TELECOM CUSTOMER CHURN PREDICTION

Comparative study of sampling methods for better predictions

Submitted by –

Group 15 :

- Chandra Kishor Mishra
- Harish Rajan
- Justin Seale
- Shubhinder Singh Rana
- Sujay Shrivastava

Table of Contents

Abstract

Churn, in the context of telecom companies, refers to the attrition or loss of customers who no longer subscribe to their services. This report aims to understand the concept of churn and its implications in the telecom industry. The ability to accurately identify customers who are likely to churn is crucial for telecom companies, as they can proactively engage with these customers by offering discounts, better plans, or incentives to prevent them from leaving.

Challenges

Retaining customers for a longer period of time presents a significant challenge in the telecom industry. While acquiring new customers may be relatively easier, maintaining customer loyalty is essential. The primary challenges in maintaining customer loyalty include:

- Revenue Loss: High churn rates directly impact revenue as the loss of customers leads to decreased revenue generation.

- Word of Mouth: Customers leaving a telecom provider can influence the decisions of other customers, potentially leading to a ripple effect of churn.

- Customer Lifetime Value (CLV): CLV is a metric that represents the total revenue a business can expect from a single customer over their relationship with the company. Factors such as ease of switching providers, quality of service, customer dissatisfaction, and attractive offers can affect CLV.

Addressing Technical Aspects

One of the technical challenges in churn analysis is dealing with class imbalance, where the data has disproportionate class proportions, such as in a 75:25 ratio. This imbalance poses two problems: over-biasness in the dataset and difficulties in achieving accurate predictions.

To mitigate these challenges, it is recommended to handle class imbalance before training a predictive model. Common methods for addressing class imbalance fall into the following categories:

- Over-sampling: This involves increasing the number of instances in the minority class by duplicating or synthesizing data points to balance the dataset.

- Under-sampling: This approach involves reducing the number of instances in the majority class to balance the dataset.

- Combination: A hybrid approach that combines over-sampling and under-sampling techniques to achieve a balanced dataset.

Result

Churn analysis plays a vital role in the telecom industry, as customer retention significantly impacts revenue and customer satisfaction. By understanding the factors contributing to churn and addressing the challenges associated with class imbalance, telecom companies can develop effective strategies to retain customers, improve customer loyalty, and optimize their business outcomes.

Introduction

This project report aims to explore and tackle the challenges associated with predicting customer churn, thereby enabling service providers to take pre-emptive actions to retain their existing customer base. We will leverage advanced data science techniques and innovative sampling methodologies to improve the effectiveness of churn models and mitigate the inherent class imbalance problem commonly encountered in such analyses.

One of the primary technical challenges in predicting churn lies in the presence of class imbalance within the dataset. This imbalance occurs when the number of churned customers is significantly lower than the number of loyal customers. Traditional predictive models tend to struggle with accurate predictions due to the lack of sufficient representation from the minority class. Therefore, our project will focus on employing various sampling techniques to address this issue, ensuring a more balanced and reliable churn prediction model.

Furthermore, our project also aims to overcome the limitations posed by random guessing and tame any potential biases within the predictive model. By leveraging the power of cutting-edge data science methodologies, we intend to develop a highly accurate churn prediction model that outperforms random guesswork, helping service providers identify customers at risk of attrition well in advance.

The objective of this project is to assist service providers in making informed decisions and implementing targeted strategies to retain customers before they churn. By effectively predicting churn and taking proactive measures to retain customers, businesses can optimize their customer retention efforts, boost customer lifetime value, and ultimately enhance their overall competitiveness in the market.

In the subsequent sections of this report, we will delve into the details of the technical challenges associated with churn prediction, discuss the data science formulation and methodologies employed, and present our findings, recommendations, and potential areas for future research. Through this project, we endeavour to contribute to the body of knowledge on customer retention and provide valuable insights that can help businesses foster stronger customer relationships and achieve sustainable growth

In addition to the challenges mentioned, there are several other factors driving the urgency for improving customer retention strategies. With the rise of digital transformation and the increasing accessibility of alternative service providers, customers now have a plethora of options

at their fingertips. This abundance of choices makes it even more challenging for businesses to differentiate themselves and maintain customer loyalty.

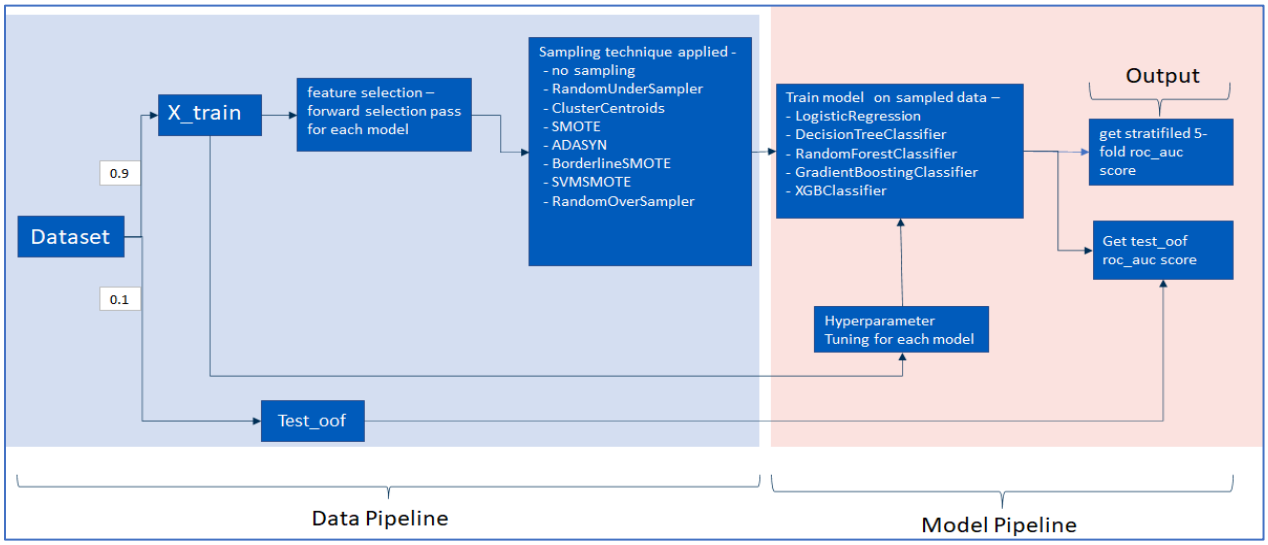
Moreover, the cost associated with acquiring new customers is often significantly higher than retaining existing ones. By focusing efforts on preventing churn, organizations can not only reduce customer acquisition costs but also capitalize on the potential revenue growth from long-term customer relationships. Thus, accurately predicting churn and implementing effective retention strategies have become vital for organizations seeking to maximize profitability and sustain their market position.

The advancements in data science and machine learning techniques have opened new opportunities for businesses to gain deeper insights into customer behaviour and preferences. By harnessing the power of predictive modelling, organizations can analyse vast amounts of historical customer data to identify patterns, indicators, and risk factors associated with churn. This allows for the development of targeted retention strategies that cater to the specific needs and preferences of individual customers, enhancing the overall customer experience.

To address the technical challenges of class imbalance in churn prediction, various sampling techniques can be employed. These techniques, such as oversampling the minority class or under sampling the majority class, aim to create a more balanced dataset that better represents the real-world distribution of churn and non-churn instances. By applying these sampling techniques, we can improve the model's ability to accurately predict churn for both minority and majority classes, leading to more reliable and actionable insights.

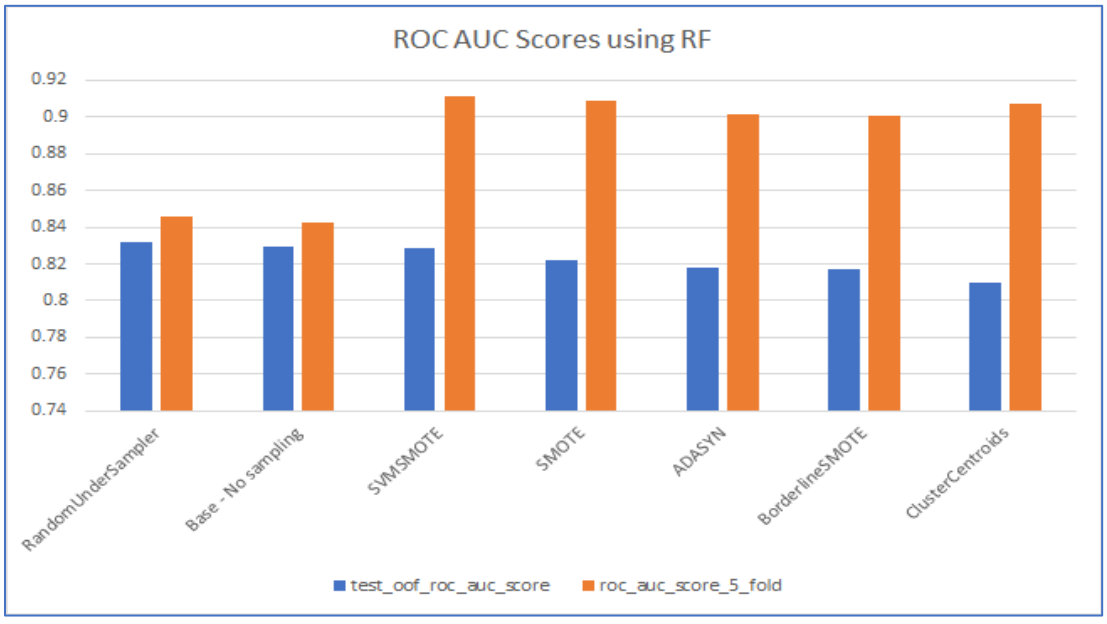
Through this project, we aim to contribute to the growing body of knowledge in the field of customer retention by developing an advanced churn prediction model that effectively addresses the challenges of class imbalance and model bias. By doing so, we hope to provide organizations with a powerful tool to proactively identify and retain customers at risk of churn, ultimately enhancing customer satisfaction, loyalty, and long-term business success.

Methods



1. **Dataset Split:** The dataset used for analysis is divided into two subsets. A test set comprising 704 records is set aside for out-of-fold (OOF) evaluation, while the remaining 6,339 records (referred to as X_train) are used for model training.
2. **Feature Selection**: For each modeling technique, a forward feature selection method is employed to identify the most relevant features. This step ensures that the subsequent evaluation of sampling techniques focuses on the selected feature set for fair comparisons.
3. **Sampling Techniques**: Several sampling techniques are applied to address the class imbalance in the dataset. These methods are employed for downsampling: RandomUnderSampler and ClusterCentroids. For upsampling, these techniques are utilized: SMOTE, ADASYN, BorderlineSMOTE, SVMSMOTE, and RandomOverSampler. The baseline case involves no sampling, serving as a reference point for comparison.
4. **Parameter Tuning**: After selecting the appropriate features, model hyperparameters are tuned to optimize performance. This step is carried out for each modeling technique and sampling combination. We used grid-search and hyperopt optimization to find the tuned parameters.
5. **Model Fitting and Evaluation**: The models are trained on the different sampled datasets, and the stratified 5-fold ROC AUC scores are computed. The predictions are then made on the OOF test set, and the performance is evaluated by comparing the predicted values against the true values.
6. **Analysis**: The obtained ROC AUC scores for each model and sampling technique combination are analyzed to determine the impact of different sampling methods on the predictive performance.

Results



Findings

This study investigated the performance of different machine learning techniques for customer churn prediction in imbalanced datasets. The results showed that undersampling techniques consistently outperformed other approaches in most classification models. This can be attributed to the balanced

representation of both churn and non-churn instances during training, allowing the models to learn from a more equitable distribution.

- Random Forest, a popular bagging-based ensemble method, demonstrated superior performance in predicting customer churn. Bagging methods, which combine multiple models trained on different subsets of the data, prove advantageous in class imbalance scenarios. By generating diverse subsets through bootstrapping, bagging reduces the variance and bias associated with imbalanced classes, resulting in improved predictions.
- Upsampling techniques did not consistently yield better performance on unknown data despite demonstrating an improvement in cross-validation scores. This finding can be attributed to the limitation of upsampling in the generalization of models to unseen instances. Upsampling artificially increases the number of minority class samples, potentially causing overfitting and reduced performance on unknown data. The lift in cross-validation scores during upsampling is likely a result of the models being evaluated on more balanced data, rather than indicating a better fit to the underlying distribution of the target variable.

Recommendations

- Prioritize undersampling techniques over up sampling techniques for imbalanced datasets. Undersampling techniques consistently outperformed other approaches in most classification models for customer churn prediction. This is because undersampling techniques balance the representation of both churn and non-churn instances during training, allowing the models to learn from a more equitable distribution.
- Consider using bagging-based methods, such as Random Forest, for customer churn prediction. Bagging methods, which combine multiple models trained on different subsets of the data, prove advantageous in class imbalance scenarios. By generating diverse subsets through bootstrapping, bagging reduces the variance and bias associated with imbalanced classes, resulting in improved predictions.
- Be cautious when interpreting the lift in cross-validation scores during upsampling. Upsampling techniques did not consistently yield better performance on unknown data despite demonstrating an improvement in cross-validation scores. This is because upsampling artificially increases the number of minority class samples, potentially causing overfitting and reduced performance on unknown data. The lift in cross-validation scores during upsampling is likely a result of the models being evaluated on artificial data which is easy to score on, rather than better generalizable fit which works well on unseen data.
- Explore alternative approaches, such as cost-sensitive learning or ensemble methods specifically designed for imbalanced datasets, to further enhance churn prediction accuracy. There are several alternative approaches that can be used to improve churn prediction accuracy in imbalanced datasets. These approaches include cost-sensitive learning and ensemble methods specifically designed for imbalanced datasets.

Conclusion

With the advancement in the field of AI, we see many techniques being published to address the class imbalance by generating synthetic data. We analyze the effectiveness of these techniques for churn prediction. Using our comparative case study, we build on ideas which can further be used to address the challenges in churn prediction and how it can be utilized further augmenting better decision making to business objectives. Given more time and resources there are multiple ways the model could have been improved as well as various applications for the model.

- One way to improve the model would be to add additional demographic factors and see how customer churn is affected. An example of this would be to add economic conditions such as: salary, the state of the economy, and debt. Another example would be geographical factors such as climate and living area (suburban, rural, city, etc.).
- The effects of real-time, unique events can even be implemented into the model. For example, during COVID a lot of people were forced to communicate electronically so it may be beneficial to see how the customer churn was affected by that. This could be done by implementing a factor that takes into account that gas prices and people's travel dropped significantly during that time period. Another example would be predicting the customer churn during a recession. To do this factors that consider the drop of employment and consumer spending can be implemented into the model.
- Potential applications for the model given more time and resources include: using the model to provide insight for marketing teams and management, integrating the model into existing systems, generalizing the model for use in multiple markets, and altering the model for real-time scoring.
- The customer churn can help predict future profit by not only predicting which customers are likely to leave but by also predicting how likely a customer is to sign up for a certain extra service (ex. device protection, tech support, and high-speed WIFI). The model can also be used to determine which extra services affect customer churn and are worth marketing more. If there are preexisting systems in place, the model can be incorporated into them to aid in accuracy and effectiveness.
- The model can be implemented into the customer retention system that automatically sends out ads to customers deemed as "flight risks" or sends discounts when the model predicts that customers with certain criteria churn at a certain time. Another system it could be used in an employee reward program where the model is used to predict the customers least likely to churn and the employee that gets them to sign up gets a bonus or incentive.
- The model can also be generalized so that it can be used not only for telecom churn but for all markets and industries that are based off of membership retention and subscription. Examples of these are: insurance companies, gym memberships, and streaming services. To do so, a majority of the factors used in the model would have to be more universal so that it can be related to various markets. There could possibly be some unique factors that are relatable

between markets that can be used. For example, time in gym and time watching a show are not the same but they can be standardized and used as a factor that works for both markets.

Finally, the model can be altered so that it operates with real-time data. This means that the model is continuously fed up to date data and outputting live predictions based on the data. In this case the model would receive live data and output churn risk scores for each customer. This would be a large-scale endeavor because it would mean that the model would have to have information for every single customer at all times.

References

- Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and Andreas K. Frank. "SMOTE: Synthetic minority over-sampling technique." Journal of artificial intelligence research 16.1 (2002): 321-357.
- Fernandez, Raquel, and Joao Gama. "On the importance of class imbalance in data mining." Proceedings of the 15th international conference on machine learning. 1998.
- Japkowicz, N. and S. Kotlarski. "The class imbalance problem: A systematic study." Intelligent Data Analysis 6.5 (2002): 429-449.
- Zhang, J., G. H. Huang, and Y. Yang. "Cost-sensitive learning for imbalanced data classification." IEEE Transactions on Knowledge and Data Engineering 21.9 (2009): 1536-1549.
- https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0191-6