# Report

### **Executive Summary (~300-500 words)**

This project's primary objective was to predict customer churn and mitigate the loss of high-value clients for a telecom company. Initially, the dataset comprised 99,999 rows and 225 columns. However, 40 features exhibited over 70% missing values, prompting an initial step of imputing them with zeroes, specifically in columns related to service recharging. Non-contributory elements like IDs and dates were subsequently dropped. Categorical columns underwent a meticulous imputation process using the Multiple Imputation by Chained Equations (MICE) technique. This rigorous data-cleaning process resulted in a refined dataset of 30,001 rows and 196 columns. To enhance predictive potential, a new variable 'churn' was introduced as the focal prediction variable. Subsequently, the data underwent thorough preprocessing to ensure suitability for various machine-learning models.

A critical aspect of the project involved addressing class imbalance. Strategies such as random under-sampling, random over-sampling, and no balancing were employed, each fed into distinct machine learning models—Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), and Random Forest. The focus during the evaluation was on minimizing both false positives and false negatives. Notably, Logistic Regression emerged as the standout performer, achieving an impressive recall of 0.86. Further optimization via the Grid Search method led to a refined model with a recall of 0.862, accomplished with hyperparameters c=0.01, max\_iter=300, and penalty=L2.

In contrast, when the data was introduced to a neural network with varied parameters, the resulting model demonstrated lackluster performance on the test data, registering an accuracy of just 0.55 compared to the Logistic Regression model. Consequently, it was determined that the Logistic Regression model represented the most effective tool for predicting potential customer churn.

Exploring feature importance revealed several key predictors significantly influencing the likelihood of churn. The top five predictors included: the total recharge amount for July, the total recharge amount for June, the Service scheme to avail services of Facebook for August, the Total recharge for mobile internet for August, and the Service scheme to avail services of Facebook for July.

In conclusion, this project executed a systematic and comprehensive exploration of customer churn prediction within the telecom industry. Through meticulous data cleaning, strategic model selection, and focused evaluation, the Logistic Regression model emerged as the preeminent tool for forecasting potential churn. Additionally, the identification of significant predictors offers valuable insights for targeted retention efforts. This holistic approach lays the groundwork for the telecom company to effectively safeguard its high-value customer base.

### **Business Insights and Recommendations (~500-1000 words)**

***Explain your insights into the profitability of the telecommunication service program and assess the potential impact of your model by considering misclassification costs.***

The logistic regression model demonstrates exceptional performance in both precision and recall, yielding an F1 Score of approximately 0.85. This indicates its proficiency in correctly identifying both positive (churn) and negative (non-churn) cases. With a precision score of around 0.84, the model assures that 84% of positive predictions are accurate, instilling confidence in its ability to reliably predict churn. Moreover, the recall score of roughly 0.86 emphasizes the model's skill in correctly identifying 86% of actual positive cases, crucial in avoiding the risk of overlooking potential churners.

In terms of accuracy, the model performs admirably, achieving 83.4% on the training set and 83.3% on the validation set. Its high precision and recall scores demonstrate a well-balanced approach, effectively minimizing both false positives and false negatives. This indicates the model's suitability for making precise predictions in the context of telecommunications services. This is vital for identifying potential churners while simultaneously avoiding unnecessary expenses related to retention efforts.

Considering the misclassification costs, it's evident that the cost of misclassifying a customer who eventually churns as "Not Churn" (false negatives) is substantially higher at $500 compared to the cost of misclassifying a customer who doesn't churn as "Churn" (false positives) at $300. This underscores the significance of minimizing false negatives, as losing a high-value customer carries a much greater financial impact than incurring costs associated with offering special plans or discounts.

In summary, the model exhibits significant potential to enhance the profitability of the telecommunication service program. By accurately identifying potential churners and minimizing misclassifications, the model enables the company to allocate resources more efficiently, thereby reducing unnecessary expenses and retaining valuable customers. Additionally, ongoing monitoring and refinement of the model can further enhance its performance over time. The achieved lowest misclassification cost on the training data, amounting to $965,300 when the cut-off threshold was set at 0.440, highlights the model's efficacy in identifying potential churners while minimizing false classifications. However, it's important to note that the model's performance on the validation set resulted in a misclassification cost of $1,284,000 indicating potential room for improvement in generalization to new data.

***Recommend strategies to leverage NLP techniques, recent advancements in LLMs and Generative AI to analyze customer interaction data for predicting churn. Propose innovative solutions by considering resource limitations, data privacy concerns, and implementation challenges.***

In the ever-evolving landscape of the telecom industry, customer retention stands as a pivotal goal. Our approach is a strategic amalgamation of real-time feedback analysis and proactive engagement strategies, all underpinned by a steadfast commitment to navigating resource constraints, data privacy, and implementation challenges with utmost care and precision.

Commencing with the initial stages, we meticulously collect and cleanse data, employing an array of techniques such as the removal of stop words and tokenization. This foundational step ensures that the data is refined and ready for in-depth analysis.

To capture the subtleties of customer sentiments, we leverage a diverse set of Natural Language Processing (NLP) techniques. This encompasses the application of Supervised Learning, a well-established method, alongside sophisticated deep learning models like transformers. By fine-tuning these models on the dataset, we elevate their accuracy in discerning sentiments, culminating in the ability to categorize interactions into positive, neutral, or negative sentiments. This provides us with a nuanced understanding of customer emotions.

A pivotal component of our strategy is the real-time feedback system, which spans a multitude of communication channels such as email, chatbot, and phone calls. This dynamic system swiftly categorizes interactions based on predefined sentiment thresholds. It acts as a sentinel, flagging customers who may be teetering on the edge of churning. This immediacy empowers us to intervene promptly, mitigating potential churn and fortifying customer relationships.

Embracing recent strides in Large Language Models (LLMs) and Generative AI, our approach attains further sophistication. Through the simulation of customer interactions, we gain the ability to forecast potential churn scenarios. For instance, if a simulated interaction hints at a customer's inclination towards churning, this insight serves as a clarion call for proactive measures. This proactive approach is manifested through the crafting of personalized offers and loyalty programs, each tailored to the individual, with the aim of securing their enduring loyalty.

Even after the deployment of our model, vigilance remains paramount. We steadfastly monitor and evaluate its performance, ensuring it remains finely attuned to the ever-evolving landscape of customer behaviour patterns.

In addressing the formidable challenges posed by resource limitations and data privacy concerns, we have adopted a decentralized approach. Models are trained on localized data sources, thereby safeguarding customer information. Only model updates are shared, minimizing the risk associated with data transmission. Furthermore, we embrace transfer learning, a technique that involves fine-tuning pre-trained models on a smaller dataset. This not only reduces computational demands but also diminishes reliance on extensive cloud resources.

Privacy, a cornerstone of our approach, is a paramount concern. To this end, we implement robust privacy techniques that anonymize customer interaction data. Additionally, we establish stringent data governance policies, meticulously defining access privileges and usage protocols. Routine audits and compliance checks serve as a bulwark, ensuring that these policies are upheld with unwavering diligence.

In this comprehensive approach, we forge a strategic path forward, one that seamlessly integrates advanced NLP techniques with innovative solutions, deftly navigating challenges while placing the utmost emphasis on safeguarding customer privacy and trust. This approach ensures not only customer retention but also the cultivation of enduring customer loyalty.

# Bibliography

S. Ranjan, S. Sood and V. Verma, "Twitter Sentiment Analysis of Real-Time Customer Experience Feedback for Predicting Growth of Indian Telecom Companies," 2018 4th International Conference on Computing Sciences (ICCS), Jalandhar, India, 2018, pp. 166-174, doi: 10.1109/ICCS.2018.00035.

Meskó, B., Topol, E.J. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. npj Digit. Med. 6, 120 (2023). <https://doi.org/10.1038/s41746-023-00873-0>

P. Silva, C. Goncalves, C. Godinho, N. Antunes and M. Curado, "Using NLP and machine learning to detect data privacy violations", Proc. IEEE Conf. Comput. Commun. Workshops, pp. 972-977, Jul. 2020.

[Transformers: State-of-the-Art Natural Language Processing](https://aclanthology.org/2020.emnlp-demos.6) (Wolf et al., EMNLP 2020)