Member Cancellation Prediction

# Introduction

Enhancing Customer Retention through Data Science

In the ever-evolving landscape of customer-centric businesses, retention remains a cornerstone of sustainable growth. Our recent observations indicate a decline in retention rates, prompting an in-depth analysis and strategic response. This work aims to harness the power of data science, deep learning, and explainable AI to not only understand the nuances of customer churn but also to provide actionable insights and solutions that seamlessly integrate with business operations.

Our approach is structured into four distinct phases:

1. **Problem Understanding:**

* This foundational phase is dedicated to comprehensively understanding the business challenge. By delineating the problem, we set a clear direction for our data-driven solutions.
* Prior to diving into the data, we engage in a brainstorming session to hypothesize potential factors influencing customer decisions, ensuring a holistic approach to the subsequent analysis.

1. **Exploratory Data Analysis (EDA):**

* Here, we delve into the data, understanding its structure and inherent patterns.
* The insights derived from this phase not only provide a deeper understanding of customer behavior but also offer actionable recommendations for the business team.

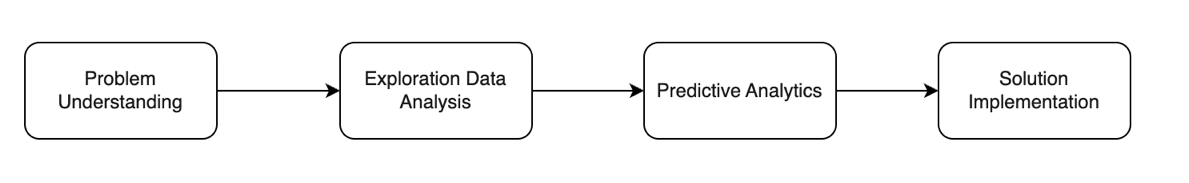
1. **Predictive Analytics:**

* Our journey begins with feature engineering, drawing from insights in the EDA phase to construct meaningful predictors.
* We then transition into model building, selecting algorithms tailored to predict potential churn. The models undergo rigorous offline evaluations, with performance enhancements achieved through hyperparameter tuning.
* A key highlight of this phase is error analysis, allowing us to pinpoint specific scenarios where our model excels or requires further refinement.

1. **Proposed solutions implementation:**

* Apart from the findings we present, suggesting new integrated functions for the business. This ensures that our data-driven solutions not only provide insights but also facilitate more effective and streamlined business processes.

In the subsequent sections, we'll delve deeper into each phase, unraveling the methodologies, findings, and recommendations. We hope this exploration offers valuable insights and paves the way for enhanced customer retention.



[Introduction 1](#_Toc1802509817)

[Phase 1: Problem Understanding 2](#_Toc1867201432)

[Problem Definition 3](#_Toc1407296687)

[Hypothesis Generation 3](#_Toc1757506129)

[Phase 2: Exploration Data Analysis 6](#_Toc1174944352)

[Data Collection and Understanding 7](#_Toc803934375)

[Hypothesis Validation 10](#_Toc466560938)

[Phase 3: Predictive Analytics 17](#_Toc1445884020)

[Feature Engineering 17](#_Toc1634570673)

[Model Building 17](#_Toc289123522)

[Model Evaluation 18](#_Toc451850253)

[Hyperparameter tuning 19](#_Toc892329211)

[Error analysis 20](#_Toc981622658)

[Phase 4: Solution Implementation 22](#_Toc936730894)

[1. Simple integration 22](#_Toc2060098343)

[2. Deploy solution with explainer 22](#_Toc928231036)

[3. Chatbot assistance 23](#_Toc1977719196)

[Discussion and Further Improvement 25](#_Toc401785280)

[Conclusion 26](#_Toc1838699211)

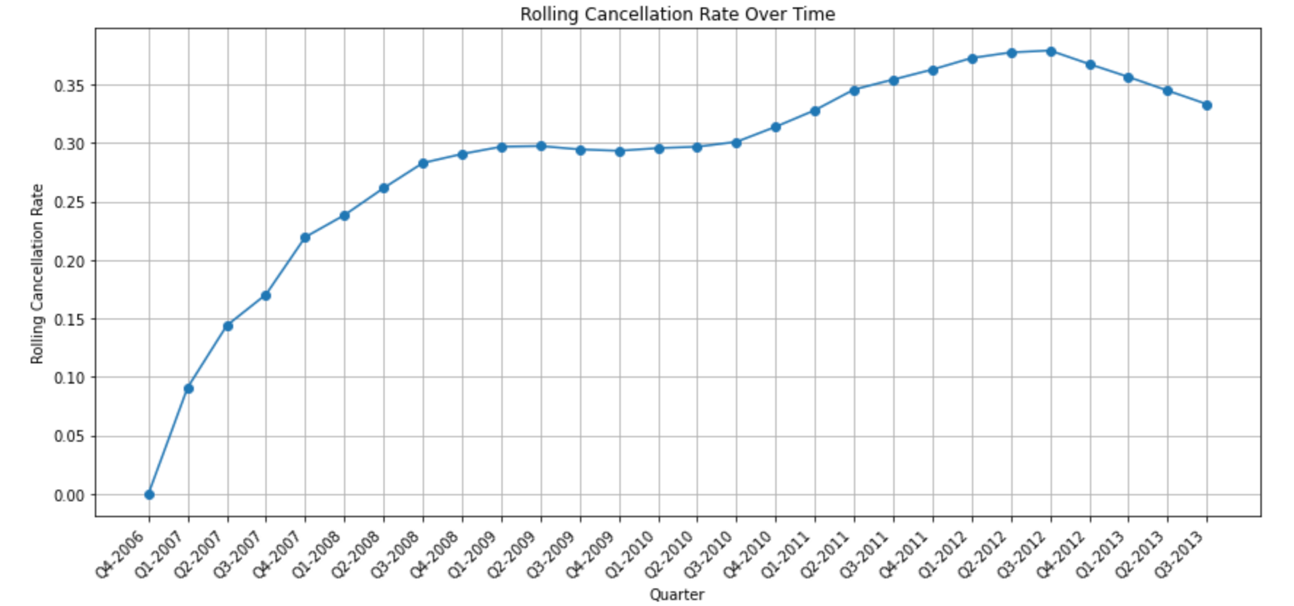
# Phase 1: Problem Understanding

## Problem Definition

The objective of this step is to ascertain whether there's a genuine issue at hand. Sometimes, perceptions can be misleading, so it's crucial to validate them with data. Moreover, it's essential to identify and agree upon the key metrics that will indicate success

Our first task is to determine if there's been a recent surge in the cancellation rate.

The chart presented below illustrates the cancellation rate trend over successive quarters. From this, we can deduce a noticeable uptick in cancellations. For instance, the cancellation rate in Q3-2013 stood at 33%, which is more than three times the rate in Q1-2017, which was 9%. Despite a minor dip recently, the overarching trend indicates a significant rise.



In this context, our primary success metric is straightforward: the "**churn rate.**" Our goal is to minimize this rate. Overlooking this metric can sometimes lead to ambiguity about the project's direction or uncertainty about the tangible impact of our data science efforts. Establishing this metric upfront ensures we have a clear benchmark for evaluation, especially during A/B testing phases when deploying our data solutions.

Given the problem, our data science approach will focus on **developing a predictive model to identify potential churn candidates.**

We'll employ **binary classification techniques** for this purpose. Potential models include tree-based algorithms like Random Forest, KNN, XGBoost, and LightGBM. Alternatively, advanced methods such as deep learning can be applied to the dataset. The specific methodology adopted will be elaborated upon in the subsequent sections

## Hypothesis Generation

In this phase, we meticulously evaluate all potential factors that might contribute to user churn. Undertaking this process ensures that we don't overlook any crucial elements. By compiling a list of inquiries, we set a focused path for the subsequent data exploration stage. The outcomes of our brainstorming are detailed below

|  |  |  |  |
| --- | --- | --- | --- |
| Aspect | Feature | Hypothesis | Data available checking |
| Personal Details of the Member | Age | * Younger members might have different priorities or financial constraints. * Older members might not find the amenities or services relevant | **Available** |
| Gender | * Services might appeal more to one gender than the other. | **Available** |
| Marital status | * Single members might have different needs than married ones or those with families. | **Available** |
| Annual Income | * Members with lower incomes might find the fees burdensome. * Members with higher incomes might seek more exclusive or premium services | **Available** |
| Occupation | * Certain professions might have time constraints or work-related pressures affecting membership usage. | **Available** |
| Membership Details | Membership Package | * Some packages might not provide enough value for their cost. * Lack of clarity or hidden terms in packages. | **Available** |
| Duration of Membership | * Short-term members might not see long-term benefits. * Long-term members might feel a need for change or find better alternatives. | **Available** |
| Annual Fees | * Perception of not getting value for money. * Sudden increases in fees without noticeable improvements in services. | **Available** |
| Payment Mode | * Inflexible payment options. * High charges or fees for certain payment modes. | **Available** |
| Services and Amenities | Quality of Services | * Decline in service quality or standards. * Inconsistent service delivery. | Not Available |
| Range of Amenities | * Lack of diverse facilities or activities. * Outdated or poorly maintained facilities. | Not Available |
| Staff and Customer Service | * Unfriendly or unhelpful staff. * Slow response to queries or complaints. | Not Available |
| Lack of Feedback Mechanisms | * Members might feel their concerns are not heard. | Not Available |
| Ineffective Communication | * Not informing members about changes, renovations, or new services. * Overloading members with non-relevant communications. | Not Available |
| External Factors | Competition | * Emergence of new clubs with better facilities or lower prices. * Aggressive marketing or promotions by competitors. | Not Available |
| Economic Factors | * Economic downturn leading to reduced discretionary spending. * Job losses or financial difficulties among members. | Not Available |
| Seasonal Factors | * Some seasons might see reduced activities or engagements. | Not Available |
| Miscellaneous | Agent Influence | * Higher agent performance, better retention | **Available** |
| Contract and Terms | * Rigid contract terms or penalties for early termination. * Hidden clauses or charges that members were unaware of. | Not Available |

After considering various factors and examining the data at hand, we've curated the following hypotheses to investigate in the subsequent phase:

1. Do younger members tend to cancel the service more frequently than older ones?
2. Is the service more favored by men than women? Moreover, do single members tend to remain longer than married ones?
3. Are members with a higher income less likely to cancel our service?
4. Are certain professions more prone to canceling?
5. Do members who bring along additional persons exhibit a higher retention rate?
6. Does the cancellation rate vary across different membership packages?
7. Are members subscribed to longer-term services less likely to churn?
8. Is there a higher cancellation tendency among members with lower annual fees compared to those with higher fees?
9. Do members who opt for annual payment methods show lower churn rates compared to others?
10. Does better agent performance correlate with improved retention?
11. Are newer members less prone to churning than long-standing ones?

# Phase 2: Exploration Data Analysis

## Data Collection and Understanding

In this phase, we delved into understanding the underlying structure and relationships between the dataset's features. A thorough comprehension of the data ensures a more effective analysis. If you're already acquainted with this dataset, you might want to bypass this section and proceed directly to the hypothesis validation segment to view our findings.

1. **Overview of cancelled rate**

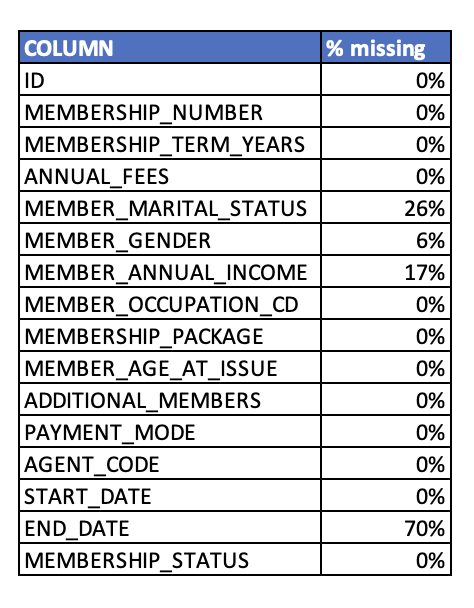
Initially, we aim to evaluate the label rate of this dataset to establish a benchmark for our analysis. The dataset indicates a cancellation rate of approximately 30%. Therefore, in the subsequent findings, any group or segment with a rate exceeding 30% is deemed to have a heightened likelihood of churning."



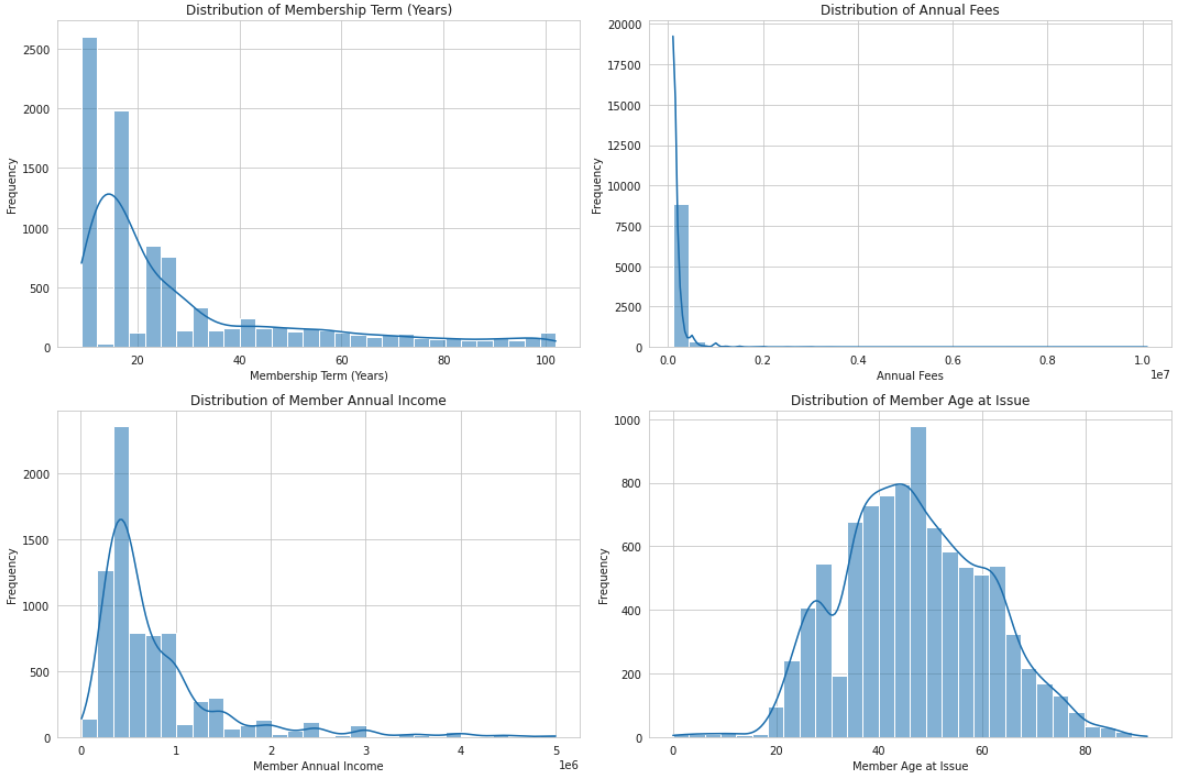
1. **Missing values in the data**

Most of features are not missing, but there 3 features with missing values:

MEMBER\_MARITAL\_STATUS, MEMBER\_GENDER and MEMBER\_ANNUAL\_INCOME

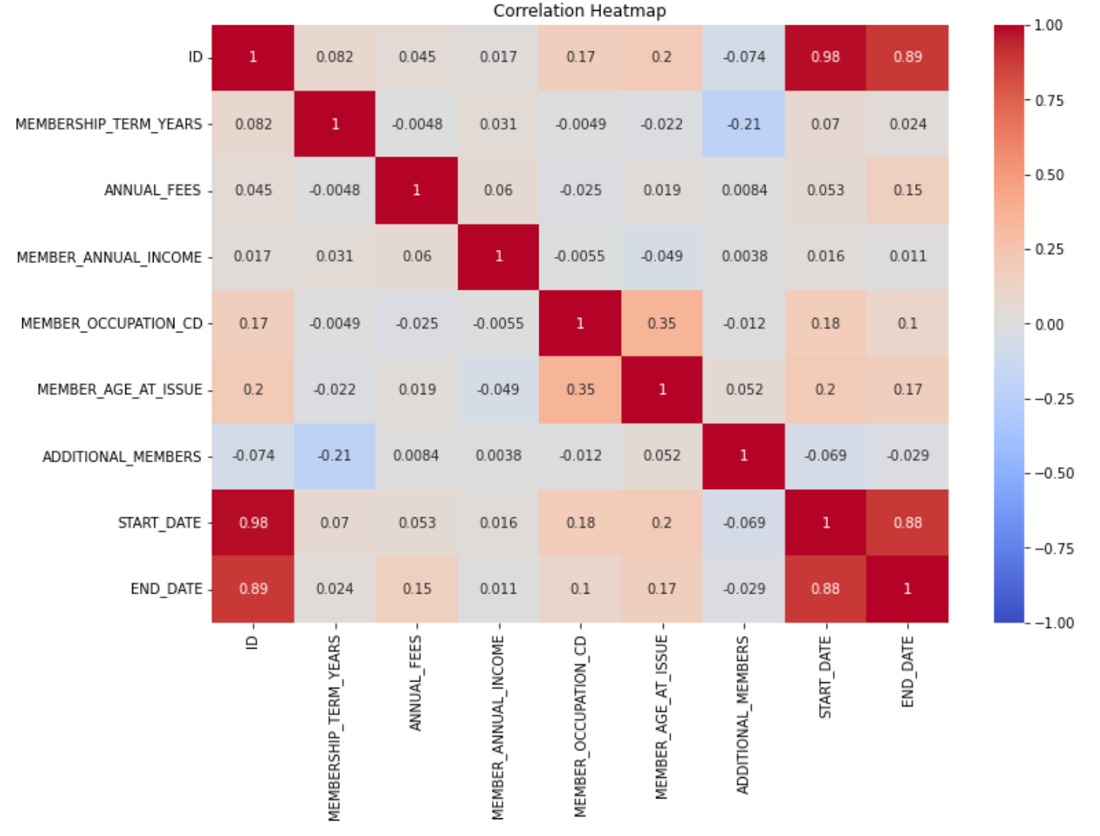


1. **Distribution of continuous features**



* MEMBERSHIP\_TERM\_YEARS: A large number of memberships have terms around 10 years, with a smaller number having terms around 40 years. Other durations are less common.
* ANNUAL\_FEES: The majority of members pay annual fees in the range of 100,000 to 200,000. Few memberships have much higher fees.
* MEMBER\_ANNUAL\_INCOME: Most members have an annual income below 1,000,000. We've filtered out extreme values for better visualization.
* MEMBER\_AGE\_AT\_ISSUE: The distribution indicates that the age at which members get their memberships is widely spread, with peaks around the mid-30s, mid-40s, and late 50s.

1. **Correlation between continuous independent features**



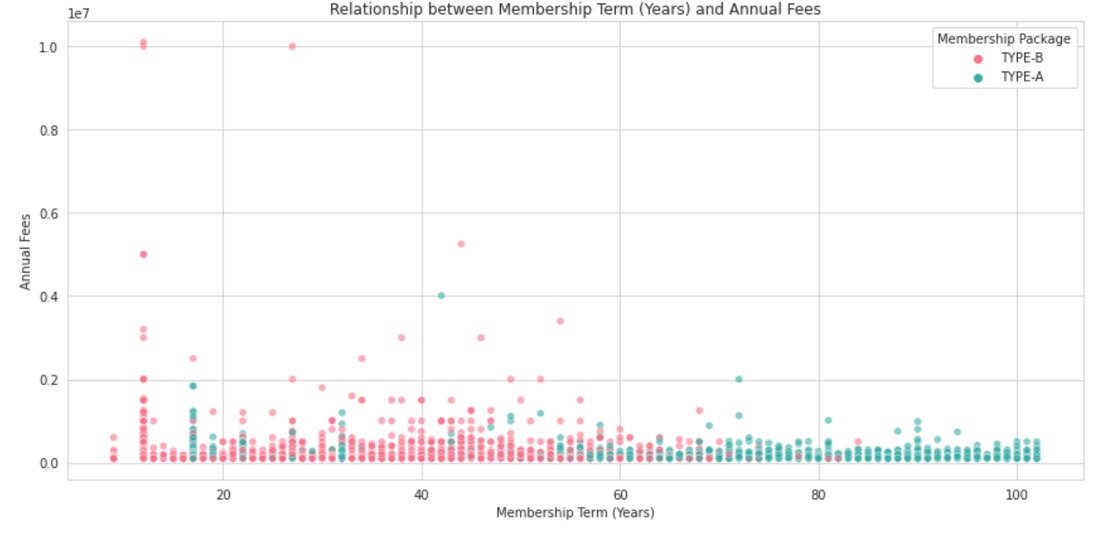
* Most of the variables don't show strong correlations with each other
* The correlations are relatively low, indicating that these variables are largely independent of each other.

1. **Distribution of categorical features.**



* Most of our members are Male
* Our members are using TYPE-B package than the TYPE-A

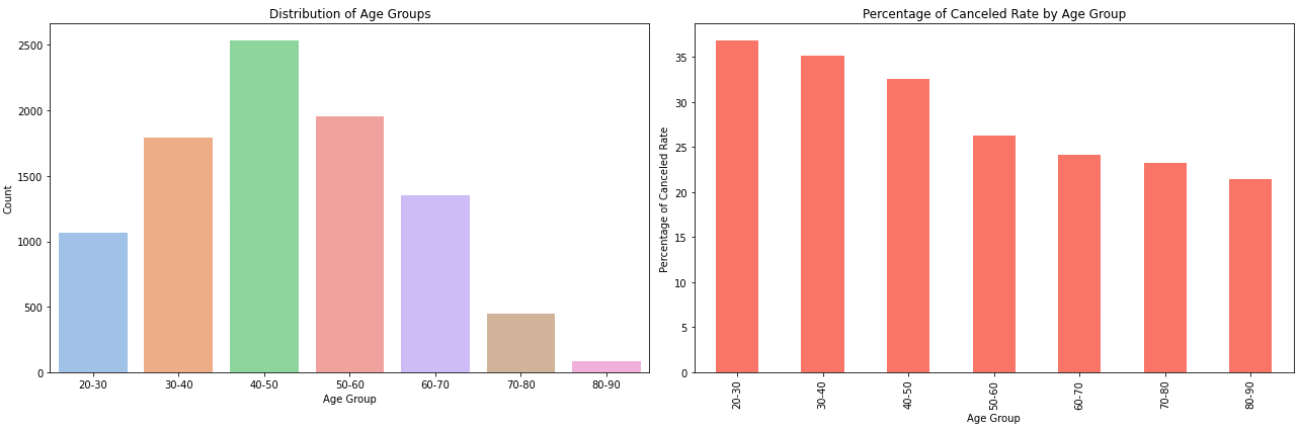
1. **Relationship between membership term and annual fees**



* Package type B are longer in term than TYPE A
* There is no difference in annual fees of the 2 types of packages

## Hypothesis Validation

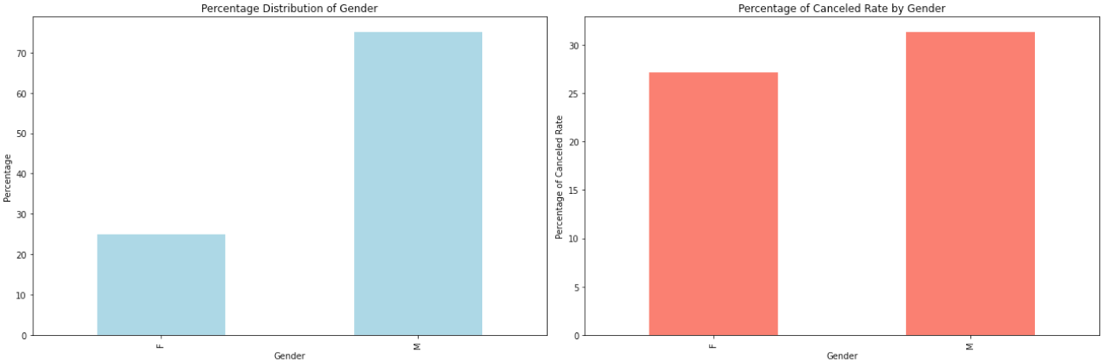
1. **Do younger members tend to cancel the service more frequently than older ones?**



Key takeaways:

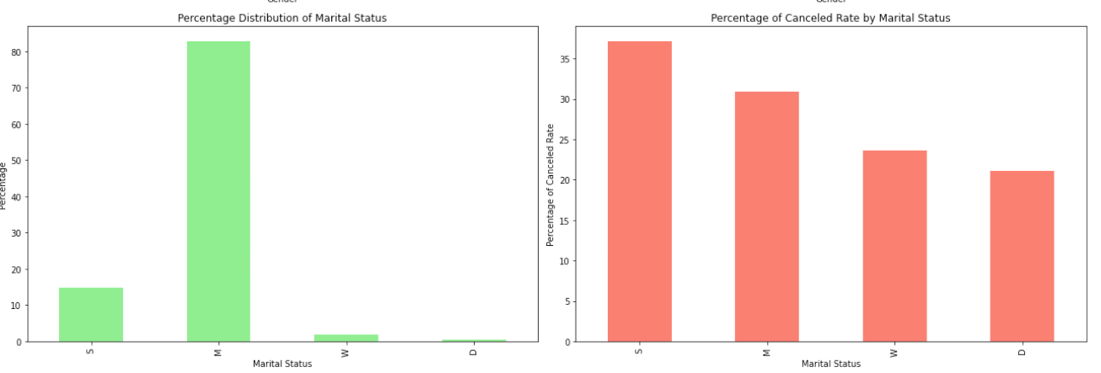
* The majority of members, at the time of joining, fall within the age range of 30 to 60 years.
* A noticeable trend is that younger members exhibit a higher inclination towards canceling our service

1. **Is the service more favored by men than women? Moreover, do single members tend to remain longer than married ones?**



Key takeaways:

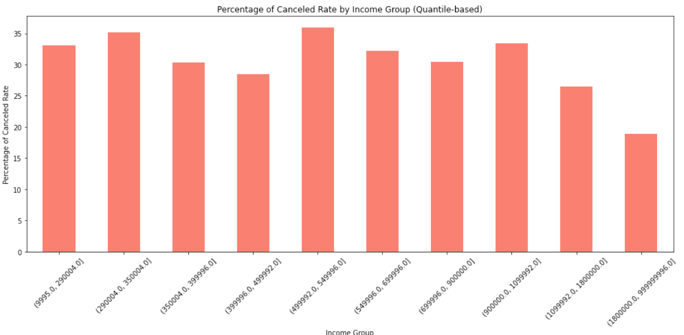
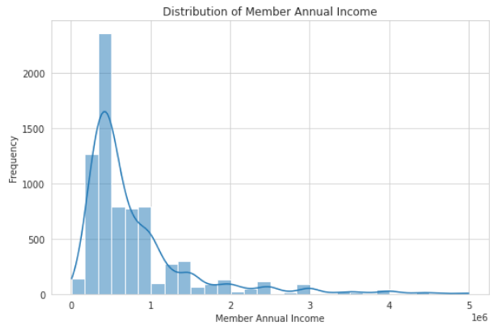
* Most of our members are man
* There is no significant different between the gender groups regarding cancellation rate



Key takeaways:

* Interestingly, there is a higher cancelled rate for the group single and married and lower for widow and divorced.
* But widows and divorces accounted for a small proportion. Most of our members are married or single

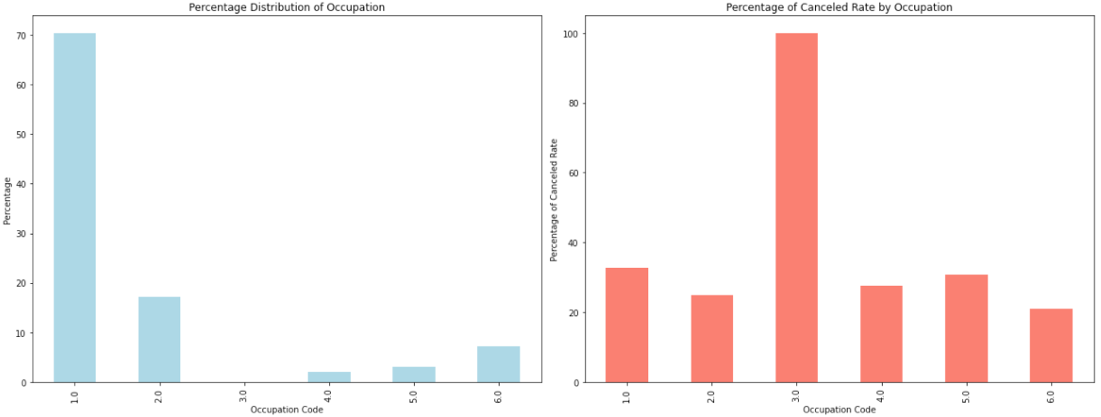
1. Are members with a higher income less likely to cancel our service?



Key takeaways:

* Most of our members have annual income below 1 million
* And there is just a slight difference for the groups under 1 million in income. And for the group above 1 million, they tend to retain better

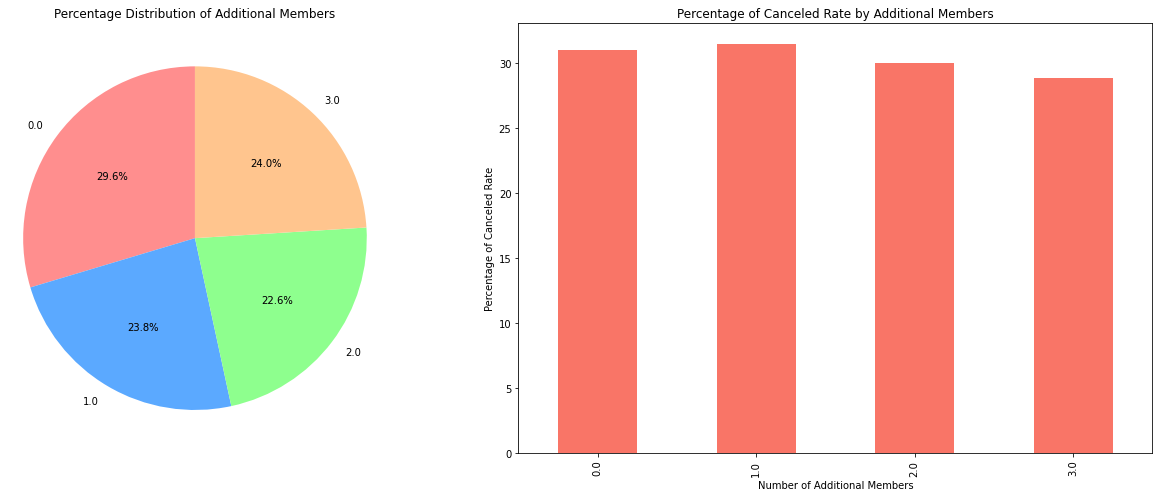
1. Are certain professions more prone to canceling?



Key takeaways:

* Most of our members have annual income below 1 million
* And there is just a slight difference for the groups under 1 million in income. And for the group above 1 million, they tend to retain better

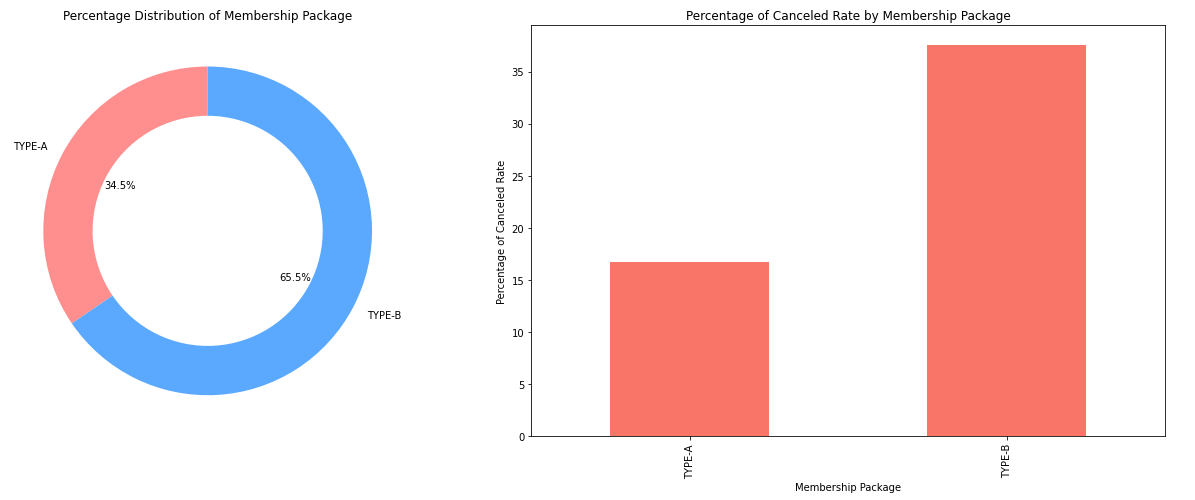
1. Do members who bring along additional persons exhibit a higher retention rate?



Key takeaways:

* There is no difference between for the group bring along additional persons as we assume
* There is evenly distribution between group bring additional persons

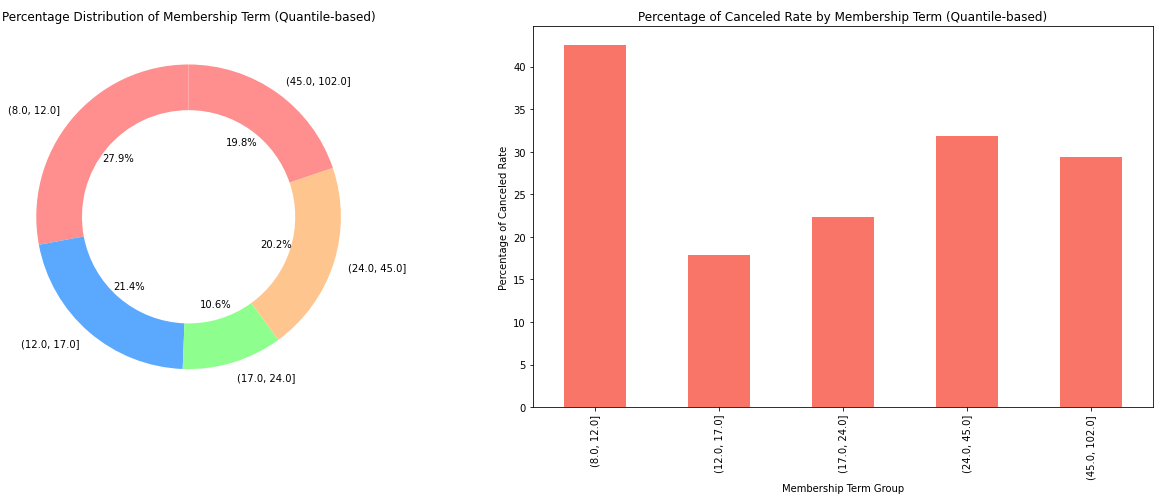
1. Does the cancellation rate vary across different membership packages?



Key takeaways:

* In response to the hypothesis, the data indeed verifies that members with the TYPE\_B package have a noticeably higher churn rate than those with TYPE\_A. The cancellation rate for TYPE\_B stands out as significantly greater than that of TYPE\_A.
* Notably, the TYPE\_B package is more popular, with the majority of our members opting for it.

1. **Are members subscribed to longer-term services less likely to churn?**

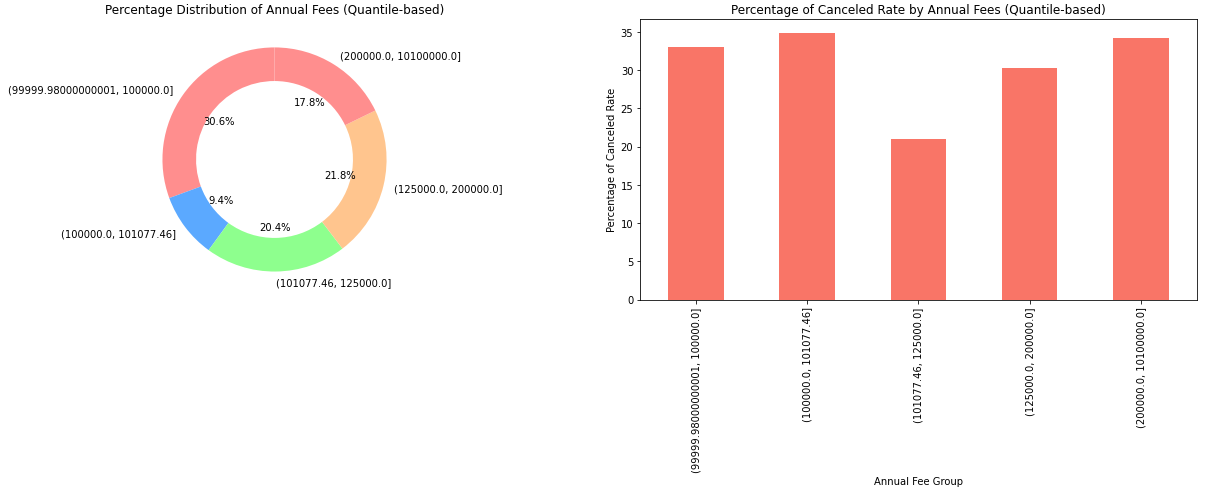


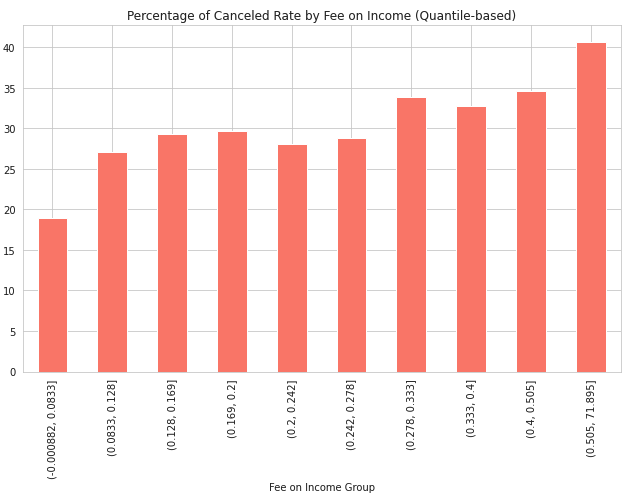
Key takeaways:

* Members who opt for shorter subscription terms exhibit a heightened propensity to cancel. This pattern suggests that those who commit for brief periods may either be testing the service or are less invested, leading to a greater likelihood of discontinuation
* Our membership displays an even distribution across various subscription durations

1. **Is there a higher cancellation tendency among members with lower annual fees compared to those with higher fees?**

There is a good retention group with an annual fee of 100 –125K. We want to know what is the best threshold to set the pricing based on their income. We can see that the fee on income ratio should be around under 10% or 13% for a better retention rate.

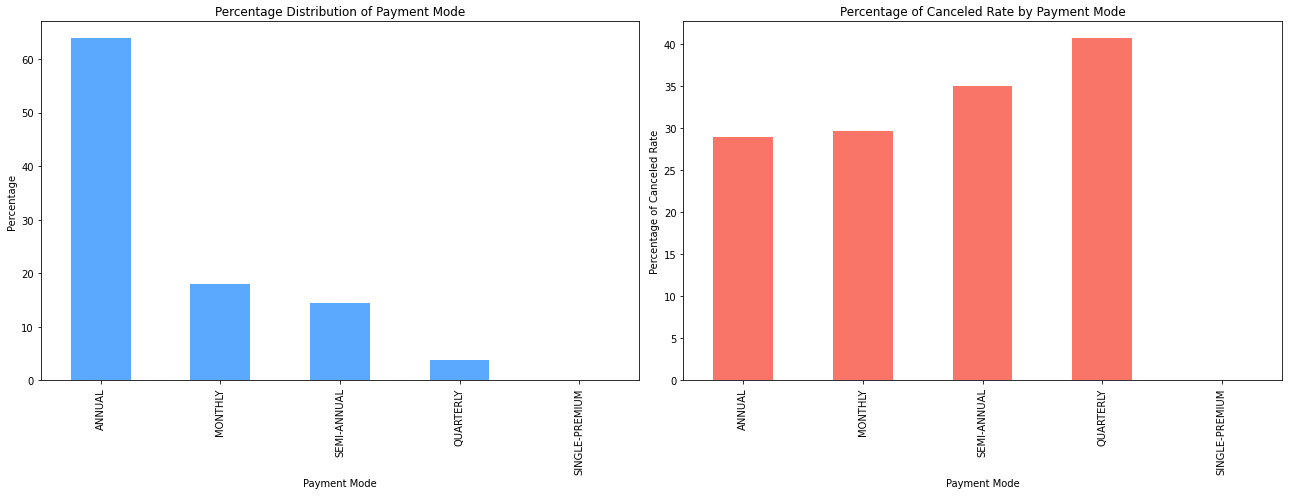




Key takeaways:

* The best group of annual fees is from 100 – 120K
* To enhance retention, it's advisable to set the annual fee at approximately 10% to 12% of a member's income.

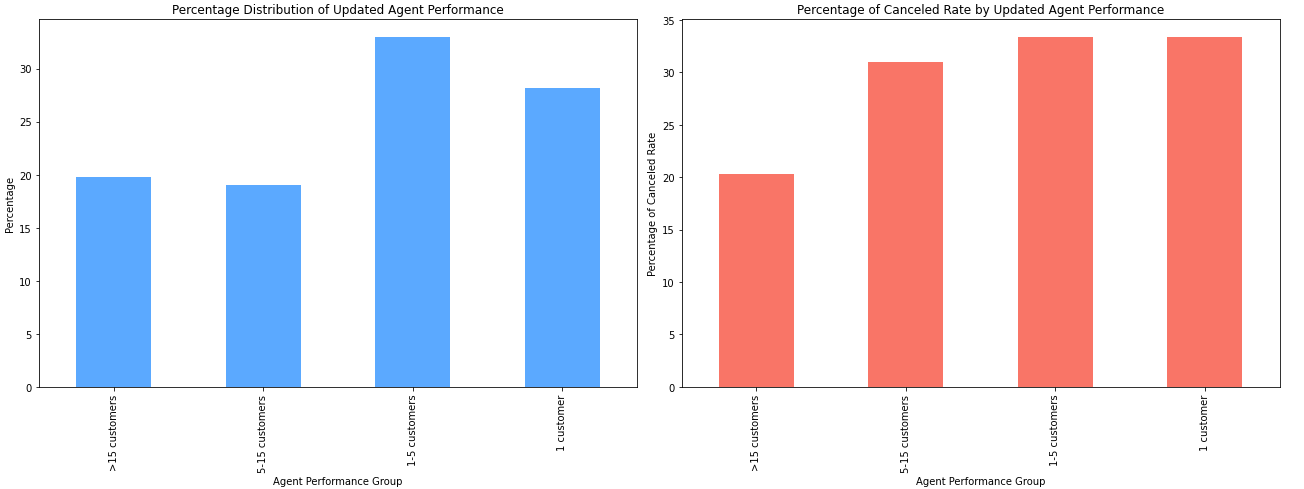
1. **Do members who opt for annual payment methods show lower churn rates compared to others?**



Key takeaways:

* Most of the members are paid annually, monthly or semi-annually.
* The group paid annually or monthly is bring the best in retention rate (with low cancelled rate)

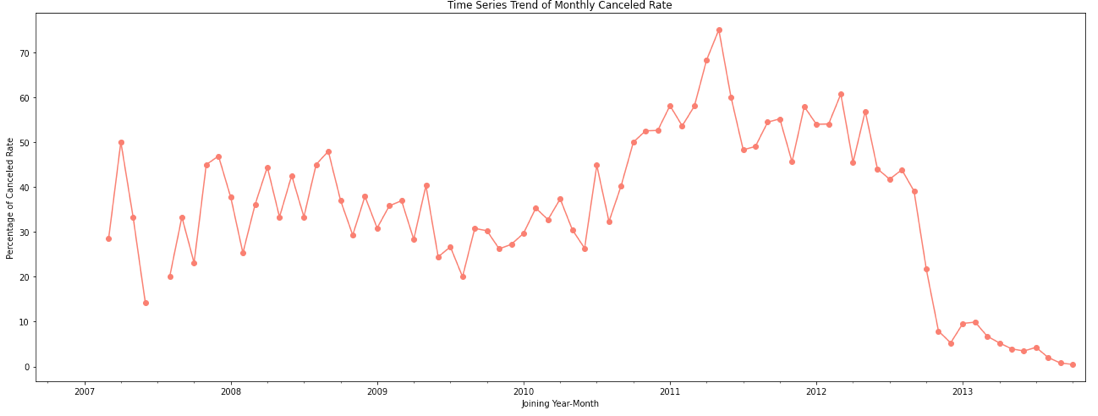
1. **Does better agent performance correlate with improved retention?**

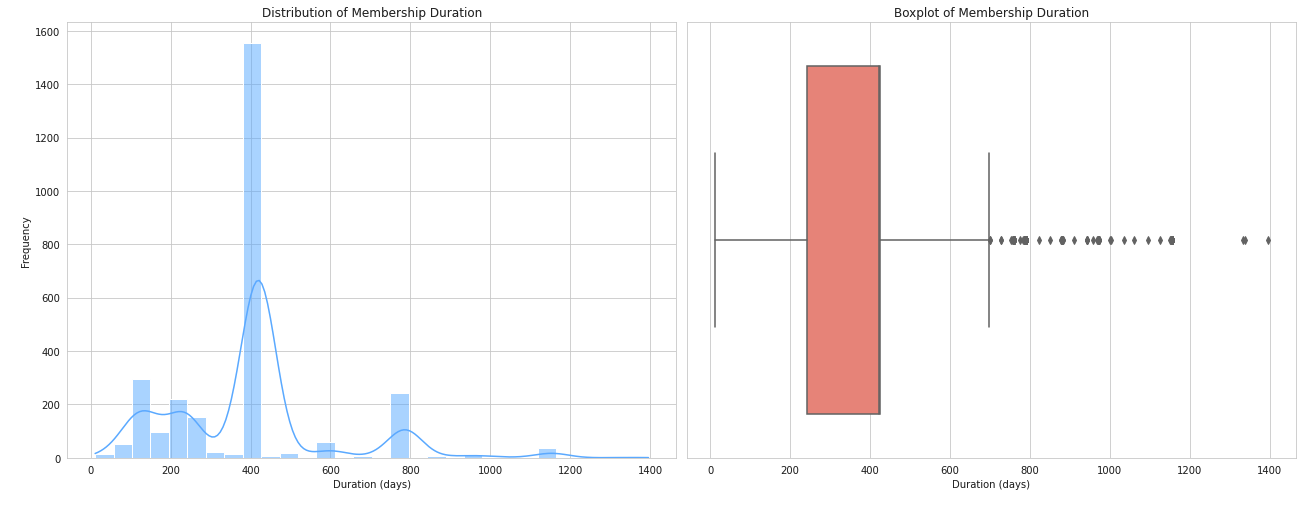


Key takeaways:

* Approximately 20% of our agents are classified as 'high-performing', having successfully onboarded more than 15 members each.
* Members introduced by these top-tier agents demonstrate a notably reduced rate of cancellation, underscoring the impact of agent performance on member retention

1. **Are newer members less prone to churning than long-standing ones?**

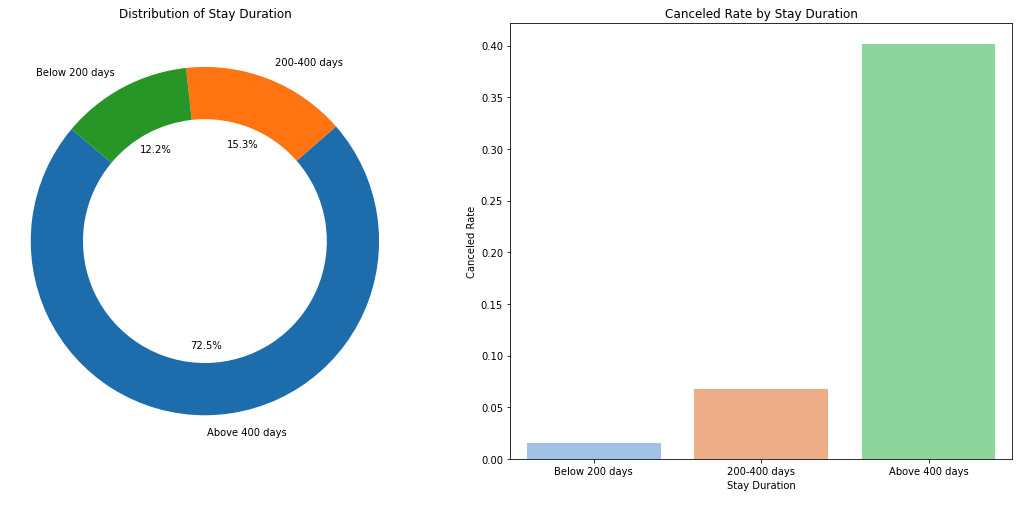




Key takeaways:

* We observe a pronounced spike in the cancellation rate for members who joined us between 2010 and 2012.
* Recent members exhibit a diminished churn rate. This can likely be attributed to their novelty; they've only recently joined, and typically, members take some time before deciding to churn. As evident in the subsequent chart, the majority of our members tend to cancel after approximately 400 days (around a year) of availing our service.

To make this trend clearer, we group our members into 3 groups based on their duration from joining date to the current date (suppose it was 20140101). And we can see that close to 30% of our members are relatively new, having been with us for less than 400 days. This newer cohort exhibits a markedly lower cancellation rate when compared to the more tenured group that has been with us for more than 400 days



# Phase 3: Predictive Analytics

## Feature Engineering

**Basic featuring:** As the basic steps we do some feature engineering such as filling missing values on the existing set of features (not yet adding new derived features) and one-hot encoding for categorical features.

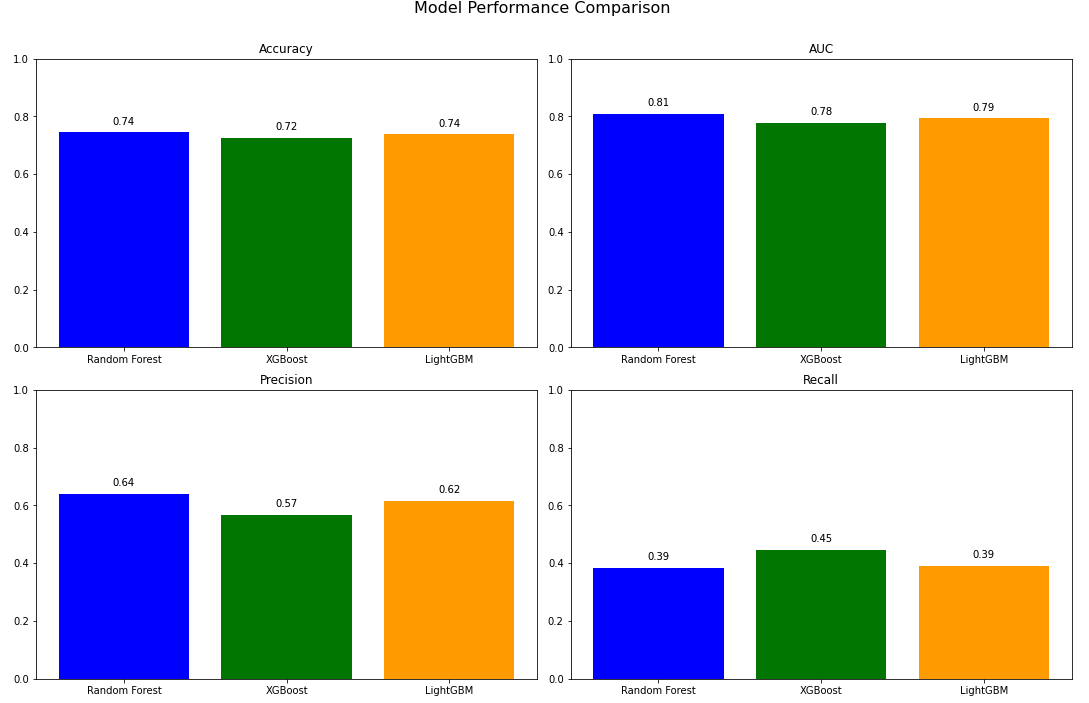
**Generating new features:** based on the analysis result, we come up with some new features such as

* Number of from their joining date to the current date (20140101)
* Fee to income ratio
* Agent performance
* Start year/month/day
* Combine start year and age

**Feature embedding:** By applying deep learning to learn the underlying patterns and then extract the embedded vectors for those categorical features. And instead of using traditional encoding methods such as one-hot encoding or label encoding, we use the embedding vectors.

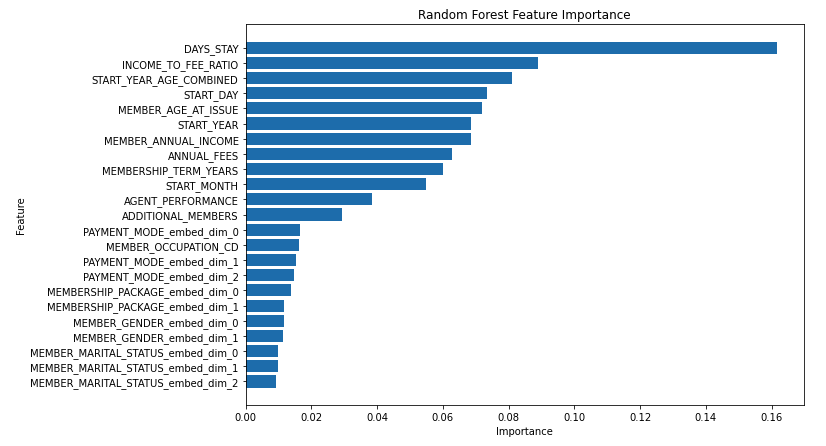
## Model Building

With the same set of features, we have built multiple models and compare their performance on various metrics. And from the result as the chart below, Random Forest is a good choice to go

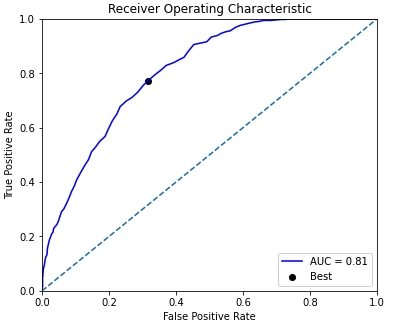


## Model Evaluation

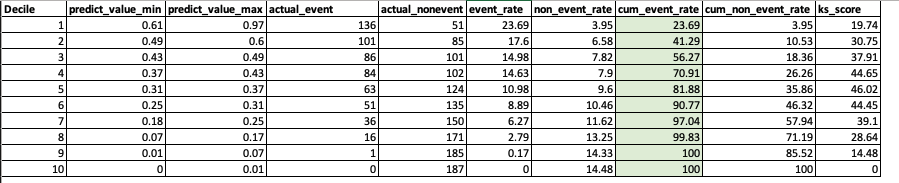
Firstly, we assess the feature importance. And then check the AUC of the model and the top feature importance are number of days from issue date to the current date (20140101), income to fee ratio and so on



And our model performs quite well with 0.81 AUC which is considered a good threshold.



By ranking our model from highest to lowest cancellation score, we see that if we select to target top 40%, we might cover 71% of the cancellation cases.



## Hyperparameter tuning

We use the grid search to search the hyperparameters in the range as below. Unfortunately, the performance has not improved much after this step.

param\_grid = {

'n\_estimators': [10, 50, 100, 200, 30],

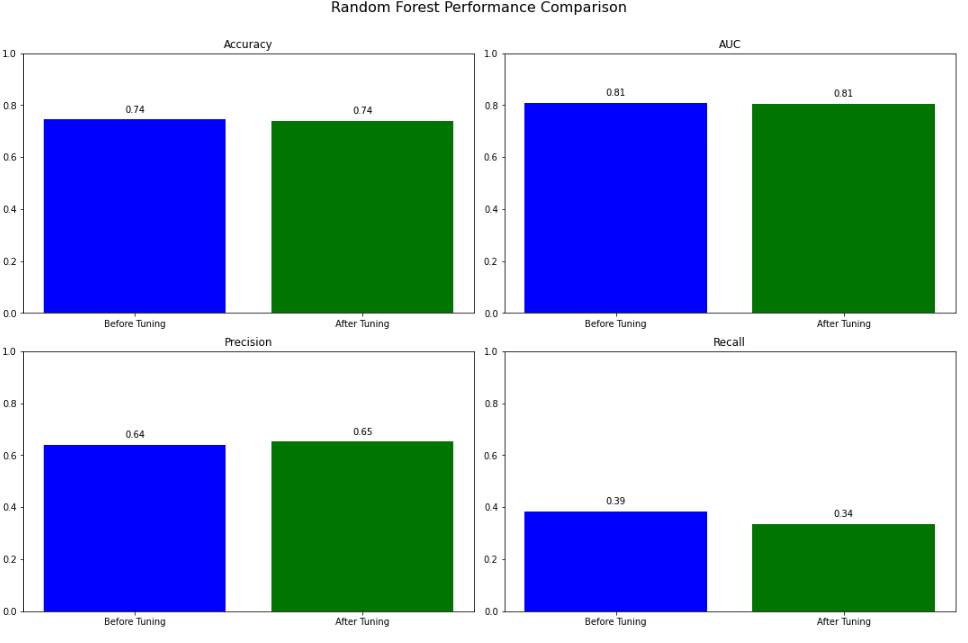
'max\_depth': [None, 10, 20, 30],

'min\_samples\_split': [2, 5, 10],

'min\_samples\_leaf': [1, 2, 4],

'bootstrap': [True, False]

}



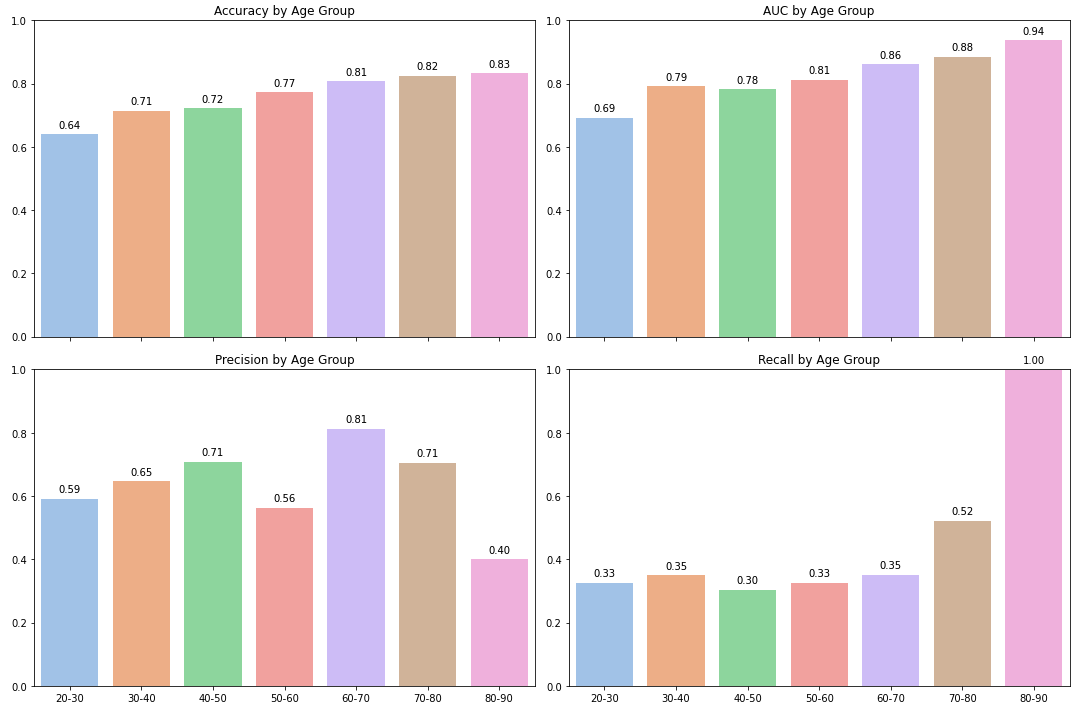
## Error analysis

Error analysis involves examining the circumstances, reasons, and moments when models falter. It encompasses the steps of identifying, scrutinizing, and interpreting incorrect ML predictions, providing insights into areas where the model excels or falls short.

Error cohort analysis is one of the most powerful steps within error analysis in ML. With this method, we segment our users into different sets and see if our model performs consistently in the groups.

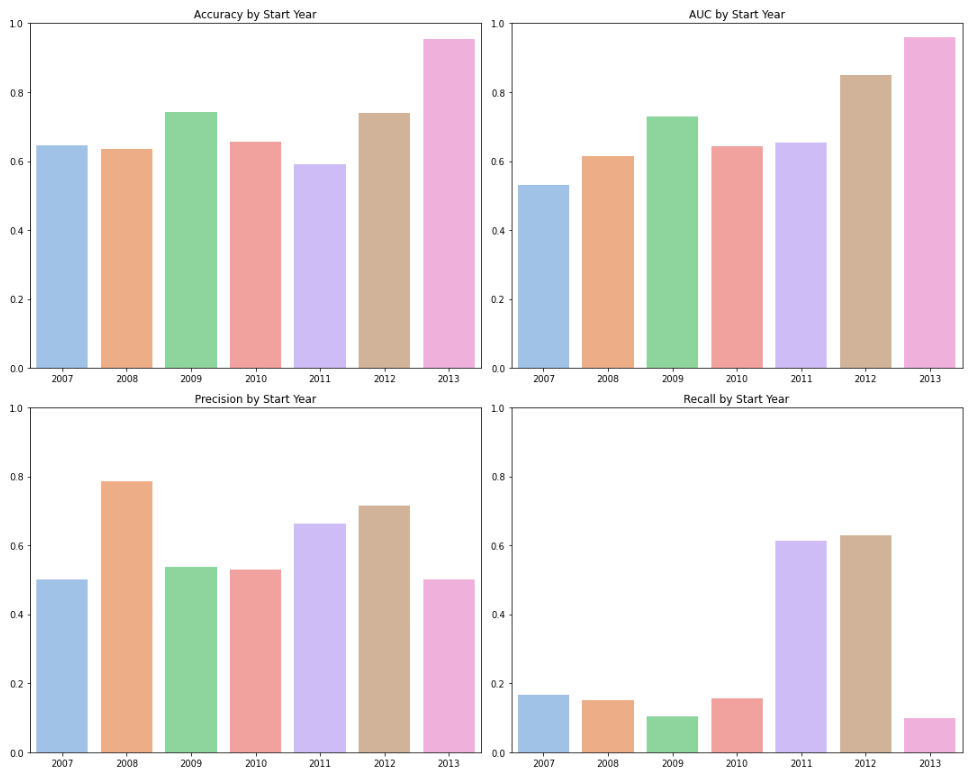
1. Segment by age group

With this segment, we can see that our model performs better in the higher age group.



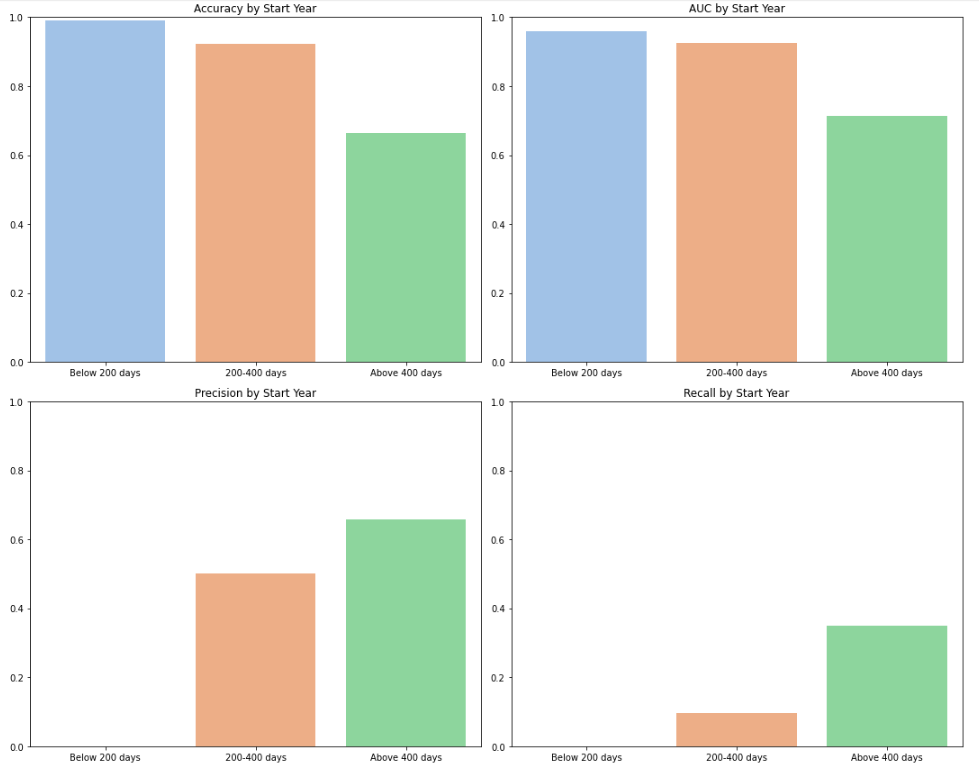
1. Segment by start year

With this view, we can see that our model performs differently in different groups. The better performance for the group of users has joined recently.



1. Segment by member tenure (duration from joining date to current)

Our model did really well in the group under 400 days (about 1 year) in tenure. But regarding precision and recall for the groups, there should be further analysis to understand why they are nearly zero as current performance.



# Phase 4: Solution Implementation

## Simple integration

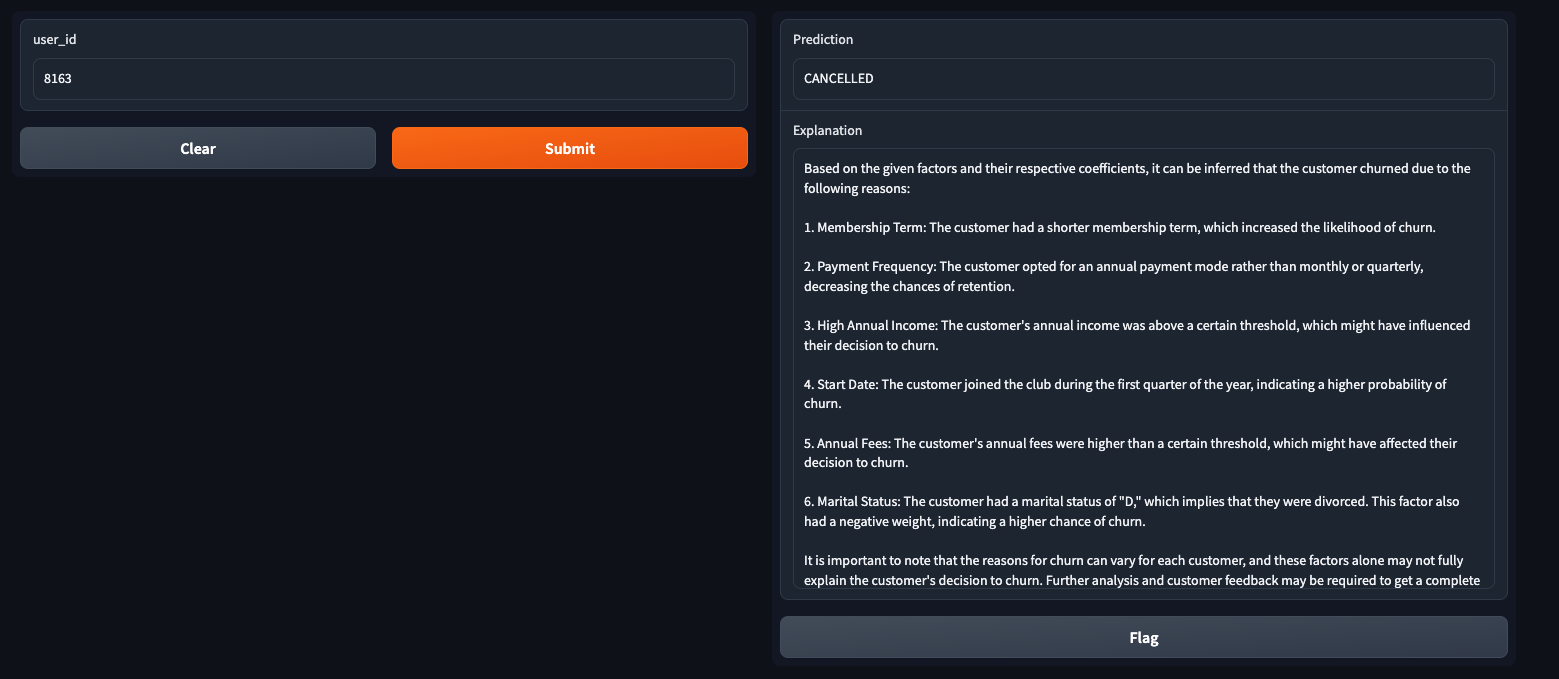
The first and most common approach is to use the predictive model to output the churn score or churn probability. From this score, the business team will rank and act such as nurturing call or offer promotion, etc. The size of action pool or cut-off threshold will be depended on business’s capacity.

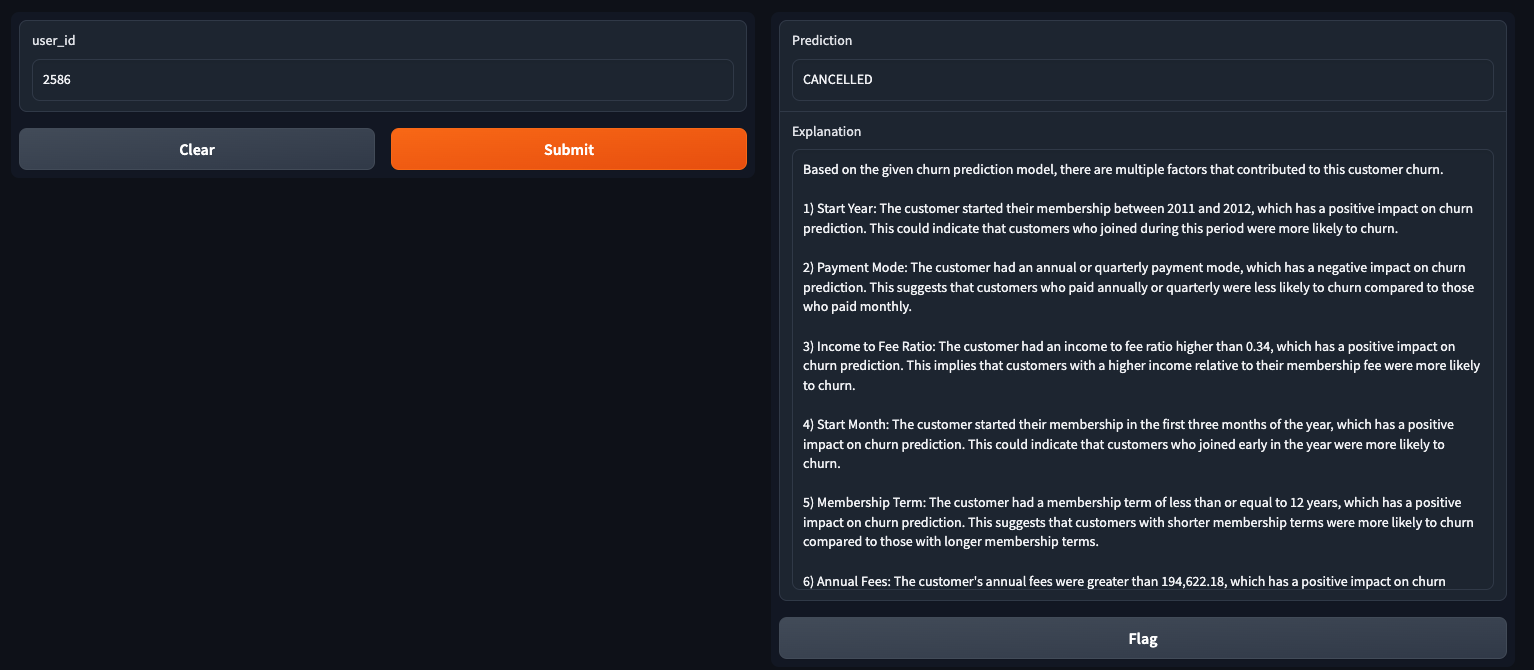
## Deploy solution with explainer

We can go further by applying explainers to help the customer service team, for example, know why one member is churned, so that they can have suitable scripts or even customized offer for the users.

Below is the demonstration we have built, after having cancellation model, we know who will be likely to churn and from this we apply LIME to know the reason. And then we translate the LIME result into readable format in natural language, in practice we can tailor the message to be suitable. In the sake of demonstration, we used Open AI’s API to translate the LIME result to readable format to support business acting effectively

With this demo, we input the member id, our application will turn out prediction cancelled or not. And provide the detailed reasons for the cancellation decision (the demo video is included in the documentation folder).

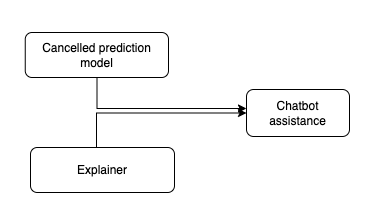




## Chatbot assistance

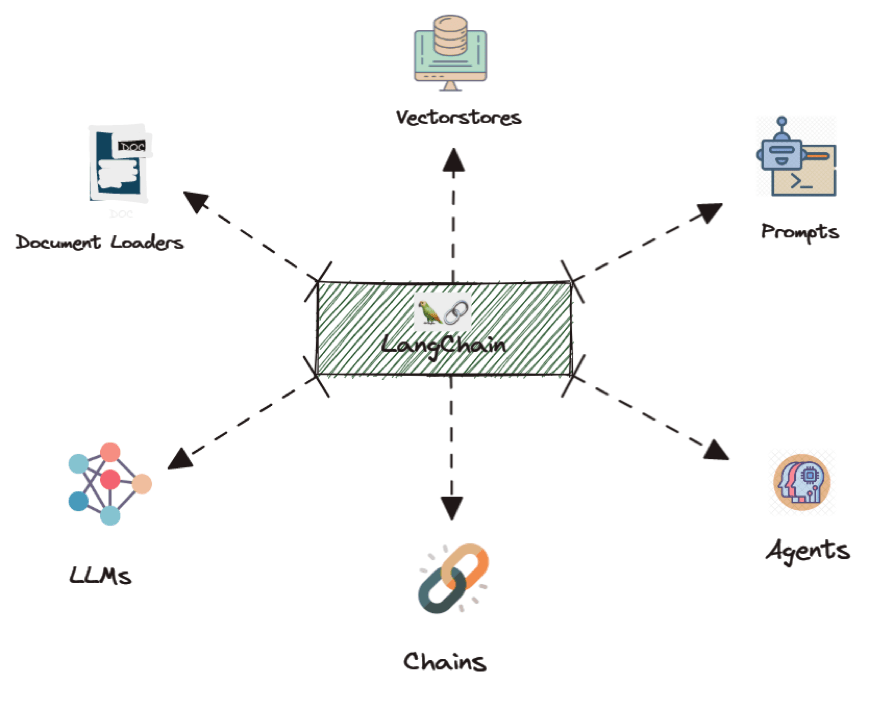
With the advancement of Generative AI, large language model. We would like to build a data product which uses tech advanced techniques with the hope of improving our business operation.

We can combine the output of cancelled prediction model, explainer as well as our local dataset. With the combination we can apply LLM to build effective chatbot for preventing churn effectively



We might be concerned about data privacy, but with the approach of finetuning model or open source LLMs, we can host it with our local data without any concerns about privacy.

Below demonstration is built based on the Langchain framework. The architecture of this framework is as below.



And I capture the result of the chat scenario, firstly I ask for the list of churn members, then following up by asking the reason why a member is churned and keep asking about the user information (the demo video is included in the documentation folder).







# Discussion and Further Improvement

While we've made commendable progress in this project, there remain several unresolved issues and queries that warrant discussion. These aspects can be potential avenues for enhancing the project further.

**Business Domain Expertise:** Despite our extensive analyses, our limited domain expertise in this field might have resulted in overlooking certain factors. Collaborating with experts in this domain could provide invaluable insights and significantly bolster our project.

**Feature Engineering:** Our approach to feature engineering has been multifaceted, leading to the creation of numerous composite features that rank highly in terms of importance. We've also experimented with deep learning for embedding categorical data. While this method is advantageous for datasets with vast categorical features, in our case, the categorical features are limited, making embedded vectors comparably effective to one-hot encoding. Moreover, while we initiated the use of autoencoders for imputing missing values, this aspect remains unfinished due to time constraints. Delving deeper into autoencoders could yield promising results.

**Model Construction & Refinement:** Our initial grid search for optimal parameters didn't yield significant performance enhancements. Exploring broader parameter ranges or employing alternative search techniques like Random or Bayesian searches might yield better outcomes. The potential of applying deep learning to tabular data is also worth exploring. Leveraging self-supervised learning through platforms like PyTorch Tabular, which supports denoising autoencoders, can be beneficial.

**Error Analysis:** Our meticulous error analysis scrutinized the model's performance across segments. However, there should be some actions to improve the model's performance based on the findings. Apart from that a comprehensive error analysis should delve into both global and local explanations, discerning the reasons behind specific predictions. Counterfactual and adversarial analyses using synthetic data, perhaps generated via GANs, could provide insights into the model's performance in unique or edge-case scenarios. This proactive approach can aid in better understanding our model and preemptively addressing potential pitfalls.

**Solution Deployment:** Our proposed solutions are ready for practical implementation. However, to determine their alignment with business requirements, further discussions are essential. For instance, employing explainable AI to aid sales or customer service teams requires an assessment of the accuracy of our model explanations. Additionally, the deployment of LLMs for chatbots necessitates exploration, especially in terms of production readiness. In essence, transitioning our solutions to a production environment entails significant testing and integration efforts.

**Consolidation & Presentation of Findings:** Investing more time in refining our visualizations and structuring this document would enhance its clarity and impact.

# Conclusion

In this project, we embarked on a comprehensive journey that spanned from understanding the core issues to conducting exploratory data analysis, predictive analytics, and finally, crafting potential solutions aligned with our goal of reducing churn.

A recap of our analytical journey reveals several pivotal insights:

* **Younger Members:** This demographic is more susceptible to churn. A tailored strategy might be required to better engage and retain these younger members.
* **Package Preferences:** The cancellation rate for 'Package type B' is alarmingly higher than 'Package type A'. It's imperative to delve deeper and pinpoint what aspects of 'Package type B' might not be resonating with our members.
* **Fee-to-Income Ratio:** Striking an optimal balance, where the fee constitutes around 10% of a member's income, seems to be a sweet spot for maximizing retention.
* **Subscription Duration:** Longer subscription terms correlate with higher retention. Encouraging members to opt for extended durations could allow them to fully realize the value of our services.
* **Membership Tenure:** Interestingly, newer members showcase a lower propensity to churn. This raises a question: Should our focus shift more towards our long-standing members?
* **Agent Performance:** The proficiency of agents plays a crucial role in influencing the cancellation rate. Enhancing the capabilities of this group could significantly bolster our retention metrics.

As for actionable solutions that can be integrated into business operations:

* We can adopt a conventional method, ranking members based on their predicted churn scores.
* Implementing our model, complete with its explanations on a platform, could empower the business team to take more informed actions.
* Alternatively, harnessing the capabilities of LLMs presents an innovative avenue for effectively mitigating churn.