Final Project Report

Analysing Feature Selection for Churn Prediction Models: Exploring the Impact of ML Model Building Blocks

Chen Ofir

A report submitted in part fulfilment of the degree of

MSc in Industrial Engineering and Management- Business Analytics

Supervisor: Eran Toch

Department of Industrial Engineering and Management

Tel Aviv University

Contents

Abstract

In today's fiercely competitive business landscape, customer churn poses a significant challenge for organizations across various industries. Accurately predicting and mitigating customer churn has become crucial for businesses striving to maintain profitability and foster long-term customer relationships. Machine learning (ML) models, such as Optimove's CatBoost, have emerged as valuable tools for churn prediction due to their ability to leverage vast amounts of customer data and identify patterns indicative of churn behavior. However, the black-box nature of these models can limit our understanding of the features that contribute most significantly to accurate predictions, therefore when trying to improve models the data teams have no lead on what type of data points have better potential of leading to model improvement.

In this research we analyzed 85 Catboost models generated by Optimove for Churn prediction and aim to explore the impact of features selected by the model and their effectiveness. By analyzing the lift in churn prediction for Catboost models generated on different data sets from the marketing industry, we seek to uncover insights into the specific attribute types that enhance the model's predictive capabilities. Understanding the role of these features can enable businesses to make informed decisions for developing and exposing additional data points or improving customer-focused strategies, ultimately leading to better churn prediction and retention efforts.

Introduction

In today's business landscape, companies prioritize robust CRM strategies, with customer churn prevention as a vital element. Predicting Churn accurately is crucial as retaining existing customers, rather than solely focusing on acquiring new ones, holds the key to a significant portion of a company's revenue. Losing valuable customers not only affects a company's revenue but also undermines its growth potential, as acquiring new customers may be a much more expensive process, that might become not profitable with low retention margins [1]. To combat churn effectively, businesses increasingly turn to data-driven approaches, utilizing data mining concepts, machine learning, and meta-heuristic algorithms, aiming to accurately estimate customer survival, hazard functions, identify potential churn, approximate churn timing, and enhance services based on hybrid models to mitigate churn.

Various churn prediction models have been developed by different authors, employing diverse methods and datasets. For instance, J. Burez et al. utilized random and advanced undersampling along with Gradient Boosting Machine and WRF on European churn modeling datasets, achieving improved prediction accuracy despite individual issues. Effendy et al. combined sampling with WRF on categorical churn data, enhancing accuracy and resolving imbalanced data problems. Meanwhile, Lu et al. employed logistic regression with Gentle AdaBoost on telecom data to accurately define high-risk customer groups. These models present advantages such as increased prediction accuracy, but some also come with drawbacks like reduced overall performance or limitations in addressing specific challenges. [2]

In this research we will be focusing on Catboost ML churn prediction model, which has demonstrated promising results in churn prediction, empowering businesses to proactively engage with customers on the verge of churning. Optimove’s Churn prediction provides heuristic results for the end users, to identify the groups of customers with higher risk of churn, and by that communicate with these audiences differently, making sure that customers with higher risk will not be missed and fall between the cracks among wider groups. [3]

While CatBoost has proven accurate in identifying potential churners, its inner workings remain opaque, making it challenging to discern which customer attributes and behaviors play a pivotal role in churn prediction. This black-box nature prevents businesses from gaining actionable insights into the underlying drivers of customer churn, limiting their ability to implement targeted strategies for customer retention. Therefore, it becomes imperative to investigate and interpret the impact of the features selected by the CatBoost model on its predictive performance.

This research aims to analyze the performance of 85 models and estimate the fundamental building blocks of Optimove's CatBoost churn prediction model. Additionally, we will explore various techniques to identify variations in the importance of different feature types. These insights will aid in identifying the key features to prioritize when expanding the list of attributes on which the model is built.

This research project aims to make several important contributions to the field of churn prediction and customer retention. First, it seeks to uncover the features that significantly impact the churn prediction performance of the Optimove CatBoost model. Second, it aims to provide practical insights into the design and enhancement of customer-focused strategies by recommending specific features that should be considered or improved to enhance churn prediction capabilities. Ultimately, the project strives to advance the understanding of churn prediction in the context of machine learning models and their potential for driving customer retention efforts.

Literature Review

Churn Prediction

The contemporary business landscape is marked by intense competition and saturated markets, prompting companies to seek revenue through long-term customer relationships. This shift towards a customer-centric approach, fueled by the abundance of customer-transactional data, has elevated Customer Relationship Management (CRM) to the forefront of marketing strategies, prompting significant investments in CRM. Organizations recognize that their existing customer base is a valuable asset, and it has been established that retaining and satisfying existing customers is more profitable than constantly acquiring new customers, given the high attrition rate associated with the latter. Predicting customer churn, a vital metric in CRM, enables the formulation of targeted retention strategies to limit losses and enhance marketing decisions [1].

To answer questions about which customers are likely to churn and why, a classification of customers is essential. Churn prediction involves identifying customers who are likely to churn in the near future, primarily based on historical data that includes information about previous instances of churn. This approach entails comparing the characteristics of past churners with those of current customers. The process identifies potential churners by identifying customers whose attributes align with those of earlier churn cases. This classification process, also referred to as pattern recognition, discrimination, or supervised learning, enables businesses to proactively address customer churn [4].

CatBoost

Gradient boosting is a robust machine-learning technique, known for achieving outstanding results across diverse practical tasks, particularly in scenarios involving heterogeneous features, noisy data, and complex dependencies such as web search, recommendation systems, and weather forecasting. Its theoretical foundation demonstrates how powerful predictors can be built through the iterative combination of weaker models (base predictors) in a greedy manner, akin to gradient descent in a function space. While most widely used implementations of gradient boosting employ decision trees as base predictors, these methods struggle when dealing with categorical features, which are often present in real-world datasets and hold significance for accurate prediction. This paper introduces a novel gradient boosting algorithm named CatBoost, capable of effectively handling categorical features during training, providing advantages over existing implementations like XGBoost, LightGBM, and H2O. CatBoost offers both CPU and GPU implementations, with the GPU version showcasing faster training times compared to other leading GBDT GPU implementations, while the CPU scoring implementation also proves superior on similar ensemble sizes. CatBoost is released as open source, further contributing to the advancement of gradient boosting techniques in the machine-learning community. [5]

Like many machine learning algorithms, Catboost can be sensitive to the features (input variables) it's provided with during training. The choice of features can significantly impact the model's performance, generalization, and its ability to handle the underlying patterns in the data, feature selection, engineering, scaling, and understanding the domain context are critical to ensuring CatBoost's optimal performance. Proper preprocessing and careful consideration of the features can help mitigate issues related to sensitivity and lead to a more robust and accurate model. [6]

Optimove

Optimove stands as a real-time Customer Data Platform (CDP), distinguished by its Relationship Marketing Hub meticulously crafted to propel quantifiable growth through the orchestration, measurement, and optimization of intelligent campaigns. The foundation of Optimove's prowess lies in its commitment to a scientific-first methodology. This manifests in the platform's capacity to seamlessly integrate diverse customer data into a singular panorama, thus facilitating a holistic comprehension of customer dynamics. Within this contextual framework, the platform autonomously excavates invaluable insights, thereby augmenting the efficacy of campaigns. A defining facet is the platform's flexibility, allowing enterprises to tailor their key performance indicators (KPIs) for maximal real-world business impact.

As attested by a cohort exceeding 500 brands across multifarious industries, Optimove has solidified its standing as an indispensable asset. Evidencing its efficacy, it has gained prominence among global pioneers spanning Retail, E-Commerce, Gaming, Entertainment, Financial Services, and more. While it's important to acknowledge that Optimove doesn't claim omniscience in eradicating customer abandonment, it excels at identification. By preemptively recognizing potential churn-prone customers, it aids in the recalibration of marketing strategies, resulting in the mitigation of churn. Moreover, the platform proactively proposes measures during the nascent phases of the customer lifecycle, accentuating loyalty, engendering higher engagement, and curtailing initial abandonment.

Of significant note is Optimove's computation of an individualized churn criteria for each customer. This metric underpins a bespoke approach to rekindling relationships with customers who have lapsed. The establishment of criteria for defining churn is an undertaking that Optimove meticulously tailors based on end-user behavior. This process is a collaborative endeavor, working with the marketing teams of its clients. The outcome is a meticulously crafted churn definition that harmonizes data analysis with a holistic business perspective. [7]

Research Objective

This research project aims to shed light on the influence of specific features selected by the Optimove CatBoost churn prediction model. By comparing the lift in churn prediction for different models and revealing groups of features which influence positively on the model accuracy, we seek to identify the key attributes that significantly contribute to the model's performance. These findings will not only enhance our understanding of customer churn dynamics but also enable businesses to develop tailored data points to improve the model's accuracy even further, enable to bridge the model's building blocks to the marketing teams and expedite the model engagement which might lead to wider use of the model outputs.[8]

Methodology

The methodology for this research project involves analyzing churn prediction models implemented for 85 different clients' database. Each model represents a dataset of one of Optimove's clients, mainly from the eComm and Gaming industries, and has its own set of potential attributes, consisting of hundreds of different attributes, and millions of records, while each record represents an end-user single customer view. As part of the model generation, a feature selection is conducted out of the list of optional ones and are ranked based on the level of relevancy for the model training. For this research purposes, the top 5 ranked features selected by the model's feature selection process are considered. These features are categorized into six groups: Geo Demographic, Product Preference, Recency, Frequency, Monetary, and Real-time Events. The correlation between these features and the lift in churn prediction is then examined.

- Geo Demographic: This group focuses on attributes related to the geographic and demographic information of customers. It may include variables such as location, age, gender, income level, occupation, balance, name, region, and other demographic factors that can provide insights into the customer's profile and potential churn behavior.
- Product Preference: This group encompasses attributes that capture customer preferences and behaviors regarding specific products or services offered by the business. It can include variables such as preferred product categories, purchase history, brand affinity, platform preference or any other indicators of customer preferences in relation to the offerings.
- Recency: The recency group refers to attributes related to the temporal aspect of customer behavior. It involves measuring the time elapsed since a customer's last interaction or purchase. This information can help identify customers who have become less engaged or are at higher risk of churning due to extended periods of inactivity. Recency can be a continuous attribute that measure days since last different types of activities such as making a purchase, depositing, or placing a bet, depending on the vertical. In addition, there are attributes that groups clusters based on different timeframes since last activity such as weeks since last activity or different bespoke cluster groupings.
- Frequency: The frequency group focuses on attributes that reflect the frequency of customer interactions or transactions. It measures how often a customer engages with the business, such as the number of purchases made, logins to the website or physical store, or interactions with customer support. Higher frequency may indicate higher engagement and lower churn propensity; for example, a customer with a frequency of using the platform every 2 days during the last 3 months will have a very high survivability rate due to high engagement, as opposed to one that engage with the platform twice a month.
- Monetary: The monetary group involves attributes that quantify the monetary value associated with customer transactions or interactions. It includes variables such as total purchase value, average transaction amount, customer lifetime value, or any other financial indicators that provide insights into the customer's spending patterns and potential churn risk based on their monetary contribution.
- Real-time Events Gathered: This group captures attributes related to real-time events or actions taken by customers that might impact their churn likelihood. These events are not received by conventional data delivery cycles and are acquired independently of standard site logins, encompassing even seemingly routine engagements like page visits. This capability endows

Optimove with a nuanced understanding of customer behavior, enabling the platform to digest underlying customer behavior, thereby enhancing its capacity for effective churn prediction and mitigation.

Selected Sample Size

The sample size of 85 models was chosen based on the work of Looney (2018) [9], which states that the minimum sample sizes needed for a significant Pearson correlation at a 0.05 level of significance, a power of 0.80, and for different effect sizes. For an effect size of 0.30, the minimum sample size needed is 85. We assume that the study's power is 0.80, which means that there is an 80% chance of detecting a true effect size of 0.30, which indicates a solid enough sample size.

We also decided to use Spearman's rank correlation coefficient rather than Pearson's correlation coefficient because our dataset consisted of only 85 samples, and we wanted to minimize the influence of outliers on our correlation analysis. Spearman's rank correlation is a non-parametric measure that assesses the monotonic relationship between variables based on their ranks, making it less sensitive to extreme values compared to Pearson's correlation, which assumes linearity and can be affected by outliers. This choice allowed us to obtain a more robust and reliable measure of association given the limitations of our small sample size [10]

Churn Prediction - Lift Calculation

As part of the onboarding process of a new client to Optimove, the data onboarding team analyzes the behavior of th end users and defines migration rules to define on which point an Active customer should be flagged as churned, it varies between industries and between different clients on the same industry, but once define a threshold, it is a fixed rule of defining a churned customer. For example, in eComm industry a period of inactivity of 90 days will indicate churn in many cases, but there are some slow pace eComm clients with a migration rule of 180 days of inactivity to flag a customer as churned.

The diversity in churn definition isn't solely constrained to varying industries but also arises from the distinctive behaviors exhibited by end users within each client. This differentiation is evident even within the same industry, such as selling socks or cars, as the customer engagement, buying cycles, and inactivity patterns can significantly differ based on the specific brand, product offering, customer base, and business objectives. Hence, optimizing the churn definition for each client within the same vertical ensures a tailored approach that accurately captures the dynamics of their unique customer behavior, thus enhancing the precision and effectiveness of churn identification and retention strategies.

When dealing with rare events like customer churn in predictive modeling, the lift metric becomes a suitable indicator of accuracy. The rarity of churn instances makes traditional accuracy metrics less informative Lift overcomes this limitation by considering the relative performance of a model or strategy compared to a baseline, reflecting the model's ability to identify churn instances more effectively than random chance. [11]

The lift calculation is a measure used to assess the effectiveness of a churn prediction model in identifying customers at risk of churning. It compares the actual churn rate of a specific decile, with the average churn rate of the entire audience [12]. In our experiment we inherit the suggested methodology and compare the top 10% of customers predicted to be at the highest risk of churn with the entire audience.

To calculate the lift, we followed the below steps:

1. Identify the top 10% of customers who are predicted to be at the highest risk of churn based on the churn prediction model's scores or probabilities.
2. Calculate the actual churn rate within this group. This is done by determining the proportion of customers within this top 10% group who actually churned during a specific time period.
3. Calculate the average churn rate of the entire audience or customer base over the same time period. This involves determining the proportion of customers who churned out of the total population.
4. Divide the actual churn rate of the top 10% group by the average churn rate of the audience.

$$Lift = \frac{\%Churn_{top\ 10\%}}{\%Churn_{All\ Population}}$$

The resulting value is the lift, which indicates the degree to which the churn prediction model outperforms random chance in identifying customers at risk of churn. The lift calculation provides insights into the model's ability to prioritize customers who are more likely to churn, allowing businesses to allocate resources and implement targeted retention strategies for these high-risk customers. A lift value greater than 1 suggests that the model is effective in identifying customers with a higher churn propensity compared to the overall average churn rate. For example, a lift of 1.5 means that the top 10% group has a churn rate 1.5 times higher than the average churn rate. Figure 1 emphasizes the calculation of 4 ratios conducted by the proportion of top 10% lift vs top churn %, highlighting the lift as an output of the two variables.

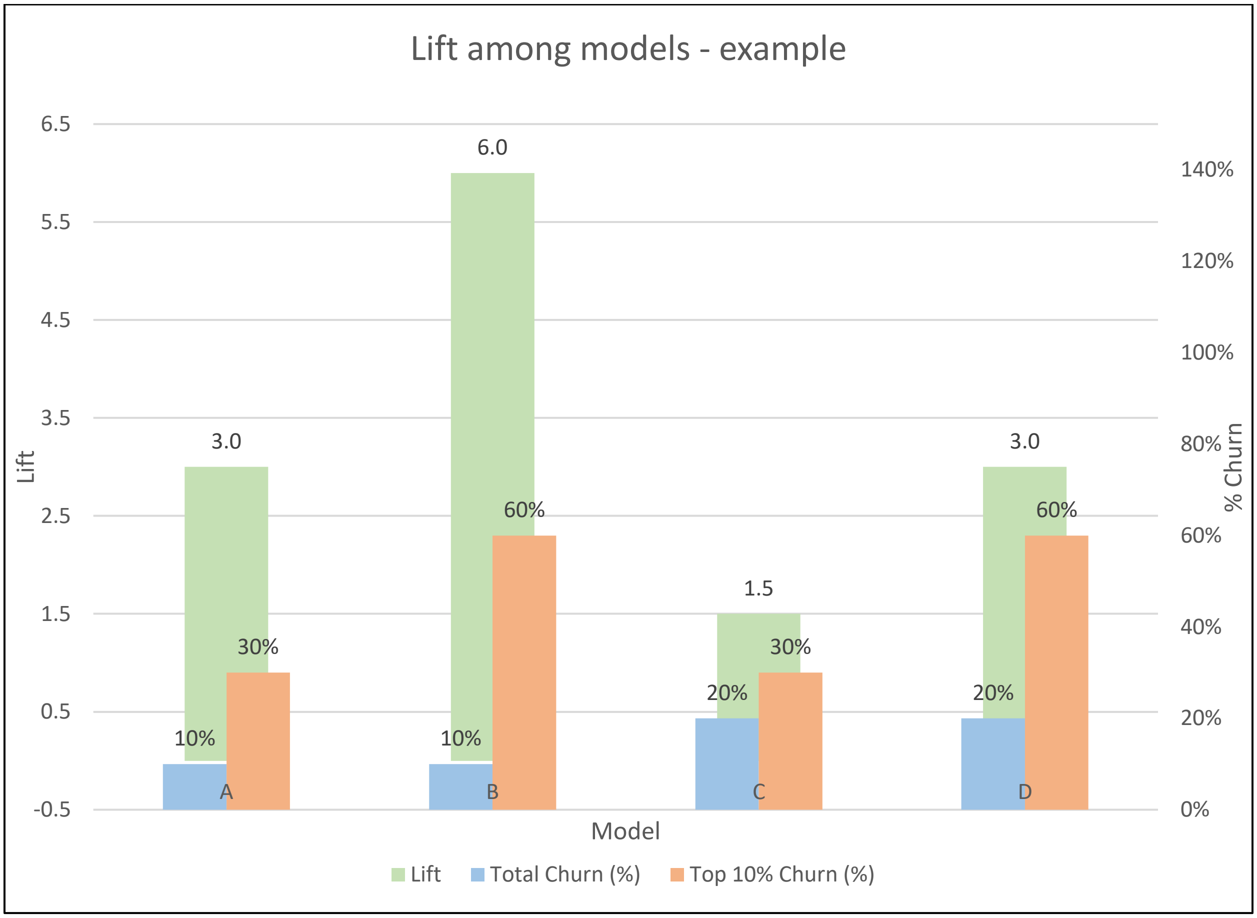

Figure 1 – List among models – example

By examining the correlation between the selected features and the lift, this study aims to identify the extent to which each category of features influences the churn prediction accuracy. The analysis will provide insights into the relationship between specific attributes and churn, shedding light on the features that significantly contribute to the model's predictive capabilities. By that we will have better understanding of what types of features worth adding to a dataset in the purpose of fine-tunning an existing model.

In this experiment we used a normalized lift value to avoid exposing clients' actual lift calculations - By dividing the lift value of a specific client by the average lift of all observed models, the client's lift is scaled and presented relative to the overall performance. This normalization ensures that the lift values remain confidential while still allowing for meaningful comparisons. The normalized lift allows for meaningful insights and comparisons while maintaining privacy and confidentiality, making it a valuable tool in evaluating churn prediction models and retaining client satisfaction.

$$Normalized\ Lift = \frac{Lift_{model}}{Lift_{All\ models}}$$

For instance, let's consider the example in figure 2 below, illustrating the normalized lift values for different clients. E.g. In Figure 2, the average life is 3.7 and by that a lift value of 4 will be normalized to 1.1, the pre-normalized values represents the actual lift of a model, while the post-normalized value represents the value we will use in the coming analyses.

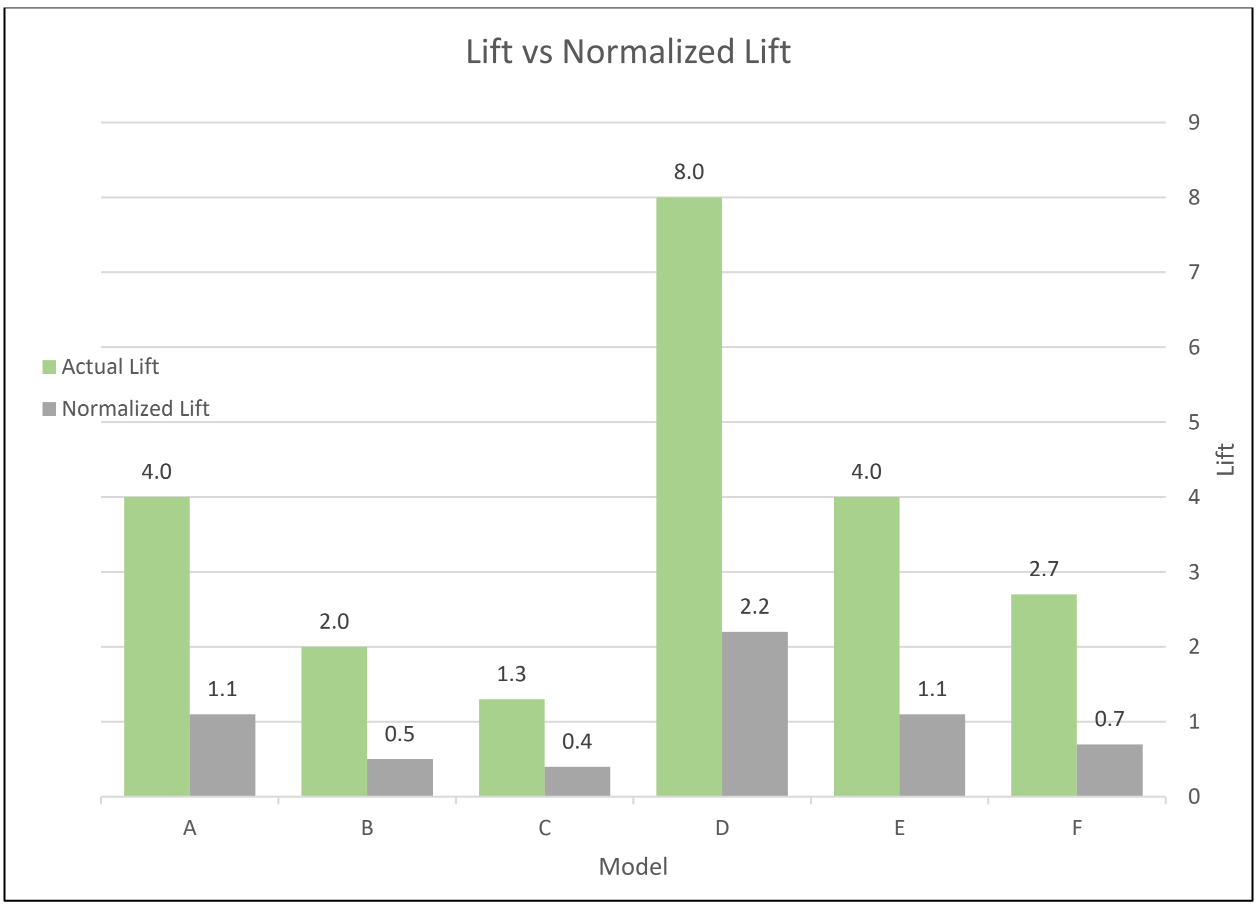

Figure 2 – An example of lift vs Normalized Lift.

Results

Initially, a benchmark analysis of churn rates was conducted, projecting 1 month ahead, across all models within the sample. The findings revealed a median rate of 30.5%, with an average of 33%. Subsequently, a graphical representation reveals that churn rate and the model's normalized lift value was constructed. Notably, a distinct counter-directional pattern emerged, indicating that heightened actual churn rates were consistently accompanied by diminishing normalized lift values. This trend is visually expounded in Figure 3 below, underscoring the inverse relationship between augmented normalized lift and decreased actual churn rates.

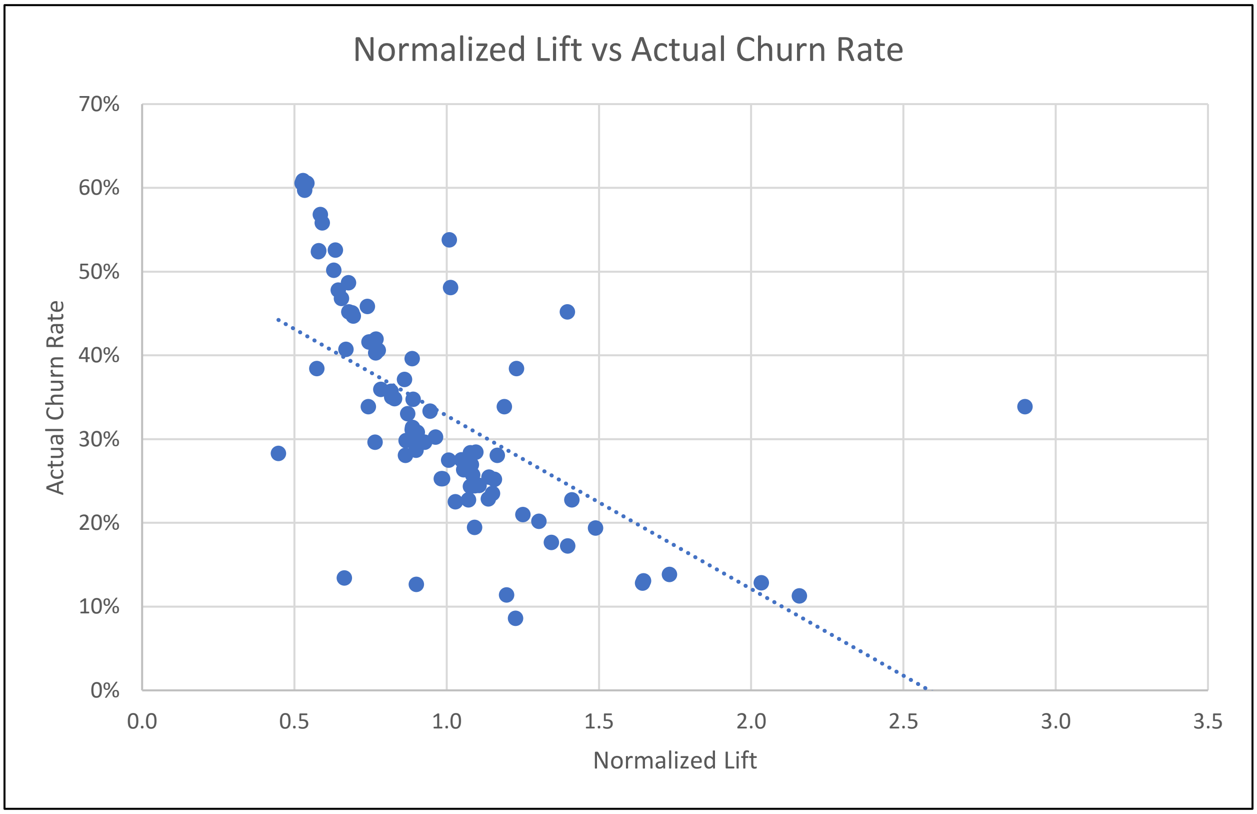

Figure 3 – Trend of Normalized lift against Actual Churn Rate

Continuing from the previous analysis, as presented in figure 4, our investigation progressed by segmenting observations into four distinct groups, stratified according to the genuine churn rate observed across the entire population. These groups were then juxtaposed against their respective normalized lift values. Remarkably, a congruent pattern emerged, a phenomenon that aligns logically when we factor in the lift's intrinsic elements.

Recognizing the rationale behind segregating models based on actual churn rate, we acknowledge that this avenue would indeed be a logical progression. However, due to the constraints imposed by the available sample size, regrettably, this avenue couldn't be explored within the scope of this research endeavor.

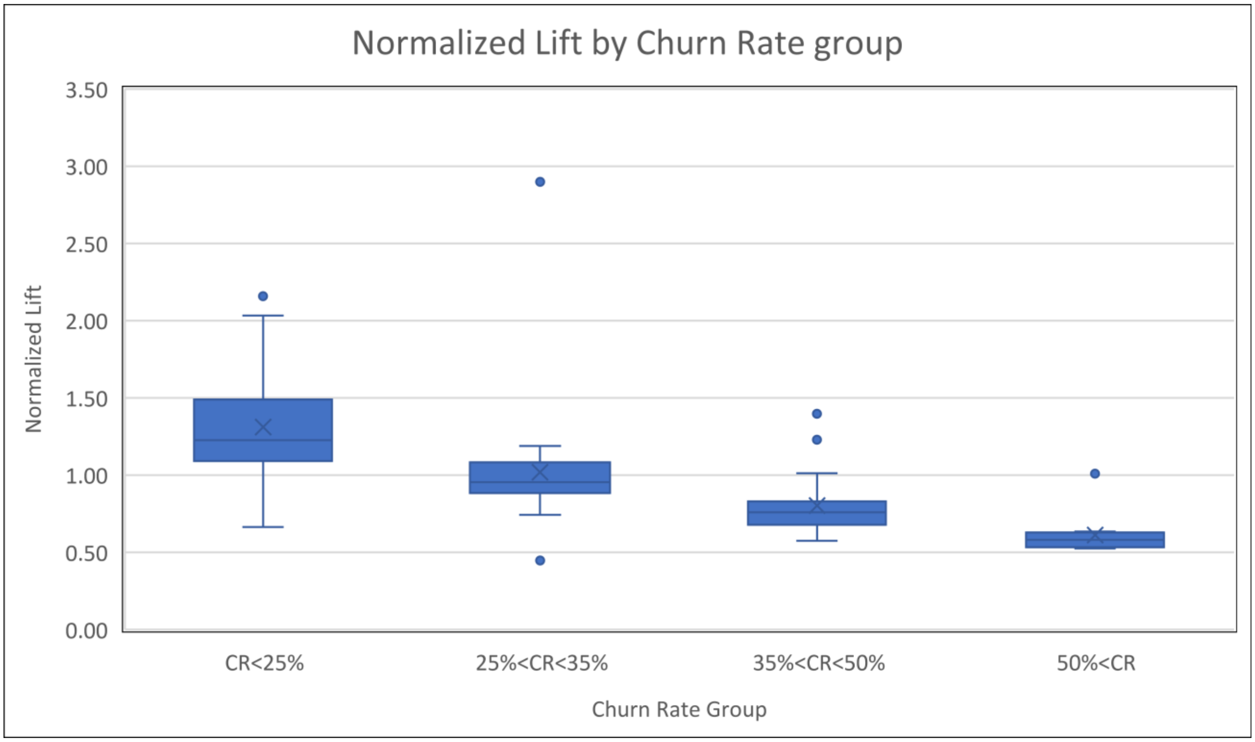

Figure 4 – Boxplot of all observations' Normalized Lift by Churn Rate

In CatBoost, feature importance is assessed by comparing the performance of the model under two scenarios: one where the feature is included (normal scenario) and the other where the feature is removed from all trees in the ensemble (model without the feature). The difference in the chosen metric, often the loss function, between these two scenarios indicates the feature's importance. The larger the difference, the more significant the feature is for the model's predictive performance. While the precise method of deriving the model without the feature isn't explicitly documented in the CatBoost documentation, it's likely based on a procedure that selectively excludes the feature during tree construction or incorporates a modification in the ensemble calculation to achieve the comparison. Although the specific details of this process may not be explicitly stated, the overall concept provides insight into how CatBoost assesses feature importance in its models.

Figure 5, presented below, provides a tangible illustration of a feature importance outcome procured through the Catboost model. This depiction offers insight into the significance attributed to various features within the model's predictive framework.

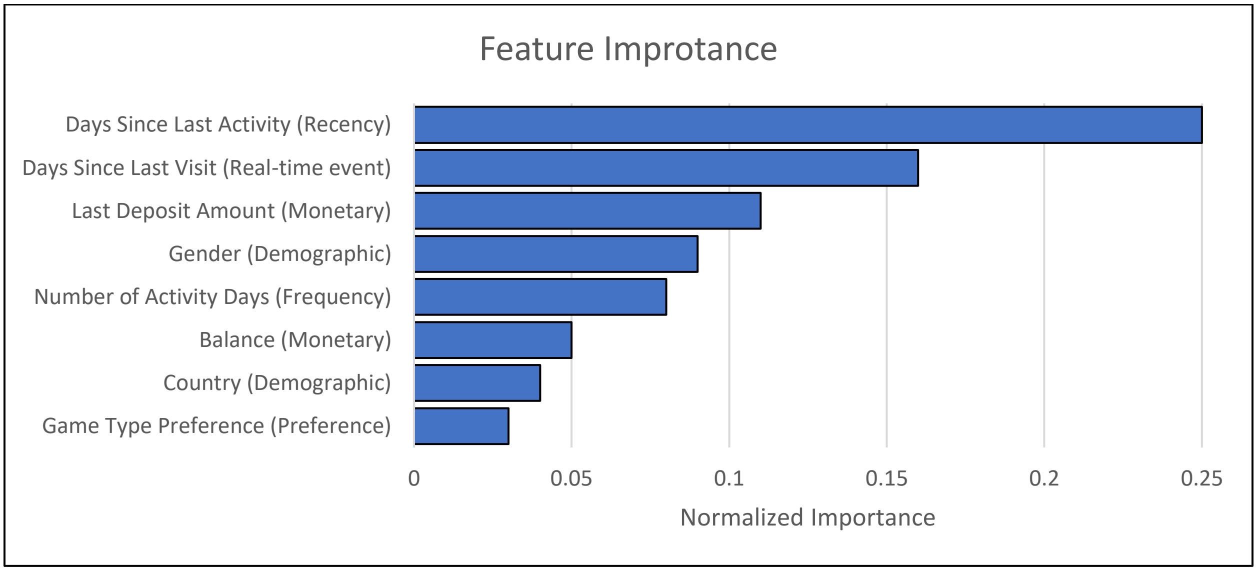

Figure 5 – Feature importance example, features ordered by importance, feature type in brackets.

Utilizing linear regression, an in-depth examination of the correlation between all feature types (Geo Demographic, Recency, Frequency, monetary, Real-Time events, Preference) and the normalized lift was undertaken. To ensure the model's integrity, we adopted a rigorous methodology that included partitioning the dataset into a training set and a testing set, maintaining an 80/20 distribution ratio. This division strategy allowed the model to learn from the training data while keeping unseen testing data for validation. As presented in figure 6, the outcomes of this comprehensive analysis were strikingly robust, culminating in an exceptionally high confidence level for the model. In addition, the resultant r-square value is 0.456, which signifies the model's robustness, denoting a potent explanatory framework. Notably, each feature within the analysis garnered a positive coefficient, underpinning their collective contribution to the augmentation of lift in a favorable direction.

	Coefficient	Std Error	t	P>|t|	[0.025	0.975]
Const	-0.7119	1.571	-0.453	0.652	-3.854	2.430
GeoDemographic	0.3090	0.312	0.991	0.325	-0.14	0.932
Recency	0.3663	0.314	1.167	0.248	-0.261	0.994
Frequency	0.3807	0.314	1.214	0.230	-0.247	1.008
Monetary	0.2422	0.321	0.753	0.454	-0.401	0.885
Real Time Events	0.8799	0.329	20.672	0.010	0.221	1.539
Preference	0.2996	0.323	0.927	0.358	-0.347	0.946

Figure 6 – OLS Regression results

Next, we conducted a Spearman correlation analysis involving the feature types of top 5 ranked features that were selected. The motivation behind this approach stems from our belief that while all features positively contribute to the anticipated future lift, certain features exhibit stronger correlations and, consequently, wield more significant impacts. By subjecting these top-ranking features to a correlation test, we gain insights into their individual influence on the predicted normalized lift values.

Due to the limited number of observations, we ran the correlation on the entire population altogether to avoid segmenting too small sample size, and the results of the analysis revealed significant correlations between the selected features and the lift in churn prediction. As seen in figure 7, the most positively correlated features were found to be Recency (0.27), Frequency (0.25), and Real-time Events (0.22). This suggests that these attributes have a stronger influence on predicting churn and should be considered as important factors when selecting the top 5 features for the model.

On the other hand, the features Monetary (-0.23), Geodemographic (-0.21), and Preference (-0.19) demonstrated negative correlations with the lift. This indicates that these attributes have a weaker impact on churn prediction and including them among the top 5 features may not yield improved results for the model - it doesn’t mean that these attributes should be removed, as it had been selected as top 5 features in the existing dataset, but it’s likely that adding in a new data source tagged as ‘Recency’ ‘Frequency’ or ‘Real-time’ that will make its way to the top 5 features will improve the model more than adding in a new ‘Monetary’, ‘Geodemographic’ or a ‘preference’ one.

It is essential to note that within our analysis, a Spearman correlation ranging from 0.2 to 0.3 is indicative of a weak to moderate correlation. This suggests the potential for the observed relationship to be influenced by chance rather than a robust connection. Nevertheless, it's important to recognize that correlations within the moderate range carry an elevated likelihood of attaining statistical significance. This insight underscores the nuanced approach necessary for interpreting these correlations, wherein both the magnitude of the correlation and its statistical significance must be carefully weighed to gauge the potential impact and relevance of the identified features on the predicted normalized lift values.

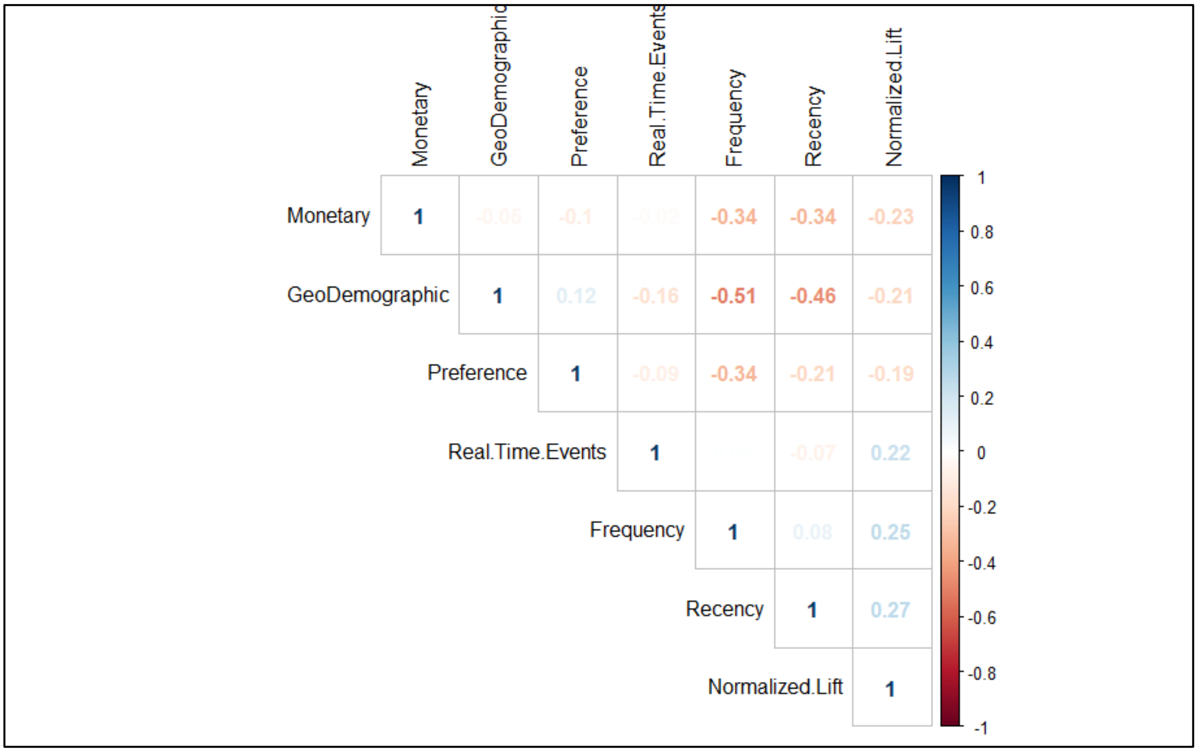

Figure 7 - Spearman's Correlation Matrix results.

The results align with the business understanding that high Recency, high frequency and frequent site visits (real-time events) capture the customer's behavioral patterns and engagement with the business. These factors provide insights into the customer's level of interest and satisfaction with the service, making them highly relevant predictors of churn, as opposed to monetary values which can be referred to a possibility of One-time High-value Customers who might have made significant monetary transactions or product references in the past, but it doesn't guarantee their continued loyalty. Geodemographic factors are more related to the characteristics of the customer rather than their behavior. While they can provide valuable insights into customer segments, they might not be as indicative of individual churn probability compared to the actual interaction patterns.

While monetary values and product preferences are highly relevant for future value prediction and can provide some valuable insights in churn prediction models, they might not be as strongly correlated to churn probability as recency, frequency, and site visits, which directly reflect a customer's active engagement and behavior with the service. As a result, businesses often prioritize the analysis of the engagement metrics when identifying and mitigating customer churn.

Discussion

Attribute selection serves as a pivotal compass in the data mining process, especially when faced with extensive datasets. It requires a delicate equilibrium between incorporating relevant attributes and pruning those that might introduce noise or redundancy. This process involves leveraging domain expertise and data-driven insights to discern the attributes most likely to impact analysis outcomes. Striking this balance is critical to avoid overwhelming the model with irrelevant data or omitting vital insights, ensuring that the selected attributes collectively contribute to meaningful patterns and informed decision-making. [13]

The results of this study shed light on the importance of feature selection in optimizing churn prediction models using Optimove's CatBoost algorithm. By analyzing data from 85 models, we gained valuable insights into the correlation between selected features and the lift in churn prediction, with the ability of enabling a more accurate way to define the list of attribute to take part in the feature selection process, and allow a new approach of adding fields into data sets, concentrating in ones that are more likely to provide superior results.

Although all features had a positive impact on the predicted value, we further looked at the spearman correlation, and interpret the results as an indication that Recency, Frequency, and Real-time Events highlight a more significant impact on predicting churn. These findings align with previous research indicating that recent customer behaviors, frequency of interactions, can serve as reliable indicators of churn risk [14], although we expected monetary values to be part of this groups as well, but the result of this research shows otherwise. The inclusion of the first 3 feature types in the top five attributes selected for the churn prediction model is crucial for improving its accuracy and performance, therefore while looking for adding new attributes to the data set for model learning and improvement, these should be prioritized on top of other type of attributes to be added.

On the other hand, the negative correlations found for Monetary, Geo Demographic, and Preference features suggest that these attributes have lower influence on churn prediction, although still positive, as having these as part of the top 5 feature indicates less features that are from the superior type on these high ranks. This implies that allocating resources towards enhancing these aspects may not result in substantial improvements in churn prediction accuracy. However, it is important to consider the specific context of the business and customer base, as the influence of these features may vary in different industries or segments.

These findings have another practical implication for businesses aiming to improve customer retention strategies. By focusing on the influential features identified in this study, such as Recency, Frequency, and Real-time Events, businesses can develop targeted interventions to mitigate churn risks. For example, strategies could involve personalized offers based on recent customer behavior, increasing engagement through frequent communication, and responding promptly to real-time events that indicate potential dissatisfaction or intent to churn.

It is essential to acknowledge the limitations of this study. The analysis was based on a specific and limited dataset and focused on the features selected by Optimove's CatBoost algorithm. The results may not be directly generalizable to all churn prediction models or industry sectors. Additionally, the study did not explore the interactions between features or evaluate the performance of the entire churn prediction model comprehensively.

To further enhance churn prediction models, future research could delve into exploring feature interactions and incorporating additional data sources or feature engineering techniques. Additionally, conducting experiments and validating the findings in different business contexts would strengthen the robustness and applicability of the results.

In addition, as the analysis discovered a clear negative correlation between the Churn Rate and the actual lift, we would suggest a further experiment with significantly higher number of observations, splitting the experiment per group of Churn Rate, as suggested in figure 4, to reduce the implication of the wider audience churn rate and potentially discover different findings per group.

Conclusions

In conclusion, this research project explored the impact of selected features on the churn prediction model implemented using Optimove's CatBoost algorithm. By analyzing data from 85 clients/models, we were able to identify the most influential features and their correlation with the lift in churn prediction.

Our findings revealed that Recency, Frequency, and Real-time Events emerged as the most positively correlated features, indicating their significant impact on predicting churn. These attributes provide valuable insights into recent customer behaviors, engagement levels, and real-time events that contribute to churn risk. Therefore, it is recommended to prioritize these features when considering adding attributes to enrich the model set of features for the churn prediction model. Conversely, Monetary, Geodemographic, and Preference features demonstrated lower correlations with the lift, therefore it is advised to put less effort in adding such attribute, in case of feature engineering for model improvement.

It is important to note that while these findings provide insights into feature importance, the specific characteristics and dynamics of each business and customer base may vary. Therefore, it is advisable to conduct further research and experimentation to validate these results in different contexts.

Overall, this study contributes to the understanding of feature selection in churn prediction models and provides guidance on the key attributes that can improve the performance of Optimove's churn prediction algorithms. By leveraging the identified influential features, businesses can optimize their customer churn prediction and by that improve customer retention efforts, increase customer loyalty, and maximize the value of their customer base.

References

[1] Coussement, K., Benoit, D. F., & Van den Poel, D. (2010). Improved marketing decision making in a customer churn prediction context using generalized additive models. Expert systems with Applications, 37(3), 2132-2143.

[2] Ahmed, A., & Linen, D. M. (2017, January). A review and analysis of churn prediction methods for customer retention in telecom industries. In 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 1-7). IEEE.

[3] Jain, N., Tomar, A., & Jana, P. K. (2021). A novel scheme for employee churn problem using multi-attribute decision making approach and machine learning. Journal of Intelligent Information Systems, 56, 279-302.

[4] Lazarov, V., & Capota, M. (2007). Churn prediction. Bus. Anal. Course. TUM Comput. Sci, 33, 34.

[5] Dorogush, A. V., Ershov, V., & Gulin, A. (1810). CatBoost: Gradient boosting with categorical features support. arXiv 2018. arXiv preprint arXiv:1810.11363.

[6] Jabeur, S. B., Gharib, C., Mefteh-Wali, S., & Arfi, W. B. (2021). CatBoost model and artificial intelligence techniques for corporate failure prediction. Technological Forecasting and Social Change, 166, 120658.

[7] https://www.optimove.com/faq/what-is-optimove

[8] Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: an interdisciplinary review. Journal of big data, 7(1), 1-45.

[9] Looney, S.W (2108). Sample Size for Correlation Coefficient Inference. SM Journal of Biometrics & Biostatistics, 3(1), 1027.

[10] De Winter, J. C., Gosling, S. D., & Potter, J. (2016). Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. Psychological methods, 21(3), 273.

[11] Luo, M., Wang, Y., Xie, Y., Zhou, L., Qiao, J., Qiu, S., & Sun, Y. (2021). Combination of feature selection and catboost for prediction: The first application to the estimation of aboveground biomass. Forests, 12(2), 216.

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

[13] Kononenko, I., & Hong, S. J. (1997). Attribute selection for modelling. Future Generation Computer Systems, 13(2-3), 181-195.

[14] Ibitoye, A. O., ONIME, C., & Zaki, N. D. (2022). Socio-Transactional Impact of Recency, Frequency, and Monetary Features oN Customers' Behaviour in Telecoms' Churn Prediction. Iraqi Journal for Computer Science and Mathematics, 3(2), 101-110.