the churn dataset

We have also found that customers without a life partner has a higher attrition rate when compared to their counterparts implying that they tend to be less likely to switch between telecom providers.

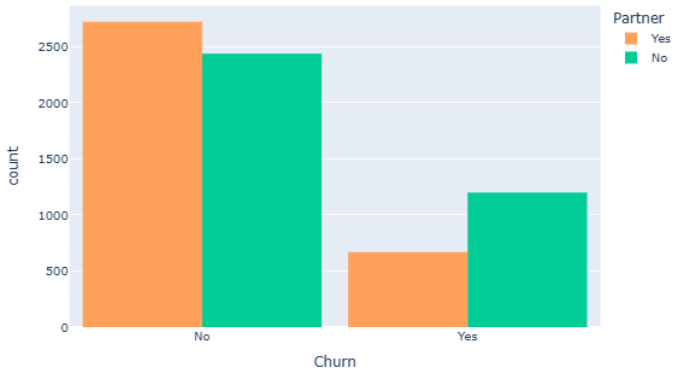

Fig. 5. Churn count by relationship status

V. Modelling and Analysis

In this analysis, higher risk customers have been defined as Yes = 1 (Positive Class), and No = 0 (Negative Class). Classifier models such as logistic regression, support vector classifier (SVC), decision trees such as random forest, AdaBoost, GradientBoost and CatBoost (due to high number of categorical features) were implemented.

Each of these models utilized the pre-processed dataset including features that were label encoded and one-hot encoding for training and testing. The train-test was split into 80-20 proportion for initially training and fitting the model. Due to the dataset imbalance between Yes (73.4%) and No (26.6%) churn labels, we planned to evaluate each model's performance on a 10 fold cross-validation and understand it's predictive capability on unseen test data. Cross-validation model evaluation was performed based on metrices like balanced accuracy, precision, recall (sensitivity), F1-score, AUC-ROC (Area under the receiver operating characteristic curve) and confusion matrix. Special metrics such as geometric mean combined sensitivity and specificity of the model indicating its ability to perform well across both classes especially in cases of imbalanced dataset. Balanced accuracy was also used for fair evaluation between both classes so that the majority class does not dominate the metric unlike accuracy.

A. *Logistic Regression*

The logistic regression model acts as the base reference model for the project due to its simplicity and interpretability and uses a linear decision boundary between churners and non-churns with cut-off at 0.5.

TABLE I. Logistic Regression Evaluation Metrics

Confusion Matrix				
Churn	***Predicted: Yes***		***Predicted: No***	
Actual: Yes	286		62	
Actual: No	339		720	
Other Evaluation Metrics				
Class	***Precision***	***Recall***	***F1-score***	***Geometric Mean***
0	0.680	0.680	0.680	0.748
1	0.460	0.822	0.588	0.748
AUC**: 0.835**; Balanced Accuracy: 0.751				

- Logistic regression parameters such as class weight were set to "balanced" to handle class imbalance, and maximum iterations were set to 1000 with a random state of 50.
- The logistic regression model does not handle non-linear features well and performed worse in comparison to advanced machine learning techniques and decision trees capable to classify and capture non-linear relationships.

B. *Support Vector Classifier (SVC)*

- The SVC model was set with a linear kernel function for mapping data to higher dimensions with class weights set to "balanced" to account for the class imbalance along with a random state of 50.
- Despite its high recall rate, the model only works with small to medium datasets and is computationally expensive and has weak training rates on larger datasets.
- SVC also is not suitable for highly imbalanced data without proper data pre-processing

TABLE II. SVC Confusion Evaluation Metrics

Confusion Matrix				
Churn	***Predicted: Yes***		***Predicted: No***	
Actual: Yes	299		49	
Actual: No	377		682	
Other Evaluation Metrics				
Class	***Precision***	***Recall***	***F1-score***	***Geometric Mean***
0	0.644	0.644	0.644	0.744
1	0.442	0.859	0.584	0.744
AUC**: 0.837**; Balanced Accuracy: 0.752				

C. *Random Forest*

- Random forest is intrinsically suited for multiclass problems. The number of trees was set to 500, while maximum features were set to the square root of the total number of features with a limit of 30 leaf nodes to prevent overfitting. A random state of 50 was used.
- High AUC comparable to AdaBoost and Gradient Boosting. Moderate precision, outperforming SVC but weaker compared to Gradient Boosting and AdaBoost. This model had the lowest number of false positive among all the models identifying non-churn cases accurately and ideal for minimizing false positives.

TABLE III. Random Forest Evaluation Metrics

Confusion Matrix				
Churn	***Predicted: Yes***		***Predicted: No***	
Actual: Yes	175		173	
Actual: No	94		965	
Other Evaluation Metrics				
Class	***Precision***	***Recall***	***F1-score***	***Geometric Mean***
0	0.911	0.911	0.911	0.677
1	0.509	0.503	0.506	0.677
AUC:** 0.839**; Balanced Accuracy: 0.707				

D. CatBoost Classifier

- CatBoost natively handles categorical variables without data pre-processing. Parameters were set with max. iterations = 500, with a depth of 6, learning rate 0.01, a log loss function to prevent overfitting and weights set to handle the data imbalance.
- This algorithm works well on imbalanced datasets through its loss functions and can manage larger datasets accommodating non-linear relationships.
- This classifier had one of the best performances in capturing the balance between sensitivity and specificity. Maintained a good recall and Balanced accuracy. However, the model took more computational time in comparison to other models.

TABLE IV. CATBOOST CLASSIFIER EVALUATION METRICS

Confusion Matrix				
Churn	***Predicted: Yes***		***Predicted: No***	
Actual: Yes	272		96	
Actual: No	246		795	
Other Evaluation Metrics				
Class	***Precision***	***Recall***	***F1-score***	***Geometric Mean***
0	0.764	0.764	0.764	0.751
1	0.526	0.739	0.632	0.751
***AUC:** 0.835**; Balanced Accuracy:** 0.751*				

E. Gradient Boosting Classifier

- Gradient Boosting set with 100 boosting iterations, learning rate of 0.1 and a maximum depth of 3 had the highest AUC among all models that were evaluated, slightly overcoming AdaBoost and Random Forest. In terms of balanced accuracy, it outperformed the Random Forest Model.
- Metrics indicated that the model had potential to miss more actual churn cases making it less suitable for churning prediction since recall is a priority in the customer churn attribution to the business.
- This model is more suited when AUC and Precision are critical for the task and balances precision and recall without skewness towards one metric.

TABLE V. GRADIENT BOOSTING CLASSIFIER EVALUATION METRICS

Confusion Matrix				
Churn	***Predicted: Yes***		***Predicted: No***	
Actual: Yes	193		155	
Actual: No	113		946	
Other Evaluation Metrics				
Class	***Precision***	***Recall***	***F1-score***	***Geometric Mean***
0	0.893	0.893	0.893	0.708
1	0.631	0.555	0.592	0.708
***AUC:** 0.843**; Balanced Accuracy:** 0.724*				

F. AdaBoost Classifier

- AdaBoost that was set with a maximum of 100 weak leaners with a random state of 50, was chosen to study whether the algorithm was able to build a more robust model in comparison to the other models. It had a decent recall but not as great compared to Random Forest and CatBoost. However, it was able to minimize false positives better than logistic regression and CatBoost.
- It had one of the highest AUC scores and highest precision comparable to gradient boosting and random forest. Showed excellent ability to differentiate churn and no-churn cases due to its high AUC. A high AUC, precision and geometric mean make it one of the most balanced models making it a viable choice when false positives are to be minimized.

TABLE VI. ADABOOST CLASSIFIER EVALUATION METRICS

Confusion Matrix				
Churn	***Predicted: Yes***		***Predicted: No***	
Actual: Yes	198		150	
Actual: No	114		945	
Other Evaluation Metrics				
Class	***Precision***	***Recall***	***F1-score***	***Geometric Mean***
0	0.764	0.892	0.892	0.713
1	0.635	0.569	0.600	0.713
***AUC:** 0.840**; Balanced Accuracy:** 0.731*				

VI. RESULTS

Based on our analysis between different machine learning techniques, we have primarily identified that for a telecom provider it is more critical to capture all customers who are at risk of churn in comparison to customers who are not at risk. This is due to the excessive cost involved in losing a customer compared to offering better retention offers for low-risk customers. Since a high recall rate is desirable in our context, the CatBoost algorithm has been chosen as the best model to identify churn cases. This is because CatBoost can inherently take categorical data without data pre-processing and is robust to imbalanced datasets in comparison to all other models. CatBoost exceptionally handles categorical features and converges quicker in comparison to techniques like Random Forest. Even though SVC has outperformed CatBoost in terms of recall, it does not work well with large datasets and is prone to errors when dealing with imbalanced classes. CatBoost also has good balanced accuracy and balances both recall and precision decently. SVC is also good at handling non-linear relationships, however, CatBoost also models non-linear interactions, making its advantageous over SVC.

VII. CONCLUSION AND FUTURE WORKS

CatBoost, Gradient Boosting, AdaBoost, Random Forest have proved to have better performance even with default parameters compared to Logistic Regression and SVC. Our next steps would be to further refine these modelling techniques by performing hyperparameter optimization using Bayesian techniques or grid search (14). This project has only experimented primarily on decision trees and boosting techniques, however, delving into deep learning methods such as neural networks may improve prediction rates and provide a more robust model. To interpret customer churn predictions, more focus should be placed on tools like SHAP or LIME. (15)

REFERENCES

[1] Vafeiadis, T., et al. (2015). "A comparison of machine learning techniques for customer churn prediction." Simulation Modelling Practice and Theory, 55, 1-9.

[2] Hadden, J., et al. (2007). "Computer assisted customer churn management: State-of-the-art and future trends." Computers & Operations Research, 34(10), 2902-2917.

[3] Dasgupta, P., et al. (2008). "Social network analysis for predicting customer churn." Proceedings of the 2008 ACM symposium on Applied computing, 1325-1329.

[4] Lemmens, A., & Croux, C. (2006). "Bagging and boosting classification trees to predict churn." Journal of Marketing Research, 43(2), 276-286.

[5] Xie, Y., et al. (2009). "Customer churn prediction using improved balanced random forests." Expert Systems with Applications, 36(3), 5445-5449.

[6] Oztekin, A., et al. (2009). "A data mining-based framework for customer value assessment in subscription services." European Journal of Operational Research, 199(2), 508-519.

[7] Smith, K. A., et al. (2011). "A neural network approach to churn prediction in the mobile phone market." Expert Systems with Applications, 39(11), 9900-9905.

[8] Burez, J., & Van den Poel, D. (2009). "Handling class imbalance in customer churn prediction." Expert Systems with Applications, 36(3), 4626-4636.

[9] Neslin, S. A., et al. (2006). "Defection detection: Measuring and understanding the predictive accuracy of customer churn models." Journal of Marketing Research, 43(2), 204-211.

[10] Minelli, M., et al. (2013). Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses. Wiley.

[11] Kumar, V., & Reinartz, W. (2016). Customer Relationship Management: Concept, Strategy, and Tools. Springer.K. Elissa, “Title of paper if known,” unpublished.

[12] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[13] Y. Yor

[14] Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. Advances in Neural Information Processing Systems, 25, 2951–2959.

[15] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

DOI: 10.1023/A:1010933404324

VIII. APPENDIX

A. TEAM CONTRIBUTION

All group members performed the exploratory data analysis and data visualizations individually and selected machine leanring techniques during preliminary analysis were delegated to each member. All members were entrusted to perform required data pre-processing relevant to their machine learning algorithms. The final code was compiled, and relevant data insights were derived from individual members.

1. Dani Alex Parayil: Exploratory data analysis (EDA), data visualization and setting up CatBoost algorithm. Preparation of Final report and slide decks for presentations.
2. Garima Gambhir: EDA, preliminary data analysis, visualization and data preprocessing for Logistic Regression.
3. Ritika Pandey: EDA, preliminary data analysis and data preprocessing for SVC model. Preparation of Final report, and project proposal
4. Semal Shastri: EDA, preliminary data analysis and data preprocessing for Random Forest and decision trees.
5. Sumedha: EDA, data preprocessing for Gradient Boost and AdaBoost techniques. Code compilation and formatting for final submission.