# Table of Contents

# Introduction

In the dynamic landscape of customer relationship management, understanding and predicting customer churn has become paramount for businesses aiming to optimize customer retention strategies. Customer churn, denoting the rate at which customers discontinue their association with a company, poses a formidable challenge across various industries. Recognizing the patterns and factors influencing customer churn is pivotal for organizations striving to enhance customer retention strategies. To confront this challenge head-on, we present a sophisticated Logistic Regression model—an advanced predictive tool designed to forecast customer churn and furnish data-driven insights that empower businesses to make informed decisions.

This report focuses on the prediction of customer churn using a Logistic Regression model. Our primary objective is to construct a precise and effective model capable of estimating the likelihood of customer churn through an in-depth analysis of a dataset enriched with diverse features. Within this dataset lie a myriad of attributes, and our meticulous feature selection and analysis have identified the top 10 features that wield significant influence over predicting customer churn. Simultaneously, we aim to empower decision-makers with actionable insights, allowing them to proactively navigate potential churn risks and implement targeted retention strategies.

As we embark on this analytical journey, the subsequent sections will unravel the methodology employed, elucidate the key features identified, and delve into the model's overall performance. By the end of this report, our aspiration is that organizations will be better equipped to anticipate and respond to customer churn, fostering enduring customer relationships in an increasingly competitive market.

# Exploratory Data Analysis

Within this section, our primary aim is to unravel the intricacies of the dataset, shedding light on pivotal variables that significantly influence our predictive model. Our goal is to deliver a succinct yet comprehensive summary of the data, strategically emphasizing pertinent variables and essential statistics. Through this exploration, we lay the groundwork for a nuanced understanding of the dataset's landscape, incorporating detailed examinations of highlighted features, a comprehensive data summary, and an exploration of correlations.

## Highlighted Features

In this section, we delve into the heart of our analysis by spotlighting key features that offer invaluable insights into the intricate landscape of customer churn. Each selected feature contributes a unique perspective, empowering our model to discern and interpret patterns indicative of potential churn. The subsequent sections will unveil detailed descriptions of these features, elucidating their pivotal roles in enhancing the predictive prowess of our model.

Our meticulous feature selection process involved stringent testing to ensure the meaningful contribution of the chosen variables to the prediction of customer churn. Techniques such as imputation addressed missing values, categorical features were appropriately encoded through the creation of dummy variables, and the dataset's imbalance was rectified using SMOTE, bolstering the robustness of our predictive model. Here, we present an ordered overview of the selected features, accompanied by concise descriptions outlining their respective roles within the dataset:

1. imp_AccountAge: Imputed data representing the mean values of the Account Age feature for the instances with missing values.
2. SubscriptionType_Premium: A binned categorical feature representing customers with a premium subscription.
3. PaymentMethod_Credit card: A binned categorical feature representing customers using credit cards for payment.
4. SubtitlesEnabled_Yes: A binned categorical feature representing customers with enabled subtitles.
5. SupportTicketsPerMonthBin_(0, 2]: A binned feature representing customers with 0 (inclusive) to 2 (exclusive) support ticket submissions per month.
6. UserRatingBin_(0, 2]: A binned feature representing customers with ratings falling between 0 (inclusive) and 2 (exclusive).
7. Gender_Female: A binned categorical feature representing female customers.
8. ContentType_TV Shows: A binned categorical feature representing customers who primarily watch TV shows.
9. ContentType_Movies: A binned categorical feature representing customers who primarily watch movies.
10. ParentalControl_Yes: A binned categorical feature representing customers with enabled parental control.

## Summary of the Data

In this pivotal section, we distill the essence of our dataset, providing a comprehensive overview enriched with key statistical insights. Illustrated in Figure 1 are the distributions of crucial features, presenting visual insights that illuminate the data's nuances. Notably, the distribution of the imp_AccountAge feature exhibits a nearly uniform pattern, with the mean conspicuously standing out. This prominence is attributed to our imputation approach, wherein missing values were replaced with the mean of all values, underscoring a substantial presence of missing Account Age values in the dataset.

Features such as SubtitlesEnabled_Yes, Gender_Female, and ParentalControl_Yes are discrete, maintaining values of 0 or 1, revealing a relative balance between these values. In contrast, other categorical features showcase a higher prevalence of 0's compared to 1's, indicating distinct patterns within the dataset. This visual exploration lays the groundwork for a nuanced understanding of the dataset's intricacies, setting the stage for subsequent in-depth analyses.

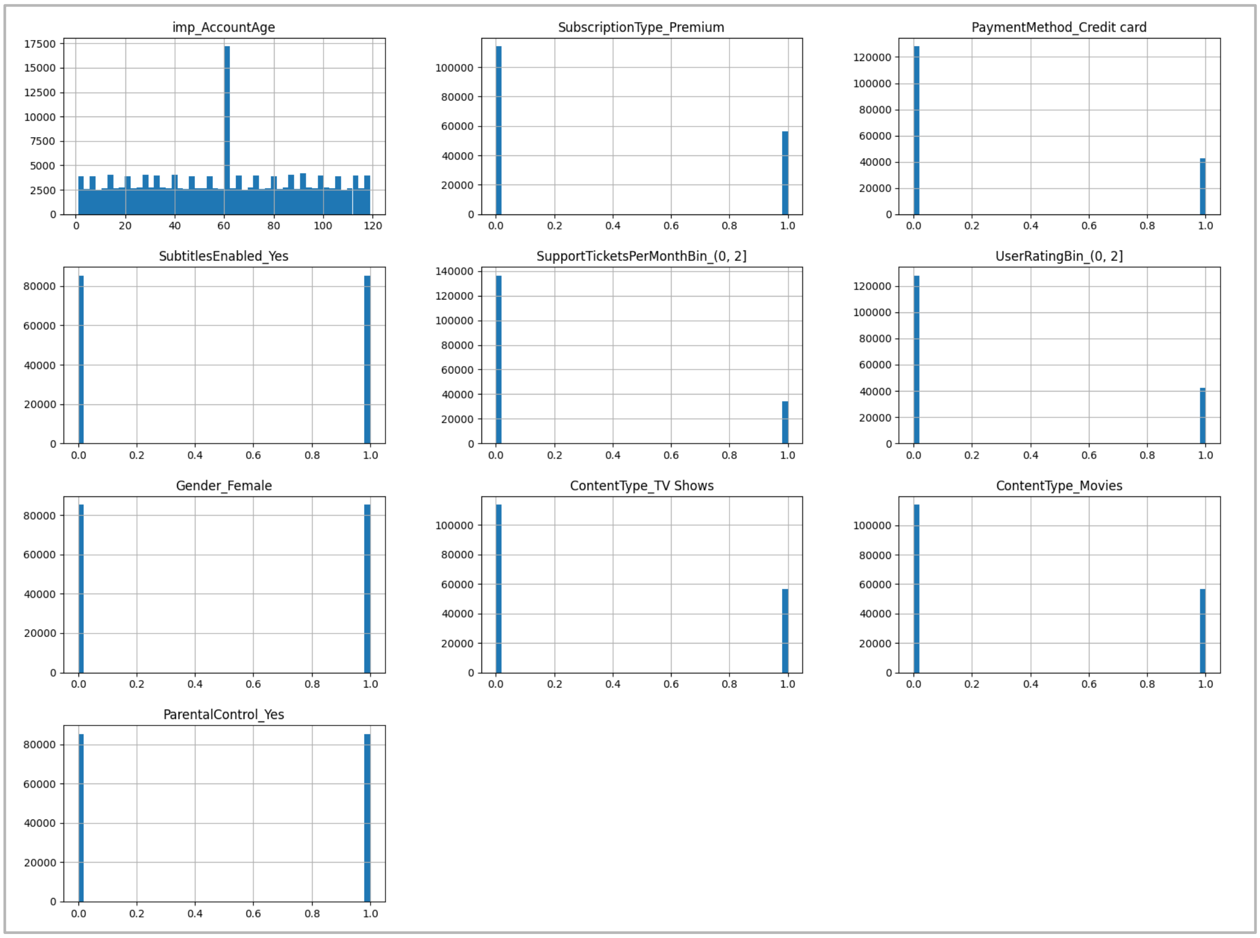

*Figure 1 —Distribution for each key feature in the dataset*

The next set of boxplot and histogram illustrations compares key feature values with the target Churn values to explore their relationships. For instance, Figure 2 below reveals the relationship between imp_AccountAge and Churn, illustrating a distinct distribution difference concerning the target values. For instance, when the Churn value is 0, the middle 50% of the Account Age values lie in the range of 48 to 90, with a median value around 60 years old. Conversely, when the Churn value is 1, the middle 50% shifts to the range of 20 to 65, with a median value around 40 years old.

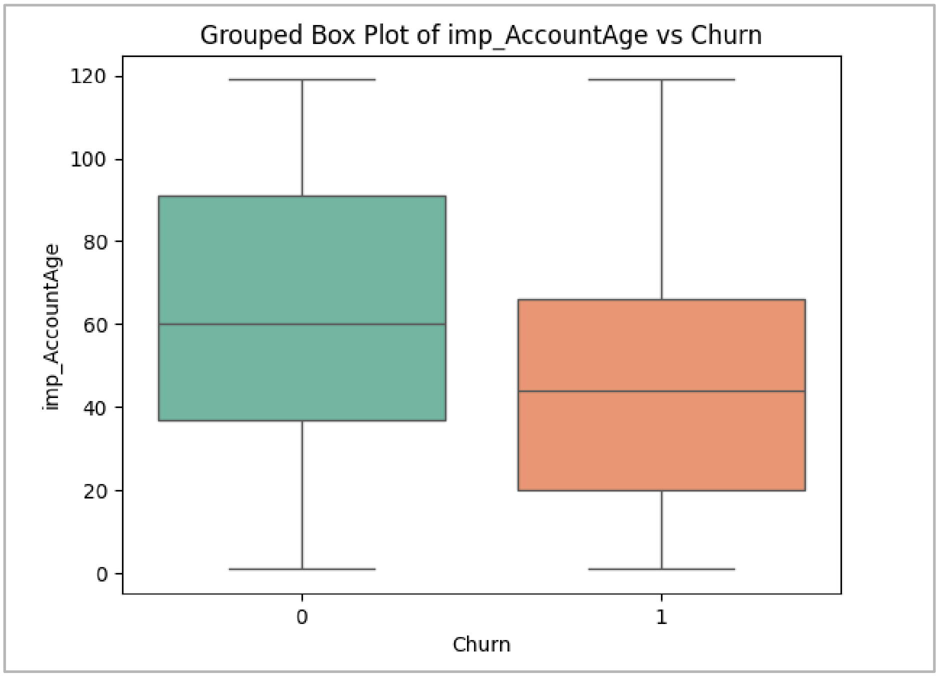

*Figure 2 — Group boxplot of imp_AccountAge vs Churn*

Expanding on this relationship, Figure 3 provides a histogram plot, reinforcing how a higher range of Account Age values tends to correlate with a Churn value of 0, while a lower range aligns with Churn 1. This alignment with real-world expectations is notable, suggesting that seniors and adults that are in the higher Account Ages, exhibit preferences that lean towards loyalty towards the company or the product.

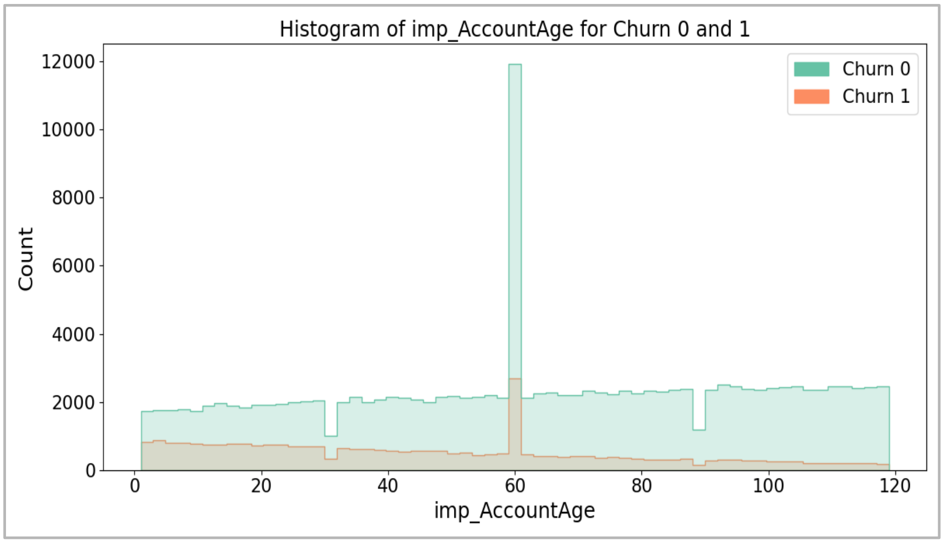

*Figure 3 — Histogram distribution of imp_AccountAge vs Churn*

The subsequent six bar plots in Figure 4 below meticulously compare the distribution of key features—PaymentMethod_Credit card, SubscriptionType_Premium, UserRatingBin_(0, 2], SupportTicketsPerMonthBin_(0, 2], ContentType_TV Shows, and ContentType_Movies—against the target Churn. Notably, a class imbalance surfaces, with higher bars for feature values 0 in both Churn cases. This imbalance prompts consideration of resampling or SMOTE techniques to mitigate potential impacts on logistic regression model performance. Additionally, the higher bars for feature value 0 might also indicate that these features have more importance or relevance in predicting the target variable and may be worth exploring feature importance techniques to quantify the impact of these features on the model.

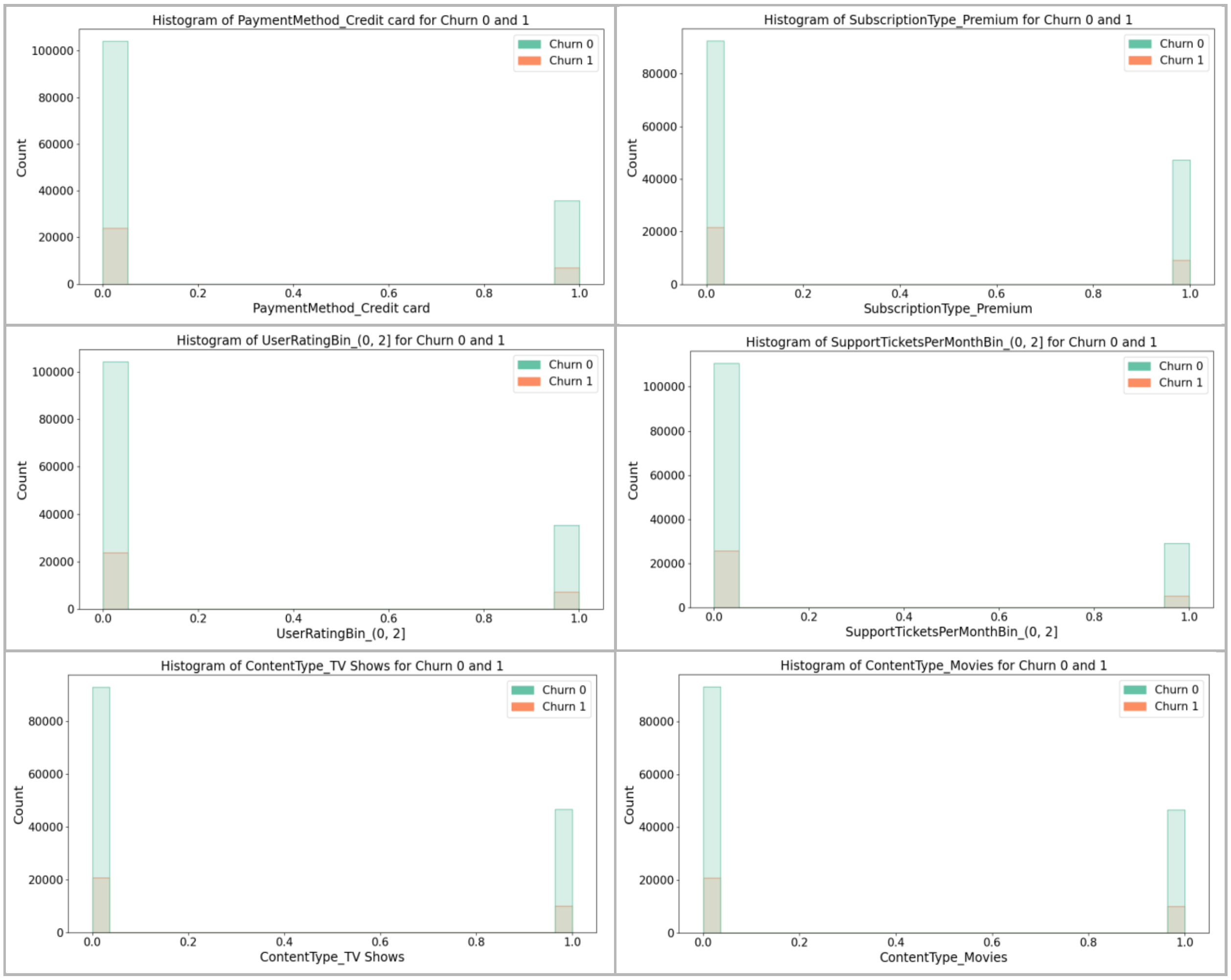

*Figure 4 — Histogram distribution of PaymentMethod_Credit card, SubscriptionType_Premium, UserRatingBin_(0, 2], SupportTicketsPerMonthBin_(0, 2], ContentType_TV Shows, ContentType_Movies vs Churn*

Concluding this exploration, the final three features in Figure 5 below present a more stable distribution. While the bars remain relatively equal in height for both feature values in both Churn cases, the lower height of bars when the Churn value is 1 suggests potential class imbalance and stronger discriminative power for the negative class. In other words, the features are more informative when predicting instances where the target is 0. Delving further into feature importance techniques could quantify these features' impact on the overall model.

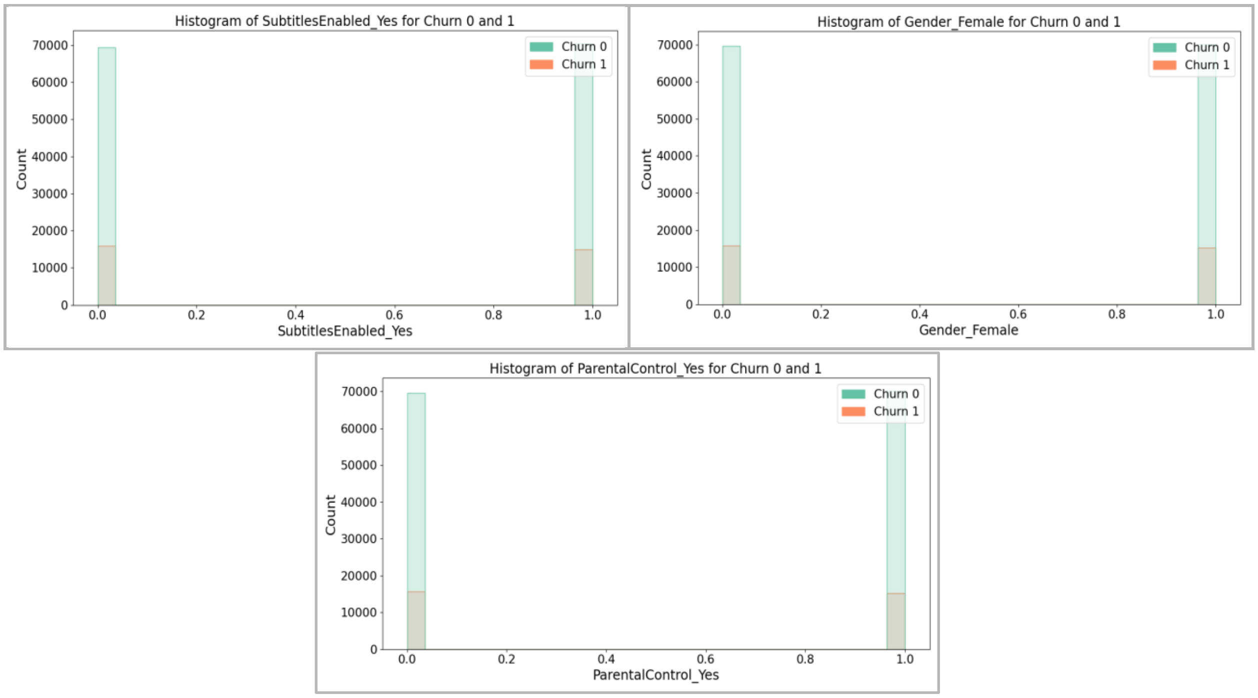

*Figure 5 — Histogram distribution of SubtitlesEnabled_Yes, Gender_Female, ParentalControl_Yes vs Churn*

This data-rich journey through our dataset aims to provide a robust foundation for subsequent analyses, ensuring a holistic understanding of the intricate interplay between our highlighted features and the target variable, Churn.

## Correlation

In this section, we turn our attention to the correlation dynamics between our target variable, Churn, and each key feature. Figure 6 below unveils a correlation heatmap, providing a visual representation of the relationships within the dataset. Notably, the correlation scores are close to 0, indicating a lack of linear correlation with the target variable.

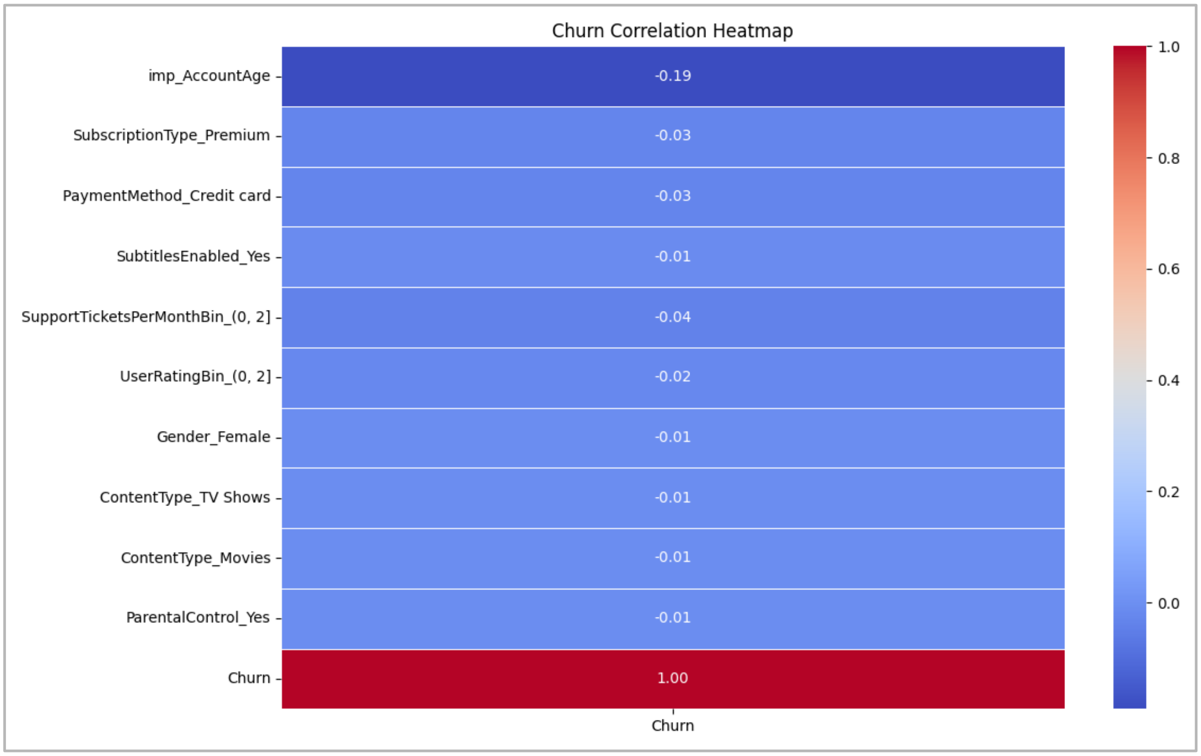

*Figure 6 —Correlation heatmap for each key feature*

It is crucial to recognize that, in the realm of logistic regression, the absence of a significant linear relationship is not necessarily a drawback. Logistic regression thrives in capturing complex, non-linear associations between features and the target variable, a nuance not fully reflected in traditional correlation metrics. Therefore, while the heatmap may depict minimal linear correlation, it doesn't diminish the model's capacity to capture intricate patterns that contribute to predicting customer churn.

This observation aligns with the inherent nature of logistic regression, where the emphasis lies on capturing the probability of an event rather than establishing linear relationships. Consequently, the absence of strong correlations in the heatmap doesn't undermine the model's efficacy. In subsequent sections, we will delve into the model's performance metrics and evaluate its ability to discern nuanced patterns, showcasing the resilience of logistic regression in handling the intricacies of predicting customer churn.

## Model Analysis Breakdown

In this pivotal section, we delve into a comprehensive analysis of the models crafted for predicting Churn, shedding light on model comparisons and the meticulous selection process.

## Model Comparisons

After delving into the intricate correlations between variables and our target variable, Churn, three distinctive logistic regression models were meticulously developed. Table 1 below offers an exhaustive statistical comparison of these models, employing five splits of cross-fold validation. Each model incorporates a unique feature selection technique, showcasing the top 10 best-selected features in descending order of significance. The table includes average accuracy, precision, recall, and F1 scores, along with their corresponding standard deviations.

| **Model** | **A** | **B** | **C** |
|---|---|---|---|
| **Feature Selection Type** | **Chi-Squared** | **Forward Feature Selection** | **RFE (Recursive Feature Elimination)** |
| **# of Features** | 10 | 10 | 10 |
| **Features Names** | TotalCharges, ContentDownloadsPerMonth, imp_AccountAge, imp_ViewingHoursPerWeek, imp_AverageViewingDuration, SubscriptionType_Premium, PaymentMethod_Credit card, GenrePreference_Action, UserRatingBin_(0, 2], SupportTicketsPerMonthBin_(0, 2] | imp_AccountAge, SubscriptionType_Premium, PaymentMethod_Credit card, SubtitlesEnabled_Yes, SupportTicketsPerMonthBin_(0, 2], UserRatingBin_(0, 2], Gender_Female, ContentType_TV Shows, ContentType_Movies, ParentalControl_Yes | PaymentMethod_Credit card, DeviceRegistered_Computer, DeviceRegistered_Mobile, DeviceRegistered_TV, DeviceRegistered_Tablet, GenrePreference_Action, GenrePreference_Comedy, GenrePreference_Drama, GenrePreference_Fantasy, GenrePreference_Sci-Fi |
| **Average Accuracy** | 0.7407465731362514 | 0.7458466053469811 | 0.8694642281951255 |
| **Std of Accuracy** | 0.001969591429086249 | 0.0007226152202316106 | 0.00196428401333571 |
| **Average Precision** | 0.6962169998454628 | 0.7077131623048167 | 1.0 |
| **Std of Precision** | 0.0029711471700655787 | 0.0022980608538012088 | 0.0 |
| **Average Recall** | 0.8542216468971716 | 0.8376393147306619 | 0.738931603006427 |
| **Std of Recall** | 0.0017630484475535714 | 0.0011036912620462607 | 0.0035843623933332081 |
| **Average F1-Score** | 0.7671645655831629 | 0.76721120507357 | 0.8498636269523617 |
| **Std of F1-Score** | 0.0020524946777347223 | 0.0011622791871395816 | 0.0023677503864745343 |

*Table 1 — Comparison table for three multiple logistic regression models*

### Model Evaluation

Among the three models, Model B, utilizing forward feature selection, emerges as the preferred choice for several compelling reasons. While Model C exhibits superior average accuracy, precision, and F1-score with impressive numerical values, a perfect precision score and a standard deviation of 0 in average precision raise concerns about potential overfitting, leading to the disregard of Model C.

Comparing Models A and B, both demonstrate relatively similar statistics, boasting high average recall scores crucial for identifying all positive cases. However, upon scrutinizing the standard deviation of the scores, Model B emerges as the most consistent performer among the three models. This consistency serves as a positive indicator of stability in performance.

In summary, Models A and B strike a balance between precision and recall without succumbing to overfitting. Nevertheless, Model B stands out as the preferred option due to its remarkable consistency and stability across various metrics. This strategic selection ensures a robust and reliable predictive model for anticipating customer churn.

## Conclusion

In concluding our analysis, let's unravel how Model B, our standout performer, navigates the realm of predicting customer churn. We'll delve into its strengths, acknowledging its accomplishments, while also identifying areas for potential enhancement in future iterations.

Model B's predictive prowess in anticipating customer churn is noteworthy, marked by a well-calibrated balance and consistent performance. With an impressive average accuracy of 74.58%, Model B strategically balances precision and recall, crucial metrics for identifying positive cases. Its superior consistency, reflected in minimal standard deviations across various metrics, positions it as a stable and reliable performer. Leveraging the top 10 features, the model's feature selection process empowers it to discern intricate patterns influencing customer churn.

The equilibrium achieved by Model B between precision and recall signifies its ability to identify positive cases without compromising overall accuracy. Additionally, the model showcases exceptional consistency, a critical aspect underscoring its performance stability.

While Model B stands out as a robust predictive tool, there is potential for refinement. Exploring additional relevant features or fine-tuning existing ones could enhance the model's understanding of nuanced patterns in customer behavior. Adjusting hyperparameters or exploring alternative algorithms might unlock additional predictive power. Furthermore, a commitment to continuously updating and enriching the dataset with fresh insights is pivotal for improving the model's adaptability to the evolving landscape of customer behaviors.

As we chart our course forward, recognizing the strengths and areas for improvement in Model B sets the stage for refining predictive capabilities. This ensures a resilient model that can effectively navigate the dynamic landscape of customer relationship management, offering enduring value to businesses seeking to optimize their customer retention strategies.